FUENLABRADA SCHOOL OF ENGINEERING

BIOMEDICAL ENGINEERING DEGREE

END OF DEGREE PROJECT

APPLICATIONS OF NEXT-GENERATION SEQUENCING TO AML DIAGNOSIS

Author: Génesis Belén Chérrez Amaya
Tutor: José Felipe Ortega Soto
Co-tutor: José Vicente Die

Academic Course 2023/2024

“Nothing in life is to be feared, it is only to be understood.
Now is the time to understand more, so that we fear less.”
Marie Curie.

Acknowledgements

To my parents because without them none of this would have happened.
To my sister who stood next to me during the entire process.
To my teachers which provide me with the accurate knowledge and background to start this project.
To my mentor Felipe, who gave me the advice and continue teaching me patiently and seeking for the best of this project we made.
To my tutor Luis at CNIO for providing me the instructions necessary to carry out the biological aspect of this study.
To my friends along the journey, it was nice having you by my side.
To my roles models who I look up to every day.
To the hard work, resilience and patience dedicated to this investigation.

And last but not least, to what’s next to come.

Abstract

Acute Myeloid Leukaemia (AML) is an aggressive cancer of the monocyte cells that grows rapidly and, therefore, normally requires immediate treatment. Advances associated with Next-Generation Sequencing (NGS) and the decreasing cost have paved the way for more efficient standard molecular profiling procedures applicable to solid tumours and hematologic malignancies. One of the most important advantages of employing NGS and RNA-Seq is the unprecedented ability to provide comprehensive and detailed cancer genetic profiling.

This project conveys key concepts regarding NGA and RNA sequencing (RNA-Seq) as the groundwork for understanding their revolutionary impact on cancer diagnosis, prognosis, and treatment. The resulting output has the desired characteristics: high-throughput, specific to the analysis of the genetic materials, allowing the detection of known and new genetic mutations. This all impacts the fact that it is crucial for accurately diagnosing AML, coordinating treatment decisions, and developing targeted therapy that targets the individual genetic makeup of a patient's tumour.

We first introduce crucial concepts and technologies related to NGS and RNA-seq for understanding the requisites, workflow and purposes from the biological point of view provided by the internship working side by side with the laboratory clinicians at the Oncologic National Centre and the partnership with Roche Hyper Design Tool sequence panelling which offers the experimental results we were seeking for and serve as the technical aspect of our study. Several algorithms were computationally tested in the Python and R programming languages to customise this work. Along this way, the teams at Roche and CNIO transmitted fundamental concepts, procedures, and best practices to accomplish these tasks.

Due to the lack of previous similar work, we found it necessary to add a bioinformatic context and deepen our knowledge of this subject by introducing salient concepts and procedures (defined in the glossary) for accurate follow-up and understanding of the study. Besides, we explore the applications of pairwise sequence in NGS theoretically, with particular consideration for aligning biological sequences (DNA, RNA or protein sequences) to find out that, indeed, identifying somatic and germline genetic variants along hereditary human genetic tendencies in AML displays significative predictive, diagnostic and prognosis advanced which can revolutionise our medical knowledge and clinical practice, in other words, push the boundaries of personalised medicine.

Results from this investigation underscore the efficiency of NGS and RNA-seq in ensuring diagnostic AML accuracy. Applying our self-created sequence panelling tool for AML biomarkers, following the inestimable guidance from the Roche team, it is safe to say that we can successfully draw findings on critical genetic variations associated with AML, outlining the potential of the practices of NGS technologies as a substitute to the standard typical technologies use mostly nowadays. Notably, the research also highlighted the selection of target biomarkers related exclusively to AML treatment and improvement to understand the disease in the patients who suffer from it.

To conclude, this project recaps findings that support advocating for the integration and implementation of NGS and RNA-Seq in clinical practice. Such an integration can play an essential role in AML diagnosis and treatment, leading to a shift towards more personalised, precise, and effective cancer diagnosis, treatment and care. Ultimately, it promises better outcomes, improved life expectancy and better quality of life for patients.

Resumen

La leucemia mieloide aguda (LMA) es un cáncer agresivo de las células monocíticas que crece rápidamente y, por lo tanto, normalmente requiere tratamiento inmediato. Los avances asociados con la secuenciación de próxima generación (NGS) y la disminución de los costos han allanado el camino para procedimientos de perfil molecular estándar más eficientes, aplicables a tumores sólidos y malignidades hematológicas. Una de las ventajas más importantes de emplear NGS y RNA-Seq es la capacidad sin precedentes de proporcionar un perfil genético del cáncer comprensivo y detallado.

Este proyecto transmite conceptos clave sobre la secuenciación de próxima generación (NGA) y la secuenciación de ARN (RNA-Seq) como base para comprender su impacto revolucionario en el diagnóstico, pronóstico y tratamiento del cáncer. El resultado obtenido tiene las características deseadas: alto rendimiento, específico para el análisis de materiales genéticos, permitiendo la detección de mutaciones genéticas conocidas y nuevas. Todo esto impacta en el hecho que es crucial para diagnosticar con precisión la LMA, coordinar las decisiones de tratamiento y desarrollar terapias dirigidas que apunten a la composición genética individual del tumor de un paciente.

Primero, introducimos conceptos y tecnologías cruciales relacionados con NGS y RNA-seq para comprender los requisitos, el flujo de trabajo y los propósitos desde el punto de vista biológico proporcionado por la pasantía, trabajando junto a los clínicos de laboratorio en el Centro Nacional Oncológico y la asociación con Roche Hyper Design Tool sequence panelling, que ofrece los resultados experimentales que buscábamos y sirve como el aspecto técnico de nuestro estudio. Se probaron computacionalmente varios algoritmos en los lenguajes de programación Python y R para personalizar este trabajo. A lo largo de este proceso, los equipos de Roche y CNIO transmitieron conceptos fundamentales, procedimientos y mejores prácticas para llevar a cabo estas tareas.

Debido a la falta de trabajos similares previos, encontramos necesario agregar un contexto bioinformático y profundizar nuestro conocimiento en este tema introduciendo conceptos y procedimientos destacados (definidos en el glosario) para un seguimiento y comprensión precisos del estudio .Además , exploramos las aplicaciones de la secuencia por pares en NGS teóricamente , con particular consideración para alinear secuencias biológicas (secuencias de ADN , ARN o proteínas) para descubrir que , efectivamente, la identificación de variantes genéticas somáticas y de línea germinal a lo largo de tendencias genéticas humanas hereditarias en la LMA muestra avances significativos en predicción , diagnóstico y pronóstico que pueden revolucionar nuestro conocimiento médico y práctica clínica , en otras palabras, empujar los límites de la medicina personalizada .

Los resultados de esta investigación subrayan la eficiencia de NGS y RNA-seq para asegurar la precisión del diagnóstico de LMA. Aplicando nuestra herramienta de panelización de secuencias creada para biomarcadores de LMA, siguiendo la guía invaluable del equipo de Roche, es seguro decir que podemos extraer hallazgos exitosamente sobre variaciones genéticas críticas asociadas con LMA , delineando el potencial de las prácticas de tecnologías NGS como un

sustituto de las tecnologías típicas estándar utilizadas mayormente hoy en día .Notablemente, la investigación también destacó la selección de biomarcadores objetivo relacionados exclusivamente con el tratamiento de LMA y la mejora en la comprensión de la enfermedad en los pacientes que la padecen.

Para concluir, este proyecto resume hallazgo que apoyan la defensa de la integración e implementación de NGS y RNA-Seq en la práctica clínica. Tal integración puede jugar un papel esencial en el diagnóstico y tratamiento y cuidado del cáncer más personalizado, preciso y efectivo. En última instancia, promete mejores resultados, mayor esperanza de vida y mejor calidad de vida para los pacientes.

Contents

List of Figures

List of tables

List of Acronyms

AML	Acute Myeloid Leukaemia
AI	Artificial Intelligence
BLAST	Basic Local Alignment Tool
BLOSUM	Block Substitution Matrix
CEBPA	CCAAT Enhancer Binding Protein Alpha
cDNA	Complementary DNA
DNA	Deoxyribonucleic Acid
EDP	End of Degree Project
FLT3	FMS-like tyrosine kinase 3
GO	Gene Ontology
KEGG	Kyoto Encyclopaedia of Genes and Genomes
M5C	5-Methylcytosine
M6A	N6-Methyladenosine
ML	Machine Learning
MRD	Minimal Residual Disease
NCBI	National Centre for biotechnology Information
NGS	Next Generation Sequencing
NLM	National Library of Medicine
NPM1	Nucleophosmin1
NGS	Next-Generation Sequencing
PAM	Point Accepted Mutation
PCR	Polymerase Chain Reaction
PML-RARA	Promyelocityc Leukaemia-Retinoic Acid Receptor
qOCR	Quantitively Polymerase Chain Reaction
RUNX1	Ribonucleic Acid

SN	Single Nucleotide Polymorphism
TP53	Tumour Protein p53
VAST	Vector Alignment Search Tool
WES	Whole Exome Sequencing
WGS	Whole Genomic Sequencing

1. Introduction and Goals

1.1 Context and motivation

In this chapter, we dive into the fascinating world of genetic sequencing, untangling the difficulties of DNA, RNA, and proteins. This examination is not simply theoretical; it denotes a pursuit to comprehend the building blocks of life. Our journey is a reflective introspection at decades of scientific progression, tracing the advancement of sequencing technologies from the initiation to this day and age.

We start off with a synopsis of conventional sequencing approaches, emphasizing the ground- breaking labour of Frederick Sanger. His technique, though world-shuttering in its time, encountered restrictions in scalability and efficacy, prompting the necessity for novelty. This historical background assesses the stage for the establishment of Next -Generation Sequencing (NGS), a technology that has remodel the scenery of genomic investigation. NGS counters as a model of contemporary scientific accomplishment, proposing the competence to sequence millions of fragments alongside, massively outpacing the abilities of previous systems.

The narrative then changes to bioinformatics, an academy area where biology and computer science collides. In this area, databases such as the National Centre for Biotechnology Information (NCBI) are not plain static repositories but dynamical resources that operate as the spine for genetic research, enabling access to a majority collection of genomic sequences, genetic variations, and common scientific literature.

We dispose precise prominence on sequence alignments, a procedure at the core of our genetic scrutiny. This technique entails aligning sequenced genetic snippets against reference genomes, a vital phase in unravelling genetic data. Tools like Basic Local Alignment Search (BLAST) are drawn special attention to their part in recognise homologous sequence, sustaining in the study of genetic structures and functions.

The usage of these means is embodied in the study of illnesses such as Acute Myeloid Leukaemia (AML) where genetic markers and biomarkers are identifies, giving meaningful insights about the disease mechanism and probable treatment pathways.

This chapter is beyond an overview, it is a roadmap. It delineates our goals and objectives, enclosing the succeeding chapters and the global track of our research. Our aim is to bridge the fissure between conventional genetic knowledge and the borderlines merges because of NGS. By doing so, we expect to impact significantly to the genomic area, solving crucial biological interrogations and paving the way for medical improvements that profit human race.

1.2 Goals

This project aims to contribute towards developing NGS techniques superior to those based on the RNA-seq, which are currently most common for accurate diagnosis and treatment of AML. Part of this work has been undertaken during a 3-month internship at CNIO, learning and testing different alternative techniques. Eventually, the main objective is finding a viable approach, including gene panel sequencing procedures and identifying the respective biomarkers and associated data preprocessing tasks described in this document.

The main objectives of this project were clearly established at the beginning of the research internship at CNIO, and they can be broken down into three distinct milestones to ensure an effective and systematic approach:

- **Objective 1: Evaluate the efficiency of next-Generation Sequencing (NGS)compared to RNA-seq in AML diagnosis.** This objective focused on detailed comparison between NGS and RNA-seq to identify which technology offers superior accuracy and comprehensiveness in diagnosing AML. NGS is hypothesized to provide better results due to its advanced sequencing capabilities and cost-effectiveness. By conducting experimental studies and data analysis, the aim is to validate these claims. The comparison will look at sequence coverage, precision in biomarker identification, and overall cost benefits. The findings are expected to support the adoption of NGS as the preferred method for clinical genomic analysis in AML.
- **Objective 2: Development of an integrated methodology for identifying AML biomarkers using NCBI Database and bioinformatics Tools.** This objective involves developing a robust methodology that leverages NCBI databases and bioinformatics tools, such as Biopython, to identify and analyse specific biomarkers of AMNL. The integration of these tools is expected to enhance the efficiency and accuracy of analysing large volumes of genetic data. This methodology will be tested for its ability to reliably identify critical biomarkers that are essential for diagnose and treatment AML. By establishing a standard protocol, the goal is to provide clear guidance for future researchers and clinicians. This will ultimately improve the accuracy and effectiveness of AML biomarker identification.

- **Objective 3: Analysis of sequencing coverage and performance of sequence design tools in detecting AML Diagnostic Biomarkers.** The objective here is to assess the sequencing coverage and performance of specific tools , like the Hyper Design Roche Tool , in identifying AML diagnostic biomarkers .Evaluating these tools is crucial for understanding their effectiveness in sequencing and data analysis , particularly in clinical diagnostics .The study will involve using these tools to design specific sequencing panels and then evaluating the coverage and precision of the resulting data .This analysis will highlight the strengths and weaknesses of current tools , providing insights for optimization .

1.3 Project scheduling

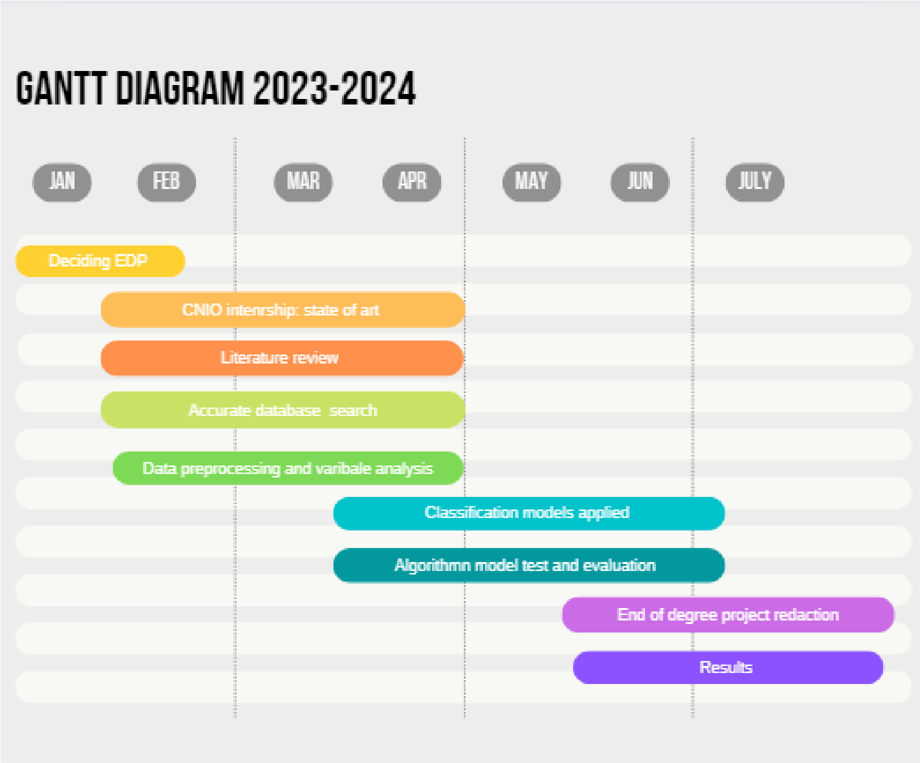


Figure 1.1 Gantt diagram for this EDP. The horizontal bars show the estimated time. Source: own work

1.4 Document structure

The rest of this document is structured as follows:

- In Chapter 1 we have provided a general introduction to this project, including its contextual framework and main objectives.

- During Chapter2 we explore sequencing techniques evolution and applications of bioinformatic beside the role of the databases in genetic research focusing specially on sequence alignment techniques for AML study.

- In Chapter 3 we focus on NCBI databases highlighting the process of using the resources and methods for the analysis and interpretation of genetic information relevant to AML diagnosis and research.

- The Chapter 4 evaluates the performance of sequencing techniques efficiency in improving AML diagnosis through NGS data analysis. For that, we guide the project by a confusion matrix were metrics like precision, recall and accuracy are taking into account. Besides, a quantitively approach is used for the determination of precise accuracy following and overall systematic methodology.

- In Chapter 5 we summarise findings and insights gained throughout the whole project as well as the lessons learned and future works.

2. Background and related work

This section of basic concepts introduces key elements for fully understand the complexities of genetic analysis. It explains also concepts related to genomics, transcriptomics and proteomics form a functional, comparative, and structural aspects.

2.1 Next-Generation Sequencing

Nowadays a new technique emerged which transform the assemblage if genomic data: the Next Generation Sequencing (NGS). The primary dissimilarity with the standard classic technique relies on the fact that it allows the customer massively parallel sequencing, which much less cost and depends on a sequencing library instead of a chain-termination sequencing.

NGS is applied in various usages such as sequencing the genome of an organism and researching of the genome diversity, relying in the whole genome sequencing [1], gene-expression profiling by RNA sequencing (RNA-seq) which we will explain further on the following sections.
It is also used for detecting variants in protein-coding region by sequencing the exome, studying specific panels of genes by targeted gene sequencing identification of bacteria recovered from environmental samples using metagenomic sequencing, and studying genetic modifications.
NGS initializes first a breakdown of the RNA into tiny subparts, normally between 50 and 300 base pairs in length. These fragments are unified to adapters and amplified through several methods to allow sequencing. The done library of subparts is then loaded into a sequencer, then each fragment is read in parallel.

One possible usage of this procedure enables the generation of enormous datasets that can be inspected for research purposes, in our case, the early detection of AML [2]. NGS is a mean to sequence DNA or RNA, meanwhile the whole genome sequencing technique (WGS) is utilized for sequencing the entire genome. Its usefulness is the fact that it can identify genetic variation such as insertions and deletions.

To sequence the protein-coding regions of the genome, known as the exome, which constitutes about 1-2 % of the genome, the Whole Exome Sequencing (WES)techniques used. This technique is essential for identifying genetic variations that might contribute to diseases like AML (Acute Myeloid Leukaemia)

2.2 NGS and RNA-seq differences

For clarification purposes the following table is displayed emphasizing the main differences between the two procedures we are going to discuss about in the whole project: NGS and RNA-Seq providing therefore a solid base for the next chapters understanding and follow up[1].

Table 2.1:An overview of the main differences between NGS and RNA sequencing techniques.

