



UNIVERSITA' DEGLI STUDI DI TORINO

Ph.D. School "Health and Life Sciences"

Ph.D. Programme "Molecular Medicine"

Ph.D. THESIS

*CUTTING-EDGE MOLECULAR TECHNOLOGIES TO ADDRESS EFFECT OF
GENOMICS AND EPIGENOMICS ON CLONE PRODUCTIVITY IN CHO-K1
SYSTEMS*

XXXIII CYCLE

Tutors:

Luisa Marena, Ph.D.

Chiar.ma Prof.ssa Fiorella Altruda

Candidate:

Dott.ssa Melissa Alzona

ACADEMIC YEARS: 2017-2021

Table of Contents:

1. Introduction	3
1.1. Manufacturing process for therapeutic protein production	3
1.1.1. Cell Line Development	5
1.2. Genetic parameters as predictive marker for best clone selection in bioproduction	6
1.2.1. Droplet Digital PCR (ddPCR) strategy to evaluate transgene copy number and mRNA ..	6
1.2.2. Transgene copy number and expression as predictive genetic marker	8
1.2.3. NGS based methods to identify relationships with best clone features	8
1.2.4. Possible use of Transcriptomics data in best clone selection	12
1.2.5. DNA methylation could regulate expression by inducing high rate of productivity	13
1.2.6. Chromatin rearrangements and histone marks could regulate expression by inducing high rate of productivity	16
2. Aim and Experimental Design	20
3. Materials and Methods	22
3.1. DNA and RNA extraction	22
3.2. Gene Copy Number determination using ddPCR	22
3.3. Quantification of target mRNA expression using ddPCR	23
3.4. RNA Sequencing and bioinformatic analysis	24
3.5. DNA methylation analysis using Whole Genome Bisulfite Sequencing (WGBS)	24
3.6. Histone marks analysis by ChIP-Seq	25
4. Results	27
4.1. Transgene Copy Number Evaluation in Good and Bad producer clones	27
4.2. Transgene expression correlates with productivity and copy number analysis	28
4.3. Transgene expression, obtained using RNA-Seq method, is correlated to clone productivity	31
4.4. Transcriptomic studies for evaluation of genes/pathways related to clone productivity ..	31
4.5. Differential DNA methylation studies for evaluation of genes/pathways related to clone productivity	38
4.6. ChIP-Seq analysis studies for evaluation of genes/pathways related to clone productivity	48
5. Discussion	54
6. Conclusion and Future Perspective	57
7. References	58

1. Introduction

1.1. Manufacturing process for therapeutic protein production

Over the last three decades, recombinant proteins gained increased importance in therapeutic applications, and over this period a huge number of biotherapeutic proteins, such as Monoclonal Antibody (mAb) are approved or launched into clinical trials and market¹.

Biotherapeutic proteins are produced based on recombinant DNA technology. Due to the higher dimension and higher complexity of those drugs with respect to non-protein drugs, Biotherapeutics proteins need biological systems for their production, such as bacteria, yeast and mammalian cell line. Currently, the large majority of biopharmaceuticals are produced in mammalian cell lines because of their ability to generate proteins similar to the naturally occurring human form. From the biochemical point of view: therapeutic protein should have a human-specific glycosylation pattern and post-translational modification to not generate allergic reaction once administered^{1,2}.

Chinese Hamster Ovary (CHO) cell line represents the expression system of choice in the manufacturing process of biopharmaceuticals because of their suitability and because satisfying important requirements for the manufacturing process. In particular, CHO cells make glycosylation patterns comparable to the human ones, are less susceptible to the human infective virus and can grow to high densities and in suspension in serum-free media, and finally, they can produce therapeutic proteins at high titer (1-10g/L)^{3,4,7}.

The manufacturing process for therapeutic protein production in mammalian cell lines follows a well-established process (Figure 1). In the first phase, during the Cell Line Development, the gene of interest, that codes for the therapeutic protein, is cloned into a plasmid vector and transfected into CHO cells^{1,2}. Secondly, the transfected cells are subjected to selective conditions to recover those that have stably integrated the transgene. Dihydrofolate reductase (DHFR) and Glutamine Synthetase (GS) are the systems normally used for the selection phase because allow transgene amplification and consequently increase the recombinant protein expression¹. Following the selection step, the recovered heterogeneous cell populations are grown, expanded and single-cell sorted to obtain clonal cell line to be further expanded^{1,2}. At this step, thousands of clones are produced and evaluated for recombinant protein expression and cell growth. After the Clone Screening phase, only one clone is selected as the best candidate for the Master Cell Bank (MCB), based on growth, productivity, stability and product quality. Aliquots of the best clonal populations are produced and stored in liquid nitrogen^{1,2}. Finally, MCB scale-up into a bioreactor, where it's called Working Cell Bank (WCB), to produce the therapeutic protein. In the end, the therapeutic protein is purified through several steps on chromatographic columns and supplements are added to obtain the Drug product^{1,2}.

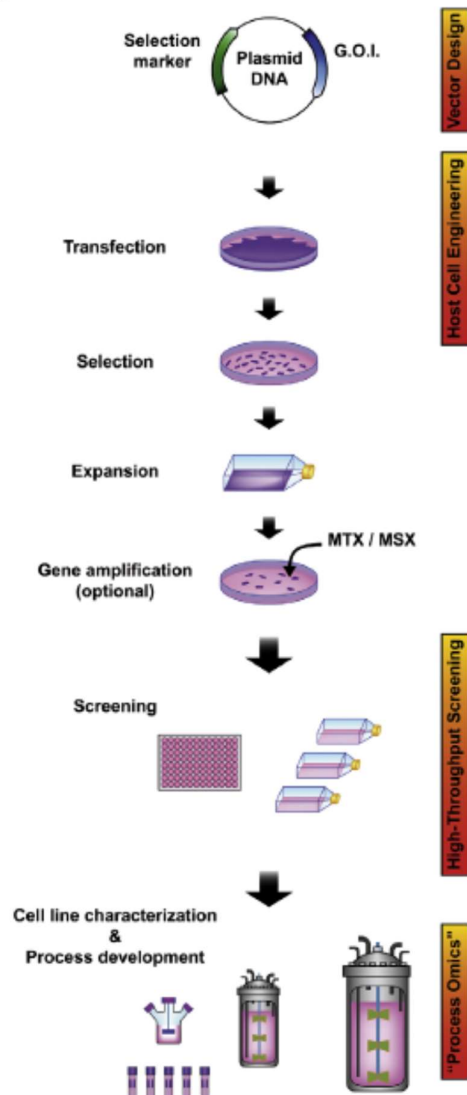


Figure 1: workflow representation of the manufacturing process: the plasmid vector is transfected into host cell line, then after steps of selection and expansions, thousands of clones are produced. Clones produced are screened for recombinant protein production and growth. Finally, the best clone is selected as the MCB (vials indicates aliquot of MCB that are stored in liquid nitrogen). The MCB is growth into Bioreactor to obtain the WCB¹.

Despite the well-establish approach, the manufacturing process remains time-consuming and quite expensive. The process previously described requires more than six months to produce a therapeutic protein. This is particularly inconvenient considering the multiple candidate drugs that are tested in Clinical Trials. It's, therefore, crucial to optimize the manufacturing process, from the gene sequence of plasmid vector to the production in bioreactor^{1,2,5}.

The goal of the development process is to maximize the final protein yield and reduce time and cost, maintaining critical quality attributes of the product. Moreover, even though about 70% of recombinant proteins are produced in CHO cells, still little is known about the molecular basis for their properties regarding adaptability and productivity^{3,4,5,6}.

For these reasons, many efforts are in place for the development process, which requires cutting edge technologies in all phases of the manufacturing and becomes particularly important for the high-throughput

screening of clones produced during the Cell Line Development, as representative for the large-scale bioproduction. In particular, in our work, we will exploit the use of the Droplet Digital PCR (ddPCR) and the Next Generation Sequencing (NGS) in the clone screening phase of the manufacturing process³.

1.1.1. Cell Line Development:

Cell Line Development takes many months from the construction of the expression vector to the best clone selection for the MCB.

The expression vector is generated in order to guarantee high and stable expression of the construct which contains the DNA sequence of the therapeutic protein.

Strong promoters and enhancers are normally used to drive the expression of recombinant protein. Moreover, because the splicing process is known to enhance mRNA nuclear export and stability, introns are introduced in the construct sequence.

Transgene expression is also influenced by the context of the integration site: transgene could be affected by gene silencing due to epigenetic modification (DNA methylation and chromatin condensation), leading to a decrease in the Transgene mRNA expression. For this reason, UTR (Untranslated Regions that flank the coding sequence) or UCOE (Ubiquitous Chromatin Opening Elements) sequence could be inserted in the expression vector in order to avoid gene silencing events⁷.

Once the vector is transfected into the host cells, selection must be applied to select only cells that stably integrate the construct. The most used selection markers are antibiotic resistance genes. A systems for selection includes Dihydrofolate reductase (DHFR) or Glutamine synthase (GS), which are inserted into the plasmid sequence because allow transgene amplification and consequently increase the transgene copy number. In fact, by exposing cell clones to Methotrexate (MTX) and methionine sulfoximide (MSX), respectively inhibitor of DHFR and GS, cells are induced to amplify the selection marker and therefore the gene of interest that is linked^{1,2}. Transgene amplifications occur at the same locus, meaning that there are many tandem copies in the insertion sites. Therefore, the higher Copy Number leads to an increase in transgene mRNA transcription and a final increase in therapeutic protein production^{1,2}.

Following selection, the heterogeneous cell populations are recovered, and consequently must be single-cell sorted and thousands of clones are produced and screened, leading to costly and time-consuming efforts. Therefore, current researches focus on the development of high-throughput screening technologies and platform to allow early identification of high-producing clonal cell lines (Figure 2)^{3,7}.

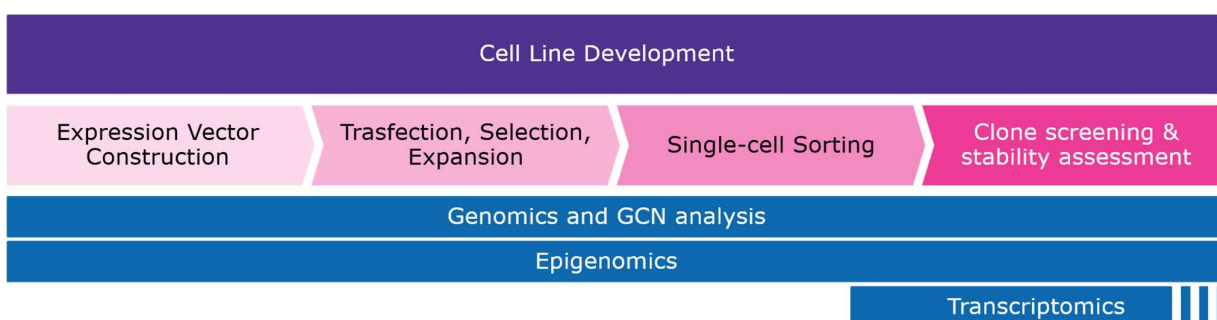


Figure 2: schematic represent of Cell Line Development. All steps could be supported by molecular biology studies: all the phases could be enhanced by analysis of Gene Copy Number and Epigenomics. Moreover, Transcriptomics could help in identification of productivity and stability marker from clone Screening to process development during manufacturing³.

Highly suitable candidate clones are selected based on high cell viability and growth, high productivity, high quality product characteristics and long-term expression stability. If the best candidate lacks those characteristics could impact future productivity in bioreactor and quality of the final product, leading to a loss in time and cost for Pharma company.

Currently, the Clone Screening phase is a multidisciplinary study involving cellular biology, metabolomics and proteomics, respectively for evaluation of cell viability and growth, quantification of protein production and protein expression stability³. For these studies, clones derived from single-cell sorting are grown in a multi-well plate, then survived clones were split and cultured in a Fed-Batch flask for viability, stability and productivity analysis. Fed-Batch flasks are useful to mimic bioreactor conditions at small-scale, in order to early identify the best clone for large-scale production⁸.

Despite many efforts being in place to identify the best candidate, there is a lack of consistent methods to predict the production capacity and growth characteristics of clonal cell lines at the production scale. Moreover, knowledge of the molecular aspects that drives productivity and stability of clones are still not well understood. Recent papers show the possibility to improve the Clone Screening phase using molecular biology techniques in order to identify genetic parameters and biomarkers important to produce the therapeutic protein^{8,9,10,15}.

1.2. Genetic parameters as predictive marker for best clone selection in bioproduction:

During these last years, the identification of biomarkers able to predict productivity and expression stability clone screening of the manufacturing process are becoming important^{3,4,6}. In particular, researchers are now focusing on the development and validation of genetic parameters using molecular biology-based methods. NGS and Droplet Digital PCR (ddPCR) based methods are recently applied to the clone screening phase because of their high-throughput capability⁹⁻¹⁰.

The ddPCR is mostly applied to the Transgene Copy Number analysis and Transgene mRNA expression, which could be directly related to final productivity^{11,12,13,14}. On the other hand, NGS is applied to genome-scale studies, such as Genomics, Transcriptomics and Epigenomics (Figure2). These fields, in fact, could be applied to investigate and understand the overall cellular physiology and mechanism related to the productivity^{3,4}.

1.2.1. Droplet Digital PCR (ddPCR) strategy to evaluate transgene copy number and mRNA:

Droplet Digital Polymerase Chain Reaction (ddPCR) is the third PCR generation, where single gene analysis is combined with quantitative end-point analysis.

While in the first PCR generation, a single gene is amplified by primer forward and reverse and then analyzed by electrophoresis, the second PCR generation, called Real-Time PCR used SYBR Green (as DNA intercalant) or Fluorogenic Probe to quantify the amount of the target sequence present in the sample in real-time analysis. However, this approach requires the use of a standard curve, leading to relative quantification of the target sequence¹¹.

ddPCR is an end-point PCR reaction that allows the absolute quantification of target nucleic acid sequences to be obtained without the use of a standard curve. The absolute quantification of the nucleic acid molecules is possible thanks to the sample distribution system, based on the concept of limiting dilution, or the distribution of the nucleic acid molecules present in the sample, into many different partitions, to obtain even only one or zero molecules by partition¹¹.

In order to obtain the partition of the sample, Droplet Digital PCR uses an approach based on the formation of water-oil emulsion nanoliter-sized droplets, within which the sample is randomly distributed. Subsequently, in each of these droplets, a PCR reaction occurs with which the DNA molecules, if present, are amplified¹¹.

The PCR reaction in the Digital system, as in Real-Time PCR, is performed using specific primers for the target and a fluorescent probe. Therefore, in each partition/droplet, if a molecule of the target DNA is present and the PCR reaction has taken place, a fluorescent signal is emitted by the probes¹¹.

After the PCR reaction, the droplets are analyzed individually by reading the fluorescence signal emitted by the probes (Figure 3a). Based on the number of positive or negative droplets for the fluorescent signal of the probes relating to the target sequence and thanks to subsequent statistical calculations, based on the Poisson distribution, the absolute quantity, in copies/μL, of the target gene present in the sample is obtained (Figure 3b)¹¹.

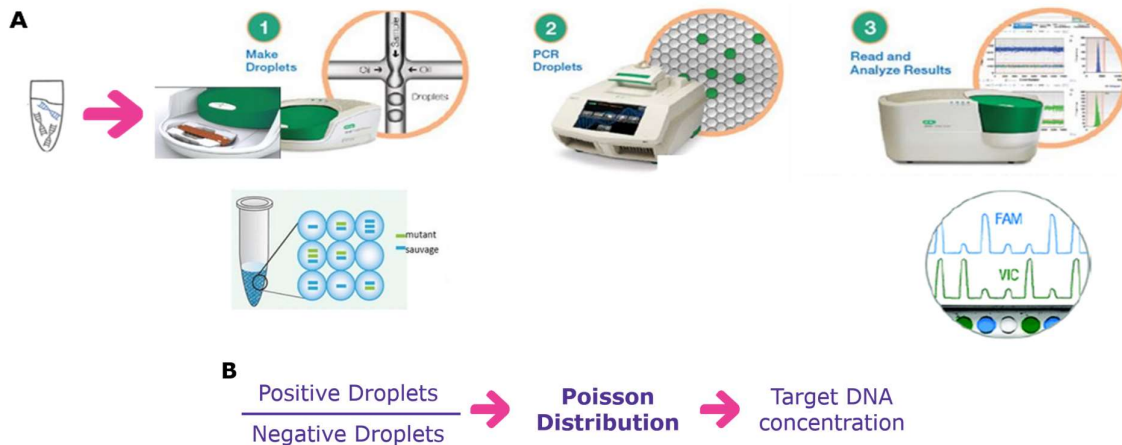


Figure 3: a) schematic workflow for ddPCR experiment. Sample is added to a ddPCR master mix that is partitioned in water-oil emulsion to create 20.000 droplets. Target DNA present in partition is amplified through PCR cycles, and finally positive droplets are read by Droplet Reader instrument, giving the absolute quantification of the target sequence. b) Positive Droplets obtain from the analysis are divided by number of Negative Droplets. Then, thanks to the Poisson Distribution statistics, it's possible to estimate the target DNA/RNA amount present in the sample.

ddPCR was developed to provide high-precision absolute quantification of the nucleic acid target sequence with wide range of applications, spacing from research to clinical diagnostic and monitoring.

In our Merck Laboratory placed in Ivrea we take advantage of the ddPCR technology QX200 (BioRad) for the Gene Copy Number (GCN), which is the number of genes copies that are present in the genome, and RNA expression analysis, in support to both Biological Quality Control group in Ivrea and Cell Line Development group in Vevey facilities.

Analysis of GCN involves determining the number of copies of a given target locus with respect to an invariant reference locus. In this method, the target and the reference gene are amplified using duplex assays and read by Droplet Reader for quantification^{12,13}. From these two quantities, in copies/μL, the number of copies of the target gene is calculated based on the formula¹¹:

$$GCN = \frac{T}{R} \times N$$

Where T is the concentration in copies/μl of the Target gene, R is the concentration in copies/μl of the Reference gene, and N is the number of copies of the Reference locus. Since we are studying molecules expressed in the CHO cell line, which present a diploid genome, N is equal to 2¹¹.

Similar to GCN analysis, evaluation of target mRNA expression involves determining the number of copies of a given target RNA with respect to housekeeping RNA. In this method, both target and reference RNAs are retro-transcribed to obtain the cDNA and directly amplified in the same thermocycling program, using a duplex assay. From the concentration in copies/μL estimated, the ratio between target and reference is calculated from the software, resulting in the expression value of the target gene^{11,14}.

$$ratio = \frac{T}{R}$$

Where T is the concentration in copies/μl of the Target gene and R is the concentration in copies/μl of the Reference gene.

1.2.2. Transgene copy number and expression as predictive genetic marker:

As discussed before, to increase the productivity of the therapeutic protein, different strategies are used to increase the Transgene Copy Number, leading to a heterogeneous cell population, where cells have different numbers of copies of the construct of interest^{10,15}.

It's known that the loss of transgene copies during culturing is one of the major causes of loss of productivity. Causes of loss in transgene copy could be searched for Chromosomal instability that leads to translocation and re-arrangement of the construct¹⁵.

A recent paper shows that high Copy Number is normally related to high productivity, through high production of recombinant mRNA, thus clonal cell lines should be screened and evaluated for GCN to identify the best candidates at the beginning of the clone screening¹⁵.

Nevertheless, the loss of copy number could not be the only reason for the loss of productivity, Loss recombinant protein production could primarily reflect the reduction of transgene expression, leading in a decrease of recombinant mRNA production. Moreover, the transgene expression could be influenced by the chromatin context in which the construct is inserted and chromatin modification. Modification surrounding the insertion site could lead to close chromatin conformation and DNA methylation, making the construct inaccessible to transcription factors and RNA polymerase. Consequently, inaccessibility of the construct leads to a decrease in construct transcription and gene silencing^{8,15}.

Causes of gene silencing events could be found in culture conditions, supplements added to the media and nutrient availability. Cells are constantly adapted to the environment by modifying their expression pattern to maintain their homeostasis, impacting the transgene expression^{5,11}.

For this reason, it's important to evaluate the production of the recombinant mRNA at the beginning of the production, with the aim of selecting the best clone able to grow and produce recombinant protein at high titer in defined conditions (growth in suspension, serum-free media, etc...) ^{3,8,15}.

1.2.3. NGS based methods to identify relationships with best clone features:

Since the advent of sequencing technology, scientists have continued inventing and refining techniques for assessing DNA and RNA sequence.

Today's complex genomic questions demand a depth of information beyond the capacity of traditional DNA sequencing technologies. For this reason, the massively parallel sequencing technology known as next-generation

sequencing (NGS) has revolutionized the biological sciences and has transformed the way scientists think about genetic information. With its ultra-high throughput, scalability, and speed, NGS enables researchers to perform a wide variety of applications and study biological systems³.

Specialized methods are now available for studying Genome Sequence, Sequence Rearrangements, Transcriptome, Methylation patterns, DNA–protein interactions, and more³.

The high throughput makes the NGS technology attractive from Pharmaceutical companies, where could be applied in many fields, spacing from Research and Development, to Manufacturing Process and Biopharmaceuticals Quality Control.

In principle, the concepts behind Sanger sequencing and NGS technologies are similar. In both NGS and Sanger sequencing DNA polymerase catalyzes the incorporation of fluorescent nucleotides one by one onto a growing DNA template strand during a sequential cycle of DNA synthesis. During each cycle, each incorporated nucleotide is identified by its fluorescent tag. While the Sanger method sequence only a single DNA fragment at a time, NGS extends this process across billions of fragments simultaneously¹⁶.

Basically, the Illumina NGS workflow is composed of four phases: Library preparation, Cluster Generation, Sequencing and Bioinformatic Analysis (Figure 4)¹⁶.

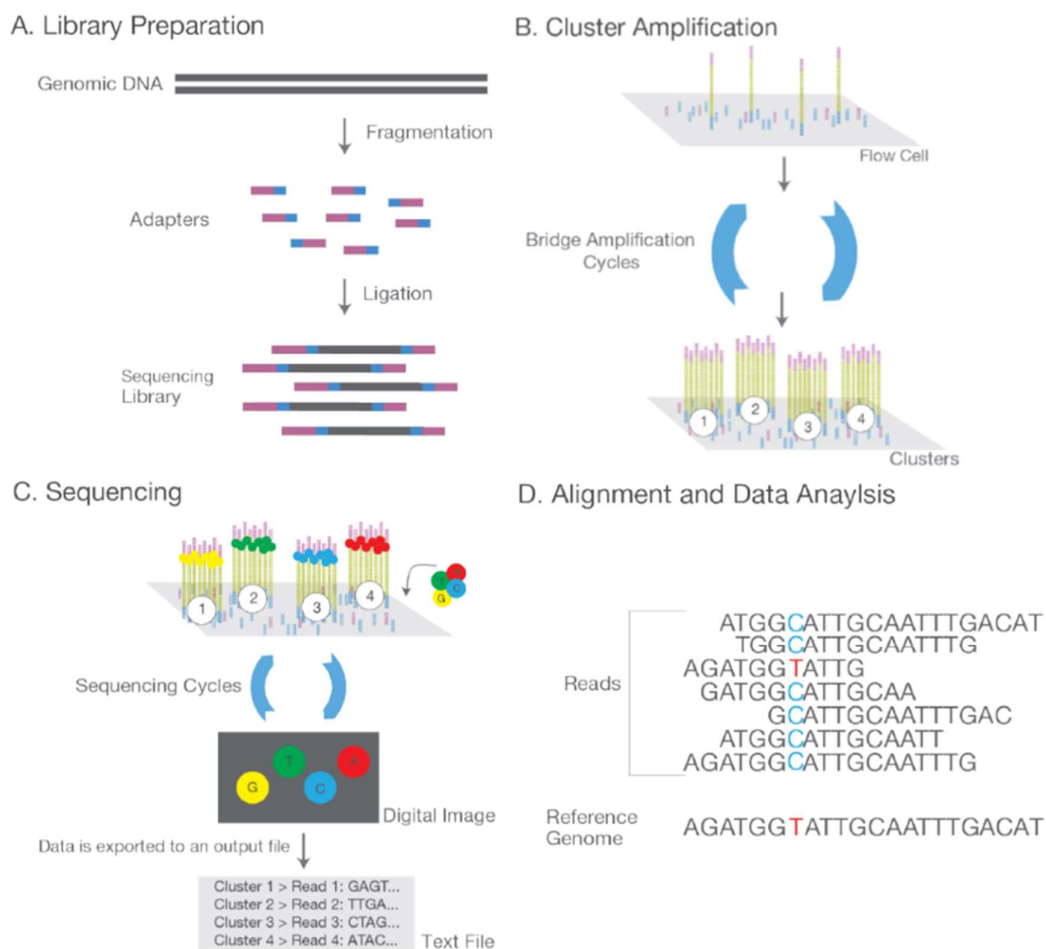


Figure 4: Next Generation Sequencing workflow includes four steps: **a)** gDNA or cDNA samples are fragmented and ligated, at both ends, to adapters and index sequences to prepare NGS library. **b)** Library are loaded onto flow cell, where fragments are captured from oligos present on the flow cell surface. Then, libraries are amplified into clonal clusters by bridge amplification. **c)** Fluorescently labeled dNTPs are incorporated onto the DNA template while fluorescence signal is detected for every single cluster. Sequencing cycles are repeated to obtain the sequence called "read". **d)** Reads obtained from a sequencing run are aligned to the reference genome for further bioinformatics analysis¹⁶.

