UNIVERSITA' DEGLI STUDI DI TORINO

Ph.D. School "Health and Life Sciences"
Ph.D. Programme "Molecular Medicine"

Ph.D. THESIS

ADVANCEMENT OF NGS BASED METHODS IN SUPPORT TO THE CHARACTERIZATION OF BIOTECHNOLOGICAL CELL LINES

XXXIII CYCLE

Tutors:
Dott.ssa Alice Praduroux
Chiar.ma Prof. Fiorella Altruda

Candidate:

Dott.ssa Carola Veglia

ACADEMIC YEARS: 2017-2021

TABLE OF CONTENTS

Advancement of Next Generation Sequencing based methods in support to the characterization of biotechnological cell line

1. INTRODUCTION

Pharmaceutical Drugs can be distinguished in chemicals and biologicals. Currently, there is no simple way to define all the drugs that are reported as biologics. Biologics are created by either a microorganism or mammalian cells and are large complex molecules, most of which are proteins or polypeptides.
Biologicals are defined as follows by the International Conference on Harmonization (ICH) Q5D guideline:

"*Biotechnological/biological products" refers to any products prepared from cells cultivated from cell banks with the exception of microbial metabolites such as, for example, antibiotics, amino acids, carbohydrates, and other low molecular weight substances*" (ICH Q5D 1997).

While Chemical compounds production is standardized and generally better characterized by analytical methodologies, biotechnological drugs are produced in biological systems. Most biologicals are produced by the DNA recombinant technology: one or more genes, coding for a human protein with a therapeutic effect, are integrated in a vector and then expressed in a specific cell system. Biologicals must be processed under tightly controlled conditions as multiple factors can influence mammalian and microbial cell banks growth, correct protein folding and the production rate. Checking cell growth conditions allows the production of a consistently safe, pure and potent product.
The production and the commercialization of the biological products must follow strict guidelines, the so-called Good Manufacturing Practices (GMP). Health Authorities (HAs), as EMA (European Medicines Agency) and FDA (Food and Drug Administration), provide these guidelines with the scope of guarantee the efficacy and the safety of a drug.

1.1 PHARMACEUTICAL DRUG PRODUCTION

Biotechnological drugs production starts with the isolation of the DNA coding for the molecule of interest, from human cells. This happens through the targeting of the genomic region containing the sequence of interest with the hybridization of a single-strand probe, and the subsequent cutting of the fragment extremities with specific restriction enzymes. That sequence is then engineered to be inserted inside an expression vector, containing variable regulatory elements, including a viral promoter and enhancer, the gene of interest without introns and at least one or more introns usually located between the promoter and the coding region. These introns ensure an efficient cytoplasmic transport and an efficient mRNA translation. All those regulatory elements vary in function of the cellular system used for the protein production. Today, 60-70% of the recombinant therapeutic proteins are produced in mammalian CHO (Chinese Hamster Ovary) cell lines, nevertheless, other cell lines, such as those derived from mouse myeloma (NS0), baby hamster kidney (BHK), human embryo kidney (HEK-293) and human retinal cells have gained regulatory approval for recombinant protein production [1]. Moreover, E. Coli bacterial cell banks are also used for the production of less complex recombinant therapeutic proteins.

These cell lines represent the main expression system to produce biopharmaceuticals thanks to their high productivity and ability to guarantee the right folding and the correct post-translational modifications of the expressed protein [2]. The transfection of the expression vector inside mammalian cells is mediated by several methodologies, like electroporation, lipofection and the calcium phosphate one. Inside the host cell, the expression construct frequently remains in the cytoplasm giving a transient expression of the protein. Much more rarely, it integrates into the host cell genome giving a stable protein expression. To increase the chance of a stable integration, the vector is linearized before transfection. Nuclear ligase binds covalently the linearized plasmid and the enzymes involved in non-homologous recombination determine the random integration within the genome. In very rare cases, the integration also occurs with the homologous recombination mechanisms. The integration site can critically affect the recombinant gene expression. In fact, if integration occurs within inactive heterochromatin regions, the transgene is not expressed or expressed at the very low level. To ensure a strong expression of the transgene is essential that plasmid integration occur in the euchromatic regions of the host genome. This phenomenon is known as position effect [3]. However,

the integration of the plasmid within the euchromatin regions is not sufficient to ensure a stable transgene expression over time: the recombinant gene could be silenced as a result of modifications such as histone hypoacetylation, and/or methylation of the CpG islands of the promoter [1]. For a long-term recombinant protein production, clones, in which the expression construct is stably integrated within their genome (transfectants), are selected by means of a so-called marker of selection.

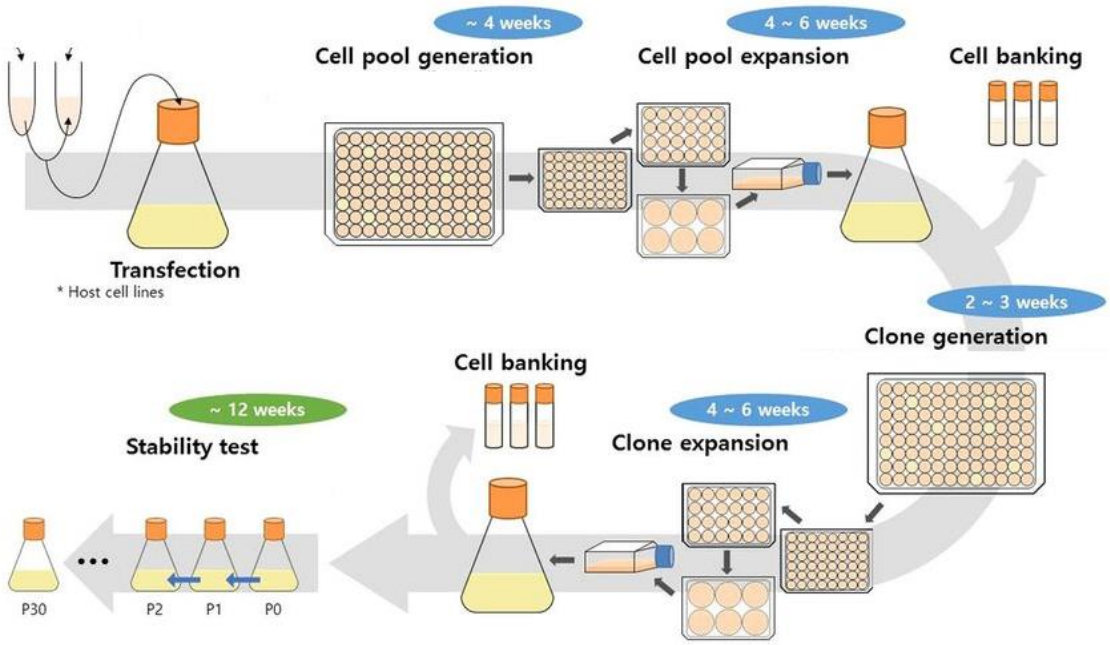


Figure 1. Schematic representation of a small-scale recombinant cell bank development. A mammalian host cell line is transfected with the recombinant vector containing the transgene encoding for the therapeutical protein. After four weeks, cells which have integrated the expression vector, are plated, and expanded (6 weeks). After that, a process of clone generation starts. Cells are plated in a 96-wells plate, to reach the concentration of 0.5 cell/well, in order to generate totally clonal subpopulation from each well. After the expansion, subpopulation are submitted to stability tests, and to the selection of the "best clone".

The selection marker can be located on the same construct of the recombinant gene (polycistronic transcript), or alternately on a different vector. Currently, the most used system for the selection of transfectants in mammalian cells is the one based on DHFR (dihydrofolate reductase) use, an enzyme involved in nucleotide metabolism [4]. Another

method of selection is the system based on glutamine synthetase (GS) [4]. After the selection of transfectants, the polyclonal culture is subjected to limiting dilutions in order to isolate individual clone, each of which will be transferred to a different container and then expanded to obtain a monoclonal population. Afterwards, a screening is needed to identify the transfected subclone with the best performances in terms of cell growth, genomic stability and protein production rate. Ones the "Best Clone" is selected, it finally undergoes to a phase of big scale production into bioreactors, to produce huge amount of recombinant protein.

1.2 BIOPHARMACEUTICAL QUALITY CONTROL AND GMP

The complex production process of biologics points out more challenges respect to the chemical drugs. In fact, biotechnological products are molecules with high molecular weight, one or more subunits and high structural complexity. During the production of biologicals, since live systems are involved, tiny differences in the condition and execution of the production can have a great impact on the quality of the final product. This is the reason why a large and tightened panel of quality control tests are put in place to guarantee the clonality and stability of recombinant drugs.

Both the production and the quality control testing must be compliant with the so-called Good Manufacturing Practices (GMP), provided by Health Authorities like FDA (Food and Drug Administration), and EMA (European Medicines Agency). GMP describes a set of principles and procedures that helps ensure a high quality of therapeutic goods. A basic tenet of GMP is that quality cannot be tested into a single batch of product but must be built into each batch of product during all stages of the manufacturing process. This assures to the customer the efficacy, safety, quality and absence of contamination of the final product all the time.

At Merck-Ivrea Site, many molecular assays are performed to guarantee drugs quality control. In particular, the one treated in this thesis, is the panel of Genotypic Characterization tests.

1.2.1 CELL BANK GENOTYPIC CHARACTERIZATION

Genotypic Characterization represents a category of the many quality control tests that must be performed to guarantee recombinant molecules stability. Their objective is to ensure that during the manufacturing of the biologicals drug no modifications occurred on cell growth conditions and on the identity of the coding and regulatory nucleic acid sequence of the desired recombinant protein. The absence of any significant difference is checked by the comparison of MCB and Extended cell bank (ExCB). If no difference is observed between the point before building the bioreactor (MCB) and the point before stopping the bioreactor production (ExCB), it can be inferred that no modifications on the process happened.

Genotypic characterization tests are performed both on in clinical trial molecules and on already on the market molecules, but the panel of assays varies between the two clinical phases. For what concerns phase I-II only three tests are required by health authorities (The Gene Copy Number, the Restriction Enzyme Analysis and the Sequencing). Phase III and sold molecules are checked by all genotypic characterization assays.

At Merck-Ivrea site, the panel of Genotypic Characterization tests is the following:

Figure 2. Genotypic characterization tests panel. Number 1 describes the FISH test; the number 2 describes the Gene Copy Number test; the number 3 describes the Restriction enzyme analysis; the number 4 describes the mRNA size test; and the panel number 5 describes the sequencing test.

I. **FISH** (Fluorescence In Situ Hybridization) is used to visualize the gene integration by means of a specific fluorescent probe.

II. **mRNA Sequencing** is carried out by Sanger sequencing to investigate the sequence of the coding regions of the gene/s of interest.

III. **DNA Sequencing** is also performed by Sanger sequencing to investigate the flanking control regions of the gene/s. These regions contain elements with an important role in transcription, translation and stability of the coding sequence as promoter, enhancers, splicing sequences and the polyadenylation site;

IV. **Gene Copy Number** of the gene of interest is established by Southern blot or qPCR;

V. **Restriction Enzyme Analysis** is performed by Southern blot to assess the integrity of expression construct;

VI. **mRNA size** and unicity of the gene of interest is managed by Northern blot.

In this thesis, the determination of the nucleic acid sequence of coding and flanking control regions 5' and 3' will be described in detail.

1.2.2 MAMMALIAN AND BACTERIAL NUCLEIC ACIDS SEQUENCE DETERMINATION OF CODING AND FLANKING CONTROL REGIONS WITH SANGER TECHNOLOGY

The determination of the nucleic acid sequence of coding and flanking regions 5' and 3' allows to identify if any mutation occurred at the DNA or RNA level during the production of recombinant cell banks. It is divided in two different tests, one structured to evaluate the nucleic acid stability of transgene coding region (test B) and the other one designed to evaluate the genetic stability of control flanking regions (test C). Both tests require that an initial vial of MCB, working cell bank (WCB) and ExCB, received from Merck manufacturing sites, has to be propagated to obtain enough cell pellets to perform genotypic characterization tests. For test B, when pellets are obtained, the total RNA is extracted using the RNeasy Mini Kit (Qiagen). The process of extraction is based on a column-dependent retention of ribonucleic acids with a subsequent elution of RNA in

water. The extraction process is followed by a quantification by spectrophotometer NanoDrop (Thermo Fisher) and an evaluation of sample integrity by electrophoretic analysis, in which the two most abundant ribosomal RNA forms should be recognized as specific bands.

A Polymerase Chain Reaction is performed to amplify the entire coding region of transgene, from the transcription start site (ATG) to the stop codon (TGA). To do that, a set of primers and the optimization of the assay thermal profile are designed to obtain the best results in terms of accuracy and sensibility.

The PCR product is then checked by electrophoresis and bands are analysed comparing their expected base pairs (bp) length to a molecular standard marker. A purification step is performed before the sanger sequencing inside the Ab3500/3500xl sequencer (Applied Biosystems). Sequences are analysed with SeqScape V3.0 software (Applied Biosystems). Reads are aligned with the reference coding sequence and checked for any mutation, heterozygosis insertions or deletions. If none of these mutations occurred, the final results are released to the manufacturing site.

For test C, once cell pellets are obtained, the total DNA is extracted using the DNA Blood Mini Kit (Qiagen). Also in this case, the process of extraction is based on a column-dependent retention of nucleic acids with a subsequent elution of DNA in water. The extraction process is followed by a quantification by spectrophotometer NanoDrop and an evaluation of sample integrity by electrophoretic analysis, in which a single high weight band (> 8000 bp) should be well recognizable and defined.

A Polymerase Chain Reaction is performed to amplify the regulatory regions that surround the coding region of transgene, including the TATA BOX (representing the core promoter region of the gene, where the TATA-binding protein binds to start gene transcription), the ATG (representing the Open Reading Frame of gene transcript), the TAA (representing the Stop Codon of gene transcription) and finally the POLYA site (representing the Polyadenylation signal that allows the synthesis of messenger RNA).

The PCR product is checked by electrophoresis and bands are analysed comparing their expected bp length to a molecular standard marker. A purification step is performed before the Sanger sequencing inside the Ab3500/3500xl sequencer. Sequences are examined with SeqScape V3.0 software. Reads are aligned with the reference coding

sequence and checked for any mutation, heterozygosis insertions or deletions. If none of these mutation occurred, the final results are released to the manufacturing site.

Sanger sequencing is nowadays considered the gold standard for genotypic stability quality control testing by Health Authorities. However, in the last decade, Next Generation Sequencing (NGS) technologies have gained steps forward because of its high throughput and reliability. This is the reason why Merck is focusing its efforts to substitute old-fashion Sanger technology with the Illumina platform based on NGS. The development of a new method based on NGS, requires big efforts to demonstrate its robustness, most of all because samples preparation and data analysis almost totally differ from the validated method. For that reason, an initial setup and validation of the new method based on NGS technology, must be performed to justify method change and to validate it for quality control testing and results release.

2. NEXT GENERATION SEQUENCING

2.1 NEXT GENERATION SEQUENCING OVERVIEW

Next Generation sequencing is a high-throughput technique able of sequencing multiple DNA molecules in parallel. The success of NGS technology is due to its capacity to sequence millions of DNA fragments at the same time, generating large amount of data within relatively short time [5]. Another advantage of this technology is the high sensitivity that allows the detection of rare DNA base-changes, therefore making possible the study of single nucleotide variants at low frequencies [6]. Moreover, the development of NGS has reduced sequencing cost enabling the widespread use of this technology [7]. Among the several NGS platforms currently available, Illumina is the most widely adopted technology and Illumina's sequencers use a sequencing-by-synthesis method producing short sequences (also called reads).

The Illumina's workflow involves four major steps as indicated in the figure below: library preparation, cluster amplification, sequencing and data analysis.

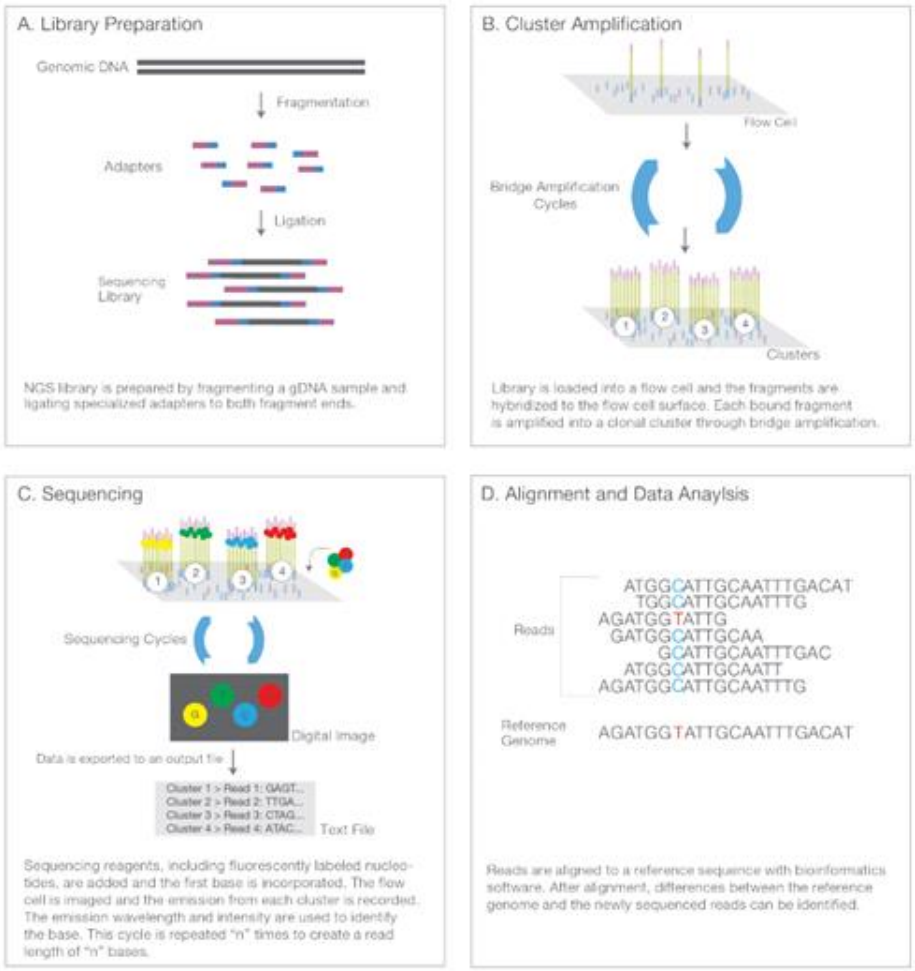


Figure 3. Next generation sequencing overview (Illumina) Illumina NGS includes four steps: (A) library preparation, (B) cluster generation, (C) sequencing, and (D) alignment and data analysis.

NGS libraries are prepared by random fragmentation of the sample (DNA, cDNA or amplicon) followed by the adapter ligation at 5′ and 3′ positions (Figure 3A). Illumina adapters are oligonucleotide sequences composed by two binding regions which bind their complementary oligos on the flowcell and index that acts as a "barcode sequence" for each read. After that through a polymerase chain reaction (PCR) the fragments are amplified. The prepared libraries are loaded into a flowcell coated with oligos complementary to the library adapters. In the NGS sequencer, through a process called "Bridge Amplification", each fragment is amplified forming distinct clonal clusters [8]. In particular, single-molecule clusters are generated in the range of millions to billions in each channel of the flowcell. Each cluster will contain multiple identical copies of the same library fragment. This process is required to boost the fluorescent signal used to

read the sequence nucleotide (Figure 3B). The sequencer can then analyse the sequence information of all the clusters simultaneously. For the sequencing phase, Illumina uses the sequencing by synthesis (SBS) technology (Illumina). This technology employs one-channel, two-channel and four-channel methods for the detection of single nucleotide. The four-channel SBS uses 4 types (A/T/C/G) of reversible dye terminators, each one tagged with different fluorophores and blocked at 3' position [9]. The sequencing process occurs in multiple cycles, each one reading a single nucleotide. Each cycle includes multiple steps. First, a fluorescently labelled nucleotide is added to the growing nucleic acid chain based on the sequence of the template. Second, the clusters are excited by a laser source in order to record the addition of the nucleotide. This fluorescence is detected by a camera that takes a picture of the flow cell after each synthesis. Lastly, the 3' blocking group and fluorophore are cleaved to allow the incorporation of the next fluorescence nucleotide by DNA polymerase (Figure 3C). Instead of using a dye for each base, in two-channels SBS, 2 fluorescent dyes are used, while in one-channel SBS only one dye is used [10].

Illumina NGS systems support both Single-End (or Single Read) and Paired-End (PE) sequencing. In the Single-End sequencing the instrument reads a fragment from only one end, generating a single sequence for each DNA molecule. In the PE sequencing, instead, both ends of the fragment are sequenced, generating two sequences for each molecule. At the end of the sequencing of the forward strands, the newly synthesized reverse strands are regenerated by bridge amplification. The forward strands are removed, leaving attached on the flow cell only the newly synthesized reverse strands to be sequenced and produce paired end sequence data. PE sequencing allows the production of high-quality reads alignment because the sequences in pair can span longer distances, increasing the accuracy of the alignment (Illumina).

