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

Kaas, Christian Schrøder; Bolt, Gert; Hansen, Jens J; Andersen, Mikael Rørdam; Kristensen, Claus

Published in:
Biotechnology Journal

Link to article, DOI:
10.1002/biot.201400667

Publication date:
2015

Document Version
Peer reviewed version

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]
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

Correspondence: Mr. Christian S. Kaas, Mammalian Cell Technology, Novo Nordisk A/S, A9.2.36, Novo Nordisk Park, 2760 Maaloev, Denmark
E-mail csrk@novonordisk.com
Keywords: Cell Culture; CHO cells; Coagulation Factor VIII; Gene delivery; Next-generation sequencing
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]

Abstract

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 2-3 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.

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 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 $3x10^5$ cells/ml in 30ml of HyClone CDM4CHO media (Thermo) supplemented with Penicillin/Streptomycin (Gibco) and MTX, and cultured for 72 hours 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 sub clones (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 $3x10^5$ 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 5 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 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) 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) 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).

2.3 RNA purification and next-generation sequencing

After 48 hours of cultivation 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). 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, Denmark) using 8 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 normalised 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 hours 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 1µg 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 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 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 mio cells using DNeasy Blood & Tissue Kit (Qiagen) following manufacturer's instructions. 100ng of gDNA were used as template for a 25µl 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 hours 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 (Figure 1). From 62 clones, 14 clones exhibiting a dynamic range of FVIII productivities and were selected for further analysis (Figure 1). These clones were named clone 1-14 based on descending FVIII productivity. The three highest yielding clones (clone 1-3) were sub-clones of the same clonal cell line. Clone 12-14 did not produce detectable FVIII levels and were classified as non-producers.

When seeded at $3x10^5$ cells/ml, most clones entered the exponential growth phase within 27 hours of cultivation and had not entered the stationary phase 73 hours after seeding (Figure 2A). The selected clones all had doubling times of 30-40 hours. During cultivation, cell growth medium glucose and glutamine levels gradually decreased, while lactate levels increased (Figure 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 hours of cultivation (Figure 2C) before stagnating or dropping. Based on these growth and production characteristics, we decided to extract RNA for analysis of the transcriptome after 48 hours 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.

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 (Figure 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/48h (Pearson correlation = 0.03) nor as ng FVIII/cell/48h (Pearson correlation = 0.24). In contrast the expression level of *dhfr* mRNA showed a small tendency to correlate negatively (Pearson correlation = -0.55) (Figure 3B). Among the three non-producing 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 (Figure 3B). The expression level of *dhfr* in the non-producing clones was similar to or higher than those of the FVIII producing clones (Figure 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 (Figure 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] (Figure 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 (Figure 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 (Figure 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 transcribed (Figure 3D). The read depth distribution of the three non-producing 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 (Figure 3D). In contrast, the *dhfr* ORF was transcribed in all 3 clones, explaining the capacity of the non-producing 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 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 472bp 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 (Figure 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 (Figure 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* RNA's 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.

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 non-conservative recombination resulting in truncations 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 shows 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 suggests that the tripartite leader can indeed increase the yields of recombinant FVIII translated from a lower level of *F8*-mRNA (Figure 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 Figure 3C top). Cell line development typically involves the generation of non-clonal 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.

Acknowledgement

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.

Conflict of interest

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. *Pediatric clinics of North America* 2013, 60, 1419-1441.

[2] Pipe, S.W., Recombinant clotting factors. *Thrombosis and haemostasis* 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. *Proceedings of the National Academy of Sciences* 1986, 83, 2979-2983.

[6] Pipe, S.W., Functional roles of the factor VIII B domain. *Haemophilia : the official journal of the World Federation of Hemophilia* 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. *Proceedings of the National Academy of Sciences* 1986, 83, 5939-5942.

[10] Wurm, F.M., Production of recombinant protein therapeutics in cultivated mammalian cells*. Nature biotechnology* 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. *Molecular cell* 2005, 17, 405-416.