The Library preparation step prepares DNA or RNA samples to be processed with a sequencer: genomic DNA or cDNA derived from retrotranscribe RNA is randomly fragmented while adapter sequences are attached to the 5' and 3' end of each fragment. In the Illumina sequencing workflow, these adapters contain complementary sequences that allow the DNA fragments to bind to the glass support for sequencing, called flow cell. Fragments derived from the same sample are uniquely identified using index sequences, or “barcodes”, that are added to each fragment of the sample. Now the fragmented and ligated DNA is called library. To save resources, multiple libraries can be pooled together and sequenced in the same run, a process known as multiplexing, where barcodes are used to distinguish between the libraries during data analysis¹⁶.

Following, libraries are loaded onto the flow cell for the sequencing run. During the Cluster Generation phase, fragments are hybridized to oligo sequence present on the flow cells, complementary to an adapter sequence. Then, each fragment is amplified into a distinct clonal cluster through bridge amplification. Subsequently, when cluster generation is completed, single-stranded DNA templates are ready for sequencing. Illumina platform uses the Sequencing By Synthesis (SBS) technology (Figure 5): at each sequencing cycle, the DNA polymerase catalyzes the incorporation of a single reversible labeled dNTP. Once dNTPs are incorporated, the flow cell is imaged, and the fluorescence emission is recorded for each cluster. Next, the fluorescence terminator is cleaved to permit a new sequencing cycle and incorporation of a new single dNTPs. Those sequencing cycles are repeated to create a sequence called “read”. After reading the forward DNA strand, the reads are washed away, and the process repeats for the reverse strand. This method is called paired-end sequencing¹⁶.

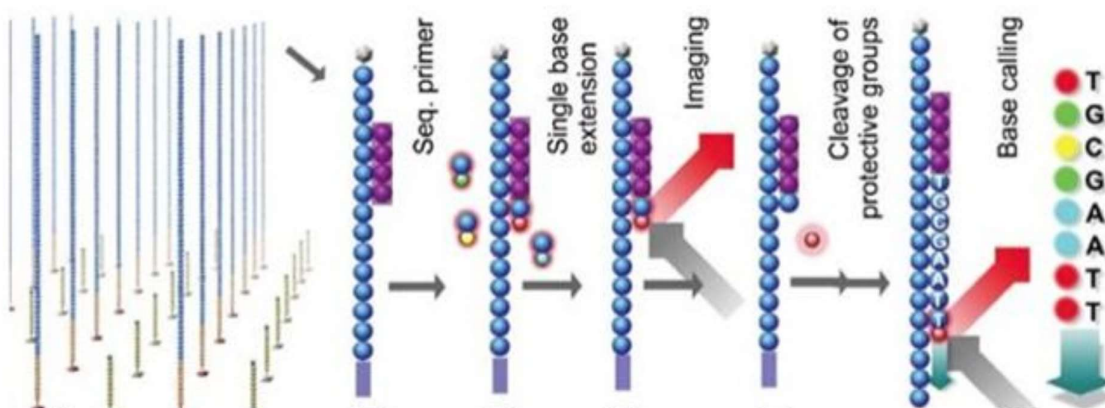


Figure 5: Illumina Sequencing By Synthesis: once a cluster is created, a sequencing primer hybridizes to the DNA template and DNA polymerase starts to catalyze dNTP incorporation. One single nucleotide is incorporated for each cycle thanks to a fluorophore, which also acts as a terminator. After imaging of the flow cell, the fluorophore terminator is cleaved away to permit the incorporation of the following dNTP in the next sequencing cycle. Finally, all fluorescences detected are collected into sequencing reads¹⁶.

After sequencing, the instrument software identifies nucleotides (a process called base calling) and the predicted accuracy of those base calls. The resulting “reads”, are aligned to reference genome for different kinds of bioinformatic analysis pipeline, such as mutation detection, phylogenetic analysis, read counting for RNA-Seq method or methylation analysis¹⁶.

Illumina offers a wide range of instruments for scalable and flexible analysis in terms of cost and quantity of generated data. MiSeq, NextSeq500 and NovaSeq6000 are the Illumina platform currently used in Merck Ivrea site (Figure 6). These three instruments covered wide-range of applications, spacing from viral genomic sequencing, amplicon sequencing that requires low data output, transcriptome analysis, whole-genome sequencing and DNA methylation analysis¹⁶.



Figure 6: The Illumina platform, present in Merck Ivrea site: MiSeq, NextSeq500 and NovaSeq 6000. MiSeq has a low data output, up to 50 million reads per run, allowing sequencing of small genome or amplicon. NextSeq has a major data output, up to 400 Million reads per run, allowing sequencing of exome and transcriptome. NovaSeq 6000 has the greatest data output, up to 10 billion reads per run: such high output allow researcher to sequence different whole-genome sequencing or DNA methylation analysis.

The Illumina Sequencer NextSeq500 has a fast and integrated workflow that allows sequencing of exome and small genome sequencing and transcriptome sequencing in a single run. Up to 400 Million reads are generated in little more than 12 hours for 75 cycle sequencing run. The NextSeq500 has highly accurate SBS chemistry, due to two-color sequencing, used to simplify fluorescent nucleotide reading. Clusters are identified using red and green filters: green clusters are interpreted as T (Thymine), while red clusters are interpreted as C (Cytosine). Clusters observed both as red and green are identified as A (Adenine), while clusters without any signal are interpreted as G (Guanine)¹⁶.

As the NextSeq500, the Illumina Sequencer NovaSeq6000 uses the two-color based sequencing, however, it also has integrated the patterned flow cell technology, in which billions of nano-wells to increase cluster generation and finally delivers a significant increase in sequencing reads and a total output of the system (Figure 7)¹⁶.

Instruments like the MiSeq and NextSeq500 use non-patterned flow cells where clusters are generated and spread over all the surface. Therefore, these instruments require more time to identified and map sequencing clusters. The NovaSeq6000 advanced flow cell uses distinct nano-wells at fixed locations for cluster generation to make more efficient use of the flow cell surface area and save time.

Since the clusters can only form in the nano-wells, the flow cell is less susceptible to overloading and reduces the chance of multiple library fragments amplifying in a single cluster. In other words, patterned flow cells allow DNA clusters originating from a single DNA template, therefore it increases the final output of each sequencing run¹². The NovaSeq6000 output generates up to 20 billion reads and supports multiple read lengths. For this reason, this instrument supports several applications, such as whole-genome methylation (Whole Genome Bisulfite Sequencing), Chromatin Immune-precipitation sequencing (ChIP-Seq), transcriptome and many others¹⁶.

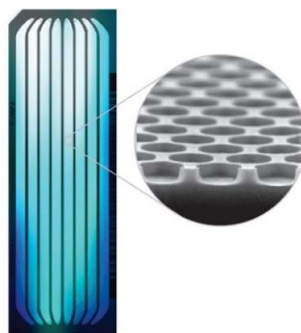


Figure 7: Patterned flow cell in which each lane is covered by billions of nano-wells at fixed location to allow monoclonal cluster generation and faster cluster identification. This advanced technology finally increases sequencing reads and the total output of the system.

As already indicated in the previous paragraph, Biopharmaceuticals production requires cutting-edge and high throughput technology to identify the best clone for MCB at early stages.

The Next Generation Sequencing approach offers both high-throughput and accuracy in analysis, allowing application such as transcriptomic and epigenomics studies.

1.2.4. Possible use of Transcriptomics data in best clone selection:

Transcriptome sequencing allows a snapshot of the whole RNA molecules rather than a predetermined subset of genes, like in qPCR studies. Moreover, while qRT-PCR can only detect known sequences, RNA sequencing (RNA-Seq) using NGS can detect both known and novel transcripts. Since RNA-Seq does not require any prior knowledge about the sequence and consequently does not require predesigned probes, the data sets are unbiased, allowing for hypothesis-free experimental design^{3,17}.

All RNA-Seq methods are based on the use of a reverse transcriptase to convert the RNA to cDNA. Two techniques are used in this first step: random primers or oligo(dT) primers for this reaction. Oligos(dT) are used to isolate mRNA from the whole transcript, thanks to the annealing of their T stretch to the mRNA polyA. Random primers are used in the case of whole RNA sequencings such as mRNA, miRNA and lncRNA³. RNA-Seq quantifies transcripts abundance using read-counting methods, where reads obtained from sequencing are aligned to a reference sequence. For this reason, having access to a high-quality reference genome is crucial in Transcriptomic analysis^{3,17}.

In the past three decades, many pharma companies look at the transgene mRNA expression as marker for biopharmaceuticals production, because the decrease in transgene mRNA expression directly affects the final protein synthesis. However, many reports show the importance of studying RNA expressions profile in order to better understand how gene expression could finally regulate clone productivity and cell growth¹⁸.

With the publication of *Cricetulus griseus* and CHO-K1 genome in 2011, including an annotated RefSeq genome, transcriptome studies have made great progress^{3,17}.

Transcriptomic data published during this year gathered around differential expression under different culture conditions (temperature shift, continuous manufacturing, chemical treatment, etc...), or at different time-point^{18,19,20,21,22}.

The increase of high-throughput technologies inherently generates a big plethora of transcriptomic data, that were collected in different databases, including NCBI and UniProt^{3,17}.

Transcriptomics studies in CHO cells provide information about the abundance of whole RNA transcripts, leading to functional class analysis on several genes instead of expression analysis of individual genes. During this year, in fact, many papers were published looking to transcriptome data using functional analysis revealing that genes involved in protein processing, secretion, signaling pathways and cell cycle were differentially expressed upon different treatment. Moreover, similar functional categories are altered in high and low producers cell clone comparison^{18,19,20,21,22}.

Interestingly, it was found that recombinant protein RNA was already at high levels and had a little increase upon Methotrexate (MTX)-induced amplification of DHFR. Consequently, it was hypothesized that the amplification step selects those cells who have high expression of genes that are required for protein production and secretion¹⁸.

All of those transcriptome data sets were combined to generate a hyper-productivity gene data set. The majority of genes are involved in protein synthesis, metabolism, cell cycle and transcription regulation (Figure 8)¹⁸.

This knowledge leads to awareness that to increase clone productivity, not only the recombinant protein expression should be enhanced, but all of the genes required for cell production machinery should be highly expressed¹⁸.

In addition, functional category studies on transcriptome data are also useful to analyze the glycosylation pattern of the final product^{18,19}.

By this way, Transcriptome analysis could be used to identify a predictive biomarker that correlates with good productivity, cell growth, high-quality product, or to recognize dysregulations that could affect the final productivity of clones produced during the early stage of the manufacturing process¹⁸.

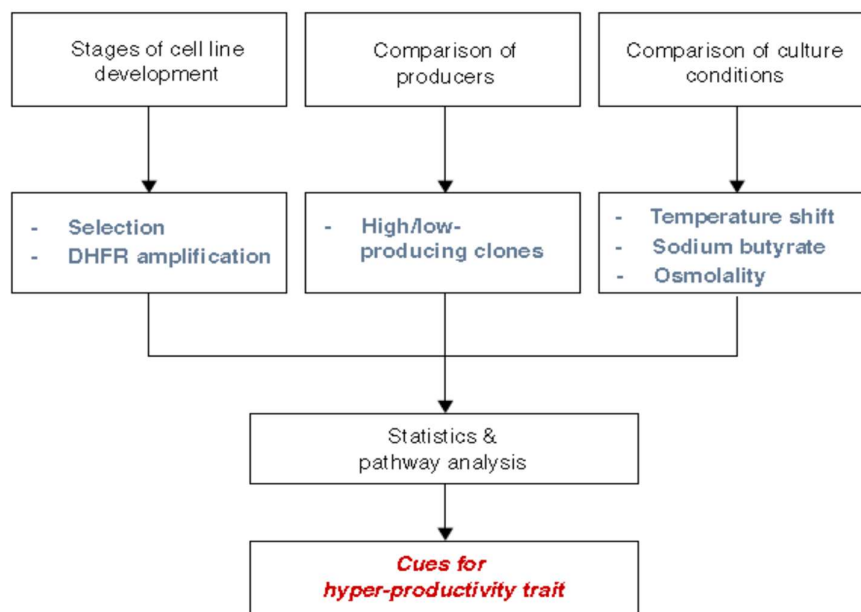


Figure 8: A schematic representation of meta-analysis of transcriptome data: several studies on differential expression shows difference at RNA level between Selection or non-selection process, between high and low producer clones and between different cell culture conditions. Combining differential expression analysis, a pivotal set of genes are identified for hyper-productivity traits¹⁸.

Finally, using a large set of Transcriptomic data it's possible to identify pivotal genes for CHO cell engineering. Zing Finger Nuclease (ZFN), Transcriptional activator-like effector nuclease (TALEN) and CRISPR/Cas9 technology have already been applied to knock out genes such as fut8 and mgat1, or genes involved in apoptotic pathways, such as bak and bax^{3,7,18,19}.

While Transcriptomic data greatly facilitate the development in the early production stage of the manufacturing process, the exploration of the CHO cell epigenome will lead to a deeper understanding of gene expression regulation, affecting cell productivity and stability^{18,19}.

1.2.5. DNA methylation could regulate expression by inducing high rate of productivity:

It was already known that transcript expression level is regulated by DNA accessibility e chromatin remodeling. Epigenetics indicates changes in gene function and expression that are mitotically and/or meiotically heritable without altering the DNA sequence. Therefore, Epigenetics studies comprise the study of histone post-translational modification, chromatin remodeling and DNA methylation²³.

DNA sequence could be methylated at Cytosine residues, regulating chromatin conformation and consequently gene expression level²³.

DNA methylation is known since the 70s when it was shown to contribute to X chromosome silencing in humans. DNA methylation takes place at position 5 of Cytosine residues (5-methyl-Cytosine, 5mC) in the context of CpG dinucleotides^{23,24}.

In the majority of the mammalian genome, CpG dinucleotides are relatively low, but are particularly dense in specific regions called CpG islands (CGI): CGIs are defined as regions more than 500bp long with at least 55% of CG content²⁵. CGI are found in the context of regulatory regions (promoter and enhancer), however, 70% of CGIs overlap with promoter regions with the function to induce gene expression or silencing (Figure 9). Methylated CGIs are usually found in CpG poor promoter and correlates with gene silencing, while CpG dense promoter are often unmethylated and correlate with gene expression (Figure 9)²⁵.

Most of the mammalian genome shows a relatively low level of CGIs methylation, while small part of CGIs are methylated, indicating a bimodal distribution of 5mC across the whole genome^{26,27}.

Patterns of DNA methylation have high information content about the ongoing activity of transcription factors and thus can help to understand specific aspects of gene regulation. To explain DNA methylation repressive function, in fact, CpG methylation level should be recognized by functional Methylation Binding Domain (MBD)- containing protein, called DNA methylation readers, which allow positioning of a repressive transcription factor^{23,24}.

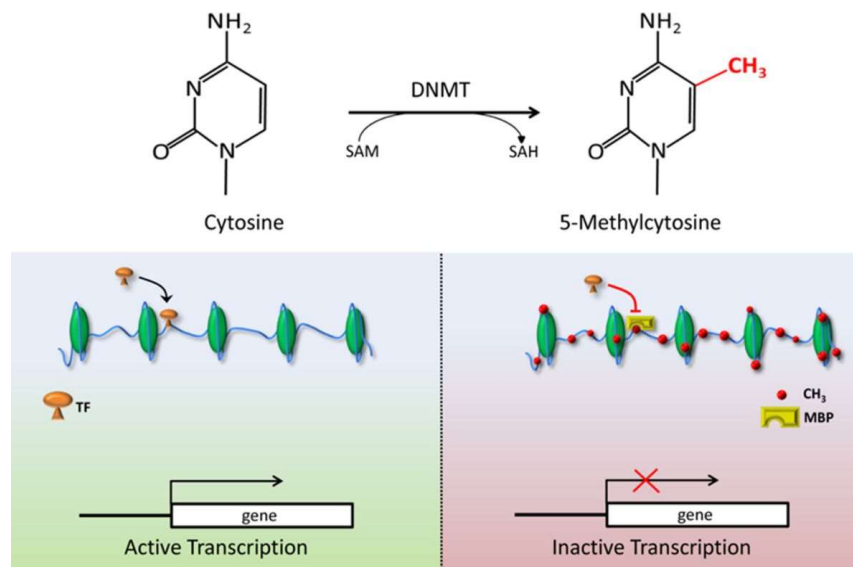


Figure 9: Cytosine could be methylated at position 5 of the pyrimidine ring. The 5-methylcytosine (5mC) is associated with gene repression when founded at level of regulatory region (promoter and enhancer). In fact, 5mC is read by Methyl Binding Domain (MBD) protein that in turn recruit repressive transcription factor, leading to close chromatin conformation (red part of the picture). On the contrary, unmethylated cytosines are associated with actively transcribe genes and open chromatin conformation, since Transcription Factor (TF) and the Transcription machinery are recruited. (green part of the picture)²⁵.

In addition, to the common 5mC, another type of modification at the same position are also been observed. One of the most important is the 5-hydroxymethylcytosine (5hmC). During active DNA demethylation, the 5mC is oxide to produce 5hmC with other intermediate states, such as 5-formylcytosine and 5-carboxilcytosine, to come back to native unmethylated cytosine (Figure 10A)^{28,29}. Several studies indicate that the 5hmC does not act simply as an intermediate state, but it has a functional role in gene regulation. While 5mC found at the level of regulatory regions correlates with gene repression, the 5mC is poorly found at the

level of gene bodies, where is correlated with gene expression. On the contrary, the 5hmC is mostly found in gene bodies of active genes (Figure 10B). Moreover, it was unclear whether the 5hmC plays a role in regulating transposon activity^{29,29}.

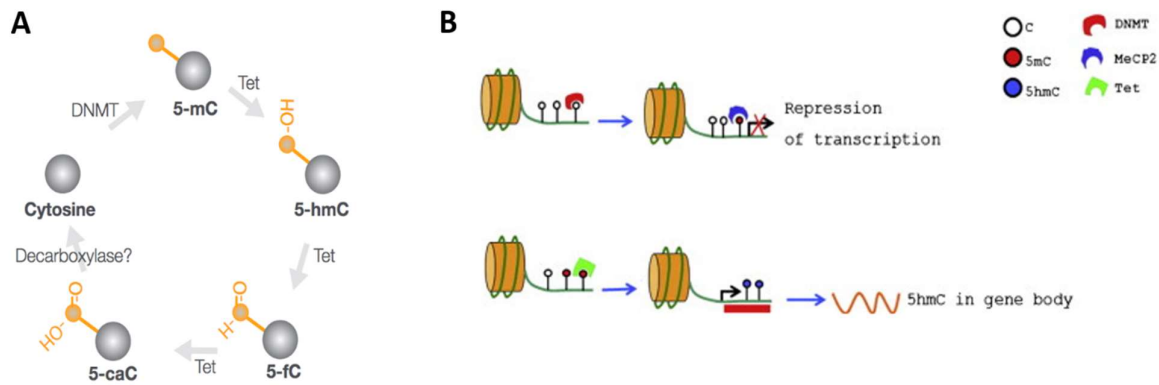


Figure 10: a) Mechanism of active DNA de-methylation. First, the TET enzyme oxidizes 5mC to 5-hydroxymethylcytosine (5hmC), then the 5hmC is converted to 5-formylcytosine and 5-carboxylcytosine to come back to unmethylated states. **b)** While 5mC is correlated to gene repression, 5hmC is correlated to gene expression. 5hmC is in fact related to gene bodies of actively transcribed genes^{28,29}.

Genome-wide mapping of DNA methylation at single-base resolution in humans, in particular in embryonic stem cells, revealed DNA methylation also in a non-CG context. It was observed that DNA methylation could occur also in CHG or CHH context, where H could be A or T nucleotide; however, 5mC in these contexts accounts for only 0.02% of total DNA methylation in the adult human genome²⁵.

Moreover, DNA methylation does not occur exclusively in CpG islands. The so-called CpG island shores refer to regions with a lower content in CG, which is found to be almost a 2kb long region. They are normally found up to 2 kilobases upstream from CGIs. As CpG island, the CpG island shores methylation is associated with transcriptional repression of downstream genes, and it was shown that most of the tissue-specific DNA methylations are found at CpG island shores instead of CpG island (Figure 11)²⁵.

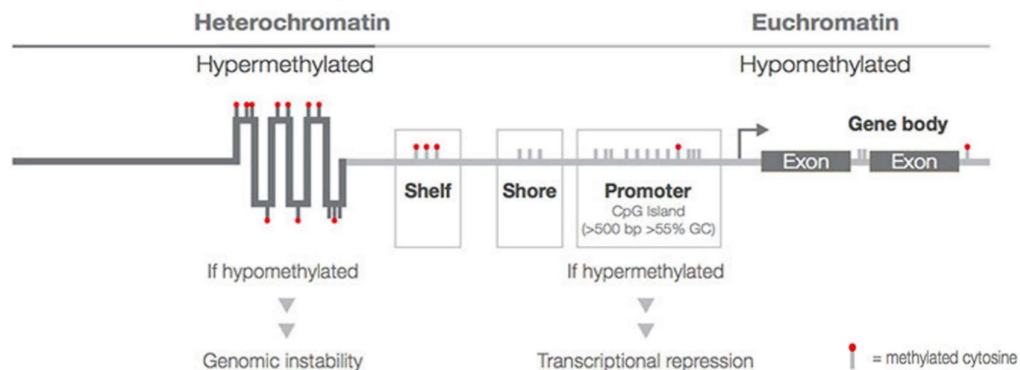


Figure 11: DNA Methylation occurs at Cytosine residue, mostly in the context of CpG island. 70% of the CpG island overlap with regulatory regions, such as promoter or enhancer. When CGIs are hypomethylated, the downstream genes are expressed, while hypermethylation of CpG island leads to transcriptional repression. 5mC could be also found at the level of gene bodies, where it's correlated to gene expression. Heterochromatic regions show global DNA hypermethylation. Hypomethylation of these regions leads to genomic instability²⁵.

While the advent of NGS and completion of the Human Genome project in 2001 enables epigenetics studies in humans, only a few decades ago the scientific community address epigenetic phenomena to

biopharmaceuticals production, which is in part due to the fact that the CHO genome is available since 2011^{3,26,30}.

DNA methylation studies in the CHO cells are emerging as a crucial element to better understand molecular basis CHO cell-based biopharmaceutical production, both at the level of single-gene and genome-wide methylation^{3,15,26,29}.

Many papers show that DNA methylation contributes to loss in clone productivity of monoclonal antibodies in CHO process. Initially, several studies address the loss of productivity and clone instability to DNA methylation at the level of CMV promoter or regulatory element of the transgene sequence^{31,32,35}. Then, it was clear that specific DNA methylation patterns in the CHO cell line could impact clone productivity, growth and expression stability^{15,33,34,3,25,26,27}.