The NGS data analysis process can be divided into three major steps: primary, secondary and tertiary data analysis. During the primary analysis a base-calling algorithm converts digital images to FASTQ files, providing sequences and associated quality score to each read. In the secondary analysis, short reads are aligned against a reference sequence (reads mapping) or used to build longer sequences (de-novo assembly) (Fig. 3D). Lastly, collected data are interpreted through bioinformatics, integrating public database information or other sample-related information.

2.2 NGS vs SANGER TECHNOLOGY

In principle, the concepts behind Sanger vs. NGS technologies are similar. In both NGS and Sanger sequencing, DNA polymerase adds fluorescent nucleotides one by one onto a growing DNA template strand. Each incorporated nucleotide is identified by specific fluorescent tag.

The critical difference between Sanger sequencing and NGS is the sequencing volume. While the Sanger method only sequences a single DNA fragment at a time, NGS is massively parallel, sequencing millions of fragments simultaneously per run. This high-throughput process can potentially translate into sequencing hundreds to thousands of genes at one time. NGS also offers greater discovery power to detect novel or rare variants with deep sequencing.

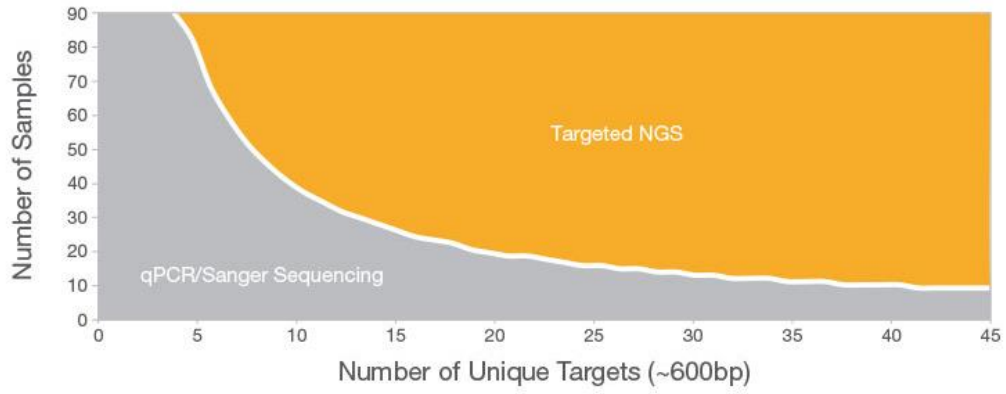


Figure 4. The area above the line represents higher cost-effectiveness with targeted DNA sequencing compared to Sanger sequencing.

From a cost-effective and a timing point of view, Sanger technology can be affordable for the processing of low numbers of targets and samples, but it has a low sensitivity and a low discovery power, in respect to NGS technology. From a technique point of view, Sanger's samples processing has many critical steps that can highly impact the goodness of cycle sequencing samples, and so the reliability and accuracy of the electrophoretic run on the sequencer. NGS, otherwise, is characterized by a higher sequencing depth that enables a higher sensitivity, a higher discovery power and a higher mutations resolution. The overall throughput data produced with the same amount of input DNA, is larger

respect to Sanger sequencing, providing more reliable and feasible results. The higher cost-effectiveness for high number of targets and samples, makes NGS technology more promising and fascinating respect to the Sanger based method and this fact, lets pharmaceutical companies, invest for its implementation inside their quality control laboratories.

2.3 NGS APPLICATIONS TO QUALITY CONTROL TESTING

Thanks to the advancement of innovative technologies, Health Authorities are progressively fostering the substitution of the old-fashioned methodologies currently used for quality control testing, in favour of the new ones. Next Generation Sequencing has many application fields and for that reason is becoming the gold standard that many pharmaceutical companies want to implement to successfully ameliorate and accelerate both the cell line development process and the quality control tests' workflow.

Between the huge panel of quality control tests performed by Merck, NGS is already used for the following purposes: to guarantee the genotypic stability of the mammalian cell bank producing recombinant proteins; to verify the biosafety of biotechnological drugs and to confirm the species of origin of mammalian cell lines used to produce recombinant proteins and exclude contaminations. However, in Merck company there are many other assays that still must be validated to guarantee drugs safety.

3. PROJECT PURPOSE

As previously mentioned, the cell line development process is characterized by three main critical steps: the expression vector design, the cell bank transfection and vector integration and finally the best clone selection. All these phases can highly impact the goodness of the cell line development workflow.

Expression vector design changes in function of the host cell line to be transfected (i.e., mammalian or microbial cell bank); it has many regulatory elements that can determine the transcription and the expression level of the recombinant gene; and lastly it has many non-coding regions (introns) that should stabilize the coding sequences but are not totally characterized. The second critical step is represented by the transfection method. It could be performed through different methodologies, including the most common used like electroporation/lipofection and calcium phosphate-based methods. Despite recent

innovative techniques of genome editing, most of the time the integration of the recombinant construct inside the host cell genome happens randomly [2]. As mentioned before, this event could bring to a high or low level of gene expression, depending on the integration sites' methylation and chromatin conformation. This aspect highly impacts the cell line development process and the best clone screening. This last phase is still performed by serial dilution of the original recombinant cells, in order to reach the concentration of 0.5 cell/well and guarantee the obtainment of a clonal population [11] . At this point, the estimated clonal population is then screened for cell growth, genomic stability and protein production rate. Evidently, manually performed serial dilutions could represent a source of error if a 0.5 cell/well is not reached before the expansion, and this could consequently bring to heterogeneous and not totally clonal MCB with differences at genome level. The aim of this project is the setup of new methods based on NGS to support the characterization of biotechnological cell lines, from the expression vectors characterization to the assessment of MCBs clonality.

3.1 PLASMID DATABASE STORAGE

The first part of the manuscript will focus on the characterization of whole expression vectors, exploiting the high throughput coverage achieved by NGS. The objective is to sequence both coding and non-coding regions of the expression vectors used to transfect recombinant cell banks. The obtained output data will be useful to totally characterize the recombinant plasmids, in order to understand if any mutation occurred during the transfection and production phases and use them as reference sequence for future bioinformatic analysis. The generation of a database containing the reference sequences of all the expression vectors used to transfect host cell lines to produce Merck molecules, can be fruitful both for the cell line development process and for the genotypic stability evaluation of its coding and regulatory regions. The approach is to generate a library from pure plasmids, avoiding the critical step of PCR amplification and potential correlated artefacts. The aim of this study is to determine the input DNA conditions for library preparation and to set the best coverage rate. In this first manuscript topic, given the confidentiality of the data shown, molecules and expression plasmids name, transgene sequences and all the additional sensitive information are not shown or specified, to protect Merck intellectual property.

3.2 DEVELOPMENT AND VALIDATION OF A METHOD FOR BACTERIAL CELL BANK GENETIC STABILITY EVALUATION USING NEXT GENERATION SEQUENCING

To evaluate the NGS specificity and limit of detection (LOD) of single point mutations, deletions and heterozygosis, an approach mimicking a production process of bacterial recombinant cell banks was used. Four different expression vectors, each containing a mutated form of the INF gamma gene. E. coli competent cells were transformed separately with the four expression vectors, which were then purified and mixed in different percentages to evaluate the specificity and the Limit of Detection of the method.

3.3 CELL BANK CLONALITY ASSESSMENT COMBINING TLA AND SANGER TECHNOLOGIES

The third part of the manuscript will focus on the clonality assessment of a recombinant cell bank, using the Target Locus Amplification technique (TLA) combined with NGS and Sanger sequencing. This assay was performed to confirm the integration sites of an expression vector inside a host cell genome and evaluate the clonality of the MCB, identifying the integration sites' conservation in 30 subclones generated by serial dilution of the original MCB vial. This new high-throughput technique can be implemented to the genotypic characterization panel, to support the cell line development process during the screening of the best clones on Phase I molecules. Moreover, it can easily and feasibly determine the clonality of a cell bank population, by evaluating the conservation of the transgene insertion sites inside the cell genome.

This method is new for Merck and represents a challenging perspective to implement quality control tests' panel and replace Sanger sequencing with gold-standard technologies.

4. PLASMID STORAGE DATABASE

4.1 INTRODUCTION

The production of biotechnological molecules exploits the potential of cell-based mechanisms to produce a big quantity of totally functional human proteins. This is possible thanks to the high replication level of mammalian and bacterial cells in bioreactors and their capacity to properly mediate protein folding and assembly.

Cell growth and protein production rate are the two factors that highly impact recombinant drug production and are the main parameters that biotechnological pharma are trying to standardize and optimize to reduce time and costs during the drug development.

Many factors can impact recombinant gene expression: the vector in which the transgene sequence is inserted; all the regulatory regions inside the expression vector; the insertion site of the transgene within the genome of the cells. Genetic engineering represents the step forward to standardize and optimize all those critical factors that currently are still not totally characterized.

Merck expression vectors used to transfect cell banks are properly designed to contain all the regulatory regions needed for transgene transcription. Generally, plasmids are standardized and contain intronic regions that interspace the exonic ones from the transcription start site. Except for the known regulatory elements that flank the coding sequence (TATA box and POLYA region) which are always checked for their genotypic stability, all the intronic sequences, restriction sites and accessory regions inside the expression vectors, remain still unchecked and not totally characterized after the rump up of the production. This is the reason why, the whole expression vector sequencing through the massive NGS technology could represent a starting point for the generation of a database containing the reference sequences of all Merck molecules. The complete vectors nucleic acid informations will be useful for many Merck functions, from the cell line development department, to the production and the quality control activities.

4.2 MATERIALS AND METHODS

Plasmids quantification

Four plasmids expressing the transgene of four different Merck molecules (molecules' names not shown), were chosen to be sequenced for the Plasmid Database Storage project. At Ivrea Merck site, expression vectors are used as positive controls for QC testing and are stocked at -80°C.

In this context, an aliquot of each plasmid was withdrawn and quantified at the Nanodrop spectrophotometer. Two subsequent quantifications were done and the average between the two measurements was considered to dilute in molecular biology water each sample, to reach the final concentration of 1 ng/µl. This quantity was used as starting input DNA for libraries preparation.

Library preparation

Libraries were generated using Nextera XT DNA Library Prep Kit (Illumina), following the manufacturer's instruction. 5 ng of each plasmid were used as input DNA. The first step of the Nextera XT provides the DNA fragmentation by bead-linked transposomes and the tagmentation of fragments with sequencing adapters.
After an incubation of 5 minutes at 55°C, the tagmented DNA was amplified by limited-cycle PCR program. Contextually, index adapters were added on both ends of the fragments. At the end of the protocol, samples were cleaned up using the 1X AMPure XP beads (Beckman Coulter) according to the manufacturer's instructions.
Purified libraries were quantified by Qubit 3.0 Life Technologies) and sized using the Agilent 2100 Bioanalyzer. Libraries molar calculation was performed using validated spreadsheet. In order to be sequenced, the mean of the fragments size had to be higher than or equal to 200 bp and the molarity had to be greater than or equal to 2 nM. NGS libraries where then diluted to 2nM and prepared for sequencing.

Library sequencing

The sequencing was performed on an Illumina MiSeq sequencer generating paired-end 2 x 150 bp reads. Given the relative amount of data to be generated during the sequencing, a 300-cycle (MiSeq Reagent Kit V2) was used, allowing a maximum of generated reads per run up to 15 million and 0.3 Gb of output data. All the data obtained by the instrument were collected in a server and subsequently analysed with a custom bioinformatic pipeline.

Bioinformatic analysis

Data generated from the sequencing, were analysed to check the presence of mutations (intended as SNPs, insertion, deletion or heterozygosis) compared to the expected plasmid reference sequence provided by the sponsor site. The bioinformatic analysis was performed using a custom pipeline. It allows the comparison between a reference sequence and the samples analysed to identify the presence of mutations. It performs an alignment of all reads generated by the MiSeq instrument (Illumina), based on the similarity between the red sequence and the reference one. Successively, the pipeline qualitatively counts the coverage of each nucleic acid and establishes the results acceptance, based on the cut-off values established as 1500x coverage. The analysis can be performed both on the entire sequence of the generated samples, or focusing on specific regions of interest (flanking, coding regions) by the definition of a BED file, in which the coordinates of those positions are defined. At the end of the analysis, the pipeline generates a report that contains the information relative to potential mutations: the position of the mutation; the coverage of the mutation, intended as the number of reads that identify the region; the reference, intended as the expected nucleic acid in that position and the percentage of appearance of each of the four nucleic acids. Moreover, the pipeline generates a plot describing the coverage and the cut-off values along the entire reference sequence.

4.3 RESULTS AND DISCUSSION

Four plasmids expressing biotechnological Merck molecules, were processed to obtain a high throughput sequencing of their whole sequences. Samples were processed as mentioned in the below paragraphs, starting from their pure quantification and proceeding with library preparation. PCR amplification step was not performed to avoid PCR correlated artifacts and to obtain a whole vector sequencing, including both coding and non-coding regions. The objective of this first part of the project was to setup and identify the optimal DNA input concentration to obtain a good vector coverage, and to store the new data obtained from the NGS analysis for future purposes.

After the sequencing on the MiSeq instrument, the data analysis was performed using the custom pipeline. The generated reads were aligned to the reference sequence of each vector, provided by the manufacturing site. Moreover, a BED file was generated, to only

tag on the exonic regions of the expression vectors and here focusing the bioinformatic analysis to identify in those critical regions the presence of mutations.
Here reported, the obtained output results for each of the four molecules, generated from the pipeline.

RESULTS:

Site	position	coverage	reference	A	T	C	G	insertion	deletion
"No Mutation Detected"									

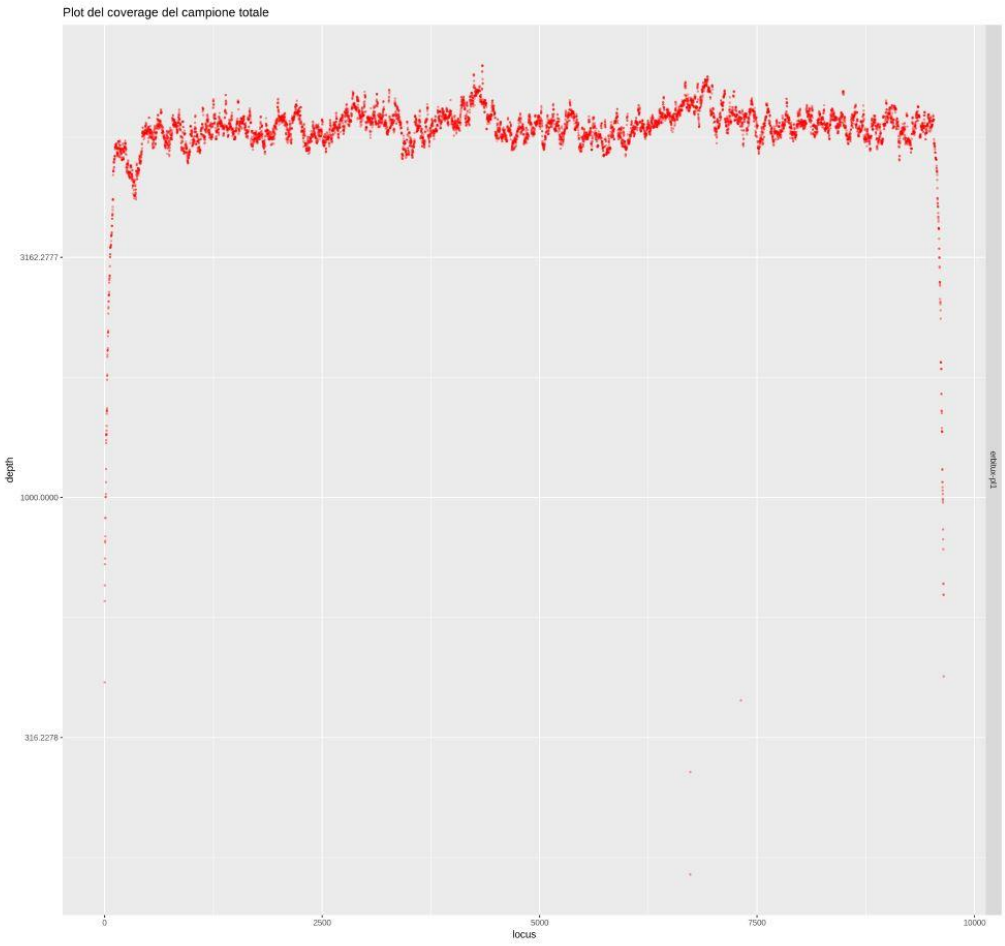


Figure 5. Plot results of molecule 1. A bioinformatic analysis was performed considering the entire vector's length. No mutation was detected by the pipeline, with a plot description of the sequence coverage along the entire vector's nucleic acids coordinates. In this case a second analysis was performed inputting a BED file with transgene's coding sequence coordinates only. Also in this second case, no mutation was detectable (data not shown).

RESULTS:

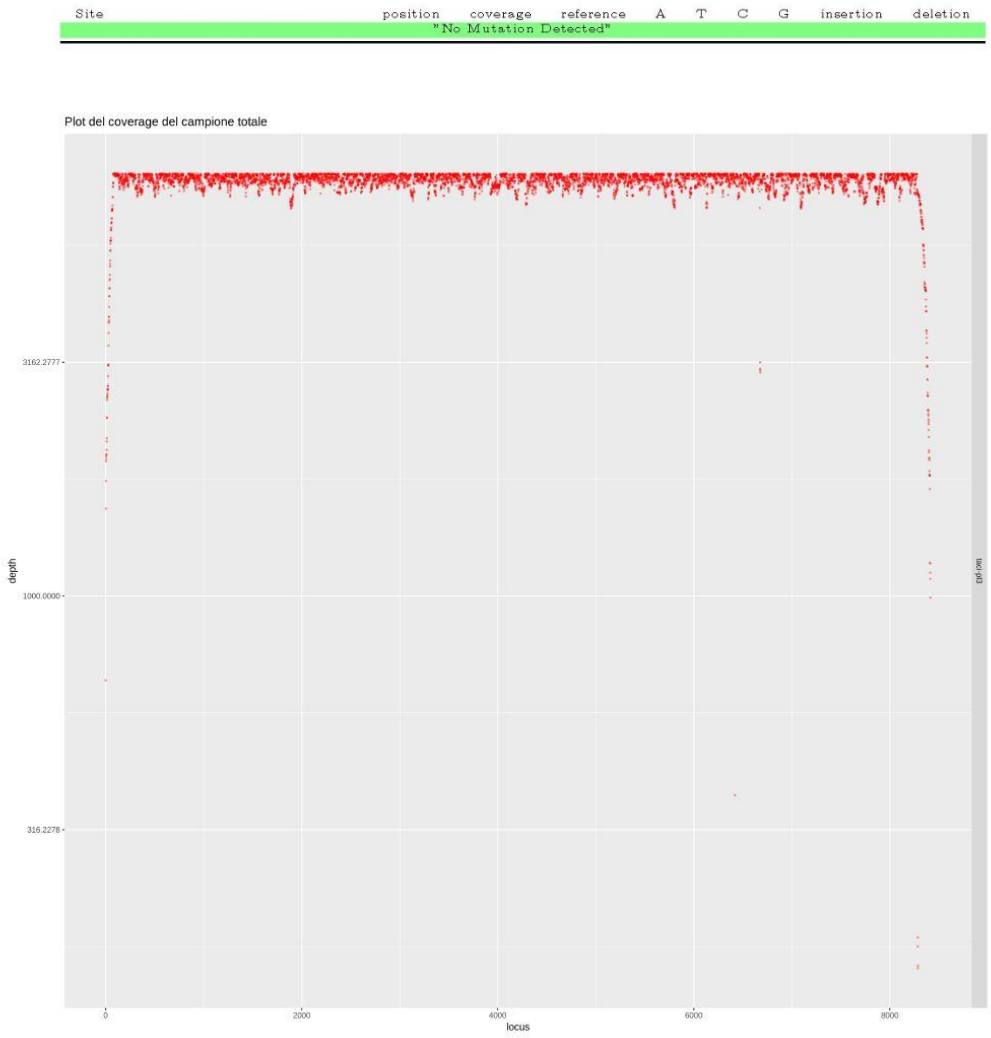


Figure 6. Plot results of molecule 2. A bioinformatic analysis was performed considering the entire vector's length. No mutation was detected by the pipeline, with a plot description of the sequence coverage along the entire vector's nucleic acids coordinates. In this case a second analysis was

performed inputting a BED file with the transgene's coding sequence coordinates only. Also in this second case, no mutation was detectable (data not shown).