	NGS	RNA-Seq
Definition	Encompasses a range of high throughput sequencing technologies utilized across diverse genomic applications such as DNA and RNA sequencing	A specialized application within NGS specifically designed for analysing and quantifying RNA molecules within a biological sample
Target Molecules	Applicable for sequencing DNA, RNA, or other nucleic acids	Focus on sequencing RNA molecules, offering insights into gene expression, alternative splicing, and transcriptome dynamics.
Purpose	Whole genome sequencing, targeted sequencing, epigenetic analysis, and metagenomics	Primarily employed for analysing and quantifying RNA transcripts. Enables the study of gene expression, identification of novel transcripts, and understanding alternative splicing patterns
Information obtained/output	Information about the sequence and organization of DNA or other nucleic acids. Data in the form of DNA sequence which can be aligned to a reference genome	Provides insights into the identity, quantity, and structure of RNA molecules, encompassing mRNA, non-coding RNA, and splice variants. Generates data detailing the abundance and structure of RNA transcripts facilitating the reconstructing of the transcriptome
Application	Genomic, epigenomics, metagenomics	Widely utilized for gene expression profiling, analysis of differential gene expression, discovery of novel transcripts and investigation of RNA-related biological processes
Challenges	Related to data storage, computational demands, and handling large datasets	Present specific challenges like managing biases introduced during library preparation, precise unification of transcript abundance and identification of novel transcripts
Integration with other omics data	Integrates with other omics data, including proteomics and metabolomics, contributing to a holistic understanding	Integrated with other omics data to provide a comprehensive view of biological processes.

A more straightforward schema is shown in the Figure2.1 which consists of a comparative of the key aspects between both methods for the user to see clearly the steps involved in each process.

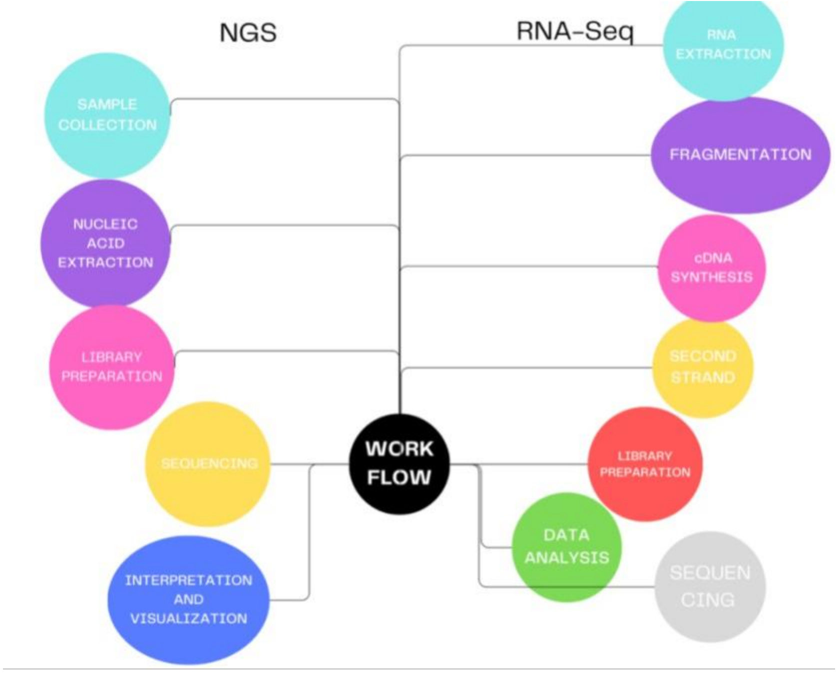


Figure 2.1: Comparison of NGS and RNA-seq workflows techniques. Source: own work.

NGS library preparation demands division of DNA and linking specific adaptor oligos to target sequenced pieces. The initial step is to fragment the DNA to the optimal length settled by the NGS sequencing technique. Considering such a fragmentation does not result in homogenous, blunt-ended fragments. Therefore, we need to check if every DNA fragment is free from overhangs and contains 5` phosphate and 3` hydroxyl groups through a process called end repair, illustrated in Figure 2.2.

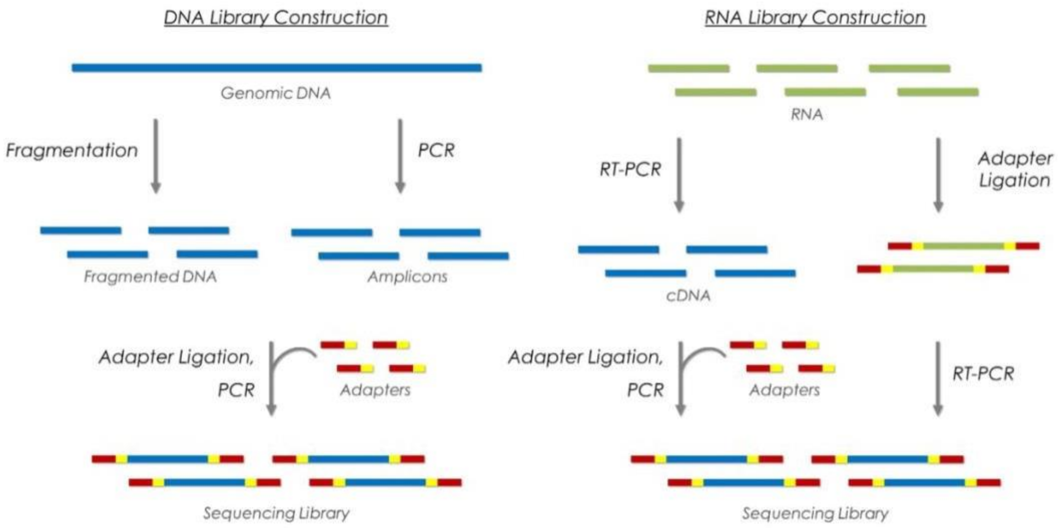


Figure 2.2: Differences between DNA and RNA library preparation in NGS. [3].

However, in RNA library preparation previous steps must be done, enrichment of mRNA by de3pleting ribosomal RNA and selection polyadenylated RNA, and the conversion in mRNA to cDNA.

- Enrichment of poly-A capture: Once the mRNA is transcribed from protein to coding genes, it undergoes several post-transcriptional adjustments to be ready for translation. The poly-A capture enrichment method involves strands, which contain oligo-dT (single-stranded sequence of deoxy thymine) on the superficial layer that can encapsulate mRNA through interactions with poly-A tails. The merge both is carried away to supress rRNA and elute the poly-A mRNA from the fibres. The resulting RNA is prepared then to undergoes the RNA library preparation.

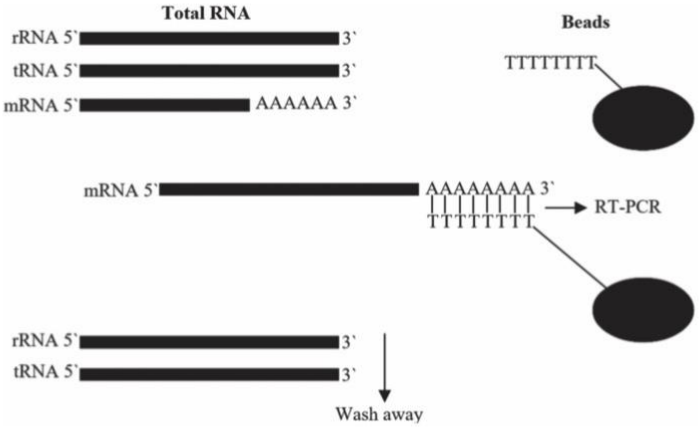


Figure 2.3: Enrichment with poly A – capture.[6].

2.3 RNA-seq workflow

1. **RNA Library Preparation:** the taken divided RNA is the extracted fragmented RNA is switched into a cDNA library that undertakes the following steps described below in Figure 2.4 where short-read sequencing(black), long-read cDNA sequencing(green) and long-read direct RNA-seq(blue) are compared when each is going to the library preparation methods concluding the bias and difficulty of said process changes accordingly to the approach we encounter.

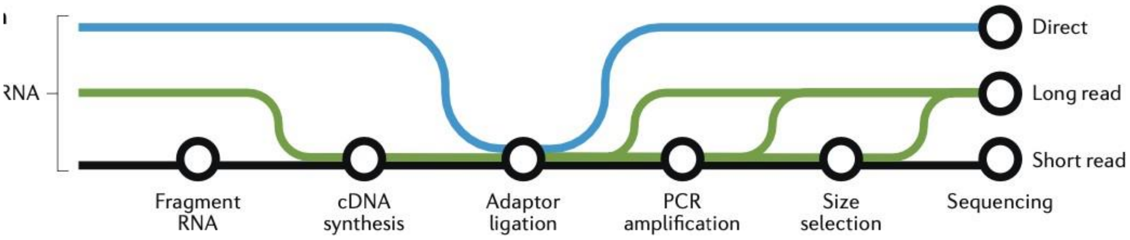


Figure 2.4: Short-read, long-read and direct RNA-seq technologies and workflows[5].

2. The short-read and long-read cDNA methods contains many steps in common during their protocols, however all methods demand an adaptor ligation stage, besides sample quality and computational issues affect through the overall preparation.

3. cDNA synthesis: Reverse transcriptase and arbitrary primers are used to transform the enriched

mRNA molecules into complementary copies of cDNA.

4. Adapter ligation: to allow sequencing on the sequence platforms, adapters are ligated to cDNA segments. Sequences in this adaptor enable the fragments to attach to the sequencing flow cell and are also used in multiplexed sequencing runs to identify the samples.

5. PCR amplification: To produce enough material for sequencing, adapter-ligated cDNA fragments are amplified using PCR in this process. Additionally, index sequences are added to the adapters by PCR amplification, enabling simultaneous sequencing of samples.

6. Size selection: long read and short read sequencing can be differentiated.

2.3.1 Short-read vs long-read cDNA sequencing

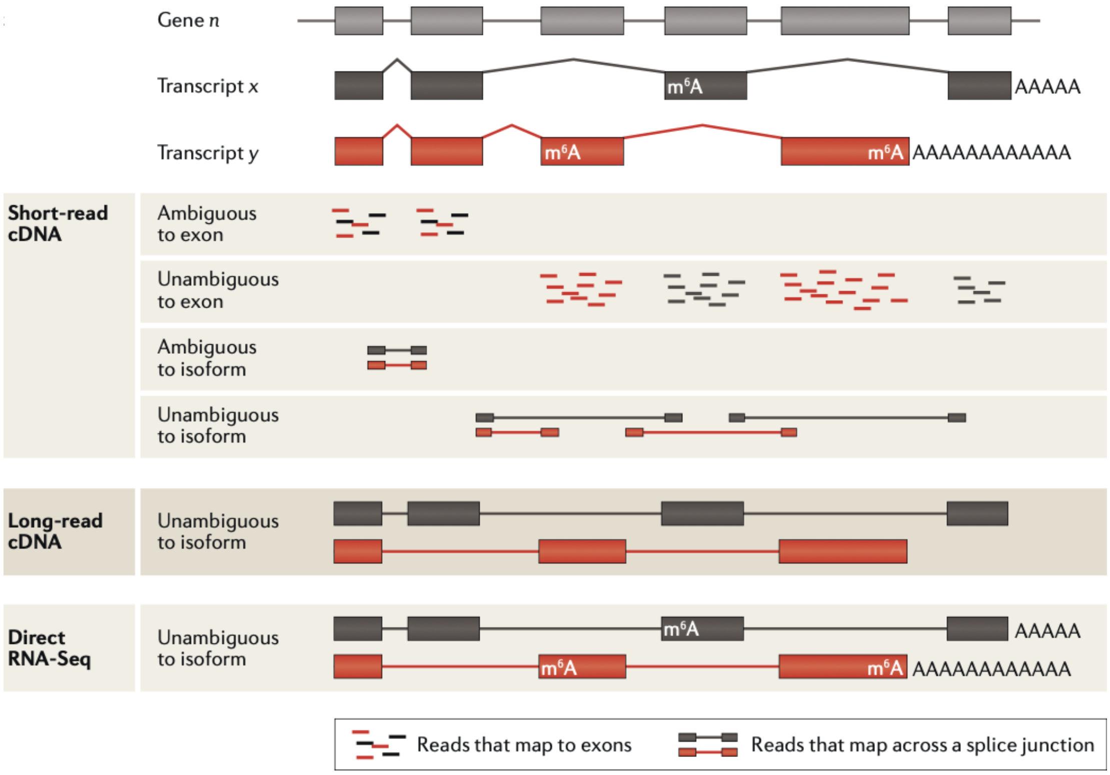


Figure 2.5: Comparison of short read, long-read and direct RNA-seq analysis[1].

Figure 2.5 illustrates how RNA is broken down into tiny fragments for short-read sequencing, also referred to as next generation sequencing, which is sequenced in parallel. Millions of fragments are simultaneously generated in high-throughput sequencing using this technology. However, the low read length makes it difficult to assemble complicated transcripts or map reads to reference genomes.

Due to their brief duration, short reads might not span the whole exon or intron during the alignment process leading to unclear alignment. The readings are then aligned to a cDNA reference sequence obtained from the annotated transcriptome to resolve this problem.

Once the reads are aligned to the cDNA reference, they can be used to infer isoform expression(mRNA isoforms present in a sample using cDNA sequencing data). This is particularly relevant for understanding the advantages of sequencing technologies in accurately mapping and quantifying isoforms. However, a single read may align to multiple isoforms, making it difficult to determine the exact isoform that the read originated from.

Because the short reads are so tiny, they might not cover the whole exon or intron during alignment, which could lead to unclear alignment. After that, a cDNA reference sequence obtained from the annotated transcriptome is aligned to the readings to resolve this problem. Because only the exonic sections and splicing junctions are present in the cDNA reference, this increases alignment accuracy.

These readings can be utilised to infer isoform expression once they have been aligned to the DNA reference. It could be challenging to identify the precise isoform from which a single read came, though, as it might align to several isoforms. To solve this problem, each read can be assigned to a distinct isoform according to its alignment pattern using computational methods that account for splicing junctions and exon-exon boundaries.

On the other hand, long-read sequencing, also known as third-generation sequencing, obtains longer pieces of RNA ranging from several hundred to tens of thousands of base pair. This approach provides complete transcripts, detects alternative splicing events, and identifies novel transcripts. However, long-read sequencing is more expensive and has higher error rates than short-read sequencing.

Long- read sequencing concerts RNA molecule into cDNA and sequence them to produce longer reads, typically thousands of base pairs long. This allows for complete coverage of exons or even multiple exons, resulting in clear targeting of the transcriptome. Unlike short-read sequencing, long-read sequencing does not require targeting to a reference genome or transcriptome, as longer reads can be targeted directly to the transcriptome.

Once reads are first aligned to a reference transcriptome or assembled again to create a transcriptome, these longer reads can span multiple exons during the alignment process, facilitating identification of the complete transcript including alternative splicing events. Compared to short-read sequencing, long-read sequencing can detect novel transcripts, alternative splicing events, and isoform diversity with higher accuracy.

After the measurements are aligned, one can use them to determine the abundance of each isoform. Each read can be assigned to a specific isoform based on its alignment pattern, allowing for more accurate quantification of isoform expression. Additionally, long-read sequencing can detect novel isoforms that may be missed by short -read sequencing. Long-read sequencing has revolutionized RNA sequencing, and its application is rapidly expanding in the field of transcriptomics, enabling more comprehensive understanding of complex gene expressions profiles.

The choice of sequencing method depends on the research question and characteristics of the RNA sample being sequenced. Library preparation, sequencing, and data analysis are important steps in both methods, with workflows depending on the sequencing platform and software used.

The analysis of gene expression and splicing events can be done in a variety of ways using RNA sequencing (RNA-Seq), although customary RNA-seq techniques have specific limits, like the powerlessness to straightforwardly recognize RNA changes and decide the course of record for non-polyadenylated RNA.

To address these impediments other techniques called RNA-seq Progressed (RASL-seq) were introduced. This strategy is a change of customary RNA -seq that empowers the immediate recognition of RNA adjustments and assurance of the course of record for non-polyadenylated RNA.

The RASL-seq convention involves ligation of 5'-phosphorylated adapter RNA to the 3`end of RNA fragments. The adapter has a degenerate sequence at the 5´end that allows ligation of adapters with RNA fragments with various structures the 3`end, including poly(A), poly(G), and non-polyadenylated RNA.

Connector bound RNA parts are opposite interpreted and enhanced utilizing a preliminary that ties to the connecter. Short-read sequencing technology is used to deuce the resulting cDNA, which reveals RNA modifications and the direction of non-polyadenylated RNA transcription.

For example, RASL-seq can identify RNA adjustments straightforwardly, for example, m6A and m5C, which are basic controllers of RNA handling and dependability. In contrast, to detect RNA modifications, conventional RNA -seq methods necessitate additional processing steps like immunoprecipitation[6].

In addition, it was discovered that RASL-seq can determine the direction of transcription for non-polyadenylated RNA, which is essential for understanding how eukaryotic cells regulate gene expression. Poly(A) tail selection, on the other hand, it can introduce bias toward polyadenylated RNA in conventional RNA-seq methods.

Nonetheless, there are certain restrictions to RASL-seq. For instance, the method only works with short-read sequencing technology, which can make it challenging to precisely assemble complex transcripts and requires high quality RNA sample[7].

In Figure 2.6 we can see that all means, including strengthening, determination, ligation, and elution, can be completed physically or on a tweaked Biomek FX robot. One of the focusing on oligos (the upstream one) contains a 5´phospate. As indicated, a particular universal primer is present in each oligo. After integrating standardized identifications during PCR, the items are pooled, filtered, a sequenced in Illumina sequencer (GA II or HiSeq 2000). The ligated region from the P5 primer is read in the first sequencing run, and the bar-coded region from the index primer is found in the second [6].