A first DNA methylation landscape was studied in CHO DP-12 cells where global DNA hypomethylation was found (61%) compared to other mammalian genomes or embryonic stem cells. However, with respect to other mammalian genomes, there is a high percentage of CpG islands that are highly methylated (55%)^{26,30}. The analysis of genomic features revealed that CGIs and UTRs region are less methylated than gene bodies, which are predominantly methylated, in accordance with other mammalian genomes^{26,30,36}. Moreover, these studies revealed genomics regions with only partial DNA methylation, which are called Partial Methylated Domains (PMDs). Only a minority of all annotated genes overlap the PMDs and weakly or not expressed genes are enriched in the Gene Ontology (GO) category related to neurological and synaptic process, indicating that PMDs methylation could be related to tissue-specific gene regulation^{30,36}. On the contrary, highly expressed genes are over-represented in GO category related to cell growth and productivity, such as cell cycle, energy metabolism, transcription and translation^{30,36}. All of those findings are also supported and confirmed by transcriptomics studies, as discussed in the previous paragraph (1.2.4).

Genome-wide DNA methylations were also analyzed under different culture conditions revealing low differentially methylated regions with respect to different cell passaging or during batch cultivation. On the contrary, high differentially methylated regions were found in response to subcloning or media changes. These findings indicate that DNA methylation changes play an important role in the phenotype characterization of clones^{8,15}.

1.2.6. Chromatin rearrangements and histone marks could regulate expression by inducing high rate of productivity:

The genome is organized into a complex nucleoprotein structure, called Chromatin, that permits DNA packaging in the cell nucleus. The nucleosome is the fundamental repeating chromatin subunit comprised of eight histones encompassed by 146bp of DNA. Each nucleosome is linked to the next by 50bp DNA linker. The histone octamer is composed of two copies each of histones H2A, H2B, H3 and H4, while the histone H1 is involved in higher chromatin folding and structure (Figure 12)²⁵.

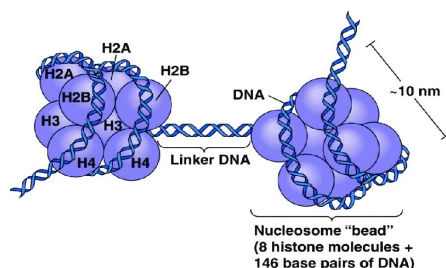


Figure 12: Schematic nucleosome structure. A nucleosome is composed by two copies of each core histone protein (H2A, H2B, H3 and H4) and 146bp of DNA helix.

Higher-order chromatin folding is organized in transcriptionally active (euchromatin) or silent regions (heterochromatin). At the chromatin level, different epigenetic events could affect the transcriptional repression of the transgene. In transcriptionally active regions, chromatin takes an open conformation to permit accessibility to transcription factors and transcriptional machinery and transcribe genes. On the contrary, in Transcriptionally silent regions, chromatin takes close and compact conformation, where high histone protein density and presence of repressive transcription factors, that could be recruited by DNA methylation, don't permit the accessibility to the transcriptional machinery. Therefore, nucleosomes are directly responsible for modulation in DNA accessibility to regulatory protein and genes activation or repression (Figure 13)^{25,37}.

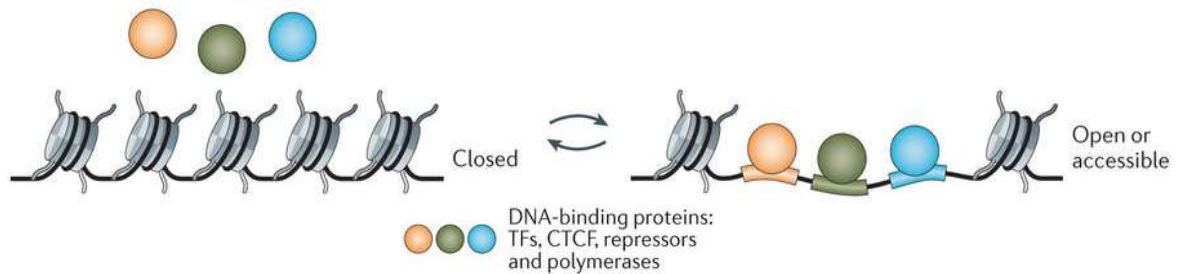


Figure 13: Chromatin accessibility. Close chromatin conformation, with a high density of nucleosomes, does not permit Transcription Factor binding and consequently, transcription machinery is not recruited. On the other hand, open chromatin conformation permits the Transcription Factor binding and recruitment of Transcription machinery.

Replacement of nucleosomal protein with histone variants or Post-Translational Modification (PTM) of amino-acids residues in the N-terminals of histones, which is extruded by nucleosome, can affect epigenetic regulation of chromatin structure and gene function (Figure 14)³⁷.

Histone PTMs include acetylation, methylation and ubiquitination on lysine residues (K), methylation and citrullination on arginine residue, and phosphorylation on serine-threonine and tyrosine³⁷.

In particular, methylation and acetylation on specific lysine residue can determine Chromatin state by directly influencing structures or serve as signals to readers proteins of histone modification. Major methylation and acetylation sites on histone H3 comprise K4, K9, K27 and K36. All acetylation markers are correlated with transcriptional activation, being localized to Transcription Start Site (TSS) and enhancer of potentially actively transcribed genes^{37,38}. On the other hand, the correlation between methylation on H3 with gene expression depends on residue and on its level. In fact, Lysine methylation can occur at three different levels, by replacing hydrogen on the primary amine, therefore lysine can carry mono-methylation (me1), di-methylation (me2) and tri-methylation (me3) (Figure 14)^{37,38}.

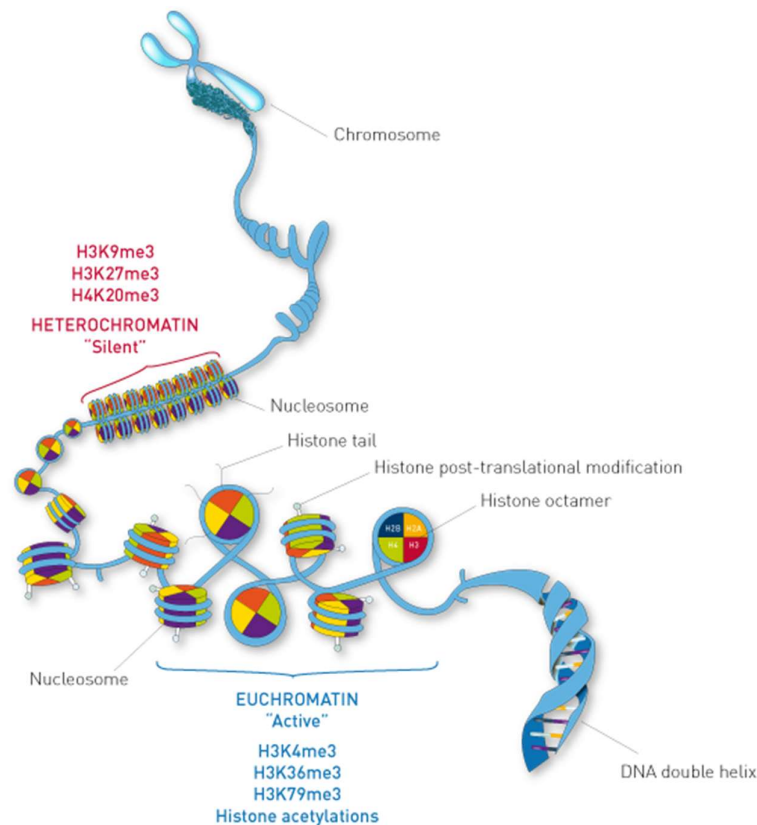


Figure 14: The general structure of chromatin. Silent chromatin or Heterochromatin is marked with its specific repressive histone marks, with a higher density structure. On the contrary, active chromatin or euchromatin is marked with activating histone marks and the higher structure takes an open conformation to accommodate transcription factor and transcription machinery^{37,38}.

Although histone modifications are not necessarily prerequisite codes, they function as positively and/or negatively regulators of gene expression and they may serve as good epigenetic indicators of open or close chromatin state in specific genomic regions. Recently, it was clear that six emerging classes of histone modification could be identified for epigenome profiling by the International Human Epigenome Consortium³⁸.

The six classes of histone modification are associated with different genomic features, such as TSS of actively transcribing genes, enhancer or gene body. Thanks to the analysis of histone modification occupancy, it's possible to identify actively transcribed genes, active enhancers or repressed genes.

Among these histone modifications, trimethylated H3K4 (H3K4me3) is a marker associated with actively transcribed genes and it's enriched around the TSS. Moreover, TSS of actively transcribed genes shows also enrichment in the H3K27 acetylation (H3K27ac) (Figure 15)³⁸.

The same histone mark, H3K27ac, is also a marker of active enhancer, with the monomethylated H3K4 (H3K4me1) (Figure 14). Co-occupancy of these last two histone marks is a better indication of active enhancers since H3K27ac could be found both at the level of TSS and enhancer³⁸.

Trimethylated H3K36 (H3K3me3) is associated with gene bodies of actively transcribed genes, and it's distributed along the decondensed open chromatin regions (Figure 15)³⁸.

Repressed genes could be indicated by two different histone marks, trimethylated H3K9 (H3K9me3) and H3K27 (H3K27me3), however, they indicate two different mechanisms of silencing and are mutually exclusive and do not exist in the same locus (Figure 15). H3K9me3 is normally associated with constitutive heterochromatin, where gene repression is stable, also marked with DNA methylation and repressive

Transcription factor. H3K27me3 instead is associated with facultative heterochromatin, where gene repression could be reverted based on environmental conditions and cell differentiation³⁸.

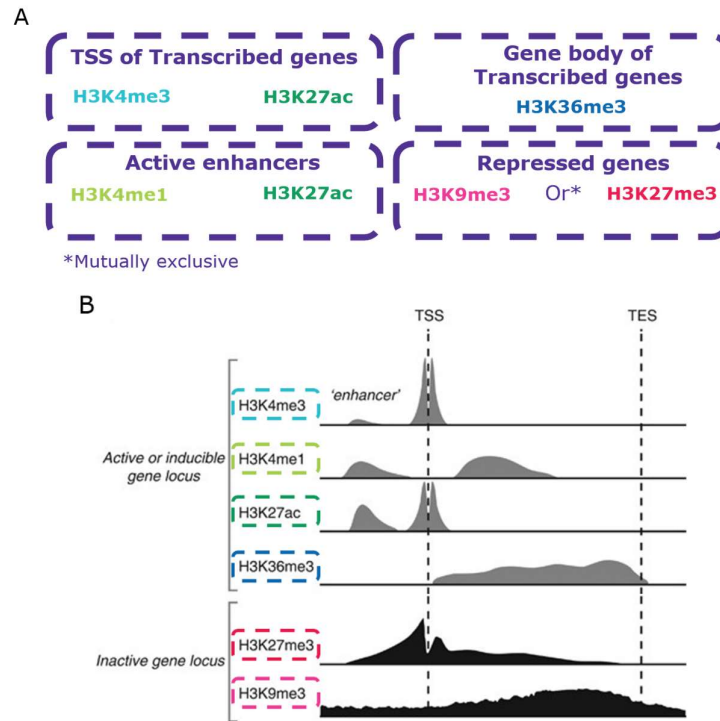


Figure 15: Six emerging classes of Histone H3 modifications that influence the human epigenome profile. **A)** Transcription Start Site (TSS) of actively transcribed genes are marked by H3K4me3 and H3K27ac, while active enhancers are marked by co-occupancy of H3K27ac and H3K4me1. Gene bodies of actively transcribed genes are associated with H3K36me3. Repressed genes are marked with H3K9me3 or H3K27me3 in a mutually exclusive manner. **B)** Distribution of histone mark modifications along with genes. TSS: Transcription Start Site; TES: Transcription end site³⁸.

Although many of work was already done for the human epigenome, and a lot of studies indicate the direct relationship between chromatin conformation and gene transcription, these mechanisms and relationships have been recently studied for the CHO cell line and bioproduction^{26,27}.

It was recently shown that the bioproduction process could be affected by epigenetic mechanisms and chromatin states¹¹. Some studies indicate the loss in monoclonal antibody productivity in CHO cell line by a decrease in H3 acetylation, while in follow-up study was shown that the enrichment of stable and good producer clones is possible selecting those clones with the highest level of H3 acetylation²⁷. Therefore, a lot of studies were performed to demonstrate the use of the Ubiquitously Acting Chromatin-Opening Element (UCOE) vector for transgene transfection, which is able to maintain stable transgene expression by avoiding repressive histone marks and keeping open chromatin conformation^{3,7}.

Other studies revealed that CHO cells could adapt to short-term and long-term changes by different but interacting mechanisms, that include DNA methylation and the presence of the 6 histone marks.

All of these recent publications indicate the need of deep investigation of chromatin state regulation and histone marks analysis, in correlation with respect to Transcriptome and Methylome data.

2. Aim of the Project and Experimental Design

In this study, CHO-K1 SV cells are used as host cell lines and engineered for the production of a mAb.

Construct designed to codify the mAb comprise Heavy and Light Chain sequence driven by two separate strong promoters. Moreover, the same construct contains the selection marker, linked to Light Chain with an IRES (internal ribosome entry site) sequence, that allows the co-expression of different genes under the control of the same promoter³⁹.

Cell Line Development group, based in the Vevey facility, took care of the transfection of the gene of interest and steps of selection and expansion of the heterogeneous cell population. The Cell Line Development group also deals with single-cell sorting and culture of clones in Fed-Batch flask, that mimic the cell growth in a bioreactor. During Fed-Batch culture in the Clone screening phase, clones are firstly evaluated for their viability, stability and productivity. However, these analyses are not sufficient to clearly identify the Best Clone for the manufacturing process. Moreover, during the early phase of the production, there isn't enough material for quality analysis of the final product while long-term production stability in bioreactors could not be predicted.

For these reasons, this project would investigate the effectiveness of genetic parameters and molecular biology technologies during the Clone Screening phase. With the purpose to help in the selection of the best clone, this project would set the stage for the use of molecular biology analysis and high-throughput analysis, such as the Droplet Digital PCR and the Next Generation Sequencing, during production phases. Importantly, this project has the aim to gain more knowledge on what distinguished “Good Producer clone” from “Bad producer clone” through the analysis of Copy Number, Transcriptome, Methylation and Chromatin region analysis.

In this study, the productivity of different clones, in terms of mAb titer, is evaluated through PA-HPLC and iQue technologies by Cell Development Group. Results are reported in Figure 15, where the productivity of the clones is reported on the y-axis as mg/L. Each clone is analyzed using three biological replicates cultured in a Fed-Batch flask.

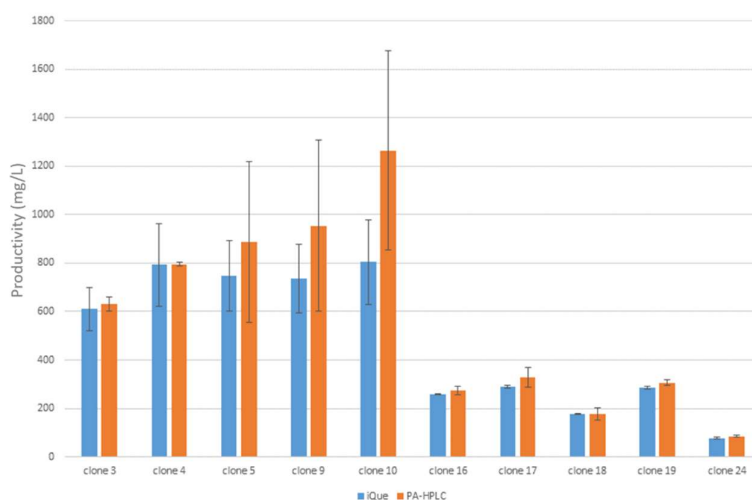


Figure 16: Several clones are evaluated for their productivity in terms of mg/L of mAb produced using three biological replicates. As shown in the graph, Productivity is evaluated by 2 different technologies: iQue (blue) and PA-HPLC (orange). The different clone are reported on the x axis, while the productivity (mg/L) is reported on the y axis. Data are represented as Mean with Standard Deviation.

Based on productivity data, 4 clones are selected for this study: 2 clones that exhibit high productivity and 2 clones that exhibit low productivity. Clones 3 and 9 are selected as “Good Producer clones”, while clones 18

and 24 are selected as “Bad producer clones”. Other clones are discarded from further analysis due to high standard deviation between biological replicates and evidence of non-clonal population (data not shown). In table 1, candidate clones for this study are reported with their productivity characteristic. Based on clone productivity data, Clone 24 is defined as “Worst clone”, while Clone 18 is defined as “Bad clone”. Despite Clone 9 showing higher productivity respect for Clone 3, these two clones are defined as “Good clone” and “best clone” respectively. In fact, Clone 9 shows expression instability later in culture, probably due to clipping phenomena on the final protein (data not shown).

Table 1:
Scheme of clones analyzed

Clone ID	Feature	Category
Clone 3	Best Clone	Good Producer Clones
Clone 9	Good Clone	
Clone 18	Bad Clone	Bad Producer Clones
Clone 24	Worst Clone	

In this study, we carry out experiments using three biological replicates (Fed-Batch culture) for each clone analyzed. Every sample is tested for Transgene Copy Number evaluation, Target mRNA expression, Transcriptome analysis, Methylome analysis and Chromatin analysis (Figure 16).

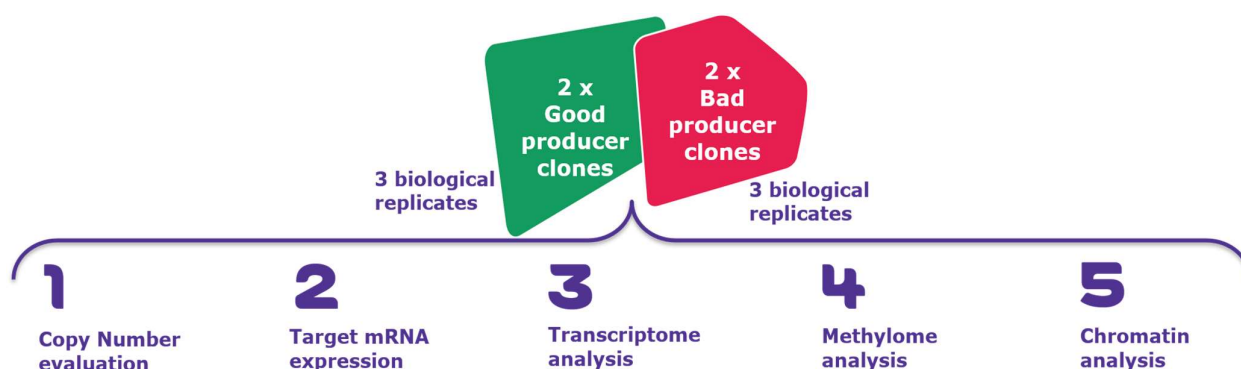


Figure 16: Picture shows the experimental design used for this study: two “Good Producer Clones” and two “Bad Producer Clones” are analyzed using three biological replicates each. Each clone is evaluated for Transgene Copy Number, target mRNA expression, Transcriptome and Methylome analysis.

As a result of these analyses, it’s possible to elucidate the relationship between the Transgene copy number and the Transgene expression with productivity. Moreover, Transcriptome analysis could help in the identification of molecular markers and patterns of expression related to productivity, as secretory or translational pathways. Finally, Methylome analysis could clarify the impact of DNA Methylation and Chromatin environment on clone productivity, mechanism of gene silencing events and expression stability.

3. Materials and Methods:

3.1. DNA and RNA extraction:

Clones derived from CHO-K1 SV host cells, with integrated construct, are cultured in Fed-Batch. For each clone analyzed (Clone 3, 9, 18 and 24), 3 biological replicates are prepared from the Cell Line Development group, based in Merck Vevey site.

1x10⁷ cell pellets are prepared for each biological replicates of each clone and stored at -80°C until DNA and RNA extraction.

Genomic DNA is extracted from cell pellets using the QIAmp Blood Mini Kit (51104, QIAGEN Group). Each cell pellet is split into 3 columns for DNA extraction and then processed as manufactured instruction. DNA derived from each column is finally eluted in 150µl of Molecular Biology water (W4502, Sigma-Aldrich). Elute DNA derived from the same biological replicates are pooled and quantified using NanoDrop ONE instrument and Qubit 2.0 fluorimeter (Life Technologies). Qubit quantifications are performed using the kit dsDNA Broad Range (Q32850, Life Technologies).

All samples show 260/280 (purity respect to protein content) and 260/230 (purity respect to organic compound) NanoDrop ratios ranging from 1,8 and 2,2.

Total RNAs are extracted from cell pellets using the RNeasy Mini Kit (74104, QIAGEN Group). Each cell pellet is split into 2 columns for RNA extraction and then processed as manufactured instruction. Total RNA is finally eluted in 30µl of Molecular Biology water (W4502, Sigma-Aldrich). RNA derived from the same biological replicates are pooled together and quantified using NanoDrop ONE instrument and Qubit 2.0 fluorimeter (Life Technologies). Qubit quantifications are performed using the kit RNA Broad Range (Q10210, Life Technologies).

All samples show 260/280 and 260/230 NanoDrop ratio ranging from 1,8 and 2,2.

RNA integrity number (RIN) is checked using a Bioanalyzer instrument (Agilent) with the RNA Nano kit (5067-1511, Agilent). All samples have RIN > 8.

3.2. Gene Copy Number determination using ddPCR:

Gene Copy Number (GCN) is determined for all samples using the QX200 Droplet Digital PCR system (BioRad). The ddPCR reaction mix is prepared using the ddPCR Supermix for Probes (no dUTP) as manufactured instruction (1863024, BioRad). Each set of primer is used at the final concentration of 900nM, while TaqMan Probes are added at 250nM as final concentration. Target assays are called LC and HC while sequence assay related to Reference genes, Ap3b1 and B2M, are reported in Table 2.

Table 2:

5'-3' sequence of primer and probes for ddPCR experiment

Assay name	ID primer	Sequence 5'-3'	Fluorescent Probe
B2M	B2M for	ACCCACTGCGACCGATAAAAT	VIC
	B2M rev	GACCTTGGGCTCCTTCAGAGT	
	B2M probe	TGCCTGCAGAGTTACA	
Ap3b1	Ap3b1 for	TCCTTGGTGGGTGGTGT	VIC
	Ap3b1 rev	CCGCAGCCAGCCAAAA	
	Ap3b1 probe	CTGGTTGCTACACTTGG	
Actin	Act for	TGCCCTGAGGCTCTTTTCC	VIC
	Act rev	TCGTGGATGCCACAGGATT	
	Act probe	CCTTCCTTCCTGGGTATG	

20ng of genomic DNA of each sample is added to the reaction mix with 2 Units of MseI restriction Enzyme (BA0525L, NEB). For each reaction mix, a No Template Control is prepared.

Reaction mix and AutoDG Oil for Probes (1864110, BioRad) are prompted to create the water-oil emulsion using the Automated Droplet Generator (Bio-Rad). Droplets are amplified using the C1000 Touch Thermal Cycler (Bio-Rad) using the protocol in Table 3. The annealing temperature is defined after a thermal gradient PCR reaction.

Table 3:
Cycling conditions for ddPCR Supermix for Probes

Cycling step	Temperature (°C)	Time	Number of cycles	Ramp rate
Enzyme activation	95	10min	1	2°C/sec
Denaturation	94	30sec	40	
Annealing/Extension	57	1min		
Enzyme deactivation	98	10min	1	
Hold	4	Infinite	1	

After thermal cycling, droplets are read using the QX200 Droplet Reader using the CNV2 analysis of the QuantaSoft software.

The GCN automatically resulting from the software, is plotted as mean and SEM.

3.3. Quantification of target mRNA expression using ddPCR:

Target mRNA expression is quantified for all samples using the QX200 Droplet Digital PCR system (Bio-Rad). The ddPCR reaction mix is prepared using the One-Step RT ddPCR Advance Kit for Probes as manufactured instruction (1864021, Bio-Rad). Each set of primer is used at the final concentration of 900nM, while TaqMan Probes are added at 250nM as final concentration. Target assays are called LC and HC while sequence assay related to Reference genes, Ap3b1 and B2M, are reported in Table 2.

25pg of total RNA of each sample is directly added to the reaction mix. Reaction mixes and AutoDG Oil for Probes (1864110, BioRad) are prompted to create the water-oil emulsion using the Automated Droplet Generator (BioRad). Droplets are amplified using the C1000 Touch Thermal Cycler (BioRad) using the protocol in Table 4. The annealing and Retro-Transcriptase temperature are defined after a thermal gradient PCR reaction.