[12] Davies, M.V., Kaufman, R.J., Internal translation initiation in the design of improved expression vectors. *Current Opinion in Biotechnology* 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. *Journal of biotechnology* 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. *Biotechnology and bioengineering* 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. *Biotechnology journal* 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. *Nature biotechnology* 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. *Nature biotechnology* 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. *Pharmaceutical Bioprocessing* 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 research* 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. *Biotechnology and bioengineering* 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 : the official journal of the World Federation of Hemophilia* 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. *Human gene therapy* 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. *Journal of thrombosis and haemostasis* 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. *Nature biotechnology* 2014, 32, 1019-1025.
[28] Trapnell, C., Pachter, L., Salzberg, S.L., TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)* 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 (Oxford, England)* 2010, 26, 139-140.
[30] Ihaka, R.,Gentleman, R., R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics* 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. *Nature protocols* 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. *Biotechnology and bioengineering* 1998, 58, 73-84
[34] Bishop, J.O., Smith, P., Mechanism of chromosomal integration of microinjected DNA. *Molecular biology & medicine* 1989, 6, 283–98.
[35] Calos, M.P., Lebkowski, J.S., Botchan, M.R., High mutation frequency in DNA transfected into mammalian cells. *Proceedings of the National Academy of Sciences* 1983, 80, 3015–3019.
[36] Würtele, H., Little, K.C.E. & Chartrand, P., Illegitimate DNA integration in mammalian cells. *Gene therapy* 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 research* 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. *Journal of virology* 1990, 64, 2669-2677.
[39] Logan, J., Shenk, T., Adenovirus tripartite leader sequence enhances translation of mRNAs late after infection. *Proceedings of the National Academy of Sciences of the United States of America* 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 and Purification* 2010, 69, 9-15.

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

Name	RNAseq	DNAseq	Chromosome
Clone 1	303 - 7,347	1 - 7,776	2 [a)]
Clone 8	1,507 - 7,327	1,313 - 8,262	X
Clone 12	2,717 - 8,185	2,705 - 8,261	7 [a)]
Clone 13	3,897 - 7,405	3,854 - 7,643	5 [a)]
Clone 14	6,387 - 7,285	583 - 1,055, 6,388 - 7,646, 4,382 - 4,549, 8,757 - 8,900	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.