RESULTS:

Site	position	coverage	reference	A	T	C	G	insertion	deletion
"No Mutation Detected"									

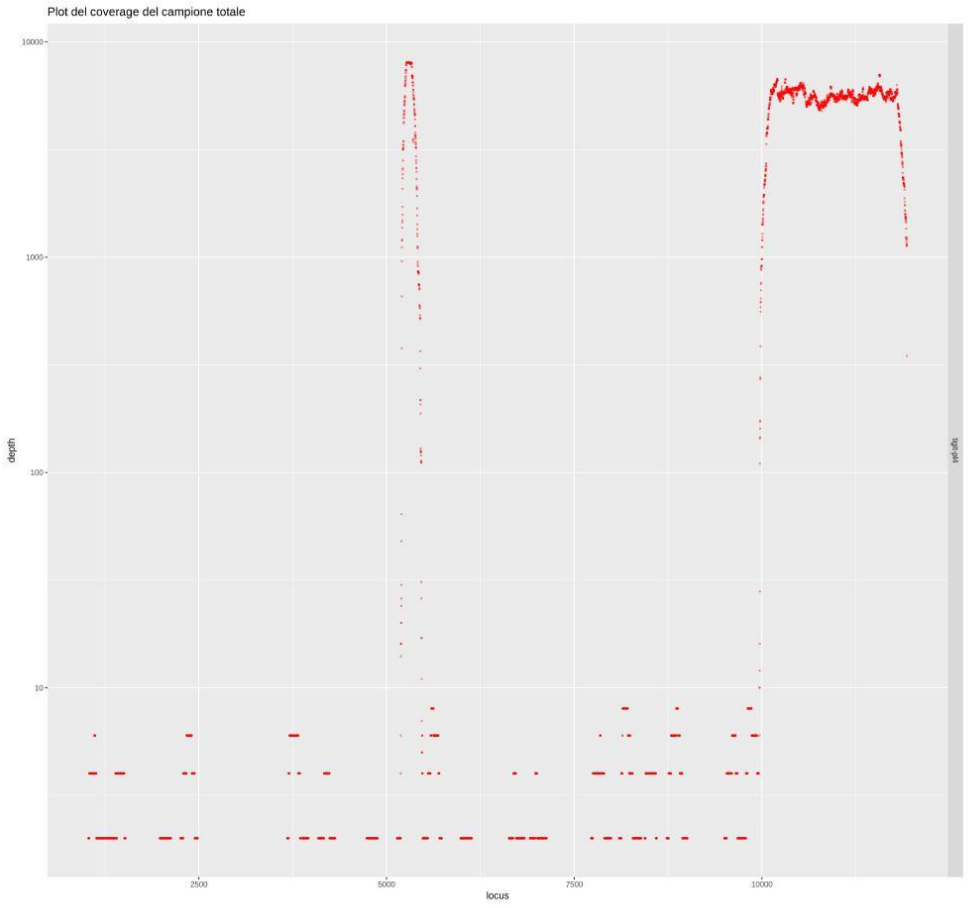


Figure 7. Plot results of molecule 3. A bioinformatic analysis was performed considering the entire vector's length. No mutation was detected by the pipeline, with a plot description of the sequence coverage along the entire vector's nucleic acids coordinates. In this case, the whole vector coverage is not homogeneous, as in some regions, a low number of reads aligned to the reference sequence. Also in this case, a second analysis was performed inputting a BED file with the transgene's coding sequence coordinates only. Also in this second case, no mutation was detectable (data not shown).

RESULTS:

Site	position	coverage	reference	A	T	C	G	insertion	deletion
"No Mutation Detected"									

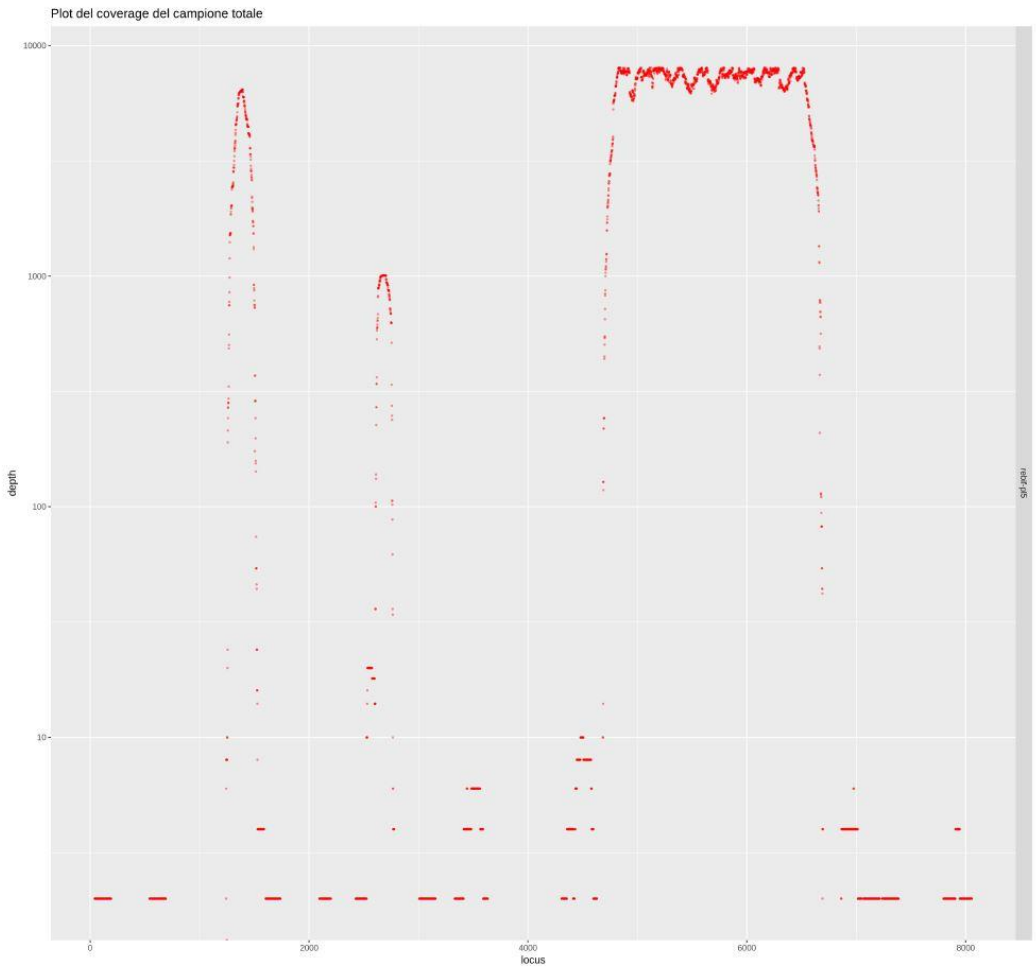
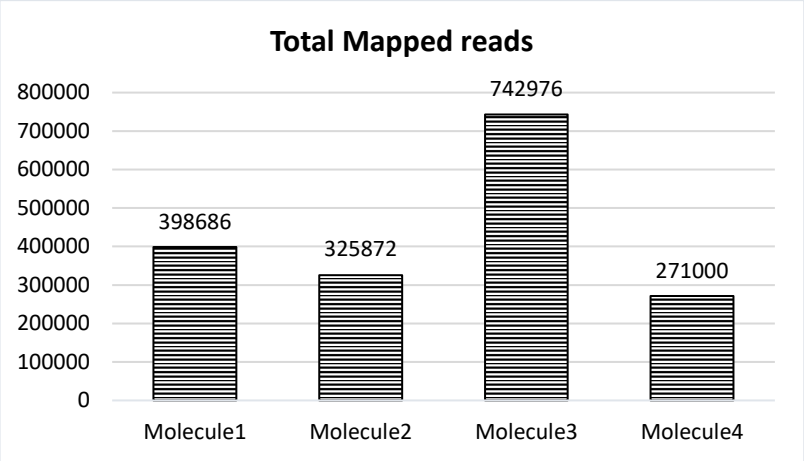


Figure 8. Plot results of molecule 4. A bioinformatic analysis was performed considering the entire vector's length. No mutation was detected by the pipeline, with a plot description of the sequence coverage along the entire vector's nucleic acids coordinates. In this case, the whole vector coverage

is not



homogeneous, as in some regions, a low number of reads aligned to the reference sequence. Also in this case, a second analysis was performed inputting a BED file with the transgene's coding sequence coordinates only. Also in this second case, no mutation was detectable (data not shown).

Figure 9. Total mapped reads table. In this graph, the total number of reads aligned to each molecule reference sequence are described. Molecule 1,2 and 4 are characterized by a comparable number of reads generated during the sequencing. Molecule 3, has a higher number of total reads but, as depicted in figure, their distribution is not uniform along vector sequence (as for Molecule 4).

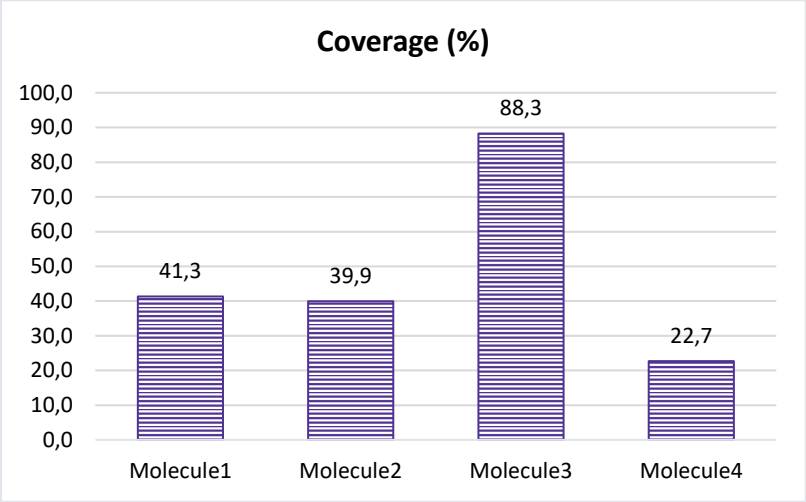


Figure 10. Percentage of total coverage. In this graphic, the total coverage for each molecule, expressed in percentage, is shown. Molecule 1 and 2 have a comparable percentage of coverage along their entire sequence (41,3% and 39,9%), while molecule 3 and 4 have respectively the higher (88,3%) and the lower (22,7%) percentage of coverage respect to the others samples.

As described in the plots relative to the bioinformatic analysis, the sequencing of Molecule 1 and 2 produced a similar number of reads (**Figure 9**), which are equally distributed along the entire vector sequence (**Figure 5 and 6**). Molecule 3 is characterized by a higher coverage (**Figure 9**), but the generated reads are not homogeneously distributed along its sequence (**Figure 7**). Molecule 4, otherwise, is characterized by the lowest number of total reads (**Figure 9**), but similarly to Molecule 3, ha a heterogeneous coverage along vector sequence (**Figure 8**). The percentage of coverage of each molecule is proportional with the total reads and is depicted in **Figure 10**. The coverage enrichment in some regions is probably due to the certified presence, both in molecule 3 and in molecule 4, of many intronic regions that regularly interspace the exonic ones. Those not-transcribed sequences, are sometimes constituted by repetitive elements, which are not always easily accessible and probably are not efficiently amplified by Taq polymerase during libraries amplification. Also, the insertion of labelled terminator nucleotides during the

sequencing itself, and the subsequent fluorescence reading by the instrument, is not always flawless when the sample is enriched by repetitive sequences. All those factors can feasibly be the cause of the irregular coverage along molecule 3 and 4. Nevertheless, the transgenes coding sequences of all the four molecules were properly covered, and considering the average number of reads per sample, a coverage-cutoff value was set at 1500 x reads. The bioinformatic analysis was performed aligning samples reads to the proper vector sequence, provided by the manufacturing site. Each analysis was performed imputing a set of coordinates (BED), restricted to the transgene coding sequence. In that way, the base calling realized by the custom bioinformatic pipeline, was focused only on the regions of interest, excluding those fragments along the vector that were not properly covered. The reads belonging to each sample, were automatically aligned to their reference sequences, and the pipeline was costume to report mutations, when a nucleic acid substitution was detected, comparing samples reads with their reference sequence. During previous setup experiments on this method conducted at Merck Ivrea site, a limit of detection was already established at 4% (data not shown). Therefore, also for these experiments, the same cutoff value was exploited, meaning that only the mutations appearing with a frequency $\geq$ of 4% were considered statistically significant and stated by the pipeline. In all the four samples, no statistically significant mutations were detected on transgenes when compared with their expected reference sequences, providing feasible results about the conserved genomic sequences of vectors used to transfect recombinant cell lines. This new approach for vector sequencing could represent a promising aspect that can be useful both for the manufacturing sites, and for the quality control departments. In the first case, the statement of the entire plasmid sequence, could be important to be aware of the real starting vector sequence used to produce Merck recombinant molecules. In the second case, the use of the plasmid as positive control for the genomic stability evaluation of recombinant molecules, is already widely used. The introduction of a direct plasmid sequencing that surrounds the amplification and the subsequent purification steps, could represent a promising upgrade to the target sequencing methodology, avoiding artefacts due to the numerous assay steps.

5. DEVELOPMENT AND VALIDATION OF A METHOD FOR BACTERIAL CELL BANK GENETIC STABILITY EVALUATION USING NEXT GENERATION SEQUENCING

5.1 INTRODUCTION

The production of biotechnological drugs exploits cellular proper replicational and translational mechanisms to express recombinant proteins. Most Merck biotechnological products are developed to be produced in mammalian cell lines. The massive use of mammalian cell lines is basically due to guarantee all the correct post-translational modifications needed to obtain a totally functional recombinant protein and administer them to patients. Nevertheless, mammalian cell lines are not the only one used to produce biotechnological drugs. In fact, also bacterial cell lines are exploited for this purpose, because of their exponential capacity of easily replicate and grow.

Bacteria lack a membrane-bound nucleus and other internal structures and are therefore ranked among the unicellular life-forms called prokaryotes. Prokaryotic cells (i.e., Bacteria and Archaea) are fundamentally different from the eukaryotic cells, that constitute other forms of life, as defined by a much simpler design . The most-apparent simplification is the lack of intracellular organelles, which are features characteristic of eukaryotic cells. In addition, prokaryotic cells are usually much smaller than eukaryotic cells. The small size, simple design, and broad metabolic capabilities of bacteria, allow them to grow and divide very rapidly and to inhabit and flourish in almost any environment. Bacteria have one circular chromosome that contains all of their genetic information, and their mRNAs are exact copies of their genes and are not modified.

Bacteria, can be forced to produce a foreign protein, by a mechanism called transformation. Transformation is the process by which foreign DNA (plasmid) is introduced into a cell. Transformation of bacteria is important not only for studies in bacteria but also because these prokaryotes cells are used for both storing and replicating plasmids. For this reason, nearly all plasmids (even those designed for mammalian cell expression) carry both a bacterial origin of replication and an antibiotic resistance gene for use as a selectable marker in bacteria. Scientists have made many genetic

modifications to create bacterial strains that can be more easily transformed and that will help to maintain the plasmid without rearrangement of the DNA. Additionally, specific treatments have been discovered that increase the transformation efficiency and make bacteria more susceptible to either chemical or electrical based transformation, generating what are commonly referred to as 'competent cells'. In Merck company, E. coli bacterial cell lines are mostly used both to propagate expression vectors containing recombinant genes, and afterwards to produce biotechnological drugs. As biological systems, bacteria must undergo to all the quality control analysis to verify that no modification occurred during the production phases.

For what concerns cell line characterization tests, the recombinant gene sequencing is performed in Merck Ivrea, to guarantee the genetic stability of the nucleic acid sequence that is translated into the recombinant protein. Currently, the validated method approved by regulatory authorities, is based on Sanger technology but the introduction to next generation sequencing based methods is now being considered.

The setup and validation of the sequencing method based on NGS, was evaluated through the analysis of two parameters, established on what reported in ICH Q2 guidelines, considering that the genotypic characterization based on transgenes sequencing, is considered a *limiting essay for impurities detection*. The two parameters that must be evaluated for a future validation of a method are: the specificity and the limit of detection (LOD) of the method. The specificity is intended as the capability of the method to recognize the searched impurity (mutation) inside a heterogeneous sample (with more than one mutation). This parameter has the scope of demonstrate that the experimental procedures and the experimental workflow of the test do not interfere with the capability of the method based on NGS to detect any type of mutation. The other parameter analysed is the LOD, which represents the lowest amount of impurity detectable by the method. The characterization of this value allows to determine what is the minimum percentage of a mutation detectable inside a heterogeneous sample. Both parameters are fundamental to determine the feasibility of the new method.

To evaluate the specificity and the limit of detection of the sequencing method based on NGS, two different approaches were used, designed to simulate the production phases of new molecules using bacterial cell banks. As mentioned for the previous topic, given the

confidentiality of the data shown, molecules and expression plasmids name, transgene sequences and all the additional sensitive information are not shown or specified, to protect Merck intellectual property.

5.2 MATERIALS AND METHODS

Recombinant vectors design

For the assay development, the wild type human gene of INF gamma was selected by extrapolating its nucleic acid coding sequence from NCBI database (NCBI Reference Sequence: NG_015840.1). Then, three different mutations were introduced inside the coding sequence: a single nucleotide polymorphism (SNP) (position 255, T>C) ; a three base deletion (DEL) (position 133-135, - GTA); and one bases insertion (INS) (position 376, T>C). Eurofins genomics company provided the certified wild type (WT) and mutated forms of INF gamma gene, inserted into a standard vector specific for bacteria transformation (plasmid backbone: pEX-K168). Transgene sequence was not optimized for protein production in bacterial cells, as this would have meant base changes for functional protein translation. A kanamycin resistance sequence was also inserted inside the plasmid to select the bacteria population that have correctly integrated the transgene.

Mix preparation

Recombinant plasmids were mixed to create heterogeneous samples, each containing a different percentage of mutation. In the first experiment, after an initial quantification with Qubit fluorometer (Thermo Fisher), each recombinant plasmid was diluted at an initial concentration of 1 ng/µl, by resuspending it in TE1X buffer. After that, pure plasmids were mixed to obtain heterogeneous samples, each containing a different percentage of INF gamma mutated form. Here below, a representative table of mix concentration.

Plasmids name	Mix1 (100%)	Mix2 (25%)	Mix3 (10%)	Mix4 (5%)	Mix5 (1%)
IFNg WT	100 µl	25 µl	70 µl	85 µl	97 µl

IFNg SNP	/	25 µl	10 µl	5 µl	1 µl
IFNg DEL	/	25 µl	10 µl	5 µl	1 µl
IFNg INS	/	25 µl	10 µl	5 µl	1 µl

Table 1. Heterogeneous mix production. Plasmids were brought to the same initial concentration of 1 ng/µl and then mixed to create standard heterogeneous samples in which the mutations are present in a known percentage. Mix1: contains the WT recombinant plasmid in a percentage of 100%. Mix2: contains each recombinant plasmid in a percentage corresponding to 25%. Mix3: contains the mutated recombinant plasmids in a percentage of 10% each, with the left 70% of the wild type form. Mix4: contains the mutated recombinant plasmid in a percentage of 5% each, with the left 85% of the wild type form. Mix5: contains the mutated recombinant plasmids in a percentage of 1% each, with the left 97% of the wild type form.

E. Coli transformation

In the second experiment the recombinant vectors were used to transform E. coli BL21 competent cells, that were then plated on an LB-agarose petri with kanamycin antibody. Bacteria were grown for 24 h and after that three colonies were selected for each transgene (WT and mutated). Those transformed bacteria were then propagated overnight in LB broth with kanamycin. The day after density and turbidity of bacteria solution were checked and standardized to bring each transgenic bank at the same concentration. At this point transformed bacteria were mixed together in different percentage, to create heterogeneous samples containing the WT and mutated forms of INF gamma.
Here are reported the mix concentration:

Plasmids name	**Mix1 (100%)**	**Mix2 (25%)**	**Mix3 (10%)**	**Mix4 (5%)**	**Mix5 (1%)**
INFg WT	100 µl	25 µl	70 µl	85 µl	97 µl
INFg snp	/	25 µl	10 µl	5 µl	1 µl
INFg del	/	25 µl	10 µl	5 µl	1 µl
INFg ins	/	25 µl	10 µl	5 µl	1 µl

Table 2. Heterogeneous mix production. Transformed bacteria were brought to the same initial concentration and then mixed to create standard heterogeneous samples in which the mutations are present in a known percentage. Mix1: contains the bacteria transformed with the WT recombinant plasmid, in a percentage of 100%. Mix2: contains bacteria transformed with each recombinant plasmid in a percentage corresponding to 25%. Mix3: contains bacteria transformed with the mutated recombinant plasmids in a percentage of 10% each, with the left 70% of the wild type form. Mix4: contains bacteria transformed with the mutated recombinant plasmid in a percentage of 5% each, with the left 85% of the wild type form. Mix5: contains bacteria transformed with the mutated recombinant plasmids in a percentage of 1% each, with the left 97% of the wild type form.