Table 4:
Cycling conditions for One-Step RT ddPCR Advance Kit for Probes

Cycling step	Temperature (°C)	Time	Number of cycles	Ramp rate
Reverse Transcription	45	60	1	2°C/sec
Enzyme activation	95	10min	1	
Denaturation	95	30sec	40	
Annealing/Extension	57	1min		
Enzyme deactivation	98	10min	1	
Hold	4	Infinite	1	

After thermal cycling, droplets are read using the QX200 Droplet Reader using the CNV1 analysis of the QuantaSoft software.

The ratio Target/Reference automatically resulting from the software, is plotted as mean and SEM.

3.4. RNA Sequencing and bioinformatic analysis:

Libraries derived from RNA samples are prepared using the TruSeq Stranded Total RNA library preparation kit as manufactured instruction (20020599, Illumina). 1 µg of total RNA of each sample is used for the library preparation. Since all samples show RIN>8, the 8 minutes Fragmentation protocol is used.

The First Strand synthesis step is performed using the Superscript II reverse Transcriptase that convert the RNA strand in a single strand DNA (18064014, Invitrogen). The single strand DNA is subsequently converted into cDNA during the library preparation. The libraries prepared are quantified using the Qubit 3.0 fluorimeter with the dsDNA High Sensitivity kit (Q32851, Life Technologies). Libraries length is checked using the Bioanalyzer instrument (Agilent), with the DNA High Sensitivity kit (5067-4626, Agilent).

Paired-end sequencing was performed in a 2x75bp run on the Illumina NextSeq500. NextSeq500/550 high Output Flow cell v2, 300 cycles (20024908, Illumina) was used for the sequencing.

Raw sequencing data (.bcl format) were de-multiplexed by the barcodes and converted to FastQ files using bcl2fastq2 conversion software (Illumina). The FastQ data obtained from each sequenced sample are checked and select for quality using 3 tools: clumpify.sh, FastQ and seqtk sample. Following, remaining reads are aligned to the CriGri-PICR (GCA_003668045.1) reference genome using STAR alignment tool. The resulting mapped reads are assigned to specific gene annotation using gtfToGenePred and abundance is calculated by FeatureCount. Library size is normalized and differentially expressed genes are found using DESeq2.

A hybrid reference sequence is created using the CriGri-PICR reference genome whit mAb sequence including both Heavy and Light chain construct. Reads are mapped on the hybrid reference genome using STAR, and resulting mapped reads are used to calculate differential expression using DESeq2.

3.5. DNA methylation analysis using Whole Genome Bisulfite Sequencing (WGBS):

Genomic DNA, extracted from each sample is processed for the WGBS using the Ultralow Methyl-Seq with True Methyl OxBS kit as manufactured instruction (0536-32, NuGEN). 1 µg of extracted genomic DNA is spiked with 1% (w/w) of the Lambda DNA - N6-methyladenine-free (BN3013S, NEB), which serves as unmethylated control during bisulfite treatment.

First, diluted and spiked DNA is mechanically fragmented for 2 minutes using S220 Sonicator (Covaris) to obtain 200bp DNA fragments. DNA fragmentation is performed in 130 µl volume. Purified DNA is quantified using Qubit fluorimeter with the dsDNA High Sensitivity kit. DNA fragment size is checked using Bioanalyzer with DNA 12000 kit (5067-1508, Agilent).

Each fragmented sample is split into two aliquots: 2/3 of the total DNA quantity is used as OxBS aliquot, while 1/3 of the total DNA quantity is used as BS aliquot. The BS aliquot s Bisulfite is converted in order to transform 5mC and 5hmC into C while C are converted into T. On the contrary, the OxBS aliquot is both oxidized and Bisulfite converted in order to transform the 5hmC into T.

Since OxBS aliquot undergoes oxidation extra-process, the DNA material is more degraded with respect to BS aliquot, therefore OxBS aliquot is amplified with 12 PCR cycles during library amplification, while the BS aliquot is amplified using 10 PCR cycles.

The library is finally quantified using Qubit Fluorimeter with the dsDNA High Sensitivity kit. Libraries length is checked using the Bioanalyzer instrument, with the DNA High Sensitivity kit.

Paired-end sequencing was performed in a 2x100bp run with 10% PhiX control v3 (FC-110-3001, Illumina) on the Illumina NovaSeq6000. Ultralow Methyl-Seq with True Methyl OxBS kit requires different sequence primer 1, that are loaded on the reagent cartridge. All libraries are loaded on the S2 Flow cell v2, 200 cycles (20028315, Illumina).

Raw sequencing data (.bcl format) were de-multiplexed by the barcodes and converted to FastQ files using bcl2fastq2 conversion software (Illumina). The FastQ data obtained from each sequenced sample are then trimmed using TrimGalore! Tool⁴⁰.

Following, remaining reads are aligned to the CHO-K1 reference genome (UCSC) using Bismark alignment tool (UCSC), which is able to identify methylated bases with respect to a reference genome sequence.

Bismark is also able to calculate bisulfite conversion rate on Lambda DNA - N6-methyladenine-free sequence⁴⁰.

The resulting mapped reads are analyzed for differential methylation using MethylKit tool⁴⁰.

The resulting regions, derived from 5mC and 5hmC datasets, are divided in 5kb genomic windows and filtered for >10 methylated sites and > 10 hydroxi-methylated sites. Those filtered regions are analyzed using UCSC genome browser with the aim to find orthologous annotated genes in Mus musculus genome since the CHO-K1 genome is poorly annotated.

3.6. Histone marks analysis by ChIP-Seq:

Freshly cultured cells are collected into falcon tube and centrifuge to remove the supernatant. Cells are resuspended in DBPS to obtain a concentration of 20x10⁶ cells/ml. A 37% Formaldehyde solution (F8775, Sigma) are added to a final concentration of 1% to cross-link the sample (make a covalent bond between histone and DNA) and incubate for 8 minutes at RT. Then 1,25M Glycine (G8790, Sigma) is added to each sample to a final concentration of 0.125M to quench formaldehyde and incubate for 5 minutes at RT.

Cell are centrifuged and washed with PBS. Washed cells are resuspended in 1ml of PBS and stored at -80°C upon shipment to Diagenode SA for ChIP-Seq analysis service, who perform the ChIP-Seq analysis.

The chromatin is prepared by Diagenode ChIP-seq/ChIP-qPCR Profiling service (Diagenode Cat# G02010000) using the iDeal ChIP-seq kit for Histones (Diagenode Cat# C01010059).

Chromatin is sheared using Bioruptor Pico sonication device (Diagenode Cat# B01060001) combined with the Bioruptor Water cooler for 4 cycles using a 30'' [ON] 30'' [OFF] settings. Shearing is performed in 1.5 ml Bioruptor Pico Microtubes with Caps (Diagenode Cat# C30010016) with 1 million in 100µl. An aliquot of this chromatin is used to assess the size of the DNA fragments obtained by High Sensitivity NGS Fragment Analysis Kit (DNF-474) on a Fragment Analyzer™ (Agilent).

ChIP was performed using IP-Star Compact Automated System (Diagenode Cat# B03000002) following the protocol of the kit. 1µg of chromatin is immunoprecipitated using the following antibodies: H3K4me3 (C15410003), H3K27me3 (C15410195), H3K27ac (C15410196) and Rabbit IgG (C15410206). The 1% of chromatin is set as Input. ChIP efficiency using KAPA SYBR FAST (Sigma-Aldrich) on LightCycler 96 System (Roche) and results were expressed as % recovery = 2^{-(Ct_input-Ct_sample)}. Primers used are the following (Table 5):

Table 5:

5'-3' sequence of primer and probes for ChIP-Seq validation experiment

Sequence Name	Sequence
384 GAPDH F	CCC GTT TCC GAC CGT CC
384 GAPDH R	TGT TCT AGA GAC AGC CGC AT
386 ACTB 2 F	CCG TCT TCC CAT CCA TCG TG
386 ACTB 2 R	GGA AAA CCC CAG CAT CGC TA
387 MBex2 1 F	GAG TGG CTG AGC TGG GAT TT
387 MBex2 1 R	CAC ATG CCA CCA AGC ACA AG

The library preparation is conducted by Diagenode ChIP-seq/ChIP-qPCR Profiling service (Diagenode Cat# G02010000). Libraries are prepared using IP-Star Compact Automated System (Diagenode Cat# B03000002) from input and ChIP'd DNA using MicroPlex Library Preparation Kit v3 /96 rxns (Diagenode Cat# C05010002) with 24 UDI for MicroPlex v3 - Set I and Set II (Diagenode Cat# C05010008 and C05010009). Libraries are purified using Agencourt AMPure XP (Beckman Coulter) and quantified using Qubit™ dsDNA HS Assay Kit (Thermo Fisher Scientific, Q32854). Fragment size is analyzed by High Sensitivity NGS Fragment Analysis Kit (DNF-474) on a Fragment Analyzer™ (Agilent).

Sequencing is performed in paired-end 50 bp on an Illumina Novaseq. Quality control of sequencing reads is performed using FastQC. Trimming of the adaptors is performed with either cutadapt or Trim_Galore!. Trimmed reads are aligned to the reference genome (CriGri) obtained from the UCSC genome browser using BWA software v.0.7.5a. Samples are filtered for regions blacklisted by the ENCODE project. PCR duplicates and multi mapping reads are removed using samtools. Alignment coordinates are converted to BED format using BEDTools v.2.17 and peak calling was performed using Epic2 with optimized parameters for H3K27ac and H3K27me3 and H3K4me3.

The Differential binding analysis is performed with the R/Bioconductor package: DiffBind.

Pathway analysis is performed with the R package ReactomePA and the Mus musculus KEGG database on the associated with significantly differentially accessible regions between Clone 3 and Clone 24 o the H3K4me3, H3K27me3 and H3K27ac histone marks.

4. Results

4.1. Transgene Copy Number Evaluation in Good and Bad producer clones:

In this study, we investigate the Transgene Copy Number of 4 different CHO-K1 SV clones that are produced following transfection of a plasmid expression vector containing cDNA of mAb's Heavy and Light Chain.

As discussed before, the 4 clones are previously evaluated for their productivity by the Cell Line Development group based in the Vevey facility. 2 Good producer and 2 Bad producer clones are then selected for this study.

Set of Primers and Probes are specifically designed for Heavy and Light chain sequence of the mAb's construct. The Copy Number of the integrated Transgene is determined using ddPCR technology.

The Copy Number is calculated based on concentration in copies/ μ l of an invariant reference locus. In fact, also a set of primers and probes are designed for 2 different Reference genes: AP-3 complex β 1 (Ap3b1) and β -2-microglobulin (B2M).

As shown in Figure 17, the copies/ μ l obtain for each reference gene, Ap3b1 and B2M, are stable around 60-90 copies/ μ l, independently from the clone analyze and type of duplex assay use (Light Chain or Heavy Chain assay). As expected, a variable number of copies/ μ l is found for the Light Chain (LC) and Heavy Chain (HC) analyzing different clone.

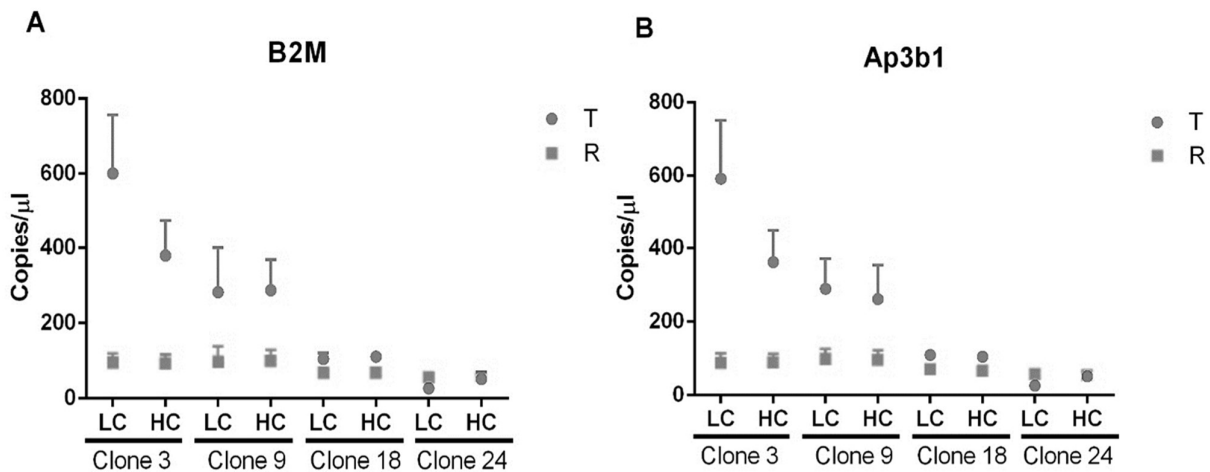
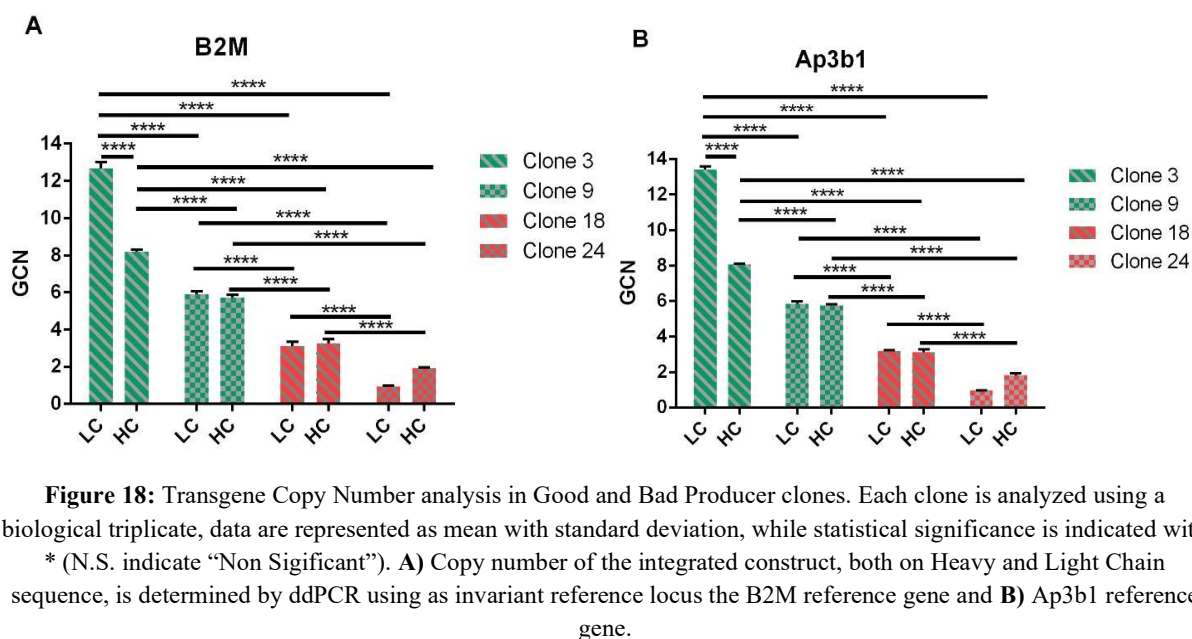


Figure 17: Copies/ μ l of genes obtained using ddPCR in all clones analyzed. **A)** Copies/ μ l obtained for B2M gene are indicated with square (R). HC and LC copies/ μ l are indicated with dot (T). **B)** Copies/ μ l obtained for Ap3b1 gene are indicated with square(R). HC and LC copies/ μ l are indicated with dot (T).

In Figure 18, the resulting copy number of the transgene is reported with respect to Ap3b1 and B2M reference genes, for both Heavy and Light Chain (HC and LC respectively).



Good producer clones, Clone 3 and 9, exhibit higher copy numbers with respect to Bad producer clones, Clone 18 and Clone 24, independently by the reference gene used. Specifically, transgene copy numbers ranged from 6 to 8 copies in Good producer clones, while there are 2-3 copies in bad producer clones. These results agree with the productivity data shown in Figure 16 (chapter 2).

Since the sequences of Heavy and Light Chain are present on the same plasmid, the copy number expected should be the same. Interestingly, Clone 3 and Clone 24 shows different copy number between HC and LC. Clone 3, the highest productive clone, shows 8 copies of HC and 13 copies of LC, therefore higher LC copies with respect to HC. On the contrary, Clone 24, the least productive clone, exhibits a higher HC copy number with respect to LC. Those differences in the Transgene copy number could be attributable to a partial integration of the construct in the CHO genome.

It's important to notice that Clone 3 exhibits a higher copy number with respect to Clone 9, which shows higher productivity. This result seems to not correlate with productivity data. However, further analysis demonstrates that Clone 9 partially lost its productivity, indicating its instability of protein expression in long-term culture (data not presented in this work).

All of these results indicate that there is a correlation between productivity and transgene copy number, nevertheless, it could be possible that the higher copy number isn't totally responsible for the higher productivity. For this reason, other molecular mechanisms should be investigated.

4.2. Transgene expression correlates with productivity and copy number analysis:

Transgene copies could go through gene silencing events, not permitting the mRNA transcription by the RNAPolymerase. For this reason, we decide to investigate the mRNA expression of mAb transgene using the same Good and Bad producer clones previously analyzed.

In order to evaluate the transgene mRNA expression, we take advantage of the ddPCR. Primers and Probes used for Copy number determination of Heavy and Light Chain, were previously designed on the coding sequence of the construct, thus could be reused for mRNA quantification. For target mRNA quantification, a

house-keeping gene is needed: here, we designed assays for Actin (Act) and β -2-microglobulin (B2M) coding sequence (the assay used for B2M gene is the one used for Copy number analysis).

As shown in Figure 19, the copies/ μ l obtained for Act reference genes are stable around 130-200 copies/ μ l, independently from the clone analyzed and type of duplex reaction assay used (Light Chain or Heavy Chain assay). Similarly, the copies/ μ l obtained for B2M reference genes are stable around 20-35 copies/ μ l, independently from the clone analyzed and the type of duplex reaction. As expected, a variable number of copies/ μ l is found for the Light Chain and Heavy Chain analyzing different clones.

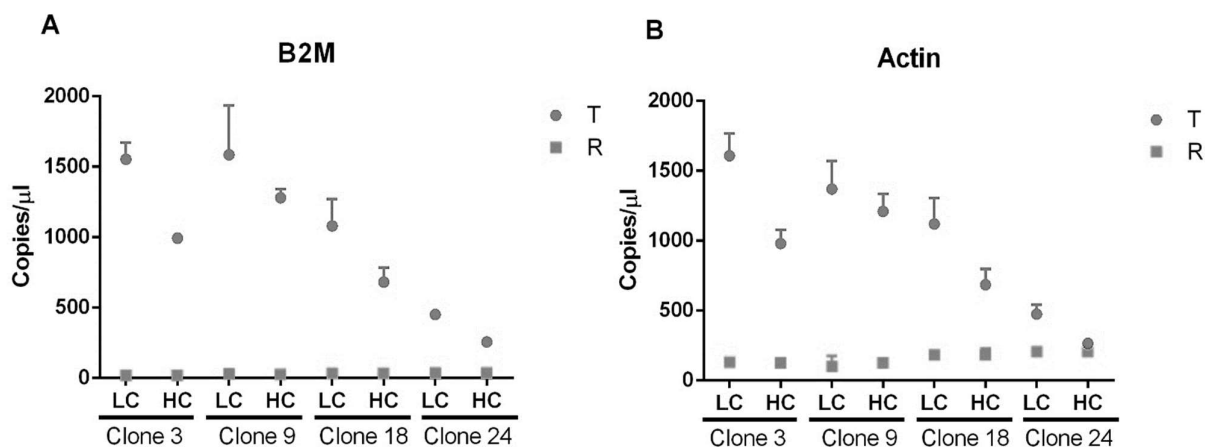


Figure 19: Copies/ μ l of RNA obtained using ddPCR in all clones analyzed. **A)** Copies/ μ l obtained for B2M gene are indicated with square (R). HC and LC copies/ μ l are indicated with dot (T). **B)** Copies/ μ l obtained for Actin gene are indicated with square (R). HC and LC copies/ μ l are indicated with dot (T).

These results indicate a different expression level of B2M and Act reference gene in CHO-K1 SV host cell, but stably expressed between clones analyzed.

The ratio of expression can vary based on the housekeeping gene (B2M or Act) used for the analysis. In fact, more the housekeeping gene is expressed, lower is the ratio found. However, since it's a relative analysis, it isn't important how much a house-keeping gene is expressed, as long as is stably expressed between samples analyzed.

In Figure 20, the ratio of mRNA expression of the transgene is reported with respect to Act and B2M house-keeping genes, for both Heavy and Light Chain (HC and LC respectively).

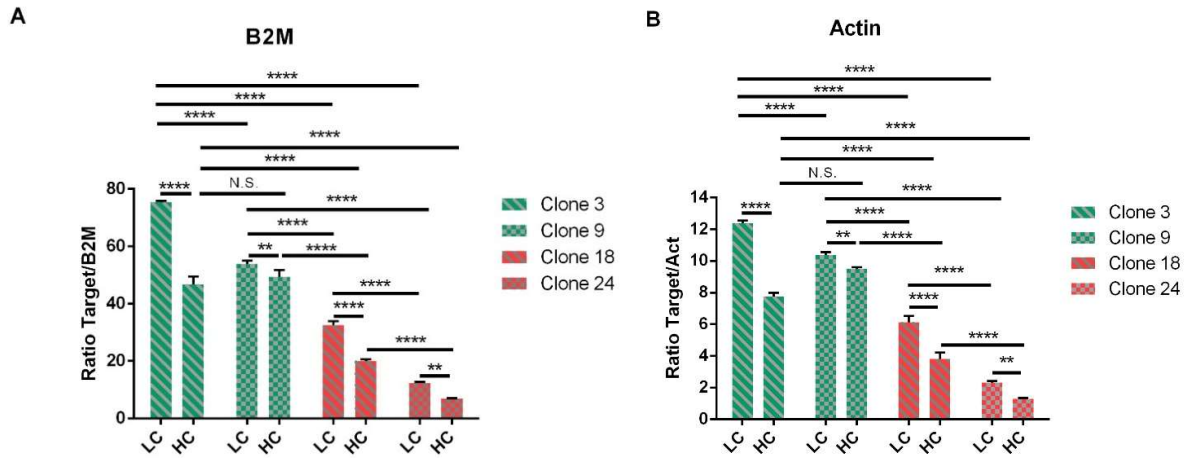


Figure 20: Heavy and Light chain mRNA expression analysis in Good and Bad Producer clones. Each clone is analyzed using a biological triplicate, data are represented as mean with standard deviation, while statistical significance is indicated with * (N.S. indicate “Non Significant”) **A)** mRNA expression of HC and LC is determined by ddPCR using as house-keeping gene the B2M reference gene and **B)** Act reference gene.

We found significant differences between Good and Bad producer clones both in terms of Light and Heavy Chain mRNA expression, independently by the reference gene used. In particular, we found higher HC and LC mRNA expression in Good Producer Clones, Clones 3 and 9, with respect to Bad Producer Clones, clones 18 and 24. These results are consistent with productivity data shows in chapter 2, and with the Transgene Copy Number data found previously.

Heavy and Light chains are inserted in the same expression vector, however, the two constructs are under the control of two separate strong promoters. Thus, it's expected that LC and HC mRNA expression values could be different in the same clone.

All the analyzed clones show higher LC expression with respect to HC, as indicated in Figure 21. A very small difference in expression is presented by the Clone 9, where the ratio LC/HC is about 1.

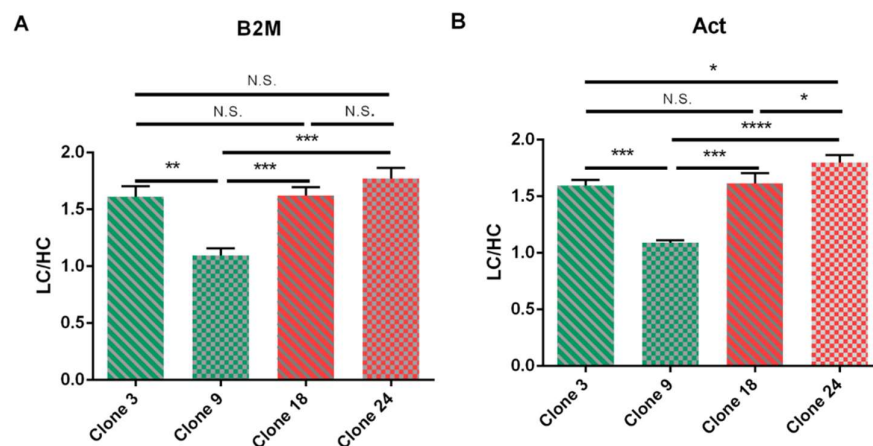


Figure 21: Heavy and Light Chain mRNA expression analysis in Good and Bad Producer clones, shows as ratio LC/HC. The value of LC mRNA expression is divided by HC obtained value for each clone analyzed. Data are represented as mean with standard deviation, while statistical significance is indicated with * (N.S. indicate “Non Significant”). **A)** mRNA expression of Heavy and Light Chain is determined by ddPCR using as house-keeping gene the B2M reference gene and **B)** Act reference gene.

