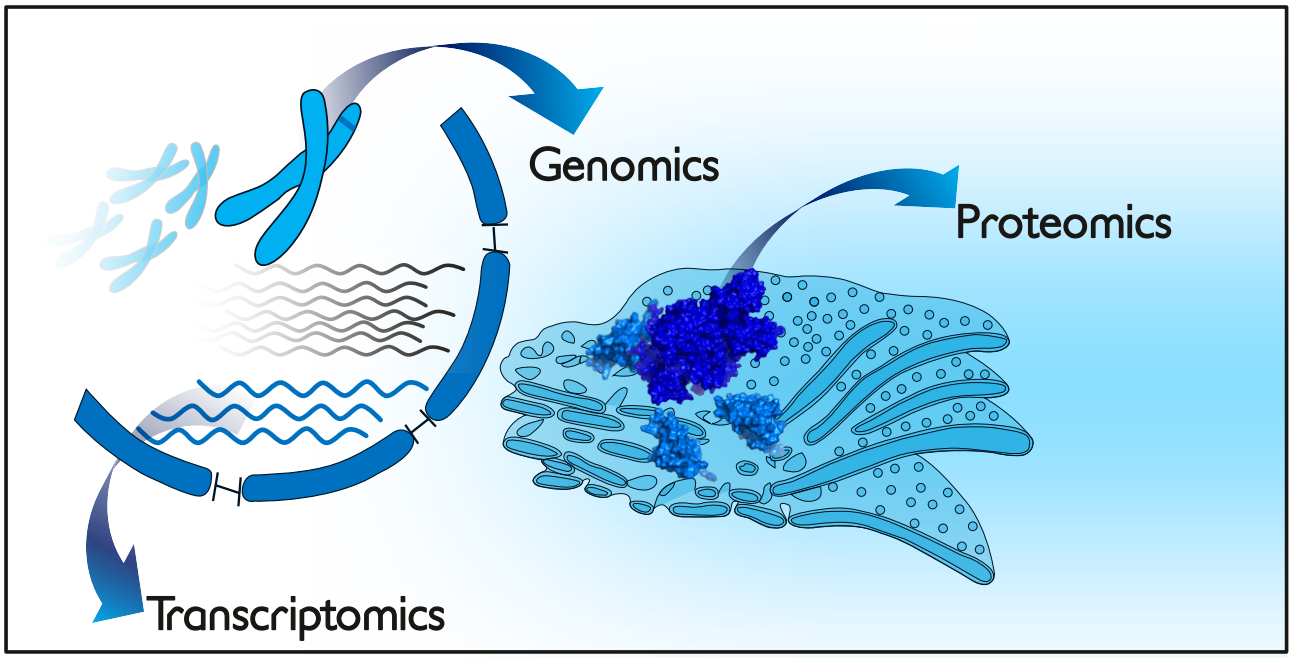
Characterization of Chinese Hamster Ovary Cells Producing Coagulation Factor VIII Using Multi-omics Tools

Kaas, Christian Schrøder

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record

Characterization of Chinese Hamster Ovary Cells Producing Coagulation Factor VIII Using Multi-omics Tools



Christian Schrøder Kaas
Ph.D. Thesis
June 2015

Dedicated to my dear wife Sara, my son Oscar and my son *in utero*

Characterization of Chinese Hamster Ovary Cells Producing Coagulation Factor VIII Using Multi-omics Tools

Ph.D. thesis

Christian Schrøder Kaas

20th of June 2015

Mammalian Cell Technology

Novo Nordisk A/S

Department of Systems Biology

Technical University of Denmark

Supervisors:

Mikael Rørdam Andersen

Technical University of Denmark

Claus Kristensen

Copenhagen University

Gert Bolt

Novo Nordisk A/S

Abstract

The first public draft of a genome from Chinese hamster ovary (CHO) cells was published in 2011, an entire decade after the first draft of the human genome. This publication of a relevant CHO reference genome, in combination with the fact that the cost for DNA sequencing has dropped more than 10,000 fold over the last couple of years due to the revolution of next-generation sequencing (NGS), has dramatically accelerated CHO-omics from virtually non-existent to a vibrant growing field.

The aim of this thesis was to investigate the impact of coagulation factor VIII (FVIII) production in CHO cells using omics tools. A wide range of methods were applied including whole-genome sequencing, targeted genome sequencing, mRNA sequencing, miRNA sequencing and mass spectrometry based shotgun proteomics on a number of clones in order to get a more holistic picture of the inner workings of these CHO transfectants.

From the whole-genome sequencing of two CHO genomes (CHO DXB11 and the FVIII producing transfectant: F435) it was observed that roughly 20% of the genes in the genome were haploid and roughly 10% had a copy number of three or higher indicating extensive rearrangements compared to the Chinese hamster origin. The transcriptome of 14 clones producing a dynamic range of FVIII was analyzed using RNA sequencing revealing an unexpected degree of 5' truncations of the transgene in 11 of the 14 clones. These truncations were validated using targeted genome sequencing, which also mapped the transgene insertion site in a number of clones. Furthermore, the RNA sequencing data was combined with proteomics data to investigate which impact FVIII biosynthesis exerts on the CHO cells. This revealed a dose-dependent induction of the unfolded-protein response, endoplasmic reticulum stress and oxidative stress which further lead to degradation of FVIII by the endoplasmic-reticulum-associated protein degradation pathway. This is to our knowledge, the first time that such extensive omics tools have been applied to a broad panel of CHO cells producing a very complex protein. The holistic view obtained for the FVIII producing cells provide a much clearer picture of the metabolic burden associated with FVIII secretion, than could be obtained using previous indirect methods.

The data and methods presented in this thesis suggest initial steps, which may be refined towards full utilization of omics technologies for analysis and engineering of industrially relevant CHO cells. Full implementation of such tools for generating specifically engineered CHO production cell lines may allow significant cost-reductions in production of complex biopharmaceuticals such as FVIII.

Dansk sammenfatning

Det første offentligt tilgængelige genom fra Chinese hamster ovary (CHO) celler blev frigivet i 2011, et helt årti efter frigivelsen af det humane genom. Tilgængelighed af dette relevante CHO reference genom, i kombination med et fald på mere end 10.000 gange i prisen for DNA sekventering over de sidste par år, har på dramatisk vis accelereret CHO-omics fra at være stort set ikke-eksisterende, til et felt i kraftig vækst.

Formålet med denne afhandling var at benytte omics værktøjer til at undersøge den påvirkning, som produktion af koagulationsfaktor VIII (FVIII) har på CHO-celler. Et bredt udvalg af værktøjer blev brugt der iblandt fuld-genom sekventering, målrettet genomsekventering af transgenområder, mRNA-sekventering, miRNA-sekventering og massespektrometri-baseret shotgun proteomics til analyse af en række kloner, for at danne et mere holistisk billede af disse CHO transfektanters karakteristika.

Fra fuld-genom sekventeringen af to CHO genomer (CHO DXB11 og en FVIII producerende transfektant: F435) blev det observeret, at omkring 20% af generne i genomet var haploide og ca. 10% havde et kopinummerantal på tre eller højere, hvilket indikerer omfattende ændringer i genomet i forhold til det oprindelige genom fra en kinesisk hamster. Transkriptomet af 14 kloner, der producerer FVIII over et stort spænd, blev analyseret med RNA-sekventering og dette viste en uventet grad af 5' trunkeringer af transgenet i 11 ud af 14 kloner. Disse trunkeringer blev valideret ved brug af målrettet genomsekventering, som også blev brugt til at kortlægge transgenintegrationsområdet i en række af klonerne. RNA-sekventeringsdataen blev desuden kombineret med proteomdata med det formål, at undersøge den effekt FVIII biosyntese har på CHO cellerne. Denne analyse viste en dosisafhængig induktion af unfolded-protein response, endoplasmatisk reticulum stress og oxidativt stress i cellerne som resulterede i endoplasmic-reticulum-associated protein degradation. Så vidt vides, er dette første gang at så omfattende værktøjer er benyttet til analyse af et bredt panel af CHO celler, der producerer et komplekst protein. Det holistiske billede, der males af de FVIII producerende celler ud fra disse analyser, giver et klarere billede af den metaboliske byrde, der er associeret med FVIII sekretion, end hvad der kan opnås med mere indirekte metoder.

Data og metoder, der præsenteres i denne tese, udgør de første skridt mod fuld anvendelse af omics værktøjer til analyse og udvikling af industrielt betydningsfulde CHO celler. Fuldstændig implementering af sådanne værktøjer vil kunne bruges til at udarbejde specialdesignede CHO produktionscellelinjer, der måske vil tillade signifikante reduktioner i produktionsprisen for komplekse lægemidler som fx FVIII.

Preface

This PhD dissertation was conducted as a collaborate effort of the department of Mammalian Cell Technology, Global Research Unit, Novo Nordisk A/S (NN) and the Network Engineering of Eukaryotic Cell Factories group at the Technical University of Denmark (DTU). The project was initiated May 1st 2012 and was solely funded by Novo Nordisk A/S. The practical part was carried out at NN while data analysis was carried out at NN and DTU. Six months of work was carried out as a visiting researcher at Johns Hopkins University from January - July 2014 under the supervision of Professor Michael J. Betenbaugh. The supervisor at NN was department director Claus Kristensen (from May 2012 - September 2014) and later Principal Scientist Gert Bolt (September 2014 – June 2015) and supervisor at DTU was Associate Professor Mikael R. Andersen

The thesis (excluding the publications, which are under copyright) can be downloaded as a pdf at www.bit.ly/CSRKthesis

Christian Schrøder Kaas

June 2015

Acknowledgements

There is a long list of people that I need to thank for invaluable help over the last three years. I tried to remember as many as I could but if I left somebody out please ascribe it to PhD deadline stress and nothing personal.

I would like to thank Claus Kristensen for the numerous hours you spent listening to my nerdy findings, overambitious ideas and for believing in them by freeing the capital needed to investigate them. Thanks for sending me to conferences/meetings/courses at Vienna, Lille, Vevey, Baltimore, Barcelona, Cambridge and Hillerød in order to discuss my ideas with some of the leading scientists in the area. I have nothing but gratitude for your style as supervisor always encouraging me, but still never afraid of pointing out when an experiment needed to be discarded. Your feedback on setting precise criteria for a given experiment and evaluating whether or not to scrap the experiment midway or continuing, is something I will definitely use onwards in my scientific carrier. Thank you for opening doors for me at Novo by presenting me to a multitude of experts and vouching for me in order to use equipment in other departments. Especially, I am thankful for the fact that you used your spare time continuing to supervise me after you left Novo Nordisk.

Mikael Rørdam Andersen: Thank for always having a fresh input when I present my data. You are extremely creative when it comes to data interpretation and comes up with suggestions for further experiments, whether it is adding an extra step in my protocol for RNA extraction or improving a script with a perl one-liner. It has truly been a pleasure working under your supervision.

Principal Scientist Gert Bolt: The most extensive Factor VIII database found in human form. So far I have yet to find any significant information on FVIII that he had not read already. Thanks for answering numerous questions and your invaluable help in drafting together the RNA truncation paper found in this thesis. I am very grateful that you were willing to step in for Claus as supervisor for the conclusion of my PhD thesis.

To all of department 279 at Novo Nordisk but particularly Else Jost Jensen, Gedske Thygesen and Jens Jacob Hansen: I am very grateful for the hours you all spent teaching (and re-teaching) me tips and tricks for working with suspension culture CHO cells. The time I spent in the lab would at least need to be increased 5-fold if you had not been there to answer my questions and tell me about the shortcuts in cloning etc. that you used routinely in the lab. Also special thanks to post doc Shamim Rahman for collaboration on several exiting genome editing projects still under development.

I would like to thank Caroline Bøtter-Jensen (CBJE) for invaluable help with purifying FVIII from my cell culture media. She went beyond what was expected of her and explained in great

detail the intricate nature of protein purification. I apologize for the data not to be present in the thesis for IP reasons.

Anahita Zamani Mohammadian: the first student to entrust me to supervise. Even though she ended up having to spend too much time tackling technical qPCR issues and selection pressure in the Icosagen system it was a lot of fun working with a person, who is so much more detail oriented than me. Due to the problems with the Icosagen selection pressure against FVIII plasmids most of her data is not in thesis but a bit can be found in chapter 3.

To all of the Network Engineering of Eukaryotic Cell Factories group at DTU but particularly Anne Mathilde Lund, Daniel Ley and Yuzhou Fan: Some of the selected few other PhD students out there working in the same area. It has always been great to discuss CHO NGS data with people, who know the frustrations of e.g. working with tools made specifically for human genome on the poorly assembled CHO genome. Even though I spent most of my time in Måløv and you in Lyngby it has been great to have companions at courses and conferences understanding the life and pressure of being a PhD student.

Professor Michael J Betenbaugh: Thanks for opening up your laboratory for me at Johns Hopkins University. It was an interesting six months of dissecting Chinese hamsters, analyzing the proteome of my FVIII cells and finding copy number variations in the sequenced CHO genomes. I would like to thank Joseph Priola, Lena Leberbauer, Amit Kumar, Kelley Heffner, Jimmy Kirsch and Deniz Baycin Hizal for guiding me and allowing me to experience the life as a researcher in the US.

I would like to thank my wife through seven years and friend through thirteen years: Sara Schrøder Kaas for absolutely invaluable help and support over the last three years. You have been there for better and for worse (e.g. took a sabbatical to be a stay-at-home-mom while we were in the USA for six months) and even though you are absolutely puzzled by the fact that I can spend weeks looking at the same data matrix you have always supported me. Lastly, I would like to thank my kid Oscar for always keeping up enthusiasm and taking on life with a smile.

Abbreviations

AEC adenylate energy charge

BDD-FVIII B-domain deleted Coagulation Factor VIII

CDS coding DNA sequence

CGCDB CHO gene co-expression database

CHO Chinese Hamster Ovary

CN copy number

CNV copy number variation

DE differentially expressed

EMS Ethyl methanesulfonate

ER endoplasmic reticulum

ERAD Endoplasmic-reticulum-associated protein degradation

FASP Filter aided sample prep

Fc fold-change

FPKM Fragments per kilobase of exon per million fragments mapped

FVIIa activated Factor VII

FVIII Coagulation Factor VIII

FVIIIa Activated Coagulation Factor VIII

GEM genome-scale metabolic model

GO Gene Ontology

GOI gene of interest

HCP host cell proteins

IRES an internal ribosomal entry site

iTRAQ isobaric tags for relative and absolute quantification

IPA Ingenuity Pathway Analysis

IVC integral of viable cells

LT Mouse Polyomavirus large T

mAbs monoclonal antibodies

MALDI TOF matrix-assisted laser desorption/ionization time of flight

MS mass spectrometry

MTX methotrexate

NGS Next-Generation Sequencing

OPLS ortho-phospho-l-serine

PacBio Pacific Biosciences

PDI Protein disulfide isomerase

PEG polyethylene glycol

q_P specific productivity

qPCR quantitative Polymerase Chain Reaction

qRT-PCR quantitative Reverse Transcription Polymerase Chain Reaction

RNAseq RNA sequencing

ROS reactive oxygen species

SDS Sodium dodecyl sulfate

SILAC stable isotope labeling with amino acids in culture

SNPs single nucleotide polymorphisms

TLA Targeted Locus Amplification

TMT tandem mass tags

UPR unfolded protein response

UTR untranslated regions

VCP Valosincontaining protein

vWF von Willebrand factor

Table of content

Structure of the thesis

In order to provide a framework for the reader, the following structure of the thesis was chosen: **Chapter 1** provides the background knowledge, which will be necessary in order to grasp the concepts utilized in this thesis. In **Chapter 2** the data from whole-genome sequencing of two CHO genomes are presented and compared to the genomes, which are currently publicly available, specifically in relation to changes in gene copy numbers. In **Chapter 3** the focus turns more specifically on the *F8* transgene in 14 CHO transfectants and how RNA sequencing data was mined in order to discover significant truncations on the transgene. **Chapter 4** focuses on the knowledge gained from proteomics and RNA sequencing in characterizing CHO cell lines producing FVIII. Furthermore, attempts for finding targets for improving the FVIII productivity on signatures from RNA sequencing were investigated. In **Chapter 5** an overview of the current state of CHO proteomics are given. The topic of **Chapter 6** is the utilization of omics data from CHO for generation of genome-scale metabolic models. **Chapter 7** contains an introduction into analyses of CHO NGS data also available at http://wiki.bio.dtu.dk/CHOomics and finally **Chapter 8** lists the conclusions and future perspectives of the work presented in this thesis.

Chapter 1 - Introduction and Background

1.1 Hemophilia A

Approximately 350,000 males world-wide suffer from hemophilia A, which is caused by genetic abnormalities of the *F8* gene encoding coagulation factor VIII (FVIII). The gene is located on chromosome X and is comprised of 26 exons spanning more than 180kb. The most commonly found genetic defects are an inversion of intron 22 [1] and an inversion of intron 1 [2] constituting 30-50% and 2-5% respectively of the cases of severe hemophilia A. The remaining polymorphisms found are a diverse spread of different nonsense, missense, splice-site mutations and deletions/insertions [3].

Without proper treatment of severe hemophilia A a person will have a life expectancy of 11 years. A doctor [4] describing patients with hemophilia A in the late 1960ies wrote: *"The typical hemophiliacs whom I saw ... were children who looked severely crippled. They had terrible joint contractures of their elbows, knees, and ankles from repeated hemorrhages and all the inflammatory reaction of the joints to get the blood out of there. They were terribly scarred. The cartilage was gone in most cases, and they had bone rubbing against bone.". "Virtually all of those who made it into their late teens and adulthood had so much chronic joint pain that they were addicted, for good reason, to Demerol and other narcotic pain medicines."* It was not until the 1970s that successful treatment with FVIII purified from blood donors were used to treat patients, but due to the low levels of FVIII in the blood, samples from up to 30,000 donors were pooled to make one vial of FVIII used for treatment [4]. The treatment was a success and the patients were able to live near-normal lives, but between 1981 and 1985 more than half of the members of the hemophilia community in the United States (approximately 20,000 people) became infected by human immunodeficiency virus due to lack of screening of blood donors [5]. Because of this disaster there was a huge push to develop a recombinant FVIII treatment in order to avoid donor blood in hemophilia A treatment.

The *F8* cDNA sequence was identified and the gene was heterologously expressed in 1984 [6,7]. Since then, the global FVIII marked has reached a value of 6.2 billion USD in 2013 of which 76% was from recombinant FVIII. Recombinant FVIII is sold under different names by Baxter, Wyeth, Aventis Behring and since 2013: Novo Nordisk. Out of the 350,000 patients estimated world-wide approximately 70,000 receive treatment. The main barrier for treatment is the cost which is approximately 100,000 USD per year per patient [5].

1.2 Expression of the FVIII protein

Following the identification of the *F8* cDNA sequence the biosynthesis of FVIII has been studied extensively with the aim of optimizing recombinant FVIII production and gaining a better understanding of the underlying cause of hemophilia A. FVIII is produced by a variety of cell types *in vivo*, but the primary source for FVIII production appear to occur in the liver by hepatocytes and sinusoidal endothelial cells [8,9]. There is currently no established cell line that naturally expresses FVIII and thus the knowledge gained concerning the expression and secretion of the protein is found solely from analysis of various mammalian cell lines transfected with *F8* cDNA [10]. Heterologous expression of human FVIII are though found to be at levels three order of magnitude lower than similarly sized secreted glycoproteins [11,12].

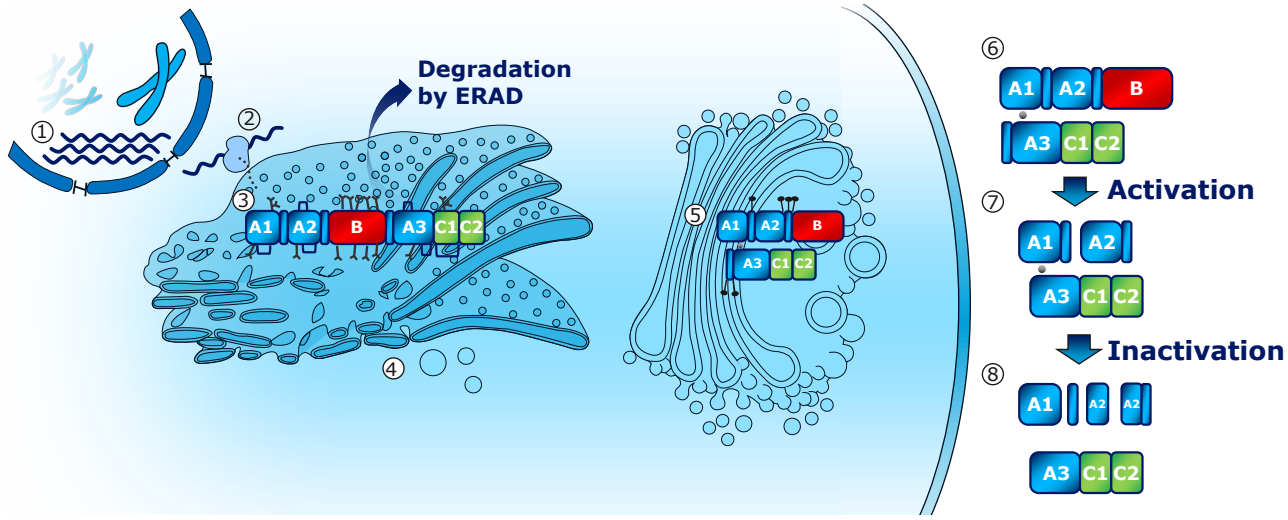


Figure 1 Synthesis and post-translational modification of FVIII
F8 mRNA is transcribed **(1)** and translated into the endoplasmic reticulum **(2)** where N-glycans are added and disulphide bridges are formed **(3)**. The protein is shipped **(4)** from the endoplasmic reticulum to the Golgi apparatus **(5)** where it is cleaved and sulphanations are added before secretion from the cell **(6)**. Upon secretion the FVIII protein can be processed by thrombin to yield activated FVIII **(7)** which subsequently lead to inactivated FVIII **(8)** by several different routes. Cleavage patterns of FVIII from [13], and disulfide bridges from [14].

The knowledge gathered concerning the expression and secretion of FVIII can be summarized as follows (see Figure 1): After transcription of the *F8* gene, the translation is directed into the endoplasmic reticulum (ER) due to the presence of a 19 amino acid-long signal peptide encoded in the 5' end of the mRNA. Following removal of this signal peptide, a single-chain protein of 2332 amino acids is synthesized with the domain structure A1-a1-A2-a2-B-a3-A3-C1-C2. The A domains share 40% amino acid identity with each other and likewise 40% amino acid identity is seen among the C domains. The B-domain consists of 908 amino acids and is without amino acid homology to any other known protein [13]. N-glycans are added to the FVIII peptide chain in order to stabilize the folded domains and to allow ER chaperones, such as the luminal calreticulin and the

membrane protein calnexin to interact with it. More than 75% of the N-glycans attachment sites of FVIII are located to the B-domain and attachment of N-glycans are found to be critical for FVIII secretion, as inhibition of N-glycosylation using tunicamycin dramatically reduced secretion [15]. FVIII is found to cycle through interactions with calnexin/calreticulin and the chaperone immunoglobulin-binding protein (BiP) possibly aiding in catalyzing correct folding [12,16]. The calnexin/calreticulin cycle function as a quality control step for glycoproteins by retaining proteins in the ER until they are correctly folded [17]. Unfortunately, it appears that approximately a third of the translated FVIII accumulate as non-disulfide-linked aggregates in the ER due to ATP depletion [18]. This furthermore activates the unfolded protein response (UPR) in the cell and cause oxidative stress [19]. In the case of incorrect folding after several cycles of calnexin/calreticulin interaction, the protein is sent for degradation by proteasomes in the cytoplasm by Endoplasmic-reticulum-associated protein degradation (ERAD) [17]. These implications of FVIII productions are further elaborated in Chapter 4.

After the protein is correctly folded it is able to interact with the soluble protein: multiple coagulation factor deficiency protein 2 (MCDF2). The transmembrane protein partner: Mannose-binding lectin 1 (LMAN1) binds MCDF2 and selectively package FVIII into vesicles, which are subsequently transported to the Golgi apparatus. It has been found that mutations in LMAN1 or MCFD2 cause combined deficiency of Coagulation Factor V and FVIII associated with the plasma levels of 5%-30% of normal levels [20]. Thus, the selective transport of FVIII from ER to the Golgi relies on the function of these proteins.

Upon arrival to the Golgi apparatus the FVIII protein is cleaved into a heavy chain (200 kDa) and a light chain (80 kDa), which are held together by a divalent metal ion. The protein furthermore undergo modification of the N-glycans, is O-glycosylated and furthermore sulphated on specific tyrosine residues.

Following release into circulation, FVIII binds a protein called von Willebrand factor (vWF). This allows the proteins to form a tight non-covalent complex, which shield FVIII from premature activation and clearance. FVIII becomes activated upon cleavage by thrombin by proteolysis at three residues: Arg_{372}, Arg_{740} and Arg_{1689} thus releasing the full B-domain and a3 region (binding vWF) from the protein [21]. Activated FVIII (FVIIIa) is subsequently able to act as a cofactor in the blood clotting cascade by interacting with coagulation factor IXa on the surface of activated platelets.

The activity of FVIIIa is rapidly lost either due to dissociation of the A2 domain from the rest of the subunits or by degradation due to cleavage by proteases such as thrombin, activated Coagulation Factor IX, activated Coagulation Factor X or activated protein C [13].

1.3 Engineering FVIII for recombinant protein production

In order to produce FVIII recombinantly at a competitive price compared to FVIII from donor blood, extensive efforts were undertaken in order to optimize production of FVIII from the cDNA sequence. It was found that the *F8* coding region contained stretches, which had a deleterious effect on *F8* mRNA accumulation and caused transcriptional silencing [22,23]. As the B-domain was found not to be essential for FVIII procoagulant activity [24], yet constituted 40% of the transcript, it was attempted to replace the fragment with a smaller region, which still ensure cleavage. It was seen that by removing the B-domain, the level of *F8* mRNA could be increased 20-fold compared to full-length *F8* mRNA. Unfortunately, the yield of secreted B-domain deleted (BDD) FVIII only doubled compared to secretion of full-length FVIII [13,25]. This indicates that the B-domain somehow have an effect on efficient biosynthesis of the protein, as higher levels of mRNA did not drastically increase the level of secreted protein. The recombinant FVIII investigated in this thesis is a FVIII protein with a B-domain, which has been truncated from 908 to only 21 amino acid residues [26]. In other attempts to optimize the protein for improved yields, Swaroop *et al* created a missense mutation (F309S) in order to decrease affinity for BiP binding in the ER. This was found to lead to more efficient secretion, although the interaction of heavy and light chain was also weakened [27].

A major obstacle for using FVIII variants, as the one created by Swaroop [27], for actual FVIII therapy in humans, is the risk that the hemophilia A patient will develop inhibiting antibodies. The FVIII protein has been found to be very immunogenic in hemophilia A patients, with a tendency for patients with a large deletions in the *F8* gene to often develop inhibitors (40% chance) and less so in patients with only small missense mutations in the *F8* gene (5% chance) [3]. Once a patient has developed antibodies against FVIII, additional FVIII administered to the blood will be rendered inactive and removed within minutes. This currently leave only one viable solution to stop the bleeding in a patient who has developed inhibiting antibodies: circumventing the steps of the blood coagulation cascade involving FVIII by administering activated Factor VII (FVIIa) [28] or FEIBA [29]. Unfortunately, this treatment is very costly as 250,000 USD worth of FVIIa can routinely be used during a single surgical operation on a hemophilia A patient with inhibitors [30]. As the FVIII drugs on the marked have a proven track record in regard to inhibitor development, it would be detrimental for a new FVIII drug if the levels were only slightly higher. It would thus be of much

lower risk to optimize the cell line for better production of FVIII compared to changing the protein only to fail in a costly clinical trial several years down the road. This way several iterations can be run quickly in order to gain a better understanding of what constitute a good FVIII cell line. A prime example for this strategy is seen in the two-fold optimization of the cellular FVIII productivity [31] by co-expression of the chaperone: heat shock protein 70 (Hsp70). This caused a significant delayed induction of apoptosis and furthermore improved the percentage of correctly folded FVIII secreted into the media (the specific activity). It was observed that less FVIII was found sequestered on the surface of these cells compared to normal FVIII producing cells [32].

The problem of FVIII bound to the surface of production cells are most profound when cells are grown in a serum-free setting where up to 90% of the FVIII can be found bound to the cells [34], but have been found to be alleviated by co-expression of vWF [35]. A successful strategy for keeping FVIII in the media and off the cell membrane is the addition of ortho-phospho-l-serine (OPLS). This molecule, which binds FVIII, is able to keep FVIII from binding the cell surface. It was found that addition of OPLS, 24 hours prior to harvesting FVIII from a culture of CHO cells, lead to a 50% increase in yield [36]. The idea was further developed by creating a cell line co-expressing a protein called lactadherin, which bind the cell surface competing with FVIII. This lead to improved yields by increasing the amount of FVIII found in the media compared to on the cell surface [37].

Purified FVIII injected into a patient has a half-life of approximately 8-12 hours, thus requiring three injections per week in order to sustain a plasma level of FVIII of 1% compared to that of a normal person [38]. The next generation of recombinant FVIII therapy entering the marked the years to come, are specifically aiming at improving the half-life of FVIII in the patient, thus reducing the amount of FVIII needed per patient per week. Biogen Idec is aiming for improved half-life through covalently linking FVIII to a human IgG1 Fc domain [39] whereas the strategy of Bayer, Baxter and Novo Nordisk is anchoring a large polyethylene glycol (PEG) molecule [40-42]. Using this strategy the half-life of FVIII was increased from 8.3 to 18 hours in Cynomolgus monkeys [41], which could bring down the number of weekly dosages from three to two per patient.

Finally, ongoing attempts are being made to effectively cure hemophilia A using gene therapy. The efforts are hindered by 1) the relatively large size of the *BDD-F8* coding region of 4.4kb, compared to the packaging capacity of the popular adeno-associated viruses, 2) the inefficient biosynthesis of FVIII mentioned above and 3) the problem of gene therapy of viral capsid-mediated cytotoxicity [11]. The problem of capsid induced cytotoxicity is further intensified by the general issue of inhibitor development, even when administered directly as purified protein from

recombinant sources [3]. The hypothesis is that infected cells might present misfolded FVIII at cell surface and elicit an immune response [43,44]. Headway has though been achieved with gene therapy of Factor IX for hemophilia B patients. Six patients have for more than two years successfully been producing Factor IX at 1-6% of normal levels [45,46].

For the time being gene therapy might not be considered the *safe choice* even though it might present itself at one point as the *cheaper choice*. For this reason gaining a better understanding of what constitutes an efficient cell line for FVIII production will benefit recombinant FVIII therapy on the short timeline, but could also shed light on optimizing gene therapy on the longer time line as well. Due to the extensive post-translational modifications required for production of active FVIII, the protein needs to be produced in a complex mammalian cell line, of which Chinese Hamster Ovary (CHO) cells are by far the most popular cell factory. Gaining a better understanding of CHO cells producing FVIII will therefore be the most straightforward approach for effectively reducing the production cost for hemophilia A therapy.

1.4 The Chinese Hamster Ovary cells

The global market for biopharmaceuticals is currently 140 billion USD of which the majority of proteins requiring post-translational modifications are produced in CHO cells [47]. The Chinese hamster (*Cricetulus griseus*) was originally brought to the USA to be used instead of mice for typing pneumococci [48]. Theodore Puck found in 1957 that a cell line of near diploid karyotype could be generated with relative ease, from a trypsinized sample of ovarian tissue from Chinese hamsters [49]. The cell line immortalized by an unknown spontaneous mechanism and was able to grow for months without showing signs of senescence, which are normally seen in cell lines derived from humans [50]. They were widely distributed to many laboratories around the world and described as “hardy” and growing well even with very low fetal bovine serum concentrations in the medium [51]. The first CHO cell line adapted to growth in media completely without serum was obtained in 1977 [52], thus paving the way for growth in cheap defined media and easing downstream purification of secreted proteins. It was later found that CHO happened to produce proteins with a glycosylation pattern similar to that of humans [53] and is not infected by a wide range of viruses dangerous to humans [54]. This has led to CHO cells being the *de facto* standard for protein production in mammalian cells. CHO is currently used for production of more than 40 biopharmaceuticals including monoclonal antibodies, hormones, cytokines and blood-coagulation factors. The history of the different lineages of CHO are excellently summarized in a recent review

[51] and shown in Figure 2 below. The genome sequencing of CHO DXB11 and F435 is described in Chapter 2 of this thesis.

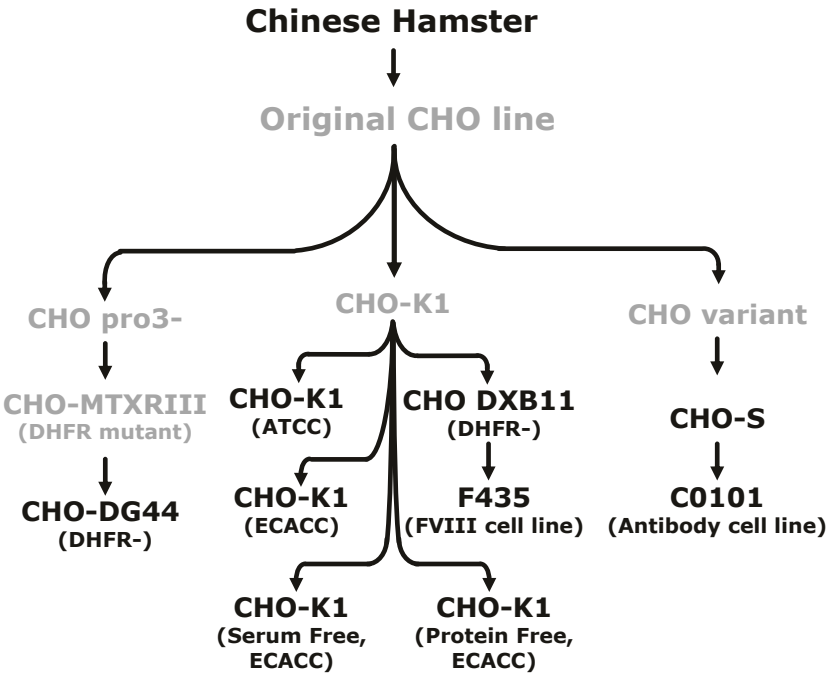


Figure 2 Overview of the lineages of the most commonly used CHO cell lines.
Cell lines which have been genome sequenced at the time of writing have been highlighted. ATCC: American Type Culture Collection. ECACC: European Collection of Cell Cultures. Figure redrawn from [55].

Another advantage of CHO cells is several well characterized methods for transfection such as calcium phosphate transfection, electroporation, lipofection or polymer-mediated gene transfer [56,57]. Upon transfection a selection system is advantageous in order to enrich the cellular population of desired transfectants. One of the most commonly used selection markers for cell line generation in CHO cells is the *dhfr*-selection system, which is used in chapter 3 and 4 of this thesis. Dihydrofolate reductase (DHFR) is an enzyme, which is able to catalyze the conversion of dihydrofolic acid to tetrahydrofolic acid – an essential cofactor carrier. Inhibition of DHFR, or knockout of the *dhfr* gene (as in the CHO DXB11 and CHO DG44 cell lines), thus leads to auxotrophy for glycine, hypoxanthine and thymidine. A favorable approach for heterologous expression of a gene of interest (GOI), would therefore be to couple expression of the GOI with the *dhfr* gene from one biscistronic mRNA by placing an internal ribosomal entry site (IRES) in between the two genes. Following transfection of $DHFR^{-/-}$ cells, only cells producing DHFR (and thus the protein of interest) will be able to survive the selection pressure from the media lacking glycine, hypoxanthine and thymidine [58]. As the transfected DNA, encoding *dhfr*, is not propagated in the cells, it will be diluted out over time (Figure 3A). In the case a copy of the transgene is integrated into the genome of the cell and expression of DHFR is high enough to

sustain viability for the cell, a stable cell line can be isolated (Figure 3B). By adding the folic acid analog methotrexate (MTX) to the growth media, the level of DHFR required to sustain viability for the cell go up as MTX will bind DHFR inhibiting its activity. As the gene of interest and *dhfr* are coupled that will lead to selection for clones with high expression levels of transgene. The MTX concentration can be increased for several rounds of selection leading to cells with many copies of the inserted transgene and *dhfr*, thus allowing the cells to produce sufficient levels of DHFR to survive [58] (Figure 3C). Transient transfection can thus be used for a short burst of production lasting a few days before the transgene is diluted out, whereas an amplified cell line with high productivity can sustain production for a much longer period of time, but it usually requires several months of selection pressure from transfection until the cell line can be isolated.

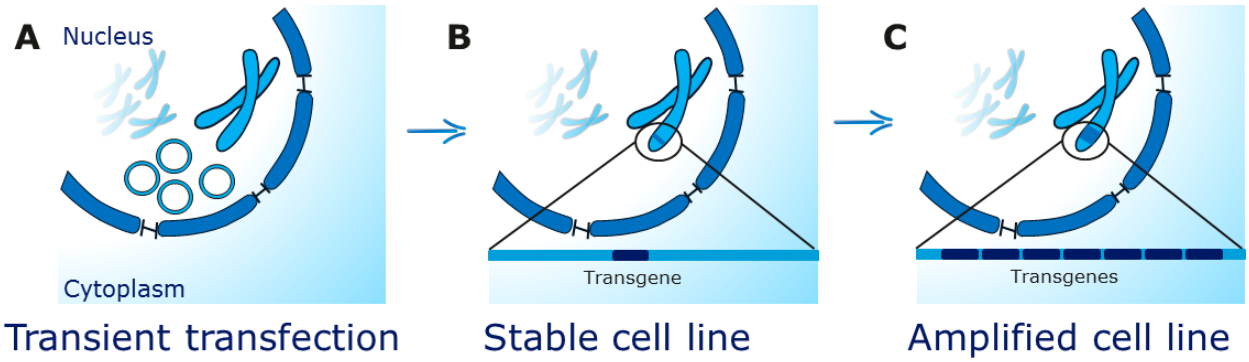


Figure 3 Transient transfection versus generation of a stable cell line.
(A) In a transient transfection the transgenic DNA is transcribed but not replicated within the cell, which leads to the transgene being diluted at each cell division. **(B)** In rare instances the transgene is inserted into the genome leading to a stable cell line. **(C)** By addition of MTX clones can be selected for, which has undergone amplification of the transgene.

1.5 CHO Systems biology by omics technologies

Over the course of the past three decades massive improvements have been achieved within media and process optimization of the CHO cells leading to typical antibody yields in 1986 of 0.05 g/L compared to 5-10 g/L today [57]. Although such results are impressive, they are most often product specific and are thus very labor intensive [59]. In contrast, only very limited success stories have been published within genetically engineering CHO through metabolic engineering towards higher protein productivities [60,61], compared to the results seen in microbial cell factories where entire metabolic pathways have been engineered [62-64]. A major reason for CHO lacking behind microbial cell factories in cell line engineering, is the fact that until recently the omics toolbox available to study the cell has been virtually empty. "Omics" as a term are typically defined as the bioinformatics study of all constituents of a biological system considered collectively, e.g., all genes

of an organism – genomics. A list of the most commonly investigated omics is given in Figure 4. As proteomics will be covered at lengths in chapter 5 of this thesis the two main omics which will be covered in detail will be CHO genomics and transcriptomics

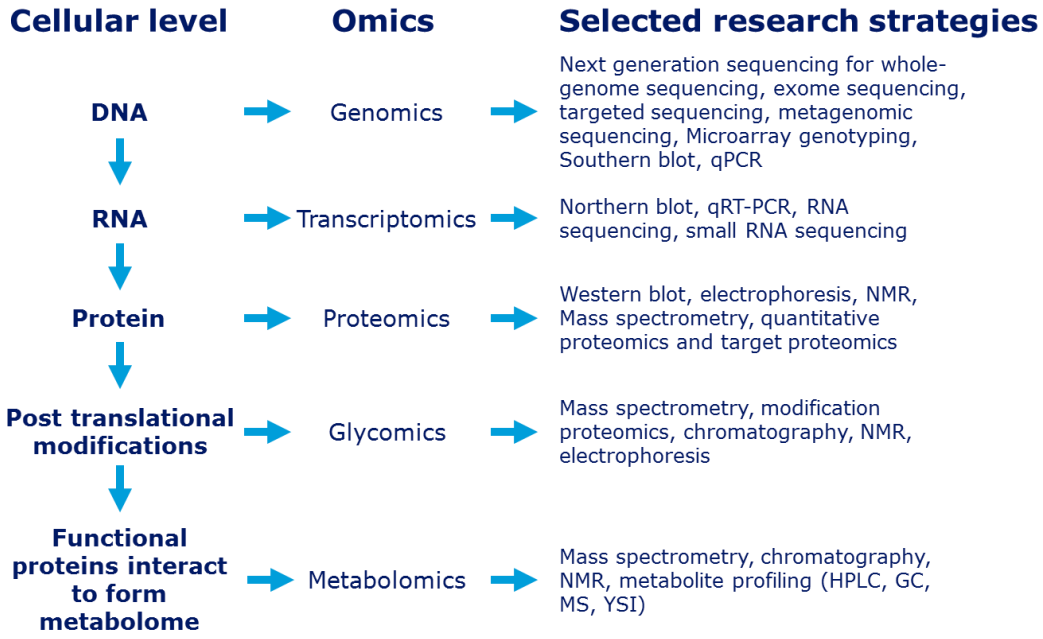


Figure 4 Different layers of cellular processes, the corresponding omics field and suggested research strategies for investigation of the given omics field. Figure modified from [59].

1.5.1 The current state of CHO genomics

Not a single CHO genome was publicly available until 2011, where two different groups released sequence from CHO-K1 [54] of ~100x coverage and a CHO DXB11 transfectant at ~1x coverage [65]. This was an entire decade after the publication of the first draft of the human genome [66] and 15 years after the genome of the first eukaryote, *Saccharomyces cerevisiae* [67]. Prior to the publication of the CHO-K1 genome most early genome-based studies of CHO cells were performed using the genome sequences from other mammals, for example, human, mouse or rat [68,69], which generally limited the possible experiments and interpretation of the results. From the genome sequence of CHO-K1, it was found that the cell line contained 24,383 genes associated with 21 chromosomes with a total of 2.45 Gb of genomic sequence. The genome sequence has been made accessible at the online database www.CHOgenome.org [70,71] as well as download from Genbank. One of the main advantages of expressing heterologous protein in CHO has been the possibility of human like post-translational modifications of recombinant proteins [53]. Analysis of the CHO-K1 genome revealed that homologs existed for 99% of the genes in the human genome

associated with glycosylation. Of these genes, merely 53% were detected to be transcribed. Furthermore, analysis of the transcriptome revealed that numerous genes associated with viral entry were not expressed, thus explaining the resistance of CHO cells to viral infection [54].

In 2013 two groups [55,72] published the genomic sequence of the Chinese hamster, from which the CHO cell line was originally extracted in 1957 [49]. The data from the two independent sequencing efforts are at the time of writing being merged with 3rd generation sequencing data from Pacific Biosciences [73] into a single well-characterized genome. The hope is that this genome will be able to be used for the future study of CHO cells lines as a standard reference. The current version of the Chinese hamster genome is composed of 53,000 separate fragments (scaffolds) and the hope is that by merging the sequencing data, that this number of fragments can be dramatically reduced in order to get a more holistic view on the localization of each gene in in genome.

The genomic sequence from a number of CHO cell lines including the industrially relevant CHO-S and CHO DG44 were also published in 2013 [55]. The genomic sequences for these cell lines are currently only available as raw sequence reads at the NCBI Short Read Archive. Analysis showed that more than 3.7 million point mutations were identified in these cell lines compared to the Chinese hamster [55] highlighting the fact that CHO genomes and the Chinese hamster are not identical and a need exist for further sequencing of genomes from additional popular CHO cell lines.

1.5.2 State of CHO transcriptomics

In order to analyze the portion of the genome that is actively transcribed under a specific set of conditions, methods such as quantitative Reverse Transcription PCR (qRT-PCR), microarrays and more recently RNA sequencing can be used. As stated above the possibilities for studying the transcriptome were limited prior to the publication of the genome although microarrays did exist, which covered a subset of the CHO transcripts. In an interesting study, Clarke et al [74] analyzed microarray data from 121 different conditions run by Pfizer on pre-genome microarrays. From the massive dataset it was possible to identify several gene clusters, which were associated with culture growth rate and productivity. The first large scale RNA sequencing paper covering CHO cells came out just two months after the CHO-K1 genome in 2011 [75]. By sequencing CHO samples from different genetic backgrounds and several cultivation conditions, they were able to *de novo* assemble the reads into 29,184 different transcripts. The transcripts identified in this paper are at the time of writing the source for CHO genes at Genbank and not the sequence from the CHO genomes. A transcriptome database for CHO RNA sequencing data has recently been developed and is available at https://gendbe.cebitec.uni-bielefeld.de/cho.html [76].

The CHO microRNA transcriptome has been thoroughly mapped [77] and the first papers analyzing the impact on apoptosis, growth and protein expression based on overexpression of different miRNAs have been published [78-80]. Notably it was found that 184 different miRNAs were able to induce up to two-fold improvements in protein expression of their transgene (secreted alkaline phosphatase) [80]. Due to the small size of miRNAs it is possible to synthesize the oligos and conduct large scale screens for a fraction of the price of mRNA overexpression. Due to these advantages and the results published so far, the strategy for cell line optimization through miRNA modulation hold great promise. A problem though still facing this area of research is the difficulty of predicting the miRNA target sites in the transcriptome. miRNA typically bind to the untranslated regions (UTR) of transcripts and these regions are currently very poorly annotated. Headway has though recently been achieved with a study conducting 5' enriched RNA sequencing for identification of transcriptional start sites, which should improve the knowledge of UTR positions in the CHO transcriptome [81].

1.6 Next-generation sequencing

The reason for the sudden increase in CHO sequencing data in the last couple of years can be explained by the revolution of Next-generation Sequencing (NGS) which has reduced the sequencing cost per megabase with more than 10,000 fold over the past 14 years (Figure 5).

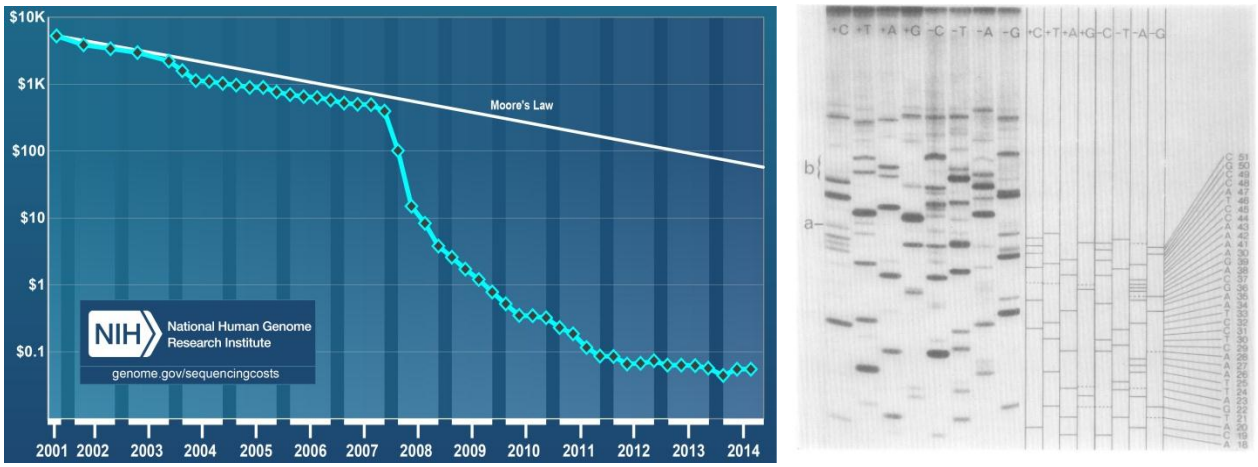


Figure 5 Sequencing before and now
Left: Cost of sequencing one megabase of DNA over the course of the last 14 years. Image from www.genome.gov/sequencingccosts. **Right:** One sequenced read using original Sanger sequencing. Figure from [82]

The rapid cost-reduction for sequencing has allowed project to be carried out today at a cost only a tiny fraction of what it would have been a few years earlier. First-generation sequencing was invented back in 1970s by Frederick Sanger [82,83] and relied on DNA polymerase extending a

sequencing primer annealing to a DNA fragment of interest. By splitting the reaction into several lanes and adding different nucleotide mixtures, the limiting nucleotide for a given lane could be used to deduce the base pair at that given position (Figure 5). The interpretation of the gel was fairly complex so it was common practice that one person would read the sequence from the combination of the lanes and another person would write down the sequence. This is why a sequenced DNA fragment is commonly referred to as a *read*. In 1986 Smith et al. [84] made a significant improvement by replacing the radioactive labels used in the Sanger method with fluorescent tags. This allowed the reaction to be run in a single vial sequencing up to 500–800bp allowing for high-throughput sequencing. The sequencing of the human genome was done using this approach, which was the predominant way for small and large scale DNA sequencing until a publication in Nature in 2005 [85], which described the whole genome sequencing of *Mycoplasma genitalium* in one four-hour run yielding a 40-fold coverage of the genome with 99% or better accuracy carried out by the 454 Life sciences Corporation. This technology was further made obsolete by the advent of Illumina Solexa solid state sequencing allowing for sequencing of 600 Gbp in a single run [86].

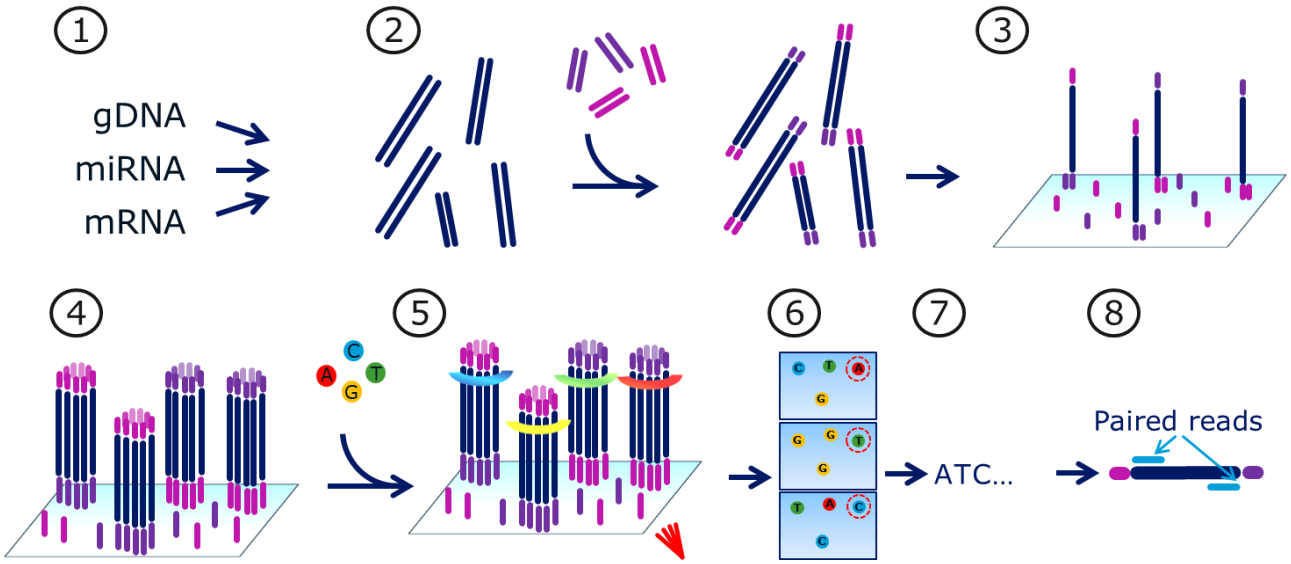


Figure 6 A schematic overview of Illumina sequencing
1) Samples consisting of RNA are reverse transcribed into cDNA and **2)** subsequently standardized adapters are added. **3)** The DNA fragments are seeded on a flow cell, **4)** amplified using primers found attached to the flow cell and **5)** sequenced in parallel one bp of the time by release of a flourophor. **6)** The color of each polony is registered for each cycle and **7)** the sequence for each fragment is deduced from the color. **8)** Each fragment is sequenced from each end resulting in two paired reads from the same fragment. Figure inspired by [87].

The method of Illumina sequencing can be summarized as follows (see Figure 6).

1. For genome sequencing, the DNA sample is digested into small DNA fragments of only a couple of hundred bps (in the case of sequencing of mRNA or miRNA samples, several steps precede purifying the desired sample and reverse transcribing it to cDNA).

2. Standardized adapters are added to the fragments.
3. Following addition of adapters the fragments are added to a flow cell containing oligos complimentary to the adapters, which have been added to the fragments.
4. The sample fragments are then amplified by PCR in order to get a stronger signal. In contrast to a regular PCR reaction the primers are in this case found on the flow cell in the vicinity of the fragment attached to the glass surface. By only being able to amplify using adapters found near the fragment, the PCR products are found in a localized area of the flow cell called a polony of approximately 1 million copies (like a bacterial colony that amplify on an agar plate from a single bacterium to an entire colony within a confined area).
5. Sequencing is carried out by adding all four nucleotides at the same time containing chain terminators so only one nucleotide can be added to each fragment in each cycle. Each nucleotide is labelled with a dye that can subsequently be read and following removal of the chain terminator the cycle is repeated.
6. For each cycle a high resolution picture is taken monitoring the color of each polony.
7. By using image analysis collecting the color of each 2D position on the flow cell the identity of the nucleotide is sequenced at each cycle can be deduced and thus allowing for approximately ~100bp of 1 billion template fragments to sequenced on the flow cell in parallel.
8. After the first round of sequencing the chemicals are washed off and then a new sequencing primer is added which align to the adapter in the other end of the fragment. Subsequently, ~100bp of the other end is sequenced yielding two read pairs originating from the same fragment with a read pair distance between them depending upon the length of the original fragment. The two reads from a single fragment is called *paired reads*.

A significant advantage of this technology compared to competing technologies, is found in the fact that all fragments are read in parallel without getting out of synchronization allowing large stretches of the same nucleotide to be read with large confidence. The vast majority of the NGS papers that are published these years have been done using the Illumina technology.

Third generation sequencing technologies such as the technique invented by Pacific Biosciences (PacBio) are currently gaining adoption by allowing sequencing without PCR amplification in the library preparation and producing reads that are up to 20kb in length contrast to the 100bp on the Illumina platform [86]. The reads can be assembled with much higher level of confidence if larger overlaps exists between the fragments and thus longer reads allow for better assembly of regions containing repeats. The price for sequencing using the PacBio technology is approximately 10x

higher per megabase than Illumina sequencing and for this reason it is common practice to get raw sequence with high depth coverage for each bp on the genome using the Illumina technology and then a low coverage of the expensive PacBio data in order to reliably assemble the genome getting the best of both worlds on a limited budget.

1.6.1 Analysis of NGS data

A typical workflow for analysis of raw data from a transcriptome and genome sequencing experiment are shown in Figure 7. In both cases, the reads are aligned to a reference genome and the number of reads aligning to a given gene is monitored.

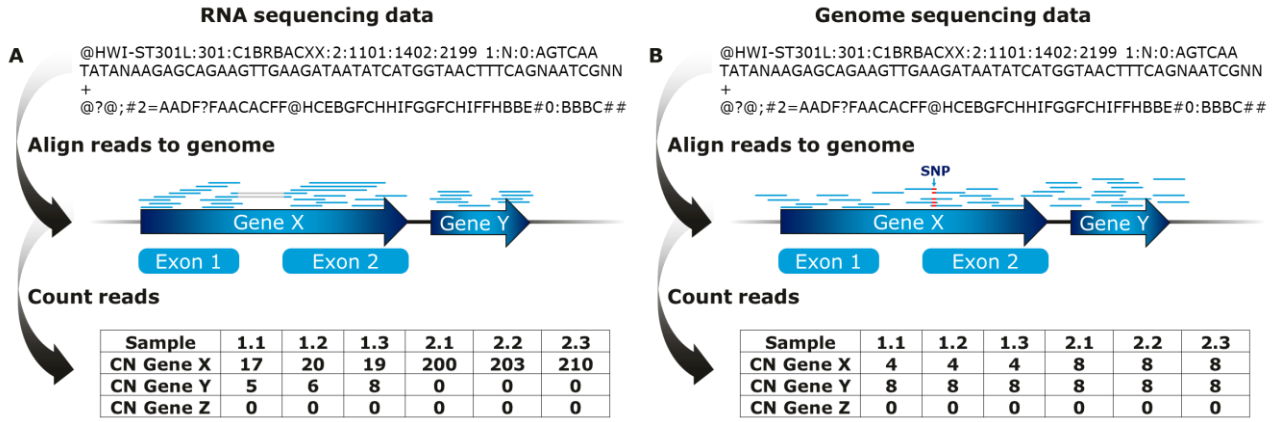


Figure 7 Representation of NGS workflows
A) An Ilumina read from a RNA sequencing experiment is shown in the fastq format: first a line containing a unique identified, then a line containing the sequenced fragment and finally a line containing the quality score for each sequenced base. Sequenced reads are aligned to genomic reference sequence and reads spanning an intron are seen aligning to exonic sequence on both sides of the intron. Following alignment the number of read pairs aligning to each gene in the genome is counted and used, after normalization, to identify genes, which are differentially expressed. **B)** Illumina reads from a genome sequencing experiment. SNPS are detected as discrepancies between the reference genome and several reads. From the sequencing depth it can be seen which genes have reduced or amplified copy number in the genome.

In the case of RNA sequencing special algorithms, such as Tophat [88], are used allowing for efficient alignment of reads that span over introns, which are not observed in genome sequencing data. After alignment, the number of reads that have aligned to each gene in the genome can be counted. In order to compare expression levels between samples, the number of reads counted for a given gene is normalized [89] between samples and subsequently used for deduction of which genes are differentially expressed. For alignment of reads from a genome sequencing experiment (Figure 7B), the alignment is more straightforward as gaps are not expected. After alignment of the reads single nucleotide polymorphisms (SNPs) can be found as discrepancies between multiple reads and the reference genome at a given position. Depending on how many reads agree with the reference

genome and the alternate sequence it can be deduced whether the mutation is homozygous or heterozygous. With a similar method as for detection of differentially expressed genes, copy number changes can de spotted based on differences in the read depth from gene to gene. In chapter 7 detailed descriptions are given on the bioinformatics pipelines used in this thesis.

1.7 References

1. Lakich D, Kazazian HH, Antonarakis SE, Gitschier J: **Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A.** *Nature genetics* 1993, **5:** 236-241.

2. Bagnall RD, Waseem N, Green PM, Giannelli F: **Recurrent inversion breaking intron 1 of the factor VIII gene is a frequent cause of severe hemophilia A.** *Blood* 2002, **99:** 168-174.

3. Ghosh K, Shetty S: **Immune response to FVIII in hemophilia A: an overview of risk factors.** *Clinical reviews in allergy & immunology* 2009, **37:** 58-66.

4. Levine PH: **HIV infection in hemophilia.** *Journal of clinical apheresis* 1993, **8:** 120-125.

5. Keshavjee S, Weiser S, Kleinman A: **Medicine betrayed: hemophilia patients and HIV in the US.** *Social Science & Medicine* 2001, **53:** 1081-1094.

6. Wood WI, Capon DJ, Simonsen CC, Eaton DL, Gitschier J, Keyt B *et al.*: **Expression of active human factor VIII from recombinant DNA clones.** *Nature* 1984, **312:** 330-337.

7. Toole JJ, Knopf JL, Wozney JM, Sultzman LA, Buecker JL, Pittman DD *et al.*: **Molecular cloning of a cDNA encoding human antihaemophilic factor.** *Nature* 1984, **312:** 342-347.

8. Hollestelle MJ, Thinnes T, Crain K, Stiko a, Kruijt JK, van Berkel TJ *et al.*: **Tissue distribution of factor VIII gene expression in vivo--a closer look.** *Thrombosis and haemostasis* 2001, **86:** 855-861.

9. Wion KL, Kelly D, Summerfield J, Tuddenham EGD, Lawn RM: **Distribution of factor VIII mRNA and antigen in human liver and other tissues.** *Nature* 1985, **317:** 726-729.

10. Plantier JL, Guillet B, Ducasse C, Enjolras N, Rodriguez M-H, Rolli V *et al.*: **B-domain deleted factor VIII is aggregated and degraded through proteasomal and lysosomal pathways.** *Thrombosis and haemostasis* 2005, **93:** 824-832.

11. Brown HC, Wright JF, Zhou S, Lytle AM, Shields JE, Spencer HT *et al.*: **Bioengineered coagulation factor VIII enables long-term correction of murine hemophilia A following liver-directed adeno-associated viral vector delivery.** *Molecular Therapy - Methods & Clinical Development* 2014, **1:** 14036.

12. Kaufman RJ, Pipe SW, Tagliavacca L, Swaroop M, Moussalli M: **Biosynthesis, assembly and secretion of coagulation factor VIII.** *Blood coagulation & fibrinolysis : an international journal in haemostasis and thrombosis* 1997, **8 Suppl 2:** S3-14.

13. Pipe SW: **Functional roles of the factor VIII B domain.** *Haemophilia : the official journal of the World Federation of Hemophilia* 2009, **15:** 1187-1196.

14. Lenting PJ, van Mourik JA, Mertens K: **The life cycle of coagulation factor VIII in view of its structure and function.** *Blood* 1998, **92:** 3983-3996.

15. Dorner AJ: **The relationship of N-linked glycosylation and heavy chain-binding protein association with the secretion of glycoproteins.** *The Journal of Cell Biology* 1987, **105:** 2665-2674.

16. Marquette K, Pittman DD, Kaufman RJ: **A 110-amino acid region within the A1-domain of coagulation factor VIII inhibits secretion from mammalian cells.** *Journal of Biological Chemistry* 1995, **270:** 10297-10303.

17. Ellgaard L, Helenius A: **Quality control in the endoplasmic reticulum.** *Nature reviews Molecular cell biology* 2003, **4:** 181-191.

18. Tagliavacca L, Wang Q, Kaufman RJ: **ATP-Dependent Dissociation of Non-Disulfide-Linked Aggregates of Coagulation Factor VIII Is a Rate-Limiting Step for Secretion ΓÇá.** *Biochemistry* 2000, **39:** 1973-1981.

19. Malhotra JD, Miao H, Zhang K, Wolfson A, Pennathur S, Pipe SW *et al*.: **Antioxidants reduce endoplasmic reticulum stress and improve protein secretion.** *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105:** 18525-18530.

20. Zhang B, McGee B, Yamaoka JS, Guglielmone H, Downes K, Minoldo S *et al*.: **Combined deficiency of factor V and factor VIII is due to mutations in either LMAN1 or MCFD2.** *Blood* 2006, **107:** 1903-1907.

21. Pittman DD, Kaufman RJ: **Proteolytic requirements for thrombin activation of anti-hemophilic factor (factor VIII).** *Proceedings of the National Academy of Sciences of the United States of America* 1988, **85:** 2429-2433.

22. Lynch CM, Israel DI, Kaufman RJ, Miller AD: **Sequences in the coding region of clotting factor VIII act as dominant inhibitors of RNA accumulation and protein production.** *Human gene therapy* 1993, **4:** 259-272.

23. Hoeben RC, Fallaux FJ, Cramer SJ, van den Wollenberg DJ, van Ormondt H, Briet. *et al*.: **Expression of the blood-clotting factor-VIII cDNA is repressed by a transcriptional silencer located in its coding region.** *Blood* 1995, **85:** 2447-2454.

24. Toole JJ, Pittman DD, Orr EC, Murtha P, Wasley LC, Kaufman RJ: **A large region (approximately equal to 95 kDa) of human factor VIII is dispensable for in vitro procoagulant activity.** *Proceedings of the National Academy of Sciences* 1986, **83:** 5939-5942.

25. Pittman DD, Alderman EM, Tomkinson KN, Wang JH, Giles AR, Kaufman RJ: **Biochemical, immunological, and in vivo functional characterization of B-domain-deleted factor VIII.** *Blood* 1993, **81:** 2925-2935.

26. Thim L, Vandahl B, Karlsson J, Klausen NK, Pedersen J, Krogh TN *et al*.: **Purification and characterization of a new recombinant factor VIII (N8).** *Haemophilia : the official journal of the World Federation of Hemophilia* 2010, **16:** 349-359.

27. Swaroop M, Moussalli M, Pipe SW, Kaufman RJ: **Mutagenesis of a potential immunoglobulin-binding protein-binding site enhances secretion of coagulation factor VIII.** *The Journal of biological chemistry* 1997, **272:** 24121-24124.

28. Erhardtsen E: **Pharmacokinetics of Recombinant Activated Factor VII (rFVIIa).** *Seminars in Thrombosis and Hemostasis* 2000, **Volume 26:** 0385-0392.

29. Turecek PL, V+íradi K, Gritsch H, Schwarz HP: **FEIBA: mode of action.** *Haemophilia : the official journal of the World Federation of Hemophilia* 2004, **10 Suppl 2:** 3-9.

30. Ponder KP: **FIXing factor VIII inhibitors.** *Blood* 2012, **119:** 325-326.

31. Ishaque A, Thrift J, Murphy JE, Konstantinov K: **Over-expression of Hsp70 in BHK-21 cells engineered to produce recombinant factor VIII promotes resistance to apoptosis and enhances secretion.** *Biotechnology and bioengineering* 2007, **97:** 144-155.

32. Ishaque A, Thrift J, Murphy JE, Konstantinov K: **Cell surface staining of recombinant factor VIII is reduced in apoptosis resistant BHK-21 cells.** *Journal of biotechnology* 2008, **137:** 20-27.

33. Kavakli K, Yang R, Rusen L, Beckmann H, Tseneklidou-Stoeter D, Maas Enriquez M: **Prophylaxis vs. on-demand treatment with BAY 81-8973, a full-length plasma protein-free recombinant factor VIII product: results from a randomized trial (LEOPOLD II).** *Journal of thrombosis and haemostasis : JTH* 2015, **13:** 360-369.

34. Kolind MP, Nørby PL, Flintegaard TV, Berchtold MW, Johnsen LB: **The B-domain of Factor VIII reduces cell membrane attachment to host cells under serum free conditions.** *Journal of biotechnology* 2010, **147:** 198-204.

35. Kaufman RJ, Wasley LC, Davies MV, Wise RJ, Israel DI, Dorner AJ: **Effect of von Willebrand factor coexpression on the synthesis and secretion of factor VIII in Chinese hamster ovary cells.** *Molecular and cellular biology* 1989, **9:** 1233-1242.

36. Kolind MP, Nørby PL, Berchtold MW, Johnsen LB: **Optimisation of the Factor VIII yield in mammalian cell cultures by reducing the membrane bound fraction.** *Journal of biotechnology* 2011, **151:** 357-362.

37. Johnsen LB, KOLIND MP, Nørby PL. Method for production of factor VIII. 30-5-2013. Patent application: EP2520586 A1.

Ref Type: Patent

38. Manco-Johnson MJ, Abshire TC, Shapiro AD, Riske B, Hacker MR, Kilcoyne R *et al*.: **Prophylaxis versus Episodic Treatment to Prevent Joint Disease in Boys with Severe Hemophilia.** *N Engl J Med* 2007, **357:** 535-544.

39. Powell JS, Josephson NC, Quon D, Ragni MV, Cheng G, Li E *et al*.: **Safety and prolonged activity of recombinant factor VIII Fc fusion protein in hemophilia A patients.** *Blood* 2012, **119:** 3031-3037.

40. Saenko EL, Pipe SW: **Strategies towards a longer acting factor VIII.** *Haemophilia : the official journal of the World Federation of Hemophilia* 2006, **12 Suppl 3:** 42-51.

41. Stennicke HR, Kjalke M, Karpf DM, Balling KW, Johansen PB, Elm T *et al*.: **A novel B-domain O-glycoPEGylated FVIII (N8-GP) demonstrates full efficacy and prolonged effect in hemophilic mice models.** *Blood* 2013, **121:** 2108-2116.

42. Mei B, Pan C, Jiang H, Tjandra H, Strauss J, Chen Y *et al*.: **Rational design of a fully active, long-acting PEGylated factor VIII for hemophilia A treatment.** *Blood* 2010, **116:** 270-279.

43. Selvaraj SR, Pipe SW: **Gene therapy : molecular engineering of factor VIII and factor IX.** 2014:298-307.

44. High KA: **Update on progress and hurdles in novel genetic therapies for hemophilia.** *Hematology / the Education Program of the American Society of Hematology American Society of Hematology Education Program* 2007, **2007:** 466-472.

45. Nathwani AC, Tuddenham EGD, Rangarajan S, Rosales C, McIntosh J, Linch DC *et al.*: **Adenovirus-Associated Virus Vector–Mediated Gene Transfer in Hemophilia B.** *N Engl J Med* 2011, **365:** 2357-2365.

46. High K: **The gene therapy journey for hemophilia : are we there yet ? The gene therapy journey for hemophilia : are we there yet ?** 2013, **120:** 4482-4487.

47. Walsh G: **Biopharmaceutical benchmarks 2014.** *Nature biotechnology* 2014, **32:** 992-1000.

48. Jayapal KP, Wlaschin KF, Hu W, Yap MG: **Recombinant protein therapeutics from CHO cells-20 years and counting.** *Chemical Engineering Progress* 2007, **103:** 40.

49. Puck TT: **Genetics Of Somatic Mammalian Cells: III. Long-term Cultivation Of Euploid Cells From Human And Animal Subjects.** *Journal of Experimental Medicine* 1958, **108:** 945-956.

50. Hayflick L, Moorhead PS: **The serial cultivation of human diploid cell strains.** *Experimental cell research* 1961, **25:** 585-621.

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

52. Hamilton WG, Ham RG: **Clonal growth of chinese hamster cell lines in protein-free media.** *In vitro* 1977, **13:** 537-547.

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

54. Xu X, Nagarajan H, Lewis NE, Pan S, Cai Z, Liu X *et al.*: **The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line.** *Nature biotechnology* 2011, **29:** 735-741.

55. Lewis NE, Liu X, Li Y, Nagarajan H, Yerganian G, O'Brien E *et al.*: **Genomic landscapes of Chinese hamster ovary cell lines as revealed by the Cricetulus griseus draft genome.** *Nature biotechnology* 2013, **31:** 759-765.

56. Norton PA, Pachuk CJ: *Gene Transfer and Expression in Mammalian Cells*, 38 edn. Elsevier; 2003.

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

58. Urlaub G, Chasin L: **Isolation of Chinese hamster cell mutants deficient in dihydrofolate reductase activity.** *Proceedings of the National Academy of Sciences of the United States of America* 1980, **77:** 4216-4220.

59. Datta P, Linhardt RJ, Sharfstein ST: **An 'omics approach towards CHO cell engineering.** *Biotechnology and bioengineering* 2013, **110:** 1255-1271.

60. Zhang F, Sun X, Yi X, Zhang Y: **Metabolic characteristics of recombinant Chinese hamster ovary cells expressing glutamine synthetase in presence and absence of glutamine.** *Cytotechnology* 2006, **51:** 21-28.

61. Kim SH, Lee GM: **Down-regulation of lactate dehydrogenase-A by siRNAs for reduced lactic acid formation of Chinese hamster ovary cells producing thrombopoietin.** *Applied Microbiology and Biotechnology* 2007, **74:** 152-159.

62. Runguphan W, Keasling JD: **Metabolic engineering of Saccharomyces cerevisiae for production of fatty acid-derived biofuels and chemicals.** *Metabolic Engineering* 2014, **21:** 103-113.

63. Wu H, Karanjikar M, San KY: **Metabolic engineering of Escherichia coli for efficient free fatty acid production from glycerol.** *Metabolic Engineering* 2014, **25:** 82-91.

64. Nielsen J, Keasling JD: **Synergies between synthetic biology and metabolic engineering.** *Nature biotechnology* 2011, **29:** 693-695.

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

66. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J *et al.*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:** 860-921.

67. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H *et al.*: **Life with 6000 genes.** *Science (New York, N Y)* 1996, **274:** 546, 563-546, 567.

68. Baik JY, Lee MS, An SR, Yoon SK, Joo EJ, Kim YH *et al.*: **Initial transcriptome and proteome analyses of low culture temperature-induced expression in CHO cells producing erythropoietin.** *Biotechnology and bioengineering* 2006, **93:** 361-371.

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

70. Hammond S, Kaplarevic M, Borth N, Betenbaugh MJ, Lee KH: **Chinese hamster genome database: an online resource for the CHO community at www.CHOgenome.org.** *Biotechnology and bioengineering* 2012, **109:** 1353-1356.

71. Kremkow BG, Baik JY, MacDonald ML, Lee KH: **CHOgenome.org 2.0: Genome resources and website updates.** *Biotechnology journal* 2015.

72. Brinkrolf K, Rupp O, Laux H, Kollin F, Ernst W, Linke B *et al.*: **Chinese hamster genome sequenced from sorted chromosomes.** *Nature biotechnology* 2013, **31:** 694-695.

73. Borth N: **Opening the black box: Chinese hamster ovary research goes genome scale.** *Pharmaceutical Bioprocessing* 2014, **2:** 367-369.

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

75. Becker J, Hackl M, Rupp O, Jakobi T, Schneider J, Szczepanowski R *et al.*: **Unraveling the Chinese hamster ovary cell line transcriptome by next-generation sequencing.** *Journal of biotechnology* 2011, **156:** 227-235.

76. Rupp O, Becker J, Brinkrolf K, Timmermann C, Borth N, Pühler A *et al.*: **Construction of a public CHO cell line transcript database using versatile bioinformatics analysis pipelines.** *PloS one* 2014, **9:** e85568.

77. Hackl M, Jakobi T, Blom J, Doppmeier D, Brinkrolf K, Szczepanowski R *et al.*: **Next-generation sequencing of the Chinese hamster ovary microRNA transcriptome: Identification, annotation and profiling of microRNAs as targets for cellular engineering.** *Journal of biotechnology* 2011, **153:** 62-75.

78. Jadhav V, Hackl M, Druz A, Shridhar S, Chung CY, Heffner KM *et al.*: **CHO microRNA engineering is growing up: recent successes and future challenges.** *Biotechnology advances* 2013, **31:** 1501-1513.

79. Fischer S, Buck T, Wagner A, Ehrhart C, Giancaterino J, Mang S *et al.*: **A functional high-content miRNA screen identifies miR-30 family to boost recombinant protein production in CHO cells.** *Biotechnology journal* 2014, **9:** 1279-1292.

80. Fischer S, Handrick R, Aschrafi A, Otte K: **Unveiling the principle of microRNA-mediated redundancy in cellular pathway regulation.** *RNA biology* 2015, **12:** 238-247.

81. Jakobi T, Brinkrolf K, Tauch A, Noll T, Stoye J, Pühler A *et al.*: **Discovery of transcription start sites in the Chinese hamster genome by next-generation RNA sequencing.** *Journal of biotechnology* 2014, **190:** 64-75.

82. Sanger F, Coulson AR: **A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase.** *Journal of molecular biology* 1975, **94:** 441-448.

83. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors.** *Proceedings of the National Academy of Sciences of the United States of America* 1977, **74:** 5463-5467.

84. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR *et al.*: **Fluorescence detection in automated DNA sequence analysis.** *Nature* 986, **321:** 674-679.

85. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA *et al.*: **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, **437:** 376-380.

86. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR *et al.*: **A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers.** *BMC genomics* 2012, **13:** 341.

87. Mardis ER: **Next-generation DNA sequencing methods.** *Annual review of genomics and human genetics* 2008, **9:** 387-402.

88. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq.** *Bioinformatics (Oxford, England)* 2009, **25:** 1105-1111.

89. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N *et al.*: **A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis.** *Briefings in bioinformatics* 2013, **14:** 671-683.

Chapter 2 – CHO genomics

In this chapter, the conclusions generated from whole-genome-sequencing (WGS) of two CHO genomes (CHO DXB11 and a FVIII producing cell line called F435) will be presented with specific focus on the detection of copy number changes in the genome. It was shown that the copy number directly correlated with the sequencing depth revealing an unexpected degree of haploidy. Furthermore, when analyzing the genes from each chromosome separately, it was found to reveal unique signatures for each chromosome. Following the published work from BMC genomics, a short update is provided concerning the importance of the 3rd generation sequencing efforts of the *C. griseus* genome currently underway.

Kaas et al. BMC Genomics (2015) 16:160
DOI 10.1186/s12864-015-1391-x

BMC Genomics

RESEARCH ARTICLE Open Access

Sequencing the CHO DXB11 genome reveals regional variations in genomic stability and haploidy

Christian Schrøder Kaas[1,2,3*], Claus Kristensen[4], Michael J Betenbaugh[3] and Mikael Rørdam Andersen[2]

Abstract

Background: The DHFR negative CHO DXB11 cell line (also known as DUX-B11 and DUKX) was historically the first CHO cell line to be used for large scale production of heterologous proteins and is still used for production of a number of complex proteins.

Results: Here we present the genomic sequence of the CHO DXB11 genome sequenced to a depth of 33x. Overall a significant genomic drift was seen favoring GC → AT point mutations in line with the chemical mutagenesis strategy used for generation of the cell line. The sequencing depth for each gene in the genome revealed distinct peaks at sequencing depths of 0x, 16x, 33x and 49x coverage corresponding to a copy number in the genome of 0, 1, 2 and 3 copies. This indicate that 17% of the genes are haploid revealing a large number of genes which can be knocked out with relative ease. This tendency of haploidy was furthermore shown to be present in eight additional analyzed CHO genomes (15-20% haploidy) but not in the genome of the Chinese hamster. The *dhfr* gene is confirmed to be haploid in CHO DXB11; transcriptionally active and the remaining allele contains a G410C point mutation causing a Thr137Arg missense mutation. We find ~2.5 million single nucleotide polymorphisms (SNP's), 44 gene deletions in the CHO DXB11 genome and 9357 SNP's, which interfere with the coding regions of 3458 genes. Copy number variations for nine CHO genomes were mapped to the chromosomes of the Chinese hamster showing unique signatures for each chromosome. The data indicate that chromosome one and four appear to be more stable over the course of the CHO evolution compared to the other chromosomes thus might presenting the most attractive landing platforms for knock-ins of heterologous genes.

Conclusions: Our studies reveal an unexpected degree of haploidy in CHO DXB11 and CHO cells in general and highlight the chromosomal changes that have occurred among the CHO cell lines sequenced to date.

Keywords: Copy number variations (CNVs), CHO DXB11, CHO cells, *C. griseus*, Single nucleotide polymorphisms (SNPs)

Background

The global market for biopharmaceuticals is currently 140 billion USD of which the majority of proteins requiring post-translational modifications are produced in Chinese Hamster Ovary (CHO) cells [1]. CHO cells have a long history as a production organism in industry due to their ability to grow in suspension without serum and to be scalable to large production volumes. Furthermore, CHO cells are able to produce proteins with a glycosylation pattern similar to that of humans [2] and are not infected by a wide range of viruses dangerous to humans [3]. So far more than 40 biopharmaceuticals including monoclonal antibodies, hormones, cytokines and blood-coagulation factors have been produced in CHO cells.

The CHO cell line was originally isolated in 1957 by T. Puck [4] and ten years later the CHO-K1 cell line was derived from this ancestral host [5]. In order to facilitate creation of stable cell lines producing a gene of interest a selection system was needed. The CHO DXB11 cell line was created with the goal of developing a stable CHO cell line with a DHFR negative phenotype as DHFR can catalyze the conversion of dihydrofolic acid to

* Correspondence: csrk@novonordisk.com
[1]Mammalian Cell Technology, Global Research Unit, Novo Nordisk A/S, A9.2.36, Novo Nordisk Park, 2760, Måløv, Denmark
[2]Network Engineering of Eukaryotic Cell Factories, Technical University of Denmark, Kgs Lyngby, Denmark
Full list of author information is available at the end of the article

tetrahydrofolic acid – an essential cofactor carrier. CHO-K1 cells were first exposed to a round of random chemical mutagenesis using Ethyl methanesulfonate (EMS) to generate the UKB25 cell line (*dhfr*+/*dhfr*-) [6] followed by a second round of mutagenesis using γ-radiation before isolation of the CHO DXB11 cell line (*dhfr*-/*dhfr*-) [7]. The DXB11 cell line was not mentioned by name in the original paper [6] and the name was not published until 1982 where the gene structure was more thoroughly investigated [7]. During the period between these two papers other laboratories used the cell line under the names CHO K1 DUX-B11 [8] and DUKX-CHO [9], explaining the origin of other names commonly used to describe the CHO DXB11 cell line. Further details concerning the clonal history of the CHO cell lines can be found in a recent review [10]. Historically, CHO DXB11 was the first CHO host cell for large scale production of a protein product (human tissue plasminogen activator [10,11]) and it is still being used for production of several protein products on the market.

Recently the genomic sequence of the Chinese hamster (*C. griseus*) [12,13] and seven CHO cell lines [12] were released making genomic comparisons of CHO cells possible for the first time. The first attempt to analyze the genomic information of the CHO DXB11 cell line was done in 2005 when Wlaschin *et al.* extracted 4608 expressed sequencing tags from CHO DXB11 RNA in order to create a CHO specific cDNA microarray [14]. This work furthermore lead to sequencing of the CHO mitochondrial genome. A 1x coverage of the genome of a CHO DXB11 transfectant producing human secreted alkaline Phosphatase was released back in 2011 [15] the same year as the CHO-K1 ATCC sequence was made public [3]. They furthermore reported that the *dhfr*-gene was detected albeit showing low coverage.

In this work, the genome of the CHO DXB11 cell line was sequenced with the goal of making this genome publicly available alongside the list of previously sequenced CHO genomes [12]. The genome was analyzed in order to validate the genomic cause of the DHFR negative phenotype of the cell line and the overall genome composition was compared to the currently sequenced CHO genomes. We found unique patterns for the evolution of each of the chromosomes from the Chinese hamster to each of the CHO cell lines and a surprising degree of haploidy.