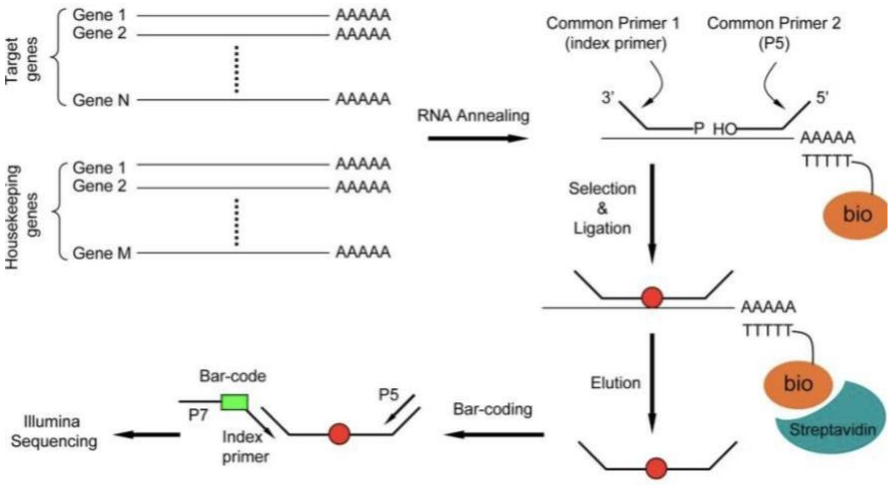
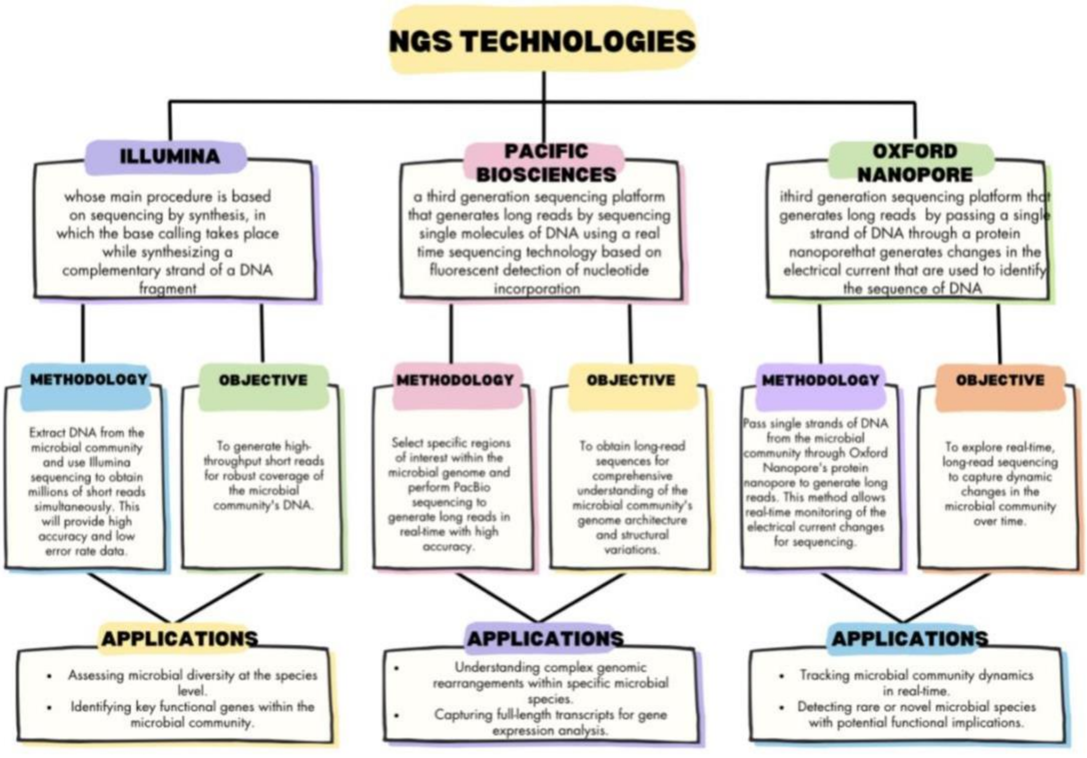


Figure 2.6: Overview of the RASL-seq technology[6].

2.3.2 NGS Technologies

Figure 2.7 summarises available NGS technologies from a general point of view regarding methodology, objectives, and application. Meanwhile, Table 2.2 is more specific directly linked to the operational and technical specifications.


NGS TECHNOLOGIES

ILLUMINA

whose main procedure is based on sequencing by synthesis, in which the base calling takes place while synthesizing a complementary strand of a DNA fragment

METHODOLOGY

Extract DNA from the microbial community and use Illumina sequencing to obtain millions of short reads simultaneously. This will provide high accuracy and low error rate data.

OBJECTIVE

To generate high-throughput short reads for robust coverage of the microbial community's DNA.

APPLICATIONS

- Assessing microbial diversity at the species level.
- Identifying key functional genes within the microbial community.

PACIFIC BIOSCIENCES

a third generation sequencing platform that generates long reads by sequencing single molecules of DNA using a real time sequencing technology based on fluorescent detection of nucleotide incorporation

METHODOLOGY

Select specific regions of interest within the microbial genome and perform PacBio sequencing to generate long reads in real-time with high accuracy.

OBJECTIVE

To obtain long-read sequences for comprehensive understanding of the microbial community's genome architecture and structural variations.

APPLICATIONS

- Understanding complex genomic rearrangements within specific microbial species.
- Capturing full-length transcripts for gene expression analysis.

OXFORD NANOPORE

ithird generation sequencing platform that generates long reads by passing a single strand of DNA through a protein nanoporethat generates changes in the electrical current that are used to identify the sequence of DNA

METHODOLOGY

Pass single strands of DNA from the microbial community through Oxford Nanopore's protein nanopore to generate long reads. This method allows real-time monitoring of the electrical current changes for sequencing.

OBJECTIVE

To explore real-time, long-read sequencing to capture dynamic changes in the microbial community over time.

APPLICATIONS

- Tracking microbial community dynamics in real-time.
- Detecting rare or novel microbial species with potential functional implications.

Figure 2.7: An overview is shown of the three main sequencing technologies for RNA-seq. Source: own work.

Table 2.2: An overview is shown of the three main sequencing technologies for RNA-SEQ [1].

	Illumina	**Pacific Biosciences**	**Oxford Nanopore**
OPERATIONAL SPECIFICATIONS			
Input DNA/RNA requirements	100 ng to several mg DNA or RNA	1-10 ug of DNA	1-2 ug of DNA or RNA
Sequencing output	Up to 3 billion reads per run	Up to 1 billion reads per run	Up to 30 Gb of data per run
Read length	Up to 300 nucleotides	Up to 100 kb	Up to 2Mb
TECHNICAL SPECIFICATIONS			
Turnaround time	Several days to a few weeks	Several days to a week	Several days
Accuracy	>99.9%	>99%	>90%
Error rate	<1%	<1%	10-15%
Cost per base	Low to moderate	Low to moderate	High

2.4 State of the art in RNA -Seq analysis

Read alignments include the arrangement of short DNA or RNA successions to a reference genome or transcriptome. In case that high-throughput sequencing output various short reads, in the field of genomics and transcriptomics, it is crucial to align them to a reference sequence for data analysis purposes. Since reading alignments outline meaningful data about orientation, precision, and location of those reads in the referenced, downstream analysis must be ensure like expression quantification, differential gene expression analysis and /or variant calling.

Expression quantification consists of checking abundance or level of gene expression in a sample. It is a process whose primary aim is to estimate quantity of RNA transcript and their corresponding expression levels. For this data related to RNA-seq is of crucial use in the aligning process. It enables medical professionals and research to compare functional and expression levels across groups and determine genes that are up or down regulated, all this thanks to valuable insights into the function of the genes under different conditions [8].

Differential gene expression analysis study and make a comparative among gene expressions levels of different samples groups to determine genes significant expression variation [8]. Its main goal is to remark genes whose expression variates significantly between two or more experimental circumstances. It takes part in the understanding of biological processes, setting potential biomarkers and detangling the genetic mechanisms underlying several phenotypes or diseases. Normally utilized statistics methods for differential gene expression analysis involving algorithms like edgeR or negative binomial test.

Variant calling associate genes with its biological functions the goal is to show a context regarding functionality and deep into the roles of genes following predicted functions criteria, protein domains, pathways, and interactions. Because of that it lies on databases and tools created specific for this process which contains valuable information from sources like the Gene Ontology (GO) Kyoto Encyclopaedia of Genes and Genomes (KEGG) [7].

Pathway analysis is a computational method that identifies biological pathways or molecular networks significantly enriched with differentially expressed genes or other genomic features. It facilitates understanding functional implications of gene expression changes at a systems level. Pathway analysis tools utilise curated databases of biological pathways and conduct statistical tests to identify pathways that are overrepresented or significantly influenced by the differentially expressed genes. This analysis assists in unravelling the underlying biological mechanisms, identifying key pathways involved in specific conditions or diseases, and generating hypotheses for further experimental validation.

2.5 Applications in AML diagnosis

Acute Myeloid Leukaemia (AML) is a blood cancer targeting myeloid cells, characterised by the uncontrolled growth and accumulation of anomalous white blood cells in the bone narrow [9]. Initial symptoms include fatigue and infection, due to shortage of normal blood cells.

AML development is believed to be triggered by a combination of genetic and environmental factors causing mutations in the DNA of bone marrow cells [2]. Such mutations may accelerate the growth and uncontrollable division of cells, leading to the creation of abnormal cells that, eventually, reside in the bloodstream and bone marrow.

Known risk factors for AML include:

- Age: the risk of AML increases with age, with most cases diagnosed in individuals over the age of 60.
- Gender: AML affects slightly more males than females.
- Ethnicity: There are certain differences in the incidence of AML among different ethnic's groups, with higher rates reported in individuals of African descent compared to those of European descent.
- Exposure to chemicals and radiation has been linked to an increased risk of AML.
- Family history: Individuals with a family history of AML or other blood disorders are at an increased risk for developing AML.
- Previous cancer treatments: Individuals who have received chemotherapy or radiation therapy for other cancers are at an increased risk for developing AML.
- Genetic disorders: Individuals with certain genetic disorders, such as Down syndrome, are at an increased risk for developing AML.

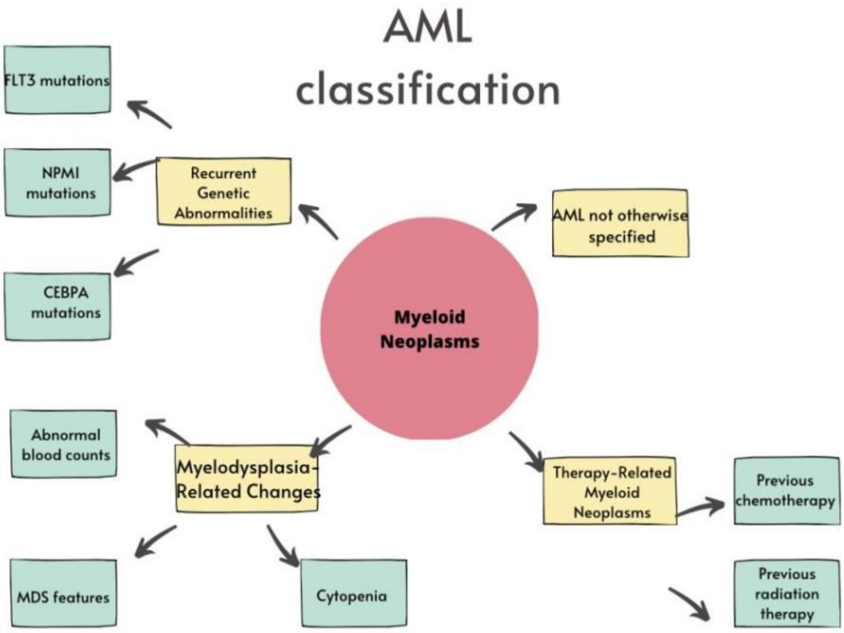

AML classification
FLT3 mutations
NPMI mutations
Recurrent Genetic Abnormalities
CEBPA mutations
AML not otherwise specified
Myeloid Neoplasms
Abnormal blood counts
Myelodysplasia-Related Changes
Therapy-Related Myeloid Neoplasms
Previous chemotherapy
MDS features
Cytopenia
Previous radiation therapy

Figure 2.8: Acute myeloid leukaemia classification. Source: own work

For the sake of contextualising the development of AML-related scientific progress Figure 2.9 presents a timeline encompassing important aspects such as treatment, prognosis and diagnosis evolution. It also emphasizes significant milestones, like the discovery of bone marrow-resident cells that can cause AML relapse and NGS advancements which have enhanced our understanding of AML over the years [2,9,7].

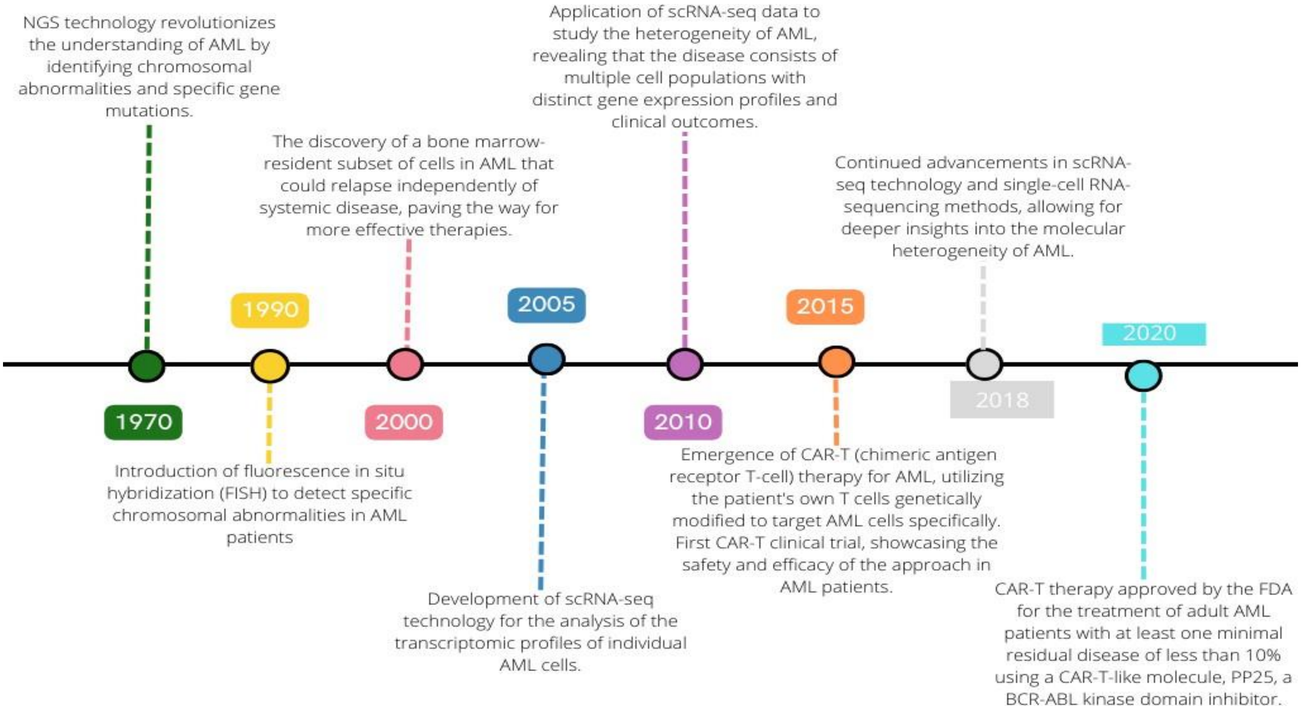

NGS technology revolutionizes the understanding of AML by identifying chromosomal abnormalities and specific gene mutations.
Application of scRNA-seq data to study the heterogeneity of AML, revealing that the disease consists of multiple cell populations with distinct gene expression profiles and clinical outcomes.
The discovery of a bone marrow-resident subset of cells in AML that could relapse independently of systemic disease, paving the way for more effective therapies.
Continued advancements in scRNA-seq technology and single-cell RNA-sequencing methods, allowing for deeper insights into the molecular heterogeneity of AML.
1990
2005
2015
2020
1970
2000
2010
2018
Introduction of fluorescence in situ hybridization (FISH) to detect specific chromosomal abnormalities in AML patients
Emergence of CAR-T (chimeric antigen receptor T-cell) therapy for AML, utilizing the patient's own T cells genetically modified to target AML cells specifically. First CAR-T clinical trial, showcasing the safety and efficacy of the approach in AML patients.
Development of scRNA-seq technology for the analysis of the transcriptomic profiles of individual AML cells.
CAR-T therapy approved by the FDA for the treatment of adult AML patients with at least one minimal residual disease of less than 10% using a CAR-T-like molecule, PP25, a BCR-ABL kinase domain inhibitor.

Figure 2.9: AML timeline. Source: own work

2.6 AML clinical praxis

Since AML is regarded as an aggressive form of leukaemia, the optimum course of treatment depends on an early diagnosis and prompt intervention. Chemotherapy is the standard treatment for acute myeloid leukaemia (AML), while it is occasionally combined with other medicines including radiation therapy or stem cell transplantation [9]. The course of treatment will be determined by several variables, including traits that biomarkers can identify.

To assist in the diagnosis, monitoring, and treatment of disease, biomarkers, biological molecules, or genetic alterations that may be tested in an individual´s blood, tissue, or other body fluids, are employed [10] AML biomarkers for frequently altered genes include the following ones.

FLT3

In the context of AML, FLT3 (FMS-like tyrosine kinase 3) is a biomarker linked to a more aggressive disease progression and poor prognosis [11]. Hematopoietic stem cells express a receptor protein that is encoded by the FLT3 gene, and this protein is essential for controlling the proliferation and survival of these cells.

FLT3 gene mutations are frequently seen in AML patients, and these mutations can produce a mutant protein that is perpetually active and encourages leukemic cell proliferation and survival. Internal tandem duplications (ITDs) and point mutations in the tyrosine kinase domain (TKD) are the two FLT3 mutations most frequently found in AML.

Certain biomarkers' techniques, including PCR and NGS, can identify the existence of FLT3 mutations. It is crucial to find FLT3 mutations in AML for varied reasons. Given that FLT3 mutated AML is linked to a more aggressive course of the disease and resistance to conventional chemotherapy, it first aids in the identification of individuals who are at a substantial risk of relapse or who have a poor prognosis. Furthermore, it offers a possible therapeutic target for the creation of innovative FLT3 inhibitors, which have demonstrated encouraging outcomes in clinical trials.

Lastly, the status of FLT3 mutations can be utilised to track how well an AML patient is responding to treatment and identify early relapses.

NPM1

NPM1 Nucleophosmin 1 is a protein that is essential to produce ribosomes, DNA repair, and the control of gene expression, among other biological function [10] NPM1 mutations are among the most prevalent genetic anomalies in acute myeloid leukaemia (AML), representing about 30% of cases [2].

A tiny DNA fragment is often inserted into the NPM1 gene in AML NPM1 mutations, producing a mutant protein that is devoid of a particular area known as the nucleolar localization signal (NLS). The cytoplasm where this mutant NPM1 protein is mislocalized, as opposed to the nucleus, where it should be. The mislocalization of NPM1 leads to abnormal gene expression, modified cell signalling pathways, and enhanced leukemic cell survival and proliferation.

When identifying NPM1 mutations in AML patients, the NPM1 biomarker is employed. Finding NPM1 mutations corresponds to a better prognosis because AML with NPM1 mutations usually has a lower recurrence rate and longer overall survival than AML without NPM1 mutations.

CEBPA

CEBPAS´s function lies on the differentiation of granulocytic cells in the bone marrow, especially in the production of neutrophils The coding of a factor involves directly in the transcription, which is essential for accurate myeloid development. It is expressed in myeloid precursor cells, including common myeloid progenitors (CMPs), granulocyte-macrophage progenitors (GMPs), and megakaryocyte-erythroid progenitors (MEPs).

As most biomarkers, it can be mutated leading to alteration in functions. Since these mutations are directly related to an AML syntype more responsive to chemotherapy and longer survival, therefore leading to a better overall prognosis[10].

It is worth to remark the double mutations (CEBPAdm) whose patients have better prognosis because of the double mutations and ends up with the production of a truncated form of CEBPA protein, which retains its DNA binding ability but lacks the negative regulatory domain. The truncated protein promotes the differentiation of granulocytic cells and suppresses cell proliferation, leading to better clinical outcomes.