Library preparation and sequencing

For both experiments, plasmid extraction was performed using the Plasmid mini kit column based system (Qiagen).

The obtained eluted DNA was quantified by Qubit fluorometer. 5 ng of input DNA was used to prepare a library for the sequencing with the Nextera DNA flex kit provided by Illumina. An already validated protocol was followed and the final concentration of each mix was quantified by Qubit and Bioanalyser (Agilent). A Nano V2- 300 cycles cartridge was used for the sequencing on MiSeq machine provided by Illumina and data were analysed using an in-house bioinformatic pipeline to detect all the expected mutations and their percentage of appearance.

5.3 RESULTS AND DISCUSSION

Four expression vectors were engineered to express the WT and three mutated form of the INF gamma gene. The recombinant plasmids where used to transform E. Coli competent cells, which were selected for their conferred resistance to kanamycin antibiotic. The recombinant bacteria were then mixed in well-known percentages before the DNA extraction and the subsequent library preparation and sequencing.

The sequenced samples were analysed with a costume bioinformatic pipeline (the same depicted before), to evaluate firstly the presence of all the different mutations and their frequency of appearance. This was basically done to evaluate the specificity and the limit

of detection of the sequencing method, in order to exploit it for future quality control analysis.

Here reported, the obtained results for Mix 2 and 3, containing respectively 25% and 10% of mutations.

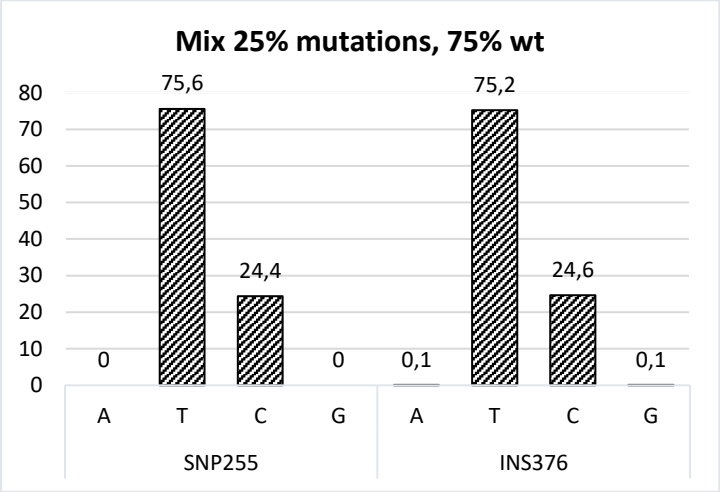


Figure 11. The table describes the frequency of mutations appearance in the Mix 25% SNP and 25% INS. The graph shows the percentage of appearance of the single nucleotide polymorphism in position 255, and the base insertion in position 376. The percentage of appearance of the mutated nucleotide corresponds to the expected one (≈ 25%).

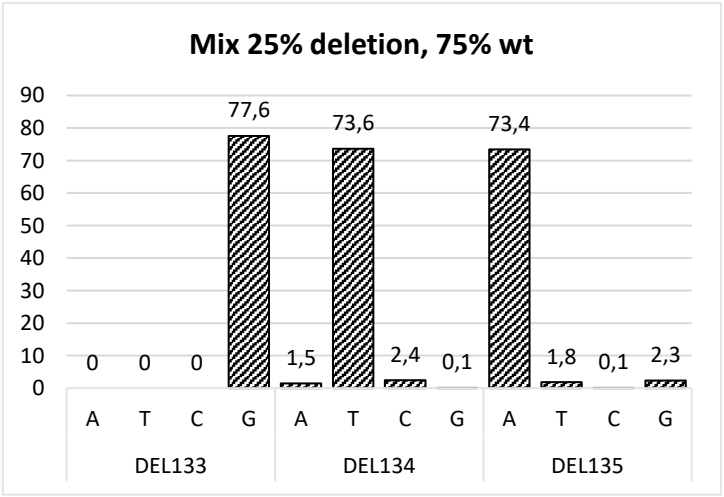


Figure 12. The table describes the frequency of mutations appearance in the Mix 25% DEL. The graph shows the percentage of appearance of the nucleotides deletion in position 133-134 and 135. The percentage of appearance of the deleted nucleotides correspond to the expected one (≈ 25%).

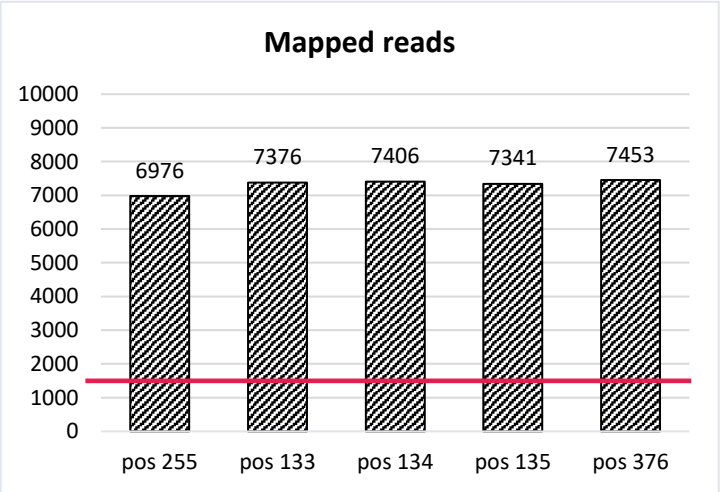


Figure 13. Graphic of the total mapped reads on each investigated position. The table describes the total number of reads covering the positions in which the mutation (255), the deletion (133-135) and the insertion (376) are expected.

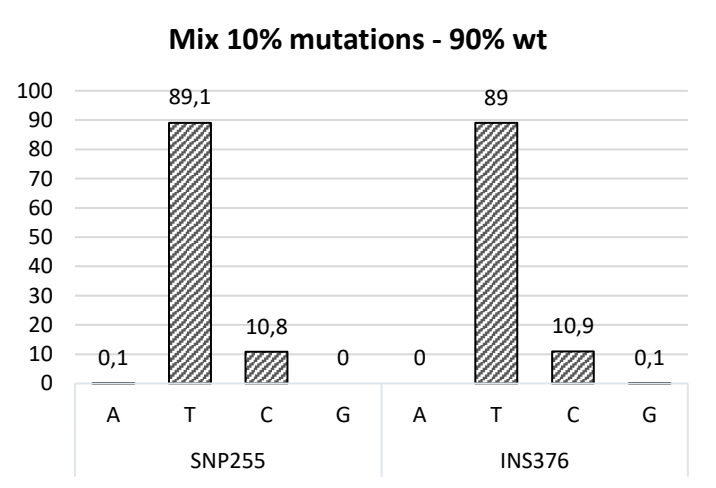


Figure 14. The table describes the frequency of mutations appearance in the Mix 10% SNP and 10% INS. The graph shows the percentage of appearance of the single nucleotide polymorphism in

position 255, and the base insertion in position 376. The percentage of appearance of the mutated nucleotide corresponds to the expected one (≈ 10%).

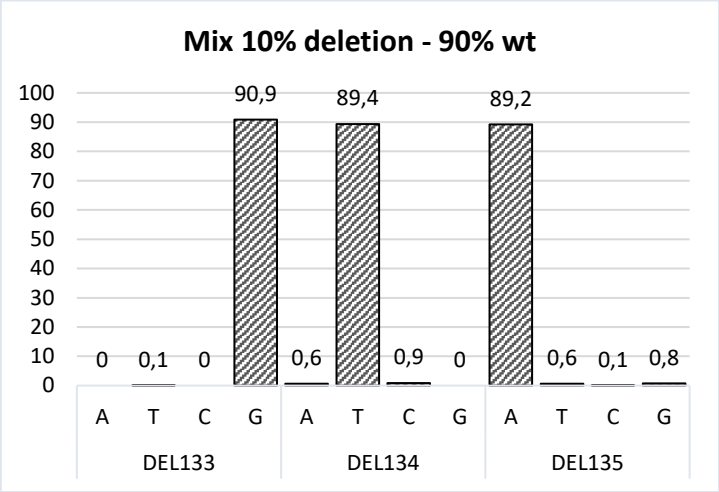


Figure 15. The table describes the frequency of mutations appearance in the Mix 10% DEL. The graph shows the percentage of appearance of the nucleotides deletion in position 133-134 and 135. The percentage of appearance of the deleted nucleotides correspond to the expected one (≈ 10%).

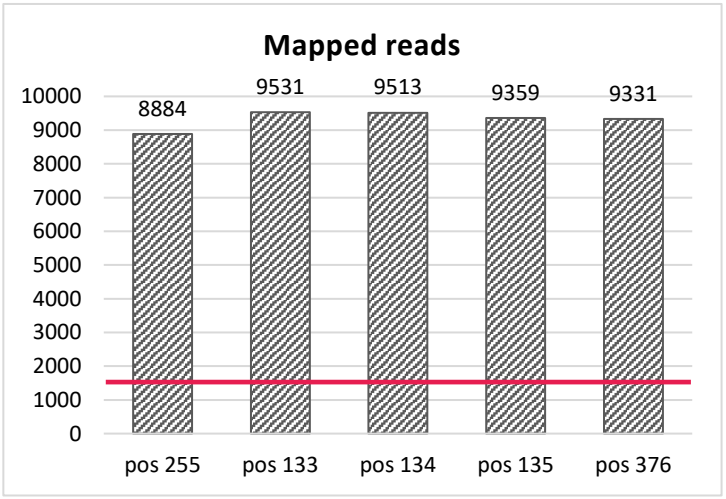


Figure 16. Graphic of the total mapped reads on each investigated position. The table describes the total number of reads covering the positions in which the mutation (255), the deletion (133-135) and the insertion (376) are expected (≈ 10%).

In the second part of the manuscript, the specificity and the limit of detection of a next generation sequencing method for bacterial genetic stability evaluation, was investigated. Firstly, four expression vectors containing different mutated forms of the human INF gamma gene were designed and produced. At the beginning pure plasmids were mixed in well-known percentages (**Table 1**) and sequenced to verify that the certified mutations were presents and detectable by the investigated method (data not shown). After the mutations check, E. Coli competent cells were transformed with each mutated plasmid and then mixed in well-known percentages. Plasmids were extracted, libraries were prepared, and the sequencing was performed. After that a bioinformatic analysis was conducted to verify the frequency of appearance of each expected mutation. The analysis of Mix 1 (corresponding to the 100% of INF gamma WT form) confirmed the expected results. In fact, in Mix 1 the absence of any mutation was demonstrated (Data not shown). Mix 2 and 4 were likewise analysed. **Figure 11** depicts the frequency of appearance of the SNP in position 255 and the INS in position 376 in Mix 2 (25% of mutations). In both cases, mutations are specifically identified by the bioinformatic pipeline and detected at the expected frequency. **Figure 12** shows the frequency of appearance of the three bases deletion, and also in this case the base calling frequency corresponds to the expected ones. The same concordant output resulted after the analysis of Mix 3 (10% of mutations) (**Figure 14**, **Figure 15** and **Figure 16**) but were no more visible in Mix 4 and 5 (containing the 5% and 1% of mutations, respectively). These analyses confirmed the specificity of the new method based on NGS and set its LOD at 10%.

The setup of an NGS based method for the genotypic stability evaluation of bacterial cell banks used to produce biotechnological molecules, represents an innovative assay to substitute the currently validated Sanger method. First, it exploits bacterial cells simplicity, to easily extract recombinant plasmids which do not integrate inside host cells genomes. Secondly, it exploits the hight throughput of NGS technology, combined with the limited dimensions of expression vector, to avoid PCR amplification in correspondence of the transgene coding sequence, promoting the potential whole vector sequencing, and escaping experimental artefacts. During the setup of this new method, its specificity was demonstrated, employing certified expression vectors containing ad hoc mutated forms of the human INF gamma gene. The limit of detection, was evaluated

through the analysis of heterogeneous mix, containing well-known percentages of each mutated E. Coli and the final LOD value was set at 10%. Overall, the setup experiments conducted in this contest, can serve as foundation to validate a new innovative method for the genotypic stability evaluation of cell banks used to produce Merck recombinant molecules. It is certainty more reliable and cost-effective respect to Sanger based methods, allowing the release of safer and solid results on the nucleic acids sequences of therapeutical molecules.

6 CELL BANK CLONALITY ASSESSMENT COMBINING TLA AND SANGER TECHNOLOGIES

6.1 INTRODUCTION

As already explained in this manuscript, biotechnological drug production requires the massive use of biological systems because of their capacity to mediate proper folding and assembly of totally functional recombinant proteins. Working with biological systems, like mammalian and microbials cell banks, means that both the transgene integration and expression can highly vary in function of many factors. The most relevant one is represented by the integration site of the expression vector around cell genome. In fact, chromatin differential methylation and nucleic acid sequences composition, influence gene expression and for that reason the transgene integration site affects its final translation into protein [12]. Moreover, as the integration happens randomly during cell transfection, different fusion sites can be identified between treated cells, yielding to heterogeneous populations of recombinant cells [13]. Differential transgene integration sites can influence the clonality of the cell lines used to produce the biotechnological products, thus making them not compliant with the ICH requirements. For that reason, is becoming increasingly important to verify, through new analytical methodologies, the clonality of productive cell banks, to respond to regulatory requirements and to release safe products to patients [11], [14].

Currently, to overcome the possibility of producing proteins deriving from cells with potential genome differences, serial dilutions of transfected cells are performed to achieve the concentration of 0.5 cell/well [15]. Among the propagated cells, the best clone selection process is performed, to finally select the optimal Master Cell Bank.

Thanks to the new sequencing technologies' advancement, it is now possible to identify vector insertion sites around cell genome, allowing the understanding of chromatin conformation of the region of interest and confirming if the integration site is common to all clones deriving from the same MCB [16]. In that context, the aim of this second part of the manuscript will focus on the assessment of the clonality of the MCB used to produce a Phase I Merck molecule. Also in this third manuscript topic, given the confidentiality of the data shown, molecules and expression plasmids name, transgene sequences and all the additional sensitive information are not shown or specified, to protect Merck intellectual property.

The molecule investigated in this manuscript is a therapeutical fusion protein, comprised of three polypeptide chains produced using the NS0-LD cell line. The expression cell line was generated by sequentially transfecting and selecting stable clones using two expression plasmids: plasmid 1 (expressing Light Chain- LC and Heavy Chain- HC) and plasmid 2 (expressing an additional construct to finalize the fusion protein). This cell line was cloned using a single round of limiting dilution to generate the Master Cell Bank.
In this study, the clonality of the MCB was assessed using a combination of multiple technologies. First, the MCB vector integration sites were identified using the Targeted Locus Amplification (TLA) technology combined with NGS. TLA offers the possibility to selectively amplify, with a single primer pair, tens to hundreds of kilobases of neighbouring sequences of a locus of interest [17], allowing the detection and characterization of vector integration sites. Second, all the sequences depicting the fusion events between the host genome and the vector are verified with PCR and Sanger sequencing of the MCB. Finally, PCR is used to verify the presence of all the identified integration sites in 30 subclones derived from the MCB.
The expected result for a clonal cell line was the identification of the same vector integration sites in the parental and derived clones. The clonality of the MCB was therefore confirmed if the plasmid integration sites were detected in all the derived subclones.

6.2 MATERIALS & METHOD

Sample Preparation

A frozen vial of the MCB was thawed to generate the cells necessary for the study. To this purpose, cells were propagated in DMEM F12 - HAM (Sigma-Aldrich) with 10% Fetal Bovine Serum (FBS, Hyclone) and 10% Penicillin/Streptomycin (Sigma-Aldrich), then collected to get samples for TLA analysis, PCR protocol, and Sanger sequencing. Since the TLA protocol kit requires 3 to 5 x 10^6 viable cells as input (Cergentis), cell aliquots containing 5 x 10^6 cells were prepared following manufacturer's recommendations. Starting from fresh cultured cells at first passage, three replicates were produced and used for TLA analysis. Viable cells were counted using a Nucleocounter (Chemometec) system as the difference between the total number of the cells and the non-viable cells. Cells were centrifuged 10 minutes at 250g, the supernatant discarded, and the cell pellet was resuspended in a pre-cold freezing buffer (culture medium mixed with 10% DMSO).

TLA Template Preparation

The TLA kit includes all the reagents necessary for the TLA template generation, except for the culture medium (DMEM F12-HAM), the molecular biology grade water, the 2-propanol and the ethanol. A schematic representation of the overall TLA protocol is shown in **Figure 17**.

To remove debris and spent media components, samples were centrifugated and the supernatant was discarded. Cell pellets were then washed resuspending in DMEM F12-HAM medium, centrifuging (2 minutes at 250 g), and discarding the supernatant. After these initial washing steps, cells were fixed resuspending in culture medium, with addition of Fixation Buffer (FB). After an incubation, fixation was stopped adding the Quenching Buffer (QB). Cells were then centrifuged, and the pellets were washed resuspending in Restriction Buffer (RB). Cells pellets were then permeabilized adding the RB and the Permeabilization Buffer (PB) and incubating 15 minutes at 65°C, while shaking. Then, permeabilization was stopped adding the Neutralization Buffer (NB) and incubating 30 minutes at 37°C. First round of DNA restriction was conducted adding the Restriction Enzyme 1 (RE1, *NlaIII*) and incubation overnight at 37°C with shaking.

The next day, RE1 was inactivated. A ligation was then conducted adding a 10X Ligation Buffer (LB), and a Ligase (LIG), and incubating for 6 hours at 20°C. Then, to allow DNA

de-crosslinking, proteinase k was added, and samples were incubated overnight at 65°C, shaking. DNA was then purified with paramagnetic isolation. After DNA elution form beads, the purified nucleic acids were quantified with the Qubit fluorometer system and stored at -20°C.

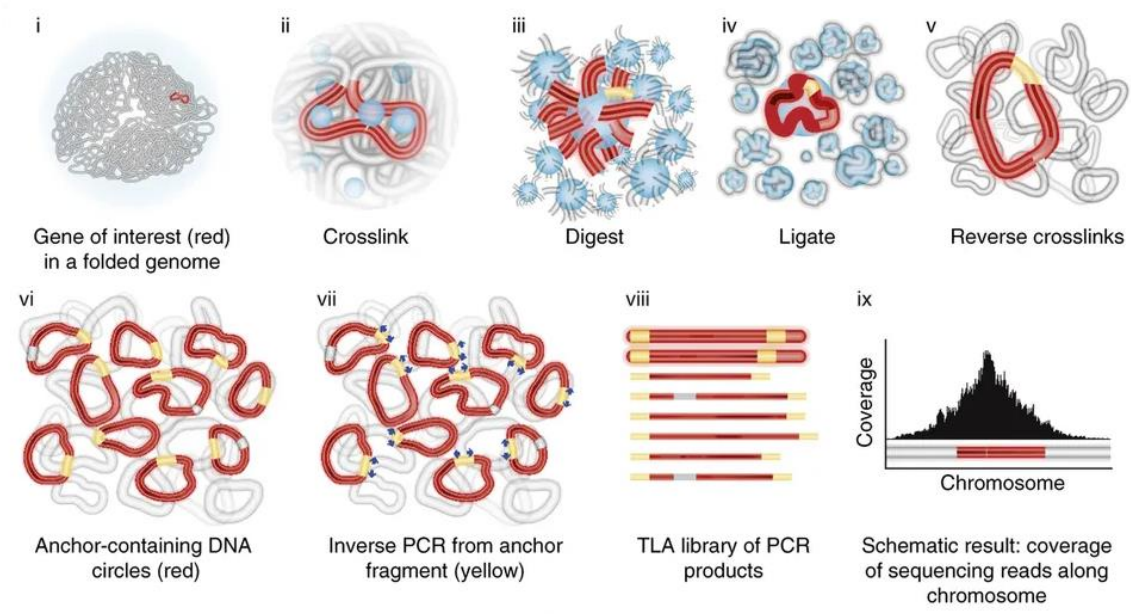


Figure 17. Overview of the steps included in the TLA protocol (de Vree PJ et al., 2014). i) Initial status: neighboring sequences that form a gene or genetic locus (red) are in close spatial proximity; ii) fixation to cross-link DNA sequences that are in close proximity. iii) DNA digestion with a frequently cutting restriction enzyme that recognizes a 4-base pair site (NlaIII, or RE1); iv) ligation to obtain large DNA circles containing multiple crosslinked RE1 restriction fragments; v) Different copies of a locus (from different cells) result in DNA circles composed of different combinations of restriction fragments. vi) Limited trimming with a compatible, but less frequently cutting enzyme (NspI, or RE2; it shares a core recognition sequence (CATG) with R1) and then re-ligation, creating PCR-amplifiable DNA circles; vii) Fragments with a region of interest (the anchor sequence, yellow) are selectively PCR-amplified with anchor-specific inverse PCR primers (blue arrows); viii) PCR with a single primer set at the anchor point will result in the amplification of many RE1 fragments across tens to hundreds of kilobases of surrounding DNA, resulting in a sample highly enriched for locus-specific sequences; ix) Amplicon are prepared with standard library preparation methods and sequenced with NGS, resulting in a collection of NGS reads that span tens of kilobases.