Figure legends

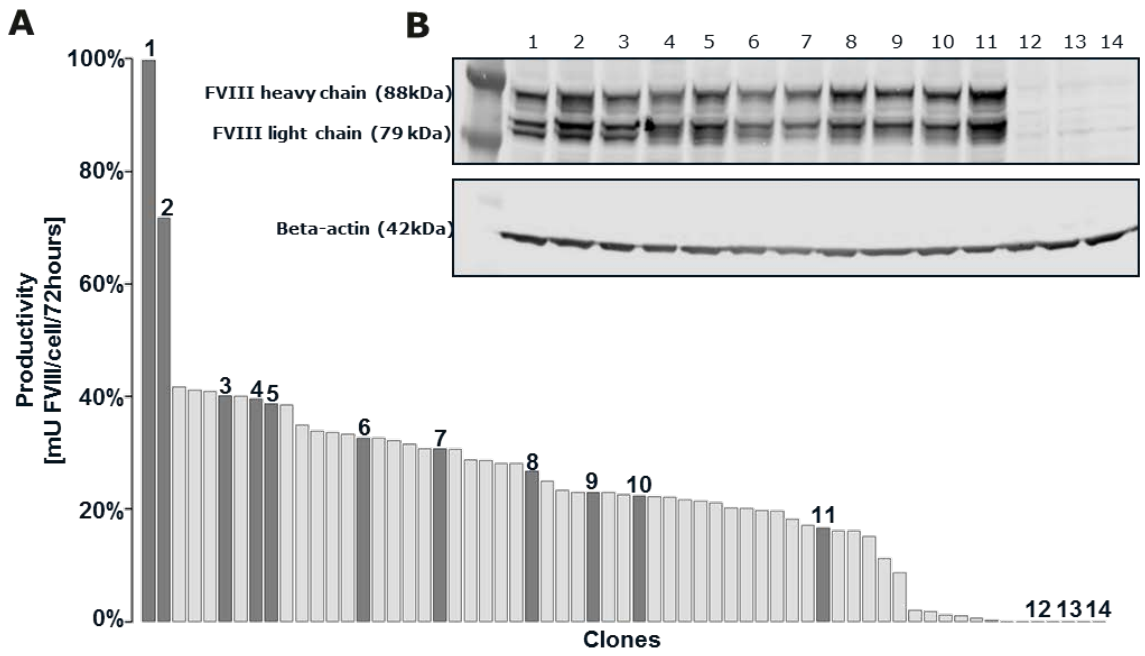

Figure 1. CHO DXB11 cells were transfected with a plasmid encoding FVIII and were subsequently single cell diluted. (A) The productivity of 62 CHO clones. Each bar represents an individual clone and clones numbered 1-14 were selected for further analysis. FVIII productivity was calculated as units recombinant 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.

