UNIVERSITY OF TURIN

DOCTORAL SCHOOL IN LIFE AND HEALTH SCIENCES

Ph.D. Program in Biomedical Sciences and Oncology

XXIX Cycle

Next generation sequencing methods development for quality control in biopharmaceuticals

Candidate: Torre Serena

Tutor: Cabodi Sara

Site: Istituto di ricerche Biomediche "A.MARXER" R.B.M. - Merck (Ivrea)

Supervisor: Fabio La Neve

Academics Years: 2013/2017

S.S.D.: CHIM/09

Contents

ACRONYMS

ATCC	American Type Culture Collection
BH	Bulk Harvest
bp	Base pair
BPS	Bioprocess Science Department
BWA	Burrows-Wheeler Aligner
C°	Celsius
CBER	Center of Biologics Evaluation and Research
CFR	Code of Federal Regulations
CHO	Chinese Hamster Ovary cell line
cm	centimeters
cDNA	comlementary DNA
db	database
DNA	Deoxyribonucleic Acid
dNTP	Deoxyribonucleotide triphosphate
ds	double strand
dscDNA	double strand complementary DNA
DVP	Division of Viral Products
dUTP	Deoxyuridine triphosphate
EoPCB	End of Production Cell Bank
FeLV	Feline Leukemia Virus
FDA	Food and Drug Administration
g	gram
Gb	Giga base (1 billion of bp)

GPEx	Gene Product Expression
HAP	Human Antibody Production tests
HC	heavy chain
ICH	International Conference on Harmonization
Kb	Kilo base (1 thousand of bp)
LC	light chain
LOD	Limit of Detection
mAb	monoclonal Antibody
MAP	Mouse Antibody Production tests
MCB	Master Cell Bank
Mbp	Mega base (1 million of bp)
MDS	Multi-Dimensional Scaling
mL	millilitre
MLV	Moloney Leukemia Virus
mm	millimeters
mM	milliMolar
NA	Nucleic Acid
NCBI	National Center for Biotechnology Information
NEB	New England BioLabs
ng	nanogram
NGS	Next Generation Sequencing
ONT	Oxford Nanopore Technologies
PacBio	Pacific Bioscience
PBS	Phosphate Buffered Saline
PCR	Polymerase Chain Reaction

PCV	Porcine Virus
PhD	Philosophiæ Doctor
PF	Passing Filter
PFGE	Pulsed-field gel electrophoresis
RFLP	restriction fragment length polymorphism
qPCR	quantitative Polymerase Chain Reaction
RIN	RNA Integrity Number
rpm	revolutions per minute
RNA	Ribonucleic Acid
rRNA	ribosomal RNA
RT	Retrotranscription
Seq	Sequencing
SBS	Sequencing by synthesis
ss	single strand
TGS	third generation sequencing
TU	Titration Unit
UDP	Uridine Diphosphate
USB	Universal Serial Bus
WCB	Working Cell Bank
µg	microgram
µl	microliter

BACKGROUND

The great majority of biopharmaceuticals, approved by regulatory authorities, are produced by recombinant DNA technology in various expression system. The most widely used hosts are *mammalian cells* (e.g. *Chinese Hamster Ovarian cells)* that are difficult to maintain but enable to obtain a higher degree of protein quality [1].

The biopharmaceutical manufacturing process is complex and focused on increasing final protein yield, assuring the quality of the products and reducing costs. The industries are driven in these directions by improving manufacturing techniques, including advancements in cloning, media formulation, removal of host cell components and downstream purifications (i.e. industrial scale chromatography). [68] The starting point of the production is a Master Cell Bank (MCB) and a Working Cell Bank (WCB) in which the construct of interest is inserted; these cells, are placed inside a bioreactor to produce the final protein by a fermentation process. While the production process takes place, both protein of interest and waste products are produced inside the bioreactor (all these elements are called Bulk). At the end of the production, the final protein is purified to remove any impurities and contaminants and the End of Production Cell Bank cells (EoPCB) are isolated for further stability tests (Figure 1).

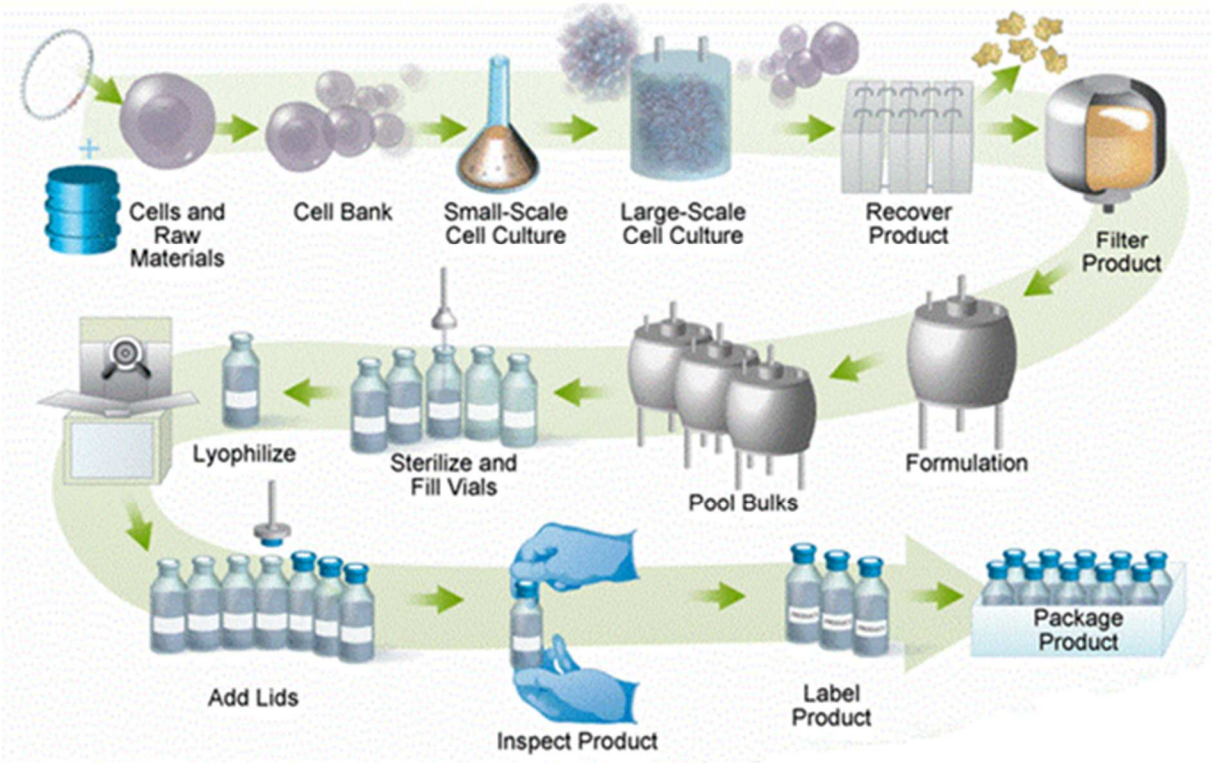

Figure 1. Biopharmaceutical process [69]

Unlike chemical synthesis processes, biopharmaceutical manufacturing needs high technology for quality assurance. For this reason, different guidelines have been created over the time for biological quality control

processes [2,3]. ICH Topic Q5A (R1) document *"is concerned with testing and evaluation of the viral safety of biotechnology products derived from characterized cell lines of human or animal origin"* [4] establishing a list of viral tests required for biotech drugs release. The risk of contamination is a frequent possibility in biotechnology products derived from cell lines; production environment (e.g. operators), raw materials or production cell substrate represent important sources of contamination. Moreover, to guarantee the integrity and stability of the transgene among the MCB and WCB cell lines, the ICH Topic Q5B provides a guidance for *"the characterization of the expression construct used for the production of recombinant DNA protein products"* [5]. Traditional *in vitro* and *in vivo* tests represent important testing tools; however, some limitations are still present, particularly in term of time for execution, limited host range and low sensitivity. To overcome these limitations, the introduction of new cutting-edge technologies is required [6]. The Next Generation Sequencing (NGS) has been introduced on the market as a powerful tool to achieve a higher throughput and a reduction of time and costs. Different NGS technologies have been settled since the first instruments were developed. The sequencing-by-synthesis (SBS) technology from Illumina is nowadays the most successful massive parallel short reads sequencing method. On the opposite side, single-molecule sequencing with zero-mode waveguide-based readout (Pacific Bioscience) is the most widely adopted long read technology. By side, the nanopore-based paradigm for single molecule real-time sequencing has been recently proposed by Oxford Nanopore (ONT) and the platform is gaining adoption over the time due to its competitive costs and ease of use.

TECHNICAL INTRODUCTION

Among the NGS platforms developed in the last decade, Illumina is the most widely adopted technology. The Illumina platforms are based on the proprietary technology “Sequencing by Synthesis” (SBS) that enables the detection of single nucleotides as they are incorporated into DNA strands [7]. Illumina NGS workflow is divided in three steps: library preparation, cluster generation and sequencing. The library consists of DNA fragments derived from whole genome, reverse transcribed RNA or amplicons in which index-adapter are ligated at the ends. By means of a complementary bound, the library is hybridized to oligonucleotides, covalently linked to an optically transparent solid surface, the flow cell. With such oligonucleotides acting as primers, the templates are extended by a polymerase, producing covalently attached single molecules. Free ends of the bound templates hybridize to neighboring complementary adapters to form U-shaped bridges, that is extended to create a double-stranded DNA bridge (thus called bridge amplification). At the end of the extension, the dsDNA bridge is denatured again to have two complementary single-stranded DNA fragments covalently attached to the adapters, ready to repeat the reaction. Cluster generation, by means of several bridge amplification cycles, generates thousands of clonal fragments per cluster. For the subsequent sequencing step, a series of fluorescently labeled reversible terminators are used and incorporated by DNA polymerase, for all the cluster fragment simultaneously. The fluorescent bases are imaged at each cycle after blocking dNTPs are added and then unblocked before the next addition cycle. Since all four reversible - terminator-bound dNTPs are present during each sequencing cycle, natural competition minimizes incorporation bias [8]. In case of paired end sequencing, both ends of the DNA fragment are sequenced. After the completion of the sequencing of the forward strands, the newly synthesized reverse strands are regenerated by bridge amplification. The forward strands are removed by cleavage leaving only the newly synthesized reverse strands attached to the flow cell to be sequenced as before to produce paired end sequence data (Figure 2).

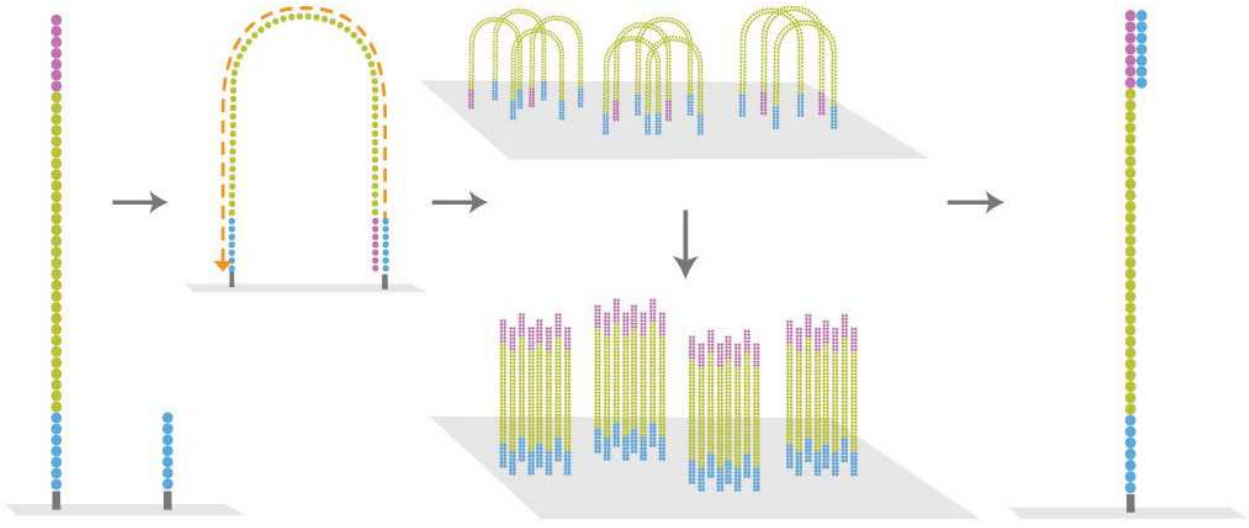

Figure 2: Cluster Generation on Illumina flow cell surface by Bridge Amplification

The result is true base-by-base sequencing that enables accurate data for a broad range of applications. The sequencing data produced are then submitted to the bioinformatics analysis. It consists in three main phases:

- **Primary analysis**: this phase is mandatory to transform the raw data generated by the Illumina instruments into base calls and afterwards reads.

- **Secondary analysis**: in the secondary analysis, the reads generated from primary analysis can be aligned to e reference genome or a genome database or used to perform a *de novo* assembly or clustering, depending on the biological expectation.

- **Tertiary analysis**: is based on the interpretation of the bioinformatics result.

Even if the short reads Illumina sequencing is a robust and reliable method applied in different fields of molecular biology, a new powerful technology based on very long reads sequencing is emerging. Oxford Nanopore Technologies Ltd has developed a disruptive technology platform for the direct electronic-based analysis of single molecules. The MinION is the world's first mobile sequencer [9]. It is a small device (10 cm x 2 cm x 3.3 cm; approximately 90 g) and is powered by a computer USB port (Figure 3). The instrument is based on using nanopores to detect a molecule. A nanopore is a biological nano-scale hole present inside a polymer membrane, the membrane is embedded in a physiological solution and across it a voltage is set. The resulting ionic current is detected by a sensor array chip that consists of several microscaffolds, each of which supports the membrane and the embedded pore and corresponds to its own electrode. When a molecule passes through the nanopore or near its aperture a disruption of the ionic current occurs. This change in the electrical signal is detected and used to identify the molecule in question. For DNA strand is possible to discriminate each single base due to their characteristic electrical signal [10] (Figure 4).

Figure 3: The MinION: the first palm sized sequencer

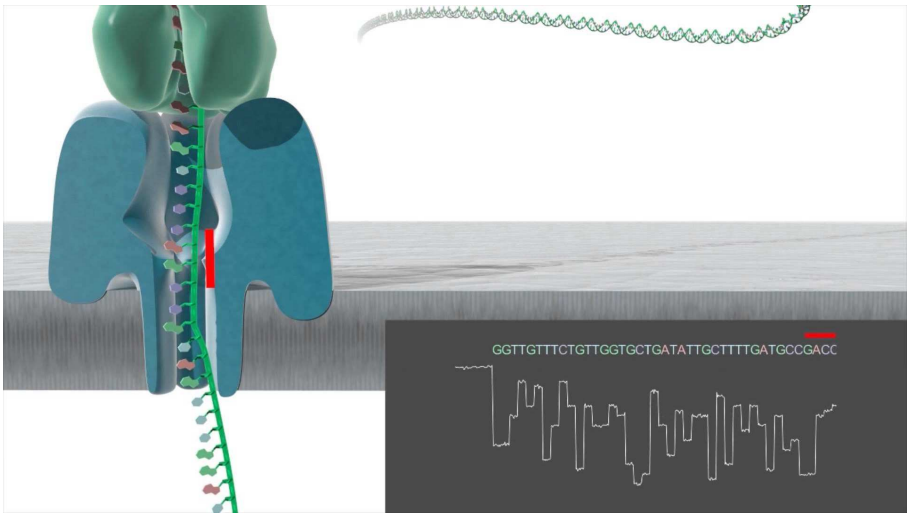

Figure 4: The Oxford Nanopore Technology

Compared to the other NGS technologies, the MinION can measure single molecules directly, without the need for nucleic acid amplification, fluorescent/chemical labelling or optical instrumentation. Moreover, the other important features of this technology are i) the low cost (1000$ for the MinION instrument), ii) the possibility to get very long sequences (up to 200kb) and iii) to perform the data analysis in real time. Meanwhile the sequencing is performed, the data analyses can be carried concurrently and the user can decide to stop the run when the biological question is answered reducing time for both sequencing and data analysis.