Results

Sequencing depth per gene predicts gene haploidy and polyploidy

In order to gain insight into the genotype of the industrially relevant cell line CHO DXB11, as well as addressing some of the genomic consequences associated with creating a stable CHO transfectant the two genomes were sequenced. Genomic DNA was extracted from adherently growing CHO DXB11 cells as well as F435 cells, which is CHO DXB11 cells transfected with the gene encoding Coagulation Factor VIII [16] and subsequently adapted to suspension culture growth. Both cell lines were sequenced using the Illumina HiSeq 2000 platform yielding 0.5 bn paired reads for CHO DXB11 and 0.3 bn reads for F435. The reads were aligned to the Chinese hamster genome with a median depth of 33x and 17x respectively. The sequencing depth in CHO DXB11 for each of the 20661 genes found in the *C. griseus* genome were calculated and plotted, showing distinct peaks at sequencing depths of 0x, 16x, 33x and 49x coverage corresponding to a copy number in the genome of 0, 1, 2 and 3 copies (Figure 1A). The *dhfr* gene was found at a depth of 2.19 in the CHO-K1 ATCC genome, 0 in the CHO DG44 genome, 1.29 in the F435 genome and 1.06 in the CHO DXB11 genome in accordance with one allele being lost by gamma radiation in CHO DXB11 (L.A. Chasin, personal communication). Seven other genes found flanking *dhfr* on the same scaffold are also seen to be present in only one copy in the genomes of CHO DXB11 and F435. The *dhfr* gene from the remaining allele is found to be transcribed as seen from RNA sequencing data from F435 (Figure 1B) but contain a homozygous G410C point mutation located in the *dhfr* coding region causing a Thr137Arg missense mutation. This threonine is conserved from *C. elegans* to mouse, rat, hamster and human. In the crystal structure of murine DHFR [17] (96.3% similarity) the threonine is found in the cleft binding dihydrofolic acid right next to the active site, thus supporting the hypothesis that this mutation is able to effectively inactivate DHFR. In the process of deleting *dhfr* in the CHO DG44 genome it is found that four of the flanking genes were deleted (Zfyve16, Fam151b, Ankrd34b and Msh3) (Figure 1B).

A significant drift in single nucleotide polymorphisms is observed

In addition to the single nucleotide polymorphism (SNP) found in the *dhfr* gene, a total of 2,496,390 SNPs were found in the CHO DXB11 genome when aligned to the *C. griseus* genome (Table 1). For the CHO-K1 ATCC genome a higher total number of SNPs were detected but more SNPs were found in the coding regions of the CHO DXB11 genome (Table 1). 91% of the mutations interfering with translation in CHO-K1 ATCC were also found in CHO DXB11. All SNPs found in CHO-K1 ATCC and CHO DXB11 genes are listed in Additional file 1: Tables S3 and S4.

By comparing all the SNPs from the available CHO genomes to that of *C. griseus* a significant drift favoring GC → AT point mutations is evident for the two CHO DXB11 cell lines compared to the four sequenced CHO-K1 cell lines (Figure 1C), probably caused by the chemical mutagen used in the creation of the CHO DXB11

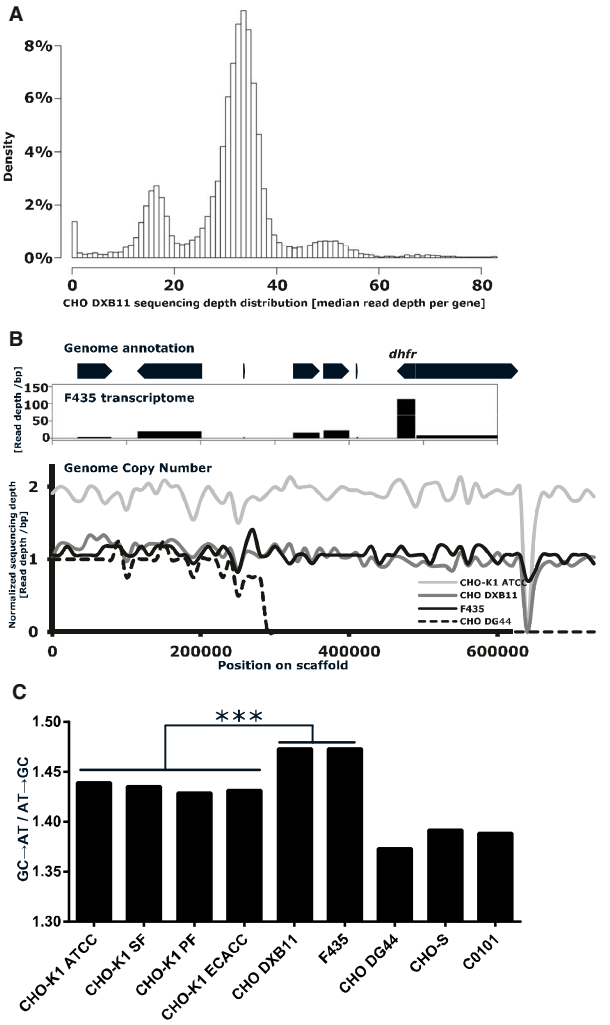


Figure 1 Copy number distribution for the CHO DXB11 genome. A) Distribution of the sequencing depth of the 20661 genes in the CHO DXB11 genome. District peaks are seen at depth of 0x, 16x, 33x and 49x corresponding to a copy number in the genome of 0, 1, 2 and 3 copies. **B)** Sequencing depth of the scaffold carrying *dhfr*. The entire 0.7 Mb scaffold is seen to be present in two copies in CHO-K1 ATCC, one copy in CHO DXB11 and roughly half the scaffold was deleted completely in CHO DG44. A peak in mRNA levels indicates that *dhfr* is seen to be transcriptionally active in F435 from the one allele that still contains the gene. **C)** The ratio of GC → AT and AT → GC mutations in CHO cell lines versus the *C. griseus* genome. A significant bias towards GC point mutations are observed in the two CHO DXB11 cell lines compared to the four CHO-K1 cell lines. Statistical significance level: ***$p < 0.001$.

cell line. However, different SNP biases were seen for CHO DG44 and CHO-S/C0101 respectively probably due to the distinct evolution of these cell lines (Figure 1C).

Copy number variation signature is chromosome dependent

By comparing the sequencing read depth for each gene from the CHO-K1 ATCC to the CHO DXB11 genome, it is seen that 96% of the genes are found in similar depths in both cell lines as expected, following the diagonal of the plot in Figure 2. No genes are observed to be deleted in the CHO-K1 ATCC genome but present in the CHO DXB11 genome as expected due to the origin of the CHO DXB11 cell line from CHO-K1. However, a total of 506 genes are seen to have reduced copy numbers in the CHO DXB11 genome compared to the CHO-K1 ATCC genome, whereas 389 genes have increased copy number. The copy number data for the CHO-K1 ATCC and the CHO DXB11 genome were furthermore mapped to the individual chromosomes of *C. griseus* (Figure 2) revealing a unique signature for each chromosome differing from the signature of the genome as a whole. It was not possible to separate chromosome nine from chromosome ten when sequencing the *C. griseus* genome and for this reason these two are listed together [13]. On chromosome two, where *dhfr* is situated, 89 genes have been reduced from diploid to haploid. Reductions are furthermore seen on chromosome five, six and seven. On chromosome three 158 genes are found to be triploid in CHO DXB11 versus diploid in CHO-K1 ATCC.

By estimating the extent of copy number variations (CNVs) between the currently sequenced CHO cell lines, a phylogenetic tree can be drawn, which accurately recapitulates the overall cell line history [10] (Figure 3). A heat map showing the extent of genes found to have reduced copy numbers between the different cell lines reveal specific patterns for each chromosome (Figure 4). From this, it can be seen that each chromosome is shown to have evolved differently across the cell lines and exhibits unique patterns. On chromosome 9/10, ~70% of the genes have been reduced from diploid in *C. griseus* to haploid in all cell lines except for C0101, CHO-S and F435 where only ~50% of the genes are haploid (Additional file 2: Figure S3). Chromosome X is seen to contain only 9% haploid genes in CHO-S whereas ~70% of the genes on this chromosome are found to be haploid in CHO DG44. Chromosome five appears to have undergone changes especially in F435 (amplification) and in DG44 (amplification of some genes and reduction of others)

Table 1 Overview of SNPs and indels in the CHO DXB11 and CHO-K1 ATCC genome

	CHO K1 ATCC	CHO DXB11
SNPs	2,527,490	2,496,390
Intronic SNPs	639,171	636,613
SNPs in CDS regions	19,096	21,142
SNPs missense/nonsense	8,195	9,357
Indels	341,848	315,422
Indels in CDS regions	211	259
Frameshifting indels	170	197

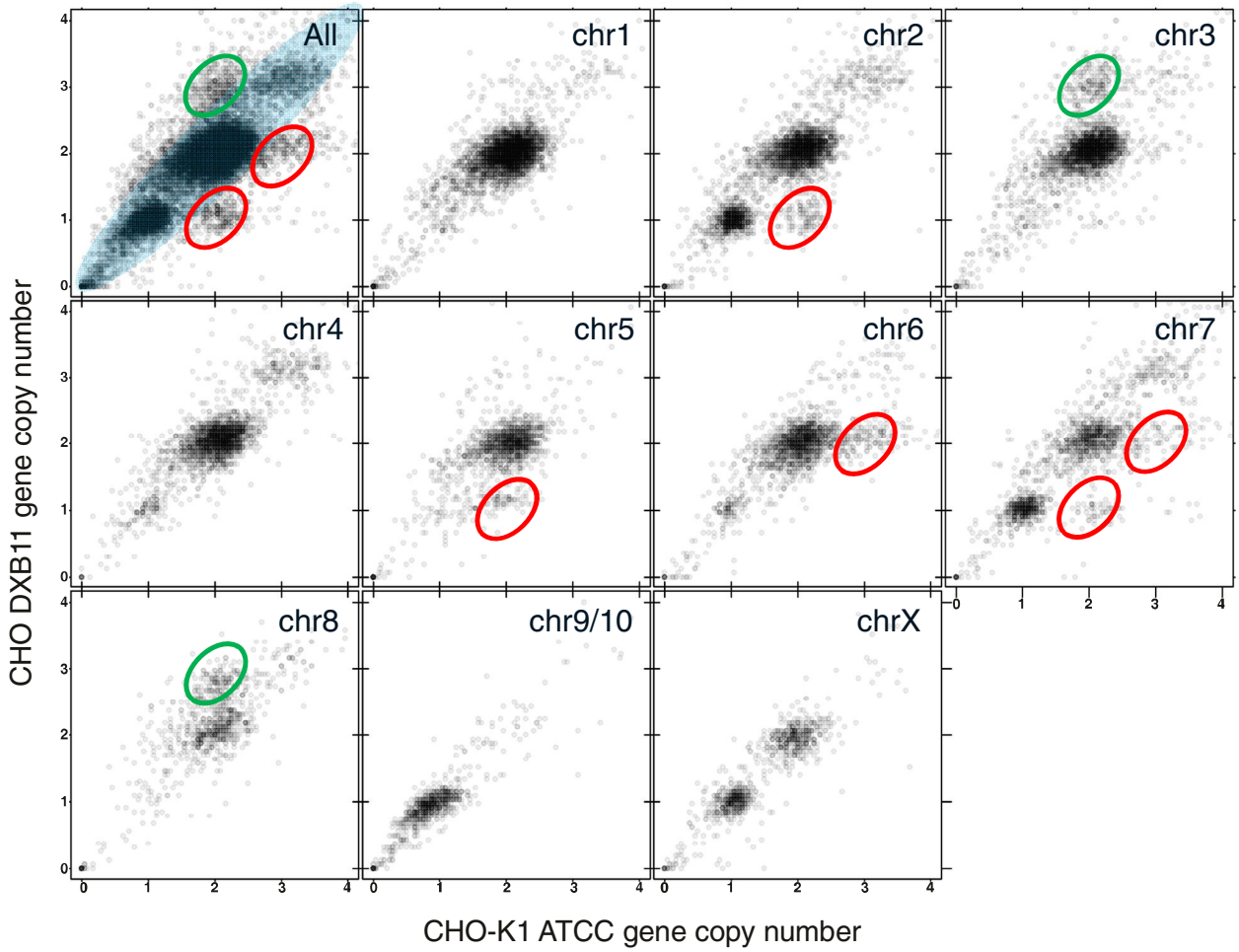


Figure 2 Copy number plot for the CHO-K1 ATCC versus CHO DXB11 genomes. The normalized sequencing depth of each gene in the *C. griseus* genome as found in CHO-K1 ATCC and CHO DXB11 listed by chromosome. Top left plot shows the distribution across all chromosomes. Red circle indicate genes, which have been reduced in copy number (e.g. from diploid to haploid), a green circle indicate duplication of genes and the blue circle in the top left plot indicate genes without change in copy number.

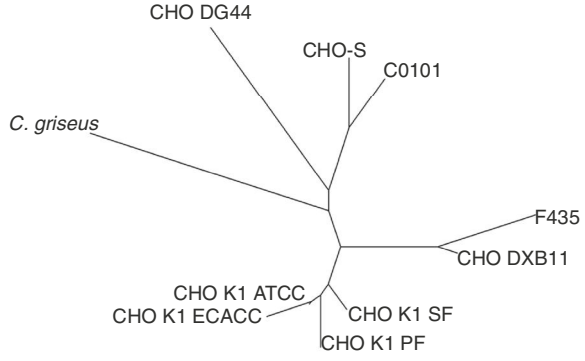


Figure 3 Phylogenetic tree based on copy number variation between cell lines.

(Additional file 2: Figure S2). Chromosomes one and four seem to be the most stable chromosomes for all the CHO cell lines. Overview of the number of genes found to be deleted, haploid, diploid or amplified in each genome are listed in Table 2.

In addition to chromosome specific changes in copy number it was investigated whether specific gene functions were seen to be enriched among genes with CNV's. 135 Gene Ontology (GO) terms were found to be significantly enriched with either amplified for reduced genes compared to *C. griseus*. The GO-terms were found to be conserved among all nine CHO cell lines (Additional file 1: Table S7, Additional file 2: Figure S4). The 3rd most significant GO-function were genes with transcription factor activity where CHO DXB11 and F435 had amplification of 49 and 51 genes respectively out of 432, whereas the other 8 CHO cell lines have 27 ± 3 genes amplified in this category. In the same category CHO DG44 were found to

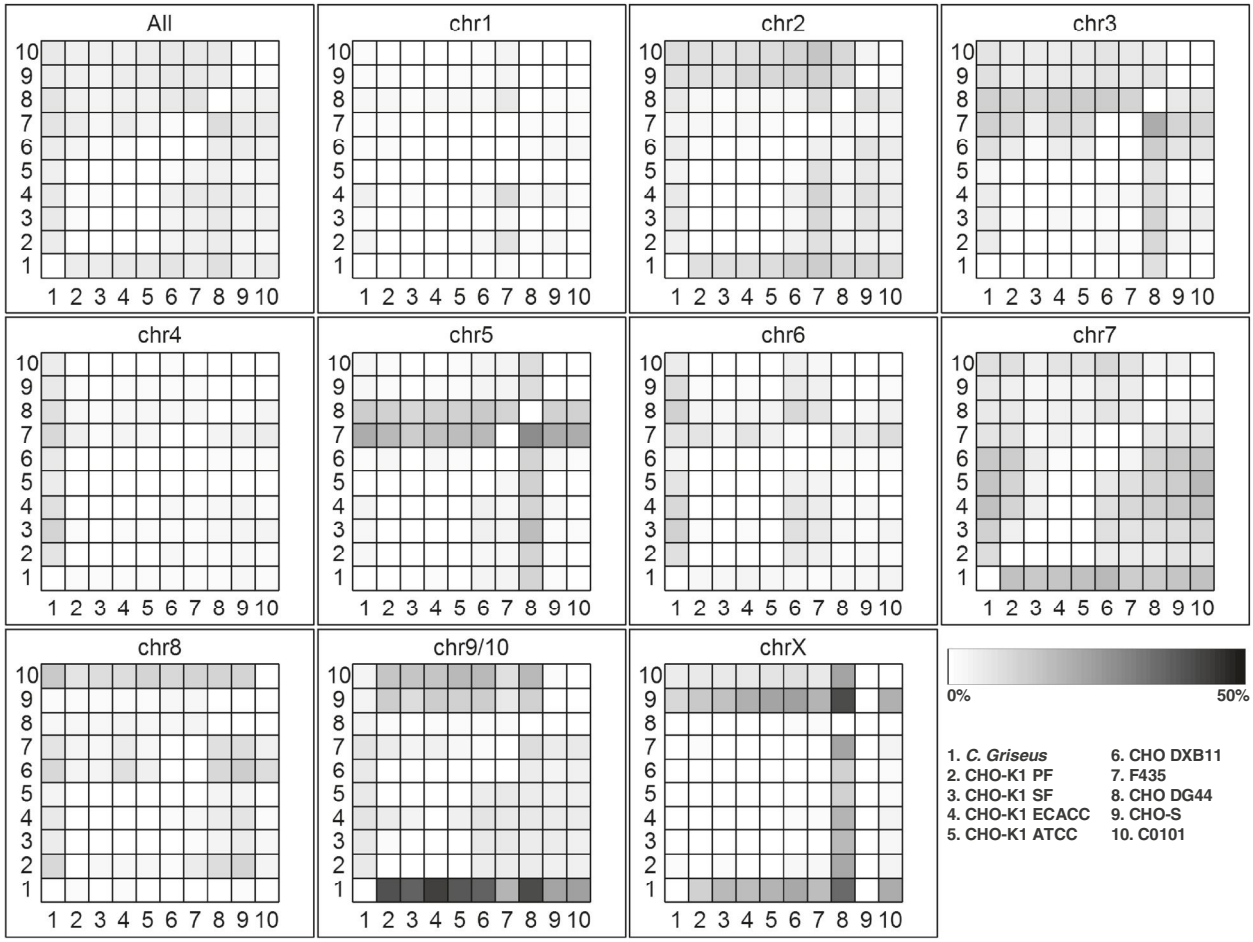


Figure 4 Heat maps showing the copy number variations among the 10 analyzed genomes. The intensity of coloring indicate the percentage of genes found on the given chromosome that has undergone a copy number reduction. Reduction in copy number in all cell lines except one is revealed by a vertical line and amplification of genes as a horizontal line.

Table 2 Overview of copy number estimation in the sequenced cell lines

	CN = 0	CN = 1	CN = 2	CN > 2
C. griseus	0	0	20,661	0
CHO-K1 ATCC	30	3,773	15,305	1,553
CHO-K1 ECACC	57	4,039	13,310	3,255
CHO-K1 PF	54	3,356	15,059	2,192
CHO-K1 SF	48	3,073	15,453	2,087
CHO DXB11	44	3,586	15,267	1,764
F435	53	3,888	14,306	2,414
CHO DG44	62	4,219	13,967	2,413
CHO-S	47	3,024	15,603	1,987
C0101	37	3,544	15,088	1,992

have 36 genes which had been reduced in copy number where the remaining eight cell lines had 16 ± 2 genes reduced.

Discussion

The genetic drift of SNPs from *C. griseus* to CHO DXB11 exhibited a significant disparity compared to the drift from *C. griseus* to the four CHO-K1 cell lines. The drift can likely be explained by the fact that the CHO DXB11 cell line was treated with the chemical mutagen EMS. The ethyl group of EMS is able to alkylate guanine forming O-6-ethylguanine which during replication commonly is paired with thymine and not cytosine [18,19]. Thus, an increased number of GC → AT mutations are expected and found in the genome of this cell line. It was seen that *dhfr* contained a missense mutation as well as the loss of an allele. These findings confirm observations by Lawrence Chasin in 1982 (personal communication, unpublished results). The threonine on position 137, which is mutated into an arginine in DXB11, is not found

in any of the annotated domains of the protein but is found to be highly conserved. The amino acid is located close to the active site and it can thus be hypothesized that the polar arginine is able to interfere with the structure of the binding cleft leading to inactivation of the enzyme. Due to the fact that only one allele of *dhfr* has been deleted and the other is transcriptionally active, it should be possible to find revertants as background in a transfection experiment. Indeed, revertants have been detected albeit at a frequency less than 10e-8 (L.A. Chasin, personal communication).

SNPs are the most frequent type of genetic polymorphism found when resequencing genomes from a common ancestor [20] and single mutations in the coding regions can result in significant changes for the phenotype of the cell lines. For sequenced diploid genomes without known and validated SNPs, it has previously been seen to filter out SNPs with a depth less than half the mean depth of the genome [12]. But in the case of CHO DXB11 (or any of the other sequenced CHO cell lines) that practice would result in SNPs found in most haploid genes (17% of the genes) to be remain undetected. For that reason a more lenient filter has to be applied in the current study for homozygous SNPs, which can be found in the haploid genes, and a more stringent filter requiring higher depth for heterozygous SNPs. The need to check for mutations in the relevant cell line before designing a laborious knock down, knockout or PCR primer solely based on the *C. griseus* genome sequence is highlighted by the 9357 SNPs, which were found in the CHO DXB11 genome within the coding regions of 3458 genes. All genes containing SNPs as well as genes containing indels for CHO-K1 ATCC and CHO DXB11 are listed in Additional file 1: Table S2. SNPs located in the other sequenced cell lines are not included due to the fact that the current lower sequencing depth of these resulted in a ~20% probability of a correct SNP call versus >99% in CHO DXB11 and K1 ATCC.

The copy number of a gene has previously been found to correlate well with the sequencing depth [21,22] and this correlation was used to determine the copy number of the 20661 genes both in the *C. griseus* genome and in the nine sequenced CHO cell lines. Based on analysis of the sequencing depth per gene it was found that only wild type *C. griseus* had a single peak centered on a copy number of two whereas the sequenced CHO cell lines derived from this organism show distinct peaks revealing large number of genes only present in one copy or amplified to three or more copies. Interestingly, the data indicates that 15-20% of the genes found in the nine CHO cell lines sequenced to date are haploid. Historically, the CHO cells were often regarded functionally haploid at many genetic loci making these cell lines ideal for investigation of molecular functions in a eukaryotic cell model [14,23,24]. This current data explain this perceived haploid phenotype and this knowledge can also be advantageous for choosing a knock-out target in a specific pathway or for elucidating target region for a knock in. As chromosome one and four have the lowest level of CNVs across the nine CHO cell lines, some of which have encountered heavy mutation pressure over a long period of time [10], these chromosomes may be considered attractive landing platforms for knock-ins of heterologous genes.

Based on the current sequencing data from CHO DXB11 and available data from other CHO cell lines a phylogenetic tree was created based on the copy variations between the cell lines. The tree reflects the clonal history of the cell lines (see recent review for details [10]) and correlates well with a phylogenetic tree recently published based on SNPs [12]. The CHO-K1 cell lines lie in one cluster neighboring the CHO DXB11 cluster and distant from the CHO DG44 and CHO-S/C0101 branches. The clonal history of the F435 cell line by transfection of CHO DXB11 is apparent from the phylogenetic tree as F435 emerges out of the same branch as CHO DXB11 (Figure 3). Nonetheless, a total of 907 genes are found to have undergone a CNV in the process from transfection of a pool of CHO DXB11 cells, amplification of the insert and subsequent adaptation to suspension culture growth.

To give a more precise estimate of the genomic differences the sequencing depth across the coding regions of the genes (1.7 kb on average) were measured. This allowed for normalization using hundreds of sequencing reads compared to looking at SNPs only supported by a dozen or so reads at most. Furthermore, the coding DNA sequences (CDS's) are the most uniquely defined elements of the genome. For this reason, assessment of the CNVs are able to provide a more detailed look into the cell lines even when only very low sequencing depths are available. This is highlighted by the clustering of the CHO-K1 cell lines closely together, even though CHO-K1 ATCC has been sequenced to a depth of 45x which is ~6x that of the other three CHO-K1 cell lines and C0101 is sequenced to a depth 3x that of CHO-S.

Each chromosome showed a distinct signature of CNVs giving for the first time an insight into CHO chromosomal genome stability from next-generation sequencing data. In the future this method could be used in combination with FISH to validate hypotheses on e.g. the range of genetic reductions and rearrangements on a particular chromosome (e.g. chromosome two containing *dhfr*) from CHO-K1 ATCC to CHO DXB11. The method also revealed a large number of haploid genes on chromosome 9/10 which seem to have been reduced in the earliest CHO cell lines prior to evolving into the cell lines sequenced today. Some of these mutations might have been critical for establishing the independent immortal first CHO cell lines.

Each rearrangement event that has occurred in the evolution from CHO-K1 to CHO DXB11 may have had an impact on a multitude of genes as seen by all eight genes on the scaffold holding *dhfr* were found to be haploid (Figure 1B) but probably caused by one single deletion event during UV radiation. Therefore, the number of genes with altered CNVs is no true indication of the number of genomic rearrangements that has occurred as one large rearrangement events could impact dozens of genes. Due to the short lengths of the genomic scaffolds in the current versions of the *C. griseus* and CHO-K1 genomes, it is not yet feasible at this time to piece together the rearrangement history of the CHO cell lines, but 3rd generation sequencing could permit the construction of a *C. griseus* genome with a reduced number of scaffolds [25]. Once a more complete scaffold is available, CNV data could be used to make genomic based chromosomal maps which currently are only done using FISH [26-28].

With improved genome constructions possible in coming years, it will be informative to elucidate more detailed genomic differences between CHO DG44 and CHO DXB11 as they are the two DHFR negative CHO cell lines most widely used today. As described earlier a comparison of found SNPs is not practical with the available data but it is seen from the copy numbers that 4219 genes in CHO DG44 are haploid versus 3586 in CHO DXB11. In addition, there are 44 deleted genes in CHO DXB11 versus 62 in CHO DG44. This difference can be ascribed to the harsh UV treatment that the cells were exposed to in the process of creating the CHO DG44 cell line compared to the relatively mild UV treatment of the CHO DXB11 cell line. The availability of additional sequence information from CHO-S and CHO DG44 among other would greatly improve the possibilities for comparing the genomic differences across a wider range of different CHO cell lines in the coming decades. This comparison could be highly informative about the evolutional path and diversity that exists across CHO cell hosts.

Analysis for enrichment of GO-terms revealed that a large portion of the changes in e.g. transcription factor copy number must have occurred in the early CHO cell. It appears that CHO DXB11 and F435 have further amplified ~20 transcription factors and CHO DG44 has reduced approximately the same number. Further studies of the transcriptome and proteome should be able to reveal the effects of these genomic changes and link differences in genotype and phenotype.

Conclusions

In this work we have described the full genome sequencing of CHO DXB11 including SNPs and CNVs which differ between this cell line and the other CHO genomes that have been sequenced to date. The DHFR negative phenotype of the cell line was verified based on the lack of one allele and a missense mutation in the other transcriptionally active allele. The analysis of the CNVs revealed a large number of genes that were found to be haploid in the CHO cell lines which is important for correct SNP detection and detection strategy for knock out verification. It furthermore revealed unique patterns for the evolution of each of the chromosomes from the Chinese hamster to each of the sequenced CHO cell lines with chromosome one and four showing the lowest level of change.

Methods

Cell culture and genome extraction

CHO DXB11 cells were thawed from an in-house master cell bank. The cell bank was generated in 2000 from a vial of CHO DXB11 from L.A. Chasin, Columbia University. The cells were passaged in alpha MEM media with 10% FBS, 1% NEAA, 1% P/S. A second in-house suspension culture adapted CHO DXB11 cell line transfected with a plasmid encoding coagulation factor VIII coupled to *dhfr* was grown in HyClone CDM4CHO media supplemented with 1% Penicillin/Streptomycin and 100 nM MTX. Genomic DNA from both cell lines was extracted from 2 mio cells using DNeasy Blood & Tissue Kit (Qiagen) following manufacturer's instructions. gDNA library and next-generation sequencing were performed by AROS Applied Biotechnology (Aarhus, Denmark) in a Illimina Hisq 2000 system for paired-end sequencing.

NGS data treatment

The FASTQC tool (www.bioinformatics.bbsrc.ac.uk/projects/fastqc/) was used to evaluate the quality of the fastq files before and after treatment. The FASTX Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) was used to remove the adaptamers (fastx_trimmer) and trim the ends for bps with a quality score lower than 20 (fastq_quality_trimmer). An in-house algorithm was used to intersect the read-pairs after quality trimming. The reads were aligned to the *C. griseus* genome (downloaded from Genbank as assembly GCF_000419365.1) using BWA (version 0.6.2). The RealignerTargetCreator from GATK (version 1.6) was used to realign the reads in problematic regions and duplicate reads were removed using Picard MarkDuplicates (http://picard.sourceforge.net/).

The depth of reads at each position on the genome for identification of deleted genes was calculated using genomeCoverageBed from BEDTools (version 2.16.2). The depth for each gene was found by extracting the depth for each position in the coding region subsequently calculating the median using a custom script. The depths of each gene were normalized by 0.5x the median depth of all genes. The data were analyzed in R [29]. The depths are listed in Additional file 1: Table S5 and an overview of all genes in *C. griseus* with Genbank IDs, GC content and

chromosome number is listed in Additional file 1: Table S6. The aligned reads from CHO DXB11 to the *C. griseus* genome was uploaded to the SRA under experiment ID: SRX689758. For direct download of raw reads from all public CHO cell lines currently in the SRA see Additional file 1: Table S1.

SNP calls

SNPs were detected by samtools mpileup and bcftools. SNPs present in the coding region were found using CLC genomic workbench (CLC Bio, version 7). Haploid SNPs were detected by the filter: A minimum depth of 0.25x the median depth measured in the CDS regions, 90% of the reads calling the SNP had to differ from the reference sequence. Heterologous SNPs: A minimum depth of 0.75x the median depth measured in the CDS regions, a minimum of 40% of the reads should agree with the reference and 40% with the called allele. All SNPs interspaced by less than 5 bp to another SNP or an indel were filtered away as well as SNPs found when aligning the raw *C. griseus* reads to the *C. griseus* genome. Mutational bias was calculated by extracting reference and non-reference bases from the filtered SNP files and calculating the occurrence of different nucleotide transitions. Significance of bias were calculated as standard student t-test assuming equal variance comparing the observations for GC → AT/AT → GC for CHO-K1 PF, CHO-K1 SF, CHO-K1 ECACC and CHO-K1 ATCC versus CHO DXB11 and F435.

Phylogenetic tree based on CNV data

Raw sequencing reads from CHO-K1 [3] and other CHO cell lines [12] were downloaded from the SRA (Additional file 1: Table S1). Reads were trimmed and intersected as described above and aligned to the *C. griseus* genome using BWA. Depths were estimated for each gene in each cell lined as described above and the occurrence of CNVs were estimated as the number of genes differing by more than 0.95 in normalized sequencing depth between two cell lines. A distance matrix were calculated, a phylogenetic tree was constructed using R with the package ape and phangorn using the neighbour joining algorithm. The tree was bootstrapped 100 times and the consensus tree was used.

Gene copy number

The absolute copy number for each gene in each cell line was calculated as above by normalizing the read depth to the median. Genes were considered to be deleted if the read depth were 0 in a given cell line but > 0.95 in *C. griseus* (as 159 genes were ~0 in all cell lines incl. *C. griseus*). Haploid: depth higher than 0 but lower than 1.3 based on local minimum in CHO DXB11 between the haploid and diploid peak. Diploid: higher than 1.3 and lower than 2.7 based on local minimum in CHO DXB11 between the diploid and triploid. Triploid or higher: depth higher than 2.7.

Chromosomal changes

All genes from the *C. griseus* genome (assembly GCF_000419365.1) as listed in the genome annotation file were downloaded from Genbank. The chromosome sorted *C. griseus* genome was downloaded from Genbank as Accession APMK00000000. The scaffold name and chromosome number was extracted from the fasta header. All genes were blasted against the chromosome sorted genome and the best hit (cutoff E-value = 0.05) was used as indicator for the chromosomal location.

RNA sequencing

In order to deduce the transcriptional activity of *dhfr* from F435, RNA was extracted. A sample was taken 48 hours into the cultivation from 2x10e6 cells and RNA was extracted using TRIzol (Invitrogen) and the RNeasy Cleanup kit (Qiagen) following manufacturer's instructions. RNA integrity was confirmed on an Agilent 2100 Bioanalyzer using total RNA nano chips (Agilent technologies, Santa Clara, Ca, USA). RNA concentration was measured using a NanoDrop spectrophotometer (NanoDrop Technologies). Multiplexed cDNA library generation and next-generation sequencing were performed by AROS Applied Biotechnology (Aarhus, Denmark) in an Illumina Hiseq 2000 system for paired-end sequencing. The FASTQC tool (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to evaluate the quality of the fastq files before and after treatment. The FASTX Toolkit (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to remove the adaptamers (fastx_trimmer) and trim the ends for base pairs with a quality score lower than 20 (fastq_quality_trimmer). An in-house algorithm was used to intersect the read-pairs after quality trimming. The reads were aligned to the CHO-K1 genome (downloaded from Genbank as assembly GCF_000223135.1) using tophat2 [30]. RNA expression from each of the genes on the NW_003614442.1 scaffold containing *dhfr* was calculated using genomeCoverageBed from BEDTools (version 2.16.2) and the mean expression level was used as indicator for expression.

GO-term enrichment

A list of GO-terms associated with the CHO genome was downloaded from CHOgenome.org (http://www.chogenome.org/files/CHO_GO_Functions_12Sep13.txt), the list was rearranged and imported into R. Fisher's exact test were used to find GO-terms enriched for either reduced or amplified genes for each of the CHO cell lines compared to *C. griseus* by > 0.95 difference in normalized copy number. All GO-terms, which had a p-value < 0.01

in just one cell line for either amplification or reduction in CN were included. Data listed in Additional file 1: Table S7.

Additional files

Additional file 1: Table S1. SRA. **Table S2.** SNP qeneoverview. **Table S3.** Exon SNP's DXB11. **Table S4.** Exon SNP's K1ATCC. **Table S5.** Sequencing depth. **Table S6.** Cqriseus overvie. **Table S7.** GO-terms.

Additional file 2: Figure S1. Read depth analysis of the 20661 in the C. griseus genome. For most of the genomes distinct peaks can be seen for genes present in one, two and three copies. Shoulders can be seen in the graphs for the cell lines sequenced at a lower depth. Only one peak is seen in wild type C. griseus as expected. **Figure S2.** The normalized sequencing depth of each gene in F435 and CHO DG44. Top left plot shows the distribution across all chromosomes. Compared to Figure 2 this reveals a much larger difference in CN. **Figure S3.** Distribution of CN across chromosomes for each cell line. **A)** Percentage of haploid genes **B)** percentage of diploid genes **C)** Percentage of genes, which are triploid or higher. **Figure S4.** Significant GO-terms in correlation to changes in CN Visualization of the 135 GO-terms, which are either significantly enriched in genes with CN reductions or amplifcations (Fisher's exact test). GO-terms are visualized in dark blue (p-value < 0.01), light blue (p-value < 0.05) or white (p-value > 0.05). Data attached in Additional file 1: Table S7.

Competing interest
The authors declare that they have no competing interests.

Authors' contributions
CSK carried out the cell work, data analysis and drafted the manuscript. CK, MBJ and MR provided guidance throughout the data analysis and manuscript formulation. All authors read and approved the final manuscript.

Acknowledgements
We would like to acknowledge Dr. Nathan Lewis for kindly providing raw reads from the *C. griseus* sequencing [12], Dr Lawrence Chasin for providing essential input on the history of the CHO DXB11 cell line and finally the reviewers for providing essential input for improving the paper.

Author details
[1]Mammalian Cell Technology, Global Research Unit, Novo Nordisk A/S, A9.2.36, Novo Nordisk Park, 2760, Måløv, Denmark. [2]Network Engineering of Eukaryotic Cell Factories, Technical University of Denmark, Kgs Lyngby, Denmark. [3]Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, MD, USA. [4]Institute of Cellular and Molecular Medicine, University of Copenhagen, Copenhagen, Denmark.

Received: 28 November 2014 Accepted: 24 February 2015
Published online: 08 March 2015

References
1. Walsh G. Biopharmaceutical benchmarks 2014. Nat Biotechnol. 2014;32:992–1000.
2. Kim JY, Kim YG, Lee GM. CHO cells in biotechnology for production of recombinant proteins: current state and further potential. Appl Microbiol Biotechnol. 2012;93:917–30.
3. Xu X, Nagarajan H, Lewis NE, Pan S, Cai Z, Liu X, et al. The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. Nat Biotechnol. 2011;29:735–41.
4. Puck TT. Genetics Of Somatic Mammalian Cells: III. Long-term Cultivation Of Euploid Cells From Human And Animal Subjects. J Exp Med. 1958;108:945–56.
5. Kao FT, Puck TT. Genetics of somatic mammalian cells, VII. Induction and isolation of nutritional mutants in Chinese hamster cells. Proc Natl Acad Sci USA. 1968;60:1275–81.
6. Urlaub G, Chasin L. Isolation of Chinese hamster cell mutants deficient in dihydrofolate reductase activity. Proc Natl Acad Sci U S A. 1980;77:4216–20.
7. Graf Jr LH, Chasin LA. Direct demonstration of genetic alterations at the dihydrofolate reductase locus after gamma irradiation. Mol Cell Biol. 1982;2:93–6.
8. Gasser CS, Simonsen CC, Schilling JW, Schimke RT. Expression of abbreviated mouse dihydrofolate reductase genes in cultured hamster cells. Proc Natl Acad Sci. 1982;79(21):6522–6.
9. Kaufman RJ, Sharp P. Amplification and expression of sequences cotransfected with a modular dihydrofolate reductase complementary dna gene. J Mol Biol. 1982;159:601–21.
10. Wurm F. CHO Quasispecies - Implications for Manufacturing Processes. Processes. 2013;1:296–311.
11. Kaufman RJ, Wasley LC, Spiliotes AJ, Gossels SD, Latt S, Larsen GR, et al. Coamplification and coexpression of human tissue-type plasminogen activator and murine dihydrofolate reductase sequences in Chinese hamster ovary cells. Mol Cell Biol. 1985;5:1750–9.
12. Lewis NE, Liu X, Li Y, Nagarajan H, Yerganian G, O'Brien E, et al. Genomic landscapes of Chinese hamster ovary cell lines as revealed by the Cricetulus griseus draft genome. Nat Biotechnol. 2013;31:759–65.
13. Brinkrolf K, Rupp O, Laux H, Kollin F, Ernst W, Linke B, et al. Chinese hamster genome sequenced from sorted chromosomes. Nat Biotechnol. 2013;31:694–5.
14. Wlaschin KF, Nissom PM, Gatti MDL, Ong PF, Arleen S, Tan KS, et al. EST sequencing for gene discovery in Chinese hamster ovary cells. Biotechnol Bioeng. 2005;91:592–606.
15. Hammond S, Swanberg JC, Kaplarevic M, Lee KH. Genomic sequencing and analysis of a Chinese hamster ovary cell line using Illumina sequencing technology. BMC Genomics. 2011;12:67.
16. Thim L, Vandahl B, Karlsson J, Klausen NK, Pedersen J, Krogh TN, et al. Purification and characterization of a new recombinant factor VIII (N8). J World Federation Hemophilia. 2010;16:349–59.
17. Cody V, Pace J, Rosowsky A. Structural analysis of a holoenzyme complex of mouse dihydrofolate reductase with NADPH and a ternary complex with the potent and selective inhibitor 2,4-diamino-6-(2'-hydroxydibenz[b, f] azepin-5-yl)methylpteridine. Acta Crystallogr D Biol Crystallogr. 2008;64:977–84.
18. Heflich RH, Beranek DT, Kodell RL, Morris SM. Induction of mutations and sister-chromatid exchanges in Chinese hamster ovary cells by ethylating agents. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 1982;106:147–61.
19. Flibotte S, Edgley ML, Chaudhry I, Taylor J, Neil SE, Rogula A, et al. Whole-genome profiling of mutagenesis in Caenorhabditis elegans. Genetics. 2010;185:431–41.
20. Brookes AJ. The essence of SNPs. Gene. 1999;234:177–86.
21. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet. 2009;41:1061–7.
22. Medvedev P, Fiume M, Dzamba M, Smith T, Brudno M. Detecting copy number variation with mated short reads. Genome Res. 2010;20:1613–22.
23. Jeggo PA, Holliday R. Azacytidine-induced reactivation of a DNA repair gene in Chinese hamster ovary cells. Mol Cell Biol. 1986;6:2944–9.
24. Siminovitch L. On the nature of hereditable variation in cultured somatic cells. Cell. 1976;7:1–11.
25. Borth N. Opening the black box: Chinese hamster ovary research goes genome scale. Pharma Bioprocessing. 2014;2:367–9.
26. Cao Y, Kimura S, Itoi T, Honda K, Ohtake H, Omasa T. Construction of BAC-based physical map and analysis of chromosome rearrangement in Chinese hamster ovary cell lines. Biotechnol Bioeng. 2012;109:1357–67.
27. Toledo F, Buttin G+, Debatisse M. The origin of chromosome rearrangements at early stages of AMPD2 gene amplification in Chinese hamster cells. Curr Biol. 1993;3:255–64.
28. Day J. Recombination involving interstitial telomere repeat-like sequences promotes chromosomal instability in Chinese hamster cells. Carcinogenesis. 1998;19:259–65.
29. Ihaka R, Gentleman R. R: A Language for Data Analysis and Graphics. J Computational Graph Stat. 1996;5:299–314.
30. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25:1105–11.

2.2 Outlook and future perspectives

Following the work published in BMC Genomics [1], the method used for deduction of copy numbers has been further improved making identification of local changes in copy number detectable (Figure 1). At the time of writing, the final stage of sequencing is taking place at Johns Hopkins and the University of Delaware (personal communication: Kelley Heffner), in order to update the assembly of the *C. griseus* genome using 3rd generation sequencing reads [2]. This assembly is expected to combine most of the 53,000 scaffolds currently making up the *C. griseus* genome into a more coherent representation. If this is achieved, an unprecedented level of detail will be able to be detected for copy number changes and will allow identification of which chromosomal regions have been amplified and reduced.

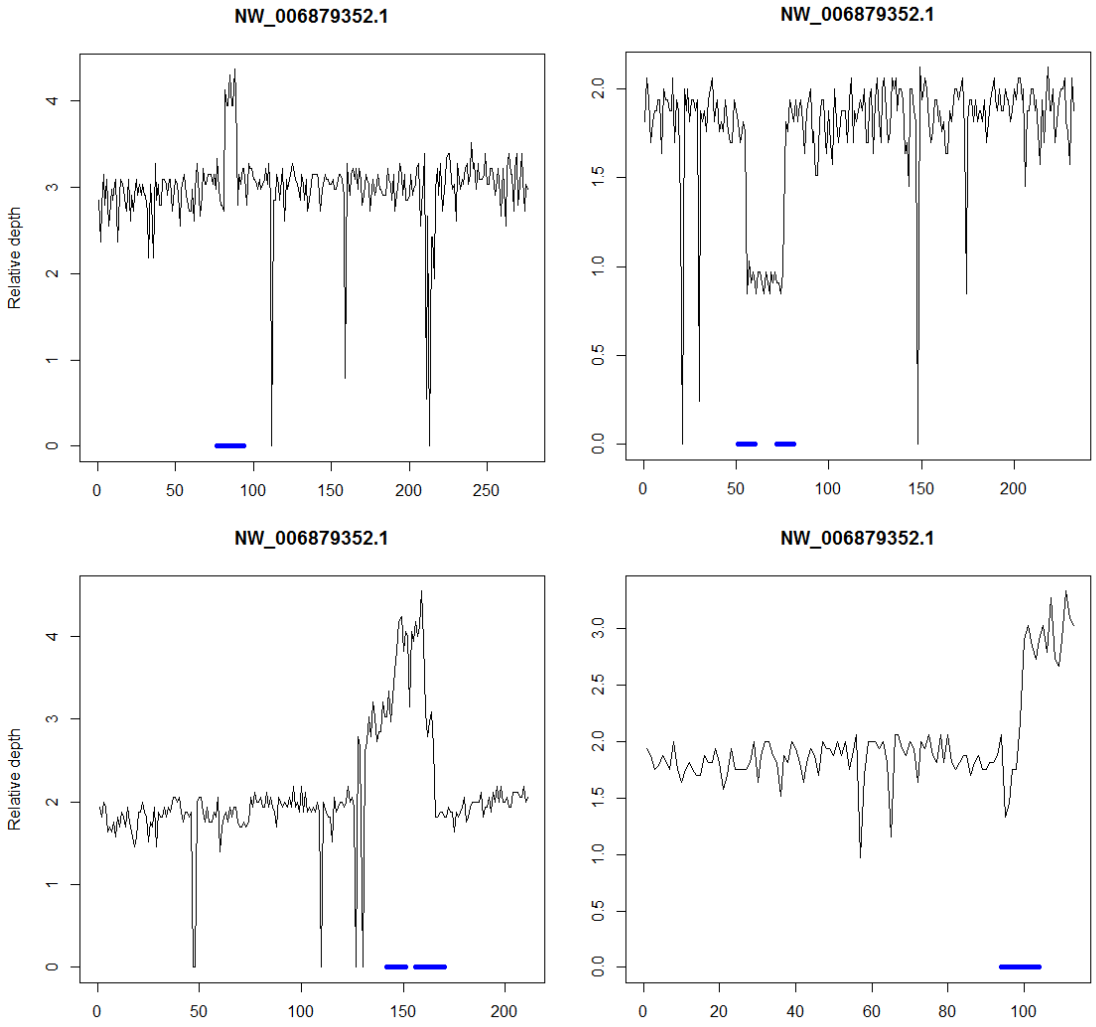


Figure 1 Copy number across four scaffolds from chromosome 2
Local differences in copynumber can be detected with relative ease providing an estimate for the number of rearrangement events in the different CHO genomes. Blue line indicate a detected difference in copy number. X-axis represent the position on the scaffold given with a scale of 10,000bp.

From the data presented previously [1] it was apparent that large regions of the CHO genomes are present in a copy number higher than two, which to some extent can explain the large number of scaffolds in the CHO-K1 ATCC sequencing project [3]. The reason why the scaffolds cannot be assembled is probably not solely a matter of assembly of short reads in regions containing repeats, but also a matter of trying to assemble a complex quadroploid stretch of DNA and demand a haploid representation in fasta format. This also reveals the need for creation of a well annotated *C. griseus* genome, because it is probably the closest organism we can sequence, which is stable in terms of copy number. This will allow for a haploid representation in fasta-format with the lowest number of scaffolds and thus provide a better overview.

2.2.1 References

1. Kaas CS, Kristensen C, Betenbaugh MJ, Andersen MR: **Sequencing the CHO DXB11 genome reveals regional variations in genomic stability and haploidy.** *BMC genomics* 2015, **16:** 1391.

2. Borth N: **Opening the black box: Chinese hamster ovary research goes genome scale.** *Pharmaceutical Bioprocessing* 2014, **2:** 367-369.

3. Xu X, Nagarajan H, Lewis NE, Pan S, Cai Z, Liu X *et al.*: **The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line.** *Nature biotechnology* 2011, **29:** 735-741.

Chapter 3 – Transgene expression in CHO

In this chapter, the focus will be moved from the genome as a whole to specifically the *F8* transgene in CHO transfectants producing FVIII. Transcriptomics data has most commonly been achieved using microarrays: predefined oligos, which based on annealing with sample, can be used to deduce expression level. One of the advantages of RNA sequencing is that whatever is in the sample will be sequenced, thus allowing for insights into the composition and expression of the transgene used for transfection of the cell line in addition to the general transcriptome. First, the paper: *"Deep sequencing reveals different compositions of mRNA transcribed from the F8 gene in a panel of FVIII-producing CHO cell lines"* is attached and next data from the characterization of the *F8* transgene in the Clone 1 genome (also referred to as F435 in this thesis) is shown. Finally, data is presented from co-expression of the *F8* transgene and three different protein disulfide isomerases in the Icosagen system revealing substantial selection pressure in the cell pools against the *F8* transgene favoring plasmids containing protein disulfide isomerases.

Biotechnology Journal
www.biotechnology-journal.com

Biotechnol. J. 2015, 10

DOI 10.1002/biot.201400667

Research Article

Deep sequencing reveals different compositions of mRNA transcribed from the *F8* gene in a panel of FVIII-producing CHO cell lines

Christian S. Kaas[1,2], Gert Bolt[1], Jens J. Hansen[1], Mikael R. Andersen[2] and Claus Kristensen[1,3]

[1] Mammalian Cell Technology, Novo Nordisk A/S, Maaloev, Denmark
[2] Department of Systems Biology, Technical University of Denmark, Kgs Lyngby, Denmark
[3] Department of Cellular and Molecular Medicine, University of Copenhagen, Copenhagen, Denmark

Coagulation factor VIII (FVIII) is one of the most complex biopharmaceuticals due to the large size, poor protein stability and extensive post-translational modifications. As a consequence, efficient production of FVIII in mammalian cells poses a major challenge, with typical yields two to three orders of magnitude lower than for antibodies. In the present study we investigated CHO DXB11 cells transfected with a plasmid encoding human coagulation factor VIII. Single cell clones were isolated from the pool of transfectants and a panel of 14 clones representing a dynamic range of FVIII productivities was selected for RNA sequencing analysis. The analysis showed distinct differences in *F8* RNA composition between the clones. The exogenous *F8-dhfr* transcript was found to make up the most abundant transcript in the present clones. No correlation was seen between *F8* mRNA levels and the measured FVIII productivity. It was found that three MTX resistant, non-producing clones had different truncations of the *F8* transcripts. We find that by using deep sequencing, in contrast to microarray technology, for determining the transcriptome from CHO transfectants, we are able to accurately deduce the mature mRNA composition of the transgene and identify significant truncations that would probably otherwise have remained undetected.

Received	19 JAN 2015
Revised	27 MAR 2015
Accepted	08 MAY 2015
Accepted article online	11 MAY 2015

Keywords: Cell culture · CHO cells · Coagulation factor VIII · Gene delivery · Next-generation sequencing

1 Introduction

Coagulation factor VIII (FVIII) is a 170–280 kDa glycoprotein that plays a vital role in blood coagulation. Deficiency of FVIII results in the congenital bleeding disorder Haemophilia A which can be treated by infusion of plasma-derived or recombinant FVIII [1]. Recombinant production of FVIII is preferred due to virus risk associated with blood products. However, recombinant FVIII is one of the most challenging therapeutics to produce due to the size of the FVIII protein, the multitude of post-translational modifications, and the low yields obtained [2].

Endogenous FVIII is synthesized as a single-chain precursor peptide that is processed into a heterodimer consisting of a heavy and a light chain connected by a metal ion bridge [3, 4]. Due to heterogeneous proteolytic processing of the heavily glycosylated B domain, in the C-terminal part of the heavy chain, the molecular weight of the heavy chain is highly variable [5]. The function of the B domain remains to be determined, as FVIII with and without B domain has the same haemostatic effect [6]. Shortening of the 904 amino acid B domain to a minimal B domain of 10–30 amino acids reduces the complexity and increases the yield of recombinant FVIII [7–9]. Still, typical yields of recombinant FVIII from standard mammalian production cell lines such as Chinese hamster ovary (CHO) cells are 100–1000-fold lower than the yields

Correspondence: Christian S. Kaas, Mammalian Cell Technology, Novo Nordisk A/S, A9.2.36, Novo Nordisk Park, 2760 Maaloev, Denmark
E-mail: csrk@novonordisk.com

Abbreviations: FVIII, coagulation factor VIII; **CHO,** Chinese hamster ovary; **FPKM**, fragments per kilobase of exon per million fragments mapped; **IRES**, internal ribosomal entry site; **TLA**, targeted locus amplification.

routinely obtained with recombinant antibodies from the same cells [6].

Generation of mammalian cell lines for production of recombinant protein typically involves the transfection of a suitable cell line with an expression construct encoding the gene of interest and the subsequent selection and screening to isolate cell clones that have incorporated the exogenous expression cassette into their genome. Different clones exhibit a wide variation in productivity, genetic stability, and performance in a production setting [10]. The productivity of mammalian cell clones are known to be influenced by the composition of the expression vector, the site of insertion in the host cell genome, the metabolic and growth characteristics of the clone, and the cultivation conditions [10]. Several different plasmid systems exist. For the current study we employed use of an IRES element [11] to allow for coupled translation of both a gene of interest and a selection marker from the same bicistronic mRNA [12].

In recent years, microarray analysis has allowed investigation and comparison of the transcriptome of mammalian production cell lines [13–15]. With the advances within deep sequencing and the publication of CHO genomes [16–18] (see recent review [19]) it is now possible to analyze the transcriptome with a much higher level of detail [20–22].

In the present study, we compare the growth, metabolic profile and FVIII productivity of a selection of CHO cell clones and utilize deep sequencing to relate these features to the transcription of the individual elements in the vector cassette directing the expression of FVIII and the selection marker. To our knowledge, this is the first report on the use of deep sequencing to reveal detailed variations of the mRNA transcribed from the exogenous expression cassette in mammalian production cell lines.

2 Materials and methods

2.1 Cell culture

CHO DXB11 cells were transfected with a plasmid encoding FVIII by electroporation (GenePulser Xcell, Biorad). The plasmid contained the adenovirus-2 major late promotor, the adenovirus-2 late mRNA tripartite leader, a 3' end and 5' end intron sequence with splice junction, the ORF of *F8* with a minimal artificial B-domain [23], an internal ribosomal entry site (IRES), the *dhfr* ORF, and a poly A signal, similar to constructs used elsewhere [24, 25]. Cells were adapted to MTX and subsequently single cell cloned by limiting dilution. Productivity of 59 clones was measured by seeding at 3×10^5 cells/mL in 30 mL of HyClone CDM4CHO media (Thermo) supplemented with Penicillin/Streptomycin (Gibco) and MTX, and cultured for 72 h before measuring FVIII chromogenic activity (see assays). In order to get the broadest range of productivities, a mother clone (clone 3) from a previous transfection created under similar conditions as above as well as two subclones (clone 1 and 2) created by single cell dilution of clone 3, were included in the panel of clones. The 14 selected clones seeded at 3×10^5 cells/mL in 75 mL of HyClone CDM4CHO (Thermo) supplemented with Penicillin/Streptomycin (Gibco), L-glutamine supplement (SAFC Biosciences) and MTX. The cells were grown in an orbital shaker at 36.5°C at a shaking speed of 125 rpm and 8.0% CO_2. Nine clones (1–3, 6–8 and 12–14) were grown in triplicates and the remaining five clones (4, 5, 9, 10 and 11) in single cultures bringing the total up to 32 cultures. The cultures were randomized prior to sampling in order to avoid bias.

2.2 Assays

Cell number and viability were measured using a Vicell XR system (Beckman Coulter). Metabolites were measured using a BioProfile 100 Plus (Nova Biopmedical). The activity of FVIII was measured as chromogenic activity using an in-house version of the Coatest SP (Chromogenix, Instrumentation Laboratory, Milano, Italy) [26]. Protein lysates for Western blotting were prepared by spinning down 10^6 cells and resuspending the pellet in 200 µL of Mammalian Protein Extraction Reagent (M-PER) (Thermo) following manufacturer's instructions. 30 µg of each sample was used for Western blotting. Gels were submitted to western blotting using Novex/NuPage blotting system (Invitrogen). Primary antibodies used were 1:500 dilution of sheep polyclonal anti-human factor VIII (CL20035AP, Cedarlane labs, Burlington, ON, Canada) and 1:1000 dilution of rabbit anti-beta-actin antibody (Cell Signaling Technology, Danvers, Ma, USA). Secondary antibodies used were 1:20 000 dilution of Donkey anti-Rabbit IRDye® 800 CW (LI-COR® Biosciences, Lincoln, Ne, USA) and 1:10 000 dilution of Alexa Fluor® 680 donkey anti-sheep IgG (Invitrogen, Carlsbad, Ca, USA). The ladders used were Full Range Rainbow™ recombinant protein molecular weight marker (Ge lifescience, Piscataway, NJ, USA).

2.3 RNA purification and next-generation sequencing

After 48 h of cultivation RNA was extracted from 2×10^6 cells using TRIzol (Invitrogen) the RNeasy Cleanup kit (Qiagen) following the manufacturer's instructions. RNA integrity was confirmed on an Agilent 2100 Bioanalyzer using total RNA nano chips (Agilent Technologies, Santa Clara, Ca, USA). RNA concentration was measured using a NanoDrop spectrophotometer (NanoDrop Technologies). Multiplexed cDNA library generation using the TruSeq RNA Sample Preparation Kit v2 (Illumina, Inc., San Diego, CA) and next-generation sequencing were performed by AROS Applied Biotechnology (Aarhus, Den-

mark) using eight samples per lane in an Illumina Hiseq 2000 system for paired-end sequencing. Clone 1, 8, 12, 13 and 14 were processed and analyzed by Cergentis (Utrecht, Netherlands) with their Targeted Locus Amplification (TLA) Technology [27] for targeted sequencing of the transgene and transgene integration site

2.4 Treatment of next-generation sequencing data

The FASTQC tool (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to evaluate the quality of the fastq files before and after treatment. The FASTX Toolkit (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to remove the adaptamers (fastx_trimmer) and trim the ends for basepairs with a quality score lower than 20 (fastq_quality_trimmer). An in house algorithm was used to intersect the read-pairs after quality trimming. The reads were aligned to the CHO-K1 genome (downloaded from Genbank as assembly GCF_000223135.1), combined with an extra scaffold containing the transgene, using tophat2 [28] and reads were counted using HTseq (http://www-huber.embl.de/users/anders/HTSeq). The samples were normalized using EdgeR [29] in R [30] and the CPM values given in EdgeR were normalized by gene length in order to calculate the FPKM-values. The Pearson correlation was calculated for expression level of *F8* and *dhfr* to the FVIII productivity defined as chromogenic activity measured at 48 h into the cultivation normalized by the viable cell count. RNA composition across the transgenes was measured by GenomeCoverageBed from BEDTools (version 2.16.2).

2.5 qRT-PCR

cDNA was produced from 1 µg RNA from each of the 32 samples sent to RNA sequencing using SuperScript III first-strand synthesis supermix (Invitrogen, Carlsbad, CA). Primers designed for the tripartite leader (transgene position 1119–1136 and 1193–1215), the *F8* ORF (transgene position 1676–1695 and 1752–1775), the *dhfr* ORF (transgene position 6750–6769 and 6886–6905) and *gapdh* (forward: AACTTTGGCATTGTGGAAGG and reverse: ACACGTTGGGGGTAGGAACA). The qRT-PCR reaction was run as 20 µL reaction using Quantifast SYBR green PCR Master Mix (QIAGEN, Germany) following manufacturer's instructions on a Stratagene MX3000P real-time PCR system (Stratagene). Primer efficiencies were calculated based on five consecutive five-fold dilutions of cDNA sample yielding efficiencies of 101% for *gapdh*, 107% for *dhfr*, 96% of *F8* and 93% for tripartite leader. The relative expression ratio for the tripartite leader, *F8* and *dhfr* was calculated compared to sample 1 from clone 1 as described elsewhere [31].

2.6 Genomic PCR

CHO DXB11 cells were thawed from an in-house master cell bank. The cells were passaged in alpha MEM media with 10% FBS, 1% NEAA, 1% P/S. Genomic DNA from this mother cell line and from clone 3, clone 7 and clone 8 was extracted from 2×10^6 cells using DNeasy Blood & Tissue Kit (Qiagen) following manufacturer's instructions. 100 ng of gDNA were used as template for a 25 µL PCR reaction using KOD Xtreme™ Hot Start DNA Polymerase (Milipore) following manufacturer's instructions. Primers were designed for *gapdh* (same as used for qRT-PCR) and for spanning from the transgene to the surround genome: PCR1 (forward primer aligning to transgene position 4694–4713 and reverse: GCAAAGAATGATCCCAGCTT), PCR2 (forward primer aligning to transgene position 4781–4800 and reverse: CCTTCCCTCCTCTCTTCCTG)

3 Results

3.1 RNA was extracted from exponentially growing cell 48 h after seeding

CHO DXB11 cells were transfected with a plasmid encoding human coagulation factor VIII (the *F8* gene) and a *dhfr* selection marker from a bicistronic mRNA. The pool of transfectants were amplified and subsequently sorted into single cells. A selection of CHO DXB11 clones were analyzed for FVIII productivity and growth rate (Fig. 1). From 62 clones, 14 clones exhibiting a dynamic range of FVIII productivities were selected for further analysis (Fig. 1). These clones were named clone 1–14 based on descending FVIII productivity. The three highest yielding clones (clone 1–3) were subclones of the same clonal cell line. Clone 12–14 did not produce detectable FVIII levels and were classified as nonproducers.

When seeded at 3×10^5 cells/mL, most clones entered the exponential growth phase within 27 h of cultivation and had not entered the stationary phase 73 h after seeding (Fig. 2A). The selected clones all had doubling times of 30–40 h. During cultivation, cell growth medium glucose and glutamine levels gradually decreased, while lactate levels increased (Fig. 2B). The glucose and glutamine levels dropped most rapidly in the medium of the faster growing clones (data not shown). For the FVIII producing clones, FVIII levels continuously increased during the first 100 h of cultivation (Fig. 2C) before stagnating or dropping. Based on these growth and production characteristics, we decided to extract RNA for analysis of the transcriptome after 48 h of cultivation. At this time point, all cultures were in the exponential growth phase, the growth medium was not depleted for glutamine, lactate levels were still modest, and the clones had released a dynamic range of FVIII to the growth medium.

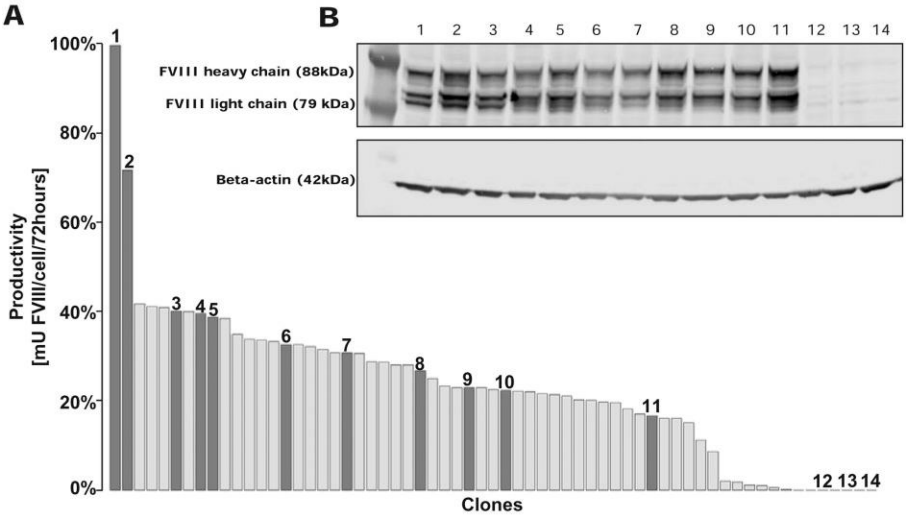


Figure 1. CHO DXB11 cells were transfected with a plasmid encoding FVIII and were subsequently single cell diluted. (**A**) The productivity of 62 CHO clones. Each bar represents an individual clone and clones numbered 1–14 were selected for further analysis. FVIII productivity was calculated as units recombinant FVIII (COA) released per cell per day, and is shown as percentage of the clone with highest productivity. (**B**) Western blot showing presence of FVIII protein in cell lysates from clone 1–14.

Following RNA extraction the samples were analyzed by next generation sequencing. The three highest FVIII producing cell lines, three of the medium producing cell lines and the three cell lines with no productivity of FVIII were analyzed in triplicate in order to assess the reproducibility. On average, 31.3 million read pairs were sequenced from each library. Hereof, 2–6% of the reads aligned to the exogenous *F8* and *dhfr* sequences, making *F8* and *dhfr* mRNA the most abundant transcripts in the present clones.

3.2 The transgene expression level did not correlate with amount of secreted protein

The expression level of *F8* and *dhfr* were quantified as sequenced fragments per kb of exon per million fragments mapped (FPKM), a normalized value allowing comparison of mRNA/cDNA levels from genes of different lengths and from RNA samples giving rise to different numbers of reads [32]. The clones analyzed in triplicate exhibited only little internal variation both with regards to the expression level of *F8* and *dhfr* and with regards to the FVIII productivity, demonstrating the reproducibility and robustness of the assays. For each of the FVIII producing clones (clone 1–11), the expression ratio of *F8* and *dhfr* was close to one (Fig. 3B). This is in agreement with the bicistronic mRNA expression strategy utilized in the present clones and described in greater detail below.

The expression of *F8* RNA was found not to correlate with the productivity neither as expressed as mU FVIII/cell/48 h (Pearson correlation = 0.03) nor as ng FVIII/cell/48 h (Pearson correlation = 0.24). In contrast the expression level of *dhfr* mRNA showed a small tendency to correlate negatively (Pearson correlation = –0.55) (Fig. 3B). Among the three nonproducing clones (12, 13, and 14), the *F8* and *dhfr* expression levels were highly variable. Interestingly, the *F8* RNA expression level of clones 12 and 13 were not reduced compared to the highest producing clones, whereas that of clone 14 was substantially reduced compared to all other clones (Fig. 3B). The expression level of *dhfr* in the nonproducing clones was similar to or higher than those of the FVIII producing clones (Fig. 3B). In conclusion, the levels of *F8*- or *dhfr*-coding mRNAs do not seem to determine FVIII productivity. The findings were validated using qRT-PCR making cDNA from the same RNA samples sent for RNA sequencing showing the same overall level of *F8* and *dhfr* mRNA among the clones (Fig. 3C).

3.3 The transgene was found to be truncated in a number of clones

In order to further characterize the mRNAs transcribed from the *F8* and *dhfr* expression cassette, the read depth distribution over the *F8-dhfr* expression cassette was analyzed for each of the 14 clones. Starting upstream, the present expression cassette comprises the adenovirus-2 major late promoter, the adenovirus-2 late mRNA tripartite leader, a 3' end and 5' end intron sequence with splice junction, the ORF of *F8* with a minimal artificial B-domain [23], an internal ribosomal entry site (IRES), the *dhfr* ORF, and a poly A signal, similar to constructs used elsewhere [24, 25] (Fig. 3A). The read depth distribution was essentially identical for the three highest yielding clones (clone 1–3). The entire *F8* and *dhfr* ORFs and the connecting IRES element were transcribed (Fig. 3D top). At the 5' end, the tripartite leader and the UTR upstream of the *F8* ORF, but not the separating intron sequences were transcribed, suggesting that the intron sequences were indeed removed from the nascent mRNA by splicing (Fig. 3D top). Likewise, the read depth distribution was very similar among the eight lower producing clones (clone 4–11), but differed from the three highest producing clones, as the tripartite leader did not appear to be

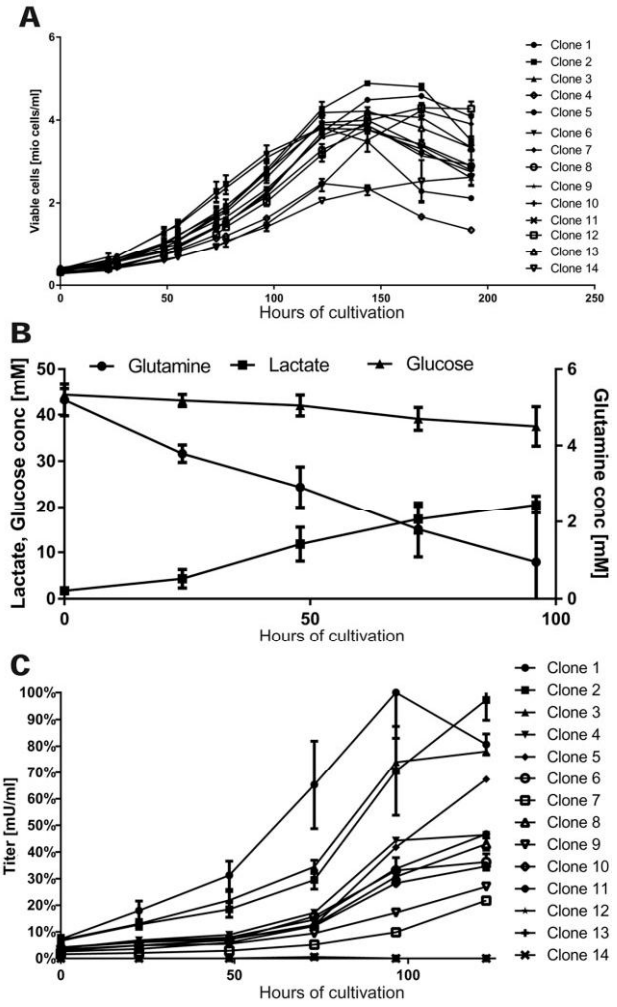


Figure 2. Growth and FVIII production of 14 selected CHO DXB11 clones. (**A**) Cultures were set up seeded at 3×10^5 cells/mL and monitored during cultivation for viable cell count. Error bars indicate standard deviation of biological replicates. (**B**) The glutamine, glucose, and lactate levels found in the growth medium (data from all cultures collapsed into one graph). Error bars indicate standard deviation from cultures from all 14 clones. (**C**) Production titer measured as units of active FVIII in the growth medium relative to maximum. Error bars indicate standard deviation of biological replicates.

transcribed (Fig. 3D). The read depth distribution of the three nonproducing clones (clones 12–14) explained the failure of these clones to produce FVIII. Clone 12 and 13 only partially transcribed the *F8* ORF, as an upstream portion containing the initiation codon was not transcribed. Clone 14 did not appear to transcribe any part of the *F8* ORF (Fig. 3D). In contrast, the *dhfr* ORF was transcribed in all 3 clones, explaining the capacity of the nonproducing clones to survive MTX selection.

In order to validate the truncations suggested by RNA sequencing, the transgene regions integrated into the genome of clone 1, 8, 12, 13 and 14 were sequenced using TLA. It was possible to identify a specific insertion locus

Table 1. Detected ranges of transgene in RNAseq data and by targeted sequencing

Name	RNAseq	DNAseq	Chromosome
Clone 1	303–7347	1–7776	2[a)]
Clone 8	1507–7327	1313–8262	X
Clone 12	2717–8185	2705–8261	7[a)]
Clone 13	3897–7405	3854–7643	5[a)]
Clone 14	6387–7285	583–1055, 6388–7646, 4382–4549, 8757–8900	5

a) A certain region is most likely the insertion locus but no breakpoint reads in an individual scaffold in the (incomplete) CHO-K1 reference genome was identified.

on the CHO-K1 reference genome for two of the five clones. The fragments found in the genome corresponded well with the results found by RNA sequencing (Table 1). Clone 1 was shown to have all of the expected sequence inserted. In the genome of clone 14 472 bp of the promoter region was found flanking the entire *dhfr* ORF thus explaining the transcription of the truncated transcript from the promoter present on the plasmid. Clone 8 was found to be devoid of promoter and tripartite sequence. Surprisingly, at the locus found by targeted sequencing to be the insertion site of the transgene, RNA sequencing data found expression from only clone 4–11 (Fig. 4A). The region downstream of the insertion locus was transcribed, but the expression level rapidly dropped to zero right at the suggested insertion site indicating that this region of the genome is the promoter region for the inserted transgene. Using genomic PCR it was found that the transgene had indeed integrated into this region at the same position in clone 7 and clone 8 (Fig. 4B).

4 Discussion

In the present study, we describe deep sequencing analysis of mRNA transcribed from the exogenous *F8* and *dhfr* expression cassette of clonal CHO cell lines exhibiting a dynamic range of recombinant FVIII productivities. The expression vector allows the synthesis of a single bicistronic mRNA comprising the adenovirus-2 late mRNA tripartite leader followed by both the *F8* and the *dhfr* coding ORFs separated by an IRES element, allowing translation of both ORFs from the same mRNA. In the present clones, the expression ratio of *F8* and *dhfr* RNAs was close to 1, suggesting that the *F8* and *dhfr* expression is indeed coupled. For comparison, *dhfr* transcripts were more abundant than transcripts from the gene of interest in a previous next-generation sequencing study on a BHK recombinant protein production cell line [21], suggesting that in the latter cell line, expression of *dhfr* and the gene of interest were less strictly coupled.

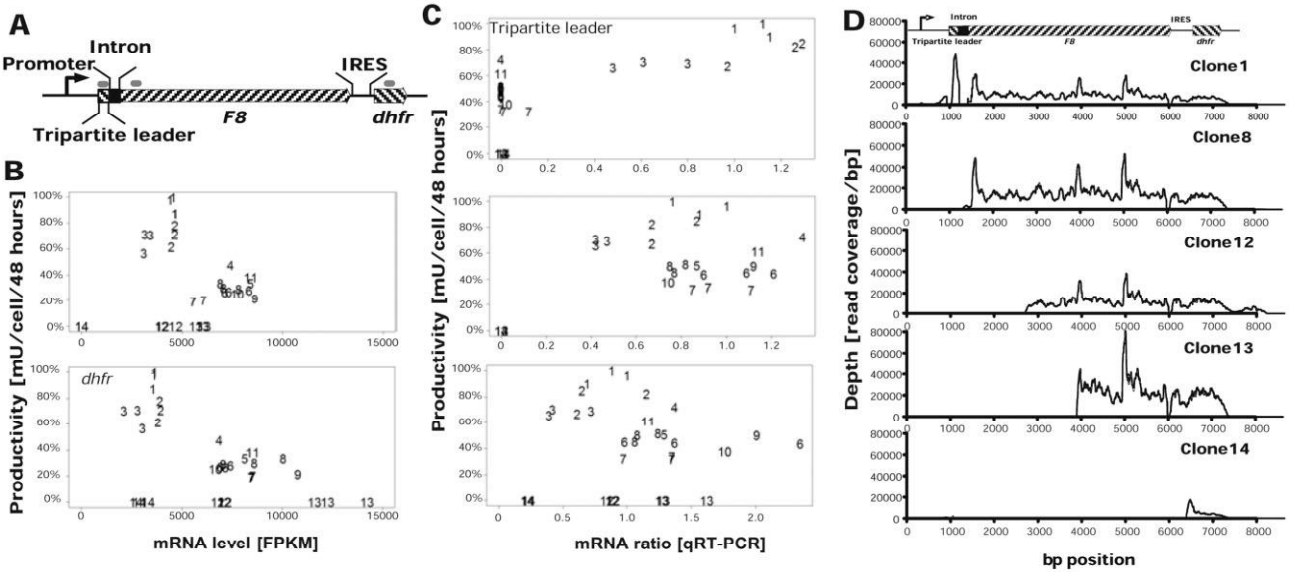


Figure 3. *F8* and *dhfr* mRNA levels across the 14 analyzed clones. (**A**) Overview of transgene composition. qPCR primer targets used in (**C**) are marked with grey lines. (**B**) *F8* and *dhfr* expression level based on RNA sequencing in the CHO DXB11 clones (1–14) versus productivity. FVIII productivity is calculated as units recombinant FVIII released per cell within the first 48 h of cultivation, and is shown as percentage of the clone with highest productivity. (**C**) Tripartite leader, *F8* and *dhfr* expression level based on qRT-PCR in the CHO DXB11 clones (1–14) versus productivity. (**D**) Read depth distribution over each basepair of the *F8-dhfr* expression cassette on the transfected plasmid. Typical results show for each group. From the top: clone 1 (high FVIII producing cell line), clone 8 (medium FVIII producing cell line), clone 12 (nonproducing cell line), clone 13 (nonproducing cell line) and clone 14 (nonproducing cell line). Representative clones for the various RNA signatures shown.

Among the clones analyzed in the present study, the level of *F8-dhfr* coding mRNAs did not determine FVIII productivity. The clones were all adapted to growth in the presence of MTX. In our hands, amplification using MTX is required to reach the optimal FVIII productivity, so the level of *F8* coding mRNA obviously plays a role for FVIII productivity, just as seen with antibodies [33]. In the present study however, the *F8-dhfr* mRNA level required for growth in the presence of MTX most likely exceeded the level required for saturating the FVIII production machinery of the highest yielding clones. Instead, the composition of the mRNA molecules transcribed from the *F8-dhfr* expression cassette appeared to play a key role in determining the FVIII productivities of the clones. The read depth distribution over the *F8-dhfr* expression cassette describes the number of reads aligning with the individual nucleotides in the expression cassette. Thus, the read depth shows the composition of the *F8-dhfr* coding mRNAs based on thousands of mRNA molecules from each clone. In clones not producing detectable FVIII, the mRNA transcribed from the *F8-dhfr* coding expression comprised the entire *dhfr* ORF, but a 5' end fragment of the *F8* ORF or the entire *F8* ORF was missing. This explains the capacity of these cells to grow in the presence of MTX without producing FVIII.

The Targeted Locus Amplification Technology has previously been used to elucidate transgene insertion loci and transgene composition [27]. In the present study TLA analysis were able to provide an explanation for the expression of truncated *F8-dhfr* in clone 14 as a fragment of the promoter had been integrated on chromosome five flanked by *dhfr* leading to expression of a truncated transcript. In the case of clone 8, 12 and 13 no fragments of the promoter was detected thus suggesting that an endogenous promoter was used to induce transcription. In the case of clone 8, it was seen that the region flanking the insertion site is expressed and that the expression level drop to zero at exactly the position suggested as the transgene insertion locus. Thus it seems that this region, which appears to be silent in the other clones, is able to induce transcription in order to allow the cells to survive under selection. Peculiarly, it is seen that all eight lower producing clones show the same signature from this region indicating that all these clones are subclones from a single cell, which inserted the truncated transgene into this particular locus. A similar mechanism explaining the transcription of truncated *F8-dhfr* in clone 12 and 13 is still to be elucidated but it was found that at the suggested insertion regions these clones showed expression profiles different from all other clones (data not shown), which could indicate a similar mechanism as in the case of clone 8.

Whether the plasmid was truncated prior to integration or during integration is not possible to tell. Earlier reports have found transgenes can undergo homologous but nonconservative recombination resulting in trunca-

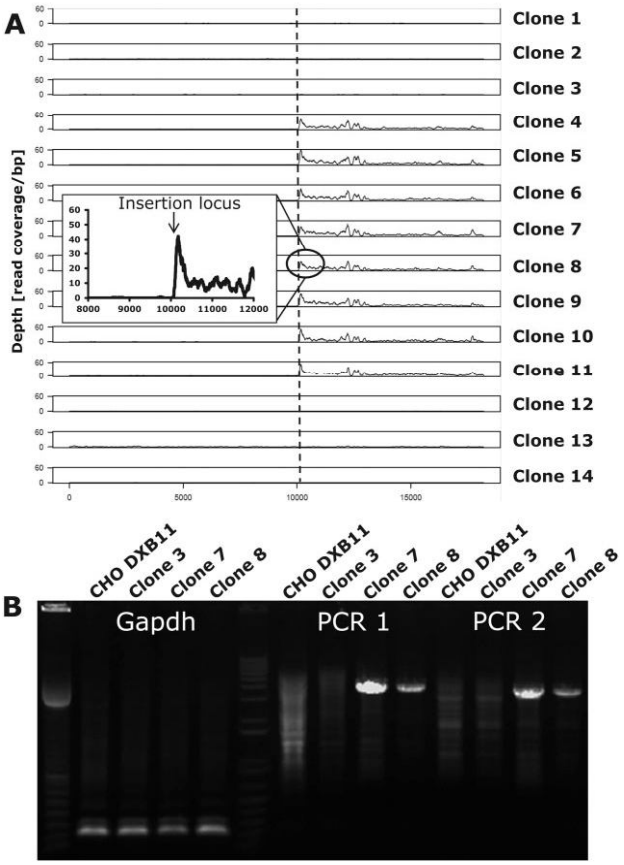


Figure 4. The transgene integration site of clone 8. (**A**) Read depth distribution for each base pair of the 20 kb region surrounding the suggested insertion site of clone 8 (vertical line) on scaffold NW_003615608.1 position 140838 in RNA sequencing data for all 14 clones. (**B**) Genomic PCR amplifying a 222 bp region of *gapdh* as positive control in CHO DXB11, clone 3, clone 7 and clone 8. PCR 1 and 2 both span from the transgene to flanking genomic region.

tions followed by illegitimate DNA integration by a DNA repair-mediated integration process into the host genome [34–37]. The most important aspect in relation to the data at hand is the consequence of only applying selection pressure for one specific part of the transgene, namely transcription of the *dhfr* ORF. The data show that in the case the cells are in any way able to produce the selection marker without the transgene there is a clear selection pressure favoring these, thus a dual selection system conferring selection for both the region downstream and upstream of the *F8* ORF could reduce the probability of encountering instances such as the one described here.

Among the FVIII producing clones, the three highest yielding clones distinguished themselves from the lower yielding clones by the presence of *F8-dhfr* coding mRNA with the adenovirus-2 late mRNA tripartite leader. The tripartite leader is reported to optimize the translation efficiency by translation independent from CAP-binding proteins [38, 39], and the present findings suggest that the tripartite leader can indeed increase the yields of recombinant FVIII translated from a lower level of *F8*-mRNA (Fig. 3B). It has been shown that introduction of the tripartite leader upstream of firefly luciferase or interferon gamma in transiently transfected CHO-K1 cells lead to an increase in protein productivity of 3.6 and 7.6-fold respectively although it decreased the productivity of an antibody in the same cell line indicating that the impact of addition of the tripartite leader is gene-specific [40].

The current results suggest that development of mammalian production cell lines may be facilitated by testing for the presence of the expected 5' mRNA end encoding the gene of interest (as in Fig. 3C top). Cell line development typically involves the generation of nonclonal pools of transfected cells that are resistant to selection and cloning of one or more pools. When choosing among several cell pools, we often choose to clone the pool with the highest yield of our protein of interest. However, the yield of the cell pool does not give an impression on the distribution of the yields among the individual clones in the pool. Testing cell pool mRNA for a probe annealing to the expected transcriptional start site, may provide a better basis for choosing cell pools containing the highest yielding clones as it may indicate proper integration into the genome without extensive truncation of the transfected plasmid.

In conclusion, deep sequencing is an extremely powerful and robust method for quantification and analysis of the mRNAs transcribed from the gene of interest and the selection marker inserted in cell lines for production of recombinant proteins. The readout is based on several thousand mRNA molecules transcribed from the same expression cassette, giving a detailed and highly representative picture of the mRNA population revealing the transcriptional start end termination site as well as intron splicing. Using RNA sequencing in contrast to microarrays thus allow for an insight into the use of the plasmid used for heterologous expression in addition to information regarding the general transcriptome.

The authors would like to thank laboratory technician Gedske Thygesen for producing the 59 CHO clones with different productivities of FVIII and Dr. Hanni Willenbrock Thomsen for assistance with the RNA sequencing data treatment. The work was funded by Novo Nordisk A/S. CSK: Carried out experimental work, data analysis and drafted the manuscript. GB: Drafted the manuscript and contributed to the data analysis. JJH, MRA and CK: contributed to the experimental design stage, data analysis and manuscript formulation. All authors read and approved the final manuscript.