From these results, we can deduce that the productivity of clones is correlated to the mRNA expression, which in turn is correlated to the number of copies of the inserted constructs. On the other hand, results indicate that correlation between productivity and mRNA level is mediated by both Heavy and Light chain transcripts.

4.3. Transgene expression, obtained using RNA-Seq method, is correlated to clone productivity:

The level of transcript mRNA expressed in different clones and clone mAb productivity is further investigated using RNA-Seq. The RNA reads sequenced from each biological replicate of Good Producer and Bad Producer clones were aligned to the sequence of the mAb construct and the resulting mapped sequences are used to calculate the differential expression of Heavy and Light Chain.

As shown in Table 6, Light and heavy chain sequences are found to be highly expressed in the Good Producer clones with respect to the Bad producer clones. In particular, Light Chain exhibit 1,84 Fold-change, while the Heavy Chain shows 1,65 Fold-change.

Table 6:
Light chain and Heavy chain expression level trough RNA-Seq

Transcript ID	Base Mean	log2 Fold-change	p-adjusted value
mAb_Light Chain	312841,3082	1,84	4,4x10 ⁻⁴
mAb_Heavy Chain	1205109,359	1,65	1,99x10 ⁻⁵

These results agree with the mRNA expression obtained using the ddPCR (Rif. paragraph 4.2).

4.4. Transcriptomic studies for evaluation of genes/pathways related to clone productivity:

Comparison between Good and Bad producer clones was made in the global Transcriptome to further find characteristics linked to the clone productivity, cell growth and stability.

The global transcriptome is obtained from Good and Bad producer clones: total RNA, extracted from 3 biological replicates for each clone, were sequenced using the Illumina NextSeq500 platform.

Differentially expressed genes were found using Good Producer clones as “Case”, while Bad Producer clones are used as “Control”.

Genes obtained from all biological replicates were firstly analyzed using the Principal Component analysis (PCA) analysis: 64% of the variance (overall variability) is explained by the first principal component, while 21% of the variance is explained by the second principal component (Figure 22).

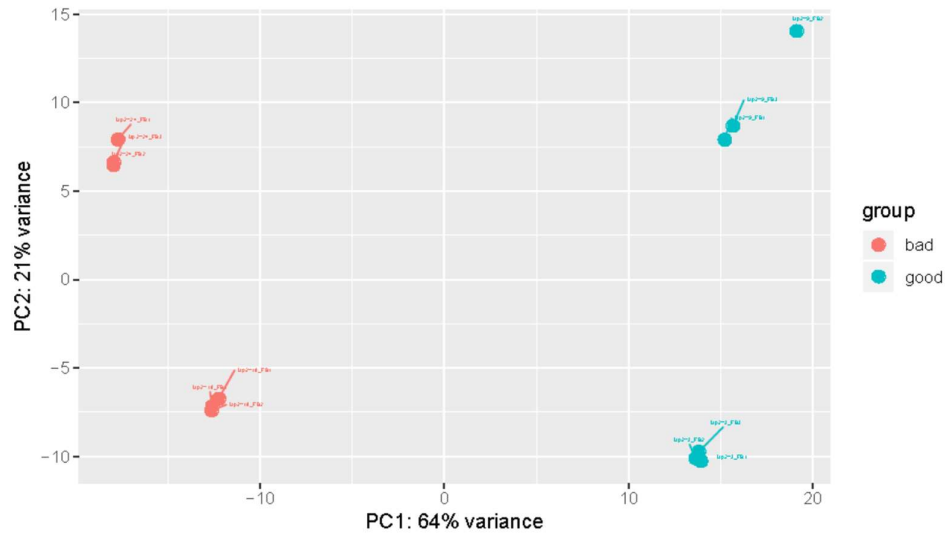


Figure 22: Principal Component Analysis of Transcriptome data: Good Producer clones are indicated in light blue, while Bad Producer clones are colored in red. The 64% of the variance is explained by the first principal component on x-axis, while 21% of the variance is explained by the second principal component, on y-axis. Biological replicates cluster and separate between different clones. The biological replicates FB2 of the Clone 9 is out of the cluster.

Biological replicates of each clone were found to cluster and separate from other clones. The only exception is represented by the Clone 9, where the biological replicates FB2 is out of the cluster. For this reason, this biological replicate is excluded from further analysis.

PCA analysis proved the statistical difference between Transcriptome data derived from different clones.

Differentially expressed genes (DEGs) identified were plotted using Volcano Plot (Figure 23).

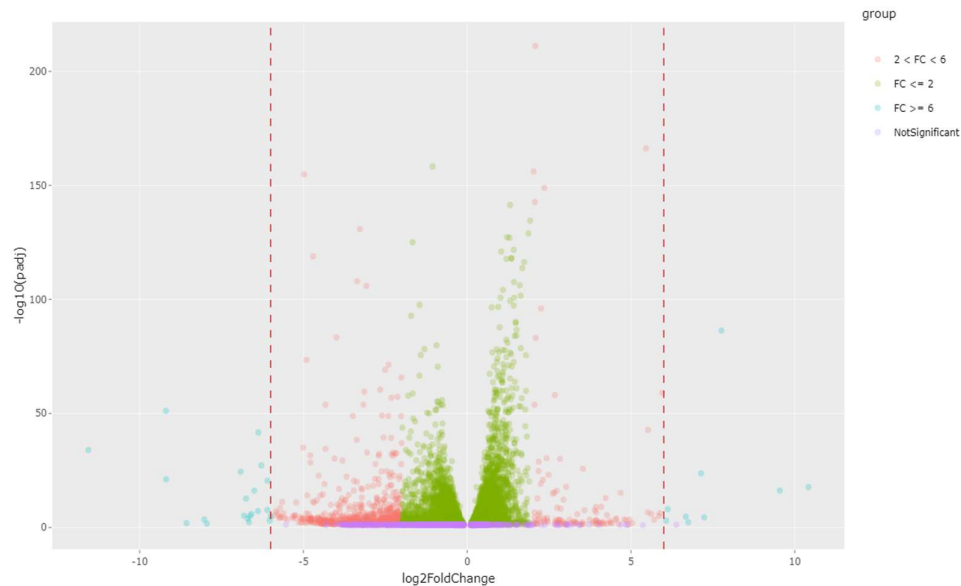


Figure 23: Differentially expressed genes (DEGs) shown in Volcano Plot: log2 Fold-change on x-axis is plotted with $-\log_{10}(\text{padj})$ on y-axis. This plot shows the distribution of highly DEGs. Dots in purple indicate non-significant genes ($p\text{-adj} > 0.01$). Dots in green indicate DEGs genes with low log2 Fold-change (≤ 2): those genes are excluded from further analysis. Dots in red and light blue are DEGs genes with log2 Fold Change > 2 . Those genes are further investigated using PANTHER tool for Gene Ontology analysis.

With a p-value adjusted minor than 0,01 ($p\text{-adj} < 0,01$), 8331 genes were found to be differentially expressed between Good and Bad producer clones. From total DEGs, 4542 are found to be down-regulated, while 3789 are Up-regulated in Good Producer clone with respect to Bad Producer clone. Using a cut-off $> 2 \log_2$ Fold-change, 858 genes are down-regulated, while 140 are up-regulated in good Producer clones. Those genes are further investigated for Gene Ontology analysis using the tool PANTHER.

To find out the highly expressed genes, the cut-off is set up at $> 6 \log_2$ Fold-change: 21 genes are found to be highly down-regulated, while 10 genes were identified as highly expressed genes in Good Producer clones. Those genes are treated as Top DEGs, and singularly analyzed.

In Table 7a and 7b, Top DEGs are reported with their $\log_2$ fold-change and gene name. These tables are also indicate GO annotation and gene function as reported from GeneCards (www.genecards.org).

The top highly expressed gene in Good Producer clone with respect to Bad producer clone is Dermatopontin (DPT), which is involved in cell-matrix interaction and cell proliferation inhibition (Table 7A). Following that, CMKLR1, BDKRB2 and SEMA4D are found to be highly expressed (Table 7A) and associated with the cell-signaling pathway (G-protein coupled receptor). PMEPA4 instead is involved in TGF- β signaling, thereby could be involved in the regulation of cell proliferation (Table 7A). Interestingly, also COL6A3 is found to be highly expressed in Good Producer clones. This gene is known to act as an endopeptidase inhibitor, which could be an important feature for recombinant protein production.

Table 7A:

Differentially expressed genes with $\log_2$ Fold Change $> +6$.

Gene ID	$\log_2$ Fold-Change	Encoded Protein/RNA	GO annotation / Gene function (www.genecards.org)
DPT	10,42	Dermatopontin	Cell-matrix interactions and matrix assembly. Enhances TGFB1 activity. Inhibits cell proliferation
CMKLR1	9,541	Chemerin Chemokine-Like Receptor 1	G protein-coupled receptor activity and chemokine receptor activity
SULF2	7,761	Sulfatase 2	Calcium ion binding and arylsulfatase activity.
BDKRB2	7,236	Bradykinin Receptor B2	G protein-coupled receptor activity and protease binding
CLMP	7,139	CXADR Like Membrane Protein	This gene encodes a type I transmembrane protein. It may have a role in cell-cell adhesion.
LRRC6	6,747	Leucine Rich Repeat Containing 6	The protein encoded by this gene contains several leucine-rich repeat domains and appears to be involved in the motility of cilia.
SEMA4D	6,672	Semaphorin 4D	Signaling receptor binding and transmembrane signaling receptor activity
PMEPA1	6,385	Prostate Transmembrane Protein, Androgen Induced 1	WW domain binding and R-SMAD binding; Functions as a negative regulator of TGF- β signaling and thereby probably plays a role in cell proliferation, differentiation, apoptosis, motility, extracellular matrix production and immunosuppression.
CYP2C27	6,123	Cytochrome P450 2C27	Catalyzes the hydroxylation of tolbutamide and the N-demethylation of aminopyrine and benzphetamine.
COL6A3	6,072	Collagen Type VI Alpha 3 Chain	Serine-type endopeptidase inhibitor activity;

In opposition, the Top down-regulated gene in the Good Producer clone is Mex-3 RNA Binding Family member C (MEX3C). This gene could bind RNA and DNA molecules and able to function as shuttles between the nucleus and cytoplasm. Several genes that are found to be down-regulated, have GPCR (G-protein coupled receptor) activity, such as NPY1R, P2RY14 and CCKAR. Moreover, UBQLN3 is found to be down-regulated and it's associated with protein proteasome degradation. Other down-regulated genes are related to tissue-specific signaling pathways, such as KIAA1549L and VMN2R26 (Table 7B).

Table 7B:

Differentially expressed genes with log2 Fold Change < -6.

Gene ID	log2 Fold-Change	Encoded Protein/RNA	GO annotation /Gene function (genecards)
MEX3C	-11,56	Mex-3 RNA Binding Family Member C	Nucleic acid binding and RNA binding; the protein interacts with mRNA via the KH domains and shuttles between the nucleus and cytoplasm.
SERPINB6	-9,194	Serpin Family B Member 6	Serine-type endopeptidase inhibitor activity and protease binding; Regulation of serine proteinases. Inhibitor of cathepsin G.
KIAA1549L	-8,567	KIAA1549 Like	Signaling via VEGFR2
CD200	-8,027	CD200 Molecule	Innate Immune System and Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell; Type I membrane glycoprotein. Immune-suppression.
NPY1R	-7,946	Neuropeptide Y Receptor Y1	G protein-coupled receptor activity and peptide YY receptor activity; Mobilization of intracellular calcium and inhibition of adenylate cyclase activity
MEP1A	-6,916	Meprin A Subunit Alpha	Metalloendopeptidase activity/Metalloproteinase activity.
VMN2R26	-6,812	Vomeroneasal type-2 receptor 26	Expressed in the basal epithelium of the vomeronasal organ.
GABRA1	-6,75	Gamma-Aminobutyric Acid Type A Receptor Subunit Alpha1	Drug binding and extracellular ligand-gated ion channel activity;
EPN3	-6,698	Epsin 3	Lipid binding and EH domain binding.
AWAT2	-6,66	Acyl-CoA Wax Alcohol Acyltransferase 2	Transferase activity, transferring acyl groups other than amino-acyl groups and long-chain-alcohol O-fatty-acyltransferase activity.
UBQLN3	-6,631	Ubiquilin 3	Binding: Link ubiquitination machinery to the proteasome to affect in vivo protein degradation.
P2RY14	-6,501	Purinergic Receptor P2Y14	G protein-coupled receptor activity and UDP-activated nucleotide receptor activity; Purinergic receptor for UDP-glucose and other UDP-sugars coupled to G-proteins.
WNT10A	-6,388	Wnt Family Member 10°	Signaling receptor binding and frizzled binding; Wnt/beta-catenin signaling pathway.
ANKAR	-6,376	Ankyrin And Armadillo Repeat Containing	Binding and nuclear hormone receptor binding.
IL33	-6,283	Interleukin 33	Cytokine activity; Activates NF-kappa-B and MAPK signaling pathways in target cells. Acts as a chromatin-associated nuclear factor with transcriptional repressor properties.
ARFGEF3	-6,102	ARFGEF Family Member 3	Binding and ARF guanyl-nucleotide exchange factor activity.

CCKAR	-6,096	Cholecystokinin A Receptor	G protein-coupled receptor activity and cholecystokinin receptor activity
CD34	-6,027	CD34 Molecule	Transcription factor binding and sulfate binding
MTTP	-6,002	Microsomal Triglyceride Transfer Protein	Protein heterodimerization activity and lipid transporter activity

Next, to exploit a larger set of DEGs, the cut-off $> 2 \log_2$ Fold-change is set up. Annotated DEGs were analyzed using PANTHER tool⁴¹. This tool is able to identify the Gene Ontology (GO) category enriched in a set of genes, with respect to the reference gene list: since the *Cricetulus griseus* reference list it's not present on PANTHER dataset, this analysis is performed using *Mus musculus* gene data.

DEGs gene was analyzed for Biological Process by Panther. Up and Down-regulated genes are uploaded separately on this tool to identify different GO categories between the two datasets. Those two datasets exhibit similar GO category enrichment in Biological Process, however looking in detail into the Cellular Process (GO:0009987), some different category comes evident. Specifically, in Up-regulated genes, the Protein Folding Category (GO:0006457) is over-represented, while Down-regulated genes show the Gene Silencing category (GO:0016458) (Figure 24 A and B). Moreover, in Down-regulated genes, categories related to tissue-specific pathways are over-represented, such as Myelination and Melanosome organization. These results could indicate that clones down-regulate tissue-specific processes to promote protein production-related processes.

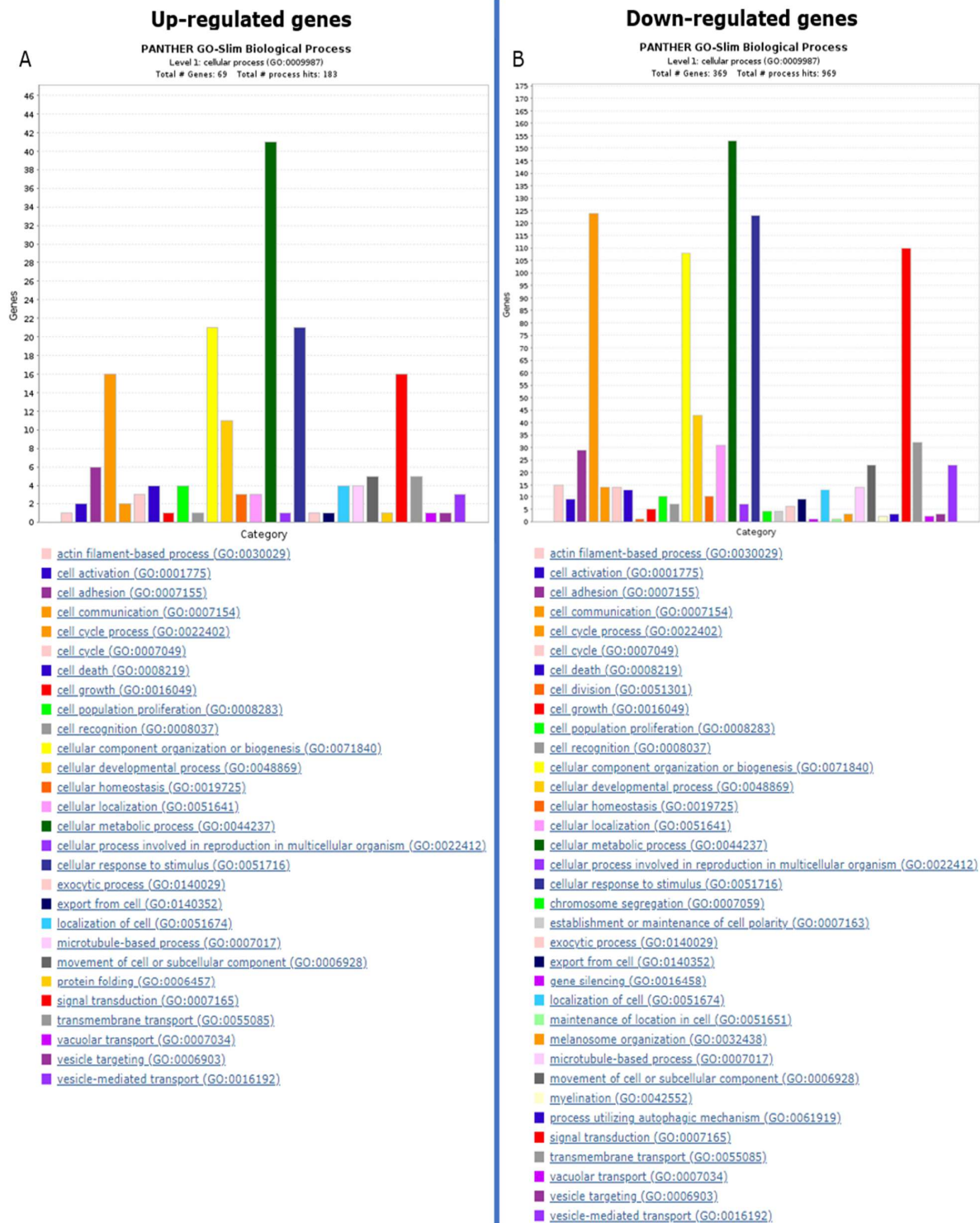


Figure 24: DEGs analyzed for Biological Process using Panther. Each bar of the histogram indicates an over-represent GO category. **a)** Cellular Process (GO:0009987) of Up-regulated genes. **b)** Cellular Process (GO:0009987) of Down-regulated genes. **c)** Signal transduction category (GO:0007165) of Up-regulated gene. **d)** Signal transduction category (GO:0007165) of Down-regulated gene. **e)** Detail analysis of the Vesicle-mediated Transport (GO:0016192) inside the Cellular Process category (GO:0009987) of Up-regulated genes. **f)** Detail analysis of the Vesicle-mediated Transport (GO:0016192) inside the Cellular Process category (GO:0009987) of Down-regulated genes.

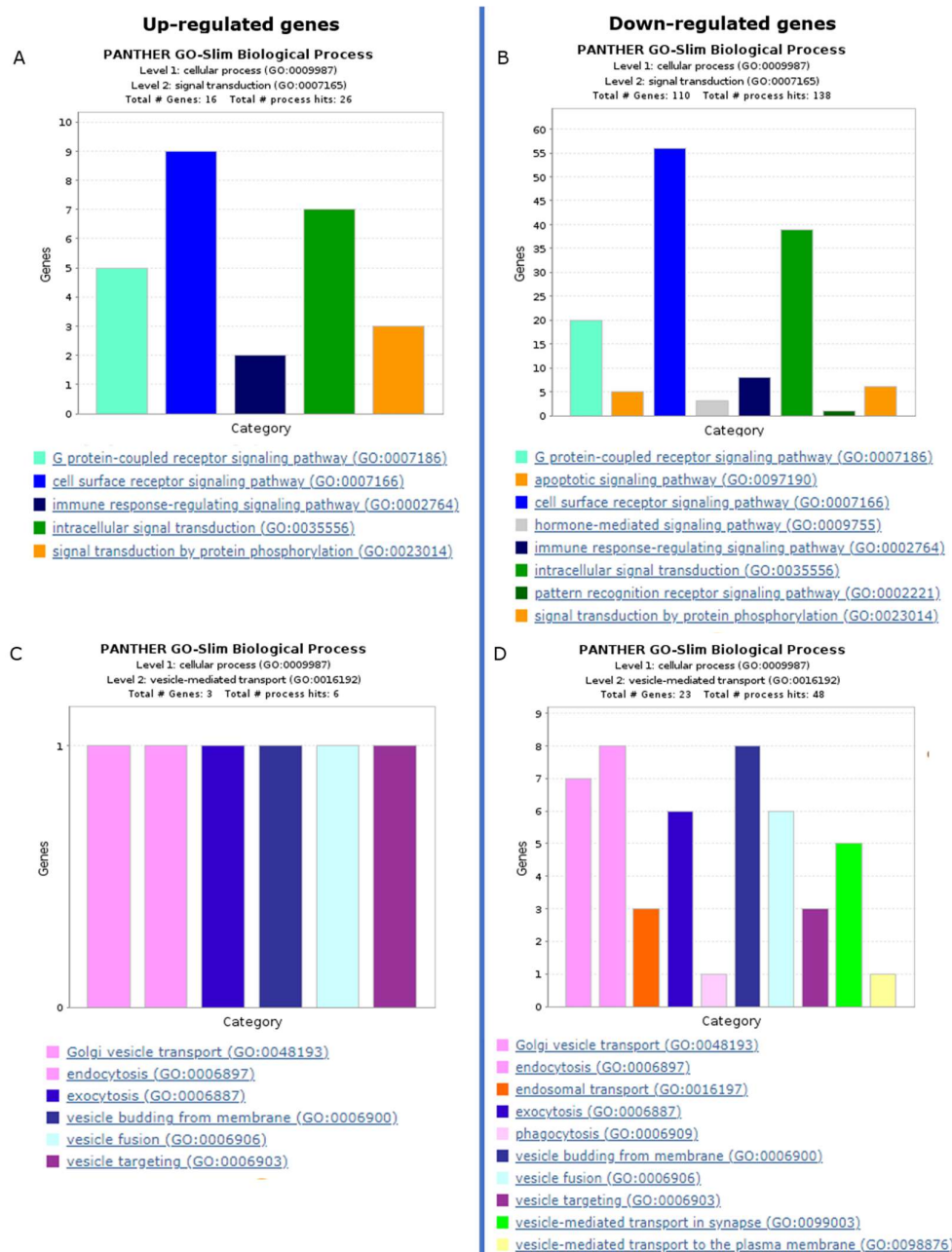


Figure 25: DEGs analyzed for Biological Process using Panther. Each bar of the histogram indicates an over-represent GO category. **A)** Signal transduction category (GO:0007165) of Up-regulated gene. **B)** Signal transduction category (GO:0007165) of Down-regulated gene. **C)** Detail analysis of the Vesicle-mediated Transport (GO:0016192) inside the Cellular Process category (GO:0009987) of Up-regulated genes. **D)** Detail analysis of the Vesicle-mediated Transport (GO:0016192) inside the Cellular Process category (GO:0009987) of Down-regulated genes.

In addition Figure 25 A and B show in detail the Signal transduction Category (GO:0007165), where the category of Apoptotic Signaling Pathway (GO:0097190) is over-represented in down-regulated genes. The G-protein Coupled Receptor Signaling Pathway (GO:0007186) is over-represented in both up and down-regulated gene datasets, as observed using the more stringent cut-off > 6 Fold-change.