OBJECTIVE OF THE THESIS

The purpose of the PhD project was to develop methods based on NGS to apply in the biopharmaceutical quality control. Particularly, Illumina technology methods were developed and applied for i) the preliminary phase of the cell line process development, ii) to guarantee the quality during the upstream and iii) the in-process phase of the biopharmaceutical production.

For the preliminary phase study, a transcriptome analysis of CHO cell line, using the Illumina NextSeq500 platform, was performed to investigate the role of supplement media components in the post-translational modifications, focusing the attention on the glycosylation process. Regarding the upstream and the in-process phases two different NGS methods were developed respectively to check the clonality of MCB and to verify the absence of adventitious viral contamination during the biomanufacturing process. Moreover, the application of the Nanopore sequencing technology by mean of the MinION instrument for the viral qualification was evaluated.

Implementation of these technologies can provide a higher performance in the analysis of the molecules and the respective host cells used for the production with unprecedented potentials for biosafety and characterization of the final products with in-process monitoring (Figure 5).

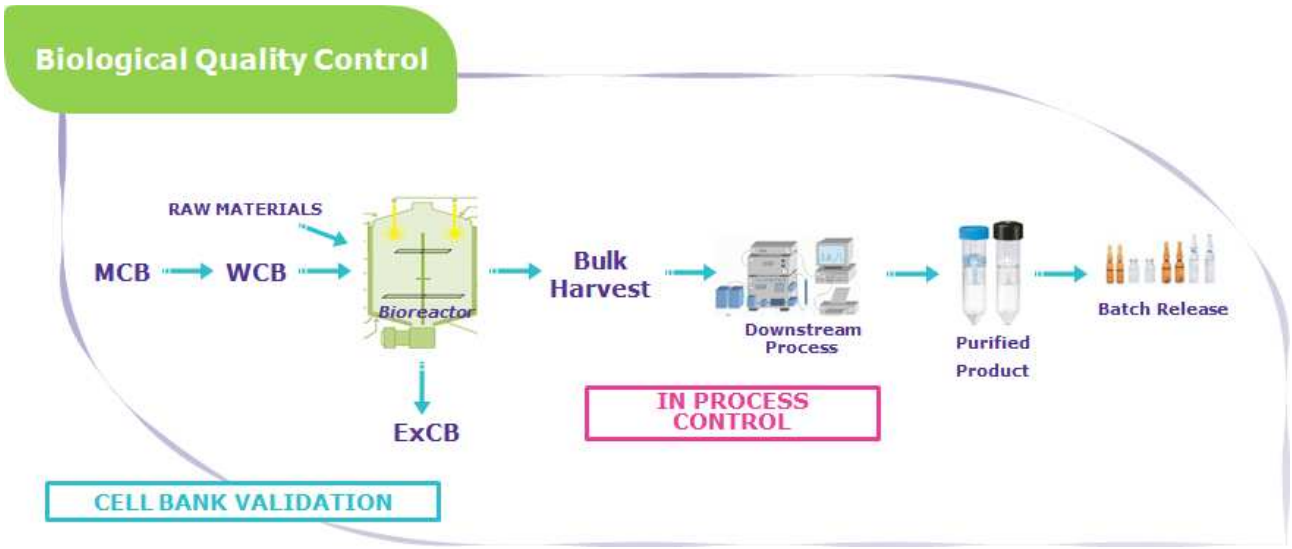

Figure 5. Biological Quality Control Activities

CHAPTER 1

TRANSCRIPTOME CELL LINE CARACTHERIZATION

INTRODUCTION

One of the most important objectives of the biopharmaceutical industry during the process development is the identification of factors that affect recombinant protein quality. Particularly, post-translational modifications play a key role in the protein solubility and stability, enzymatic activity, cellular processing, secretion, both clearance and half-life, as well as efficacy and biological activity. [11-14].

In the past decade, the main tool to understand biological behavior of producing cell lines and production processes was the analysis of the transcriptome using microarrays [15,16]. However, the accuracy of the technique depends on prior knowledge of the sequence and the affinity of probes for hybridization. The advance of the NGS, especially the RNA-Seq, allowed to overcome these constraints and expanded the knowledge in cell culture bioprocessing [17]. Compared to microarrays, RNA-Seq methods cover all aspects of the transcriptome without any a priori knowledge of it, allowing for the analysis of novel transcripts, splice junctions and noncoding RNA [18].

In this study, performed in collaboration with the Bio Process Science Department (BPS - Merck) in Vevey (Switzerland), differential expression of glycosylation genes in CHO K1 cell cultures used for the manufacturing process was analyzed. Particularly, the role of the Raffinose - a naturally occurring trisaccharide composed of galactose, glucose, and fructose - was evaluated to understand how this alternative media component impact the post-translational modifications (data published as "*Cell culture media supplemented with raffinose reproducibly enhances high mannose glycan formation*", [18]). Several studies described the effect of Raffinose on various metabolic pathways of different cell types; in humans for example Raffinose intake was correlated with leukotoxic effects and oxidative stress [62].

MATERIALS AND METHODS

SAMPLE PREPARATION

A media design approach was applied to modulate glycosylation in CHO cell culture [18]. Briefly, CHO K1 cell lines were treated in microscale cultures using the Raffinose as alternative supplement. Four different concentrations of Raffinose were tested: 10mM, 30mM, 50mM, 100mM. A condition with no treatment of the cells was used as control. Three replicates for each condition were analyzed.

RNA EXTRACTION

Total RNA extraction cellular pellet was performed with affinity columns using the RNeasy Mini Kit (QIAGEN) according to manufacturer's instructions and internal working procedure. Briefly, the method

consisted of a first step of lysis, followed by the binding of RNA to the column membrane and different washing steps. The cellular pellet was disrupted by means of a lysis buffer and homogenized by passing the lysate through a blunt 20-gauge needle (0.9 mm diameter) fitted to an RNase-free syringe. All volume was transferred to a spin column and centrifuged at 10,000 rpm for 30 seconds promoting selective binding of RNA to the membrane column. Contaminants were removed using two different buffers provided by the kit. An enzyme treatment was employed to remove any DNA residual prior to a final elution step using RNase-free water. RNA quantification was evaluated using Fluorometer Qubit® 2.0 and the RNA quality (RIN value > 8) was assessed by Agilent 2100 Bioanalyzer using 6000 RNA Nano kit.

LIBRARY PREPARATION

Libraries were prepared using TruSeq Stranded Total RNA Sample Preparation Kit according to Illumina's protocol [19] and internal working instructions starting from 1µg of extracted total RNA (Figure 6).

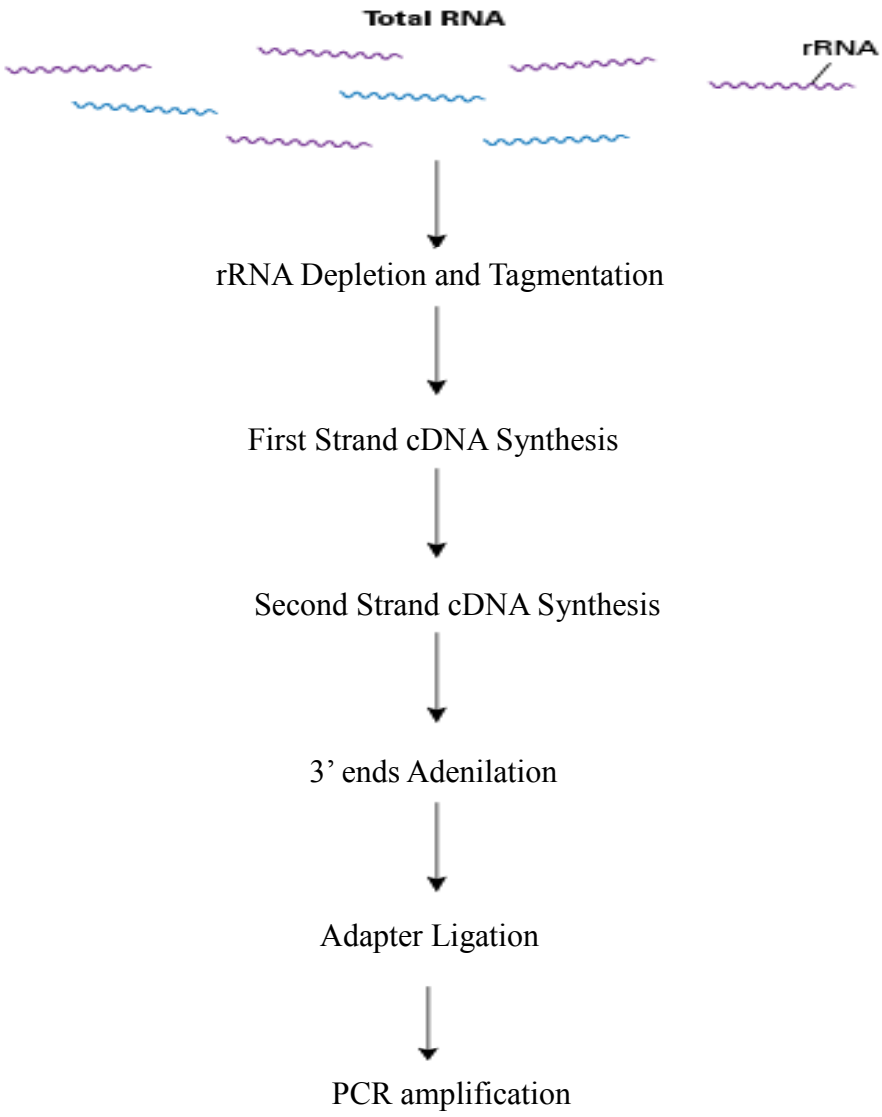

Figure 6: The TruSeq Stranded Total RNA Steps

Briefly, a ribosomal RNA depletion was carried out using biotinylated, target-specific oligos combined with Ribo-Zero rRNA removal beads. Following purification, RNA was fragmented and primed for cDNA synthesis. The cleaved RNA fragments were converted into first strand cDNA using SuperScript II reverse transcriptase and random primers, followed by second strand cDNA synthesis using DNA Polymerase I and RNase H. The Strand specificity was achieved by replacing dTTP with dUTP in the Second Strand Marking Mix (SMM). The incorporation of dUTP in second strand synthesis quenched the second strand during amplification, since the polymerase used in the assay will not incorporate past this nucleotide. AMPure XP beads are used to separate the ds cDNA from the second strand reaction mix. At the end of this process, you have blunt-ended cDNA. A single 'A' nucleotide is added to the 3' ends of the blunt fragments and multiple indexing adapters are ligated to the ends of the ds cDNA, preparing them for hybridization onto a flow cell. The adapter-ligated cDNA fragments were then amplified on thermal cycler using the following program:

- 98°C for 30 seconds
- 15 Cycles of PCR:
    - 98°C for 10 seconds
    - 60°C for 30 seconds
    - 72°C for 30 seconds
- 72°C for 5 minutes

Hold at 4°C.

The final library was purified using AMPure XP beads (Beckman Coulter), quantified by QUBIT® 2.0 Fluorometer (Invitrogen) and size distribution was assessed by Agilent 2100 Bioanalyzer (using High Sensitivity DNA kit).

SEQUENCING

Samples were sequenced on Illumina NextSeq500 instrument from both ends of DNA fragments to produce "paired end" reads using a High Output Kit (2x76 cycles). Equal volumes of libraries were mixed to create two multiplexed pools sequenced in two different runs (9 sample/run). The instrument performs both cluster generation and sequencing of samples enables reducing cycles and data processing time. The clusters were imaged using 2-channel sequencing chemistry and filter combinations specific to each of the fluorescently labeled chain terminators. The two-channel sequencing uses only two images: an image from a red channel and an image from a green channel. The intensity emitted by each base is as follows: T emits 100% green intensity, C emits 100% red intensity, A emits 50% green and 50% red intensities and G is dark and does not emit any intensity.

DATA ANALYSIS

Data obtained from the RNA Sequencing were used to perform a differential gene expression analysis between the control and the different experimental conditions.

In the first step, base call files obtained from raw data generated from NextSeq500 platform were converted to sequence reads (fastq files) using *bcl2fastq* (version 2.15.0.4) [50]. After, fastq files were filtered for the quality using a tool able to trim the reads by quality parameters [21,22]. Trimmomatic tool (version 0.27) performed a dynamic trimming of reads with low quality bases (known as quality trimming) at sequence ends coupled with adapter leftover removal and a read filtering to eliminate reads falling below a residual length threshold. After the first data processing, reads for each condition were mapped to the genome sequence leveraging transcriptome annotations of *CHOprj* [23] using *tophat* tool (version 2.0.13) [27] and the embedded bowtie2 alignment tool [51]. For tophat standard options were used. Gene level raw counts of mapped reads were performed by *HTSeq* (version 0.6.1p1) [24,25]. Data results were managed using *Bioconductor-DESeq* package (version 1.14.0) [26] to identify a differential expression between control sample and mix of biological and technical replicates of the treated samples.

RESULTS

By comparing control samples (no treatment) against the other conditions (10mM Raffinose, 30mM Raffinose, 50mM Raffinose, 100mM Raffinose) a total list of about 200 genes differentially expressed was found (data not shown).

Focusing only on the genes involved in the glycosylation we found a differential expression only for a subset of genes and especially for the 100mM Raffinose condition compared to the control. Despite a lenient statistical threshold (Bonferroni-adjusted p-value < 0.1) was used to extract significant genes, a coherent trend of expression was found among the different conditions (Figure 7), confirming the direction of expression feedback (average R^2 of 0.77 with 14 out of 20 genes with $R^2 > 0.70$). In the Table 1 is indicated a list of all genes differential expressed for each condition; the red ones indicate the down regulated, the green the up regulated genes (both categories are showed in a gradient color).