CSK, GB, JJH, and CK are employees of Novo Nordisk A/S. MRA has no financial or commercial conflicts of interest.

5 References

[1] Kumar, R., Carcao, M., Inherited abnormalities of coagulation: Hemophilia, von Willebrand disease, and beyond. *Pediatr. Clin. North Am.* 2013, *60*, 1419–1441.
[2] Pipe, S. W., Recombinant clotting factors. *Thromb. Haemost.* 2008, *99*, 840–850.
[3] Kaufman, R. J., Wasley, L. C., Dorner, A. J., Synthesis, processing, and secretion of recombinant human factor VIII expressed in mammalian cells. *J. Biol. Chem.* 1988, *263*, 6352–6362.
[4] Lenting, P. J., van Mourik, J. A., Mertens, K., The life cycle of coagulation factor VIII in view of its structure and function. *Blood* 1998, *92*, 3983–3996.
[5] Andersson, L. O., Forsman, N., Huang, K., Larsen, K. et al., Isolation and characterization of human factor VIII: molecular forms in commercial factor VIII concentrate, cryoprecipitate, and plasma. *Proc. Natl. Acad. Sci. U.S.A.* 1986, *83*, 2979–2983.
[6] Pipe, S. W., Functional roles of the factor VIII B domain. *Haemophilia* 2009, *15*, 1187–1196.
[7] Burke, R. L., Pachl, C., Quiroga, M., Rosenberg, S. et al., The functional domains of coagulation factor VIII:C. *J. Biol. Chem.* 1986, *261*, 12574–12578.
[8] Pittman, D. D., Alderman, E. M., Tomkinson, K. N., Wang, J. H. et al., Biochemical, immunological, and in vivo functional characterization of B-domain-deleted factor VIII. *Blood* 1993, *81*, 2925–2935.
[9] Toole, J. J., Pittman, D. D., Orr, E. C., Murtha, P. et al., A large region (approximately equal to 95 kDa) of human factor VIII is dispensable for in vitro procoagulant activity. *Proc. Natl. Acad. Sci. U.S.A.* 1986, *83*, 5939–5942.
[10] Wurm, F. M., Production of recombinant protein therapeutics in cultivated mammalian cells. *Nat. Biotechnol.* 2004, *22*, 1393–1398.
[11] Fernandez, J., Yaman, I., Huang, C., Liu, H. et al., Ribosome stalling regulates IRES-mediated translation in eukaryotes, a parallel to prokaryotic attenuation. *Mol. Cell* 2005, *17*, 405–416.
[12] Davies, M. V., Kaufman, R. J., Internal translation initiation in the design of improved expression vectors. *Curr. Opin. Biotechnol.* 1992, *3*, 512–517.
[13] Clarke, C., Doolan, P., Barron, N., Meleady, P. et al., Large scale microarray profiling and coexpression network analysis of CHO cells identifies transcriptional modules associated with growth and productivity. *J. Biotechnol.* 2011, *155*, 350–359.
[14] Seth, G., Philp, R. J., Lau, A., Jiun, K. Y. et al., Molecular portrait of high productivity in recombinant NS0 cells. *Biotechnol. Bioeng.* 2007, *97*, 933–951.
[15] Trummer, E., Ernst, W., Hesse, F., Schriebl, K. et al., Transcriptional profiling of phenotypically different Epo-Fc expressing CHO clones by cross-species microarray analysis. *Biotechnol. J.* 2008, *3*, 924–937.
[16] Lewis, N. E., Liu, X., Li, Y., Nagarajan, H. et al., Genomic landscapes of Chinese hamster ovary cell lines as revealed by the *Cricetulus griseus* draft genome. *Nat. Biotechnol.* 2013, *31*, 759–765.
[17] Xu, X., Nagarajan, H., Lewis, N. E., Pan, S. et al., The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. *Nat. Biotechnol.* 2011, *29*, 735–741.
[18] Kaas, C. S., Kristensen, C., Betenbaugh, M. J., Andersen, M. R., Sequencing the CHO DXB11 genome reveals regional variations in genomic stability and haploidy. *BMC Genomics* 2015, *16*, 160.
[19] Kaas, C. S., Fan, Y., Weilguny, D., Kristensen, C. et al., Towards genome-scale-models of the Chinese hamster ovary cells: Incentives, status, and perspectives. *Pharm. Bioprocess.* 2014, *2*, 437–448.
[20] Birzele, F., Schaub, J., Rust, W., Clemens, C. et al., Into the unknown: Expression profiling without genome sequence information in CHO by next generation sequencing. *Nucleic Acids Res.* 2010, *38*, 3999–4010.
[21] Johnson, K. C., Yongky, A., Vishwanathan, N., Jacob, N. M. et al., Exploring the transcriptome space of a recombinant BHK cell line through next generation sequencing. *Biotechnol. Bioeng.* 2014, *111*, 770–781.
[22] Zeck, A., Regula, J. T., Larraillet, V., Mautz, B. et al., Low level sequence variant analysis of recombinant proteins: An optimized approach. *PLoS One* 2012, *7*, 40328.
[23] Thim, L., Vandahl, B., Karlsson, J., Klausen, N. K. et al., Purification and characterization of a new recombinant factor VIII (N8). *Haemophilia* 2010, *16*, 349–359.
[24] Connelly, S., Smith, T. A., Dhir, G., Gardner, J. M. et al., In vivo gene delivery and expression of physiological levels of functional human factor VIII in mice. *Hum. Gene Ther.* 1995, *6*, 185–193.
[25] Orlova, N. A., Kovnir, S. V., Vorobiev, I. I., Yuriev, A. S. et al., Stable expression of recombinant factor VIII in CHO cells using methotrexate-driven transgene amplification. *Acta Naturae* 2012, *4*, 93–100.
[26] Ovlisen, K., Kristensen, A. T., Valentino, L. A., Hakobyan, N. et al., Hemostatic effect of recombinant factor VIIa, NN1731 and recombinant factor VIII on needle-induced joint bleeding in hemophilia A mice. *J. Thromb. Haemos.* 2008, *6*, 969–975.
[27] de Vree, P. J. P., de Wit, E., Yilmaz, M., van de Heijning, M. et al., Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping. *Nat. Biotechnol.* 2014, *32*, 1019–1025.
[28] Trapnell, C., Pachter, L., Salzberg, S. L., TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* 2009, *25*, 1105–1111.
[29] Robinson, M. D., McCarthy, D. J., Smyth, G. K., edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010, *26*, 139–140.
[30] Ihaka, R., Gentleman, R., R: A language for data analysis and graphics. *J. Comput. Graph. Stat.* 1996, *5*, 299–314.
[31] Kubista, M., Andrade, J. M., Bengtsson, M., Forootan, A. et al., The real-time polymerase chain reaction. *Molecular aspects of medicine* 2006, *27*, 95–125.
[32] Trapnell, C., Roberts, A., Goff, L., Pertea, G. et al., Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 2012, *7*, 562–578.
[33] Kim, S. J., Kim, N. S., Ryu, C. J., Hong, H. J. et al., Characterization of chimeric antibody producing CHO cells in the course of dihydrofolate reductase-mediated gene amplification and their stability in the absence of selective pressure. *Biotechnol. Bioeng.* 1998, *58*, 73–84.
[34] Bishop, J. O., Smith, P., Mechanism of chromosomal integration of microinjected DNA. *Mol. Biol. Med.* 1989, *6*, 283–98.
[35] Calos, M. P., Lebkowski, J. S., Botchan, M. R., High mutation frequency in DNA transfected into mammalian cells. *Proc. Nat.l. Acad. Sci. U.S.A.* 1983, *80*, 3015–3019.
[36] Würtele, H., Little, K. C. E. & Chartrand, P., Illegitimate DNA integration in mammalian cells. *Gene Ther.* 2003, *10*, 1791–1799.
[37] Yan, B. W., Zhao, Y. F., Cao W. G., Li, N. et al., Mechanism of random integration of foreign DNA in transgenic mice. *Transgenic Res.* 2013, *22*, 983–992.
[38] Dolph, P. J., Huang, J. T., Schneider, R. J., Translation by the adenovirus tripartite leader: Elements which determine independence from cap-binding protein complex. *J. Virol.* 1990, *64*, 2669–2677.
[39] Logan, J., Shenk, T., Adenovirus tripartite leader sequence enhances translation of mRNAs late after infection. *Proc. Nat.l. Acad. Sci. U.S.A.* 1984, *81*, 3655–3659.
[40] Ho, S. C. L., Yap, M. G. S., Yang, Y., Evaluating post-transcriptional regulatory elements for enhancing transient gene expression levels in CHO K1 and HEK293 cells. *Protein Expression Purif.* 2010, *69*, 9–15.

3.2 Composition of the transgene composition in the genome of Clone 1

3.2.1 Introduction

Following transfection of a transgene into CHO cells the gene can be incorporated into the genome. By applying selection pressure over longer periods of time only the cells having incorporated the transgene and actively transcribing the selection marker are able to survive. In the case of the DHFR selection system, MTX addition can be used to increase the cellular need for DHFR inducing higher transcription levels of the transgene and the selection marker. This is mostly achieved by the cell through amplification of the transgene copy number in the genome [1]. Here we present an analysis in the composition and location of the *F8* transgene in the Clone 1 genome, which is a CHO DXB11 cell line, which after transfection underwent amplification with MTX.

3.2.2 Results and Discussion

In the pursuit of identifying the location of the inserted transgene in the Clone 1 genome (also referred to as F435 in a recent paper [2]) the raw WGS reads were aligned to the CHO-K1 ATCC genome and subsequently the alignment file was searched for break-spanning reads, which are paired reads aligning to the transgene with one pair and on the genome with the other pair.

Table 1 List of scaffolds found to have break-spanning reads aligning to the transgene and the genome.

Ranking	Reads	Scaffold name	Ranking	Reads	Scaffold name
1	260	NW_003613906.1	6	3	NW_003613600.1
2	34	NW_003614308.1	7	3	NW_003613653.1
3	5	NW_003613609.1	8	3	NW_003613658.1
4	4	NW_003613585.1	9	3	NW_003613679.1
5	3	NW_003613583.1	10	3	NW_003613683.1

In Table 1 it is seen that scaffold NW_003613906.1, found on the *C. griseus* chromosome 2, holds the largest number of break-spanning reads and was thus investigated more thoroughly.

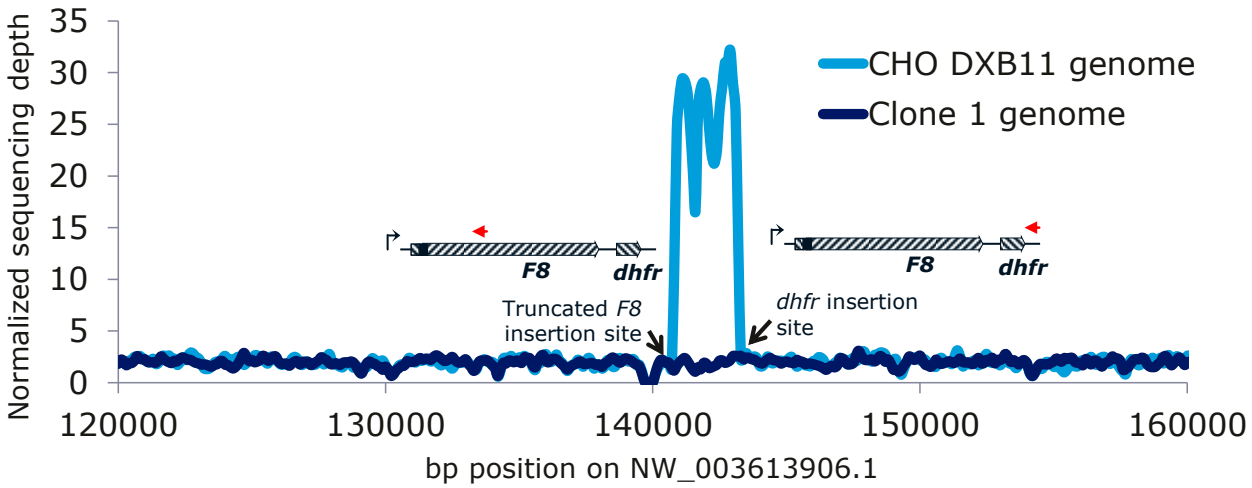


Figure 1 Sequencing depth for 40kb flanking the two suggested insertion locations.
It is seen that the 2.5kb flanking the insertion are found in much higher depth indicating amplification of this region.

From the depth distribution (Figure 1) it is seen that reads aligning to a 2.5kb region on NW_003613906.1 (pos 140800-143300) are vastly overrepresented in sequencing results from the Clone 1 genome compared to the CHO DXB11 genome with a median depth of 27.7 in this region compared to 2.0 in the rest of the scaffold. The 260 break-spanning reads found to align to this scaffold was *de novo* assembled and the contigs were searched against NW_003613906.1 and the transgene sequence. One contig revealed the transition from position 7664 on the plasmid on the minus strand into pos 143061 on NW_003613906.1 indicating that 114bp of the plasmid had been trimmed away from the site used for linearization compared to what was found in the genome. Another contig surprisingly revealed the transition from 1.2kb into the FVIII CDS onto the genome at pos 140979 indicating the integration of a truncated plasmid only containing promoter and 1.2kb of the FVIII CDS.

Copy number analysis of the transgene

In order to get a better picture of the transgene, the depth of coverage was calculated for *F8* and *dhfr* transgene regions (Table 2).

Table 2 The normalized read count for *F8* and *dhfr* listing the minimum, maximum, median, mean and standard deviation of read depth measured spanning their CDS

	min	max	median	mean	st.dev	st.dev/mean
F8	26.8	103.6	55.6	63.1	17.4	27.6%
dhfr	82.6	103.6	92.0	92.2	5.4	5.8%

Due to the large standard deviation for the depth of the *F8* coding region the entire region was plotted and it was seen that the first 1,200bp of the *F8* CDS had a copy number of 90 whereas the rest of the *F8* CDS only have a copy number of 54 (Figure 2A). All of *dhfr* has a copy number of 92. The mRNA depth distribution is seen to be uniform over the entire transcript, which indicates that both *F8* and *dhfr* are only expressed in the form covering the entire transcript (Figure 2B). Thus, the extra copies of *dhfr* and the start of *F8* either transcriptionally silent or degraded quickly. The insertion event described in the previous section showed that the left insertion was from 2.8kb into the plasmid on the minus strand, which then explain why the first 1.2kb of FVIII are found in a higher copy number than the rest of the gene. As to *dhfr* we know that the insertion on the right side of the 2.5kb span of genomic sequence held the *dhfr* sequence and this is needed to be transcribed in order to select by MTX inducing amplification of this region of the genome. This consequently indicates that full length transgene are also present on the left side of the amplified region. The remaining copies might be present next to the truncated *F8* gene as they are found in almost identical number of copies. It could be speculated that the second insertion 2.5kb from the full length transgene might aid in some way at keeping the region open and e.g. function as promoter decoys which the cell can methylate without affecting FVIII production.

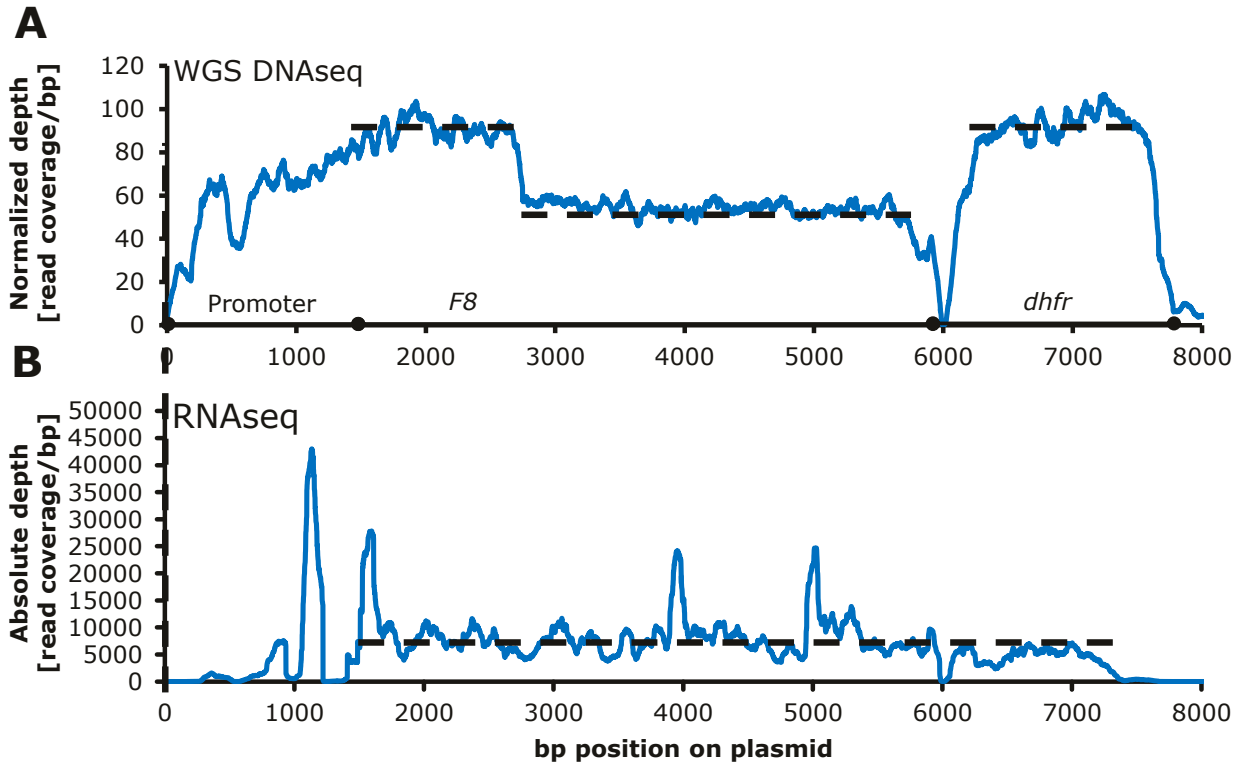


Figure 2 Sequencing depth for the transgene sequence
A) Depth measured from WGS genome sequencing data from Clone 1 and **B)** from RNAseq data

Targeted sequencing of the transgene

In combination with the work published earlier [3] Clone 1 was sent for targeted sequencing [4]. The coverage of the transgene is visualized in Figure 3A. Two libraries were made enriching for fragments centered on position 5110 and another on position 6760, in both cases the enrichment of these positions in the sequencing data is apparent (Figure 3B). The two libraries both validated the integration from pos 1-7776 seen from WGS (with a small exception for coverage of the region 5707-6476 due to a technical bias).

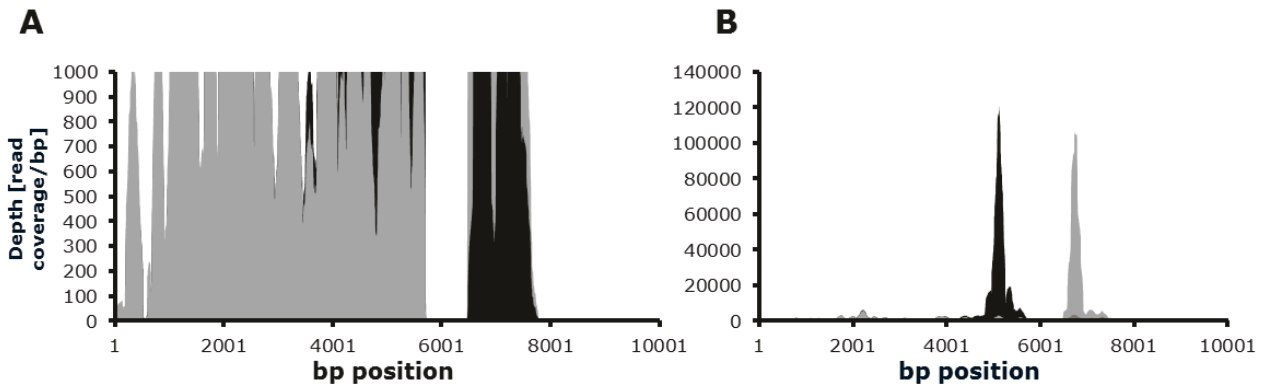


Figure 3 Sequencing depth for the transgene sequence in the targeted sequencing data of Clone 1.
A) sequencing reads are seen to cover pos 1-7776 of the transgene. **B)** Two libraries were made primed for sequence from position 5110 and another on position 6760 and reads covering these positions are seen to be vastly overrepresented compared to reads aligning to the rest of the transgene. Library A shown in black and library B shown in grey.

As to the insertion site, the targeted sequencing data detected the same insertion site as by WGS, but due to the nature of the TLA targeted sequencing method, it would be expected that the surrounding region of the scaffold would be heavily sequenced, as it is found flanking the primed DNA [4]. In contrast, only the 2.5 kb found in between the two inserts are seen in the targeted sequencing data and no reads align to other regions of the scaffold (Figure 4A). From each of the two libraries approximately 350 mio reads were sequenced, allowing accurate investigation of the transgene and the suggested insertion site with 600 times less sequenced reads as when used for WGS. SNP detection can be carried out by high accuracy due to the very high coverage of the transgene in targeted sequencing compared to WGS, but only pos 710 and 878 of the promoter region was found to contain mutations. It has recently been shown by Zhang et al [5] that using standard SNP detection tools on RNAseq data, that they were able to detect point mutations in the IgG transgene of clones and thus use RNAseq data for cell line screening early in the process.

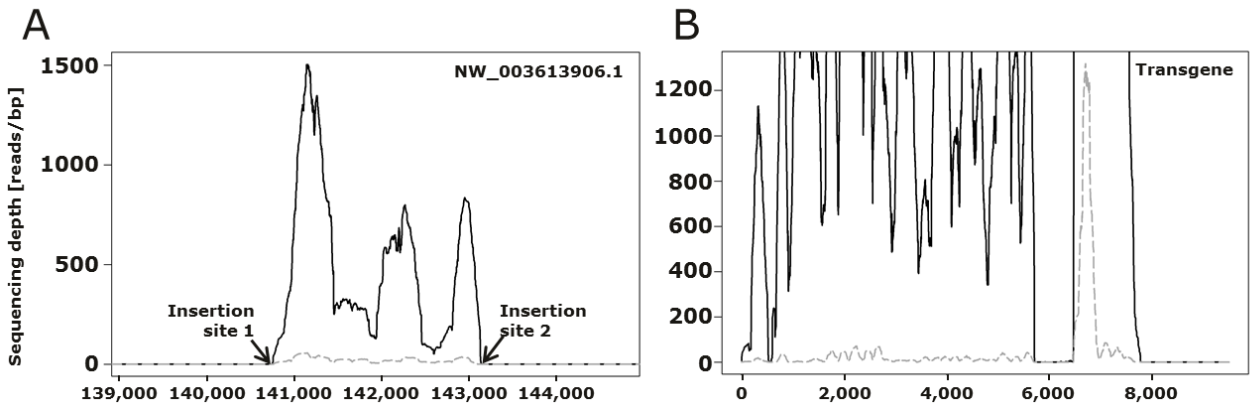


Figure 4 Sequencing depth on scaffold NW_003613906.1 using targeted sequencing data
A) Depth measured at scaffold NW_003613906.1 and B) transgene. Reads aligning to position are shown in black and break-spanning reads aligning to position shown in dashed grey.

The transcriptomics landscape of the suggested insertion site

The area suggested as integration site is found in an intron between exon 7 and 8 of the gene LOC100774471 (sodium- and chloride-dependent glycine transporter 1-like), which is a transcriptionally active gene (Figure 5B). Looking specifically into transcription of each of the exons no difference is seen in the transcription of the gene in Clone 1 compared to other FVIII clones. If the transgene had been inserted and amplified 30 times in one location the combined length would be more than 600 kb and would thus expect to impact the transcription level from exon 8-13 compared to 1-7.

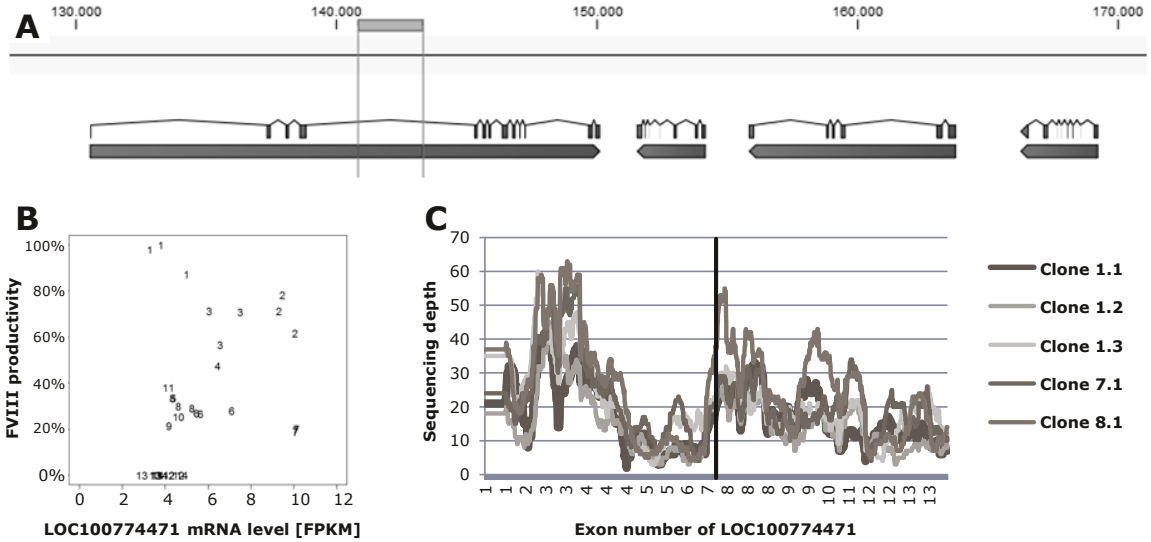


Figure 5 Transcription of the gene harboring the suggested integration site.
A) The site is found in the gene: LOC100774471 between exon 7 and 8. **B)** The gene is not differentially expressed in the Clone 1 and the **C)** The expression of each exon is similar in Clone 1 compared to other FVIII clones with no differences seen in the signature after exon 7.

3.2.3 Conclusion

Based on the knowledge gained from WGS, mRNAseq and targeted sequencing data, the most likely explanation is that the 2.5kb fragment from scaffold NW_003613906.1 was integrated into the genome somewhere else with transgenes flanking it: one side with a truncated and a complete transgene and one complete on the other side (see Figure 6). After insertion, selection by MTX induced amplification of the region in the genome. Due to the stoichiometric relationship of the model it would be expected to find a sequencing depth of 30x for the NW_003613906.1 fragment, 60x for the majority of the *F8* transgene and 90x for the start of the transgene and *dhfr*, which correspond quite well with the depths measured. As the targeted sequencing data did not yield any reads aligning to the NW_003613906.1 scaffold outside the 2.5 kb region and the transcriptome did not show any sign of disruption in Clone 1 it would appear that the insertion has occurred somewhere else. Due to the short fragments sequenced using the Illumina technology it is not possible to conclude more at the present time, but sequencing with PacBio or southern blots might provide further insights. SNP detection did not prove to be relevant in this given case, but all three omics technologies allow for such analysis. Though, SNP detection by WGS require high depth over the entire genome and would thus be the most expensive option of the three to answer that question.

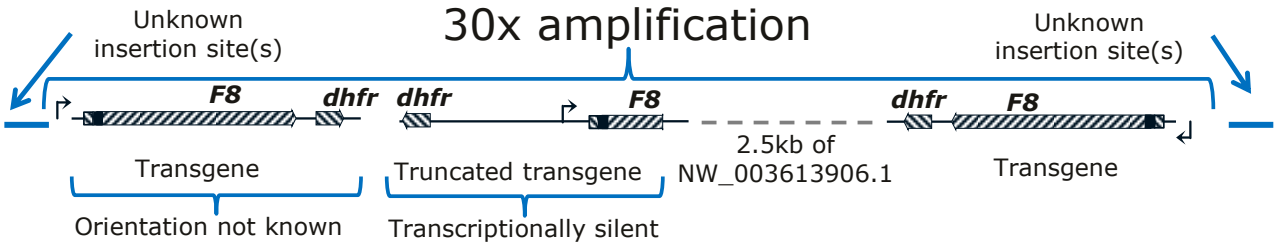


Figure 6 a model of the transgene composition in the Clone 1 genome

3.2.4 Materials and methods

Deduction of sequencing depth

Genomic DNA was extracted from Clone 1 cells, sequenced and aligned to the CHO-K1 ATCC genome as described elsewhere [2]. From the alignment bam file reads were extracted aligning to the heterologous sequence and CHO scaffolds by a custom script. The depth of sequencing coverage was measured for the transgene and the NW_003613906.1 and normalized as described elsewhere [2]. Break-spanning reads were assembled *de novo* into contigs using CLC genomic workbench (CLC Bio, version 7.5) under standard settings. The contigs were exported and blasted against the transgene sequence and CHO genome sequence using BLAST+ [6].

Targeted sequencing

Targeted sequencing of the transgene in the Clone 1 genome was carried out as described elsewhere [3]. The reads were aligned to the CHO-K1 ATCC genome and break-spanning reads were extracted. The sequencing depth was measured using genomeCoverageBed from BEDTools (version 2.16.2) measuring both the depth for all aligned reads and for a subset only containing break-spanning reads.

RNA sequencing

RNA was extracted from Clone 1 cells in the exponential growth phase, sequenced and aligned to the CHO-K1 ATCC genome out as described elsewhere [3]. The location of the suggested insertion site was visualized using CLC genomic workbench (CLC Bio, version 7.5). The depth of sequencing was measured using genomeCoverageBed from BEDTools (version 2.16.2) specifically for the exon sequences and visualized using Microsoft Excel.

3.2.5 References

1. Urlaub G, Chasin L: **Isolation of Chinese hamster cell mutants deficient in dihydrofolate reductase activity.** *Proceedings of the National Academy of Sciences of the United States of America* 1980, **77:** 4216-4220.

2. Kaas CS, Kristensen C, Betenbaugh MJ, Andersen MR: **Sequencing the CHO DXB11 genome reveals regional variations in genomic stability and haploidy.** *BMC genomics* 2015, **16:** 1391.

3. Kaas CS, Bolt B, Hansen JJ, Andersen MR, Kristensen C: **Deep sequencing reveals different compositions of mRNA transcribed from the FVIII gene in a panel of FVIII producing CHO cell lines.** *Biotechnology journal* 2015.

4. de Vree PJP, de Wit E, Yilmaz M, van de Heijning M, Klous P, Verstegen MJAM *et al.*: **Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping.** *Nature biotechnology* 2014, **32:** 1019-1025.

5. Zhang S, Bartkowiak L, Nabiswa B, Mishra P, Fann J, Ouellette D *et al.*: **Identifying low-level sequence variants via next generation sequencing to aid stable CHO cell line screening.** *Biotechnology progress* 2015.

6. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K *et al.*: **BLAST+: architecture and applications.** *BMC bioinformatics* 2009, **10:** 421.

3.3 Semi-stable expression of transgenes in the Icosagen system

3.3.1 Introduction and Background

In order to study targets for cell line engineering, the gene of interest are generally co-expressed in a cell line to investigate the impact on the productivity or product quality. This can be achieved either by 1) transiently expressing one or several genes until the plasmids containing the genes are diluted out of the cellular population or 2) spending months generating a stable cell line with the target gene(s) stably integrated into the genome. In order to introduce an alternative, the QMCF system from Icosagen promises to generate a pool of semi-stable cells expressing a gene of interest within two weeks of transfection [1]. The system rely on the mouse polyomavirus large T (LT) antigen to initiate replication of the plasmid containing the transgene and the EBNA-1 protein (from the Epstein-Barr virus) to bind the plasmid to the host chromosome ensuring approximately equal distribution of plasmid from mother to daughter cells [2]. The CHOEBNALT85 cell line is derived from CHO-S and has been engineered to stably express the EBNA-1 protein allowing for binding of plasmids, containing the EBNA binding site, to CHO host chromosomes (Figure 1).

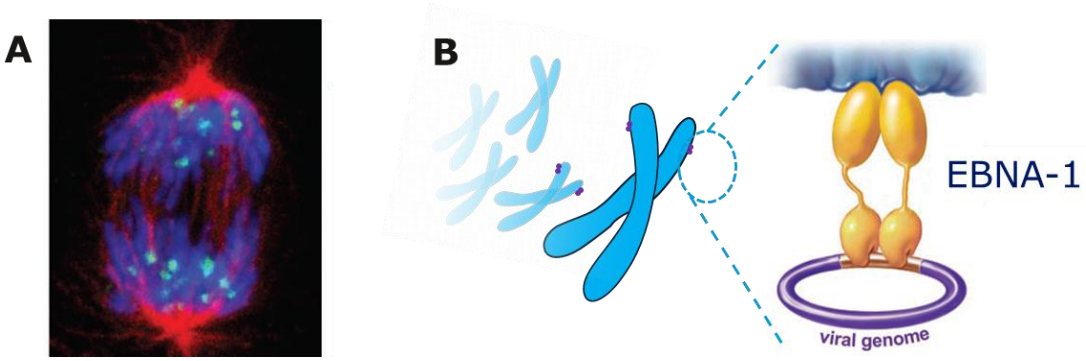


Figure 1 Viral genome propagated into both daughter cells following mitosis due to direct binding of viral genome to host chromosomes by the Epstein-Barr E2 protein. The mechanism above was used for development of the Icosagen system. Figure rearranged from [3].

Protein folding in the cell is assisted by several chaperones able to catalyze the reactions so that correct folding is induced within a short timeframe. Secreted and cell-surface proteins are commonly stabilized by disulfide bonds and the construction of these are catalyzed by Protein disulfide isomerases (PDIs) [4]. It has previously been shown that overexpression of PDI in CHO cells lead to an increase in antibody productivity of 15-37% [5,6], whereas it did not impact the productivity of thrombopoietin [5] and caused reduction in productivity when co-expressed with tumor necrosis factor receptor:Fc fusion protein by retaining the protein in the ER [7]. To my knowledge, the co-expression of PDI with FVIII has not been described in literature. Here data is

presented on FVIII production co-expressed with combinations of three different PDI's (PDI, PDIA3 and PDIA6) using the Icosagen system in order to investigate the robustness of the system for high throughput hypothesis testing and investigate the impact of PDI co-expression for improved FVIII production.

3.3.2 Results and Discussion

In order to investigate the impact of co-expressing PDI, PDIA3 and PDIA6 in CHO cells producing FVIII, the genes were synthesized and inserted into a plasmid compatible with the Icosagen CHOEBNALT85 system. Cells were transfected with different combination of plasmids and the expression of transgenes was validated using western blotting (Figure 2).

A

	1	2	3	4	5	6	7	8	9	10	11
PDI	0.3µg	0	0	1.5µg	0	0	0.3µg	0.3µg	0.3µg	0	0
PDIA3	0	0.3µg	0	0	1.5µg	0	0.3µg	0.3µg	0	0.3µg	0
PDIA6	0	0	0.3µg	0	0	1.5µg	0.3µg	0	0.3µg	0.3µg	0
FVIII	4µg	4µg	4µg	4µg	4µg	4µg	4µg	4µg	4µg	4µg	4µg

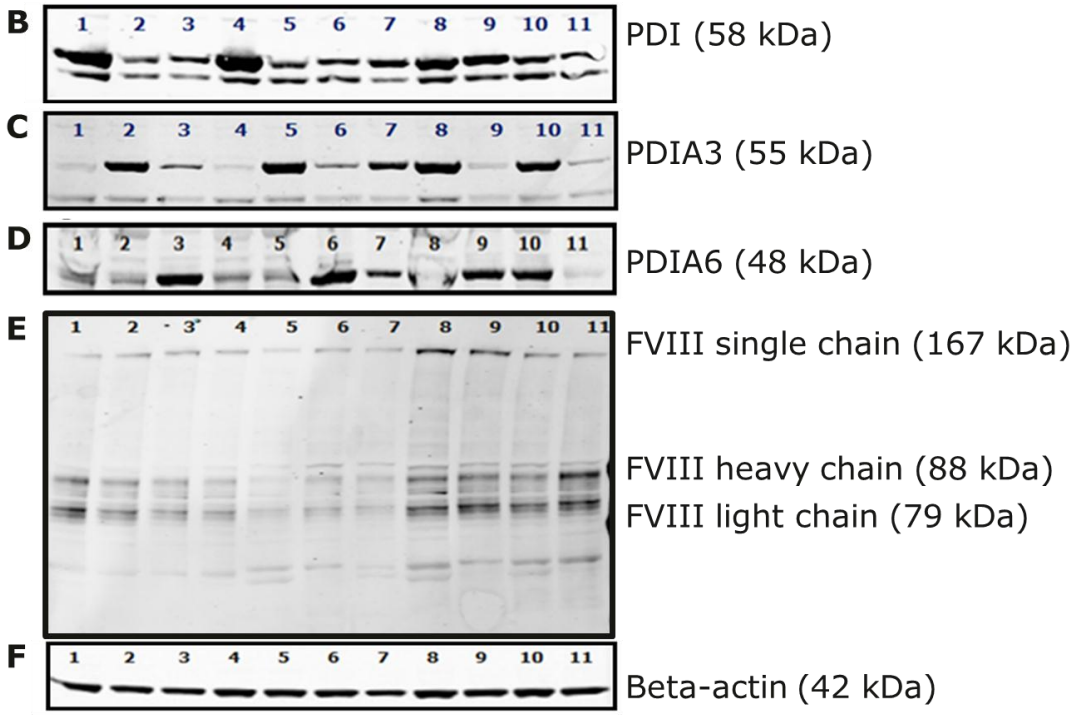


Figure 2 Western blots showing overexpression of PDIs and FVIII
A) Information regarding the amount and plasmid(s) used for transfection of cells. Western blotting staining for **B)** PDI **C)** PDIA3 **D)** PDIA6 **E)** FVIII and **F)** Beta-actin. Picture rearranged with permission from [8]

Following four days of production the productivity was calculated and compared to the plasmids levels in the cell measured by qPCR and *F8* expression levels measured by qRT-PCR (Figure 3A). Pool 11 only transfected with plasmid containing *F8* were found to yield the highest level of *F8* on DNA, RNA and expressed FVIII compared to the other ten pools. Interestingly, Pool 9 expressing *F8* as well as PDI and PDIA6, was found to have a FVIII productivity of 87% to Pool 11 but only

containing a third of the *F8* DNA and RNA-level observed in Pool 11. It has been shown previously that the *F8* expression level does not appear to correlate with FVIII productivity [9]. Due to the severe fluctuations in *F8* expressions levels among the pools it cannot be concluded whether or not the PDIs elicit a significant difference in FVIII productivity.

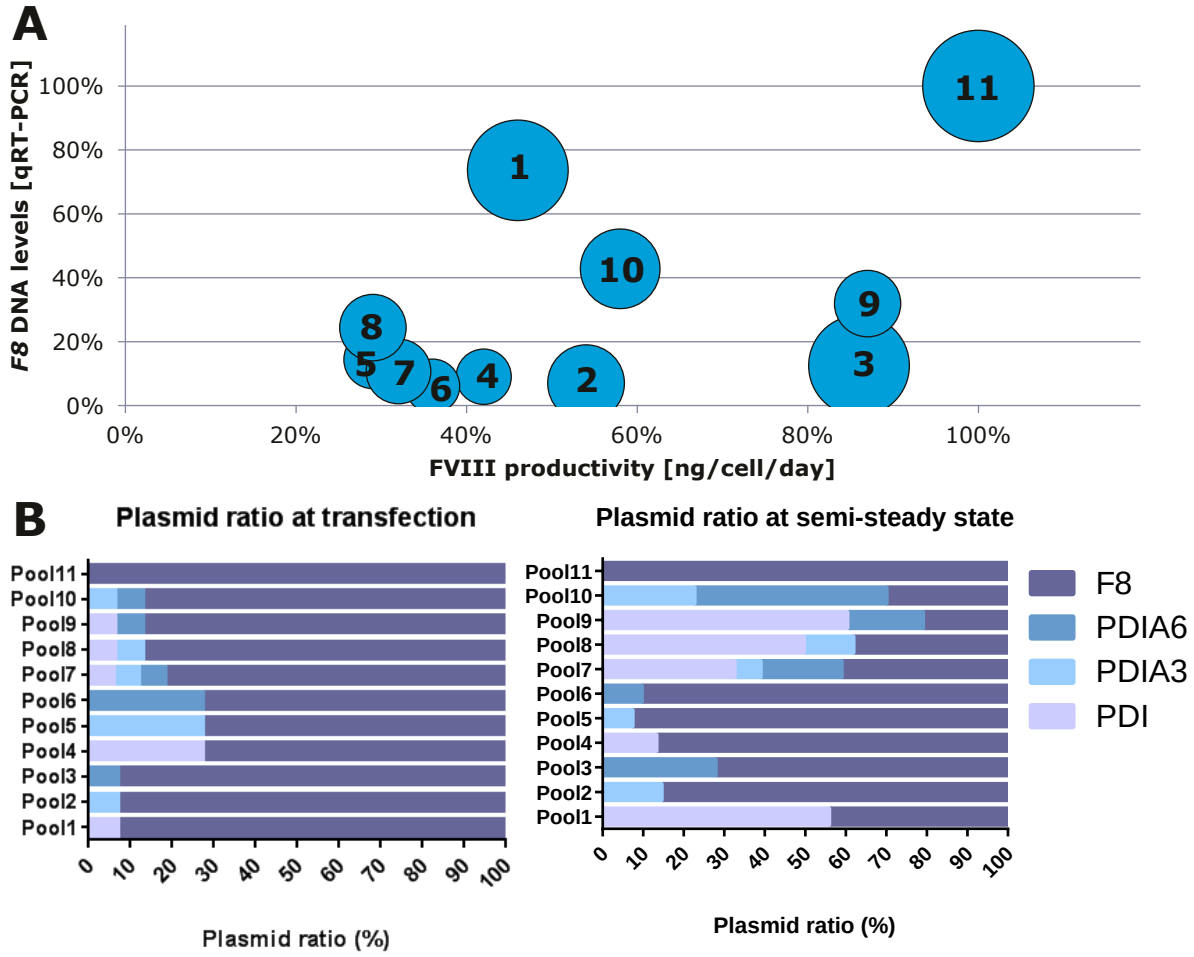


Figure 3 Fluctuations in the *F8* plasmid levels following transfection.
A) Productivity was calculated as ng of FVIII per cell per day relative to the highest producing pool. DNA levels were estimated based on *F8* levels from genomic DNA extraction normalized to pool 11. Diameter of circles indicates *F8* expression levels based on qRT-PCR normalized to pool 11. **B)** Levels of F8, PDI, PDIA3 and PDIA6 transgene plasmid measured by qPCR from extracted genomic DNA.

The ratios of transfected plasmid found in the end of the production phase were found to differ markedly from the ratios used for transfection (Figure 3B). It was seen that plasmids encoding PDI and to a lesser extent PDIA6 and PDIA3 dominated in pools co-expressing these genes and only in Pool 11 encoding only FVIII did the *F8* plasmid dominate. It thus appears based on the differences between the two time points, that a clear selection pressure exists to avoid plasmids encoding *F8*. This is found to be in contrast to data from earlier work in our lab co-expressing antibodies and chaperones of similar size as the PDIs in the same cell lines, where the antibody productivities did not change between pools (data not shown). The problem stem from the fact that only a single

selection marker is available at the present time for the Icosagen CHOEBNALT85 system. In the case more selection markers had been available, one could have been used for the *F8* transgene and another for the PDIs, but that would probably also have led to different equilibria being set from pool to pool within the PDIs. Another solution would be to insert all relevant genes into one plasmid, but that would require several time-consuming cloning steps, and it might prove to be unstable due to the size of the plasmid.

3.3.3 Conclusions

The concentration of episomal *F8* transgene as well as the *F8* expression level was seen to fluctuate noticeably from cell line to cell line showing a clear selection pressure favoring one plasmid over the other, in the case of transfecting multiple plasmids at the same time with the same selection marker. For this reason the Icosagen system cannot be recommended for co-expression studies of *F8*.

3.3.4 Acknowledgements

I would like to acknowledge Anahita Zamani Mohammadian for carrying out the majority of the experimental work described in this section further described in [8].

3.3.5 Materials and Methods

Gene synthesis and generation of transfectants

The CHO genes encoding: PDI, PDIA3 and PDIA6 were synthesised codon optimized for CHO and inserted directly into the QMCF-5 vector (Icosagen, Estonia) by the synthesis vendor (Genescript, NJ, USA). CHOEBNALT85 cells were transfected by electroporation (GenePulser Xcell, Biorad) with 4µg of plasmid encoding FVIII, Herring Sperm DNA Solution and the following concentrations of PDI plasmids: Pool 1: 0.3µg plasmid encoding PDI, Pool 2: 0.3µg plasmid encoding PDIA3, Pool 3: 0.3µg plasmid encoding PDIA6, Pool 4: 1.5µg plasmid encoding PDI, Pool 5: 1.5µg plasmid encoding PDIA3, Pool 6: 1.5µg plasmid encoding PDIA6, Pool 7: 0.3µg Plasmids encoding PDI/PDIA3/PDIA6, Pool 8: 0.3µg Plasmids encoding PDI/PDIA3, Pool 9: 0.3µg Plasmids encoding PDI/PDIA6, Pool 10: 0.3µg Plasmids encoding PDIA3/PDIA6, Pool 11: N/A. Cells were transferred to 1:1 CD CHO and SFM II media (Gibco) with puromycine, and selection with G418 was added after 24 hours. The cells were passaged every 4 days for 16 days until transferred to a 30 ml culture seeded with $3x10^5$ cells/ml in an orbital shaker at 36.5˚C at a shaking speed of 125 rpm and 8.0% CO_2. Following three days of growth the incubation

temperature was lowered to 30°C and after 4 days of growth samples were taken for DNA purification, RNA purification, western blotting and FVIII quantification.

Western blotting

Protein lysates for Western blotting were prepared by spinning down 10e6 cells and resuspending the pellet in 200 µl of Mammalian Protein Extraction Reagent (M-PER) (Thermo) following manufacturer's instructions. 30 µg of each sample was used for Western blotting. Gels were submitted to western blotting using Novex/NuPage blotting system (Invitrogen). Primary antibodies used were 1:500 dilution of sheep polyclonal anti-human factor VIII (CL20035AP, Cedarlane labs, Burlington, ON, Canada), 1:1000 dilution of PDI, mAb (1D3) Mouse IgG1 (Enzo® Life Sciences, Farmingdale, Ny, USA), 1:167 dilution of Goat anti-Human PDI A3 (RayBiotech, Inc., Norcross, Ga, USA), 1:100 dilution of Rabbit anti-PDI A6 antibody (Pierce Biotechnology , Rockford, Il USA) and 1:1000 dilution of rabbit anti-beta-actin antibody (Cell Signaling Technology, Danvers, Ma, USA). Secondary antibodies used were 1:20000 dilution of Donkey anti-Rabbit IRDye® 800 CW (LI-COR® Biosciences, Lincoln, Ne, USA) 1:20000 dilution of Donkey anti-Mouse IR-Dye® 680 LT (LI-COR® Biosciences, incoln, Ne, USA), 1:15000 dilution of Donkey anti-Goat IRDye® 800 CW (LI-COR® Biosciences, Lincoln, Ne, USA) and 1:10000 dilution of Alexa Fluor® 680 donkey anti-sheep IgG (Invitrogen, Carlsbad, Ca, USA). The ladders used were Full Range Rainbow™ recombinant protein molecular weight marker (GE lifescience, Piscataway, NJ, USA)

Estimation of plasmid and expression levels

Genomic and plasmid DNA purification was extracted from $2x10^6$ cells using DNeasy Blood & Tissue Kit (Qiagen, Germantown, Md, USA) following the manufacturer´s instructions. RNA was extracted from $2x10^6$ cells using TRIzol (Invitrogen) the RNeasy Cleanup kit (Qiagen) following the manufacturer's instructions. RNA integrity was confirmed on an Agilent 2100 Bioanalyzer using total RNA nano chips (Agilent Technologies, Santa Clara, Ca, USA). RNA concentration was measured using a NanoDrop spectrophotometer (NanoDrop Technologies). cDNA was produced from 1µg RNA using SuperScript III first-strand synthesis supermix (Invitrogen, Carlsbad, CA). Primers used for *F8* ORF (transgene position 1676-1695 and 1752-1775), *gapdh* (Forward: AACTTTGGCATTGTGGAAGG and reverse: ACACGTTGGGGGTAGGAACA), PDI (Forward: CTGCAGCCGAAACACTGA and reverse: CACATCCTTAAAGAATCCGATGA), PDIA3 (Forward: TCCCTTCTCCATACGAGGTG and reverse: CAGTTTCTTGTTGGCAGGACT), PDIA6 (Forward: GGAGGATTCGGATCACCAG and reverse: CAGCAGTGCGAATTTCATCT),

QMCF-5 plasmid backbone (Forward: CATCAGCCATGATGGATACTTTC and reverse: GGCAGGATCTCCTGTCATCT)

The qPCR reactions were run as 20µl reaction using Quantifast SYBR green PCR Master Mix (QIAGEN, Germany) following manufacturer's instructions on a Stratagene MX3000P real-time PCR system (Stratagene). Primer efficiencies were calculated based on 5 consecutive 5-fold dilutions of cDNA samples and used to calculate relative expression ratio for each gene relative to *gapdh* and normalized compared to pool 11 as described elsewhere [10]. For calculation of plasmid ratios the total level of plasmid was deduced based on primers binding to the plasmid backbone normalized to *gapdh* and compared to levels of relevant transgenes normalized to *gapdh*.

Specific productivity

FVIII antigen was measured using an ELISA kit (VisuLize, FVIII Antigen Kit) from Affinity Biologicals (Ancaster, ON, USA) as described by the manufacturer. The titer was normalised to the number of viable cells at the time of FVIII extraction as measured on a Vicell XR system (Beckman Coulter), the number of days days after the temperature drop and finally to the highest producing pool.

3.3.6 References

1. Silla T, Tagen I, Kalling A, Tegova R, Ustav M, Mandel T *et al*.. Viral expression plasmids for production of proteins, antibodies, enzymes, virus-like particles and for use in cell-based assays. 31-3-2011. Patent:US20110076760.

Ref Type: Patent

2. Silla T, Hääl I, Geimanen J, Janikson K, Abroi A, Ustav E *et al*.: **Episomal maintenance of plasmids with hybrid origins in mouse cells.** *Journal of virology* 2005, **79:** 15277-15288.

3. McBride AA, Oliveira JG, McPhillips MG: **Partitioning Viral Genomes in Mitosis: Same Idea, Different Targets.** *Cell Cycle* 2014, **5:** 1499-1502.

4. Wilkinson B, Gilbert HF: **Protein disulfide isomerase.** *Biochimica et biophysica acta* 2004, **1699:** 35-44.

5. Mohan C, Park SH, Chung JY, Lee GM: **Effect of doxycycline-regulated protein disulfide isomerase expression on the specific productivity of recombinant CHO cells: thrombopoietin and antibody.** *Biotechnology and bioengineering* 2007, **98:** 611-615.

6. Borth N, Mattanovich D, Kunert R, Katinger H: **Effect of increased expression of protein disulfide isomerase and heavy chain binding protein on antibody secretion in a recombinant CHO cell line.** *Biotechnology progress* 2005, **21:** 106-111.

7. Davis R, Schooley K, Rasmussen B, Thomas J, Reddy P: **Effect of PDI overexpression on recombinant protein secretion in CHO cells.** *Biotechnology progress* 2000, **16:** 736-743.

8. Mohammadian AZ: *Master thesis: Optimization of disulfide bond formation for heterologous production of coagulation factor FVIII in Chinese hamster ovary cells*. Technical University of Denmark; 2014.

9. Kaas CS, Bolt B, Hansen JJ, Andersen MR, Kristensen C: **Deep sequencing reveals different compositions of mRNA transcribed from the FVIII gene in a panel of FVIII producing CHO cell lines.** *Biotechnology journal* 2015.

10. Kubista M, Andrade JM, Bengtsson M, Forootan A, Jonák J, Lind K *et al*.: **The real-time polymerase chain reaction.** *Molecular aspects of medicine* 2006, **27:** 95-125.

Chapter 4 – Impact of FVIII production on the CHO cell

In this chapter, the general transcriptome and proteome of the 14 CHO transfectants producing FVIII described in the previous chapter will be investigated. It was found that expression of the *F8* transgene has a severe impact on the state of the cells causing them to respond to the pressure by upregulating the unfolded protein response, the oxidative stress pathway and degrade the FVIII protein through the Endoplasmic-reticulum-associated protein degradation pathway. The preliminary manuscript attached is aimed for submission to Biotechnology and Bioengineering in the autumn of 2015.

4.1 Manuscript 1: Characterization of Chinese Hamster Ovary cells producing Coagulation Factor VIII using transcriptomics and proteomics

Christian Schrøder Kaas[1,2], Anne Mathilde Lund[2], Deniz Baycin-Hizal[3], Michael J Betenbaugh[3], Mikael Rørdam Andersen[2], Gert Bolt[1], Claus Kristensen[4]

1 Mammalian Cell Technology, Global Research Unit, Novo Nordisk A/S, Måløv, Denmark.

2 Network Engineering of Eukaryotic Cell Factories, Technical University of Denmark, Kgs Lyngby, Denmark.

3 Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, MD, USA.

4 Institute of Cellular and Molecular Medicine, University of Copenhagen, Copenhagen, Denmark

4.1.1 Abstract

Coagulation factor VIII (FVIII) is one of the most complex proteins produced heterologously in large scale as a biopharmaceutical. There is currently no established cell line that naturally expresses FVIII and thus the knowledge gained concerning the expression and secretion of the protein is found solely from analyses of various mammalian cell lines transfected with *F8* cDNA. We have recently published RNA sequencing data from 14 CHO DXB11 clones expressing a dynamic range of recombinant FVIII, characterizing the transgene composition of the clones. Here we attempt to characterize the global impact FVIII biosynthesis exerts on the transcriptome and proteome of a number of CHO transfectants producing FVIII. A FVIII dose-dependent induction of the unfolded protein response, the NRF2-mediated oxidative stress response and the endoplasmic reticulum stress pathway is observed. Xbp1 were found to be spliced in the FVIII producing cells causing induction of the unfolded protein response. Proteomic data from three clones were found to correlate with a Pearson's correlation coefficient of 0.38, but a correlation of 0.84 was found between the two methods when analysing the subset of genes correlating with the FVIII productivity. Finally it was attempted to overexpress targets based on the data. Overexpression of PDI, NRF2, Bach1, LAMN1, Calr, PDI/PDIA6, and Calr/PDIA3 was found in all instances to yield significant reductions in FVIII productivity when co-expressed, but a minor increase in FVIII productivity were seen by co-expression of PDIA6, FKBP2 and the miRNA: let-7f-2 in a transient setting.

4.1.2 Introduction and background

Approximately 350,000 males world-wide suffer from hemophilia A, which is caused by genetic abnormalities of the *F8* gene encoding coagulation factor VIII (FVIII). There is currently no established cell line that naturally expresses FVIII and thus the knowledge gained concerning the expression and secretion of the protein is found solely from analysis of various mammalian cell lines transfected with *F8* cDNA (Plantier et al. 2005). Heterologous expression of human FVIII are though found to be at levels three order of magnitude lower than similarly sized secreted glycoproteins (Brown et al. 2014;Kaufman et al. 1997).

Following transcription of the *F8* gene, the translation is directed into the endoplasmic reticulum (ER) where it upon folding binds ER chaperones, such as immunoglobulin-binding protein (BiP) (Marquette et al. 1995), calreticulin and calnexin (Pipe et al. 1998). The calnexin/calreticulin cycle function as a quality control step for glycoproteins by retaining proteins in the ER until they are correctly folded (Ellgaard and Helenius, 2003). Unfortunately, it appears that a substantial portion of the translated FVIII accumulate as aggregates in the ER probably due to ATP depletion (Tagliavacca et al. 2000). This further activates the unfolded protein response (UPR) in the cell and cause oxidative stress (Malhotra et al. 2008).

We have previously described selection and growth characterization of 14 CHO DXB11 transfectants producing FVIII in a dynamic range (Kaas et al. 2015). The 14 clones were grown in batch cultures and samples for RNA sequencing were drawn 48 hours into the cultivation when all cultures were in the exponential growth phase. No correlation was seen between expression of the *F8* transcript and the level of functional secreted FVIII. The transgene compositions of the clones were thoroughly characterized revealing severe truncations of the transgene in clone 12-14 explaining the lack of FVIII production and Clone 3-11 were found not to transcribe a tripartite leader in the 5' end of the transcript expected to impact the level of FVIII protein produced from each *F8* transcript. Here we present, to our knowledge, the first description of the global impact that FVIII biosynthesis exert on the transcriptome and proteome of CHO cells.

4.1.3 Results and Discussion

Characterization of expression patterns from FVIII cell lines

14 CHO DXB11 transfectants (named based on descending FVIII productivity) were grown under batch conditions, RNA was extracted 48 hours into the cultivations and used for RNA sequencing

(RNAseq). The overall variation within the dataset was visualized based on multidimensional scaling of the entire RNAseq dataset (Figure 8A). It is seen that the three top producing clones separate on the first dimension from the rest, which can be explained by the fact that these three clones did not originate from the same transfection reaction from the rest of the clones. On the second dimension clone 12 and 13 separate from the rest as controls versus the FVIII producers. The third control cell line, clone 14, is seen to clump together with the FVIII producers in this dimension.

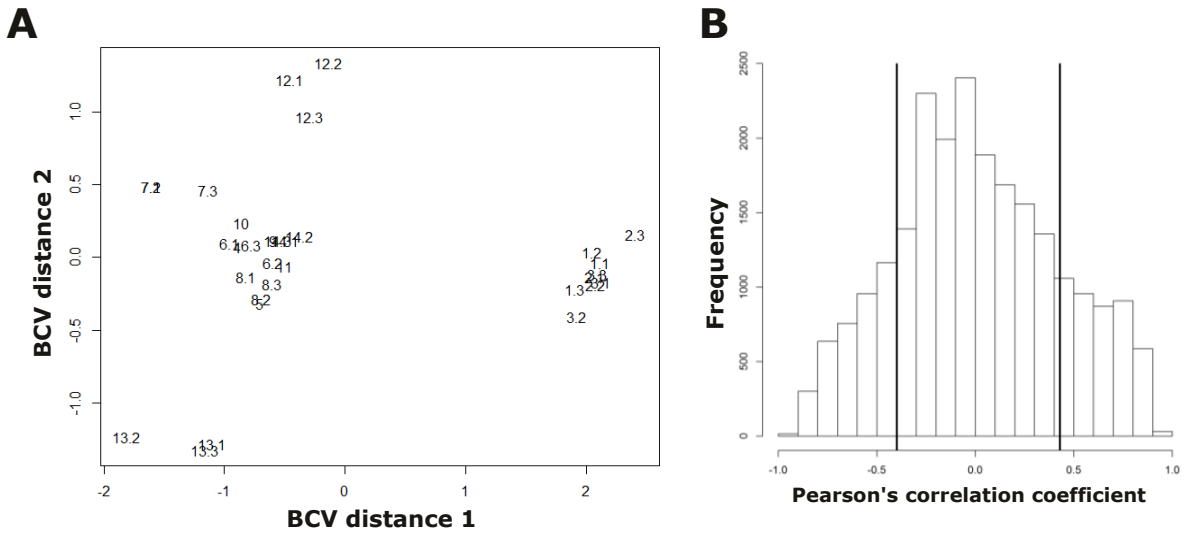


Figure 8 Characterization of the RNAseq dataset.
A) Multidimensional scaling of the biological coefficient of variation for the RNAseq dataset from all samples. **B)** A histogram showing the distribution of pearson's correlation coefficients for each gene to the FVIII productivity of the cell line. Lines mark one standard deviation from the mean.

By mapping the Pearson's correlation coefficient for the expression level of each gene to the amount of secreted FVIII found at the time of RNA extraction, a list of 504 genes with a correlation higher than one standard deviations from the mean (Figure 8B) and a fold change higher than two from control to high producer were extracted (Supplementary table 1-2). Notably, BiP (HSPA5), which is known to bind FVIII in the ER, was found to be upregulated 3.3 fold from the top FVIII producer (Clone 1) versus a control (Clone 12) with a p-value of 7.6E-108. BiP was furthermore found to correlate with a Pearson correlation coefficient of 0.87 indicating a dose-dependent induction of the expression. In contrast, the BiP level did not correlate to the level of *F8* mRNA (the Pearson's correlation coefficient = 0.08), which could indicate that the translational efficiency is different from clone to clone. It has previously been found that overexpression of BiP decreased FVIII productivity and down-regulation of BiP using shRNA lead to a 50% increase of FVIII yield

(Brown et al. 2011), which represent a complex interplay between FVIII productivity and BiP expression.

The Ingenuity Pathway Analysis (IPA) software package was used to identify biological pathways significantly enriched for the 504 genes correlating with productivity (Table I, full table in Supplementary Table 3). Notably, the three highest ranking pathways are the unfolded protein response (UPR), the endoplasmic reticulum stress pathway and the NRF2-mediated oxidative stress response indicating a dose dependent relationship between induction of these pathways and the amount of FVIII that successfully pass through the secretion machinery of the CHO cell.

Table I List of the top 10 enriched pathways as found by IPA for genes correlating with FVIII productivity.

Ranking	Ingenuity Canonical Pathways	p-value	DE genes	Pathway size	CHO Genes
1	Unfolded protein response	1.66E-07	10	42	Calr, Edem1, Traf2, Dnajc3, Srebf1, Hsp90b1, Hspa5 (BiP) , Ppp1r15a, Sel1l, LOC103163461
2	Endoplasmic Reticulum Stress Pathway	6.31E-05	5	19	Calr, Traf2, Dnajc3, Hsp90b1, Hspa5 (BiP)
3	NRF2-mediated Oxidative Stress Response	1.23E-04	13	133	Chcr1, Acta2, Ppib, Pik3cb, LOC100753467, Dnajc3, LOC100750937, LOC100766772, Dnajb11, Ephx1, Gclm, Dnajb14, Herpud1
4	Amyotrophic Lateral Sclerosis Signaling	2.29E-04	9	51	Grin3b, Akt3, Grin2d, Xiap, Bcl2l1, Gria4, Vegfb, Pik3cb, Birc3
5	TNFR2 Signaling	3.24E-04	5	29	Traf2, Tnfaip3, Xiap, Nfkbie, Birc3
6	Role of Osteoblasts, Osteoclasts and Chondrocytes in Rheumatoid Arthritis	8.13E-04	13	219	Nfatc4, Traf2, Akt3, Dkk3, Itga3, Nfkbie, Pik3cb, Birc3, Il18, Csf1, Xiap, Il33, Apc2
7	PI3K Signaling in B Lymphocytes	1.58E-03	9	128	Nfatc4, Akt3, Fcgr2b, C3, Pdia3, Blnk, Atf5, Nfkbie, Pik3cb
8	Agranulocyte Adhesion and Diapedesis	2.34E-03	11	189	Il18, LOC100763261, Mmp9, Itga3, Cxcl3, Acta2, Sdc4, Cxcl1, Mmp17, Il33, LOC100769029
9	CD40 Signaling	2.45E-03	6	65	Traf2, Fcer2, Tnfaip3, Ptgs2, Nfkbie, Pik3cb
10	Molybdenum Cofactor Biosynthesis	2.57E-03	2	4	Mocs3, Nfs1

Xbp1 splicing

There were seen to be a significant overlap (p-value 6.09E-10) between 21 genes found to correlate with productivity (Srebf1, Sec11c, Sdf2l1, Rpn2, Ppib, Piga, Pdia3, Pdia4, Pdia6, Hyou1, Hspa5 (BiP), Hsp90b1, Herpud1, Fkbp2, Fkbp11, Edem1, Dnajc3, Dnajb11, Ddost, Calr, Cxcl1), and genes found to be up-regulated by the transcription factor Xbp1(s) (Sriburi, 2007). The transcription

factor Xbp1 is normally translated into an in-active isoform but in the case unfolded proteins accumulate in the ER lumen, BiP will bind the proteins and no longer bind sequester IRE1α (encoded by Ern1). Once liberated from BiP, IRE1α will homodimerize and use its RNase activity to splice the transcript of Xbp1 by removal of a 26bp intron inside exon3 causing the translation of spliced Xbp1 (called Xbp1(s)) (Zhang and Kaufman, 2006). The ratio of Xbp1(s) versus total Xbp1 transcripts is shown in Figure 9 to be a mean of 40±13% in FVIII producing cell lines versus 14±10% in control cells (p-value 5.6E-06) indicating that the majority of BiP is indeed bound in the FVIII producing cell lines (Clone 1-11) causing induction of the unfolded protein response. Earlier work showed that overexpression of Xbp1(s) in CHO-K1 cells lead to an increase in SEAP production (2x) (Tigges and Fussenegger, 2006), but overexpression in a FVIII cell line has been shown to yield conflicting results with no impact (Campos-da-Paz et al. 2008) or positive impact on FVIII biosynthesis (Brown et al. 2011).

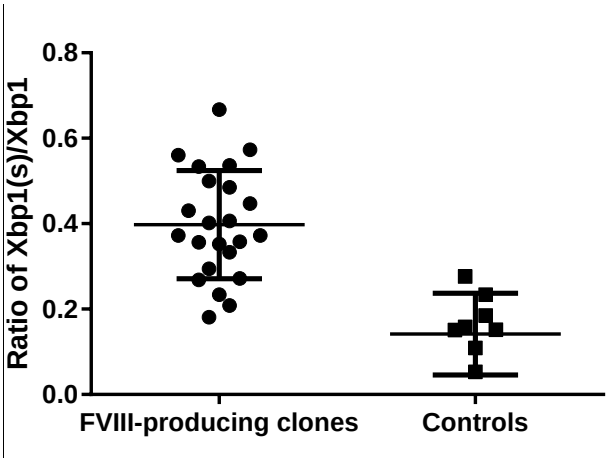


Figure 9 Ratio of spliced Xbp1(s) versus total expression of Xbp1.
Ratios shown of Xbp1(s) compared to total Xbp1 transcript from FVIII producing clones (Clone 1-11) versus the controls (Clone 12-13). See Supplementary Figure 1 for further details.

Network-based analysis of transcriptional response to FVIII expression

In order to gain a better insight into the specific impact that FVIII expression puts on the CHO secretion machinery, a CHO-specific reconstruction of the secretion network (Lund *et al*, manuscript in preparation) was used to evaluate the data (shown in Figure 10A and listed in Supplementary Table 4). The network consists of 654 genes known to be relevant for secretion linked by 166 functions. Only genes known to be present in the CHO-K1 genome (Xu et al. 2011) were included in the network. When analysing the subnetwork associated with the function "recognition of misfolding" (Figure 10B) it is seen that the CRT/CNX-complex consisting of Calreticulin/ERP57 (PDIA3)/Calnexin is up-regulated. This complex is essential for quality control of glycoproteins such as FVIII and it is seen that Calreticulin and PDIA3 correlates with the FVIII

productivity whereas Calnexin is only slightly induced (as Calr and PDIA3 expression is induced by Xbp1(s)). PDIA3 has been shown to be vital in assembly of the HMC complex (Garbi et al. 2006) and furthermore have been found to mediate the substrate recognition of the Calreticulin/PDIA3/Calnexin complex (Oliver et al. 1999). It would thus be interesting to investigate whether PDIA3 alone or in combination with calreticulin could be a limiting factor in FVIII biosynthesis.

PDI expression and a possible link to oxidative stress

Genes coding for PDI's such as PDIA3 (ERP59), PDIA4 (ERP72) and PDIA6 (TXNDC7) were found to be up-regulated in FVIII cells as a part of the UPR (Figure 10B). This tendency is interesting as disulfide bond formation catalyzed by PDIs produce reactive oxygen species (ROS) and it has been shown that production of FVIII lead to oxidative stress due to ROS (Malhotra et al. 2008). From the current dataset it is clear that there is not a single gene in the oxidative phosphorylation or glycolysis pathway that is seen to be differentially expressed in the FVIII producing cell lines (data not shown), which would otherwise be considered the primary source for ROS. It is estimated that approximately 25% of the ROS generated by the cell is a byproduct from disulfide bond formation in the ER (Malhotra and Kaufman, 2007) and due to the presence of the retained FVIII in the ER, overexpression of PDI could thus result in a futile cycle of disulfide bond formation and breakage leading to elevated ROS production, oxidative stress and subsequently apoptosis.

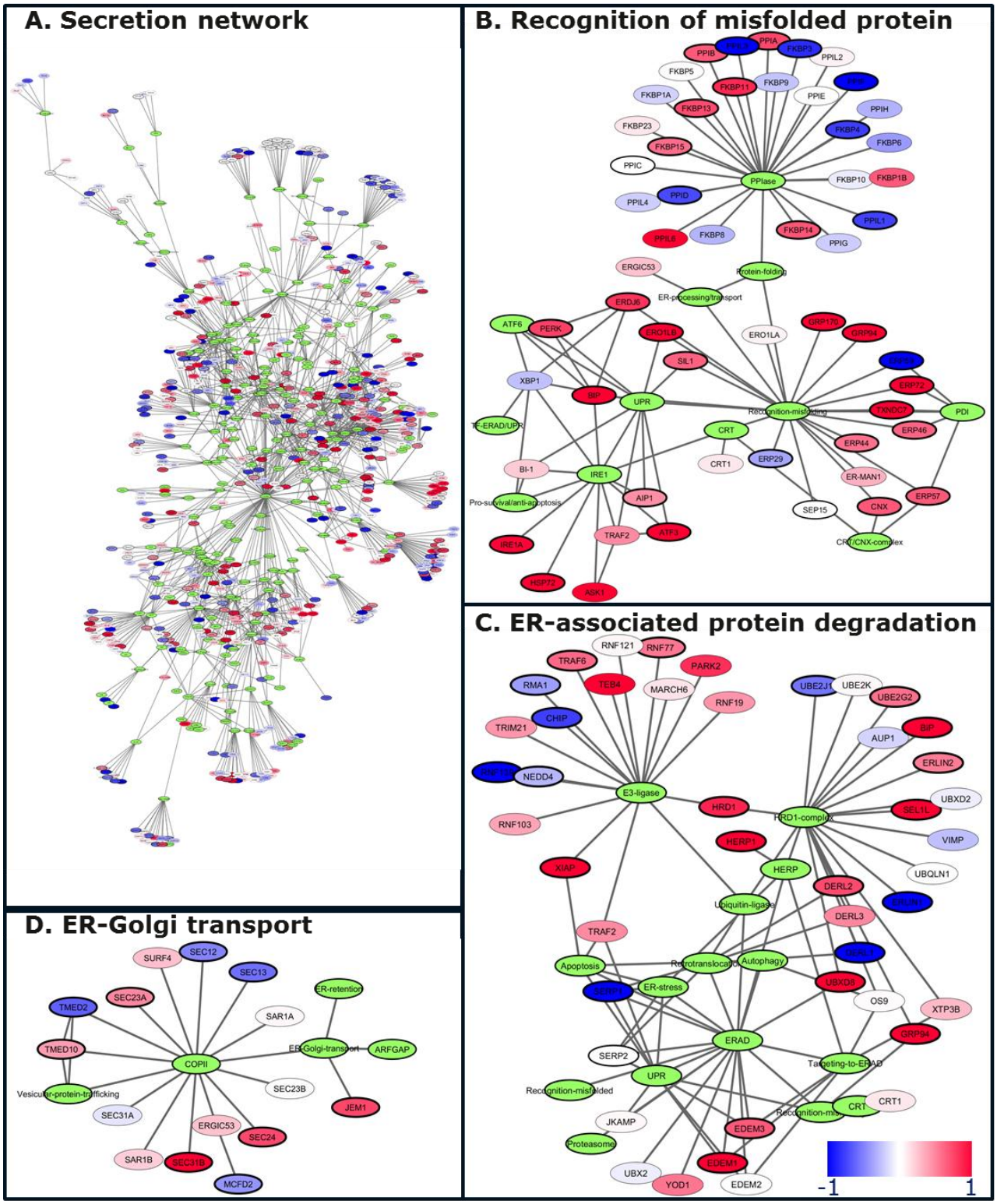


Figure 10 CHO specific secretion network overlaid with RNAseq data
A) CHO secretion network consisting of 654 genes and 166 functions. **B)** Subnetwork consisting of genes relevant for detection of misfolded protein. **C)** Subnetwork consisting of genes relevant for ERAD. **D)** Subnetwork consisting of genes relevant for ER-Golgi transport. Colors indicate log2(Foldchange) between all FVIII producing clones (Clone1-11) versus controls (Clone 12-14).

Endoplasmic-reticulum-associated protein degradation actively degrade FVIII

Several members of the endoplasmic-reticulum-associated protein degradation (ERAD) pathway such as EDEM1, EDEM2 and DERL2 were found to be significantly up-regulated in FVIII producing cells (Figure 10C). These findings correspond well with the fact that pulse chase experiments has previously shown that significant portions of synthesized FVIII is held back in the ER and degraded instead of being secreted (Pipe et al. 1998). EDEMs have been found to bind glycoproteins failing quality control from the Calreticulin/Calnexin/PDIA3 complex and delivering them for translocation to the cytosol and subsequent degradation by the 20S proteasome, by retrotranslocation channels such as the one encoded by DERL2 (Vembar and Brodsky, 2008;Oda et al. 2006). Overexpression of EDEM1 and DERL2 has previously been shown to result in decreased half-life of misfolded proteins in the ER (Molinari et al. 2003;Oda et al. 2006). Addition of lactacystin (an inhibitor of the 20S proteasome) resulted in successfully stalling ERAD and caused accumulation of FVIII in the ER, but it did not impact the secretion level of FVIII indicating that increasing the half-life of FVIII in the ER will not impact the productivity of a cell line (Pipe et al. 1998).

Transport from the ER to the golgi reveal patterns different from IgG production

Within the expression pattern of the genes known to be relevant for ER to Golgi transport it is seen that LMAN1 (ERGIC53) and MCFD2 are constitutively expressed showing no significant induction in high FVIII producing cell lines (Figure 10D). Previous studies showed that addition of cycloheximide (a strong inhibitor of protein biosynthesis) to a cell line producing FVIII-GFP lead to the majority of FVIII being contained in the ER close to exit sites for more than 6 hours, which is very slow compared to just GFP containing an ER-export signal sequence (Heinz et al. 2009). LAMN1 and MCFD2 are known to be essential for correct transport from the ER to the cis-golgi complex (Zhang et al. 2006) and it could thus be suggested that LMAN1 and MCFD2 constitute a bottleneck in FVIII biosynthesis as the production of FVIII must be expected to be several orders of magnitude higher in a stable FVIII producing cell line like Clone 1 compared to native expression of *F8* in hepatocytes and sinusoidal endothelial cells (Hollestelle et al. 2001).

Proteome data correspond well with transcriptomics data

In order to validate the results above the proteomes from clone 1, 7 and 12 were investigated. The clones were grown under the same experimental conditions as for the RNA extraction and protein lysates were extracted 48 hours into the cultivation when the cells were in the exponential growth

phase. There were no significant differences detected in neither growth rate nor productivity (data not shown). Following treatment and digestion the lysates were divided and measured over two mass spectrometry (MS) runs identifying 5062 and 6152 proteins with an FDR < 1% respectively (Supplementary Table 5). 4187 proteins were detected in both runs and used for further analysis. Notably, BiP were found to be upregulated by 2.8 fold from the Clone 12 to Clone 1, which correspond well with the results from RNAseq. The oxidative stress response (p-value 3.24E-07) was found to be the most significantly enriched pathway in the two FVIII producing clones (Clone 1 and 7) versus the control (Clone 12). By comparing the 15 genes found to be differentially expressed in the oxidative stress pathway either by RNAseq (14 genes) or MS (11 genes) it is seen that the two methods generally agree in respect to fold-changes with some discrepancies for genes with low expression levels in the RNA sequencing data (Gstm3, GSTM1 and DNAJB13) (Table 2).

Among the 4187 proteins detected, there was found to be a correlation of 0.38 in fold-changes between Clone 1 and Clone 12 in RNAseq versus MS (Figure 4). By removing the 4% of the genes with very low expression levels in the RNAseq experiment the correlation between the two methods was raised to 0.50. The relation was fitted using least-square regression yielding a slope of 0.406 (p-value < 2E-16), which is very close to the slope of 0.4 found earlier for the correlation between RNAseq expression levels and protein identities probably indicating translational regulation (Baycin-Hizal et al. 2012). Among the 504 genes correlating positively with productivity in the RNAseq data 101 were found in proteomics as well and among the 647 genes correlating negatively with the productivity, 163 proteins were found. These 264 genes correlated by 0.84 between the two methods revealing a large overlap in differentially expressed genes found through both methods and showing that discrepancies between the two methods to a large extent is found in the noise of housekeeping genes.

Table II Overview of differentially expressed genes from the oxidative stress pathway as detected by transcriptomics and proteomics.

		Transcriptome data				Proteome data	
	Gene Name	p-value	Intensity (FPKM)	Fold Change	Pearson cor	p-value	Fold Change
Gstm3	glutathione S-transferase, mu 3	3.4E-164	3.4	86.8	0.88	4.9E-02	-1.3
Gsta1	glutathione S-transferase alpha 1	3.1E-65	358.2	21.3	0.87	5.5E-06	4.2
GSTM1	glutathione S-transferase mu 1	2.0E-54	3.0	9.3	0.69	2.1E-01	-2.6
HERPUD1	homocysteine-inducible, endoplasmic reticulum stress-inducible, ubiquitin-like domain member 1	2.2E-22	38.8	4.3	0.83	2.5E-03	3.4
EPHX1	epoxide hydrolase 1, microsomal (xenobiotic)	4.1E-11	5.4	3.0	0.72	1.4E-03	2.0
PIK3CB	phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit beta	4.3E-28	23.6	2.7	0.87	4.9E-03	1.6
FTL	ferritin, light polypeptide	3.6E-10	7.2	2.7	0.43	#Not detected	

DNAJC3	DnaJ (Hsp40) homolog, subfamily C, member 3	1.7E-44	109.3	2.4	0.92	#Not detected	
MRAS	muscle RAS oncogene homolog	9.6E-25	16.6	2.3	0.77	#Not detected	
MAFF	v-maf avian musculoaponeurotic fibrosarcoma oncogene homolog F	8.2E-10	5.4	2.2	0.66	#Not detected	
DNAJB11	DnaJ (Hsp40) homolog, subfamily B, member 11	3.2E-31	105.0	2.2	0.90	3.1E-02	1.8
CBR1	carbonyl reductase 1	7.4E-10	17.9	2.1	0.85	#Not detected	
MAFK	v-maf avian musculoaponeurotic fibrosarcoma oncogene homolog K	1.3E-10	21.5	2.0	0.37	#Not detected	
GSTA2	glutathione S-transferase alpha 2	4.1E-33	239.8	2.0	0.75	#Not detected	
MGST1	microsomal glutathione S-transferase 1	6.8E-21	597.9	-1.8	-0.69	5.0E-03	-2.4
PMF1/PMF1-BGLAP	polyamine-modulated factor 1	2.3E-16	35.3	-2.0	-0.80	#Not detected	
DNAJC16	DnaJ (Hsp40) homolog, subfamily C, member 16	4.1E-16	24.8	-2.1	-0.50	#Not detected	
DNAJB2	DnaJ (Hsp40) homolog, subfamily B, member 2	6.4E-14	50.7	-2.1	-0.81	#Not detected	
AKR1A1	aldo-keto reductase family 1, member A1 (aldehyde reductase)	4.0E-50	548.0	-2.2	-0.69	1.1E-02	-2.1
AKT1	v-akt murine thymoma viral oncogene homolog 1	3.2E-52	233.5	-2.3	-0.88	1.4E-03	-2.2
PRDX1	peroxiredoxin 1	1.1E-47	1589.2	-2.3	-0.49	3.6E-03	-2.0
CUL3	cullin 3	5.0E-20	79.2	-2.4	-0.86	2.6E-02	-1.4
AOX1	aldehyde oxidase 1	1.4E-19	3.5	-4.1	-0.60	#Not detected	
DNAJB13	DnaJ (Hsp40) homolog, subfamily B, member 13	3.0E-40	3.8	-8.2	-0.77	1.2E-01	2.0

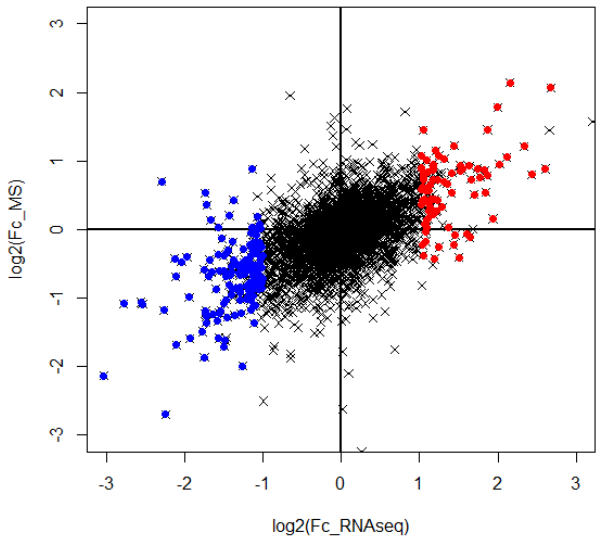


Figure 11 Correlation of transcriptomics and proteomic data.
Fold change in expression of 4187 genes found in both the proteome and transcriptome datasets. Transcripts correlating positively with productivity shown in red and transcripts correlating negatively with the productivity in blue.

Transient and stable overexpression of target genes

Based on the large number of chaperones found to correlate with the FVIII productivity, it was hypothesized that these might be limiting factors for successful folding of FVIII and they were thus chosen to be overexpressed. The PDIs: PDI (ERP59), PDIA3 (ERP57), PDIA6 (TXNDC7), two PPiases catalyzing proline cis-trans isomerization: FKBP13 (FKBP2) and FKBP11 as well as Calreticulin were chosen for overexpression. Furthermore, the inducer of the oxidative stress response, NRF2 (previously overexpressed in neurons by (Shih et al. 2003)), as well as two transcription factors with unknown targets (TGIF2 and Bach1) and the two ER-golgi transporters: LAMN1 and MCFD2 were overexpressed. Finally, miRNAseq was conducted from Clone 1, 7 and 13 identifying 748 expressed miRNAs in all samples. 4 miRNAs (let-7d, let-7f-2, mir-181c and mir-34b) were found to be differentially expressed and correlated with the FVIII productivity and were overexpressed as short-hairpin RNAs coupled to GFP.