In Figure 25 C and D, the Vesicle-mediated Transport category (GO:0016192) is reported. In Down-regulated gene datasets, categories related to the Endosomal transport process are over-represented: Endocytosis (GO:0006897), Endosomal Transport (GO:0016197) and Phagocytosis (GO:0006909). Moreover, also the category related to Vesicle-mediated Transport in Synapse (GO:0099003) is over-represented in down-regulated gene datasets. These results could indicate that clones down-regulate the Endosomal process and tissue-specific vesicle pathway in order to promote protein secretion and exocytosis process.

4.5. Differential DNA methylation studies for evaluation of genes/pathways related to clone productivity:

In parallel to gene expression analyses, we aimed to identify genomic regions showing differences in DNA methylation between Good and Bad producer clones to further investigate characteristics linked to clone productivity, cell growth and stability.

Whole Genome Bisulfite Sequencing (WGBS) is used as it's considered the 'gold standard' for assaying DNA methylation because it provides global coverage at single-base resolution⁴².

To perform WGBS on Good and Bad producer clones, the genomic DNA is first randomly fragmented to the desired size (200 bp) using the Sonicator instrument. Once the fragmented DNA is end-repaired, each sample is divided in two aliquots for analysis of the 5mC and 5hmC. 5hmC in genomic DNA is oxidized to 5-formylcytosine (5fC). Unlike 5mC and 5hmC, 5fC is sensitive to deamination by bisulfite, therefore bisulfite treatment of the oxidized DNA converts the 5fC's to uracil. Then, using sodium bisulfite treatment, unmethylated cytosines are converted into thymine (through uracil intermediate) while methylated cytosines remain protected. The oxidized aliquot is used for the analysis of the true level of 5mC, while the 5hmC level is obtained by the reads subtraction between oxidized and non-oxidized aliquot (Figure 26). Finally, DNA fragments are adapter ligated and library prep is amplified.

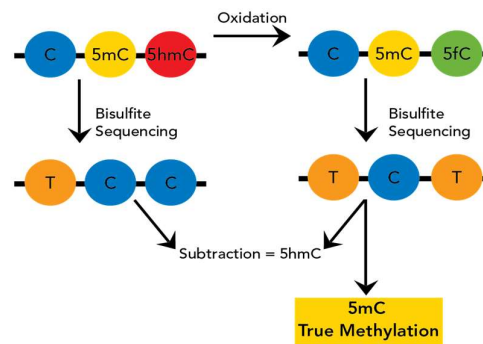


Figure 26: Scheme for analysis of DNA methylation and hydroxy-methylation. After DNA fragmentation, the sample is divided into two aliquots: the first is oxidized and bisulfite converted, while the second is only bisulfite converted. In the first aliquot, the 5hmC is oxidized to 5-formylcytosine (5fC); 5fC is sensitive to deamination by bisulfite, therefore bisulfite treatment of the oxidized DNA converts the 5fC's to uracil, as unmethylated cytosine. On the contrary, 5mC is converted into a non-methylated cytosine. Both 5hmC and 5mC are insensitive to bisulfite conversion, therefore they are converted into non-methylated cytosine, while methylated cytosines are converted into uracil, and finally into thymine. The oxidized aliquot is used for the analysis of the true level of 5mC, while the 5hmC level is obtained by the reads subtraction between oxidized and non-oxidized aliquot.

As summarized in Table 8, more than 100 million reads were obtained for each sample, and almost 99% are unique mapped reads. For bisulfite treatment quality control, the conversion rate of unmethylated spike-in Lambda DNA is determined, which is above 99% for all samples.

Table 8:

Sequencing run summary with total number of reads and Lambda conversion rate results.

Sample	5mC / 5mC+5hmC	Total number of reads	Unique Mapped reads	Lambda Conversion rate (%)
Clone 3_FB1	5mC	1,26E+08	9,79E+07	99,12
	5mC+5hmC	1,82E+08	1,40E+08	99,23
Clone 3_FB2	5mC	1,45E+08	1,13E+08	99,21
	5mC+5hmC	1,81E+08	1,38E+08	99,19
Clone 3_FB3	5mC	1,43E+08	1,12E+08	99,18
	5mC+5hmC	1,91E+08	1,46E+08	99,18
Clone 9_FB1	5mC	1,19E+08	9,50E+07	99,22
	5mC+5hmC	2,24E+08	1,73E+08	99,17
Clone 9_FB2	5mC	1,3E+08	1,02E+08	99,24
	5mC+5hmC	1,91E+08	1,47E+08	99,13
Clone 9_FB3	5mC	1,6E+08	1,26E+08	99,30
	5mC+5hmC	2,16E+08	1,63E+08	99,23
Clone 18_FB1	5mC	1,27E+08	9,95E+07	99,25
	5mC+5hmC	2E+08	1,51E+08	99,09
Clone 18_FB2	5mC	1,17E+08	9,19E+07	99,30
	5mC+5hmC	1,66E+08	1,25E+08	99,21
Clone 18_FB3	5mC	1,15E+08	9,04E+07	99,30
	5mC+5hmC	2,14E+08	1,62E+08	99,24
Clone 24_FB1	5mC	1,29E+08	1,01E+08	99,03
	5mC+5hmC	1,52E+08	1,16E+08	99,10
Clone 24_FB2	5mC	1,61E+08	1,25E+08	99,31
	5mC+5hmC	1,77E+08	1,37E+08	99,28
Clone 24_FB3	5mC	1,36E+08	1,05E+08	99,29
	5mC+5hmC	1,69E+08	1,28E+08	99,24

Differentially methylated positions are found using Good Producer clones as “Control”, while Bad Producer clones are used as “Case”.

Genes obtained from all biological replicates, are firstly analyze using Cluster analysis: biological replicates of each clone were found to cluster and separate from other clones. Moreover, Good Producer clones cluster separately with respect to Bad producer clones (Figure 27).

Cluster analysis proved the statistical difference between DNA methylation sites data derived from different clones.

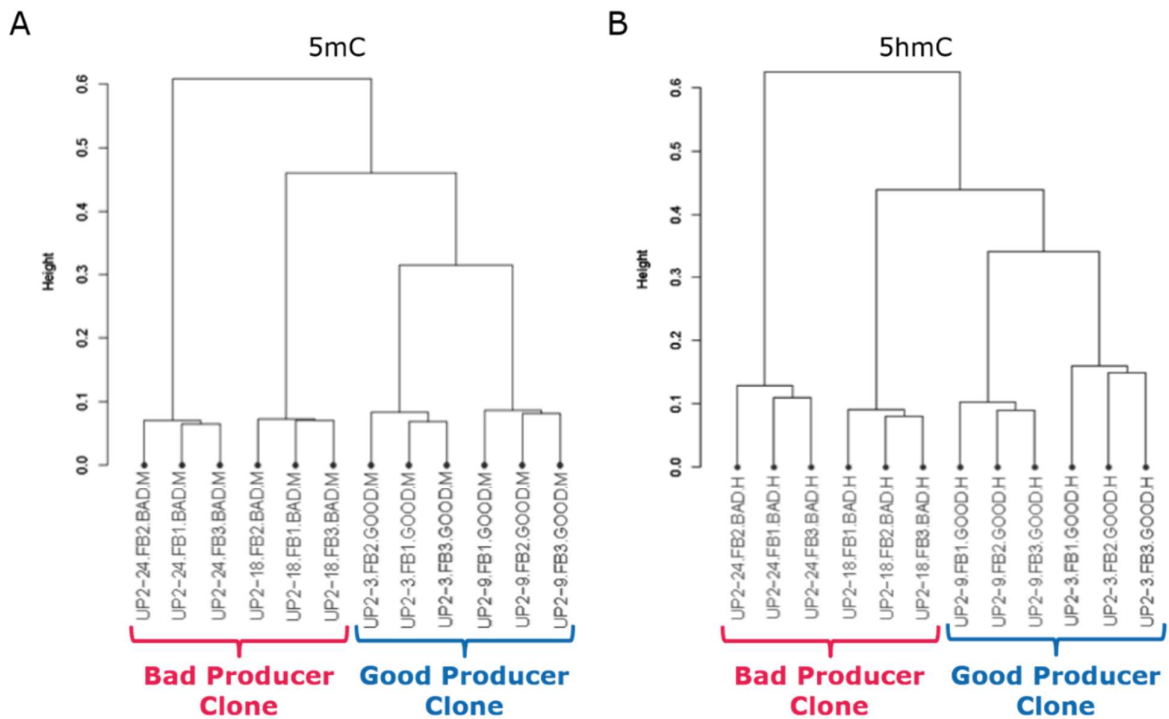


Figure 27: Cluster analysis of DNA methylation data: Good Producer clones are indicated in blue, while Bad Producer clones are colored in red. Biological replicates cluster and separate between different clones both analyzing samples for 5mC and 5hmC.

Both Good and Bad Producer Clones show global methylation around 78%, which agrees with the global methylation level reported for a wide variety of cell type.

5mC and 5hmC sites are analyzed in CpG, CHG and CHH context. As expected, the large majority of methylated sites, around 70%, are found in the CpG context, while a small part of the 5mC and 5hmC are found in CHG and CHH context. In fact, higher methylated levels are found in CHG and CHH context only in the staminal end embryonic cell line.

Interestingly, the Best producer Clone, Clone 3, shows the lowest methylation level in the CpG context, < 70%, with respect to the other clones (Table 9).

Table 9:

Summary of 5mC and 5hmC methylated site in different context

Sample	5mC / 5mC+5hmC	Total methylated C (%)	CpG methylated site (%)	CHG methylated site (%)	CHH methylated site (%)	% of methylated C in unknow context
Clone 3_FB1	5mC	76,28	69,6	0,74	0,68	5,26
	5mC+5hmC	75,66	68,88	0,63	0,58	5,57
Clone 3_FB2	5mC	76,08	69,46	0,64	0,59	5,39
	5mC+5hmC	76,08	69,26	0,66	0,61	5,55
Clone 3_FB3	5mC	76,39	69,77	0,68	0,63	5,31
	5mC+5hmC	76,62	69,51	0,68	0,62	5,81
Clone 9_FB1	5mC	81,15	74,31	0,62	0,58	5,64
	5mC+5hmC	81,17	73,99	0,68	0,62	5,88
Clone 9_FB2	5mC	81,72	74,93	0,59	0,56	5,64
	5mC+5hmC	81,73	74,57	0,72	0,66	5,78
Clone 9_FB3	5mC	81,03	74,69	0,54	0,5	5,3
	5mC+5hmC	80,77	73,89	0,62	0,57	5,69
Clone 18_FB1	5mC	77,83	70,98	0,58	0,53	5,74
	5mC+5hmC	79,22	72,16	0,76	0,69	5,61
Clone 18_FB2	5mC	76,76	70,36	0,53	0,49	5,38
	5mC+5hmC	78,38	71,65	0,63	0,57	5,53
Clone 18_FB3	5mC	77,65	71,3	0,52	0,49	5,34
	5mC+5hmC	78,63	71,94	0,6	0,54	5,55
Clone 24_FB1	5mC	78,40	71,11	0,83	0,77	5,69
	5mC+5hmC	78,32	71,06	0,76	0,7	5,80
Clone 24_FB2	5mC	77,92	71,84	0,53	0,49	5,06
	5mC+5hmC	78,46	71,83	0,57	0,52	5,54
Clone 24_FB3	5mC	79,32	73,39	0,54	0,51	4,88
	5mC+5hmC	77,31	70,87	0,61	0,56	5,27

A global analysis of DNA methylation on the level of single CpG dinucleotides showed a general bimodal distribution of methylation for all samples. In fact, both Good and Bad producer clones have more than 50% of all CpGs being largely methylated (90-100% of reads showing methylated sites), and almost 10% of unmethylated CpGs ($\leq 10\%$ of reads showing methylated sites).

The remaining methylated CpG sites showed an intermediate level of methylation, from 20% to 80% of CpGs methylation (Figure 28). This bimodal distribution is also confirmed in case of CHO-DP12 cells³⁰.

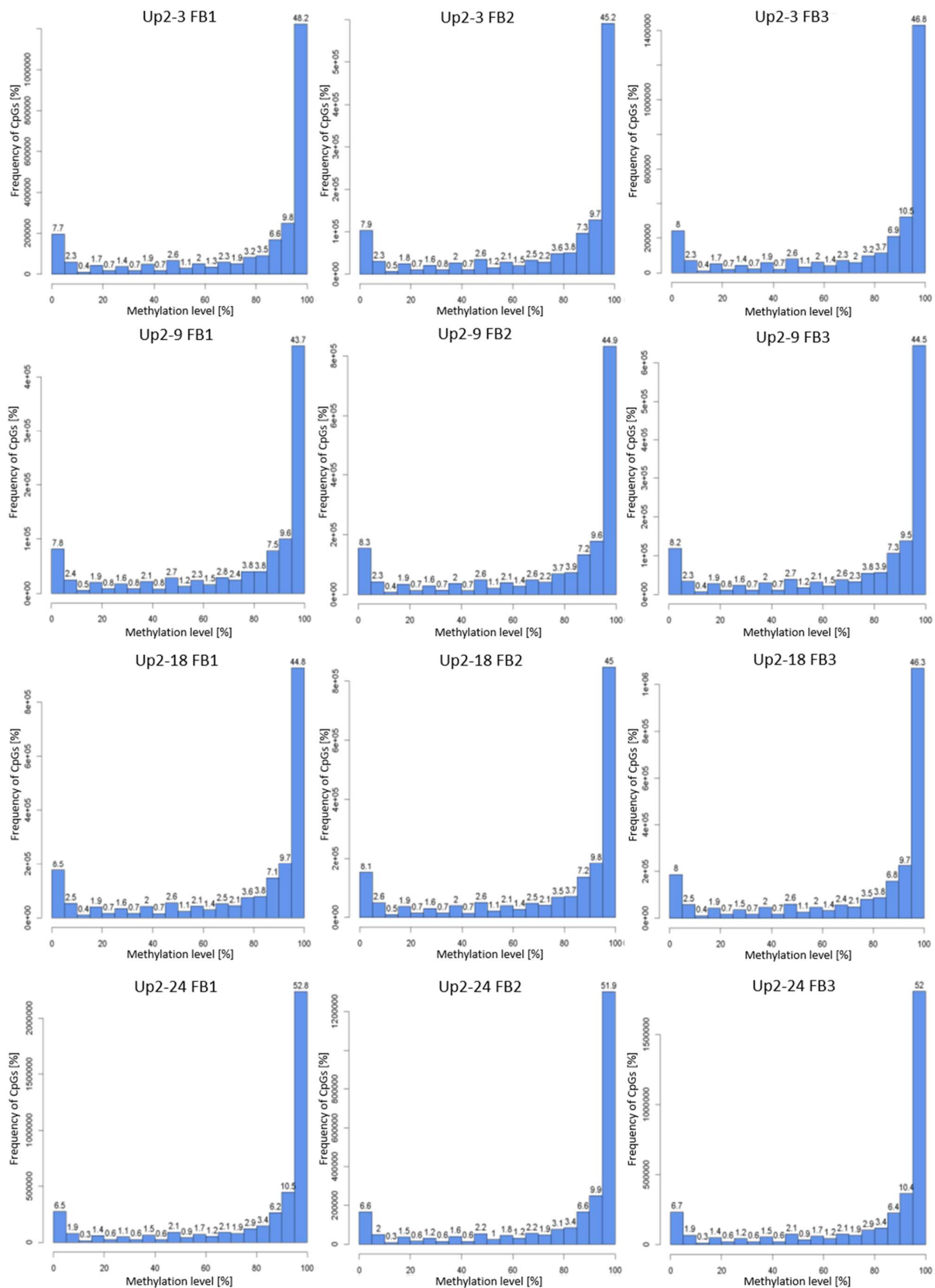


Figure 28: Global distribution of DNA methylation level: methylation percentage of CpG dinucleotides for each replicate of clones analyzed.

To further characterize the CHO-K1 SV clones, differentially methylated positions were found using Good Producer clones as “Control”, while Bad Producer clones are used as “Case”. Accordingly, “Hypermethylated regions” refers to higher methylation level of genomic region in Bad Producer clones with respect to Good Producer Clones, while “Hypomethylated regions” refers to lower methylation level in Bad producer Clones, respect to Good Producer clones.

Despite CHO-K1 reference genome is poorly annotated, Differentially Methylated sites (DMS) and Differentially Hydroxy-methylated sites (DHMS) are analyzed regarding annotated genomic features: CpG islands, CpG shores, untranslated regions (UTRs), promoter, exonic and intronic regions.

As shown in Table 10A, only 63 of a total of 3652 differentially methylated positions are found to be annotated, while only 18 of a total of 1831 differentially hydroxy-methylated positions.

Since the CHO-K1 genome is poorly annotated, most of the differentially methylated and hydroxy-methylated sites fall in intergenic or “other” regions.

Data from Table 10B shows that there isn’t a significant difference, in terms of DNA hydroxy-methylation between Good and Bad Producer Clones, while a very low difference is found for DNA methylation at level exonic features.

Interestingly, higher differential DNA methylation and hydroxy-methylation were found at the level of CpG island shores with respect to CpG island (Table 10C), indicating possible transcriptional repression related to tissue-specific genes. These also confirm that higher methylated residues are found in CG poor regions with respect to CG reach regions.

Table 10:

A: Number of Differentially Methylated and Hydroxy-methylated annotated sites, respect to total.

B: Percentage of differential Methylated and Hydroxy-methylated positions in promoter, exonic, intronic and intergenic genomic regions.

C: Percentage of differential Methylated and Hydroxy-methylated positions in CpG island and CpG shores.

A	N° differential Methylated sites	N° annotated sites (CHO)
5hmC	1831	18
5mC	3652	63

B	% Promoter	% exon	% intron	% intergenic
5hmC	0	0	0	100
5mC	0	0,03	0	99,97

C	% CpGi	% Shores	% Other
5hmC	0,11	1,53	98,36
5mC	0,27	2,77	96,96

The number of DMS and DHMs are also analyzed with respect to their distance from annotated Transcription Start Sites (TSS). As expected, a higher number of DMS were found respect DHMS, however both of them are mostly found in 100Kb distance from TSS. As shown in Table 10, 48,72% of DMS were found at 100Kb distance from TSS, while only 10,26% of DMS are found at 10Kb distance. Accordingly, 71,43% of DHMS are found in 100Kb windows from TSS, while only 28,57% are found at 10Kb from TSS (Table 11).

Promoter regions are normally reached in CpG Island and stand up to 5Kb distance from TSS of a gene, while CpG Island Shores are present at 2Kb distance upstream to the CpG island regions, indicating the methylation level within the 10Kb distance from TSS. Therefore, the higher DMS found at 100Kb distance from TSS could be addressed to enhancer regions.

Table 11:

The number of Differentially Methylated and Hydroxy-methylated annotated sites in respect to 100kb and 10kb distance from TSS.

	N° of total hits	Hits at 100kb distance	% 100kb distance	Hits at 10kb distance	% 10kb distance
5hmC	7	5	71,43	2	28,57
5mC	156	76	48,72	16	10,26

As we found indications of heterogeneous DNA methylation pattern in our clones, we plotted DMS along all the CHO-K1 scaffolds to visualize differential methylation distribution. In this analysis, we selected scaffolds with at least 10 DMS found in Bad clones (Clones 18 and 24) with respect to Good clones (Clones 3 and 9). As shown in Figure 29, Bad Producer Clones has a higher number of scaffolds that exhibit hyper-methylated sites with respect to the number of scaffolds that exhibit hypo-methylation. Moreover, Bad Producer clones show higher DMS for each scaffold with respect to Good Producer Clones. These results indicate global higher DNA methylation in Bad producer clones with respect to Good Producer clones.

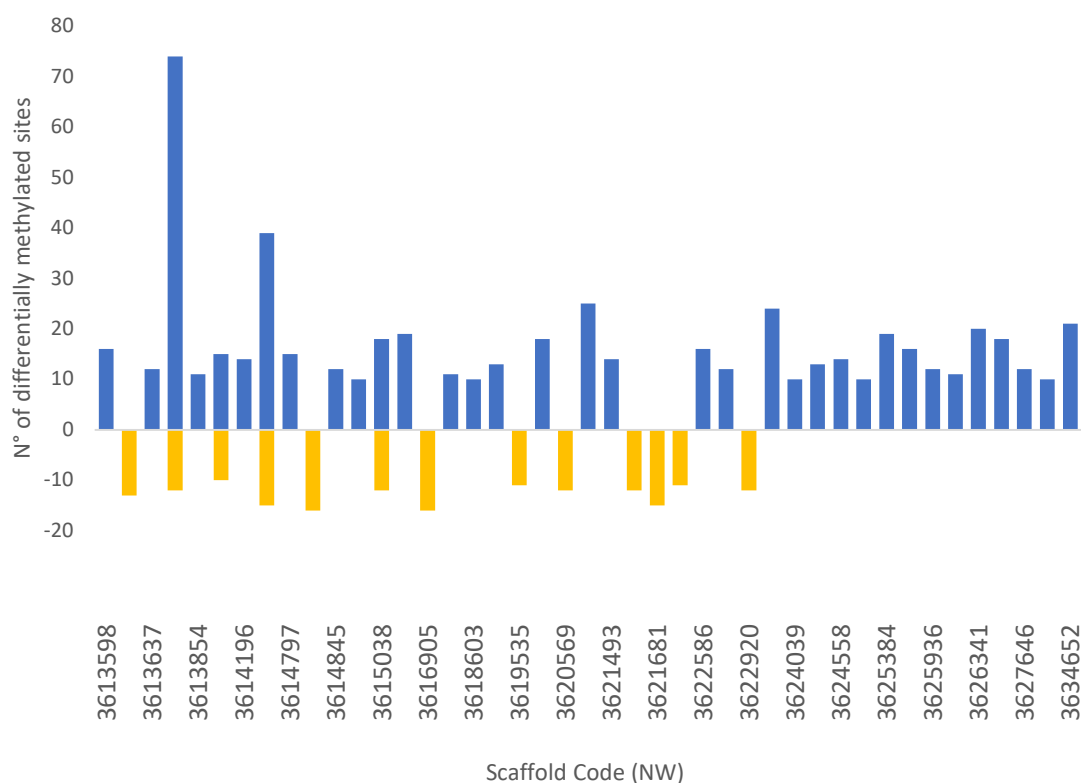


Figure 29: Graphical representation of DMS for each scaffold. Only scaffolds with more than 10 DMS are reported in this analysis. Blue bar indicated hypermethylated sites in Bad Producer with respect to Good Producer Clones, while the yellow bar indicates hypo-methylated positions in Bad Producer with respect to Good Producer clones.

With the aim to isolate specific differentially methylated and hydroxy-methylated genomic regions, we defined 5kb long regions containing at least 10 DMS or DHMS.

The analysis performed reveals 21 regions for the DNA methylation data set, while 14 regions are identified for the hydroxy-methylation data set. Interestingly, 9 regions are overlapping between the two data sets, indicating the coexistence of 5mC and 5hmC (Figure 30 and green boxes in Table 12).

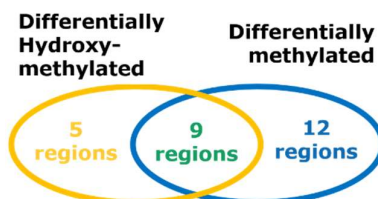


Figure 30: Venn Diagram of methylated (blue) and hydroxy-methylated regions (yellow). In green, are indicated the overlapping region between the two data set.

All the regions, except one, show hypermethylation of all DMS and DMHS in Bad Producer clones, with respect to Good Producer clones (Table 12A and B).

To better understand the functional activity of those selected regions, we search for an orthologous sequence in *Mus musculus* using the UCSC Genome Browser (University of California Santa Cruz Genomics Institute).

As summarized in Table 12A, the majority of the orthologous regions found are correlated to the intergenic sequence. However, few regions contain part of genes or promoter regions related to specific genes, such as SLC43a2, SQLE and MYC. As shown in Table 12A, SCL43a2 is involved in amino-acid transport, SQLE in sterol biosynthesis, and MYC is involved in cell growth. Since those genes are found to be Hyper-methylated in Good Producer Clones, these results are in agreement with their poor productivity.

Table 12A:

5kb regions with at least 10 DMS.