ID	Gene Name	Raffinose concentration			
		10mM	30mM	50mM	100mM
B3galt2	beta-1,3-galactosyltransferase 2	1.30	1.10	**1.36**	0.90
B3gat3	beta-1,3-glucuronyltransferase 3	-0.09	-0.13	-0.17	**-0.37**
B4galt3	beta-1,4-galactosyltransferase 3	-0.19	-0.29	-0.29	**-0.51**
Chpf	chondroitin polymerizing factor	-0.36	-0.34	-0.34	**-0.46**
Chst11	carbohydrate (chondroitin 4) sulfotransferase 11	0.36	0.52	0.55	**0.51**
Galk1	galactokinase 1	-0.31	-0.42	-0.43	**-0.51**
Gla	galactosidase alpha	-0.13	0.08	0.23	**0.53**
Gns	glucosamine (N-acetyl)-6-sulfatase	0.42	0.59	0.94	**0.94**
Hyal1	hyaluronoglucosaminidase 1	0.72	0.36	0.58	**0.64**
Mgat5	mannosyl (alpha-1,6-)-glycoprotein beta-1,6-N-acetyl-glucosaminyltransferase	0.20	0.30	0.33	**0.33**
Neu1	neuraminidase 1 (lysosomal sialidase)	-0.12	0.19	0.42	**0.88**
Ogt	O-linked N-acetylglucosamine (GlcNAc) transferase	**0.57**	0.46	**0.46**	0.31
Pigq	phosphatidylinositol glycan anchor biosynthesis class Q	0.02	-0.13	-0.12	**-0.36**
Rpn2	ribophorin II	-0.11	-0.22	-0.24	**-0.34**
Slc35a4	solute carrier family 35 member A4	-0.24	-0.39	-0.35	**-0.64**
Slc35d1	solute carrier family 35 member D1	0.25	0.43	0.44	**0.45**
Slc35f2	solute carrier family 35, member F2	-0.14	-0.31	-0.33	**-0.49**
Slc35f5	solute carrier family 35 member F5	0.11	0.16	0.27	**0.43**
St8sia6	ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase 6	0.51	0.90	**1.21**	**1.91**
Ugcg	UDP-glucose ceramide glucosyltransferase	0.35	0.49	0.46	**0.58**

Table 1. List of differentially expressed genes involved in the glycosylation (numbers are referred to the log2foldchange, highlighted values are the statistical significative value with Bonferroni adjusted pvalue < 0.1)

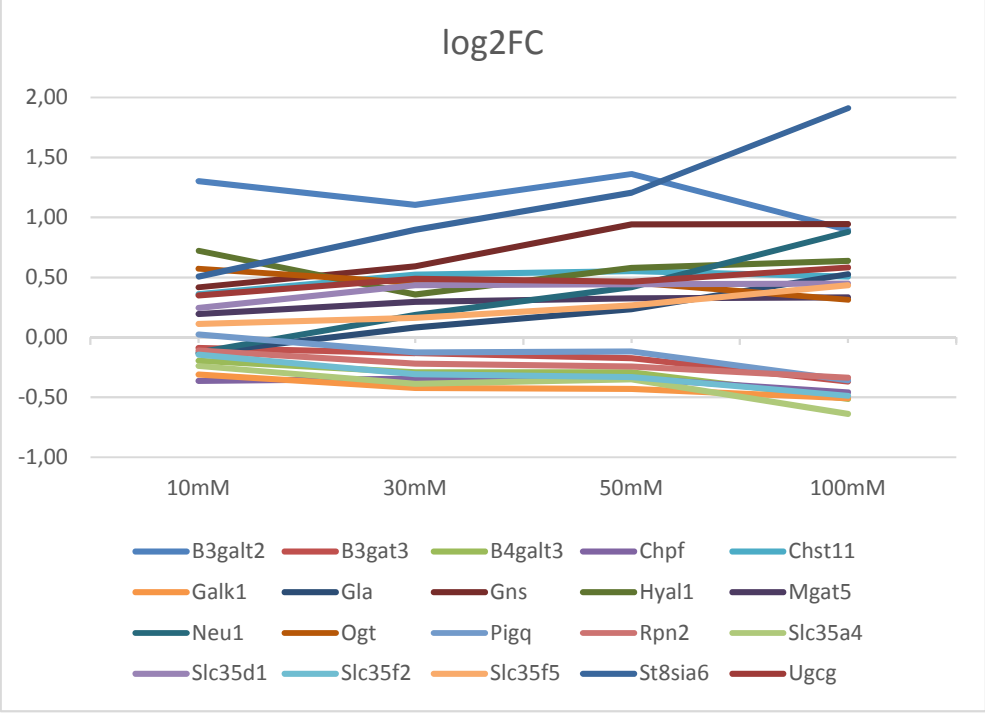

Figure 7. Expression of glycosylation genes among the different concentrations expressed as log2 fold-change

The expression of one of the most important enzymes, the mannosyl (alpha-1,6-)-glycoprotein beta-1,6-N-acetyl-glucosaminyltransferase (Mgat5), that plays a pivotal role in the regulation of the biosynthesis of glycoprotein oligosaccharides [60] at the higher concentration of 100mM was upregulated. While the beta-1,4-galactosyltransferase 3 (B4GalT3) expression was downregulated as well as the UDP-galactose transporter, solute carrier family 35, member A4 (Slc35a4) for all the Raffinose concentrations, the beta-1,3-galactosyltransferase 2 (B3galt2) had a fluctuating trend among the different concentrations with a particular lower regulation at the 100mM Raffinose concentration. This is surprising since Galactosylation increased at higher Raffinose concentration. Looking for the genes involved in the initial step of the glycosylation pathway no differential expression was found for Mannosidase I (ManI), Mannosidase II (ManII), Mgat1 and Mgat2 enzymes. The expression of the UDP-glucuronic acid and UDP-GalNAc transporter genes (gene: Slc35d1) [61] was upregulated. Sialyltransferase (SialT) gene expression was highly upregulated, whereas the levels of sialic acid remained low (data not shown).

CONCLUSIONS

The use of NGS technology for the differential expression identification in cell line development was investigated.

Differential expression glycosylation analysis was carried out on CHO K1 cell line by Next Generation Sequencing. Different treatment conditions (10mM Raffinose, 30mM Raffinose, 50mM Raffinose, 100mM Raffinose) were evaluated and compared to a control condition (no treatment).

Transcriptomics analysis showed that Raffinose supplementation influenced the expression levels of different glycosylation related genes particularly for the 100mM Raffinose condition, however other consensus patterns were observed at lower concentrations.

Particularly, the GalT gene was found downregulated, while the SialT gene was strongly upregulated. Based on these results we can hypothesize that on one hand the downregulation of the β-1,4-GalT 3 gene along with the upregulation of galactosidase α gene and, on the other hand, the steric hindrance effects in the CH2-domain of the mAb and the considerably higher SialT gene expression still resulted in a negligible effect on the entire sialylation process. Although the enzyme was potentially present in higher concentrations, its accessibility was either unfavorable, or other unknown parameters hampered the attachment of sialic acid to the galactose moiety of the oligosaccharide backbone.

Even if further investigations are required to identify the underlying mechanism at the gene level and resulting real protein level, considering the substrate transport into the Golgi apparatus and the GalT activity, these results highlight the potential of cell culture medium supplementation to alter glycosylation patterns of

recombinant proteins. Changing the environment in which the cells are cultured is a rather straightforward approach that allows to finetune within the potential of the selected cell line.

Unlike microarrays, the Next Generation Sequencing technology represented an optimal tool to identify differential expressions in a fast and time-efficient way.

CHAPTER 2

CLONALITY ASSESSMENT OF MCB USED FOR THE PRODUCTION

INTRODUCTION

During the past 20 years, CHO cells have been the most commonly used platform to express various forms of therapeutic proteins and antibodies [29]. In general, the process of cell lines development involves transfection of the host CHO cell line with a transgene that expresses the target protein followed by plating and selection of the transfected cells. The growing clones can then be ranked and successively expanded based on protein expression and, after further evaluations, the top clone can be chosen as production line.

The proof of clonality for these cells, represent an important request by regulatory agencies [30, 31]. Recently, different methods have been developed to further assure that cell lines are clonal; these methods include but are not limited to: a single-cell cloning step, two rounds of subcloning by limiting dilution, flow cytometry mediated single cell sorting and deposition, or microfluidics based cell printing [32]. However, due to technological limitations of each of these methods, they are often being used in combination with high throughput imaging of freshly plated microwell plates to provide proof of clonal derivation [33,34].

Even if these methods are well recognized by the Health Authorities, in the last few years the agencies are increasing the demand for a better characterization of the clonality. Particularly, the focus is now direct on the characterization of the insertion site(s) of the gene of interest [59]. In the project here described, a novel method based on Next Generation Sequencing to determine the clonality of a cell bank by multiple analysis of the insertion site(s) of the gene of interest is proposed.

MATERIALS AND METHODS

SAMPLE PREPARATION

Cell lines expressing the antibody of interest were created by an external company with a particular system called Gene Product Expression (GPEx). This system uses replication-defective retroviral vectors (Figure 8a), derived from Moloney murine leukemia virus (MLV) and pseudotyped with vesicular stomatitis virus G protein (VSV-G), to stably insert single copies of genes into dividing cells. Retrovectors deliver genes coded as RNA that, after entering the cell, are reverse transcribed to DNA and integrated stably into the genome of the host cell (Figure 8b). Two enzymes, reverse transcriptase and integrase, provided transiently in the vector particle, perform this function.

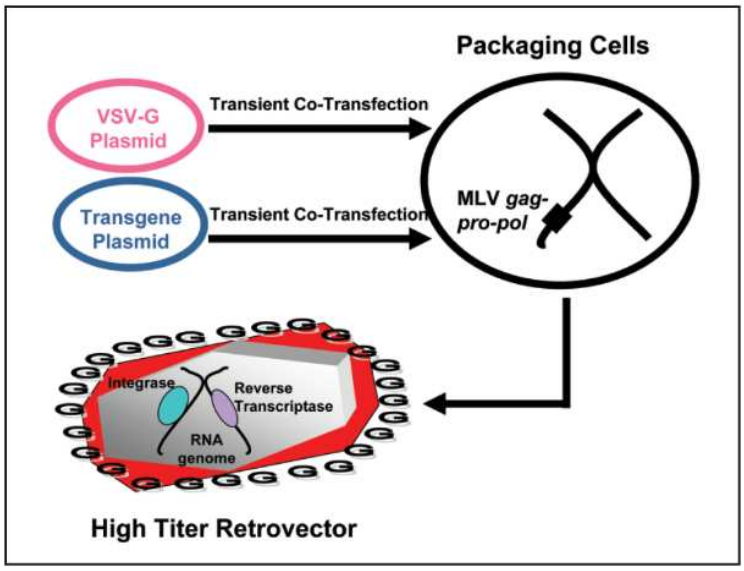

Figure 8a: Retrovector Production Process

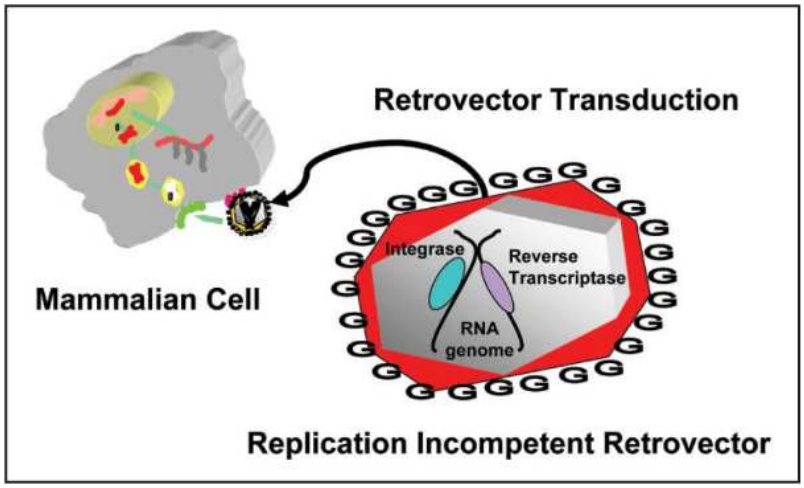

Figure 8b: GPEx Cell line engineering

For generating antibody-producing cell lines, an initial transduction of CHO cells was performed using a retrovector containing the light chain (LC) gene (transductions are performed at a multiplicity of infection of at least 1,000 retrovector particles per cell). The LC-expressing pool of cells was then transduced with a retrovector containing the heavy chain (HC) gene. The cell pool should be subject to multiple series of transductions for both LC and HC. The result of this transduction step is the multiple insertion in the CHO genome of LC and HC genes. Upon completion of the transductions, the resulting pool of cells are submitted to one round of limiting dilution subcloning. Single cell clones (based on the observation of the colonies) are isolated, and potential manufacturing cell lines are selected based on multiple criteria (growth, level of expression, product quality, etc.).

Due to the complexity of the cell bank and the fact that only a limited dilution step was performed, the NGS was applied as a tool to confirm the clonality of the Cell Bank generated. To perform the analysis, the MCB was subcloned to generate 30 subclones that were used for the analysis. In addition, the analysis of two divergent clones was performed to check the specificity of the method. The divergent clones are 2 clones of the same cell bank: this means that the insertion sites of the genes should be in different position of the genome, compared to the MCB and its subclones.

DNA EXTRACTION

Total DNA extraction of the samples was performed with affinity columns using the QIAamp Blood DNA Mini kit (QIAGEN®) according to manufacturer's instructions and internal working instructions. This method has a first step of lysis, followed by the binding of the DNA to the column membrane. Remaining contaminants and enzyme inhibitors were removed by washing steps before the DNA was eluted in water. To remove any RNA residual RNase enzyme incubation was employed.

Cell pellet was resuspended in 200 µl or 600 µl of Phosphate Buffered Saline (PBS) according to sample concentration and split in two aliquots. 200 µl of Buffer AL and 20 µl of proteinase K were added at each sample. Samples were mixed thoroughly by vortexing and incubated at 56°C for 10 minutes. Then 200 µl ethanol (96–100%) was added and the mixture was transferred into the DNeasy Mini spin column placed in a 2 ml collection tube. Samples were washed with two different washing buffers provided by the kit. Then, samples were centrifuged for a minute at 13,000 rpm to remove any residual ethanol. Elution was performed in 150-300 µl of water which was added directly onto the DNeasy membrane. Eluates of the same clone were combined. Each sample was incubated with RNase (Roche) at 37°C for 30 minutes to remove any RNA residual. At the end of the incubation period, samples were quantified by Fluorometer Qubit® 2.0.

LIBRARY PREPARATION

Library preparation was carried out by Nextera Mate Pair (Illumina) according to manufacturer's instructions and internal working instructions [35] (Figure 9).