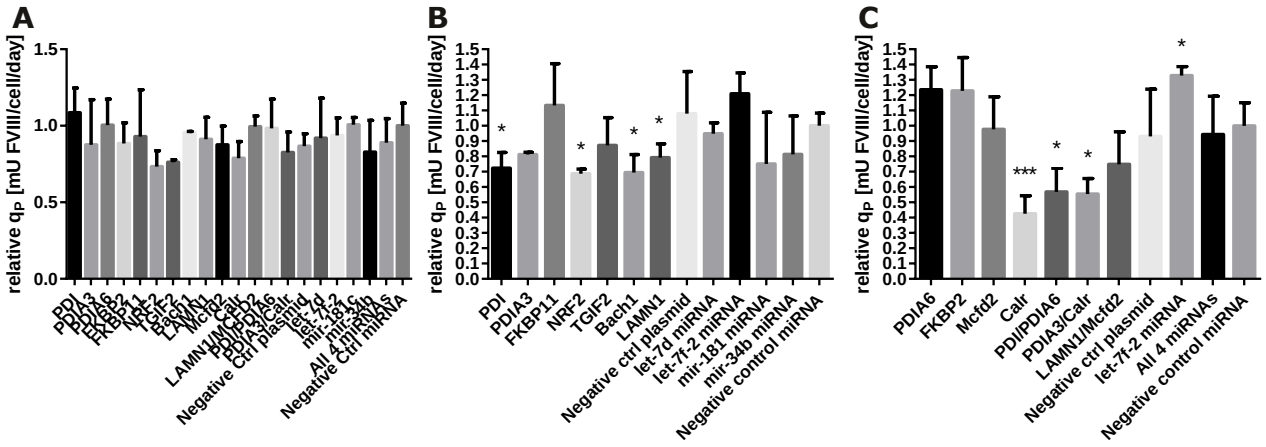


Figure 12 Specific productivity of FVIII transient transfectants.
A) Specific productivity made relative to the productivity in the negative control miRNA cell line after 72 hours of growth in Clone 7 transiently co-expressing target genes. Specific productivity made relative to the productivity in the negative control miRNA cell line after in Clone 1 transiently co-expressing target genes for **B)** 48 hours and **C)** 72 hours. Error bars indicate standard deviation within biological triplicates. * p-value < 0.05. *** p-value < 0.01.

Transient overexpression of the targets above in clone 7 did not yield any significant changes in FVIII productivity (Figure 2A)**,** but it was found that overexpression of the miRNA let-7f-2 lead to a significant increase in specific productivity of 33% (Figure 5C) when overexpressed in Clone 1. Overexpression of PDIA6 and FKBP2 lead to an increases in productivity of 24% and 23% respectively, but was not significant due to a relatively high standard deviation for the productivity of the control. Overexpression of PDI, NRF2, Bach1, LAMN1, Calr, PDI/PDIA6, and Calr/PDIA3 was found in all instances to yield significant reductions in FVIII productivity. In the cells overexpressing Calr and Calr/PDIA3 it was furthermore found to cause significant reductions in the

growth rate and cell size (data not shown). The transgene expression was measured using qRT-PCR showing clear induction of transgene in the relevant transfectants (Supplementary Figure 6). In order to validate the findings from the transient overexpression experiment, stable cell pools were generated by transfecting Clone 1with plasmids encoding PDIA6, FKBP2 and the miRNA let-7f-2. Furthermore, an IgG cell line and an EPO-producing cell line were also transfected with plasmid encoding the miRNA let-7f-2. The cell pools grew under selection with a growth rate unchanged from the untransfected mother cell line. The growth characteristics, product titer and specific productivity of the transfectants were unexpectedly unchanged compared to the mother cell line (Supplementary Figure 7).This difference might It needs to be stated that the targets above were only co-expressed under the control of one strong CMV promoter and the targets might thus prove successful under the control of a weaker promoter bringing the dose into a more favorable region for the cell (Tagkopoulos, 2013).

4.1.4 Conclusion

Using both transcriptomics and proteomics we have, to our knowledge for the first time, described the cellular response to heterologous production of FVIII. The data indicate a dose-dependent induction of the UPR, ER-stress and oxidative stress pathways largely induced by BiP binding of unfolded FVIII. It was not seen to be possible to rescue unfolded FVIII solely by overexpression of single chaperones or proteins essential of transport from ER to the Golgi. It appears from the literature and from the data presented here that a bottleneck exists in the ER which is not able to promote folding of a significant portion of the *F8*-mRNA being translated. It is suggested from the data at hand that due to induction of the UPR, PDIs are overexpressed leading to generation of ROS due to a futile cycle of disulphide isomerization on the FVIII retained in the ER, which cause apoptosis.

4.1.5 Materials and methods

RNA sequencing data analysis

14 CHO clones were grown and RNA was extracted 48 hours after seeding. See our earlier work for further information (Kaas et al. 2015). Following identification of the expression levels for all genes in the CHO genome, the pearson's correlation coefficient was calculated for each gene to the productivity (coagulation activity measured at 48 hours into the cultivation normalised by the viable cell count) using R (Ihaka and Gentleman, 1996). Genes were considered to correlate significantly with productivity with Persons's correlation > 0.81 or < -0.81 (constituting two standard deviations from the mean of all measured correlations). Determination of differentially expressed genes was

carried out by EdgeR as genes with a ± 2.0 fold changes, expression levels > 1 FPKM (as elsewhere (Mortazavi et al. 2008)) and a p-value < 0.01.

Ingeniuity pathway analysis

The proteome sequences from mouse, rat and human were downloaded from the Ensembl Biomart (Kinsella et al. 2011). Homologs were defined as the closest hit for each protein in CHO using blast+. FPKM-values, p-value, fold change ratio and the Ensembl-ID's were introduced to IPA and the list of significantly enriched pathways was exported.

Sample preparation for proteomics analysis

Clone 1, 7 and 12 were grown in biological triplicates in 75 ml cultures with same experimental conditions and media as when extracting RNA samples for RNA sequencing (Kaas et al. 2015). After 48 hours of growth the cells 2-3e6 cells were spun down for 10 minutes at 1200RPM, washed with PBS twice followed by 5min of centrifugation at 5000RPM. The pellets were resuspended in SDS-lysis buffer made up by: 2% SDS (w/v), 0.1 mM phenylmethylsulfonyl fluoride (Sigma-Aldrich) and 1 mM EDTA (Sigma-Aldrich), adjusted to pH=8 with triethylammonium bicarbonate (Sigma-Aldrich). The suspension was sonicated on ice for 60 seconds three times. The protein concentration was measured with the BCA assay (Thermo scientific). 200 µg of protein lysate was reduced in TCEP (final concentration of 10mM) at

at 60°C for 1 hour with shaking at 750 rpm. The lysate is cooled to room temperature and a final concentration of 17.05 mM iodoacetamide is added in order to alkylate the sample. 90 microgram of lysate was diluted by 9M sequanol grade urea (Thermo Scientific) in order to decrease the SDS concentration to 0.09% and left to react in the dark for 1 hour.The sample was enriched on the membrane of a spin filter (Amicon Ultra-0.5 mL, Milipore) by centrifugation at 14000 xg for 10 minutes retaining all fragments > 10kDa. The membrane was washed with 9M urea and spun down at 14000 xg for 10 minutes three times. The membrane was further washed two times with 50mM TEABC. The lysate on the membrane was digested by adding 8.5µg Trypsin/Lys-C mix (Promega) dissolved in 50mM TEABC, pH set to 7.8 and left to incubate over night at 37°C. The filter was turned around and the lysate was spun down at 1000 rpm for 2 minutes. The filer was subsequently turned back, 300µl TEABC was added and the filter was left for 5 minutes before the filter was spun down at 14,000 xg for 10 minutes. The two lysates were combined and digestion was validated using silver staining. Lysates were dried in a speed vac (Savant, Thermo Scientific) for ~3hours until dry. The samples were resuspended in 80% anhydrous acetonitrile (Sigma-Aldrich)

and split into two samples of equal volume and dried using a speed vac. Samples were resuspended in 100µl 100mM TEABC and 40µl anhydrous acetonitrile. TMT labels were added to the samples using the TMT Mass Tagging Kits and Reagents kit (Thermo scientific) following manufacturer's instructions. The nine lysates were run on two TMT runs. Run1: Clone1.1, Clone1.2, Clone7.1, Clone7.1, Clone12.1, Clone12.2. Run2: Clone1.1, Clone1.3, Clone7.1, Clone7.3, Clone12.1, Clone12.3.

Fractionation of peptides and LC-MS/MS analysis

TMT labeled peptides were fractionated on basic RPLC (bRPLC) column: XBridge C18, 5 µm 100 x 2.1 mm analytical column (Waters, Milford, MA); XBridge C18, 5 µm 2.1 x 20 mm Guard column (Waters, Milford, MA). Total 24 fractions were generated by concatenation of 96 bRPLC fractions. Peptide samples were analyzed on Q-Exactive instrument (Thermo scientific) interfaced with Proxion nanoflow LC system. Peptides were fractionated by reverse-phase HPLC on a 75 µm x 15 cm PicoFrit column with a 15 µm emitter (PF3360-75-15-N-5, New Objective) in-house packed with Magic C18AQ (5 µm, 120Å) using 0-60% acetonitrile/0.1% formic acid gradient over 90 min at 300 nl/min. ESI voltage 2.4 kV. Survey scans (full ms) were acquired from 350-1,800 m/z with up to 15 peptide masses (precursor ions) individually isolated with a 2 Da window with offset 0.5 Da and fragmented (MS/MS) using a collision energy of 30 and 30 s dynamic exclusion. Precursor and the fragment ions were analyzed at 70,000 and 35,000 resolution (at m/z =200), respectively.

Identification of Chinese Hamster proteins

Peptide sequences were identified from isotopically resolved masses in MS and MS/MS spectra extracted with and without deconvolution using Thermo Scientific MS2 processor and Xtract software. Data was searched against *Cricetulus griseus* database (RefSeq annotation of *C. griseus* GCF_000419365.1, downloaded 1st of September 2014) with oxidation on methionine, deamidation on residues N and Q , (as different variable modifications) and carbamidomethyl on cysteine (fixed), 6-plex TMT on lysine and N-term (fixed) as modifications using Sequest software interfaced in the Proteome Discoverer 1.4 (Thermo scientific) workflow. Mass tolerances on precursor and fragment masses were 25 ppm and 0.05 Da, respectively. Only protein with at least one unique peptide identified with FDR<0.01 was included in the dataset.

Identification of significantly enriched pathways

Only proteins detected in both proteome runs (4187 proteins) were used for analysis. The relative proteins concentrations were used to calculate the significance using unpaired two sample student t-test assuming unequal variance using R. The fold change ratio was calculated for each protein using the mean value across the biological replicates. The nearest homologous protein in human, rat or mouse was found, as given above for the RNA sequencing data. A protein was set to be differentially expressed if the fold change were higher than ± 2.0 with a p-value < 0.05. The p-value, fold change ratio and the Ensembl-ID's were introduced to IPA and the list of significantly enriched pathways was exported**.**

CHO specific secretion network

Based on literature search a secretion network was manually built around the presence of secretion relevant genes in the CHO genome. Network was visualized in Cytoscape version 3.2.1(Shannon et al. 2003). Color of nodes were set based on log2(foldchange) from Clone 1-11 versus Clone 12-14. Thickness of lines encircling nodes were increased by p-value when < 0.05.

miRNA sequencing data analysis

miRNA library preparation was carried out by AROS Biotechnology by the TruSeq Small RNA Sample Preparation Kit (cat RS-200-0012, Illumina) using the same totalRNA samples used for mRNA sequencing. Only biological triplicates from Clone 1, 7 and 13 were used for miRNA sequencing. RNA PCR primer, index 1 sequence was removed from the raw sequencing data using CLC genomic workbench (version 7.5) and the miRNAs were search against miRBase (Release 21) for *Cricetulus griseus, Rattus norvegicus, Homo sapiens* and *Mus musculus* . miRNAs having a Pearson correlation > 0.8 to the productivity and p-value < 0.05 (student t-test) between Clone 1 versus clone 8 and clone 8 versus clone 13 were chosen for cloning. miRNA stemloop sequence was extracted from miRBase (Griffiths-Jones et al. 2006) and synthesized as two complimentary oligoes. The oligoes were inserted as decriebed elsewhere (Jadhav et al. 2013) into the BLOCK-iT^{TM} Pol II miR expression system (Invitrogen Inc., Carlsbad, CA) downstream of GFP. Primers used for the 4 insertion events: let-7d-fw TGCTGTCGGTTTGTAGGCAGTGTAATTAG CTGATTGTACCGCGGTGCTGACAATCACTAACTCCACTGCCATCAAAACAAGGC. let-7d-rv CCTGGCCTTGTTTTGATGGCAGTGGAGTTAGTGATTGTCAGCACCGCGGTACAATC AGCTAATTACACTGCCTACAAACCGAC. let-7d-fw TGCTGCTCCTAGGAAGAGGTAGTAG GTTGCATAGTTTTAGGGCAGGGATTTTGCCCACAAGGAGGTAACTATACGACCTGCTG CCTTTCTTAGGGCCTT. let-7d-rv CCTGAAGGCCCTAAGAAAGGCAGCAGGTCGTATAGT

TACCTCCTTGTGGGCAAAATCCCTGCCCTAAAACTATGCAACCTACTACCTCTTCCTAG GAGC. let-7f-fw TGCTGTCTATCAGAGTGAGGTAGTAGATTGTATAGTTGTGGGGTAGT GATTTTACCCTGTTCAGGAGATAACTATACAATCTATTGCCTTCCCTGAGGAGTAGAC. let-7f-rv CCTGGTCTACTCCTCAGGGAAGGCAATAGATTGTATAGTTATCTCCTGAACAG GGTAAAATCACTACCCCACAACTATACAATCTACTACCTCACTCTGATAGAC. mir-181c-fw TGCTGTTTGGGGGAACATTCAACCTGTCGGTGAGTTTGGGCAGCTCAGACAAACCA TCGACCGTTGAGTGGACCCCGAGGCCTG. mir-181c-rv CCTGCAGGCCTCGGGGTCCACT CAACGGTCGATGGTTTGTCTGAGCTGCCCAAACTCACCGACAGGTTGAATGTTCCCCCA AAC.

Gene synthesis and transient overexpression

The CHO genes encoding: PDI, PDIA3, PDIA6, FKBP2, FKBP11, NRF2, TGIF2, Bach1, LAMN1 and MCDF2 were synthesised (Genescript, NJ, USA) further codon optimized for CHO. Calr was not synthesized but amplified from cDNA extracted from CHO DXB11 cells. The genes were subsequently PCR amplified using Phusion polymerase master mix (New England Biolabs) and inserted into a pTT22 backbone with puromycine selection using In-Fusion™ cloning (Clontech, CA, USA) following manufacturer's instructions (see Supplementary Figure 8). Correct insertion was validated using Sanger sequencing. All transient transfections were performed using the Nucleofector II (Lonza, Basel, Switzerland) with Nucleofector Kit V using program U-024. 2 million exponentially growing Clone 1 or Clone 7 cells were spun down for 8 min at 100 x g and resuspended in Nucleofector Solution (including supplement) and the molar equivalent of 4µg of DNA from a 6kb plasmid was added. After transfection the cells were left to grow for 48 or 72 hours in 4.5 ml of HyClone CDM4CHO media (Thermo). The viable cell density was measured at 0, 24, 48 and 72 hours using the Vicell XR system (Beckman Coulter) and RNA was extracted from approximately 2mio viable cells at the end of the experiment. Transfection efficiency was estimated using GFP Transfection Efficiency Assay (Chemometec, Allerod, Denmark) using samples transfected with miRNA vectors containing emGFP. Cells were counted every 24 hours and samples were extracted at 0h and 72h for ELISA and chromogenic activity using an in-house version of the Coatest SP (Chromogenix, Instrumentation Laboratory, Milano, Italy).

Estimation of the specific productivity

The integral of viable cells (IVC) for the first 72 hours of growth where calculated as below, where X is viable cell density as measured by Vicell XR system (Beckman Coulter) and t is the amount of time given in days.

$$IVC = \sum^{n} \frac{X_{i+1} + X_i}{2} \times (t_{i+1} - t_i)$$

The specific productivity of the transfectants was calculated as below using the product titer at two time points divided by the IVC.

$$q_P = \frac{CP_{i+1} - CP_i}{IVC}$$

qRT-PCR validation of transgene expressing in transfectants

Expression level from transgene and endogenous versions of the gene were measured using SYBR green qPCR. RNA was extracted 72h hours after transfection using $1.5x10^6$ cells using TRIzol (Invitrogen) and Direct-zol™ RNA MiniPrep Kit (Zymo Research, CA, USA) following the manufacturer's instructions. 1.5µg of RNA from each sample was used for cDNA synthesis using the iScript cDNA Synthesis Kit. Relevant reactions were set up using QuantiFast® SYBR® Green PCR kit and run on a Mx3000P qPCR System. Expression levels were plotted following the ΔΔCt method normalized to *gapdh*. Primers used for detection of endogenous genes: Fkbp2 (Forward: GGACCACTGTCCCATCAAGT and reverse: TGAAAACAAAGGGCTGGTTC), BiP (Forward: CCTATTCCTGGGTTGGTGTG and reverse: TTGGAGGTGAGCTGGTTCTT), Calr (Forward: GGAACCTGCCGTCTATTTCA and reverse: CCCGTAGAATTTGCCAGAAC), *gapdh* (Forward: AACTTTGGCATTGTGGAAGG and reverse: ACACGTTGGGGGTAGGAACA), PDI (Forward: CTCAAGTGAGGTGGCTGTCA and reverse: CGTGATTCCAAAAGGGATGT), PDIA3 (Forward: GCTTGCCCCTGAGTATGAAG and reverse: TAAGGGTTGGGTAGCCACTG), PDIA6 (Forward: GTACCCAAACCCTCCAATCC and reverse: TCTTCTGAAGCTGGCTGACA) Primers used for detection of transgenes: F8 (transgene position 1676-1695 and 1752-1775), Fkbp2 (Forward: GCACAGGACAGGTCATCAAA and reverse: CTTTTCTCCCTCGCACATTC), GFP(miRNA) (Forward: AAGTCGTGCTGCTTCATGTG and reverse: GAACGGCATCAAGGT GAACT), PDI (Forward: CTGCAGCCGAAACACTGA and reverse: CACATCCTTAAAGAA TCCGATGA), PDIA3 (Forward: TCCCTTCTCCATACGAGGTG and reverse: CAGTTTCTTG TTGGCAGGACT), PDIA6 (Forward: GGAGGATTCGGATCACCAG and reverse: CAGCAG TGCGAATTTCATCT)

Stable overexpression in cell pools

GFP and the linked miRNA (let-7f-2 and negative control miRNA) were moved from the BLOCK-iT™ Pol II miR expression plasmids into the pTT22 plasmid backbone with puromycine selection using In-Fusion™ cloning (Clontech, CA, USA) following manufacturer's instructions. Correct insertion was validated using Sanger sequencing. 10e7 of exponentially growing FVIII producing cells (Clone 1) were transfected with 10µg of plasmid containing let-7f-2 miRNA/negative control miRNA/ PDIA6/FKBP2/ empty pTT22 vector respectively, with three independent electroporation reactions (GenePulser Xcell, Biorad). Furthermore, 10e7 of 26.27.1E7 cells (in-house CHO-K1-SV cell line producing the model antibody B72.3 (Colcher et al. 1981)) and C7 cells (in-house CHO-K1 cell line (Ley et al. 2015) producing Erythropoietin) were transfected with 10µg of let-7f-2 miRNA/negative control miRNA vector respectively with three independent electroporation reactions. All plasmids were linearized prior to transfection with ApaLI (New England Biolabs).

The puromycine concentration was slowly raised until the GFP transfection rates (in relevant cell lines) were >80% measured using the GFP Transfection Efficiency Assay (Chemometec, Allerod, Denmark). The FVIII cells were grown in HyClone CDM4CHO media (Thermo) supplemented with Penicillin/Streptomycin (Gibco), MTX and the final puromycine concentration was 5.0 µg/ml. 26.27.1E7 pools were grown in CD CHO medium (Gibco), supplemented with Penicillin/Streptomycin (Gibco), MSX and 8.0 µg/ml puromycine. C7 pools were grown in in-house made Q-CM105 media with 4mM gln, lipid mix, Vitamin K, Soy hydrolysates and 5.5 µg/ml puromycine.

The stable pools were grown in an orbital shaker at 36.5˚C at a shaking speed of 250 rpm and 8.0% CO2 as 15ml cultures in 50 ml tubes (Cultiflask 50, Sartorius AG, Germany). Samples were drawn each day of cultivation, spun down at 15,000 xg and stored at -80 C° until analysis. Antibody titer analysis was performed on undiluted samples using Dip and Read™ Protein A (ProA) Biosensors measured on an Octet RED96 (ForteBio, CA, USA) following manufacturer's instructions. EPO titer analysis was done on samples diluted 10E-5 using Human Erythropoietin Quantikine IVD ELISA Kit (R&D Systems, MN, USA) following manufacturer's instructions. The EPO titer in mIU/ml was converted to mg/l by the correlation 150 units/µg (R&D Systems, MN, USA). FVIII quantification was carried out as described earlier and made relative to the value of the highest producing sample.

4.1.6 Acknowledgements

We would like to thank Dr. Birgitte Friedrichsen for providing the IgG producing 26.27.1E7 cell line and Daniel Ley / Dr. Ali Kazemi Seresht for providing the EPO producing C7 cell line.

4.1.7 Authors' contributions

CSK carried out the cell work, data analysis and drafted the manuscript. AML generated the CHO specific secretion network and used it to mine the RNA sequencing data. DBH and MJB provided essential guidance through the proteomics work. CK, MRA and GB provided guidance throughout the data analysis and manuscript formulation.

4.1.8 References

Plantier JL, Guillet B, Ducasse C, Enjolras N, Rodriguez M-H, Rolli V, Négrier C. 2005. B-domain deleted factor VIII is aggregated and degraded through proteasomal and lysosomal pathways. Thrombosis and haemostasis 93:824-832.

Brown HC, Wright JF, Zhou S, Lytle AM, Shields JE, Spencer HT, Doering CB. 2014. Bioengineered coagulation factor VIII enables long-term correction of murine hemophilia A following liver-directed adeno-associated viral vector delivery. Molecular Therapy - Methods & Clinical Development 1:14036.

Kaufman RJ, Pipe SW, Tagliavacca L, Swaroop M, Moussalli M. 1997. Biosynthesis, assembly and secretion of coagulation factor VIII. Blood coagulation & fibrinolysis : an international journal in haemostasis and thrombosis 8 Suppl 2:S3-14.

Marquette K, Pittman DD, Kaufman RJ. 1995. A 110-amino acid region within the A1-domain of coagulation factor VIII inhibits secretion from mammalian cells. Journal of Biological Chemistry 270:10297-10303.

Pipe SW, Morris JA, Shah J, Kaufman RJ. 1998. Differential interaction of coagulation factor VIII and factor V with protein chaperones calnexin and calreticulin. The Journal of biological chemistry 273:8537-8544.

Ellgaard L, Helenius A. 2003. Quality control in the endoplasmic reticulum. Nature reviews Molecular cell biology 4:181-191.

Tagliavacca L, Wang Q, Kaufman RJ. 2000. ATP-Dependent Dissociation of Non-Disulfide-Linked Aggregates of Coagulation Factor VIII Is a Rate-Limiting Step for Secretion. Biochemistry 39:1973-1981.

Malhotra JD, Miao H, Zhang K, Wolfson A, Pennathur S, Pipe SW, Kaufman RJ. 2008. Antioxidants reduce endoplasmic reticulum stress and improve protein secretion. Proceedings of the National Academy of Sciences of the United States of America 105:18525-18530.

Kaas CS, Bolt B, Hansen JJ, Andersen MR, Kristensen C. 2015. Deep sequencing reveals different compositions of mRNA transcribed from the FVIII gene in a panel of FVIII producing CHO cell lines. Biotechnology journal.

Brown HC, Gangadharan B, Doering CB. 2011. Enhanced biosynthesis of coagulation factor VIII through diminished engagement of the unfolded protein response. The Journal of biological chemistry 286:24451-24457.

Sriburi RB. 2007. Coordinate regulation of phospholipid biosynthesis and secretory pathway gene expression in XBP-1(S)-induced endoplasmic reticulum biogenesis. J Biol Chem 282:7024-7034.

Zhang K, Kaufman RJ. 2006. The unfolded protein response: a stress signaling pathway critical for health and disease. Neurology 66:S102-S109.

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

Campos-da-Paz M, Costa CS, Quilici LS, de Carmo Simões I, Kyaw CM, Maranh+úo AQ, Brigido MM. 2008. Production of recombinant human factor VIII in different cell lines and the effect of human XBP1 co-expression. Molecular Biotechnology 39:155-158.

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

Garbi N, Tanaka S, Momburg F, H+ñmmerling GJ. 2006. Impaired assembly of the major histocompatibility complex class I peptide-loading complex in mice deficient in the oxidoreductase ERp57. Nature immunology 7:93-102.

Oliver JD, Roderick HL, Llewellyn DH, High S. 1999. ERp57 functions as a subunit of specific complexes formed with the ER lectins calreticulin and calnexin. Molecular biology of the cell 10:2573-2582.

Malhotra JD, Kaufman RJ. 2007. The endoplasmic reticulum and the unfolded protein response. Seminars in cell & developmental biology 18:716-731.

Vembar SS, Brodsky JL. 2008. One step at a time: endoplasmic reticulum-associated degradation. Nature reviews Molecular cell biology 9:944-957.

Oda Y, Okada T, Yoshida H, Kaufman RJ, Nagata K, Mori K. 2006. Derlin-2 and Derlin-3 are regulated by the mammalian unfolded protein response and are required for ER-associated degradation. The Journal of Cell Biology 172:383-393.

Molinari M, Calanca V, Galli C, Lucca P, Paganetti P. 2003. Role of EDEM in the release of misfolded glycoproteins from the calnexin cycle. Science (New York, N Y) 299:1397-1400.

Heinz S, Schüttrumpf JC, Simpson JC, Pepperkok R, Nicolaes GA, Abriss D, Milanov P, Roth S, Seifried E, Tonn T. 2009. Factor VIII-eGFP fusion proteins with preserved functional activity for the analysis of the early secretory pathway of factor VIII. Thrombosis and haemostasis 102:925-935.

Zhang B, McGee B, Yamaoka JS, Guglielmone H, Downes K, Minoldo S, Jarchum G, Peyvandi F, De Bosch NB, Ruiz-Saez A, Chatelain B, Olpinski M, Bockenstedt P, Sperl W, Kaufman RJ, Nichols WC, Tuddenham EGD, Ginsburg D. 2006. Combined deficiency of factor V and factor VIII is due to mutations in either LMAN1 or MCFD2. Blood 107:1903-1907.

Hollestelle MJ, Thinnes T, Crain K, Stiko a, Kruijt JK, van Berkel TJ, Loskutoff DJ, van Mourik J. 2001. Tissue distribution of factor VIII gene expression in vivo--a closer look. Thrombosis and haemostasis 86:855-861.

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

Shih AY, Johnson DA, Wong G, Kraft AD, Jiang L, Erb H, Johnson JA, Murphy TH. 2003. Coordinate Regulation of Glutathione Biosynthesis and Release by Nrf2-Expressing Glia Potently Protects Neurons from Oxidative Stress. J Neurosci 23:3394-3406.

Tagkopoulos I. 2013. Microbial factories under control: Auto-regulatory control through engineered stress-induced feedback. Bioengineered 4:1-4.

Ihaka R, Gentleman R. 1996. R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics 5:299-314.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods 5:621-628.

Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, Kersey P, Flicek P. 2011. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database : the journal of biological databases and curation 2011:bar030.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research 13:2498-2504.

Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. 2006. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic acids research 34:D140-D144.

Jadhav V, Hackl M, Druz A, Shridhar S, Chung CY, Heffner KM, Kreil DP, Betenbaugh M, Shiloach J, Barron N, Grillari J, Borth N. 2013. CHO microRNA engineering is growing up: recent successes and future challenges. Biotechnology advances 31:1501-1513.

Colcher D, Hand PH, Nuti M, Schlom J. 1981. A spectrum of monoclonal antibodies reactive with human mammary tumor cells. Proceedings of the National Academy of Sciences of the United States of America 78:3199-3203.

Ley D, Seresht AK, Engmark M, Magdenoska O, Nielsen KF, Kildegaard HF, Andersen MRr. 2015. Multi-omic profiling of EPO-producing Chinese hamster ovary cell panel reveals metabolic adaptation to heterologous protein production. Biotechnology and bioengineering.

Chapter 5 – Insights into CHO Proteomics

In this chapter, a review is presented, which was published as chapter 19 in the book: *Cell Engineering, Animal Cell Culture* (ISBN: 978-3-319-10319-8). The review concerns the current knowledge within proteomics to study cell culture and discuss in detail the state-of-the-art of proteomics methods used today.

Chapter 19
Proteomics in Cell Culture: From Genomics to Combined 'Omics for Cell Line Engineering and Bioprocess Development

Kelley Heffner, Christian Schroeder Kaas, Amit Kumar, Deniz Baycin-Hizal, and Michael Betenbaugh

Abstract The genetic sequencing of Chinese hamster ovary cells has initiated a systems biology era for biotechnology applications. In addition to genomics, critical 'omics data sets also include proteomics, transcriptomics and metabolomics. Recently, the use of proteomics in cell lines for recombinant protein production has increased significantly because proteomics can track changes in protein levels for different cell lines over time, which can be advantageous for bioprocess development and optimization. Specifically, the identification of proteins that affect cell culture processes can aid efforts in media development and cell line engineering to improve growth or productivity, delay the onset of apoptosis, or utilize nutrients efficiently. Mass-spectrometry based and other proteomics methods can provide for the detection of thousands of proteins from cell culture and bioinformatics analysis serves to identify and quantify protein levels. Optimizations of sample preparations and database development, including a detailed CHO proteome now available, have improved the quantity and accuracy of identified proteins. The applications are widespread and expanding, thus suggesting numerous applications of proteomics and combined 'omics experiments in coming years.

Keywords Proteomics • Genomics • Transcriptomics • Metabolomics • Bioprocess development • Cell line engineering • Bioinformatics

K. Heffner • D. Baycin-Hizal • M. Betenbaugh (✉)
Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, MD, USA
e-mail: beten@jhu.edu

C.S. Kaas
Network Engineering of Eukaryotic Cell Factories, Technical University of Denmark, Kgs Lyngby, Denmark

A. Kumar
Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, MD, USA

Biotechnology Core Laboratory, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD, USA

M. Al-Rubeai (ed.), *Animal Cell Culture*, Cell Engineering 9,
DOI 10.1007/978-3-319-10320-4_19

19.1 Introduction

Chinese hamster ovary (CHO) cells are the production host of choice for many recombinant proteins. Their growth in suspension cell culture is scalable for high-density production of biotherapeutics. As mammalian cells, CHO cells provide glycosylation processing that is typically compatible with humans. In 2012, the majority of top biologics were produced in CHO cells as shown in Table 19.1 (Lawrence and Lahteenmaki 2014). The top biotherapeutic is Humira, produced using the CHO expression system for the treatment of rheumatoid arthritis (Lawrence and Lahteenmaki 2014). Characteristics such as human-compatibility and manufacturing scalability help explain CHO cells' dominating use in biotechnology applications.

Table 19.1 Top selling drugs of 2013

Drug	Company	Indication	Production host
Humira (adalimumab)	AbbVie	Rheumatoid arthritis (RA), juvenile rheumatoid arthritis, Crohn's disease, psoriatic arthritis (PA), psoriasis, ankylosing spondylitis, ulcerative colitis (UC), Behçet syndrome	CHO
Enbrel (etanercept)	Amgen	RA, psoriasis, ankylosing spondylitis, PA, juvenile rheumatoid arthritis	CHO
Lantus (insulin glargine)	Sanofi	Diabetes mellitus type 1	E. coli
Rituxan (rituximab)	Roche	RA, chronic lymphocytic leukemia/small cell lymphocytic lymphoma, non-Hodgkin's lymphoma, antineutrophil cytoplasmic antibodies associated vasculitis, indolent non-Hodgkin's lymphoma, diffuse large B-cell lymphoma	CHO
Remicade (infliximab)	J&J	RA, Crohn's disease, psoriasis, UC, ankylosing spondylitis, Behçet syndrome, PA	CHO
Avastin (bevacizumab)	Roche	Colorectal cancer, non–small cell lung cancer, renal cell cancer, brain cancer (malignant glioma; anaplastic astrocytoma, glioblastoma multiforme)	CHO
Herceptin (trastuzumab)	Roche	Breast cancer, gastric cancer	CHO
Gleevec (imatinib)	Novartis	Chronic myelogenous leukemia, gastrointestinal stromal tumor, acute lymphocytic leukemia, hypereosinophilic syndrome, mastocytosis, dermafibrosarcoma protuberans, myelodysplastic syndrome, myeloproliferative disorders	Small molecule
Neulasta (pegfilgrastim)	Amgen	Neutropenia/leukopenia (NL)	E. coli
Copaxone (glatiramer acetate)	Teva	Multiple sclerosis	Small molecule
Revlimid (lenalidoamide)	Celgene	NL	CHO

An improved understanding of CHO cell physiology resulted from the recently completed genome sequence of CHO cell lines and hamsters (Xu et al. 2011; Lewis et al. 2013; Brinkrolf et al. 2013). From this information, a variety of methods have been used to quantify the genome, transcriptome, proteome, and metabolome. These data sets offer new insights into cell physiology. To complement these studies, Baycin-Hizal et al. completed the proteome of the CHO cell line, including information on intracellular, secreted, and glyco-proteins (Baycin-Hizal et al. 2012). This study complemented the results from the CHO genome (Xu et al. 2011) and provided a codon frequency analysis of the differences between CHO cells and humans for improved expression of therapeutics (Baycin-Hizal et al. 2012). Additionally, this study integrated proteomic and transcriptomic data to analyze pathway changes, such as enrichment of protein processing and apoptosis, and depletion of steroid hormone and glycosphingolipid metabolism (Baycin-Hizal et al. 2012). The complete CHO proteome will enhance our capacity for bioprocess development by increasing knowledge of the most-widely used production host.

Proteins have diverse functions in the cell and are involved in growth, signaling, regulation, and metabolism. Their rapidly changing levels provide important information about subtle changes in the cell that may not be detected at the transcriptome or genome level. A variety of methods are used to generate large, complex data sets, which require processing and analysis to reveal useful information about the cellular phenotype. Production of biologics, such as monoclonal antibodies (mAbs), for therapeutic use requires development of a scalable and consistent bioprocess, ensuring high yield and purity of the biotherapeutics. Quantification of protein levels provides a clearer understanding of cell physiology and can lead to improvements in cell culture for biotechnology applications. Proteomics can also be used during development in order to identify proteins that affect cell culture growth, apoptosis, recombinant protein productivity, and product quality. This information can suggest methods to improve the bioprocess through cell line engineering and rational media formulation.

Recently, combined ‘omics approaches have become more widespread in application. These approaches provide useful insights because direct one-to-one correlations between approaches rarely exist. The combination improves the reliability and accuracy of the results compared to either approach alone.

This chapter highlights how both proteomics and genomics can be used for cell culture applications. Proteomics can provide identification and quantification of thousands of cellular proteins and can be used to increase the understanding of production hosts such as CHO cells. In recent years, the number of published proteomic data sets has fluctuated but expanded gradually as shown in Fig. 19.1. Due to the availability of the CHO genome and proteome and introduction of new techniques to increase protein identification and accuracy, the number of applications of CHO proteomics is likely to grow even larger in coming decades.

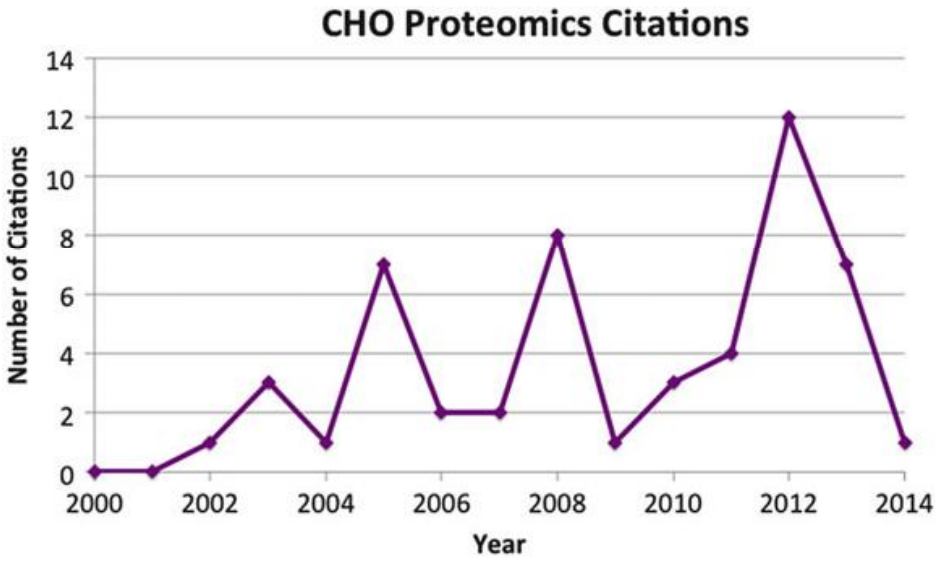


Fig. 19.1 Increase in the number of PubMed citations over time. The number of citations annually is determined by searching "Chinese hamster ovary proteomics" for the specified year (search performed February 2014)

19.2 Genomics

Compared to the genomes of several other mammals such as human (Lander et al. 2001), mouse (Waterston et al. 2002) and rat (Gibbs et al. 2004), the first draft of the CHO-K1 genome was not publicly available until recently (Xu et al. 2011). The sequence revealed 24,383 predicted genes associated with 21 chromosomes with a total of 2.45 gigabytes of genomic sequence (Xu et al. 2011). To facilitate accessibility of the genomic sequence the online database www.CHOgenome.org has been created (Hammond et al. 2012). Before the first publicly available CHO genome came out, the microRNA transcriptome (Hackl et al. 2011) and first transcriptome (Becker et al. 2011) were published as well as several papers using microarrays from mouse without knowledge of the genetic modifications found in hamster yielding suboptimal results (Baik et al. 2006; Yee et al. 2008; Tabuchi et al. 2010; Hernandez Bort et al. 2012). With the genomic sequence available for the CHO-K1 cell line it has now been possible to identify miRNAs (Hackl et al. 2012) and use the sequence for siRNA design (Fischer et al. 2013) and genome editing using Zing finger nucleases (Gaj et al. 2012). One of the main advantages of expressing heterologous protein in CHO has been deciphering the similarity of post-translational modifications of proteins between human and CHO (Kim et al. 2012). Analysis of the CHO-K1 genome revealed that homologs existed for 99 % of the genes in the human genome associated with glycosylation. Of these genes 53 % were detected as transcribed. Furthermore, analysis of the transcriptome revealed that numerous genes associated with viral entry were not expressed, thus explaining the resistance of CHO cells to viral infection (Xu et al. 2011).

In 2013, two groups (Lewis et al. 2013; Brinkrolf et al. 2013) published the genomic sequence of the Chinese hamster (*Cricetulus griseus*) that the CHO cell line was originally extracted from (Puck et al. 1958). The data from the two independent sequencing efforts are presently being merged into a single well-characterized genome, which will be used as the standard reference for the future study of CHO cells lines. The genomic sequence from a number of CHO cell lines (CHO-S, DG44, serum free CHO-K1, CHO protein-free, and C0101) were also published in 2013 (Lewis et al. 2013). Analysis identified more than 3.7 million point mutations in these cell lines compared to the Chinese hamster, highlighting the mutagenesis that has occurred in the process of creating the various cell lines.

19.3 Proteomics

The applications of proteomics in cell culture applications are now widespread. However the use of proteomics for cell culture has also coincided with advances in methods, such as the optimization of sample preparation, digestion, labeling, mass spectrometry (MS), and bioinformatics. Proteomics is increasingly applied to understand cell lines and aid in cell line engineering and process optimization efforts to increase cell growth, increase recombinant protein productivity, and maintain high product quality.

19.3.1 Optimization of Proteomics Methods

Following the initial proteomics experiments in CHO cells, there have been considerable refinements in the methods in order to improve the recovery of cell proteins and enhance their identification. The ability to elucidate increasing numbers of proteins with high accuracy is dependent on optimized sample preparation methods. Proteomics methods include extraction, reduction, alkylation, digestion, and peptide fractionation prior to quantification with MS as shown in the workflow in Fig. 19.2. After cell culture pellets are generated, the cells are lysed and proteins extracted. Following reduction, alkylation, digestion, and fractionation, the individual peptides can be identified. This requires the use of search engines and databases to match the mass spectra to specific peptide sequences and finally proteins. In recent years, there have been numerous improvements to proteomic methods including the optimization of sample preparation, digestion, labeling, and mass spectrometric analysis.

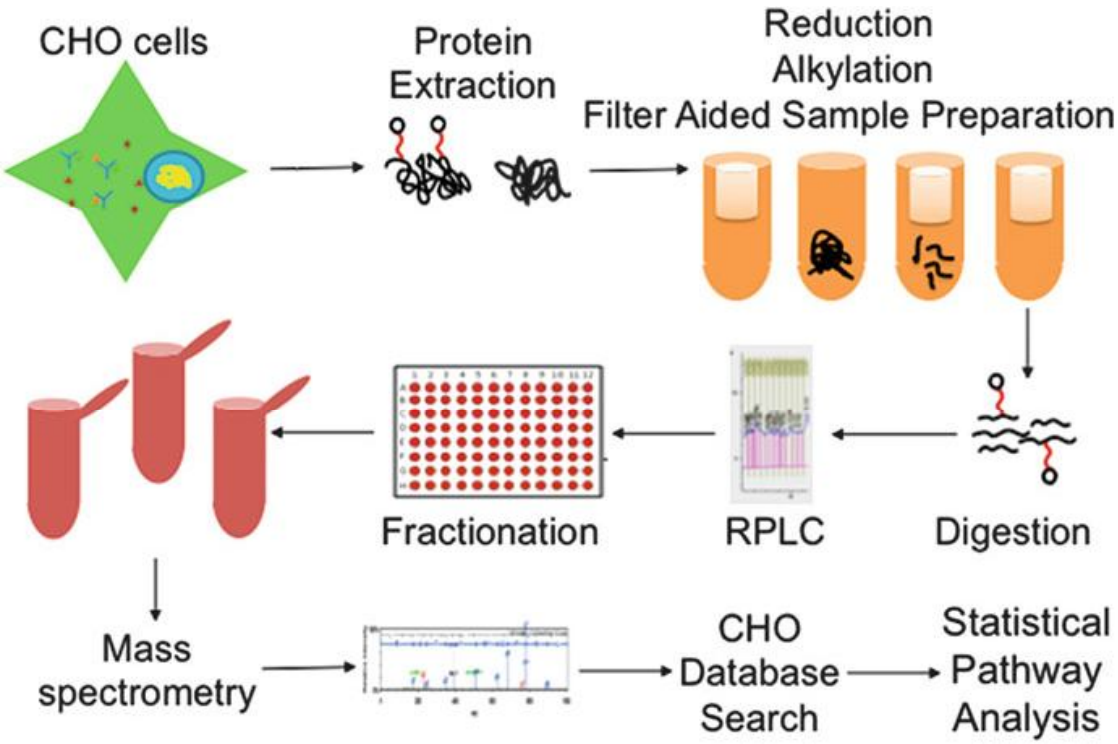


Fig. 19.2 Overview of proteomics workflow. After proteins are extracted from cell culture, sample preparation involves reduction, alkylation, filter aided sample preparation, and fractionation. Peptides are injected into the mass spectrometer and the resultant peaks are analyzed. Proteins are identified and quantified using CHO-specific databases

19.3.1.1 Sample Preparation Methods and Improvements

Initial sample preparation involves using different extraction and digestion techniques. One method involves the recovery of proteins from two-dimensional (2D) gels which serve as an initial separation technique. In order to improve protein recovery, the concentration of solubilizers such as urea, DTT, CHAPS, and SDS have been optimized (Valente et al. 2012). Maintaining solubility enables recovery of proteins with diverse physical and chemical properties. The optimum solubilizing factors for CHO cells include DTT, urea-DTT cross-interaction, and urea-CHAPS cross-interaction for improved protein recovery (Valente et al. 2012). A final solution composition of 8 M urea, 32.5 mM DTT, and at least 2 % CHAPS was selected for maximal recovery of CHO cell lysates (Valente et al. 2012).

Besides the extraction efficiency, digestion efficiency is another criteria for increasing the number of proteins identified. In-gel and in-solution digestion methods have separate and distinct advantages. In-gel digestion involves solubilizing proteins with detergent and separating proteins by gel electrophoresis. Isolated proteins are then digested from the gel and quantified by MS. In-solution digestion involves extracting proteins with strong reagents and digesting in the solution. There are advantages and disadvantages of each method. In-gel digestion protects

against impurities but the protein yield is typically poor in comparison to in-solution digestion. On the other hand, in-solution digestion is easier to implement but there is greater risk of impurities or incomplete solubilization. Sodium dodecyl sulfate (SDS) is one of the principal detergents used for full extraction of the cell lysates including insoluble membrane proteins; however, SDS has to be removed prior to MS analysis. Filter aided sample prep (FASP) was developed to remove the SDS prior to digestion and MS (Wisniewski et al. 2009). The addition of this step allows for more complete coverage of the proteome, and the inclusion of FASP was used to maximize the protein recovery during the CHO proteome analysis prior to trypsin digestion (Baycin-Hizal et al. 2012).

Both in-gel and in-solution digestions continue to be used for biotechnology applications and there are various examples. In-gel digestion was used to identify a number of proteins during cell line engineering efforts (Baik et al. 2008, 2011; Van Dyk et al. 2003) and also aided in quantifying differences between cell lines (Kuystermans et al. 2010; Beckmann et al. 2012). In-gel digestion was also used to identify phosphorylated proteins from CHO-K1 cell culture (Hayduk et al. 2004) and to elucidate the proteome of the CHO DG44 cell line (Lee et al. 2010).

In-solution digestion can help to address some of the limitations of in-gel digestion such as difficulties in separating proteins with low molecular weight, high molecular weight, or hydrophobic properties. Meleady used in-solution digestion to profile protein levels in cell lines with or without miR-7 overexpression (Meleady et al. 2012a). In-solution digestion was also used to identify secreted proteins from the CHO-S and CHO DG44 cell lines (Slade et al. 2012).

In other experiments, a combination of both in-gel and in-solution digestions are used. Both digestion techniques were used to prepare protein fractions prior to MS analysis (Baycin-Hizal et al. 2012) and were also used directly prior to LC/MS injections (Meleady et al. 2012a). Although digestion methods can vary, these can be important for preparing peptide samples for MS injection. An optional step that is now widely used for comparative proteomics is the labeling of digested peptides as will be discussed in more detail below.

Following digestion, fractionation can be applied to enhance protein identification. In one example, basic reversed phase liquid chromatography was used to separate samples into 96 fractions that were then combined into 48 fractions for MS analysis (Baycin-Hizal et al. 2012). For some applications, proteins related to a specific organelle or intracellular and extracellular compartments can be targeted. Two recent examples include the secretome and mitotic spindle, as discussed next.

The secretome includes host cell proteins (HCP) that must be removed prior to formulation of the final drug product. Identification of secreted proteins is often limited by their low abundance. A design of experiments approach was used to optimize sample preparation methods and to increase protein recovery for HCP identification (Valente et al. 2014). Precipitation parameters such as the precipitant chemical, precipitant concentration, and incubation length, were evaluated for both gel-based and shotgun proteomics (Valente et al. 2014). The results were used to optimize a method for identification of HCPs, which differ in physical and chemical properties, as well as their physiological function; this method used methanol

precipitation to identify 178 HCPs, including clusterin, beta-actin, glyceraldehyde-3-phosphate dehydrogenase, and immunoglobulin superfamily member 8 (Valente et al. 2014). Optimization of sample preparation is critical to maximize the recovery of low abundance secreted proteins.

The mitotic spindle proteins were also characterized in CHO cells (Bonner et al. 2011). Cell division is an important event to study, as it relates to the growth of cell lines. Isolation of the spindle aids in identification of factors affecting division and thus growth. After synchronizing all cells in metaphase, over 1,100 proteins were identified, of which 239 proteins were cell division factors and 841 proteins were involved in early stages of division (Bonner et al. 2011). Of the proteome, 11 % of proteins localized to the membrane, 7 % were associated with microtubules, and 3 % were associated with actin (Bonner et al. 2011). Identification of cell division factors may aid bioprocess development due to their importance in cell growth and ultimately recombinant protein yields.

19.3.1.2 Protein Labeling

Labeling strategies are used in proteomics to provide relative quantification in addition to the identification of proteins. Recent methods for labeling proteins include stable isotope labeling with amino acids in culture (SILAC), isobaric tags for relative and absolute quantification (iTRAQ), and tandem mass tags (TMT). Various labeling methods as well as label-free methods are now used for relative quantification.

In SILAC, proteins are labeled when amino acids from the culture medium are incorporated into the cell. Incorporation of amino acid labeling into proteins can be detected by MS. SILAC experiments involve cell adaptation to labeled media, cell growth, protein identification by MS, and data analysis (Harsha et al. 2008). Both in-gel and in-solution digestion can be used to extract proteins prior to MS. SILAC labeling experiments provide information on protein dynamics because they quantitate the incorporation of labeled amino acids. However, this method introduces experimental errors, such as incomplete amino acid isotope incorporation and sample mixing errors (Park et al. 2012). To reduce error, label-swap replication experiments were used to average ratios of individual replicates and validated with a triplet experiment (Park et al. 2012). The approach corrected for incomplete labeling and arginine to proline conversion, thus providing consistent experimental ratios (Park et al. 2012).

An alternative labeling method is iTRAQ, which is used to label peptides before running the sample on MS. The N-terminus and primary amine groups of digested peptides are covalently labeled in this technique (Wiese et al. 2007). The isobaric mass design of the labeling reagents provides the quantification in the MS/MS spectra with tag-specific reporter ions (Wiese et al. 2007). The benefit of this method is that cells do not require adaptation to labeled medium because the labeling is performed on protein extracts. It is important to equalize masses of iTRAQ reagent added across samples because the quantification is relative.

Besides iTRAQ, the mTRAQ technique can allow direct quantification at MS1 level by providing mass differences in precursor ions (Mertins et al. 2012). A comparison of iTRAQ versus nonisobaric labeling (mTRAQ) revealed that iTRAQ labeling identifies over double the total number of proteins and approximately triple the phosphopeptides as compared to mTRAQ (Mertins et al. 2012). Additionally, kinase identification was significantly increased using iTRAQ, thus improving knowledge about protein-protein interactions involved in recombinant protein production (Mertins et al. 2012). In addition to iTRAQ, TMT is another widely used isobaric labeling technique, which provides the measurement of the intensities at the peptide fragmentation level. The greatest advantage of these labeling techniques is they can provide multiplexing of up to eight samples (iTRAQ) and ten samples (TMT), which significantly decreases not only LC-MS/MS run time but also the variations between the samples in the whole process (Megger et al. 2013). All of these proteomics experiments can yield novel insights into CHO cells that lead to a better understanding of their use as biopharmaceutical production hosts.

19.3.1.3 MS

High quality digested peptides are injected into the MS for protein identification. Components of the MS instrument include the source, which produces gas phase ions from the sample; mass analyzer, which resolves the ions based on mass to charge ratio; and detector, which detects ions that have been resolved by the analyzer. Both tandem MS and matrix-assisted laser desorption/ionization time of flight (MALDI TOF) MS have been used to identify and quantify proteins in cell lines with different growth or productivity characteristics (Doolan et al. 2010; Beckmann et al. 2012; Van Dyk et al. 2003; Baik et al. 2008, 2011; Lee et al. 2010; Hayduk et al. 2004).

Peaks from the MS/MS spectra are next identified as specific peptides that are ultimately attributed to specific proteins. Proper identification requires the use of search engines and databases. Different search engines such as TagRecon and MyriMatch (Baycin-Hizal et al. 2012) have been combined to match the CHO genome database. Currently, many open source search engines such as X!Tandem and OMSSA, as well as many proprietary identification programs such as Mascot and SEQUEST, exist for matching the identification and quantification of MS/MS spectra. Due to algorithm differences, identifications from each search engine may show slight variances; for that reason coupling of multiple search engines can increase the confidence levels.

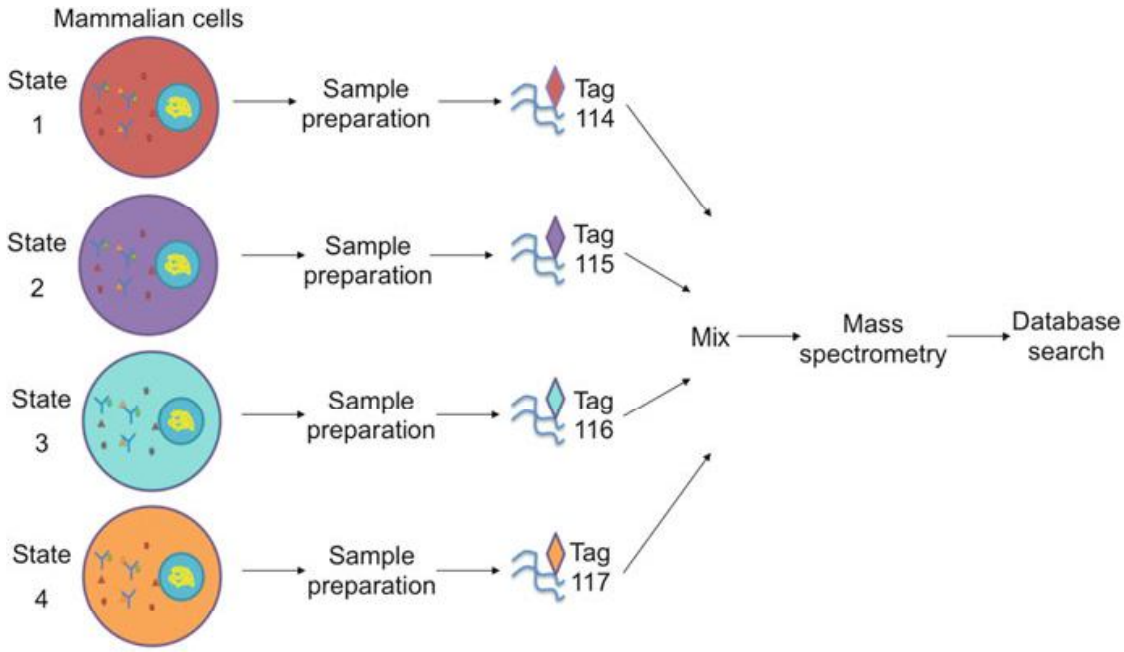


Fig. 19.3 Overview of comparative proteomics. Comparative proteomics can be used to identify differences in protein expression levels between culture conditions. Following sample preparation, peptides are labeled separately with unique tags and then mixed in equal amounts prior to MS injection. Peptides are identified and quantified using CHO-specific databases

19.3.2 Proteomics for Bioprocess Development

Recently, proteomics has been applied to identify proteins that play key roles in recombinant protein production for use in bioprocess optimization. A comparison of high and low expressors as shown in Fig. 19.3 can provide significant insights and understanding of cell properties that are important for improving product yields.

In recent years, comparative and label-free proteomics have been used in a number of studies to gain insights about CHO cells used in bioprocessing. Table 19.2 summarizes some of the recent publications in this area.

19.3.2.1 Proteomics Analysis to Increase Cell Growth Rate and Viable Cell Density

A critical bioprocess development goal is to maximize growth rate in order to increase the amount of biotherapeutic produced over a given time period. An approach combining transcriptomics and proteomics was enacted to identify candidates for a high growth phenotype (Clarke et al. 2012). The proteomics results indicated that 285 proteins were differentially expressed between cell lines with fast and slow growth rates (Clarke et al. 2012). Benefits of combining the 'omics approaches include accounting for low abundance protein expression and

Table 19.2 Summary of CHO proteomics for bioprocess development

Reference	Purpose	Method	Conclusion
Baik et al. (2008)	To determine differences in proteins between CHO cell cultures with or without sodium butyrate supplementation	Used in-gel digestion for proteins. Used MALDI TOF MS and MS/MS to identify proteins	Identified increased levels of GRP78 and peroxiredoxin following treatment with sodium butyrate. Phosphopyruvate hydratase levels decreased with sodium butyrate treatment
Baik et al. 2011	To determine differences in protein levels during adaptation to serum-free medium	Used in-gel digestion for proteins. Used MALDI TOF MS and MS/MS to identify proteins	Identified increased levels of HSP60 and HSC70 in serum-free media. Subsequent overexpression resulted in improved cell concentration during serum-free adaptation
Baycin-Hizal et al. (2012)	To identify the proteome, secretome, and glycoproteome of CHO-K1 cell line	Used SPEG to generate glycoprotein fractions and in-gel and in-solution digestion for proteins. Used MS/MS to measure proteome, secretome, and glycoproteome	Identified important proteins in CHO-K1 cell line and combined proteomic and transcriptomic data for improved analysis
Carlage et al. (2009)	To compare the proteomes of high and low producing cell cultures over time	Proteins were digested and identified by LC-MS	Found differentially expressed proteins between cultures. Eukaryotic translation initiation factor 3 and ribosome 40S were upregulated and vimentin, annexin, and histones were downregulated in the high producer
Carlage et al. (2012)	To determine the changes in the proteome over time for a CHO cell culture overexpressing Bcl-xL	Used iTRAQ labeling and LC-MS to identify proteins	Identified proteins with changing levels over time from exponential to stationary transition and related the differences to cell growth and apoptosis
Clarke et al. (2012)	To determine important proteins that elucidate how miRNAs affect CHO cell growth	Used LC-MS to identify proteins	Compared miRNA, mRNA, and protein expression levels and identified processes

(continued)

Table 19.2 (continued)

Reference	Purpose	Method	Conclusion
			regulating cell growth, such as ribosome synthesis, translation, and mRNA processing
Doolan et al. (2010)	To combine transcriptomics and proteomics to identify important proteins that relate to high growth rate	Used in-gel digestion for proteins. Used MALDI TOF MS to identify proteins	Identified valosin containing protein as important regulator of cell growth. Overexpression resulted in improved growth with no decrease in viability
Dorai et al. (2013)	To identify proteins that affect high productivity in bioreactor cell cultures	Collected spent media samples for host cell proteins. Used in-gel digestion for proteins. Used LC-MS to identify proteins	Comparison of high and low expressing clones revealed 180 differentially expressed proteins. Identified proteins related to cytoskeletal organization, protein synthesis, metabolism, and growth
Kang et al. (2013)	To identify differences between cell lines that affect antibody production	Used LC-MS/MS shotgun proteomics to identify protein expression differences between cell lines	Identified proteins with positive correlation to productivity, including DHFR, adaptor protein complex subunits AP3D1 and AP2B2, DNA repair protein DDB1, and ER translocation subunit SRPR
Kim et al. (2011)	To determine the effect of hydosylate supplementation on antibody productivity in serum-free cell cultures	Used 2D gel electrophoresis combined with nano LC-ESI-QTOF MS/MS to identify proteins	Found significant changes in protein expression in serum-free medium supplemented with hydrosylates, such as upregulation of metabolic, cytoskeletal organization, and growth regulated proteins
Kuystermans et al. (2010)	To determine the effect of cMyc on proteome	Used in-gel digestion for proteins. Used MS/MS to identify proteins	Found increase in nucleolin and decreased in regulation of proteins related to matrix and cell adhesion. Also found increased ATP synthetase and mitochondrial protein levels

(continued)

Table 19.2 (continued)

Reference	Purpose	Method	Conclusion
Lee et al. (2010)	To identify the proteome of CHO DG44 cell line	Used in-gel digestion for proteins. Used MALDI TOF MS and MS/MS to identify proteins	Improved protein identification by enrichment of medium and low abundance proteins. Most identified proteins function in energy metabolism
Lim et al. (2013)	To identify growth factors that can be used as supplements in serum-free cell cultures to improve antibody production	Used MS shotgun proteomics to identify proteins in spent medium	Identified 290 secreted proteins from CHO cell culture including 8 novel growth factors. Used growth factors as medium supplements to increase cell growth rate
Meleady et al. (2011)	To compare the proteomes of high and low producing cell cultures over time	Used in-gel digestion for proteins. Used LC-MS to identify proteins	Identified 89 proteins with differential expression between high and low producer cell lines
Meleady et al. (2012a)	To determine the effect of miR-7 overexpression on proteome	Used in-solution digestion for proteins. Used label-free LC-MS to identify proteins	Identified 93 decreased proteins and 74 increased proteins resulting from miR-7 overexpression. Decreased proteins related to protein translation and DNA/RNA processing. Increased proteins related to protein folding and secretion
Meleady et al. (2012b)	To improve identification of proteome by using multiple databases	Used in-gel and in-solution digestion for proteins. Used MALDI TOF MS and electrospray ion trap MS to identify proteins	Improved protein identification by 40–50 % through multiple CHO-specific databases
Slade et al. (2012)	To develop a method for identifying secreted proteins	Cell culture medium supplemented with GalNAz and enriched by copper catalyzed click chemistry. Used iTRAQ labeling and MS to identify proteins	Identified differences between CHO-S and CHO DG44 secreted proteins. Found 70 % similarity between cell lines
Van Dyk et al. (2003)	To identify important proteins that affect protein productivity in butyrate and zinc treated culture	Used in-gel digestion for proteins. Proteins identified with MALDI-TOF MS	Identified increased expression of GRP75, enolase, and thioredoxin in response to media supplements

(continued)

Table 19.2 (continued)

Reference	Purpose	Method	Conclusion
Wei et al. (2011)	To identify the proteome of CHO cells during prolonged cultivation	Used in-gel digestion for proteins. Electrospray ionization tandem MS used to identify proteins	After prolonged cultivation, identified 40 proteins with different expression levels related to cytoskeletal proteins, chaperones, and metabolic enzymes

identifying mRNA post-translational processing, thus reducing error (Clarke et al. 2012). The most significantly enriched gene ontology terms were translational elongation, translation, generation of precursor metabolites and energy, oxidation-reduction, and aerobic respiration (Clarke et al. 2012). Combining 'omics strategies improves the confidence in the findings from either data set and provides useful information about the culture conditions that improve growth.

Another study used a combined transcriptomics and proteomics approach to compare cell lines with fast and slow growth (Doolan et al. 2010). Valosin-containing protein (VCP) was shown to have a significant effect on cell growth and viability. This finding was confirmed by silencing VCP expression, which resulted in decreased viable cell density and viability (Doolan et al. 2010).

Proteomics was also used to compare CHO cell lines with and without the cMyc gene, which had been previously shown to improve cell growth rate and concentration (Kuystermans et al. 2010). Over 100 proteins were differentially expressed between cultures. Culture performance improved as measured by the increase in nucleolin (important for growth, preventing apoptosis, protein productivity, and energy utilization) and decreases in regulation of adhesion proteins (Kuystermans et al. 2010). Specific components of ATP synthetase were up-regulated, which may indicate a change in energy utilization to release more ATP in cMyc cultures (Kuystermans et al. 2010). These identified proteins may now be further investigated through hypothesis-driven research approaches.

One way to understand cell death is to elucidate what happens at the transition of cell culture from exponential growth to stationary phase. Comparative proteomics, using iTRAQ labeling, identified 59 proteins with significantly different protein expression levels between the exponential and stationary phases (Carlage et al. 2012). Some of these proteins included binding immunoglobulin protein, protein disulfide isomerase, DNA replication licensing factors MCM2 and MCM5, transglutaminase-2, and clusterin (Carlage et al. 2012). These time points were compared in order to identify proteins associated with cell growth and apoptosis. By classifying differentially expressed proteins, it was found that both growth and apoptotic proteins are expressed highly during the stationary phase (Carlage et al. 2012). Results from this study will facilitate a better understanding of the dynamic changes that occur during the different stages of cell culture and may be

used to prolong cell growth and protein production during the later stages of a bioprocess.

Another proteomics study examined protein samples collected throughout the cell culture process (Wei et al. 2011). This study also identified 40 differentially expressed proteins between exponential and stationary phases (Wei et al. 2011). The results indicate that over time, apoptosis occurs due to the initiation of the unfolded protein response (Wei et al. 2011). Thus, delaying apoptosis can increase protein production. This was validated in other studies including one in which co-expression of the anti-apoptotsis gene Bcl-xL increased membrane protein expression in CHO cells (Ohsfeldt et al. 2012).

19.3.2.2 Proteomics Analysis to Increase Recombinant Protein Production

An important goal in bioprocess development is to maximize the recombinant protein yields. In one study, proteomics was used to identify proteins related to improved protein productivity observed following supplementation of butyrate and zinc sulfate to the culture medium (Van Dyk et al. 2003). Increased expression was observed for metabolic and chaperone proteins including GRP75, enolase, and thioredoxin (Van Dyk et al. 2003). These proteins were thus identified as potential cell engineering targets for improving recombinant protein production.

Another study compared high and low producing cell cultures. Proteomics analysis showed that over 30 proteins were differentially expressed between high and low producing cultures (Carlage et al. 2009). In the high producing cell line, eukaryotic translation initiation factor 3 and ribosome 40S were upregulated, whereas vimentin, annexin, and histone H1.2/H2A were downregulated (Carlage et al. 2009). Additionally, the chaperone binding immunoglobulin protein was upregulated in the high producing cell line to suggest that the unfolded protein response occurs as a consequence of endoplasmic reticulum stress (Carlage et al. 2009).

A combination of proteomics and transcriptomics was used to determine the effect of low temperature and sodium butyrate on recombinant protein production in CHO cells (Kantardjieff et al. 2010). This approach relied on the transcriptomic information to identify hundreds of differentially expressed genes between treatments. From this data, proteomics was used to further identify different protein levels. Butyrate treatment and low temperature were shown to improve recombinant protein production rates by improving cell secretory capacity (Kantardjieff et al. 2010). Enriched pathways included Golgi processing, cytoskeleton binding, and GTPase mediated signal transduction (Kantardjieff et al. 2010).

In more recent years, the development of the CHO genome database has led to improvements in protein identifications. Different cell culture conditions were used for proteomics experiments in order to compare high and low producing cell lines (Dorai et al. 2013). From 180 differentially expressed proteins, 12 proteins exhibited differential expression levels over the culture duration in bioreactors,

including ADP-ribosylation factor protein, V-type proton ATPase, colony stimulating factor 1, and angiopoietin 4 (Dorai et al. 2013). These differentially expressed proteins included functions related to growth, metabolism, organization, and protein synthesis (Dorai et al. 2013), which can be used for process optimization by genetic engineering or media supplementation strategies.

Meleady also investigated differences in protein levels between high and low producing cultures (Meleady et al. 2011). Results identified 89 differentially expressed proteins between the high and low producing cultures (Meleady et al. 2011). In particular, 12 proteins were shown to differ in expression level in the same direction, including aldose reductase-related protein 2, annexin, eukaryotic translation initiation factor, glucose-6-phosphate 1-dehydrogenase, endoplasmin, and nuclear migration protein (Meleady et al. 2011). Proteins that were expressed at higher levels in the high-producing cell line included proteins involved in translation and folding (Meleady et al. 2011), which are associated with recombinant protein production.

Overexpression of miRNAs may also improve process development by regulating protein productivity. A proteomic comparison of a cell line overexpressing miR-7 compared with a control cell line revealed differences in protein expression that contributed to higher productivity in the miRNA-engineered cell line (Meleady et al. 2012a). Overexpression of miR-7 resulted in 93 downregulated proteins and 74 upregulated proteins (Meleady et al. 2012a). Proteins with decreased levels were involved in protein translation and DNA/RNA processing and those with increased levels were related to protein folding and secretion (Meleady et al. 2012a), representing potential cell line engineering targets. In addition to protein production, overexpression of miRNAs can affect other processes such as cell growth and apoptosis (Druz et al. 2011, 2013).

The genetic factors contributing to high productivity of a cell line were studied by maintaining the same process conditions (Kang et al. 2013). Results from proteomics indicated a positive correlation between productivity and expression of dihydrofolate reductase, adaptor protein complex subunits, DNA repair proteins, and the endoplasmic reticulum translocation complex components (Kang et al. 2013). Both transcriptomics and proteomics data were combined in order to improve understanding of the high production phenotype (Kang et al. 2013). Differences in expression suggest important roles for these proteins in high producing clones and targets of opportunity for cell line engineering.

19.3.2.3 Proteomics to Optimize Media Formulations

The growth and productivity of a cell line is highly dependent on the media composition. Media formulation is a key component of process development and methods to adjust the media in order to improve product yields and quality are highly sought. The formulation of cell culture medium can have a significant effect on cell growth, viability, and protein production rates. Proteomics analysis can

provide detailed information about cell protein levels that can aid in subsequent medium development.

Recently, proteomics has been used to profile differences between medium formulations in order to correlate improved cell growth with protein expression levels. In one case, proteomics was used to help identify proteins helpful for adaptation from serum-bearing to serum-free media (Baik et al. 2011). Results indicated that two molecular chaperones and four de novo nucleotide synthesis related proteins were significantly increased in the serum-free cell culture (Baik et al. 2011).

Subsequently, two of the chaperones identified (HSP60 and HSC70) were overexpressed and this increased the cell growth rate up to 15 % and decreased adaptation time up to 33 % (Baik et al. 2011). Thus, quantification of protein levels for different media formulations can provide insights into cell line engineering strategies to improve cell growth and medium adaptation.

The secretome consists of extracellular proteins processed through the secretory pathway. These extracellular molecules are in low abundance, but may be involved in diverse biological processes. In one approach, secreted proteins from conditioned media samples were identified in order to develop a serum-free media formulation (Lim et al. 2013). Supplementation of identified growth factors, such as fibroblast growth factor 8, growth regulated alpha protein, hepatocyte growth factor, and macrophage colony stimulating factor 1, to the serum-free cloning media formulation led to increased cell growth (Lim et al. 2013).

Analysis of the secretome can also aid in the identification of proteins that accumulate in the medium over time. N-azido-galactosamine labeling was used to tag the mucin-type O-linked glycans of secreted proteins in order to enable their identification in cell-conditioned media (Slade et al. 2012). This method helped to identify secreted proteins in low abundance and minimize the number of background proteins (Slade et al. 2012). The secretomes of CHO-S and CHO DG44 cell lines were compared and it was observed that 171 proteins were identified in both cell lines (Slade et al. 2012). Close to 70 % of the proteins identified were the same between the CHO-S and CHO DG44 cell lines (Slade et al. 2012). However, there were also 96 proteins unique to CHO DG44 and 85 proteins unique to CHO-S (Slade et al. 2012). Proteins observed at different levels were related to adhesion, cell growth, and proteases (Slade et al. 2012). Important secreted proteins may be investigated as medium supplements in the future. It is also critical that secreted proteins are identified and removed from the final biotherapeutic drug product.

In another experiment, label-free comparative proteomics was used to identify differences between cells cultivated in serum-free medium formulations with or without hydrosylates (Kim et al. 2011). The changes in protein expression upon addition of hydrosylates, containing blends of peptides, free amino acids, vitamins, and trace elements, helped to explain the increased recombinant protein expression (Kim et al. 2011). Proliferative proteins were upregulated whereas pro-apoptotic proteins were downregulated in the culture containing hydrosylates (Kim et al. 2011).

Proteomics can thus be used to identify differences in protein expression between media formulations and to help optimize formulations for high growth and recombinant protein production. These studies can also help elucidate cell engineering targets as well as potential novel media supplements.

19.3.2.4 Systems Biology

A full characterization of the proteome significantly aids efforts to understand cell physiology through the analysis of metabolic pathways. The complete proteome was recently elucidated for the CHO-K1 cell line (Baycin-Hizal et al. 2012). Analysis of genomics, transcriptomics and proteomics data at the systems biology level using KEGG pathway analysis revealed that pathways for protein processing and apoptosis were enriched in CHO-K1, whereas pathways for steroid hormone and glycosphingolipid metabolism were depleted (Baycin-Hizal et al. 2012). The complementary glycoproteomics analysis elucidated major cell adhesion and membrane proteins, which are difficult to identify with intracellular proteomics approaches. Furthermore, Baycin-Hizal et al. used both genomics and proteomics information to generate the codon usage preference tables for CHO cells.

In another experiment, various databases were used to improve the confidence of identified CHO cell proteins (Meleady et al. 2012b). By using multiple databases, it was possible to increase the number of identified proteins by over 40 % (Meleady et al. 2012b). Different methods were used across locations and the results were compared to improve confidence. These studies are facilitating the development of a reliable proteomic profile of CHO cell physiology going forward.

19.3.3 Database Development

In recent years, CHO-specific databases have been developed to improve sequence information availability for protein identification. Prior to this, CHO-specific sequences were not publicly available, requiring researchers to rely on cross-species information. Following the completed genome of the CHO-K1 cell line, the draft Chinese hamster genome was published (Xu et al. 2011; Brinkrolf et al. 2013; Lewis et al. 2013). The published data sets from these genomes served to provide unifying information about the genes and variations between CHO cell lines. Such genomic data sets are now available online and have been compiled into databases such as www.chogenome.org as shown in Fig. 19.4. Similar efforts have been made to establish large-scale proteomic databases for CHO cells, including http://chogenome.org/proteome.php, as shown in Fig. 19.5. This database can be used to find detected proteins and associated accession numbers which includes 6,163 entries. The proteins of interest can be found either by protein name or the accession number, as reported in the work by Baycin-Hizal et al. 2012. Upon clicking on the accession number, users can find

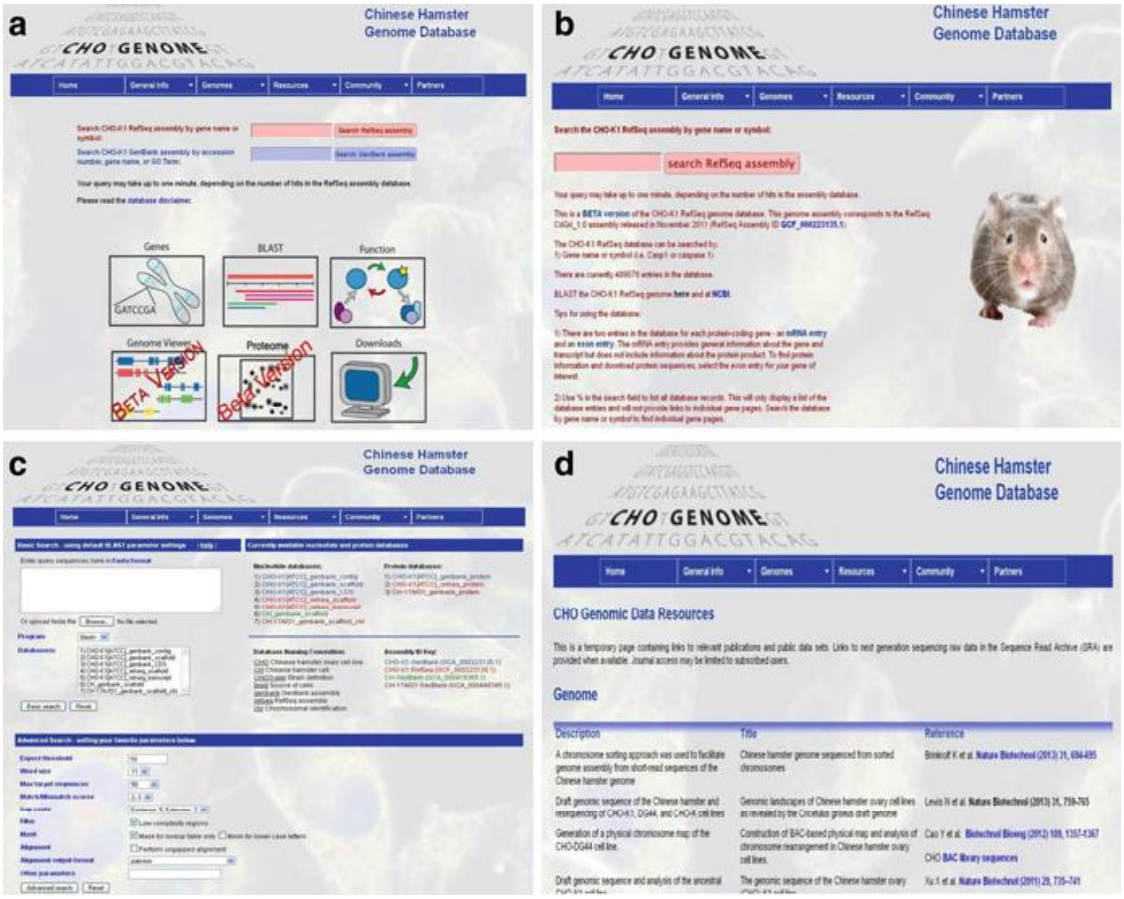


Fig. 19.4 Screenshots of www.chogenome.org, (**a**) shows the home page of the website with links to various other pages, (**b**) shows the genes page where user can search the RefSeq assembly, (**c**) shows the BLAST tool implemented in the website using which user can run CHO specific BLAST queries, and (**d**) shows the webpage from where user can download public datasets and can find links to the relevant publications

more detailed information on the webpage, such as the protein's SwissProt annotation, GO annotation, Kegg annotation, identified peptides sequences, false discovery rate, and more (Baycin-Hizal et al. 2012). The widespread accessibility and ease of use makes the CHO genome website a useful tool for proteomics analysis.

An evaluation based on the CHO databases increased protein identification by 282 proteins, a 40–50 % increase in the total number of proteins identified at that time (Meleady et al. 2012b). CHO sequence information was combined from the SwissProt database, CHO-K1 draft genome (Xu et al. 2011), and the Bielefeld-BOKU-CHO database (Meleady et al. 2012b). Because more peptides matched, there was increased confidence in the results. The increased proteins identified were related to protein translation and energy metabolism, both important for bioprocess development (Meleady et al. 2012b). The compilation of data sets from proteomics experiments around the world allows for data sharing and database generation. Multiple groups can use data to improve proteomics results.

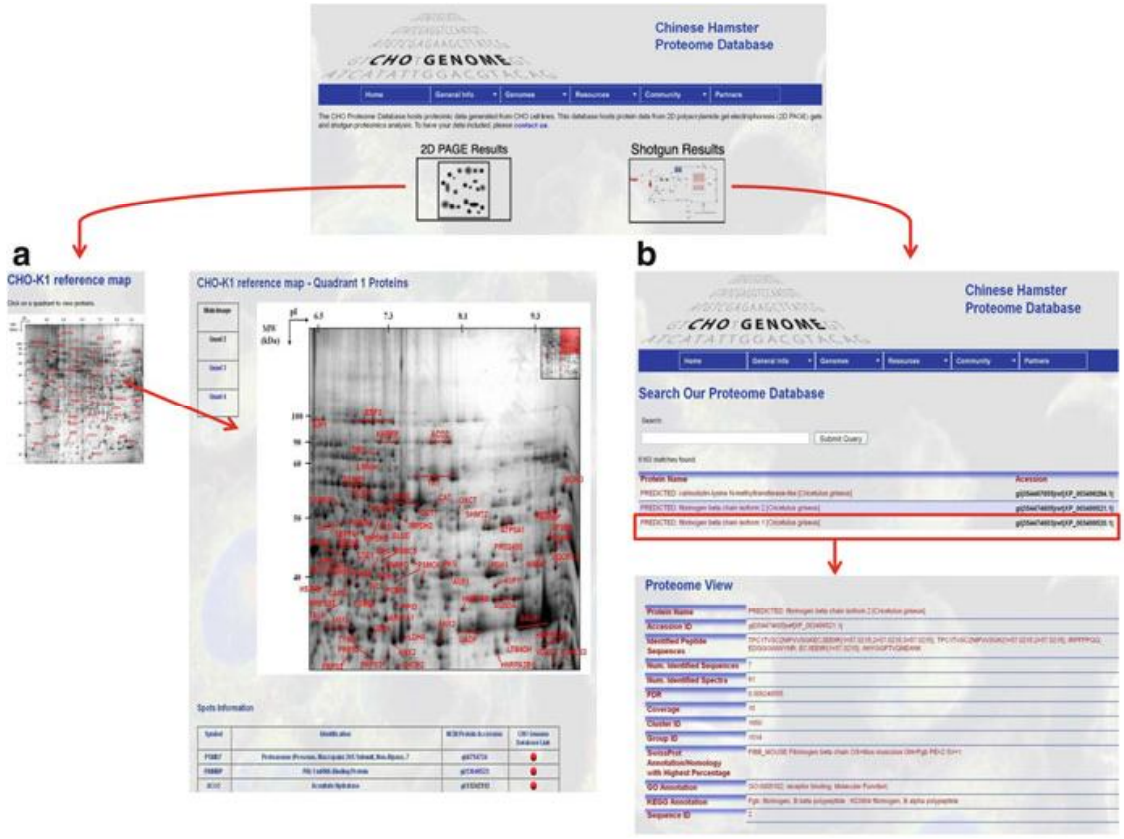


Fig. 19.5 An example output for CHO proteome portion of the www.chogenome.org website. It provides identified proteins in CHO-K1 proteome by 2D polyacrylamide gel electrophoresis as well as by shotgun proteomics. (**a**) Example output for CHO-K1 gel and identified proteins. (**b**) The protein name and accession number after searching for a protein name which can be used to view experimental details about the corresponding protein

19.3.4 Combined Metabolomics and Proteomics

In addition to combined genomics and proteomics or transcriptomics and proteomics, there have recently been examples of combined metabolomics and proteomics. In this case, protein identification confidence increases as protein expression is correlated to changes in metabolite pools. A combination of proteomics and metabolomics was used to increase understanding of CHO cells during prolonged cultivation (Beckmann et al. 2012). The observation that prolonged culture results in increased cell growth was related to an increase in adenylate energy charge (AEC) and differential protein expression (Beckmann et al. 2012). The increased AEC relates to a high energetic state of the cell. Additionally, 43 differentially expressed proteins were identified (Beckmann et al. 2012). Some of the proteins, including phosphoglycerate mutase, phosphoglycerate kinase and pyruvate kinase isozymes, were associated with the improved growth rate (Beckmann et al. 2012). Other differentially expressed endoplasmic reticulum stress proteins were related to cell robustness to changing environmental conditions (Beckmann et al. 2012). The

combined approach was useful for determining correlations between the different protein and metabolite levels and the observed cell phenotype.