Green rows: common regions containing both DMS and DHMS; Yellow row: hypomethylated region; White rows: hypermethylated regions

Scaffold (NW)	start	End	n° of DMS	% Differential Methylation	UCSC annotation (Mus musculus)	Sequence biotype	GO category / Gene function (Genecards)
003615517v1	185697	189368	10	36,71			
003616958v1	83531	88856	10	50,76	SLC43a2		aminoacids transport
003617513v1	13817	18610	10	44,56		Repetitive region	
003619521v1	17803	18779	22	50,82			
003619648v1	17018	17920	18	42,15			
003621335v1	2246	6650	30	44,34			
003621580v1	1346	7467	18	47,56			
003622586v1	2528	4568	21	45,86			
003622859v1	1980	4337	13	37,41			
003623338v1	2246	3293	25	-34,35			
003623351v1	294	1152	26	38,20	SQLE	Last exon	Steroid biosynthesis

003623357v1	233	2159	10	39,47			
003623420v1	969	1861	14	45,73			
003625187v1	358	2331	18	47,18			
003625384v1	420	2311	39	48,93			
003625970v1	509	1336	11	43,12	MYC		cell growth
003626749v1	403	1480	12	36,74			
003626835v1	1662	2036	22	45,13			
003630050v1	409	1482	22	38,35			
003632385v1	245	1016	14	48,83			
003635404v1	176	741	10	67,61			

Table 12B summarizes the orthologous regions with at list DHMS, in *Mus musculus*. Even if the majority of the selected regions are related to intergenic regions, two genes are identified: SQLE and MYC. Since the DNA hydroxy-methylation is related to demethylation pathways and gene expression, these results could indicate the changing in the expression of these two genes.

Table 12B:

5kb regions with at least 10 DHMS.

Green rows: common regions containing both DMS and DHMS; White rows: hypermethylated regions

Scaffold (NW)	start	end	n° of DHMS	% Differential Methylation	UCSC annotation (Mus musculus)	Sequence biotype	GO category / Gene function (Genecards)
003618166v1	24034	43099	12	33,86			
003619521v1	17978	19390	18	57,85			
003619648v1	17206	17920	10	40,43			
003621335v1	4915	6609	28	43,69			
003621493v1	4381	8108	14	37,13			
003622859v1	1980	4274	10	35,63			
003622748v1	1333	1719	10	40,23			
003623351v1	271	1152	11	35,19	SQLE	last exon	Steroid biosynthesis
003624558v1	263	2287	10	42,02			
003625187v1	358	2331	15	50,7			
003625384v1	420	2311	36	52,85	MYC		cell growth
003625970v1	509	1336	12	44,72			
003626749v1	340	1480	12	36,09			
003626835v1	1737	1970	14	48,27			

It's known that DNA methylated regions could interact and regulate other regions at several kilobases of distance. For this reason, we look for the nearest genes of each selected region using the UCSC genome browser on *Mus musculus* (Figure 31). As shown in Table 13 we found two regions with neighbor genes. The region NW 003615517v1, from 185697 to 189368 (Figure 31A), has two neighbor genes: *Cdh5*, that are involved in intercellular junction and has a role in calcium-binding activity, and *Rpl22*, which is a structural component of Ribosomes. The second region, for which a neighbor gene is found, is NW 003617513v1, from 13817 to 18610 (Figure 31B); the nearest gene is *ASAP1*, which is involved in GTPase pathways.

Table 13:

Nearest genes reported for the previously selected regions (5kb regions with 10 DMS and DHMS).

Scaffold (NW)	start	end	Nearest genes	Distance	GO category / Gene function (Genecards)
003615517v1	185697	189368	<i>Cdh5</i>	2kb upstream	Intercellular junction/Calcium binding
003615517v1	185697	189368	<i>Rpl22</i>	10kb downstream	Structural constituent of ribosomes
003617513v1	13817	18610	<i>ASAP1</i>	5kb downstream	ARF GTPase activating proteing

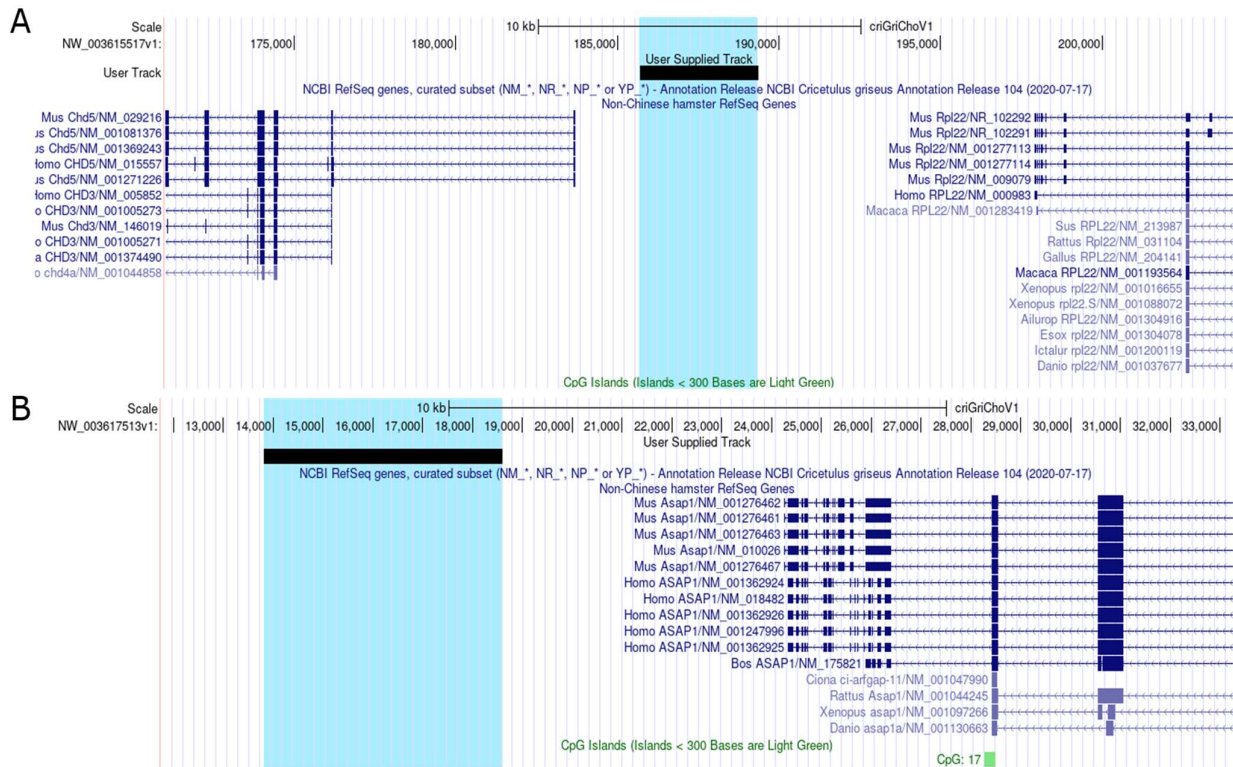


Figure 31: Regions previously selected, are analyzed for orthologous regions using UCSC genome browser. The region of interest is highlighted in light blue. The nearest genes are indicated in blue. **A)** NW_003615517v1, from 185697 to 189368. This region has two neighbor genes: this region is located downstream to *Rnpl22* and upstream to *Cdh5*. **B)** NW_003617513v1, from 13817 to 18610. This region is located downstream to *ASAP1*.

3.6. ChIP-Seq analysis studies for evaluation of genes/pathways related to clone productivity:

Results obtained from RNA-Seq and Methyl-Seq prompted further investigation relating to the role of the chromatin state in transgene productivity and stability. We decide to investigate the chromatin state at genome-wide level between the best clone, Clone 3, and the worst clone, Clone 24.

We carried out the Chromatin Immunoprecipitation followed by sequencing (ChIP-Seq), which is the most common method used to map proteins such as histones or transcription factors associated with genomic regions.

Briefly, chromatin is fixed using formaldehyde that is able to make covalent bound between chromatin protein and DNA. Then, the chromatin is isolated and fragmented to obtain mono- or di-nucleosomes (around 200bp DNA fragments). Once chromatin is fragmented, the antibody against a specific chromatin protein epitope is added for immunoprecipitation. The specific genomic regions bound to the protein of interest (ChIP enriched DNA) are finally purified for Library Preparation and sequencing (Figure 32).

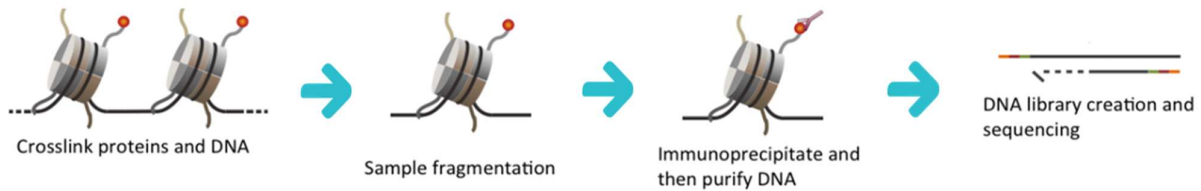


Figure 32: General Workflow of ChIP-Seq experiment. First, proteins are cross-linked to DNA. Second, the chromatin is fragmented to obtain mono-or di-nucleosome. Third, specific antibodies are used to immunoprecipitate specific proteins or modifications. Finally, the specific DNA regions isolated are purified and processed for Library prep and sequencing.

In order to understand open and close chromatin regions in the best producer clone respect the worst producer clone, we analyze only 3 of the six predictive histone marks: H3K27ac, which marks active promoter, TSS and enhancer, H4K4me3 that specifically marks the TSS of actively transcribed genes, and finally H3K27me3 that marks repressed genes.

This specific analysis is performed by Diagenode SA (Seraing, Belgium).

In order to evaluate the ChIP efficiency and specificity, the ChIP enriched DNA is analyzed by qPCR using primers for enriched regions (positive control) and unenriched regions (negative control) for the protein of interest. The promoter of GAPDH (GAPDH-TSS), a constitutively expressed gene, was used as positive control region for ChIP with H3K4me3 and negative control region for ChIP with H3K27me3 (Figure 33A and B). Myoglobin exon 2 (MBex2), a constitutive repressed gene, was used as negative control locus for H3K4me3 and H3K27ac and positive control region for ChIP with H3K27me3 (Figure 33A, B and C). And finally, the promoter of ACTB (ACTB-TSS), another constitutively expressed gene, was used as the positive control region for ChIP with H3K27ac (Figure 33A). Moreover, one more sample is processed in parallel using the negative control IgG in order to exclude a specific signal from the analysis.

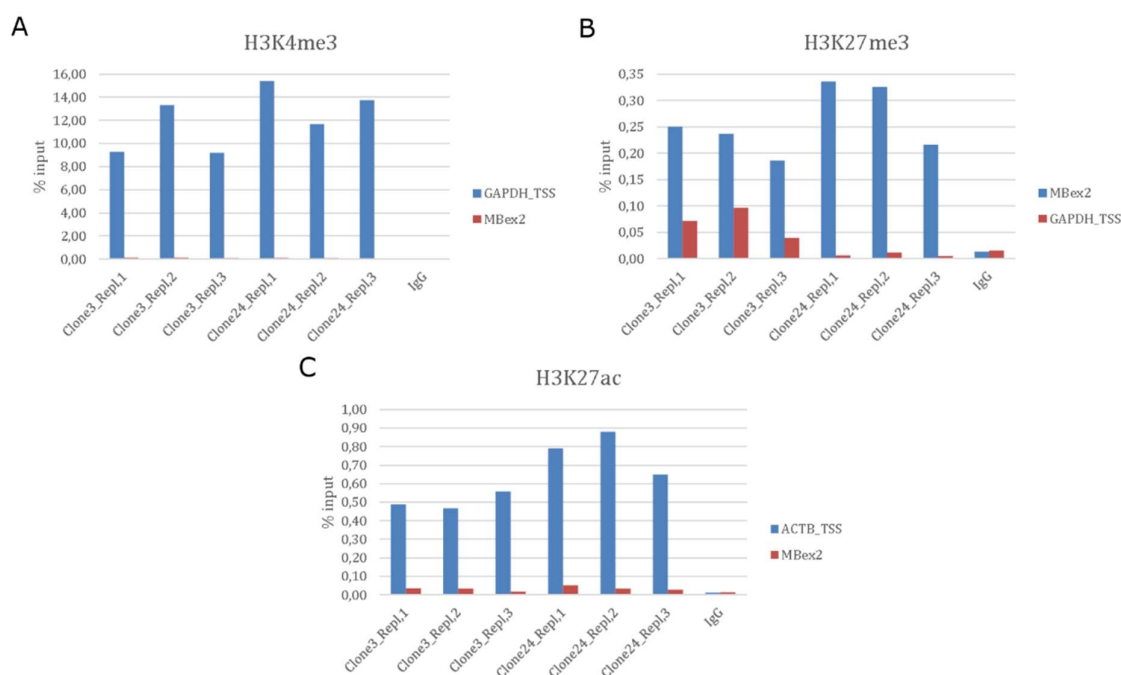


Figure 33: These figures show the recovery expressed as a percentage of input (the relative amount of immunoprecipitated DNA compared to input DNA after qPCR analysis). **A)** % of input related to H3K4me3 ChIP DNA. GAPDH_TSS is used as positive control locus, while MBex2 is used as negative control locus. **B)** % of input related to H3K27me3 ChIP DNA. GAPDH_TSS is used as negative control locus, while MBex2 is used as positive control locus. **C)** % of input related to H3K27ac ChIP DNA. ACTB_TSS is used as positive control locus, while MBex2 is used as negative control locus.

The ratios of the recovery for the positive regions over the background (corresponding to the recovery for the negative regions) were calculated for each mark (table 14). These ratios give an estimation of the specificity of the signal.

Table 14:

Recovery ratio for target regions respect to background. The ratio indicate the specificity of the signal.

Sample	Replicate	H3K4me3	H3K27me3	H3K27ac
		GAPDH_TSS/MBex2	MBex2/GAPDH_TSS	ACTB_TSS/MBex2
Clone 3	1	75.3	3.5	14.2
Clone 3	2	104.7	2.5	13.8
Clone 3	3	96.7	4.8	31.7
Clone 24	1	132.1	56.1	15.3
Clone 24	2	120.7	29.3	26.2
Clone 24	3	291.0	45.1	25.0
IgG		1.2	0.9	0.9

The results obtained reveal that efficient ChIP results were obtained for the 3 marks of interest: H3K4me3, H3K27me3 and H3K27ac.

As summarized in Table 15, more than 50 million reads were obtained for each sample, and more than 50% are unique mapped reads. Peak calling analysis is used to identify areas in a genome that have been enriched with aligned reads as a consequence of ChIP-sequencing. The peak calling was performed and optimized for H3K27ac, H3K27me3 and H3K4me3. The Table 15, summarize also the number of peaks detected for each sample. In Table 15, the Reads in peak (RiP) are also reported, which is the total number of reads falling inside the regions detected as peaks, while the Fraction of reads in peak (FRiP) is referring to the proportion of RiP over the total number of mapped reads.

Table 15:

Sequencing run summary with number of peak detected for each sample.

Sample	Mark	Total Reads	% Unique Mapped Reads	Peaks	Average Peak width	Average Peak Score	Reads in Peak (RiP)	Fraction of reads in peak (FRiP)
Clone3_FB1	H3K4me3	93,477,050	66.93	28,895	3044	250	33,556,322	56
Clone3_FB2	H3K4me3	87,608,668	70.88	28,847	3108	251	34,726,466	56
Clone3_FB3	H3K4me3	87,863,696	70.86	29,116	3159	252	34,649,729	56
Clone24_FB1	H3K4me3	93,817,552	67.52	38,127	2518	269	36,862,811	59
Clone24_FB2	H3K4me3	109,121,832	67.96	41,784	2476	251	40,871,762	56
Clone24_FB3	H3K4me3	88,982,134	70.74	38,14	2339	281	38,473,564	62
Clone3_FB1	H3K27me3	109,578,162	62.00	2,595	226053	63	28,457,124	42
Clone3_FB2	H3K27me3	115,685,864	57.98	2,921	198573	72	30,169,453	45
Clone3_FB3	H3K27me3	100,297,608	60.43	3,087	185312	80	28,678,333	48
Clone24_FB1	H3K27me3	121,077,150	48.05	3,207	153105	106	31,853,976	55
Clone24_FB2	H3K27me3	106,708,698	50.07	3,091	159970	97	26,446,095	50
Clone24_FB3	H3K27me3	128,983,120	46.57	3,211	154414	103	32,255,457	54
Clone3_FB1	H3K27ac	82,921,146	52.61	27,174	3910	213	16,999,512	40
Clone3_FB2	H3K27ac	91,989,754	39.76	30,617	3043	230	15,406,969	43
Clone3_FB3	H3K27ac	90,203,266	51.38	27,126	4072	226	20,093,443	44
Clone24_FB1	H3K27ac	63,052,068	36.14	37,119	2598	238	8,711,035	39
Clone24_FB2	H3K27ac	45,724,376	46.41	28,813	3797	244	8,873,578	42
Clone24_FB3	H3K27ac	69,832,964	55.18	34,832	3442	192	12,064,031	32

For each mark, H3K4me3, H3K27me3 and H3K27ac, a differential binding analysis was performed between Clone 3 and Clone 24.

Identified peaks, obtained from all biological replicates, were firstly analyzed using PCA analysis for each histone mark. Samples analyzed for H3K4me3 marks show 69% of variance in the first principal component (PCA1), while 10% of variance is explained by the second principal component (PCA2, Figure 34A). In the case of the H3K27me3, 95% of variance is explained by the first principal component and only 2% is explained by the second principal component (Figure 34B). Finally, the analysis of H3K27ac marks in the same samples shows 71% of variance in the first principal component, while 9% of variance is explained by the second principal component (Figure 34C). For all of the histone marks analyzed, biological replicates for Clone 3 cluster separately respect biological replicates of Clone 24.

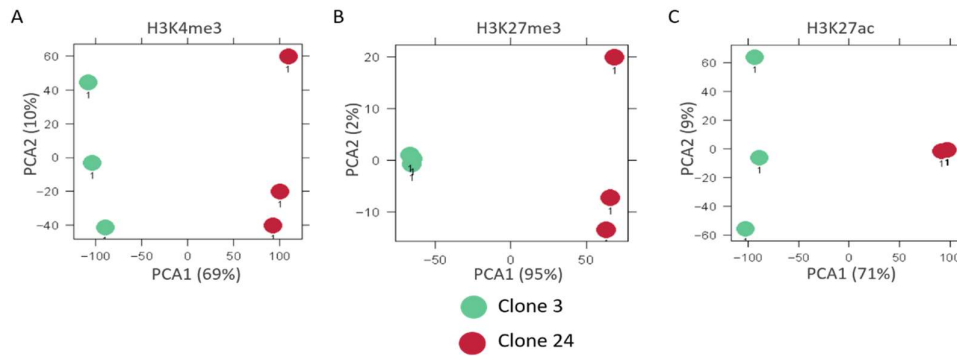


Figure 34: Principal Component Analysis of the three histone marks analyzed: H3K4me3, H3K27me3 and H3K27ac. The Best Producer Clone, Clone 3 is indicated in green, while the Worst Producer Clone, Clone 24, is indicated in red. The corresponding percentage of explained variance is indicated in the x-axis for the first principal component (PCA1) and on the y-axis for the second principal component (PCA2) **A)** For H3K4me3 marks shows 69% of variance in the first principal component, while 10% of variance is explained by the second principal component. **B)** In the analysis of the H3K27me3 mark, 95% of variance is explained by the first principal component and only 2% is explained by the second principal component. **C)** The analysis of H3K27ac marks shows 71% of variance in the first principal component, while 9% of variance is explained by the second principal component.

To further characterize CHO-K1 SV from the epigenetic and chromatin point of view, the identified peak for the H3K4me3, H3K27me3 and H3K27ac histone marks were analyzed for the differential binding analysis. Differentially Bound Regions (DBR) identified were plotted using the MA plot, where the log₂ Fold-change is plotted on y-axis, while the log concentration is plotted on x-axis (Figure 35). The DBRs were identified using the p-value adjusted used is minor than 0,01 ($p\text{-adj} < 0,01$) for all the histone marks. Related to the analysis of the H3K4me3 histone marks, 8751 regions were found to be statistically differentially bound between Clone 3 and Clone 24. Next, 1559 regions were found to be differentially bound in the H3K27me3 analysis. Lastly, 6193 regions were identified as DBRs (Table 16). Considering the cut-off of log₂ Fold-change ≤ -1 and log₂ Fold-change ≥ 1 , 4651 regions were differentially bound in the H3K4me3 analysis, 770 regions were identified for the H3K27me3 mark, while in the H3K27ac the DBRs were 3940 (Table 16). Those regions are further investigated for the annotated features on the *Cricetulus griseus* genome and Pathway analysis.

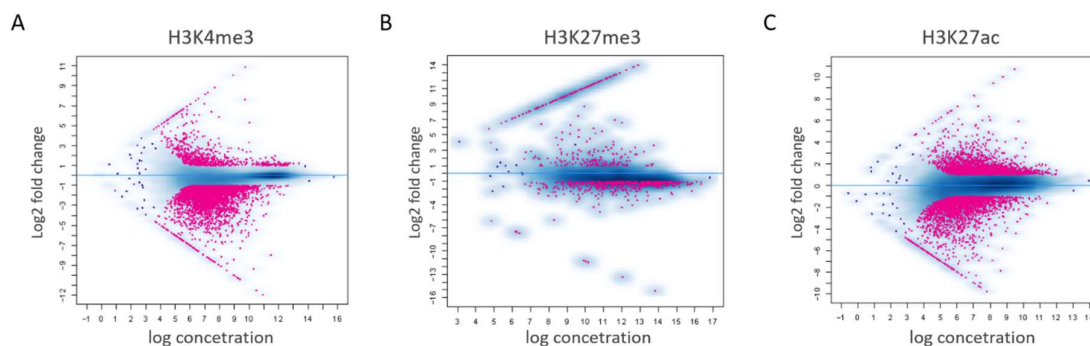


Figure 35: Differentially bound regions (DBRs) between Clone 3 and Clone 24 shown in MA Plot: log₂ Fold-change on y-axis is plotted with log₁₀ (padj) on x-axis. Dots in blue indicate non-significant regions ($p\text{-adj} > 0,01$) and log₂ Fold-change < 1 or > 1 ($-1 < \log_2 \text{Fold-change} < 1$). Dots in pink indicates significant bound regions with $p\text{-adj} > 0,01$ and log₂ Fold-change ≤ -1 and log₂ Fold-change ≥ 1 . Those genes, indicated in pink, are further investigated. **A)** DBRs between Clone 3 and Clone 24 related to the H3K4me3 analysis. **B)** DBRs between Clone 3 and Clone 24 related to the H3K27me3 analysis. **C)** DBRs between Clone 3 and Clone 24 related to the H3K27ac analysis.

Table 16:

Number of differentially bound regions of each histone mark analysis respect to p-adj and log2 Fold-change cut-off

Histone mark	Differentially Bound regions	
	p-adj > 0,01	p-adj > 0,01 and log2 Fold-change ≤ -1 and log2 Fold-change ≥ 1
H3K4me3	8751	4651
H3K27me3	1159	770
H3K27ac	6193	3940

To further investigate and identify open and close chromatin regions in the best and worst producer clones, Differentially Bound Regions for each of the histone marks are analyzed regarding annotated genomic features: promoter, exon, introns, intergenic regions and from 1 to 5kb upstream of the TSS regions (1to5kb).

Figure 36A shows the genomic distribution of differential bound peak with respect to the H3K4me3 marker. The majority of DBR found are annotated as intergenic regions, Introns and Exons. Moreover, a little enrichment is also found in the region upstream of the TSS.

As shown in Figure 36B, the H3K27me3 mark is homogeneously spread on all the annotated genomic features with 20% of a differential bound peak for each region, indicating chromatin region with low accessibility. These results agree with the H3K27me3 distribution and its correlation with gene repression and facultative heterochromatin.

Results obtained for the H3K27ac histone mark indicate that the majority of DBRs are found at the level of Intergenic regions, Introns and Exons. Enrichment of peaks were also found in the region upstream of the TSS (Figure 36C). These results are in line with the ENCODE classification, where the H3K27ac mark normally identifies active enhancer region and actively transcribe genes.