A second round of DNA restriction was conducted mixing 10 µg of previously isolated DNA, with 10X RB and Restriction Enzyme 2 (RE2, *NspI*), and incubating at least for one hour at 37°C. At the end of the incubation, the enzyme was inactivated at 65°C for

25 minutes. Samples were then transferred into 5 ml tubes, adding LB, LIG. Samples were mixed via inversion and incubated for at least one hour at room temperature.
The resulting DNA was once again purified with magnetic beads and purified as described before (see above). The resulting DNA, also called TLA template, was quantified with Qubit fluorometer and stored at -20°C.
To assess the quality of the TLA preparation after these initial steps of the protocol, aliquots of the intermediate DNA samples were collected after the steps of cell permeabilization ("Undigested controls"), RE1 digestion ("Digested Controls"), and de-crosslinking ("Ligation Controls") and these samples were run on a 0,5% w/v agarose gel (data not shown).

PCR Amplification and Purification

Among the three generated TLA templates, one of them was chosen to perform a PCR amplification using eight primer pairs designed on the two different expression plasmids (1 and 2), as reported in Table 2.

.

Set Name	Specificity	Primer Name	Primer Sequence	Orientation on vector sequence
Set 1	Plasmid 1 and 2	Plasmid 2_P1_Fw	TATTTACGGTAAACTGCCCA	+
		Plasmid 2_P1_Rv	TTATGTAACGCGGAACTCC	-
Set 2	Plasmid 2	Plasmid 2_P2_Fw	CAATATCACGGGTAGCCAAC	+
		Plasmid 2_P2_Rv	TCGAAATGACCGACCAAG	-
Set 3	Plasmid 2	Plasmid 2_P3_Fw	TATGTCCTGATAGCGGTCC	+
		Plasmid 2_P3_Rv	GCCTTCTTGACGAGTTCT	-
Set 4	Plasmid 1	Plasmid 1_P4_Fw	CTATAGCAAGCTCACCGTG	+
		Plasmid 1_P4_Rv	GTCATCTCCTCCCGTGAT	-
Set 5	Plasmid 1	Plasmid 1_P5_Fw	CGGAGAACAACTACAAGACC	+
		Plasmid 1_P5_Rv	TACAGCGGTCACTCTCAG	-
Set 6	Plasmid 1	Plasmid 1_P6_Fw	GGCCTCCAGAAAGACCTC	+
		Plasmid 1_P6_Rv	CAATGGTAAACAGGCCTCC	-
Set 7	Plasmid 2	Plasmid 2_P7_Fw	ACCATCCAAGTCAAAGAGTT	+
		Plasmid 2_P7_Rv	ACCAATCCAATTCTACGACA	-
Set 8	Plasmid 2	Plasmid 2_P8_Fw	GGACAACAAGGAGTATGAGTAC	+
		Plasmid 2_P8_Rv	TTAGAACCTCGCCTCCTTT	-

Table 3. TLA PCR Primer sets specificity, sequence, and annealing sites.

PCR amplification was conducted according to manufacturer instructions (Cergentis). The kit for TLA included the necessary reagents for the amplification, except for the eight

primers pairs, the molecular biology grade water, the 2-propanol, and the ethanol. PCR amplification was performed with the following program:

1) 98°C for 30 seconds;
2) 98°C for 5 seconds;
3) 55°C for 5 seconds;
4) 72°C for 2 minutes;
   Back to step 2) for 34 cycles;
5) 72°C for 5 seconds;

PCR product was purified with magnetic beads and a paramagnetic purification was completed eluting nucleic acids in TE 1X buffer. To assess the PCR product quality, purified DNA was run on agarose gel (1% w/v, data not shown) and quantified with Qubit fluorometer.

Library Preparation

Libraries were prepared using the Nextera Flex kit (Illumina) according to the manufacture's procedures. 100 ng of DNA of each amplicon was tagmented by adding the tagmentation master mix and incubated at 55°C for 15 minutes. After the completion of the tagmentation reaction, the Tagmentation Stop Buffer was added to each reaction, and samples were incubated at 37°C for 15 minutes in a thermal cycler. Subsequently, samples were placed into a magnetic rack for three minutes to separate the beads from the supernatant. Supernatant was then discarded, and the beads were washed following Illumina's guidelines. Finally, tagmented DNA was amplified by a limited-cycle PCR reaction. Beads were resuspended into 40 µl of PCR master mix and 5 µl of i7 adapters necessary for sequencing and samples barcoding. The PCR was conducted using the following program:

1) 68°C for 3 minutes;
2) 98°C for 3 minutes;
3) 98°C for 45 seconds;
4) 62°C for 30 seconds;
5) 68°C for 2 minutes;

Back to step 3) for 4 cycles;

6) 68°C for 1 minute;

Finally, a double-sided bead purification procedure (Sample Purification Beads, SPBs) was used to clean up the PCR product by removing fragments too short or too long from the final library. Finally, the quality of the library was verified using Agilent 2100 Bioanalyser to calculate library fragment size distribution and the fluorometer Qubit to determine library concentration.

Next Generation Sequencing

Sequencing was carried out on an Illumina MiSeq instrument in paired-end mode (2 X 151 cycles) using a MiSeq V3 Reagent kit 600 cycles (for run number 1) and a MiSeq v2 Reagent kit 300 cycles (for run number 2). For each run, libraries were pooled together after being normalized to the final concentration of 2 nM. To prevent focusing and phasing problems, typical of libraries with low diversity, a 1% control library of genomic DNA from bacteriophage PhiX was added to the pools to increase the genetic diversity of the library.

Bioinformatic Analysis

Because the TLA primer sets were designed on the vector sequences, each vector integration should be detectable by the presence of a high and broad sequencing coverage peak in the genomic region corresponding to the vector integration site. The bioinformatic approach for the detection of the vector integrations was therefore based on the detection of regions having high and broad sequencing coverage peaks as well as fusion sequences between the genome and the vector.

The following simplified bioinformatic steps were conducted during data analysis on each sample analysed:

1. Random sorting of the reads, trimming and sampling of 1.5 million sequencing reads.
2. The reference mouse genome (GRCm38.p6/mm10, including only complete chromosomal sequences) and the reference sequences of plasmid 1 and 2 were indexed.

3. Read alignments against the reference mouse genome and plasmids sequences were conducted using TLA-Tools (Cergentis; v0.5), a bioinformatic pipeline written in Perl.
4. The TLA-Tools (Cergentis, v0.5) was used to generate whole-genome coverage plots that were then inspected to detect the candidate integrations regions.
5. The candidate genomic regions identified as vector integration regions were followed up with a visual inspection to confirm that their coverage distribution was consistent with the expected pattern of a vector integration (high and broad coverage peak). For this step, read alignments in BAM format were visualized using IGV (Integrative Genomics Viewer).
6. Sequences depicting the fusion events between the mouse genome and vectors were detected using TLA-Tools (Cergentis, v0.5). The fusion sequences reported were identified applying the following three filtering criteria:
    a. detected in at least two independent primer sets;
    b. located in the genomic proximities of a candidate integration site (often fusion sequences are also located at the highest point of the peak and in correspondence with a coverage drop);
    c. not located in proximity (<5 bp) with a CATG (the sequence cleaved by restriction enzymes used in the TLA protocol).

PCR and Sanger Analysis of the Fusion Sequences

Subclone Generation and Statistical Analysis

The generation of subclones from the MCB was previously performed at Merck. Cells from one vial of the MCB were expanded in culture for four days, then seeded into 96-well plates at a concentration of 0.5 cells/well. Cell growth was monitored by microscopic examination after 24 hours from cells seed. Following expansion, 30 subclones were randomly selected for use in this study. The amount of subclones to be analysed was determined assuming a binomial distribution of the clone and hypothetically divergent subclone. Giving the binomial distribution, the analysis of 30 subclones was determined to have a 95.8% probability of detecting a cell subpopulation with a frequency of 10%. Analysing 30 subclones, the sensitivity of the method is therefore approximately 10%.

DNA Extraction

DNA was extracted from a single vial of MCB ($1x10^7$ cells/vial) and from the NS0 host cell bank ($5x10^6$ cells/vial). The extraction was performed using the QIAAMP Blood DNA Extraction Kit provided by Qiagen based on a column retention system. Eluted nucleic acids were quantified averaging two subsequent measurement on a Nanodrop spectrophotometer. For the extraction, a 260/280 ratio in the range of 1.7 – 2.1 (limits included) was used as acceptance criteria. DNAs were diluted in molecular biology grade water and brought to the concentration of 50 ng/µl. The integrity of the extracted DNA was evaluated using gel electrophoresis, loading 500 ng of sample on a 1x agarose gel. The bands lengths were standardized with a 1kbp molecular marker and were considered valid if they appeared as a single band with no smear and at high molecular weight (>8 kbp).

PCR Protocol for the Analysis of Fusion Sequences

All the PCR detailed in this protocol were performed starting from 100 ng of genomic DNA. Amplification reactions were conducted using the FastStart High Fidelity PCR System (Roche) and reagents concentration was calculated considering a final reaction volume of 40 ul. Amplification products were then loaded on a 1x agarose gel with a PCR marker to assess the band size. In particular, PCR reactions were considered valid if the obtained band length fell within the range of expected molecular weight (bp) ± 10%.
For the analysis of the MCB, 11 individual PCR were performed to verify the 11 fusions sequences identified by TLA at the genomic location of the vectors integration sites. For the analysis of the 30 subclones, 4 PCR assays were conducted per subclone. The thermal profile used for PCR amplification is reported in **Table 4**. Primer sets used for the analysis of the MCB and the 30 subclones, as well as their corresponding amplification product lengths are detailed in **Table 5**. For the investigation of subclone 6, different primer sets (**Table 6**) and thermal profiles (**Table 7**) were used.

Stage 1	95°C for 10 minutes
Stage 2 x 5 cycles	95°C for 1 minute
	68-66°C for 45 seconds
	72°C for 1 minute

Stage 3 x 25 cycles	95°C for 1 minute
	63-61°C for 45 seconds
	72°C for 1 minute
Stage 4	72 °C for 10 minutes
Stage 5	4 °C forever

Table 4. Thermal profile used for PCR amplification.

Assay	Primer Sequence	Expected size of the Amplification Band
Fusion1	GAGTATTCAACATTTCCGTGTCG	557
	GCTTGAGAAATGGCTCAGC	
Fusion2	TTCATCCATAGTTGCCTGACTC	416
	CACTTGGCTTGCTTTGTGC	
Fusion3	TCTCAGCGATCTGTCTATTTCG	570
	TGTCTTTCTGGCCTTGACG	
Fusion4	AGTATTCAACATTTCCGTGTCG	433
	TTTCTGCATTGTAACCTGAAGC	
Fusion5	AGTATTCAACATTTCCGTGTCG	424
	GCACATCTTGCACTTCTCAG	
Fusion6	ATCCGGGAAGGCATAAGC	532
	CCCTTCCCAGGATCATAGTC	
Fusion7	CGTTTGGTATGGCTTCATTCAG	552
	ACAGGGTCTTTGCCTTGC	
Fusion8	TCTCAGCGATCTGTCTATTTCG	568
	TTGTCCTTGATCTTTGAGCAAC	
Fusion9	GTGAGGCACCTATCTCAGC	473
	CCTGTTGTCCCTGACCTC	
Fusion10	TCGCCGCATACACTATTCTC	563
	ACTCATACACACGTATTGAAGC	
Fusion11	AACTCCCAATCTTCTCTCTGC	487
	TTCCTGTGGGGTTAGAAGC	

Table 5. Primer sequences and expected band size of the PCR assays used to confirm the fusion sequences identified by TLA.

Assay	Primer sequence	Expected Size of the Amplification Band
PCR specific	GAGTATTCAACATTTCCGTGTCG	327
	ATAGCCCAAGACCACTTGC	
PCR extraband_1	GAGTATTCAACATTTCCGTGTCG	352
	TCACAGTCAATCATTGACATGC	

Table 6. Clone 6 PCR Primer Sequence and expected band size.

PCR specific

Steps	T°C/time
Stage 1	95°C 10 minutes
Stage 2 x 5 cycles	95°C 1 minute
	68°C 45 seconds
	72°C 1 minute
Stage 3 x 25 cycles	95°C 1 minute
	63°C 45 seconds
	72°C 1 minute
Stage 4	72°C 10 minutes
Stage 5	4°C forever

PCR Extraband_1

steps	T°C/time
Stage 1	95°C 10 min
Stage 2 x 5 cycles	95°C 1 minute
	67°C 45 seconds
	72°C 1 minute
Stage 3 x 30 cycles	95°C 1 minute
	62°C 45 seconds
	72°C 1 minute
Stage 4	72°C 10 minutes
Stage 5	4°C forever

Table 7a and 7b. Thermal profiles used for PCR amplifications of the PCR specific and extra band.

PCR products were checked on a 1x agarose gel and the remaining volumes were purified. For each sample, if a single specific band was observed on the gel, purification was done by pooling the remaining volumes of each replicate of the same PCR, using the MinElute PCR Purification Kit (Qiagen). Alternatively, for samples displaying non-specific bands on gel, the band of interest was purified from agarose gel using the MinElute Gel Extraction Kit (Qiagen). Finally, the purified DNAs were diluted in water and quantified averaging two subsequent measurement on a Nanodrop spectrophotometer.

Sanger Sequencing Protocol for the Analysis of Fusion Sequences

The ABI Prism BigDye Terminator Ready Reaction Cycle Sequencing Kit v.3.1 (Applied Biosystems - Life Technologies) was used to perform the cycle sequencing reaction. This reaction principle is based on the incorporation of fluorescently labelled nucleotide terminators, allowing to read the nucleic acid sequences using an automatic sequencer.
A pre-mix including buffer, BDT mix, and water was prepared and then distributed into all the tubes. The same primers employed for the PCR reactions were used for the elongation (see Table 5 for the primer sequences), sequencing in both forward and reverse orientation. Moreover, each primer was processed in duplicate, to obtain coverage on the region of interest with at least four electropherograms. For each set of cycle sequencing reactions, a tube containing the DNA of the pGEM-3Zf (+) plasmid was included as reaction positive control. Table 8 reports the thermal profile used for the cycle sequencing reactions.

The obtained labelled fragments were purified using the Big Dye 2.0 Spin Kit (Qiagen) and 3 µl of each purification product were loaded on a 96-wells plate with 12 µl of formamide for the sequencing analysis. Samples were denatured at 95 °C for 3 minutes before loading the plate in the sequencer. For the sequencing, the AB3130 Genetic Analyser was used (Applied Biosystem) and the analysis of raw data were performed using the SeqScape 3.0 software (Applied Biosystem). Fragments obtained were aligned with the respective PCR reference sequence to detect the coverage and the potential presence of single point mutations, heterozygosis, deletions, or insertions at the breakpoints.

Stage 1: for 25 cycles	96°C, 10 seconds
	55°C, 5 seconds
	60°C, 4 minutes
Stage 2	4°C, forever

Table 8. Thermal profile used for the cycle sequencing reactions.

6.3 RESULTS AND DISCUSSION

Identification of Plasmid 1 and Plasmid 2 Insertion Regions

The analysis of the insertion sites for the plasmid 1 identified a single candidate insertion region named SITE A in all the TLA primer sets specific for plasmid 1 (sets 4, 5 and 6). Based on the mouse genome coverage of these three datasets, the insertion region is located in the terminal end of chromosome 12 at approximately chr12:114 Mbp (**Figure 18**). Moreover, the analysis of primer set 1 (specific for plasmid 1 as well as plasmid 2), confirmed this region as the region of insertion of the plasmid (**Figure 19**). Of note, “SITE

A" is located in a repetitive sequence of a complex region of the mouse genome, explaining the asymmetric coverage peak obtained at the plasmid integration site.

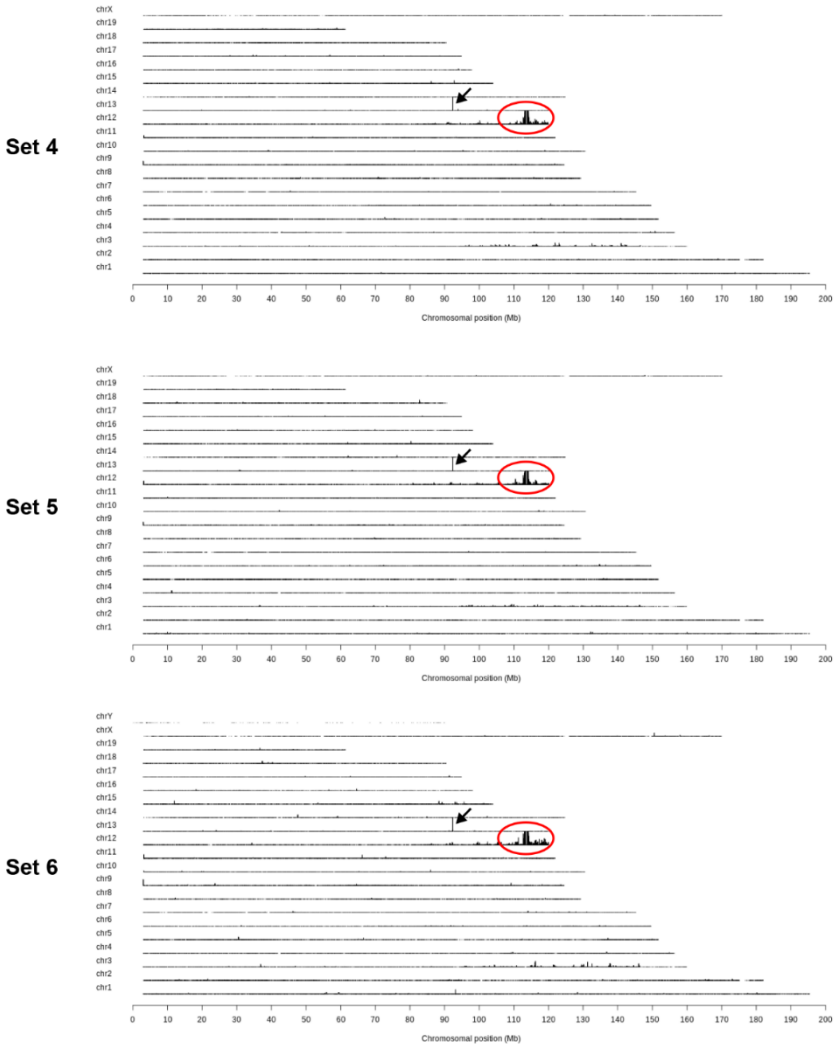


Figure 18. Whole-genome coverage distribution obtained with TLA primer sets specific for plasmid 1 (sets 4, 5, and 6) on the MCB. The plots consistently show the presence of a coverage peak on chr12 (SITE A, red circles) as well as a spike of coverage on chr13 (murine Dhfr gene, black arrows). Y axes are set to a maximum of 100'000 X.

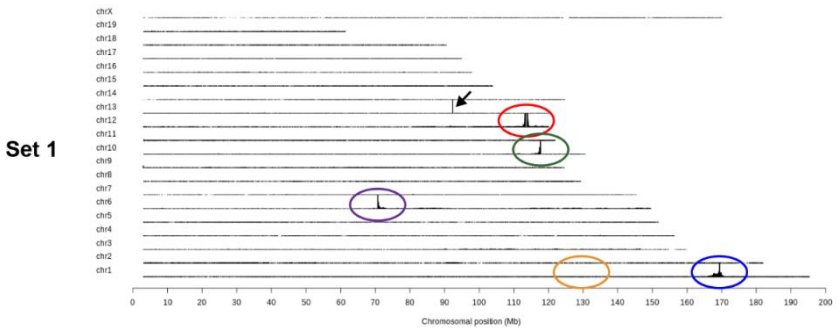


Figure 19. Whole-genome coverage distribution obtained with TLA primer set 1 on the MCB. This primer set is specific for both Plasmid 1 and Plasmid 2. The plot present four coverage peaks on chr12 (Site A, red circle), chr1 (SITE B, blue circle), chr6 (SITE D, purple circle), and chr10 (SITE E, green circle), as well as a spike of coverage on chr13 (murine Dhfr gene, black arrow). For comparison, the position of SITE C on chr1 is indicated by a yellow circle. Y axis is set to a maximum of 100'000 X.

A second coverage signal was detected on chr13 (approximately 92 Mbp, corresponding to the coding region of the murine *DHFR* gene) in primer sets 1, 4, 5, and 6 (**Figure 18** and **19**). This signal however turned out to be an artefact due to the presence of the *DHFR* gene in the Plasmid 1 and was therefore not followed up.

