UNIVERSITA' DEGLI STUDI DI TORINO

Ph.D. School in Health and Life Sciences
Ph.D. Programme in Complex Systems for Life Sciences

Development of a Next-Generation Sequencing Method for Yeast cell bank identity confirmation in biopharmaceutical production

XXXII CYCLE
ACADEMIC YEARS: 2016-2019

Chiara Celli

Advisor: *Ph.D. Fabrizio Lecce*
Coordinator: *Prof. Michele De Bortoli*

Table of Contents

Introduction

Pharmaceutical drug production and GMP

Pharmaceutical drugs can be sub-classified into two categories based on the production process: chemical and biological drugs. Nevertheless, biological drugs are becoming more and more relevant such as Panitumumab (Amgen), Tocilizumab (Roche) or Cetuximab (Merck KGaA). Consequently, the impact of biological products is important not only from an economic point of view but also in terms of treated diseases.

Currently, there is no simple way to define all the drugs that are reported to be 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*"[1].

ICH guidelines provide broad guidance on appropriate standards for the derivation of human and animal cell lines and microbes used to:

- prepare biotechnological/biological products;
- prepare and characterize cell banks used for production.

Types of biological products are blood derivatives, proteins, whole blood, human tissue xenotransplantation, products, vaccines (preventive and therapeutic) allergenic, blood extracts, components cellular and gene therapies[2].

However, most of the biologicals are produced by biotechnological processes and with the DNA recombinant technology: one or more genes coding for human proteins with a therapeutic effect is integrated into a vector and then expressed in a bacteria, yeast or mammalian cell system. Biologicals must be processed under tightly controlled conditions/controls throughout production to consistently produce a safe, pure, and potent product, and avoid the introduction of environmental contamination[2].

In contrast to most drugs that are chemically synthesized and have a known structure, most biological products are complex mixtures that are not easily identified or characterized. As with small-molecule drugs, research and development of biologics is expensive and risky, often ending in failure.

Chemical drugs are often purer and better characterized by current analytical technology than biologics. The activity of a biological agent may be affected by the cell system in which it is produced, the fermentation media or operating conditions.

The production process of chemical drugs is relatively well defined, which allows these drugs to be produced in uniform large quantities. However, biologics have a complex production process that tends to yield small quantities. It is difficult to scale up biologics from laboratory quantity to larger-scale batches and maintain product purity and batch-to-batch equivalence.

Recombinant DNA requires isolating the DNA from human cells and potentially modifying that DNA segment, inserting it into bacteria or a mammalian cell, and getting that organism or cell to express it. Several steps are involved in the development process: locating genes that code for proteins, cloning genes, reproducing the proteins associated with the genes, determining the role of the proteins in the disease process, and then developing a potential therapy.

There is a greater potential for immune reactions to biologics than to chemical drugs. The molecules in chemical drugs are too small to be considered immunogenic and generally are not recognized by the immune system as "invaders."

The complex production process of biologics points out more challenges compared to chemical drugs. In fact, despite the chemical drugs, biotechnological products are molecules with high molecular weight, with one or more subunits and high structural complexity. The comparison of the different complexity between chemical and biological drugs is easy and evident in Figure 1 [3].

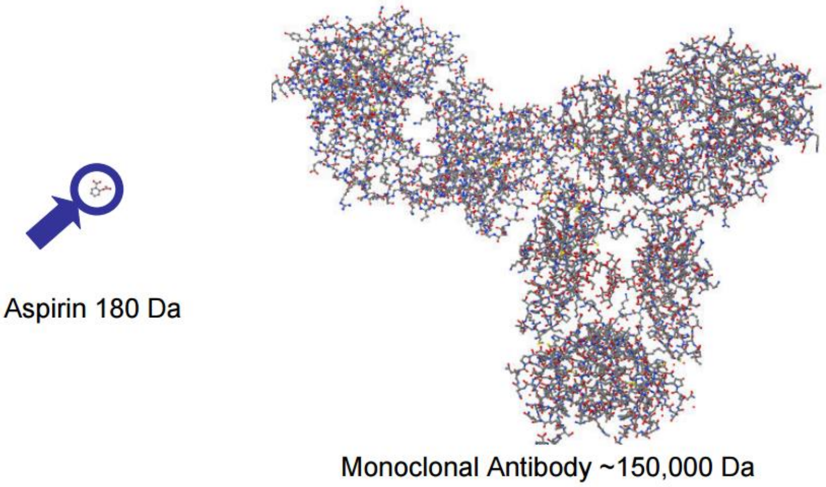


Figure 1 - Size and Complexity of Proteins[4]

Furthermore, the biologicals production is an extremely complex process compared to chemical reactions that can be easily standardized and reproduced. On the other hand, during the biologicals production, since live systems are involved, tiny differences in the condition and execution of the production can highly impact the final product quality. Due to this reason, a big effort is required to standardize as much as possible each production process step and tests used to characterize the final product. It is important to have a biological drug with desired quality in terms of efficacy and safety.

Consequently, the production and the commercialization of the biological products must follow strict guidelines, the so-called Good Manufacturing Practices (GMP). GMP is a group of national operative rules that aim to guarantee the safety and quality of pharmaceutical products.

Health Authorities (HA), as EMA (European Medicines Agency) and FDA (Food and Drug Administration), provide guidelines that guarantee the efficacy and the safety of a drug. Furthermore, audits are periodically settled by HA to verify the compliance with these guidelines. If GMP requirements are not completely fulfilled and severe lack are found during an inspection, HA has the faculty to interrupt the biopharmaceutical production and, even in some cases, close the entire production site.

GMP describes a panel of principles and procedures that help ensure that therapeutic goods are of high quality. 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. Another aspect typical of the GMP guidelines is the traceability of the whole process: every single part and step of the production must be traced and documented in a proper manner.

The document regarding the GMP was redacted by the World Health Organization (WHO).

Nowadays, it represents one of the most used reference followed by pharmaceutical companies and by regulatory agencies. The original version was, in some cases, reviewed: e.g. European countries follow a more stringent version with respect to that redacted from WHO. Such version is called Eudralex and includes the 2003/94/EC Directive [5]. Conversely United States, Canada, Australia, Japan, follow the version issued by the FDA (Food and Drug Administration) [6]. The GMP system is not static: it adapts and evolves along with new technologies and new challenges thus ensuring an increase in security level.

However, a harmonization process in production and quality control of the drug was needed due to the existence of complex and diversified guidelines in the different Countries. Therefore, the ICH guidelines (International Conference on Harmonization of technical requirements for registration of pharmaceuticals for human use) were defined. Such guidelines bring together the authorities responsible for drug regulation in Europe (EMA), USA (FDA) and Japan (Ministry of Health of Japan) and experts from the pharmaceutical industry. ICH procedures aim to achieve greater harmonization in the interpretation and application of technical guidelines and requirements for registration of pharmaceuticals for human use.

In a GMP environment, there is the need to validate the entire production process and qualify the critical steps: production and analytical methods used for the control process, instruments/ systems/ software used, and training of staff involved must be documented, updated and verifiable. All phases of setup and validation of the method have to be set in GMP compliance, all process registered to ensure the correct storage of data products during tests, both in electronic or paper form.

Analytical methods require specific guidelines, such as ICH Q2 (R1) *"Validation of Analytical Procedures: Text and Methodology"*[7], in which all parameters that need to be tested during validation of a new method are defined.

Quality Control and Regulatory Authorities

Biopharmaceutical production is a complex process, that needs a strict Quality Control system during all production phases. Recombinant drugs are usually produced using cell banks with different origins, such as Mammalian, Bacterial or Yeast cell banks.

The biopharmaceutical production cycle starts on a small scale from a host cell line, in which one or more copies of a recombinant plasmid are inserted.

The plasmid contains the coding sequence of a protein (active principle), which will become a drug. The cells are cultured under defined and controlled conditions and grown in large-scale, up to a specified number of steps. During the growth, cells replicate and produce large quantities of active principle. At the end of the production process, the active principle is present inside the broth. It will be subsequently extracted, filtered and concentrated to the final formulation of the biotechnological drug, ready to be commercialized on the market.

For each production cycle (batch of production) it is necessary to test all parameters defined on guidelines before the commercialization.

The main national regulatory authorities are US-FDA (Food and Drug Administration) and EMEA (European Medicinal Evaluation Agency), that suggest type of tests to be performed, such as cell-substrate origin, composition and culture media origin, the recombinant product nature and the main stages of isolation and purification process. The two authorities are also available to discuss specific cases with the producer before the beginning of industrial production phases.

The Quality Control of a recombinant cell bank is carried out at Merck Company with the purpose to verify some characteristics, such as Purity, Identity and Genetic Stability and are critical aspects in Pharma manufacturing.

1. **Purity tests** are performed to verify/confirm, into the recombinant cell bank, the absence of contaminants (accidentally introduced during the process) like bacteria, viruses, mycoplasma, bacteriophages, and fungi.
2. **Identity tests** are intended to verify/confirm that, during the entire production process, the cell banks maintain intact the typical inherent characteristics of the original cell line, such as morphology, vitality of belonging strain and typical biochemical markers.
3. **Genetic Stability tests** verify that the cell banks used have not mutations at the construct level, at 5’-3’ flanking regions or into the sequence of the product.

All these evaluations are necessary to have a complete overview of the entire production process and final drug; therefore it is crucial to develop analytical tests in order to evaluate the safety of biotechnological products in Good Manufacturing Practices (GMP), according to Regulatory Authorities rules.

The guideline requires that Identity and Genetic Stability tests, on microbial cell banks (bacterial and yeast), have to be carried out during the entire production cycle, in order to ensure that the recombinant cell bank used, has not met cross-contamination with other microorganisms and thus maintains constant Genus, Species and Strain.

For this purpose, during a production cycle, samplings of microbial cell line are made at specific stages/times of the production, as follows:

- **Master Cell Bank (MCB):** initial stage of the cell line with uniform composition, derived from a single clone developed in defined culture conditions, storage and dispensing.
- **Working Cell Bank (WCB):** development stage of the cell line, corresponding to the actively producing stage, deriving from one MCB aliquot.
- **Post Production Cell Bank (PPCB):** the latest stage of the cell line at the end of their production activity, becoming from WCB expansion.

Yeast Cell Banks

The methylotrophic yeast *Komagataella* (*Pichia*) *pastoris* has become one of the most used cell factories to produce recombinant proteins over the last three decades. This success story is linked to its specific physiological traits, such as the ability to grow at high cell density in inexpensive culture medium and to secrete proteins at high yield[8].

Komagataella (Pichia) pastoris has many biotechnological applications. Initially, *P. pastoris* has been developed as a source of single-cell protein because the species can grow on either glucose or methanol, and high cell densities can be maintained under fermentation conditions[9].

Later, a gene expression system was developed to produce large quantities of medically and industrially important proteins[10].

Thus, there are many reasons for the popularity of the *P. pastoris* expression system, but two aspects of the species are most compelling and have contributed to its utility:

1. *P. pastoris* assimilates methanol, as the expression system is linked with alcohol oxidase, which is abundantly produced in the presence of methanol[11]. This is an unusually efficient and tightly regulated promoter from the alcohol oxidase I gene (AOX1) that is used to drive the expression of the foreign gene[12]. The AOX1 promoter is strongly repressed in cells grown on glucose and most other carbon sources but is induced over 1000-fold when cells are shifted to a medium containing methanol as a sole carbon source. The ability to repress expression of the foreign protein is advantageous if the protein is toxic to the cell (as many recombinant proteins are when synthesized at high levels);

2. *P. pastoris* expression system prefers a respiratory rather than a fermentative mode of growth. Fermentation techniques were developed for maintaining extremely high cell densities in excess of 100 g/l dry weight. In fact, fermentation products include ethanol and acetic acid, which quickly reach toxic levels in the high cell density environment of a bioreactor with strongly fermentative organisms. As a result of these features, a growing number of researchers are employing this methylotroph to produce proteins.[13]

In the biotechnological drug production using recombinant yeast cell banks, such as *Pichia pastoris* system, the identity of cell lines at the species and strain level must be confirmed. This control ensures that the recombinant yeast cell line used for drug production is the same in terms of genus, species and strain of origin during the entire manufacturing process.

The verification of the cell line identity is currently performed through two assays: sequencing of rDNA regions based on Sanger technology and SNPs identification through the Random Amplification of Polymorphic DNA (RAPD) methods. Combining these two techniques allows the identification of the cell line genus, species and strain of origin.

Microbial cell bank genus and species confirmation by Sanger sequencing method

The *MicroSeq* method is considered, by the Health Authorities, the "gold standard" for species identification. MicroSeq is based on Sanger technology and identifies the genus and species of cells that constitute a microbial cell bank used to produce biotechnological drugs, in order to confirm their identity.

In particular, the identification of microbial cells by DNA sequencing is carried out for the first microbial clone, engineered for the biotechnological drug production (Master Cell Bank, MCB), for the microbial clone, after several replication steps, during the drug production process (Working Cell Bank, WCB), and for the microbial clone at the end of its production phase (End of Production/Post Production Cell Bank, EoP/PPCB).

The purpose of the identification is to confirm the identity of the microbial cell bank during all production cycle stages. Microbial samples are cultured in liquid media, then the total DNA is extracted using the QIAamp DNA mini kit (Qiagen). Subsequently, the DNA is quantified and subjected to a PCR reaction using universal primers that can determine the amplification of a specific portion based on the type of microorganism. For bacterial identification, the primers target is the 16s ribosomal DNA, where the amplification size is around 500 bp, while for yeast, fungi or mold identifications the primers target is the D2 LSU ribosomal DNA, where the amplification size is around 300 bp. These portions, whose nucleotide composition is highly specific to each microorganism, are sequenced and subsequently identified using specific software, which determines the genus and species by comparing the sequence with sequences contained within a reference database.

For the amplification steps and DNA sequencing, the kit "*MicroSeq 500 16S rDNA PCR kit*", "*MicroSeq 500 16S rDNA Sequencing kit*" or "*MicroSeq D2 rDNA Fungal PCR kit*", "*MicroSeq D2 rDNA Fungal Sequencing kit*" (Applied Biosystems) are used. For the identification analysis, the specific software is the "MicroSeq ID Analysis Software" (Applied Biosystems).

The analysis result is the genus and species of the microorganism of interest, with an indication of the homology percentage (Match %) of the analyzed sequence with one or more reference sequences contained into the database. Higher the Match% and greater is the homology of the sequence to be identified with the reference sequence, and therefore a more accurate identification.

The analysis software also provides, for each identification, a numerical score (Specimen Score) that refers to the quality of the analyzed nucleotide sequence. Higher Specimen Score value corresponds to a greater identification accuracy of the microorganism at the genus and species level.

The method has been classified, in accordance with ICH Q2 (R1) guidelines[7], as identification tests. For its use, as part of the validation exercise as identification test for environmental samples and as In Process Control (IPC), the validated parameters were: Specificity, Precision (Repeatability), Accuracy and Robustness.

Microbial cell bank strain confirmation by RAPD method

The Random Amplification of Polymorphic DNA (RAPD) method is currently used as an identity test with the aim to detect the presence of any microbial cell contaminants within an engineered microbial cell bank used to produce biopharmaceutical drugs and belonging to a previously defined strain. In particular, the presence of any contaminant strains within a microbial bank is detected by comparing the genetic profile obtained by the RAPD method of the first engineered microbial clone for the biotech drug production (Master Cell Bank, MCB), of the engineered microbial clone after undergone several replication steps during the drug production process (Working Cell Bank, WCB) and the microbial clone at the end of its production phase (End of production/Post Production Cell Bank, EoP/PPCB).

As part of this test, the genus and species of each microbial cell bank should be previously tested and confirmed by the Sanger sequencing method and subsequently, the cell bank is analyzed in order to obtain one specific genetic profile ("pattern").

The obtained genetic profile thus allows to define the belonging strain of the microorganism. It must be confirmed, by the comparison among profiles, that the strain has remained "pure" during its entire life (MCB, WCB, and EOP/PPCB), without any contamination occurred by other microorganisms having different strain.

Microbial samples are cultured in liquid media, then the total DNA is extracted. A PCR reaction is performed using six decanucleotide primers, guanine and cytosine-rich, which link polymorphic regions (SNPs) within the extracted genomic DNA. The PCR reaction originates to an amplification pattern, consisting of a series of DNA bands, separable and visible on agarose gels, that can be analyzed using an imaging system.

The amplification reaction is performed by the *Ready-to-go RAPD Analysis kit* (GE Healthcare), which contains reagents and primers for the PCR reaction. The obtained genetic amplification profile is specific to each microbial strain. The Master Cell Banks (MCBs) amplification profile is compared to the Working Cell Banks (WCB) and/or End of production/Post Production Cell Bank (EOP/PPCB) amplification profiles, having the purpose of confirming its identity and excluding any other microorganisms' contamination among the production cycle stages.

The method involves the use of a negative control, represented by culture media, and a positive control, represented by DNA of *E. coli BL21* or *E. coli C1a* (provided in the Ready-to-go RAPD Analysis kit), that are processed together with the samples.

The profiles similarity or diversity among samples and controls is assessed comparing the amplification bands gained from the same primer-type reaction (e.g. comparison of all the amplification profiles gained from primer 1, then comparison of all those gained from primer 2, then among those from primer 3, primer 4, primer 5 and finally among those from primer 6). In this comparison context, the number of different amplification bands found between samples and controls is defined for each primer. The sum of all different bands across all amplification profiles (from primer 1 to primer 6) determines the total differences number between patterns.

In order to correctly compare the amplification profiles, the amplification products got from each primer have to be run on the same agarose gel, to apply the same electrophoresis conditions and limit differences that may be caused by changes in running settings.

In addition, two or more samples are considered as belonging to the same strain when the test is valid, and their amplification band profile differs by less than 6 bands from the reference host cell. Such different bands can come from amplification obtained from one or more primers.

Two or more samples are considered as belonging to a different strain when the test is valid, and their amplification band profile differs by 6 or more bands. Such different bands can come from amplification obtained from one or more primers.

The method has been classified, in accordance with ICH Q2(R1) guidelines[7], as a limit test for impurities. As part of the validation activities, the parameters tested were the Specificity, the Limit of detection (LOD), Robustness, Intermediate Precision, Precision (Repeatability), and Accuracy.

As part of the validation protocol, the tests required in the method validation were performed as an identity test to identify the strain of the microbial cell banks used for recombinant drugs production.

Following the validation process results, the RAPD method found to be specific, robust, precise, repeatable and accurate. The defined Limit of detection, understood as the smallest percentage of the contaminant microbial strain detectable within the microbial strain being analyzed, is 25% compared to the total amount of analyzed sample.

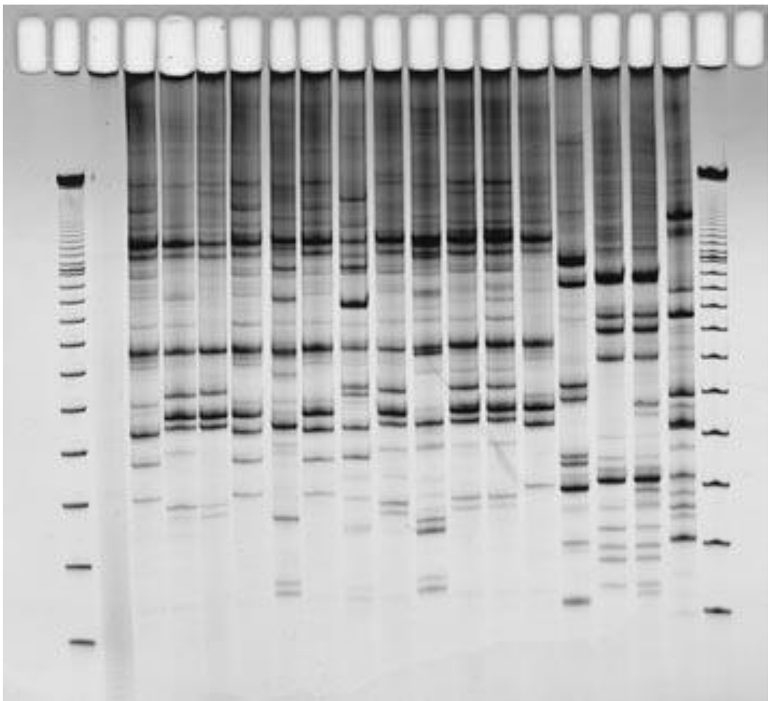


Figure 2 - Example of a RAPD analysis of microbial strains

The Health Authorities consider the combination of these two assays (Sanger sequencing and RAPD) as the "gold standard" for the cell line identification at the species and strain level. However, both techniques exhibit various drawbacks.

Nowadays, a more innovative, powerful and straightforward technology is available: Next Generation Sequencing (NGS), which is transforming the landscape of clinical microbiology and public health laboratories[14,15].

NGS represents the new "state of the art" in the sequencing field. This technique solved many problems of Sanger sequencing, increasing throughput while reducing time spent per run drastically[16]. A variety of applications in both DNA and RNA sequencing are possible: whole genome, exome, targeted and de-novo sequencing can be listed as main DNA sequencing applications, while total RNA, mRNA and small RNA sequencing are the basic applications in RNA sequencing[17,18,19].

Ph.D. Project

During the First Ph.D. year, the work was focused on the development of a new method to confirm the identity of recombinant yeast cell banks. The development included the study of the operational conditions, reagents, volumes and instruments necessary to standardize the method and the obtained results.

Firstly, the feasibility study tests were performed using *Pichia pastoris X-33* (*Komagataella pastoris*) standard strain (Invitrogen) and negative control, consisting of YPD Medium spiked with *Escherichia coli*. Moreover, among the tested recombinant yeast cell banks, some were considered as "real samples", such as Master Cell Bank (MCB) and End of Production Cell Bank (EoPCB) belonging to *Pichia Pastoris X-33* strains engineered with two different Molecules A and B[1].

Secondly, the aim was drawing a method able to distinguish different species and strain. For this reason, a group of standard strains (i.e. ATCC – DBVPG [(2)]) were chosen and divided into two subgroups: one phylogenetically closer to *Pichia Pastoris X-33* and another one phylogenetically more distant. Genome information was acquired through NCBI web database[20].

The Next Generation Sequencing (NGS) technology, MiSeq platform (Illumina), was used to determine the genus, species and strain of yeast cell banks. Two different instruments for DNA quantification: NanoDrop1000 (Thermo Fischer Scientific) and Qubit (Life Technologies) were compared. The library preparation protocol using Nextera Kit (Illumina) was optimized to increase the efficiency and standardize the library preparation. In details, the ideal reagent and DNA concentrations were defined with the aim to obtain the best average library size to optimize sequencing results. Moreover, the average library size calculation was standardized using both Qubit (Life Technologies) and Bioanalyzer (Agilent) instruments and values were re-elaborated using a validated Excel sheet. During the experimental design, two different genomic DNA extraction kits, YeaStar™ Genomic DNA Kit (Zymo Research) and Yeast DNA extraction kit (Thermo Fischer Scientific) have been evaluated to guarantee routine activities also in case of discontinuing of one extraction kit.

After setting the technical conditions, one of the main difficulties met was related to the standardization of the bioinformatics data analysis. Furthermore, give to all operators the possibility to perform their analysis by themselves, without Bioinformatic competencies was also challenging.

The aim was the standardization systems for Bioinformatics Data Analysis to be used in a GMP environment. Therefore, customized pipelines were developed and then validated in collaboration with a Bioinformatic programmer. These two pipelines are:

- "*DrAll - 1.0 version*" pipeline was developed for species identification: all the reads deriving from the sample were aligned with a genomes sequence database;
- "*BeTween - 2.0 version*" pipeline was developed for strain confirmation: the sample was aligned with the expected reference strain sequence for the strain confirmation.

For the pipeline designs, several basic tools (open source) present on the web were used such as Bowtie, Perl, Python and LaTeX. The yeast genome database used for the "*DrAll*" pipeline was customized starting from the information published on NCBI. We distinguished between chromosomal (Chrs) and mitochondrial (Mt) sequences. The genomes database was called "*Yeast_database*" and was stored on a dedicated server with only read access by operators in compliance with data integrity requirements.

[1] Molecules A and B: Due to not disclosure agreement Merck molecules are renamed using "A" and "B" code;

[2] Vegetal Biology Department of Perugia;

During the Second Ph.D. year, experimental conditions, reagents, volumes, instruments, acceptance criteria and method workflow were defined. In addition, the validation strategy was discussed and established according to the International Council for Harmonization (ICH) guideline[7].

One of the most critical steps in a method development is the definition of the "acceptance criteria". Acceptance criteria are specified indicators or measures employed in assessing the ability of a component, structure, or system to perform its intended function[21].

Analyzing all the results obtained during the method development, and considering that these parameters should be universally valid despite the sample origin could variate, the following acceptance criteria were established to confirm the genus, species and strain of a sample:

- For the genus and species identification: the sequenced sample genome must obtain a percentage coverage (%cov) ≥ 85 %;
- For the strain confirmation: the sequenced sample must obtain a number of mismatches ≤ 2.

Eight tests were carried out during the method setup. The collected results are summarized in the following paragraphs.

The genomic DNA extraction efficiency has been evaluated comparing two different extraction kits: YeaStar™ Genomic DNA Kit (Zymo Research) and YEAST Extraction DNA kit (Thermo Fisher Scientific). Both kits were effectively able to extract genomic DNA from yeast starting from 2 mL sample volume. Nevertheless, the YEAST Extraction DNA kit (Thermo Fisher Scientific) has been selected for validation scopes since the extraction efficiency is higher. Consequently, a higher gDNA concentration is available for the subsequent steps.

The Nextera (Illumina) library preparation protocol was optimized to obtain a library with higher average sizes (in bp). To that end, the starting gDNA amount was increased and the Tagmentation DNA Enzyme1 (TDE1) volume was modified as follows:

- Original protocol: gDNA starting concentration: 50 ng; 5µl of TDE1;
- Optimized protocol: gDNA starting concentration: 75 ng; 2µ of TDE1 + 3µl of H2O ultrapure.

Larger libraries (size average range: 600-1000 bp) were obtained with the new protocol and a better reading quality on the MiSeq sequencer (Illumina) was observed. Furthermore, the MiSeq run settings were modified using a 2x151 bp reading to further implement the sequencing quality results.

The culture broth formulation (YPD broth), normally used for yeast cell banks growth, contains yeast extract that could negatively impact on the sample sequencing result. Due to this reason, the ability of the YPD formulation to produce DNA sequences was evaluated.

The YPD medium was contaminated with *Escherichia coli* bacterial cells to increase the DNA extraction efficiency. This sample led to a library with yeast genomic DNA fragment traces. In addition, the reads from YPD broth were not enough to ensure a coverage percentage value in compliance with the acceptance criteria (Cov% above 85%).

The method ability to detect other yeast contaminants in a mixture of two different yeast cell lines (different genus and species, or different strains) was evaluated. A mixture of *Pichia pastoris* X-33 with *Saccharomyces cerevisiae S288c* (for Genus and Species) was analyzed, using *DrAll* pipeline, and a mixture of *Pichia pastoris X-33* with *Pichia pastoris GS115* (for Strain) was analyzed using *BeTween* pipeline. The obtained result shows the 10% limit of detection as threshold. To confirm the LOD value, a mixture of *Pichia pastoris X-33* with *Pichia pastoris KM71H* has been analyzed during the method validation tests.

The robustness tests of the "Genus, species and strain identification of yeast cell banks by next-generation sequencing method" were defined through a risk-based evaluation, in which each experimental phase has been considered. This risk assessment evaluated each critical step analyzing whether changes in specific parameters can influence the method performance and defined which phase needs to be tested during the robustness validation tests. The evaluation of the experimental phases has been done according to the Failure Modes, Effects and Criticality Analysis (FMECA) model[22],[23].

For each step have been considered:

- the **probability** that a parameter variation foreseen by the procedure will occur;
- the **severity** that this variation would involve;
- how the consequences of this variation can be detected (**detectability**).

Based on the risk analysis, the following steps were identified as critical and have been analyzed deeply in the method validation:

1. **Extracted DNA quantification**. An inaccurate quantification of genomic concentration could have a direct impact on the average size of the produced library and, consequently, on the read quality produced by MiSeq sequencer (Q30%). Two libraries were prepared starting from modified gDNA concentrations compared to the standard condition. The first sample was processed starting from a lower concentration (50 ng) and the second from a higher concentration (100 ng) compared to the defined condition of 75 ng.

2. **Library preparation.** The library preparation step is fundamental to obtain a clusters density in the optimal range (defined in the acceptance criteria as Cluster Density: 600 e 1900 K/mm2). Variation in one library preparation step (e.g. Tagment DNA enzyme 1 volume) could affect the result and not meet the defined acceptance criteria. Thus, the method ability to be unaffected by predetermined Tagment DNA enzyme 1 (TDE) volume changes was tested. A range of TDE1 volumes has been verified checking if it produces valid libraries to be loaded on MiSeq sequencer. Two samples were fragmented using 1,6 µl or 2,4 µl of Tagment DNA enzyme 1 volume compared to the defined condition working at 2 µl.

Some robustness tests were performed during validation activities with certified standard microorganisms. At the end of the method setup, the experimental standard procedure and the validation protocol were written. Moreover, a complete software validation for the two bioinformatic pipelines (*DrAll v1.0* and *BeTween v2.0*) was performed ensuring the results consistency. The pipeline configuration was performed following also all the rules to ensure Data Integrity.

During my Third and last Ph.D. year, the method validation strategy has been designed and the validation exercises performed.

The method is classified as an “identification test”, although it could be defined as a “limit assay for impurities” too. Following the ICH guideline for an identification test and limit assay for impurities, the method robustness, specificity and limit of detection (LOD) should be determined and verified during validation tests[7].

The test could be divided in 3 phases:

- 1° phase: Identify yeast cell bank Genus and Species;
- 2° phase: Confirm the yeast cell bank Strain;
- 3° phase (optionally): Compare different recombinant cell bank production stages (MCB, WCB and PPCB / EoPCB) results.

Specificity: Meant as the assay ability to discriminate genus and species and strains genomes comparing the sample reads with reference genomes. For the genus and species determination, the specificity was verified by comparing different standard samples genomes with all the yeast database genomes; for the strain determination, the specificity was verified by comparing different strains with the reference host cell genome.

Robustness: Meant as the ability of the method to remain unchanged despite the introduction of small deliberate changes in the parameters of execution, was verified introducing some variations in some critical steps. For the robustness evaluation, some of the variants were tested in the validation tests, while others have been tested during the preliminary setup activities.

In addition, the following parameter required for impurity limit tests has also been tested:

Limit of Detection (LOD): Meant as the lower % of microbial contaminants detectable in the sample, was verified analyzing the sequences obtained from a mixture of 2 different genus, species or 2 different yeast strains at different percentages.

Materials and Methods

Materials

Yeast Cell Lines

The following yeast cell lines were used in the validation exercise, as example of cell lines used to create engineered yeast cell banks:

- X-33 *Pichia pastoris* Yeast Strain (Invitrogen)
- GS115 *Pichia pastoris* Yeast Strain (Invitrogen)
- KM71H *Pichia pastoris* Yeast Strain (Invitrogen)
- *Saccharomyces cerevisiae* S288c (ATCC204508)
- *Saccharomyces cerevisiae* CBS1171 (ATCC18824)
- *Cyberlindnera jadinii* (ATCC18201)
- *Candida glabrata* (DBVPG3178)
- *Eremothecium gossypii* (ATCC10895)
- Recombinant *Pichia pastoris X-33* cell bank (manufacturing site)

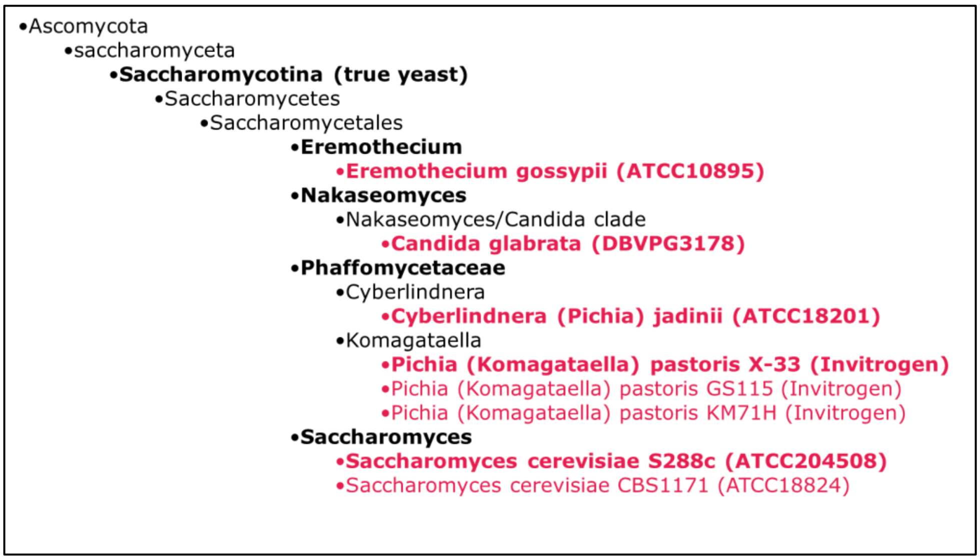
Pichia pastoris has been earlier reassigned in literature to the genus *Komagataella* following phylogenetic analysis of gene sequences.[11]

Below is reported the phylogenetical tree of the selected microorganisms (Figure 3). The selection includes a group of standard strains (i.e. ATCC - DBVPG) based on the phylogenetically distance with the test item *Pichia pastoris X-33* and an engineered cell bank provided by the manufacturing site.