PML-RARA

Acute Promyelocytic leukaemia (APL) [2] is regulated and detected by the PML-RARA biomarker, it born when the PML (promyelocytic leukaemia) gene on chromosome 15 and the RARA (retinoic acid receptor alpha) gene on chromosome 17 are merged as the consequence of a chromosomal translocation.

The abnormal proliferation and differentiation of myeloid cells is a cause of the produced PML- RARA fusion gene that generates a protein guilty of disrupting the normal PML and RARA functions. This process starts with the protein interfering in the differentiation of promyelocytes, type of immature myeloid cells. If these said cells are accumulated in the bone marrow and blood usually the symptoms which would take place are anaemia, bleeding, and an increased risk of infection. Secondly, apoptosis in inhibited by this biomarker, in other words, programmed cells start whose aim is to get rid of damaged cells.

However, it is also useful for being a target in therapeutic intervention in APL. Medications like ATRA (can produce differentiation of promyelocytes into mature granulocytes, this motivates the reduction of abnormal cells and a decreasing of symptoms. ATRA binds to RARA portion of the PML-RARA protein and then degradation of the fusion protein starts, ending with the restoration of normal myeloid differentiation.

TP53

TP53 tumour suppressor gene regulates the cell cycle and prevents the development of cancer Regarding AML TP53 mutations assure poor prognosis and deficient resistance to chemotherapy.

Genomic stability and prevention of accumulate mutations in the cell is ensure thanks to T53 as it acts as a transcription factor that helps in the regulation of the expression if a wide range of proteins mainly involved in cell cycle control, DNA repair and apoptosis.

Normal healthy cell's function is reduced to detect and repair DNA damage. After its detection, it triggers a series of downstream signalling events that ends up in repair or apoptosis. So, it is safe to say that TP53 prevent more mutations and cancer development.

When it comes to AML, TP53 mutation normally produce a loss of its tumour suppressors functions, which are linked to more aggressive disease and resistance to chemotherapy, as these cells are less likely to undergo apoptosis [2].

RUNX1

RUNX1, also known as AML1, factor oversees most of the hematopoietic stem cell differentiation, especially in the regulation of myeloid lineage and including the creation of erythrocytes, megakaryocytes, and platelets [3]. It is also playing a key role in regulation of the proliferation and survival of those cells. It is frequently mutated in AML and related with poor diagnosis and increase risk of relapse.

In AML, RUNX1 mutations can lead to a loss of its tumour suppressor function, resulting in the accumulation of genetic mutations and the development of cancer. The mutations in RUNX1 can disrupt normal myeloid differentiation and promote leukemogenesis.

10-15 % of the AML cases are due to RUNX1 mutations, these are much more frequent in patients with intermediate-risk cytogenetics. They can occur as a cause point mutations or as chromosomal translocations involving the RUNX1 gene.

In the following table a detailed look at main biomarkers essential in diagnosis, treatment, and prognosis of Acute Myeloid Leukaemia (AML), displaying its roles, implications, and potential therapeutic targets. Clinicians can use these research results to provide guidance in treatment strategies and further research into AML ´s molecular underpinnings.

Table 2.3: summary of significance of each biomarker in the context of AML, genetic implications and how they influence treatment decisions and prognosis. Source: own work.

Acronym	Complete Name	Target
FLT3	FMS -like tyrosine kinase 3	Associated with poor prognosis, aggressive disease, targeted for FLT3 inhibitors
NPM1	Nucleophosmin 1	Crucial in ribosome biogenesis and gene regulation, mutations associated with favourable prognosis in AML
CEBPA	CCAAT Enhancer Binding Protein Alpha	A transcriptome factor essential for myeloid development, often indicating better prognosis in AML
PML-RARA	Promyelocytic Leukaemia-Retinoic Acid Receptor Alpha	A fusion gene target for ATRA which induces differentiation of promyelocytes
TP53	Tumour Protein p53	Tumour suppressor gene regulating cell cycle, mutations in AML, related with poor prognosis and chemotherapy resistance
RUNX1	Runt-related transcription factor 1	A transcript factor important for hematopoietic stem cell differentiations, mutations, and poor prognosis in AML

The implementation of an AML molecular diagnosis test through NGS and RNA-seq is a multi- step process. Figure 2.10 provides a conceptual overview of this process, which starts with the preparation of the AML cell RNA sample, ensuring high-quality RNA for accurate outcomes. The RNA undergoes reverse transcription and fragmentation to create a cDNA library, sequenced using technologies like Illumina.

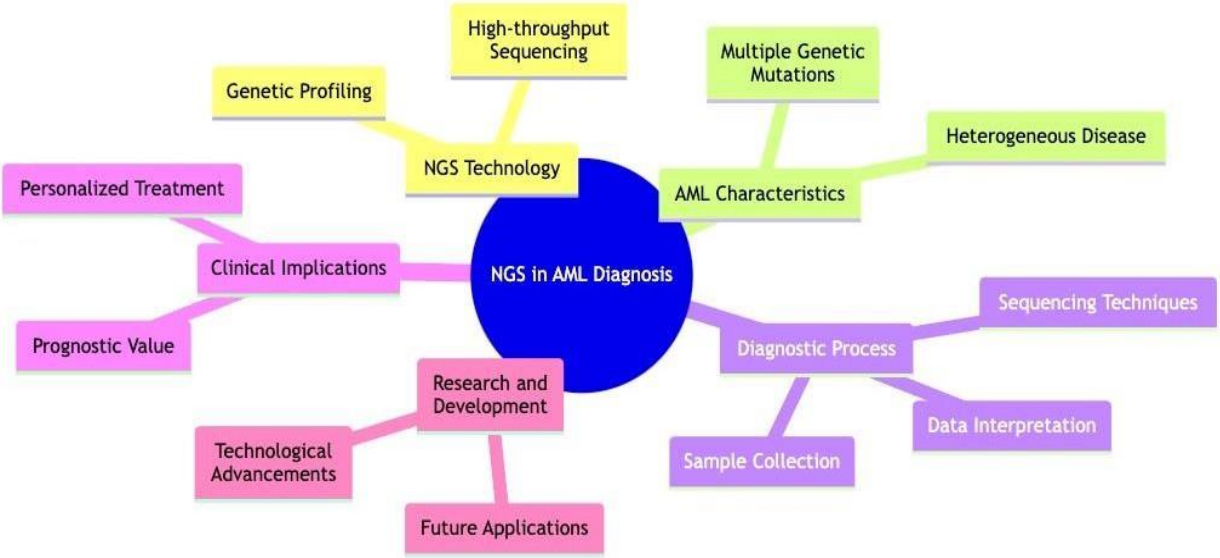


Figure 2.10: Illustrates a conceptual map for NGS. Source: own work

The selection of a gene panel, including biomarkers such as FLT3 and RUNX1, is critical for the diagnosis and understanding of AML's biological and treatment implications. The DNA libraries are sequenced with NGS technology, and the data produced are analysed through alignment with a reference genome.

Figure 2.11 outlines the workflow from sample collection to personalized medicine, showing the methodical steps required for molecular diagnosis and subsequent treatment planning. The sequence data analysis quantifies transcript abundance and compares expression levels to identify significant variations.

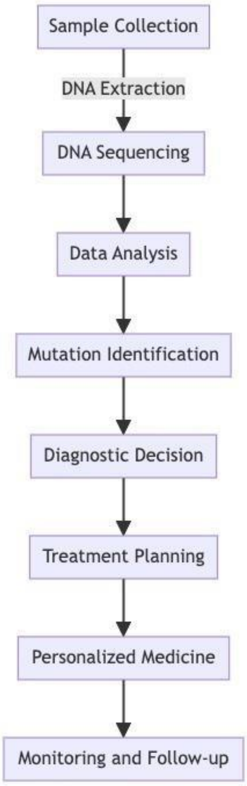


Figure 2.11: Molecular diagnosis workflow. Source: own work

In Figure 2.12 depicts how NGS technology enables comprehensive genetic analysis, which identifies mutations specific to AML. These findings inform prognosis and treatment options, leading to personalized medicine and improved patient outcomes. The data analysis includes variant interpretation, integrating clinical significance and additional information for a thorough understanding of the genetic variations.

A comprehensive report is generated, detailing the gene panel, sequence data, variant calling, and interpretation, culminating in a validated diagnostic decision. The entire process is concluded by integrating NGS results with clinical data, providing a foundation for treatment decisions and patient management following the steps below [12]:

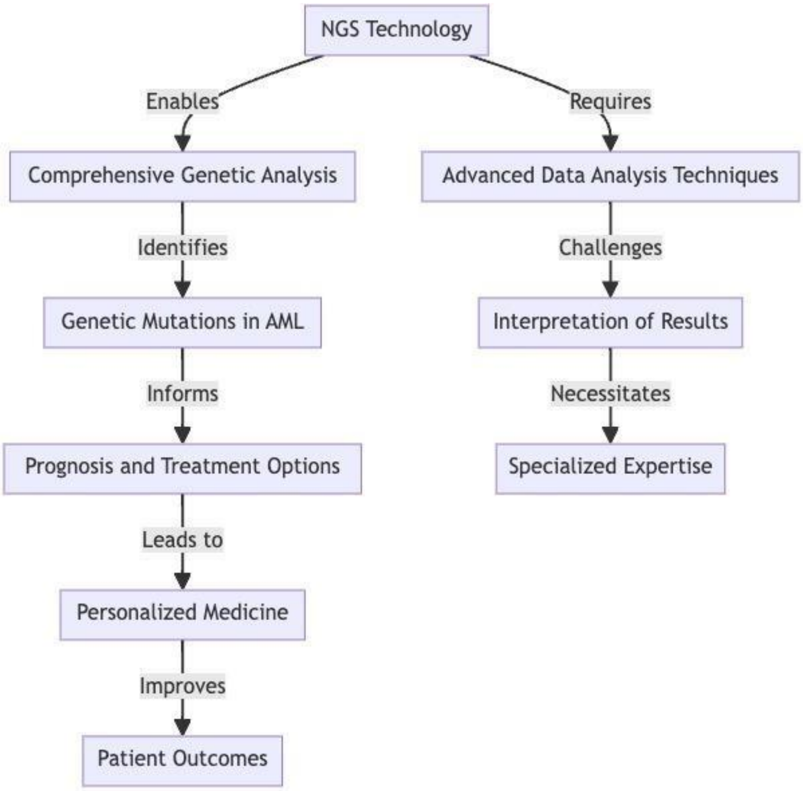


Figure 2.12: NGS workflow. Source: own work

- Preparation of sample: AML cell RNA is obtained using standard procedures to ensure high-quality and sufficient RNA for accurate outcomes.

  - Library creation: The RNA is transformed into a cDNA library through reverse transcription and fragmentation. The Illumina or equivalent technology is used to sequence the library.

- Gene panel selection: A gene panel that includes the biomarkers FLT3, NPM1, CEBPA, PML-RARA, TP53, and RUNX1, as well as other genes relevant to AML biology and treatment, is selected. The panel is customized to include sufficient coverage of each gene of interest.

- Sequencing: The DNA libraries are sequenced using next-generation sequencing (NGS) technology, in our case, we created a new tool from scratch collaborating and personalizing the Roche Hyper design Tool for our specific target: AML

- Data analysis: The sequences produced are aligned to a reference genome, and specialized software is utilized to quantify the abundance of transcripts. The expression levels of FLT3, NPM1CEBPA, PML-RARA, TP53, and RUNX1 are determined and then compared between AML samples and healthy controls or different AML subtypes.

- Variant interpretation: The variants are classified based on their clinical significance according to established guidelines. Additional information such as population frequency, functional assays, and clinical correlation is considered to aid in variant interpretation.

- Reporting: A comprehensive report is generated that includes information on the gene panel, sequencing data, variant calling, variant interpretation, and clinical significance of variants.

- Validation: Validation of the identified transcript variations is carried out using alternative techniques such as qPCR or western blotting.

- Integration: The results of the NGS analysis are integrated with clinical and pathological data to guide treatment decisions and patient management.

3. Methodology

This chapter outlines the methodological foundation of our study, where we emphasize the use of Next Generation Sequencing (NGS) in the investigation and diagnosis of Acute Myeloid Leukaemia (AML). Moreover, it provides a thorough explanation of our methods and strategies, including the procedures and instruments that make up the core of our investigation.

Using the database of the National Centre for Biotechnology information (NCBI) is a key component of our research. These databases contain a multitude of biological data, including protein structures, genetic variants, genomic sequences, and a substantial amount of scholarly literature. They are an essential research tool as they supply the data that we need to analyse and understand the results[8].

We delve into the particularities of using NCBI Entrez system, a corporate interface which is made for efficient access to said databases. For facilitating connection between databases through neighbouring and hard links, by streaming the retrieval of a biological data, it is converted into a very efficient and sophisticating tool. This approach enables a comprehensive and interconnected understanding of genetic information.

It was also briefly introducing the sequence alignment techniques to give biological and clinical context, even though when it came to practice it is not crucial in our investigation but important enough to remark it. Sequence alignment, a fundamental aspect of bioinformatics, means rearranging DNA and/ or protein sequences depending on the scope of the project to identify regions of similarity. Said similarities can give valuable information regarding gene functionalities, evolution, etc. Pairwise and sequence alignments were evaluated to ensure accuracy and relevance of our discoveries[13].

Moreover, assessing the effectiveness of our project is crucial, to do so we rely on metric such as, precision, recall and execution time. This allows us to validate our methods and result reliability, especially in the context of diagnosing and understanding AML[11].

We can conclude that this section is a comprehensive guide to our entire methodology process, which gives a detailed and achievable understanding of the tools and process we used along the way. It also set this tone for the next subsequent chapter which are about methodologies, outcomes and give reliable information about the understandings of AML and potential advancements in its diagnosis and treatment.

3.1 Overall process workflow

Figures 3.1 and 3.2 depict the process of diagnosis and management of AML in clinical algorithms. They set a visual representation that map the decision-making process for clinicians from the first suspicious of disease to the final strategies treatment based on biomarker classification. The importance of molecular testing, risk, stratification, and prognostic factors that determine treatment choices, aiming to achieve remission and avoid relapse.

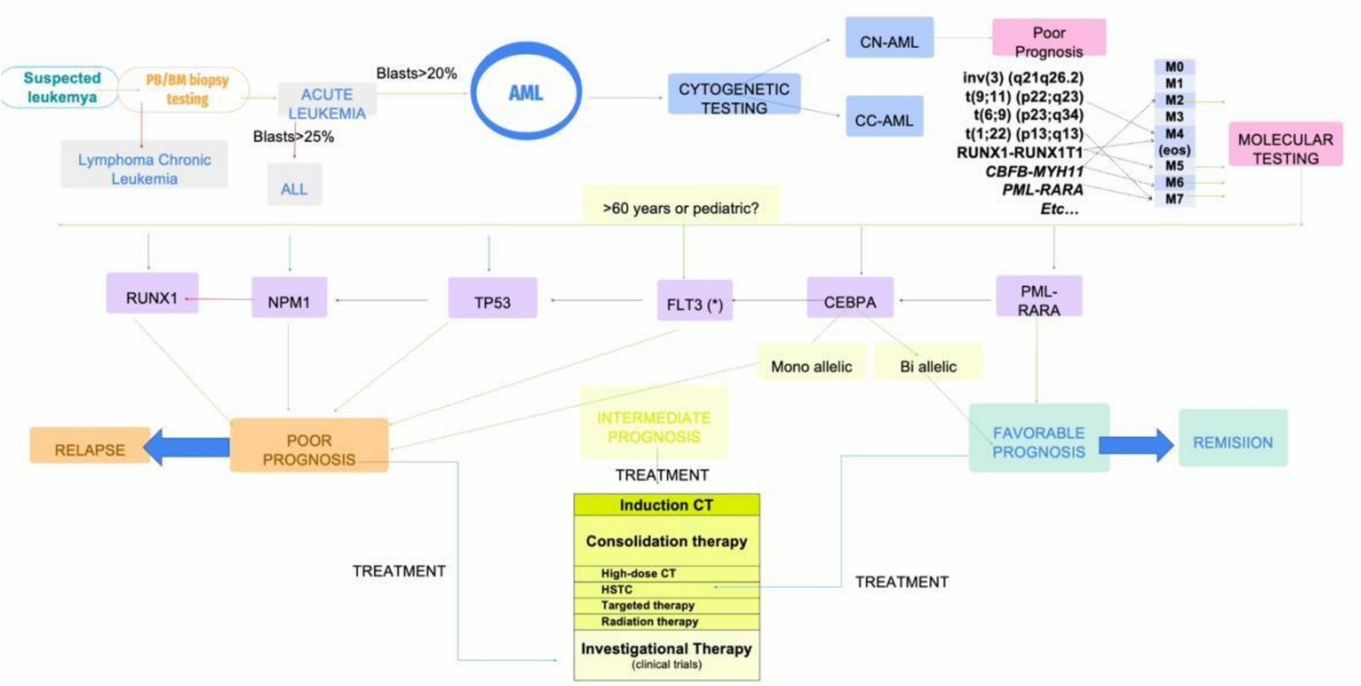


Figure 3.1: Clinical praxis overview: Biomarker classification and its repercussion in the detection, prognosis, and treatment of acute myeloid leukaemia. Source: own work