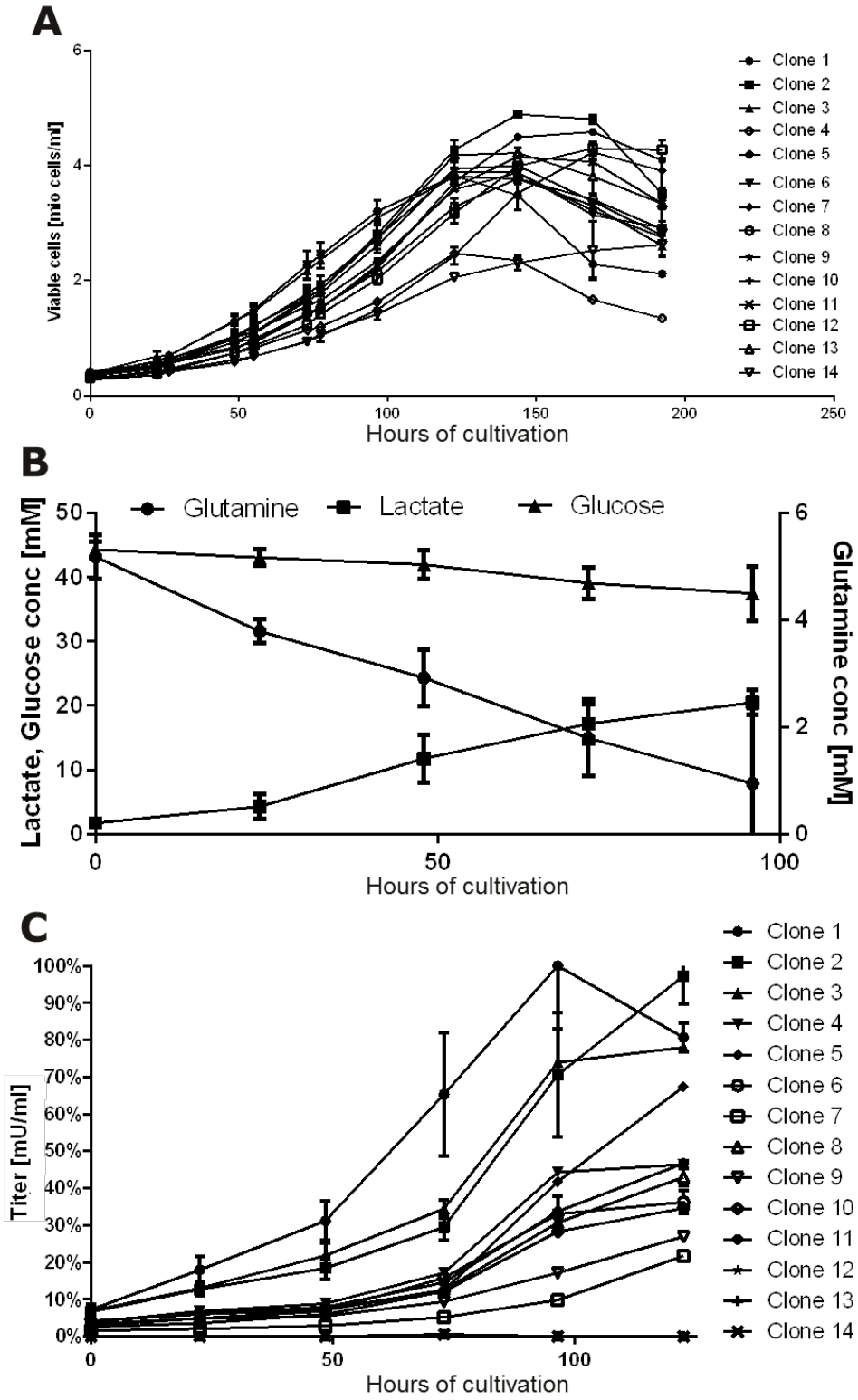

Figure 2

Figure 2. Growth and FVIII production of 14 selected CHO DXB11 clones. (A) Cultures were set ups seeded at 3x10e5 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.

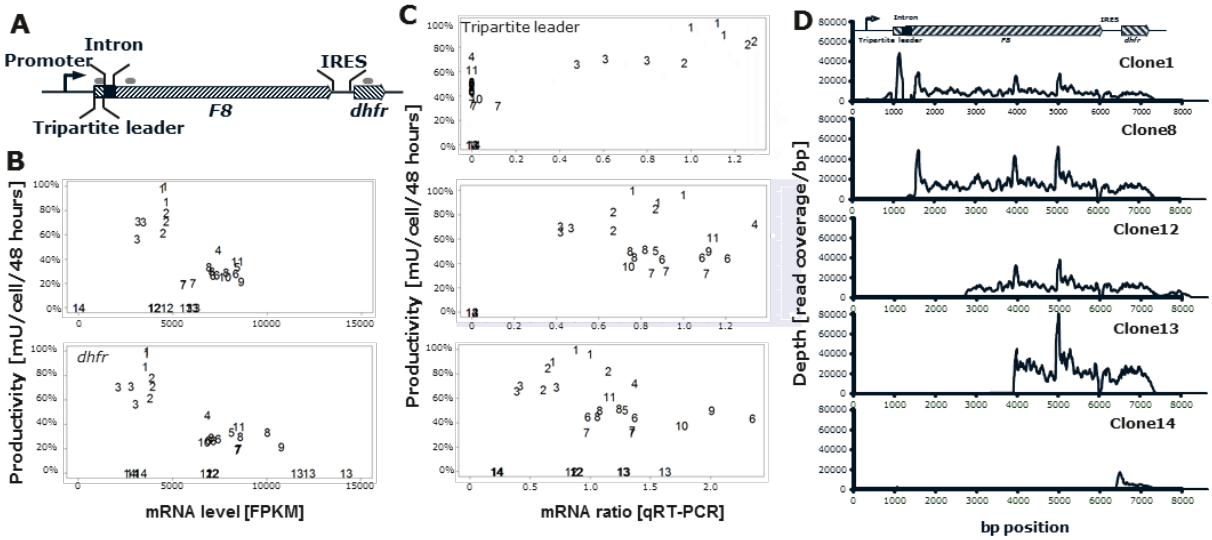

Figure 3. *F8* and *dhfr* mRNA levels across the 14 analyzed clones. (A) Overview of transgene composition. qPCR primer targets used in (C) are marked with grey lines. (B) *F8* and *dhfr* expression level based on RNA sequencing in the CHO DXB11 clones (1-14) versus productivity. FVIII productivity is calculated as units recombinant FVIII released per cell within the first 48 hours 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 (non-producing cell line), clone 13 (non-producing cell line) and clone 14 (non-producing cell line). Representative clones for the various RNA signatures shown.