The analysis of the insertion sites of the Plasmid 2 identified four candidate integration regions located in three different chromosomes, but not all these regions were consistently present in all the Plasmid 2-specific primer sets (set 2, 3, 7 and 8), as visible in the mouse genome coverage of these three datasets (**Figure 20**). In addition, the analysis of primer set 1 (specific for both Plasmid 1 as well as Plasmid 2), confirmed three of these candidate integration regions.

The four candidate integration regions identified for Plasmid 2 were the following:

SITE B – this candidate integration region was identified in all the relevant primer sets (1, 2, 3, 7 and 8) and is located approximately at **chr1:169 Mbp** of the mouse genome. The asymmetrical sequencing coverage peak at the integration site and the positions of the fusion sequences suggest the presence of an inversion of genomic sequences at the integration site

SITE C – this candidate integration region was identified in only two primer sets (2 and 3) and is located approximately at **chr1:129 Mbp** of the mouse genome. The identification of this signal in only two out of five primer sets suggested the presence of

a partial vector integration in this genomic region of the mouse genome. This hypothesis was further investigated and confirmed (data not shown).

SITE D – this candidate integration region was identified in all the relevant primer sets (1, 2, 3, 4 and 7) and is located approximately at **chr6:70 Mbp** of the mouse genome

SITE E – this candidate integration region was identified in all relevant primer sets (1, 2, 3, 4 and 7) and is located approximately at **chr10:117 Mbp** of the mouse genome. Nevertheless, the analysis of the coverage distribution revealed a highly asymmetric coverage peak at this location. In addition, a fusion sequence depicting a breakpoint from **chr10:117710736(-) to chr6:70725414(+)** was identified. Together, these two pieces of evidence (the shape of the coverage peak and the fusion sequence) strongly indicate the presence of a genomic rearrangement (translocation between chr10 and chr6) placing this genomic region in close proximity with the plasmid integration in SITE D and explaining the sequencing coverage generated at SITE E. This region is therefore not an independent integration site of vector Plasmid 2, instead it is part of SITE D.

All the candidate integration regions identified in the study were confirmed by visual inspection of the coverage distribution. In all the reported regions, a significant coverage signal indicated the presence of an integrated vector, as displayed and summarized in **Figure 21**.

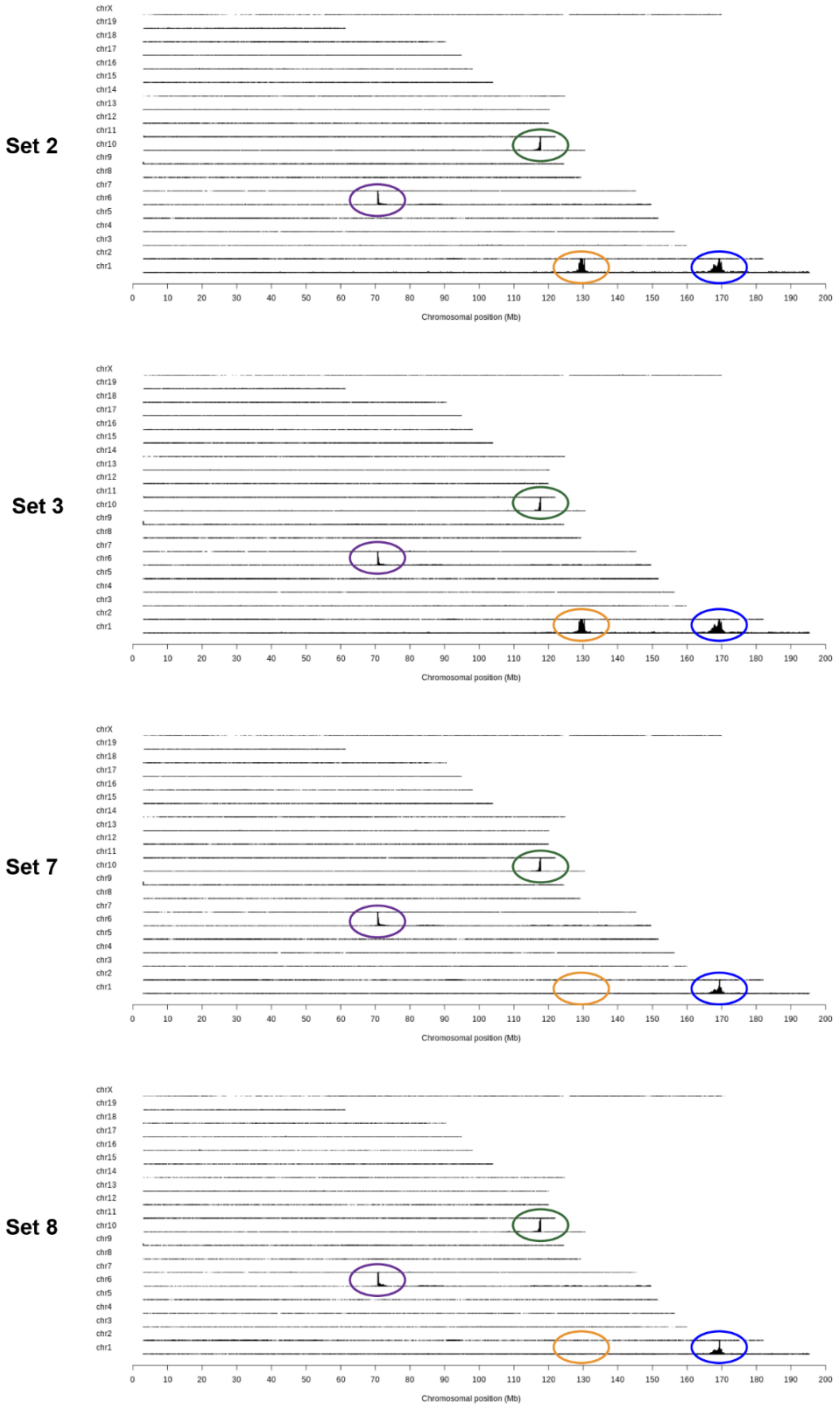


Figure 20. Whole-genome coverage distribution obtained with TLA primer sets specific for Plasmid 2 (sets 2, 3, 7 and 8) on the MCB. The plots consistently show the presence of 3 coverage peaks on chr1 (SITE B, blue circles), chr6 (SITE D, purple circles) and chr10 (SITE E, green circles). In addition, primer sets 2 and 3 display a coverage peak located on chr1 (SITE C, yellow circles). Y axes are set to a maximum of 100'000 X.

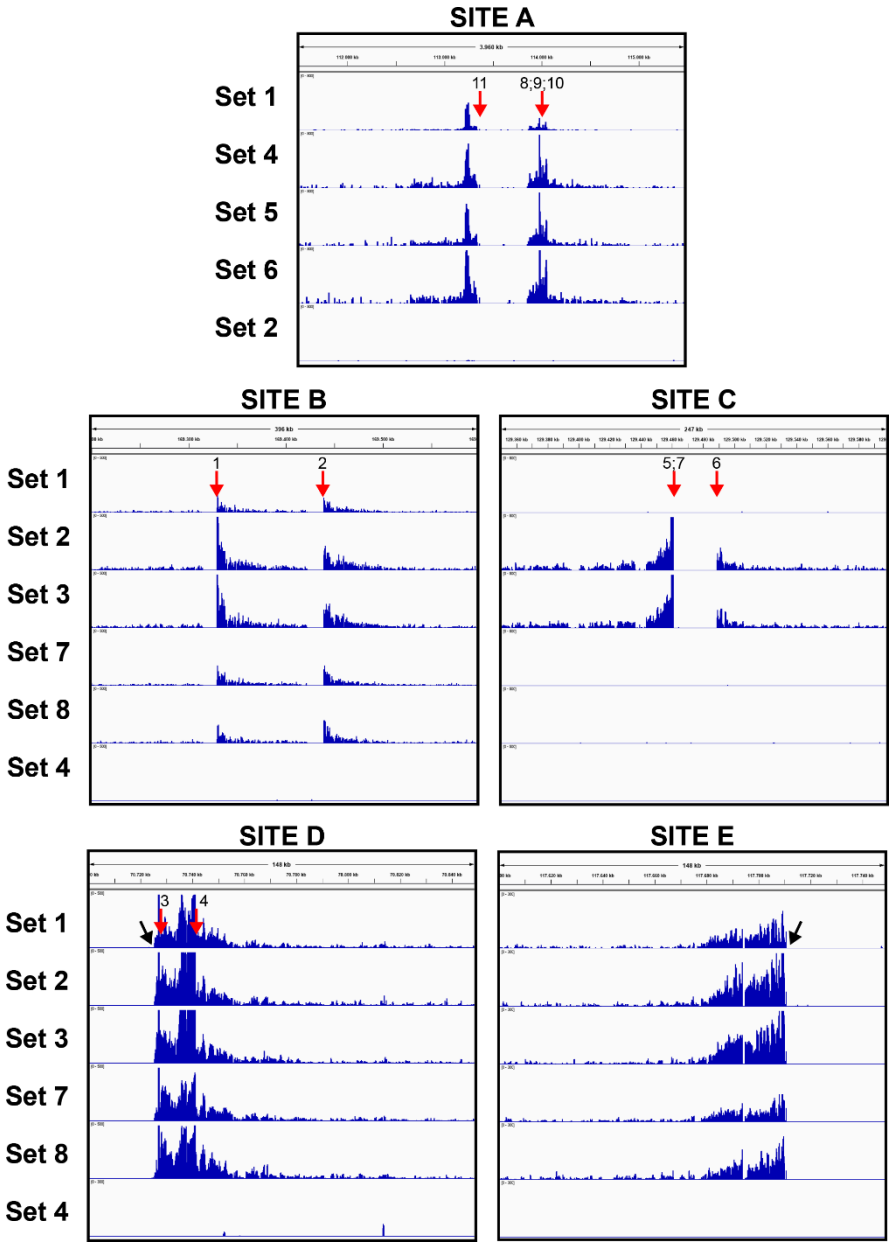


Figure 21. Coverage distribution at the 5 sites identified by the TLA experiments. The plots display the coverage distribution at the genomic regions corresponding to SITE A (chr12:111'500-115'500 kbp), SITE B (chr1:169'200-169'600 kbp), SITE C (chr1:129'350-129'600 kbp), SITE D (chr6:70'000-70'850 kbp), and SITE E (chr10:117'600-117'750 kbp). For each site, the results obtained with TLA primer sets specific for the integrated vector (either Plasmid 1 or Plasmid 2) are displayed. In addition, TLA primer set 2 is displayed as negative control for SITE A and TLA primer set 4 is displayed as negative control for SITEs B, C, D, E. Red arrows indicate the genomic positions of the eleven plasmid-genome fusion sequences; black arrows indicate the breakpoint of the fusion event between chromosomes 6 and 10. Y axis settings were adapted to best display the coverage distribution at the five sites. In particular, Y axis of SITE E is set to a maximum of 300X; for all the other graphs, the Y axes are set to a maximum of 500X.

The analysis of the candidate integration regions of Plasmid 1 and Plasmid 2 identified multiple fusion sequences depicting the exact integration site of the plasmids at base-pair

resolution. During the analysis, a total of 11 candidate transgene-genome fusion sequences (indicated by red arrows in **Figure 21**) were identified at the four vector integration sites, as reported in Table 13. Of note, the expected number of fusion sequences for a single non-complex vector integration is two: one on the 5' side and one on the 3' side of the vector sequence. However, in some cases, more than two fusion sequences for an integration site are reported in Table 13. One reason is because the same fusion sequence is indicated multiple times; fusion 8 and 9 are in fact a single fusion sequence having its genomic side identical to two different genomic locations (the read is in a repetitive region of the mouse genome). Another reason is because sequences with limited evidence were included for follow-up investigation (*i.e.* fusions with limited sequencing coverage, such as fusion 7 and fusion 11).

Site	Fusion Name & Breakpoints	TLA Primer Sets	Sequence
SITE B	Fusion 1 Plasmid 2	1;2;3;7;8	CGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAG CACTTTTAAAGTTCTGCTATGTGGCGCGGTATTATCCCGTGTTGACGCCGG GCAAGAGCAACTCGGTCGCCGCATACACTGCTCTGTGGGTGGCGCTATTCC CTTAGGCAAGTGGTCTTGGGCTATCTTAAAATGCTAGCTAAGTATGAGCCT ATGAATGAGAGAGCAAGACGCATTTCTCCATG
	Fusion 2 Plasmid 2	1;2;3;7;8	TTTATCAGCAATAAACCAGCCAGCCGGAAGGGCCGAGCGCAGAAGTGGTCC TGCAACTTTATCCGCCTCCATCCAGTCTATTAATTGTTGCCGGGAAGCTAG AGTAAGTAGTTCGCCAGTTAATAGTTTGCGCTATCCCATTTACATGAGGTG TAAATGTGTTTTCACA
SITE D	Fusion 3 Plasmid 2	1;2;3;7;8	CACCGGCTCCAGATTTATCAGCAATAAACCAGCCAGCCGGAAGGGCCGAGC GCAGAAGTGGTCCTGCAACTTTATCCGCCTCCATCCAGTCTATTAATTGTT GCCGGGAAGCTAGAGTAAGTAGTTCGCCAGAGCATAGTTTAATCTCATAAA CAAGATAATAAGCAAAACAAAACATTTTTTCATCCATG
	Fusion 4 Plasmid 2	1;2;3;7;8	TTTGCGGCATTTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAA GTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCGAACTG GATCTCAACAGCGGTAAGATCCTTGAGAGTCTTTCCCTGCCAACTTAATGA ACTCTTACACGGGACCAAATTACTCGACCCCAGTGTAGACTCGCTCTCTTG ACACTAACAGTAAATGACCATG
SITE C	Fusion 5 Plasmid 2	2;3	TGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCG AACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTGGCGACCGAGTTGCT CTTGCCCGGCGTCAACACGGGCCAGCTTCAAATGTCAAGGGTTTCAGACCA GGACAACCTAGAGGGACAGTCTGCTTTGTGCTCCTTTGGAGTAAGTTTTCC TTGGCTATGTCCCC
	Fusion 6 Plasmid 2	2;3	CATGGCCGCAGAACCACGCTTTGTTGGTCAGGTAATCGACATTGATTATTG ACTAGTTATTAATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATAT ATGGAGTTCCGCTTTTGGGGTGTATTTACAGTTTTAATAATTTTAGCCTAT TGCAAAAATGTCTACCAGTTTATTCCATAAAAGGTGAAATAATTTGTATTT ATATTTTTTCCAAAGTCACACCAGAAAATTAACTAATTCATA
	Fusion 7 Plasmid 2	2;3	AGTGTATGCGGCGACCGAGTTGCTCTTGCCCGGCGTCAACACGGGCCAGCT TCAAATGTCAAGGGTTTCAGACCAGGACAACCTAGAGGGACAGTCTGCTTT GTGCTCCTTTGGAGTAAGTTTTCCTTGGCTATGTCCCCCTCCAGTACATG
SITE A	Fusion 8 Plasmid 1	1;4;5;6	GCAATAAACCAGCCAGCCGGAAGGGCCGAGCGCAGAAGTGGTCCTGCAACT TTATCCGCCTCCATCCAGTCTATTAATTGTTGCCGGGAAGCTAGAGTAAGT AGTTCGCCAAATATTATTTAATAAAATGTTGCTTGGGGCTCGGCATG
	Fusion 9 Plasmid 1	1;4;5;6	GCAATAAACCAGCCAGCCGGAAGGGCCGAGCGCAGAAGTGGTCCTGCAACT TTATCCGCCTCCATCCAGTCTATTAATTGTTGCCGGGAAGCTAGAGTAAGT AGTTCGCCAAATATTATTTAATAAAATGTTGCTTGGGGCTCGGCATG
	Fusion 10 Plasmid 1	1;4;5;6	CATGTAACTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACCA AACGACGAGCGTGACACCACGATGCCTGCAGCAATGGCAACAACGTGAGAT TAAAAACCAACCTTAAACTAATATTTGGTGTGAAGCATG
	Fusion 11 Plasmid 1	4;5	CATGCGTGGTGGTGGACGTGAGCCACGAAGACCCTGAGGTCAAGTTCAACT GAAAAGCCCTCCCAGCCCCCATCGAGAAAACCATCTCCAAAGCCAAAGGTG GGACC

Table 8. List of fusion sequences identified at the candidate vectors integration sites. For each fusion sequence, the name of the TLA primer sets in which the sequence was detected The plus and minus signs indicate the orientation of the sequences. A plus indicate a 5'→3' orientation from the breakpoint, while a minus indicate 3'→5' orientation from the breakpoint. Pink bases, vector sequences; purple bases, genomic sequences; green bases, inserted sequences (neither plasmid, nor genomic); light blue, homologous bases (both plasmid and genomic).

Confirmation of the Fusion Sequences Detected in the MCB

To verify the vector integration sites detected with TLA, a series of 11 PCRs were performed on the MCB to confirm the fusion sequences depicting the insertion sites between the recombinant vector and the mouse genome. For this purpose, specific primer sets were designed to amplify the DNA extracted from the MCB as already explained in the previously section.

PCR Analysis of the MCB

The DNA extraction of the MCB resulted in genomic DNA passing internal settled quality criteria (data not shown). The PCR analysis of the MCB confirmed only part of the fusion sequences identified with TLA (**Figure 22**).

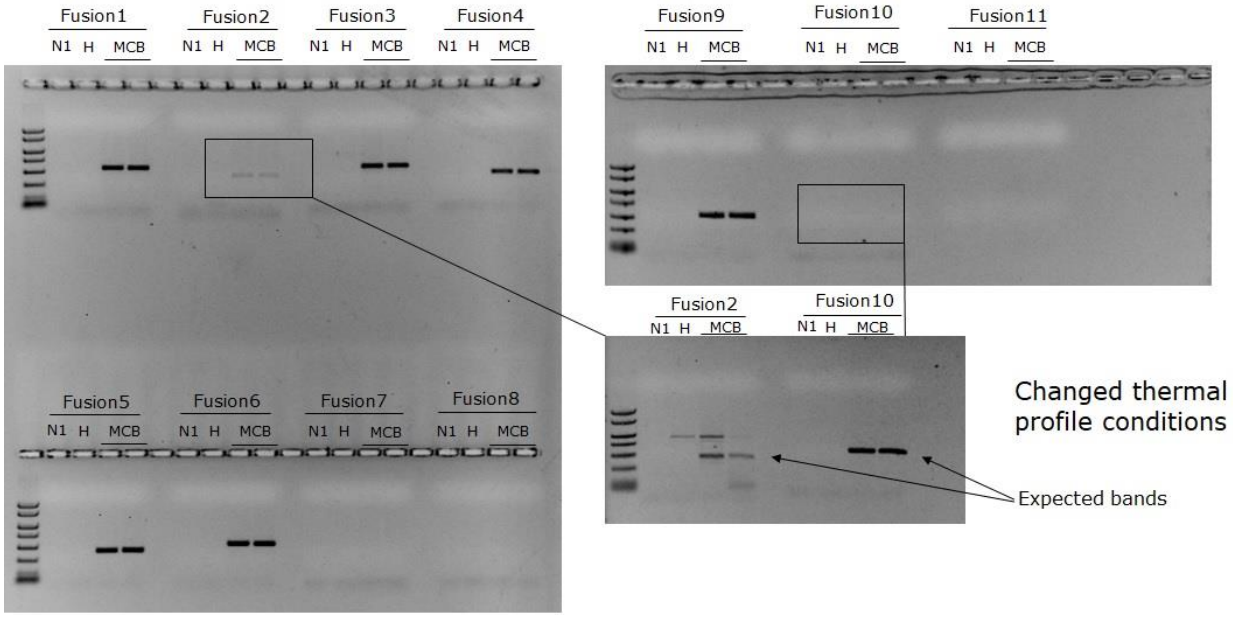


Figure 22. Results of the PCR amplification obtained with the PCR assays designed to confirm fusion sequences from 1 to 11. Samples correspond to: N1, reaction negative control; H, host cell negative control; MCB, a duplicate of MCB.

In particular, only the PCR assays designed based on fusion sequences 1, 2, 3, 4, 5, 6, 9 and 10 resulted positive for the PCR amplification. The assays fusion 2 and fusion 10 were repeated a second time to optimize the thermal profile, by increasing the number of elongation cycles from 25 to 30. Instead, the assays for fusion sequences 7, 8, and 11 were not confirmed by PCR amplification.

Therefore, PCR amplification confirmed only two fusion sequences for each of the vector integration sites identified by TLA. Of note, this number corresponds to the number of fusion sequences expected for a non-complex vector integration (one on the 5' and the other on the 3' of the vector sequence).

Sanger Sequencing of the MCB

Each of the eight amplicons that were amplified by PCR underwent Sanger sequencing to confirm the fusion sequences detected by TLA. For each assay, the resulting electropherograms were aligned to the fusion sequences obtained by TLA (reference sequences reported in Table 13 extended to include the entire predicted amplicon sequence). The sequence analysis was performed for all the eight fusion sequences and the relative reports were generated with the software SeqScape 3.0 (Applied Biosystems). In all cases, no mutations from the expected fusion sequences identified with TLA were detected, thereby further confirming the sequences of the eight breakpoints detected with TLA and confirmed with PCR.