In fact, based on the method objectives, the aim is demonstrating the ability of the method to distinguish among different genus and species and strains.

Therefore, the selected cell lines can be divided into two subgroups:

- First group phylogenetically closer each other, composed by yeasts belonging to the same Genus and Species but different Strain:
    - *P. pastoris X-33, GS115* and *KM71H;*
    - *S. cerevisiae S288c* and *CBS1171.*
- Second group phylogenetically distant to *Pichia pastoris X-33*, composed by yeasts belonging to different genus and species such as *Saccharomyces cerevisiae*, *Cyberlindnera jadinii*, *Candida glabrata* and *Eremothecium gossypii*.



- Ascomycota
  - saccharomyceta
    - **Saccharomycotina (true yeast)**
      - Saccharomycetes
        - Saccharomycetales
          - **Eremothecium**
            - **Eremothecium gossypii (ATCC10895)**
          - **Nakaseomyces**
            - Nakaseomyces/Candida clade
              - **Candida glabrata (DBVPG3178)**
          - **Phaffomycetaceae**
            - Cyberlindnera
              - **Cyberlindnera (Pichia) jadinii (ATCC18201)**
            - Komagataella
              - **Pichia (Komagataella) pastoris X-33 (Invitrogen)**
              - Pichia (Komagataella) pastoris GS115 (Invitrogen)
              - Pichia (Komagataella) pastoris KM71H (Invitrogen)
          - **Saccharomyces**
            - **Saccharomyces cerevisiae S288c (ATCC204508)**
            - Saccharomyces cerevisiae CBS1171 (ATCC18824)

Figure 3 – Yeast taxonomic classification

Reagents and Consumables

2, 10, 20, 100, 200 and 1000 µl single channel pipettes
Tips with filter
0.2, 0.5, 1.5 and 2 mL tubes
Tube racks
Magnetic rack
Ultrapure water (MilliporeSigma)
Ethanol (Merck)
Yeast DNA Extraction Kit (Thermo Fisher Scientific):
Y-PER^{TM} Reagent
DNA Releasing Reagent ADNA Releasing Reagent B
Protein Removal Reagent
RNase A (Roche)
Nextera kit (Illumina):
Tagment DNA buffer
Tagment DNA enzyme 1
Nextera PCR master mix
PCR primer cocktail
Resuspension buffer
AMPure XP Beads (Beckman coulter)
Nextera index (Illumina)
Qubit® 2.0 dsDNA HS/BR assay kit (Life technologies)
Bioanalyzer chip: DNA 12000 series II (Agilent)

Instruments

Thermoblock heater
Heated orbital shaker
Tube centrifuges
Vortex

PE GeneAmp 9700 Thermal Cycler (Applied Biosystems)
Biohazard laminar flow hood
NanoDrop ND-1000 Spectrophotometer
Qubit® 2.0 Fluorimeter (Life Technologies)
+4°C refrigerator
≤ -20°C freezer
MiSeq (Illumina)
Bioanalyzer (Agilent)

Method workflow

The method workflow can be summarized in the following steps:

1. Genomic DNA Extraction
2. DNA Quantification
3. Library Preparation
4. Library Assessment
5. Sample loading on MiSeq platform
6. Data analysis

Genomic DNA extraction

Sample preparation starts with extraction of genomic DNA from the yeast cell bank, using the “Yeast DNA Extraction Kit” (Thermo Fisher Scientific) or alternatively, can be done with the “YeaStar™ Genomic DNA Kit” (Zymo Research).

Extraction is performed starting with the whole content of a cryovial (about 2 mL) directly prepared and shipped from the manufacturing site.

gDNA Extraction using the “Yeast DNA Extraction Kit” (Thermo Fisher Scientific)

For the genomic DNA extraction proceed as follow:

Centrifuge the cells at 5000xg for 5 minutes at room temperature and discard the supernatant. Resuspend the pellet with 500 µL of Y-PER reagent (about 8 µL/ 1 mg of pellet) and mix by vortexing or pipetting until it becomes homogenous. Incubate at 65°C for 10 minutes.

Centrifuge at 13000g for 5 minutes, and then discard the supernatant. Add 400 µL of DNA releasing reagent A and 400 µL of DNA releasing reagent B and mix by vortexing or pipetting until it becomes homogenous. Incubate at 65°C for 10 minutes.

Add 200 µL of Protein Removal Reagent and mix. Centrifuge at 13000g for 5 minutes. Transfer the supernatant to a new 1.5 mL tube. Add 600 µl of Isopropyl Alcohol and mix by inverting. Precipitate the genomic DNA by centrifuging at 13000g for 10 minutes. Remove the supernatant having care not to remove also the pellet which is transparent and difficult to see. Add 1 mL of 70% ethanol to the pellet.

Mix and centrifuge at 13000g for 1 minute. Invert the tube and dry from ethanol residues. Resuspend in 50 µL of molecular biology grade water. The pellet should dissolve completely in 5 minutes.

Gently mix the bottom of the tube or pipette slowly. Wash the sides of the tube to keep all the genomic DNA to the bottom. Get rid of any RNA by adding 1 µL of RNAse every 50 µL of extract, previously diluted 1:10, and incubate the samples at 37°C+/-1°C for 30 minutes;

Store the extracted DNA in a 2-8°C refrigerator for no longer than one working day, or at ≤ -15°C.

gDNA Extraction using the "YeaStar™ Genomic DNA Kit" (Zymo Research)

Transfer 1 mL of yeast cells to a 1.5 mL tube, centrifuge at 500g for 2 minutes. Remove the supernatant and add the remaining volume of the culture to the same tube. Centrifuge again at 500g for 2 minutes and remove the supernatant.

For each pellet, add 120 µL of YD Digestion Buffer and 5 µL of R-Zymolyase and resuspend the pellet by vortexing. Incubate the samples at 37°C +/-1°C for 60 minutes. Add 120 µL of YD Lysis Buffer and mix the contents of the tube.

Centrifuge at 9391g (corresponding to 10000 rpm on the Centrifuge 5424 - Eppendorf) for 2 minutes. Transfer the supernatant which has formed to a Zymo-spin column and centrifuge at 9391g for 1 minute.

Add 300 µL of DNA Wash Buffer to each column and centrifuge at 9391g for 1 minute. Add a further 300 µL of DNA Wash Buffer and centrifuge at 9391g for 1 minute. Transfer the Zymo-spin column to a new 1.5 mL tube. Add 60 µL of ultrapure water, incubate at room temperature for 1 minute and then centrifuge at maximum speed for 10 seconds to elute the DNA.

Store the extracted DNA in a 2-8°C refrigerator for no longer than one working day, or at ≤ -15°C.

DNA Quantification

Following extraction of the genomic DNA, treat the sample with 1.2 µL of RNase A (Roche), previously diluted 1:10, to get rid of any RNA in the extract.

Quantify the concentration of extracted DNA using the Qubit® 2.0 (dsDNA BR Assay Kit or dsDNA HS Assay Kit) and prepare 75ng of DNA in 20 µL of ultrapure water for each sample.

Use Illumina Experiment Manager (IEM) software to check the indices to be used. Choose a different combination of index depending on the number of samples to be run together in the same sequencer flow cell, and check validity using the IEM program.

Library preparation

The Nextera protocol allows to prepare libraries to be sequenced using Illumina technology "Sequencing by Synthesis" (SBS). Libraries can be generally sequenced through one of the Next-generation Illumina sequencers: HiSeq, NextSeq 500 or MiSeq.

The method consists of dsDNA fragmentation through an enzyme reaction performed by Transposomes, binding of indexed adapters to fragmented DNA and the library amplification. During library preparation, a couple of indexed adapters is bind to each sample (Figure 4).

If during the sequencing phase, multiple libraries analysis is necessary, it is possible to create a pool and use the Illumina Experiment Manager software to verify that the indexes mixture could be correctly read by the sequencer.

The Nextera protocol has a total starting DNA amount of 50ng. DNA used for the library has to be double-stranded to enhance transposons fragmentation. In fact, these enzymes are able to produce dsDNA fragments with a length of about 300bp. It is, therefore, necessary to consider that the amplicons length must be greater than 300bp.

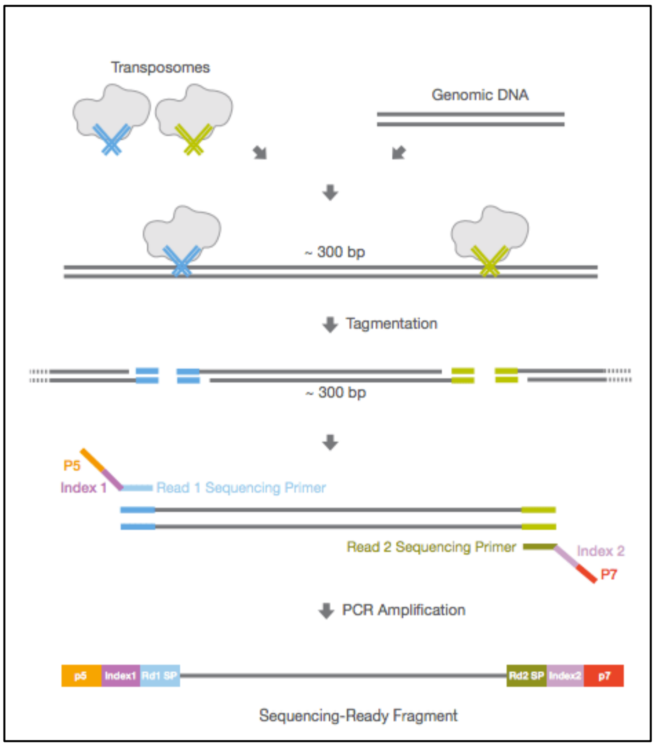


Figure 4 – Nextera protocol reaction scheme[24]

Library preparation using Nextera protocol consists of 4 steps:

- DNA Tagmentation
- Cleaning fragmented DNA using magnetic beads
- Fragmented DNA amplification
- Library purification using magnetic beads

The total starting DNA request by Nextera protocol is 50ng. However, during the method development was optimized to start with 75ng of genomic DNA in a total of 20µL for each reaction tube (3,75 ng/µL in reaction).

DNA tagmentation

The first step involves the fragmentation of the starting dsDNA through the transposons. Thaw at room temperature and then keep the following reagents in ice:

Tagment DNA Buffer (TD)

Tagment DNA Enzyme (TDE1)

Before you begin, you need to mix the reagents by reversing the tubes 3-5 times, centrifuge briefly. Then proceed as follows:

Add 20 µL of DNA to the concentration of 3.75ng/µL (75ng tot) in 0.2ml test tubes appropriately identified. Add 25 µL of TD buffer in each sample. Add 2 µL of Tagment DNA Enzyme 1 (TDE1) and 3 µL of ultrapure water. Mix to make the solution homogeneous. Place the samples in the thermal cycler and incubate with the program:

Total Volume: 50 µL
55 °C for 5 min

10 °C for ∞

If the source material is made up of amplicons, the incubation can be modified as follows:

Total Volume: 50 µL
58 °C for 10 min
10 °C for ∞

Cleaning fragmented DNA using magnetic beads

The fragmented DNA is purified through the AMPure XP Beads (Figure 5).

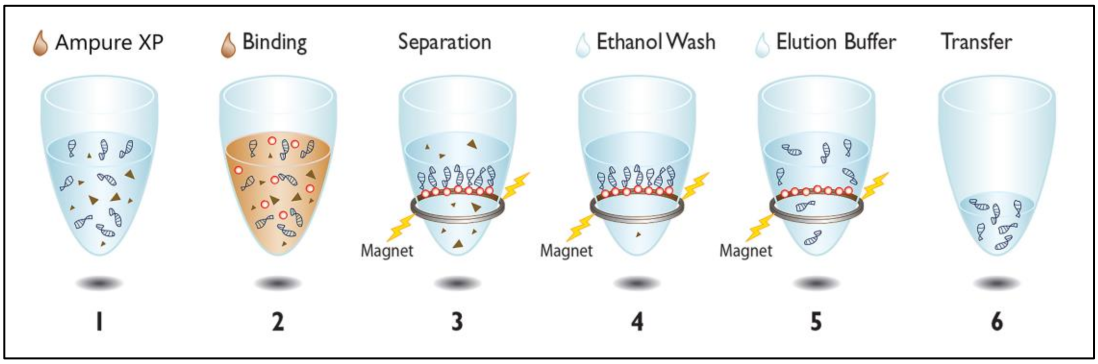


Figure 5 - AMPure XP Beads workflow[25]

Before starting, the AMPure XP Beads are kept at room temperature for 30 min before use. Vortex the beads making sure that are well resuspended. Thaw the Resuspension Buffer (RSB) and keep it at room temperature until the end of the procedure. Prepare an 80% fresh EtOH solution. Calculate 400 µL for each sample.

After incubation, transfer samples from 0.2ml test tubes to 1.5ml (low retention) appropriately identified. Add 45 µL of beads to each sample and mix. Incubate the samples for 10 min at room temperature. Then place the samples in the magnetic rack for 5 min or until the liquid becomes clear (the beads collect near the magnet).

Remove all the supernatant from each sample gently without touching the beads.

Then two washing steps with 80% EtOH solution are needed. Add 80% of the EtOH solution to each sample by not touching the beads. Leave at room temperature for 30 seconds. Remove the supernatant without touching the beads. Add again 200 µL of 80% EtOH solution by not touching the beads. Leave at room temperature for 30 seconds. Remove the supernatant by not touching the beads. Make sure to have removed all the ethanol excess. Leave the samples in the rack with an open cap to dry for 15 min. Remove the samples from the magnetic rack and suspend the beads in 22.5 µL of Resuspension buffer. Mix gently and centrifuge at low speed and for a few seconds the samples. Incubate samples for 5 min at room temperature. Place the samples in the magnetic rack for 2 min. Appropriately identify new 1.5 ml test tubes and transfer 20 µL of supernatant. Throw away the 1.5 ml tubes with the remaining beads.

Fragmented DNA amplification

Fragmented and purified DNA is now ready to be tied with Indexed Adapters and amplified.

If during the sequencing phase the libraries will be analyzed together in the same lane you will need to use the Illumina Experiment Manager software to verify that the indexes reading on the sequencer is Valid.

Thaw the following reagents at room temperature:

- Nextera PCR Master Mix (NPM)
- PCR Primer Cocktail (PPC)
- Index 1 (orange cap)
- Index 2 (white cap)

Mix the reagents by reversing the tubes 3-5 times, centrifuge briefly. Identify the 0.2 ml tubes appropriately. Add 5 µl of the Index 2 (white cap) to each sample. Add 5 µL of the Index 1 (orange cap) to each sample. Add 15 µl of NPM to each sample that contains the indexed adapters. Add 5 µl of PPC in each sample. Transfer 20 µl of fragmented and purified DNA to the solution and mix to make it homogeneous.

If necessary, centrifuge briefly. Place the samples in the thermal cycler and incubate with the program:

Total Volume: 50 µl
72°C for 3 min
98°C for 30sec
98°C for 10sec
63°C for 30sec } 10 cycles
72°C for 3 min
4°C ∞

Proceed to the next step. Alternatively, the reaction can be stored overnight in the heater or in the refrigerator for maximum two days before being purified.

Library purification using magnetic beads

After amplification, transfer samples from 0.2 ml test tubes to 1.5 ml (low retention) appropriately identified. Add 45 µL of beads to each sample and mix. Incubate the samples for 10 min at room temperature. Then place the samples in the magnetic rack for 5 min or until the liquid becomes clear (the beads collect near the magnet).

Remove all the supernatant from each sample gently without touching the beads.

Then two washing steps with 80% EtOH solution are needed. Add 80% of the EtOH solution to each sample by not touching the beads. Leave at room temperature for 30 seconds. Remove the supernatant without touching the beads. Add again 200 µL of 80% EtOH solution by not touching the beads. Leave at room temperature for 30 seconds. Remove the supernatant by not touching the beads. Make sure to have removed all the ethanol excess.

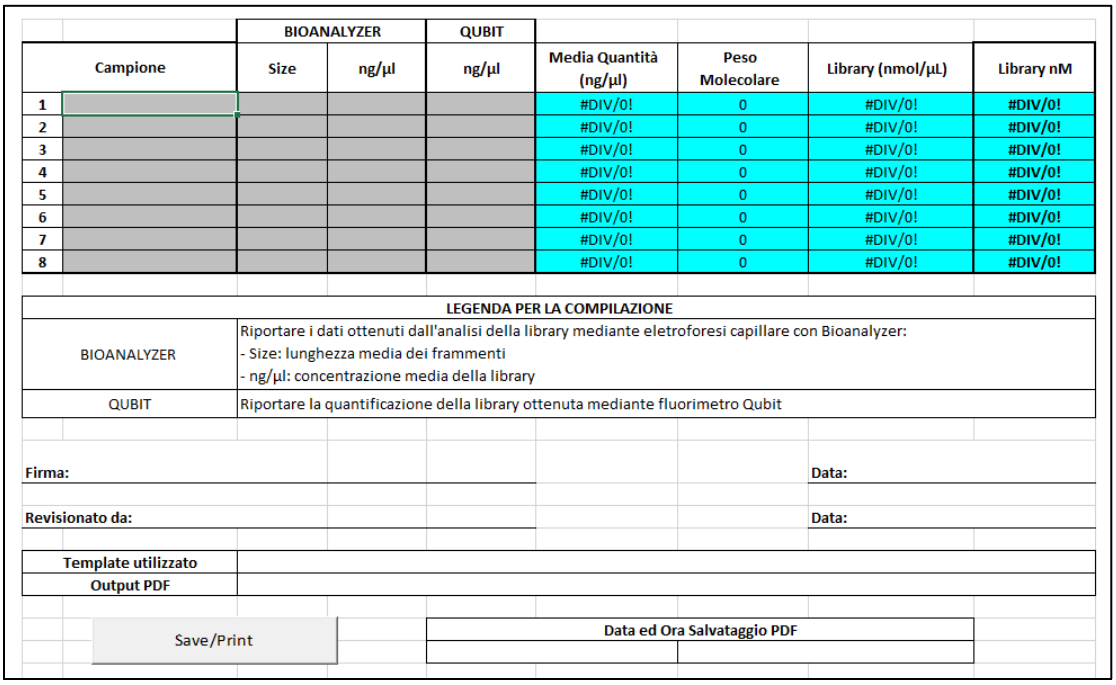
Remove the samples from the magnetic rack and suspend the pellets in 32.5 µL of Resuspension buffer. Mix gently and centrifuge at low speed and for a few seconds the samples. Incubate the samples for 2 min at room temperature. Place the samples in the magnetic rack for 2min or until the liquid becomes clear (the marbles collect near the magnet). Appropriately identify new 1.5 ml test tubes and transfer 30 µL of supernatant.

The generated libraries can be stored for up to a month at -20 °C. Libraries must be evaluated qualitatively and quantitatively before they are sequenced.

Library assessment

When preparation has been completed, quality must be assessed through quantitation with the Qubit® 2.0 (dsDNA HS Assay Kit) and the mean size of fragments produced must be checked using 1µl of library for analysis on the Bioanalyzer with Agilent DNA 12000 kit.

Use the following spreadsheet to calculate the mean of the concentrations read with the Qubit® 2.0 and those read with the Bioanalyzer. (Figure 6)



		BIOANALYZER		QUBIT				
	Campione	Size	ng/µl	ng/µl	Media Quantità (ng/µl)	Peso Molecolare	Library (nmol/µL)	Library nM
1					#DIV/0!	0	#DIV/0!	#DIV/0!
2					#DIV/0!	0	#DIV/0!	#DIV/0!
3					#DIV/0!	0	#DIV/0!	#DIV/0!
4					#DIV/0!	0	#DIV/0!	#DIV/0!
5					#DIV/0!	0	#DIV/0!	#DIV/0!
6					#DIV/0!	0	#DIV/0!	#DIV/0!
7					#DIV/0!	0	#DIV/0!	#DIV/0!
8					#DIV/0!	0	#DIV/0!	#DIV/0!

LEGENDA PER LA COMPILAZIONE	
BIOANALYZER	Riportare i dati ottenuti dall'analisi della library mediante eletroforesi capillare con Bioanalyzer: - Size: lunghezza media dei frammenti - ng/µl: concentrazione media della library
QUBIT	Riportare la quantificazione della library ottenuta mediante fluorimetro Qubit

Firma: Data:

Revisionato da: Data:

Template utilizzato	
Output PDF	

Data ed Ora Salvataggio PDF	

Figure 6 - Excel macro spreadsheet

If the produced libraries meet the acceptance criteria shown in the paragraph below, reagents can be prepared for loading on the MiSeq sequencer (Illumina).

Library acceptance criteria

Average Library Size greater than or equal to 200 bp.
Library concentration greater than or equal to 2 nM.

Sample loading on MiSeq platform

The Illumina MiSeq is a next-generation sequencer that is able to take advantage of the sequencing methodology called SBS (Sequencing by Synthesis). The MiSeq system integrates cluster generation, sequencing, and data analysis functions into a single tool.

With this new sequencing technology, DNA molecules and primers are first covalently adapted to a flow cell and amplified through a procedure called cluster amplification. Such a procedure requires a chemistry called bridge amplification.

Once the clusters are generated on the flow cell, the actual sequencing of the DNA fragments amplified using reversible marked terminator nucleotides takes place.

In each sequencing cycle, polymerase incorporates complementary nucleotide from the sequencing primer.

There are two CCD cameras on the instrument that can capture the fluorescence from fluorescent-marked terminator nucleotides. Subsequently, the fluorochrome is removed as well as the blockage of the base thanks to a chemical reaction. This allows the bond with a new nucleotide.

To determine the nucleotide sequence, the four terminated marked bases are added in the reaction to each cycle and the unincorporated nucleotides are removed by washing. Unlike other technologies, a single nucleotide is inserted at each cycle, thus increasing read accuracy and reducing the number of embedding errors. Cluster identification grid generation is the process by which cluster locations on the entire flow cell surface are defined according to X and Y coordinates.

After the cluster location grid is generated, images produced at each subsequent imaging cycle are aligned with the grid. The intensities of the individual clusters in all four nucleotide color channels can be extracted and base identifications can be derived from the normalized cluster intensities.

The number of sequencing cycles that can be performed on the machine depends on the type of protocol you choose, for example in a 2x150 pair-end protocol, the total number of cycles performed is 300.

In Figure 7, a brief diagram of how sequencing takes place in the MiSeq tool is represented.

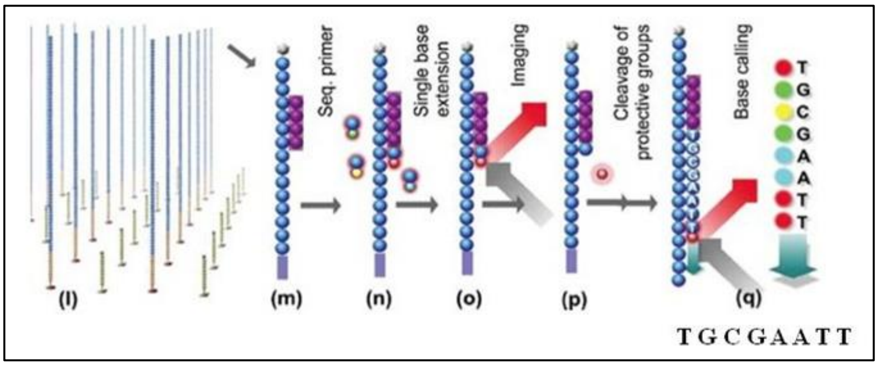


Figure 7 – Diagram of MiSeq sequencing[26]

Load the sample libraries on the MiSeq sequencer (Illumina) as described below.

The reagent cartridge is stored at -20 °C. It is, therefore, necessary to thaw it before use. This phase can be carried out in two ways: keeping the cartridge in the refrigerator at a temperature between 2 °C and 8 °C starting from the day before the start of the run or keeping it in a water bath at room temperature, without submerging it, until complete thawing.

In the package together with the cartridge is also included the HT1 buffer which is used to dilute the libraries. Once defrosted, this buffer must be kept at a temperature between 2 °C and 8 °C until it is used.

When the reagent cartridge is completely thawed and ready for use, you can load the libraries. Gently shake the cartridge three or four times to mix the components well before loading the libraries. Gently tap the cartridge on the bench to reduce air bubbles in the reagents.

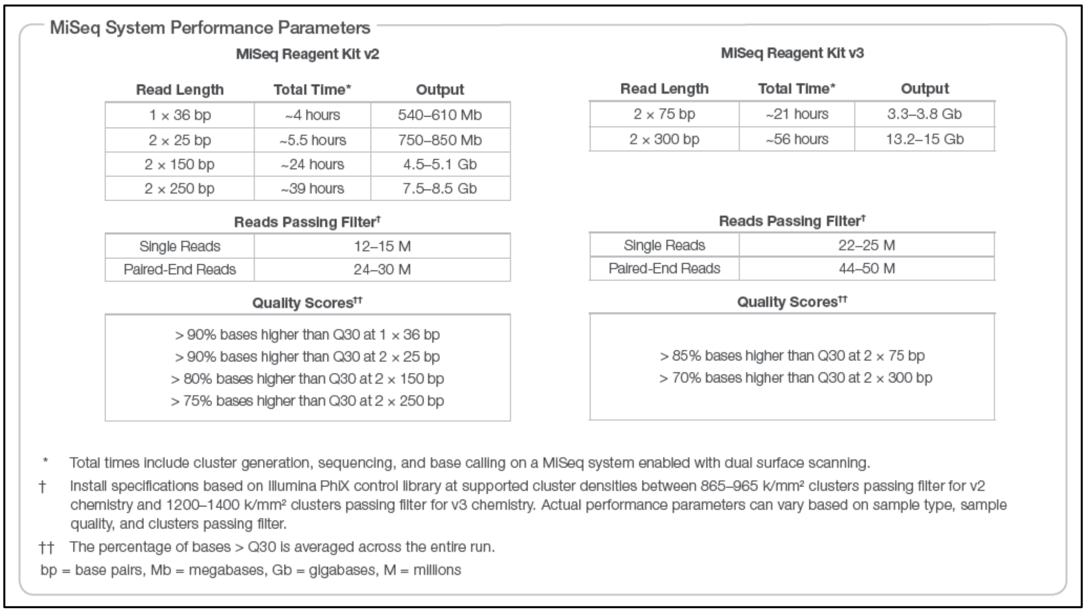
Before loading the libraries into the cartridge, it is necessary to create the pool of samples (since the instrument's flow cell has only one lane), denature it and dilute it. The libraries must initially be diluted to the concentration of 2 nM in buffer EB. The same amount of each library is taken to form the pool (for example 5 µl). The pool is denatured in a 0.2N NaOH solution (prepared at the time of use) for 5 minutes at room temperature (the denaturation concentration is 1 nM, then 10 µl of pool + 10 µl of NaOH).

At the end of the incubation the pool is diluted in HT1 buffer up to the loading concentration (for example 10 pM); these steps are made on ice.

Once denatured and properly diluted, the pool is ready to be loaded onto the cartridge. An adequate volume must be prepared since 600 µl of the pool must be loaded.

Use a 1 ml tip to drill the aluminum capsule that seals the tank marked with the words Load Samples. It is important that no other reagent positions are punctured. The other reagent positions are automatically punched during the sequencing run.

Samples must be run on a flow cell that produces at least a theoretical total of 13.2-15 Gb (with a read length of 2X300 bp). As run parameters in the MiSeq sequencer sample sheet, set a read length of 2x150bp, to have a theoretical 1.1-1.8 Gb per sample, loading 4-6 samples per Flow cell (Figure 8).



MiSeq System Performance Parameters

MiSeq Reagent Kit v2

Read Length	Total Time*	Output
1 × 36 bp	~4 hours	540–610 Mb
2 × 25 bp	~5.5 hours	750–850 Mb
2 × 150 bp	~24 hours	4.5–5.1 Gb
2 × 250 bp	~39 hours	7.5–8.5 Gb

Reads Passing Filter†

Single Reads	12–15 M
Paired-End Reads	24–30 M

Quality Scores††

> 90% bases higher than Q30 at 1 × 36 bp
> 90% bases higher than Q30 at 2 × 25 bp
> 80% bases higher than Q30 at 2 × 150 bp
> 75% bases higher than Q30 at 2 × 250 bp

MiSeq Reagent Kit v3

Read Length	Total Time*	Output
2 × 75 bp	~21 hours	3.3–3.8 Gb
2 × 300 bp	~56 hours	13.2–15 Gb

Reads Passing Filter†

Single Reads	22–25 M
Paired-End Reads	44–50 M

Quality Scores††

> 85% bases higher than Q30 at 2 × 75 bp
> 70% bases higher than Q30 at 2 × 300 bp

* Total times include cluster generation, sequencing, and base calling on a MiSeq system enabled with dual surface scanning.

† Install specifications based on Illumina PhiX control library at supported cluster densities between 865–965 k/mm² clusters passing filter for v2 chemistry and 1200–1400 k/mm² clusters passing filter for v3 chemistry. Actual performance parameters can vary based on sample type, sample quality, and clusters passing filter.

†† The percentage of bases > Q30 is averaged across the entire run.

bp = base pairs, Mb = megabases, Gb = gigabases, M = millions

Figure 8 - Performance parameters of Illumina flow cells[26]

Data analysis

The electronic data produced by the instrument are analyzed using bioinformatics pipelines. Pipelines are programs consisting of a series of instructions and/or computer commands given to the server to perform operations that compare the genomic sequence produced by the sample being tested with sequences in a database.

The host cell, communicated by the Sponsor site, is considered as a reference cell line. Therefore, the nucleotide sequence of its genome is used as the reference for comparison with the samples being tested.

Yeast Genus and Species determination

The genus and species of the sample are determined by comparing the sample genomic sequence with all the different yeast genus and species sequences present into the Yeast database.

DrAll v1.0 bioinformatics pipeline aligns all the fragments that make up a library according to the similarity between the sequence read and the corresponding one in the database. This pipeline is applied to obtain a coverage %, which correspond to the coverage percentage over the whole reference genome in the database.

The coverage % obtained allows the sample to be identified. The database consists of the genomic sequences of yeasts taken from the NCBI database[20].

If this alignment meets the acceptance criteria set for this test, the genus and species to which the test sample belongs can be attributed.

To attribute the identity of a sample with respect to genus and species, all the genomes that meet the following condition is considered:

- **Coverage %** greater than or equal to 85%;

To analyze the genus and species, it should first be checked that the genomic sequence of the yeast used for the production of the recombinant cell bank to be tested (host cell) is contained in the database. If this microorganism is not in the database, the host cell specified by the Sponsor site should be sequenced and added to the database. Samples can then be analyzed.

Yeast Strain determination

The sample Strain is determined by the comparison between the genomic sequences obtained from the various samples (e.g. MCB, WCB and PPCB/EoPCB) and the respective host cell indicated by the manufacturing site. The aim is to identify any mismatches or variations in one or more nucleotides of the genome, by "*BeTween 2.0 version*" bioinformatics pipeline.

The *BeTween v2.0* pipeline aligns all the reads obtained from sequencing according to the similarity between the sample sequence and the corresponding one of the reference host cell (see also paragraph ***BeTween 2.0 version)***.

BeTween v2.0 first compares each sample with the host cell sequence and assesses any mutations (Mismatch). Subsequently, a comparison can be made between the different samples tested to produce a result consisting of the number of Mismatches between, for example, the MCB, WCB and PPCB/EoPCB.

If this analysis meets the test acceptance criteria, the strain of the test sample can be confirmed.

In addition, *BeTween v2.0*, can exclude the reads corresponding to the sequence of the transgene (plasmid + construct) used during engineering of the cell line from all the sample reads. This function may or may not be used depending on the sample to be tested and allows a specific comparison of the sample genomic sequence with the host cell.