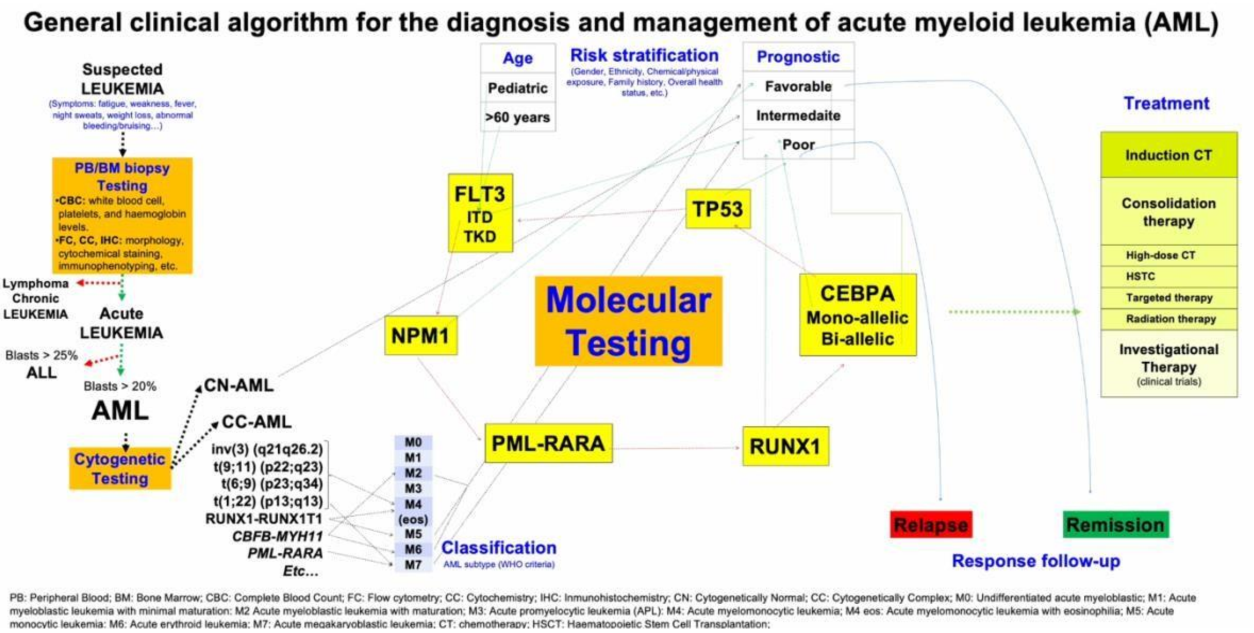

General clinical algorithm for the diagnosis and management of acute myeloid leukemia (AML)
Suspected LEUKEMIA
(Symptoms: fatigue, weakness, fever, night sweats, weight loss, abnormal bleeding/bruising…)
PB/BM biopsy Testing
•CBC: white blood cell, platelets, and haemoglobin levels.
•FC, CC, IHC: morphology, cytochemical staining, immunophenotyping, etc.
Lymphoma
Chronic LEUKEMIA
Acute LEUKEMIA
Blasts > 25%
ALL
Blasts > 20%
AML
Cytogenetic Testing
CN-AML
CC-AML
inv(3) (q21q26.2)
t(9;11) (p22;q23)
t(6;9) (p23;q34)
t(1;22) (p13;q13)
RUNX1-RUNX1T1
CBFB-MYH11
PML-RARA
Etc…
M0
M1
M2
M3
M4
(eos)
M5
M6
M7
Classification
AML subtype (WHO criteria)
Age
Pediatric
>60 years
Risk stratification
(Gender, Ethnicity, Chemical/physical exposure, Family history, Overall health status, etc.)
Prognostic
Favorable
Intermedaite
Poor
FLT3
ITD
TKD
TP53
NPM1
Molecular Testing
CEBPA
Mono-allelic
Bi-allelic
PML-RARA
RUNX1
Treatment
Induction CT
Consolidation therapy
High-dose CT
HSTC
Targeted therapy
Radiation therapy
Investigational Therapy
(clinical trials)
Relapse
Remission
Response follow-up
PB: Peripheral Blood; BM: Bone Marrow; CBC: Complete Blood Count; FC: Flow cytometry; CC: Cytochemistry; IHC: Inmunohistochemistry; CN: Cytogenetically Normal; CC: Cytogenetically Complex; M0: Undifferentiated acute myeloblastic; M1: Acute myeloblastic leukemia with minimal maturation: M2 Acute myeloblastic leukemia with maturation; M3: Acute promyelocytic leukemia (APL): M4: Acute myelomonocytic leukemia; M4 eos: Acute myelomonocytic leukemia with eosinophilia; M5: Acute monocytic leukemia: M6: Acute erythroid leukemia: M7: Acute megakaryoblastic leukemia: CT: chemotherapy: HSCT: Haematopoietic Stem Cell Transplantation:

Figure 3.2: General clinical algorithm for the diagnosis and management of AML. Source: own work

In acute myeloid leukaemia (AML), RNA-seq has the capability to identify specific transcript variations linked to certain AML subtypes, such as FLT3, NPM1CEBPA, PML-RARA, TP53, and RUNX1. In Figure 3.3 is an outline of the general process:

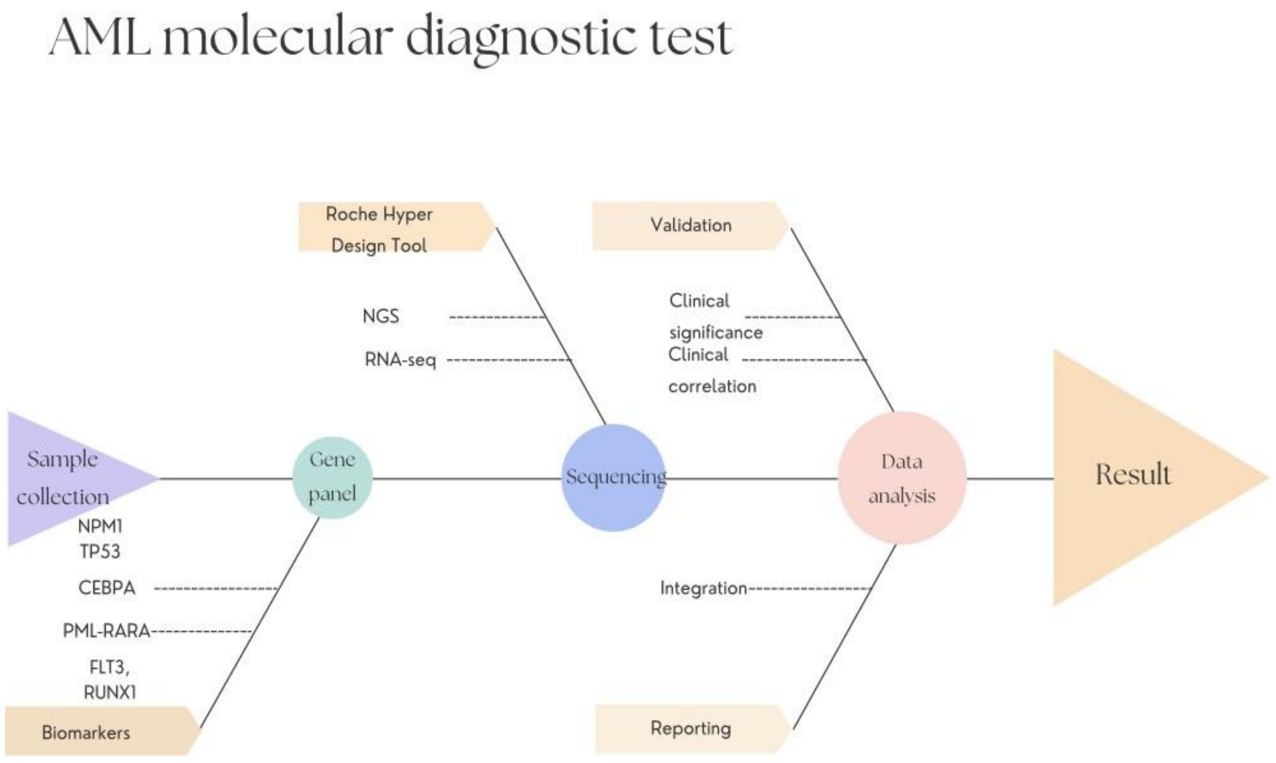
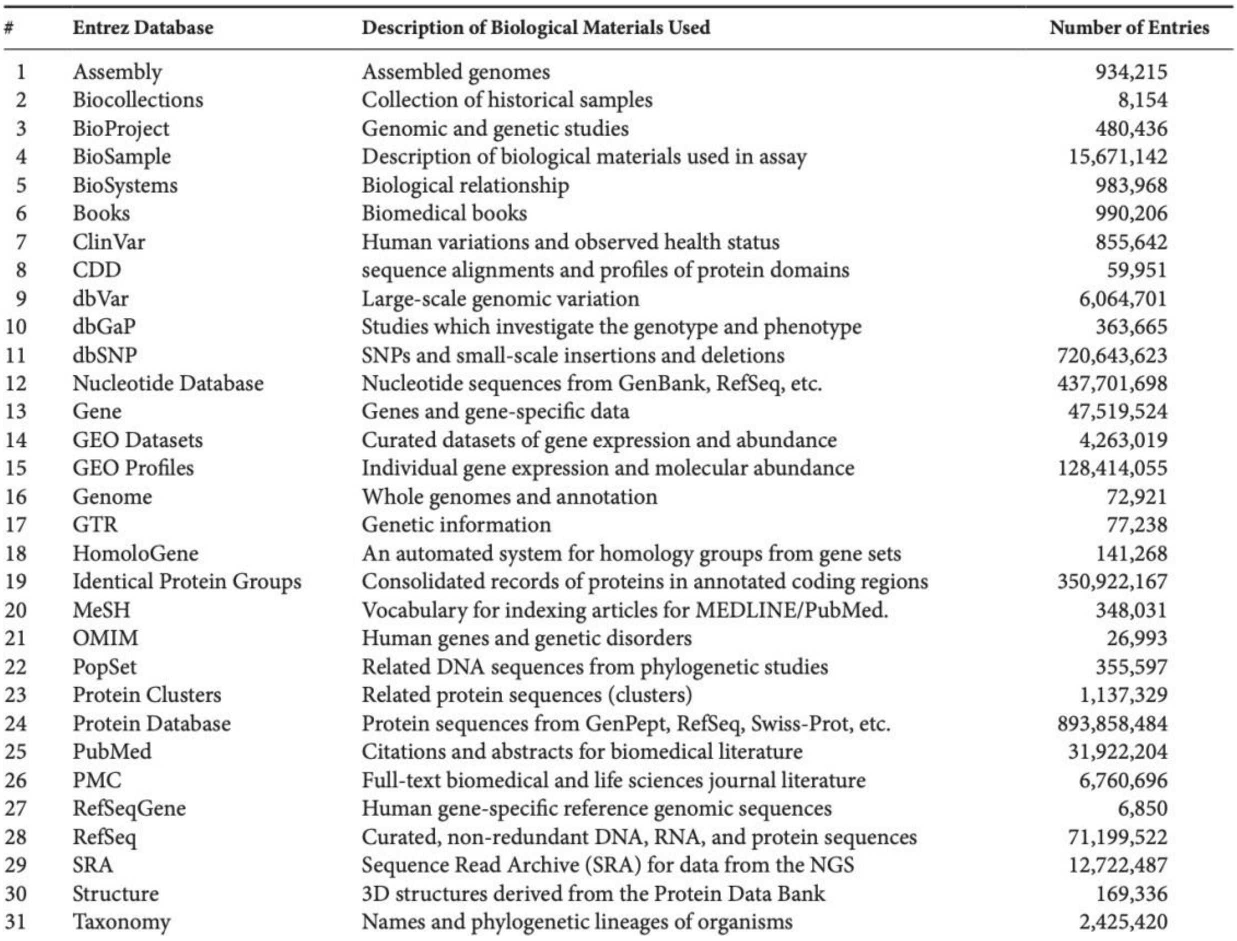


Figure 3.3: Overview of implemented AML molecular test. Source: own work.

3.2 NCBI Databases

The National Centre for Biotechnology (NCBI), part of the United States National Library of Medicine (NLM), documents biological information critical for advancing scientific research and healthcare.

In Figure 3.4 we can see the interconnection and correlation between the collection of databases maintained and included in NCBI. Information related to genomic sequences, genetic variations and protein structures can be found in these databases regarding the extensive range of biological data. They act as very useful repositories for researchers, scientists and healthcare professionals worldwide [14].



#	Entrez Database	Description of Biological Materials Used	Number of Entries
1	Assembly	Assembled genomes	934,215
2	Biocollections	Collection of historical samples	8,154
3	BioProject	Genomic and genetic studies	480,436
4	BioSample	Description of biological materials used in assay	15,671,142
5	BioSystems	Biological relationship	983,968
6	Books	Biomedical books	990,206
7	ClinVar	Human variations and observed health status	855,642
8	CDD	sequence alignments and profiles of protein domains	59,951
9	dbVar	Large-scale genomic variation	6,064,701
10	dbGaP	Studies which investigate the genotype and phenotype	363,665
11	dbSNP	SNPs and small-scale insertions and deletions	720,643,623
12	Nucleotide Database	Nucleotide sequences from GenBank, RefSeq, etc.	437,701,698
13	Gene	Genes and gene-specific data	47,519,524
14	GEO Datasets	Curated datasets of gene expression and abundance	4,263,019
15	GEO Profiles	Individual gene expression and molecular abundance	128,414,055
16	Genome	Whole genomes and annotation	72,921
17	GTR	Genetic information	77,238
18	HomoloGene	An automated system for homology groups from gene sets	141,268
19	Identical Protein Groups	Consolidated records of proteins in annotated coding regions	350,922,167
20	MeSH	Vocabulary for indexing articles for MEDLINE/PubMed.	348,031
21	OMIM	Human genes and genetic disorders	26,993
22	PopSet	Related DNA sequences from phylogenetic studies	355,597
23	Protein Clusters	Related protein sequences (clusters)	1,137,329
24	Protein Database	Protein sequences from GenPept, RefSeq, Swiss-Prot, etc.	893,858,484
25	PubMed	Citations and abstracts for biomedical literature	31,922,204
26	PMC	Full-text biomedical and life sciences journal literature	6,760,696
27	RefSeqGene	Human gene-specific reference genomic sequences	6,850
28	RefSeq	Curated, non-redundant DNA, RNA, and protein sequences	71,199,522
29	SRA	Sequence Read Archive (SRA) for data from the NGS	12,722,487
30	Structure	3D structures derived from the Protein Data Bank	169,336
31	Taxonomy	Names and phylogenetic lineages of organisms	2,425,420

Figure 3.4: Entrez databases and number of records[10].

3.2.1. NCBI Entrez relationships between entries

Entrez was created to address the need of finding all the information about a particular biological entity without visiting and querying individual databases, one by one. This solution was ground-breaking, and it soon became one of the most widely used interfaces to gather information about biological databases in the NCBI Entrez system.

As mentioned above, Entrez is an interface not a database. Users can use its components to access databases that can be traversed, like PubMed records, nucleotides, protein sequence data, information on conserved protein, domains and three-dimensional structure information. The revolutionary approach is the access by just the execution of a single query , thanks to the fact that two connections are used between database entries : neighbouring and hard links [14,15].

The possibility for a certain database to be connected to one another is due to neighbouring. The same user can ask Entrez to search for all papers with the desired information which coincides in subject matter to the original paper. Likewise, the same process can be achieved with the sequences list regarding the original sequence.

These neighbouring relationships are created based on statistical similarity. Hard links are enhanced between entries in several databases. A reasonable connection between entries is necessary. Hard links exist only when there is a logical connection between entries. For example, if a PubMed entry contains a sequencing description of chromosomal regions and a gene of interest, a hard link can be established between the PubMed entry and the corresponding nucleotide entry for that specific gene. The Basic Logical Alignment Tool (BLAST) is one of the most used techniques for achieving similarity between sequences. Its purpose is to find high-scoring segment pairs to detect biological sequence data are compared using this method [1].

Its capabilities involve finding the best region of local alignment between a query sequence and its target and determine if other plausible alignments are feasible. It starts with seeding the search with a query word to discover all the matches that be relevant in a biological and informative aspect. VAST (Vector Alignment Search Tool) is also widely used for comparing molecule structure using a specific set of coordinates.

Regarding the comparison methodology, it considers vectors obtained by calculating whether how the C-terminal connects to N-terminal or, in other words, regarding the position of the secondary structural elements, and previously first alpha and beta strands that compose 3D-coordinates are identified. When comparison ends, optimization begins by aligning the vector for pairs of structural elements, following relative orientation, type, and connectivity criteria.

Finally, to build more accurate template-based protein models by conformational sampling these vectors, which are already aligned, undergo a refinement procedure to optimize the structural alignment.

3.2.2 The Entrez Discovery Pathway

With the main purpose to understand the repercussion of neighbouring and hard links concepts in bioinformatics with the Entrez system, in this section we can distinguish practical examples fully explained.

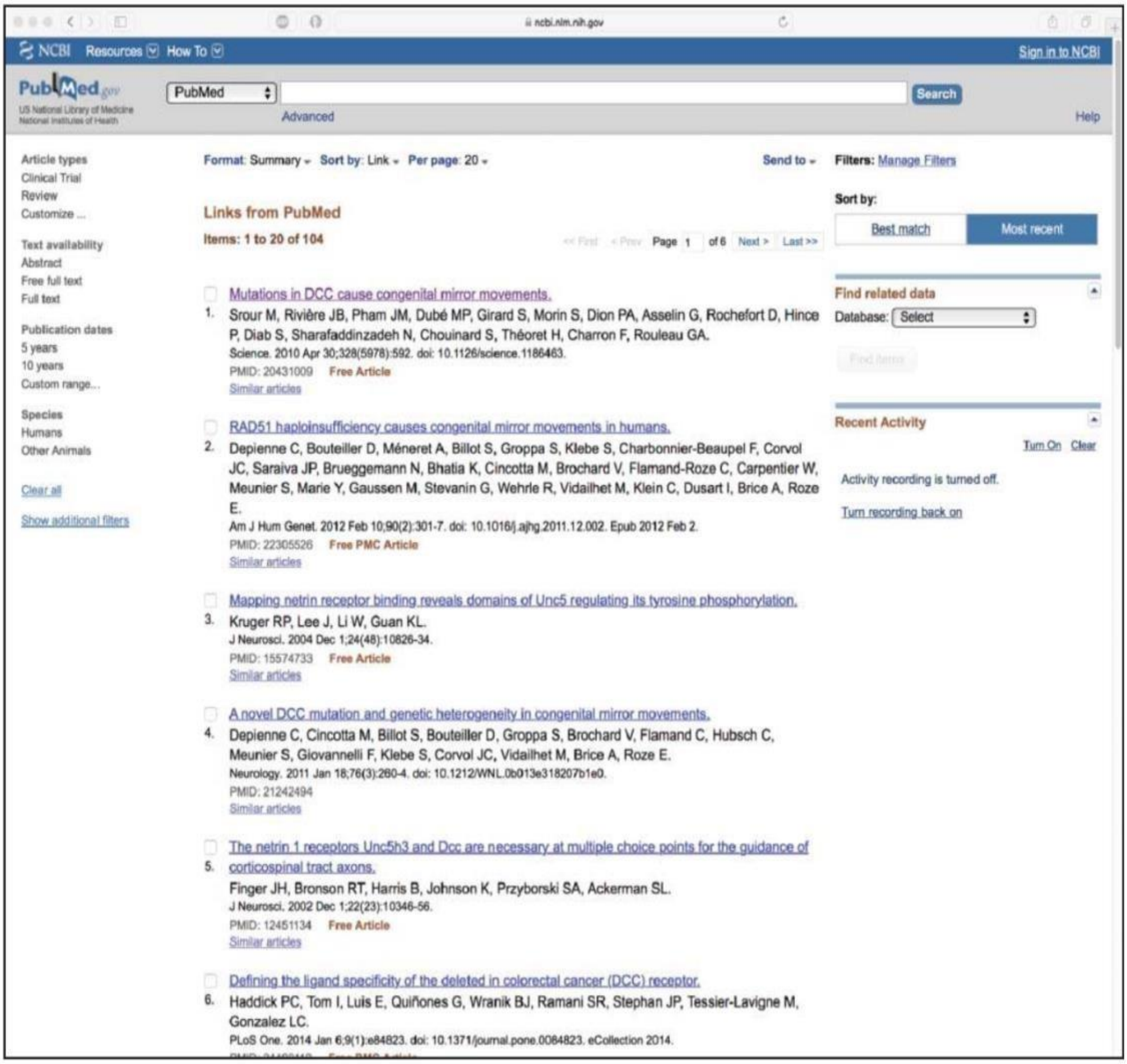
In Figure 3.6 is illustrated the Boolean operators to query Entrez and use individual search terms by analysing the existence of a query interface at the top of NCBI home page for the database user selection to search from a pull-down menu and a text box.