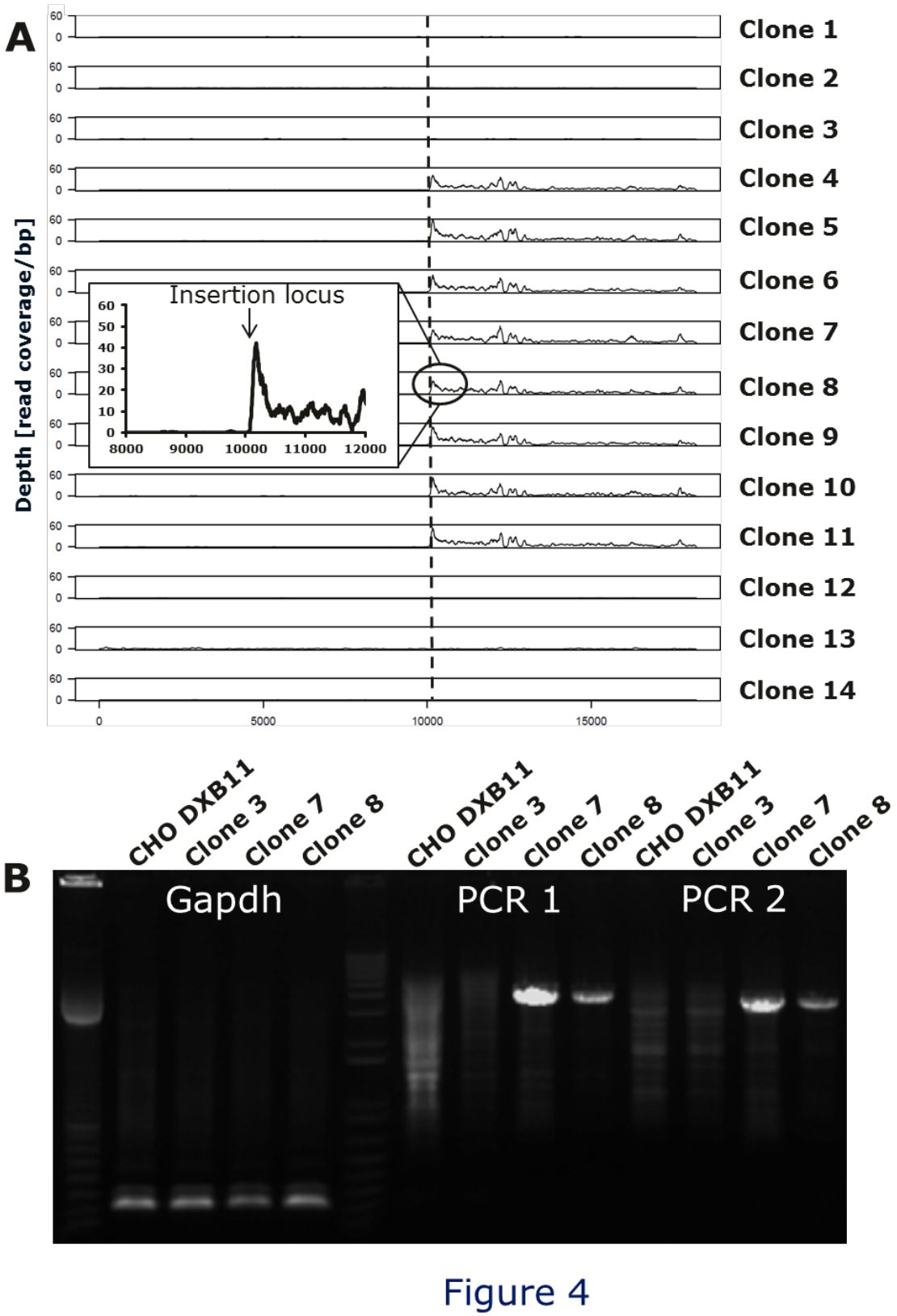

Figure 4

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 222bp 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.