The Strain identity is attributed when the sample meets the following acceptance criteria:

- **Mismatches** fewer than or equal to 2;

Bioinformatic Pipelines

General Data Integrity rules

A described in FDA Guidance for Industry Data Integrity and Compliance with cGMP:

"*Data integrity refers to the completeness, consistency, and accuracy of data. Complete, consistent, and accurate data should be attributable, legible, contemporaneously recorded, original or a true copy, and accurate (ALCOA). Data integrity is critical throughout the CGMP data life cycle, including in the creation, modification, processing, maintenance, archival, retrieval, transmission, and disposition of data after the record's retention period ends. System design and controls should enable easy detection of errors, omissions, and aberrant results throughout the data's life cycle.*"[27]

In order to follow the Data Integrity rules and being GMP compliant, to access the computer, each operator must use the personal *User Name* and *Password*. For each pipeline, the information related to the operator name and login date/time are automatically registered into the system.

Data Integrity guideline required the operator to renew the password every 3 months, the system is set in order to automatically remind this action to the operator. As well as, for safety reason, each user can have access exclusively to the pipelines has been authorized, located in the "Pipelines" folder present on the computer desktop.

Sample Sheet generation

Before to start the bioinformatic pipeline, the creation of a sample sheet text file is needed. The sample sheet file is required to input basic information on samples to be analyzed.

Open the .txt file called **template** located at the path **Pipelines /SampleSheet/template.txt.** This file could be opened in "Read-only" mode, thus to modify the content save it as the test name into the folder **Pipelines/SampleSheet/Test.**

Fill the .txt file with sample information, as follow: Run Folder name, sample name_ sample sheet Illumina position (e.g. *221015_M1234_96328521   Hostcell_S1*).

Space is not allowed into the sample sheet file. Use the tab command as separator between Run Folder name and Sample name. Each sample is named with the same name present in the Illumina Sample Sheet. When the sample sheet is filled, save and close.

DrAll 1.0 version

The bioinformatic pipeline *DrAll 1.0 version* has been designed to identify genus and species of recombinant Yeast cell banks, applied in the biotech pharma production.

The raw data produced by MiSeq sequencer are analyzed with the *DrAll* pipeline, which compares sample genomic sequences with an internal genomic database. The database contains Yeast validated genomic sequences downloaded from NCBI web site[20].

The host cell line, indicated by the sponsor site, is the reference standard corresponding to the first cell line used before the recombination step. For each study in which a new host cell line has to be tested, the introduction of the reference genome sequence into the database is necessary. If the reference sequence is not already present into this database, the standard sequence could be obtained during the qualification test or, if available, the standard sequence could be downloadable from the NCBI web site.

Sample genus and species determination is possible through the comparison of the sample genome sequence with the yeast genome sequences present into the database.

DrAll v1.0 performs an alignment of all reads produced by the MiSeq sequencer on the database's sequences, based on the level of similarity. At higher levels of deep coverage, each base is covered by a greater number of aligned sequences reads, so base calls can be made with a higher degree of confidence. The average deep coverage (Vertical coverage) for each sample sequence should be at least 130X in order to be considered relevant for the analysis.

At the end of the comparison, the pipeline generates a report containing the genus and species name and the homology percentage (Horizontal coverage or *Coverage* %) of one or more genome sequences present into the database.

Following are described the working instructions to use the bioinformatic pipeline *DrAll 1.0 version*.

The operator, to start the *DrAll 1.0 version* pipeline, have to double click on the executable file present at the path: **Pipelines/DrAll/file.exe** (Figure 9).

Figure 9 - DrAll pipeline executable icon

At the start, the tool requires to input the following information (Figure 10):

- **Sample_sheet**: text file previously created by the operator. The Sample sheet should contain samples names and position defined in the IEM sample sheet (see also Sample Sheet generation paragraph);
- **TITLE**: Test code of the analysis.

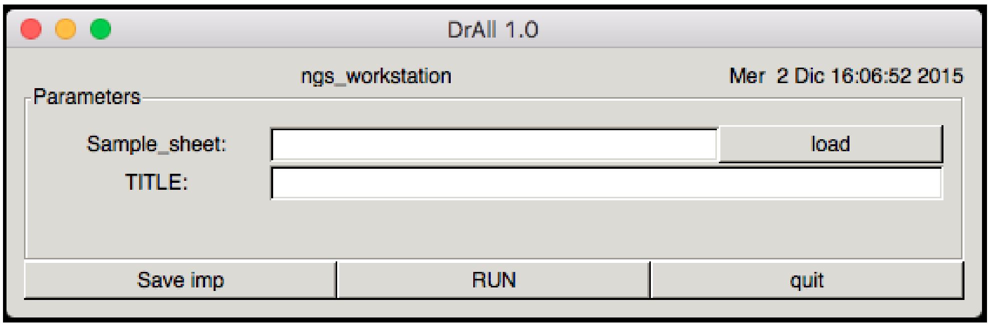
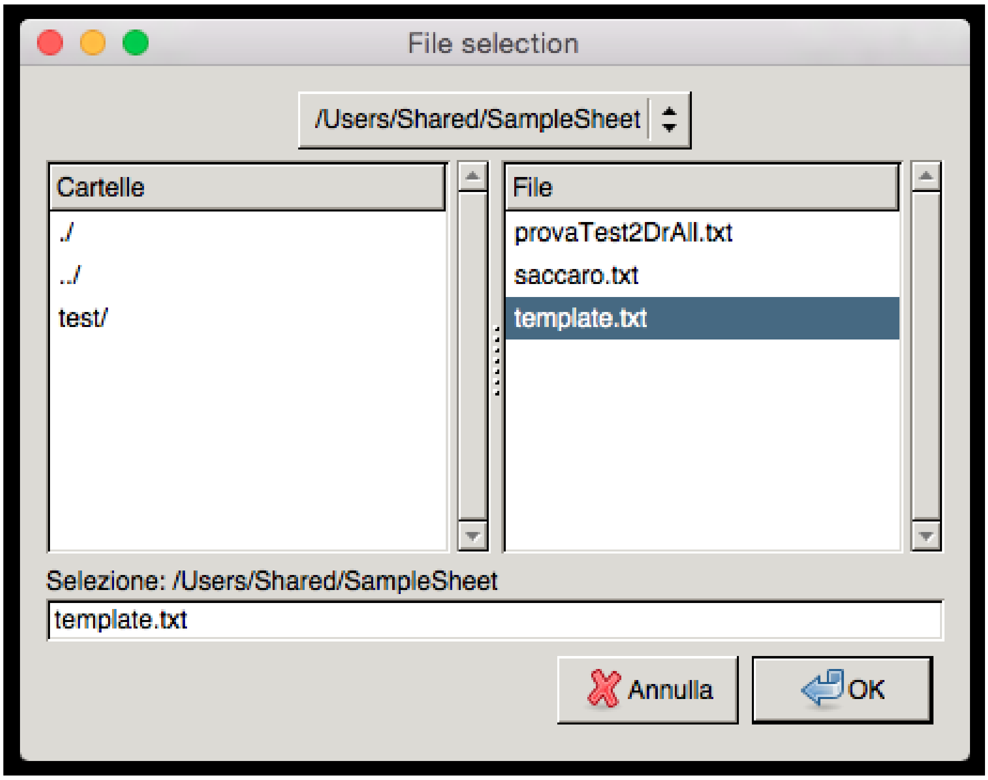


Figure 10 - DrAll Graphical User Interface

Use the "Load" button to open a secondary window and select the sample sheet path (Figure 5).

The reference genome database, called "***yeast_database.fa***", is automatically loaded by the pipeline. The *yeast_database.fa* is located in the server folder: **/home/biomol/database/…**



Selezione: /Users/Shared/SampleSheet

template.txt

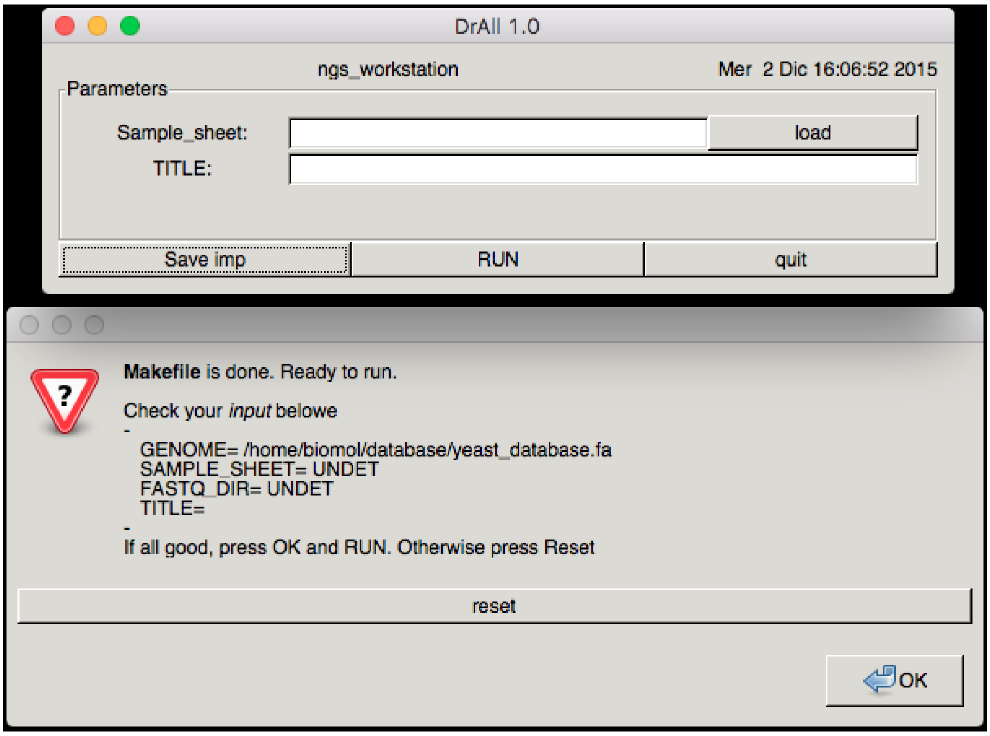
Figure 11 - Secondary window for the sample sheet selection

In case of a new cell line (host cell) has to be analyzed and the related reference *fasta* sequence is not available into the database, the System Admin has to create a *denovo* sequencing starting from the reads produced during the qualification study.

The *denovo* sequence has to be stored into the server folder: **/home/biomol/database**/, following the naming rules: "*Genus_species_strain.fa*" and updating the "*yeast_database.fa*" file content.

At the end of the input, save using the "Save imp" button (Figure 12). A wrap-up window will appear reporting all the information previously inserted.

If a correction is needed, use to "reset" button, and "OK" to come back at the first window, or confirm the information with "OK" button to proceed with the data analysis.



Makefile is done. Ready to run.

Check your *input* belowe

```
-
  GENOME= /home/biomol/database/yeast_database.fa
  SAMPLE_SHEET= UNDET
  FASTQ_DIR= UNDET
  TITLE=
-
If all good, press OK and RUN. Otherwise press Reset
```

Figure 12 - Imputed File confirmation window

At this stage, the *DrAll v1.0* pipeline is ready, and it is possible to start the run using the "RUN" button (Figure 10).

The pipeline, during the run, read the Run Folder Name and Samples Name information indicated in the sample sheet file and process the *fastq* raw data stored into the server. Raw data are automatically saved by the MiSeq sequencer at path: **/sequencer/MiSeq/RunFolderName/ Data/Intensities/BaseCalls**.

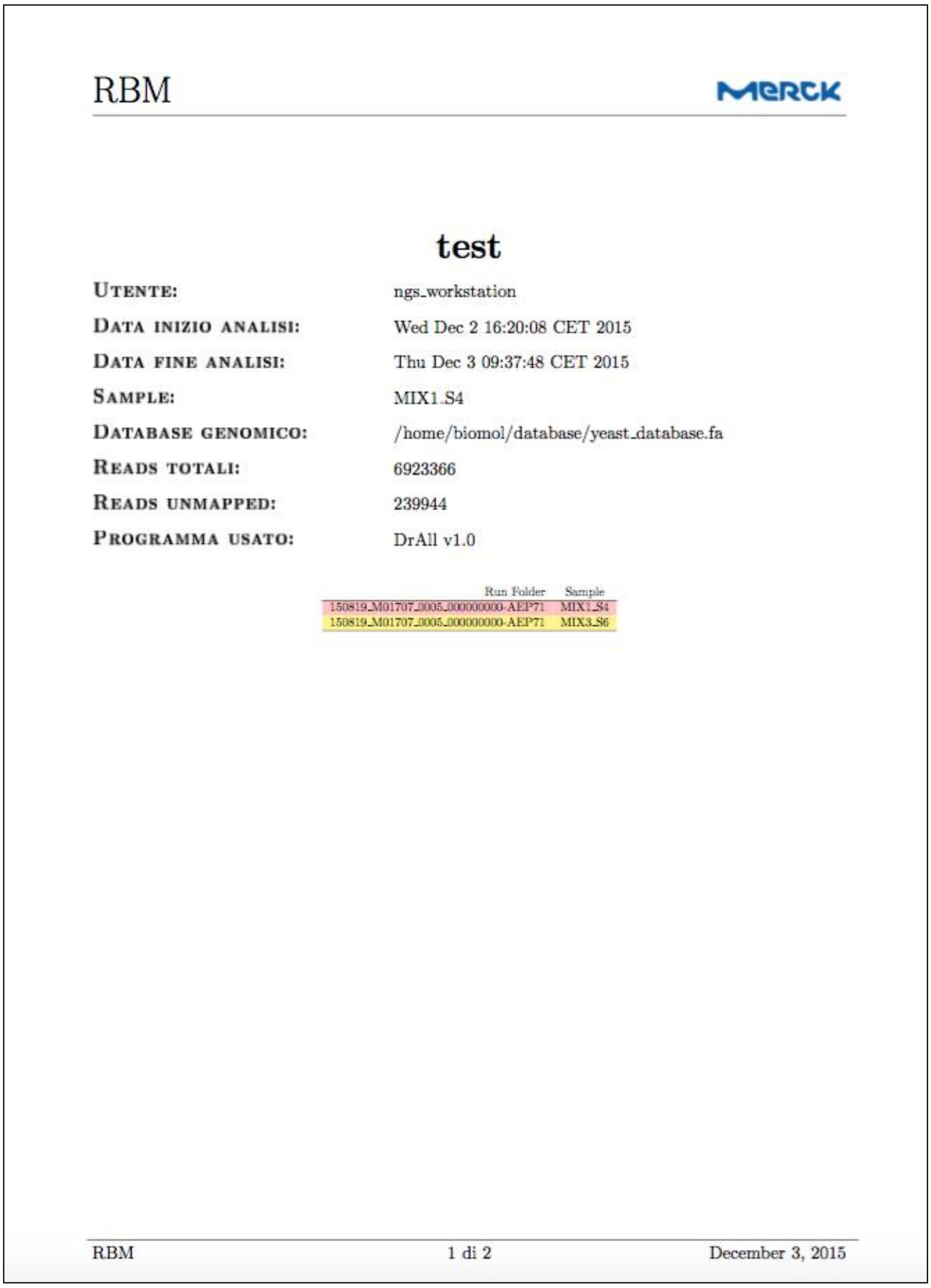
A bioinformatic function into the script allows all the sample reads to be aligned exclusively to only one reference genome present into the database.
For each run, *DrAll* pipeline generates a "makefile". The makefile is copied both in the server working folder and in the results folder on the user Desktop.

At the end of the data analysis, *DrAll v1.0* creates a report in .pdf and automatically save it on the server at path: **/home/biomol/analysis/results_monthdayear_hour** and on the workstation desktop, to visualize and print the report are available at the path: **/Users/(user_name)/Desktop/ results_ monthdayear_hour /*.pdf**

The report contains the following information (Figure 13):

- Project Code (reported by the operator during the initial analysis settings,);
- User Name: The name of the operator who performed the analysis;
- Date and time start analysis;
- Date and time end analysis;
- Sample: The name of the sample being analyzed;
- Genomic database: Database used for analysis;
- Total Reads: Total number of reads that passed the quality filter;
- Unmapped Reads: Number of reads that passed the quality filter but not aligned with the database;

- Program used: The name of the pipeline used for analysis;
- Summary table of information contained within the text file loaded by the operator ("Run Folder Name" – "Sample name_Sample position in the sample sheet";
- Table of results;
- The date the .pdf report was generated.



RBM MERCK

test

UTENTE:	ngs_workstation
DATA INIZIO ANALISI:	Wed Dec 2 16:20:08 CET 2015
DATA FINE ANALISI:	Thu Dec 3 09:37:48 CET 2015
SAMPLE:	MIX1_S4
DATABASE GENOMICO:	/home/biomol/database/yeast_database.fa
READS TOTALI:	6923366
READS UNMAPPED:	239944
PROGRAMMA USATO:	DrAll v1.0

Run Folder	Sample
150819_M01707_0005_000000000-AEP71	MIX1_S4
150819_M01707_0005_000000000-AEP71	MIX3_S6

RBM 1 di 2 December 3, 2015

Figure 13 - Fac-Simile of the result report generated by DrAll v1.0 pipeline

BeTween 2.0 version

The bioinformatic pipeline *BeTween 2.0 version* has been designed to identify mismatches between the sample sequence and the reference sequence into the database. By identifying mismatches it is possible to confirm the strain of belonging of engineered yeast cell banks, used for the production of biotech drugs.

After the genus and species determination, in the frame of the Strain identification, the genomic sequences obtained from the various production stages (MCB, WCB and PPCB/EoPCB) are compared with the related host cell, for the presence of mismatches, i.e. changes in gene material from one or more nucleotides.

BeTween 2.0 version aligns all the library fragments between the sample sequence and the reference host cell, based on the similarity.

Then compares each yeast cell bank production stage (MCB, WCB, and PPCB/EoPCB) with the host cell sequence and evaluates its mismatches. Each mismatch must have more than 20X of vertical coverage to be considered by the pipeline as a real mismatch.

In addition, after the first comparison, it evaluates the different production stages each other and produces a result consisting of the mismatch number that differentiates between MCB, WCB, and PPCB/EoPCB.

Following are described the working instructions to use the bioinformatic pipeline *BeTween 2.0 version*.

To start the *BeTween 2.0 version* pipeline, the operator has to double click the executable icon present at the path: **Pipelines/BeTween/file.exe** (Figure 14).

Figure 14 - BeTween pipeline executable icon

Starting the tool, it is required to input the following information (Figure 15):

- **Sample_sheet**: text file previously created by the operator. The Sample sheet should contain samples names and position defined in the IEM sample sheet (see also Sample Sheet generation paragraph);
- **Reference Genome**: Database containing the *fasta* file of the reference genome to be used for alignment;
- **Reference Align**: the *.sam/.bam* alignment file for the reference host cell genome to be used for comparison between strains;
- **Reference Transgene**: the *.fasta/.fa* file for the sequence of the transgene (plasmid - construct) used to engineer the Yeast cell bank;
- **TITLE**: Test code of the analysis;
- **Merge Report**: It is used when there are multiple samples to compare each other (e.g. MCB vs WCB vs EOPCB). When ticked, the pipeline performs the first comparison with the host cell and then compares the resulting reports;

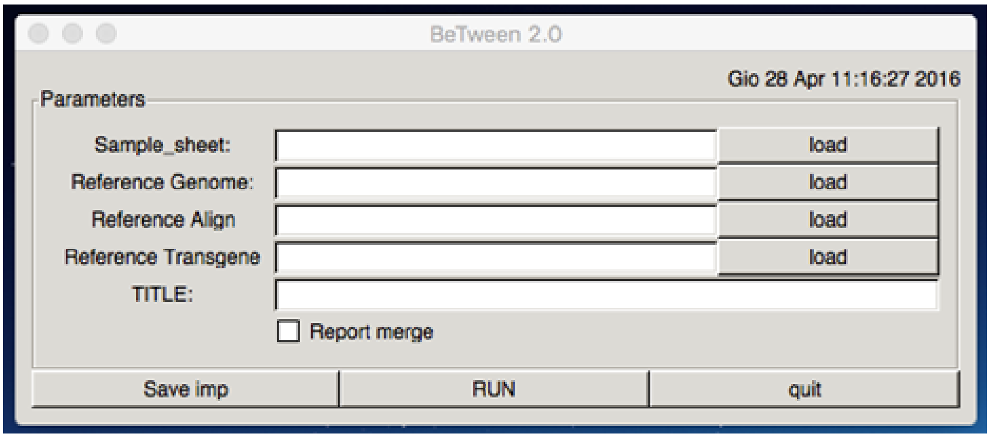
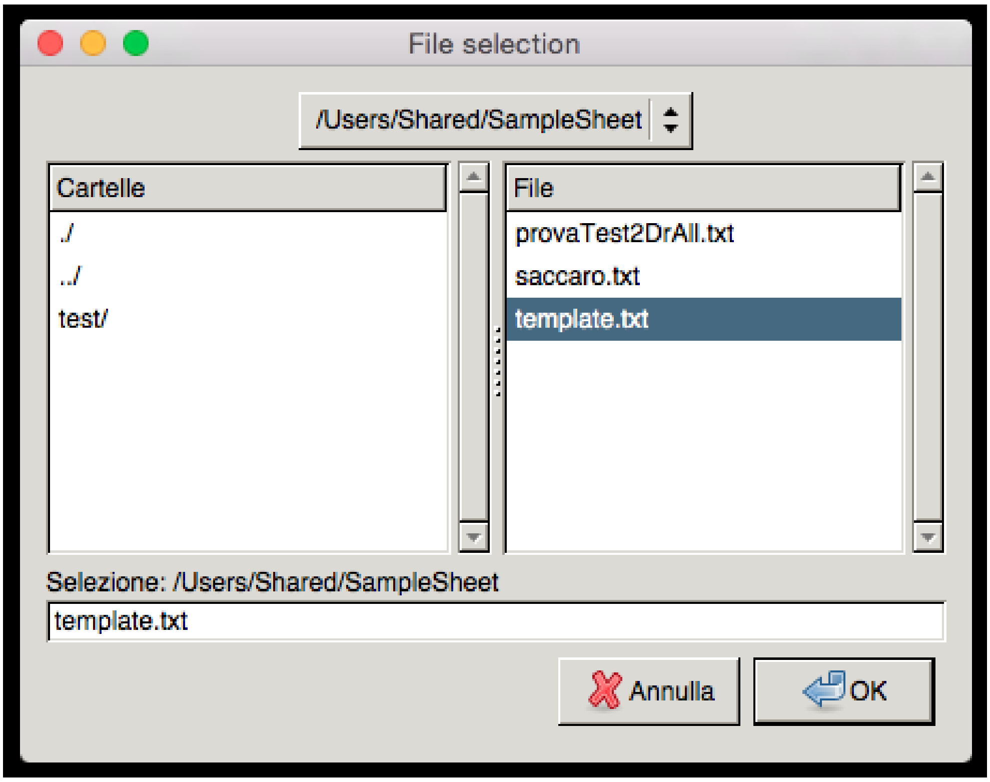


Figure 15 – BeTween Graphical User Interface

Use the "Load" button to open the secondary window and select the loading location of the Sample sheet (Figure 16).

The *fastq* raw data produced by the MiSeq sequencer are saved into the Illumina data storage platform.



Selezione: /Users/Shared/SampleSheet

template.txt

Figure 16 - Secondary window for the sample sheet selection

The operator has to select the reference genomes database to be used for analysis and is located in the server folder at the path: **/home/biomol/database/...**

The database (Reference Genome) contains only the mitochondrial (MT) and chromosomal (CHRS) *fasta* sequences belonging to the genus and species of the sample being analyzed.

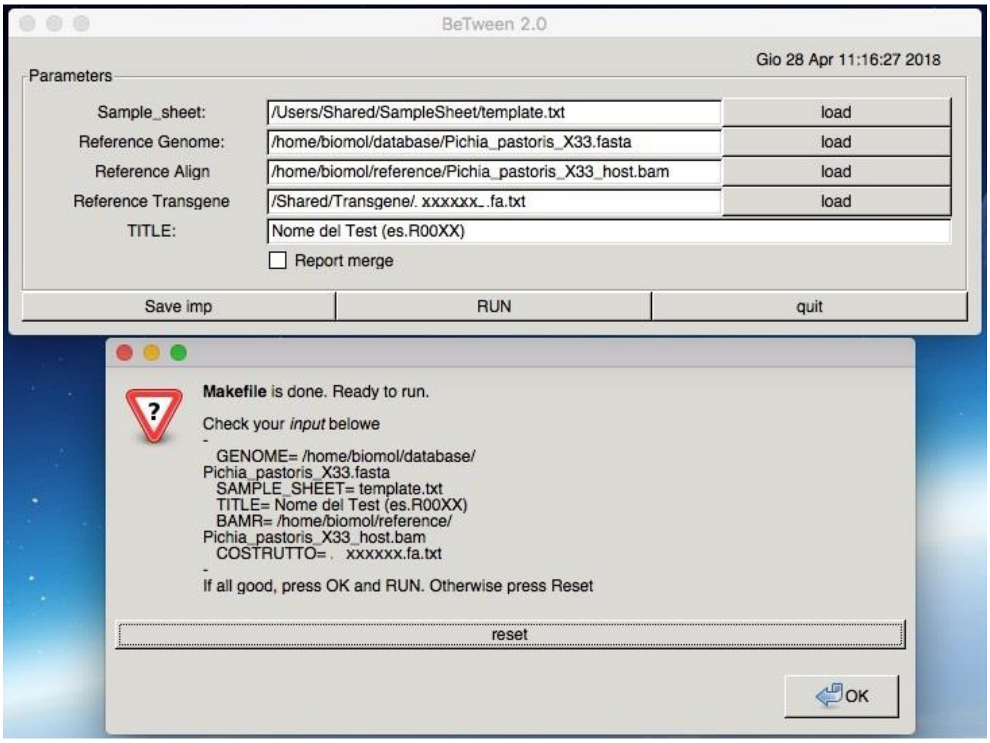
The pipeline allows to load, using the "load" button, the host cell database that is intended to use.

If a new cell line analysis is needed (e.g. host cell), but the related *fasta* sequence is not present in the database, the system administrator has to create the bam alignment file. The *.bam* file is created starting from the reads obtained during the qualification study. Subsequently, the system administrator has to save the *.bam* file at the folder path: /home/biomol/reference/, giving it the name: "genus_species_strain.bam".

The Reference Align is located in the folder on the server, at the path: **/home/biomol/reference/...** The Reference Transgene is loaded onto the workstation.

The constructs sequences are saved in *.fasta/.fa* format within the shared folder: **Pipelines/Transgene/.fa**. After the run is started, a copy of the uploaded file in the Reference Transgene field is automatically saved from the pipeline to a server folder, within the path: **/home/biomol/ROIseq/...**

When the requested information has been loaded, save using the "Save imp" button (Figure 17). A summary window of previously uploaded information appears. In case a correction is needed, use the "reset" button, and then "OK" to return to the data entry window or continue with the analysis confirming the information using only the "OK" button.



Gio 28 Apr 11:16:27 2018

Sample_sheet: /Users/Shared/SampleSheet/template.txt
Reference Genome: /home/biomol/database/Pichia_pastoris_X33.fasta
Reference Align /home/biomol/reference/Pichia_pastoris_X33_host.bam
Reference Transgene /Shared/Transgene/. xxxxxx_.fa.txt
TITLE: Nome del Test (es.R00XX)
□ Report merge

```
Makefile is done. Ready to run.

Check your input belowe
-
  GENOME= /home/biomol/database/
Pichia_pastoris_X33.fasta
  SAMPLE_SHEET= template.txt
  TITLE= Nome del Test (es.R00XX)
  BAMR= /home/biomol/reference/
Pichia_pastoris_X33_host.bam
  COSTRUTTO= .  xxxxxx.fa.txt
-
If all good, press OK and RUN. Otherwise press Reset
```

Figure 17 - Imputed File confirmation window

At this point, the system is ready and it is possible to start the analysis by pressing the "RUN" button (Figure 15).

When analyzing bioinformatics data, the pipeline takes the stroke folder name information on the sequencer and the name of each sample, contained in the Sample sheet, and searches the data, in *fastq* format, in the respective folders. The raw data of the races are contained at the path: **/sequencer/MiSeq/run folder/Data/Intensities/BaseCalls**.

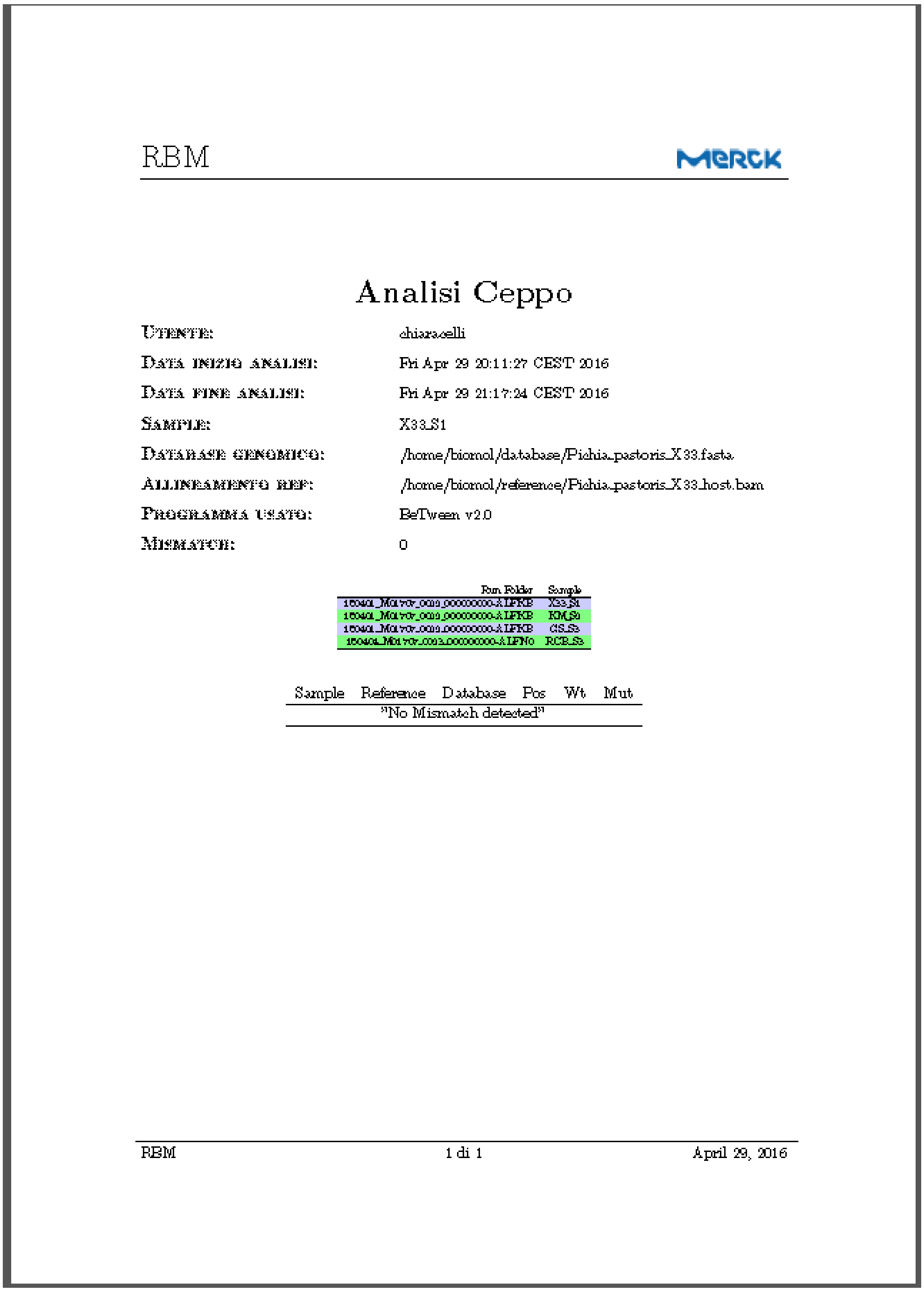
For each analysis, initiated using the RUN button, the pipeline uses the information loaded by the user to create a file called "makefile". The makefile is copied both within the workbook on the server and in the results folder located on the desktop.

The Reference Transgene field may not be filled in in case you want to analyze a standard strain, and therefore without transgene (e.g. host cell).

In case you start a trace by loading the Reference Transgene file, the first operation that the pipeline performs is to align all read of the sample on the transgene sequence. Whenever reads that align correctly on the transgene are automatically discarded. Unmatched reads that belong to the genome of the cell line investigated are maintained and then aligned and compared with Reference Genome and Reference Align.