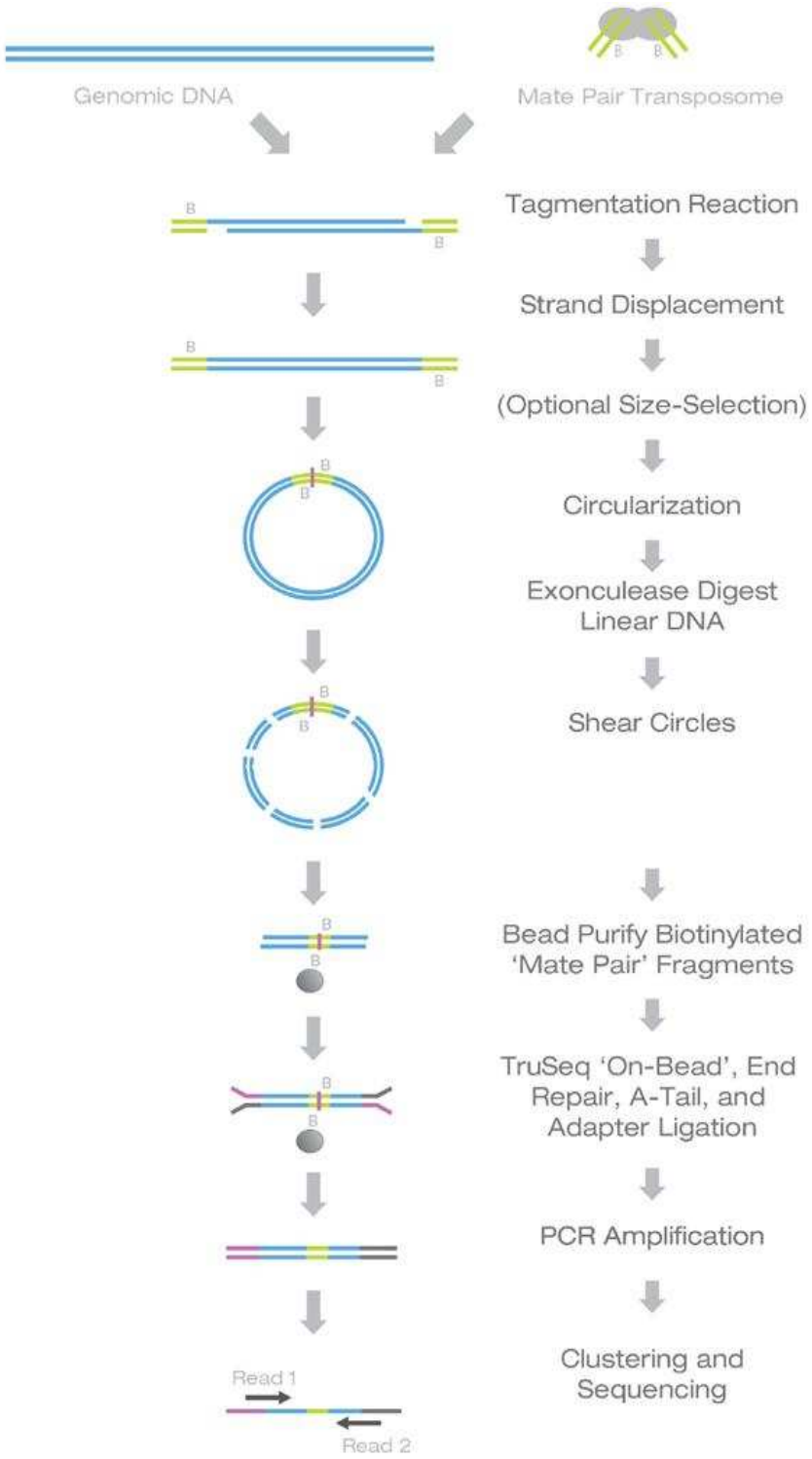

Figure 9: Nextera Mate Pair Protocol (Illumina)

Briefly 4µg of genomic purified DNA sample is simultaneously fragmented and tagged with a biotinylated mate pair junction adaptor using a specially formulated transposome. The tagmentation step left a short single stranded sequence gap in the tagmented DNA; so a polymerase extension with strand displacement of one adapter strand was carried out in order to fill gap. Next, DNA fragments of ~10,000bp was selected from agarose gel. The size-selected fragments were circularized in a blunt ended intramolecular ligation, with an overnight incubation at 30°C and any linear molecules remaining in the circularization reaction were removed by DNA exonuclease treatment. The exonuclease and ligase enzymes were both inactivated by heat treatment and the addition of 12µl of Stop Ligation Buffer. The large circularized DNA fragments were sheared by Covaris S220 instrument to smaller sized dsDNA fragments with 3' or 5' overhangs. Prior to convert the overhangs resulting from fragmentation into blunt ends, sheared DNA fragments containing the biotinylated junction adapters were purified by binding to streptavidin magnetic beads. Unbiotinylated molecules were removed through a series of washes and the fragments still bound to the streptavidin beads were end-repaired and A-tailed. Then, the Illumina oligo adapters were attached by ligation. PCR was carried out to enrich and amplify those DNA fragments that had adapter molecules on both ends. PCR was performed on thermal cycler using the following program:

- 98°C for 30 seconds
- 15 Cycles of PCR:
    - 98°C for 10 seconds
    - 60°C for 30 seconds
    - 72°C for 30 seconds
- 72°C for 5 minutes
- Hold at 4°C.

An AMPure bead purification step was used to clean up the PCR reaction and remove the smallest fragments from the final library. Library control quality was checked by Agilent 2100 Bioanalyzer instrument to calculate library medium size while quantification was verified by Fluorometer Qubit® 2.0.

BIOINFORMATIC ANALYSIS

ANALYSIS CONCEPT

The strategy applied to investigate the clonality of the MCB is based on a comparison of the insertion sites of the transgenes in the genomic DNA of the MCB, the 30 subclones derived from the MCB and the divergent clones plus a statistical analysis of the probability of the results occurring by chance. Several different bioinformatics tools were applied to determine the boundaries of the transgene inserts in the MCB, the subclones and the divergent clones. Determination of the boundaries allows the identification of the univocal position of the transgenes inside the CHO (hamster cell line) genome.

Once determined, the boundaries were analyzed by mean of statistical tools. The bioinformatics concept as illustrated in Fig. 10 was used to determine the boundaries.

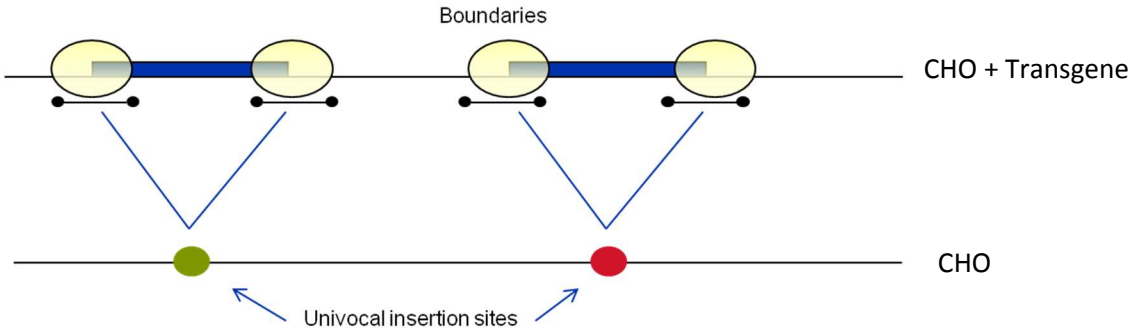

Figure 10. Concept of the bioinformatics analysis

The paired end analysis was used considering separately Read 1 (R1) and Read 2 (R2).

- Mapping of R1 and R2 separately to the transgene plasmid sequence (Figure 11, step A)
- Once mapped four kind of files were obtained:
    - Read 1 mapped on transgene plasmid sequence;
    - Read 1 unmapped on transgene plasmid sequence;
    - Read 2 mapped on transgene plasmid sequence;
    - Read 2 unmapped on transgene plasmid sequence;
- The list of MAPPED reads for Read 1 and Read 2 was created and the corresponding reads from the UNMAPPED reads was taken. See
-
- 11, step B/C.
- The UNMAPPED reads that had the corresponding paired read mapped on the plasmid were then mapped on the CHO-reference genome [36],

- 11, step D
- The alignment to the CHO reference provided estimates of the boundaries of the plasmid insertion site within the CHO reference genome
- The insertion regions were identified based on position on the different scaffold of the CHO reference genome by means of Geneious software V. 8.0

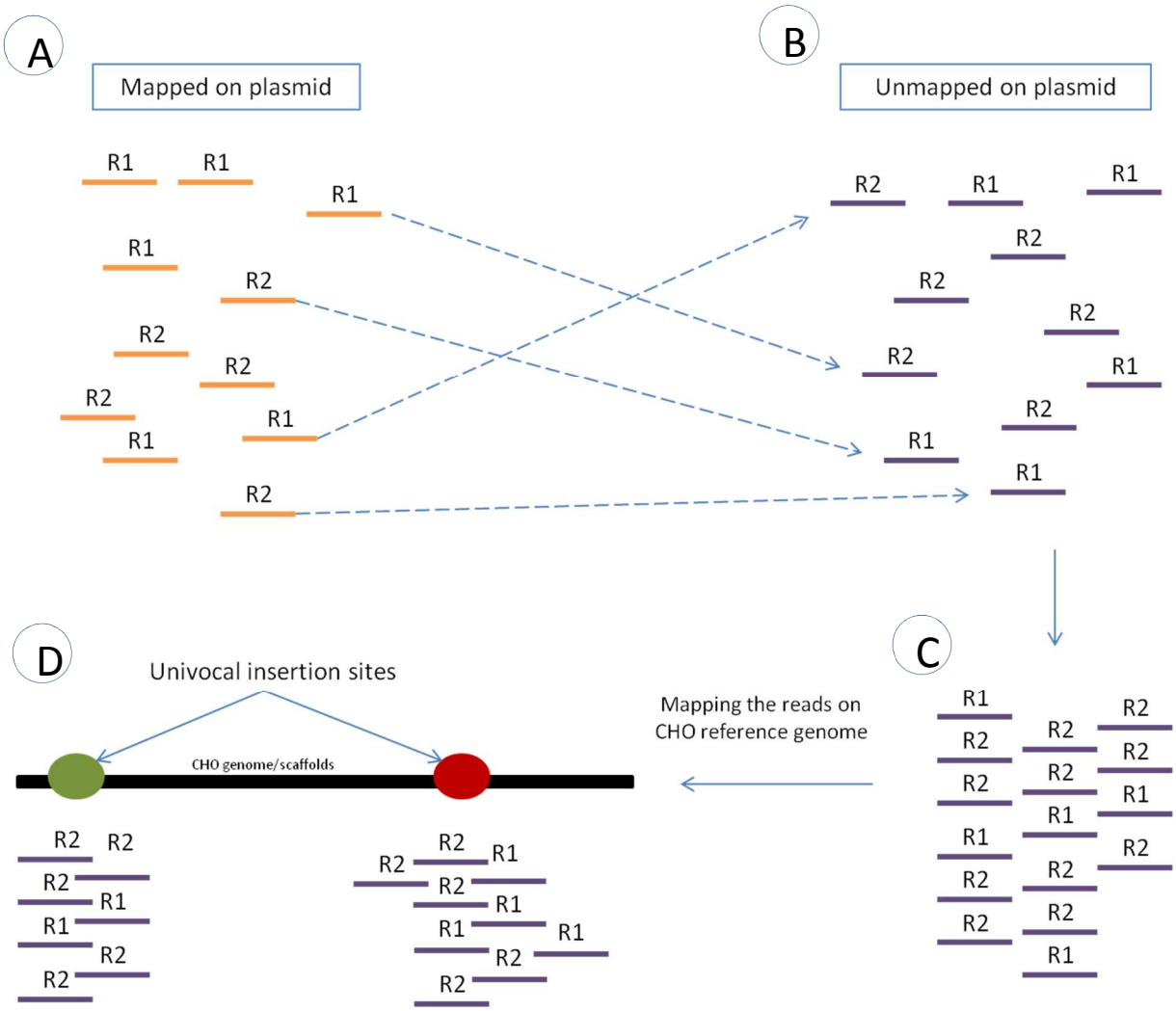

Figure 11. Bioinformatic analysis workflow

BIOINFORMATIC TOOLS

The following bioinformatic tools were used:

- For the conversion of the raw data from the sequencer (bcl files) to fastq files:
    - **Illumina CASAVA V. 1.8.2**

- For the indexing of the plasmid and the CHO reference genome:
    - **Burrows-Wheeler Aligner (BWA) V. 0.6.1-r104** [37]
        - bwa -a is database.fna

- For the alignment of the reads (in single end):

- **Burrows-Wheeler Aligner (BWA) V. 0.6.1-r104**
  - bwa aln -t database.fna reads.fastq > reads.sai
  - bwa samse database.fna reads.sai > reads.sam

- **samtools V. 0.1.18** [38]
  - samtools view –buS reads.sam > reads.bam

- **bamtools V. 2.1** [39]
  - bamtools filter –isMapped true/false -in reads.bam –out readsaligned.bam

- For the data interpretation and analysis
  - **Geneious V. 8.0**

Once the mapped reads were obtained, the coordinates on the CHO reference genome (boundaries of the plasmid) were taken in consideration.

SELECTION OF THE INSERTION REGIONS

After the selection of the boundaries for all the 30 subclones the following approach was used to identify the most representative insertion regions:

- Analysis of 1 clone that was sequenced with a deeper coverage (~50x)
- Identification of insertion regions (~20kb) by:
  - the **number of reads** (depth of coverage): the insertion regions with the highest number of reads were selected.
  - the **dispersion** (percentage of coverage of a particular region of the genome) of the reads among the insertion region: insertion regions with the highest dispersion percentage were selected.

As result of this analysis, on a total of 120 insertion sites identified for both HC and LC, 20 insertion regions were identified for LC and 20 for HC and used to create a pattern of presence/absence among all the samples analyzed. These 20 insertion regions represent the most represented and robust insertion sites.

STATISTICAL ANALYSIS

All 30 samples, the MCB and the divergent clones were analyzed to find the 20 insertion regions for LC and 20 insertion regions for HC. The pattern created for each sample (presence/absence of each insertion region) was then compared to the other samples. Like ‘bands on gel’ (Figure 12) molecular

analyses (i.e. PFGE or RFLP), each sample was identified by a binary code pattern, representing the list of present (1, marked in black) or absent (0, marked in red) insertion region, compared to the all 20 insertion regions previously identified.

Similarity (expressed as 1 – the distances among samples) was calculated using cluster analysis approach. In more details, distances were evaluated using Dice coefficient which consider:

$$Dd(A,B) = \frac{2|A \cap B|}{|A| + |B|}$$

In this case, if it is considered:

M11 as the number of ISs present both in sample A and B

M10 as the total number of ISs present in sample A

M01 as the total number of ISs present in sample B

Dice distance is defined as:

$$Dd(A,B) = 1 - \frac{M_{11}}{M_{01} + M_{10}}$$

To graphically show the distance among all the samples, Multi - Dimensional Scaling (MDS) approach was used.

An MDS is a method that represents measurements of similarity (or distance) among pairs or objects as distances between points in a low-dimensional multidimensional space.

Choosing to adapt distances to a 2-dimensions space, the objects are included into a two-dimensional scatterplot. In this case, axes are called “Dimension 1” and “Dimension 2”, and they have no meaning but they reflect the adaptation of differences between samples to a two-dimensional space.

RESULTS

CREATION OF THE PATTERN "BANDS ON A GEL"

Once the insertion regions were identified as described a pattern of presence/absence of insertion region for each sample analyzed was carried out (Figure 12) and used to compare all the subclones, the MCB and the divergent clones to each other.