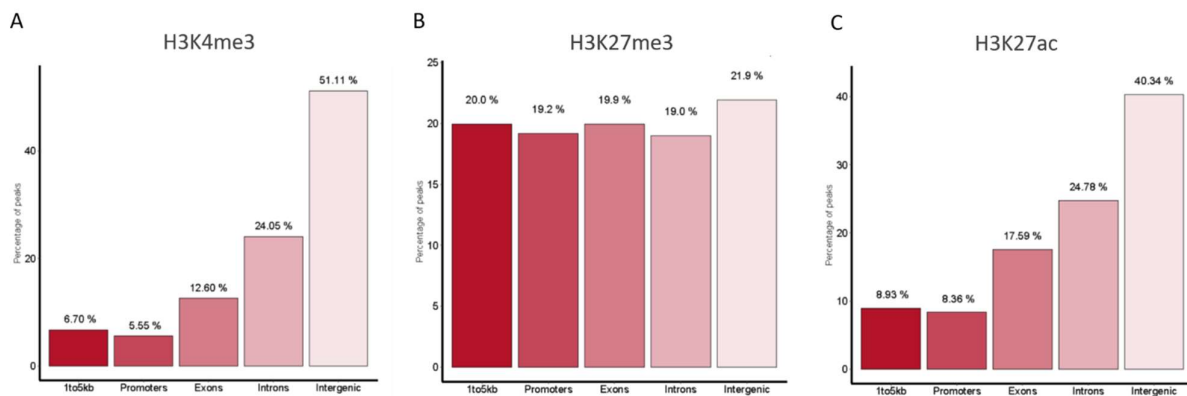


Figure 36: Genomic distribution of DBRs peak with respect to promoters, exons, introns, intergenic regions and from 1 to 5Kb upstream of the TSS regions (1to5kb). The analysis is performed for each histone mark: **A)** H3K4me3, **B)** H3K27me3 and **C)** H3K27ac.

To further investigate DBR and its function with respect to specific histone mark analysis, the peaks obtain were analyzed to find out associated genes. In the H3K4me3 analysis, 2149 were found to be associated to the DBR between Clone 3 and Clone 24, 4075 genes are identified for the H3K27me3 mark, and 2417 genes are associated to differential bound peak in H3K27ac analysis.

Those genes are used for the pathway analysis using KEGG database. However, since *Cricetulus griseus* reference list it's not present on KEGG database, this analysis is performed using *Mus musculus* gene data.

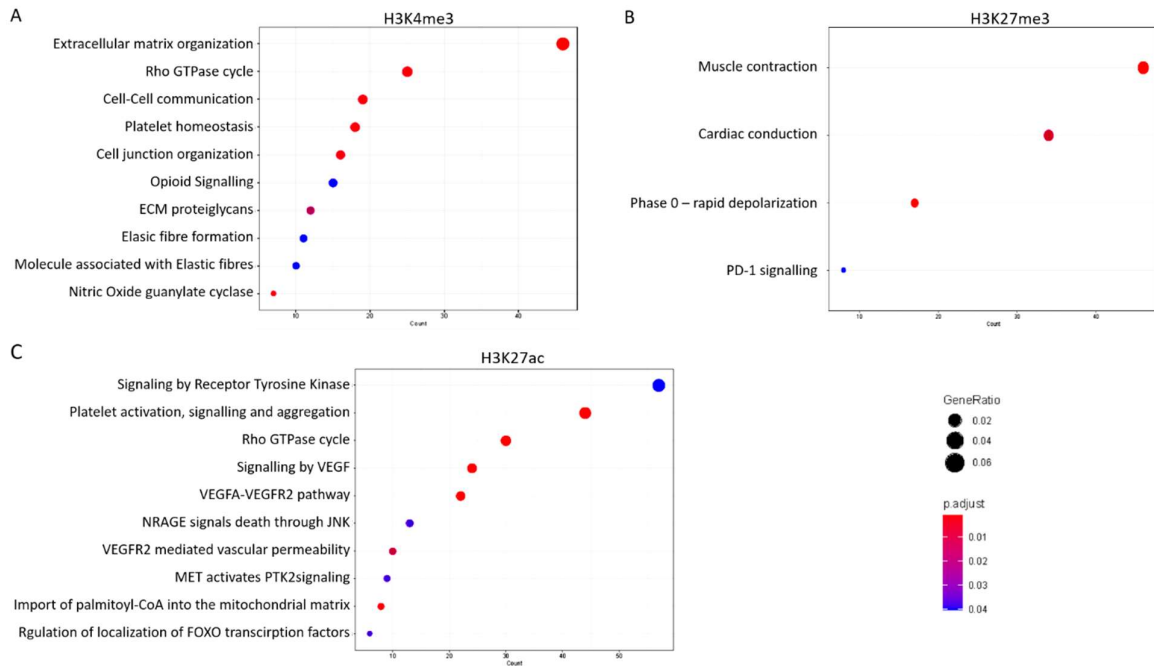


Figure 37: Pathway analysis on DBR associated genes performed using mouse KEGG database. For each dot in the graph, the related enriched pathway is indicated on the y axis. Dimension of each dot in the graph is related to the gene ratio obtained (from 0.02 to 0.06), while dot colored is referred to the p-value adjusted (red for $p\text{-adj} < 0,01$ and blue for $p\text{-adj} < 0,05$)

The Pathway Analysis shows that higher differences between Clone 3 and Clone 24 at chromatin level are found inaccessible regions respect to close chromatin domains. In particular, the H3K4me3 analysis shows a high enrichment in “Extracellular matrix organization”, “Rho GTPase cycle” and “cell-cell communication pathways”. In the same analysis, is also evident the enrichment of other pathways related to the cell-matrix, cell junction and cell signaling. Interestingly, the “Platelet homeostasis pathway” is also enriched (Figure 37A). These data are confirmed by the analysis of the other open chromatin region marker, H3K27ac. In this analysis, there is high enrichment of “Signaling by Receptor Tyrosine Kinase”, “Platelet activation, signaling and aggregation”, and “Rho GTPase cycle”. The other pathways that are found enriched in the H3K27ac analysis are related to vascular signaling and extracellular matrix pathways (Figure 37C). On the contrary, the Pathway analysis of the inaccessible region marker, the H3K27me3 histone mark, shows different pathway enrichment, such as “Muscle contraction”, “Cardiac conduction”, “Phase 0 – rapid depolarization” and “PD-1 signaling”.

All of these data are in agreement with Transcriptome data, where highly Differentially expressed genes are related to cell signaling pathways, GTPase signaling and extracellular matrix.

5. Discussion

Despite the well-established approach, the manufacturing process remains time-consuming and quite expensive. During the last few years, the growing demand for biopharmaceuticals puts the focus on the need to speed up the processes of drug development and their production^{1,2}.

Currently, the Clone screening phase is one of the most time-consuming steps of the manufacturing process, and despite many efforts being in place to identify the best candidate, it's still difficult to predict the production capacity and growth characteristics of clonal cell lines at the production scale^{1,2,3}.

Moreover, knowledge of the molecular aspects that drive the productivity and stability of clones are still not well understood¹.

For these reasons, in this work, we investigate the effectiveness of genetic parameters and molecular biology technologies during the high-throughput screening of clones produced during the Cell Line Development, as representative for the large-scale bioproduction. With the purpose to help in the selection of the best clone, this project would set the stage for the use of molecular biology high-throughput analysis, such as the Droplet Digital PCR and the Next Generation Sequencing, during production phases.

Here, we provide the first integrative studies that underline differences between good and bad producer clones for the identification of biomarkers or new targets for CHO cell engineering. While recent articles provide GCN analysis, Transcriptome and Methyome data, and studies on chromatin organization^{8,15,30,35,36}, our study provides the first complete picture, from a genetic point of view, of good and bad producer clones, using a wide variety of molecular biology approaches.

We first select 4 clones, derived from a single transgene insertion process in CHO-K1 SV cells, based on their productivity data: 2 of the clones are good producers, while the others are bad producer clones. Then, we analyzed them for different genetic parameters, such as the Gene Copy Number, transgene mRNA expression and Transcriptome, Methyome and we also look to chromatin organization.

We found that the good producer clones have the higher copy number, while bad producer clones show a low range of transgene copies. These results indicate a possible correlation between the number of transgene copies that are integrated into the CHO-K1 SV genome and productivity data.

Despite both coding regions are in the same plasmid vector, it was also shown for Clone 3 that the Light Chain copy number is higher with respect to the Heavy Chain region. This result suggests a partial integration or a partial loss of the construct sequence in some of the copies that are inserted in the CHO genome. However, to better understand how and where the transgene is inserted, other technologies are required, such as the Nanopore sequencing or Target Locus Amplification (TLA) method.

The genomic context in which the construct is inserted, it's important because is able to modify their expression and stability during cell culture in bioreactor^{8,15,26}. In spite of the use of a strong promoter, like the CMV promoter, to decrease the risk of gene silencing, the transgene copies could undergo the effect of the genomic region in which are inserted, and finally influence the transgene expression²⁶.

Therefore, in our study, we also analyzed the transgene mRNA expression using the ddPCR. Correlating with Copy Number analysis and productivity data, the higher expression of the transgene mRNA is found in Good Producer clones while compared with Bad producer clones, which exhibit a low ratio for the transgene mRNA. Moreover, higher Light Chain expression was found both in Good and Bad producer clones, except for the Clone 9. The higher expression of the Light Chain with respect to the Heavy Chain is possible because of the control of two separate CMV promoters. However, the ratio between Light Chain and Heavy Chain expression is around 1, indicating that the productivity of the clone is mediated by both Heavy and Light Chain.

In addition to these results, RNA-Seq analysis performed on those clones confirmed the higher expression of the transgene in Good Producer clones with respect to the Bad ones.

It was shown from several articles that the productivity of cell banks is not only related to the direct production of the transgenic protein, in fact, it was demonstrated that also all the cell machinery involved in expression and production are required to support recombinant protein production. For this reason, in this work, we investigated the entire Transcriptome of the selected clones^{18,20}.

Thanks to Transcriptome data, several top upregulated and downregulated genes were identified in Good Producer clones with respect to Bad Producer clones. Most of them are related to G-Protein Coupled Receptor (GPCR) signaling, suggesting a role in cell proliferation and metabolism regulations. Moreover, different proteases are downregulated in Good Producer clones: these genes likely avoid cutting phenomena on the recombinant protein production. Interestingly, many downregulated genes are also involved in tissue-specific pathways and unrelated to productivity functions.

Using the PANTHER tool, we were able to discriminate between different GO categories, specifically in Biological Process. Since the network of vesicle-mediated and intra-cellular transport was reported to be critical for clone productivity, we focused on those related GO categories. The category related to Endosomal transport comes evident only by analyzing the GO category associated with down-regulated genes, indicating that clones down-regulate the Endosomal process and tissue-specific vesicle pathway in order to promote protein secretion and exocytosis process.

This result was probably due to a repression of retrograde transport in Good Producer Clones, in order to enhance the secretory capacity. However, to clearly address this effect, specific studies should be performed, for example, silencing or overexpressing specific genes involved in these pathways.

Epigenetic silencing of the recombinant genes can be listed among the prime causes leading to reduced recombinant protein production. Hence, in this work, we analyze epigenetic modification affecting both DNA methylation and Chromatin structure²⁶.

Using the Whole Genome Bisulfite Sequencing (WGBS) to provide an affordable method for assaying DNA methylation at single base-pair resolution, we have assayed the DNA methylation in Good and Bad Producer clones. Moreover, thanks to the Oxidation step, performed prior to the Bisulfite conversion, we were also able to identify the true level of the 5mC and to isolate the signal from the 5hmC.

Previous studies have shown that promoter hypermethylation contributes to loss in recombinant protein production, while later a CHO cells DNA methylation landscape was published³⁰. The global characterization of DNA methylation in the CHO-DP12 cell genome indicated heterogeneous arrangements of hyper and hypomethylated regions, leading to a bimodal distribution of DNA methylation³⁰. In our study, we confirm the bimodal distribution of DNA methylation in both Good and Bad producer clones, indicating that is a normal state of the CHO genome as other mammalian cells. Therefore, differences that characterize Good producer clones should be found in specific regions of the CHO genome and not in the overall DNA methylation state. Despite the CHO genome is poorly annotated, in our study, we investigate the differential DNA methylation between Good and Bad producer clones at the functional annotated region in the genome, such as CpG island, promoter, exon, intron and intergenic regions. Moreover, we found the highest difference at the level of the CpG shores, which are normally found at 2kb upstream of a CpG island and hence is the 10Kb distance from the TSS.

Isolating regions of 5kb with more than 10 DMS, we identified a few genes involved in steroid biosynthesis and amino-acid transport. Notably, some of the regions found are overlapping with regions that contain more than 10 DHMS. Since the presence of the 5hmC is related to gene expression, this data could indicate that overlapping regions between high DMS and DHMS are regions that are changing their expression level indicating those regions that could change in expression during cell culturing of clones.

Moreover, using the UCSC genome browser we also look to regions in close proximity to DHMS and DMS with the aim to identify possible genes that could be influenced. We found only 3 genes, that have a role in cell-matrix and as GTPase signaling. These results are in agreement with RNA-Seq data, where we found GO category related to GPCR signaling and matrix proteases. Moreover, it was also possible to identify the Rpl22

gene, which is a ribosome constituent, indicating that the translational machinery is critical for recombinant protein production.

Within the various epigenetic signatures, histone marks represent the highly used markers to understand the chromatin structure in human cells. The ENCODE project help in defining the cross-talk of histone modification with the expression or silencing effects³⁸. To elucidate the role of the chromatin structure we analyzed the H3K4me3, H3K27me3 and H3K27ac, respectively corresponding to TSS of expressed genes, repressed genes and regulatory regions of expressed genes.

As done for differentially methylated regions, we have analyzed the same annotated functional region also for the differential binding region in the 3 histone marks. According to ENCODE project, the H3K27me3 marks are spread on the promoter, exon, intron and intergenic region, indicating their repressive condition. On the contrary, markers of open chromatin regions, H3K4me3 and H3K27ac, shows differences in the differentially binding regions. These results could indicate that those markers, H3Kme3 and H3K27ac, should be much more predictive to discriminate between Good and Bad Producer clones, respect the H2K27me3 marks.

Taking advantage of KEGG database, we also identify different pathways associated with the Differentially bound Region in the 3 histone marks analyzed. As obtained from Transcriptome data, the pathways identified have a role in cell signaling with GTPase protein and in a cellular matrix organization and also relating to vascular function.

These results might be investigated in detail to clarify their relationship with clone productivity and their possible use as genetic markers.

6. Conclusion and Future Perspective

Clone screening is still one of the longest steps for biopharmaceuticals production. It requires many efforts to identify the best producer clone to be used as MCB. Therefore, the most powerful and cutting-edge technologies were used to improve knowledge and clarify the molecular mechanisms that characterize good producers with respect to bad producer clones.

In this work technologies like the Droplet Digital PCR and the Next Generation Sequencing were used to investigate different genetic parameters that could be used in the Clone screening phase.

Importantly, this project had the aim to gain more knowledge on what distinguished “Good Producer clone” from “Bad producer clone” through the analysis of Copy Number, Transcriptome, Methylome and Chromatin region analysis.

After carrying out this broad variety of genetic analyses, we were able to identify possible predictive markers and some genetic signatures for the productivity of the clone during the clone screening phase of the manufacturing process. However, detailed studies to clarify interaction and cross-talk between specific pattern of genes found in our analysis and the clone productivity, are still necessary.

Altogether, our results show the importance of the use of genetic parameters, spacing from genomics, transcriptomics and epigenomics, to predict transgene expression and hence the recombinant protein production at the early phases of the manufacturing process, helping in saving cost and time.

With our work, we have set the stage for molecular biology technologies to be used as a screening method in the early phases of the manufacturing process. Moreover, these new tools could be used for the development of new analytical methods in the quality control context and production monitoring during the manufacturing process.

Taking advantage of Transcriptome data, our work represents a good starting point for the identification of gene editing targets. Genes involved in the productivity and stability of the recombinant protein could be overexpressed or silenced to increase CHO-K1 host productivity, thanks to different gene-editing technologies such as the CRISPR-Cas9. Moreover, based on ChIP-Seq data, it's possible to identify open chromatin regions, that are more prone to transgene expression. Those regions could be used for the targeted insertion of the transgene of interest with the CRISPR-Cas9 or Transposase systems, avoiding the risk of generating non-productive clones and lowering the times for the clone screening.

7. References

1. Matasci, M., Hacker, D. L., Baldi, L. & Wurm, F. M. Protein therapeutics., Recombinant therapeutic protein production in cultivated mammalian cells : current status and future prospects. 37–42 (2008). doi:10.1016/j.ddtec.2008.12.003
2. Wurm, F. M. Production of recombinant protein therapeutics in cultivated mammalian cells. **22**, 1393–1398 (2004).
3. Stolfi, G. *et al.* CHO-Omics Review : The Impact of Current and Emerging Technologies on Chinese Hamster Ovary Based Bioproduction. **1700227**, 1–14 (2018).
4. Jesus, M. De & Wurm, F. M. European Journal of Pharmaceutics and Biopharmaceutics Manufacturing recombinant proteins in kg-ton quantities using animal cells in bioreactors. *Eur. J. Pharm. Biopharm.* **78**, 184–188 (2011).
5. Hacker, D. L., Jesus, M. De & Wurm, F. M. 25 years of recombinant proteins from reactor-grown cells — Where do we go from here ? *Biotechnol. Adv.* **27**, 1023–1027 (2009).
6. Bandaranayake, A. D. & Almo, S. C. Recent advances in mammalian protein production. *FEBS Lett.* **588**, 253–260 (2014).
7. Datta, P., Linhardt, R. J. & Sharfstein, S. T. An ‘ Omics Approach Towards CHO Cell Engineering. doi:10.1002/bit.24841
8. Hernandez, I. *et al.* Epigenetic regulation of gene expression in Chinese Hamster Ovary cells in response to the changing environment of a batch culture. doi:10.1002/bit.26891
9. Lattenmayer, C. *et al.* Characterisation of recombinant CHO cell lines by investigation of protein productivities and genetic parameters. **128**, 716–725 (2007).
10. He, L., Winterrowd, C., Kadura, I. & Frye, C. Transgene Copy Number Distribution Profiles in Recombinant CHO Cell Lines Revealed by Single Cell Analyses. **109**, 1713–1722 (2012).
11. Droplet Digital PCR: Application Guide. BioRad, Bulletin 6407 (Ver. B)
12. Lin, H., Okumura, T., Yatsuda, Y., Ito, S. & Al, L. I. N. E. T. Application of Droplet Digital PCR for Estimating Vector Copy Number States in Stem Cell Gene Therapy. **27**, 197–208 (2016).
13. Berman, J. *et al.* Droplet Digital TM PCR : High-Resolution Copy Number Variation Analysis. 6–7
14. Ra, N., Dreo, T., Gutierrez-aguirre, I., Blejec, A. & Ravnkar, M. Reverse transcriptase droplet digital PCR shows high resilience to PCR inhibitors from plant , soil and water samples. 1–10 (2014). doi:10.1186/s13007-014-0042-6

15. Veith, N., Ziehr, H., Macleod, R. A. F. & Reamon-buettner, S. M. Mechanisms underlying epigenetic and transcriptional heterogeneity in Chinese hamster ovary (CHO) cell lines. *BMC Biotechnol.* 1–16 (2016). doi:10.1186/s12896-016-0238-0
16. A Beginner's guide to NGS (Illumina), [Next-Generation Sequencing for Beginners | NGS basics for researchers \(illumina.com\)](https://www.illumina.com/genomics/solutions/next-generation-sequencing-for-beginners.html)
17. Monger, C. *et al.* Towards next generation CHO cell biology : Bioinformatics methods for RNA-Seq-based expression profiling. 950–966 (2015). doi:10.1002/biot.201500107
18. Vishwanathan, N., Le, H., Le, T. & Hu, W. ScienceDirect Advancing biopharmaceutical process science through transcriptome analysis. *Curr. Opin. Biotechnol.* **30**, 113–119 (2014).
19. Sha, S., Bhatia, H. & Yoon, S. An RNA-seq based transcriptomic investigation into the productivity and growth variants with Chinese hamster ovary cells. *J. Biotechnol.* **271**, 37–46 (2018).
20. Gram, H., Fastrup, H. & Min, G. Improving the secretory capacity of Chinese hamster ovary cells by ectopic expression of effector genes : Lessons learned and future directions. **35**, 64–76 (2017).
21. Doolan, P. *et al.* Transcriptomic analysis of clonal growth rate variation during CHO cell line development. *J. Biotechnol.* **166**, 105–113 (2013).
22. Shridhar, S. *et al.* Transcriptomic changes in CHO cells after adaptation to suspension growth in protein-free medium analysed by a species-specific microarray. **257**, 13–21 (2017).
23. Schübeler, D. Function and information content of DNA methylation. (2015). doi:10.1038/nature14192
24. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, (2009).
25. Portela, A. & Esteller, M. review Epigenetic modifications and human disease. *Nat. Publ. Gr.* **28**, (2010).
26. Wippermann, A. & Noll, T. DNA methylation in CHO cells. *J. Biotechnol.* **258**, 206–210 (2017).
27. Feichtinger, J. *et al.* Comprehensive Genome and Epigenome Characterization of CHO Cells in Response to Evolutionary Pressures and Over Time. **113**, 2241–2253 (2016).
28. Shi, D., Ali, I., Tang, J. & Yang, W. New Insights into 5hmC DNA Modification : Generation , Distribution and Function. **8**, 1–11 (2017).
29. Booth, M. J., Branco, M. R., Ficz, G., Oxley, D. & Krueger, F., Quantitative Sequencing of 5-Methylcytosine and 5-Hydroxymethylcytosine at Single-Base Resolution. **8333**, 934–938 (2012).
30. Wippermann, A., Rupp, O., Brinkrolf, K., Hoffrogge, R. & Noll, T. The DNA methylation landscape of Chinese hamster ovary (CHO). *J. Biotechnol.* **199**, 38–46 (2015).

31. Patel, N. A. *et al.* Antibody Expression Stability in CHO Clonally Derived Cell Lines and Their Subclones: Role of Methylation in Phenotypic and Epigenetic Heterogeneity. (2018). doi:10.1002/btpr.2655
32. Osterlehner, A., Simmeth, S., Go, U., Development, E. & Gmbh, R. D. Promoter Methylation and Transgene Copy Numbers Predict Unstable Protein Production in Recombinant Chinese Hamster Ovary Cell Lines. **108**, 2670–2681 (2011).
33. Spencer, S., Gugliotta, A., Koenitzer, J., Hauser, H. & Wirth, D. Stability of single copy transgene expression in CHOK1 cells is affected by histone modifications but not by DNA methylation. *J. Biotechnol.* **195**, 15–29 (2015).
34. Yang, Y., Chusainow, J. & Yap, M. G. S. DNA methylation contributes to loss in productivity of monoclonal antibody-producing CHO cell lines. *J. Biotechnol.* **147**, 180–185 (2010).
35. Moritz, B., Woltering, L., Becker, P. B. & Ulrich, G. High Levels of Histone H3 Acetylation at the CMV Promoter Are Predictive of Stable Expression in Chinese Hamster Ovary Cells. (2016). doi:10.1002/btpr.2271
36. Wippermann, A., Rupp, O., Brinkrolf, K., Hoffrogge, R. & Noll, T. Integrative analysis of DNA methylation and gene expression in butyrate-treated CHO cells. *J. Biotechnol.* **257**, 150–161 (2017).
37. Calo, E. & Wysocka, J. Review Modification of Enhancer Chromatin: What? How, Why? *Mol. Cell* **49**, 825–837 (2013).
38. Kimura, H. Histone modifications for human epigenome analysis. 439–445 (2013). doi:10.1038/jhg.2013.66
39. Chai, Y., Ge, M., Wei, T., Jia, Y., Guo, X. & Wang, T. Human rhinovirus internal ribosome entry site element enhances transgene expression in transfected CHO-S cells. 8:6661 (2018). doi:10.1038/s41598-018-25049-9
40. Gosdschan, A., Yusuf, D. & Grüning, B. Strategies for analyzing bisulfite sequencing data. **261**, 105–115 (2017).
41. Mi, H., Muruganujan, A., Huang, X., Guo, X. & Thomas, P. D. Protocol Update for large-scale genome and gene function analysis with the PANTHER. *Nat. Protoc.* (2016). doi:10.1038/s41596-019-0128-8
42. Adusumalli, S., Feroz, M., Omar, M., Soong, R. & Benoukraf, T. Methodological aspects of whole-genome bisulfite sequencing analysis. **16**, (2018).