After data processing is complete, *BeTween v2.0* generates a .pdf report that is automatically saved both on the server within the path: **/home/biomol/analysis/results_between _yearmonthday_hour** and on the workstation, for viewing/ print reports within the path: **/Users/(user_name)/Desktop/results_BETWEEN_ yearmonthday_time/.pdf**

The generated report contains the following information (Figure 18):



RBM

Analisi Ceppo

Utente:	chiaracelli
Data inizio analisi:	Fri Apr 29 20:11:27 CEST 2016
Data fine analisi:	Fri Apr 29 21:17:24 CEST 2016
Sample:	X33_S1
Database genomico:	/home/biomol/database/Pichia_pastoris_X33.fasta
Allineamento ref:	/home/biomol/reference/Pichia_pastoris_X33_host.bam
Programma usato:	BeTween v2.0
Mismatch:	0

Run Folder	Sample
160401_M01707_0019_000000000-ALFKB	X33_S1
160401_M01707_0019_000000000-ALFKB	KM_S2
160401_M01707_0019_000000000-ALFKB	CS_S3
160404_M01707_0093_000000000-ALFN0	RCB_S3

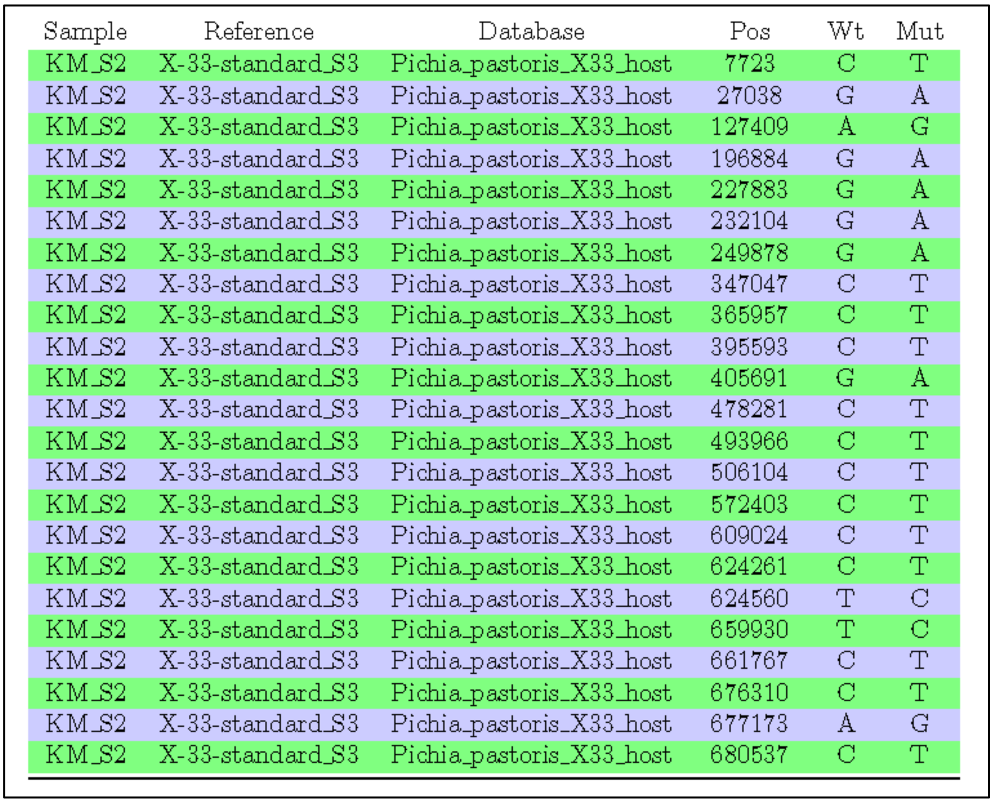
Sample	Reference	Database	Pos	Wt	Mut
"No Mismatch detected"					

RBM 1 di 1 April 29, 2016

Figure 18 – Fac-Simile of the result report generated by BeTween v2.0 pipeline

- Project Code (reported by the operator during the initial analysis settings,);
- User Name: The name of the operator who performed the analysis;
- Date and time start analysis;
- Date and time end analysis;
- Sample: The name of the sample being analyzed;
- Genomic database: Database used for analysis;
- Ref alignment: reference genome .bam file (host cell genome);
- Transgene: *.fasta/.fa* file of the reference transgene (plasmid sequence / construct);
- Program used: The name of the pipeline used for analysis;
- Mismatch: Number of Mismatches detected during sample (Sample) and host cell (Reference) comparison;
- Summary table of information contained within the text file loaded by the operator ("Run Folder Name" – "Sample name_Sample position in the sample sheet";
- Table of results;
- The date the .pdf report was generated.

The table containing the results of the pipeline consists of the following 6 columns (Figure 19):



Sample	Reference	Database	Pos	Wt	Mut
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	7723	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	27038	G	A
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	127409	A	G
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	196884	G	A
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	227883	G	A
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	232104	G	A
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	249878	G	A
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	347047	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	365957	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	395593	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	405691	G	A
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	478281	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	493966	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	506104	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	572403	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	609024	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	624261	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	624560	T	C
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	659930	T	C
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	661767	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	676310	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	677173	A	G
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	680537	C	T

Figure 19 - Example of a results table generated in the .pdf report

- Sample: The name of the sample being analyzed, entered by the user within the sample sheet;
- Reference: Name of the host cell genome used for comparison with The Sample;
- Database: Database used for alignment of Sample and Reference genomes;
- Pos: Position within the genome where the pipeline detects a Sample Mismatch relative to the Reference;
- Wt: Type of nucleotide read at position Pos on the Reference genome;
- Mut: Type of nucleotide read at position Pos on the Sample genome;

Report comparison (Report merge)

If the operator has checked the Report Merge box (Figure 20), the pipeline automatically performs a comparison of the results.

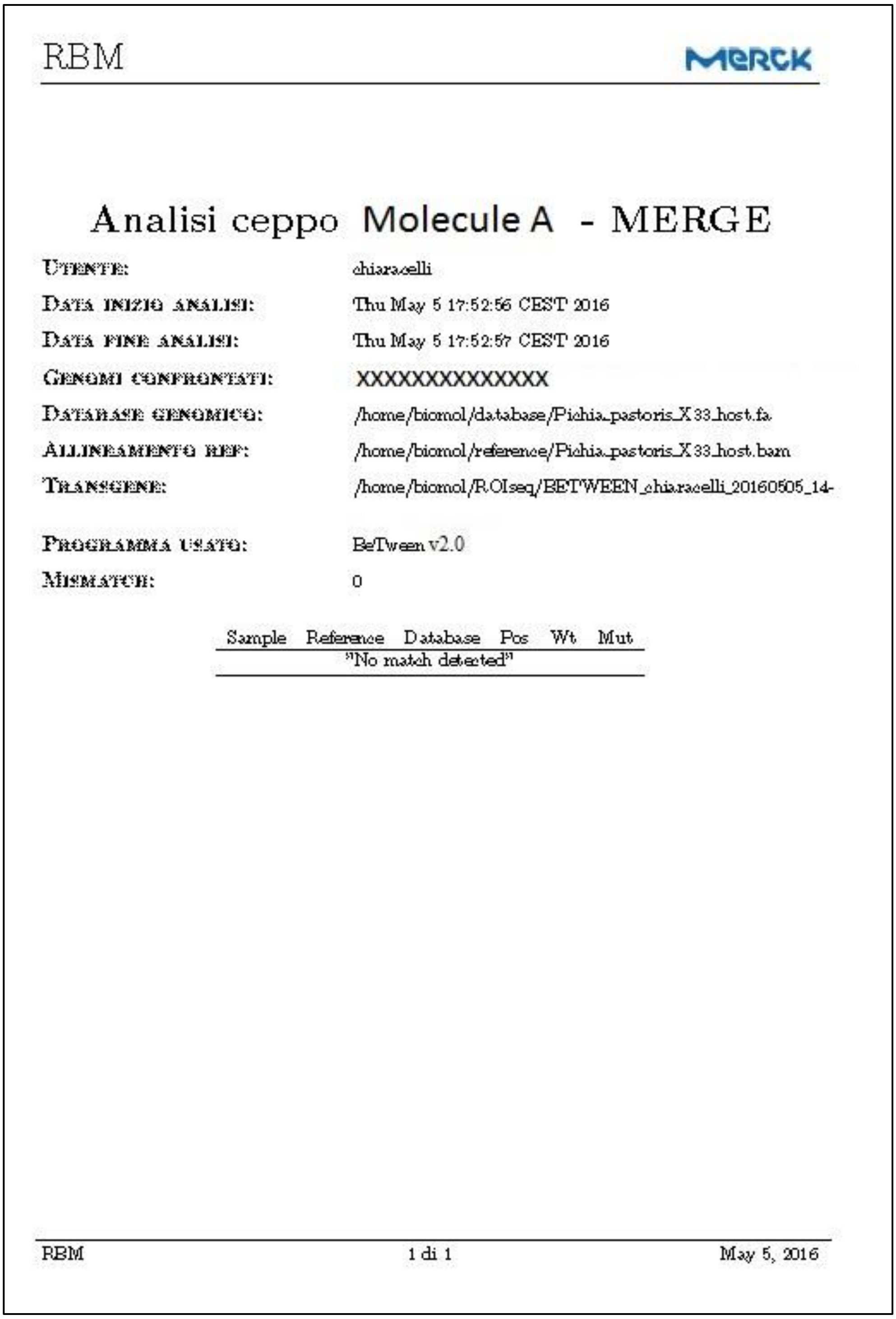
During the second analysis, all detected mismatches of each sample with respect to the cell host are compared to each other (e.g. MCB vs WCB vs EoPCB/PPCB).

The bioinformatics function in the pipeline works in such a way that every same mismatch in all three production stages is automatically deleted from the results list. This feature has been inserted for the purpose of highlighting only mismatches that differentiate between the different samples analyzed.

At the end of the data processing, a second report is generated, in .pdf format, showing only the mismatches detected between different analyzed samples (e.g. WCB, WCB, and PPCB/EoPCB).

Within the report that is generated at the end of the second analysis, the entry "Sample" does not appear, but the entry "Compared genomes". This field contains a list of all the sample names compared.

The table containing the results of the pipeline consists of the following 6 columns (Figure 14):



RBM MERCK

Analisi ceppo Molecule A - MERGE

UTENTE:	chiaracelli
DATA INIZIO ANALISI:	Thu May 5 17:52:56 CEST 2016
DATA FINE ANALISI:	Thu May 5 17:52:57 CEST 2016
GENOMI CONFRONTATI:	XXXXXXXXXXXXXX
DATABASE GENOMICO:	/home/biomol/database/Pichia_pastoris_X33_host.fa
ALLINEAMENTO REF:	/home/biomol/reference/Pichia_pastoris_X33_host.bam
TRANSGENE:	/home/biomol/ROIseq/BETWEEN_chiaracelli_20160505_14-
PROGRAMMA USATO:	BeTween v2.0
MISMATCH:	0

Sample	Reference	Database	Pos	Wt	Mut
"No match detected"					

RBM 1 di 1 May 5, 2016

Figure 20 – Fac-Simile of the report merge generated by BeTween v2.0 pipeline

- Sample: Name of the samples being compared (more than one name may appear in this column when the same mutation is present in more than one sample);
- Reference: Name of the host cell genome used for comparison with the Sample;
- Database: Database used for alignment of Sample and Reference genomes;
- Pos: Position within the genome where the pipeline detects a Sample Mismatch relative to the Reference;
- Wt: Type of nucleotide read at position Pos on the Reference genome;
- Mut: Type of nucleotide read at position Pos on the Sample genome;

Reference sequences creation using Bioinformatic tools

During Next Generation Sequencing (NGS) tests, the Illumina platforms available in the Molecular Biology group produce electronic raw data that can be processed using bioinformatics pipelines. For bioinformatics processing one or more reference sequences have to be created to make comparisons and analyze.

All data processing that takes place directly in the Illumina Data Analysis and Storage Platform can only be performed by users with an Administrator profile.

Administrator can:

- Manage, create and delete users;
- Manage user privileges;
- Create/run pipelines;
- Install and Update software;
- Access, create and delete data folders and files on the server;

The following paragraphs will describe how to create the following types of file:

- Fasta
- Consensus
- De Novo

Creating a *.fasta* file

fasta files are text files containing sequences of interest. Various types of sequences may be contained (e.g. amino acids, genomic RNA/DNA).

A *.fasta* file should have the following features:

>Title
ATGCTGATGTTCCAT (sequence)

The “>” denotes the beginning of a sequence. Several sequences may be found in a *.fasta* file.

It is possible to start from a sequence in .doc/.pdf/.excel/.txt format to create a *.fasta* file to be used in specific bioinformatics pipelines, for example, for creating Construct sequences.

To generate a .fasta file:

Open a new text file using WordPad/Notepad/Notepad++, so that it does not contain formatting. Start the first line with the symbol “>” and write the information desired concerning the sequence (do not leave spaces between the symbol “>” and the first word used). Starting on the second line Copy and Paste the sequence of interest from the original text (e.g. *.doc, .pdf, .xlsx, .txt*). Save the file specifying the format with the extension *.fa* or *.fasta*.

Creating a *Consensus* reference sequence

If a reference sequence is to be created starting from sample reads, and there is already a base sequence to use for alignments, a consensus sequence can be created.

First of all, the sample reads are selected according to their quality characteristics.

Then, the consensus sequence is created by aligning the reads having good quality with respect to the above parameters, for which no base reference sequence is available, on a sequence genetically closest to the sequence to be created.

The outcome of this alignment is a new sequence that integrates the data from the original sequence with the differences made by the new reads.

Basic rules to create a Consensus sequence

The basic rules to follow to prepare a consensus sequence are:

- Run the Trimmomatic tool using the command "trimmomatic_bit" version pair ends (PE). According to the settings of the Illumina platforms, Trimmomatic is set using the "phred33" quality matrix.
- The reads are selected according to the following quality parameters:
  - The first and the last three bases of each read are trimmed if they are below a certain quality, using the parameters "TRAILING=3" and "LEADING=3". This is done since these regions may have low quality due to technical reasons related to reading and not connected with the sequence.
  - The central part of each read is read using the command: "SLIDINGWINDOW=10:25" in a 10 base reading window whose mean quality must not fall below Q25,
  - Subsequently, all the reads are assessed for the presence of sequences attributable to the adaptors. The adaptor sequences are discarded since they may be included in sequencing if the size of libraries is small. The adaptors depend on the type of sequencing done and their sequences can be downloaded directly from the Illumina web site[26].
  - After making all the selections and trims, reads under a certain length are completely discarded. The minimum length required is 65% of the original length of the reads.
- When processing is completed, Trimmomatic produces new copies of *.fastq* files that will replace the original .fastq files.
- The files produced by Trimmomatic are then aligned on the reference sequence most similar to the target using the BowTie2 tool.
- BowTie2 is used keeping the default settings. The pairs of reads are processed in Unpaired mode to produce a .sam alignment file.

BowTie2 default settings for DrAll pipeline are:

- For the reference index → `bowtie2-build -f genome.fa genome`
- For the samples alignment → `bowtie2 -x genome --rg-id sample --rg "SM:sample" -U R1.fastq,R2.fastq > Sample_align.sam`

BowTie2 default settings for BeTween pipeline are:

-
  ```
  bowtie2 -x costrutto --rg-id sample --rg "SM:sample" -1 R1.fastq -2 R2.fastq | samtools view -bhS - | samtools sort - > Sample_costrutto_align.bam
  ```

-
  ```
  bamToFastq -i Sample_costrutto_align.bam -fq Sample_R1_unalign.fq -fq2 Sample_R2_unalign.fq
  ```

-
  ```
  bowtie2 -x genome --rg-id sample --rg "SM:sample" -U Sample_R1_unalign.fq,Sample_R2_unalign.fq > Sample_align.sam
  ```

- The .sam file and the reference are loaded on Geneious software, where its function called "Generate consensus sequence" builds a new sequence based on the initial reference but changed in those portions in which the sample reads vary. The minimum concordant cover required for the sequence of a base to be changed is 2.

- The Consensus, file created is a *.fasta* file that is then renamed and transferred to the appropriate folder on the server, where it can be used as a reference in the pipelines for subsequent bioinformatics analyses.

Creating a *De Novo* reference sequence

If a reference sequence needs to be created starting from sample reads, but there is no base sequence on which performing an alignment, a De Novo sequence can be created.

First of all, the sample reads are selected according to their quality characteristics.
Then, the De Novo sequence is created by concatenating the reads using the Bruijn Graph algorithm based on similar portions within the reads. This algorithm allows a longer concatenated sequence to be created.

The outcome of creation of a De Novo is a series of sequences: Contig (set of reads without interruptions), Scaffold (set of Contigs with interruptions) and Reads it has not been able to insert into any Contig. Reads, Contig and Scaffold are *.fasta* format files.

Basic rules to create a De Novo sequence

To prepare a De Novo sequence use Trimmomatic to select the reads according to their quality and the presence of adaptors (a *.fasta* file containing the sequences of the adaptors can be downloaded from the Illumina web site[26]). Use the most suitable trimming tool depending on the expected size of the genome to be assembled. The most common tools available online and used to create De Novo sequences are:

- ABySS[28]
- SOAPdenovo[29]
- Velvet[30]
- ALLPATHS-LG[31]
- Celera Assembler[32]
- Trinity[33]

A De Novo is a *.fasta* file consisting of several sequences that can be renamed and transferred into certain pathways on the Server, for use as reference files when analyzing using bioinformatics pipelines.

Method Validation

The validation of an analytical method is intended to demonstrate, through the execution of specific tests and in a suitably documented way, that this method always leads to the expected result, in accordance with the quality requirements of GMPs and is suitable for its intended purpose.[7]

As part of the validation process for each analytical method, specific parameters have been considered, related to the characteristics of the method itself and representative in order to demonstrate its efficiency.

Analytical methods can be methods described within the Pharmacopoeia[34], methods that have been transferred to the GHO Ivrea or methods that have been directly developed and set up within the GHO Ivrea.

The strategy used to validate analytical methods depends on the category of membership of each method with regard to the parameters being verified and also changes based on the fact that the method is Pharmacopoeia, transferred or developed internally to the BQC.

For this reason, the test activities, regardless of the type of method to be validated and its origin, may not meet all the above requirements and may not "cover" all the parameters listed in the following paragraphs. In this case, within the specific protocols, it is necessary to detail this need and describe and motivate the chosen test strategy.

Classification of Analytical Methods

Analytical methods can be grouped into the four most common types of analytical procedures:

- Identification test;
- Quantitative assay for impurities' content;
- Limit test for the control of impurities;
- Quantitative tests of the active principle in samples of drug substance or drug product or other selected component(s) in the drug product, and "bioassay"

Identification test

Identification tests allow you to identify an analytic within a sample.

Identification is usually achieved by comparing an element feature to be identified with that of a reference standard. Identification tests can be based on the use of physical-chemical or immunological methods, or molecular biology and cell biology techniques.

Quantitative assays for impurities' content

Quantitative assays for impurities allow to detect the presence of a certain "impurities" within a sample and quantify it. The term "impurities" refers to any foreign contaminant element present within a sample.

Quantitative assays for impurities can be based on the use of physical-chemical or immunological methods, or on the use of molecular biology and cell biology techniques.

Limit test for the control of impurities

Impurities limit tests allow for the detection of a contaminant foreign element within a sample when the contaminant is present in an amount greater than a certain specific threshold for each method. This is variable based on the characteristics of the analytical method used to locate it.

Limit tests for impurities can be based on the use of physical-chemical or immunological methods, or on the use of molecular biology and cell biology techniques.

Quantitative assays and Bioassay

Quantitative assays and bioassay allow you to quantify a given analyte within a sample. For example, they are used for measuring and quantifying active ingredients and/or specific components within drugs.

These tests may be based on physical-chemical or immunological methods, or on the use of molecular biology and cell biology techniques.

Bioassay or biological assay use in vivo and/or in vitro techniques using animal models or cellular systems and allow to measure the biological activity (power) of a given active ingredient within a sample.

Object parameters

The parameters that must be checked as part of the validation of an analytical method depend on the category of the method itself, so it must be properly classified before you define the tests to run for the verification of its efficiency.

In order to be able to define how a method is validated, it is always very important to refer to the guidelines in the ICH and FDA guidelines, and you can also apply a risk assessment process, according to the guidelines of the ICH Q9 guideline[35].

The parameters to be verified during the validation are:

- Specificity;
- Linearity;
- Application interval;
- Accuracy;
- Detection limit;
- Quantify limit;
- Robustness.

The procedure for evaluating the parameters listed above depend on the characteristics of the test in question and the objective for which it is used and should be described in detail within the individual validation protocols.

It is pointed out that the tests used at Merck Ivrea are applied for the analysis of biological samples. For this reason, given the inherent complexity of this type of analysis, it is not always possible to perform test validation considering all required parameters (e.g. due to a limited amount of available sample).

Any changes to the validation strategy with respect to the parameters in question are detailed and justified within the validation protocols.

Specificity

The term specificity refers to the ability of the test to discriminate between the analytic subject of the test and all other possible elements present within the sample.

This parameter must be evaluated for all test categories (Table 1).

In the case of identification tests, it is necessary, if applicable, to demonstrate their ability to discriminate between similar elements that could be present in this sample.

In some cases, it may not be possible to demonstrate the absolute specificity for an analyte: in such cases, it is possible to compensate through the combined use of one or more specific tests, in order to obtain the level of discrimination necessary.

However, this compensation must always be well supported and accompanied by an adequate assessment of the risks associated with the use of a non-specific method.

Linearity

The concept of linearity refers to the linear "dose-response" correlation between the analytic being analyzed and the measurement obtained as part of the test and is expressed by an appropriate mathematical function.

A minimum of 5 concentrations of test item is recommended to investigate linearity. Alternative approaches are accepted if justified. In particular, the number of concentrations may be different for the assays described in pharmacopeia.

Linearity should be assessed in the context of quantitative assays for impurities and quantitative assays and "bioassay" (Table 1).

Application range

The application range is usually derived from the results of the linearity analysis carried out during validation and is correlated with its scope. This parameter defines the range by which the results of the measurements taken are characterized by an appropriate level of linearity, accuracy, and precision.

The tests for which the application interval evaluation is required are quantitative assays for impurities and quantitative and bioassay assays (Table 1).

Accuracy

The accuracy indicates how far the measured result deviates from its corresponding reference value. The parameter must be validated in the case of quantitative assays for impurities and quantitative assays and "bioassay" (Table 1).

In the case of quantitative assays and "bioassay", the accuracy verification can be made by calculating the percentage difference between the result obtained and the corresponding value of the reference standard or, if a reference standard is not percentage recovery of a known quantity of analytic in the sample.

In the case of quantitative tests for impurities, it is possible to verify the accuracy of the test using contaminated samples with known quantities of the analyte that must be identified and quantified.

It is recommended to test accuracy using at least 9 determinations for a minimum of 3 concentrations (e.g. 3 replicated for each of the 3 concentrations) within the range of application of the method.

If this is not possible, the verification can be done by comparing the results obtained by the test to be validated with those obtained using another validated procedure.

Precision

This parameter should be tested when validating quantitative assays for impurities and quantitative assays/"bioassay" (Table 1).

The following parameters are included within the concept of precision:

- Repeatability;
- Intermediate precision;
- Reproducibility.

Repeatability

Repeatability is the level of variability of the results calculated multiple times within the same analytical session. It must be verified using an appropriate number of determinations that cover the range of application of the method or an appropriate number of 100% determinations of the concentration to be tested. It is recommended to test at least 9 determinations using a minimum of 3 concentrations (e.g. 3 replicated for each of the 3 concentrations) within the range of application of the method, or 6 determinations 100% of the concentration to be tested.

Intermediate precision

The evaluation of the intermediate accuracy of a test allows us to determine the variability of the results as a result of changes such as, for example, on the day of execution, the operator performing the test, the instrumentation used.

How you test this parameter depends on the circumstances and how you use this test. It is not necessary to study individual variations individually and it is recommended to use an experimental matrix design.

Reproducibility

Reproducibility assesses the level of variability of the results of a method as a result of the need to perform such a method at another laboratory or with a different analytical approach, such as applying robotization to a method performed manually.

Limit of Detection (LOD)

This parameter expresses the slightest concentration at which a qualitative and/or quantitative analysis can be conducted to determine whether an analyte is present.

The detection limit can be assessed according to different approaches, for example, based on visual assessments, the ratio of signal to background noise, the standard deviation of a response or gradient or negative control, or on a calibration curve.

In general, other types of approaches can also be used, depending on the characteristics of the test to be validated.

Quantifying limit

The quantifiable limit applies in the case of quantitative tests for impurities and is defined as the minimum amount of analyte present in a sample and that can be quantified by testing, that is, the limit of concentration up to which it is possible to obtain a quantitative measure with relative confidence.

The approaches to its determination can be multiple, depending on the characteristics of the test. You can use the following methods for assessing the detection limit.

In some cases, you can define the quantification limit as the lowest value in the application range.

Robustness

Robustness indicates the ability of an analytical method not to be affected by deliberately introduced operational changes.

Thus, it can be considered as the ability of the method not to be significantly affected by changes in analytical conditions (e.g. in the composition and suppliers of the reagents used, in the volumes of the reagents, in the temperature and in the times of incubation, in the analytical instruments used, etc.) It can be evaluated and documented both during the development of the method and during its validation.

Based on the FDA's Guidance for Industry "Analytical Procedures and Methods Validation for Drugs and Biologics"[36], it is recommended that you evaluate the robustness of a method during its development in order to evaluate, in advance to the other parameters to be tested during validation exercise, the performance of the method and reduce the risks of a failure of the validation itself.

It is not necessary to individually evaluate the changes when an experimental design based on a risk assessment approach is applied to the robustness test. The experimental design allows us to combine the variations by reducing the number of tests to be carried out allowing, at the same time to evaluate the interactions between them.

It is necessary to first explain the individual steps of the method (the steps that can have an impact on the results produced) and for those steps check the criticality considering the following parameters:

- Severity (the impact that a problem relating to the passage considered could have on the results produced);
- The possibility of detecting the possible problem of the tested step is detectable;
- Probability (the possibility and frequency with which the problem might occur).

These evaluations must be formalized using a matrix schema, and for each step, the risk priority index is calculated using the values in the matrix, which can be defined as low, medium, high, or very high. Steps that are assessed at medium or high risk and very high should be involved in assessing the robustness of the method. The risk assessment is compiled by the staff who develop/optimize the method.

Once the limits to be tested for each factor have been established (e.g. incubation time limit of 20 and 28 hours compared to a target incubation time set in 24 hours), the software defines the number, type and order of the tests to be performed to have an analysis statistically significant and, based on the results obtained in the tests, determines whether and factors (and/or their interactions) tested within their limits significantly influence the method.

If the method provides a uniquely qualitative result (e.g. positive or negative result), the software can only be used to construct the matrix of the tests to be carried out and the influence of the changes to the passage will be defined when a different result occurs in the given test than expected.

The approach just described for robustness assessment should be used for all developed methods from scratch. If this approach is not possible, it should be described as rational within the validation protocol.

If you need to perform robustness tests later than validating a method (e.g. introducing additional tools or critical reagents that are not expected in the validation you perform), those tests will be handled in an addendum to the described in the protocol and reports.

Table 1 - Validation object parameters for different test types

Method *Parameters*	*Identification test*	*Quantitative assay for impurities*	*Limit test for the control of impurities*	*Quantitative assays and Bioassay*
Specificity	+	+	+	+
Accuracy	-	+	-	+
Repeatability	-	+	-	+
Intermediate precision	-	+	-	+
Reproducibility	-	+/-	-	+/-
Limit of Detection	-	+	+	-
Limit of Quantification	-	+	-	-
Linearity	-	+	-	+
Range of Applicability	-	+	-	+
Robustness	+	+	+	+

Legend: + Evaluation object parameter
- Parameter not being evaluated
+/- Parameter whose evaluation is not mandatory

Method Validation strategy

The method, which allows the whole genome of a yeast cell bank to be tested against those of other yeasts in a database, is classified as an **Identification test**, but also has potential as a **Limit test for the control of impurities**.

The test is divided into 2 phases:

- The first, aimed at identifying the genus and species of the yeast cell bank being tested;
- The second, with the aim of confirming the strain of the yeast cell bank, achieved by comparing the sample with the host cell (cell line used for the producing biopharmaceuticals, before being recombined with the gene of interest). Subsequently, the results from the various manufacturing stages of the recombinant cell bank (MCB, WCB e PPCB/EoPCB) can be compared.

The parameters evaluated during the method validation were:

- Specificity
- Robustness

And in addition, the following parameter required for Limit Tests for Impurities was tested:

- Limit of Detection (LOD)

Specificity of the method is intended as the ability of the assay to distinguish the genome of yeast genera, species and strains of interest from those of other yeasts. To determine the genus and species, the specificity of the method will be checked by making a comparison between the genome of various samples and all the genomes in a database, while for identification of the strain the specificity of the method will be checked by comparing various strains against the genome of the reference cell line.

Limit of Detection (LOD) of the method, intended as the lowest concentration of contaminating microorganisms detectable in the sample of interest, will be checked by analyzing the genomic sequence obtained from mixes with varying percentages of 2 different genus and species or 2 different strains of yeast.

Robustness of the method, intended as the ability of the method to remain unchanged despite the introduction of small deliberate changes in the run parameters, will be checked by introducing some changes into critical steps of the method. For the method robustness check, some of the variations will be tested under this validation protocol, while other variations will be tested during preliminary setup activities.

Verification of the method’s Specificity

The objective of this test is to check the ability of the method to discriminate genomes belonging to different genus, species, and strains of yeast by comparing whole genomic sequences.

This test has two aims to:

- Check the ability of the method to discriminate between yeasts having different genus and species using the *DrAll* bioinformatics pipeline.
- Check the ability of the method to discriminate between two yeasts belonging to the same genus and species but different strains, using the *BeTween* bioinformatics pipeline.

Acceptance criteria

All the assays done as part of this test must be valid, as defined in *Method Workflow* paragraph. The expected results from this test are that:

1. the method and the *DrAll* program must be able to discriminate between microorganisms belonging to different genus and species and to identify microorganisms belonging to the same genus and species as being alike.

Table 2 – Summary of genus and species tested for the method’s specificity verification

Test samples	*Expected Result* *(Genus and Species)*
X-33, Pichia pastoris Yeast Strain (Invitrogen)	*Pichia Pastoris*
KM71H, Pichia pastoris Yeast Strain (Invitrogen)	*Pichia Pastoris*
GS115, Pichia pastoris Yeast Strain (Invitrogen)	*Pichia Pastoris*
Saccharomyces cerevisiae S288c (ATCC204508)	*Saccharomyces cerevisiae*
Cyberlindnera jadinii (ATCC18201)	*Cyberlindnera jadinii*
Candida glabrata (DBVPG3178)	*Candida glabrata*
Recombinant P. pastoris X-33 cell bank (RCB)	*Pichia Pastoris*
Saccharomyces cerevisiae CBS1171 (ATCC18824)	*Saccharomyces cerevisiae*

2. The method and the *BeTween* program must be able to discriminate between two or more microorganisms belonging to different strains and to identify microorganisms belonging to the same strain as being alike.

Table 3 - Summary of strains tested for the method’s specificity verification

Test samples	*Expected Result* *(Strain)*
GS115, *Pichia pastoris* Yeast Strain (Invitrogen) aligned on X-33 *Pichia pastoris* Host Cell	Different strains More than 2 mismatches
KM71H, *Pichia pastoris* Yeast Strain (Invitrogen) aligned on X-33 *Pichia pastoris* Host Cell	Different strains More than 2 mismatches
X-33, *Pichia pastoris* Yeast Strain (Invitrogen) aligned on X-33 *Pichia pastoris* Host Cell	Same strains Fewer than 2 mismatches
Recombinant *P. pastoris* X-33 cell bank (RCB) aligned on X-33 *Pichia pastoris* Host Cell	Same strains Fewer than 2 mismatches
Saccharomyces cerevisiae CBS1171 (ATCC18824) aligned on *Saccharomyces cerevisiae* S288c ATCC204508	Different strains More than 2 mismatches

The list of samples used to perform the method’s Specificity verification and the expected results are reported in Table 2 and Table 3.

Verification procedure

The samples used for the following test are:

- X-33, *Pichia pastoris* Yeast Strain (Invitrogen)
- *Saccharomyces cerevisiae* S288c (ATCC204508)
- KM71H, *Pichia pastoris* Yeast Strain (Invitrogen)
- GS115, *Pichia pastoris* Yeast Strain (Invitrogen)
- *Cyberlindnera jadinii* (ATCC18201)
- *Candida glabrata* (DBVPG3178)
- Recombinant *Pichia pastoris* X-33 cell bank (RCB)
- *Saccharomyces cerevisiae* CBS1171 (ATCC18824)

A single analysis has been performed on each sample as described below.