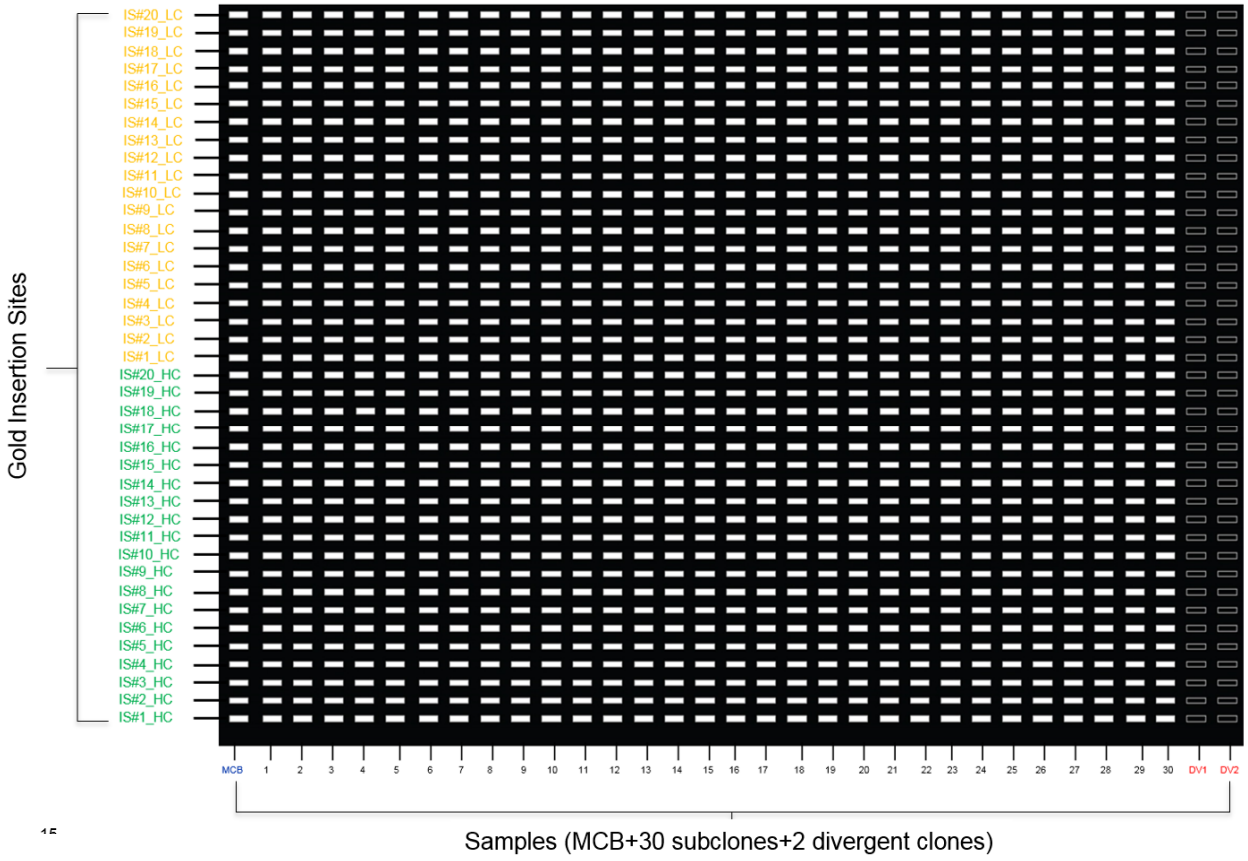

Figure 12. 40 insertion regions shared between samples. White bands represent the common scaffolds. Black bands represent absent scaffolds.

DISTANCE MATRIX

After the creation of the pattern for all the samples, a distance matrix by Multi-Dimensional Scale plot (Figure 13) was carried out to calculate the distance between the samples and the divergent clones. Three clearly different subgroups were identified. One was related to the 30 subclones and the MCB, the second one was related to one divergent clone and the last one the second divergent clone.

The distance between subclones+MCB and divergent subclones was 100%. The same results are obtained if the two divergent clones were compared to each other.

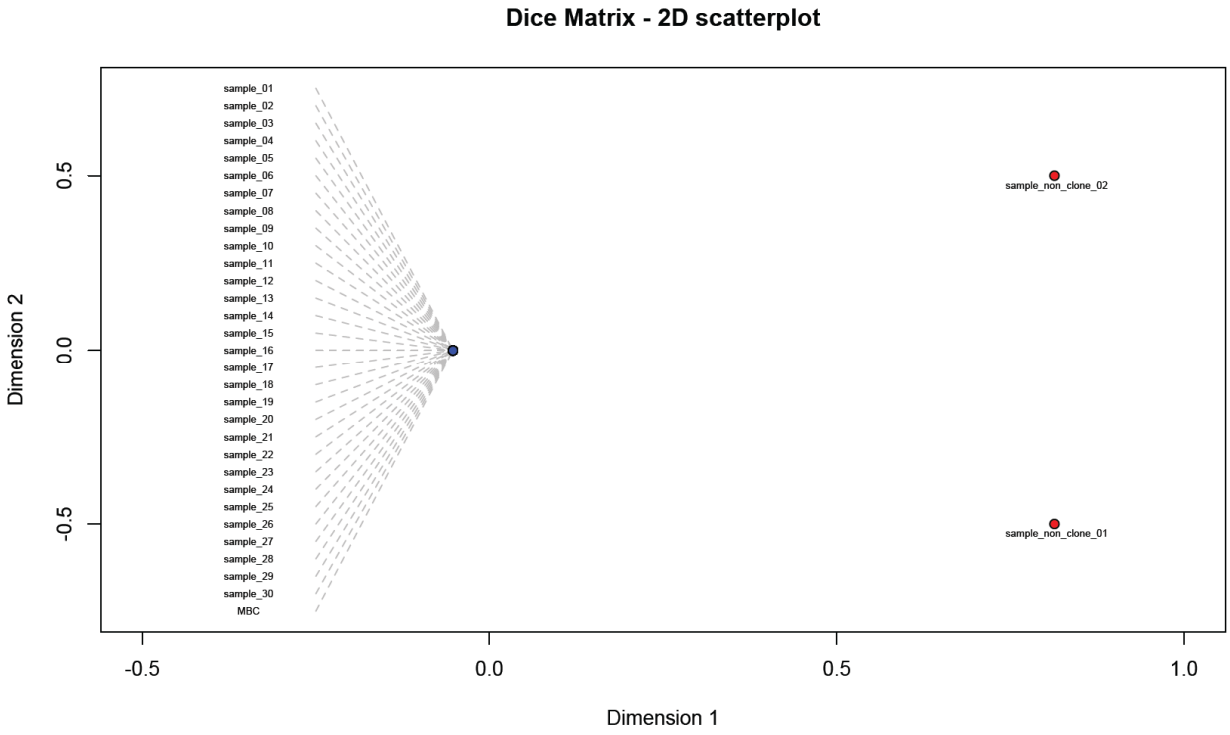

Figure 13. Multi Dimentional Scale plot.

PROBABILITY ANALYSIS

To evaluate the reliability of the bands on a gel pattern probability analysis was carried out. The probability that two subclones coming from different population share by chance 40 insertion sites was calculated by hypergeometrical function:

$$P(k) = \frac{\binom{h}{k}\binom{n-h}{r-k}}{\binom{n}{r}}$$

If considered that the retrovector has only 120 hotspots (total of the insertion sites identified for both HC and LC) available for the insertion of the transgene, the probability that 2 subclones from different population share the same 40 insertion sites is: p= $8.7*10^{-33}$ (phyper(39,40,80,40,lower.tail=F)). Alternatively, assuming only 1% of the genome to be receptive of transgene integration (245Mbp), that the analysis is carried on 20kbp windows and taking in to account possible border effects, the maximum

number of sites would be 6,125. This results in a probability of overlap-by-chance of 3.482152e-72 given the hypergeometric distribution (phyper(39,40,6085,120,lower.tail=F)).

CONCLUSIONS

The whole genome sequencing of 30 subclones derived from a MCB with inside the transgene of interest allowed the identification of the plasmid Insertion regions.

By mean of a combined bionformatic and statistical methods 40 well characterized insertion regions (20 for HC and 20 for LC) were identified and used to check the similarity of the 30 subclones and the MCB.

The coordinates of the insertion regions were used to analyze all the subclones, the MCB and the divergent clones. The results of this comparison showed that all the subclones have the plasmids inserted in the same position of the genome. A statistical tool for the similarity has been applied and showed that the MCB and the subclones have a similarity of 100%, and a distance to the divergent clones of 100%, allowing the conclusion that all 30 subclones derive from the same clone as present in the MCB. All subclones analyzed and the MCB showed a similarity of 100% and a distance to the divergent clones of 100%, concluding that all the 30 subclones belong to the same MCB.

Finally, the probability that this result was obtained by chance was an unlikely event ($p= 8.7*10^{-33}$).

This approach represents a well-establish method based on NGS to identify the insertion regions of the plasmid transgene inside the production cells and to assess the clonality on the cells during the entire production process (**Patent n°: EP96310 – Method for determining cell clonality**). Next steps will be to use the Illumina NovaSeq 6000 Sequencing System that combine a scalable throughput and flexible sequencing technology into a production-scale platform with the efficiency and cost-effectiveness of a benchtop system allowing to reduce drastically the run time of the analysis from weeks to hours.

CHAPTER 3

BIOSAFETY: A UNIVERSAL PROTOCOL FOR VIRAL CONTAMINANTS DETECTION IN BULK HARVEST

INTRODUCTION

Production of drug substance using animal derived cell carries the risk of adventitious virus contamination of the final product. Thus, the test for adventitious viruses is an essential quality control step in the manufacture of biological drugs. At present, the worldwide recognized techniques for the detection of adventitious agents are based on in vivo and in vitro tests [40]. However, these methods are limited by the restricted tropism of some viruses and may not detect some classes of viruses (non-cytopathic, non-pathogenic and non-haemadsorbing viruses) [41]. A complementary test to the cell-based approach is the use of biomolecular methods for the detection of nucleic acid of viral genomes like PCR-based tests (traditional PCR, Real Time PCR, etc.) that offer sensitive and specific detection of their target pathogens. Conversely, the backside of the use of the PCR-based methods are i. the unworkability to screen large number of viruses and ii. the need to know PCR target for the detection, meaning that what could be present inside a sample should be known a priori [42] and iii. a possible mutation occurring in the primer/probe regions may hamper the sensitivity of the assay.

Next Generation Sequencing can overcome these issues and combine the benefit of the in vitro/in vivo test such as the possibility to detect a broad range of viruses and the capability of the biomolecular methods to detect some class of viruses that did now show any effects on cells with the additional benefit to identify them without the need to create specific probes/primers. In recent years, some NGS-based methods were successfully applied in the detection of adventitious agents. One of the first case presented to the scientific community was the identification of a contamination by porcine circovirus (PCV) in two rotavirus vaccines [43,44]. This contamination was not identified by classical routine test but only with the use of the NGS approach. Recently the use of the Next Generation Sequencing was successfully applied on cell lines [45], vaccines [46] and, as in this study, on bioreactors [47].

The Next Generation Sequencing method here described is focused on a universal protocol for the detection of nucleic acids of different class of viral contaminants (dsDNA, ssDNA, dsRNA, ssRNA) potentially present in bulk harvest samples (BH) in one single analysis. Illumina platforms are used to assure a dramatic throughput/sample. Finally, the sequencing data are submitted to the bioinformatics analysis to get the final results. The workflow of this method consists of three stages:

- Sample preparation: nucleic acids extraction and reverse transcription reaction

- Library preparation and sequencing: HiSeq1000 and NextSeq500
- Data analysis: alignment of data to a curated internal viral database
-

MATERIALS AND METHODS

SAMPLE PREPARATION

During the setup of the method 42 blank samples of 4 different BH matrix were used to have a background knowledge of the endogenous retroviral components for each sample. To assess sensitivity of the method different classes of viruses (ds/ss DNA/RNA and retroviruses) were used to spike the blank samples using three different concentrations: 20 TU/mL, 10 TU/mL and 2 TU/mL.

Briefly the method consists of three steps: nucleic acid extraction and dscDNA synthesis, library preparation and sequencing and the bioinformatic analysis.

NUCLEIC ACIDS EXTRACTION AND dscDNA SYNTHESIS

NA extraction from bulk harvest was carried out with a fast procedure based on the use of a silica membrane with selective binding properties (QIAGEN) to obtain a DNA/RNA filtration and cellular components removal. The nucleic acids extraction was performed starting from 5 mL of sample and the method included 4 main steps: lysis, bind, wash and elution. Briefly, in the first step, sample was lysed under highly denaturing conditions at elevated temperatures in the presence of proteinase K and a specific buffer, which together ensure inactivation of DNases and RNases and complete release of nucleic acids. Lysates were then transferred onto a QIAamp Mini column. By a vacuum manifold, circulating nucleic acids were adsorbed from a large volume onto the small silica membrane as the lysate is drawing through by vacuum pressure. Remaining contaminants are removed during three washing steps and the nucleic acids were eluted in water. The sample was finally quantified by RNA broad range kit Qubit® 2.0 Fluorometer and employed as template for the next steps.

Pure nucleic acids were then retrotranscribed and a double strand synthesis was performed to have the complementary dscDNA. Starting from 4 µg of RNA, a retrotranscription was carried out using an internal protocol consisting of a preliminary incubation with random primer and a first strand synthesis using the ThermoFisher Scientific SuperScript® II Reverse Transcriptase. A double strand cDNA synthesis was needed to perform the library preparation. The dscDNA was quantified by Qubit® 2.0 Fluorometer and then used for the library preparation.

LIBRARY PREPARATION AND SEQUENCING

For the libraries construction, a paired-end (PE) library preparation was used to sequence both ends of the dsDNA fragments. Starting from 50 ng of dscDNA, library preparation was carried out by means of Illumina Nextera DNA Library Protocol [48]. The protocol consists of a first step of DNA fragmentation, an Indexed adapter ligation and a final step of PCR amplification. In the first step DNA was tagmented by trasposones that simultaneously added adapter sequences at the ends of fragments; after a bead-based purification, the tagmented DNA was amplified by a limited-cycle PCR program that allowed to add index (i7 and i5) to barcode samples for sequencing. At the end of the protocol a step of purification using the Beckman Coulter AMPure XP Beads was employed to select the final library. Library size is evaluated using Agilent 2100 Bioanalyzer and quantified by Qubit® 2.0 Fluorometer.

The Illumina HiSeq1000 Sequencer instrument and the TruSeq SBS Kit v3 – HS were used for the sequencing with a maximum read length of 2x100 and an Output of 300Gb. Libraries were loaded independently, on the eight lanes of the flow cell and cluster generation was performed on cBOT Illumina instrument [49]. The run time for this instrument is ten days including the sequencing of Read 1, Read 2 and the paired-end turnaround.

The four-channel sequencing works by using four images to determine which base occurs in each cluster. Each of the four DNA bases emits an intensity of a unique wavelength. Therefore, during a cycle, each cluster appears in only one of the four images. For example, when a strong intensity signal is detected in the wavelength related to the G base, a G is called. When a strong intensity signal is detected in the wavelength related to the T base, a T is called, and so forth. Four-Channel Sequencing requires all four images to build up the DNA sequence. Following denaturation of clusters, four fluorescently labeled reversible terminators are incorporated by DNA polymerase. The fluorescent base is imaged as each dNTP is added and then cleaved to allow incorporation of the next base. This process is repeated for multiple cycles of sequencing until the desired read length is reached [7](Figure 14).

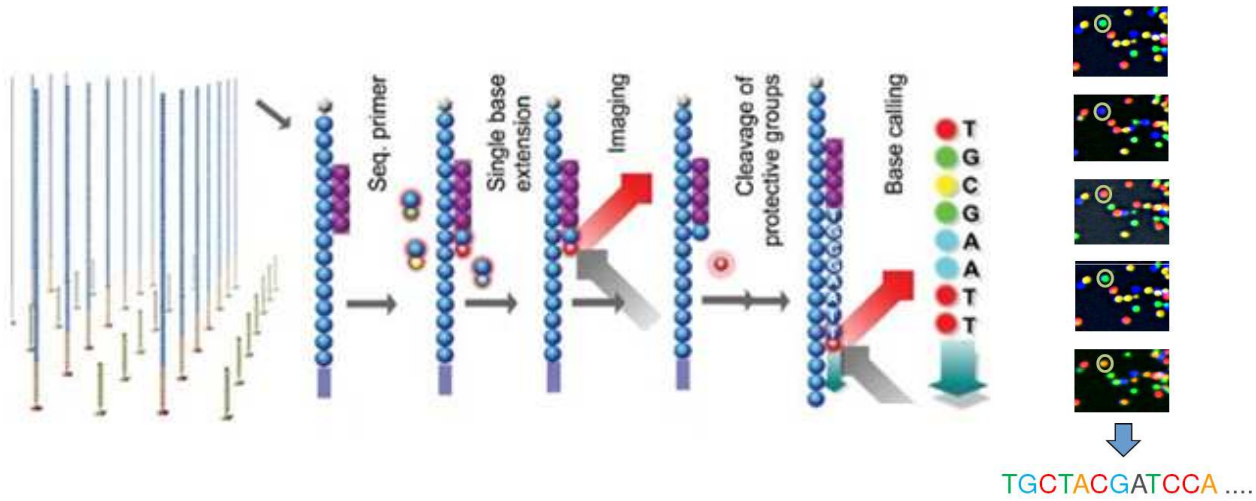

Figure 14: Illumina Sequencing by Synthesis Technology

To have a minor run time and lower cost per sample the method was also tested using the NextSeq500 instrument, in which the read length is increased from 100bp to 150bp and the run time decreased to 40 hours. To have the same quantity of data three samples were loaded on a single high-throughput flow cell and sequenced by means of the NextSeq500 High Output Kit. For each run, quality parameters were evaluated, focusing particularly on %Q30, cluster density and cluster density passing filter.