In the output result user can see the title, author, and citation of the paper to be clicked by the user and redirect it to the abstract view [16].

- Various alternative formats are available for displaying this information which can be selected using the pull down-menu, this layout consists of:
  - A two-letter code referring to the contest of each field ging down the left-hand side of the entry, entries can be saved and imported into a third-party bibliography program.
  - The Discovery column at the right-hand side enable access to the full text version and link to additional useful related information.

The similar articles section presents entry points from which users may find advantageous neighbouring and hard-link relationships previously described, as shown in Figure 3.5.

Finally, Entrez can return a list of references related to the original search paper following the criteria of statistical similarity with the parent entry (reference), a very useful time saving tool is the scanning titles which allows the user to rapidly find related information and group a relevant bibliography.



Links from PubMed

Items: 1 to 20 of 104

Activity recording is turned off.

1. Mutations in DCC cause congenital mirror movements.
Srour M, Rivière JB, Pham JM, Dubé MP, Girard S, Morin S, Dion PA, Asselin G, Rochefort D, Hince P, Diab S, Sharafaddinzadeh N, Chouinard S, Théoret H, Charron F, Rouleau GA.
Science. 2010 Apr 30;328(5978):592. doi: 10.1126/science.1186463.
PMID: 20431009 Free Article
Similar articles

2. RAD51 haploinsufficiency causes congenital mirror movements in humans.
Depienne C, Bouteiller D, Méneret A, Billot S, Groppa S, Klebe S, Charbonnier-Beaupel F, Corvol JC, Saraiva JP, Brueggemann N, Bhatia K, Cincotta M, Brochard V, Flamand-Roze C, Carpentier W, Meunier S, Marie Y, Gaussen M, Stevanin G, Wehrle R, Vidailhet M, Klein C, Dusart I, Brice A, Roze E.
Am J Hum Genet. 2012 Feb 10;90(2):301-7. doi: 10.1016/j.ajhg.2011.12.002. Epub 2012 Feb 2.
PMID: 22305526 Free PMC Article
Similar articles

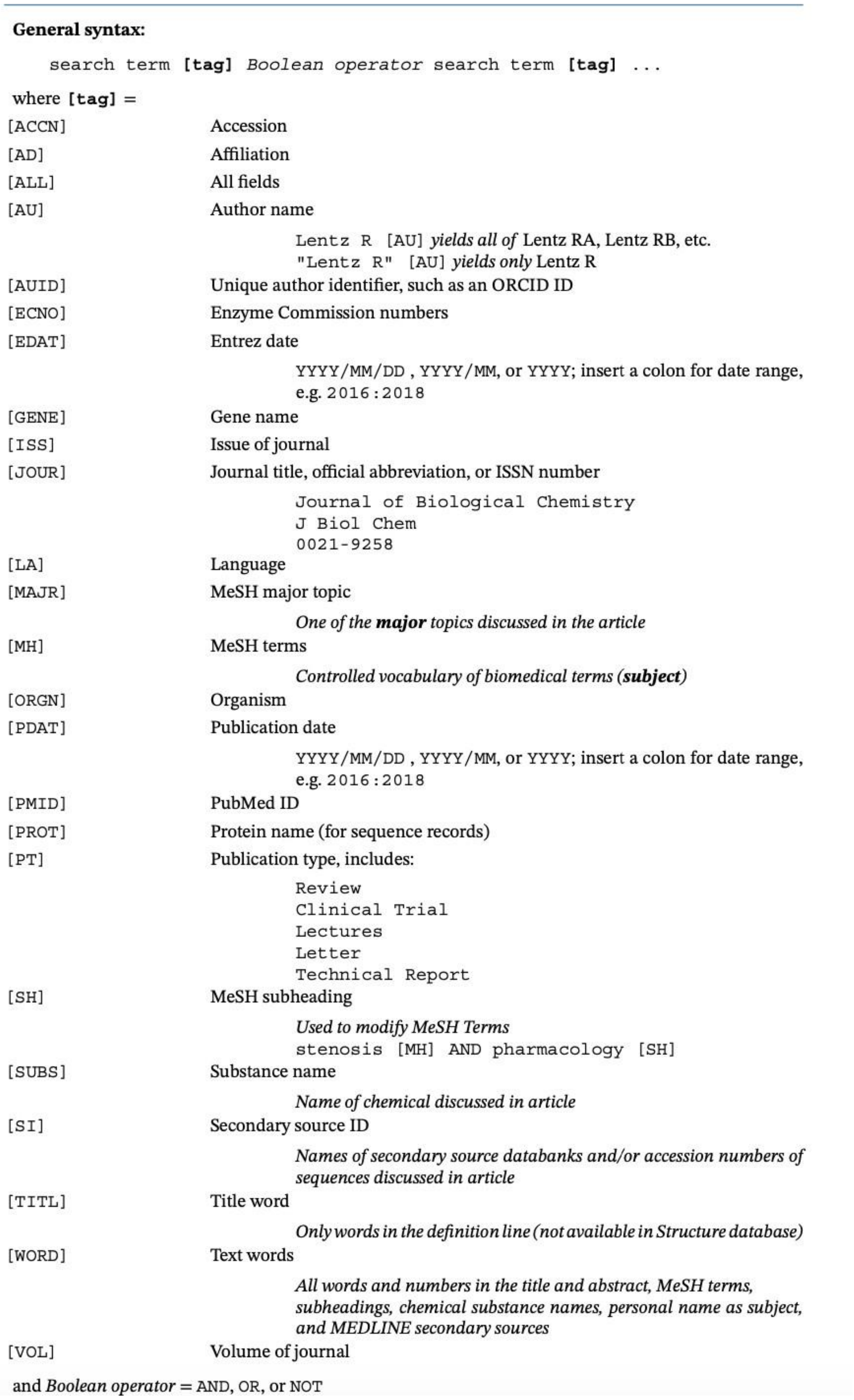
3. Mapping netrin receptor binding reveals domains of Unc5 regulating its tyrosine phosphorylation.
Kruger RP, Lee J, Li W, Guan KL.
J Neurosci. 2004 Dec 1;24(48):10826-34.
PMID: 15574733 Free Article
Similar articles

4. A novel DCC mutation and genetic heterogeneity in congenital mirror movements.
Depienne C, Cincotta M, Billot S, Bouteiller D, Groppa S, Brochard V, Flamand C, Hubsch C, Meunier S, Giovannelli F, Klebe S, Corvol JC, Vidailhet M, Brice A, Roze E.
Neurology. 2011 Jan 18;76(3):260-4. doi: 10.1212/WNL.0b013e318207b1e0.
PMID: 21242494
Similar articles

5. The netrin 1 receptors Unc5h3 and Dcc are necessary at multiple choice points for the guidance of corticospinal tract axons.
Finger JH, Bronson RT, Harris B, Johnson K, Przyborski SA, Ackerman SL.
J Neurosci. 2002 Dec 1;22(23):10346-56.
PMID: 12451134 Free Article
Similar articles

6. Defining the ligand specificity of the deleted in colorectal cancer (DCC) receptor.
Haddick PC, Tom I, Luis E, Quiñones G, Wranik BJ, Ramani SR, Stephan JP, Tessier-Lavigne M, Gonzalez LC.
PLoS One. 2014 Jan 6;9(1):e84823. doi: 10.1371/journal.pone.0084823. eCollection 2014.

Figure 3.5: Shows a practical example if neighbours to an entry found in PubMed [14].



General syntax:

`search term [tag] Boolean operator search term [tag] ...`

where **`[tag]`** =

[ACCN]	Accession
[AD]	Affiliation
[ALL]	All fields
[AU]	Author name `Lentz R [AU]` *yields all of* Lentz RA, Lentz RB, etc. `"Lentz R" [AU]` *yields only* Lentz R
[AUID]	Unique author identifier, such as an ORCID ID
[ECNO]	Enzyme Commission numbers
[EDAT]	Entrez date `YYYY/MM/DD` , `YYYY/MM`, or `YYYY`; insert a colon for date range, e.g. `2016:2018`
[GENE]	Gene name
[ISS]	Issue of journal

[JOUR]	Journal title, official abbreviation, or ISSN number
	`Journal of Biological Chemistry`
	`J Biol Chem`
	`0021-9258`

[LA]	Language
[MAJR]	MeSH major topic *One of the **major** topics discussed in the article*
[MH]	MeSH terms *Controlled vocabulary of biomedical terms (**subject**)*
[ORGN]	Organism
[PDAT]	Publication date `YYYY/MM/DD` , `YYYY/MM`, or `YYYY`; insert a colon for date range, e.g. `2016:2018`
[PMID]	PubMed ID
[PROT]	Protein name (for sequence records)

[PT]	Publication type, includes:
	`Review`
	`Clinical Trial`
	`Lectures`
	`Letter`
	`Technical Report`

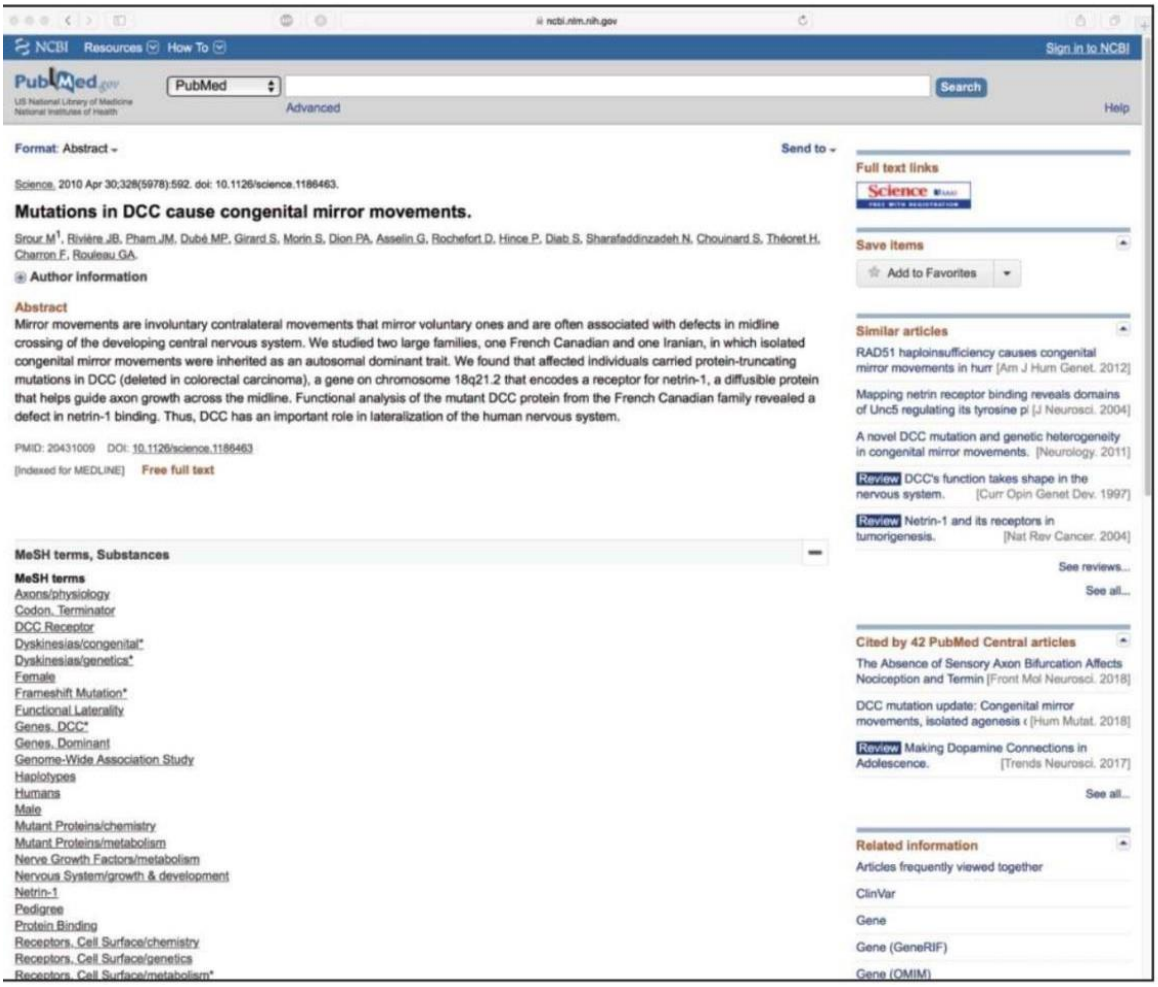
[SH]	MeSH subheading *Used to modify MeSH Terms* `stenosis [MH] AND pharmacology [SH]`
[SUBS]	Substance name *Name of chemical discussed in article*
[SI]	Secondary source ID *Names of secondary source databanks and/or accession numbers of sequences discussed in article*
[TITL]	Title word *Only words in the definition line (not available in Structure database)*
[WORD]	Text words *All words and numbers in the title and abstract, MeSH terms, subheadings, chemical substance names, personal name as subject, and MEDLINE secondary sources*
[VOL]	Volume of journal

and *Boolean operator* = AND, OR, or NOT

Figure 3.6: Shows Entrez Boolean search statements [14].

In Figure 3.7 following the explanation of the abstract view, there are displayed a series if hard- link connections at the discovery column this serve as a redirection to other databases in the Entrez to package of extensive data, in our case example the Entrez Gene feature is selected an illustrated in Figure 3.8 where information linked to sources like RefSeq is gathered [15].

All kind of data can be displayed from this attribute like genomic regions, transcripts, product of genes to expression data, phenotype, protein on protein interaction and homologies in similar sequence in selected organisms.



Science. 2010 Apr 30;328(5978):592. doi: 10.1126/science.1186463.

Mutations in DCC cause congenital mirror movements.

Srour M[1], Rivière JB, Pham JM, Dubé MP, Girard S, Morin S, Dion PA, Asselin G, Rochefort D, Hince P, Diab S, Sharafaddinzadeh N, Chouinard S, Théoret H, Charron F, Rouleau GA.

Author information

Abstract

Mirror movements are involuntary contralateral movements that mirror voluntary ones and are often associated with defects in midline crossing of the developing central nervous system. We studied two large families, one French Canadian and one Iranian, in which isolated congenital mirror movements were inherited as an autosomal dominant trait. We found that affected individuals carried protein-truncating mutations in DCC (deleted in colorectal carcinoma), a gene on chromosome 18q21.2 that encodes a receptor for netrin-1, a diffusible protein that helps guide axon growth across the midline. Functional analysis of the mutant DCC protein from the French Canadian family revealed a defect in netrin-1 binding. Thus, DCC has an important role in lateralization of the human nervous system.

PMID: 20431009 DOI: 10.1126/science.1186463

[Indexed for MEDLINE] Free full text

MeSH terms, Substances

MeSH terms

Axons/physiology
Codon, Terminator
DCC Receptor
Dyskinesias/congenital*
Dyskinesias/genetics*
Female
Frameshift Mutation*
Functional Laterality
Genes, DCC*
Genes, Dominant
Genome-Wide Association Study
Haplotypes
Humans
Male
Mutant Proteins/chemistry
Mutant Proteins/metabolism
Nerve Growth Factors/metabolism
Nervous System/growth & development
Netrin-1
Pedigree
Protein Binding
Receptors, Cell Surface/chemistry
Receptors, Cell Surface/genetics
Receptors, Cell Surface/metabolism*

Similar articles

RAD51 haploinsufficiency causes congenital mirror movements in hum [Am J Hum Genet. 2012]

Mapping netrin receptor binding reveals domains of Unc5 regulating its tyrosine pl [J Neurosci. 2004]

A novel DCC mutation and genetic heterogeneity in congenital mirror movements. [Neurology. 2011]

Review DCC's function takes shape in the nervous system. [Curr Opin Genet Dev. 1997]

Review Netrin-1 and its receptors in tumorigenesis. [Nat Rev Cancer. 2004]

See reviews...

See all...

Cited by 42 PubMed Central articles

The Absence of Sensory Axon Bifurcation Affects Nociception and Termin [Front Mol Neurosci. 2018]

DCC mutation update: Congenital mirror movements, isolated agenesis ‹ [Hum Mutat. 2018]

Review Making Dopamine Connections in Adolescence. [Trends Neurosci. 2017]

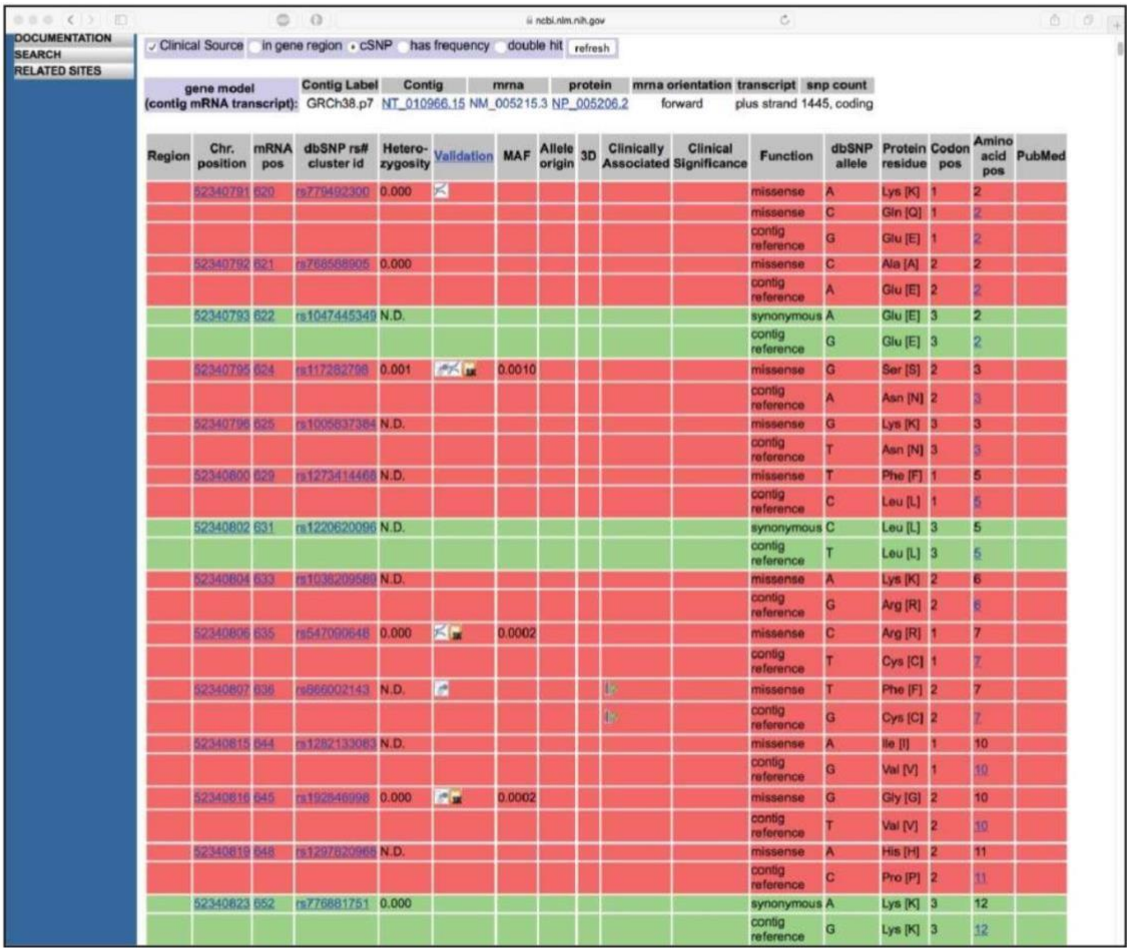
See all...