Conclusions

Current developments in biotechnology rely on the wealth of information provided by the genome, transcriptome, proteome, and metabolome. Protein levels provide clear information about cell physiology; thus, proteomic methods are important for hypothesis-driven research and development. In recent years, sample preparation methods for proteomics have been optimized in order to obtain the maximum number of correct protein identifications. This has enabled more developed databases and identification of proteins localized to cell membranes, the cytoplasm, or other organelles. A variety of different digestion techniques and proteomics techniques have been developed for CHO proteomics. In addition to this, a variety of labeling techniques such as SILAC, iTRAQ and TMT have been used to study differences between cell lines and analyze important pathways such as apoptosis, growth, and protein production. In bioprocess development, applications of proteomics include identification of accumulating or depleting protein levels important in recombinant protein production. From this information, cell line engineering approaches or medium formulation developments help to increase cell growth, prevent apoptosis, improve recombinant protein productivity, or enhance utilization of nutrients. Thus, 'omics approaches enable research that delves deeper into the workings of the cell and improve its performance for biotechnology applications.

In summary, proteomics has followed genomic investigations and surged to the front of analytic methodologies to characterize and drive strategies for improving mammalian, and especially CHO, cell performance. This information will help to improve the manufacture of biotherapeutics at large scale and with high product quality, lowering operating costs and ultimately reducing overall health care costs. In the future, proteomics analysis will become a standard yet essential element in the arsenal that biotechnologists use for understanding and optimizing bioprocesses.

References

Baik JY, Lee MS, An SR, Yoon SK, 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(2):361–371

Baik JY, Joo EJ, Kim YH, Lee GM (2008) Limitations to the comparative proteomic analysis of thrombopoeitin producing Chinese hamster ovary cells treated with sodium butyrate. J Biotechnol 133(4):461–468

Baik JY, Ha TK, Kim YH, Lee GM (2011) Proteomic understanding of intracellular response of recombinant Chinese hamster ovary cells adapted to grow in serum-free suspension culture. Biotechnol Prog 27(6):1680–1688

Baycin-Hizal D, Tabb DL, Chaerkady R, Chen L, Lewis NE, Nagarajan H, Sarkaria V, Kumar A, Wolozny D, Colao J, Jacobson E, Tian Y, O'meally RN, Krag S, Cole RN, Palsson BO, Zhang H, Betenbaugh M (2012) Proteomic analysis of Chinese hamster ovary (CHO) cells. J Proteome Res 11(11):265–5276

Becker J, Hackl M, Rupp O, Jakobi T, Schneider J, Szczepanowski R, Bekel T, Borth N, Goesmann A, Grillari J et al (2011) Unraveling the Chinese hamster ovary cell line transcriptome by next-generation sequencing. J Biotechnol 156(3):227–235

Beckmann TF, Kramer O, Klausing S, Heinrich C, Thute T, Buntemeyer H, Hoffrogge R, Noll T (2012) Effects of high passage cultivation of CHO cells: a global analysis. Appl Microbiol Biotechnol 94(3):659–671

Bonner MK, Poole DS, Xu T, Sarkeshik A, Yates Iii JR, Skop AR (2011) Mitotic spindle proteomics in Chinese hamster ovary cells. PLoS One 6(5):e20489

Brinkrolf K, Rupp O, Laux H, Kollin F, Ernst W, Linke B, Kofler R, Romand S, Hesse F, Budach WE et al (2013) Chinese hamster genome sequenced from sorted chromosomes. Nat Biotechnol 31(8):694–695

Carlage T, Hincapie M, Zang L, Lyubarskaya Y, Madden H, Mhatre R, Hancock WS (2009) Proteomic profiling of a high-producing Chinese hamster ovary cell culture. Anal Chem 81 (17):7357–7362

Carlage T, Kshirsagar R, Zang L, Janakiraman V, Hincapie M, Lyubarskaya Y, Weiskopf A, Hancock WS (2012) Analysis of dynamic changes in the proteome of a Bcl-Xl overexpressing Chinese hamster ovary cell culture during exponential and stationary phases. Biotechnol Prog 28(3):814–823

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

Doolan P, Meleady P, Barron N, Henry M, Gallagher R, Gammell P, Melville M, Sinacore M, Mccarthy K, Leonard M, Charlebois T, Clynes M (2010) Microarray and proteomics expression profiling identifies several candidates, including the valosin-containing protein (VCP), involved in regulating high cellular growth rate in production CHO cell lines. Biotechnol Bioeng 106(1):42–56

Dorai H, Liu S, Yao X, Wang Y, Tekindemir U, Lewis MJ, Wu S, Hancock W (2013) Proteomic analysis of bioreactor cultures of an antibody expressing CHO-GS cell line that promotes high productivity. J Proteomics Bioinformatics 6:99–108

Druz A, Chu C, Majors B, Santuary R, Betenbaugh M, Shiloach J (2011) A novel microRNA mmu-miR-466h affects apoptosis regulation in mammalian cells. Biotechnol Bioeng 108(7): 1651–1661

Druz A, Son YJ, Betenbaugh M, Shiloach J (2013) Stable inhibition of mmu-miR-466h-5p improves apoptosis resistance and protein production in CHO cells. Metab Eng 16:67–94

Fischer S, Wagner A, Kos A, Aschrafi A, Handrick R, Hannemann J, Otte K (2013) Breaking limitations of complex culture media: functional non-viral miRNA delivery into pharmaceutical production cell lines. J Biotechnol 168(4):589–600

Gaj T, Guo J, Kato Y, Sirk SJ, Barbas Iii CF (2012) Targeted gene knockout by direct delivery of zinc-finger nuclease proteins. Nat Methods 9(8):805–807

Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE et al (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428(6982):493–521

Hackl M, Jakobi T, Blom J, Doppmeier D, Brinkrolf K, Szczepanowski R, Bernhart SH, Honer Zu Siderdissen C, Bort JA, Wieser M et al (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(1–2):62–75

Hackl M, Jadhav V, Jakobi T, Rupp O, Brinkrolf K, Goesmann A, Puhler 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(3):151–155

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(6):1353–1356

Harsha HC, Molina H, Pandey A (2008) Quantitative proteomics using stable isotope labeling with amino acids in cell culture. Nat Protoc 3(3):505–516

Hayduk EJ, Choe LH, Lee KH (2004) A two-dimensional electrophoresis map of Chinese hamster ovary cell proteins based on fluorescence staining. Electrophoresis 25(15):2545–2556

Hernandez Bort JA, Hackl M, Hoflmayer H, Jadhav V, Harreither E, Kumar N, Ernst W, Grillari J, Borth N (2012) Dynamic mRNA and miRNA profiling of CHO-K1 suspension cell cultures. Biotechnol J 7(4):500–515

Kang S, Ren D, Xiao G, Daris K, Buck L, Enyenihi AA, Zubarev R, Bondarenko PV, Deshpande R, (2013) Cell line profiling to improve monoclonal antibody production. Biotechnol Bioeng 111(4):748–760

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

Kim JY, Kim Y, Han YK, Choi HS, Kim YH, Lee GM (2011) Proteomic understanding of intracellular responses of recombinant Chinese hamster ovary cells cultivated in serum-free medium supplemented with hydrosylates. Appl Microbiol Biotechnol 89(6):1917–1928

Kim JY, Kim YG, Lee GM (2012) CHO cells in biotechnology for production of recombinant proteins: current state and further potential. Appl Microbiol Biotechnol 93(3):917–930

Kuystermans D, Dunn MJ, Al-Rubeai M (2010) A proteomic study of cMyc improvement of CHO culture. BMC Biotechnol 10:25

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, Fitzhugh W et al (2001) Initial sequencing and analysis of the human genome. Nature 409(6822):860–921

Lawrence S, Lahteenmaki R (2014) Public biotech 2013 the numbers. Nat Biotechnol 32:626–632

Lee JS, Park HJ, Kim YH, Lee GM (2010) Protein reference mapping of dihydrofolate reductase-deficient CHO DG44 cell lines using 2-dimensional electrophoresis. Proteomics 10(12): 2292–2302

Lewis NE, Liu X, Li Y, Nagarajan H, Yerganian G, O'brien E, Bordbar A, Roth AM, Rosenbloom J, Bian C (2013) Genomic landscapes of Chinese hamster ovary cell lines as revealed by the *Cricetulus griseus* draft genome. Nat Biotechnol 31(8):759–765

Lim UM, Yap MG, Lim YP, Goh L, Ng SK (2013) Identification of autocrine growth factors secreted by CHO cells for applications in single-cell cloning media. J Proteome Res 12(7): 3496–3510

Megger DA, Pott LL, Ahrens M, Padden J, Bracht T, Kuhlmann K, Eisenacher M, Meyer HE, Sitek B (2013) Comparison of label-free and label-based strategies for proteome analysis of hepatoma cell lines. Biochim Biophys Acta 1844(5):967–976

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

Meleady P, Gallagher M, Clarke C, Henry M, Sanchez N, Barron N, Clynes M (2012a) Impact of miR-7 over-expression on the proteome of Chinese hamster ovary cells. J Biotechnol 160(3–4): 251–262

Meleady P, Hoffrogge R, Henry M, Rupp O, Bort JH, Clarke C, Brinkrolf K, Kelly S, Muller B, Doolan P, Hackl M, Beckmann TF, Noll T, Grillari J, Barron N, Puhler A, Clynes M, Borth N (2012b) Utilization and evaluation of CHO-specific sequence databases for mass spectrometry based proteomics. Biotechnol Bioeng 109(6):1386–1394

Mertins P, Udeshi ND, Clauser KR, Mani DR, Patel J, Ong SE, Jaffe JD, Carr SA (2012) iTRAQ labeling is superior to mTRAQ for quantitative global proteomics and phosphoproteomics. Mol Cell Proteomics 11(6):M111.014423

Ohsfeldt E, Huang S, Baycin-Hizal D, Kristoffersen L, Le TT, Li E, Hristova K, Betenbaugh MJ (2012) Increased expression of the integral membrane proteins EGFR and FGFR3 in anti-apoptotic Chinese hamster ovary cell lines. Biotechnol Appl Biochem 59(3):155–162

Park SS, Wu WW, Zhou Y, Shen RF, Martin B, Maudsley S (2012) Effective correction of experimental errors in quantitative proteomics using stable isotope labeling by amino acids in cell culture (SILAC). J Proteomics 75(12):3720–3732

Puck TT, Cieciura SJ, Robinson A (1958) Genetics of somatic mammalian cells III Long-term cultivation of euploid cells from human and animal subjects. J Exp Med 108(6):945–956

Slade PG, Hajivandi M, Bartel CM, Gorfien SF (2012) Identifying the CHO secretome using mucin-type-O-linked glycosylation and click-chemistry. J Proteome Res 11(12):6175–6186

Tabuchi H, Sugiyama T, Tanaka S, Tainaka S (2010) Overexpression of taurine transporter in Chinese hamster ovary cells can enhance cell viability and product yield, while promoting glutamine consumption. Biotechnol Bioeng 107(6):998–1003

Valente KN, Choe LH, Lenhoff AM, Lee KH (2012) Optimization of protein sample preparation for two-dimensional electrophoresis. Electrophoresis 33(13):1947–1957

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(1):87–99

Van Dyk DD, Misztal DR, Wilkins MR, Mackintosh JA, Poljak A, Varnai JC, Teber E, Walsh BJ, Gray PP (2003) Identification of cellular changes associated with increased production of human growth hormone in a recombinant Chinese hamster ovary cell line. Proteomics 3(2): 147–156

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P et al (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420(6915):520–562

Wei YC, Naderi S, Meshram M, Budman H, Scharer JM, Ingalls BP, Mcconkey BJ (2011) Proteomics analysis of Chinese hamster ovary cells undergoing apoptosis during prolonged cultivation. Cytotechnology 63(6):663–677

Wiese S, Reidegeld KA, Meyer HE, Warscheid B (2007) Protein labeling by iTRAQ: a new tool for quantitative mass spectrometry in proteome research. Proteomics 7(3):340–350

Wisniewski JR, Zougman A, Nagaraj N, Mann M (2009) Universal sample preparation method for proteome analysis. Nat Methods 6(5):359

Xu X, Nagarajan H, Lewis NE, Pan S, Cai Z, Liu X, Chen W, Xie M, Wang W, Hammond S (2011) The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. Nat Biotechnol 29:735–741

Yee JC, de Leon GM, Philp RJ, Yap M, Hu WS (2008) Genomic and proteomic exploration of CHO and hybridoma cells under sodium butyrate treatment. Biotechnol Bioeng 99(5): 1186–1204

Chapter 6 – System wide analysis of CHO omics data

In this chapter, a review will be presented on the topic of generating a genome-scale metabolic model specific for CHO. Following the rapid increase in genomic, transcriptomics and proteomic data from CHO it is now possible to generate a computational model so far only applied to other organisms with well-annotated genomes such as human, yeast and *E. coli*. The current status and future perspectives for this field is summarized and discussed.

Pharmaceutical BIOPROCESSING

Toward genome-scale models of the Chinese hamster ovary cells: incentives, status and perspectives

Bioprocessing of the important Chinese hamster ovary (CHO) cell lines used for the production of biopharmaceuticals stands at the brink of several redefining events. In 2011, the field entered the genomics era, which has accelerated omics-based phenotyping of the cell lines. In this review we describe one possible application of this data: the generation of computational models for predictive and descriptive analysis of CHO cellular metabolism. We describe relevant advances in other organisms and how they can be applied to CHO cells. The immediate implications of the implementation of these methods will be accelerated development of the next generation of CHO cell lines and derived biopharmaceuticals.

Christian S Kaas[1,2], Yuzhou Fan[1,3], Dietmar Weilguny[3], Claus Kristensen[2], Helene F Kildegaard[4] & Mikael R Andersen*[,1]
[1]Department of Systems Biology, Technical University of Denmark, Denmark
[2]Mammalian Cell Technology, Biopharmaceutical Research Unit, Novo Nordisk A/S, Maaloev, Denmark
[3]Symphogen A/S, Ballerup, Denmark
[4]The Novo Nordisk Foundation Center for Biosustainability, HØrsholm, Denmark
*Author for correspondence: mr@bio.dtu.dk

It is generally appreciated that cell culture based on Chinese hamster ovary (CHO) cells holds substantial economical and medical importance. The global market for biologics was US$99 billion in 2009, where 60–70% of the products were produced in CHO cells [1]. Over 40 biopharmaceuticals have been produced in CHO cells so far, including monoclonal antibodies, hormones, cytokines and blood-coagulation factors. It is furthermore evident that the impact of CHO cell culture will only increase in the immediate future: the US market for biologics alone has been climbing from US$51.3 billion in 2010 to US$63.6 billion in 2012, and expected to increase at higher rates with the US Affordable Care Act [2]. The global market for biologics is expected to rise to US$190 billion in 2015 [3], and the percentage of CHO-derived products in approved new biologics are climbing. In 2010 and 2011 combined, 14 out of 19 approved biopharmaceuticals were derived from cell culture, the majority of these using CHO cells as hosts [4].

Despite this impact, the development of CHO cell processes – although highly successful – has been mainly driven by medium development and process engineering and to a lesser extent genomic technologies such as enhanced expression technologies for heterologous proteins [5]. Metabolic engineering, such as seen in microbial cell factories [6,7], has been very limited, although with some notable exceptions (for example, see [8,9]). We will argue that this has been due to the relatively late arrival of genome sequences for CHO cell lines; even though the first CHO expressed sequence tags (EST) sequences were published in 2005 [10], the first CHO genome sequences were published in 2011 [11,12], an entire decade after the first draft publication of the human genome [13], and two decades after the genome of the first eukaryote, *Saccharomyces cerevisiae* [14]. As a result, most early genome-based studies of CHO cells were performed by using the genome sequences from other mammals, for example, human, mouse or rat [15,16], which generally limits the possible experiments and interpretation of the results.

However, there is now ample genomic information available for the CHO cell lines. The CHO-K1 genome sequence [11] has been supplemented by the 2013 release of two draft genomes for the Chinese hamster (*Cricetulus griseus*) [17,18] from which the CHO cell line was originally isolated in 1957 [19]. Additionally, draft sequences for a number of CHO cell lines including the industrially relevant CHO-S and CHO DG44 have

FUTURE SCIENCE part of fsg

Key terms

Computational framework: Modeling metabolism is typically done using linear programming, which allows optimization of fluxes to a single criterion; for example, maximum possible growth rate, which is a typical approach for microbial cultures. Alternative methods include quadratic programming, which allows optimization for two criteria. Quadratic programming methods are often used for modeling effects of gene deletions.

Cellular compartments: In a genome-scale metabolic model, cellular compartments are modeled by assigning reactions to a given compartment, and adding known and required transport reactions in and out of the compartment to the model. Predicting in which compartment a specific reaction takes place is challenging for enzymes with a low degree of characterization.

been published [17]. While this still leaves a few widely used cell lines, for example, the CHO DXB11 cell line [20,21], unsequenced, and a general need for improved genome quality, it is clear that CHO cell research has reached the genomic era.

One highly promising application of genomics-based research is the generation of genome-scale models of CHO cells (Figure 1).

Potential applications of metabolic models to CHO cell cultures

A genome-scale metabolic model (GSM) is a systematic correlation of the genomic information of an organism to a metabolic network, effectively reconstructing the metabolic network of the cell type in question. Such a network is most often built from available generic pathway databases (e.g., Kyoto Encyclopedia of Genes and Genomes [KEGG] [22]) and specific literature for the organism being modeled, combined with an annotated genome [23]. This underlying network is often called a genome-scale reconstruction or genome-scale metabolic network reconstruction (GENRE). Integration of the GENRE with a linear programming-based mathematical framework allows modeling of the metabolic fluxes of the cell, which is often predictive and nearly always helpful in data interpretation. The actual model and **computational framework** apply the laws of mass conservation and balances of metabolic fluxes around single metabolites to compute enzymatic rates for every single enzyme present in the model. These rates are seen as averages for the culture and are most often given as specific rates relative to a certain number of cells. Additional algorithms may be applied to predict the effect of, for example, gene deletions/insertions, perturbations of feeding rates/nutrient uptake or increased production rates [24]. Pioneering work and additional application such as integration of the protein secretion network and regulatory information has been driven forward in *Escherichia coli* [25–27]. As CHO cells are arguably more complex in terms of gene numbers and **cellular compartments** than *E. coli*, the work associated with building a CHO GSM is more laborious and complicated, in particular in terms of assigning correct genes to enzymatic functions, and assigning enzymatic reactions to the correct compartments. However, the algorithms and uses of these models are general, and examples of potential applications from *E. coli* are equally relevant for CHO cells. In addition to this, implementation has been performed in a wide span of eukaryotic organisms as well, several of which with a complexity and quality of annotation resembling CHO cells. Examples of eukaryotic models include eukaryotic microbes, for example, industrially relevant yeasts such as *S. cerevisiae, Kluyveromyces lactis* and *Pichia pastoris* [28–31], filamentous fungi applied for enzyme production, for example, *Aspergillus niger* [32–34], and also higher eukaryotes such as *Arabidopsis* [35], mouse hybridoma cells [36], and human cells [37,38]. In these examples, cells, arguably as complex as CHO cells, have had their metabolism reconstructed. Cells from mouse, *Arabidopsis* and human are evidently of similar or higher complexity than the CHO cell. Even eukaryotic microbes such as filamentous fungi have a more complex growth physiology, with multicellular growth compared with the relatively homogeneous CHO cells with a more uniform growth. While the current annotation of the CHO genome is far from the quality of annotation and gene characterization found for human cells or even mouse cells [39], models can to a large part be generated by inferring function by homology to organisms with better annotation, for example, mouse or human in the case of CHO.

The primary applications of these models can be divided into at least five major categories: metabolic engineering, model-directed discovery, interpretations of phenotypes, analysis of network properties and studies of evolutionary processes [40,41]. All of these applications are highly relevant and interesting for CHO cell culture in their omics-driven approach to cellular physiology (Figure 1E).

Metabolic engineering holds considerable promise for CHO cell culture, as GSMs have the possibility of predicting the effect of gene deletions, additions and over-/under-expression. Several phenotypic traits of the CHO cells are sub-optimal for prolonged culture and protein productivity. Some examples of this are the conversion of high **glycolytic flux** to lactate, or the formation of ammonium by conversion of amino acids in the medium. Both are detrimental to cell growth and product quality [42,43]. Accordingly, these processes have been subjected to metabolic

engineering with varying success [8,9], but definitive solutions have not been found at the cell engineering level, and have to a large extent been addressed and alleviated with process and medium design (see, [44] for an example of lactate production). Model-driven engineering presents an interesting angle on these problems, by generating platform cell lines incapable of producing such by-products or producing them in highly reduced amounts. Other possibilities are found in areas addressed in microbial cell factories, for example, increasing the number of sugars available for carbon catabolism to decrease the problems with high glycolytic flux [45]. GSMs also provide attractive possibilities to model decreased by-product formation [46], generally applicable to any biotechnological production process. Another tantalizing possibility is the extension of the significant advances in CHO culture medium development [5]. One could imagine that it would be interesting to perform model-guided systemic cell line engineering to tailor the cell lines to a specific medium, or a certain feeding profile. The capabilities of GSMs to model cellular metabolism on a systemic scale allow the combination of, for example, consumption rates of medium components with model predictions. The result of this would be further improvement of platform media and processes with a tailored cell line.

Model-directed discovery has been useful in many microbial systems, in particular for improving gene annotation and functional assignment [32,33,41]. With the current state of **CHO genomics** being in its infancy, the annotation of the identified genes is very preliminary. The state of the CHO genome annotation on one hand complicates accurate model reconstructions, due to the challenges of linking genes with function, but also presents opportunities. In particular, the reconstruction of CHO metabolism (GENRE) will suggest tentative assignment for a high percentage of the metabolic genes, as is seen in other eukaryotic microbes organisms [28,32,34], where the genes involved in metabolism have not been studied as extensively as in leading model organisms, for example, yeast or *E. coli* [25,29]. However, even in *E. coli* the metabolic network has been applied to find and characterize candidate genes for a specific function [47]. This type of application of the metabolic network is particularly powerful when combined with other omics-data types such as metabolomics and proteomics (see below). Here one can, for example, integrate orphan metabolites into the metabolic network and thus improve our understanding of CHO metabolism, or identify active isoenzymes for specific pathways from protein expression data coupled with tentative metabolic enzyme networks. As this application of the model is fairly independent of the computational predictive power, it is relevant to CHO cells no matter how accurate the models might become. The same holds true for the next application.

Key terms

Glycolytic flux: In this article, glycolytic flux is defined as being the conversion rate of glucose through glycolysis.

CHO genomics: Sequencing of multiple CHO cell lines and the progenitor hamster has revealed that while individual cell line genomes have a similar number of genes as the hamster, there is a large number of structural variations, in particular insertions, deletions and single-nucleotide polymorphisms. Such variation suggests that models should be tailorized to individual cell lines.

Interpretation of phenotypes is perhaps the most universally applicable use of the GSMs and GENREs. The network – irrespective of predictive power – provides a framework for interpretation of experimental data. For CHO cells, the calculation of metabolite consumption/production rates, growth rates and specific product formation rates has long been standard for cell culture medium and process design. However, a theoretical and computational framework for holistic interpretation of the data has not previously been available. These reconstructed networks can aid the interpretation of experimental data related to growth, as well as multiple types of omics data [41]. Models have proven important in interpretation of metabolic/flux data [48], transcriptomics [49] and proteomics [50]. The approaches are listed in an excellent recent review [51], which also covers mammalian cell types. A related example from medical research is the interpretation of proteomic and DNA microarray data from human macrophages through a reconstructed metabolic network of the cell type [38]. In this study, a holistic view of the process of activation of macrophages was achieved, and systemic activation and inactivation of parts of metabolism was identified.

In many cases, models are also capable of quite persuasive prediction of phenotypes. Especially interesting, considering the costs and time required to produce stable genetic changes in CHO cells, is prediction of phenotypes of genetic mutants [52,53].

Analysis of network properties is generally speaking a computational exercise, in which one analyzes the network structure of the GENRE to discover inherent features or emergent properties of the metabolic network. In some cases, this analysis has become mainly an arithmetic exercise, applying standard network topology algorithms, and has generated limited biological insight. However, in some cases, deep insights are found. The most easily applicable example in the context of CHO cell and process engineering is the application of elementary flux modes [54] to identify

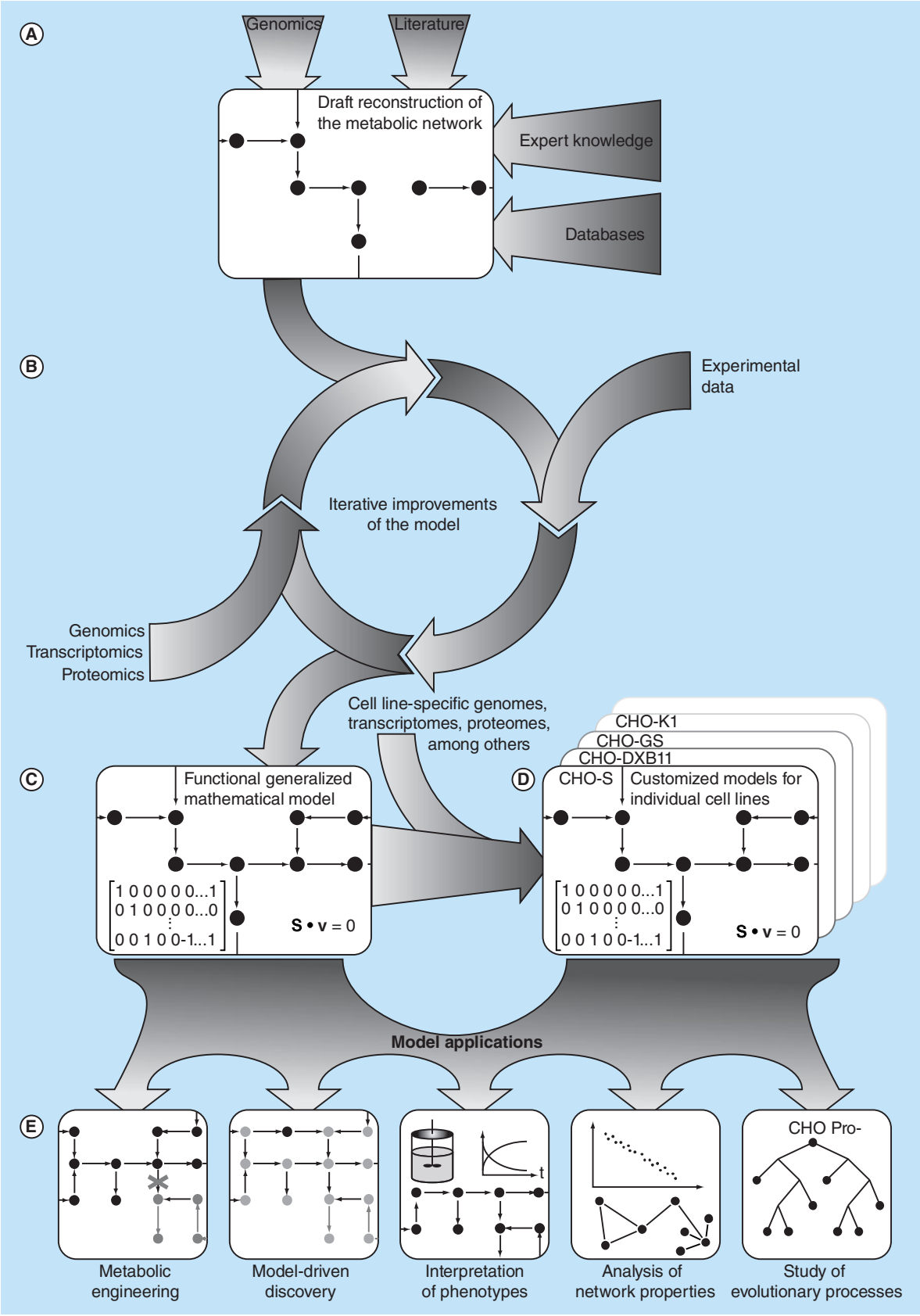

A
Genomics
Literature
Draft reconstruction of the metabolic network
Expert knowledge
Databases
B
Experimental data
Iterative improvements of the model
Genomics
Transcriptomics
Proteomics
Cell line-specific genomes, transcriptomes, proteomes, among others
C
Functional generalized mathematical model
S • v = 0
D
CHO-K1
CHO-GS
CHO-DXB11
CHO-S
Customized models for individual cell lines
S • v = 0
Model applications
E
CHO Pro-
Metabolic engineering
Model-driven discovery
Interpretation of phenotypes
Analysis of network properties
Study of evolutionary processes

Figure 1. Model building and application in Chinese hamster ovary cells (see facing page). **(A)** Initial drafting of a GSM requires genomic annotation supplemented with available literature and knowledge on the metabolism of choice. This generates a draft genome-scale metabolic network reconstruction. **(B)** Multiple iterations of model improvement and testing, supplemented with available omics and phenotypic data, generate a GSM capable of computation. **(C)** A generalized GSM for CHO cells can be tailored to specific cell lines **(D)** by the inclusion of additional omics data specific for the individual cell lines. **(E)** Both generalized and specialized CHO GSMs may be applied to engineer cell lines, interpret data and increase functional understanding of these important cell lines. CHO: Chinese hamster ovary; GSM: Genome-scale metabolic model; t: Time.

the smallest possible set of essential metabolic genes in CHO cells, such as that achieved for individual pathways in *E. coli* [55] or *S. cerevisiae* [56]. Such identification could be applied for generating cell lines with a trimmed metabolism, thus decreasing the variability of the system. This would be feasible for CHO, despite the presence of extra isoenzymes or alternate pathways in many enzymatic steps. For several enzymatic functions, only one of the isoenzymes is detected at the protein level [57]. Alternatively, one can delete steps in the enzymatic pathways/elementary flux modes, where only one enzyme exists.

Studies of evolutionary processes have been a reoccurring theme in several applications of bacterial network, in particular the *E. coli* GENRE [40,41]. In such studies, specialized models have, for example, been developed to describe specific strains of *E. coli*, and compared these to identify the genetic origin of specific phenotypes [58]. Given the availability of the genomic sequence for multiple CHO cell lines with varying properties [17] and surely more to come, such an exercise would hold exciting perspectives for interpreting these genomes. One possible application would be the genetic basis for certain metabolic features in cell lines generated by mutagenesis, and the possibility of *de novo* engineering the features into a 'clean' background. This would reduce possible complications from decreased genetic stability in mutants subjected to mutagenesis [59].

With the above-mentioned being only a small percentage of the possible applications of such models for CHO cell lines, the potential is clearly large for the generation and application of GENREs and GSMs for CHO cells.

Additional available data sources for increased applicability of CHO models

The availability of an annotated genome for CHO cells is the bare minimum of information required to generate a draft model for CHO cells. Models of microbial systems have been published based mainly on literature on characterized genes (e.g., for *Corynebacterium glutamicum* [60] or *Aspergillus niger* [32,61]), but this requires detailed legacy data for a wide selection of metabolic pathways. In general, most recent generation of GSMs is based on variations of a standardized protocol for metabolic network reconstruction and model validation published in 2010 [23], using basic genome annotation as a starting point for the organism of choice.

However, other types of omics data have proven highly valuable for model generation, validation and application. Here, the CHO field is maturing at an impressive pace, considering the quite recent publication of the first public CHO genome [11], followed by a wealth of other omics types being published in these years [62]. Here we will briefly emphasize selected studies which provide data highly applicable to CHO modeling, either due to the experimental setup of the study, the type of the data or the perspectives these offer for CHO modeling.

Genomics

A well-annotated genome with a high coverage is a crucial component in building a GSM with predictive power or a GENRE with a potential for informative data integration and interpretation. It is essential to be able to identify the genes of all major metabolic pathways in order to generate an accurate GENRE and following that a GSM. This requires a genome coverage of ideally >99% of the genes. Early EST sequencing efforts [10] identified only less than 20% of the genes, shown to be present in the CHO-K1 cell line draft genome [11]. Convenient for model construction, the genome sequence has been made accessible at the online database CHOgenome.org [63,64] as well as at the NCBI genbank. The coverage appears to be at least 99%, at least it was demonstrated that homologs existed for 99% of the genes in the human genome associated with glycosylation.

Due to the variability of the cell lines, it can be argued that it would be most appropriate to use the progenitor Chinese hamster (*C. griseus*) both as the source of a reference genome and as a scaffold for a master CHO GSM, from which specialized models can be generated for individual cell lines. This is now possible due to

Key term

Reference genome: Typically a genome sequence with a high (e.g., >99.5%) percentage of coverage, and a gene annotation of high quality. It is typically used to provide context and mapping of sequence from closely related genomes with a low sequence depth. This can be highly advantageous for especially large genomes, where sequence depth is expensive, or to be able to use the same set of reference genes in comparisons and data analysis.

the publication of draft genomes for *C. griseus* [17,18]. This publication showed that there are more than 3.7 million point mutations between the progenitor hamster and the CHO cell lines [17] and extensive chromosomal rearrangements that have occurred between CHO-K1 and CHO-DG44 [65] emphasizing the effects of the mutagenesis that occurred in the process of creating the various cell lines. Currently, in order to fully exploit these sequences, these genomes must be mapped against reference genomes with gene annotation. Improving the gene annotation for the *C. griseus* genome and CHO-K1, which has become the *de facto* reference genome for cell lines, would be beneficial to model building. Current and new genomes could thus be aligned to these references for detection of mutations.

Furthermore, the state of genome assembly should be improved. Currently the genomes for both hamster and cell lines are divided into at least 4000 contigs per genome, which means that genes for important metabolic functions may be lost in the sequencing gaps. Such gaps can to some extent be detected and fixed in the network reconstruction process [23,32]. Even so, an appropriate solution would be to apply third generation sequencing to yield longer sequenced reads that can assemble the contigs to improve the coverage of the genomes. Such efforts are in progress in the community [Borth N, Pers. Comm.], and should have a substantial positive impact on the models, which can be constructed for CHO cells.

Transcriptomics

In general, it is only a low percentage of the CHO genes which are actually expressed under normal condition, for example, only approximately 50% of the genes involved in protein glycosylation are transcriptionally active [11]. Consequently, integration of dynamic omics data such as transcriptomics and proteomics are important for accurate prediction of gene deletion/silencing effects. Several studies and tools are now available for this, including both sequencing and DNA microarray based methods. Naturally, some of the first transcriptome data were generated prior to the genome sequence based on EST sequences from CHO and mouse used for design of microarrays [66,67].

Worth particular mention is a large-scale comparison of microarray data from more than 120 individual CHO cultures [68]. The data can be accessed through the web-based CHO gene coexpression database allowing easy access to the list of genes found to coexpress with, for example, cell specific productivity and growth rate. Such data could be used for model improvement and validation. For easy and relatively inexpensive assessment of the CHO transcriptome in future experiments, a new generation of the Affymetrix® CHO DNA microarray (Affymetrix, CA, USA) has been launched with up to 26 unique sequences of each transcript with a total of more than 644,000 probes [69].

RNA-sequencing is expanding for CHO culture as in many other fields [70]. Recently, a transcriptome database for CHO RNA sequencing data has been developed and is available at GenDBE [71,72].

In summary, transcriptomics data are abundantly available, and will only increase in the coming years.

Proteomics

The CHO proteome is interesting in the context of CHO metabolic modeling as it can provide additional functional information, in some cases expanding on transcriptomic evidence. The proteome of CHO-K1 was thoroughly characterized by Baycin-Hizal in 2012 [57]. Here, 6164 proteins were detected. Of these, only 60% were also detected at the mRNA level by Xu in 2011 [11]. The functional application of the data and the need for having models specialized to individual cell lines become apparent from this study. Statistical analysis indicated that some pathways such as fatty acid metabolism, amino sugar and nucleotide sugar metabolism, which provide important precursors for recombinant protein synthesis, as well as protein processing and apoptosis, were enriched in CHO-K1 [57].

Given the principal application of CHO cells for production of secreted proteins, in this content, secretome data, such as characterized from the CHO DG44 and CHO-S cell lines by Slade [73], are interesting to incorporate. Such data can help identify secretory bottlenecks or extracellular proteases as seen for microbial cell factories [74,75].

CHO-specific protein databases have been constructed based on data from the CHO-K1 genome [11] and the CHO transcriptome [76], and have been shown to increase the number of identified proteins by 40–50% from proteomics studies compared with only using protein databases based on, for example, the murine proteome [77].

It is generally accepted that the generation of proteomics data is more technically challenging than transcriptome data, but the pilot studies within CHO cells, such as those mentioned above, show that there is clearly additional value to be gained from interrogating this data set.

Metabolomics

As mentioned in the text above, metabolomics have considerable value to add in the model building pro-

cess, as this type of data can help identify metabolic pathways, which are experimentally shown to be occurring in the cells due to the presence of metabolic intermediates or products, but the genetic basis is not necessarily clear. Being able to identify and include these pathways can increase the predictive power of the model.

For such inclusion, several studies of high-quality [78,79] and standardized protocols [79–82] have been available for several years, as metabolomics are not as such dependent on the availability of a genome.

Two studies are of particular interest in terms of getting data of a sufficient quality for model integration. The first was published by Dietmair *et al.* [83] correlating intracellular and extracellular metabolite concentrations with growth. The second is the work of Chong *et al.* [84] where intracellular metabolite profiles were obtained for eight single-cell clones with high and low production rates of monoclonal antibodies at the mid-exponential phase during shake flask batch cultures. Such studies can give insight in metabolic responses, and help validate CHO metabolic models, in that one can examine and adapt the ability of the model to predict these responses.

The application of the metabolic networks to interpret metabolomics data can also be exemplified in a 2012 study by Selvarasu *et al.* [85], where a generalized metabolic network of mammalian cells was adapted to CHO cells to aid in metabolomics data interpretation (see further details below). The coupling of the network, genome-scale-modeling and metabolomics data allowed the identification of growth-limiting factors.

Overall, the CHO field is at this point uniquely poised to utilize the substantial amounts of available omics data in building high-quality models for CHO cells. An overview is presented in Figure 1. During any future model-building efforts, one should draw upon the current availability of computational models for CHO and similar systems, and incorporate this where appropriate.

Overview of cellular modeling efforts in CHO cells & beyond

So far, no dedicated effort to building a CHO GSM *de novo* has been published. The closest example is the adaptation of a model of mouse hybridoma cells [36] to CHO cells by the addition of 35 CHO-specific metabolic reactions and subsequent model curation resulting in a model comprising 1540 reactions and 1302 metabolites [85]. This model has been further developed by other groups, although not published through a journal at this time, but is available for download from CHO.sf.net [86]. A similar approach of adapting a mouse GSM was employed by Martínez *et al.* [87] for examining the energy consumption and metabolism surrounding lactate formation and consumption in CHO cells.

Dedicated models have been developed for related cell lines in other systems, as mentioned a generic model for mouse cells, applied to mouse hybridoma cell lines [36], and a model for the HEK-293 cell line has been developed as well [88]. This study is particularly promising for CHO cell modeling, as the HEK-293 model was developed by reducing the generic model for human cellular metabolism [37] to a model specific for HEK-293 metabolism. Furthermore, this model was employed to interpret both transcriptomic, metabolomic and flux data to gain functional understanding of glucose and glutamine metabolism; both key features for CHO metabolism [88]. A similar study has been seen for baby hamster kidney (BHK) cells for interpretation of metabolomics data [89].

These models listed above represent the full list of available metabolic genome-scale models with relevance to CHO cells. However, to the best of our knowledge, a model specific for CHO cells or any specific cell line has still not been generated.

One area, where modeling in CHO cells is more developed, is the kinetic modeling of protein N-glycosylation, in particular integrated with mass spectrometry on glycans. Here, very accurate predictions and substantial networks have been generated and improved over the last two decades. The first mathematical model for protein N-glycosylation process was built in 1997 by the complementary studies of a single-compartmental model [90] and a multi-compartmental model [91]. Later work expanded upon the previous work to involve glycosylation processes as galactosylation, fucoslation, sialylation and addition of N-acetyllactosamine residues [92]. This model had up to 7565 N-glycans and 22,871 reactions included. Furthermore, two glycosylation models based on different views of protein transport across the Golgi, namely Golgi maturation mechanism and vesicular transport mechanism, were studied and compared. This model was highly expanded and sophisticated by the same group to include interpretative power of N-glycan mass spectrometry data [93]. More recently, an optimized model considering 77 N-glycans, 8 enzymes, 4 nucleotide transporters and 95 reactions with individual rate expressions were built on the basis of Golgi maturation mechanism with an improvement of taking Golgi protein recycling into account [94]. On top of that, a more comprehensive glycosylation model that links a model that described the metabolism of nucleotides and nucleotide sugars

to the previous N-glycosylation model was developed by the same group [95]. These networks have been shown to have both high predictive and interpretative power, and would be unique key features to have integrated in CHO GSMs, to the extent that it is possible. Such additions could predict effects of glycosylation engineering and/or the effect of different substrate uptake rates.

Conclusion & future perspective

With the potential of GSMs tailored to CHO cells as demonstrated above, it is not surprising that several groups in the CHO community are working on building whole and partial reconstructions of CHO metabolism, including some of the authors of this review. These groups have this year formed a consortium compiling their work, and are working toward generating a community consensus model for CHO cells [Lewis NE, Pers. Comm.]. Such models and network reconstructions are known from several other research communities, including Salmonella Typhimurium [96], yeast [29,97] and human [37] metabolism.

The future arrival of the CHO GSM will probably raise the same discussion that followed the release of the first CHO genome: How well does this model describe each of the different CHO cell lines? Each of the cell lines has undergone rearrangements and has diverse transcriptomes and for this reason several parameters will need to be investigated. Future and current sequencing projects for individual cell lines should be combined with bioreactor characterization of the cell lines and their corresponding models to gain a functional understanding of the differences (Figure 1C–D). The next years will tell whether these models will be able to model the complex behavior of the CHO cell and open up new design targets such as it has been the case in microbes. Making such specialized models will be a substantial amount of work, but this task will be made easier, if a generic CHO model of high quality based on an assembled and annotated reference genome is generated first. From that, specialized models can be made in semi-automated fashion through comparative genomics. This would have the additional advantage that annotation of the genomes of the individual cell lines would not be required (as this is currently not available [17]), but could be achieved by alignment to the reference genome.

Should the models be able to deliver on the promise and potential seen in other cells, it is bound to trigger a second wave of CHO cell line engineering. Notably CRISPR-Cas9-based genome-editing systems being made available at non-cost prohibitive prices [98] and efficient high-throughput mammalian vector design systems [99] support the development of faster and cheaper genome engineering tools to accelerate future cell line engineering efforts in CHO.

In summary, the potential of genome-scale models stands to be unleashed in CHO cells within a very short time span. Combined with the genomes ushering in the genomics era for CHO, substantial amounts of omics data are being generated, and the development of efficient genetic engineering tools of CHO cell culture will soon move into the next generation of cell line development. Such advances promise better and cheaper development of biopharmaceuticals for this important group of cell factories.

Financial & competing interests disclosure

Y Fann and D Weilguny are employees of Symphogen A/S, CS Kaas and C Kristensen are employees of Novo Nordisk A/S. While the companies have funded the salaries of the employees, the company has had no role in defining the content of the manuscript. MR Andersen and HF Kildegaard declare no competing financial interests. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Executive summary

- Genome-scale metabolic models have been applied with success in many other prokaryotic and eukaryotic cell factories.
- The Chinese hamster ovary (CHO) field now has all of the relevant information and methods needed to construct and apply such models.
- A CHO metabolic model will have applications both in design and engineering of cells, but equally important also in interpretation of omics data. The potential is large.
- Initial CHO models have adapted from models of mouse metabolism, but no *de novo* CHO models have been published at this time.
- The community is currently constructing a consensus model for CHO metabolism.
- Added value will come from generating specialized models for individual cell lines.

References

Papers of special note have been highlighted as:
• of interest; •• of considerable interest

1 Walsh G. Biopharmaceutical benchmarks 2010. *Nat. Biotechnol.* 28(9), 917–924 (2010).

2 Aggarwal RS. What's fueling the biotech engine – 2012 to 2013. *Nat. Biotechnol.* 32(1), 32–39 (2014).

3 IMS Institute for Healthcare Informatics. The global use of medicines: outlook through 2015. www.imshealth.com/ims/Global/Content/Insights

4 Walsh G. New biopharmaceuticals. *Biopharm. Int.* 25(6), 34–36 (2012).

5 Wurm FM. Production of recombinant protein therapeutics in cultivated mammalian cells. *Nat. Biotechnol.* 22(11), 1393–1398 (2004).

6 Bailey JE. Toward a science of metabolic engineering. *Science* 252(5013), 1668–1675 (1991).

7 Nielsen J. Metabolic engineering: techniques for analysis of targets for genetic manipulations. *Biotechnol. Bioeng.* 58(2–3), 125–132 (1998).

8 Kim SH, Lee GM. Down-regulation of lactate dehydrogenase-A by siRNAs for reduced lactic acid formation of Chinese hamster ovary cells producing thrombopoietin. *Appl. Microbiol. Biotechnol.* 74(1), 152–159 (2007).

• **Article is a solid example of metabolic engineering of Chinese hamster ovary (CHO) cells.**

9 Zhang F, Sun X, Yi X, Zhang Y. Metabolic characteristics of recombinant Chinese hamster ovary cells expressing glutamine synthetase in presence and absence of glutamine. *Cytotechnology* 51(1), 21–28 (2006).

•• **Expression of glutamine synthetase in CHO cells is an important technology as well as an example of a type of engineering which can be predicted using genome-scale metabolic models.**

10 Wlaschin KF, Nissom PM, de Leon Gatti M *et al.* EST sequencing for gene discovery in Chinese hamster ovary cells. *Biotechnol. Bioeng.* 91(5), 592–606 (2005).

11 Xu X, Nagarajan H, Lewis NE *et al.* The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. *Nat. Biotechnol.* 29(8), 735–741 (2011).

12 Hammond S, Swanberg JC, Kaplarevic M, Lee KH. Genomic sequencing and analysis of a Chinese hamster ovary cell line using Illumina sequencing technology. *BMC Genomics* 12(1), 67 (2011).

13 Lander ES, Linton LM, Birren B *et al.* Initial sequencing and analysis of the human genome. *Nature* 409(6822), 860–921 (2001).

14 Goffeau A, Barrell BG, Bussey H *et al.* Life with 6000 genes. *Science* 274(5287), 546–567 (1996).

15 Gibbs RA, Weinstock GM, Metzker ML *et al.* Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 428(6982), 493–521 (2004).

16 Waterston RH, Lindblad-Toh K, Birney E *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* 420(6915), 520–562 (2002).

17 Lewis NE, Liu X, Li Y *et al.* Genomic landscapes of Chinese hamster ovary cell lines as revealed by the *Cricetulus griseus* draft genome. *Nat. Biotechnol.* 31(8), 759–765 (2013).

18 Brinkrolf K, Rupp O, Laux H *et al.* Chinese hamster genome sequenced from sorted chromosomes. *Nat. Biotechnol.* 31(8), 694–695 (2013).

19 Puck TT, Cieciura SJ, Robinson A. Genetics of somatic mammalian cells. III. Long-term cultivation of euploid cells from human and animal subjects. *J. Exp. Med.* 108(6), 945–956 (1958).

20 Graf LH, Chasin LA. Direct demonstration of genetic alterations at the dihydrofolate reductase locus after gamma irradiation. *Mol. Cell. Biol.* 2(1), 93–96 (1982).

21 Urlaub G, Chasin LA. Isolation of Chinese hamster cell mutants deficient in dihydrofolate reductase activity. *Proc. Natl Acad. Sci. USA* 77(7), 4216–4220 (1980).

22 Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* 42, D199–D205 (2014).

23 Thiele I, Palsson BØ. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* 5(1), 93–121 (2010).

24 Burgard AP, Pharkya P, Maranas CD. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.* 84(6), 647–657 (2003).

25 Orth JD, Conrad TM, Na J *et al.* A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism. *Mol. Syst. Biol.* 7(535), 535 (2011).

26 Monk JM, Charusanti P, Aziz RK *et al.* Genome-scale metabolic reconstructions of multiple *Escherichia coli* strains highlight strain-specific adaptations to nutritional environments. *Proc. Natl Acad. Sci. USA* 110(50), 20338–20343 (2013).

27 Edwards JS, Palsson BO. The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl Acad. Sci. USA* 97(10), 5528–5533 (2000).

28 Chung BKS. Genome-scale metabolic reconstruction and in silico analysis of methylotrophic yeast *Pichia pastoris* for strain improvement. *Microb. Cell Fact.* 9, 50 (2010).

29 Herrgård MJ, Swainston N, Dobson P *et al.* A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat. Biotechnol.* 26(10), 1155–1160 (2008).

30 Zomorrodi AR, Maranas CD. Improving the iMM904 *S. cerevisiae* metabolic model using essentiality and synthetic lethality data. *BMC Syst. Biol.* 4, 178 (2010).

31 Dias O, Pereira R, Gombert AK, Ferreira EC, Rocha I. iOD907, the first genome-scale metabolic model for the milk yeast *Kluyveromyces lactis*. *Biotechnol. J.* 9(6), 776–790 (2014).

32 Andersen MR, Nielsen ML, Nielsen J. Metabolic model integration of the bibliome, genome, metabolome and reactome of *Aspergillus niger*. *Mol. Syst. Biol.* 4, 178 (2008).

33 Vongsangnak W, Olsen P, Hansen K, Krogsgaard S, Nielsen J. Improved annotation through genome-scale

metabolic modeling of *Aspergillus oryzae*. *BMC Genomics* 9, 245 (2008).

34 Dreyfuss JM, Zucker JD, Hood HM, Ocasio LR, Sachs MS, Galagan JE. Reconstruction and validation of a genome-scale metabolic model for the filamentous fungus *Neurospora crassa* using FARM. *PLoS Comput. Biol.* 9(7), e1003126 (2013).

35 De Oliveira Dal'Molin CG, Quek L-E, Palfreyman RW, Brumbley SM, Nielsen LK. AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis. *Plant Physiol.* 152(2), 579–589 (2010).

36 Selvarasu S, Karimi IA, Ghim G-H, Lee D-Y. Genome-scale modeling and *in silico* analysis of mouse cell metabolic network. *Mol. Biosyst.* 6(1), 152–161 (2010).

37 Thiele I, Swainston N, Fleming RMT *et al.* A community-driven global reconstruction of human metabolism. *Nat. Biotechnol.* 31(5), 419–425 (2013).

38 Bordbar A, Mo ML, Nakayasu ES *et al.* Model-driven multi-omic data analysis elucidates metabolic immunomodulators of macrophage activation. *Mol. Syst. Biol.* 8, 558 (2012).

39 Karp PD, Keseler IM, Shearer A *et al.* Multidimensional annotation of the *Escherichia coli* K-12 genome. *Nucleic Acids Res.* 35(22), 7577–7590 (2007).

40 Feist AM, Palsson BØ. The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nat. Biotechnol.* 26(6), 659–667 (2008).

41 McCloskey D, Palsson BØ, Feist AM. Basic and applied uses of genome-scale metabolic network reconstructions of *Escherichia coli*. *Mol. Syst. Biol.* 9, 661 (2013).

42 Yang M, Butler M. Effects of ammonia and glucosamine on the heterogeneity of erythropoietin glycoforms. *Biotechnol. Prog.* 18(1), 129–138 (2002).

43 Lao MS, Toth D. Effects of ammonium and lactate on growth and metabolism of a recombinant Chinese hamster ovary cell culture. *Biotechnol. Prog.* 13(5), 688–691 (1997).

44 Gagnon M, Hiller G, Luan Y-T, Kittredge A, DeFelice J, Drapeau D. High-end pH-controlled delivery of glucose effectively suppresses lactate accumulation in CHO fed-batch cultures. *Biotechnol. Bioeng.* 108(6), 1328–1337 (2011).

• **Gagnon *et al.* demonstrates how process design in many cases can improve fed-batch performance.**

45 Richard P, Verho R, Putkonen M, Londesborough J, Penttila M. Production of ethanol from L-arabinose by containing a fungal L-arabinose pathway. *FEMS Yeast Res.* 3(2), 185–189 (2003).

46 Vemuri GN, Eiteman MA, McEwen JE, Olsson L, Nielsen J. Increasing NADH oxidation reduces overflow metabolism in *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* 104(7), 2402–2407 (2007).

47 Fuhrer T, Chen L, Sauer U, Vitkup D. Computational prediction and experimental verification of the gene encoding the NAD^+/$NADP^+$-dependent succinate semialdehyde dehydrogenase in *Escherichia coli*. *J. Bacteriol.* 189(22), 8073–8078 (2007).

48 Choi HS, Kim TY, Lee D-Y, Lee SY. Incorporating metabolic flux ratios into constraint-based flux analysis by using artificial metabolites and converging ratio determinants. *J. Biotechnol.* 129(4), 696–705 (2007).

49 Lee KH, Park JH, Kim TY, Kim HU, Lee SY. Systems metabolic engineering of *Escherichia coli* for L-threonine production. *Mol. Syst. Biol.* 3, 149 (2007).

50 Lewis NE, Hixson KK, Conrad TM *et al.* Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. *Mol. Syst. Biol.* 6, 390 (2010).

51 Hyduke DR, Lewis NE, Palsson BØ. Analysis of omics data with genome-scale models of metabolism. *Mol. Biosyst.* 9(2), 167–174 (2013).

•• **Provides a solid overview of demonstrated data integration approaches for metabolic networks.**

52 Famili I, Forster J, Nielsen J, Palsson BO. *Saccharomyces cerevisiae* phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proc. Natl Acad. Sci. USA* 100(23), 13134–13139 (2003).

53 Fong SS, Palsson BØ. Metabolic gene-deletion strains of *Escherichia coli* evolve to computationally predicted growth phenotypes. *Nat. Genet.* 36(10), 1056–1058 (2004).

54 Schuster S. Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol.* 17(2), 53–60 (1999).

55 De Figueiredo LF, Podhorski A, Rubio A *et al.* Computing the shortest elementary flux modes in genome-scale metabolic networks. *Bioinformatics* 25(23), 3158–3165 (2009).

56 Jonnalagadda S, Srinivasan R. An efficient graph theory based method to identify every minimal reaction set in a metabolic network. *BMC Syst. Biol.* 8, 28 (2014).

57 Baycin-Hizal D, Tabb DL, Chaerkady R *et al.* Proteomic analysis of Chinese hamster ovary cells. *J. Proteome Res.* 11(11), 5265–5276 (2012).

58 Baumler DJ, Peplinski RG, Reed JL, Glasner JD, Perna NT. The evolution of metabolic networks of *E. coli*. *BMC Syst. Biol.* 5, 182 (2011).

59 Morgan WF, Hartmann A, Limoli CL, Nagar S, Ponnaiya B. Bystander effects in radiation-induced genomic instability. *Mutat. Res.* 504(1–2), 91–100 (2002).

60 Kjeldsen KR, Nielsen J. In silico genome-scale reconstruction and validation of the *Corynebacterium glutamicum* metabolic network. *Biotechnol. Bioeng.* 102(2), 583–597 (2009).

61 David H, Akesson M, Nielsen J. Reconstruction of the central carbon metabolism of *Aspergillus niger*. *Eur. J. Biochem.* 270(21), 4243–4253 (2003).

62 Kildegaard HF, Baycin-Hizal D, Lewis NE, Betenbaugh MJ. The emerging CHO systems biology era: harnessing the 'omics revolution for biotechnology. *Curr. Opin. Biotechnol.* 24(6), 1102–1107 (2013).

63 Hammond S, Kaplarevic M, Borth N, Betenbaugh MJ, Lee KH. Chinese hamster genome database: an online resource for the CHO community at www.CHOgenome.org. *Biotechnol. Bioeng.* 109(6), 1353–1356 (2012).

64 CHOgenome.org. www.chogenome.org/

65 Cao Y, Kimura S, Itoi T, Honda K, Ohtake H, Omasa T. Construction of BAC-based physical map and analysis of chromosome rearrangement in Chinese hamster ovary cell lines. *Biotechnol. Bioeng.* 109(6), 1357–1367 (2012).

66 Kantardjieff A, Nissom PM, Chuah SH *et al.* Developing genomic platforms for Chinese hamster ovary cells. *Biotechnol. Adv.* 27(6), 1028–1035 (2009).

67 Yee JC, de Leon Gatti M, Philp RJ, Yap M, Hu W-S. Genomic and proteomic exploration of CHO and hybridoma cells under sodium butyrate treatment. *Biotechnol. Bioeng.* 99(5), 1186–1204 (2008).

68 Clarke C, Henry M, Doolan P *et al.* Integrated miRNA, mRNA and protein expression analysis reveals the role of post-transcriptional regulation in controlling CHO cell growth rate. *BMC Genomics* 13(1), 656 (2012).

69 Affymetrix. Data Sheet GeneChip® CHO Gene 2.0 ST Array. http://media.affymetrix.com/support/technical/datasheets

70 McGettigan PA. Transcriptomics in the RNA-seq era. *Curr. Opin. Chem. Biol.* 17(1), 4–11 (2013).

71 Rupp O, Becker J, Brinkrolf K *et al.* Construction of a public CHO cell line transcript database using versatile bioinformatics analysis pipelines. *PLoS ONE* 9(1), e85568 (2014).

72 GenDBE – ProCell. A eukaryotic genome browser and annotation system. https://gendbe.cebitec.uni-bielefeld.de/cho.html

73 Slade PG, Hajivandi M, Bartel CM, Gorfien SF. Identifying the CHO secretome using mucin-type O-linked glycosylation and click-chemistry. *J. Proteome Res.* 11(12), 6175–6186 (2012).

74 Meerman HJ, Georgiou G. Construction and characterization of a set of *E. coli* strains deficient in all known loci affecting the proteolytic stability of secreted recombinant proteins. *Biotechnology* 12(11), 1107–1110 (1994).

75 Jin FJ, Watanabe T, Juvvadi PR, Maruyama J, Arioka M, Kitamoto K. Double disruption of the proteinase genes, tppA and pepE, increases the production level of human lysozyme by *Aspergillus oryzae*. *Appl. Microbiol. Biotechnol.* 76(5), 1059–1068 (2007).

76 Becker J, Hackl M, Rupp O *et al.* Unraveling the Chinese hamster ovary cell line transcriptome by next-generation sequencing. *J. Biotechnol.* 156(3), 227–235 (2011).

77 Meleady P, Hoffrogge R, Henry M *et al.* Utilization and evaluation of CHO-specific sequence databases for mass spectrometry based proteomics. *Biotechnol. Bioeng.* 109(6), 1386–1394 (2012).

78 Sellick CA, Croxford AS, Maqsood AR *et al.* Metabolite profiling of recombinant CHO cells: designing tailored feeding regimes that enhance recombinant antibody production. *Biotechnol. Bioeng.* 108(12), 3025–3031 (2011).

79 Sellick CA, Hansen R, Maqsood AR *et al.* Effective quenching processes for physiologically valid metabolite profiling of suspension cultured mammalian cells. *Anal. Chem.* 81(1), 174–183 (2009).

80 Sellick CA, Hansen R, Stephens GM, Goodacre R, Dickson AJ. Metabolite extraction from suspension-cultured mammalian cells for global metabolite profiling. *Nat. Protoc.* 6(8), 1241–1249 (2011).

81 Kronthaler J, Gstraunthaler G, Heel C. Optimizing high-throughput metabolomic biomarker screening: a study of quenching solutions to freeze intracellular metabolism in CHO cells. *OMICS* 16(3), 90–97 (2012).

82 Dietmair S, Timmins NE, Gray PP, Nielsen LK, Krömer JO. Towards quantitative metabolomics of mammalian cells: development of a metabolite extraction protocol. *Anal. Biochem.* 404(2), 155–164 (2010).

83 Dietmair S, Hodson MP, Quek L-E *et al.* Metabolite profiling of CHO cells with different growth characteristics. *Biotechnol. Bioeng.* 109(6), 1404–1414 (2012).

84 Chong WPK, Goh LT, Reddy SG *et al.* Metabolomics profiling of extracellular metabolites in recombinant Chinese hamster ovary fed-batch culture. *Rapid Commun. Mass Spectrom.* 23(23), 3763–3771 (2009).

85 Selvarasu S, Ho YS, Chong WPK *et al.* Combined in silico modeling and metabolomics analysis to characterize fed-batch CHO cell culture. *Biotechnol. Bioeng.* 109(6), 1415–1429 (2012).

•• **Possibly the best example of the application of genome-scale metabolic modeling to CHO cells.**

86 CHO model download. http://cho.sourceforge.net/

87 Martínez VS, Dietmair S, Quek L-E, Hodson MP, Gray P, Nielsen LK. Flux balance analysis of CHO cells before and after a metabolic switch from lactate production to consumption. *Biotechnol. Bioeng.* 110(2), 660–666 (2013).

• **Martínez *et al.* have a solid analysis of carbon fluxes through the CHO cell.**

88 Dietmair S, Hodson MP, Quek L-E, Timmins NE, Gray P, Nielsen LK. A multi-omics analysis of recombinant protein production in Hek293 cells. *PLoS ONE* 7(8), e43394 (2012).

89 Vernardis SI, Goudar CT, Klapa MI. Metabolic profiling reveals that time related physiological changes in mammalian cell perfusion cultures are bioreactor scale independent. *Metab. Eng.* 19, 1–9 (2013).

90 Monica TJ, Andersen DC, Goochee CF. A mathematical model of sialylation of N-linked oligosaccharides in the trans-Golgi network. *Glycobiology* 7(4), 515–521 (1997).

91 Umaña P, Bailey JE. A mathematical model of N-linked glycoform biosynthesis. *Biotechnol. Bioeng.* 55(6), 890–908 (1997).

92 Krambeck FJ, Betenbaugh MJ. A mathematical model of N-linked glycosylation. *Biotechnol. Bioeng.* 92(6), 711–728 (2005).

93 Krambeck FJ, Bennun S V, Narang S, Choi S, Yarema KJ, Betenbaugh MJ. A mathematical model to derive N-glycan structures and cellular enzyme activities from mass spectrometric data. *Glycobiology* 19(11), 1163–1175 (2009).

94 Jimenez Del Val I, Nagy JM, Kontoravdi C. A dynamic mathematical model for monoclonal antibody N-linked

glycosylation and nucleotide sugar donor transport within a maturing Golgi apparatus. *Biotechnol. Prog.* 44(0), 1–44 (2011).

95 Jedrzejewski PM, Del Val IJ, Constantinou A *et al.* Towards controlling the glycoform: a model framework linking extracellular metabolites to antibody glycosylation. *Int. J. Mol. Sci.* 15(3), 4492–4522 (2014).

96 Thiele I, Hyduke DR, Steeb B *et al.* A community effort towards a knowledge-base and mathematical model of the human pathogen *Salmonella typhimurium* LT2. *BMC Syst. Biol.* 5, 8 (2011).

97 Heavner BD, Smallbone K, Price ND, Walker LP. Version 6 of the consensus yeast metabolic network refines biochemical coverage and improves model performance. *Database (Oxford)* 2013, bat059 (2013).

98 Ronda C, Pedersen LE, Hansen HG *et al.* Accelerating genome editing in CHO cells using CRISPR Cas9 and CRISPy, a web-based target finding tool. *Biotechnol. Bioeng.* 111(8), 1604–1616 (2014).

99 Lund AM, Kildegaard HF, Petersen MBK *et al.* A versatile system for USER cloning-based assembly of expression vectors for mammalian cell engineering. *PLoS ONE* 9(5), e96693 (2014).

Chapter 7 - The CHO omics toolbox

In this chapter, a free online database (wiki.bio.dtu.dk/CHOomics) is presented, which contain step by step directions for completing bioinformatics analyses as the ones presented in this thesis. As the community analyzing CHO omics data is tiny, compared to the scientists working with omics data from human or mouse, most databases or methods do not always work easily when applied to CHO omics data. The intention of introducing these pipelines in the thesis and on the online wiki-page is to hopefully reduce the amount of time needed for newcomers to start analyzing omics data from CHO. A short paper *"Expanding the omics toolbox for CHO – a free online resource for NGS tools"* is given, which is intended to be submitted as "Communication to the editor" at Biotechnology and Bioengineering, allowing potential users of the database to cite the pipelines. The pipelines are listed in the appendix.

7.1 Manuscript 2: Expanding the omics toolbox for CHO – a free online resource for NGS tools

Christian Schrøder Kaas[1,2] and Mikael Rørdam Andersen[2]

1 Mammalian Cell Technology, Global Research Unit, Novo Nordisk A/S, Måløv, Denmark.

2 Network Engineering of Eukaryotic Cell Factories, Technical University of Denmark, Kgs Lyngby, Denmark.

7.1.1 Abstract

The CHO bioinformatics wiki (wiki.bio.dtu.dk/CHOomics) is an online resource for the Chinese hamster (*Cricetulus griseus*) and Chinese hamster ovary (CHO) cell communities. As the price has dropped dramatically for next generation sequencing during the past decade, sequencing of CHO genomes and transcriptomes are now becoming routine. In order to answer important biological questions from the data, complex bioinformatics pipelines are needed. The CHOomics bioinformatics wiki is intended to convey suggestions for how to do routine analyses such as finding differentially expressed genes in an RNAseq experiment and locate changes in copy number in a genome.

KEYWORDS: Chinese hamster ovary cell; CHO genome; CHO transcriptome; database

Chinese Hamster Ovary (CHO) cells, originally isolated by Ted Puck back in 1957 [1], have had a long history as a production organism in industry due to their ability to grow in suspension without serum and to be scalable to large production volumes. Biopharmaceuticals are currently generating sales of approximately 140 billion USD of which the majority of proteins requiring post-translational modifications are produced CHO cells [2]. Compared to the human genome [3] that has been available since 2001, CHO first entered the genomic era a decade later in 2011 [4] with several additional genomes being released recently [5-7]. For this reason tools such as standardized qPCR primers and databases of validated SNPs are not available yet making omics analysis on CHO a bit more time consuming than equivalent work in cell lines extracted from human or mouse tissue [8,9].

Following the publication of the CHO-K1 genome back in 2011 [4], the online database www.CHOgenome.org was created allowing the user community to blast the genome as well as look up the transcripts and amino-acid sequence encoded by the annotated genes [10,11]. As valuable as these first steps were, simple descriptions on how to download the CHO or *C.griseus* genome and a run simple analyzes are lacking. In order to help the newcomer to the field of CHO omics the website wiki.bio.dtu.dk/CHOomics has been created allowing access to bioinformatics pipelines used for CHO omics analyzes. In the database, suggestions are listed in a step-by-step

manner on how to accomplish the most standard bioinformatics problems in the field. The database is organized as a wiki-page allowing all registered users to correct and add information.

A "getting started" section is introduced, specifically with the purpose of introducing the genome and the annotation file, which are the core essential for the majority of omics analyses to be run. Information regarding e.g. how to access and mine the annotation file to generate a list of genes are explained.

Unfortunately, a CHO-specific SNP database does not exist at the time of writing, but using the code listed on the website, suggestions can be given specifically on how to locate SNPs and how to apply a hard-filter to the available CHO genomes, which takes into account that approximately 20% of the genes in the CHO genomes have been found to be haploid and thus cannot be mined similarly as a strictly diploid genome [6].

In the current version of www.CHOgenome.org only the sequence of the mature transcripts are listed, which make qPCR primer design of intron-spanning primers very time-consuming. At the wiki, the gene sequence for all the genes in the *C. griseus* genome can be accessed with annotation of intron and exon positions making qPCR design straight forward.

The initial focus of the website is to supply the most basic needs that a newcomer to the field might have, but the intention is that more sophisticated pipelines can be listed in the future. This way, scientists can be inspired with additional methods to be applied to their data and find solutions for the most basic problems they might encounter when analyzing CHO omics data.

7.1.2 References

1. Puck TT: **Genetics Of Somatic Mammalian Cells: III. Long-term Cultivation Of Euploid Cells From Human And Animal Subjects.** *Journal of Experimental Medicine* 1958, **108:** 945-956.

2. Walsh G: **Biopharmaceutical benchmarks 2014.** *Nature biotechnology* 2014, **32:** 992-1000.

3. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J *et al.*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:** 860-921.

4. Xu X, Nagarajan H, Lewis NE, Pan S, Cai Z, Liu X *et al.*: **The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line.** *Nature biotechnology* 2011, **29:** 735-741.

5. Lewis C, Galloway T: **Reproductive consequences of paternal genotoxin exposure in marine invertebrates.** *Environ Sci Technol* 2009, **43:** 928-933.

6. Kaas CS, Kristensen C, Betenbaugh MJ, Andersen MR: **Sequencing the CHO DXB11 genome reveals regional variations in genomic stability and haploidy.** *BMC genomics* 2015, **16:** 1391.

7. Brinkrolf K, Rupp O, Laux H, Kollin F, Ernst W, Linke B *et al.*: **Chinese hamster genome sequenced from sorted chromosomes.** *Nature biotechnology* 2013, **31:** 694-695.

8. Wang X, Spandidos A, Wang H, Seed B: **PrimerBank: a PCR primer database for quantitative gene expression analysis, 2012 update.** *Nucleic acids research* 2012, **40:** D1144-D1149.

9. Smigielski EM: **dbSNP: a database of single nucleotide polymorphisms.** *Nucleic acids research* 2000, **28:** 352-355.

10. Hammond S, Kaplarevic M, Borth N, Betenbaugh MJ, Lee KH: **Chinese hamster genome database: an online resource for the CHO community at www.CHOgenome.org.** *Biotechnology and bioengineering* 2012, **109:** 1353-1356.

11. Kremkow BG, Baik JY, MacDonald ML, Lee KH: **CHOgenome.org 2.0: Genome resources and website updates.** *Biotechnology journal* 2015.

Chapter 8 - Conclusions and Future Perspectives

The aim of this thesis was to investigate the impact of FVIII production in CHO cells using NGS tools. From sequencing the genomes of the CHO DXB11 cell line, as well as a FVIII transfectant, it was found that the DHFR$^{-/-}$ phenotype could be validated from the genome sequence and in addition to the *dhfr* gene, it was found that roughly 20% of all other the genes were also haploid. These findings might aid researchers, using genome editing tools such as Crispr/Cas9, to find targets within a given pathway that are haploid and thus facilitate knock out. It was concluded that chromosome one and four might be more stable in terms of copy number variations compared to the other chromosomes. With the advent of the improved *C. griseus* genome and possibly additional CHO genomes published in the years to come, it might be possible to more accurately locate stable islands within the CHO genome. These stable regions might present themselves to be suitable safe harbors for targeted integration of transgenes. Current protocols for generation of stable cell lines largely depends on random integration of transgenes into the CHO genome and simply screen for cells with highest productivity, but it is often observed that high producing cell lines are unstable over time due to silencing or genome instability. By targeting the transgene into a stable region of the genome it might be possible to create stable cell lines sustaining high productivity over longer periods of time despite lower gene copy number. Furthermore, the targeted approach allow for better comparisons between cell lines when different transgenes may be expressed from the same locus under the same conditions.

The genome analysis also showed that 907 genes had undergone copy number changes in the FVIII transfectant F435 cell line as compared to the CHO DXB11 host cell line. When more CHO genomes become available it will be interesting to see whether such extensive rearrangements are a common trait among transfectants, or it is found to be a special case, possibly due to the burden put onto cells when producing FVIII. It would in particular be interesting to see how many gene copy number variations that are found in a genome, if targeted integration is used, and the cell line is not subsequently amplified by MTX. CHO has been used in the industry for more than three decades, but in the years to come, we might just learn that these cells are more plastic that we even imagined, due to the new possibilities within omics for measuring these fluctuations.

Using RNAseq data it was presented in this thesis that transgene expression, splicing of introns and detection of truncations can reliably be monitored. As the data can furthermore be used for SNP detection in the transgene, it will probably soon become standard practice at companies to run RNAseq analyses of the most promising clones in order to validate that the transgene is error-free

and transcribed as designed. If the Food and Drug Administration should choose to make whole-genome sequencing, or targeted sequencing of transgene and transcripts, mandatory for production cell lines, biotech companies will soon be generating vast quantities of omics data. As NGS data can quickly be purged for transgene sequence by simply removing all reads aligning to the transgene sequence (or any other region desired to be kept secret), the possibly of donating data from private companies to the universities will become less problematic.

Using RNAseq and MS it was shown that FVIII production in CHO cells was accomplished at the cost of a high metabolic burden and stress. As the endoplasmic-reticulum-associated protein degradation pathway was found to be up-regulated there appear to be ample room for improvement of folding capacity of CHO for FVIII production. Several chaperones were overexpressed, but they did not impact the FVIII productivity significantly. In order to improve the folding capacity, it will be needed to test a broader panel of targets, combinations of targets and perhaps even different promoter strength for expression of the target chaperones. Such systematic testing will require a high throughput setup and screening assay. Given the general challenges in the gene therapy field FVIII will most likely be produced as a recombinant drug for the foreseeable future, and thus further work on cell line optimization will be carried out especially for future generations of FVIII products with longer serum half-life.

Within the next few months a CHO consensus genome-scale metabolic model is expected to be made public. In the years to come it will be interesting to see the power of the model for predictions of cell line engineering targets, as well as the potential for generating specific models for the different CHO cell lines. In order to generate a model specific for e.g. CHO DXB11 the omics data presented in this thesis could be mined in order to identify cell line specific patterns in transcriptome data and genomic copy numbers.

The future of CHO omics looks promising; with a better reference genome and annotation coming out, as well as a genome-scale metabolic model expanding the omics toolbox to a levels catching up with most other model organisms. By improving our understanding of this industrially important cell line, the prospect of generating increasingly efficient CHO cell lines for production of complex biopharmaceuticals such as FVIII, is surely promising.

About the Author

Personal information
Name: Christian Schrøder Kaas
Work email: csrk@novonordisk.com/ckaa@bio.dtu.dk
Private email: kaaschr@gmail.com
Private Cell phone: +45 6167 3410

Publications

- **Kaas, C. S.**, Kristensen, C., Betenbaugh, M.J., Andersen, M.R. Sequencing the CHO DXB11 genome reveals regional variations in genomic stability and haploidy. BMC Genomics. Submitted
- **Kaas, C. S.**, Bolt, B., Hansen, J.J., Andersen, M.R., Kristensen, C. Deep sequencing reveals different compositions of mRNA transcribed from the FVIII gene in a panel of FVIII producing CHO cell lines. Biotechnology Journal. Submitted
- **Kaas, C. S.**, Fan, Y., Weilguny, D., Kristensen, C., Kildegaard, H. F., & Andersen, M. R. (2014). Toward genome-scale models of the Chinese hamster ovary cells: incentives, status and perspectives. Pharmaceutical Bioprocessing, 2(5), 437-448.
- Heffner, K., **Kaas, C.S.**, Kumar, A., Baycin-Hizal, D., Betenbaug, M.J. (2015). Animal Cell Culture, Chapter 19 : Proteomics in Cell Culture: From genomics to combined 'omics for cell line engineering and bioprocess development.