DATA ANALYSIS

One of the critical aspects of the NGS workflow is related to the data analysis of the runs. Due to the quantity of the data generated (more than 40billions of base pairs per sample) it was important to have a well-established bioinformatic pipeline able to analyze the sequences produced by the machine. The software **bcl2fastq** (ver. 2.15.0.4) [50] was used for the conversion of .bcl file to.fastq file. The alignment of reads to the curated viral database was performed through the **bowtie2** software [51] (Figure 15).

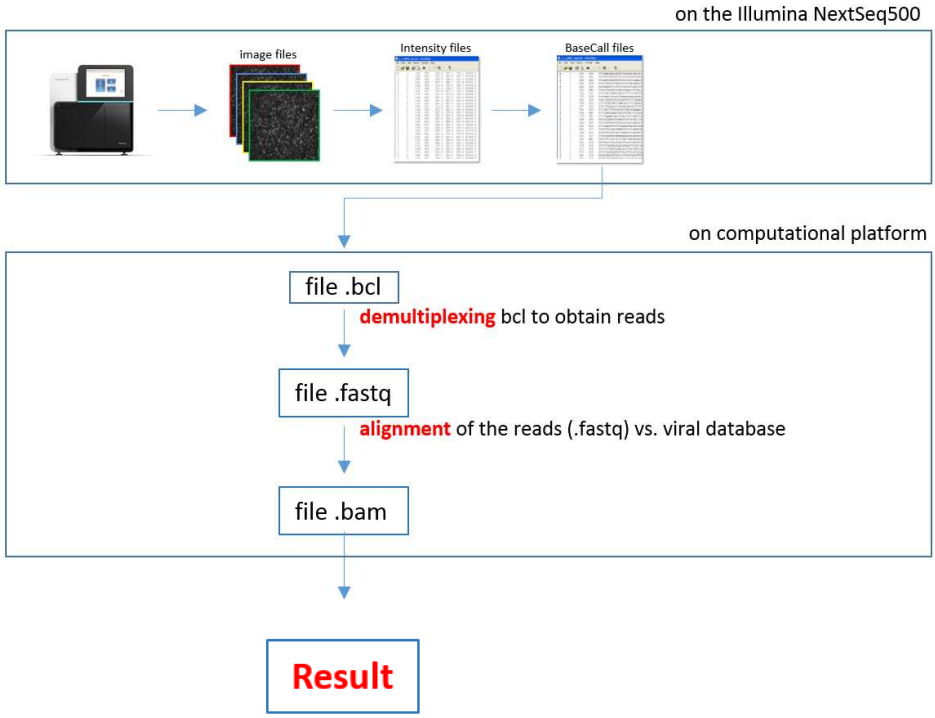

Figure 15. Schematic representation of the data analysis workflow

VIRAL DATABASE

The viral database created for the alignment of the data generated by NGS sequence is a comprehensive collection of sequences of different type of viruses.

The first version of the viral database (ViraldbI) contained 1500 viral sequences obtained from a public database (NCBI). After this first draft a risk assessment on the database was performed based on i) the availability of sequences from all the desired viruses and i) the quality (complete genome or partial sequences) and criticality (viruses with high interest for the biotech production vs rare viruses). In consequence, to fill the identified gaps the live viruses were bought, sequenced and their genome sequences inserted in the new viral db (ViraldbII). Finally, the ViraldbII contains the DNA genome sequence of:

- All viral DNA reference sequences that have as hosts vertebrates including human (font: NCBI);
- Virus present in MAP/HAP panel;
- Virus present in the 9 CFR Bovine and 9 CFR Porcine panels;
- Viruses responsible of large scale contamination in other pharma companies;
- New emerging viruses and blood borne pathogens;
- Internally sequenced virus used as internal positive controls for in-house viral safety tests;

- Internally sequenced virus whom the reference sequence was not present or present with poor quality

The viral database contains at the present 2,325 viral genome sequences of all known viruses able to infect vertebrate' cell lines (Figure 16) including new emerging viruses and adventitious agents responsible of large scale contaminations in other pharmaceutical companies.

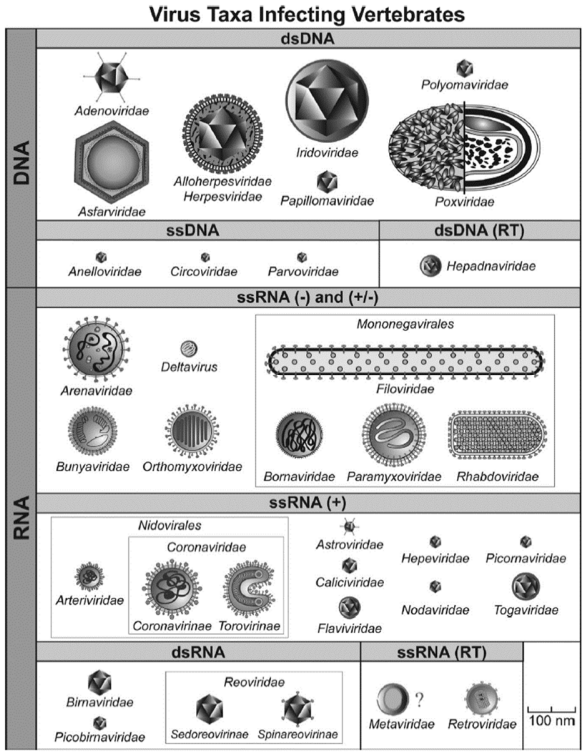

Figure 16: List of the virus Taxa able to infect vertebrates [52]

The database is routinely updated to maintain a comprehensive catalog of viruses.

The workflow of the creation and the updating of the viral database is described in Figure 17.

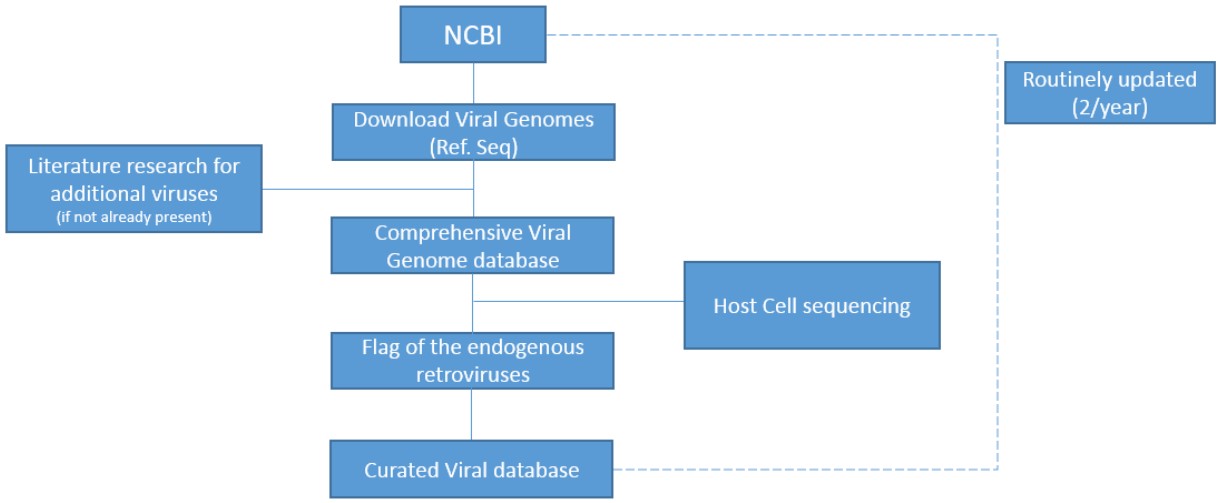

Figure 17: Workflow for creation and the update of the curated viral database for the bioinformatics analysis.

RESULTS

42 blank Bulk samples were analyzed to get the background composition for each matrix used (data not shown) ; at the same time, to assess the Limit of detection (LOD) of the method different viral concentration (20TU/mL, 10TU/mL and 2TU/mL) were used and analyzed in different bulk harvest. For the evaluation of the hits results from the analysis (Figure 18), three parameters were considered:

- reads number for each sample ≥20
- coverage ≥20%
- reads dispersion across the viral genome.

Furthermore, the database was curated and validated for the detection and the discrimination between potential contaminants and portion of viruses naturally present inside the host cells. These endogenous viral sequences are present in the database but are flagged. This flag let the analyst the possibility to discriminate between a real contamination and the presence of endogenous retroviruses. In the Figure 18 is represented the percentage of hits for the virus classes (ss/ds DNA – RNA) identified at the different viral concentrations used for the spiking.

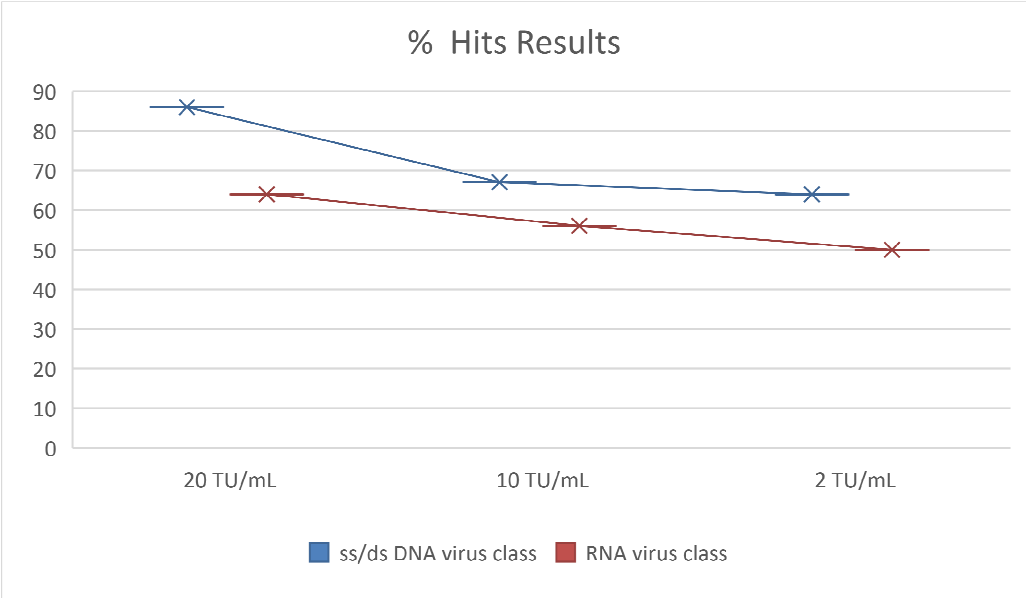

Figure 18. Percentage of positive hits for the different classes of viruses analyzed, considering the cutoff criteria of n° reads ≥20 and Coverage ≥20%

As shown in the figure 18 the method can detect both virus classes for all the spiking concentrations, however the ss/ds DNA viruses class is more detectable than RNA viruses class. Indeed the 86% of ss/ds DNA viruses were detected at 20TU/mL, the 67% of ss/ds DNA viruses were detected at 10TU/mL and the 64% of ss/ds DNA viruses were detected at 2TU/mL. On the contrary, for the RNA viruses class, the percentage for the three concentrations are respectively: 64%, 56% and 50%. This is also confirmed by the coverage obtained for each class among the three concentrations (Figure 19): the highest mean coverage (65%) was obtained for the ss/ds DNA viruses class at the concentration of 20TU/mL while a slight decrease is observed at the lower concentrations (50% and 47%). Regarding the RNA viruses class coverage there are not significant differences between the 20TU/mL and the 10TU/mL (respectively 43% and 41%) however a strongly decrease is observed at the lower concentration of 2TU/mL (26%).

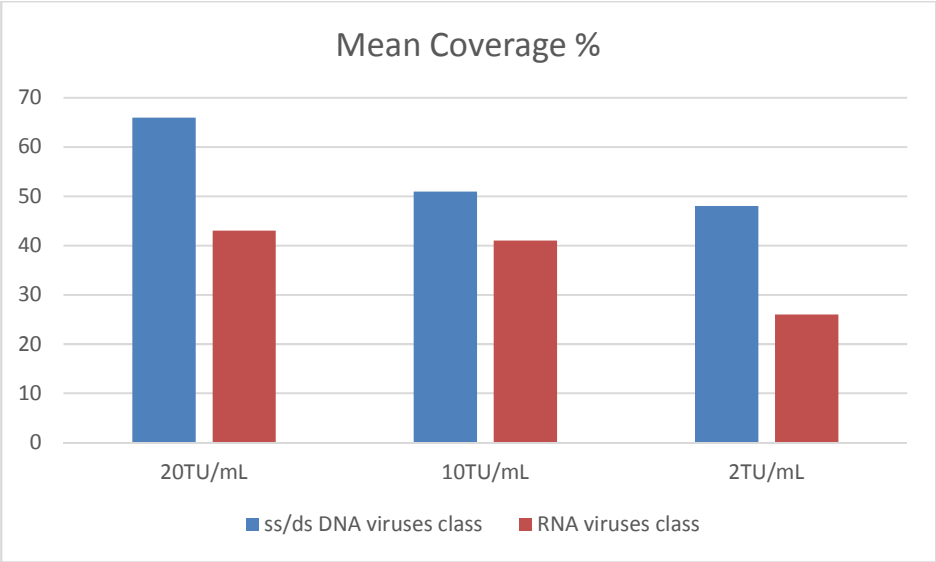

Figure 19. The mean coverage between the different classes of viruses (ss/ds DNA / RNA) for the three viral concentrations (expressed in titration unit/mL).

CONCLUSIONS

The method described consists of a generic protocol for sequencing and detecting by NGS both DNA and RNA viruses (ss/ds DNA - RNA) by mean of a bioinformatic pipeline and a curated and validated viral database created to avoid false positive results and discriminate the endogenous retroviral sequences normally present in the samples.

42 BH blank samples coming from 4 different matrixes were analyzed to create a background of the endogenous retroviral sequences. Moreover, the sensitivity of the method was assessed using different classes of viruses (ss/ds DNA - RNA) for spiking' test. Based on the results obtained, the protocol can be applied for the viral contamination detection in Bulk harvest, even if better results were obtained with the DNA viruses compare to the RNA ones. This is probably due to an efficiency of retrotranscription step that may cause a loss of viral material during the entire workflow. Moreover, a mean limit of detection of 10TU/mL was identified during the set-up activities. The analysis was performed on both the NextSeq500 and the HiSeq1000 instruments to have two comparable platforms that can be used based on the routine workload.