Related information

Articles frequently viewed together
ClinVar
Gene
Gene (GeneRIF)
Gene (OMIM)

Figure 3.7: An example of a PubMed record in Abstract format, as returned through Entrez [14].

Further down the Discovery Column are extensive lists of links to additional resources provided through NCBI and other sources. One link of note is the *SNP: Gene View* link, taking the user to data derived from dbSNP. The information found within dbSNP goes beyond just single-nucleotide polymorphisms (SNPs), including data on short genetic variations such as short insertions and deletions, short tandem repeats etc.



ncbi.nlm.nih.gov

DOCUMENTATION
SEARCH
RELATED SITES

☑ Clinical Source ○ in gene region ● cSNP ○ has frequency ○ double hit refresh

gene model (contig mRNA transcript):	Contig Label	Contig	mrna	protein	mrna orientation	transcript	snp count
	GRCh38.p7	NT_010966.15	NM_005215.3	NP_005206.2	forward	plus strand	1445, coding

Region	Chr. position	mRNA pos	dbSNP rs# cluster id	Hetero-zygosity	Validation	MAF	Allele origin	3D	Clinically Associated	Clinical Significance	Function	dbSNP allele	Protein residue	Codon pos	Amino acid pos	PubMed
	52340791	620	rs779492300	0.000							missense	A	Lys [K]	1	2	
											missense	C	Gln [Q]	1	2	
											contig reference	G	Glu [E]	1	2	
	52340792	621	rs768588905	0.000							missense	C	Ala [A]	2	2	
											contig reference	A	Glu [E]	2	2	
	52340793	622	rs1047445349	N.D.							synonymous	A	Glu [E]	3	2	
											contig reference	G	Glu [E]	3	2	
	52340795	624	rs117282798	0.001		0.0010					missense	G	Ser [S]	2	3	
											contig reference	A	Asn [N]	2	3	
	52340796	625	rs1005837384	N.D.							missense	G	Lys [K]	3	3	
											contig reference	T	Asn [N]	3	3	
	52340800	629	rs1273414468	N.D.							missense	T	Phe [F]	1	5	
											contig reference	C	Leu [L]	1	5	
	52340802	631	rs1220620096	N.D.							synonymous	C	Leu [L]	3	5	
											contig reference	T	Leu [L]	3	5	
	52340804	633	rs1038209589	N.D.							missense	A	Lys [K]	2	6	
											contig reference	G	Arg [R]	2	6	
	52340806	635	rs547090648	0.000		0.0002					missense	C	Arg [R]	1	7	
											contig reference	T	Cys [C]	1	7	
	52340807	636	rs866002143	N.D.							missense	T	Phe [F]	2	7	
											contig reference	G	Cys [C]	2	7	
	52340815	644	rs1282133083	N.D.							missense	A	Ile [I]	1	10	
											contig reference	G	Val [V]	1	10	
	52340816	645	rs192846998	0.000		0.0002					missense	G	Gly [G]	2	10	
											contig reference	T	Val [V]	2	10	
	52340819	648	rs1297820968	N.D.							missense	A	His [H]	2	11	
											contig reference	C	Pro [P]	2	11	
	52340823	652	rs776881751	0.000							synonymous	A	Lys [K]	3	12	
											contig reference	G	Lys [K]	3	12	

Figure 3.8: A section of the Database of Single Nucleotide Polymorphisms [14].

The main purpose of the sequence alignment method is the rearrange of DNA to detect similarities among regions, which are used to determine the potential evolutionary relationships. Its repercussion allows us to predict gene functions, classify proteins into distinct families, sequence annotation and novels sequence identification.

The process involves selecting the most suitable alignment algorithm, scoring scheme and statistical tests to determine the significance of the observed similarity.

We can distinguish between pairwise or multiple sequence alignment.

- Pairwise alignment takes two sequence and compare it to the reference one to align it. It is the core of widely tool know called BLAST which makes easy the search of relatable and similar sequences in the sequence databases [14].

- Multiple sequence alignment: to seek the optimal align, various sequences are analysed, normally as previous step the pairwise alignment is needed to ensure the accuracy of the output results, then we can find its application in several genomic endeavours, including phylogeny, motif discovery prediction of conserved protein regions and homology modelling all of that to predicted protein structure and functionalities [17].

3.2.3 Pairwise alignment

Mutations are the cause frequency diversity, which occurs at different genomes regions and at distinct areas. Regions can be conserved or non-protein coding regions, the first type contains slow mutations rates preserving biological functions and the second suffer a faster mutation rate.

Mutations can also have more consequences such as emerge of new phenotypes, they can be accelerated also by more factors like chemical exposure, radiations, cancer, or common ancestor malignant genes. Regarding common ancestor genes, relationships between genes were establish in homology studies within across our specie, it can be divided in tow groups: paralog or ortholog [18].

Insightful conclusions can be drawn from mutations like the emerge of new phenotypes as well as they can be accelerated by many factors like exposure to chemical, radiations, cancer, or common ancestral genes. For the last one, homology studies arise to cover the need to establish similarity relationships between genes derived form a common ancestor within or across species; they can be paralog genes (evolved from a common ancestral gene within the same species) or orthologs (genes in different species derived from a common ancestral gene).

Currently the most popular procedure for this task is pairwise sequence alignment performing a comparison of DNA or protein sequences, with global alignment (spam the length sequence entirely), local alignment (identifying specific regions of similarity).

In the following subsection we will deep in global sequence alignment, local sequence alignment and specialized local alignment tool known as BLAST.

3.2.4 Global Alignment Algorithm

To search for introns, exons, promoter regions in genes sequence, or domain and motifs in a protein sequence the best align match is the global alignment algorithm to align the entire sequence to another. Algorithms must find the best alignment among all the broad possibilities since many mutations like deletion and insertions come into play making it a must to optimize the alignments output to one. To do this, for the sake of simplicity the dynamic programming algorithm divides the whole sequence aligned into smaller subs-alignments to find the optimal one.

However, one must take into consideration the case where sequences don't have the same length, which is normally the rule rather than the exception, this can be caused by the presence of deletion or insertion, to solve this, gaps are introduced to maximize the alignment and scoring measurements quantifying the degree of similarity [13].

It is useful when the sequences have the same length and are similar, however the normal case is that the sequences have not the same length due to deletion and or insertion mutation, for that gaps may be introduced to maximize the alignment and scoring measurements to quantify the degree of similarity, as shown in Example 1:

```
MENSEL-AQY
ME—ELDAQY
```

Example 1: example of two sequences before and after length adaptation to perform global alignment.

Difficulties arise when we try to align two distant and long but related sequences. To address this case computer algorithms were developed to find optimal alignment in a suitable time.

The first one, created by Needleman and Wunsch, is a global alignment algorithm which returns a high score for similarity criteria and a low one when evolutionary distance criteria is applied. In either case, the first step is to create an alignment score matrix with $(m + 1) \times (n + 1)$ dimensions, where *m* and *n* are the length of the first and second sequence (*y* and *x*) respectively. Any *aa* (amino acids) in *y* is denoted by y_i, where $i = 1,2, \dots, m$, and x_i is any residual where $j = 1,2, \dots, n - 1$.

The *x* and *y* major sequences are divided into subsequences $\{y_1, \dots, y_i\}$ and $\{x_1, \dots, x_j\}$, respectively. The goal is to compute scores of optimal alignments and determine the path which finds the maximum score and subsequently find optimal solution regarding alignment.
The scoring matrix include PAM or BLOSUM; the following criteria are applied:

- Match pair: +1.
- Unmatched pair: -1
- Gap penalty (g): -2

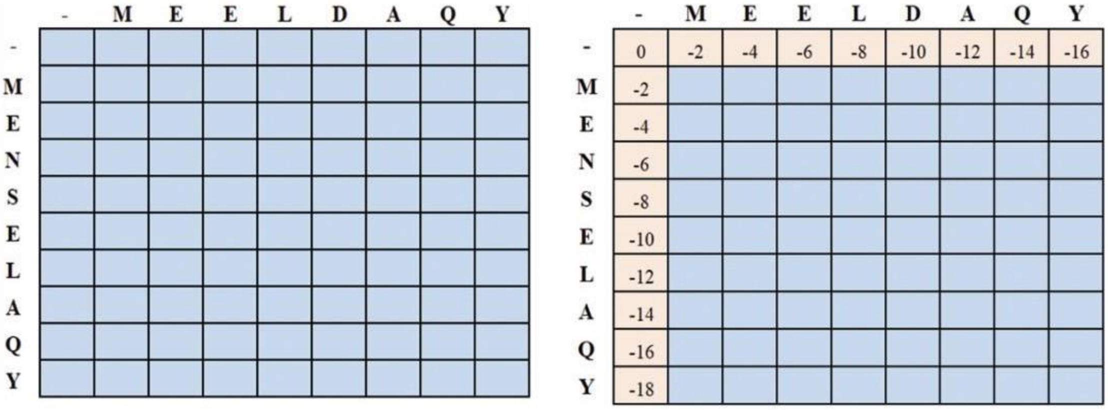
The global algorithm methods follow these procedures [2]:

1. Initialize the scoring matrix: As seen in Figure 3.9, rows and columns are filled starting from the top left corner and the value 0; then the rest of the cells are filled with alignments scores. We can obtain such scores with two formulas. For the cells of the first column:

$$S_i, 1 = (S_i - 1,1 + g),$$

whereas for the cells of the first row:

$$S_1, j = (S_i - 1,1 + g).$$



	-	M	E	E	L	D	A	Q	Y
-									
M									
E									
N									
S									
E									
L									
A									
Q									
Y									

	-	M	E	E	L	D	A	Q	Y
-	0	-2	-4	-6	-8	-10	-12	-14	-16
M	-2								
E	-4								
N	-6								
S	-8								
E	-10								
L	-12								
A	-14								
Q	-16								
Y	-18								

Figure 3.9: Initialization process matrix undergoes changes from an empty alignment scoring matrix (left) to an initialized one (right)[10].

2. Fill in the cells of the scoring matrix: the following four scenarios may arise:
   - Pair amino acids are matched.
   - Pair of amino acids are mismatched.
   - Gap may be introduced from the first sequence.
   - Gap may be introduced from the second sequence.

The dynamic programming algorithm offers a scoring formula for each of the following scenarios. To complete the filling of the remaining cells, three paths are considered within the scoring matrix:

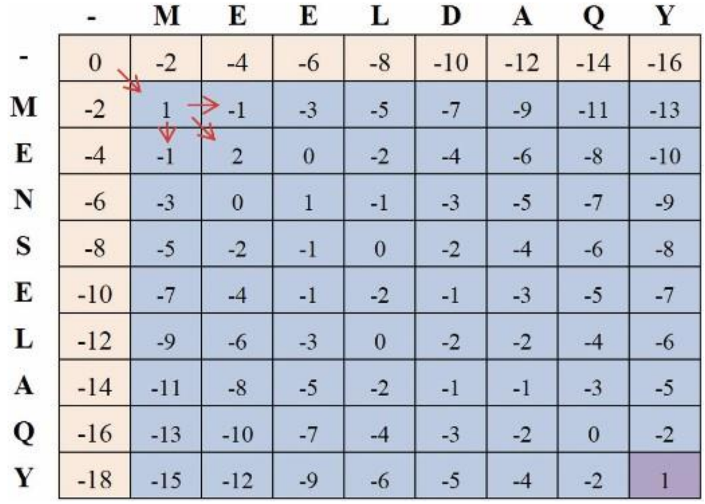
- A diagonal move from position $(i - 1, j - 1)$ to position (i, j) without any penalty for introducing gap.

- A move from any other position along column j with a penalty for introducing a gap.
- A move from any other position along row i with a penalty for introducing a gap.

The dynamic programming algorithm finds the optimal alignment score (S_{ij}) as the maximum score in the set [$S_{i-1.j-1}$ +s (y_i, x_j), $S_{i-1.j}$ +g, $S_{i.\ j-1}$ +g], which is given by:

$$S_{ij} = \text{Max}\begin{cases} S_{i-1.j-1} + s(y_i, x_j) \\ S_{i-1.j} + g \\ S_{i.j-1} + g \end{cases}$$

As the maximum score is obtained from the diagonal S1,1 + s (y2, x2) = +1 among the three scores, the next move will be diagonal from (y1, x1) to (y2, x2), resulting in a score of S2,2 = +1. This movement can be represented by an arrow, as depicted in Figure 3.10 For the alignment of (y3, x2), we observe a mismatch between the characters (E, M), with a score of s (y3, x2) = -1. Therefore, the resulting score is obtained as follows:



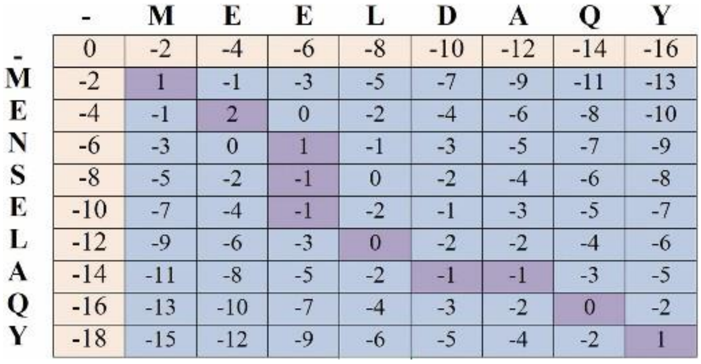
	-	M	E	E	L	D	A	Q	Y
-	0	-2	-4	-6	-8	-10	-12	-14	-16
M	-2	1	-1	-3	-5	-7	-9	-11	-13
E	-4	-1	2	0	-2	-4	-6	-8	-10
N	-6	-3	0	1	-1	-3	-5	-7	-9
S	-8	-5	-2	-1	0	-2	-4	-6	-8
E	-10	-7	-4	-1	-2	-1	-3	-5	-7
L	-12	-9	-6	-3	0	-2	-2	-4	-6
A	-14	-11	-8	-5	-2	-1	-1	-3	-5
Q	-16	-13	-10	-7	-4	-3	-2	0	-2
Y	-18	-15	-12	-9	-6	-5	-4	-2	1

Figure 3.10: The provided illustration presents a completed alignment scoring matrix depicting the optimal alignment scores for the given subsequence [14].

Each cell within the matrix contains a score representing the optimal alignment of the subsequence. Moreover, the score located in the bottom right corner of the matrix represents the overall optimal alignment score.

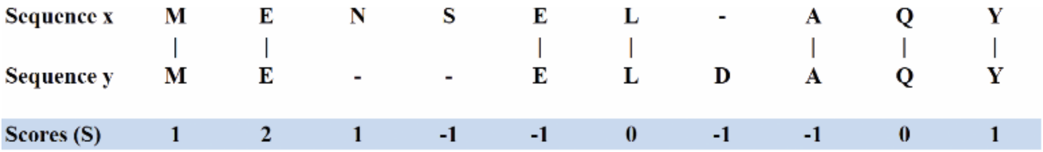
3. Traceback: the optimal alignment begins, it starts from lower right cell, which is state as the overall optimal alignment score, and retrace the path by choosing the maximum score of the adjacent cells. In our example, we can see that staring from the score + 1 cell (Y, Y) the maximum score is 0 for (Q, Q), following the process in order and continuing this reasoning we arrive to the conclusive path shown in Figure 3.11, it ends

when the cell of the first aligned pair is reached, giving the output illustrated in Figure 3.12.



	-	M	E	E	L	D	A	Q	Y
-	0	-2	-4	-6	-8	-10	-12	-14	-16
M	-2	1	-1	-3	-5	-7	-9	-11	-13
E	-4	-1	2	0	-2	-4	-6	-8	-10
N	-6	-3	0	1	-1	-3	-5	-7	-9
S	-8	-5	-2	-1	0	-2	-4	-6	-8
E	-10	-7	-4	-1	-2	-1	-3	-5	-7
L	-12	-9	-6	-3	0	-2	-2	-4	-6
A	-14	-11	-8	-5	-2	-1	-1	-3	-5
Q	-16	-13	-10	-7	-4	-3	-2	0	-2
Y	-18	-15	-12	-9	-6	-5	-4	-2	1

Figure 3.11: Represent the complete traceback path for the optimal alignment [14].



Sequence x	M	E	N	S	E	L	-	A	Q	Y
	|	|			|	|		|	|	|
Sequence y	M	E	-	-	E	L	D	A	Q	Y
Scores (S)	1	2	1	-1	-1	0	-1	-1	0	1

Figure 3.12: represent the optimal sequence alignment. [14]

3.2.5 Local Sequence Alignment and Algorithm

A simpler approach is given by the local alignment algorithm, which is used to identify specific aligned regions within two sequences, with tools like BLAST. The scoring matrix is computed in a similar fashion as the global alignment algorithm. Negative values are included, as well, indicating mismatches or gaps. The main difference is that there is no such thing as gap penalty criteria. Instead, optimal subsequence values below 0 are discarded and restarted to 0. That is, the optimal score is the one that maximizes among the four options in the Smith Waterman algorithm. It can be computed as [16,14]:

$$S_{ij} = \mathrm{Max}\begin{cases} S_{i-1.j-1} + s(y_i, x_j) \\ S_{i-1.j} + g \\ S_{i.j-1} + g \\ 0 \end{cases}$$

Let us apply the Smith-Waterman algorithm to perform a local alignment of two example protein sequences: "AQYLQNTFKV" (x) and "NPKLQNTLA" (y). We will use the identity scoring scheme with +1 for a matched pair, -0.4 for a mismatched pair, and -1.4 for a gap.