To further investigate the presence of the 4 vector integration sites (SITES A, B C, and D) in the subclones derived from the MCB, one PCR assay was selected for each of the vector integration sites. The assays were selected based on: **a**) the presence of a clear single band on gel, **b**) the good quality of the electropherograms and **c**) the sequence coverage the breakpoints. The assays of fusion sequences 1, 3, 5 and 9 were selected. The electropherograms corresponding to these fusion sequences are reported in **Figures 23**, **24**, **25**, and **26**.

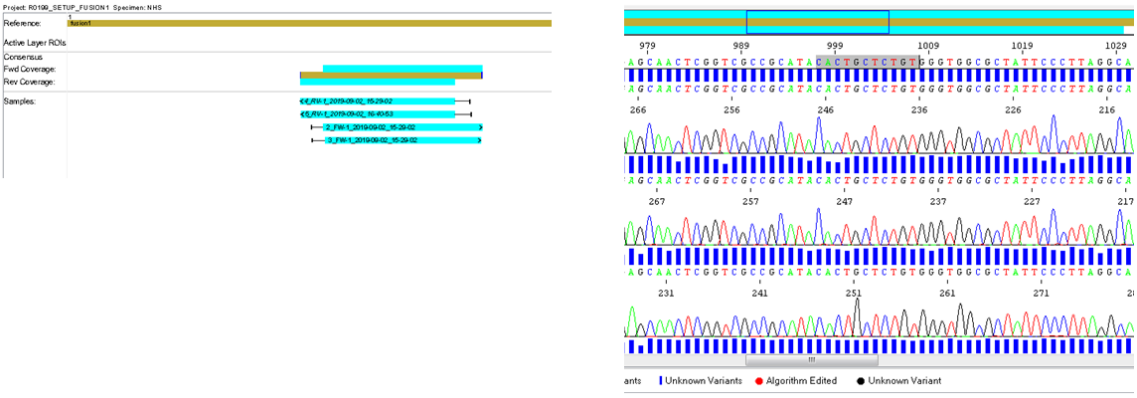


Figure 23. Fusion 1 electropherograms with the relative alignment with its PCR reference sequence.

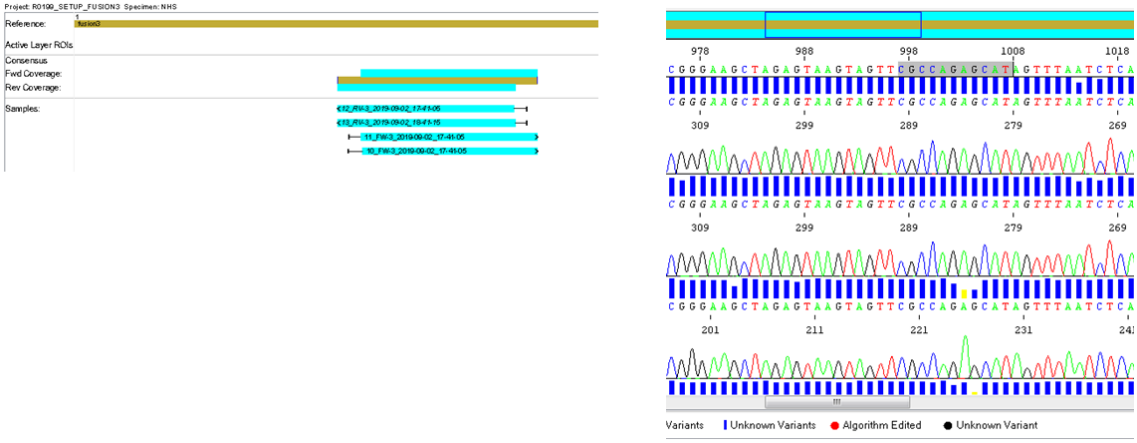


Figure 24. Fusion 3 electropherograms with the relative alignment with its PCR reference sequence

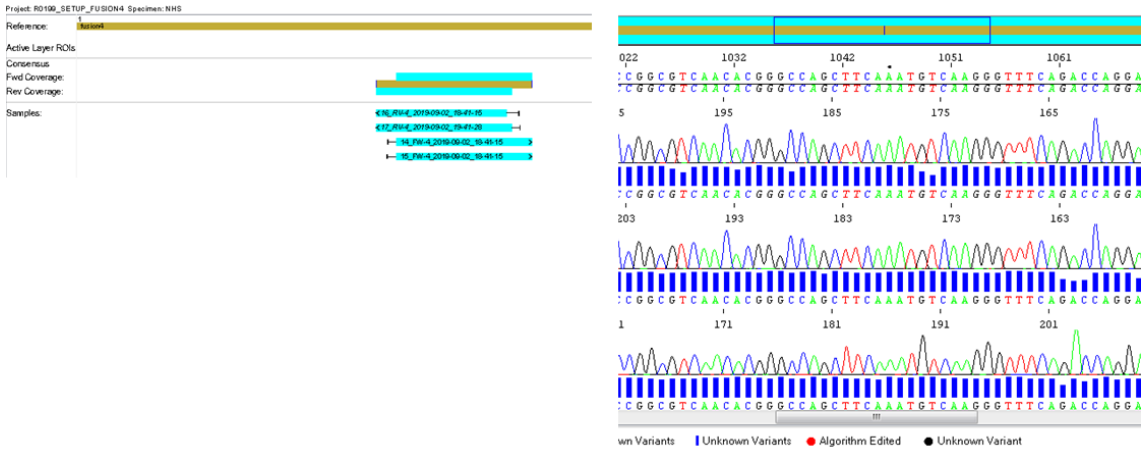


Figure 25. Fusion 5 electropherograms with the relative alignment with its PCR reference sequence

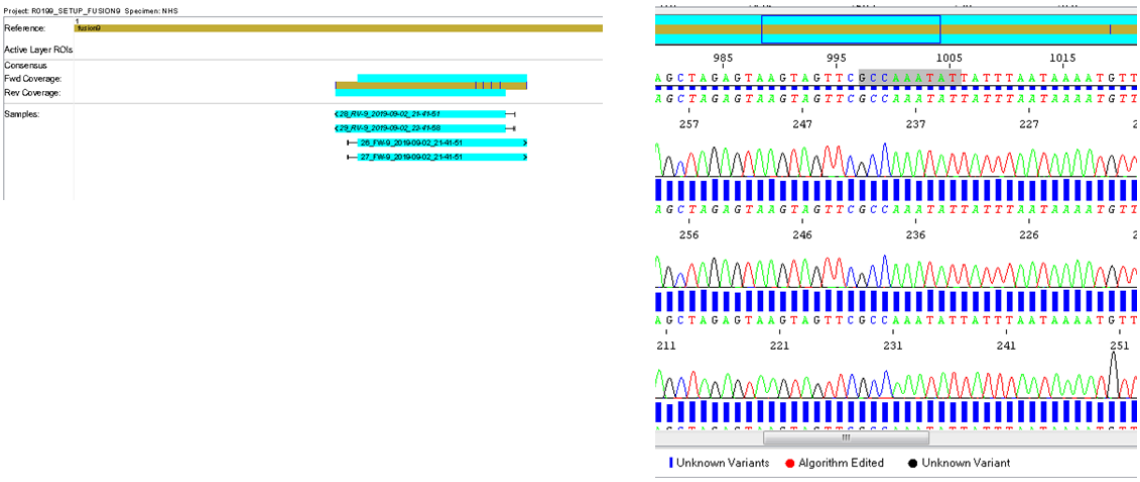


Figure 26. Fusion 9 electropherograms with the relative alignment with its PCR reference sequence

Clonality Assessment: Vector Integration Sites Analysis in the 30 Subclones

The method used for the clonality assessment of the MCB is based on the following assumption: if the MCB is a monoclonal line, subclones derived from the MCB should always display the four integration sites identified in the parental clone (MCB).

To assess the MCB clonality, the expression vector integration sites were identified in 30 subclones derived from the MCB. The 30 subclones had been previously derived from the MCB, as described in section 6.2. The presence of the 4 four integration sites was assessed using four PCR assays specific for the four vector integration sites (fusion sequences 1, 3, 5 and 9, see section 6.2. for the criteria used to select these assays).

PCR Analysis 30 subclones

The DNA extraction of the 30 subclones was conducted as described in the materials and methods section and resulted in genomic DNAs passing the internal quality criteria (data not shown). PCRs with the assays for fusion sequences 1, 3, 5 and 9 were conducted on all 30 subclones, using the reaction condition specified in section 6.2. For each PCR reaction the following samples were analysed:

- A duplicate of each subclone
- The reaction negative control N1 (PCR mix plus water)
- The host cell negative control H (PCR mix plus NS0 host cell)
- A positive control (PCR mix plus MCB).

The results of the electrophoretic analysis of the four fusion sequences are displayed in **Figures 27**, **28**, **29** and **30**.

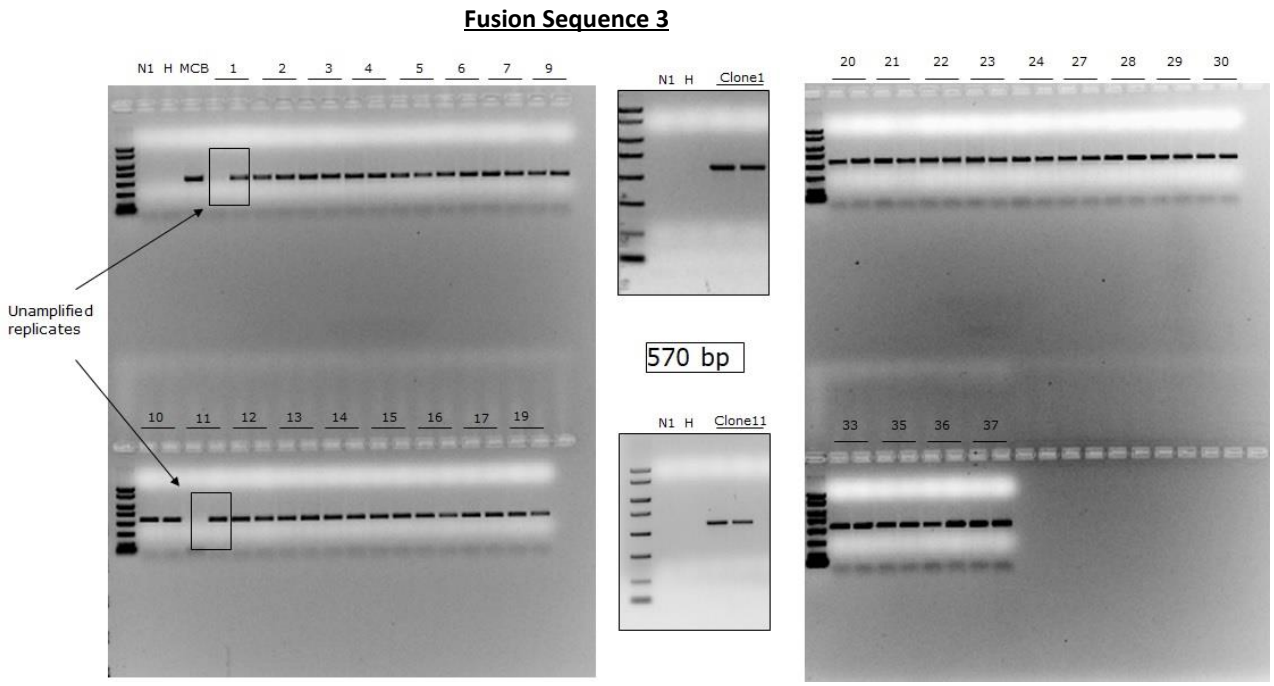


Figure 27. PCR products of the assay fusion sequence 3 on the 30 subclones. PCR Samples correspond to: N1 reaction negative control; H, NS0 host cell negative control; MCB, positive control; numbers indicate the duplicates of each subclone. The expected length of the PCRs is 570 (bp) and all detected bands are ranging between the bp value of 570 ± 10%.

Fusion Sequence 5

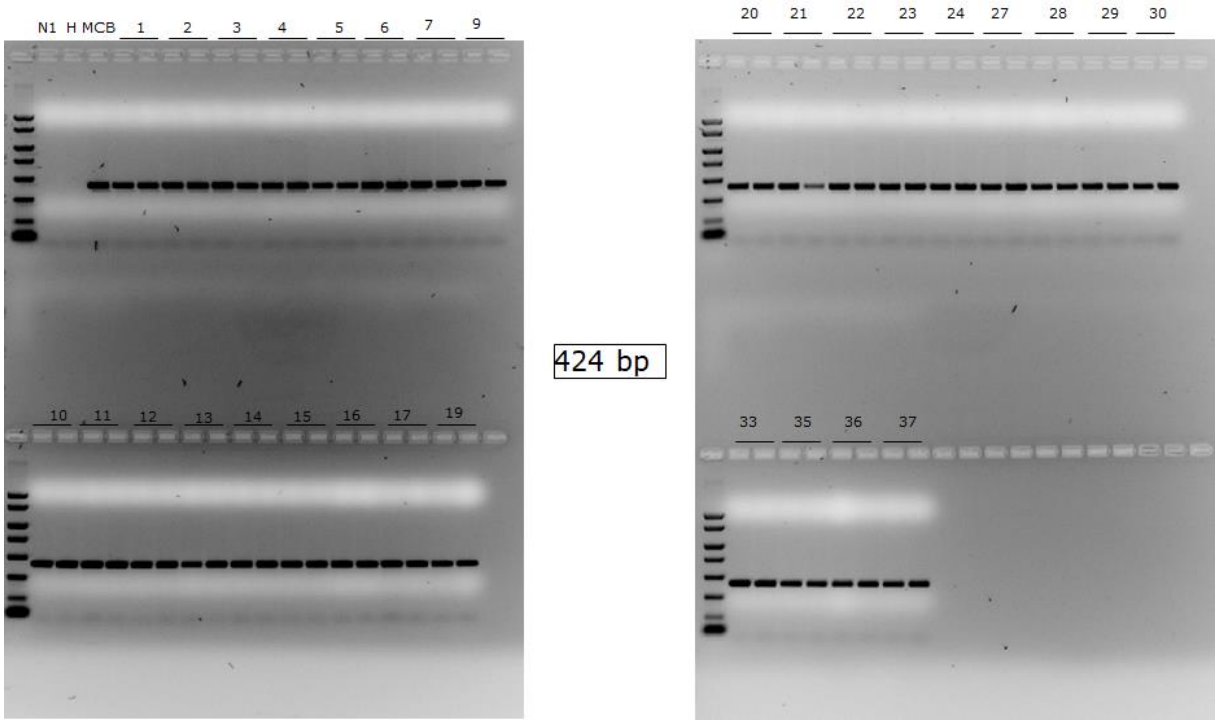


Figure 28. PCR products of the assay fusion sequence 5 on the 30 subclones. PCR Samples correspond to: N1 reaction negative control; H, NS0 host cell negative control; MCB, positive control; numbers indicate the duplicates of each subclone. The expected length of the PCRs is 424 (bp) and all detected bands are ranging between the bp value of 424 ± 10%.

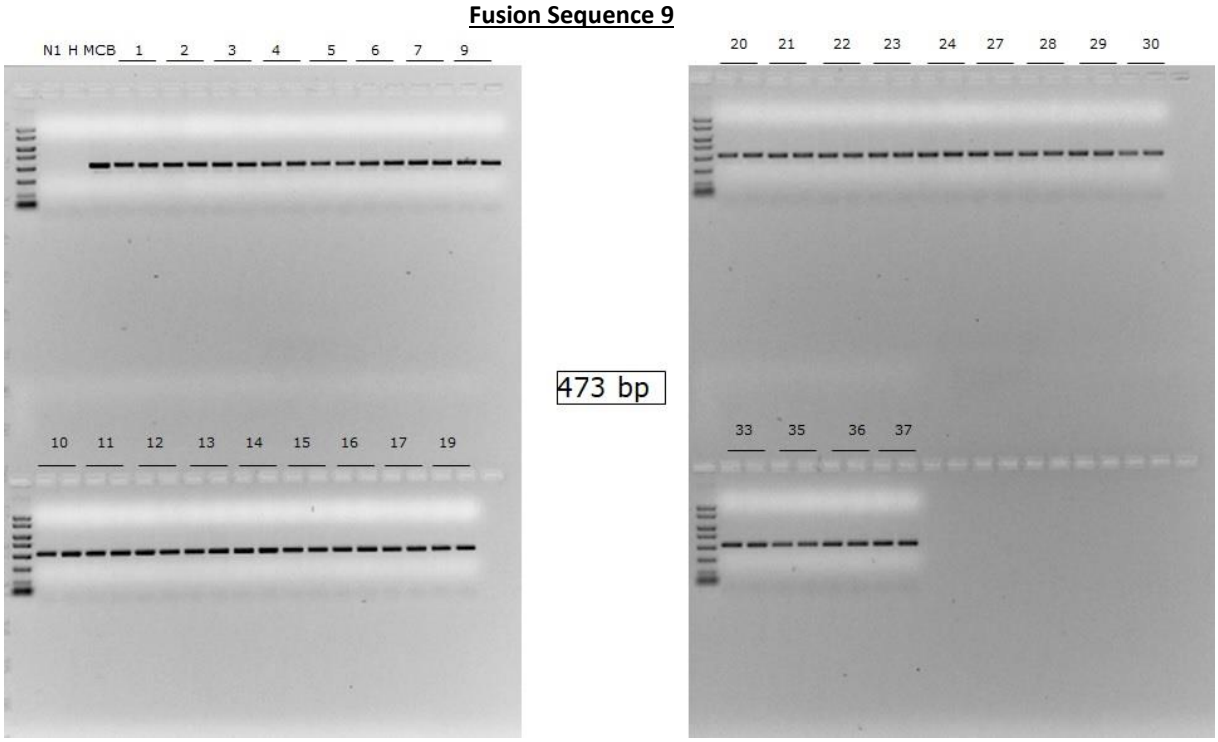


Figure 29. PCR products of the assay fusion sequence 9 on the 30 subclones. PCR Samples correspond to N1 reaction negative control; H, NS0 host cell negative control; MCB, positive control; numbers indicate the duplicates of each subclone. The expected length of the PCRs product is 473 (bp) and all detected bands are ranging between the bp value of 473 ± 10%.

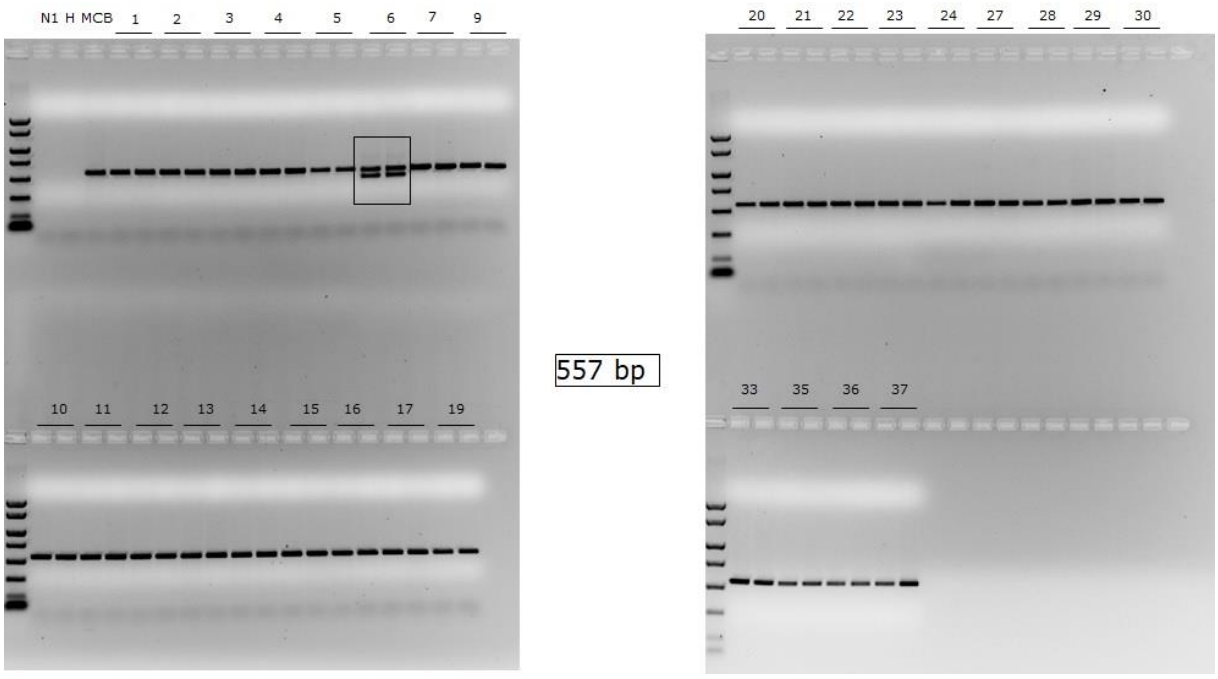


Figure 30. PCR products of the assay fusion sequence 1 on the 30 subclones. PCR Samples correspond to: N1 reaction negative control; H, NS0 host cell negative control; MCB, positive control; numbers indicate the duplicates of each subclone. The black box highlights the bands corresponding to subclone 6. The expected length of the PCRs is 557 (bp) and all detected bands are ranging between the bp value of 557 ± 10%.