Further steps will be introduced to improve the method for the RNA virus detection and to perform the validation process to use it as routinely safety test in the quality control process. Particularly, the Robustness, the Specificity and the Limit of Detection of the method will be assessed and demonstrated according to the regulatory guidelines.

In conclusion, implementation of this methodology will allow the protection of production facilities against potential contaminations with any kind of adventitious viruses and mitigate the risk of a contamination spread.

CHAPTER 4

PERFORMANCE EVALUATION OF NANOPORE SEQUENCING PLATFORM FOR OPTIMAL VIRUS QUALIFICATION

INTRODUCTION

During the last decade, next generation sequencing (NGS) has emerged as an advanced technology that can provide high speed, throughput for a range of real-world applications, including broad viral detection [53,54]. From the first next-generation high-throughput sequencing technology, rapid expansion and improved platforms have been developed. Sequencing of full-length viral genomes is a difficult task due to the limitations in the short length of reads obtained using the most common high-throughput sequencing techniques, such as Illumina platform [55]. Even if the third-generation sequencing (TGS) technology, led by Pacific Biosciences (PacBio), can provide longer reads for high quality *de novo* assembly of viral genomes [56], the high cost is still a bottleneck.

The MinION device is the world's first marketed single-molecule nanopore sequencer, launched a few years ago from Oxford Nanopore [57]. This newly developed technology offers the possibility of sequencing very long DNA fragments offering a promise of solving assembly problems for large and complex viral genomes [58], in contrast with short-read sequencing technologies. The device, given the technology, has a very low cost and the scalability to larger platforms is easily achievable.

Due to its greater features, the 4th Next Generation Platform, the MinION Nanopore Sequencer was explored focusing on full-length retroviral genome sequencing and choosing as model sample the Feline Leukemia virus.

This project was carried out in a laboratory inside the Center of Biologics Evaluation and Research within the Food and Drug Administration in Silver Spring (CBER/FDA). The CBER regulates biological products for human use under applicable federal laws, including the Public Health Service Act and the Federal Food, Drug and Cosmetic Act.

The study, performed during a 5 months internship, included the following steps:

- optimization of sample preparation step to get the highest viral RNA quantity;
- development of RT protocol to get full-length cDNA
- use of the MinION and data analysis using two different approaches

MATERIALS AND METHODS

SAMPLE PREPARATION

The Feline leukemia virus (FeLV) is a member of the Gammaretro-virus genus, from the family Retroviridae. Feline Leukemia able to infect cats [65]. The genome is approximately 8.4-kb in length and its structure includes three genes (*gag*, *pol* and *env*) flanked by un-translated regulatory sequences known as long terminal repeats (LTR). Gag encodes group-specific capsid antigens, pol encodes protease, integrase, and reverse transcriptase enzymes, and env encodes the envelope proteins [66]. For this study, a viral stock of FeLV was propagated at American Type Culture Collection (ATCC) and send to the laboratory of retrovirus inside the Division of Viral Products (DVP/CBER) to be analyzed by NGS. The stock concentration was calculated by ATCC using qPCR (.

VIRAL RNA EXTRACTION

Viral RNA extraction from viral stock was carried out using silica membrane with selective binding properties (QIAGEN Viral RNA Kit) according to the manufacturing instruction. Briefly, the viral RNA extraction was performed starting from 140 µL of sample that was lysed under highly denaturing conditions at elevated temperatures to inactivation of RNases to ensure isolation of intact viral RNA. Lysate was then transferred onto a QIAamp Mini spin column and RNA bound to the membrane. Contaminants were washed away in two steps using two wash buffer. Elution was performed in water. The sample is finally quantified by RNA High Sensitivity range kit Qubit® 3.0 Fluorometer (ThermoFisher Scientific) and used for the step of full length cDNA synthesis.

RETROTRANSCRIPTION AND dscDNA SYNTHESIS

Viral RNA was then retrotranscribed using poly-T primer by mean of Maxima First Strand cDNA Synthesis Kit (Thermo Fisher), that allows synthesis of long cDNAs up to 20kb at elevated temperatures. A double strand cDNA synthesis was carried out using SequalPrep Long PCR kit with dNTPs (Applied Biosystems) and specific primer for FeLV (Figures 20 and 21). The PCR products were analyzed by agarose gel electrophoresis (data not shown) and only long fragments (between 2.5kb and 9kb) were purified from gel using Zymoclean Gel DNA recovery kit (Zymo Research). The dscDNA obtained was quantified by Qubit® 3.0 Fluorometer and then used for the library preparation.

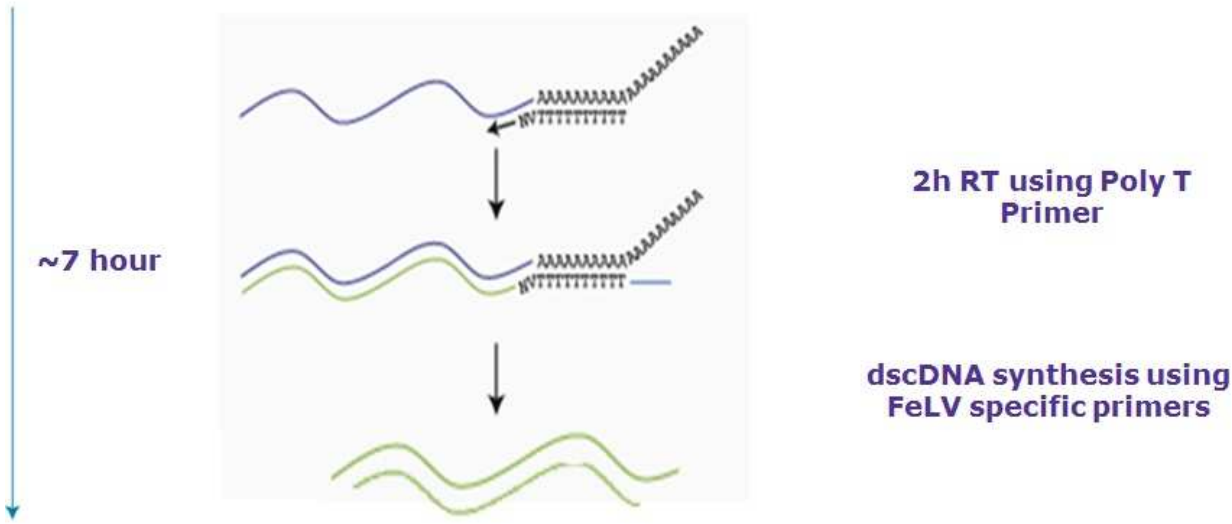

Figure 20: Full length Retrotranscription and dscDNA synthesis

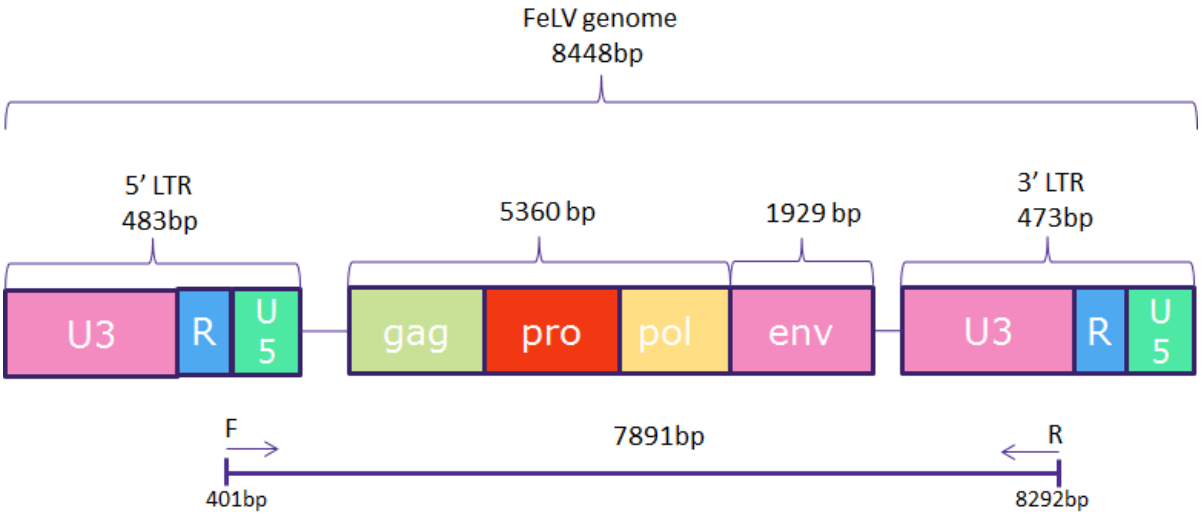

Figure 21: FeLV primer design

LIBRARY PREPARATION AND SEQUENCING

The Ligation Sequencing Kit 1D from Oxford Nanopore (SQK-LSK108) was used to prepare the library for sequencing on MinION starting from 1µg of dscDNA. Briefly the steps for library preparation consisted of:

- end-repair
- dA-tailing
- adapter Ligation

The NEBNext End Repair/dA tailing module (NEB) was used to prepare the dscDNA ends. Adapters supplied in the Ligation Sequencing Kit 1D were then ligated onto the dscDNA, the ligation was assisted by hybridization of the A and T overhangs of the DNA fragments and adapters, respectively. The 'leader adapter', consisted of two oligos with partial complementarity that form a Y-shaped structure once annealed [9] allowing the attachment of the first strand to the sequencing nanopore. The adapter was conjugated with a special protein (Motor enzyme, HP motor, Tether) that controlled tethers the DNA strand to the membrane of the MinION Flow Cell and reduces the diffusion of the DNA strands from three to two dimensions (Figure 22). A DNA control of 3.6kb amplicon lambda genome was added to sample before the library preparation.

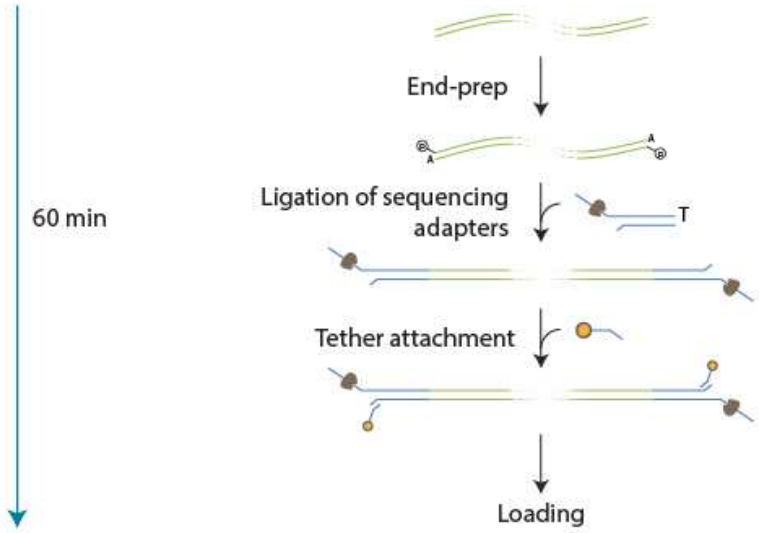

Figure 22: Ligation Sequencing 1D Protocol (Oxford Nanopore)

Final library was evaluated by Qubit 3.0 Fluorimeter and loaded on the MinION Flow cell R 9.4 version. One strand of the duplex was sequenced at a time, producing 1D reads. The nanopores sequenced the full length of fragments presented to them.

MinKNOW is the software for running the MinION sequencer. The MinKNOW software carries out several core tasks: data acquisition, real-time analysis and feedback, basecalling, data streaming, providing device control, and ensuring that the platform chemistry is performing correctly to run the samples (Figure 23).

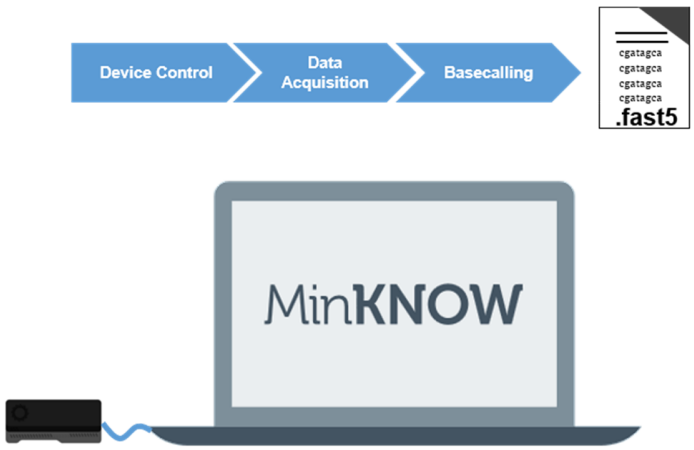

Figure 23: MinKNOW software workflow

DATA ANALYSIS

For the data analyses the fast5 files produced by the instrument were converted in fastq file using *Poretools*. In a preliminary phase, the reads were corrected using *nanoCORR* utilizing Illumina data previously obtained on the same FeLV stock (Illumina data not shown). *BWA MEM* tool was used to aligned the reads to FeLV genome reference.

Moreover, a *de novo* assembly was performed using the software *Canu*. This software performed the correction, the trimming and the assembly of the reads to unitigs. Important information needed for the software was approximate genome size (to determine the coverage in the input reads) and the technology used to generate the reads. Pilon tool was used to polish Oxford Nanopore assembly.

RESULTS AND CONCLUSIONS

A purified viral stock of FeLV was analyzed using the new pocket Oxford Nanopore Sequencer, the MinION instrument. Using the Sequencing 1D protocol and the MinKNOW 48-hour script protocol for local basecalling the sample was run for 24 hours. The total amount of data produced corresponded to about 240.000 reads (produced by the instrument in a fist5 file format).

By mean of bioinformatic tools reads were corrected and a total of final 117.558 reads were used for the alignment. The NCBI FeLV genome KP728112.1 was used as reference and the data results obtained were a total coverage of 100% with 99,88% of mapped reads, with a lower coverage zone related to the *pol/gag* (Figure 24). In the *de novo* assembly exercise, a single full-length molecule was responsible for the generation of a unitig of 7559bp after internal (ONT-only with *Canu*) error correction. The span of this sequences of consistent with the viral genome structure as it aligned on NCBI reference genome sequence KP728112.1. However, many erroneous deletions in the sequence were present due to errors brought by nanopore sequencing. By further applying a polishing step using Illumina reads, the length of such sequence was corrected to 7861bp, with only 30bp of difference to the expected amplicon length.