The test is done starting from a whole cryovial (2 mL of sample). Sample genomic DNA extraction is followed by preparation of the library and running on the MiSeq (Illumina) sequencer and finally by bioinformatics analysis.

Bioinformatics analysis has been carried on using both the *DrAll* and *BeTween* programs in order to analyze both differences in genus and species, and differences in strain.

Specifically, during identification of the strain with the *BeTween* v2.0 pipeline, the *Pichia pastoris X-33* cell bank sample (RCB) will be tested excluding the sequence of the Molecule A transgene (plasmid + construct), using the reference file MoleculeA.*fa*.

Once the libraries for each sample have been obtained, the sequences are saved on the server connected to the MiSeq sequencer. The bioinformatics data can be used for comparison and analysis both in this test and in subsequent tests, given that these sequences are independent of the method for preparing the library and of the sequencer run. Run data from different times can be compared and can be used for different types of analyses.

Verification of the method's Limit of Detection (LOD)

The objective of this test is to check the limit of detection (LOD), intended as the lowest concentration of genomic DNA of contaminating yeast (i.e. belonging to a different genus and species or to a different Strain from that of the sample of interest) detectable in the sample of interest.

Specifically, testing for the genus and species will evaluate the minimum concentration of contaminating genomic DNA able to generate a sequence that can be discriminated from the test sample. Strain testing will evaluate the minimum concentration of contaminating genomic DNA able to generate a sequence that can be discriminated from the test sample.

Acceptance criteria

All the results obtained in this test should be valid and met the defined acceptance criteria.

The limit of detection of the method for identifying the genus and species will be defined as the lowest percentage of genomic DNA belonging to the microorganism defined as "contaminating" (*S. cerevisiae S288c*), able to give rise to a recognizable, distinguishable sequence with respect to the test sample, in a total of three assays. The LOD result will be considered as the lowest percentage of contaminant found in all three repetitions of the test.

In the *DrAll* program report, show two results having a coverage% greater than or equal to 85%.

The limit of detection of the method for identifying the strain will be defined as the lowest percentage of genomic DNA belonging to the microorganism defined as "contaminating" (*P. pastoris KM71H*), able to give rise to a recognizable, distinguishable sequence with respect to the test sample, in a total of three assays. The LOD result will be considered as the lowest percentage of contaminant found in all three repetitions of the test.

The *BeTween* report shows all results having more than 2 mismatches.

Verification procedure

To verify the method limit of detection, two yeasts belonging to different genus and species and two yeasts belonging to the same genus and species but to different strains will be used.

Prepare mixes of Genomic DNA as follow:

- MIX1= Mix of 80% DNA extracted from X-33, *Pichia pastoris* Yeast Strain (Invitrogen) and 20% DNA extracted from *Saccharomyces cerevisiae* S288c;
- MIX2= Mix of 90% DNA extracted from X-33, *Pichia pastoris* Yeast Strain (Invitrogen) and 10% DNA extracted from *Saccharomyces cerevisiae* S288c;
- MIX3=Mix of 99% DNA extracted from X-33, *Pichia pastoris* Yeast Strain (Invitrogen) and 1% DNA extracted from *Saccharomyces cerevisiae* S288c;

- MIX4= Mix of 80% DNA extracted from X-33, *Pichia pastoris* Yeast Strain (Invitrogen) and 20% DNA extracted from KM71H, *Pichia pastoris* Yeast Strain (Invitrogen);
- MIX5= Mix of 90% DNA extracted from X-33, *Pichia pastoris* Yeast Strain (Invitrogen) and 10% DNA extracted from KM71H, *Pichia pastoris* Yeast Strain (Invitrogen);
- MIX6= Mix of 99% DNA extracted from X-33, *Pichia pastoris* Yeast Strain (Invitrogen) and 1% DNA extracted from KM71H, *Pichia pastoris* Yeast Strain (Invitrogen);

Mix the DNA extracts as described above, prepare a new library for each mix and run on the sequencer. Use the *DrAll* pipeline for genus and species testing while, use the *BeTween* pipeline for strain testing.

This test allows the method LOD identified during setup testing, corresponding to 10% of contaminants both as regards genus/species and strain to be checked and confirmed.

The LOD is defined as the lowest percentage of contaminant that can be distinguished from the sequence of samples named main. To be acceptable, genus and species testing should identify at least two results with a coverage% greater than or equal to 85%, corresponding to both the sequences of the main and contaminating samples. To be acceptable, strain testing should identify a result having a number of mismatches greater than or equal to 2, corresponding to the mix from the two main and contaminating samples.

Conclusively to confirm the LOD, the test described above will be completely repeated a further two times, using 3 mixes of microorganisms: one containing the percentage of contaminant identified as “at LOD”, one containing the percentage immediately above the former “above LOD” and one containing the percentage immediately lower “below LOD”.

Verification of the method’s Robustness

The objective of this test is to check the ability of the test to remain unchanged following the introduction of deliberate change to some critical steps.

The parameters to be tested as part of the robustness check were defined using a Risk Assessment approach during development activities.

The phases identified as being critical for this method were:

1. Quantitation of extracted DNA. Its high criticality is because variations in the range of concentration of genomic DNA could have an impact on the yield of library preparation, with the possibility of producing a library that does not meet the quality acceptance criteria for a “good” Library, which is therefore not suitable for loading on the sequencer. In fact, if the quantity of libraries on the flow cell is overestimated, the number of clusters will be low, and few sequencing data will be obtained but with good quality. On the other hand, if the concentration has been underestimated, there is a risk of having too many clusters and overloading the flow cell, obtaining few poor-quality data.
2. Library preparation. Its high criticality is because is a fundamental step to obtain an optimum number of clusters. Variations in the volume of *Tagment DNA enzyme 1* during library preparation could prevent the library obtained from meeting the quality acceptance criteria and therefore not being suitable for loading on the sequencer. In fact, if the quantity of libraries on the flow cell is overestimated, the number of clusters will be low, and few sequencings will be obtained but with good quality. On the other hand, if the concentration has been underestimated, there is a risk of having too many clusters and overloading the flow cell, obtaining few poor-quality data.

During the test, robustness will be assessed by identifying the range of:

- The library concentration which in any case allows a valid result to be obtained;
- The volume of *Tagment DNA enzyme* 1 (TDE1) which in any case allows a valid result to be obtained;

Acceptance criteria

All the assays done as part of this test should be valid in compliance with acceptance criteria.

The robustness range to be checked, as regards Library concentration, will be between 50 and 100 ng (target 75 ng);

The robustness range to be checked, as regards the volume of TDE1, will be between 1.6 and 2.4 µl (target 2 µl);

The results expected from this test are that the libraries produced meet the library acceptance criteria, also after the introduction of deliberate changes in the reagents used while running the test.

Verification procedure

This test has been carried out using the DNA extracted from *GS115 Pichia pastoris Yeast Strain* (Invitrogen) during specificity verification, introducing the following changes:

All the reactions described below have been done as single. See also Table 4.

- Prepare a library starting with 50 ng of genomic DNA in the reaction, following the Nextera protocol;
- Prepare a library starting with 100 ng of genomic DNA in the reaction, following the Nextera protocol;

- Prepare a library starting with 75 ng of genomic DNA in the reaction, changing the Nextera protocol to 1.6 µl of TDE;
- Prepare a library starting with 75 ng of genomic DNA in the reaction, changing the Nextera protocol to 2.4 µl of TDE;

Assess the quality of the libraries on completion of preparation. Libraries should be quantified using the Qubit® 2.0 fluorimeter with the dsDNA HS Assay Kit. Quality control of libraries is done using 1 µl of library analyzed on the Bioanalyzer using the Agilent DNA 12000 kit.

Table 4 – Summary of experimental conditions for the method's robustness verification

Sample Name	*Starting concentration of gDNA*	*TDE1 volume*	*Validity*
Library A	*50ng*	*2* µl (target)	***YES***
Library B	*100ng*	*2* µl (target)	***YES***
Library C	*75 ng* (target)	*1,6* µl	***YES***
Library D	*75 ng* (target)	2.4 µl	***YES***

Results

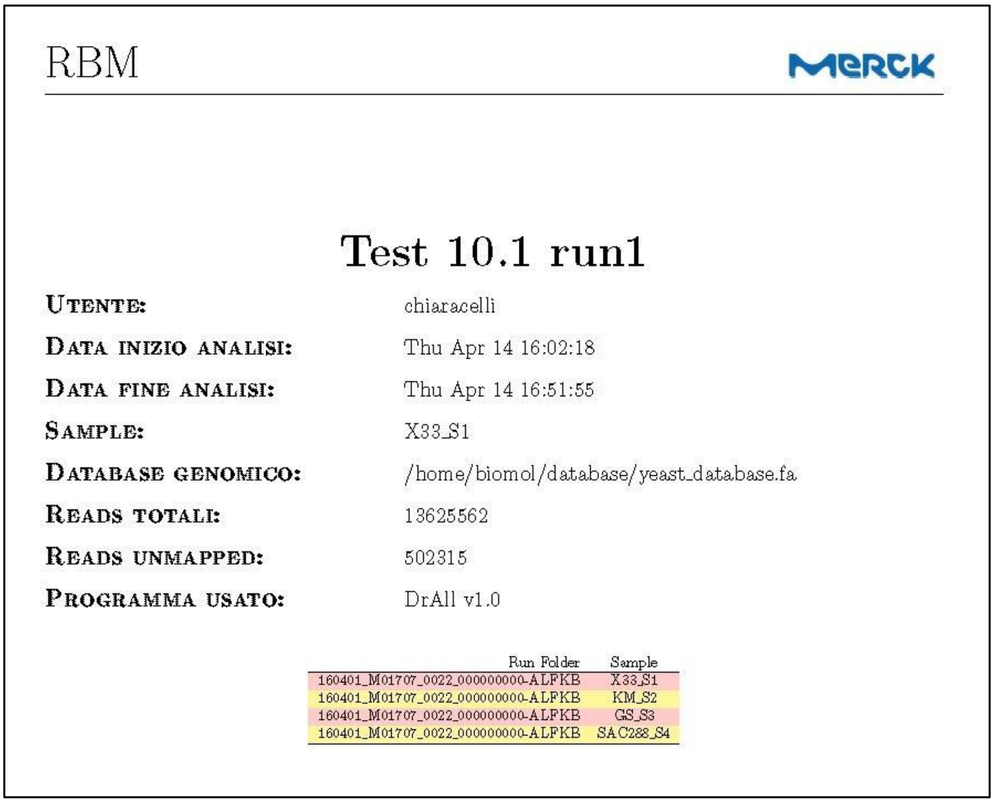
Method Specificity results

The method Specificity has been verified by preparing 8 different libraries and performing two separate MiSeq runs (Figure 21 and Figure 26). For each MiSeq run 4 libraries were pooled and analyzed in one single flow cell, as follow:

Run1: *P. pastoris X-33* (Invitrogen) - Figure 22
P. pastoris KM71H (Invitrogen) - Figure 23
P. pastoris GS115 (Invitrogen) - Figure 24
S. cerevisiae S288c (ATCC204508) - Figure 25

Run2: *Cyberlindnera jadinii* (ATCC18201) - Figure 27
Candida glabrata (DBVPG3178) - Figure 28
Recombinant *P. pastoris X-33* cell bank (RCB) - Figure 29
S. cerevisiae CBS1171 (ATCC18824) - Figure 30

All raw data produced by sequencing these yeast standard microorganisms has been further analyzed through *DrAll* pipeline to confirm the ability in genus and species identification.



RBM

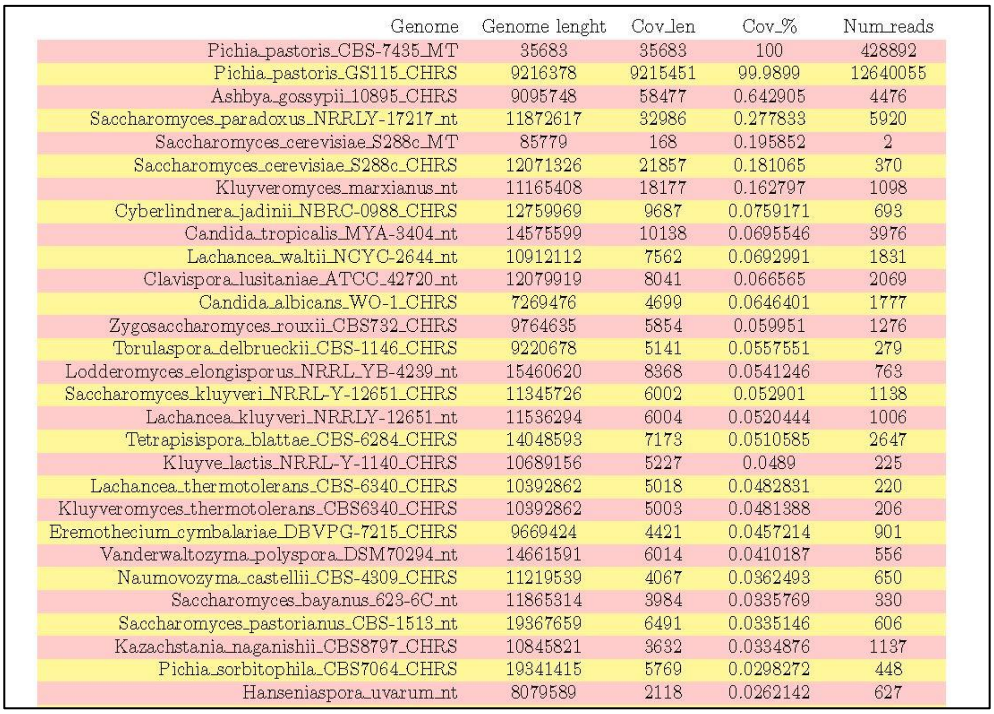
Test 10.1 run1

UTENTE:	chiaracelli
DATA INIZIO ANALISI:	Thu Apr 14 16:02:18
DATA FINE ANALISI:	Thu Apr 14 16:51:55
SAMPLE:	X33_S1
DATABASE GENOMICO:	/home/biomol/database/yeast_database.fa
READS TOTALI:	13625562
READS UNMAPPED:	502315
PROGRAMMA USATO:	DrAll v1.0

Run Folder	Sample
160401_M01707_0022_000000000-ALFKB	X33_S1
160401_M01707_0022_000000000-ALFKB	KM_S2
160401_M01707_0022_000000000-ALFKB	GS_S3
160401_M01707_0022_000000000-ALFKB	SAC288_S4

Figure 21 – Run1 result report for genus and species identification

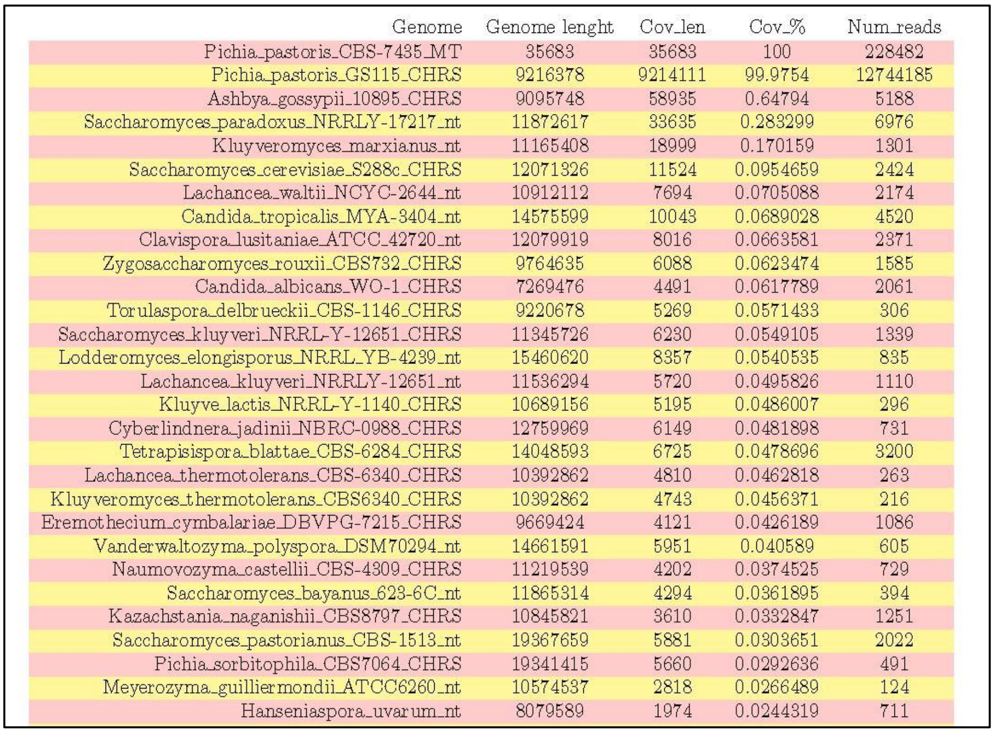
Pichia pastoris X-33 (Invitrogen)



Genome	Genome lenght	Cov_len	Cov_%	Num_reads
Pichia_pastoris_CBS-7435_MT	35683	35683	100	428892
Pichia_pastoris_GS115_CHRS	9216378	9215451	99.9899	12640055
Ashbya_gossypii_10895_CHRS	9095748	58477	0.642905	4476
Saccharomyces_paradoxus_NRRLY-17217_nt	11872617	32986	0.277833	5920
Saccharomyces_cerevisiae_S288c_MT	85779	168	0.195852	2
Saccharomyces_cerevisiae_S288c_CHRS	12071326	21857	0.181065	370
Kluyveromyces_marxianus_nt	11165408	18177	0.162797	1098
Cyberlindnera_jadinii_NBRC-0988_CHRS	12759969	9687	0.0759171	693
Candida_tropicalis_MYA-3404_nt	14575599	10138	0.0695546	3976
Lachancea_waltii_NCYC-2644_nt	10912112	7562	0.0692991	1831
Clavispora_lusitaniae_ATCC_42720_nt	12079919	8041	0.066565	2069
Candida_albicans_WO-1_CHRS	7269476	4699	0.0646401	1777
Zygosaccharomyces_rouxii_CBS732_CHRS	9764635	5854	0.059951	1276
Torulaspora_delbrueckii_CBS-1146_CHRS	9220678	5141	0.0557551	279
Lodderomyces_elongisporus_NRRL_YB-4239_nt	15460620	8368	0.0541246	763
Saccharomyces_kluyveri_NRRL-Y-12651_CHRS	11345726	6002	0.052901	1138
Lachancea_kluyveri_NRRLY-12651_nt	11536294	6004	0.0520444	1006
Tetrapisispora_blattae_CBS-6284_CHRS	14048593	7173	0.0510585	2647
Kluyve_lactis_NRRL-Y-1140_CHRS	10689156	5227	0.0489	225
Lachancea_thermotolerans_CBS-6340_CHRS	10392862	5018	0.0482831	220
Kluyveromyces_thermotolerans_CBS6340_CHRS	10392862	5003	0.0481388	206
Eremothecium_cymbalariae_DBVPG-7215_CHRS	9669424	4421	0.0457214	901
Vanderwaltozyma_polyspora_DSM70294_nt	14661591	6014	0.0410187	556
Naumovozyma_castellii_CBS-4309_CHRS	11219539	4067	0.0362493	650
Saccharomyces_bayanus_623-6C_nt	11865314	3984	0.0335769	330
Saccharomyces_pastorianus_CBS-1513_nt	19367659	6491	0.0335146	606
Kazachstania_naganishii_CBS8797_CHRS	10845821	3632	0.0334876	1137
Pichia_sorbitophila_CBS7064_CHRS	19341415	5769	0.0298272	448
Hanseniaspora_uvarum_nt	8079589	2118	0.0262142	627

Figure 22 – Genus and Species result obtained for P. pastoris X-33

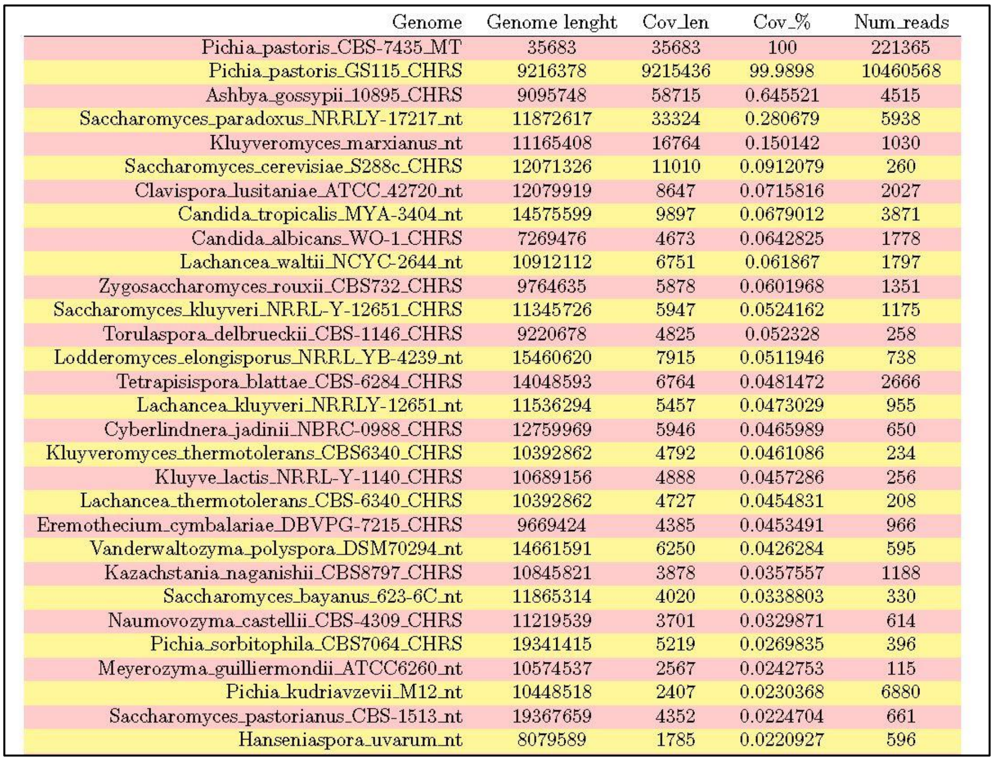
Pichia pastoris KM71H (Invitrogen)



Genome	Genome lenght	Cov_len	Cov_%	Num_reads
Pichia_pastoris_CBS-7435_MT	35683	35683	100	228482
Pichia_pastoris_GS115_CHRS	9216378	9214111	99.9754	12744185
Ashbya_gossypii_10895_CHRS	9095748	58935	0.64794	5188
Saccharomyces_paradoxus_NRRLY-17217_nt	11872617	33635	0.283299	6976
Kluyveromyces_marxianus_nt	11165408	18999	0.170159	1301
Saccharomyces_cerevisiae_S288c_CHRS	12071326	11524	0.0954659	2424
Lachancea_waltii_NCYC-2644_nt	10912112	7694	0.0705088	2174
Candida_tropicalis_MYA-3404_nt	14575599	10043	0.0689028	4520
Clavispora_lusitaniae_ATCC_42720_nt	12079919	8016	0.0663581	2371
Zygosaccharomyces_rouxii_CBS732_CHRS	9764635	6088	0.0623474	1585
Candida_albicans_WO-1_CHRS	7269476	4491	0.0617789	2061
Torulaspora_delbrueckii_CBS-1146_CHRS	9220678	5269	0.0571433	306
Saccharomyces_kluyveri_NRRL-Y-12651_CHRS	11345726	6230	0.0549105	1339
Lodderomyces_elongisporus_NRRL_YB-4239_nt	15460620	8357	0.0540535	835
Lachancea_kluyveri_NRRLY-12651_nt	11536294	5720	0.0495826	1110
Kluyve_lactis_NRRL-Y-1140_CHRS	10689156	5195	0.0486007	296
Cyberlindnera_jadinii_NBRC-0988_CHRS	12759969	6149	0.0481898	731
Tetrapisispora_blattae_CBS-6284_CHRS	14048593	6725	0.0478696	3200
Lachancea_thermotolerans_CBS-6340_CHRS	10392862	4810	0.0462818	263
Kluyveromyces_thermotolerans_CBS6340_CHRS	10392862	4743	0.0456371	216
Eremothecium_cymbalariae_DBVPG-7215_CHRS	9669424	4121	0.0426189	1086
Vanderwaltozyma_polyspora_DSM70294_nt	14661591	5951	0.040589	605
Naumovozyma_castellii_CBS-4309_CHRS	11219539	4202	0.0374525	729
Saccharomyces_bayanus_623-6C_nt	11865314	4294	0.0361895	394
Kazachstania_naganishii_CBS8797_CHRS	10845821	3610	0.0332847	1251
Saccharomyces_pastorianus_CBS-1513_nt	19367659	5881	0.0303651	2022
Pichia_sorbitophila_CBS7064_CHRS	19341415	5660	0.0292636	491
Meyerozyma_guilliermondii_ATCC6260_nt	10574537	2818	0.0266489	124
Hanseniaspora_uvarum_nt	8079589	1974	0.0244319	711

Figure 23 - Genus and Species result obtained for P. pastoris KM71H

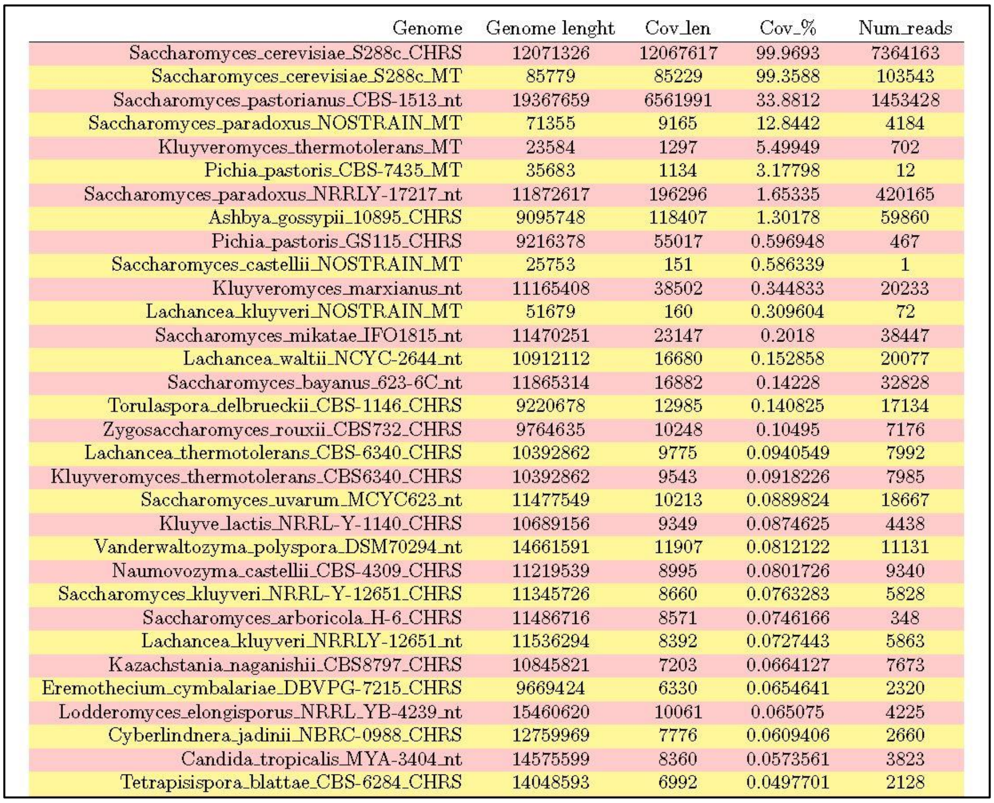
Pichia pastoris GS115 (Invitrogen)



Genome	Genome lenght	Cov_len	Cov_%	Num_reads
Pichia_pastoris_CBS-7435_MT	35683	35683	100	221365
Pichia_pastoris_GS115_CHRS	9216378	9215436	99.9898	10460568
Ashbya_gossypii_10895_CHRS	9095748	58715	0.645521	4515
Saccharomyces_paradoxus_NRRLY-17217_nt	11872617	33324	0.280679	5938
Kluyveromyces_marxianus_nt	11165408	16764	0.150142	1030
Saccharomyces_cerevisiae_S288c_CHRS	12071326	11010	0.0912079	260
Clavispora_lusitaniae_ATCC_42720_nt	12079919	8647	0.0715816	2027
Candida_tropicalis_MYA-3404_nt	14575599	9897	0.0679012	3871
Candida_albicans_WO-1_CHRS	7269476	4673	0.0642825	1778
Lachancea_waltii_NCYC-2644_nt	10912112	6751	0.061867	1797
Zygosaccharomyces_rouxii_CBS732_CHRS	9764635	5878	0.0601968	1351
Saccharomyces_kluyveri_NRRL-Y-12651_CHRS	11345726	5947	0.0524162	1175
Torulaspora_delbrueckii_CBS-1146_CHRS	9220678	4825	0.052328	258
Lodderomyces_elongisporus_NRRL_YB-4239_nt	15460620	7915	0.0511946	738
Tetrapisispora_blattae_CBS-6284_CHRS	14048593	6764	0.0481472	2666
Lachancea_kluyveri_NRRLY-12651_nt	11536294	5457	0.0473029	955
Cyberlindnera_jadinii_NBRC-0988_CHRS	12759969	5946	0.0465989	650
Kluyveromyces_thermotolerans_CBS6340_CHRS	10392862	4792	0.0461086	234
Kluyve_lactis_NRRL-Y-1140_CHRS	10689156	4888	0.0457286	256
Lachancea_thermotolerans_CBS-6340_CHRS	10392862	4727	0.0454831	208
Eremothecium_cymbalariae_DBVPG-7215_CHRS	9669424	4385	0.0453491	966
Vanderwaltozyma_polyspora_DSM70294_nt	14661591	6250	0.0426284	595
Kazachstania_naganishii_CBS8797_CHRS	10845821	3878	0.0357557	1188
Saccharomyces_bayanus_623-6C_nt	11865314	4020	0.0338803	330
Naumovozyma_castellii_CBS-4309_CHRS	11219539	3701	0.0329871	614
Pichia_sorbitophila_CBS7064_CHRS	19341415	5219	0.0269835	396
Meyerozyma_guilliermondii_ATCC6260_nt	10574537	2567	0.0242753	115
Pichia_kudriavzevii_M12_nt	10448518	2407	0.0230368	6880
Saccharomyces_pastorianus_CBS-1513_nt	19367659	4352	0.0224704	661
Hanseniaspora_uvarum_nt	8079589	1785	0.0220927	596

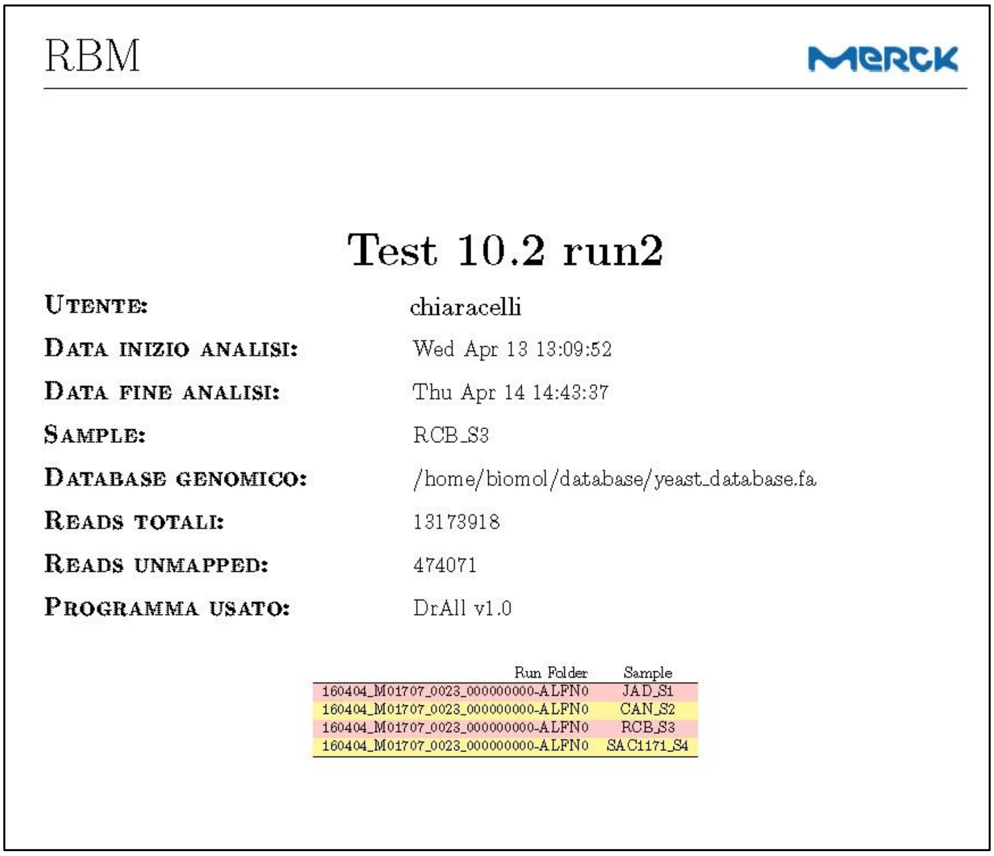
Figure 24 - Genus and Species result obtained for P. pastoris GS115