For all the four fusion sequences, each clone displayed a specific band with a bp length corresponding to the expected size observed in the MCB. For fusion sequence 3 (**Figure 27**), subclones 1 and 11 were repeated a second time because one of the two replicates didn't amplify as expected. The second amplification did confirm the presence of fusion sequence 3 in these two subclones. The identification of the four fusion sequences in all the subclones confirmed the presence of the 4 vector integrations in all the MCB-derived subclones. Moreover, in one of the subclones (subclone 6) the PCR assay for fusion sequence 1 revealed two distinct bands at different lengths. In particular, in addition to the expected band at 557 bp, a second band of approximately 470 bp was visible in subclone 6 (**Figure 30**). The origin of this band was further investigated to assess its relevance to the clonality of the MCB.

Investigation on Subclone 6

The identification of two amplification products for fusion sequence 1 in subclone 6 (the expected 557 bp and a second at approximately 470 bp, also referred to as the "*extra band*") indicated the presence of a genetic difference between the MCB and subclone 6 (Figures 14 and 15). At the same time, the presence of the expected bands for all four fusion sequences in subclone 6 clearly demonstrated the presence of the four vector integration sites in subclone 6, thereby suggesting that this clone was somehow derived from the same clone of the MCB. To explain the presence of a second amplification product, two scenarios have been hypothesized:

1) The 470 bp band in subclone 6 is the result of a genetic modification of the MCB introduced during the subcloning process (in this hypothetical scenario the MCB is monoclonal).
2) The 470 bp band in subclone 6 is the genetic marker of a rare cell subpopulation present in the MCB (in this hypothetical scenario the MCB is polyclonal).

To confirm one of these theories, an investigation on the *extra band* product was performed, characterizing its sequence and verifying its presence in the MCB.
To investigate the sequence of the 470bp amplification product, the PCR assay of fusion sequence 1 was repeated on MCB and subclone 6 and both bands visible in subclone 6 were extracted and purified directly from the agarose gel (**Figure 31**).

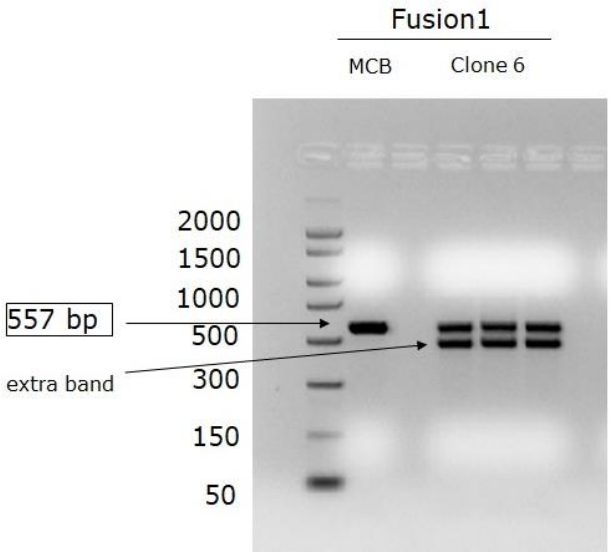


Figure 31. PCR products of the assay fusion sequence 1 on the MCB and subclone 6. PCR Samples correspond to: MCB positive control and a triplicate of clone 6.

The DNA from each band was extracted, quantified, and used to perform the cycle sequencing reaction. For the elongation of labelled fragments, the same primers of PCR Fusion1 were used (Fw_fusion1 + Rv_fusion1). Cycle sequencing products were processed on the AB3130 Genetic Analyser (Applied Biosystems) and subsequently analysed with SeqScape 3.0 software (Applied Biosystems). The sequence obtained from the expected band was aligned with the Fusion 1 reference sequence (represented by the Fusion 1 PCR) and no mutations were detected, while the sequences of the extra band were directly extrapolated from the genetic analyser, and then characterized. Of note, the sequence of the extra band product turned out to be highly similar to the fusion sequence 1 amplicon, as displayed in **Figure 32**.

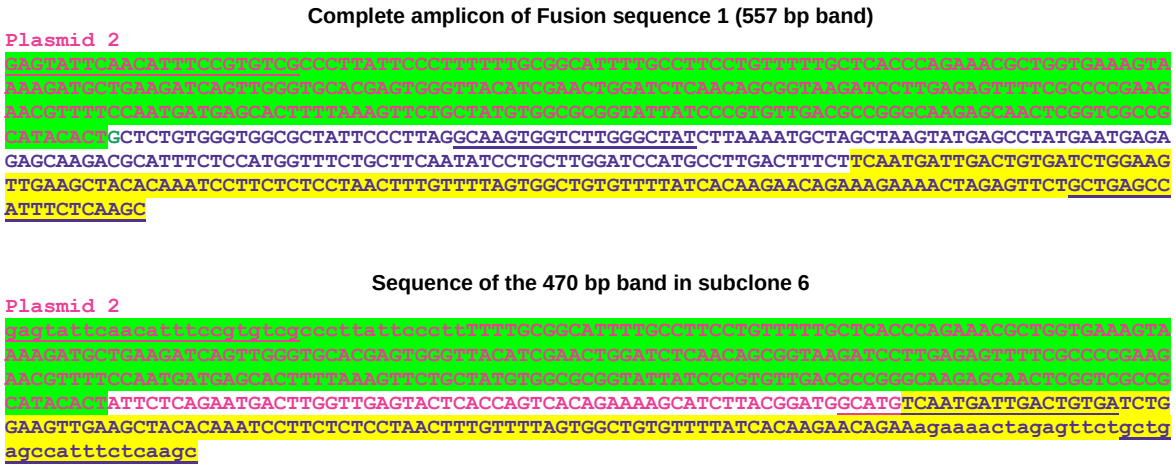


Figure 32. Theoretical sequence of the complete amplicon for the assay "Fusion sequence 1" (top) and alternative PCR product identified in subclone 6 with Sanger sequencing (bottom). Identical sequences are highlighted with the same color (green or yellow). For both sequences, the position of the vector and genomic breakpoints are indicated (note that the vector sequence is indicated as vector Plasmid 2 based on the TLA results, but the sequence is in fact present also in the Plasmid 1 vector). The plus and minus signs after the plasmid/genomic positions indicate the relative orientation of the sequences. A plus indicate a 5'→3' orientation from the breakpoint, while a minus indicate 3'→5' orientation from the breakpoint. Pink bases, vector sequences; purple bases, genomic sequences; green base, inserted sequence (neither plasmid, nor genomic). Upper case letters indicate bases that have been sequenced with four electropherograms; lower case letters indicate theoretical bases, not observed in the electropherogram. Underlined bases correspond to the position of the annealing sites of the primers sets used. The forward primer is used in all the sets, while the reverse primer is different in the three assays: "fusion sequence 1" (non-specific) , "PCR specific" (specific for the sequence on top), and "PCR extraband_1" (specific for the sequence on the bottom)

Because the PCR assay for fusion sequence 1 is specific for both the fusion sequence 1, as well as the extra band sequence, new assays specific for the two sequences were designed. In particular, the identified sequence of the *extra band* was used to design two new sets of primers: one specific for the expected band of fusion sequence 1 and one specific for the *extra band* sequence. Given the specificity of the *extra band* assay, its use on the MCB was expected to reveal the presence of the *extra band* product with higher sensitivity than the PCR assay of the fusion sequence 1.

Two distinct thermal profiles were used, one for the "PCR Specific" with 25 elongation cycles and one optimized thermal profile for the *extra band* assay ("PCR extraband_1"), with 30 elongation cycles. A second vial of subclone 6 was used as additional control. After amplification, PCR products were analysed on a 1x agarose gel and the obtained

bands were standardized to PCR marker to check the molecular weight size. As acceptance criteria, PCR were considered valid if the obtained band length falls within the range of PCR expected bp ± 10%. The obtained results are displayed in **Figure 33**.

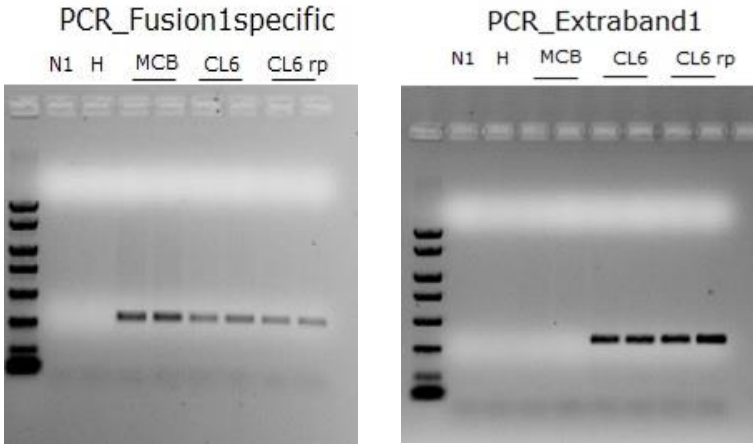


Figure 33. PCR products of the assays specific for the fusion sequence 1 and the extra band. PCR Samples: N1, reaction negative control; H, NS0 host cell negative control; MCB positive control; CL6, duplicates of subclone 6; CL6 rp, duplicates of subclone 6 repropagated.

The results obtained with the "PCR Specific" assay confirmed once again the presence of the expected fusion sequence 1 in both the MCB as well as in subclone 6, confirming the common genetic origin of the MCB and subclone 6. On the contrary, the use of the assay "PCR Extraband_1" did provide a PCR product for subclone 6 but not for the MCB, confirming the presence of this PCR amplicon in subclone 6 and its absence in the MCB. Overall, the collected evidence indicates that subclone 6 is not a rare cell subpopulation present in the MCB and instead the genetic difference observed in this subclone is most likely a mutation introduced in the subclone 6 during the subcloning process.

The results of the study confirm the capability of the TLA technology to identify the genomic location of vectors integrated in the genome of recombinant cell lines. The TLA analysis of the MCB of NHS-IL12 allowed the characterization of the integration sites of the vectors at base-pair resolution.

In total, four vector integration sites were detected: one integration site on chromosome 12 (SITE A) for the Plasmid 1 and three integration sites on chromosomes 1 and 6 (SITE B, C, and D) for the Plasmid 2. For each integration site, at list two breakpoints depicting

the fusion events between the mouse genome and the vectors were detected by TLA and then confirmed by PCR and Sanger.

The PCR analysis of 30 subclones derived from the MCB was used to assess the clonality of the MCB. The fusion sequences characteristic of the MCB vector integration sites were used as unique molecular markers to assess the presence of the four vector integrations in all the 30 subclones. The results of this analysis showed that all the subclones analysed have the two vectors inserted in the same four positions of the mouse genome as the MCB. In one of the subclones, in addition to the expected breakpoints identified in the MCB, the analysis identified a modification of the breakpoint of SITE B. Nevertheless, this modified breakpoint turned out to be a genetic mutation introduced in the subclone, likely during the subcloning process. Overall, the analysis excludes with a probability of 95.8% the presence of cells subpopulations with a frequency of 10% or more in the MCB. The analysis therefore supports the absence of cell sub-populations in the MCB with a sensitivity of 10%.

Notably, a recent commentary paper from the members of the Working Group on Clonality of the International Consortium for Innovation and Quality in Pharmaceutical Development, described the combination of TLA and PCR as an effective approach for the clonality assessment of cell banks, therefore reinforcing the validity of the method applied in this study [18].

In conclusion, the results of this study confirmed the validity of proposed approach for the clonality assessment of cell banks. The application of TLA, NGS, PCR, and Sanger sequencing allowed a detailed genetic characterization of the vector integration sites of the MCB. Moreover, the identification of identical vector integration sites in the MCB and derived subclones confirmed that all the samples analysed are derived from a single cell progenitor, thereby confirming the clonal origin of the MCB.

7. FINAL DISCUSSION AND FUTURE PERSPECTIVES

The production of biotechnological drugs represents a complex topic. The employment of biological systems for drug production, makes the entire manufacturing process not easily affordable. Health authorities require each step of the production chain, to be strictly controlled and performed following the GMPs. From the cell line development to all the quality control tests, each phase follows well-defined procedures, allowing

molecules to undergo clinical trials and finally be accepted to the global market and be administered to patients. Being compliant with health authorities' requirements, represents for biotechnological industries a major investment, and many efforts and professional skills are put in place to achieve these goals. For that reason, new technologies and instrumentations are exploited to enhance performances and release safer and reliable drugs. The quality control represents an unavoidable phase of biotechnological drugs production, because it defines molecules safety and their follow-up. For that reasons, new technologies are constantly investigated to implement the currently used methodologies and substitute them to release more reliable results. Indeed, during the last decades, Next Generation Sequencing has become the gold standard for genomes sequencing. It is widely used for research purposes, but its high throughput makes it appealing to pharmaceutical industries, to determine the genomic stability of biotechnological molecules. For that reason, many assays that are currently conducted using Sanger method, are progressively being substituted with NGS. The aim of this manuscript was to investigate NGS property and evaluate its possible application to define both molecules and cell lines genomic stability. In the first part of the manuscript, the NGS application to whole recombinant vector sequencing was investigated. It is currently widely accepted, that the expression vectors used to transfect mammalian cells for the big-scale production of biotechnological molecules (like antibodies or hormones), are often standardized, and engineered to insert a specific recombinant gene. Given that vectors sequences are maintained the same, except for the transgene insertion, their whole sequence, including introns and accessory codifying regions, is not verified during the cell line development. Whole vector sequencing can give information about: plasmid real sequence after its synthesis and engineering; plasmid genetic stability after long-term storage; plasmid whole sequence to be used as positive control and reference sequence for the genotypic characterization tests. Here, the NGS potential to produce high throughput data about vector coding sequences was described. Another relevant aspect of this technology, compared to Sanger method, is that a big amount of data is obtained from low quantity of input DNA, but most of all, without the need of amplifying a target sequence. This bypassed step, avoids possible experimental artefacts due to the amplification phase followed by its purification, which can introduce in the sample impurities, aspecific amplicons or accessory mutations due to Taq polymerase low efficiency. However, the introduction of a new analytical method in a quality control

workflow is not simple. The parameters that must be evaluated to setup and validate a new analytical method depend on the method type as detailed in the ICH Q2 R01 and could be the specificity, limit of detection, robustness, linearity, range, accuracy and precision. Since NGS sequencing method is considered a "Impurity Limit test", the specificity and the LOD were evaluated to investigate the possibility of introducing NGS based method for the genotypic characterization of bacterial cell banks used to produce Merck molecules.

In the second part of the manuscript, E. Coli competent cells and engineered recombinant vectors were used to evaluate NGS based method specificity and limit of detection. As widely described in **paragraph 5**, the capability of the method to recognise punctiform polymorphism and deletions was confirmed, while the limit of detection was settled at 10%. Also in this case, the potential of NGS was observed both in the absence of PCR amplification and in the large amount of data produced, starting from low quantity of input samples. Given the obtained results, the implementation of the new sequencing technology in substitution of Sanger method, seems to be promising and fruitful for futures genotypic characterization tests. Its potential resides in the cost-effectiveness for numerous analysed samples, and the undeniable reliability of data produced. NGS-based methods can be promising not only for the quality control of recombinant molecules but also for the preliminary phase of the cell line development. In this context, recombinant cell genomes can be sequenced to evaluate the cell bank clonality through the determination of the expression vectors' insertion sites.

In the third chapter of the manuscript, Target Locus Amplification technique was combined with NGS to determine the integration sites of a recombinant molecule among the MCB genome exploited for its big-scale production. The fusion sites discovery and its confirmation in 30 subclones derived from the MCB, confirmed that the cell bank used to produce the biotechnological molecule, derived from a single progenitor and was totally clonal. Given that the analytical method is still not validated, TLA technology is currently used only for setup experiments. Although, it is under discussion the possibility of using this technique to slim the process of drug discovery, by accelerating recombinant cells screening during the "best clone" selection phase of the cell line development. In this context, TLA will catalyse both the determination of cell bank clonality and the

confirmation of the genomic stability of the transgene that codify for the molecule of interest. This will simplify the screening process of Merck molecules undergoing to clinical trials phases.

Overall, NGS based technologies have a great potential and can be widely used for different purposes, from the genomic stability evaluation of recombinant molecules, to the confirmation of recombinant cell lines identity and exclusion of foreign nucleic acids contaminations. As already stressed in the previous paragraphs, one of the major potentials of NGS is represented by its high throughput. Big data management is not always simple and requires hi-tech server infrastructures to storage all the collected data. Moreover, from an experimental point of view, the big quantity of reads produced by the sequencing, must be properly filtered and interpreted, to avoid incorrect results release. Bioinformatic pipelines must be designed to cover their specific target, and different algorithms and commands are required for each analysis. The statical significance of data produced is another important aspect to keep in mind when the results interpretation is carried out. The more reads are produced, the more likely they are to present reciprocal incongruities and is important to interpret their truthfulness. Despite the complex aspects of data management, NGS technology represents the gold standard for genome sequencing, and its employment in different pharmaceutical contexts is constantly evolving, providing for the future, the release of safer drugs to the global market and their final administration to patients.

8. BIBLIOGRAPHY

1. Wurm, F. M. Production of recombinant protein therapeutics in cultivated mammalian cells. *Nat. Biotechnol.* **22**, 1393–1398 (2004).
2. Barnes, L. M., Bentley, C. M. & Dickson, A. J. Stability of protein production from recombinant mammalian cells. *Biotechnol. Bioeng.* **81**, 631–639 (2003).
3. Akhtar, W. *et al.* XChromatin position effects assayed by thousands of reporters integrated in parallel. *Cell* **154**, 914–927 (2013).
4. Li, F., Vijayasankaran, N., Shen, A., Kiss, R. & Amanullah, A. Cell culture processes for monoclonal antibody production. *MAbs* **2**, 466–479 (2010).
5. König, K. *et al.* Implementation of amplicon parallel sequencing leads to improvement of diagnosis and therapy of lung cancer patients. *J. Thorac. Oncol.* **10**, 1049–1057 (2015).
6. Rivas, M. A. *et al.* Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat. Genet.* **43**, 1066–1073 (2011).
7. Pareek, C. S., Smoczynski, R. & Tretyn, A. Sequencing technologies and genome sequencing. *J. Appl. Genet.* **52**, 413–435 (2011).
8. Kawashima et Al. No TitleMethod of nucleic acid amplification. Google patent (1998).
9. Chen, F. *et al.* The History and Advances of Reversible Terminators Used in New Generations of Sequencing Technology. *Genomics, Proteomics Bioinforma.* **11**, 34–40 (2013).
10. Reuter, J. A., Spacek, D. V. & Snyder, M. P. High-Throughput Sequencing Technologies. *Mol. Cell* **58**, 586–597 (2015).
11. Frye, C. *et al.* Industry view on the relative importance of 'clonality' of biopharmaceutical-producing cell lines. *Biologicals* vol. 44 117–122 (2016).
12. 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. *Biotechnol. Prog.* **34**, 635–649 (2018).
13. Scarcelli, J. J. *et al.* Analytical subcloning of a clonal cell line demonstrates cellular heterogeneity that does not impact process consistency or robustness. *Biotechnol. Prog.* **34**, 602–612 (2018).

14. Welch, J. T. & Arden, N. S. Considering “clonality”: A regulatory perspective on the importance of the clonal derivation of mammalian cell banks in biopharmaceutical development. *Biologicals* vol. 62 16–21 (2019).
15. Klottrup, K. J., Miro-Quesada, G., Flack, L., Pereda, I. & Hawley-Nelson, P. Measuring the aggregation of CHO cells prior to single cell cloning allows a more accurate determination of the probability of clonality. *Biotechnol. Prog.* **34**, 593–601 (2018).
16. Ballester, L. Y., Luthra, R., Kanagal-Shamanna, R. & Singh, R. R. Advances in clinical next-generation sequencing: Target enrichment and sequencing technologies. *Expert Review of Molecular Diagnostics* vol. 16 357–372 (2016).
17. De Vree, P. J. P. *et al.* Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping. *Nat. Biotechnol.* **32**, 1019–1025 (2014).
18. Wu, P. *et al.* Advancing biologics development programs with legacy cell lines: Advantages and limitations of genetic testing for addressing clonality concerns prior to availability of late stage process and product consistency data. *PDA J. Pharm. Sci. Technol.* **74**, 264–274 (2020).