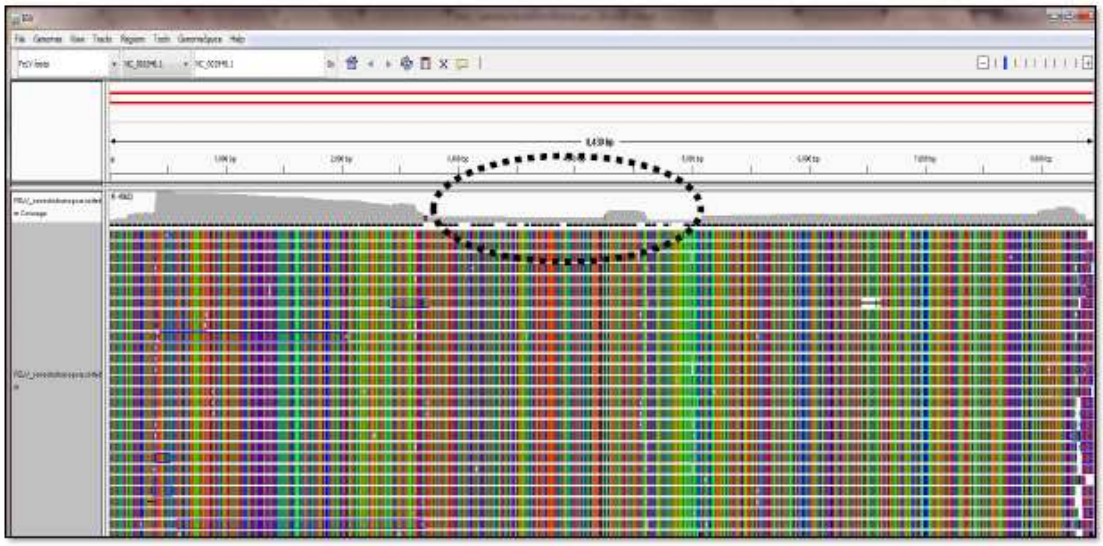

Figure 24: Mapping of reads on FeLV reference genome using BWA

Even if the accuracy of the technology needs to be improved and a surprisingly not uniform coverage of the genome was obtained, these results demonstrate that the MinION represents an easy and successful tool for the full-length retrovirus genome sequencing. Future steps will be to use the newest Oxford direct RNA sequencing protocol which for the first time allows the sequencing of an RNA molecule and not a synthetic copy, overcoming the limitation of the reverse transcriptase or PCR bias and revealing any modifications present when analyzing the raw data [67].

FINAL REMARKS

Biopharmaceutical manufacturing, because of its complex nonlinear nature, is fraught with a myriad of process variations that can impact safety and efficacy of the drug [64]. For this reason, different guidelines have been created over the time for biological quality control processes [2,3] hand to hand with the improvement of analytical methods used. Particularly, the advent of the Next Generation Sequencing (NGS) has completely revolutionized the genome field opening the way for exploring new fields of research and application [63].

In this project, different methods based on NGS were developed and applied as support for the quality control steps during the biopharmaceutical production process. Among the several sequencing systems, the most used technology (Illumina) and the emerging one (Nanopore) were evaluated and different methods based on these platforms were setting up. These methods concerned:

- the transcriptome analysis of the cell lines used during the preliminary phase of the development process to discover the candidate final clone,
- the whole genome analysis of the Master Cell Bank created during the upstream phase,
- the identification of potential adventitious viral contaminants in the Bulk harvest during the in-process phase of the biopharmaceutical production.

Furthermore, the newest Oxford Nanopore sequencing technology, due to its capability to get very ultra-long sequences, was evaluated for the full-length genome retroviral identification.

The results obtained demonstrate that, compare to other traditional methods, the NGS allows to perform a complete genome and transcriptome analysis with a higher sensitivity and throughput and without the need of a priori knowledge of your target. Moreover, implementation of these methods as routine tests for the quality control represents an important tool to evaluate and guarantee the safety and the quality of the biopharmaceuticals, reducing the cost and the time consuming of the entire biomanufacturing process.

REFERENCE

[1] Berlec, A. &. (2013). Current state and recent advances in biopharmaceutical production in Escherichia coli, yeasts and mammalian cells. Journal of Industrial Microbiology & Biotechnology, 40(3), 257–274.

[2] WHO Expert Committee on Biological Standardization. (2013). Annex 3: Recommendations for the evaluation of animal cell cultures as substrates for the manufacture of biological medicinal products and for the characterization of cell bank. WHO Technical Report Series, No 978: World Health Organization, 2013.

[3] Food and Drug Administration. (2010). Guidance for Industry: Characterization and qualification of cell substrates and other biological materials used in the production of viral vaccines for infectious disease indications. U.S. Department of Health and Human Services

[4] Use, C. for M. P. for H. (2006). European Medicines Agency. Draft Guideline on Reporting the ICH Topic Q 5 A (R1) Quality of Biotechnological Products: Viral Safety Evaluation of Biotechnology Products Derived from Cell Lines of Human or Animal Origi, (June 1996), 1–29.

[5] EMA. (2014, July). European Medicines Agency. ICH Topic Q 5 B Quality of Biotechnological Products: Analysis of the Expression Construct in Cell Lines Used for Production of r-DNA Derived Protein Products

[6] G. J. et al, (2014 May). Systematic evaluation of in vitro and in vivo adventitious virus assays for the detection of viral contamination of cell banks and biological products. Vaccine, 19;32(24):2916-26.

[7] Illumina. (n.d.). Illumina.com. Retrieved from https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html

[8] Janitz, M. (2010). Next-Generation Genome Sequencing.

[9] Quick J, Q. A. (2014). A reference bacterial genome dataset generated on the MinIONTM portable single-molecule nanopore sequencer. *GigaScience* , 3:22doi:10.1186/2047-217X-3-22.

[10] nanoporetech. (n.d.). Retrieved from https://nanoporetech.com/how-it-works

[11] Chee Furng Wong D, T. K. (2005). Impact of dynamic online fed-batch strategies on metabolism, productivity and N-glycosylation quality in CHO cell cultures. Biotechnol Bioeng., 89(2):164-177. doi:10.1002/bit.2.

[12] Chugh PK, R. V. (2014). Current scientific and regulatory considerations. Curr Clin Pharmacol. , 9(1):53-63.

[13] Pacis E, Y. M. (2011). Effects of cell culture conditions on antibody N-linked glycosylation-what affects high mannose 5 glycoform. Biotechnol Bioeng. , 108(10):2348-2358. doi:10.1002/bit.23200.

[14] Rouiller Y, P. A.-N. (2012;). Effect of hydrocortisone on the production and glycosylation of an Fc-fusion protein in CHO cell cultures. Biotechnol Prog., 28(3):803-813. doi:10.1002/btpr.1530.

[15] Ha, E. H.-J.-J. ((2006).). Microarray analysis of transcription factor gene expression in melatonin-treated human peripheral blood mononuclear cells. . Journal of Pineal Research, , 40(4), 305–311.

[16] Yee, J. C.-S. (2008). Quality assessment of cross-species hybridization of CHO transcriptome on a mouse DNA oligo microarray. . *Biotechnology and Bioengineering*,, 101(6), 1359–1365. https://doi.org/10.1.

[17] Vishwanathan, N. Y. (2015). Global insights into the Chinese hamster and CHO cell transcriptomes. Biotechnology and Bioengineering, , 112(5), 965–976. https://doi.org/10.1002/bit.25513.

[18] Monger, C. K. (2015). Towards next generation CHO cell biology: Bioinformatics methods for RNA-Seq-based expression profiling. Biotechnology Journal, 10(7), 950–966. https://doi.org/10.1002/b.

[19]truseq_stranded_total_rna_sample_preparation_guide_15031048. (n.d.). Retrieved from https://support.illumina.com/downloads/truseq_stranded_total_rna_sample_preparation_guide_15031048.html

[20] CASAVA 1.8.2 Quick Reference Guide Part # 15011197 Rev C . (2011., October).

[21] trimmomatic. (n.d.). Retrieved from http://www.usadellab.org/cms/?page=trimmomatic

[22] Trimmomatic. a flexible trimmer for Illumina sequence data. . (2014). Bioinformatics, 30:2114.

[23] CHOproj. (n.d.). Retrieved from https://gendbe.cebitec.uni-bielefeld.de/cho.html

[24] Anders S, P. P. (2015). HTSeq-A Python framework to work with high-throughput sequencing data. . Bioinformatics. , 31(2):166-169. doi:10.1093/bioinformatics/btu638.

[25] htseq. (n.d.). Retrieved from http://htseq.readthedocs.io/en/release_0.9.1/

[26] DESeq2. (n.d.). Retrieved from https://bioconductor.org/packages/release/bioc/html/DESeq2.html

[27] Retrieved from https://ccb.jhu.edu/software/tophat/index.shtml

[28] Brühlmann, D. M. (2017). Cell culture media supplemented with raffinose reproducibly enhances high mannose glycan formation. Journal of Biotechnology, 252(December 2016), 32–42. htt.

[29] Fan, L. R. (2017). Comparative study of therapeutic antibody candidates derived from mini-pool and clonal cell lines. Biotechnology Progress, 33(6), 1456–1462. https://doi.org/10.1002/b.

[30] G., W. (2014). Biopharmaceutical benchmarks . *Nat Biotechnol*, 32:992–1000.

[31] Frye C, D. R. (2016). Industry view on the relative importance of “clonality” of biopharmaceutical-producing cell lines. Biologicals., ;44:117–122.

[32] Ko, P. M. (2017). Probing the importance of clonality: Single cell subcloning of clonally derived CHO cell lines yields widely diverse clones differing in growth, productivity, and product quality. Biotechnology Progress.

[33] Evans K, A. T.-Q. (2015). Assurance of monoclonality in one round of cloning through cell sorting for single cell deposition coupled with high resolution cell imaging. Biotechnol Prog, 31.

[34] Fieder J, S. P. (2017). A single-step FACS sorting strategy in conjunction with fluorescent vital dye imaging efficiently assures clonality of biopharmaceutical production cell lines. .Biotechnol J, ;12:1700002.

[35]https://support.illumina.com/content/dam/illuminasupport/documents/documentation/chemistry_documentation/samplepreps_nextera/nexteramatepair/nextera-mate-pair-reference-guide-15035209-02.pdf. (n.d.).

[36] Xu et al., The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line, Nature Biotech.Vol.: 29, 735–741, 2011)

[37] R., L. H. (2009). Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60.

[38] L. H. et al., (2009 Aug 15). The Sequence Alignment/Map format and SAMtools . Bioinformatics , 25(16):2078-9.

[39] D. B.et al., *BAMTools: a C++ API and toolkit for analyzing and managing BAM files*.

[40] V. J. et al., (2010 Jun;). Viral nucleic acids in live-attenuated vaccines: detection of minority variants and an adventitious virus. . J Virol, 84(12):6033-40.

[41] C. e.et al., (2013). Next-generation sequencing technology in clinical virology. . Clin. Microbiol. Infect., 19:15–22.

[42] S. P. et al., (2014). Exploring the Potential of Next-Generation Sequencing in Detection of Respiratory Viruses . J. Clin. Microbiol., , vol. 52 no. 10 3722-3730.

[43] D. G. et al., (2013 Nov;). Investigation of a regulatory agency enquiry into potential porcine circovirus type 1 contamination of the human rotavirus vaccine, Rotarix: approach and outcome. . Hum Vaccin Immunother, 9(11):2398-408.25973.

[44] G. J. et al., (2014 May). Systematic evaluation of in vitro and in vivo adventitious virus assays for the detection of viral contamination of cell banks and biological products. . Vaccine, 19;32(24):2916-26.

[45] M. S al., (2014 Dec 12;). Evaluation of cells and biological reagents for adventitious agents using degenerate primer PCR and massively parallel sequencing. . Vaccine , 32(52):7115-21.

[46] O. D. al., (2011 Sep 22;). Ensuring the safety of vaccine cell substrates by massively parallel sequencing of the transcriptome. Vaccine, 29(41):7117-21.

[47] R. B. al., (2014 Nov;). Detection of adventitious agents using next-generation sequencing. PDA J Pharm Sci Technol, 68(6):651-60.

[48] Nextera® DNA Sample Preparation Guide Part # 15027987 . (2016 January).

[49] Illumina. ((2015).). Guida del sistema cBot (15006165). Retrieved from https://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/cbot/translations/cbot-system-guide-15006165-02-ita.pdf.

[50] bcl2fastq Conversion Software [WWW Document], 2016. URL http://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html.

[51] Bowtie [WWW Document], 2016. URL http://bowtie-bio.sourceforge.net/bowtie2/index.shtml.

[52] Ninth ICTV Report on Viral Taxonomy. (2012). *International Committee on Taxonomy of Viruses 2012.*

[53] R. A. al., (2012,). Application of next-generation sequencing technologies in virology. J Gen Virol , 93 (Pt 9):1853–1868.

[54] S. D. al., (2015 Aug 12;). Next-generation sequencing in clinical virology: Discovery of new viruses. World J Virol. , 4(3): 265–276.

[55] D. A. al., (2013). Next generation sequencing of viral RNA genomes. BMC Genomics , 14:444.

[56] Rhoads A, A. K. (2015;). PacBio Sequencing and Its Applications. . Genomics,Proteomics & Bioinformatics., 13(5):278-289. doi:10.1016/j.gpb.2015.08.002.

[57] Feng Y, Z. Y. (2015;). Nanopore-based Fourth-generation DNA Sequencing Technology. . Genomics,Proteomics & Bioinformatics , 13(1):4-16. doi:10.1016/j.gpb.2015.01.009.

[58] Quick J, L. N. (2016;). Real-time, portable genome sequencing for Ebola surveillance. *Nature.*, 530(7589):228-232. doi:10.1038/nature16996.

[59] Sarah Kennett, Ph.D, Establishing Clonal Cell Lines – A Regulatory Perspective Black Cell, Blue Cell, Old Cell, New Cell?. 2014, Review Chief Division of Monoclonal Antibodies OBP/OPS/CDER/FDA

[60] B4GALT3 beta-1,4-galactosyltransferase 3 [Homo sapiens (human)] [WWW Document], 2016. URL https://www.ncbi.nlm.nih.gov/gene/8703.

[61] Ishida, N., Kawakita, M., 2004. Molecular physiology and pathology of the nucleotide sugar transporter family (SLC35). Pflugers Arch. Eur. J. Physiol. 447, 768–775. http://dx.doi.org/10.1007/s00424-003-1093-0.

[62] Nieman, D.C., Scherr, J., Luo, B., Meaney, M.P., Dréau, D., Sha, W., Dew, D.A., Henson, D.A., Pappan, K.L., 2014. Influence of pistachios on performance and exerciseinduced inflammation, oxidative stress, immune dysfunction, and metabolite shifts in cyclists: a randomized, crossover trial. PLoS One 9, e113725. http://dx.doi.org/10.

[63] Jiekun Xuan, Ying Yu, Tao Qing, Lei Guo, Leming Shi, 2013 Nov 1, Next-generation sequencing in the clinic: Promises and challenges, Cancer Lett, 340(2): 284–295.

[64] Rathore, A. S., Kumar, D., & Kateja, N. (2018). Role of raw materials in biopharmaceutical manufacturing: risk analysis and fingerprinting. Current Opinion in Biotechnology, 53, 99–105. https://doi.org/https://doi.org/10.1016/j.copbio.2017.12.022

[65] Kawamura M, Watanabe S., Odahara Y., Nakagawa S., Endo Y., Tsujimoto H., Nishigaki K., 2015 Jun 2, Genetic diversity in the feline leukemia virus gag gene., Virus Res.;204:74-81.

[66] Coffin, J.M.; Garfinkel, D.J.; Kozak, C.A.; Leung, N.J.; Luciw, P.A.; Myers, G.; Pavlakis, G.N.; Payne, L.N.; Ruscetti, S.; Temin, H.M. Virion Structure. In The Retroviridae; Levy, J.A., Ed.; Plenum Press: New York, NY, USA, 1992; pp. 21–34.

[67] Direct RNA seequencing protocol, https://nanoporetech.com/rna

[68] http://cellculturedish.com/2013/06/perfusion-bioreactors-with-so-much-to-offer-they-deserve-a-closer-look/

[69] Genzyme Corporation. s.d. https://www.sec.gov/Archives/edgar/data/732485/000110465910029855/a10-9595_13defa14a.htm.