Courses and training during PhD

- Biological Interpretation of Next-Generation Sequencing Data, European Bioinformatics Institute, UK
- Project Management for PhD students, Technical University of Denmark
- Glycobiology, University of Copenhagen
- Ph.D. Course in Management in Science and Innovation, Technical University of Denmark
- Bioinformatics and UNIX, Technical University of Denmark
- DNA Microarray Analysis, Technical University of Denmark
- Biological data analysis and chemometrics, Technical University of Denmark

Teaching and supervision

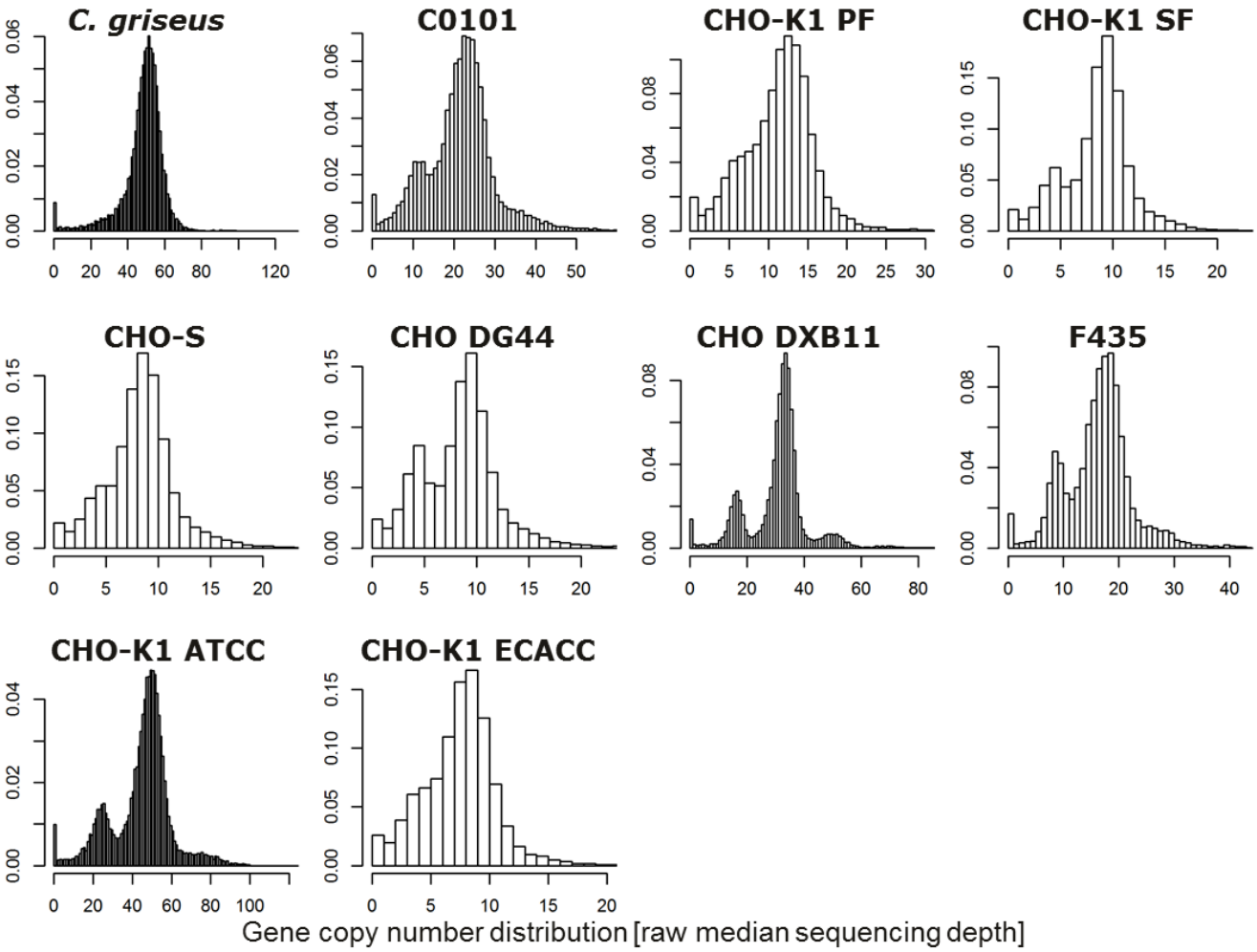
- 2012: Invited lecturer on "Mammalian cell cultivations for production of heterologous proteins", Master course at The Technical University of Denmark
- 2013: Poster presentation at Cell Factories and Biosustainability conference, Hillerød, Denmark
- 2013: Poster presentation at European Society for Animal Cell Technology conference, Lille, France
- 2013: Invited lecturer at the STAR Symposium in Bagsværd, Denmark
- 2013: Invited lecturer and scholarship recipient at Animal Cell Technology Industrial Platform meeting in Vevey, Switzerland
- 2013-2014: Master student supervisor for Anahita Zamani, DTU
- 2014: Poster presentation at CHOgenome Workshop 2014, Vienna, Austria
- 2015: Poster presentation at European Society for Animal Cell Technology conference, Barcelona, Spain

Appendix

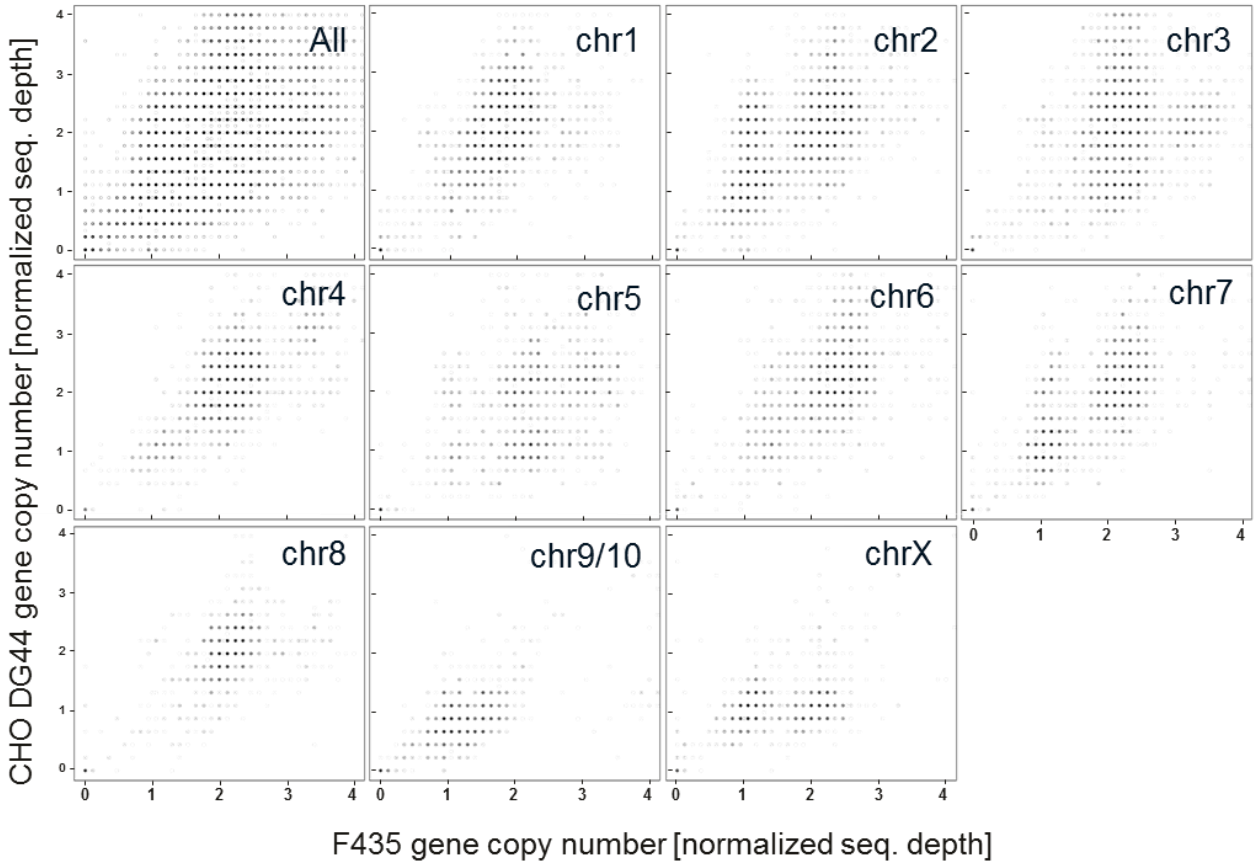
10.1 Supplementaries for Chapter 2

10.1.1 Supplementary figures

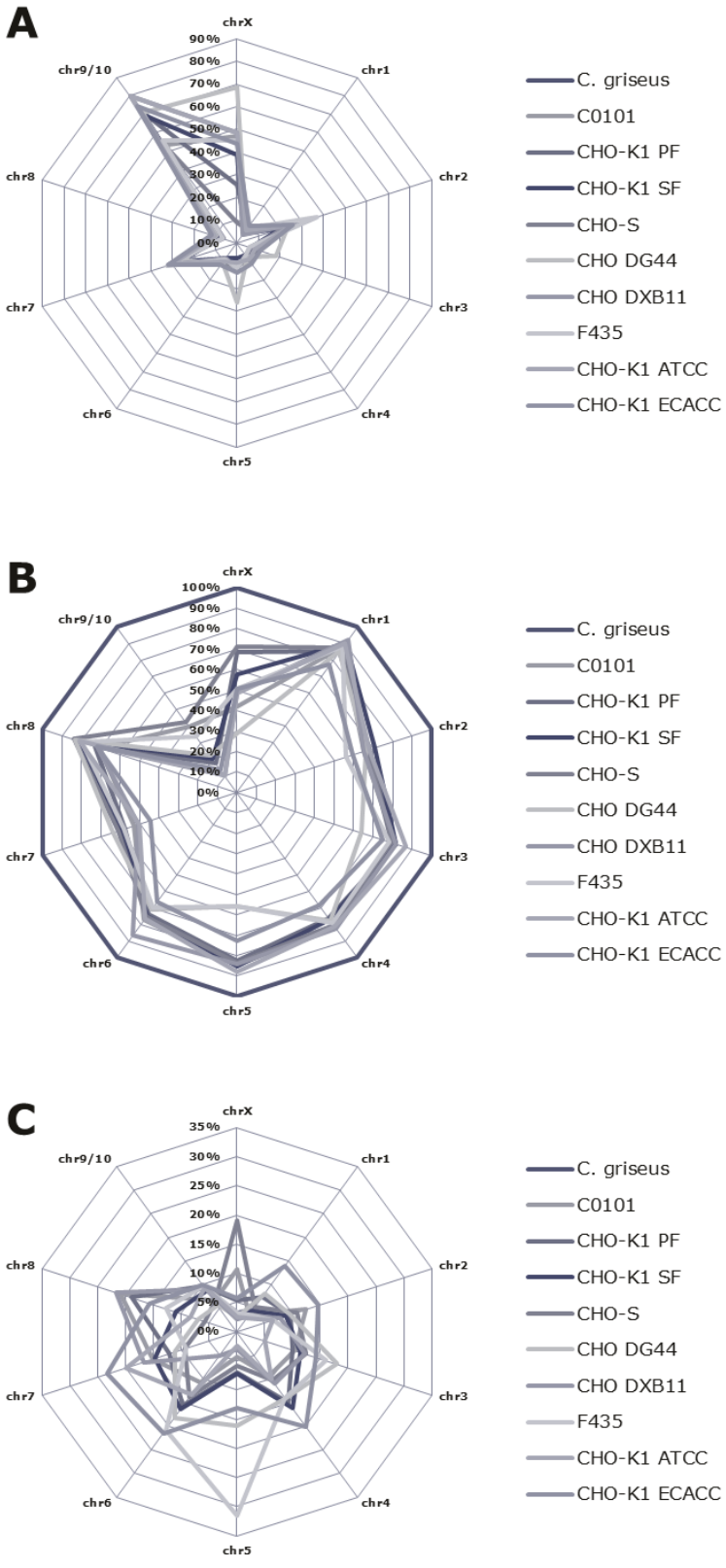
The supplementary materials shown below can be downloaded from BMC genomics (http://www.biomedcentral.com/content/supplementary/s12864-015-1391-x-s2.docx or http://bit.ly/DXB11suppl1)



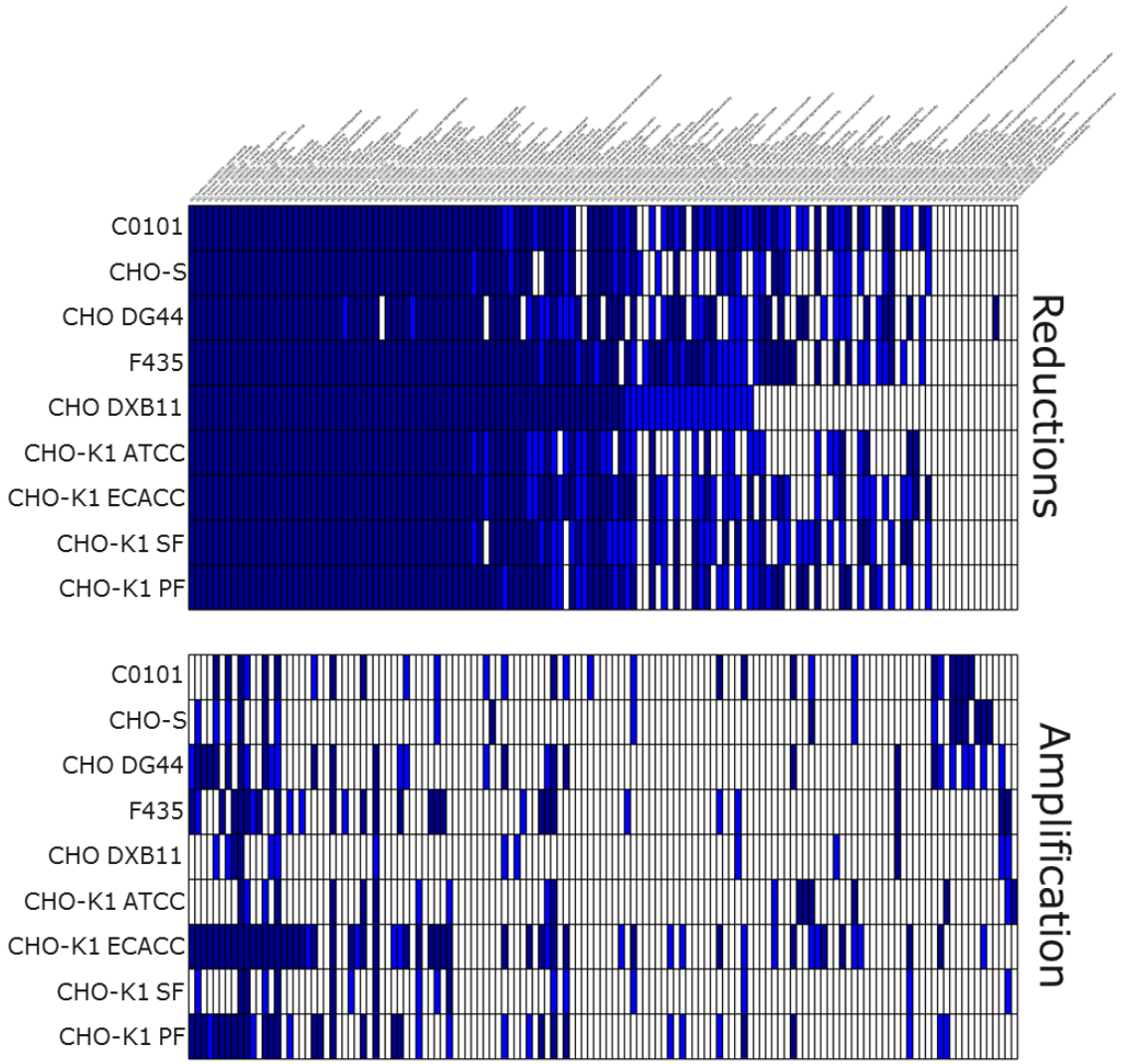
Supplementary figure 1 Read depth analysis of the 20661 in the *C. griseus* genome.
For most of the genomes distinct peaks can be seen for genes present in one, two and three copies. Shoulders can be seen in the graphs for the cell lines sequenced at a lower depth. Only one peak is seen in wild type *C. griseus* as expected.



Supplementary figure 2 The normalized sequencing depth of each gene in F435 and CHO DG44.
Top left plot shows the distribution across all chromosomes. Compared to Figure 2 this reveals a much larger difference in CN.



Supplementary figure 3 Distribution of CN across chromosomes for each cell line.
A) Percentage of haploid genes **B)** percentage of diploid genes **C)** Percentage of genes, which are triploid or higher.



Supplementary figure 4 Significant GO-terms in correlation to changes in CN
Visualization of the 135 GO-terms, which are either significantly enriched in genes with CN reductions or amplifcations (Fisher's exact test). GO-terms are visualized in dark blue (p-value < 0.01), light blue (p-value < 0.05) or white (p-value > 0.05). Data attached in Supplementary table 7.

10.1.2 Supplementary tables

The supplementary tables shown below can be downloaded from BMC genomics (http://www.biomedcentral.com/content/supplementary/s12864-015-1391-x-s1.xlsx or http://bit.ly/DXB11suppl2). Listed below the reader will find subsections of the supplementary tables in order to give an overview of what information is stored in each table.

Supplmentary table 1. SRA
A list of SRA accession number for the publicly available CHO genomes (including CHO DXB11).

CHO cell name	SRA accesion number	Original publication
CHO DXB11	SRR1561441	Kaas et al, 2015
CHO DXB11	SRR1561442	Kaas et al, 2015
CHO DXB11	SRR1561427	Kaas et al, 2015
CHO DXB11	SRR1561428	Kaas et al, 2015
CHO-K1 ATCC	SRR329939	Xu et al, 2011
CHO-K1 ATCC	SRR329940	Xu et al, 2011
CHO-K1 ATCC	SRR329941	Xu et al, 2011
CHO-K1 ATCC	SRR329942	Xu et al, 2011
CHO-K1 ATCC	SRR329943	Xu et al, 2011
CHO-K1 ATCC	SRR329944	Xu et al, 2011
CHO-K1 ATCC	SRR329945	Xu et al, 2011
CHO-K1 ATCC	SRR329946	Xu et al, 2011
CHO-K1 ATCC	SRR329947	Xu et al, 2011
CHO-K1 ATCC	SRR329948	Xu et al, 2011
CHO-K1 ATCC	SRR329949	Xu et al, 2011
CHO-K1 ATCC	SRR329950	Xu et al, 2011
CHO-K1 ATCC	SRR329951	Xu et al, 2011
CHO-K1 ATCC	SRR329952	Xu et al, 2011
CHO-K1 ATCC	SRR329953	Xu et al, 2011
CHO-K1 ATCC	SRR329954	Xu et al, 2011
CHO K1 Protein free	SRR803173	Lewis et al, 2013
CHO K1 Protein free	SRR803174	Lewis et al, 2013
CHO K1 Protein free	SRR803175	Lewis et al, 2013
CHO-K1 ECACC	SRR803176	Lewis et al, 2013
CHO-K1 ECACC	SRR803177	Lewis et al, 2013
CHO-K1 ECACC	SRR803178	Lewis et al, 2013
CHO-K1 Serum free	SRR803179	Lewis et al, 2013
CHO-K1 Serum free	SRR803180	Lewis et al, 2013
CHO-K1 Serum free	SRR803181	Lewis et al, 2013
CHO-S	SRR803182	Lewis et al, 2013
CHO-S	SRR803183	Lewis et al, 2013
CHO DG44	SRR803184	Lewis et al, 2013
CHO DG44	SRR803185	Lewis et al, 2013
C0101	SRR801491	Lewis et al, 2013
C0101	SRR801492	Lewis et al, 2013
C0101	SRR801493	Lewis et al, 2013
C0101	SRR801494	Lewis et al, 2013
C0101	SRR801495	Lewis et al, 2013
C0101	SRR801496	Lewis et al, 2013

For download from SRA, customize the link below by the acession number
ftp://ftp-trace.ncbi.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR156/SRR1561441/SRR1561441.sra

Supplmentary table 2. SNP geneoverview

List of the number of SNPs detected in each gene in the CHO DXB11 and CHO-K1 ATCC genome. Truncated view

Genbank	Gene-name	Cell line	SNPs/indels total	SNPs/indels in introns	SNPs/indels in exon (silent)	SNPs/indels in exon (missense/nonsense)
XP_007626087.1	A1cf	CHO_DXB11	5	5	0	0
XP_007609693.1	A2ml1	CHO_DXB11	149	136	5	8
XP_007613122.1	A3galt2	CHO_DXB11	12	10	1	1
XP_007627973.1	A4galt	CHO_DXB11	1	0	0	1

...

Genbank	Gene-name	Cell line	SNPs/indels total	SNPs/indels in introns	SNPs/indels in exon (silent)	SNPs/indels in exon (missense/nonsense)
XP_007626087.1	A1cf	K1ATCC	3	3	0	0
XP_007609693.1	A2ml1	K1ATCC	139	126	6	7
XP_007613122.1	A3galt2	K1ATCC	11	10	0	1
XP_007627973.1	A4galt	K1ATCC	1	0	0	1

...

Supplmentary table 3. Exon SNPs DXB11
Overview of all SNPs and indels affecting exonic sequence in CHO DXB11. Selected rows seen.

Gene	Cell line	Type	Length	Zygocity	Depth	DP4	Implication for translation	Sequencing scaffold	Transcript	Protein
Dhfr	CHO_DXB11	SNV	1	Homozygous	17	0,0,9,8	Yes	NW_006884452.1.3862471C>G	NP_001230945.1.410C>G	NP_001230945.1.Thr137Arg
Dhdh	CHO_DXB11	SNV	1	Homozygous	19	0,0,12,5	Yes	NW_006886266.1.1266320G>A	XP_007636926.1.350G>A	XP_007636926.1.Gly117Asp
Dgkz	CHO_DXB11	SNV	1	Homozygous	48	0,0,24,20	Yes	NW_006881596.1.252933C>T	XP_007629092.1.1342G>A	XP_007629092.1.Ala448Thr
Dgkq	CHO_DXB11	SNV	1	Homozygous	24	0,0,10,12	Yes	NW_006871769.1.2241378T>C	XP_007611560.1.1341A>G	XP_007611560.1.Ile447Met
Dgkk	CHO_DXB11	SNV	1	Heterozygous	42	11,9,14,4	Yes	NW_006877738.1.643720C>T	XP_007617940.1.2603C>T	XP_007617940.1.Thr868Ile
Dgki	CHO_DXB11	SNV	1	Heterozygous	34	6,9,11,8	Yes	NW_006875242.1.696470C>T	XP_007615711.1.2843C>T	XP_007615711.1.Ser948Phe
Dgki	CHO_DXB11	SNV	1	Heterozygous	39	11,14,7,6	Yes	NW_006875242.1.694750G>C	XP_007615711.1.2655G>C	XP_007615711.1.Glu885Asp

Supplmentary table 4. Exon SNPs K1ATCC
Overview of all SNPs and indels affecting exonic sequence in CHO-K1 ATCC. Selected rows seen.

Gene	Cell line	Type	Length	Zygocity	Depth	DP4	Implication for translation	Sequencing scaffold	Transcript	Protein
Dhx35	K1ATCC	SNV	1	Homozygous	47	0,0,28,14	Yes	NW_006869199.1.157649G>A	XP_007607785.1.1955+2376C>T	
Dhx35	K1ATCC	SNV	1	Homozygous	47	0,0,28,14	Yes	NW_006869199.1.157649G>A	XP_007607784.1.1961C>T	XP_007607784.1.Thr654Met
Dhx35	K1ATCC	SNV	1	Homozygous	47	0,0,28,14	Yes	NW_006869199.1.157649G>A	XP_007607783.1.2066C>T	XP_007607783.1.Thr689Met
Dhx32	K1ATCC	SNV	1	Heterozygous	43	13,9,4,16	Yes	NW_006880038.1.2114629G>T	XP_007623878.1.967G>T	XP_007623878.1.Ala323Ser
Dhx32	K1ATCC	SNV	1	Heterozygous	43	13,9,4,16	Yes	NW_006880038.1.2114629G>T	XP_007623877.1.967G>T	XP_007623877.1.Ala323Ser
Dhx29	K1ATCC	SNV	1	Homozygous	48	0,0,25,23	Yes	NW_006874248.1.306469A>G	XP_007614384.1.1523A>G	XP_007614384.1.His508Arg
Dhx15	K1ATCC	SNV	1	Heterozygous	43	8,13,10,9	Yes	NW_006879217.1.902723T>G	XP_007618842.1.2014A>C	XP_007618842.1.Asn672His

Supplmentary table 5. Sequencing depth

The absolute CN, the raw depth measured and the normalized CN value for each gene in the *C. griseus* genome in each genome

Median	2	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**

Absolute CN	Cgriseus	C0101	CHO PF	CHO SF	CHO-S	DG44	DXB11	F435	K1 ATCC	K1 ECACC
A1cf	2	3	2	2	2	3	2	2	2	2
A2ml1	2	2	2	2	2	2	2	2	2	2
A3galt2	2	2	2	2	2	2	2	2	2	2
A4galt	2	2	2	2	2	2	2	2	2	2
A4gnt	2	2	2	2	1	2	2	2	2	2
Aaas	2	2	2	2	2	5	2	2	2	2

…

Median	**2,3**	**3,3**	**2,3**	**2,4**	**2,7**	**4,7**	**2,4**	**2,6**	**2,3**	**2,5**
Raw	Cgriseus	C0101	CHO PF	CHO SF	CHO-S	DG44	DXB11	F435	K1 ATCC	K1 ECACC
A1cf	46	38	14	9	12	14	32	18	49	10
A2ml1	51	16	9	7	6	7	30	16	37	6
A3galt2	50	17	9	9	9	9	29,5	17	42	7
A4galt	58	21	13	11	11	12	39	22	53	10
A4gnt	48	18	12	9	5	7	33	19	50	7
Aaas	49	24	13	10	11	21	35	22	54	9

…

Median	2	2	2	2	2	2	2	2	2	2
Normalized	Cgriseus	C0101	CHO PF	CHO SF	CHO-S	DG44	DXB11	F435	K1 ATCC	K1 ECACC
A1cf	1,8	3,3	2,3	2	2,7	3,1	1,9	2,1	2	2,5
A2ml1	2,0	1,4	1,5	1,6	1,3	1,6	1,8	1,9	1,5	1,5
A3galt2	2,0	1,5	1,5	2,0	2,0	2,0	1,8	2,0	1,8	1,8
A4galt	2,3	1,8	2,2	2,4	2,4	2,7	2,4	2,6	2,2	2,5
A4gnt	1,9	1,6	2,0	2,0	1,1	1,6	2,0	2,2	2,1	1,8
Aaas	1,9	2,1	2,2	2,2	2,4	4,7	2,1	2,6	2,3	2,3

…

Supplmentary table 6. C. griseus overview
Lists general information for all genes in the *C. griseus* genome

Gene name	Genbank (gene)	Genbank (protein)	RNA ID	Genomic scaffold	Chromosome	GC%	Transcript length	Protein length [AA]	Gene name2
A1cf	XM_007627897.1	XP_007626087.1	rna20866	NW_006880426.1	3	47,7	1782	593	APOBEC1 complementation factor, transcript variant X2
A2ml1	XM_007611503.1	XP_007609693.1	rna3450	NW_006870833.1	8	51,5	4371	1456	alpha-2-macroglobulin-like 1, transcript variant X2
A3galt2	XM_007614932.1	XP_007613122.1	rna7152	NW_006873147.1	2	59,1	1014	337	alpha 1,3-galactosyltransferase 2
A4galt	XM_007629783.1	XP_007627973.1	rna22847	NW_006881179.1	2	52,9	1945	348	alpha 1,4-galactosyltransferase, transcript variant X1
A4galt	XM_007629784.1	XP_007627974.1	rna22846	NW_006881179.1	2	53,4	1871	348	alpha 1,4-galactosyltransferase, transcript variant X2
A4gnt	XM_007634050.1	XP_007632240.1	rna27323	NW_006883256.1	4	51,6	1020	339	alpha-1,4-N-acetylglucosaminyltransferase
Abat	XM_007638296.1	XP_007636486.1	rna31987	NW_006885922.1	7	50,9	1335	444	4-aminobutyrate aminotransferase, transcript variant X2

Supplmentary table 7. GO-terms

Lists all GO-terms with the number of genes found to have reduced or amplified copy number within this particular category and the corresponding p-value.

Number of genes reduced in CN

Overall ranking	GO-term	Genes with GO total	C0101	CHO-K1 PF	CHO-K1 SF	CHO-S	CHO DG44	CHO DXB11	F435	CHO-K1 ATCC	CHO-K1 ECACC
1	GO_function: GO:0005515 - protein binding	1160	51	52	46	44	101	71	64	55	57
3	GO_function: GO:0008270 - zinc ion binding	1054	64	56	48	54	85	73	59	64	57
5	GO_function: GO:0003700 - transcription factor activity	432	17	14	14	18	36	14	14	17	18
6	GO_function: GO:0005524 - ATP binding	935	54	57	45	49	68	76	56	58	53

p-value (Fischer Exact test) for reduced genes

Overall ranking	GO-term	Genes with GO total	C0101	CHO-K1 PF	CHO-K1 SF	CHO-S	CHO DG44	CHO DXB11	F435	CHO-K1 ATCC	CHO-K1 ECACC
1	GO_function: GO:0005515 - protein binding	1160	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
3	GO_function: GO:0008270 - zinc ion binding	1054	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
5	GO_function: GO:0003700 - transcription factor activity	432	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
6	GO_function: GO:0005524 - ATP binding	935	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%

Number of genes amplified in CN

Overall ranking	GO-term	Genes with GO total	C0101	CHO-K1 PF	CHO-K1 SF	CHO-S	CHO DG44	CHO DXB11	F435	CHO-K1 ATCC	CHO-K1 ECACC

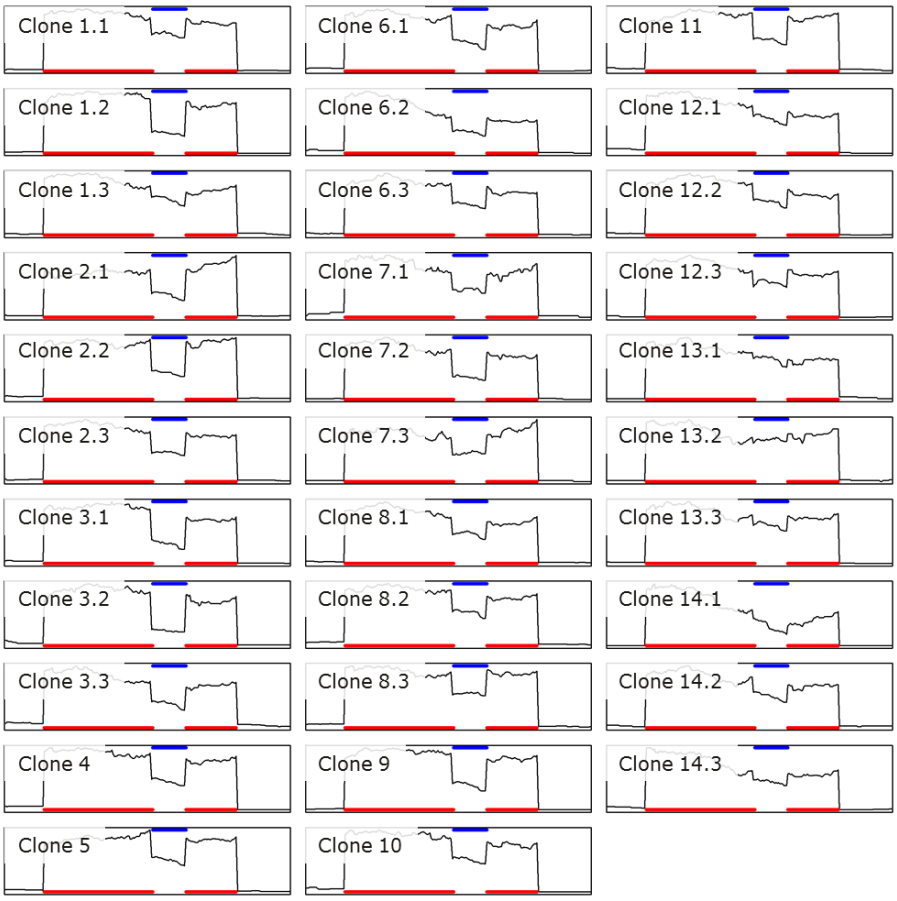
1	GO_function: GO:0005515 - protein binding	1160	101	82	107	107	114	103	105	74	104
3	GO_function: GO:0008270 - zinc ion binding	1054	88	71	98	83	97	89	105	64	83
5	GO_function: GO:0003700 - transcription factor activity	432	24	23	33	27	26	49	54	25	28
6	GO_function: GO:0005524 - ATP binding	935	83	65	84	92	99	70	83	63	78

p-value (Fischer Exact test) for amplified genes

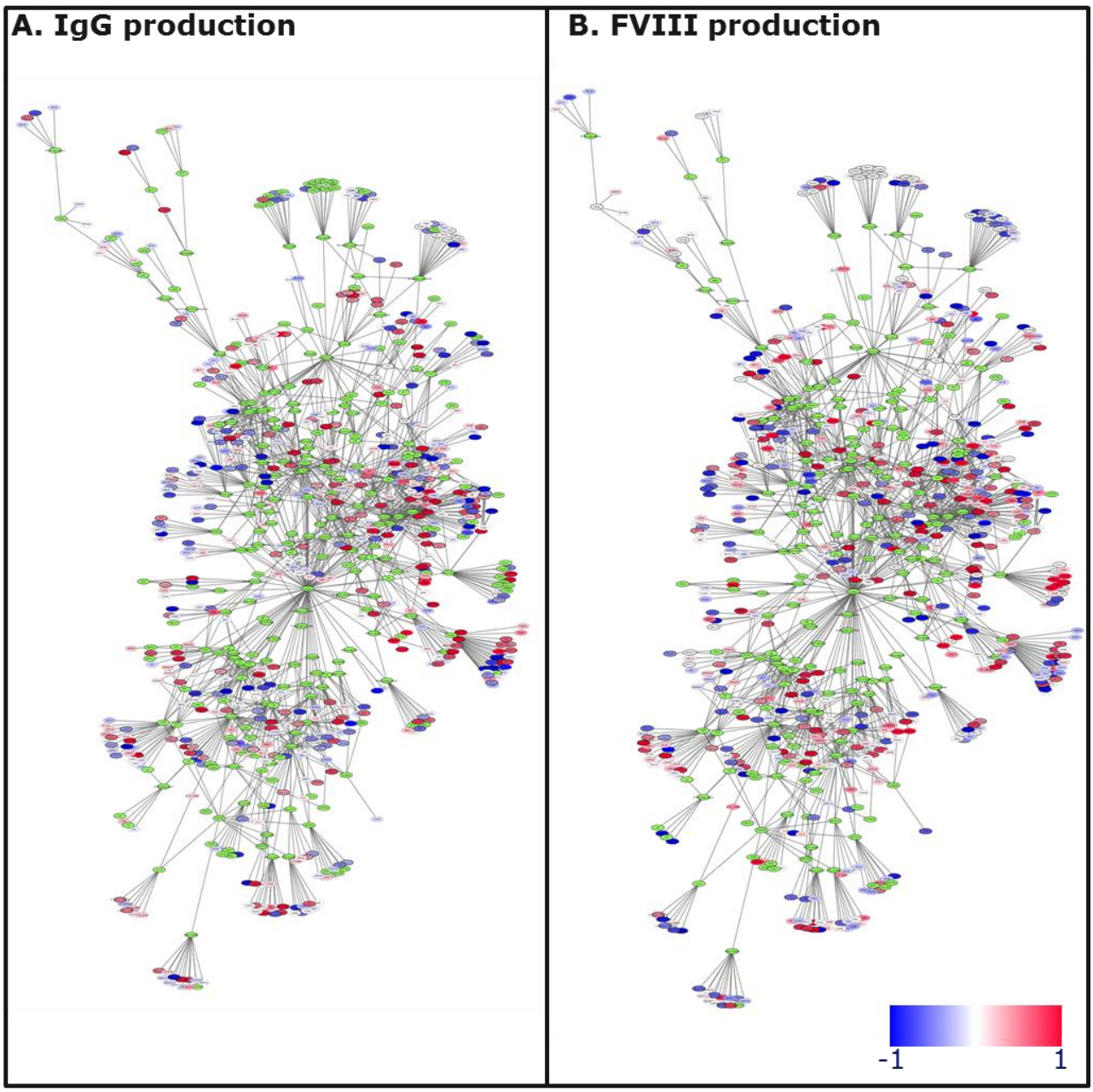
Overall ranking	GO-term	Genes with GO total	C0101	CHO-K1 PF	CHO-K1 SF	CHO-S	CHO DG44	CHO DXB11	F435	CHO-K1 ATCC	CHO-K1 ECACC
1	GO_function: GO:0005515 - protein binding	1160	28,2%	0,0%	34,0%	68,2%	4,3%	66,5%	0,4%	13,6%	0,0%
3	GO_function: GO:0008270 - zinc ion binding	1054	14,8%	0,0%	40,1%	5,3%	0,9%	95,5%	7,6%	7,2%	0,0%
5	GO_function: GO:0003700 - transcription factor activity	432	0,3%	0,0%	9,0%	1,6%	0,0%	4,5%	59,6%	19,6%	0,0%
6	GO_function: GO:0005524 - ATP binding	935	46,1%	0,0%	26,7%	82,0%	32,2%	25,5%	0,6%	40,9%	0,0%

10.2 Supplementaries for Chapter 4

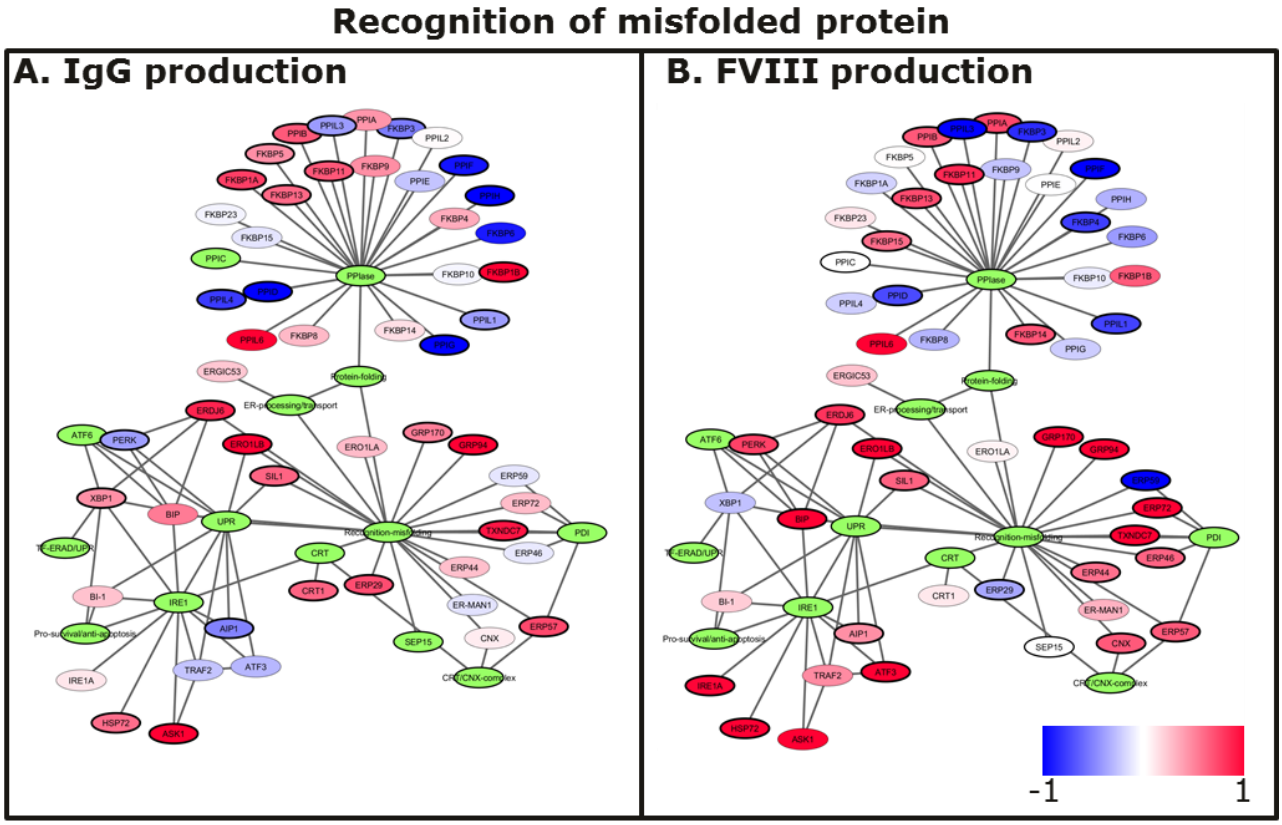
10.2.1 Supplementary figures



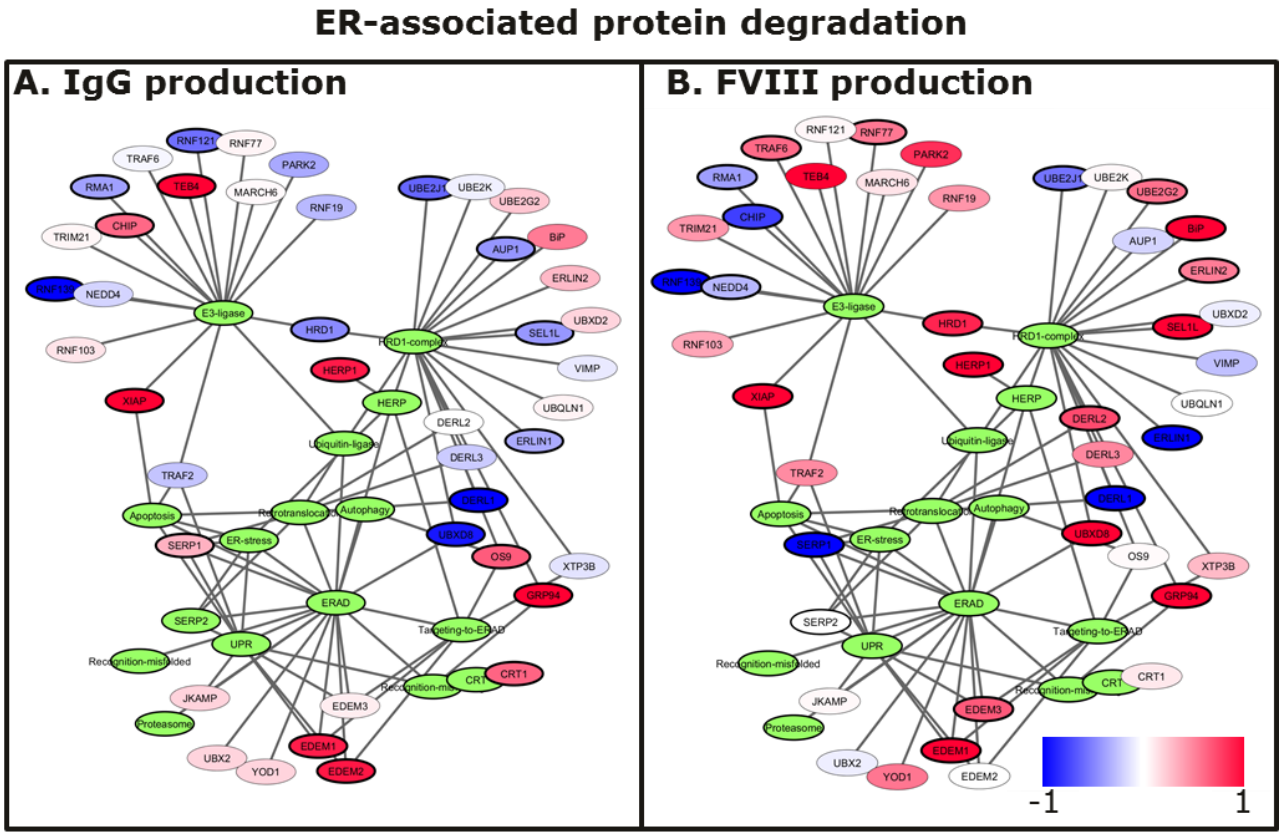
Supplementary Figure 1 Read depth distribution of exon 3 for the Xbp1 transcript. Exonuclease activity of Ern1 (IRE1α) removes a 26bp intron inside this exon (NW_006879406.1 position 31183-31209) causing a frameshift conversion. The ratio of the median read depth inside this region (blue line) was used relative to the median for the rest of the exon (red lines) as a measure of Xbp1 induction.



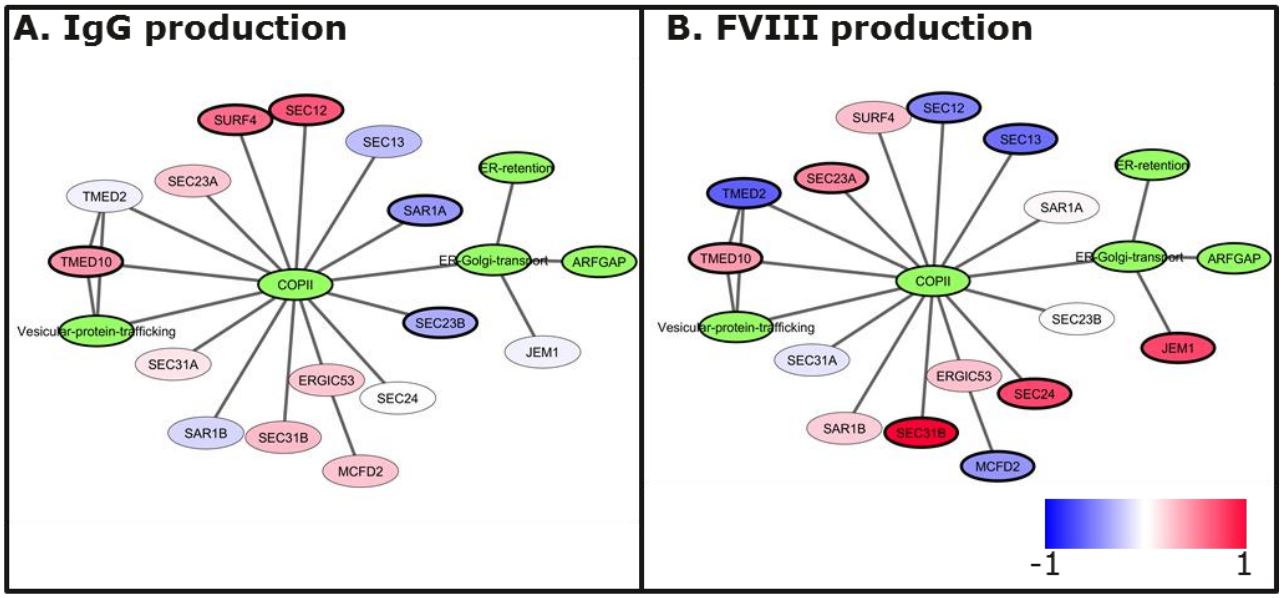
Supplementary Figure 2 CHO specific secretion network overlaid with RNAseq data.
A) Colors indicate log2(foldchange) between IgG producing CHO-K1 cells in exponential growth phase versus non-producing control CHO-K1 cells. B) Colors indicate log2(foldchange) between the FVIII producing cell lines (Clone 1-11) versus control cell lines (Clone 12-13).



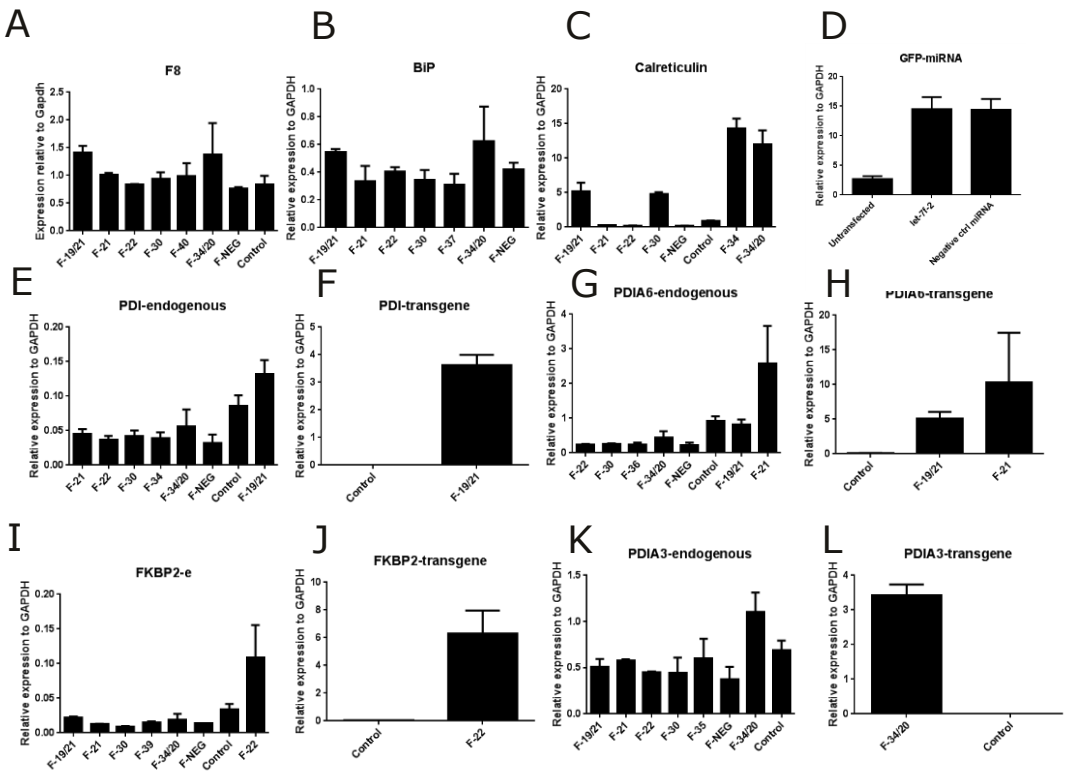
Supplementary Figure 3 subnetwork specific for genes relevant for recognition of misfolded proteins overlaid with RNAseq data.
A) Colors indicate log2(foldchange) between IgG producing CHO-K1 cells in exponential growth phase versus non-producing control CHO-K1 cells. B) Colors indicate log2(foldchange) between the FVIII producing cell lines (Clone 1-11) versus control cell lines (Clone 12-13).



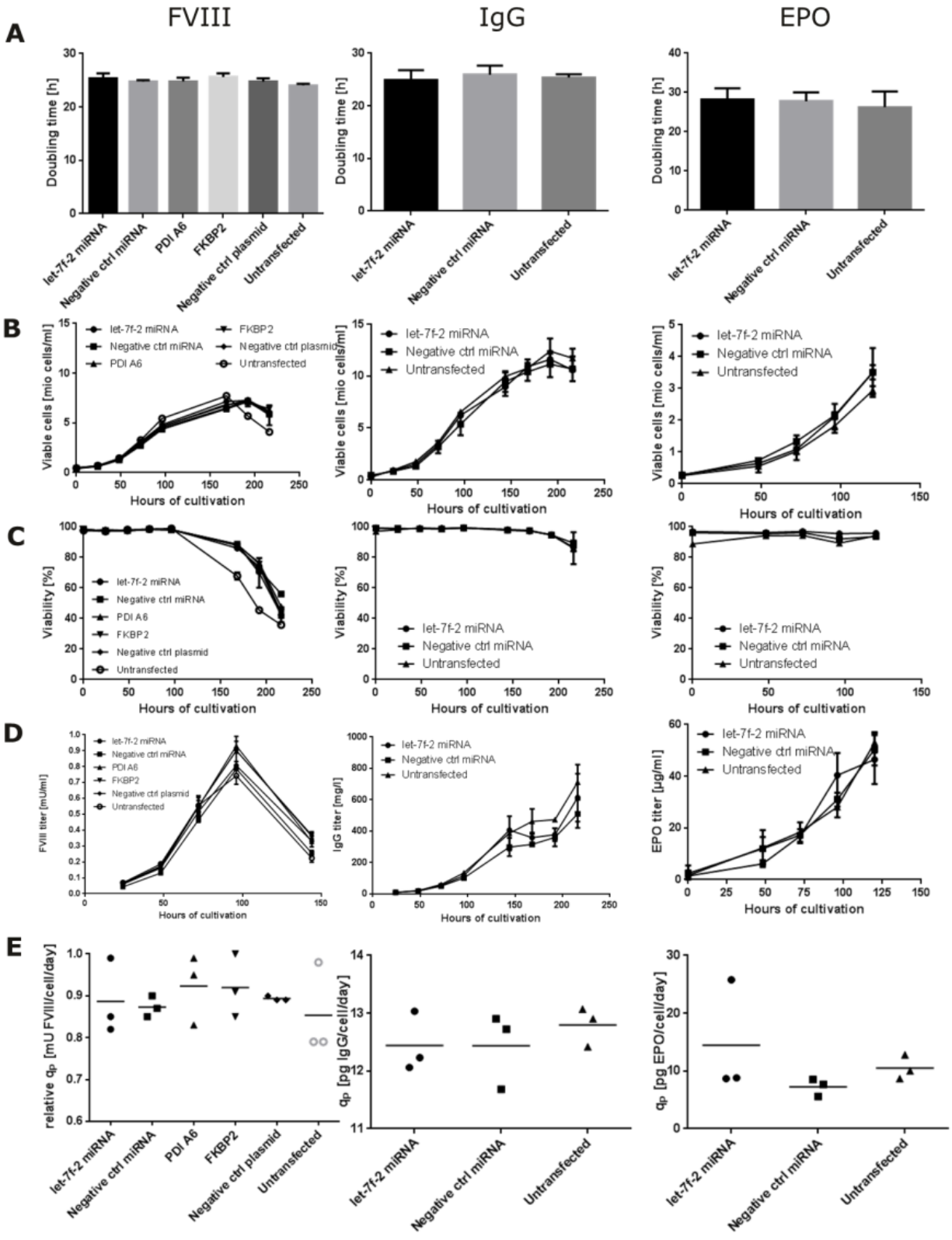
Supplementary Figure 4 subnetwork specific for genes relevant for ER-associated protein degradation overlaid with RNAseq data.
A) Colors indicate log2(foldchange) between IgG producing CHO-K1 cells in exponential growth phase versus non-producing control CHO-K1 cells. B) Colors indicate log2(foldchange) between the FVIII producing cell lines (Clone 1-11) versus control cell lines (Clone 12-13).



Supplementary Figure 5 subnetwork specific for genes relevant for ER-Golgi transport overlaid with RNAseq data.
A) Colors indicate log2(foldchange) between IgG producing CHO-K1 cells in exponential growth phase versus non-producing control CHO-K1 cells. B) Colors indicate log2(foldchange) between the FVIII producing cell lines (Clone 1-11) versus control cell lines (Clone 12-13).

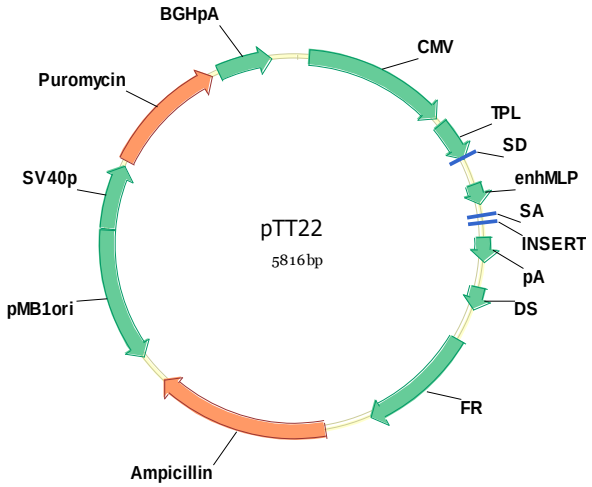


Supplementary Figure 6 Transgene expression across relevant transient transfectants. qRT-PCR expression values relative to GAPDH.



Supplementary Figure 7 Growth characteristics from stable cell pools producing FVIII, IgG or EPO respectively.
A) Doubling time, B) growth curve, C) viability and D) product titer versus hours of cultivation E) specific productivity after 72 hours of growth.

Supplementary Figure 8 Primers used for ligation-free cloning of transgenes into the pTT22 plasmid



CK105	GGGCCCTCTAGAGGGAATTC	rv	pTT22 backbone for cloning
CK106	GTTTCTGCTAGCAAGCTTGC	fw	pTT22 backbone for cloning
CK114	CCCTCTAGAGGGCCC ATGATGGATCTGGAACTGCC	fw	NRF2 into pTT22
CK115	CTTGCTAGCAGAAAC GCGCTAGCTTATCAATTCTT	rv	NRF2 into pTT22
CK116	CCCTCTAGAGGGCCC ATGTCCGTGTCCGAGAGCGC	fw	bach1 into pTT22
CK117	CTTGCTAGCAGAAAC GCGCTAGCTTATCACTCATC	rv	bach1 into pTT22
CK118	CCCTCTAGAGGGCCC ATGTACAAAGTCGGTCGGCG	fw	LMAN1 into pTT22
CK119	CTTGCTAGCAGAAAC GCGCTAGCTTATCAAAAGAAC	rv	LMAN1 into pTT22
CK251	CCCTCTAGAGGGCCC ATGCTGAGTCGTTCACTGCT	fw	PDI into pTT22
CK252	CTTGCTAGCAGAAAC GACAAAAGTCGCTTCGAAGC	rv	PDI into pTT22
CK253	CCCTCTAGAGGGCCC ATGAGGTTTAGTTGCCTGGC	fw	PDIA3 into pTT22
CK254	CTTGCTAGCAGAAAC GACAAAAGTCGCTTCGAAGCG	rv	PDIA3 into pTT22
CK255	CCCTCTAGAGGGCCC ATGACCGATACCTTTGAGCA	fw	PDIA6 into pTT22
CK256	CTTGCTAGCAGAAAC GACAAAAGTCGCTTCGAAGC	rv	PDIA6 into pTT22
CK257	CCCTCTAGAGGGCCC ATGAGGCTGTCTCGTGTCCT	fw	FKBP2 into pTT22
CK258	CTTGCTAGCAGAAAC TTATCACAGTTCGCTTCTAC	rv	FKBP2 into pTT22
CK259	CCCTCTAGAGGGCCC ATGTCCCTGCGTCCTCTGCTG	fw	FKBP11 into pTT22
CK260	CTTGCTAGCAGAAAC GACCGGCTGTACAGGATTTA	rv	FKBP11 into pTT22
CK261	CCCTCTAGAGGGCCC ATGAGCGACTCCGACCTGGG	fw	TGIF2 into pTT22
CK262	CTTGCTAGCAGAAAC TTATCATTTGGCATTCTCAC	rv	TGIF2 into pTT22
CK263	CCCTCTAGAGGGCCC ATGGCTTCACTGTGGCTGCT	fw	Mcfd2 into pTT22
CK264	CTTGCTAGCAGAAAC CTTATCACTGCAGACTTTTA	rv	Mcfd2 into pTT22
CK302	CCCTCTAGAGGGCCC ATGCTCCTTTCCGTGCCGCT	fw	Calr into pTT22
CK303	CTTGCTAGCAGAAAC CTACAGCTCATCCTTGGTCTGGC	rv	Calr into pTT22
CK365	CCCTCTAGAGGGCCC ATGGTGAGCAAGGGCGA	fw	GFP+pC30miRNA forward into pTT22
CK366	CTTGCTAGCAGAAAC CCTCCCCCATCAGCGAA	rv	GFP+pC30miRNA reverse into pTT22

10.2.2 Supplementary tables

The complete tables can be downloaded from http://bit.ly/SupTables

Subsections from each of the tables are listed below in order to provide an overview of the information content of each table

Supplementary Table 1 overview of genes correlating positively with FVIII productivity

Gene	logFC	FPKM	PValue	FDR	Homolog	IPA identifier
Matk	1.03	0.82	2.37E-11	2.51E-10	ENSG00000007264	MATK
LOC103162370	1.06	2.67	0.000208	0.000996	ENSG00000049769	PPP1R3F
LOC103162202	1.27	3.15	7.97E-06	4.74E-05	ENSG00000053438	NNAT
Pkp2	1.86	0.79	8.97E-22	2.05E-20	ENSG00000057294	PKP2
LOC103162774	1.09	17.84	4.13E-17	6.83E-16	ENSG00000058091	CDK14
Tle2	1.40	1.74	1.79E-07	1.30E-06	ENSG00000065717	TLE2
Fam107b	2.15	11.94	2.96E-71	4.84E-69	ENSG00000065809	FAM107B
Derl2	1.45	16.89	2.04E-43	1.35E-41	ENSG00000072849	DERL2
Ebf4	1.54	2.29	1.98E-32	8.04E-31	ENSG00000088881	EBF4
Cbx7	1.36	2.28	1.21E-40	7.19E-39	ENSG00000100307	CBX7

…

Supplementary Table 2 overview of genes correlating negatively with FVIII productivity

Gene	logFC	FPKM	PValue	FDR	Homolog	IPA identifier
Abca6	-3.42	0.19	1.38E-39	7.82E-38	ENSRNOG00000046890	ABCA6
Abcg1	-4.29	0.45	1.47E-26	4.50E-25	ENSMUSG00000024030	ABCG1
Abi2	-1.00	5.18	5.03E-11	5.17E-10	ENSG00000138443	ABI2
Abtb2	-1.31	0.59	8.64E-09	7.27E-08	ENSMUSG00000032724	ABTB2
Acer1	-1.99	0.66	5.83E-10	5.51E-09	ENSRNOG00000047368	ACER1
Acp2	-1.03	22.79	1.07E-39	6.11E-38	ENSRNOG00000013594	ACP2
Acsl1	-1.43	23.35	1.32E-31	5.18E-30	ENSRNOG00000010633	ACSL1
Acsl3	-1.17	21.52	3.60E-43	2.34E-41	ENSRNOG00000014718	ACSL3
Acsl6	-1.21	0.7	1.37E-16	2.18E-15	ENSRNOG00000026745	ACSL6

…

Supplementary Table 3 overview of genes enriched for genes correlating with the FVIII productivity

Ranking	Ingenuity Canonical Pathways	p-value	DE genes	Pathway size	Genes (IPA)
1	Unfolded protein response	1.66E-07	10	42	CALR,EDEM1,TRAF2,DNAJC3,SREBF1,HSP90B1,HSPA5,PPP1R15A,SEL1L,CEBPA
2	Endoplasmic Reticulum Stress Pathway	6.31E-05	5	19	CALR,TRAF2,DNAJC3,HSP90B1,HSPA5
3	NRF2-mediated Oxidative Stress Response	1.23E-04	13	133	CBR1,ACTA2,PPIB,PIK3CB,Gsta1,DNAJC3,GSTM1,GSTM5,DNAJB11,EPHX1,GCLM,DNAJB14,HERPUD1
4	Amyotrophic Lateral Sclerosis Signaling	2.29E-04	9	51	GRIN3B,AKT3,GRIN2D,XIAP,BCL2L1,GRIA4,VEGFB,PIK3CB,BIRC3
5	TNFR2 Signaling	3.24E-04	5	21	TRAF2,TNFAIP3,XIAP,NFKBIE,BIRC3

6	Role of Osteoblasts, Osteoclasts and Chondrocytes in Rheumatoid Arthritis	8.13E-04	13	105	NFATC4,TRAF2,AKT3,DKK3,ITGA3,NFKBIE,PIK3CB,BIRC3,IL18,CSF1,XIAP,IL33,APC2
7	PI3K Signaling in B Lymphocytes	1.58E-03	9	86	NFATC4,AKT3,FCGR2B,C3,PDIA3,BLNK,ATF5,NFKBIE,PIK3CB
8	Agranulocyte Adhesion and Diapedesis	2.34E-03	11	43	IL18,Ccl7,MMP9,ITGA3,Cxcl3,ACTA2,SDC4,CXCL2,MMP17,IL33,Glycam1
9	CD40 Signaling	2.45E-03	6	44	TRAF2,FCER2,TNFAIP3,PTGS2,NFKBIE,PIK3CB
10	Molybdenum Cofactor Biosynthesis	2.57E-03	2	4	MOCS3,NFS1
11	TNFR1 Signaling	3.63E-03	5	39	TRAF2,TNFAIP3,XIAP,NFKBIE,BIRC3
12	Small Cell Lung Cancer Signaling	3.89E-03	6	55	TRAF2,AKT3,PTGS2,BCL2L1,NFKBIE,PIK3CB
13	TWEAK Signaling	5.50E-03	4	28	TRAF2,XIAP,NFKBIE,BIRC3
14	Lymphotoxin β Receptor Signaling	5.50E-03	5	38	TRAF2,AKT3,BCL2L1,PIK3CB,NFKBID
15	Pancreatic Adenocarcinoma Signaling	7.24E-03	7	83	AKT3,NOTCH1,MMP9,PTGS2,BCL2L1,VEGFB,PIK3CB
16	Induction of Apoptosis by HIV1	8.71E-03	5	60	TRAF2,XIAP,BCL2L1,NFKBIE,BIRC3
17	RANK Signaling in Osteoclasts	1.07E-02	6	62	TRAF2,AKT3,XIAP,NFKBIE,PIK3CB,BIRC3
18	Role of Macrophages, Fibroblasts and Endothelial Cells in Rheumatoid Arthritis	1.12E-02	13	155	NFATC4,TRAF2,AKT3,DKK3,NFKBIE,PIK3CB,CEBPA,IL18,CSF1,PDIA3,VEGFB,IL33,APC2
19	Salvage Pathways of Pyrimidine Deoxyribonucleotides	1.15E-02	2	5	CDA,TYMP
20	Granulocyte Adhesion and Diapedesis	1.32E-02	9	38	IL18,Ccl7,MMP9,ITGA3,Cxcl3,SDC4,CXCL2,MMP17,IL33
21	Death Receptor Signaling	1.32E-02	6	67	PARP10,TRAF2,XIAP,ACTA2,NFKBIE,BIRC3
22	VEGF Signaling	1.32E-02	6	92	AKT3,BCL2L1,ACTA2,Actn3,VEGFB,PIK3CB
23	LXR/RXR Activation	1.45E-02	7	43	ABCG4,IL18,C3,SREBF1,MMP9,PTGS2,IL33

24	PPAR Signaling	1.48E-02	6	62	TRAF2,IL18,HSP90B1,PTGS2,NFKBIE,IL33
25	Aldosterone Signaling in Epithelial Cells	1.55E-02	8	106	DNAJC3,HSP90B1,PIP5KL1,HSPA5,DNAJB11,PDIA3,PIK3CB,DNAJB14
26	PI3K/AKT Signaling	1.55E-02	7	98	AKT3,ITGA3,HSP90B1,PTGS2,BCL2L1,NFKBIE,PIK3CB
27	IL-8 Signaling	1.66E-02	9	116	ANGPT2,AKT3,MMP9,GNG3,PTGS2,VASP,BCL2L1,VEGFB,PIK3CB
28	PEDF Signaling	1.70E-02	5	54	AKT3,BCL2L1,NFKBIE,DOCK3,PIK3CB
29	NF-κB Activation by Viruses	1.91E-02	5	49	TRAF2,AKT3,ITGA3,NFKBIE,PIK3CB
30	Neuropathic Pain Signaling In Dorsal Horn Neurons	1.95E-02	6	46	GRIN3B,GRIN2D,PDIA3,GRIA4,PRKACB,PIK3CB
31	Asparagine Biosynthesis I	2.09E-02	1	1	ASNS
32	Alanine Biosynthesis III	2.09E-02	1	1	NFS1
33	Ovarian Cancer Signaling	2.14E-02	7	79	AKT3,MMP9,PTGS2,MLH1,PRKACB,VEGFB,PIK3CB
34	Endometrial Cancer Signaling	2.40E-02	4	44	AKT3,MLH1,PIK3CB,APC2
35	Relaxin Signaling	2.45E-02	7	71	AKT3,MMP9,GNG3,GNAT1,NFKBIE,PRKACB,PIK3CB
36	Glutathione-mediated Detoxification	2.51E-02	3	17	Gsta1,GSTM1,GSTM5
37	Sonic Hedgehog Signaling	2.51E-02	3	22	GLIS2,PTCH1,PRKACB
38	Leukocyte Extravasation Signaling	2.51E-02	9	198	SPN,MMP9,ITGA3,VASP,ACTA2,MMP17,Actn3,PIK3CB,RAPGEF3
39	Integrin Signaling	2.75E-02	9	135	ITGA2B,AKT3,ITGA3,VASP,BCAR3,ACTA2,TSPAN7,Actn3,PIK3CB
40	Acute Phase Response Signaling	2.75E-02	8	92	TRAF2,IL18,AKT3,C3,NFKBIE,C1R,PIK3CB,IL33
41	Colorectal Cancer Metastasis Signaling	2.88E-02	10	141	AKT3,MMP9,GNG3,PTGS2,MLH1,BCL2L1,MMP17,PRKACB,VEGFB,PIK3CB
42	Role of NFAT in Regulation of the Immune Response	2.88E-02	8	92	NFATC4,AKT3,FCGR2B,GNG3,BLNK,GNAT1,NFKBIE,PIK3CB
43	CREB Signaling in Neurons	2.88E-02	8	98	AKT3,GNG3,GRIN2D,PDIA3,GNAT1,GRIA4,PRKACB,PIK3CB

44	Role of Tissue Factor in Cancer	2.95E-02	6	64	AKT3,CYR61,ITGA3,CSF1,BCL2L1,PIK3CB
45	NF-κB Signaling	3.09E-02	8	98	TRAF2,IL18,AKT3,TNFAIP3,NFKBIE,PRKACB,PIK3CB,IL33
46	Ephrin Receptor Signaling	3.16E-02	8	111	GRIN3B,AKT3,EPHB3,ITGA3,GNG3,GRIN2D,GNAT1,VEGFB
47	MIF-mediated Glucocorticoid Regulation	3.16E-02	3	15	PTGS2,NFKBIE,PLA2G4A
48	Glucocorticoid Receptor Signaling	3.24E-02	11	184	TRAF2,NFATC4,AKT3,HSP90B1,PTGS2,HSPA5,BCL2L1,NFKBIE,PRKACB,PIK3CB,CEBPA
49	Inhibition of Angiogenesis by TSP1	3.47E-02	3	20	AKT3,MMP9,THBS1
50	Protein Kinase A Signaling	3.47E-02	14	219	NFATC4,PTCH1,DUSP15,VASP,NFKBIE,PRKACB,ANAPC1,KDELR1,GNG3,PTGS2,PTPRU,PDIA3,PTPDC1,H1f0
51	Calcium Signaling	3.55E-02	8	178	CALR,GRIN3B,NFATC4,GRIN2D,ACTA2,GRIA4,PRKACB,HTR3A
52	IL-6 Signaling	3.63E-02	6	72	TRAF2,IL18,AKT3,NFKBIE,PIK3CB,IL33
53	Bladder Cancer Signaling	3.72E-02	5	43	MMP9,DAPK1,THBS1,MMP17,VEGFB
54	Neuregulin Signaling	3.89E-02	5	67	AKT3,ITGA3,HSP90B1,NRG3,MATK
55	Synaptic Long Term Potentiation	4.07E-02	6	67	GRIN3B,GRIN2D,PDIA3,GRIA4,PRKACB,RAPGEF3
56	P2Y Purigenic Receptor Signaling Pathway	4.07E-02	6	77	ITGA2B,AKT3,GNG3,PDIA3,PRKACB,PIK3CB
57	4-hydroxyproline Degradation I	4.17E-02	1	2	HOGA1
58	Estrogen-Dependent Breast Cancer Signaling	4.17E-02	4	44	AKT3,HSD17B7,PIK3CB,TERT
59	Axonal Guidance Signaling	4.27E-02	15	217	GLIS2,NFATC4,AKT3,EPHB3,MMP9,PTCH1,ITGA3,VASP,PRKACB,PIK3CB,GNG3,PLXNB2,PDIA3,GNAT1,VEGFB
60	Complement System	4.27E-02	3	13	C3,C1QBP,C1R
61	ILK Signaling	4.47E-02	8	113	AKT3,MMP9,PTGS2,FBLIM1,ACTA2,Actn3,VEGFB,PIK3CB
62	April Mediated Signaling	4.57E-02	3	38	TRAF2,NFATC4,NFKBIE
63	Salvage	4.6	5	57	CDA,GRK4,DAPK1,UPRT,GRK6

	Pathways of Pyrimidine Ribonucleotides	8E-02			
64	Molecular Mechanisms of Cancer	4.68E-02	13	243	AKT3,NOTCH1,PTCH1,ITGA3,NFKBIE,PRKACB,SYNGAP1,PIK3CB,RAPGEF3,BIRC3,XIAP,BCL2L1,GNAT1

Supplementary Table 4 genes detected by proteomics

SYMBOL	Complex	CHO GeneID	CHO GeneName	Mouse GeneName	Mouse Uniprot ID	None-IgG_logFC	None-IgG_PValue	None-IgG_FDR	None-FVIII_logFC	None-FVIII_PValue	None-FVIII_FDR
TPST1	O-sulfation	100757649	LOC100757649	Tpst1	O70281	0.95	0.00	0.00	0.49	0.07	0.18
TPST2	O-sulfation	100771893	LOC100771893	Tpst2	O88856	-0.21	0.15	0.25	-0.86	0.00	0.00
14-3-3-BETA/ALPHA	14-3-3	100757122	LOC100757122	Ywhab	Q9CQV8	0.47	0.01	0.03	-0.48	0.04	0.12
14-3-3-EPSILON	14-3-3	100753603	LOC100753603	Ywhae	P62259	-0.26	0.13	0.23	-0.90	0.00	0.00
14-3-3-ETA	14-3-3	100771398	LOC100771398	Ywhah	P68510	0.25	0.21	0.33	-1.06	0.00	0.00

...

Supplementary Table 4 genes detected by proteomics

Relative concentration normalized to Clone 1.1, p-value and fold change

	Clone1.1	Clone1.2	Clone1.3	Clone7.1	Clone7.2	Clone7.3	Clone12.1	Clone12.2	Clone12.3
Sept2	1.000	1.107	1.000	0.939	0.866	0.580	0.816	0.869	1.112	1.109	1.161	1.136
Sept3	1.000	0.874	1.000	0.822	2.245	5.439	1.870	1.948	0.307	0.343	0.331	0.356
Sept5	1.000	0.917	1.000	1.051	0.573	0.734	0.793	0.716	0.517	0.603	0.612	0.568
Sept6	1.000	0.852	1.000	0.684	0.636	0.837	0.764	0.774	0.891	1.002	1.008	1.129
Sept7	1.000	1.094	1.000	0.692	0.733	1.033	0.659	0.710	0.920	0.974	0.995	1.003
Sept9	1.000	0.877	1.000	1.049	0.994	0.780	1.245	1.200	1.632	1.704	1.817	1.754

...

	pval (1vs7)	pval (1vs12)	pval (7vs12)	pval (1+7vs12)	Fc (1vs7)	Fc (1vs12)	Fc (7vs12)	Fc (1+7vs12)
Sept2	3.6E-02	3.7E-02	1.4E-02	4.1E-03	1.292	0.896	0.693	0.794
Sept3	1.1E-01	6.2E-04	5.9E-02	2.3E-02	0.321	2.764	8.603	5.684
Sept5	3.4E-03	3.3E-05	6.3E-02	2.3E-03	1.409	1.725	1.224	1.475
Sept6	1.9E-01	2.3E-01	7.8E-03	2.3E-02	1.174	0.877	0.747	0.812
Sept7	2.3E-01	7.9E-01	1.1E-01	1.4E-01	1.208	0.973	0.805	0.889
Sept9	5.5E-01	9.0E-06	4.8E-03	9.8E-07	0.931	0.568	0.611	0.590

...

10.3 Supplementaries for Chapter 6

The sections below contain detailed descriptions on how to carry out

10.3.1 Getting started – downloading the genome and annotation file

This section describes how to download the current version of the CHO genome to your own server and prepare for using it as reference genome for analyzing e.g. RNA sequencing data.

Download of the genome and annotation file

At Genbank (http://www.ncbi.nlm.nih.gov/genome/annotation_euk/Cricetulus_griseus/101/) there are currently two different genome versions available for *Cricetulus griseus*: The CHO-K1 genome (CriGri_1.0, Assembly GCF_000223135.1), and the Chinese Hamster (C_griseus_v1.0, Assembly GCF_000419365.1). By searching Genbank for the assembly name, all 52710 scaffolds in the genome will appear as hits and can be downloaded into one fasta file[1].

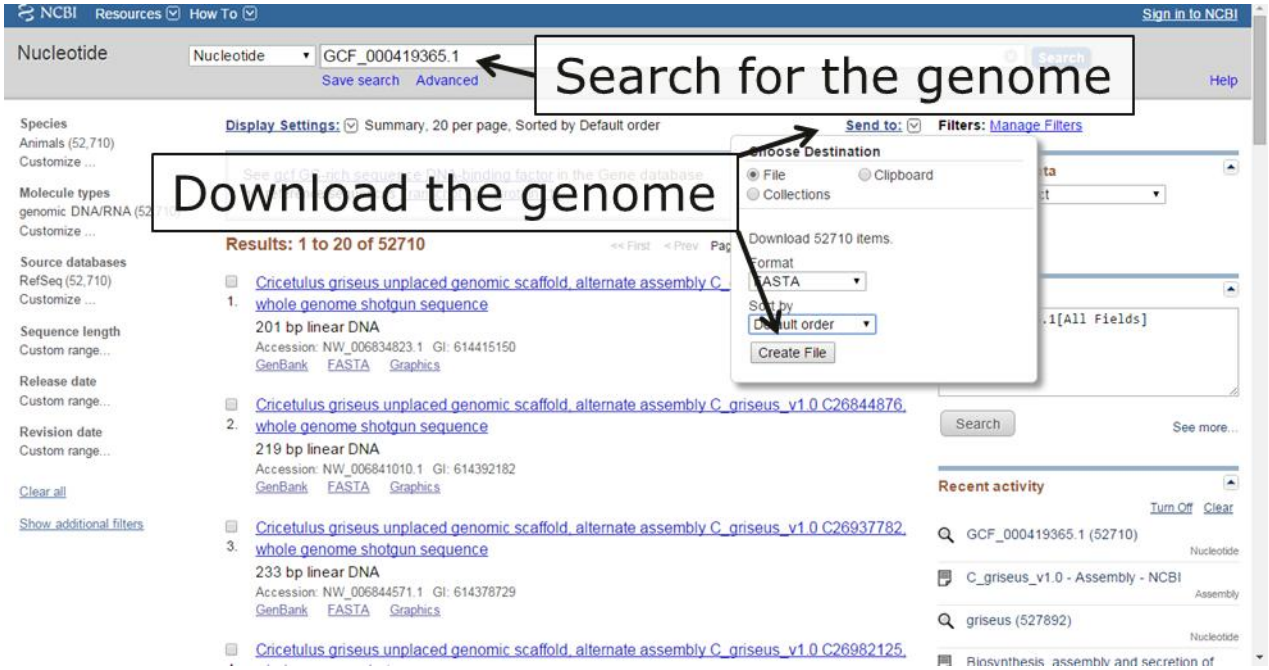


Figure 13 Directions for download of the CHO genome at NCBI Genbank

The annotation file compatible with the genome above, which contains information regarding the location and orientation of all annotated genes in the CHO genome, can be downloaded from the Genbank FTP server by typing the following in your command line environment:

wget ftp://ftp.ncbi.nlm.nih.gov/genomes/Cricetulus_griseus/GFF/alt_C_griseus_v1.0_scaffolds.gff3.gz
gunzip alt_C_griseus_v1.0_scaffolds.gff3.gz

[1] http://www.ncbi.nlm.nih.gov/nuccore/?term=GCF_000419365.1

After unzipping the file it can be opened

head -10 alt_C_griseus_v1.0_scaffolds.gff3

##gff-version 3
#!gff-spec-version 1.20
#!processor NCBI annotwriter
#!genome-build C_griseus_v1.0
#!genome-build-accession NCBI_Assembly:GCF_000419365.1
#!annotation-date 2 May 2014
#!annotation-source NCBI Cricetulus griseus Annotation Release 101
##sequence-region NW_006834731.1 1 201
##species http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=10029
NW_006834731.1 RefSeq region 1 201 . + . ID=id0;Dbxref=taxon:10029;Name=Unknown;chromosome=Unknown;collection-date=Feb-2011;country=China;gbkey=Src;genome=genomic;mol_type=genomic DNA

By searching for e.g. Dhfr one will find the table below.

grep "Dhfr" alt_C_griseus_v1.0_scaffolds.gff3

NW_006884452.1	BestRefSeq	gene	3849042	3872852	.	+	.	ID=gene24052;...
NW_006884452.1	BestRefSeq	mRNA	3849042	3872852	.	+	.	ID=rna29637;...
NW_006884452.1	BestRefSeq	exon	3849042	3849131	.	+	.	ID=id344549;...
NW_006884452.1	BestRefSeq	exon	3849436	3849485	.	+	.	ID=id344550;...
NW_006884452.1	BestRefSeq	exon	3851818	3851923	.	+	.	ID=id344551;...
NW_006884452.1	BestRefSeq	exon	3859552	3859678	.	+	.	ID=id344552;...
NW_006884452.1	BestRefSeq	exon	3862431	3862546	.	+	.	ID=id344553;...
NW_006884452.1	BestRefSeq	exon	3872015	3872852	.	+	.	ID=id344554;...
NW_006884452.1	BestRefSeq	CDS	3849046	3849131	.	+	0	ID=cds26512;...
NW_006884452.1	BestRefSeq	CDS	3849436	3849485	.	+	1	ID=cds26512;...
NW_006884452.1	BestRefSeq	CDS	3851818	3851923	.	+	2	ID=cds26512;...
NW_006884452.1	BestRefSeq	CDS	3859552	3859678	.	+	1	ID=cds26512;...
NW_006884452.1	BestRefSeq	CDS	3862431	3862546	.	+	0	ID=cds26512;...
NW_006884452.1	BestRefSeq	CDS	3872015	3872093	.	+	1	ID=cds26512;...

Each type of annotation (in this case gene, mRNA, exon and CDS) contain different pieces of information regarding the gene, which are listed in column 9 as e.g. for the exon

ID=id344550;Parent=rna29637;Dbxref=GeneID:100689028,Genbank:NM_001244016.1;Note=The RefSeq transcript has 11 substitutions compared to this genomic sequence;exception=annotated by transcript or proteomic data;gbkey=mRNA;gene=Dhfr;product=dihydrofolate reductase;transcript_id=NM_001244016.1

Thus, from the annotation file information regarding: the sequencing scaffold, position on scaffold and orientation, GenbankID for transcript and protein, CHOgenome.org gene name (Dhfr) and long gene name can all be extracted. For more information regarding the gff3 standard see[2]

[2] http://www.sequenceontology.org/gff3.shtml

Correcting headers

Following genome download your fasta file will contain headers for each sequence

```
>gi|614415150|ref|NW_006834823.1| Cricetulus griseus unplaced genomic scaffold, alternate assembly
C_griseus_v1.0 C26702588, whole genome shotgun sequence
AGGTT…
```

In order to get names that are compatible with the annotation file the header will be shortened.

```
sed "s/^>.*|ref|/>/g" sequence.fasta | cut -f 1,1 -d'|' > Cgriseus.fasta
```

The header will now look like

```
>NW_006834823.1
AGGTT…
```

Addition of transgene to the annotation file and genome

In case you are working with data from a transfected cell line the transgene sequence can be added to the genome and the positions of the genes to the annotation file. In the example below the sequence from human coagulation factor VIII and the selection marker dihydrofolate reductase (DHFR) are added. A new sequencing scaffold called P_Plasmid is made based on the known sequence of the plasmid used for transfection.

```
>P_Plasmid
catgacattaacctataaaaataggcgtatcacgaggccctttcgtcttcctgatcagaa…
```

This file was then added to the end of the genome file

```
P_Plasmid.txt >> Cgriseus.fasta
```

In order to detect expression level of this scaffold two exon sequences are added to the annotation file for each of the two ORF's following the same nomenclature as the Dhfr example given above. HTSeq, which can be used to deduce the expression level scan the genome based on ranges in the annotation file listed as exons[3]. For this reason only these two lines are added.

```
P_Plasmid RefSeq exon 6581 7144 . + 0 ID=id251660;Parent=rna251660;Dbxref=GeneID:251660;gbkey=mRNA;gene=
dhfr_p;transcript_id= rna251661
P_Plasmid RefSeq exon 102 1000 . + 0 ID=id251662;Parent=rna251662;Dbxref=GeneID:251662;gbkey=mRNA;gene=
hFVIII_p;transcript_id= rna251663
```

RNA and gene ID's were made up and checked by grep whether they were already in use. The final line of the annotation file contain 3 hashtags which indicating that the file is done. The two lines above thus needs to be added before that

```
grep -v "###" alt_C_griseus_v1.0_scaffolds.gff3 > CHOannotation.gff3
cat new_annotation.txt >> CHOannotation.gff3
echo "###" >> CHOannotation.gff3
```

[3] http://www-huber.embl.de/HTSeq/

Extraction of information from the annotation file

In order to gain an overview on the content of the annotation files a list of genes can be extracted.

```
cat CHOannotation.gff3 | sed 's/\t/,/g' |  awk -F, '$3=="exon" {gsub("," " ");print $0}' | grep "gene="  > subset
cat subset | sed 's/^.*gene=//g' | cut -f 1,1 -d';' > genes
cat subset | sed 's/^.*Genbank://g' | cut -f 1,1 -d';' | cut -f 1,1 -d',' | gawk '{print $1}' > Genbk
cat subset | sed 's/^.*Parent=//g' | cut -f 1,1 -d';' | cut -f 1,1 -d',' | gawk '{print $1}' > RNAid
paste genes Genbk RNAid | sort | uniq > Genes

cat CHOannotation.gff3 | sed 's/\t/,/g' |  awk -F, '$3=="CDS" {gsub("," " ");print $0}' | grep "gene="  > subset
cat subset | sed 's/^.*gene=//g' | cut -f 1,1 -d';' > genes
cat subset | sed 's/^.*Genbank://g' | cut -f 1,1 -d';' | cut -f 1,1 -d',' | gawk '{print $1}' > Genbk
cat subset | sed 's/^.*Parent=//g' | cut -f 1,1 -d';' | cut -f 1,1 -d',' | gawk '{print $1}' > RNAid
paste genes Genbk RNAid | sort | uniq > Proteins
```

The two files thus contain information about all transcripts and all proteins

```
paste Genes Proteins | head -1
A1cf    XM_007627897.1  rna20866        A1cf    XP_007626087.1  rna20866
```

By merging the two files above a list of all the genes in the *C. griseus* genome can be made containing both the Genbank information for the coding region and the translated protein made from it. The information of the scaffold they are placed on is furthermore included

```
cat Genes | gawk '{print $3}' > rna_genes
cat Proteins | gawk '{print $3}' > rna_protein
sdiff rna_protein rna_genes | grep -v ">" | gawk '{print $1}' > uniqueRNA

wc -l uniqueRNA > length
length=`awk '{print $1 }' < length`
for ((  i = 1 ;  i <= $length ; i++  ))
do
cat uniqueRNA | head -$i | tail -1 > gene
gene=`awk '{print $1 }' < gene`
grep -w "$gene" Genes > one
grep -w "$gene" Proteins > two
grep -w "$gene" ./CHOannotation.gff3| head -1 | gawk '{print $1}' > scaffold
paste one two scaffold | gawk '{print $1, $2, $5, $3, $7}' | sed 's/ /\t/g' >> genenames
done
printf "\tASDF\n" | awk -F'\t' '$3' genenames > listofgenes.txt
```

The list will look as the one below containing all splice variants for protein coding transcripts in the *C. griseus* genome

```
A1cf    XM_007627897.1  XP_007626087.1  rna20866        NW_006880426.1
A2ml1   XM_007611503.1  XP_007609693.1  rna3450 NW_006870833.1
A3galt2 XM_007614932.1  XP_007613122.1  rna7152 NW_006873147.1
A4galt  XM_007629783.1  XP_007627973.1  rna22847        NW_006881179.1
A4galt  XM_007629784.1  XP_007627974.1  rna22846        NW_006881179.1
```

Downloading data from the SRA

The raw sequencing reads from the currently sequenced CHO genomes such as CHO DG44 and CHO DXB11 can be downloaded from the Sequence Read Archive hosted by Genbank[4]. The reads, which are currently publically available, are listed in the table below

[4] http://www.ncbi.nlm.nih.gov/sra

Table 1 list of raw sequence from CHO genomes publicly available for download at the time of writing

CHO cell name	SRA accession number	Original publication	CHO cell name	SRA accession number	Original publication
CHO DXB11	SRR1561441	Kaas et al, 2015	CHO-K1 Protein free	SRR803173	Lewis et al, 2013
CHO DXB11	SRR1561442	Kaas et al, 2015	CHO-K1 Protein free	SRR803174	Lewis et al, 2013
CHO DXB11	SRR1561427	Kaas et al, 2015	CHO-K1 Protein free	SRR803175	Lewis et al, 2013
CHO DXB11	SRR1561428	Kaas et al, 2015	CHO-K1 ECACC	SRR803176	Lewis et al, 2013
CHO-K1 ATCC	SRR329939	Xu et al, 2011	CHO-K1 ECACC	SRR803177	Lewis et al, 2013
CHO-K1 ATCC	SRR329940	Xu et al, 2011	CHO-K1 ECACC	SRR803178	Lewis et al, 2013
CHO-K1 ATCC	SRR329941	Xu et al, 2011	CHO-K1 Serum free	SRR803179	Lewis et al, 2013
CHO-K1 ATCC	SRR329942	Xu et al, 2011	CHO-K1 Serum free	SRR803180	Lewis et al, 2013
CHO-K1 ATCC	SRR329943	Xu et al, 2011	CHO-K1 Serum free	SRR803181	Lewis et al, 2013
CHO-K1 ATCC	SRR329944	Xu et al, 2011	CHO-S	SRR803182	Lewis et al, 2013
CHO-K1 ATCC	SRR329945	Xu et al, 2011	CHO-S	SRR803183	Lewis et al, 2013
CHO-K1 ATCC	SRR329946	Xu et al, 2011	CHO DG44	SRR803184	Lewis et al, 2013
CHO-K1 ATCC	SRR329947	Xu et al, 2011	CHO DG44	SRR803185	Lewis et al, 2013
CHO-K1 ATCC	SRR329948	Xu et al, 2011	C0101	SRR801491	Lewis et al, 2013
CHO-K1 ATCC	SRR329949	Xu et al, 2011	C0101	SRR801492	Lewis et al, 2013
CHO-K1 ATCC	SRR329950	Xu et al, 2011	C0101	SRR801493	Lewis et al, 2013
CHO-K1 ATCC	SRR329951	Xu et al, 2011	C0101	SRR801494	Lewis et al, 2013
CHO-K1 ATCC	SRR329952	Xu et al, 2011	C0101	SRR801495	Lewis et al, 2013
CHO-K1 ATCC	SRR329953	Xu et al, 2011	C0101	SRR801496	Lewis et al, 2013
CHO-K1 ATCC	SRR329954	Xu et al, 2011			

Download of e.g. C0101, run1 = SRR801491 was downloaded from the ftp using the name of the file as shown underlined in the link below:

wget ftp://ftp-trace.ncbi.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR801/SRR801491/SRR801491.sra

In order to convert the SRA file into paired reads, the SRA file toolkit was downloaded from NCBI[5] and run as follows (remember full name path for the SRA file as it otherwise will not work):

~/sratoolkit.2.3.2-5-centos_linux64/bin/fastq-dump --split-3 /novo/users/csrk/Cgenomes/SRAdumps/SRR801491.sra

Asses and change the quality score of the reads

By extracting the first 1000 reads from each file Fastqc[6] can be run quickly to investigate which standard for the quality score was used for each sample.

```
head -400 K1ATCC_1_1.fastq >  test_1.fastq
fastqc test_1.fastq
```

From the output fastqc_report.html it can be seen that the file is in "Illumina 1.9"-encoding. This is the case for all reads from the CHO-K1 ATCC and CHO DXB11 genome whereas the rest from the

[5] http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=toolkit_doc&f=std
[6] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

table above is found to be in "Illumina 1.5"-encoding. If this later is shown to cause problems with downstream settings it can be changed to the same score using a small script found online[7]:

```
#!/usr/bin/perl
use strict;
use warnings;
my $count = 0;
while (<>) {
    chomp;
    if ($count++ % 4 == 3) { tr/\x40-\xff\x00-\x3f/\x21-\xe0\x21/; }
    print "$_\n";
}
```

```
perl sanger.pl 1rawread_1.fastq > 2rawread_1.fastq
```

10.3.2 Extraction and analysis of genomic copy numbers from CHO genomes

Summary: This tutorial describes how the copy number was extracted for each gene in each of the sequenced CHO cell lines. I kindly ask you cite: Kaas et al 2015, *Sequencing the CHO DXB11 genome reveals regional variations in genomic stability and haploidy* in case you will publish data based on this workflow. For the tutorial below it is assumed that the raw reads from all available CHO genomes have been downloaded as described above.