Saccharomyces cerevisiae S288c (ATCC204508)



Genome	Genome lenght	Cov_len	Cov_%	Num_reads
Saccharomyces_cerevisiae_S288c_CHRS	12071326	12067617	99.9693	7364163
Saccharomyces_cerevisiae_S288c_MT	85779	85229	99.3588	103543
Saccharomyces_pastorianus_CBS-1513_nt	19367659	6561991	33.8812	1453428
Saccharomyces_paradoxus_NOSTRAIN_MT	71355	9165	12.8442	4184
Kluyveromyces_thermotolerans_MT	23584	1297	5.49949	702
Pichia_pastoris_CBS-7435_MT	35683	1134	3.17798	12
Saccharomyces_paradoxus_NRRLY-17217_nt	11872617	196296	1.65335	420165
Ashbya_gossypii_10895_CHRS	9095748	118407	1.30178	59860
Pichia_pastoris_GS115_CHRS	9216378	55017	0.596948	467
Saccharomyces_castellii_NOSTRAIN_MT	25753	151	0.586339	1
Kluyveromyces_marxianus_nt	11165408	38502	0.344833	20233
Lachancea_kluyveri_NOSTRAIN_MT	51679	160	0.309604	72
Saccharomyces_mikatae_IFO1815_nt	11470251	23147	0.2018	38447
Lachancea_waltii_NCYC-2644_nt	10912112	16680	0.152858	20077
Saccharomyces_bayanus_623-6C_nt	11865314	16882	0.14228	32828
Torulaspora_delbrueckii_CBS-1146_CHRS	9220678	12985	0.140825	17134
Zygosaccharomyces_rouxii_CBS732_CHRS	9764635	10248	0.10495	7176
Lachancea_thermotolerans_CBS-6340_CHRS	10392862	9775	0.0940549	7992
Kluyveromyces_thermotolerans_CBS6340_CHRS	10392862	9543	0.0918226	7985
Saccharomyces_uvarum_MCYC623_nt	11477549	10213	0.0889824	18667
Kluyve_lactis_NRRL-Y-1140_CHRS	10689156	9349	0.0874625	4438
Vanderwaltozyma_polyspora_DSM70294_nt	14661591	11907	0.0812122	11131
Naumovozyma_castellii_CBS-4309_CHRS	11219539	8995	0.0801726	9340
Saccharomyces_kluyveri_NRRL-Y-12651_CHRS	11345726	8660	0.0763283	5828
Saccharomyces_arboricola_H-6_CHRS	11486716	8571	0.0746166	348
Lachancea_kluyveri_NRRLY-12651_nt	11536294	8392	0.0727443	5863
Kazachstania_naganishii_CBS8797_CHRS	10845821	7203	0.0664127	7673
Eremothecium_cymbalariae_DBVPG-7215_CHRS	9669424	6330	0.0654641	2320
Lodderomyces_elongisporus_NRRL_YB-4239_nt	15460620	10061	0.065075	4225
Cyberlindnera_jadinii_NBRC-0988_CHRS	12759969	7776	0.0609406	2660
Candida_tropicalis_MYA-3404_nt	14575599	8360	0.0573561	3823
Tetrapisispora_blattae_CBS-6284_CHRS	14048593	6992	0.0497701	2128

Figure 25 - Genus and Species result obtained for S. cerevisiae S288c



RBM

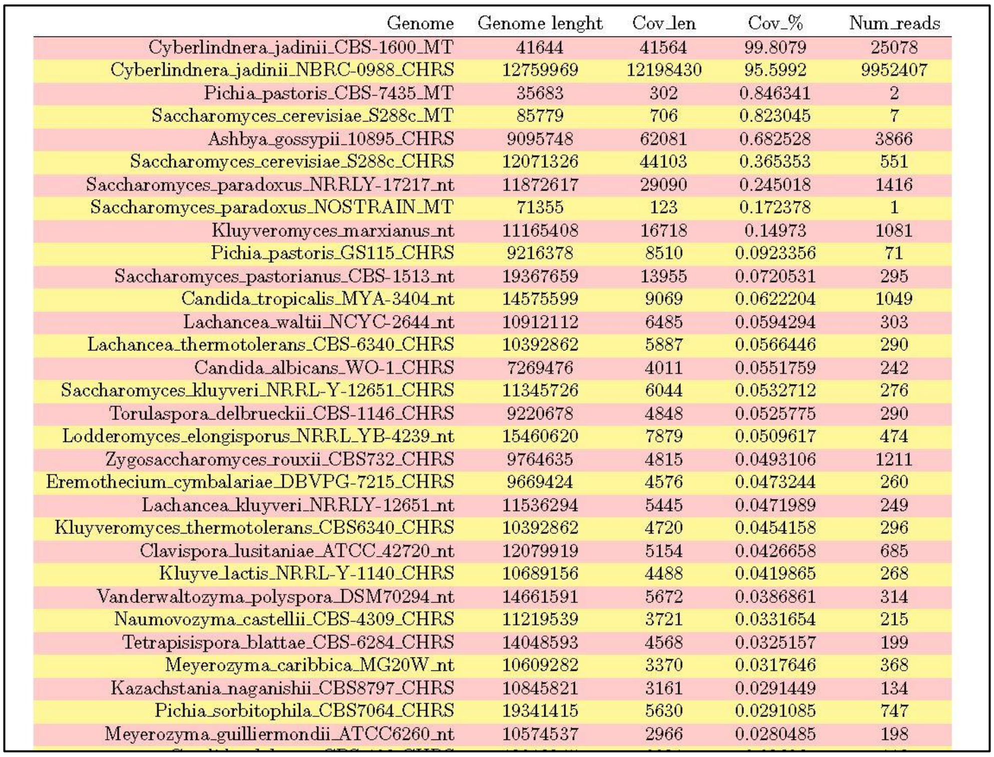
Test 10.2 run2

Utente:	chiaracelli
Data inizio analisi:	Wed Apr 13 13:09:52
Data fine analisi:	Thu Apr 14 14:43:37
Sample:	RCB_S3
Database genomico:	/home/biomol/database/yeast_database.fa
Reads totali:	13173918
Reads unmapped:	474071
Programma usato:	DrAll v1.0

Run Folder	Sample
160404_M01707_0023_000000000-ALFN0	JAD_S1
160404_M01707_0023_000000000-ALFN0	CAN_S2
160404_M01707_0023_000000000-ALFN0	RCB_S3
160404_M01707_0023_000000000-ALFN0	SAC1171_S4

Figure 26 – Run2 result report for genus and species identification

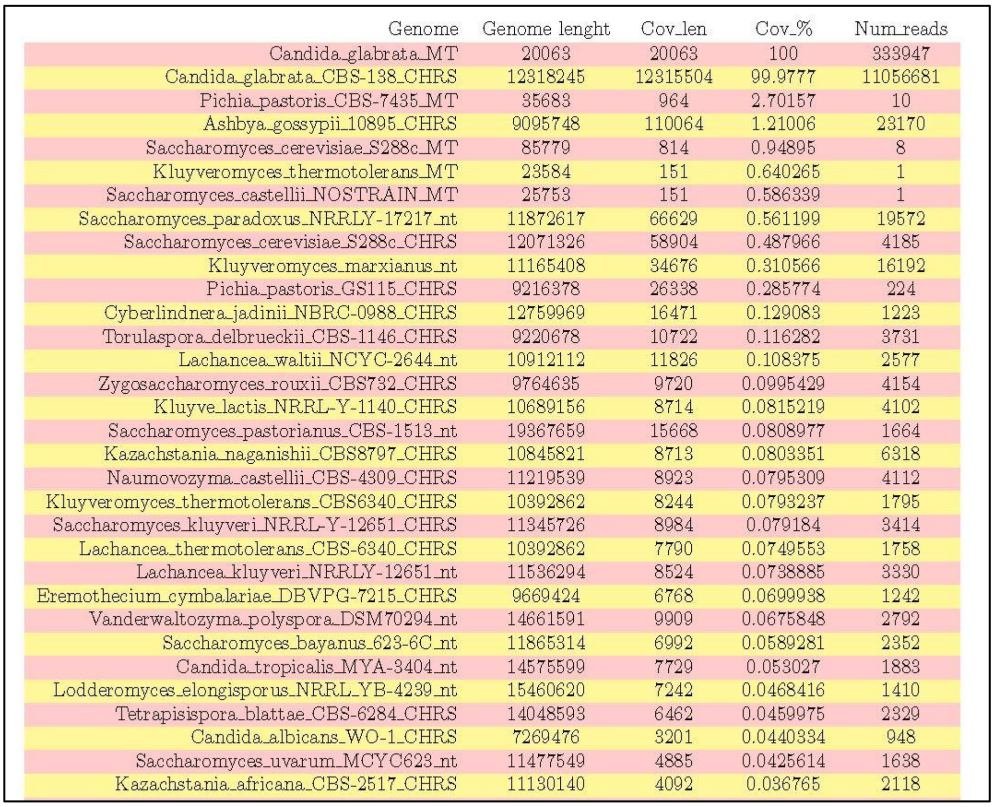
Cyberlindnera jadinii (ATCC18201)



Genome	Genome lenght	Cov_len	Cov_%	Num_reads
Cyberlindnera_jadinii_CBS-1600_MT	41644	41564	99.8079	25078
Cyberlindnera_jadinii_NBRC-0988_CHRS	12759969	12198430	95.5992	9952407
Pichia_pastoris_CBS-7435_MT	35683	302	0.846341	2
Saccharomyces_cerevisiae_S288c_MT	85779	706	0.823045	7
Ashbya_gossypii_10895_CHRS	9095748	62081	0.682528	3866
Saccharomyces_cerevisiae_S288c_CHRS	12071326	44103	0.365353	551
Saccharomyces_paradoxus_NRRLY-17217_nt	11872617	29090	0.245018	1416
Saccharomyces_paradoxus_NOSTRAIN_MT	71355	123	0.172378	1
Kluyveromyces_marxianus_nt	11165408	16718	0.14973	1081
Pichia_pastoris_GS115_CHRS	9216378	8510	0.0923356	71
Saccharomyces_pastorianus_CBS-1513_nt	19367659	13955	0.0720531	295
Candida_tropicalis_MYA-3404_nt	14575599	9069	0.0622204	1049
Lachancea_waltii_NCYC-2644_nt	10912112	6485	0.0594294	303
Lachancea_thermotolerans_CBS-6340_CHRS	10392862	5887	0.0566446	290
Candida_albicans_WO-1_CHRS	7269476	4011	0.0551759	242
Saccharomyces_kluyveri_NRRL-Y-12651_CHRS	11345726	6044	0.0532712	276
Torulaspora_delbrueckii_CBS-1146_CHRS	9220678	4848	0.0525775	290
Lodderomyces_elongisporus_NRRL_YB-4239_nt	15460620	7879	0.0509617	474
Zygosaccharomyces_rouxii_CBS732_CHRS	9764635	4815	0.0493106	1211
Eremothecium_cymbalariae_DBVPG-7215_CHRS	9669424	4576	0.0473244	260
Lachancea_kluyveri_NRRLY-12651_nt	11536294	5445	0.0471989	249
Kluyveromyces_thermotolerans_CBS6340_CHRS	10392862	4720	0.0454158	296
Clavispora_lusitaniae_ATCC_42720_nt	12079919	5154	0.0426658	685
Kluyve_lactis_NRRL-Y-1140_CHRS	10689156	4488	0.0419865	268
Vanderwaltozyma_polyspora_DSM70294_nt	14661591	5672	0.0386861	314
Naumovozyma_castellii_CBS-4309_CHRS	11219539	3721	0.0331654	215
Tetrapisispora_blattae_CBS-6284_CHRS	14048593	4568	0.0325157	199
Meyerozyma_caribbica_MG20W_nt	10609282	3370	0.0317646	368
Kazachstania_naganishii_CBS8797_CHRS	10845821	3161	0.0291449	134
Pichia_sorbitophila_CBS7064_CHRS	19341415	5630	0.0291085	747
Meyerozyma_guilliermondii_ATCC6260_nt	10574537	2966	0.0280485	198

Figure 27 - Genus and species result obtained for Cyberlindnera jadinii

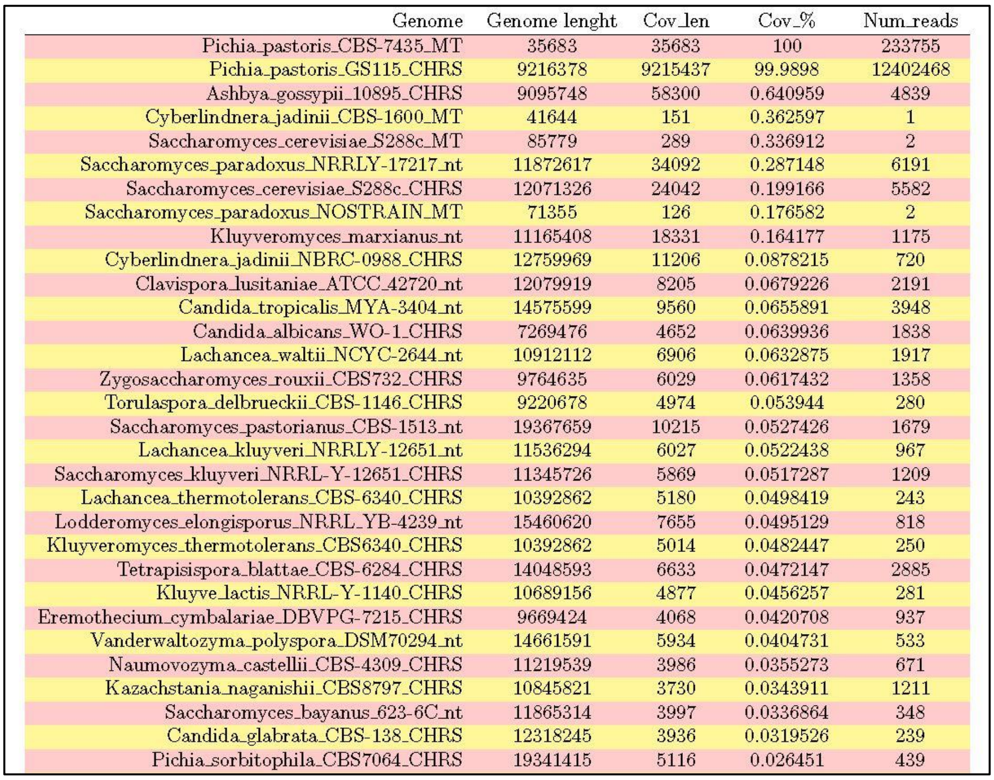
Candida glabrata (DBVPG3178)



Genome	Genome lenght	Cov_len	Cov_%	Num_reads
Candida_glabrata_MT	20063	20063	100	333947
Candida_glabrata_CBS-138_CHRS	12318245	12315504	99.9777	11056681
Pichia_pastoris_CBS-7435_MT	35683	964	2.70157	10
Ashbya_gossypii_10895_CHRS	9095748	110064	1.21006	23170
Saccharomyces_cerevisiae_S288c_MT	85779	814	0.94895	8
Kluyveromyces_thermotolerans_MT	23584	151	0.640265	1
Saccharomyces_castellii_NOSTRAIN_MT	25753	151	0.586339	1
Saccharomyces_paradoxus_NRRLY-17217_nt	11872617	66629	0.561199	19572
Saccharomyces_cerevisiae_S288c_CHRS	12071326	58904	0.487966	4185
Kluyveromyces_marxianus_nt	11165408	34676	0.310566	16192
Pichia_pastoris_GS115_CHRS	9216378	26338	0.285774	224
Cyberlindnera_jadinii_NBRC-0988_CHRS	12759969	16471	0.129083	1223
Torulaspora_delbrueckii_CBS-1146_CHRS	9220678	10722	0.116282	3731
Lachancea_waltii_NCYC-2644_nt	10912112	11826	0.108375	2577
Zygosaccharomyces_rouxii_CBS732_CHRS	9764635	9720	0.0995429	4154
Kluyve_lactis_NRRL-Y-1140_CHRS	10689156	8714	0.0815219	4102
Saccharomyces_pastorianus_CBS-1513_nt	19367659	15668	0.0808977	1664
Kazachstania_naganishii_CBS8797_CHRS	10845821	8713	0.0803351	6318
Naumovozyma_castellii_CBS-4309_CHRS	11219539	8923	0.0795309	4112
Kluyveromyces_thermotolerans_CBS6340_CHRS	10392862	8244	0.0793237	1795
Saccharomyces_kluyveri_NRRL-Y-12651_CHRS	11345726	8984	0.079184	3414
Lachancea_thermotolerans_CBS-6340_CHRS	10392862	7790	0.0749553	1758
Lachancea_kluyveri_NRRLY-12651_nt	11536294	8524	0.0738885	3330
Eremothecium_cymbalariae_DBVPG-7215_CHRS	9669424	6768	0.0699938	1242
Vanderwaltozyma_polyspora_DSM70294_nt	14661591	9909	0.0675848	2792
Saccharomyces_bayanus_623-6C_nt	11865314	6992	0.0589281	2352
Candida_tropicalis_MYA-3404_nt	14575599	7729	0.053027	1883
Lodderomyces_elongisporus_NRRL_YB-4239_nt	15460620	7242	0.0468416	1410
Tetrapisispora_blattae_CBS-6284_CHRS	14048593	6462	0.0459975	2329
Candida_albicans_WO-1_CHRS	7269476	3201	0.0440334	948
Saccharomyces_uvarum_MCYC623_nt	11477549	4885	0.0425614	1638
Kazachstania_africana_CBS-2517_CHRS	11130140	4092	0.036765	2118

Figure 28 - Genus and species result obtained for Candida glabrata

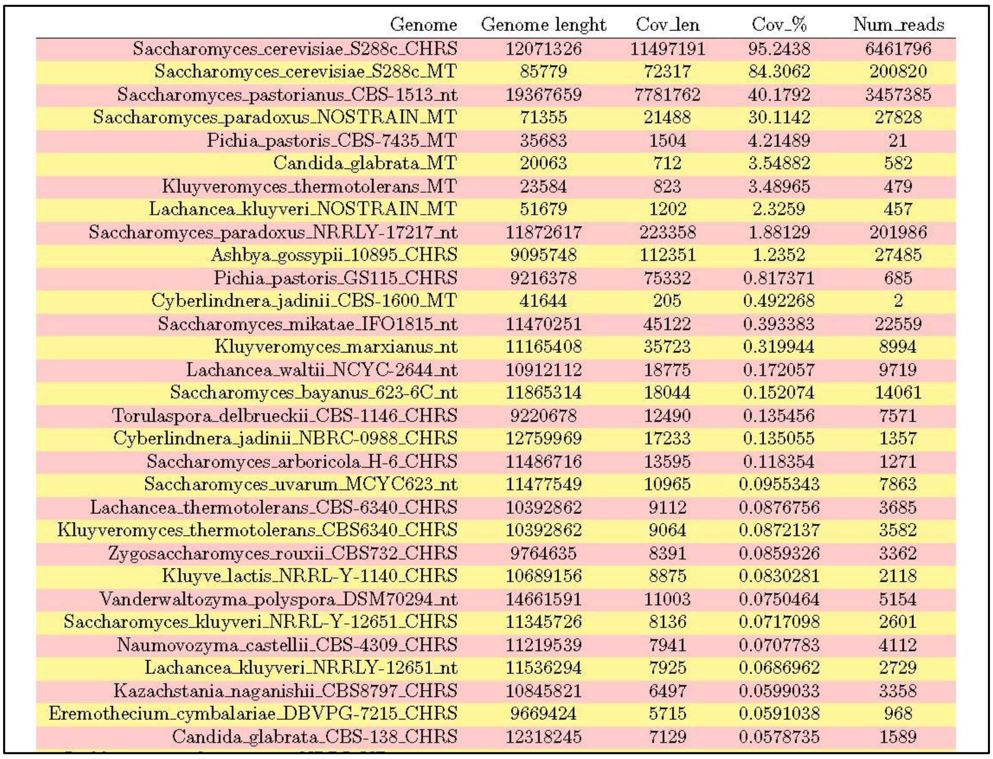
Recombinant *P. pastoris X-33* cell bank



Genome	Genome lenght	Cov_len	Cov_%	Num_reads
Pichia_pastoris_CBS-7435_MT	35683	35683	100	233755
Pichia_pastoris_GS115_CHRS	9216378	9215437	99.9898	12402468
Ashbya_gossypii_10895_CHRS	9095748	58300	0.640959	4839
Cyberlindnera_jadinii_CBS-1600_MT	41644	151	0.362597	1
Saccharomyces_cerevisiae_S288c_MT	85779	289	0.336912	2
Saccharomyces_paradoxus_NRRLY-17217_nt	11872617	34092	0.287148	6191
Saccharomyces_cerevisiae_S288c_CHRS	12071326	24042	0.199166	5582
Saccharomyces_paradoxus_NOSTRAIN_MT	71355	126	0.176582	2
Kluyveromyces_marxianus_nt	11165408	18331	0.164177	1175
Cyberlindnera_jadinii_NBRC-0988_CHRS	12759969	11206	0.0878215	720
Clavispora_lusitaniae_ATCC_42720_nt	12079919	8205	0.0679226	2191
Candida_tropicalis_MYA-3404_nt	14575599	9560	0.0655891	3948
Candida_albicans_WO-1_CHRS	7269476	4652	0.0639936	1838
Lachancea_waltii_NCYC-2644_nt	10912112	6906	0.0632875	1917
Zygosaccharomyces_rouxii_CBS732_CHRS	9764635	6029	0.0617432	1358
Torulaspora_delbrueckii_CBS-1146_CHRS	9220678	4974	0.053944	280
Saccharomyces_pastorianus_CBS-1513_nt	19367659	10215	0.0527426	1679
Lachancea_kluyveri_NRRLY-12651_nt	11536294	6027	0.0522438	967
Saccharomyces_kluyveri_NRRL-Y-12651_CHRS	11345726	5869	0.0517287	1209
Lachancea_thermotolerans_CBS-6340_CHRS	10392862	5180	0.0498419	243
Lodderomyces_elongisporus_NRRL_YB-4239_nt	15460620	7655	0.0495129	818
Kluyveromyces_thermotolerans_CBS6340_CHRS	10392862	5014	0.0482447	250
Tetrapisispora_blattae_CBS-6284_CHRS	14048593	6633	0.0472147	2885
Kluyve_lactis_NRRL-Y-1140_CHRS	10689156	4877	0.0456257	281
Eremothecium_cymbalariae_DBVPG-7215_CHRS	9669424	4068	0.0420708	937
Vanderwaltozyma_polyspora_DSM70294_nt	14661591	5934	0.0404731	533
Naumovozyma_castellii_CBS-4309_CHRS	11219539	3986	0.0355273	671
Kazachstania_naganishii_CBS8797_CHRS	10845821	3730	0.0343911	1211
Saccharomyces_bayanus_623-6C_nt	11865314	3997	0.0336864	348
Candida_glabrata_CBS-138_CHRS	12318245	3936	0.0319526	239
Pichia_sorbitophila_CBS7064_CHRS	19341415	5116	0.026451	439

Figure 29 - Genus and species result obtained for Recombinant cell bank

Saccharomyces cerevisiae CBS1171 (ATCC18824)



Genome	Genome lenght	Cov_len	Cov_%	Num_reads
Saccharomyces_cerevisiae_S288c_CHRS	12071326	11497191	95.2438	6461796
Saccharomyces_cerevisiae_S288c_MT	85779	72317	84.3062	200820
Saccharomyces_pastorianus_CBS-1513_nt	19367659	7781762	40.1792	3457385
Saccharomyces_paradoxus_NOSTRAIN_MT	71355	21488	30.1142	27828
Pichia_pastoris_CBS-7435_MT	35683	1504	4.21489	21
Candida_glabrata_MT	20063	712	3.54882	582
Kluyveromyces_thermotolerans_MT	23584	823	3.48965	479
Lachancea_kluyveri_NOSTRAIN_MT	51679	1202	2.3259	457
Saccharomyces_paradoxus_NRRLY-17217_nt	11872617	223358	1.88129	201986
Ashbya_gossypii_10895_CHRS	9095748	112351	1.2352	27485
Pichia_pastoris_GS115_CHRS	9216378	75332	0.817371	685
Cyberlindnera_jadinii_CBS-1600_MT	41644	205	0.492268	2
Saccharomyces_mikatae_IFO1815_nt	11470251	45122	0.393383	22559
Kluyveromyces_marxianus_nt	11165408	35723	0.319944	8994
Lachancea_waltii_NCYC-2644_nt	10912112	18775	0.172057	9719
Saccharomyces_bayanus_623-6C_nt	11865314	18044	0.152074	14061
Torulaspora_delbrueckii_CBS-1146_CHRS	9220678	12490	0.135456	7571
Cyberlindnera_jadinii_NBRC-0988_CHRS	12759969	17233	0.135055	1357
Saccharomyces_arboricola_H-6_CHRS	11486716	13595	0.118354	1271
Saccharomyces_uvarum_MCYC623_nt	11477549	10965	0.0955343	7863
Lachancea_thermotolerans_CBS-6340_CHRS	10392862	9112	0.0876756	3685
Kluyveromyces_thermotolerans_CBS6340_CHRS	10392862	9064	0.0872137	3582
Zygosaccharomyces_rouxii_CBS732_CHRS	9764635	8391	0.0859326	3362
Kluyve_lactis_NRRL-Y-1140_CHRS	10689156	8875	0.0830281	2118
Vanderwaltozyma_polyspora_DSM70294_nt	14661591	11003	0.0750464	5154
Saccharomyces_kluyveri_NRRL-Y-12651_CHRS	11345726	8136	0.0717098	2601
Naumovozyma_castellii_CBS-4309_CHRS	11219539	7941	0.0707783	4112
Lachancea_kluyveri_NRRLY-12651_nt	11536294	7925	0.0686962	2729
Kazachstania_naganishii_CBS8797_CHRS	10845821	6497	0.0599033	3358
Eremothecium_cymbalariae_DBVPG-7215_CHRS	9669424	5715	0.0591038	968
Candida_glabrata_CBS-138_CHRS	12318245	7129	0.0578735	1589

Figure 30 - Genus and Species result obtained for S. cerevisiae CBS117

All the results achieved during the specificity verification complied with the acceptance criteria and reflect the expected results for each of the standard microorganisms tested.

The obtained results for the method ability to correctly identify yeast genus and species is summarized in Table 5. Results and reports, as well as all Coverage % results, were obtained using the *DrAll v1.0* pipeline.

Acceptance criteria to determine the genus and species: coverage ≥85%.

The nomenclature used for the genomes in the database follows the rule: *Genus_species_strain_MT* or *Genus_species_strain_CHRS*, where “MT” refers to Mitochondrial genome and “CHRS” refers to Chromosome genome.

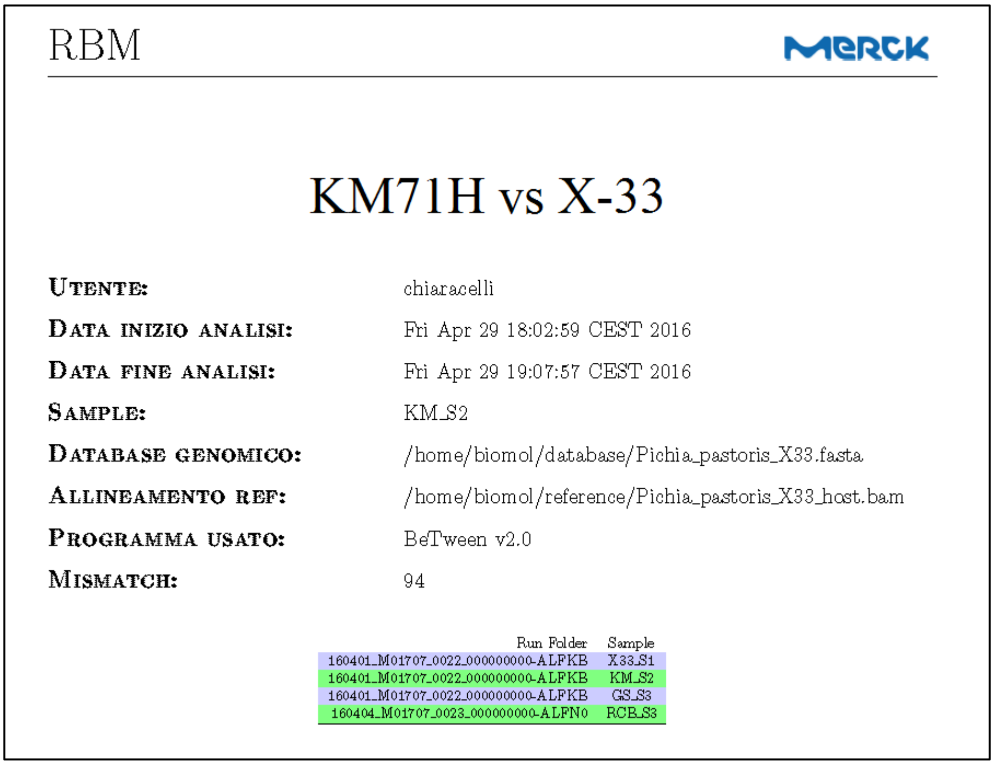
Table 5 - Specificity test results for genus and species identification

Sample Name	***ID name***	***Genus and Species Identified (Cov% ≥85.00%)***	***Cov%***
P. pastoris KM71H (Invitrogen)	***KM***	*Pichia pastoris MT*	***100***
		Pichia pastoris CHRS	***99.98***
P. pastoris GS115 (Invitrogen)	***GS***	*Pichia pastoris MT*	***100***
		Pichia pastoris CHRS	***99.99***
P. pastoris X-33 (Invitrogen)	***X-33***	*Pichia pastoris MT*	***100***
		Pichia pastoris CHRS	***99.99***
S. cerevisiae S288c (ATCC204508)	***SAC288***	*Saccharomyces cerevisiae CHRS*	***99.97***
		Saccharomyces cerevisiae MT	***99.36***
S. cerevisiae CBS1171 (ATCC18824)	***SAC1171***	*Saccharomyces cerevisiae CHRS*	***95.24***
		Saccharomyces cerevisiae MT	***84.30***
Cyberlindnera jadinii (ATCC18201)	***JAD***	*Cyberlindnera jadinii MT*	***99.81***
		Cyberlindnera jadinii CHRS	***95.60***
Candida glabrata (DBVPG3178)	***CAN***	*Candida glabrata MT*	***100***
		Candida glabrata CHRS	***99.98***
Recombinant P. pastoris X-33 Cell Bank	***RCB***	*Pichia pastoris MT*	***100***
		Pichia pastoris CHRS	***99.99***

After genus and species confirmation, the same raw data has been compared with their reference genome using the *BeTween* pipeline to detect any mismatch.