Pre-alignment trimming of reads

Trim Galore is a usefull tool, which are able to trim the paired reads, while keeping the files synchronized so the pairs are broken[8]. It can be downloaded from here[9] and the required packagages cutadapt from here[10] and fastqc from here[11]. In the code below all low quality bps are trimmed from the ends until bps with a quality >30 are found, but all reads that after trimming are shorter than 40bp are removed from the dataset. See full decription of options here[12]

```
fastqc DXB11_1.fastq
fastqc DXB11_2.fastq
mkdir ./trimmed_reads
~/software/Trimgalore/trim_galore --paired -q 30 --length 40 --fastqc --path_to_cutadapt /novo/appl/ngs/cutadapt-1.2.1/bin/cutadapt --stringency 5 -o ./trimmed_reads DXB11_1.fastq DXB11_2.fastq
```

[7] http://seqanswers.com/forums/showthread.php?t=5210
[8] http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
[9] http://www.bioinformatics.babraham.ac.uk/projects/download.html#trim_galore
[10] https://pypi.python.org/pypi/cutadapt/
[11] http://www.bioinformatics.babraham.ac.uk/projects/download.html#fastqc
[12] http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/trim_galore_User_Guide_v0.4.0.pdf

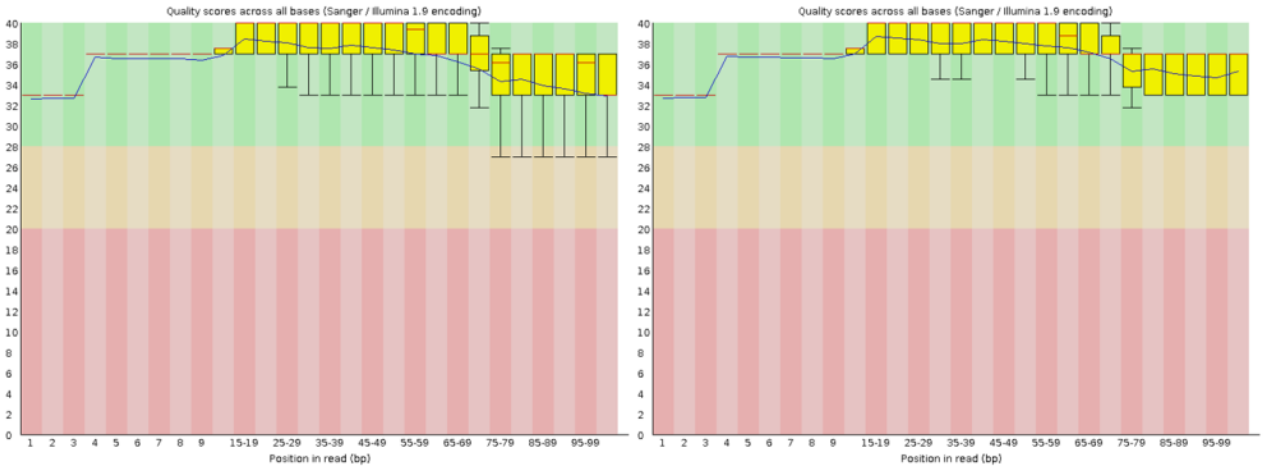


Figure 14 quality score across the reads before (left) and after (right) trimming.

Reads were aligned to the *C. griseus* genome

The trimmed reads are subsequently aligned to the *C. griseus* genome using BWA[13] and subsequently converted into a bam-file from a sam-file. First, the genome has to be indexed by BWA before alignment.

bwa index Cgriseus.fa

After the genome has been indexed the reads are alinged to the genome

bwa aln Cgriseus.fasta 1BWA_1.fastq > 1BWA_1.sai
bwa aln Cgriseus.fasta 1BWA_2.fastq > 1BWA_2.sai
bwa sampe -r "@RG\tID:libA\tSM:sample_EACC\tPL:ILLUMINA" Cgriseus.fasta 1BWA_1.sai 1BWA_2.sai 1BWA_1.fastq 1BWA_2.fastq | samtools view -Sb - > 2alignment_L.bam

All reads with a minimum mapping quality less than 30 were removed and the file was sorted and indexed

samtools view -u -q 30 2alignment_L.bam | samtools sort - 2alignment_L.sort
samtools index 2alignment_L.sort.bam

GATK was used in order to remap all the reads in region containing indels

The RealignerTargetCreater algorithm[14] was used to realign reads in problematic regions of the genome.

java -Xmx2g -jar GenomeAnalysisTK.jar -T RealignerTargetCreator -R Cgriseus.fasta -I 2alignment_L.sort.bam -o 3_L.intervals

java -Xmx2g -jar GenomeAnalysisTK.jar -T IndelRealigner -targetIntervals 3_L.intervals -R Cgriseus.fasta -I 2alignment_L.sort.bam -o 3realigned_L.bam

[13] http://bio-bwa.sourceforge.net/

[14] https://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_gatk_tools_walkers_indels_RealignerTargetCreator.php

Alignments run in parallel were merged into one bam-file

In the steps above the data from the same sample was split into multiple files and now the files are merged from several bam files into one single bam file. This is done using samtools[15]

```
samtools merge Cgris_prePicard.bam 3realigned_L.bam 3realigned_L.bam /H3/3realigned_L.bam
/H4/3realigned_L.bam /H5/3realigned_L.bam
```

PCR duplicates were removed using Picard

PCR remnants are removed using Picard[16]

```
java -Xmx2g -jar MarkDuplicates.jar INPUT=Cgris_prePicard.bam OUTPUT=Cgris_postPicard.bam
ASSUME_SORTED=FALSE METRICS_FILE=/dev/null VALIDATION_STRINGENCY=SILENT REMOVE_DUPLICATES=true
```

Calculation of the sequencing depth in genome

The depth at every position in the genome was calculated using genomeCoverageBed from bedtools[17].

```
genomeCoverageBed -d -ibam Cgris_postPicard.bam > Cgris.genomecoverage
```

In order to access the given position easily the coverage file need to be split into files each containing each contig. This is done by first merging the three columns into one and subsequently using awk to create files for each contig

```
sed 's/\t/,/g' Cgris.genomecoverage > Cgris.comma.genomecoverage
awk -F, '{print > $1.txt}' Cgris.comma.genomecoverage
```

The files are restored into having three columns again containing 1: the name of the contig, 2: the position on the contig, 3: the depth at that given position

```
for file in *.1
do
sed 's/,/\t/g' $file > $file.txt
rm $file
done
```

Depths are extracted at genomic coding regions

The listofgenes.txt file made in the section “Extraction of information from the annotation file” is used gene by gene to extract the positions in the genome that the coding region of given gene is situated (the file pos) and the depth found for each bp in these intervals are merged into one file (seq), which subsequently is sorted and the median is extracted.

```
echo '#! /usr/bin/env Rscript' > scriptR
echo 'd<-scan("stdin", quiet=TRUE)' >> scriptR
echo 'cat((median(d)/(33/2)), min(d), max(d), median(d), mean(d), sd(d), sd(d)/mean(d), sep="\t")' >> scriptR
chmod a+x scriptR

for (( k = 1 ; k <= 29983 ; k++ ))
```

---

[15] http://samtools.sourceforge.net/samtools.shtml
[16] http://broadinstitute.github.io/picard/
[17] http://bedtools.readthedocs.org/en/latest/content/tools/genomecov.html

```
do
cat ../listofgenes2 | head -$k | tail -1 > geneinfo
geneProtein=`awk '{print $3 }' < geneinfo`
scaffold=`awk '{print $5 }' < geneinfo`
grep -w "$geneProtein" ../listofgenes2 > data
grep -w "$geneProtein" ~/annotationK1.gff3 | sed 's/\t/,/g' | awk -F, '$3=="CDS" {gsub(","," ");print $0}' | gawk
'{print $1, $4-1, $5}' | sed 's/ /\t/g' | sort -k 2 -n > pos
wc -l pos > length
length=`awk '{print $1 }' < length`
cat pos | head -1 | gawk ' {print $2}' > start
cat pos | tail -1 | gawk ' {print $3}' > slut
echo "0" > seq
for (( j = 1 ; j <= $length ; j++ ))
do
cat pos | head -$j | tail -1 | gawk '{print $3}' > slutnr
cat pos | head -$j | tail -1 | awk '{a=$3-$2+1;print a}' > lengthCDS
slut=`awk '{print $1 }' < slutnr`
len=`awk '{print $1 }' < lengthCDS`
cat $scaffold* | head -$slut | tail -$len | gawk '{print $3}' >> seq
done
./test1 < seq > median
wc -l seq | gawk '{print $1}' > seqlength
paste median data start slut seqlength >> list.txt
done
```

Evaluation of data

The list.txt files from each cell line were merged into a matrix (in excel) and histograms can be made for each cell lines showing the distribution of genes across the measured depths

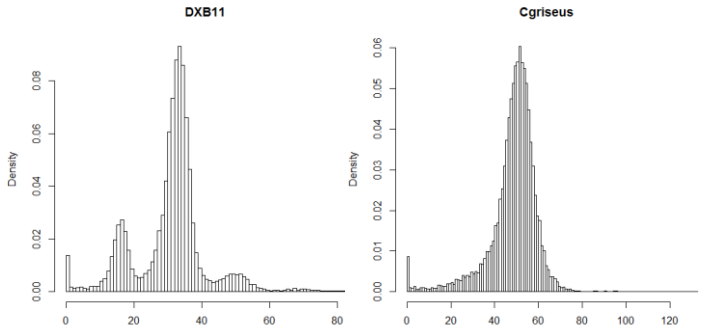


Figure 15 Histogram showing the number of genes found to have copy number of 1,2 and 3.
To the left the data from CHO DXB11 is shown and to the right the same data is represented for wild type *C. griseus* clearly showing a single peak indicating a whole diploid genome

start	end	genes	mean(sd)
0.4	0.75	193	0.795
0.75	1.25	1987	0.335
1.25	1.75	1269	0.377
1.75	2.25	8592	0.227
2.25	2.75	1030	0.211
2.75	3.25	687	0.206

Different regions from DXB11 were extracted in order to see if the standard deviation for each gene were different. It is seen that the genes from 0 depth to 0.75 seem to have pretty high standard

deviation across the CDS but the ones which have 1.5 in depth does not seem to be more noisy than genes with 2.0.

Merging into one matrix

The list.txt file generated from each analyzed cell line was renamed by their cell line name and merged into one matrix that was subsequently sorted based on the 20661 genes name and not based on the 29978 transcripts which my list originally centered on.

```
paste Cgris.txt C0101.txt CHO_PF.txt CHO_SF.txt CHO_S.txt DG44.txt DXB11.txt F435.txt K1ATCC.txt K1ECACC.txt |
gawk '{print $2, $1, $10, $19, $28, $37, $46, $55, $64, $73, $82}' | sed 's/ /\t/g' > depthmatrix
cat depthmatrix | gawk '{print $1}' | sort | uniq > genes
wc -l genes > length
length=`awk '{print $1 }' < length`
for ((  i = 1 ;  i <= $length ; i++  ))
do
cat genes | head -$i | tail -1 > gene
gene=`awk '{print $1 }' < gene`
grep -w "$gene" depthmatrix | sort -k 9 -n | tail -1 >> depthmatrix2
done

echo "Gene Cgriseus C0101 CHO_PF CHO_SF CHO_S DG44 DXB11 F435 K1ATCC K1ECACC" | sed 's/ /\t/g' >
depthmatrix.txt
cat depthmatrix2  >> depthmatrix.txt
```

The matrix is imported into R and normalized by half the median depth

```
col=as.matrix(read.table("depthmatrix.txt",header = TRUE, row.names=1))
```

	Cgriseus	C0101	CHO_PF	CHO_SF	CHO_S	DG44	DXB11	F435	K1ATCC	K1ECACC
A1cf	27	38	14	9	12	14	32.0	18	49	10
A2ml1	27	16	9	7	6	7	30.0	16	37	6
A3galt2	26	17	9	9	9	9	29.5	17	42	7
A4galt	33	21	13	11	11	12	39.0	22	53	10
A4gnt	28	18	12	9	5	7	33.0	19	50	7
Aaas	28	24	13	10	11	21	35.0	22	54	9

The list is extracted into a matrix and normalized based on half of the median for each cell line (thus leading to the mean peak having a normalized depth of 2)

```
coll=col;m2=m3=1;for(i in 1:ncol(col))
{m2[i]=median(as.vector(col[,i]), na.rm=TRUE)
coll[,i]=col[,i]/(m2[i]/2)}
```

```
> head(round(coll, digits=1))
        Cgriseus C0101 CHO_PF CHO_SF CHO_S DG44 DXB11 F435 K1ATCC K1ECACC
A1cf         2.0   3.3    2.3    2.0   2.7  3.1   1.9  2.1    2.0     2.5
A2ml1        2.0   1.4    1.5    1.6   1.3  1.6   1.8  1.9    1.5     1.5
A3galt2      1.9   1.5    1.5    2.0   2.0  2.0   1.8  2.0    1.8     1.8
A4galt       2.4   1.8    2.2    2.4   2.4  2.7   2.4  2.6    2.2     2.5
A4gnt        2.1   1.6    2.0    2.0   1.1  1.6   2.0  2.2    2.1     1.8
Aaas         2.1   2.1    2.2    2.2   2.4  4.7   2.1  2.6    2.2     2.2
```

CHO-K1 ATCC versus CHO DXB11

It is now possible to plot e.g. CHO-K1 ATCC versus CHO DXB11. col="#00000010" is used as it has transparency allowing for dense spots on the plot to be easily seen whereas outliers does not dominate the picture.

```
plot(coll[,9],coll[,7],ylim=c(0,5),xlim=c(0,5),pch=20,col="#00000010",ylab="CHO DXB11",xlab="CHO K1 ATCC")
points(coll["Dhfr",9],coll["Dhfr",7],pch=3,col="red") #DHFR
points(coll["Gapdh",9],coll["Gapdh",7],pch=3,col="blue") #Gapdh
```

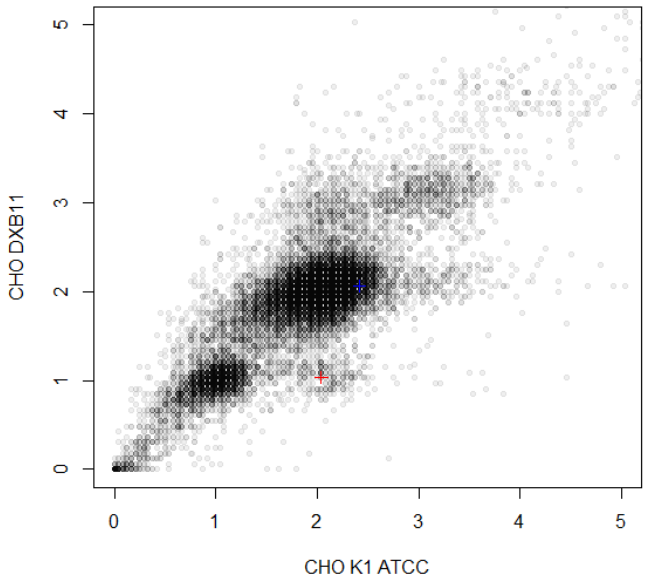


Figure 16 copy number distribution in CHO-K1 ATCC versus CHO DXB11

Construction of a phylogenetic tree

In order to construct a phylogenetic tree a distance matrix is calculated by extracting the number of genes that has a normalized copy number that differ by more than 0.95. A heatmap is made representing the same data as a phyloenetic tree (without compromising on the distances due to the contraints of a 2D drawing)

```
distance=distance2=matrix(ncol=ncol(col),nrow=ncol(col),0);for(i in 1:ncol(col))
{for(j in 1:ncol(coll))
  {col3=coll[which(abs(coll[,j]-coll[,i])>0.95),]
   distance[i,j]=dim(col3)[1]/dim(coll)[1]
   distance2[i,j]=dim(col3)[1]
  }}
colnames(distance)=colnames(distance2)=colnames(coll);rownames(distance)=rownames(distance2)=colnames(coll)
distance;distance2

Box <- function(x, y, color)
{rect(x, y, x+1, y+1, density = NULL, col = color, border = NULL, lty = par("lty"), lwd = par("lwd"))}
colfunc <- colorRampPalette(c("green", "white"))
colfunc2 <- colorRampPalette(c("white", "red"))
colors=c(colfunc2(50),colfunc(50))
plot(100,100, xlim=c(0,12),ylim=c(0,12),xaxt='n',yaxt='n',ylab="",xlab="")
for(j in 1:ncol(coll)){for(i in 1:ncol(coll)) {Box(j,i,colors[distance[j,i]*100])}}
for(i in 1:ncol(col)) {text(i+0.5,1,colnames(col)[i],cex=0.4,pos=1)}
```

```
for(i in 1:ncol(col)) {text(0.5,i+1,colnames(col)[i],cex=0.4,pos=1)}
```

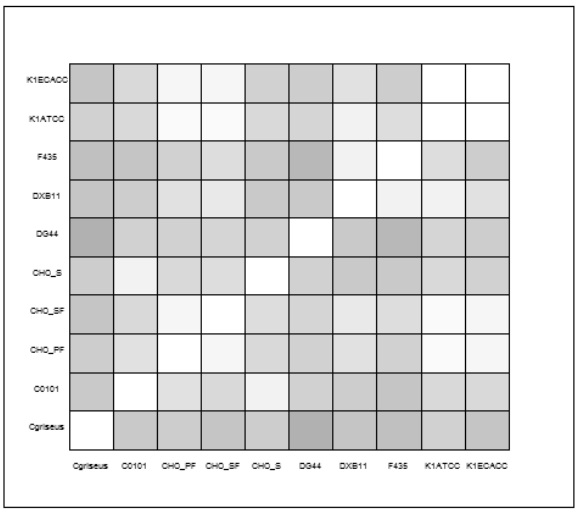


Figure 17 heatmap showing the distance between sequenced CHO genomes

In order to draw a phylogenetic tree the packagages below are loaded, a function is made able to output a distance matrix based on any matrix, 100 phylogenetic trees are drawn using the bootstrap algorithm (boot.phylo) and the concensus tree is used to represent the distance matrix. The tree is rooted in C griseus.

```
.libPaths(c("C:\\R\\Rpackages", .libPaths()))
install.packages("ape", repos="http://cran.r-project.org")
install.packages("phangorn", repos="http://cran.r-project.org")
install.packages("phyclust", repos="http://cran.r-project.org")
install.packages("MASS", repos="http://cran.r-project.org")
library(ape)
library(phangorn)
library(phyclust)

estimate_tr <- function(m)
{n=t(m)
distance=distance2=matrix(ncol=ncol(n),nrow=ncol(n),0)
for(i in 1:ncol(n))
  {for(j in 1:ncol(n))
   {col3=n[which(abs(n[,j]-n[,i])>0.95),]
   distance[i,j]=dim(col3)[1]/dim(n)[1]}
  }
colnames(distance)=colnames(distance2)=rownames(distance)=rownames(distance2)=colnames(coll)
return(nj(as.dist(distance)))
}
point_est <- estimate_tr(t(coll))
treerooted_no = root(point_est, "Cgriseus", resolve.root = TRUE, interactive = FALSE)
plot(treerooted_no, type="u", main="NJ without bootstrap")
windows()
bs <- boot.phylo(point_est, t(coll), estimate_tr, trees=TRUE, B=100,block = 1)
con <- consensus(bs$trees[[1]], p=0.5)
treerooted = root(con, "Cgriseus", resolve.root = TRUE, interactive = FALSE)
plot(treerooted , type="u", main="NJ with bootstrap")
```

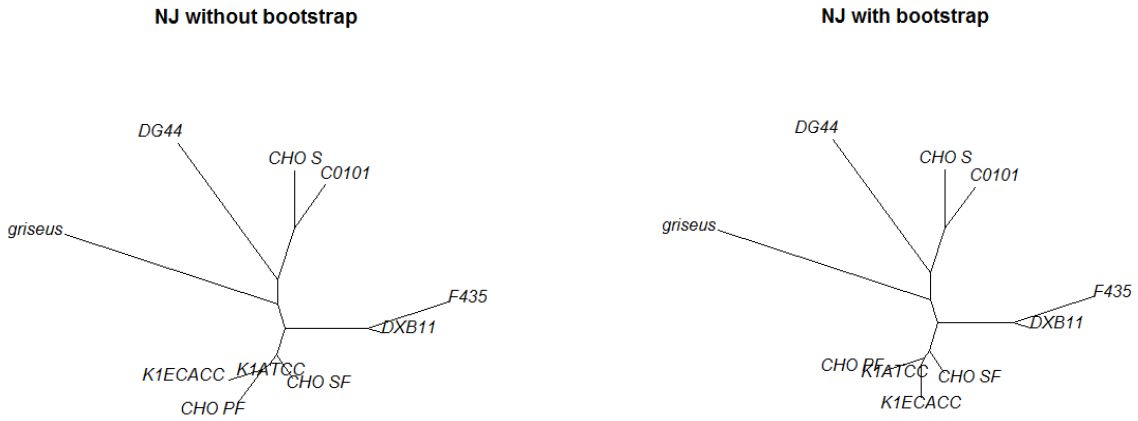


Figure 18 Phylogenetic tree showing the relationship between sequenced genomes based on copy number variations

CNVs based on chromosomal loci

It would be interessting to investigate the genomic location of the genes across the chromosomes. The *C. griseus* genome had been sequenced sorted by chromosomes[18]. This genome was downlaoded from NCBI[19] and headers were extracted using grep from the fasta file.

```
grep "^>" CHOchromosomes.fa > scaffolds.txt
head -2 scaffolds.txt
>gi|537274396|gb|KE662775.1| Cricetulus griseus strain 17A/GY chromosome 1 unlocalized genomic scaffold
chr1_scaffold_0, whole genome shotgun sequence
>gi|537274281|gb|KE662776.1| Cricetulus griseus strain 17A/GY chromosome 1 unlocalized genomic scaffold
chr1_scaffold_1, whole genome shotgun sequence
```

It was seen that the chromosome and scaffold name is listed in the header and this information was filtered out and the based on the paper "unplaced scaffold" were attributed to chromosome 9/10 (here just called 9). Chromosome X were called 0

```
cat scaffolds.txt | sed 's/^.*gb|//g' | cut -f 1,1 -d'|' > scaffolds
cat scaffolds.txt | sed 's/^.*chromosome //g' | sed 's/^.*Crice//g' | cut -f 1,1 -d'u' | sed 's/t/9/g' | sed 's/X/0/g' >
chromosome
paste scaffolds chromosome > ~/Cgenomes/borth/chromosomeoverview
```

In order to get the chromosome number for each gene in the *C. griseus* genome, the 29978 transcripts listed in the *C. griseus* genome needs to be blasted against the chromosome sorted genome. But first the sequences for those transcripts needs to be downloaded from Genbank. This was done by extracting the names of the genes from the *C. griseus* annotation file (see the section

[18] http://www.nature.com/nbt/journal/v31/n8/full/nbt.2645.html
[19] http://www.ncbi.nlm.nih.gov/nuccore/?term=APMK01000000

"Extraction of information from the annotation file") and using this list for download based on a script found online[20].

The file Genbankdownload.txt used below:

```
perl -e XXuse
LWP::Simple;getstore(YYhttp://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=nucleotide&rettype=fasta&retmode=text&id=XM_007627897.1YY,YYseqYY)XX

wc -l GCmangler2.txt > length
length=`awk '{print $1 }' < length`
for ((  i = 1 ;  i <= $length ; i++  ))
do
cat GCmangler2.txt | head -$i | tail -1 > data
Genbank=`cat data | gawk '{print $1}'`
cat ~/software/perlGenbankdownload.txt | sed "s/XM_007627897.1/$Genbank/g" |  sed "s/XX/'/g" |  sed 's/YY/"/g' > perlscript.pl
chmod a+x ./perlscript.pl
./perlscript.pl
head -1 seq | sed 's/^.*|ref|/>/g' | cut -f 1,1 -d'|' > ./$Genbank.fasta
grep -v "^>" seq >>  ./$Genbank.fasta
done
```

Having all these files in one folder the files were blasted against the *C. griseus* chromosomes one by one and the name of the contig it was found on was extracted and then used to search the list of chromosomes made above

Blastall was downloaded[21] and the CHOchromosomes.fa file was indexed

```
formatdb -i CHOchromosomes.fa -o T -p F
```

Subsequently all transcripts annotated from the C.griseus genome was blasted against to find the chromosome name.

```
for file in *.fasta
do
blastall -i $file -d ./CHOchromosomes.fa -p blastn -o $file.out
grep -A2 'Sequences producing significant alignments:' $file.out | tail -1 | gawk '{print $1}' > contig
grep -A2 'Sequences producing significant alignments:' $file.out | tail -1 | gawk '{print $2}' > Eval
grep -A2 'Sequences producing significant alignments:' $file.out | tail -1 | gawk '{print $3}' > Eval2
head -1 $file | sed 's/^.//' > genbankname
chromosome=`awk '{print $1 }' < contig`
grep -w "$chromosome" ./chromosomeoverview | gawk '{print $2}' > chromosome
genbankname=`awk '{print $1 }' < genbankname`
grep -w "$genbankname" ./listofgenes.txt | head -1 > data
paste genbankname contig chromosome Eval Eval2 data >> col
rm contig
rm genbankname
rm Eval
rm Eval2
rm chromosome
rm data
done
```

Some none-protein-coding genes were removed and a list of the gene names were extracted

[20] http://www.bioinformatics-made-simple.com/2012/07/some-easy-ways-to-download-multiple.html
[21] http://ged.msu.edu/angus/tutorials/unix-and-blast.html

```
printf "\tASDF\n" | awk -F'\t' '$10' col.txt > col2.txt
cat ./col2.txt | gawk '{print $6}' | sort | uniq > genes
```

The list of gene names were gone through one by one in order to weed out the spleice vatiants.

```
wc -l genes > length
length=`awk '{print $1 }' < length`
for ((  i = 1 ;  i <= $length ; i++  ))
do
cat genes | head -$i | tail -1 > gene
gene=`awk '{print $1 }' < gene`
grep -w "$gene" col2.txt | sort -k 4 -n | tail -1 >> col3.txt
done
```

Finally a list contaning just the gene name and the suggested chromosome number was made

```
cat ./col3.txt | gawk '{print $3}' > chr
cat ./col3.txt | gawk '{print $6}' > gens
paste gens chr | sort | uniq > ~/chroverview.txt
```

This list was imported into R and the number of genes found on each chromosome were plotted versus the lengt of each chromosome as listed in the original paper. A correlation of 0.95 were found indicating that most genes are probably correctly placed

```
chroverview = read.table("chroverview.txt")
rownames(chroverview)=chroverview[,1]
test=0;for(i in 0:9)
{test[i+1]=dim(chroverview[which(chroverview[,2]==i),])[1]}
chrlengths=c(114,492,432,247,190,190,153,125,94,52)
chrnames=c("X",1,2,3,4,5,6,7,8,"9/10")
plot(1000,1000,xlim=c(0,4000),ylim=c(0,500),xlab="Genes",ylab="Chr lengths")
for(i in 1:10) {text(test[i],chrlengths[i],chrnames[i])}
cor(test,chrlengths)
```

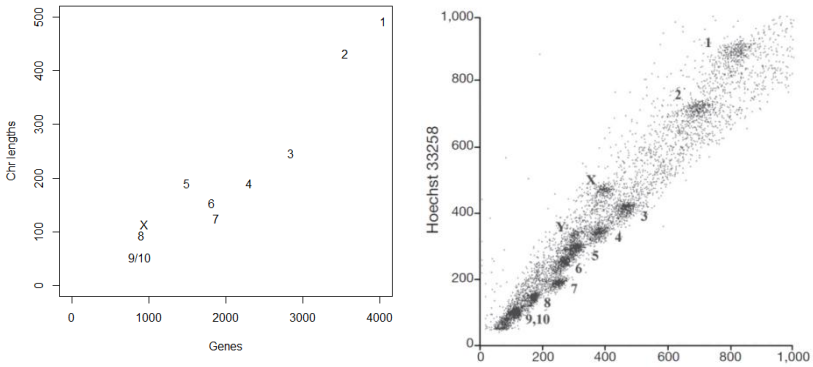


Figure 19 *C. griseus* chromosomes
Left: Chromosome length versus number of genes as found in R from input data. Right: Figure from[22] showing bivariate flow cytometric analysis of *C. griseus* chromosomes prior to isolation and sequencing.

The distribution of copy number in CHO-K1 ATCC versus CHO DXB11 were plotted for each chromosome

```
png("Figure2.png", width = 20, height = 15, units = 'cm', res = 300)
```

[22] http://www.nature.com/nbt/journal/v31/n8/full/nbt.2645.html

```
par(mfrow = c(3,4));par(mar = c(0.1, 0.1, 0.1, 0.1))
plot(coll[,9],coll[,7],ylim=c(0,4),xlim=c(0,4),pch=20,col="#00000010",ylab=" ",xlab=" ",main=" ",xaxt='n',yaxt='n')
Axis(side=1, labels=FALSE);Axis(side=2, labels=FALSE);Axis(side=3, labels=FALSE);Axis(side=4, labels=FALSE)
for(i in 1:9)
{
plot(coll[rownames(chroverview[which(chroverview[,2]==i),]),9],coll[rownames(chroverview[which(chroverview[,2]==i),]),7],ylim=c(0,4),xlim=c(0,4),pch=20,col="#00000010",ylab=" ",xlab=" ",main=" ",xaxt='n',yaxt='n')
#text(3,3,chrnames[i+1]) #show chr number on plot
Axis(side=1, labels=FALSE);Axis(side=2, labels=FALSE);Axis(side=3, labels=FALSE);Axis(side=4, labels=FALSE)
}
plot(coll[rownames(chroverview[which(chroverview[,2]==0),]),9],coll[rownames(chroverview[which(chroverview[,2]==0),]),7],ylim=c(0,4),xlim=c(0,4),pch=20,col="#00000010",ylab=" ",xlab=" ",main=" ",xaxt='n',yaxt='n')
#text(3,3,chrnames[1]) #show chr number on plot
Axis(side=1, labels=FALSE);Axis(side=2, labels=FALSE);Axis(side=3, labels=FALSE);Axis(side=4, labels=FALSE)
dev.off()
```

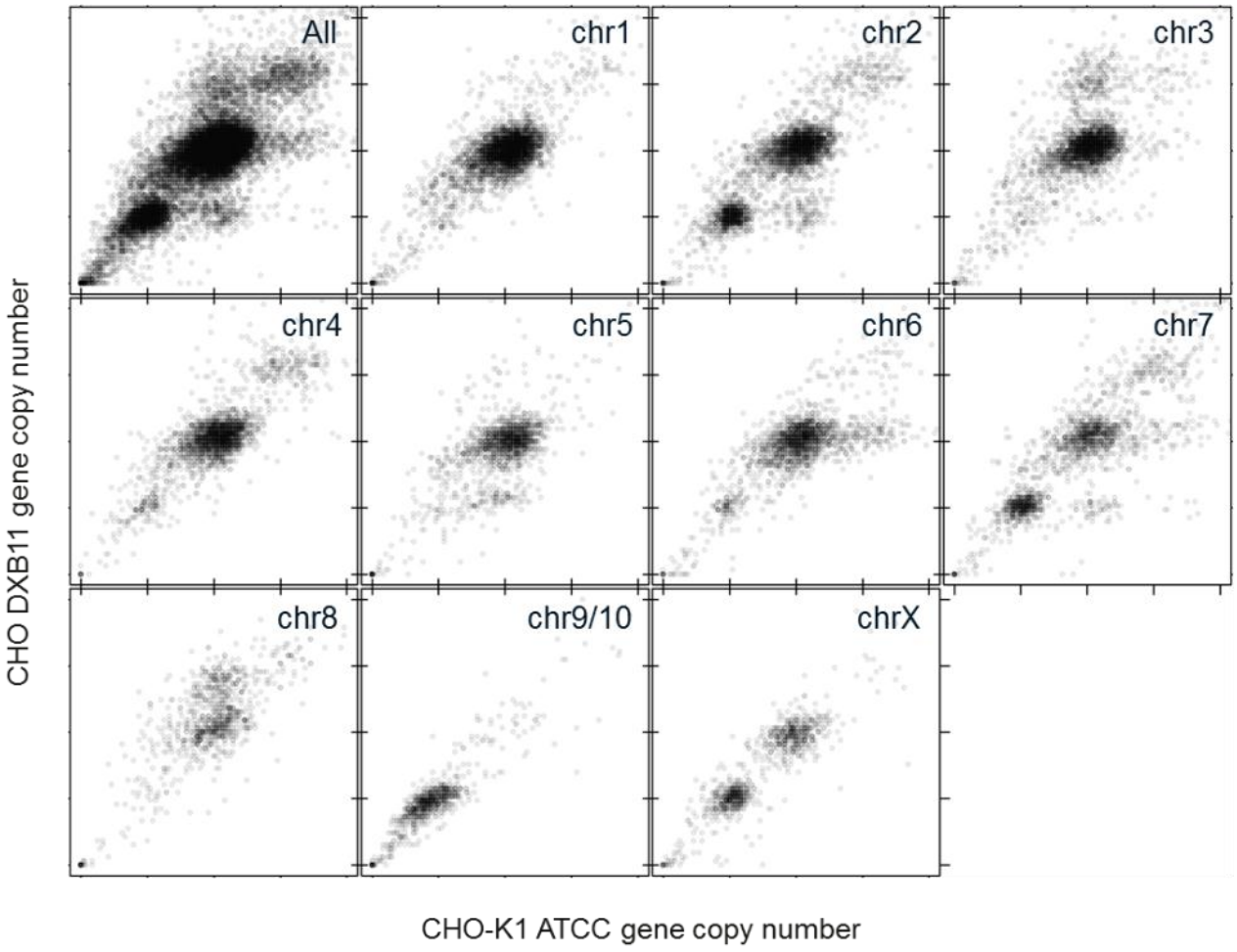


Figure 20 copy number distribution for CHO-K1 ATCC versus CHO DXB11 sorted by chromosome.
The x- and y-axes and chromosome names have been added manually in powerpoint.

Subsequently the same data was used for creating a phylogenetic tree for each chromosome

```
for(k in 0:9)
{
coll2=coll[rownames(chroverview[which(chroverview[,2]==k),]),]
distance=distance2=matrix(ncol=ncol(col),nrow=ncol(col),0);for(i in 1:ncol(col))
{for(j in 1:ncol(coll))
  {col3=coll2[which(abs(coll2[,j]-coll2[,i])>0.95),]
   distance[i,j]=dim(col3)[1]/dim(coll2)[1]
   distance2[i,j]=dim(col3)[1]}
```

```
}
colnames(distance)=colnames(distance2)=rownames(distance)=rownames(distance2)=colnames(coll)
plot(nj(as.dist(distance)), type="u", main=chrnames[k+1])
windows()
}
```

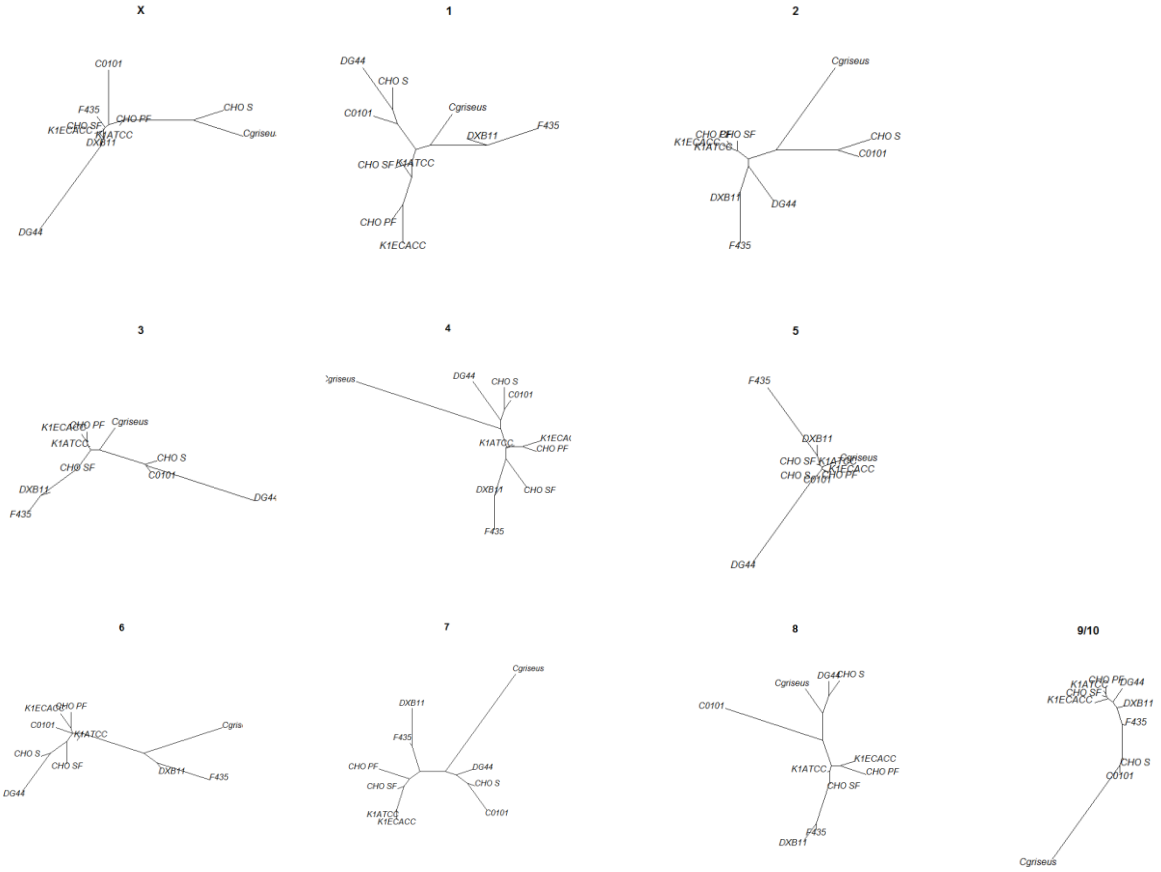


Figure 21 phylogenetic tree among the CHO genomes evalutaed for each chromosome.

The same information from above can be told in heatmaps. It seems that chromosome 1 and 4 are more stable than the other chromosomes.

```
Box <- function(x, y, color)
{rect(x, y, x+1, y+1, density = NULL, col = color, border = NULL, lty = par("lty"), lwd = par("lwd"))}
chrnames=c("chrX","chr1","chr2","chr3","chr4","chr5","chr6","chr7","chr8","chr9/10")

#rearrange the columns so heatmap is easier to read
names=colnames(coll)
coll3=cbind(coll[,1],coll[,3],coll[,4],coll[,10],coll[,9],coll[,7],coll[,8],coll[,6],coll[,5],coll[,2])
colnames(coll3)=c(names[1],names[3],names[4],names[10],names[9],names[7],names[8],names[6],names[5],nam
es[2])
head(coll3)
colnames(coll3)

png("70.png", width = 20, height = 15, units = 'cm', res = 300)
colfunc <- colorRampPalette(c("white", "white"))
colfunc2 <- colorRampPalette(c("white", "blue"))
colfunc3 <- colorRampPalette(c("blue", "green"))
colfunc4 <- colorRampPalette(c("green", "red"))
colors=c(colfunc(2),colfunc2(8),colfunc3(10),colfunc4(30))
#colors=c(colfunc(70),colfunc2(30))
par(mfrow = c(3,4));par(mar = c(0.1, 0.1, 0.1, 0.1))
```

```
#plot(1,1,xlim=c(2,11),ylim=c(2,11))
#text(6,6,"max 70%",cex=3,pos=1)
distance=matrix(ncol=ncol(col),nrow=ncol(coll3),0);for(i in 1:ncol(coll3))
{for(j in 1:ncol(coll3))
  {col3=coll3[which((coll3[,j]-coll3[,i])>0.95),1]
   distance[i,j]=length(col3)[1]/dim(coll3)[1]
  }
}
plot(100,100, xlim=c(0,12),ylim=c(0,12),xaxt='n',yaxt='n',ylab="",xlab="")
for(j in 1:ncol(coll))
{for(i in 1:ncol(coll))
 {Box(j,i,colors[distance[j,i]*100])}
}
for(i in 1:ncol(col)) {text(i+0.5,1,c(1:10)[i],cex=1.3,pos=1)}
for(i in 1:ncol(col)) {text(0.4,i+1.3,c(1:10)[i],cex=1.3,pos=1)}
text(6,12.5,"All",cex=1.3,pos=1)

for(k in 1:9)
{
coll2=coll3[rownames(chroverview[which(chroverview[,2]==k),]),]
distance=distance2=matrix(ncol=ncol(col),nrow=ncol(col),0);for(i in 1:ncol(col))
  {for(j in 1:ncol(coll3))
    {col2=coll2[which((coll2[,j]-coll2[,i])>0.95),1] #CN up
     distance[i,j]=length(col2)[1]/dim(coll2)[1]
    }
  }
plot(100,100, xlim=c(0,12),ylim=c(0,12),xaxt='n',yaxt='n',ylab="",xlab="")
for(j in 1:ncol(coll3))
{for(i in 1:ncol(coll3))
 {Box(j,i,colors[distance[j,i]*100])}
}
for(i in 1:ncol(col)) {text(i+0.5,1,c(1:10)[i],cex=1.3,pos=1)}
for(i in 1:ncol(col)) {text(0.4,i+1.3,c(1:10)[i],cex=1.3,pos=1)}
text(6,12.5,chrnames[k+1],cex=1.3,pos=1)
}
coll2=coll3[rownames(chroverview[which(chroverview[,2]==0),]),]
distance=distance2=matrix(ncol=ncol(col),nrow=ncol(col),0);for(i in 1:ncol(col))
  {for(j in 1:ncol(coll3))
    {col1=coll2[which(abs(coll2[,j]-coll2[,i])>0.95),1] #both up and down
     col2=coll2[which((coll2[,j]-coll2[,i])>0.95),1] #CN up
     #col3=coll2[which((coll2[,j]-coll2[,i])<(-0.95)),1] #CN down
   distance[i,j]=length(col2)[1]/dim(coll2)[1]
   distance2[i,j]=length(col2)[1]
  }
}
#colnames(distance)=colnames(distance2)=colnames(coll3);rownames(distance)=rownames(distance2)=colnames(co
ll3)
colnames(distance)=rownames(distance)=c(1:10)
plot(100,100, xlim=c(0,12),ylim=c(0,12),xaxt='n',yaxt='n',ylab="",xlab="")
for(j in 1:ncol(coll3))
{for(i in 1:ncol(coll3))
 {Box(j,i,colors[distance[j,i]*100])}
}
for(i in 1:ncol(col)) {text(i+0.5,1,c(1:10)[i],cex=1.3,pos=1)}
for(i in 1:ncol(col)) {text(0.4,i+1.3,c(1:10)[i],cex=1.3,pos=1)}
text(6,12.5,chrnames[1],cex=1.3,pos=1)
dev.off()
```

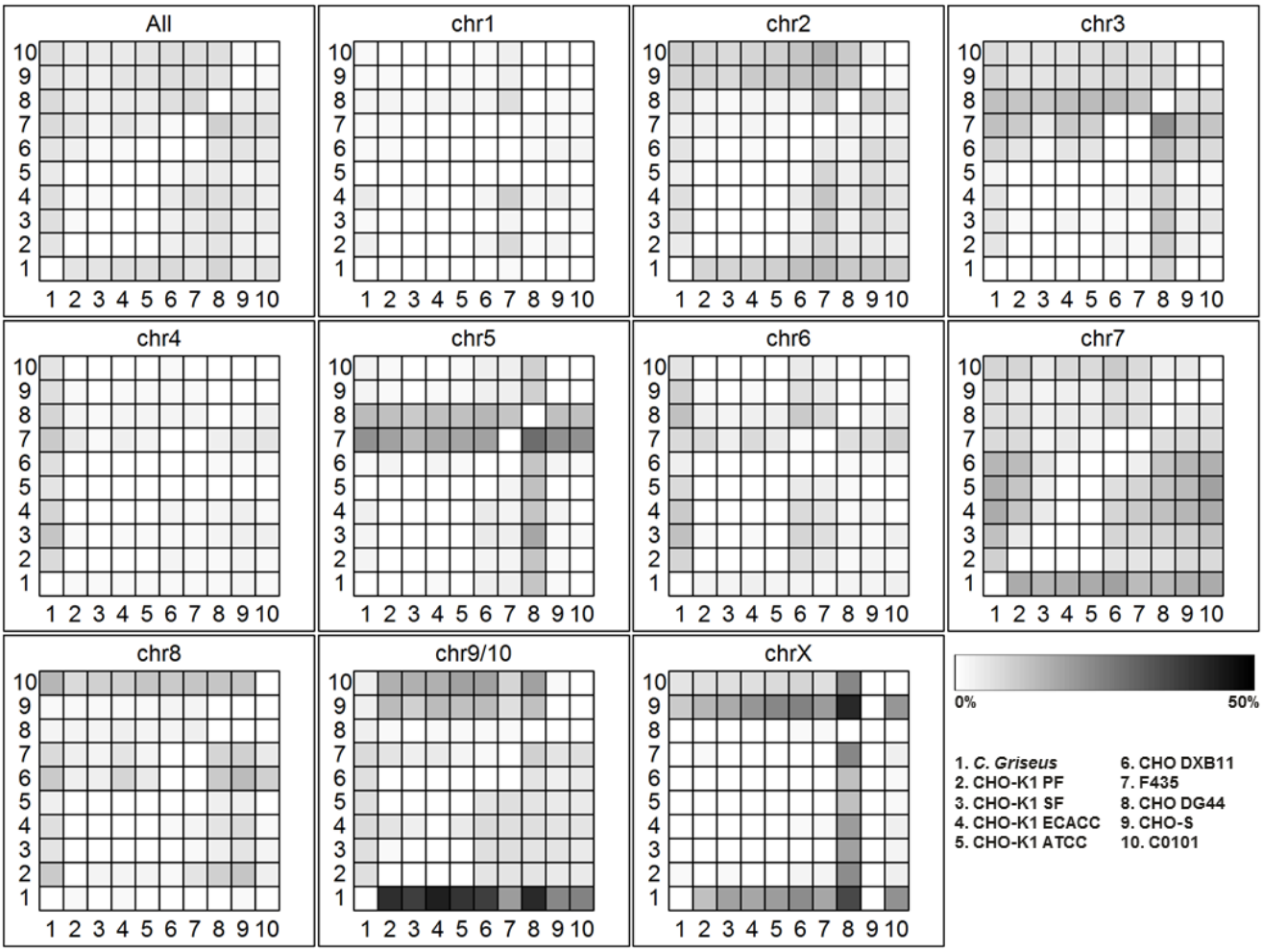


Figure 22 Heatmap visualising the CN reductions and amplifications seen between the currently sequenced CHO cell lines and *C. griseus*

Absolute copy numbers

In order to have provide an absolute copy number for each gene in each cell line it was assumed that the distribution was the same for all cell lines but CHO DXB11 were used to guide the cutoff. Genes not detected in any cell line: 159 genes had a median depth of 0 in all 10 genomes (including *C. griseus*). Those 159 genes were taken out of the dataset.

Haploid genes: depth higher than 0 but lower than 1.3 based on local minimum in CHO DXB11 between the haploid and diploid peak. Diploid: higher than 1.3 and lower than 2.7 based on local minimum in CHO DXB11 between the diploid and triploid. Triploid or higher: depth higher than 2.7. Finally all genes are assumed to be diploid in *C. griseus*

```
hist(coll[,7],breaks =c(0,1:5000/100,3000),xlim=c(2.3,3),ylim=c(0,2),main="CHO DXB11")
coll2=coll
coll2[which(coll[,1]==0),]=2 #159 genes are 0 in C.griseus so they are considered diploid (or not counted)
for(j in 2:ncol(coll))
{
coll2[which(coll[,j]==0 & abs(coll[,1]-coll[,j])<=0.95),j]=2
coll2[which(abs(coll[,1]-coll[,j])>=0.95 & coll[,j]==0),j]=0
coll2[which(coll[,j]<=1.3 & coll[,j]>0.0),j]=1
```

```
coll2[which(coll[,j]<=2.7 & coll[,j]>1.3),j]=2
coll2[which(coll[,j]<=3.5 & coll[,j]>2.7),j]=3
for(i in 4:159)
{coll2[which(coll[,j]<=(i+0.5) & coll[,j]>(i-0.5)),j]=i}
#coll2[which(round(coll2[,j], digits=0)!=coll2[,j]),]=round(coll2[,j], digits=0)
}
coll2[,1]=2 #C. griseus are assumed fully diploid
```

As a test Dhfr is investigated. It is seen that the copy number agree with the known copy number in F435, CHO DXB11, DG44 and CHO-K1

Dhfr	*C. griseus*	C0101	CHO-K1 PF	CHO-K1 SF	CHO-S	CHO DG44	CHO DXB11	F435	CHO-K1 ATCC	CHO-K1 ECACC
Raw depth	47	28	13	9	8	0	17	11	49	9
Normalized	1.8	2.4	2.2	2.0	1.8	0.0	1.0	1.3	2.0	2.3
Absolute	2	2	2	2	2	0	1	1	2	2

Finally a matrix is made summarizing the number of genes with each copy number. The number which are added to the 5th and 7th column are the number of deleted and duplicated genes as found in Lewis et al

```
cp=matrix(0,nrow=ncol(coll2),ncol=9);rownames(cp)=colnames(coll2);colnames(cp)=c(0,1,2,3,4,"N0","ND","hap","3
+")
for(j in 1:ncol(coll))
{cp[j,1]=nrow(coll2[which(coll2[,j]==0),])
cp[j,2]=nrow(coll2[which(coll2[,j]==1),])
cp[j,3]=nrow(coll2[which(coll2[,j]==2),])
cp[j,4]=nrow(coll2[which(coll2[,j]==3),])
cp[j,5]=nrow(coll2[which(coll2[,j]>=4),])}
cp[,6]=c(0,36,23,32,23,54,0,0,36,19)
cp[,7]=c(0,1935,1448,1574,2002,1881,0,0,1374,1430)
cp[,8]=cp[,2]/20661
cp[,9]=(cp[,4]+cp[,5])/20661
round(cbind(cp[,8]), digits=3)
```

	0	1	2	3	4	N0	ND	hap	3+
C. griseus	0	0	20661	0	0	0	0	0%	0%
C0101	37	3544	15088	1318	674	36	1935	17%	10%
CHO_PF	54	3356	15059	1734	458	23	1448	16%	11%
CHO_SF	48	3073	15453	1348	739	32	1574	15%	10%
CHO_S	47	3024	15603	1236	751	23	2002	15%	10%
DG44	62	4219	13967	1426	987	54	1881	20%	12%
DXB11	44	3586	15267	1323	441	0	0	17%	9%
F435	53	3888	14306	1713	701	0	0	19%	12%
K1ATCC	30	3773	15305	1085	468	36	1374	18%	8%
K1ECACC	57	4039	13310	2657	598	19	1430	20%	16%

GO-term enrichment

A list of GO-terms associated with the CHO genome was downloaded from CHOgenome.org (http://www.chogenome.org/files/CHO_GO_Functions_12Sep13.txt).

```
…
EGW13817      "Nxf2,Tap12,Nxf2b"          I79_021570    GO_component: GO:0005634 - nucleus [Evidence
IEA]; GO_process: GO:0051028 - mRNA transport [Evidence IEA]
…
```

As seen above the list contain multiple genes and multiple GO-terms on the same row. The file above was resorted into a file with one GO-term per line (but still multiple genes) using the script below

```
for ((  i = 1 ;  i <= 11691 ; i++  ))
do
cat GOliste.txt | head -$i | tail -1  > genename
cat genename | gawk '{print $1}' > gene
cat genename | sed 's/\t/\n/g' | sed 's/;/\n/g' |  sed 's/^ //g' | sed 1d > Goterms
number=`wc -l Goterms | gawk '{print $1}' `
for ((  j = 1 ;  j <= $number ; j++  ))
do
cat Goterms | head -$j | tail -1  > namespecific
paste gene namespecific >> namelist
done
done
rm genename
rm GOterm
rm names
rm namespecific
```

A list of all unique genes in the *C. griseus* genome was generated previously and used one by one to search the file made above to make a list with just the genename and the a GO-term

```
for ((  i = 1 ;  i <= 20661 ; i++  ))
do
cat uniqgenes | head -$i | tail -1  > genename
gene=`cat genename`
grep "$gene" namelist | awk '{$1=""}1' | cut -f 1,1 -d'['  | sed 's/^ //g' > Goterms
number=`wc -l Goterms | gawk '{print $1}' `
for ((  j = 1 ;  j <= $number ; j++  ))
do
cat Goterms | head -$j | tail -1  > namespecific
paste genename namespecific >> namelist3
done
done
```

The list was imported into R. The number of genes with different CN was used to find the number of genes in total amplified or reduced compared to *C. griseus*

```
GOterms=unique(read.table("namelist3"))
GoUnique=unique(GOterms[,1])

diffgenes=0;for(i in 2:10)
{
t1=i*2-3;t2=i*2-2
diffgenes[t1]=nrow(coll2[which(coll2[,i]>coll2[,1]),]) #amplification
diffgenes[t2]=nrow(coll2[which(coll2[,i]<coll2[,1]),]) #reduction
}
```

Subsequently, every single GO-term was investigated mining how many CN differences were detected for each cell line for that GO-term. The Fischer exact test was used to determine whether a GO-term was significantly enriched for amplified or reduced genes.

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\ (c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,d!\,n!}$$

Where a = Number of genes with a specific GO-term AND reduced/amplified, b= number of reduced/amplified genes, which did not have this GO-term, c = Number of genes with tis GO-term, which did not have a CN change. d = Number of genes, which did not have the GO-term nor a CN change. a+b+c+d = the 20661 genes in *C. griseus*.

```
GOlist=matrix(nrow=length(GoUnique),ncol=37,0);rownames(GOlist)=GoUnique;colnames(GOlist)=c("Genes_with_G
O",1:36)
for(j in 1:length(GoUnique))
{
subGOlist=GOterms[which(GOterms[,1]==as.character(GoUnique[j])),]
r=matrix(nrow=nrow(subGOlist),ncol=18,0)
for(i in 1:nrow(subGOlist))
 {
subsub=coll[as.character(subGOlist[i,2]),]
#C0101
if(subsub[2]-subsub[1] > 0.95)  { r[i,1] <- 1} else {r[i,1] <- 0} #amp
if(subsub[2]-subsub[1] < -0.95) { r[i,2] <- 1} else {r[i,2] <- 0} #reduk
#CHO_PF
if(subsub[3]-subsub[1] > 0.95)  { r[i,3] <- 1} else {r[i,3] <- 0}
if(subsub[3]-subsub[1] < -0.95) { r[i,4] <- 1} else {r[i,4] <- 0}
#CHO_SF
if(subsub[4]-subsub[1] > 0.95)  { r[i,5] <- 1} else {r[i,5] <- 0}
if(subsub[4]-subsub[1] < -0.95) { r[i,6] <- 1} else {r[i,6] <- 0}
#CHO_S
if(subsub[5]-subsub[1] > 0.95)  { r[i,7] <- 1} else {r[i,7] <- 0}
if(subsub[5]-subsub[1] < -0.95) { r[i,8] <- 1} else {r[i,8] <- 0}
#DG44
if(subsub[6]-subsub[1] > 0.95)  { r[i,9] <- 1} else {r[i,9] <- 0}
if(subsub[6]-subsub[1] < -0.95) { r[i,10] <- 1} else {r[i,10] <- 0}
#DXB11
if(subsub[7]-subsub[1] > 0.95)  { r[i,11] <- 1} else {r[i,11] <- 0}
if(subsub[7]-subsub[1] < -0.95) { r[i,12] <- 1} else {r[i,12] <- 0}
#F435
if(subsub[8]-subsub[1] > 0.95)  { r[i,13] <- 1} else {r[i,13] <- 0}
if(subsub[8]-subsub[1] < -0.95) { r[i,14] <- 1} else {r[i,14] <- 0}
#K1ATCC
if(subsub[9]-subsub[1] > 0.95)  { r[i,15] <- 1} else {r[i,15] <- 0}
if(subsub[9]-subsub[1] < -0.95) { r[i,16] <- 1} else {r[i,16] <- 0}
#K1ECACC
if(subsub[10]-subsub[1] > 0.95)  { r[i,17] <- 1} else {r[i,17] <- 0}
if(subsub[10]-subsub[1] < -0.95) { r[i,18] <- 1} else {r[i,18] <- 0}
 }
#
  r2=r3=0;for(i in 1:18)
  {
  r2[i]=sum(r[,i])
  a=sum(r[,i]);b=diffgenes[i]-a;c=nrow(subGOlist)-a;d=20661-a-b-c
  Fischer=matrix(c(a,c,b,d),nrow = 2)
  r3[i]=fisher.test(Fischer, alternative = "two.sided")$p.value
  }
```

```
GOlist[j,]=c(nrow(subGOlist),r2,r3)
}
```

Following evaluation of all GO-terms in all cell lines, a list of all GO-terms which had a p-value < 0.01 in just one of the 9 cell lines in either enrichment of amplified or reduced genes was made and exported for supplementary table 7.

```
test=rlist=0
for(i in 20:37)
{
t1=i-19
rlisttemp=GOlist[which(GOlist[,i]<0.01),]
test[t1]=nrow(rlisttemp)
rlist=rbind(rlist,rlisttemp)
}
nrow(unique(rlist))

sGO=r3[order(r3[,31]),20:37] #sort by reduction in CHO DXB11
names=c("C0101 duplication","C0101 reduction","CHO-K1 PF duplication","CHO-K1 PF reduction","CHO-K1 SF
duplication","CHO-K1 SF reduction","CHO-S duplication","CHO-S reduction","CHO DG44 duplication","CHO DG44
reduction","CHO DXB11 duplication","CHO DXB11 reduction","F435 duplication","F435 reduction","CHO-K1 ATCC
duplication","CHO-K1 ATCC reduction","CHO-K1 ECACC  duplication","CHO-K1 ECACC  reduction")
colnames(sGO)=names
significantGO2=cbind(sGO[,3],sGO[,5],sGO[,17],sGO[,15],sGO[,11],sGO[,13],sGO[,9],sGO[,7],sGO[,1],sGO[,4],sGO[
,6],sGO[,18],sGO[,16],sGO[,12],sGO[,14],sGO[,10],sGO[,8],sGO[,2])
names=c("CHO-K1 PF duplication","CHO-K1 SF duplication","CHO-K1 ECACC  duplication","CHO-K1 ATCC
duplication","CHO DXB11 duplication","F435 duplication","CHO DG44 duplication","CHO-S duplication","C0101
duplication ","CHO-K1 PF reduction","CHO-K1 SF reduction","CHO-K1 ECACC  reduction","CHO-K1 ATCC
reduction","CHO DXB11 reduction","F435 reduction","CHO DG44 reduction","CHO-S reduction","C0101 reduction")
names=c("CHO-K1 PF","CHO-K1 SF","CHO-K1 ECACC ","CHO-K1 ATCC","CHO DXB11","F435","CHO DG44","CHO-
S","C0101 ","CHO-K1 PF","CHO-K1 SF","CHO-K1 ECACC","CHO-K1 ATCC","CHO DXB11","F435","CHO DG44","CHO-
S","C0101")
colnames(significantGO2)=names
write.table(significantGO2,file="Sup_table7.txt",sep="\t") #all genes
```

Finally the table above were visualized as a heatmap also where p-values < 0.01 were dark blue and 0.01 < p-values < 0.05 were shown as blue.

```
Box <- function(x, y, color)
{rect(x, y, x+1, y+1, density = NULL, col = color, border = NULL, lty = par("lty"), lwd = par("lwd"))}
colfunc <- colorRampPalette(c("darkblue", "darkblue"))
colfunc2 <- colorRampPalette(c("blue", "blue"))
colfunc5 <- colorRampPalette(c("white", "white"))
colors=c(colfunc(10),colfunc2(40),colfunc5(950))

png("SupplFigure4.png", width = 30, height = 30, units = 'cm', res = 300)
plot(100,100, xlim=c(1,nrow(significantGO2)+60),ylim=c(0,25),xaxt='n',yaxt='n',ylab="",xlab="")
for(j in 1:nrow(significantGO2))
{
for(i in 1:9){Box(j+34,i,colors[round(significantGO2[j,i]*1000,0)+1])}
for(i in 10:18){Box(j+34,i+1,colors[round(significantGO2[j,i]*1000,0)+1])}
}
for(i in 1:9) {text(0.5,i+0.5,names[i],cex=0.60,pos=4)}
for(i in 10:18) {text(0.5,i+1.5,names[i],cex=0.60,pos=4)}
for(i in 1:nrow(significantGO2)) {text(i+32.25,20.1,rownames(significantGO2)[i],cex=0.2,pos=4,srt = 45)}

dev.off()
```

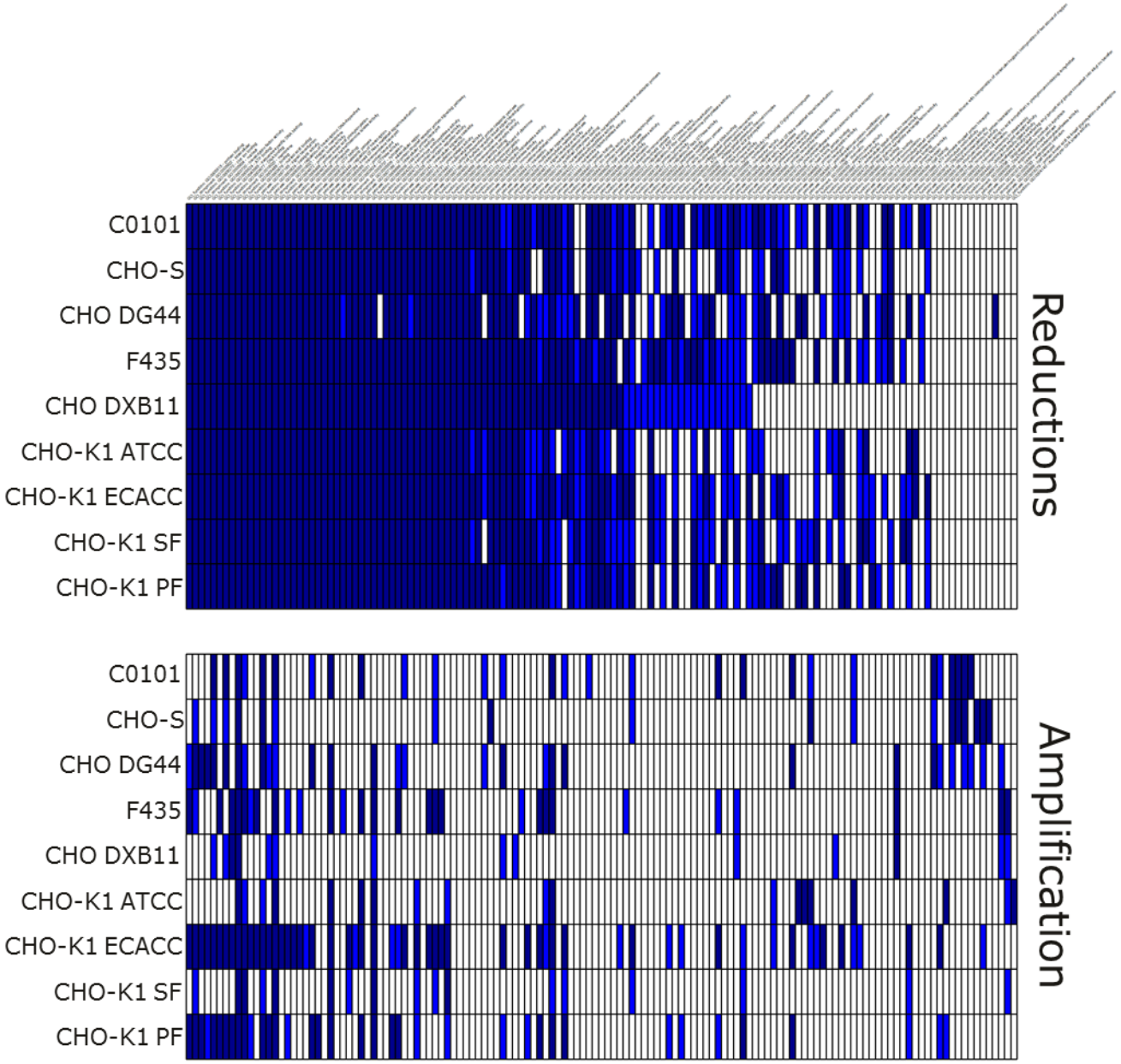


Figure 23 Significant GO-terms in correlation to changes in copy number
Visualization of the 135 GO-terms, which are either significantly enriched in genes with CN reductions or amplifcations (Fisher's exact test). GO-terms are visualized in dark blue (p-value < 0.01), light blue (p-value < 0.05) or white (p-value > 0.05).

10.3.3 Detecting genomic rearrangements from break-spanning reads

This section describes how structural variants from a resequenced genomes are located using SVdetect[23].

Pre-alignment trimming of reads

Trim Galore is a usefull tool, which are able to trim the paired reads from a genome resequencing experiment, while keeping the files synchronized so the pairs are broken[24]. It can be downloaded from here[25] and the required packagages cutadapt from here[26] and fastqc from here[27]. In the code below all low quality bps are trimmed from the ends until bps with a quality >30 are found, but all reads that after trimming are shorter than 40bp are removed from the dataset. See full decription of options here[28]

```
fastqc DXB11_1.fastq
fastqc DXB11_2.fastq
mkdir ./trimmed_reads
~/software/Trimgalore/trim_galore --paired -q 30 --length 40 --fastqc --path_to_cutadapt /novo/appl/ngs/cutadapt-1.2.1/bin/cutadapt --stringency 5 -o ./trimmed_reads DXB11_1.fastq DXB11_2.fastq
```

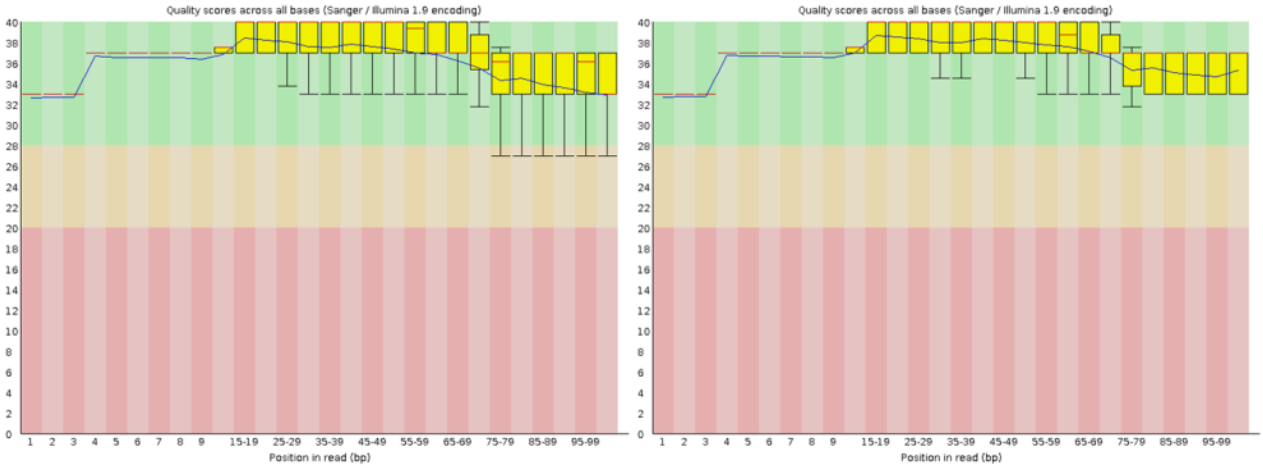


Figure 24 quality score across the reads before (left) and after (right) trimming.

Reads were aligned to the *C. griseus* genome

The trimmed reads are subsequently aligned to the *C. griseus* genome using BWA[29] and subsequently converted into a bam-file from a sam-file. First, the genome has to be indexed by BWA before alignment.

---

[23] http://svdetect.sourceforge.net/Site/Home.html
[24] http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
[25] http://www.bioinformatics.babraham.ac.uk/projects/download.html#trim_galore
[26] https://pypi.python.org/pypi/cutadapt/
[27] http://www.bioinformatics.babraham.ac.uk/projects/download.html#fastqc
[28] http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/trim_galore_User_Guide_v0.4.0.pdf
[29] http://bio-bwa.sourceforge.net/

bwa index Cgriseus.fa

After the genome has been indexed the reads are alinged to the genome

bwa aln Cgriseus.fasta 1BWA_1.fastq > 1BWA_1.sai
bwa aln Cgriseus.fasta 1BWA_2.fastq > 1BWA_2.sai
bwa sampe -r "@RG\tID:libA\tSM:sample_EACC\tPL:ILLUMINA" Cgriseus.fasta 1BWA_1.sai 1BWA_2.sai 1BWA_1.fastq 1BWA_2.fastq | samtools view -Sb - > 2alignment_L.bam

All reads with a minimum mapping quality less than 30 were removed and the file was sorted and indexed

samtools view -u -q 30 2alignment_L.bam | samtools sort - 2alignment_L.sort
samtools index 2alignment_L.sort.bam

GATK was used in order to remap all the reads in region containing indels

The RealignerTargetCreater algorithm[30] was used to realign reads in problematic regions of the genome.

java -Xmx2g -jar GenomeAnalysisTK.jar -T RealignerTargetCreator -R Cgriseus.fasta -I 2alignment_L.sort.bam -o 3_L.intervals

java -Xmx2g -jar GenomeAnalysisTK.jar -T IndelRealigner -targetIntervals 3_L.intervals -R Cgriseus.fasta -I 2alignment_L.sort.bam -o 3realigned_L.bam

Alignments run in parallel were merged into one bam-file

In the steps above the data from the same sample was split into multiple files and now the files are merged from several bam files into one single bam file. This is done using samtools[31]

samtools merge Cgris_prePicard.bam 3realigned_L.bam 3realigned_L.bam /H3/3realigned_L.bam /H4/3realigned_L.bam /H5/3realigned_L.bam

PCR duplicates were removed using Picard

PCR remnants are removed using Picard[32]

java -Xmx2g -jar MarkDuplicates.jar INPUT=Cgris_prePicard.bam OUTPUT=4Picard_L.bam.bam ASSUME_SORTED=FALSE METRICS_FILE=/dev/null VALIDATION_STRINGENCY=SILENT REMOVE_DUPLICATES=true

In the 4Picard_L.bam file made above, a certain number of paired reads do not align to the same genomic scaffolds in close proximity to each other. Some of these will be noise, but with proper filtering it should be possible to hypothesize regarding genomic rearrangements using these anomalous reads. This can be carried out using the program SVdetect[33].

[30] https://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_gatk_tools_walkers_indels_RealignerTargetCreator.php
[31] http://samtools.sourceforge.net/samtools.shtml
[32] http://broadinstitute.github.io/picard/
[33] http://svdetect.sourceforge.net/Site/Home.html

Detect break-spanning reads

The SVDetect BAM_preprocessingPairs script found in the SVdetect library was used to pre-process the bam file made above

```
perl ./SVDetect_r0.8b/scripts/BAM_preprocessingPairs.pl 4Picard.bam
```

Giving the output:

```
Total : 62889376 pairs analysed
-- 251915 pairs whose one or both reads are unmapped
-- 62637461 mapped pairs
---- 746850 abnormal mapped pairs
------ 542994 pairs mapped on two different chromosomes
------ 443784 pairs with incorrect strand orientation and/or pair order
------ 179904 pairs with incorrect insert size distance
---- 61890611 correct mapped pairs
```

The bam file was converted to a sam file

```
samtools view -h -o 4Picard.ab.sam 4Picard.ab.bam
```

Prepare chromosome list for SVdetect

A list of the chromosomes and their length (called sortedlist.txt) is required to run SVdetect. Such a list was extracted from the header of the sam file generated above

```
grep "^@SQ" DXB11_new.ab.sam > headerraw
sed 's/SN://g' headerraw | sed 's/LN://g' | gawk '{print $2, $3}' | sed 's/ /\t/g'> chr_len

length=`wc -l chr_len | awk '{print $1 }'`
for (( j = 1 ; j <= $length ; j++ ))
do
echo $j >> numbers
done

paste numbers chr_len > sortedlist.txt
```

Yielding a list looking like this:

```
1       NW_003717424.1  209
2       NW_003613580.1  8779783
3       NW_003613581.1  8081566
4       NW_003613582.1  6666273
5       NW_003613583.1  6533376
...
4390    NW_003712234.1  219
4391    NW_003713686.1  216
4392    NC_007936.1     16284
```

Calculate paired read distance

Calculate the paired end distance using Picard

```
java -jar /novo/appl/ngs/picard-tools-1.70/CollectInsertSizeMetrics.jar INPUT=./4Picard.bam ASSUME_SORTED=true
OUTPUT=./histogram.numerical.txt REFERENCE_SEQUENCE=./Cgris.fasta
HISTOGRAM_FILE=./histogram.histogram.pdf VALIDATION_STRINGENCY=SILENT
```

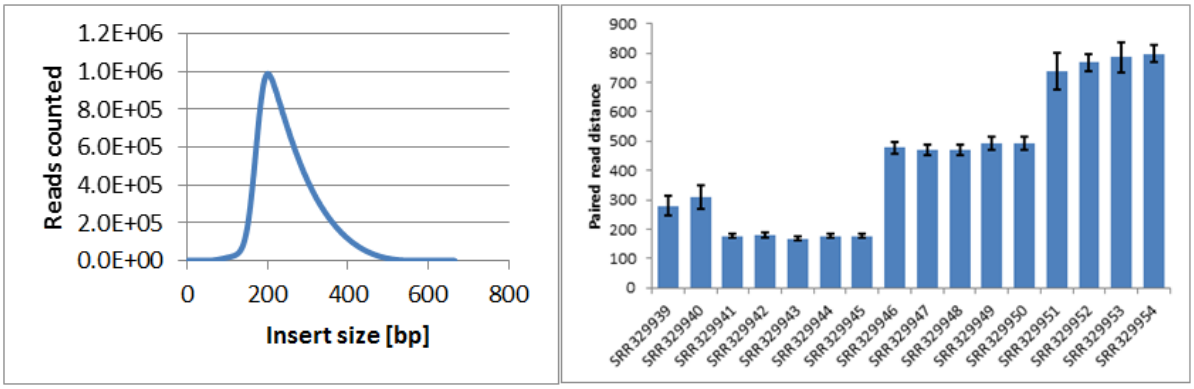


Figure 25 Paired read distance
Left: Paired read distance distribution for the CHO DXB11 genome re-sequencing data showing 250±70bp of distance.
Right: Mean paired read distance for each of the CHO-K1 ATCC SRA files.

By removing the first 11 lines in the output file (./histogram.numerical.txt) the data could be loaded into R and the mean (250), median (236) and stdev (70) was found. It was furthermore seen in the data that 99% of the reads had an insert size less than 505 bp

```
R
test = read.table("1")
x=c(rep(test[,1],test[,2]))
mean(x); median(x);sd(x)
```

From fastqc of my treated reads it was found that my paired reads were on average 86bp in length. SVdetect advice is 2μ+2stdev used for cutoff for rearrangements: 2x (250 + 86 x2) + 2x 70 = 984 = 1000bp will be used as cutoff.

In order to compare to another genome, the distance for each dataset on the SRA from the CHO-K1 ATCC genome was analysed. From CHO-K1 ATCC a subset of SRR329941-5 was taken, which all had a distance of 175±7bp

Make the SVdetect config file

I made a config file used to run SVdetect

```
<general>
input_format=sam
sv_type=all
mates_orientation=FR #always use this for paired reads
read1_length=86 # data found from fastqc output
read2_length=86
mates_file=/novo/omicsmanager/processed01/1099/data/pipe2del/SVdetect/4Picardtest/4Picard.ab.sam #full path to file containing reads with irregular paired read distance
cmap_file=/novo/omicsmanager/processed01/1099/data/pipe2del/SVdetect/4Picardtest/sortedlist.txt #full path to list of chromosomes and their length
num_threads=32 #number of threads the calculation is splitted into
</general>

<detection>
split_mate_file=0 #see next section for details. If one uses a genome with >24 chromosomes the software cannot handle to make this file itself. Set this to 0 and make it manually as explained in the next section.
```

window_size=1000 #This is important. This is the distance that is checked and set as cutoff. Reads with insert sizes larger than this are accepted
step_length=500 # how big a windows used for analysis. They recommend ¼ or ½ of the window size so ½ is used.
</detection>

<filtering>
split_link_file=0 #same reason as above. Due to >24 chromosomes this need to be 0 and made manually
nb_pairs_threshold=18 #number of reads needed to validate a specific rearrangement. The median depth of 18 is used as a position on each copy of a given chromosome is 9 but each position is covered by 9 reads but also 9 more spanning with the insert size twice the length = a depth of 18 is picked in order to be able to recognize haploid rearrangements
strand_filtering=1 #analyse the distribution of forward and reverse reads. Checks that a rearrangements are not only backed by forward reads only on one side but an equal distribution of forward and reverse reads as expected
</filtering>

Create the sam.all files due to the fragmentation of the CHO genome

If you run the cofig file above with split_mate_file=1 it will try to open all contigs at the same time (make perfect sense in the well assembled human genome, but not in a 4000+ contig CHO) = the program crashes with a "too many open files"-error.

After contacting the programmer, a solution was found by making the sam.all files by extracting all lines in the sam file that is relevant for a given scaffold and place it in the "mates"-folder. When this is done the program works fine if run with split_mate_file=0 (they have already been split)

```
cat sortedlist.txt | gawk '{print $2}' > chrlist
wc -l chrlist > length
length=`awk '{print $1 }' < length`
for ((  i = 1 ;  i <= $length ; i++  ))
do
cat chrlist | head -$i | tail -1 > scaffold
scaffold=`awk '{print $1 }' < scaffold`
grep –w "$scaffold" 4Picard.ab.sam > ./mates/4Picard.ab.sam.all.$scaffold
done
```

Run SVDetect

```
perl ./SVDetect_r0.8b/bin/SVDetect linking -conf ./configfile.txt
perl  ./SVDetect_r0.8b/bin/SVDetect filtering -conf /./configfile.txt
perl  ./SVDetect_r0.8b/bin/SVDetect links2SV -conf ./configfile.txt
perl  ./SVDetect_r0.8b/bin/SVDetect links2circos -./configfile.txt
perl  ./SVDetect_r0.8b/bin/SVDetect links2bed -conf configfileDXB11.txt
```

Filtering the SVdetect output

10,028 rearrangement events were predicted in CHO DXB11 compared to CHO-K1 ATCC. Example output below from an intra-chromosomal structural variant (same scaffold the distance is just off) and the orientation of each end is the same (indicate a deletion). 48% of the rearrangements were found to be either in the first 10% or the last 10% of the scaffolds, which could be genuine due to the assembly of the genome.