In detail, considering *Pichia pastoris X-33* and *Saccharomyces cerevisiae S288c* as standard references, the comparison of the sample strain (on the left) versus the reference genome (on the right) performed has been as follow:

P. pastoris KM71H	vs	*P. pastoris X-33* (Figure 31, Figure 32)
P. pastoris GS115	vs	*P. pastoris X-33* (Figure 33)
P. pastoris X-33	vs	*P. pastoris X-33* (Figure 34)
Recombinant P. pastoris X-33 cell bank	vs	*P. pastoris X-33* (Figure 35, Figure 36)
S. cerevisiae S288c	vs	*S. cerevisiae S288c* (Figure 37)
S. cerevisiae CBS1171	vs	*S. cerevisiae S288c* (Figure 38)



RBM

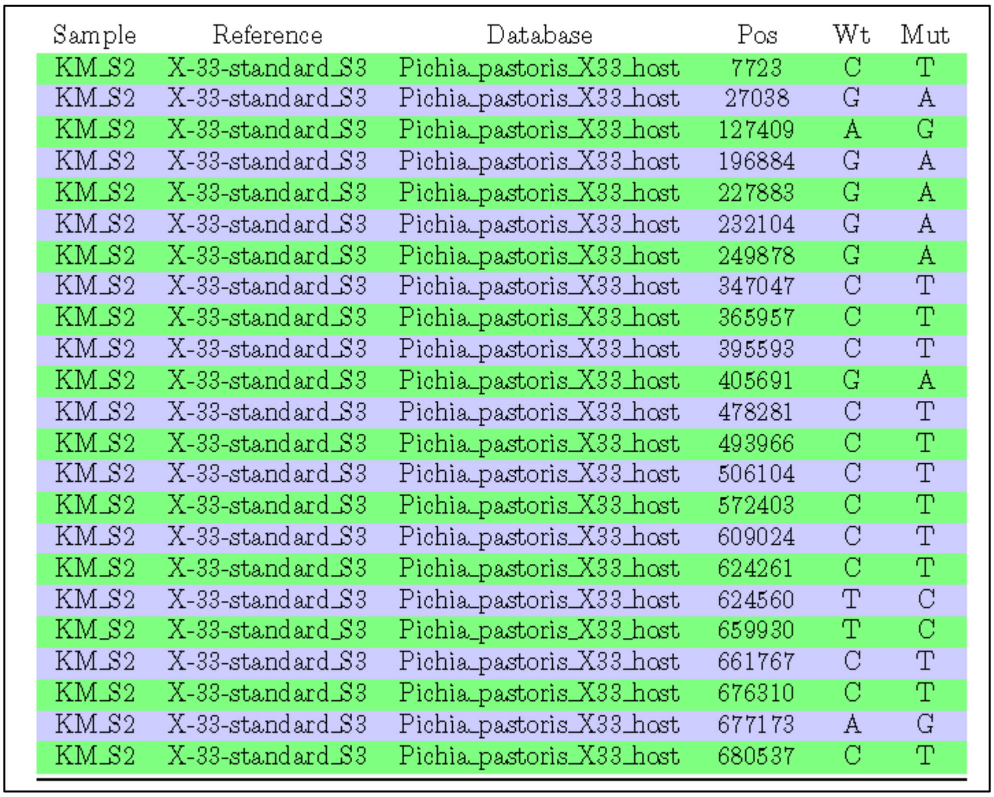
KM71H vs X-33

UTENTE:	chiaracelli
DATA INIZIO ANALISI:	Fri Apr 29 18:02:59 CEST 2016
DATA FINE ANALISI:	Fri Apr 29 19:07:57 CEST 2016
SAMPLE:	KM_S2
DATABASE GENOMICO:	/home/biomol/database/Pichia_pastoris_X33.fasta
ALLINEAMENTO REF:	/home/biomol/reference/Pichia_pastoris_X33_host.bam
PROGRAMMA USATO:	BeTween v2.0
MISMATCH:	94

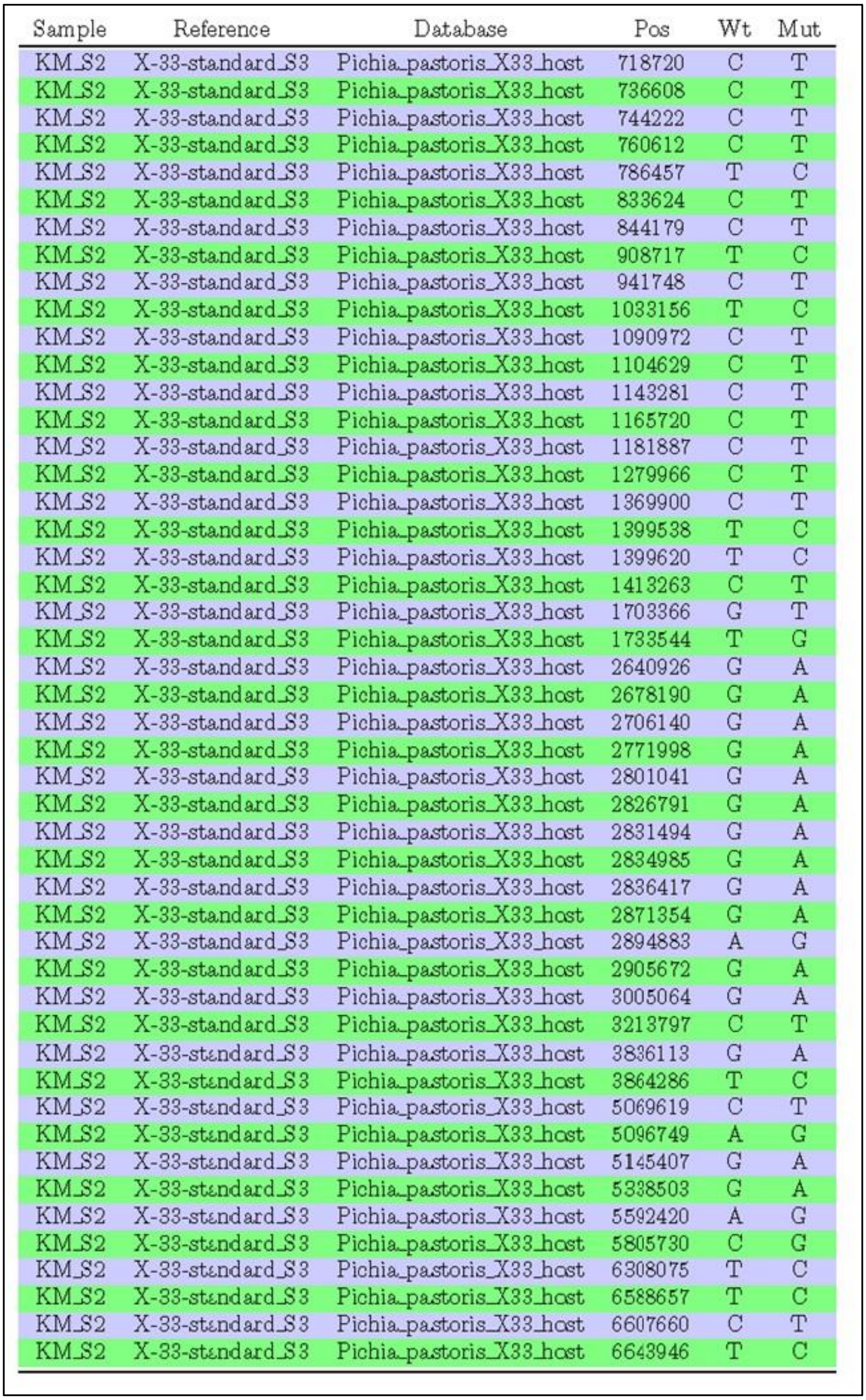
Run Folder	Sample
160401_M01707_0022_000000000-ALFKB	X33_S1
160401_M01707_0022_000000000-ALFKB	KM_S2
160401_M01707_0022_000000000-ALFKB	GS_S3
160404_M01707_0023_000000000-ALFN0	RCB_S3

Figure 31 - Strain result obtained for P. pastors KM71H vs P. pastors X-33

Here below are reported the 94 mismatches detected by the *BeTween* pipeline aligning KM71H against X-33 strain (Figure 31). In Figure 32 is also possible to see the position in the genome and the mutated nucleotide present in the KM71H strain.

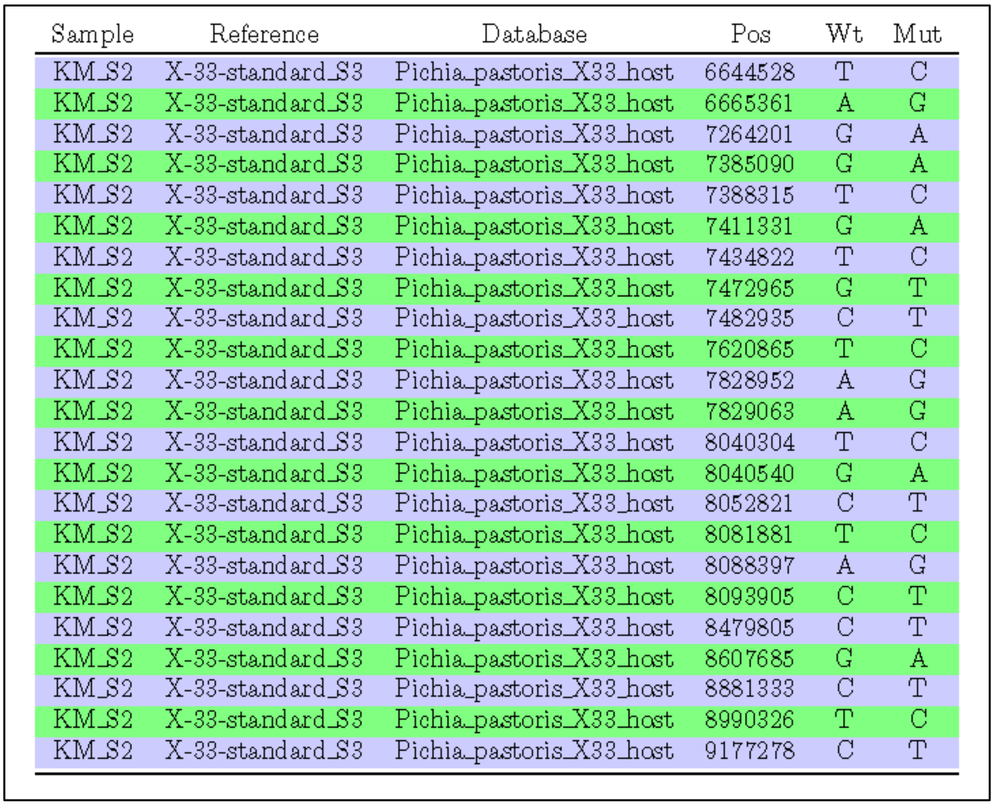


Sample	Reference	Database	Pos	Wt	Mut
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	7723	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	27038	G	A
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	127409	A	G
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	196884	G	A
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	227883	G	A
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	232104	G	A
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	249878	G	A
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	347047	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	365957	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	395593	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	405691	G	A
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	478281	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	493966	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	506104	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	572403	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	609024	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	624261	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	624560	T	C
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	659930	T	C
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	661767	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	676310	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	677173	A	G
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	680537	C	T



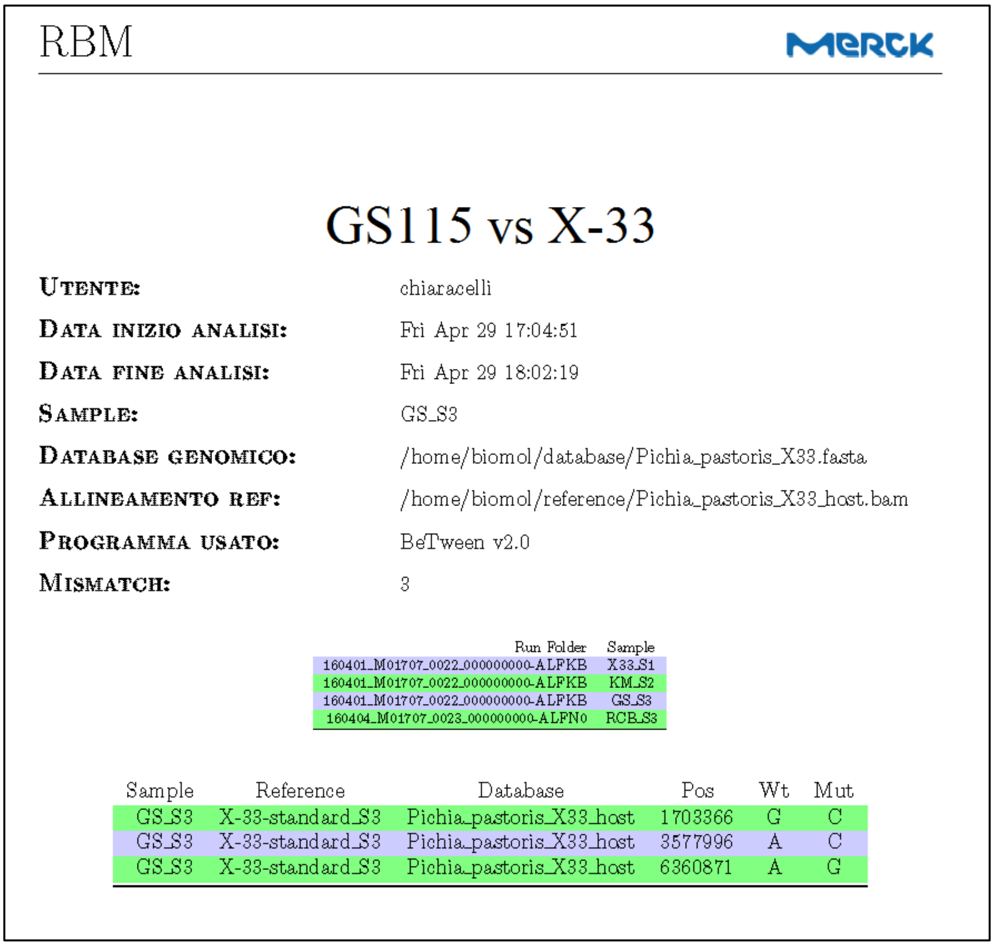
Sample	Reference	Database	Pos	Wt	Mut
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	718720	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	736608	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	744222	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	760612	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	786457	T	C
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	833624	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	844179	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	908717	T	C
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	941748	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	1033156	T	C
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	1090972	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	1104629	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	1143281	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	1165720	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	1181887	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	1279966	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	1369900	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	1399538	T	C
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	1399620	T	C
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	1413263	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	1703366	G	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	1733544	T	G
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	2640926	G	A
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	2678190	G	A
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	2706140	G	A
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	2771998	G	A
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	2801041	G	A
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	2826791	G	A
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	2831494	G	A
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	2834985	G	A
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	2836417	G	A
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	2871354	G	A
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	2894883	A	G
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	2905672	G	A
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	3005064	G	A
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	3213797	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	3836113	G	A
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	3864286	T	C
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	5069619	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	5096749	A	G
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	5145407	G	A
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	5338503	G	A
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	5592420	A	G
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	5805730	C	G
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	6308075	T	C
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	6588657	T	C
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	6607660	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	6643946	T	C

]



Sample	Reference	Database	Pos	Wt	Mut
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	6644528	T	C
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	6665361	A	G
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	7264201	G	A
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	7385090	G	A
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	7388315	T	C
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	7411331	G	A
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	7434822	T	C
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	7472965	G	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	7482935	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	7620865	T	C
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	7828952	A	G
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	7829063	A	G
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	8040304	T	C
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	8040540	G	A
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	8052821	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	8081881	T	C
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	8088397	A	G
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	8093905	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	8479805	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	8607685	G	A
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	8881333	C	T
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	8990326	T	C
KM_S2	X-33-standard_S3	Pichia_pastoris_X33_host	9177278	C	T

Figure 32 – Mismatches detected comparing P. pastoris KM71H vs P. pastoris X-33 strain



RBM

MERCK

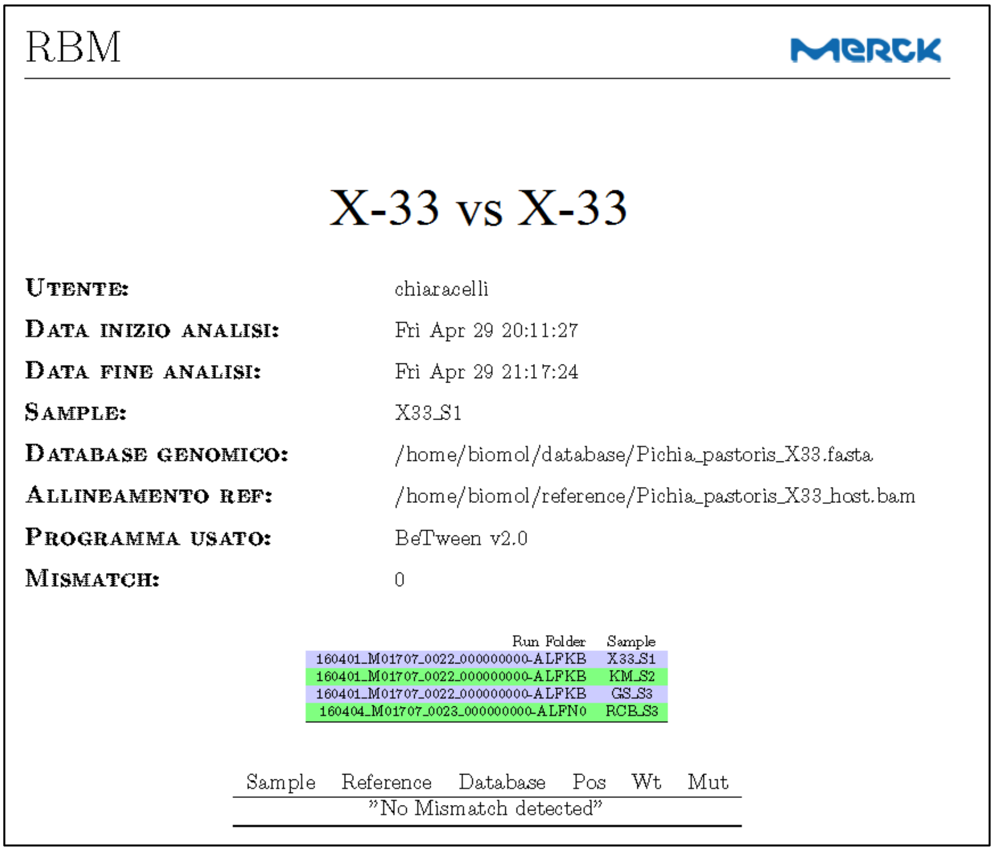
GS115 vs X-33

Utente:	chiaracelli
Data inizio analisi:	Fri Apr 29 17:04:51
Data fine analisi:	Fri Apr 29 18:02:19
Sample:	GS_S3
Database genomico:	/home/biomol/database/Pichia_pastoris_X33.fasta
Allineamento ref:	/home/biomol/reference/Pichia_pastoris_X33_host.bam
Programma usato:	BeTween v2.0
Mismatch:	3

Run Folder	Sample
160401_M01707_0022_000000000-ALFKB	X33_S1
160401_M01707_0022_000000000-ALFKB	KM_S2
160401_M01707_0022_000000000-ALFKB	GS_S3
160404_M01707_0023_000000000-ALFN0	RCB_S3

Sample	Reference	Database	Pos	Wt	Mut
GS_S3	X-33-standard_S3	Pichia_pastoris_X33_host	1703366	G	C
GS_S3	X-33-standard_S3	Pichia_pastoris_X33_host	3577996	A	C
GS_S3	X-33-standard_S3	Pichia_pastoris_X33_host	6360871	A	G

Figure 33 – Strain result obtained for P. pastors GS115 vs P. pastors X-33



RBM

X-33 vs X-33

Utente: chiaracelli

Data inizio analisi: Fri Apr 29 20:11:27

Data fine analisi: Fri Apr 29 21:17:24

Sample: X33_S1

Database genomico: /home/biomol/database/Pichia_pastoris_X33.fasta

Allineamento ref: /home/biomol/reference/Pichia_pastoris_X33_host.bam

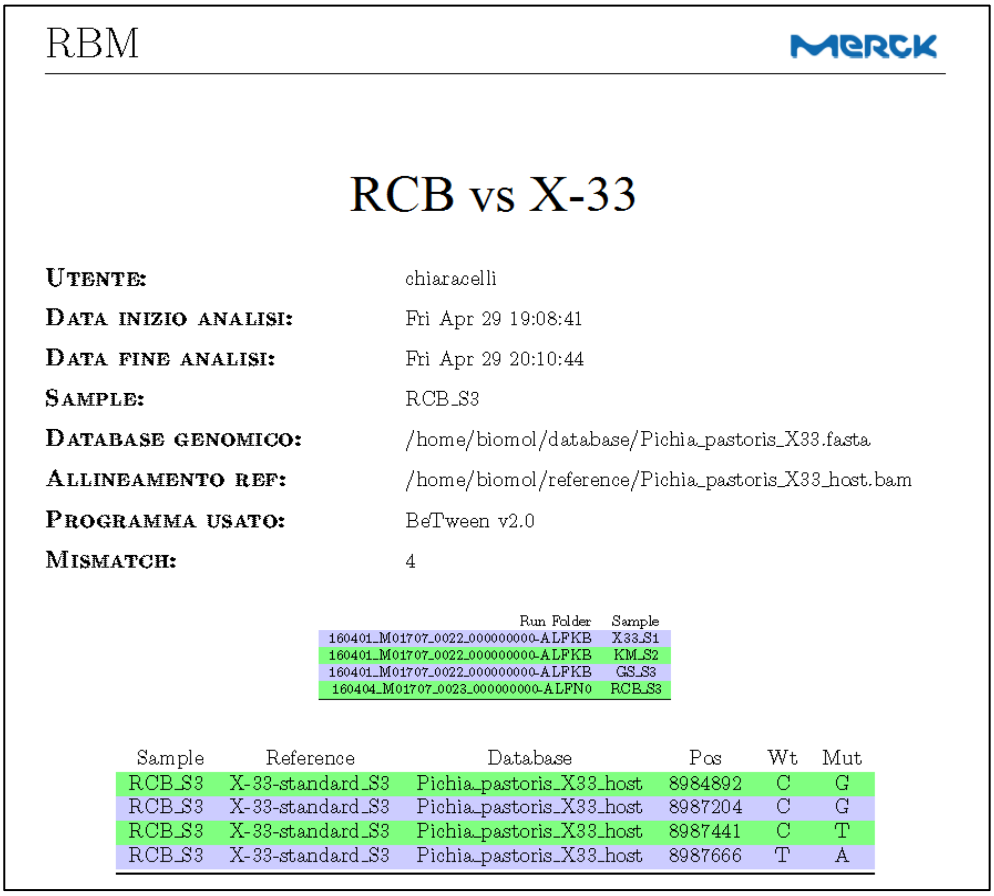
Programma usato: BeTween v2.0

Mismatch: 0

Run Folder	Sample
160401_M01707_0022_000000000-ALFKB	X33_S1
160401_M01707_0022_000000000-ALFKB	KM_S2
160401_M01707_0022_000000000-ALFKB	GS_S3
160404_M01707_0023_000000000-ALFN0	RCB_S3

Sample	Reference	Database	Pos	Wt	Mut
"No Mismatch detected"					

Figure 34 - Strain result obtained for P. pastors X-33 vs P. pastors X-33



RBM

RCB vs X-33

Utente: chiaracelli

Data inizio analisi: Fri Apr 29 19:08:41

Data fine analisi: Fri Apr 29 20:10:44

Sample: RCB_S3

Database genomico: /home/biomol/database/Pichia_pastoris_X33.fasta

Allineamento ref: /home/biomol/reference/Pichia_pastoris_X33_host.bam

Programma usato: BeTween v2.0

Mismatch: 4

Run Folder	Sample
160401_M01707_0022_000000000-ALFKB	X33_S1
160401_M01707_0022_000000000-ALFKB	KM_S2
160401_M01707_0022_000000000-ALFKB	GS_S3
160404_M01707_0023_000000000-ALFN0	RCB_S3

Sample	Reference	Database	Pos	Wt	Mut
RCB_S3	X-33-standard_S3	Pichia_pastoris_X33_host	8984892	C	G
RCB_S3	X-33-standard_S3	Pichia_pastoris_X33_host	8987204	C	G
RCB_S3	X-33-standard_S3	Pichia_pastoris_X33_host	8987441	C	T
RCB_S3	X-33-standard_S3	Pichia_pastoris_X33_host	8987666	T	A

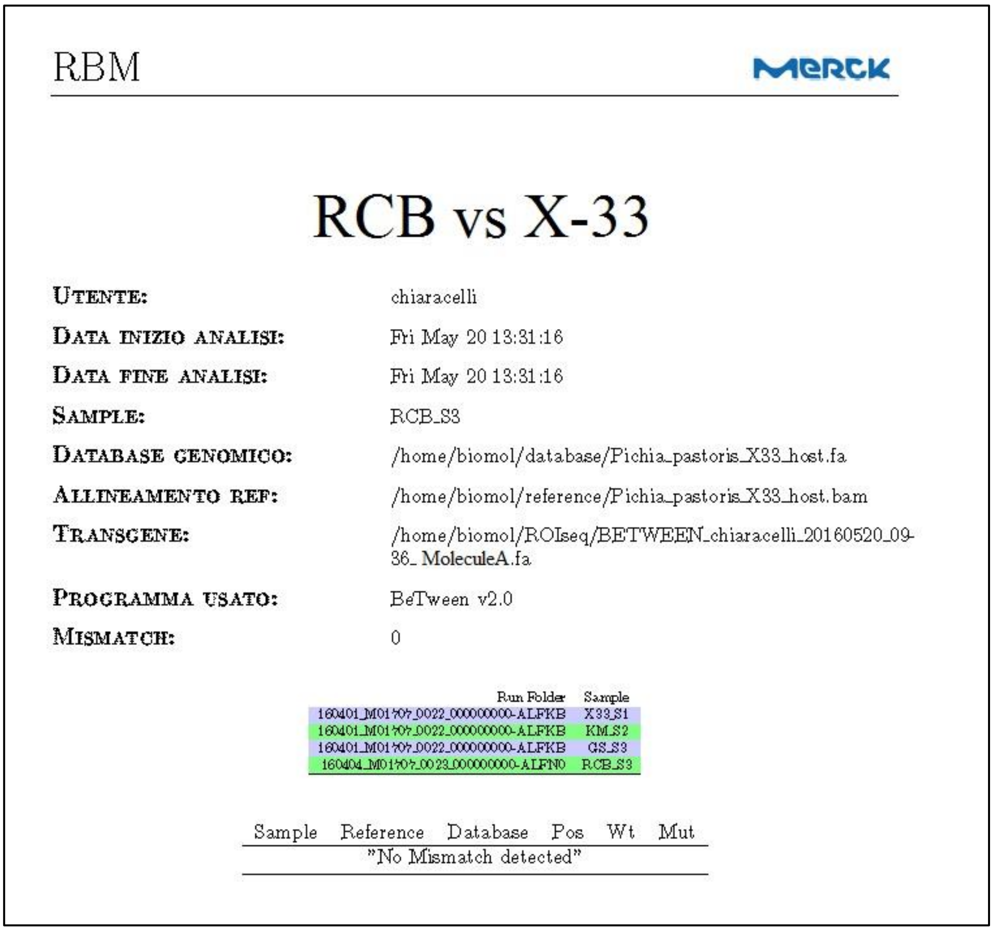
Figure 35 - Strain result obtained for Recombinant cell bank vs P. pastors X-33

In the frame of the Recombinant Cell Bank result, the first comparison showed a number of 4 detected mismatches, that is above the acceptance criteria limit set at ≤ 2.

During that experiment was considered that RCB is a *P. pastoris X-33* cell line engineered with a specific construct that is integrated in the yeast genome. For this reason, mutation detected can be associated with the insertion point in the genome.

In fact, it was performed a second analysis, introducing the Reference Transgene sequence "*MoleculeA.fa*". In this way, the *BeTween* pipeline excluded, from the first analysis, the genome portion affected by the engineering modification, and confirm the similarity between these two X-33 strains.

Therefore, the result in the report is "No mismatch detected" (Figure 36).



RBM

MERCK

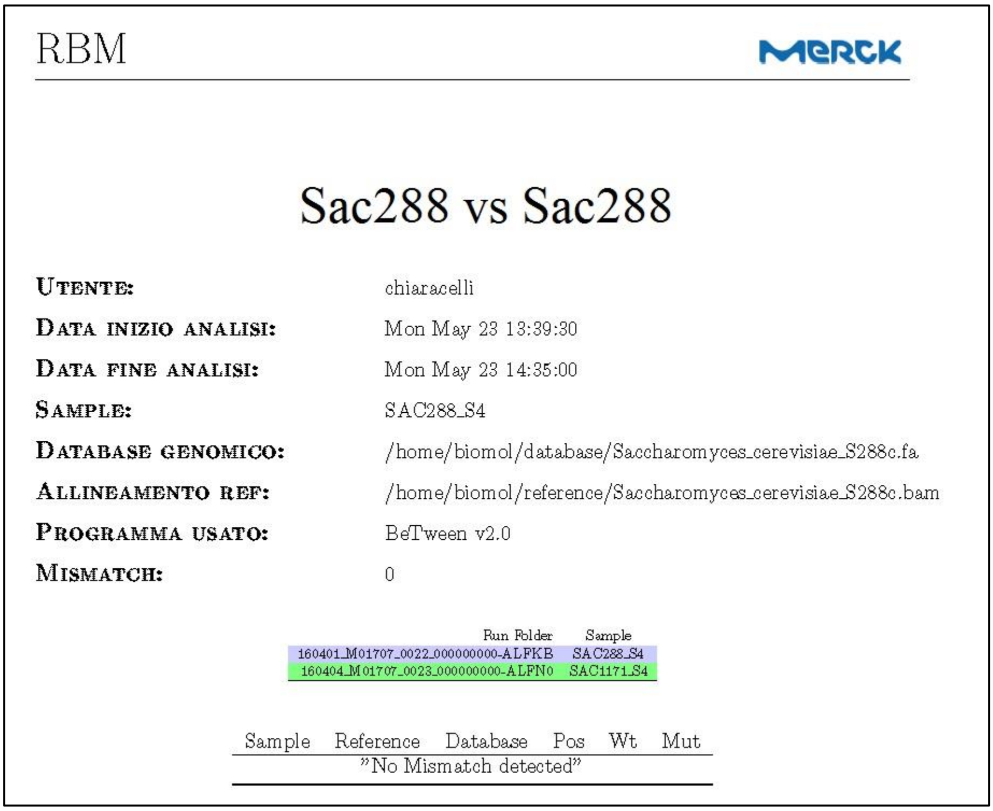
RCB vs X-33

UTENTE:	chiaracelli
DATA INIZIO ANALISI:	Fri May 20 13:31:16
DATA FINE ANALISI:	Fri May 20 13:31:16
SAMPLE:	RCB_S3
DATABASE GENOMICO:	/home/biomol/database/Pichia_pastoris_X33_host.fa
ALLINEAMENTO REF:	/home/biomol/reference/Pichia_pastoris_X33_host.bam
TRANSGENE:	/home/biomol/ROIseq/BETWEEN_chiaracelli_20160520_09-36_ MoleculeA.fa
PROGRAMMA USATO:	BeTween v2.0
MISMATCH:	0

Run Folder	Sample
160401_M01707_0022_000000000-ALFKB	X33_S1
160401_M01707_0022_000000000-ALFKB	KM_S2
160401_M01707_0022_000000000-ALFKB	GS_S3
160404_M01707_0023_000000000-ALFN0	RCB_S3

Sample	Reference	Database	Pos	Wt	Mut
"No Mismatch detected"					

Figure 36 - Strain result obtained for RCB vs P. pastors X-33 without Transgene sequence



RBM

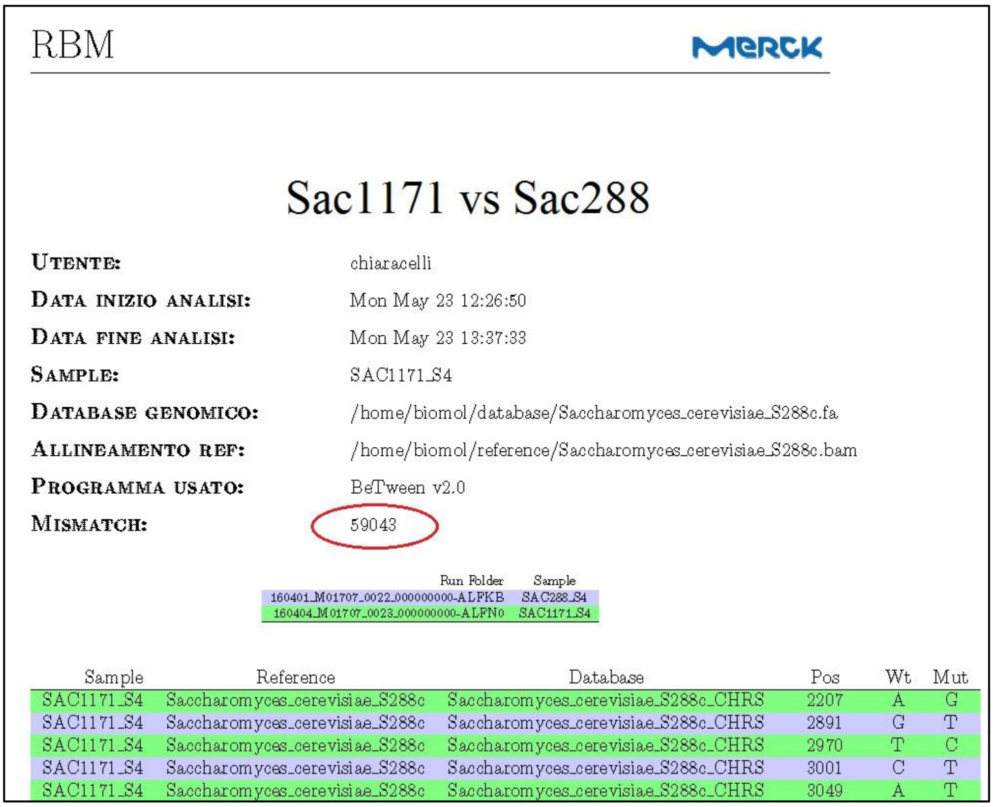
Sac288 vs Sac288

UTENTE:	chiaracelli
DATA INIZIO ANALISI:	Mon May 23 13:39:30
DATA FINE ANALISI:	Mon May 23 14:35:00
SAMPLE:	SAC288_S4
DATABASE GENOMICO:	/home/biomol/database/Saccharomyces_cerevisiae_S288c.fa
ALLINEAMENTO REF:	/home/biomol/reference/Saccharomyces_cerevisiae_S288c.bam
PROGRAMMA USATO:	BeTween v2.0
MISMATCH:	0

Run Folder	Sample
160401_M01707_0022_000000000-ALFKB	SAC288_S4
160404_M01707_0023_000000000-ALFN0	SAC1171_S4

Sample	Reference	Database	Pos	Wt	Mut
"No Mismatch detected"					

Figure 37 - Strain result obtained for S. cerevisiae S288c vs S. cerevisiae S288c



RBM

Sac1171 vs Sac288

UTENTE:	chiaracelli
DATA INIZIO ANALISI:	Mon May 23 12:26:50
DATA FINE ANALISI:	Mon May 23 13:37:33
SAMPLE:	SAC1171_S4
DATABASE GENOMICO:	/home/biomol/database/Saccharomyces_cerevisiae_S288c.fa
ALLINEAMENTO REF:	/home/biomol/reference/Saccharomyces_cerevisiae_S288c.bam
PROGRAMMA USATO:	BeTween v2.0
MISMATCH:	59043

Run Folder	Sample
160401_M01707_0022_000000000-ALFKB	SAC288_S4
160404_M01707_0023_000000000-ALFN0	SAC1171_S4

Sample	Reference	Database	Pos	Wt	Mut
SAC1171_S4	Saccharomyces_cerevisiae_S288c	Saccharomyces_cerevisiae_S288c_CHRS	2207	A	G
SAC1171_S4	Saccharomyces_cerevisiae_S288c	Saccharomyces_cerevisiae_S288c_CHRS	2891	G	T
SAC1171_S4	Saccharomyces_cerevisiae_S288c	Saccharomyces_cerevisiae_S288c_CHRS	2970	T	C
SAC1171_S4	Saccharomyces_cerevisiae_S288c	Saccharomyces_cerevisiae_S288c_CHRS	3001	C	T
SAC1171_S4	Saccharomyces_cerevisiae_S288c	Saccharomyces_cerevisiae_S288c_CHRS	3049	A	T

Figure 38 - Strain result obtained for S. cerevisiae CBS1171 vs S. cerevisiae S288c

Results obtained comparing *S. cerevisiae CBS1171* with *S. cerevisiae S288c* shows above 59’000 mismatches detected. For convenience, the report has been reported truncated (Figure 38).