INTRA NORMAL_SENSE - chrNW_003614308.1 4390-4898 - chrNW_003614308.1 7976-8846 11414 99% - - 1.000 --

In an interesting study by Mijuskovic et al[34], filtering was suggested based on 1) remove sequenced PCR duplicates using Picard, 2) remove reads with low mapping quality, 3) rearrangement found in the parental genome (running SVdetect in this case on the CHO-K1 genome using the raw CHO-K1 reads and filtering out all all rearrangements found in the parental reads) and 4) remove rearrangement within regions with simple repeats (suing repeatmasker). Using this strategy the number of SV's in their case went from 11262 to 513. In the case above it was reduced from 10,028 to 365, but validation of rearrangements using PCR did not prove to be succesfull. Even if that had been the case the amount of filtering required make the final list questionable for exactly how many false negatives are being ignored.

It is possible to detect rearrangement in complex genomes using software such as SVdetect, but due to repetitive sequences it would probably not yield a realistic result of the number of rearrangement events that has occoured.

10.3.4 SNP detection in two CHO genomes

Summary: This tutorial describes how to locate SNPs from in a CHO genome aligned to the *C. griseus* reference sequence. I kindly ask you cite: Kaas et al 2015, *Sequencing the CHO DXB11 genome reveals regional variations in genomic stability and haploidy* in case you will publish data based on this workflow.

Pre-alignment trimming of reads

Trim Galore is a usefull tool, which are able to trim the paired reads from a genome resquencing experiment, while keeping the files synchronized so the pairs are broken[35]. It can be downloaded from here[36] and the required packagages cutadapt from here[37] and fastqc from here[38]. In the code below all low quality bps are trimmed from the ends until bps with a quality >30 are found, but all reads that after trimming are shorter than 40bp are removed from the dataset. See full decription of options here[39]

```
fastqc DXB11_1.fastq
fastqc DXB11_2.fastq
mkdir ./trimmed_reads
~/software/Trimgalore/trim_galore --paired -q 30 --length 40 --fastqc --path_to_cutadapt /novo/appl/ngs/cutadapt-1.2.1/bin/cutadapt --stringency 5 -o ./trimmed_reads DXB11_1.fastq DXB11_2.fastq
```

[34] http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0048314
[35] http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
[36] http://www.bioinformatics.babraham.ac.uk/projects/download.html#trim_galore
[37] https://pypi.python.org/pypi/cutadapt/
[38] http://www.bioinformatics.babraham.ac.uk/projects/download.html#fastqc
[39] http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/trim_galore_User_Guide_v0.4.0.pdf

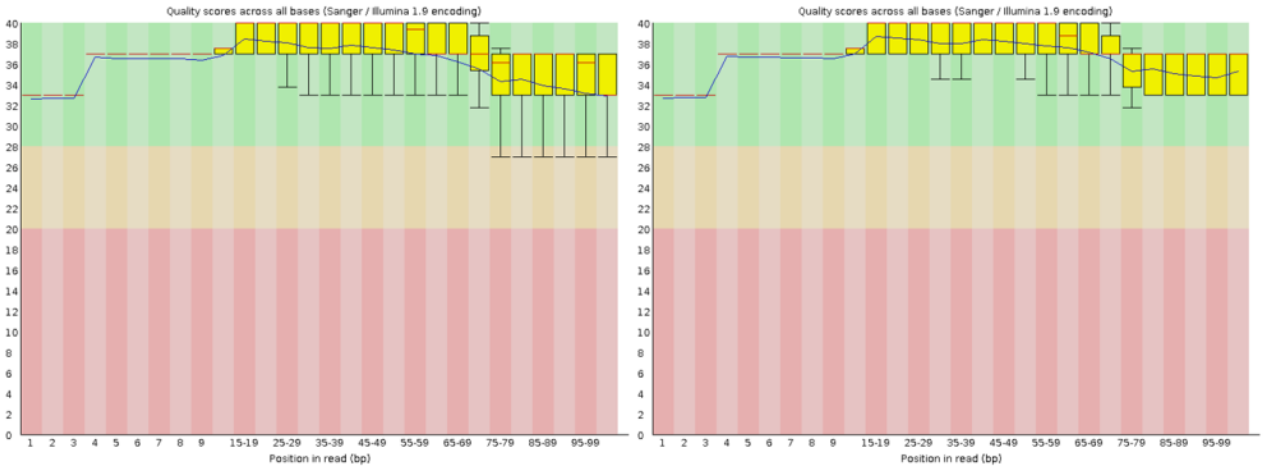


Figure 26 quality score across the reads before (left) and after (right) trimming.

Reads were aligned to the *C. griseus* genome

The trimmed reads are subsequently aligned to the *C. griseus* genome using BWA[40] and subsequently converted into a bam-file from a sam-file. First the genome has to be indexed by BWA before alignment.

bwa index Cgriseus.fa

After the genome has been indexed the reads are alined to the genome

bwa aln Cgriseus.fasta 1BWA_1.fastq > 1BWA_1.sai
bwa aln Cgriseus.fasta 1BWA_2.fastq > 1BWA_2.sai
bwa sampe -r "@RG\tID:libA\tSM:sample_EACC\tPL:ILLUMINA" Cgriseus.fasta 1BWA_1.sai 1BWA_2.sai 1BWA_1.fastq 1BWA_2.fastq | samtools view -Sb - > 2alignment_L.bam

All reads with a minimum mapping quality less than 30 were removed and the file was sorted and indexed

samtools view -u -q 30 2alignment_L.bam | samtools sort - 2alignment_L.sort
samtools index 2alignment_L.sort.bam

GATK was used in order to remap all the reads in region containing indels

The RealignerTargetCreater algorithm[41] was used to realign reads in problematic regions of the genome.

java -Xmx2g -jar GenomeAnalysisTK.jar -T RealignerTargetCreator -R Cgriseus.fasta -I 2alignment_L.sort.bam -o 3_L.intervals

java -Xmx2g -jar GenomeAnalysisTK.jar -T IndelRealigner -targetIntervals 3_L.intervals -R Cgriseus.fasta -I 2alignment_L.sort.bam -o 3realigned_L.bam

[40] http://bio-bwa.sourceforge.net/

[41] https://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_gatk_tools_walkers_indels_RealignerTargetCreator.php

Alignments run in parallel were merged into one bam-file

In the steps above the data from the same sample was split into multiple files and now the files are merged from several bam files into one single bam file. This is done using samtools[42]

```
samtools merge Cgris_prePicard.bam 3realigned_L.bam 3realigned_L.bam
/novo/omdb/pds01/PDS1099/data/pipe2del/hamsterreads/H3/3realigned_L.bam
/novo/omdb/pds01/PDS1099/data/pipe2del/hamsterreads/H4/3realigned_L.bam
/novo/omdb/pds01/PDS1099/data/pipe2del/hamsterreads/H5/3realigned_L.bam
```

PCR duplicates were removed using Picard and raw SNPs located

PCR remnants are removed using Picard[43]

```
java -Xmx2g -jar MarkDuplicates.jar INPUT=Cgris_prePicard.bam OUTPUT=Cgris_postPicard.bam
ASSUME_SORTED=FALSE METRICS_FILE=/dev/null VALIDATION_STRINGENCY=SILENT REMOVE_DUPLICATES=true
```

The Picard-treated[44] bam file was sorted based on genome

```
java -jar /picard-tools-1.70/ReorderSam.jar INPUT=Cgris_postPicard.bam OUTPUT=Cgris_postPicard.genomesort.bam
REFERENCE=Cgriseus.fasta VALIDATION_STRINGENCY=SILENT
```

Samtools mpileup[45] and bcftools[46] were used to locate SNPs

```
samtools mpileup -uf Cgriseus.fasta Cgris_postPicard.genomesort.bam | bcftools view -bvcg - > 3SNPs.raw.bcf
```

The bcf file was converted into vcf using bcftools

```
bcftools view 3SNPs.raw.bcf > name.vcf
```

Filtering based on background

The raw vcf-file from *C. griseus* was used to remove background SNPs found in *C. griseus*.

The vcf files were fist zipped and indexed

```
bgzip DXB11.vcf
tabix -p vcf DXB11.vcf.gz
```

VCF-tools[47] was used to filter out the background SNPs found in *C. griseus* reads against the *C. griseus* genome.

```
perl vcftools_0.1.12a/perl/vcf-isec -c DXB11.raw.vcf.gz Cgris.vcf.gz > DXB_uCgris.vcf
```

Hard filter

In order to filter away false positives a hard filter was applied which first filtered away all indels or SNPs found within 5bp of each other. This should remove SNPs found in regions problematic to align. Secondly homozygous SNPs were found by using the DP4-depths to find SNPs with more than 80% of the reads showing change and the sum of reference forward, reference reverse, allele

[42] http://samtools.sourceforge.net/samtools.shtml
[43] http://broadinstitute.github.io/picard/
[44] http://broadinstitute.github.io/picard/
[45] http://samtools.sourceforge.net/mpileup.shtml
[46] https://samtools.github.io/bcftools/bcftools.html
[47] http://vcftools.sourceforge.net/

forward and allele reverse being higher than the standard deviation from the haploid peak: median were in the case of DXB11 33 => ¼ x 33 = 8.25 and ¾ x 33 = 24.75. 8 were chosen as minimum for a homozygous SNP and 25 as minimum for a heterologous SNP. This was validated as realistic cutoffs by comparing to the actual depth distribution (Figure 1[48])

```
cp DXB_uCgris.vcf name.vcf
grep "^#" name.vcf > header
grep -v "^#" name.vcf > file1
cat file1 | gawk '{print $2}' > file2.1
cat file2.1 | sed 1d > file2.down
echo "1" >> file2.down
echo "1" > file2.up
head -n -1 file2.1 >> file2.up
paste file2.1 file2.up | awk '{a=$2-$1; print a;}' | awk ' { if($1>=0) { print $1} else {print $1*-1 }}' > file2.diff1
paste file2.1 file2.down | awk '{a=$2-$1; print a;}' | awk ' { if($1>=0) { print $1} else {print $1*-1 }}' > file2.diff2
paste file1 file2.diff1 file2.diff2 > file2
awk '$11 > 5' file2 | awk '$12 > 5' > file3
#Only SNPS are looked at right now
grep -v "INDEL" file3 | cut -f1-10 > file4
cat file4 | gawk '{print $8}' | sed 's/^.*DP4=//g' | cut -f 1,1 -d';' | sed 's/,/\t/g' > file5
cat file5 | awk '{a=$3+$4; b=$1+$2+$3+$4; c=((a/b)); print a,b,c}' | sed 's/ /\t/g'> file6
paste  file4 file6 | awk '$12 >= 8' | awk '$6 > 30' | awk '$13 > 0.80' > file7.homo
paste  file4 file6 | awk '$12 >= 24' | awk '$6 > 30' > file7.hetero
cat file7* | sort | uniq  > file8
grep "^#" name.vcf > header
cat file8 | cut -f1-10 >> header
cp header DXB_SNP_uCgris.vcf
#Only indels are looked at right now
grep "INDEL" file3 | cut -f1-10 > file4
cat file4 | gawk '{print $8}' | sed 's/^.*DP4=//g' | cut -f 1,1 -d';' | sed 's/,/\t/g' > file5
cat file5 | awk '{a=$3+$4; b=$1+$2+$3+$4; c=((a/b)); print a,b,c}' | sed 's/ /\t/g'> file6
paste  file4 file6 | awk '$12 >= 8' | awk '$6 > 30' | awk '$13 > 0.80' > file7.homo
paste  file4 file6 | awk '$12 >= 24' | awk '$6 > 30' > file7.hetero
cat file7* | sort | uniq  > file8
grep "^#" name.vcf > header
cat file8 | cut -f1-10 >> header
cp header DXB_indel_uCgris.vcf
```

Finding functional consequences

The vcf-files from above with SNPs and indels were imported into CLC Genomic workbench (version 7.0.4) and functional consequences were found using the "Amino acid changes" function after importing the Cgriseus.gff3 annotation file as tracks. The output from the program was exported as csv files.

The csv files were reorganized

The Indel file contained an extra column so each output was treated differently. All lines containing "SNPs" which were identical to the reference were removed

Indel output:

[48] http://www.biomedcentral.com/1471-2164/16/160

cat K1_indel_uCgris_AAC.csv | sed 's/""/"-"/g' | sed 's/","/\t/g' | sed 's/,/YYY/g' | sed 's/"//g' | gawk '{print $6,$1,$2,$3,$4,$5,$7,$9,$20,$24,$38,$33,$34}' | sed 's/ /,/g' | awk -F, '$1=="No" {print $0}' | sed 's/,/\t/g' | cut -f2-14 | sed 's/YYY/,/g' > K1_indel.txt

Table III. Example output from indels found in the CHO-K1 ATCC genome compared to *the C. griseus* genome

Chromosome	Pos	Type	Ref	Allele	Length	Type	Depth	DP4	Conceq uence	Gene concequence	Protein concequence
NW_006868235.1	2566^2567	Insertion	-	AG	2	Heterozygous	44	16,10,7,6	Yes	XP_007606577.1:c.19_20insCT	XP_007606577.1:p.Pro7fs
NW_006868510.1	286875^286876	Insertion	-	C	1	Homozygous	30	0,0,12,14	Yes	XP_007606782.1:c.146_147insG	XP_007606782.1:p.Lys49fs

SNP output:

cat K1_SNP_uCgris_AAC.csv | sed 's/""/"-"/g' | sed 's/","/\t/g' | sed 's/; /;/g' | sed 's/,/YYY/g' | sed 's/"//g' | gawk '{print $6,$1,$2,$3,$4,$5,$7,$9,$19,$23,$37,$32,$33}' | sed 's/ /,/g' | awk -F, '$1=="No" {print $0}' | sed 's/,/\t/g' | cut -f2-14 | sed 's/YYY/,/g' > K1_SNP.txt

Table IV Example output from SNPs found in the CHO-K1 ATCC genome compared to *the C. griseus* genome

Chromosome	Pos	Type	Ref	Allele	Length	Type	Depth	DP4	Concequence	Gene concequence	Protein concequence
NW_006865609.1	312	SNV	A	C	1	Homozygous	25	0,0,7,13	Yes	XP_007606494.1:c.638T>G	XP_007606494.1:p.Ile213Ser
NW_006867217.1	913	SNV	C	G	1	Heterozygous	48	9,9,14,12	Yes	XP_007606534.1:c.388G>C	XP_007606534.1:p.Gly130Arg

Filtering the output based on genes

The output was then filtered based on each gene making the output more easily searchable

```
echo "K1ATCC" > name
name=`awk '{print $1 }' < name`
wc -l ~/Cgenomes/CHOK1_version1.1/listofgenes2 > length
length=`awk '{print $1 }' < length`
for ((  i = 22055;  i <= $length ; i++  ))
do
cat ~/Cgenomes/CHOK1_version1.1/listofgenes2 | head -$i | tail -1 | gawk '{print $3}' > protein
protein=`awk '{print $1 }' < protein`
grep -w "$protein" database > hits
wc -l hits > length2
length2=`awk '{print $1 }' < length2`
for ((  j = 1 ;  j <= $length2 ; j++  ))
do
cat hits | head -$j | tail -1 > subset
scaffold=`awk '{print $1 }' < subset`
pos=`awk '{print $2 }' < subset`
ref=`awk '{print $4 }' < subset`
allele=`awk '{print $5 }' < subset`
echo "$scaffold.$pos$ref>$allele" > mutation
```

```
gawk '{print $3}' subset  > type
gawk '{print $6}' subset  > length
gawk '{print $7}' subset  > zygosity
gawk '{print $8}' subset  > depth
gawk '{print $9}' subset  > DP4
gawk '{print $10}' subset  > CDSconcequence
cat subset | sed "s/^.*$protein:c/$protein/" | gawk '{print $1}' | cut -f 1,1 -d';' > coding
grep "$protein:p" subset | sed "s/^.*$protein:p/$protein/" | gawk '{print $1}' | cut -f 1,1 -d';' > proteinmut
paste name type length zygosity depth DP4 CDSconcequence mutation coding proteinmut >> SNPs
done
grep -w "$protein" SNPs > overview
wc -l overview | gawk '{print $1}' > rnumber
gawk '{print $5}' overview | grep "-" | wc -l > intronnumber
gawk '{print $5}' overview | grep "No" | wc -l > exon_nochange
gawk '{print $5}' overview | grep "Yes" | wc -l > AAchange
#paste exon_nochange AAchange | awk '{a=$2+$1;print a;}' > exonnumber
paste protein name rnumber intronnumber exon_nochange AAchange >> Proteinoverview
done
cat SNPs | tr '[' ' ' | tr ']' ' ' | sed 's/.1. /.1./g' > SNP
```

The first couple of SNPs rearranged based on the code above (see full table in Supplementary table 4 in recent paper[49])

```
csrk@davinci:~/SNPs> head SNPs
K1ATCC  Heterozygous    51      10,11,20,9      -       NW_006880426.1.151267G>A        XP_007626087.1.868-878C>T
K1ATCC  Heterozygous    43      13,8,9,11       -       NW_006880426.1.163920G>A        XP_007626087.1.604+140C>T
K1ATCC  Heterozygous    44      12,9,7,13       -       NW_006880426.1.172161A>C        XP_007626087.1.234+603T>G
K1ATCC  Homozygous      44      0,0,24,17       -       NW_006870833.1.281202G>A        XP_007609693.1.4330+263C>T
K1ATCC  Homozygous      41      0,0,20,20       -       NW_006870833.1.281584T>G        XP_007609693.1.4228-17A>C
K1ATCC  Homozygous      50      0,0,31,16       -       NW_006870833.1.282030G>A        XP_007609693.1.4228-463C>T
K1ATCC  Homozygous      45      0,0,29,13       -       NW_006870833.1.282133G>A        XP_007609693.1.4228-566C>T
K1ATCC  Homozygous      58      0,0,22,33       -       NW_006870833.1.282211A>T        XP_007609693.1.4228-644T>A
K1ATCC  Homozygous      52      0,0,18,33       -       NW_006870833.1.282277A>G        XP_007609693.1.4228-710T>C
```

The overview generated for each gene in the genome (see full table in Supplementary table 2 in recent paper[50])

```
csrk@davinci:~/SNPs> cat Proteinoverview
XP_007626087.1  K1ATCC  3       3       0       0
XP_007609693.1  K1ATCC  128     115     6       7
XP_007613122.1  K1ATCC  10      9       0       1
XP_007627973.1  K1ATCC  1       0       0       1
XP_007627974.1  K1ATCC  1       0       0       1
XP_007632240.1  K1ATCC  11      7       3       1
XP_007637716.1  K1ATCC  3       2       1       0
XP_007637718.1  K1ATCC  3       2       1       0
XP_007623137.1  K1ATCC  6       6       0       0
```

Making an output of the changed genes

In order to obtain a fasta file with a genome as that of the reference, but containing the filtered indels and SNPs the method below can be used to create a new pseudo-genome

[49] http://www.biomedcentral.com/1471-2164/16/160
[50] http://www.biomedcentral.com/1471-2164/16/160

```
./vcf-sort DXB11.vcf > DXB11sorted.vcf
bgzip DXB11sorted.vcf
tabix -p vcf DXB11sorted.vcf.gz
cat CHOK1_ver1.1.fasta | ./vcf-consensus DXB11sorted.vcf.gz > testout.fa
```

Frameshifting mutations

In order to get lists of the genes having homozygous frameshift mutations in CHO DXB11 and K1:

```
cat DXB_indel.txt | grep "Homozygous" | gawk '{print $12}' | grep "fs" | sed 's/:p./\t/g' > DXBfs
cat K1_indel.txt | grep "Homozygous" | gawk '{print $12}' | grep "fs" | sed 's/:p./\t/g' > K1fs
sdiff DXBfs K1fs | grep "<" | gawk '{print $1, $2}' > DXBfrunique
sdiff DXBfs K1fs | grep "|" | gawk '{print $1, $2}' >> DXBfrunique

wc -l DXBfrunique > length
length=`awk '{print $1 }' < length`
for ((  i = 1 ;  i <= $length ; i++  ))
do
cat DXBfrunique | head -$i | tail -1 > protein
protein=`awk '{print $1 }' < protein`
grep -w "$protein" ~/Cgenomes/CHOK1_version1.1/listofgenes2 > hits
paste protein hits >> frameshiftDXB11
done
```

10.3.5 Extracting differentially expressed genes from an RNA sequencing experiment

In this section a pipeline is presented for obtaining a list of differentially expressed genes from raw reads obtained from RNA sequencing (RNAseq) data

Pre-alignment trimming of reads

Trim Galore is a usefull tool, which are able to trim the paired reads, while keeping the files synchronized so the pairs are broken[51]. It can be downloaded from here[52] and the required packagages cutadapt from here[53] and fastqc from here[54]. In the code below all low quality bps are trimmed from the ends until bps with a quality >30 are found, but all reads that after trimming are shorter than 40bp are removed from the dataset. See full decription of options here[55]

```
fastqc DXB11_1.fastq
fastqc DXB11_2.fastq
mkdir ./trimmed_reads
~/software/Trimgalore/trim_galore --paired -q 30 --length 40 --fastqc --path_to_cutadapt /novo/appl/ngs/cutadapt-1.2.1/bin/cutadapt --stringency 5 -o ./trimmed_reads DXB11_1.fastq DXB11_2.fastq
```

[51] http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
[52] http://www.bioinformatics.babraham.ac.uk/projects/download.html#trim_galore
[53] https://pypi.python.org/pypi/cutadapt/
[54] http://www.bioinformatics.babraham.ac.uk/projects/download.html#fastqc
[55] http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/trim_galore_User_Guide_v0.4.0.pdf

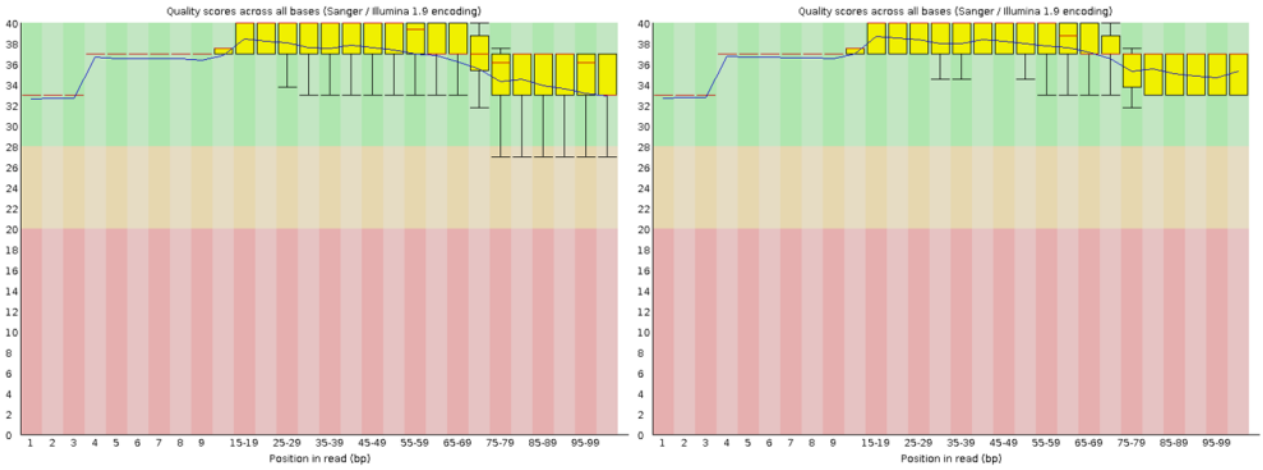


Figure 27 quality score across the reads before (left) and after (right) trimming.

Alignment to the genome

The alignment of RNAseq reads to the genome is slightly different from genomic data as the RNAseq data have had the introns removed and this are only expected to align to exonic sequence.

The genome was indexed using the bowtie-algorithm so it would work with tophat

bowtie-build Cgriseus.fasta Cgriseus

The index was tested by looking for a specific sequence only found in the LRP1-gene

bowtie -c Cgriseus tatccaggggtccagaaggtgcactttggactggctgacg

0 - NW_003614353.1 616991 CGTCAGCCAGTCCAAAGTGCACCTTCTGGACCCCTGGATA
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII 0
 # reads processed: 1
 # reads with at least one reported alignment: 1 (100.00%)
 # reads that failed to align: 0 (0.00%)
Reported 1 alignments to 1 output stream(s)

This output shows that the genome index was successfully searched and only a single hit was found as expected.

The reads are aligned as below

tophat -o F3901 #output directory
-p 32 #number of threads to run on the supercomputer
--segment-mismatches 3 #number of mismatches that is tolerated for each read
--segment-length 25 #the read can be split in two (specie sites) but min 25bp on each side
--mate-inner-dist 80 #distance between the two paired reads (consult sequencing company)
--mate-std-dev 50 #the standard dev of the above distance so 80+50 is OK
--transcriptome-index /F3901/transcriptome_idx #generates a transcriptome file that can be used with other samples with the same genome
-G ./Cgriseus.gtf #the genome annotation that is used to lead the assembly.
./bowtie_index/Cgriseus #where to look for the bowtie-indexed genome
F390_1_comm_1.fq #input fastq file 1
F390_1_comm_2.fq #input fastq file 2

Count the number of reads aligning to each gene

Expression values will be deduced using htseq-count[56], which counts the number of reads aligning to the exonic sequence.

First the accepted_hits.bam file made by tophat is sorted using samtools and subsequently converted into a sam file

```
samtools sort -n accepted_hits.bam accepted_hits.nsort
samtools view accepted_hits.nsort.bam > accepted_hits.nsort.sam
```

Then the accepted_hits.nsort.sam file is used to look for exonic sequence based on the position in the annotation file.

```
htseq-count --stranded=no --idattr=gene_name --samout=htseq-cout.accepted_hits.sam --mode=intersection-nonempty accepted_hits.nsort.sam ./feature/Cgriseus.gtf > 01.1.htseq.txt
```

The 01.1.htseq.txt looks like this

```
A1cf        0
A2ml1       0
A3galt2     0
A4galt      0
A4gnt       0
Aaas        1247
Aacs        2001
…
```

Import into R and normalization of the data

In R the datafiles above are imported and technical replicates sequenced on different lanes but from the same library are simply added together (as e.g. T1+S1) in case to generate a table of 32 columns and 26632 rows.

```
T1=as.matrix(read.table("01.1.htseq.txt",header = TRUE, row.names=1))
T2=as.matrix(read.table("02.1.htseq.txt",header = TRUE, row.names=1))
…
col=cbind(T1+S1,T2,T3+S3,T4,T5,T6,T7+S7,T8+S8,T9+S9,T10+S10,T11+S11,T12+S12,T13+S13,T14+S14,T15+S15,T16+S16,T17,T18+S18,T19+S19,T20+S20,T21,T22,T23+S23,T24,T25,T26+S26,T27,T28,T29,T30+S30,T31,T32)
colnames(col)=c(1.1, 1.2, 1.3, 2.1, 2.2, 2.3, 3.1, 3.2, 3.3, 4, 5, 6.1, 6.2, 6.3, 7.1, 7.2, 7.3, 8.1, 8.2, 8.3, 9, 10, 11, 12.1, 12.2, 12.3, 13.1, 13.2, 13.3, 14.1, 14.2, 14.3)
```

Based on the review by Dillies et al[57] the samples were normalized based on the algorithm used in EdgeR[58]. First the group is define, which is the samples are actually biological replicates. In this rame column 1-3 is sample1, 4-6 is sample 2…

[56] http://www-huber.embl.de/users/anders/HTSeq/doc/count.html
[57] http://www.ncbi.nlm.nih.gov/pubmed/22988256
[58] http://www.bioconductor.org/packages/release/bioc/html/edgeR.html

```
group <- factor(c(rep.int(1, 3),rep.int(2, 3),rep.int(3, 3),4,5,rep.int(6, 3),rep.int(7, 3),rep.int(8,
3),9,10,11,rep.int(12, 3),rep.int(13, 3),rep.int(14,  3)))
```

Subsequently the samples are normalized

```
.libPaths(c("C:\\R\\Rpackages", .libPaths())) #if you want to install the
source("http://bioconductor.org/biocLite.R");biocLite("edgeR")
library(edgeR)
x <- DGEList(counts=col,group=group)
plot(aveLogCPM(x, normalized.lib.sizes=TRUE, prior.count=2, dispersion=0.05))
y <- calcNormFactors(x)
z <- estimateCommonDisp(y) #overall dispersion of the dataset (inter-library variability)
w <- estimateTagwiseDisp(z) #dispersion per tag (gene),
```

Multidimensional scaling is used to the show the overall variation between the samples.

```
plotMDS(y, method="bcv")
```

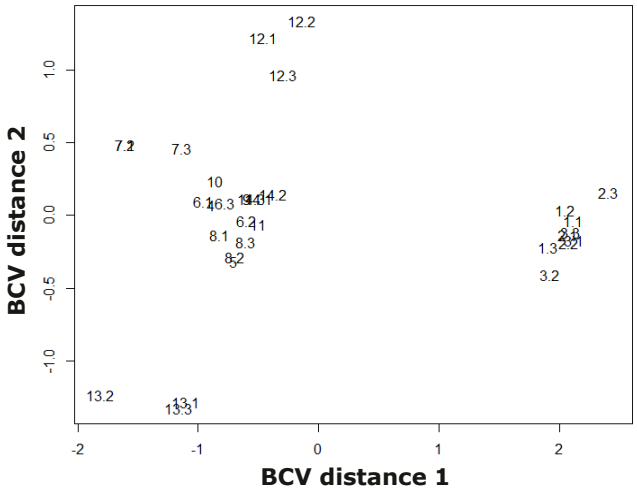


Figure 28 Multidimensional scaling of the 32 RNAseq samples

From supplementary table 6 from recent paper[59] the length of the coding region for each gene is given.

```
A1cf        593
A2ml1       1456
A3galt2     337
A4galt      348
A4galt      348
…
```

Due to alternative spleicing the same CHO gene identifier is listed with different lengths but as htseq counts for the entire gene the longest will be picked. The list (called names.txt) is imported into R and used for normalization for each gene in order to arrieve to arrive at Fragments Per Kilobase of transcript per Million mapped reads (FPKM) values instead of CPM (Counts per

[59] http://www.biomedcentral.com/1471-2164/16/160

million), which is normalized between samples but nor normalized for the length of the transcript (see EdgeR UsersGuide[60]).

```
total=cpm(w, normalized.lib.size=TRUE)
names=read.table("names.txt")
total2=total
for(i in 1:nrow(total))
{
length=max(names[which(names[,1]==rownames(total)[i]),2])
total2[i,]=total[i,]/(length*3/1000)
}
```

Identifying differentially expressed genes

From the dataset seen above the difference between sample 1 compared to a control (here sample 12) is used for calculating the p-values and fold-changes between samples

```
et <- exactTest(w, pair=c(12,1))
write.table(topTags(et,n=25029),file="pvals.csv");pvals = read.csv("pvals.csv",sep=" ")
head(pvals)
```

	logFC	logCPM	PValue	FDR
Acadm	11.007199	4.993089	0	0
Apobr	10.844896	4.650211	0	0
Zmat4	8.606725	3.640402	0	0
LOC100765542	8.238092	3.230184	0	0
LOC100769145	7.445602	3.099561	0	0
LOC103161684	6.970651	5.306510	0	0

Figure 29 All genes in the genome with the log2(fold-change) for sample 1 vs sample 12, the log2(CPM) and the probability that the difference is random

The differentially expressed genes can be isolated from the matrix by assigning cut-off such as fold-change higher than ±2, expression level above 1 and p-value < 0.01. This is plotted below

```
DE1a=pvals[which(pvals[,1]>log2(2)),]#fold-change
DE1b=pvals[which(pvals[,1]<log2(1/2)),]#fold-change
DE1=rbind(DE1a,DE1b)
DE2=DE1[which(DE1[,2]>log2(1)),]
DE3=DE2[which(DE2[,3]<0.01),]

jpeg("Degenes.jpg", width=10, height=10, units="in", res=500)
detags <- rownames(DE3)
plotSmear(et, de.tags=detags)
dev.off()
```

The list of genes are exported to a tabular seperated file

```
write.table(DE3,file="FPKM.txt", sep="\t")
```

[60] http://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf

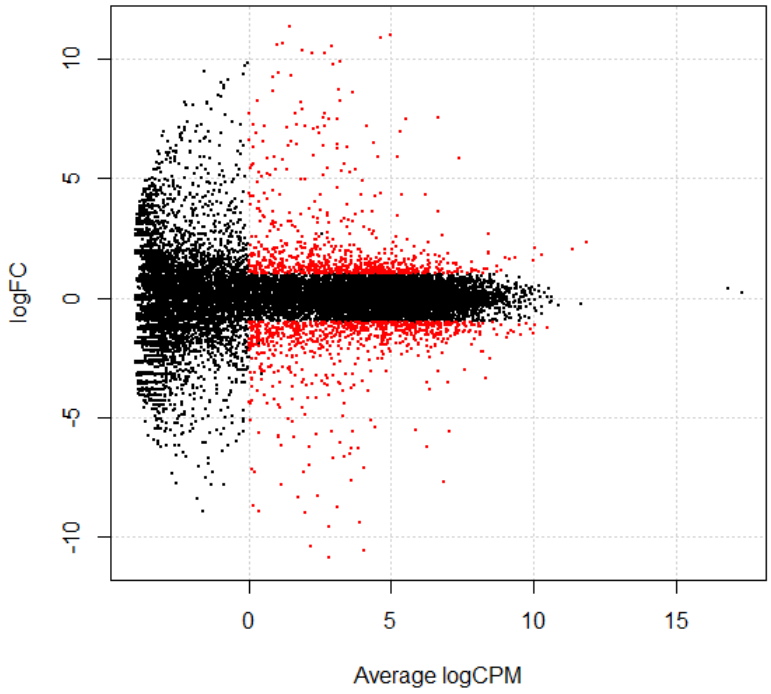


Figure 30 visulization of the differentially expressed genes

10.3.6 Analyzing miRNA sequencing data

This section contains a step-by-step guide for data analysis of miRNA sequencing data for identification of differentially expressed miRNAs.

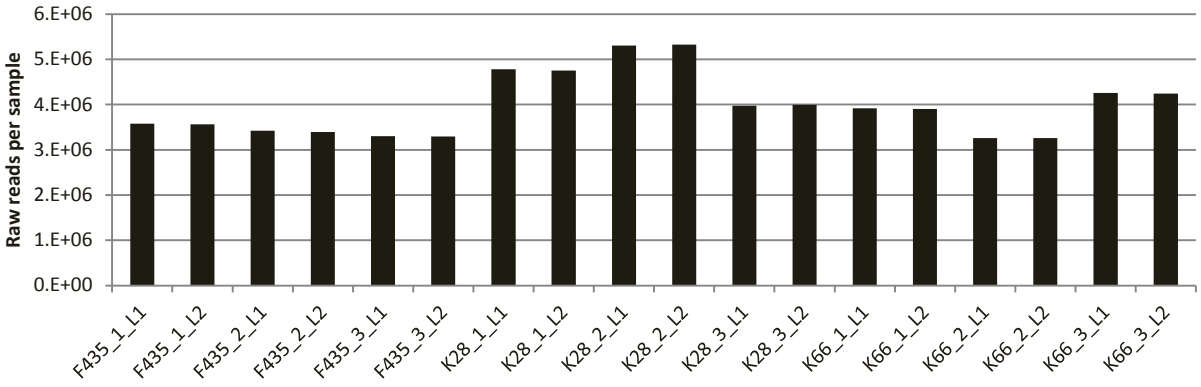
Extraction of the miRNAs

RNA was extracted from $2x10^6$ cells using TRIzol (Invitrogen) and the RNeasy Cleanup kit (Qiagen) following the manufacturer's instructions. RNA integrity was confirmed on an Agilent 2100 Bioanalyzer using total RNA nano chips (Agilent Technologies, Santa Clara, Ca, USA). RNA concentration was measured using a NanoDrop spectrophotometer (NanoDrop Technologies). Multiplexed cDNA library generation and next-generation sequencing were performed by AROS Applied Biotechnology (Aarhus, Denmark) pooling the 9 miRNA libraries together with two CHO genomes on two lanes in an Illumina Hiseq 2000.

Identification of adapter

Between 6e6 and 10e7 reads per miRNAseq library were sequenced with ~50bp per read. Due to the short nature of a mature miRNA ~30bp of the read will thus be standardized adapter found

flanking the miRNA as a result of the library preparation. The reads were analysed using fastqc which revealed that RNA PCR Primer, Index 1 were overrepresented in the dataset.



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GGAATACCGGGTGCTGTAGGCTTTTGGAATTCTCGGGTGCCAAGGAACTC	883171	24.700539475034773	RNA PCR Primer, Index 1 (100% over 26bp)
GGAATACCGGGTGCTGTAGGCTTTGGAATTCTCGGGTGCCAAGGAACTCC	208275	5.825038253252051	RNA PCR Primer, Index 1 (100% over 27bp)
AAGCTGCCAGTTGAAGAACTGTTGGAATTCTCGGGTGCCAAGGAACTCCA	156951	4.38960786885686	RNA PCR Primer, Index 1 (100% over 28bp)
GGAATACCGGGTGCTGTAGGCTTGTGGAATTCTCGGGTGCCAAGGAACTC	128904	3.605188961695846	RNA PCR Primer, Index 1 (100% over 26bp)
TACCCTGTAGAACCGAATTTGTTGGAATTCTCGGGTGCCAAGGAACTCCA	80730	2.2578578234787567	RNA PCR Primer, Index 1 (100% over 28bp)

Figure 31 top: Number of sequenced reads from each sample and sequencing lane. Bottom: Output from fastqc showing the overrepresentation of RNA PCR primer, index 1.

Online a list of the full primer sequence was found[61], which showed the sequence for that primer, which was:
CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA

Extraction and counting miRNAs in CLC genomic workbench version 7.5

The general tutorial for smallRNAseq analysis using CLC genomic workbench was followed[62][63].

- The raw reads were imported (import | Illumina)
- New | Trim adapter list | Add row | Insert name and sequence of RNA PCR primer, index 1 | Action = Discard when not found
- Toolbox | Transcriptomics Analysis | Small RNA Analysis | Extract and Count

[61] http://omicsoft.com/downloads/ngs/contamination_list/v1.txt
[62] http://www.clcbio.com/files/tutorials/Small_RNA_analysis_Illumina.pdf
[63] http://www.clcsupport.com/clcgenomicsworkbench/650/index.php?manual=Adapter_trimming.html

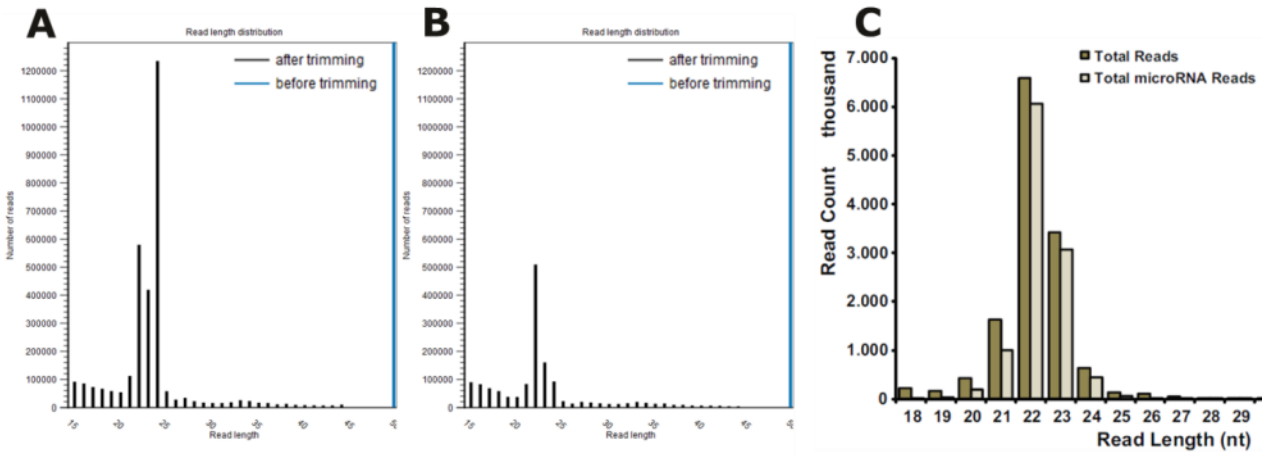


Figure 32 Left: Distribution of miRNAseq length after filtering. Middle: Distribution of miRNAseq length after filtering with inclusion of the rRNA contamination sequence in the filter. Right: Distribution of miRNAseq length as shown in[64]

From the initial filter with only RNA PCR primer, index 1 the distribution was completely skewed to "miRNAs" with a length of 24bp and not 22bp as expected (see Figure XX.C). From the count data it was seen that the vast majority of 24bp fragments were a single species of GGAATACCGGGTGCTGTAGGCTTT. This sequence was blasted and found to be 100% identical with Homo sapiens RNA, 5S ribosomal pseudogene 500 (NG_033696.1). This sequence must somehow have contaminated the library during preperation enriching for miRNAs. This 24bp sequence was added to the filter as with the RNA PCR primer (but with Action = Discard when found). And following trimming a distribution appeared centred around 22bp as expected.

Identifying miRBase miRNAs using CLC genomic workbench version 7.5

- Toolbox | Transcriptomics Analysis | Small RNA Analysis | Download miRBase
- Toolbox | Transcriptomics Analysis | Small RNA Analysis | Annotate and Merge Counts
  - Click: use miRBase, select downloaded miRBase from above.
  - From miRBase species: *Cricetulus griseus*, *Rattus norvegicus*, *Homo sapiens* and *Mus musculus* were chosen
  - Standard settings were used regarding mature length variants

A table is generated showing the expression value with the name

Feature ID	Expression values	Name	Resource
let-7a-1 (Homo sapiens)	0	let-7a-1	Homo sapiens
let-7a-2 (Homo sapiens)	4	let-7a-2	Homo sapiens

…

[64] http://www.ncbi.nlm.nih.gov/pubmed/21392545

The tables were exported as csv files and merged into one table in excel with 748 miRNAs identified across the 18 datasets. Distribution of miRNAs identified in each sample shown in table below

	miRNA			miRNA			miRNA
F435_1_1	499		K28.1.1	553		K66.1.1	507
F435_1_2	501		K28.1.2	555		K66.1.2	515
F435_2_1	477		K28.2.1	537		K66.2.1	615
F435_2_2	477		K28.2.2	539		K66.2.2	629
F435_3_1	515		K28.3.1	491		K66.3.1	542
F435_3_2	520		K28.3.2	524		K66.3.2	536

Importing data into R and normalizing

The merged table was inserted into a notepad file and all #N/A's were replaced with 0's. The file was imported into R and the reads from the sample but sequenced on different sequencing lanes were pooled into one column.

```
list = read.table("CountmiRNA.txt",header = TRUE)
rownames(list)=list[,1]
col=list[,2:19]
coll=cbind(list[,2]+list[,3],list[,4]+list[,5],list[,6]+list[,7],list[,8]+list[,9],list[,10]+list[,11],list[,12]+list[,13],list[,14]
+list[,15],list[,16]+list[,17],list[,18]+list[,19])
colnames(coll)=c("F445_1","F445_2","F445_3","K28_1","K28_2","K28_3","K66_1","K66_2","K66_3")
rownames(coll)=rownames(col)
write.table(coll[1:4,],file="example_table.txt",sep="\t")
```

	F445_1	F445_2	F445_3	K28_1	K28_2	K28_3	K66_1	K66_2	K66_3
let-7a-1//let-7a-3_hs	13	8	17	21	18	12	17	10	17
let-7a-1//let-7c-2_mm	374	269	367	363	250	201	169	195	216
let-7a-2//let-7a_mm//cg	0	0	0	0	0	0	3513	0	0
let-7a-2//let-7a_mm//rn//cg	10589	19021	0	18618	0	12229	3544	5291	0

EdgeR was loaded, three groups were defined with triplicates in each and the data was normalized.

```
.libPaths(c("C:\\R\\Rpackages", .libPaths()))
source("http://bioconductor.org/biocLite.R")
biocLite("edgeR")
library(edgeR)
group <- factor(c(rep.int(1, 3),rep.int(2, 3),rep.int(3, 3)))
x <- DGEList(counts=coll,group=group)
y <- calcNormFactors(x)
z <- estimateCommonDisp(y) #overall dispersion of the dataset (inter-library variability)
w <- estimateTagwiseDisp(z) #dispersion per tag (gene)
et <- exactTest(w)
total=cpm(w, normalized.lib.size=TRUE)
write.table(total[1:4,],file="example_table2.txt",sep="\t")
```

	F445_1	F445_2	F445_3	K28_1	K28_2	K28_3	K66_1	K66_2	K66_3
let-7a-1//let-7a-3_hs	18,9	13,3	20,0	18,6	17,6	16,7	24,5	12,3	19,4
let-7a-1//let-7c-2_mm	543,7	447,6	432,5	321,5	245,0	279,0	243,8	239,5	245,9
let-7a-2//let-7a_mm//cg	0,0	0,0	0,0	0,0	0,0	0,0	5.067,1	0,0	0,0
let-7a-2//let-7a_mm//rn//cg	15393,1	31648,7	0,0	16488,1	0,0	16972,9	5111,9	6498,3	0,0

Finding differentially expressed miRNAs

```
et <- exactTest(w, pair=c(1,3))
topTags(et,n=20)
detags <- rownames(topTags(et)$table)
cpm(w)[detags, order(w$samples$group)]  #extract de differentially expressed
summary(de <- decideTestsDGE(et, p=0.05)) #gener der er op og ned
detags <- rownames(w)[as.logical(de)]
png("Figure1.png", width = 20, height = 15, units = 'cm', res = 300)
plotSmear(et, de.tags=detags, cex =1) #plot with DE genes
abline(h = c(-2, 2), col = "blue")
dev.off()
```

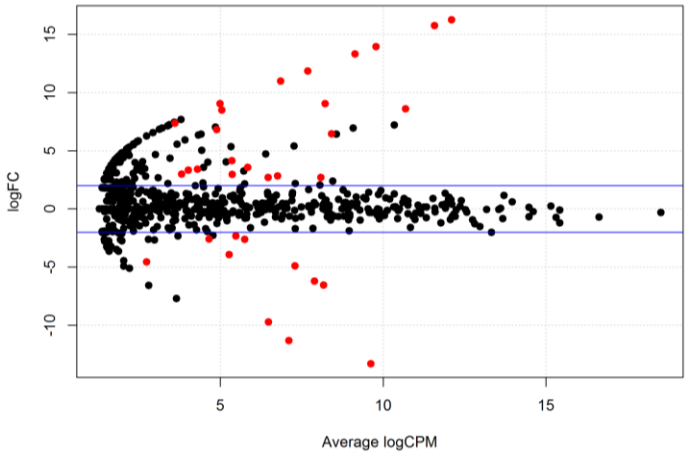


Figure 33 Visalization of the 11 miRNAs which are differentially downregulated, the 22 miRNAS which are differentially upregulated and the 715 miRNAs which are similar in F435 compared to K66.

Finding miRNAs correlating to FVIII productivity

The relative productivities of FVIII from the experiment where the total RNA was purified was imported and subsequently the information was used to calculate the spearman correlation between the expression of each miRNA and the FVIII productivity. This revealed 32 miRNAs, which correlated positively with productivity and 37 miRNAs correlating negatively with the FVIII productivity.

```
prod=t(meta[72,]/meta[63,])/max(t(meta[72,]/meta[63,]))
prodsample=c(prod[1:3],prod[15:17],prod[27:29])
```

F445_1	F445_2	F445_3	K28_1	K28_2	K28_3	K66_1	K66_2	K66_3
100%	87%	98%	20%	19%	20%	0%	0%	0%

```
cortotal=1;for(i in 1:nrow(total))
{cortotal[i]=cor(total[i,],prodsample, method ="spearman")}
cor1a=total[which(cortotal>0.80),]; nrow(cor1a)
cor1b=total[which(cortotal<(-0.80)),]; nrow(cor1b)
```

In order to filter the results even further it was investigated whether there was a significant difference in expression between the three clones. After this filter there were 8 miRNAs, which correlated positively with productivity and 12 miRNAs correlating negatively with the productivity.

```
cortotal2=cortotal3=1
```

```
for(i in 1:nrow(cor1a))
{cortotal2[i]=t.test(cor1a[i,1:3],cor1a[i,4:6])$p.value
cortotal3[i]=t.test(cor1a[i,4:6],cor1a[i,7:9])$p.value}
cor2a=cbind(cor1b,cortotal2,cortotal3)
cor3a=cor2a[which(cor2a[,10]<0.05 & cor2a[,11]<0.05),]

p2 <- function(name)
{plot(prodsample,total[name,],ylim=c(0,max(total[name,])),ylab=name,xlab="Productivity",pch=16)}
p2("mir-34b_cg")
```

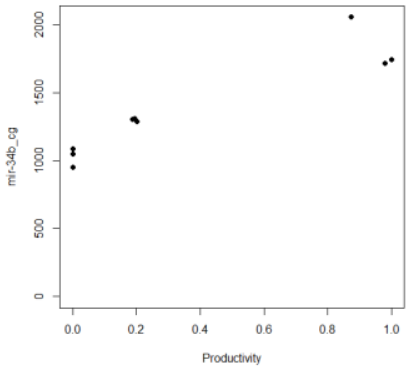


Figure 34 The expression of mir-34b_cg correlating positively with the productivity.

Genome position of the miRNA

Using blast the genomic position of the miRNAs can be found. In my particular case the target miRNAs chosen are all either found in introns of genes, which appear to be silent or in intergenic space. A lot could have been interpreted from the results if the miRNA was found to be excised from the intron of a gene with a known function.

10.3.7 Identification of potential targets for Crispr/Cas9 induced knock outs

The aim of this section was to generate a method for locating potential Crispr/Cas9 sites for knock-out of genes in *C. griseus* genome

The idea in short:

1. Create list of the target regions (exon 1-6 of all genes in the genome)
2. Make a list of all 16bp combinations (sense and antisense), which have the NGG motif
3. Create a *C. griseus* transcriptome list containing the transcribed regions of the genome
4. Search the *C. griseus* genome / transcriptome for off-target binding of each potential Crispr/Cas9 target site in order to generate an alignment score for each target

Extraction of coding region from all genes in the genome

From the *C. griseus* annotation file a list of all genes was extracted previously. One gene at the time the list of exons were extracted from the gff3 annotation file and based on the orientation of the

gene exons one was taken from end or beginning. These coordinates were used to extract the region from the genome and export it as a fasta file.

```
cat ./listofgenes | gawk '{print $3, $1, $2}' | sed 's/ /\t/g' | uniq > Genelist
wc -l Genelist > ./length
length=`awk '{print $1 }' < ./length`
for ((  i = 1;  i <= $length ; i++  ))
do
cat Genelist | head -$i | tail -1 > ./geneinfo
cat ./geneinfo | gawk '{print $1}' > ./genename
genesearch_CDS=`awk '{print $1 }' < ./geneinfo`
CHOname=`awk '{print $2 }' < ./geneinfo`
genesearch_mRNA=`awk '{print $3 }' < ./geneinfo`
grep -w "$genesearch_CDS" ../annotation.gff3 > stuff
cat stuff | sed 's/\t/,/g' | awk -F, '$3=="CDS" {gsub(","," ");print $0}' | gawk '{print $1, $4, $5}' | sed 's/ /\t/g' >
exon
 length2=`wc -l exon | awk '{print $1 }'`
 for ((  j = 1;  j <= $length2 ; j++  ))
 do
 cat exon | head -$j | tail -1 > exon.bed
 /novo/appl/ngs/BEDTools-Version-2.16.2/fastaFromBed -fi ../Cgriseus.fasta -bed exon.bed -fo ./output.fasta
 echo ">$CHOname-$genesearch_mRNA-exon$j" > $CHOname-$genesearch_mRNA-$j.fasta
 sed 1d ./output.fasta >> $CHOname-$genesearch_mRNA-$j.fasta
 done
done
```

An example is investigated

```
cat A1cf-XM_007627897.1-1.fasta
>A1cf-XM_007627897.1-exon1
TGGACCAAGCTATATCCTGTGCGCTGGACCAGTGCGCGGAGGGCTGCTTCCTTCTGGGTGCCGCTCAATCCATCCCCGGATTTGTG
ATTTGATTCCAT
```

All the files are sorted into folders

```
#organize into folders
for ((  i = 1;  i <= 10 ; i++  ))
do
mkdir exon$i
mv *.$i.fasta ./exon$i/
done
mkdir exon_rest
mv *.fasta ./exon_rest
```

Find all potential target sites

As an example exon one from Ccl24 will be used.

```
>Ccl24
CCCCCACAGCTTCATCACCAAGAGGGGCCATAAATTCTGTGGTGACCCCAAGTTGCCGTGGGTCCAGAGGCACATACGGAACCTG
GATGCCAAGAGAAAGCAGCCTTCCGCAGGTGCCAAGGCACTGGGCAAGTTTTCCATCCAGAGACACCATGGCAACAGCACTGGGG
TTTAATTGCCTCTCC
```

The length of the exon is found and a running window is made of all 33bp ranges (in this case 153). A 33bp window was chosen in order to know a bit about the base pairs flanking the 16bp target site. All of these results are reverse complimented to make sure targets on both strands are found. Finally only targets which has a “GG” in position at position 15 and 16 of the final guide RNA are included ending up with 29 targets in this exon

```
length=`sed 1d $file | wc -m`
```

```
for ((  i = 0 ;  i <= $length ; i++  ))
do
sed 1d $file | sed "s/^.\{$i\}//" | cut -c 1-33 >> sample.txt
done
head --lines=-34 sample.txt > sample2.txt
perl -0777ne's/\n //g; tr/ATGCatgcNn/TACGtacgNn/; print scalar reverse $_;' sample2.txt | sed 1d >> sample2.txt
grep ".GG.....$" sample2.txt > finalcandidates

TCACCAAGAGGGGCCATAAATTCTGTGGTGACC
ATTCTGTGGTGACCCCAAGTTGCCGTGGGTCCA
TTCTGTGGTGACCCCAAGTTGCCGTGGGTCCAG
TGACCCCAAGTTGCCGTGGGTCCAGAGGCACAT
GTTGCCGTGGGTCCAGAGGCACATACGGAACCT
TGGGTCCAGAGGCACATACGGAACCTGGATGCC
…
```

Extract the genes in the CHO genome in order to find mRNA targets

In order to not just find the number of off-targets in the genome but also which off-targets are in transcribed regions of the genome all ranges ±50bp of an mRNA annotated region are fished out (the same way as exon 1)

```
cat /novo/users/csrk/Cgenomes/CHOK1_version1.1/listofgenes | gawk '{print $1}' | uniq > Genelist
plus=+
wc -l Genelist > ./length
length=`awk '{print $1 }' < ./length`
for ((  i = 1;  i <= $length ; i++  ))
do
cat Genelist | head -$i | tail -1 > ./geneinfo
cat ./geneinfo | gawk '{print $1}' > ./genename
genename=`awk '{print $1 }' < ./geneinfo`
grep -w "$genename" ../annotationK1.gff3 | sed 's/\t/,/g' |  awk -F, '$3=="mRNA" {gsub(","," ");print $0}' > gene
orientation=`cat gene | gawk '{print $7}' | head -1`
cat gene | gawk '{print $1, $4-50, $5+50}' | sed 's/ /\t/g' |  sed 's/ /\t/g' | sort -n -k 2 | tail -1 > coordinates
/novo/appl/ngs/BEDTools-Version-2.16.2/fastaFromBed -fi ../CHOK1_ver1.1.fasta -bed ./coordinates -fo ./output.fasta
orientation=`cat gene | gawk '{print $7}' | head -1`
echo " " >> mRNA.fasta
echo ">$genename" >> mRNA.fasta
if [ "$orientation" == "$plus" ]
then
 sed 1d output.fasta >> mRNA.fasta
else
 sed 1d output.fasta > temp.txt
 perl -0777ne's/\n //g; tr/ATGCatgcNn/TACGtacgNn/; print scalar reverse $_;' temp.txt | sed 1d >> mRNA.fasta
fi
done
```

This ended up with 148 errors where the genes were within 50bp of the start/end of scaffold. The output file was split into individual fasta files and each were counted in order to find the files, which were to be rerun with the script above without the extra ±50bp

```
csplit -z mRNA.fasta '/^>/' '{*}' --suffix="%02d.fa" --prefix=RNA- -s
for file in RNA-*
do
head -1 $file > name
sed 1d $file | wc -m > number
echo "$file" > filename
paste name number filename>> list.txt
done
```

The genome and the new transcriptome were indexed making them searchable with bowtie (which is made for NGS analysis and thus optimized for aligning short reads better than e.g. blastn)

```
bowtie-build Cgriseus.fasta Cgriseus
bowtie-build mRNA.fasta mRNA
```

Search these targets against the CHO genome / transcriptome

For each of the 29 target sequences the 34bp sequence extracted above is trimmed to the 13bp target sequence and a 3bp cap is inserted to the end

TCACCAAGAGGGGCCATAAATTCTGTGGTGACC => GCCATAAATTCTGNGG

The new guide sequence is then searched against the mRNA database (shown) and subsequently the same way against the genome (not shown). The output show all target hits, the scaffold (or gene name) and the position of the start of the alignment. The final column show a list of the mismatches. All targets, which have mismatches in pos 15 and 16 (the final GG's) are removed. The number of targets with 1, 2 and 3 mismatches are counted and the targets which have 1 (shown below) and 2 (not shown) are extracted.

Output from

```
0  -  LOC103158597 28290   CCNCAGAATTTATGGC  IIIIIIIIIIIIIIII  3     0:A>C,13:A>N,15:T>C
0  -  LOC100774700 9295    CCNCAGAATTTATGGC  IIIIIIIIIIIIIIII  0     0:A>C,8:G>A,13:C>N
0  -  Kctd16       127312  CCNCAGAATTTATGGC  IIIIIIIIIIIIIIII  0     0:A>C,8:C>A,13:T>N
```

Each guide sequence are scored based on the arbitrary score: 10 points for each off-target gene with 0 mismatches, 5 points for each off-target gene with 1 mismatch, 2 points for each genome off-target with 0 mismatches and 1 point for each genomic off-target with 1 mismatch. All guide sequences are sorted so that the target with the lowest score are listed in the top the output.

```
motif=NGG
length=`sed 1d finalcandidates | wc -l`
for ((  i = 1 ;  i <= $length ; i++  ))
do
head -$i finalcandidates | tail -1 > oligo
cat oligo | sed "s/^.\{12\}//" | cut -c 1-13 > temp2
temp2=`cat temp2`
blastmotif="$temp2$motif"
echo "$temp2$motif" >  candidate
bowtie -c ./libs/mRNA $blastmotif -v 3 -a > results
cat results | grep -v "14:" | grep -v "15:" > correcthits
cat correcthits | gawk '{print $8}' | awk -F ':' '{print NF-1, NR}' | gawk '{print $1}' | sed 's/1/FXX1/g' | sed
's/2/FXX2/g' | sed 's/3/FXX3/g' > number
grep "1" number | wc -l > 1RNA
grep "2" number | wc -l > 2RNA
grep "3" number | wc -l > 3RNA
```

```
##extract hits
paste correcthits number | grep "FXX1"  > sub1hits
length2=`cat sub1hits | wc -l`
rm sample2
for ((  j = 1 ;  j <= $length2 ; j++  ))
 do
 head -$j sub1hits | tail -1 > sample
 ori=`cat sample | gawk '{print $2}'`
if [ "$ori" == "$plus" ]
   then
    cat sample | gawk '{print $3, $4+1, $4+16, 0}' >> sample2
   else
    cat sample | gawk '{print $3, $4+16, $4+1, 0}' >> sample2
   fi
done
cat sample2 | sed 's/ /;/g' | awk '$1=$1' RS= OFS=, > 1mmRNAhits
…#similar code as above for searching the genomic sequence
paste 1RNA 2RNA 1genome 2genome | awk '{a=$1*10; b=$2*5; c=$3*2; d=a+b+c+$4; print d}' > oligoscore
paste genename oligo candidate 1RNA 2RNA 3RNA 1genome 2genome 3genome oligoscore 1mmRNAhits >>
some_results
paste genename oligo candidate 1RNA 2RNA 3RNA 1genome 2genome 3genome oligoscore 1mmRNAhits 2mmRNAhits
1mmGhits 2mmGhits >> all_results
done
cat some_results | sort -k 6 -n > results_minimal_$file.txt
cat all_results | sort -k 6 -n > results_all_$file.txt
done
```

Convert output to excel

Finally the output is converted from a txt file to an excel file

converter_all script

```
#! /usr/bin/env Rscript
col=as.matrix(read.table("stdin"))
library(xlsx)
colnames(col)=c("Gene","target","GuideRNA","RNA_targets_0mismatch","RNA_targets_1mismatch","RNA_targets_2mi
smatch","genome_targets_0mismatch","genome_targets_1mismatch","genome_targets_2mismatch","score","RNA_hit
s_0","RNA_hits_1","Genome_hits_0","Genome_hits_1")
write.xlsx(col,"./output.xlsx")

for file in results_all_*
do
sed 's/\t\t/\t0\t/' $file > temp
./converter_all < temp
name=`echo "$file" | sed 's/....$//' | sed 's/^...........//'` >> names
mv output.xlsx $name.xlsx
mv $file $file.done
rm temp
done
rm names
```

Example of transposed output given below

Gene	Aadac
target	TATGTCCCTCTTCCAGGGGATATTGAGGAGCCT
GuideRNA	CCAGGGGATATTGNGG
RNA_targets_0mismatch	3
RNA_targets_1mismatch	49
RNA_targets_2mismatch	908
genome_targets_0mismatch	5
genome_targets_1mismatch	146
genome_targets_2mismatch	2793
score	431
RNA_hits_0	Aadac;116;131;0.Acp5;8012;7997;0.Srpk1;34003;33988;0

RNA_hits_1 LOC100750418;23316;23331;1.Zyg11b;41771;41786;1…
Genome_hits_0 NW_006876508.1;193104;193119;0...NW_006885091.1;1096669;1096684;0
Genome_hits_1 NW_006880411.1;33917;33932;1.NW_006880577.1;2284658;2284673;1…

10.3.8 Primer design for qRT-PCR in CHO

Due to RNA sequencing data[65] being released before the genome sequence of CHO-K1[66] the Genbank entries for each gene as stated on CHOgenome.org are only based on mature RNA and not the entire mRNA sequence containing introns. Thus, designing qRT-PCR primers spanning introns in order to desinguish between DNA and cDNA is not very straight forward. For this reason the mRNA regions of the genome including introns were extracted using the code below and used to create 29978 Genbank files, which can be opened in e.g. Vector NTI.

A list of all genes in the *C. griseus* genome was extracted from the annotation file

```
cat annotationK1.gff3 | sed 's/\t/,/g' |  awk -F, '$3=="exon" {gsub(","," ");print $0}' | grep "gene=" > sca
cat sca | sed 's/^.*gene=//g' | cut -f 1,1 -d';' > genes
cat sca | sed 's/^.*Genbank://g' | cut -f 1,1 -d';' | cut -f 1,1 -d',' | gawk '{print $1}' > Genbk
cat sca | sed 's/^.*product=//g' | cut -f 1,1 -d';' | sed 's/ /;/g' > genenames
cat sca | sed 's/^.*Parent=//g' |   cut -f 1,1 -d';' > RNAid
cat sca | gawk '{print $1}' > seqscaffold
paste genes Genbk RNAid genenames seqscaffold  | sort | uniq > exon_Genbank

cat annotationK1.gff3 | sed 's/\t/,/g' |  awk -F, '$3=="CDS" {gsub(","," ");print $0}' | grep "gene=" > sca
cat sca | sed 's/^.*gene=//g' | cut -f 1,1 -d';' > genes
cat sca | sed 's/^.*Genbank://g' | cut -f 1,1 -d';' | cut -f 1,1 -d',' | gawk '{print $1}' > Genbk
cat sca | sed 's/^.*product=//g' | cut -f 1,1 -d';' | sed 's/ /;/g' > genenames
cat sca | sed 's/^.*Parent=//g' |   cut -f 1,1 -d';' > RNAid
paste genes Genbk RNAid genenames | sort | uniq > CDS_Genbank

paste CDS_Genbank | gawk '{print $3}' | sort | uniq > RNACDS
paste exon_Genbank | gawk '{print $3}'  | sort | uniq > RNAexon
sdiff RNACDS RNAexon | grep -v ">" | grep -v "|" | gawk '{print $1}' > unique_proteinrna

wc -l unique_proteinrna > length
length=`awk '{print $1 }' < length`
for ((  i = 1 ;  i <= $length ; i++  ))
do
cat unique_proteinrna | head -$i | tail -1 > gene
gene=`awk '{print $1 }' < gene`
grep -w "$gene" CDS_Genbank > temp1
grep -w "$gene" exon_Genbank > temp2
paste temp1 temp2 | gawk '{print $1, $6, $2, $3, $9}' | sed 's/ /\t/g' >> genenames
done
printf "\tASDF\n" | awk -F'\t' '$3' genenames | sort | uniq > listofgenes
```

The list looks like this and thus contain all Genbank entries for RNA tarnscripts, CDS region and the sequencing scaffold the gene was found on for all splice variants

A1cf	XM_007627897.1	XP_007626087.1	rna20866	NW_006880426.1
A2ml1	XM_007611503.1	XP_007609693.1	rna3450	NW_006870833.1

[65] http://www.ncbi.nlm.nih.gov/pubmed/21945585
[66] http://www.nature.com/nbt/journal/v29/n8/full/nbt.1932.html

A3galt2	XM_007614932.1	XP_007613122.1	rna7152	NW_006873147.1
A4galt	XM_007629783.1	XP_007627973.1	rna22847	NW_006881179.1
A4galt	XM_007629784.1	XP_007627974.1	rna22846	NW_006881179.1
A4gnt	XM_007634050.1	XP_007632240.1	rna27323	NW_006883256.1

Next step is to extract the regions from the genome using FastaFromBed from BEDTools. This is done by first finding the coordinates in the genome from the annotation file for each gene and then using the coordinates to extract a fasta file using only this sequence from the genome.

```
cat /novo/users/csrk/Cgenomes/CHOK1_version1.1/listofgenes | gawk '{print $3, $1, $2}' | sed 's/ /\t/g' | uniq > Genelist
plus=+
wc -l Genelist > ./length
length=`awk '{print $1 }' < ./length`
for ((  i = 1;  i <= $length ; i++  ))
do
cat Genelist | head -$i | tail -1 > ./geneinfo
cat ./geneinfo | gawk '{print $1}' > ./genename
geneCHO=`awk '{print $2 }' < ./geneinfo`
geneCHOname=`awk '{print $1 }' < ./geneinfo`
genesearch=`awk '{print $3 }' < ./geneinfo`
grep -w "$genesearch" ../annotationK1.gff3 | sed 's/\t/,/g' |  awk -F, '$3=="mRNA" {gsub(","," ");print $0}' > gene
orientation=`cat gene | gawk '{print $7}' | head -1`
cat gene | gawk '{print $1, $4-1, $5}' | sed 's/ /\t/g' |  sed 's/ /\t/g' | sort -n -k 2 | tail -1 > coordinates
/novo/appl/ngs/BEDTools-Version-2.16.2/fastaFromBed -fi ../CHOK1_ver1.1.fasta -bed ./coordinates -fo ./output.fasta
orientation=`cat gene | gawk '{print $7}' | head -1`
echo " " >> mRNA.fasta
if [ "$orientation" == "$plus" ]
then
  echo ">$geneCHO-$geneCHOname-p" >> mRNA.fasta
 sed 1d output.fasta >> mRNA.fasta
else
  echo ">$geneCHO-$geneCHOname-m" >> mRNA.fasta
 sed 1d output.fasta > temp.txt
 perl -0777ne's/\n //g; tr/ATGCatgcNn/TACGtacgNn/; print scalar reverse $_;' temp.txt | sed 1d >> mRNA.fasta
fi
done
```

The mRNA.fasta file is then split into individual files given the name from the fasta header

```
csplit -z mRNA.fasta '/^>/' '{*}' --suffix="%02d.fa" --prefix=mRNA- -s
for file in *.fa
do
name=`head -1 $file | sed 's/^>//g'`
mv $file $name.fasta
done
```

To fish out the exon coordinates the R-script below is used to make the coordinates relative

```
echo '#! /usr/bin/env Rscript' > Rminus
echo 'RNA=as.matrix(read.table("RNA"))' >> Rminus
echo 'exon=as.matrix(read.table("exon"))' >> Rminus
echo 'exonminus=cbind((RNA[1,2]-exon[,2]+1),(RNA[1,2]-exon[,1]+1))' >> Rminus
echo 'write.table(exonminus,file="output.txt",sep="\t")' >> Rminus
chmod a+x Rminus

echo '#! /usr/bin/env Rscript' > Rplus
echo 'RNA=as.matrix(read.table("RNA"))' >> Rplus
echo 'exon=as.matrix(read.table("exon"))' >> Rplus
echo 'exonplus=cbind((exon[,1]-RNA[1,1]+1),(exon[,2]-RNA[1,1]+1))' >> Rplus
echo 'write.table(exonplus,file="output.txt",sep="\t")' >> Rplus
chmod a+x Rplus
```

The code below is used to fish out the genome coordinates and using the code above the coordinates are made relative to the fastafile created above

```
cat /novo/users/csrk/Cgenomes/CHOK1_version1.1/listofgenes | gawk '{print $3, $1, $2}' | sed 's/ /\t/g' | uniq > Genelist
plus=+
wc -l Genelist > ./length
length=`awk '{print $1 }' < ./length`
for ((  i = 1;  i <= $length ; i++  ))
do
cat Genelist | head -$i | tail -1 > ./geneinfo
cat ./geneinfo | gawk '{print $1}' > ./genename
geneCHO=`awk '{print $2 }' < ./geneinfo`
genesearch_tran=`awk '{print $3 }' < ./geneinfo`
genesearch_CDS=`awk '{print $1 }' < ./geneinfo`
grep -w "$genesearch_tran" ../annotationK1.gff3 > stuff1
grep -w "$genesearch_CDS" ../annotationK1.gff3 > stuff2
cat stuff1 | sed 's/\t/,/g' | awk -F, '$3=="mRNA" {gsub(","," ");print $0}' | gawk '{print $4, $5}' > RNA
cat stuff2 | sed 's/\t/,/g' | awk -F, '$3=="CDS" {gsub(","," ");print $0}' | gawk '{print $4, $5}' > exon
orientation=`head -1 stuff1 | gawk '{print $7}'`
if [ "$orientation" == "$plus" ]
then
 ./Rplus
 cat output.txt | sed 's/^"/exon/g' | sed 's/"\t/\t/g' | sed 1d > $geneCHO-$genesearch_CDS-p.txt
else
 ./Rminus
 cat output.txt | sed 's/^"/exon/g' | sed 's/"\t/\t/g' | sed 1d > $geneCHO-$genesearch_CDS-m.txt
fi
done
```

Finally the genes with only one exon needs to have the name changed.

```
for file in *.txt
do
cat $file | sed 's/exonV1/exon1/g' > temp
mv temp $file
done
```

In order to merge the coordinates in the txt file with the fasta file into one annotated Genbank file a script was written by my colleague Thomas P. Boesen

```
#!/usr/bin/env ruby
def process_file(name)
  puts 'Processing: ' + name
  name =~ /-(\w+)[^\/]*$/
  locus = $1
  # READ FILE CONTENT #
  sequence = ''
  File.open(name + '.fasta', "r") do |f|
    f.each_line do |line|
      sequence += line.strip unless line.start_with? '>'
    end
  end
#  puts "Seq length: #{sequence.length}"
  exons = []
  File.open(name + '.txt', 'r') do |f|
    f.each_line do |line|
      if line =~ /^(\w+)\s+(\d+)\s+(\d+)/
        exons.push({:label => $1, :start => Integer($2), :end => Integer($3)})
      end
    end
  end
#  exons.each {|exon|
#    puts "Exon: #{exon[:label]} #{exon[:start]} #{exon[:end]} #{sequence[exon[:start]..exon[:end]]}"
#  }
  # GENERATE GENBANK FILE #
  File.open(name + '.genbank', 'w'){|file|
```

```
    file.write "LOCUS       " + locus.ljust(17) + "#{sequence.length}".rjust(6) + " bp    DNA     circular     26-MAR-
2015\n"
    file.write "SOURCE      \n"
    file.write "  ORGANISM  \n"
    file.write "FEATURES             Location/Qualifiers\n"
    exons.each {|exon|
      file.write "     misc_feature    #{exon[:start]}..#{exon[:end]}\n"
      file.write "                     /vntifkey=\"21\"\n"
      file.write "                     /label=#{exon[:label]}\n"
    }
    file.write "ORIGIN\n"
    iBase = 0
    while iBase < sequence.length
      file.write "#{iBase+1}".rjust(9) + " #{sequence[iBase..iBase+9]} #{sequence[iBase+10..iBase+19]}
#{sequence[iBase+20..iBase+29]} #{sequence[iBase+30..iBase+39]} #{sequence[iBase+40..iBase+49]}
#{sequence[iBase+50..iBase+59]}\n"
      iBase += 60
    end
    file.write "//\n"
  }
end
def process_dir (name)
  files = Dir.glob(name + '/*.txt')
  files.each {|file|
    process_file file[0..-5]
  }
end
process_dir(ARGV[0])
#process_file(ARGV[0])
```

Output opened in Vector NTI

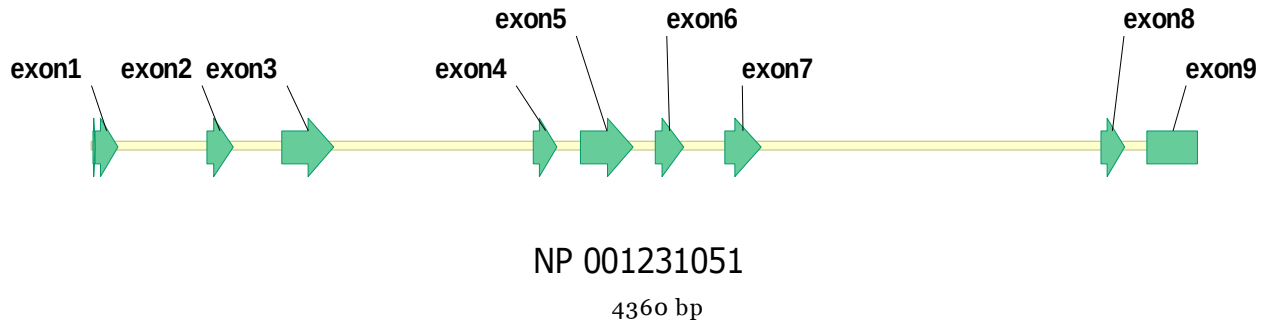


Figure 35 overview of the expon location in calreticulin (NP_001231051).

Using the files above qPCR primers were designed using Primer3 (http://bioinfo.ut.ee/primer3-0.4.0/).

10.3.9 Validation of primers against the CHO genome, transcriptome and mature transcriptome

Input a list of primers to be validated against the CHO genome/transcriptome

cat input.txt

```
CK369       GGAACCTGCCGTCTATTTCA
CK370       CCCGTAGAATTTGCCAGAAC
CK371       GGACCACTGTCCCATCAAGT
CK372       TGAAAACAAAGGGCTGGTTC
```

```
CK373           GCTTGCCCCTGAGTATGAAG
CK374           TAAGGGTTGGGTAGCCACTG
CK375           GTACCCAAACCCTCCAATCC
CK376           TCTTCTGAAGCTGGCTGACA
CK377           CTCAAGTGAGGTGGCTGTCA
CK378           CGTGATTCCAAAAGGGATGT
CK379           AGAACGGCATCAAGGTGAAC
CK380           TGCTCAGGTAGTGGTTGTCG
```

The script below imports the list above and uses bowtie to align the oligoes to the CHO genome to find potential off-targets

```
cat input.txt | tr '\015' '\012' | sort -n | uniq | sed 1d > input
length=`wc -l input | gawk '{print $1}'`
for ((  i = 1 ;  i <= $length ; i++  ))
do
oligo=`head -$i input | tail -1 | gawk '{print $2}'`
echo $oligo > name
head -$i input | tail -1 | gawk '{print $1}' > primername
bowtie -c ~/Cgenomes/CHOK1_version1.1/Crispr/libs/Cgriseus $oligo -v 2 -a > results1
bowtie -c ~/Cgenomes/CHOK1_version1.1/Crispr/libs/mRNA $oligo -v 2 -a > results2
bowtie -c ~/Cgenomes/CHOK1_version1.1/Crispr/libs/mRNAs_mature $oligo -v 2 -a > results3
cat results1 | gawk '{print $8}' | awk -F ':' '{print NF, NR}' | gawk '{print $1}' | sed 's/0/FXX0/g' | sed 's/1/FXX1/g' |
sed 's/2/FXX2/g' > number
grep "0" number | wc -l > 0g
grep "1" number | wc -l > 1g
grep "2" number | wc -l > 2g
paste number results1 | grep -v "FXX2"  > sub1
length2=`cat sub1 | wc -l`
rm sample2
for ((  j = 1 ;  j <= $length2 ; j++  ))
 do
 head -$j sub1 | tail -1 > sample
 ori=`cat sample | gawk '{print $2}'`
if [ "$ori" == "$plus" ]
                    then
                     cat sample | sed 's/FXX0/0/g' | sed 's/FXX1/1/g' | gawk '{print $4, $5+1, $5+16, $1}' >> sample2
                    else
                     cat sample | sed 's/FXX0/0/g' | sed 's/FXX1/1/g' | gawk '{print $4, $5+16, $5+1, $1}' >> sample2
                    fi
done
cat sample2 | sed 's/ /;/g' | awk '$1=$1' RS= OFS=, > 0mm_g
#
cat results2 | gawk '{print $8}' | awk -F ':' '{print NF, NR}' | gawk '{print $1}' | sed 's/0/FXX0/g' | sed 's/1/FXX1/g' |
sed 's/2/FXX2/g' > number
grep "0" number | wc -l > 0r
grep "1" number | wc -l > 1r
grep "2" number | wc -l > 2r
paste number results2 | grep -v "FXX2"  > sub1
length2=`cat sub1 | wc -l`
rm sample2
for ((  j = 1 ;  j <= $length2 ; j++  ))
 do
 head -$j sub1 | tail -1 > sample
 ori=`cat sample | gawk '{print $2}'`
if [ "$ori" == "$plus" ]
                    then
                     cat sample | sed 's/FXX0/0/g' | sed 's/FXX1/1/g' | gawk '{print $4, $5+1, $5+16, $1}' >> sample2
                    else
                     cat sample | sed 's/FXX0/0/g' | sed 's/FXX1/1/g' | gawk '{print $4, $5+16, $5+1, $1}' >> sample2
                    fi
done
cat sample2 | sed 's/ /;/g' | awk '$1=$1' RS= OFS=, > 0mm_r
#
cat results3 | gawk '{print $8}' | awk -F ':' '{print NF, NR}' | gawk '{print $1}' | sed 's/0/FXX0/g' | sed 's/1/FXX1/g' |
sed 's/2/FXX2/g' > number
grep "0" number | wc -l > 0g
grep "1" number | wc -l > 1g
```

```
grep "2" number | wc -l > 2g
paste number results3 | grep -v "FXX2"  > sub1
length2=`cat sub1 | wc -l`
rm sample2
for ((  j = 1 ;  j <= $length2 ; j++  ))
 do
 head -$j sub1 | tail -1 > sample
 ori=`cat sample | gawk '{print $2}'`
if [ "$ori" == "$plus" ]
                    then
                     cat sample | sed 's/FXX0/0/g' | sed 's/FXX1/1/g' | gawk '{print $4, $5+1, $5+16, $1}' >> sample2
                    else
                     cat sample | sed 's/FXX0/0/g' | sed 's/FXX1/1/g' | gawk '{print $4, $5+16, $5+1, $1}' >> sample2
                    fi
done
#
cat sample2 | sed 's/ /;/g' | awk '$1=$1' RS= OFS=, > 0mm_r2
paste primername name 0g 1g 2g 0r 1r 2r 0r2 1r2 2r2 0mm_g 0mm_r 0mm_r2 >> list.txt
done
```

Finally the output is converted to an excel file

```
echo '#! /usr/bin/env Rscript' > converter_all
echo 'col=as.matrix(read.table("stdin"))' >> converter_all
echo 'library(xlsx)' >> converter_all
echo
'colnames(col)=c("primername","primerseq","Genome_0mm","Genome_1mm","Genome_2mm","RNA+intron_0mm","
RNA+intron_1mm","RNA+intron_2mm","RNA-intron_0mm","RNA-intron_2mm","RNA-intron_3mm","Genome_off-
targets","mRNA_offtargets","mRNA_mature_offtargets")' >> converter_all
echo 'write.xlsx(col,"./output.xlsx")' >> converter_all
chmod a+x converter_all

for file in list.txt
do
sed 's/\t\t/\t0\t/' $file > temp
./converter_all < temp
name=`echo "$file" | sed 's/....$//' >> names
mv output.xlsx $name.xlsx
mv $file $file.done
rm temp
done
```

From the output (list.xlsx) it can be seen how well each oligo fares concerning off-target binding. From the binding positions found in the last 3 columns it can be calculated what the amplicon length will be using gDNA or cDNA as template.

Primername	CK369	CK370
Primerseq	GGAACCTGCCGTCTATTTCA	CCCGTAGAATTTGCCAGAAC
Genome_0mm	1	1
Genome_1mm	0	0
Genome_2mm	0	0
RNA+intron_0mm	1	1
RNA+intron_1mm	0	0
RNA+intron_2mm	0	0
RNA-intron_0mm	0	0
RNA-intron_2mm	0	0
RNA-intron_3mm	0	0
Genome_off-targets	NW_006879744.1;597082;597067;0	NW_006879744.1;597539;597524;0…
mRNA_offtargets	Calr;115;100;0	Calr;572;557;0,Daam1;18574;18559;3…
mRNA_mature_offtargets	NM_001244122.1;66;51;0	NM_001244122.1;170;155;0