A summary of all the results achieved for the strain identification method is reported in Table 6.

Acceptance criteria to determine the sample is belonging to the same strain as the host cell has been fixed at ≤ 2 mismatches.

Table 6 - Specificity test results for strain confirmation

Sample Name	*ID name*	*Reference Genome*	*TOTAL Mismatches*	*Strain confirmed*
P. pastoris KM71H (Invitrogen)	***KM71H***	*P. pastoris X-33*	***94***	***NO***
P. pastoris GS115 (Invitrogen)	***GS115***	*P. pastoris X-33*	***3***	***NO***
P. pastoris X-33 (Invitrogen)	***X-33***	*P. pastoris X-33*	***NO Mismatch detected***	***YES***
Recombinant P. pastoris X-33 Cell Bank	***RCB***	*P. pastoris X-33*	***NO Mismatch detected***	***YES***
S. cerevisiae S288c (ATCC204508)	***SAC288***	*S. cerevisiae S288c*	***NO Mismatch detected***	***YES***
S. cerevisiae CBS1171 (ATCC18824)	***SAC1171***	*S. cerevisiae S288c*	***59’043***	***NO***

Method Limit of Detection results

Table 7, Table 8 and Table 9 show the Limit of Detection (LOD) results obtained during the method genus and species identification tests. These results comply with the acceptance criteria and show that the method can detect a minimum of 10% of contaminant (different genus and species) in the main microorganism *P. pastoris X-33*. Considering that the LOD results could be influenced not only by the relative abundance of the strains but also by the total number of the sequenced reads, an additional check has been executed on the total number of reads achieved for each sample. All the results summarized in the tables below reached a minimum number of reads for each sample constant and greater than 4’000’000 reads that is acceptable taking into account that the method workflow is fixed and the minimum of reads sequenced is every time confirmed.

Table 7 - Limit of Detection (LOD) test results for genus and species identification – 1°

Sample Name (MIX)	***ID name***	***Genus and Species identified (Cov% ≥85.00%)***	***Cov%***
80% P. pastoris X-33 + 20% S. Cerevisiae	***MIX1***	*Pichia pastoris MT*	***100***
		Pichia pastoris CHRS	***99.99***
		Saccharomyces cerevisiae CHRS	***97.78***
		Saccharomyces cerevisiae MT	***86.61***
90% P. pastoris X-33 + 10% S. cerevisiae	***MIX2***	*Pichia pastoris MT*	***100***
		Pichia pastoris CHRS	***99.99***
		Saccharomyces cerevisiae CHRS	***98.51***
		Saccharomyces cerevisiae MT	***88.82***
99% P. pastoris X-33 + 1% S. cerevisiae	***MIX3***	*Pichia pastoris MT*	***100***
		Pichia pastoris CHRS	***99.99***
80% P. pastoris X-33 + 20% P. pastoris KM71H	***MIX4***	*Pichia pastoris MT*	***100***
		Pichia pastoris CHRS	***99.99***
90% P. pastoris X-33 + 10% P. pastoris KM71H	***MIX5***	*Pichia pastoris MT*	***100***
		Pichia pastoris CHRS	***99.99***
99% P. pastoris X-33 + 1% P. pastoris KM71H	***MIX6***	*Pichia pastoris MT*	***100***
		Pichia pastoris CHRS	***99.99***

Table 8 - Limit of Detection (LOD) test results for genus and species identification – 2°

Sample Name (MIX)	***ID name***	***Genus and Species identified (Cov% ≥85.00%)***	***Cov%***
80% P. pastoris X-33 + 20% S. cerevisiae	***MIX7***	*Pichia pastoris MT*	***100***
		Pichia pastoris CHRS	***99.99***
		Saccharomyces cerevisiae CHRS	***97.77***
		Saccharomyces cerevisiae MT	***96.01***
90% P. pastoris X-33 + 10% S. cerevisiae	***MIX8***	*Pichia pastoris MT*	***100***
		Pichia pastoris CHRS	***99.99***
		Saccharomyces cerevisiae CHRS	***98.21***
99% P. pastoris X-33 + 1% S. cerevisiae	***MIX9***	*Pichia pastoris MT*	***100***
		Pichia pastoris CHRS	***99.99***
80% P. pastoris X-33 + 20% P. pastoris KM71H	***MIX10***	*Pichia pastoris MT*	***100***
		Pichia pastoris CHRS	***99.99***
90% P. pastoris X-33 + 10% P. pastoris KM71H	***MIX11***	*Pichia pastoris MT*	***100***
		Pichia pastoris CHRS	***99.99***
99% P. pastoris X-33 + 1% P. pastoris KM71H	***MIX12***	*Pichia pastoris MT*	***100***
		Pichia pastoris CHRS	***99.99***

Table 9 - Limit of Detection (LOD) test results for genus and species identification – 3°

Sample Name (MIX)	***ID name***	***Genus and Species identified (Cov% ≥85.00%)***	***Cov%***
80% P. pastoris X-33 + 20% S. cerevisiae	***MIX13***	*Pichia pastoris MT*	***100***
		Pichia pastoris CHRS	***99.99***
		Saccharomyces cerevisiae CHRS	***98.98***
		Saccharomyces cerevisiae MT	***92.90***
90% P. pastoris X-33 + 10% S. cerevisiae	***MIX14***	*Pichia pastoris MT*	***100***
		Pichia pastoris CHRS	***99.99***
		Saccharomyces cerevisiae CHRS	***98.14***
99% P. pastoris X-33 + 1% S. cerevisiae	***MIX15***	*Pichia pastoris MT*	***100***
		Pichia pastoris CHRS	***99.99***
80% P. pastoris X-33 + 20% P. pastoris KM71H	***MIX16***	*Pichia pastoris MT*	***100***
		Pichia pastoris CHRS	***99.99***
90% P. pastoris X-33 + 10% P. pastoris KM71H	***MIX17***	*Pichia pastoris MT*	***100***
		Pichia pastoris CHRS	***99.99***
99% P. pastoris X-33 + 1% P. pastoris KM71H	***MIX18***	*Pichia pastoris MT*	***100***
		Pichia pastoris CHRS	***99.99***

Table 10 collects the Limit of Detection (LOD) results achieved during the method strain confirmation tests. These results comply with the acceptance criteria and show that the method can detect a minimum of 10% of contaminant (different strain) in the main microorganism.

Table 10 - Limit of Detection (LOD) test results for strain confirmation

Sample Name	*ID name*	*Reference Genome*	*TOTAL Mismatches*	*Strain confirmed*
80% P. pastoris X-33 + 20% P. pastoris KM71H	*MIX4*	*P. pastorisX-33*	***93***	***NO***
90% P. pastoris X-33 + 10% P. pastoris KM71H	*MIX5*	*P. pastorisX-33*	***68***	***NO***
99% P. pastoris X-33 + 1% P. pastoris KM71H	*MIX6*	*P. pastorisX-33*	***NO Mismatch detected***	***SI***
80% P. pastoris X-33 + 20% P. pastoris KM71H	*MIX10*	*P. pastorisX-33*	***92***	***NO***
90% P. pastoris X-33 + 10% P. pastoris KM71H	*MIX11*	*P. pastorisX-33*	***78***	***NO***
99% P. pastoris X-33 + 1% P. pastoris KM71H	*MIX12*	*P. pastorisX-33*	***NO Mismatch detected***	***SI***
80% P. pastoris X-33 + 20% P. pastoris KM71H	*MIX16*	*P. pastorisX-33*	***90***	***NO***
90% P. pastoris X-33 + 10% P. pastoris KM71H	*MIX17*	*P. pastorisX-33*	***79***	***NO***
99% P. pastoris X-33 + 1% P. pastoris KM71H	*MIX18*	*P. pastorisX-33*	***NO Mismatch detected***	***SI***

Method Robustness results

The method *Robustness* test aimed to demonstrate that despite deliberate changes introduced at critical steps, such as the quantitation of extracted DNA (in terms of quantity of DNA used to prepare the library) and library preparation (in terms of quantity of TDE1 enzyme), the method is robust within the ranges tested during the validation tests.

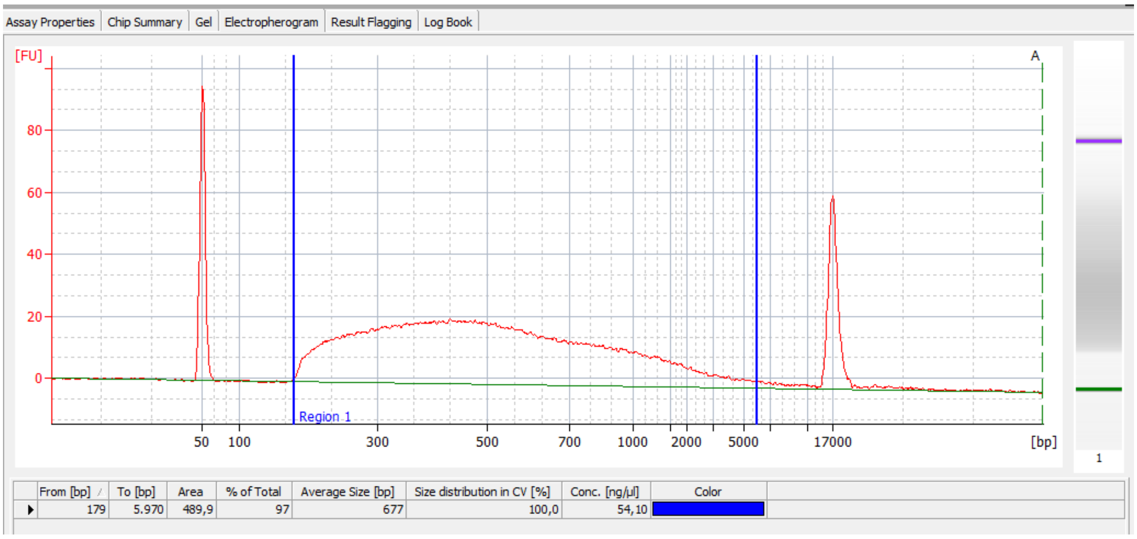
To check robustness, considering as a target the concentration of 75 ng of starting gDNA, as described in the method procedure, the starting quantity of DNA for library preparation were tested as follow:

- Library A: prepared starting with 50ng of gDNA in reaction;
- Library B: prepared starting with 100ng of gDNA in reaction.

While, considering as a target 2 µl of TDE1 enzyme, as described in the method procedure, the TDE1 volume to prepare the libraries C and D was tested in a range between 1.6 and 2.4 µl.

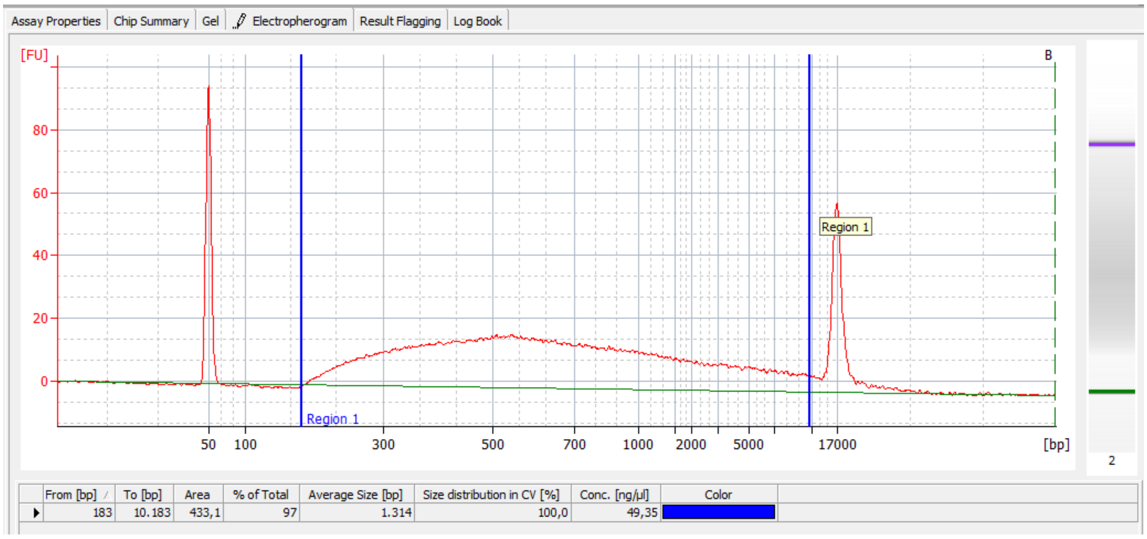
The libraries' concentration has been quantified using the Qubit® 2.0 fluorimeter with the dsDNA HS Assay Kit while the libraries' average sizes have been analyzed on the Bioanalyzer using the Agilent DNA 12000 kit.

Following, are reported the libraries' concentration and profiles (Figure 39, Figure 40, Figure 41, Figure 42).



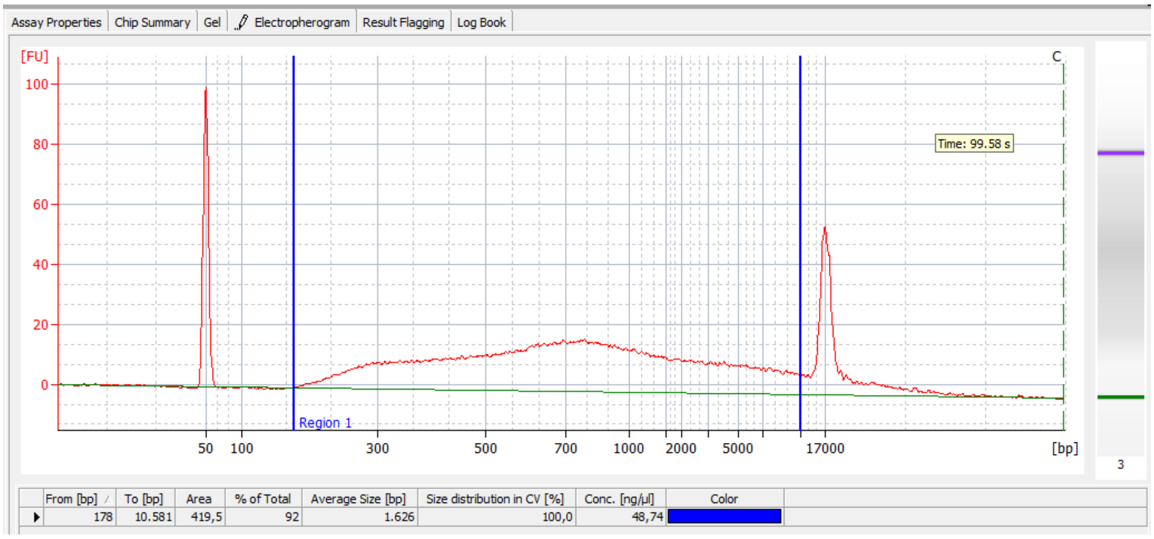
	From [bp]	To [bp]	Area	% of Total	Average Size [bp]	Size distribution in CV [%]	Conc. [ng/µl]	Color
▶	179	5.970	489,9	97	677	100,0	54,10	🟦

Figure 39 – Library A (50 ng of gDNA)



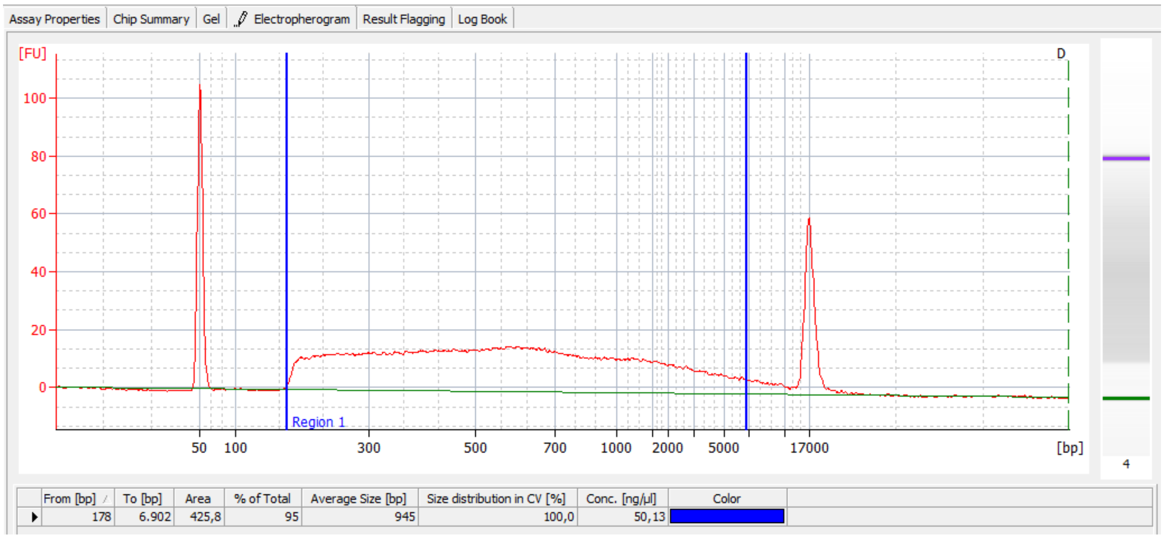
	From [bp]	To [bp]	Area	% of Total	Average Size [bp]	Size distribution in CV [%]	Conc. [ng/µl]	Color
▶	183	10.183	433,1	97	1.314	100,0	49,35	🟦

Figure 40 - Library B (100 ng of gDNA)



	From [bp]	To [bp]	Area	% of Total	Average Size [bp]	Size distribution in CV [%]	Conc. [ng/µl]	Color
▶	178	10.581	419,5	92	1.626	100,0	48,74	🟦

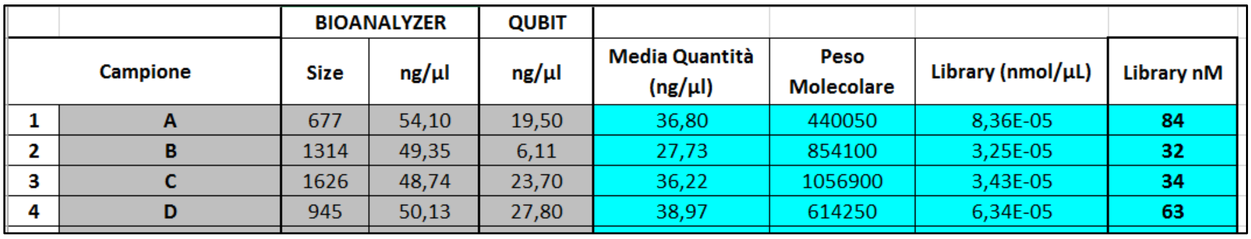
Figure 41 - Library C (1,6 µl of TDE1)



	From [bp]	To [bp]	Area	% of Total	Average Size [bp]	Size distribution in CV [%]	Conc. [ng/µl]	Color
▶	178	6.902	425,8	95	945	100,0	50,13	🟦

Figure 42 - Library D (2,4 µl of TDE1)

Table 11 - Summary of obtained results during method's robustness verification



		BIOANALYZER		QUBIT				
Campione		Size	ng/µl	ng/µl	Media Quantità (ng/µl)	Peso Molecolare	Library (nmol/µL)	Library nM
1	A	677	54,10	19,50	36,80	440050	8,36E-05	**84**
2	B	1314	49,35	6,11	27,73	854100	3,25E-05	**32**
3	C	1626	48,74	23,70	36,22	1056900	3,43E-05	**34**
4	D	945	50,13	27,80	38,97	614250	6,34E-05	**63**

The final assessment of the libraries produced was compliant to the defined acceptance criteria:
Average library size greater than or equal to 200 bp;
Library concentration greater than or equal to 2 nM.

Table 11 summarizes all the results achieved during the method Robustness tests combining Qubit and Bioanalyzer parameters. These results comply with the acceptance criteria and demonstrate that, despite deliberate changes introduced at critical steps, the method is robust within the ranges defined during the validation tests.

Additionally, based on the tests carried out during the development phase and based on the Risk Assessment it was confirmed that both the following two genomic DNA extraction kits can be used:

- YeaStar™ Genomic DNA Kit (Zymoresearch);
- YEAST Extraction DNA kit (Thermo Fisher Scientific).

Conclusion and Discussion

The drug quality assessment is crucial for patient safety and essential from a business point of view. The quality control must be performed, following strict guidelines, on the entire manufacturing process to assure the drug safety, before the drug introduction on the market.

The cell bank Identity is a test panel used to guarantee that, during the entire production process, the cell banks maintain intact the typical inherent characteristics of the original cell line, such as morphology and vitality of belonging strain and typical biochemical markers. Absence of any issues during the production is checked by comparing the cell bank before building the bioreactor (Master Cell Bank, MCB) and the cell bank before stopping the bioreactor production (Post Production Cell Bank, PPCB). It can be inferred that no modifications/cross-contamination on the process happened if no difference between MCB and PPCB are observed.

Testing activities must be carried out in GMP compliance[5] and respect acceptance criteria established during the validation of the corresponding methods.

The method currently used for the genus, species and strain identification of microbial cell banks is based on a combination of Sanger sequencing and RAPD techniques. When the methods were validated, both Sanger sequencing and RAPD were considered by Health Authorities the "gold standard". However, Next Generation Sequencing (NGS) technique is growing as a more innovative, powerful and straightforward technology that could replace old methods. NGS could increase throughput while reducing drastically the time spent per run. Thus, NGS easily allows the detection and identification of microorganisms using a culture-independent strategy[37].

Consecutively a new method, based on NGS technology, has been developed and subsequently validated according to GMP requirements, aimed to confirm the identity of recombinant yeast cell banks.

In accordance to ICH Q2 (R1) guideline, the method has been classified as an Identification Test but also has some potential as a Limit test for impurities, therefore the parameters assessed during validation were Specificity, Limit of Detection (LOD) and Robustness.

Following validation, the method was found to be Specific and Robust, with a 10% Limit of Detection of any contaminant both for genus, species and for strain identification.

The defined limit of detection understood as the smallest percentage of the contaminant microbial strain detectable within the microbial strain being analyzed, is 10% compared to the total amount of sample to be analyzed. Therefore, the detection limit is greatly improved when compared to the result previously obtained during RAPD method validation, by 25%.

This result, together with the increased sequencing throughput and reduced time spent to conduct each test represents a huge improvement in the quality control routine test. Additionally, the NGS technology allows a variety of applications in both DNA and RNA sequencing, which could be, in the future, a useful tool for Biopharma Quality Control activities.

Concluding, the validation test results show that the "Identification of the genus, species, and strain of Yeast Cell Banks by Next Generation Sequencing" method has to be considered validated and compliant to GMP requirements.

Bibliography

1. ICH - Q5D. Requirements for registration of Pharmaceuticals for human use used for production of Biotechnological / Biological products. (1997).
2. Morrow, T. & Felcone, L. H. Defining the difference: What Makes Biologics Unique. *Biotechnol. Healthc.* 1, 24–9 (2004).
3. FDA Presentation on Biologics. (2013).
4. FDA – Overview of Biological Products. Available at: www.fda.gov.
5. EudraLex - EU Legislation. *Guidelines for Good Manufacturing Practice for Medicinal Products for Human and Veterinary Use*. 4, (2015).
6. FDA. Current Good Manufacturing Practices For Finished Pharmaceuticals. (2011).
7. ICH - Q2(R1). Validation of Analytical Procedures: Text and Methodology. 1994, (2005).
8. Theron, C.W., Berrios, J., Delvigne, F. et al. A. M. B. (2018) 102: 63. https://doi. org/10. 1007/s0025.-017-8612-y. Integrating metabolic modeling and population heterogeneity analysis into optimizing recombinant protein production by Komagataella (Pichia) pastoris. *Appl. Microbiol. Biotechnol.* 102, 63–80
9. Wegner, E. H. Biochemical conversions by yeast fermentation at high cell densities - United States Patent 44114329. (1983).
10. Cregg J.M., Vedvick T.S., R. W. C. Recent advances in the expression of foreign genes in Pichia pastoris. *Biotechnol. (N Y). 1993 Aug;11(8)905-10.*
11. Kurtzman, C. P. Biotechnological strains of Komagataella (Pichia) pastoris are Komagataella pha Y i as determined from multigene. 1435–1438 (2009). doi:10.1007/s10295-009-0638-4
12. Cregg JM, M. K. Development of the methylotrophic yeast, Pichia pastoris, as a host system for the production of foreign proteins. *Dev Ind Microbiol 29* 33–41 (1988).
13. Cereghino, G. P. L., Cereghino, J. L., Ilgen, C. & Cregg, J. M. Production of recombinant proteins in fermenter cultures of the yeast. *Protein Eng.* 329–332 (2002). doi:10.1016/S0958166902003300
14. Gargis, A. S., Kalman, L. & Lubin, I. M. Assuring the quality of next-generation sequencing in clinical microbiology and public health laboratories. *J. Clin. Microbiol.* 54, 2857–2865 (2016).
15. Hrabak, J., Bitar, I. & Papagiannitsis, C. C. Combination of mass spectrometry and DNA sequencing for detection of antibiotic resistance in diagnostic laboratories. *Folia Microbiol. (Praha).* (2019). doi:10.1007/s12223-019-00757-5
16. Buermans, H. P. J. & den Dunnen, J. T. Next generation sequencing technology: Advances and applications. *Biochim. Biophys. Acta - Mol. Basis Dis.* (2014). doi:10.1016/j.bbadis.2014.06.015
17. Moran-Gilad, J. *et al.* Proficiency testing for bacterial whole genome sequencing: An end-user survey of current capabilities, requirements and priorities. *BMC Infect. Dis.* 15, 1–10 (2015).
18. LaDuca, H. *et al.* Exome sequencing covers >98% of mutations identified on targeted next generation sequencing panels. *PLoS One* 12, 1–11 (2017).
19. Hölzer, M. & Marz, M. De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. *Gigascience* 8, 1–16 (2019).
20. NCBI. Available at: http://www.ncbi.nlm.nih.gov/.
21. ICH - Q6B. Specification: Test procedures and Acceptance Criteria for Biotechnological / Biological products. (1999).
22. Mendes, M. E., Romano, P. & Sumita, N. M. Practical aspects of the use of FMEA tool in clinical laboratory risk management. 174–181 (2013).
23. Serafini, A., Troiano, G., Franceschini, E., Calzoni, P. & Nante, N. Use of a systematic risk analysis method (FMECA) to improve quality in a clinical laboratory procedure. 288–295 (2016). doi:10.7416/ai.2016.2108
24. Nextera® DNA Library Prep Reference Guide. Available at: www.illumina.com.

25. Beckma Coulter. Agencourt AMPure XP Protocol, PN B37419AA. (2013).
26. Illumina. Available at: www.illumina.com.
27. FDA. Data Integrity and Compliance With Drug CGMP Questions and Answers Guidance for Industry Pharmaceutical Quality/Manufacturing Standards (CGMP) Data Integrity and Compliance With Drug CGMP Questions and Answers Guidance for Industry. (2018).
28. Abyss. Available at: http://www.bcgsc.ca/downloads/abyss/.
29. SOAPdenovo. Available at: http://soap.genomics.org.cn/soapdenovo.html.
30. Velvet. Available at: http://www.ebi.ac.uk/~zerbino/velvet.
31. ALLPATHS-LG. Available at: ftp://ftp.broadinstitute.org/pub/crd/ALLPATHS/Release-LG/.
32. Celera Assembler. Available at: https://sourceforge.net/projects/celeragb/.
33. Trinity. Available at: https://github.com/trinityrnaseq/trinityrnaseq/wiki.
34. Strasbourg: Council of Europe. European Pharmacopoeia. (2019).
35. ICH. Quality risk management Q9 guideline. 1–20 (2005).
36. FDA. *Analytical Procedures and Methods Validation for Drugs and Biologics Guidance for Industry, U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Res. Pharmaceutical Quality* (2015).
37. Boers, S. A., Jansen, R. & Hays, J. P. Understanding and overcoming the pitfalls and biases of next-generation sequencing (NGS) methods for use in the routine clinical microbiological diagnostic laboratory. *Eur. J. Clin. Microbiol. Infect. Dis.* 38, 1059–1070 (2019).

Index of Figure

Index of Tables

Acronyms

ALCOA: Attributable, Legible, Contemporaneously recorded, Original and Accurate
AOX1: Alcohol Oxidase I
ATCC: American Type Culture Collection
bp: base pair
BQC: Biological Quality Control
BR: Broad Range
CCD: Charged Coupled Device
D2 LSU: D1-D2 region of the ribosomal DNA Large Subunit
DBVPG: Vegetal Biology Department of Perugia
DNA: Deoxyribonucleic Acid
dsDNA: double-strand DNA
EMA: European Medicines Agency
EMEA: European Medicinal Evaluation Agency
EoPCB: End of Production Cell Bank
EtOH: Ethyl Alcohol (Ethanol)
ExCB: Extended Cell Bank
FDA: Food and Drug Administration
FMECA: Failure Modes, Effects, and Criticality Analysis
Gb: Gigabase
gDNA: genomic DNA
GMP: Good Manufacturing Practices
HA: Health Authorities
HS: High Sensitivity
HT1: Hybridization Buffer
ICH: International Conference on Harmonization
IEM: Illumina Experiment Manager
IPC: In Process Control
LOD: Limit of Detection
MCB: Master Cell Bank
NCBI: National Center for Biotechnology Information
NGS: Next Generation Sequencing
PPCB: Post Production Cell Bank
RAPD: Random Amplification of Polymorphic DNA
RCB: Recombinant Cell Bank
rDNA: ribosomal DNA
SBS: Sequencing by Synthesis
SNPs: Single Nucleotide Polymorphisms
TD: Tagment DNA Buffer
TDE1: Tagment DNA Enzyme 1
WCB: Working Cell Bank
WHO: World Health Organization
YPD: Yeast extract Peptone Dextrose