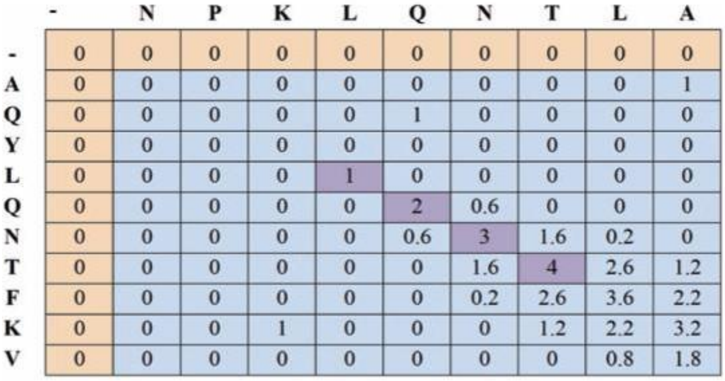
The steps to create the alignment scoring matrix are the following (see also Figure 3.13).

(1) Creating an empty matrix.

(2) Initializing the cells of the first left column and top row with zeros.

(3) Filling the remaining cells of the matrix.

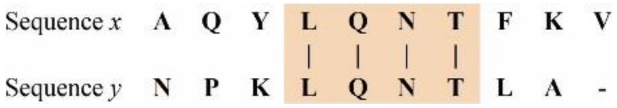
(4) Performing traceback to identify the local alignments: the traceback process starts from the highest score in the alignment matrix and proceeds diagonally up to the left, stopping when a cell with a value of zero is reached.



	-	N	P	K	L	Q	N	T	L	A
-	0	0	0	0	0	0	0	0	0	0
A	0	0	0	0	0	0	0	0	0	1
Q	0	0	0	0	0	1	0	0	0	0
Y	0	0	0	0	0	0	0	0	0	0
L	0	0	0	0	1	0	0	0	0	0
Q	0	0	0	0	0	2	0.6	0	0	0
N	0	0	0	0	0	0.6	3	1.6	0.2	0
T	0	0	0	0	0	0	1.6	4	2.6	1.2
F	0	0	0	0	0	0	0.2	2.6	3.6	2.2
K	0	0	0	1	0	0	0	1.2	2.2	3.2
V	0	0	0	0	0	0	0	0	0.8	1.8

Figure 3.13: Alignment scoring matrix for the two sequences[14].

The constraint in the formula sets negative scores to 0, and the traceback starts from the highest score, moving diagonally up and stopping at the first zero score creating the output display in Figure 3.14.



```
Sequence x   A  Q  Y  L  Q  N  T  F  K  V
                      |  |  |  |
Sequence y   N  P  K  L  Q  N  T  L  A  -
```

Figure 3.14: Represent the optimal sequence alignment[14].

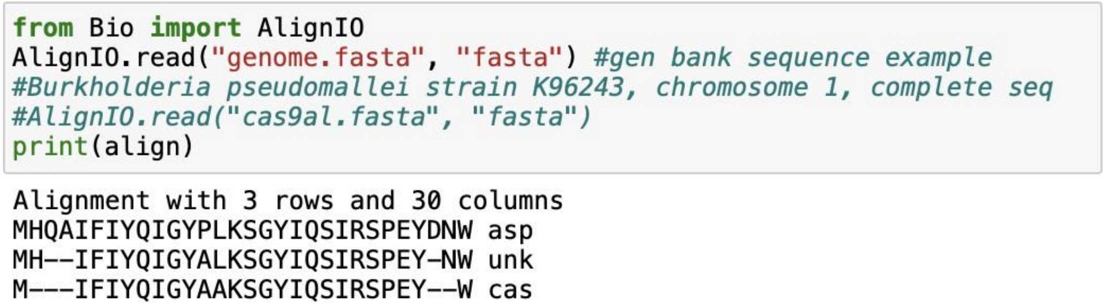
3.3 Biopython

Biopython examples were tested in this practical section to provide useful insight real life cases regarding alignments algorithms. Biopython is a python package of numerous implemented modules designed for bioinformatics applications. There are some common procedures for bioinformatics problematics which are constated and address by between programs and standard file formats to be handled by this package [15].

Biopyhton most widely used components in our project are the following.

AlignIo: is the input output interface to work with alignments sequences. The procedure is similar as SeqIO, but it returns an Alignment object, illustrated in Figure 3.15. Three main aspects can be seen:

- `read(handle, format [, sec_count]`: Inputs parameters are a file handled and the alignment format for data preparation and processing and generates an alignment output object that is return.
- The `sec_count` arguments indicate the number of sequences per alignment.
- `write(iterable, handle, format)`: converts file handle, alignment object and file format into an only file.



```
from Bio import AlignIO
AlignIO.read("genome.fasta", "fasta") #gen bank sequence example
#Burkholderia pseudomallei strain K96243, chromosome 1, complete seq
#AlignIO.read("cas9al.fasta", "fasta")
print(align)
```

```
Alignment with 3 rows and 30 columns
MHQAIFIYQIGYPLKSGYIQSIRSPEYDNW asp
MH--IFIYQIGYALKSGYIQSIRSPEY-NW unk
M---IFIYQIGYAAKSGYIQSIRSPEY--W cas
```

Figure 3.15: Code example with result expressed in expression of alignments. Source: own work.

BLAST: aims to compare a user´s query to a database of sequences. It finds match between two sequences named "hot spots", alignment process begins, and statistical information is provided like the expected value. BLAST applications include [12]:

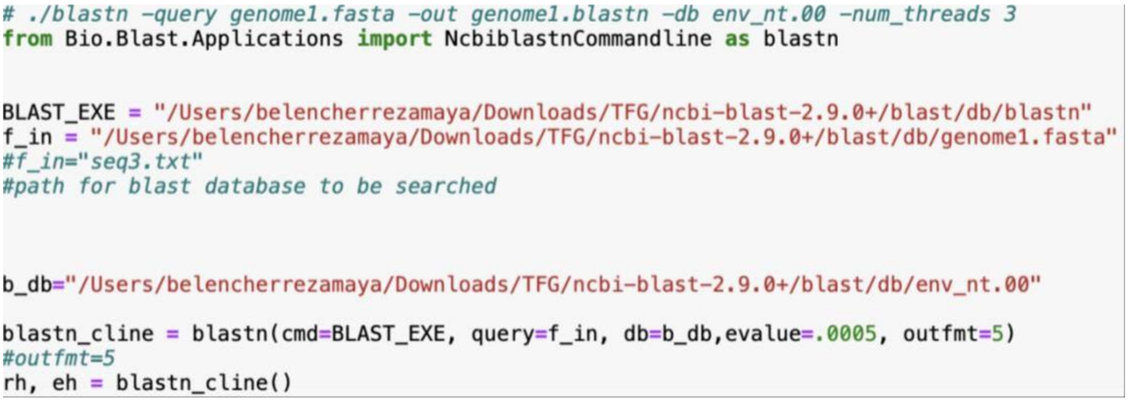
- Discovering relationships in function and evolution between sequences.
- Identifying member of gene families.
- Iterating through the species genome to find similar genes.
- Searching for regulatory patterns like promoter signals, splicing sites and transcriptor binding sites.

Table 3.3: Blast programmes[12].

Program name	Query/database combination
blastn	nucleotide vs nucleotide
blastp	protein vs protein
blastx	translated nucleotide vs protein
tblastn	protein vs translated nucleotide
Tblastx	translated nucleotide vs translated nucleotide

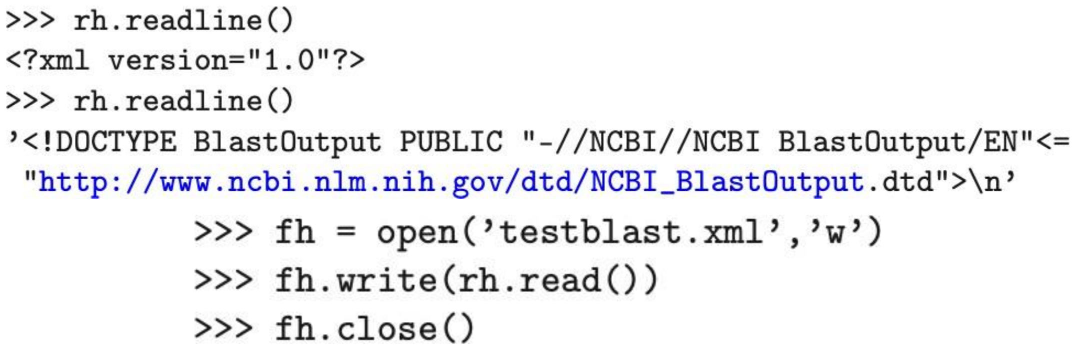
-

- BLAST Running Processing with Biopython: it has a plugin in NCBI webserver for the user to try it online, other approach may be locally on your computer, it really depends on the one desired approach. The handling of large amount of data is done locally otherwise online NCBI used could be targeted, this way we avoid the error: "CPU usage limit was exceeded "and allows us to create a custom database. This example is illustrated in Figure 3.16 [8].
- Starting a Blast Job: Biopython has an executable for every blast programme from your script with the following command described in line 5 in Figure 3.17:



```
# ./blastn -query genome1.fasta -out genome1.blastn -db env_nt.00 -num_threads 3
from Bio.Blast.Applications import NcbiblastnCommandline as blastn


BLAST_EXE = "/Users/belencherrezamaya/Downloads/TFG/ncbi-blast-2.9.0+/blast/db/blastn"
f_in = "/Users/belencherrezamaya/Downloads/TFG/ncbi-blast-2.9.0+/blast/db/genome1.fasta"
#f_in="seq3.txt"
#path for blast database to be searched



b_db="/Users/belencherrezamaya/Downloads/TFG/ncbi-blast-2.9.0+/blast/db/env_nt.00"

blastn_cline = blastn(cmd=BLAST_EXE, query=f_in, db=b_db,evalue=.0005, outfmt=5)
#outfmt=5
rh, eh = blastn_cline()
```

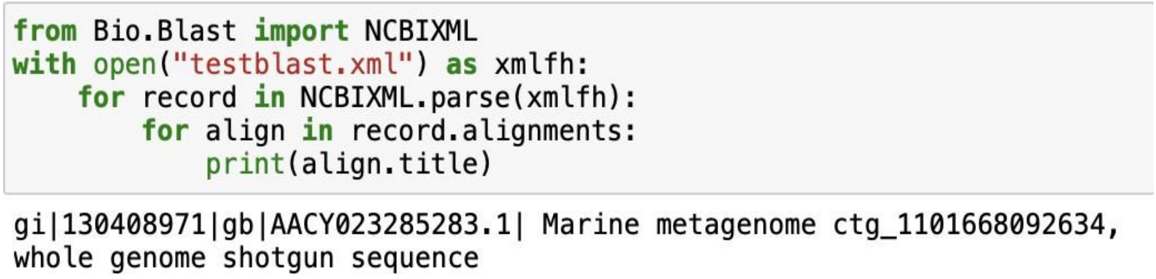
Figure 3.16: In this code a command where a local NCBI BLAST is shown. Source: own work



```
>>> rh.readline()
<?xml version="1.0"?>
>>> rh.readline()
'<!DOCTYPE BlastOutput PUBLIC "-//NCBI//NCBI BlastOutput/EN"<=
 "http://www.ncbi.nlm.nih.gov/dtd/NCBI_BlastOutput.dtd">\n'
          >>> fh = open('testblast.xml','w')
          >>> fh.write(rh.read())
          >>> fh.close()
```

Figure 3.17: This code shows the process of reading the file-like object rh, to then process it and store it. Source: own work

- Reading the Blast output: for the sake of parsing the object, a module called parse is of utility which takes an input file object with the BLAST result for each record inside the file, as seen in Figures 3.18 and 3.19.



```
from Bio.Blast import NCBIXML
with open("testblast.xml") as xmlfh:
    for record in NCBIXML.parse(xmlfh):
        for align in record.alignments:
            print(align.title)
```

```
gi|130408971|gb|AACY023285283.1| Marine metagenome ctg_1101668092634,
whole genome shotgun sequence
```

Figure 3.18: Python code to extract alignments title from a BLAST output, and its output image.

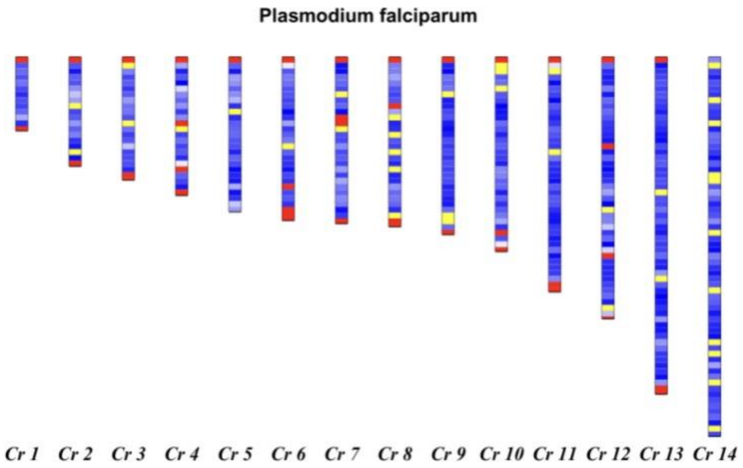


Figure 3.19: Extract alignments title from a BLAST output, and its output image [15].

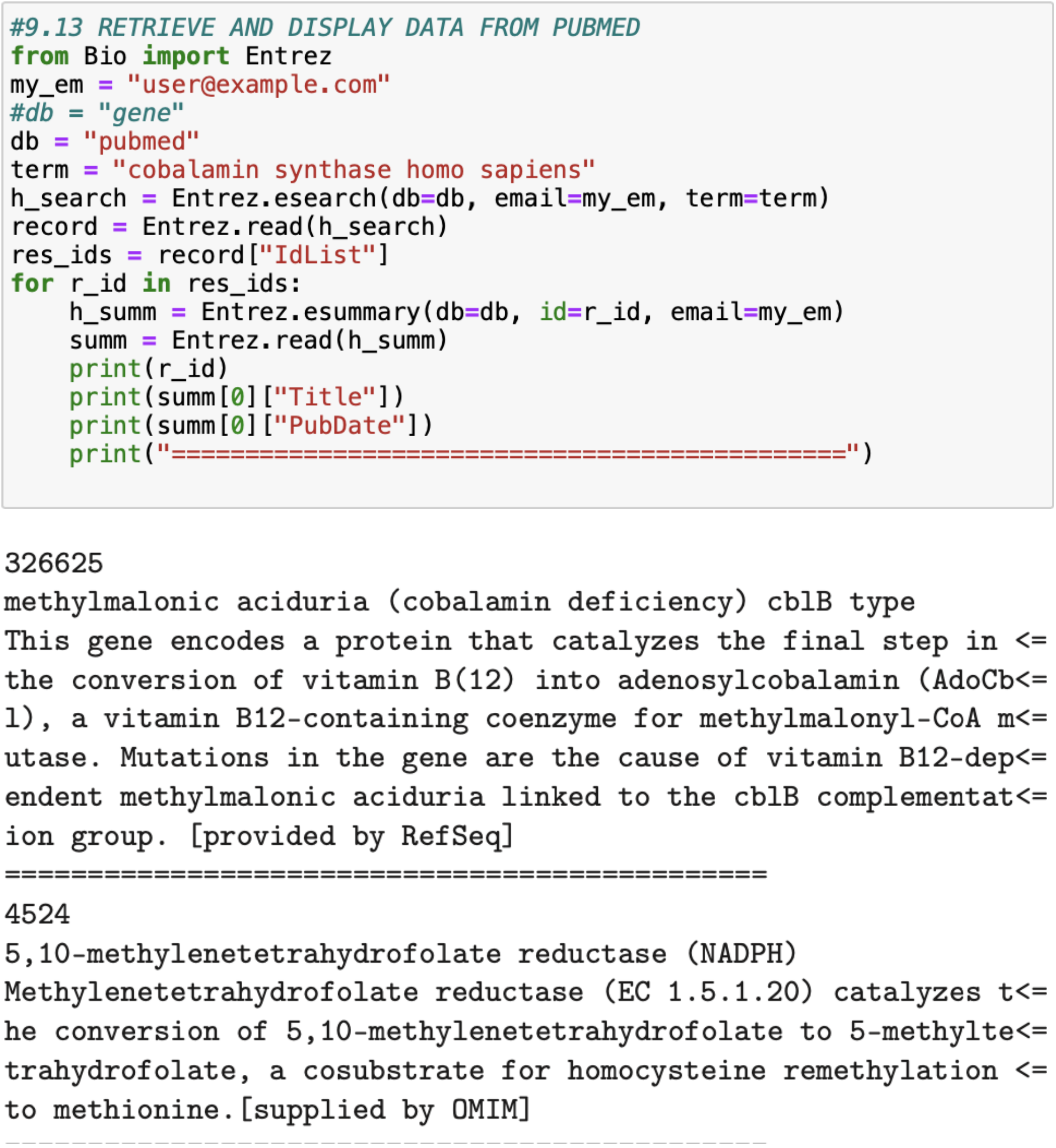
3.3.1 Biopython and entrez (Bio.Entrez)

NCBI datasets and other health science related data is integrated in Entrez, a user can access through two different methods: the standard online way or by Entrez Programming Utilities made by NCBI by which one can query the data without the need to browser and then incorporating this related info to your local scripts.

A crafted URL must be created composed by the name of the program to use in NCBI web server and required parameters, after posting, retrieving, and processing the desire information the resulted data is sent in XML format to the user by the NCBI.

Table 3.4: Bio. Entrez module utilities and attributes [15].

Name	Description
efetch	Retrieves records in the requested format form a list of one or more primary IDs or from the user´s environment.
einfo	Provides field index term counts, last update, and available links for each database.
egquery	Provides Entrez database counts in XML for a single search using Global Query.
elink	Checks for the existence of an external or Related Articles link form a list of one or more primary IDs.
epost	Posts a file containing a list of primary IDs for future use in the user´s environment to use with subsequent search strategies
esearch	Searches and retrieves primary IDs (for use in EFetch, Elink, and ESummary).
espell esummary	Retrieves spelling suggestions and documents summaries from a list of primaries IDs or from the user´s environment.
read	Parses the XML results returned by any of the above functions.



```python
#9.13 RETRIEVE AND DISPLAY DATA FROM PUBMED
from Bio import Entrez
my_em = "user@example.com"
#db = "gene"
db = "pubmed"
term = "cobalamin synthase homo sapiens"
h_search = Entrez.esearch(db=db, email=my_em, term=term)
record = Entrez.read(h_search)
res_ids = record["IdList"]
for r_id in res_ids:
    h_summ = Entrez.esummary(db=db, id=r_id, email=my_em)
    summ = Entrez.read(h_summ)
    print(r_id)
    print(summ[0]["Title"])
    print(summ[0]["PubDate"])
    print("==============================================")
```

```
326625
methylmalonic aciduria (cobalamin deficiency) cblB type
This gene encodes a protein that catalyzes the final step in <=
the conversion of vitamin B(12) into adenosylcobalamin (AdoCb<=
l), a vitamin B12-containing coenzyme for methylmalonyl-CoA m<=
utase. Mutations in the gene are the cause of vitamin B12-dep<=
endent methylmalonic aciduria linked to the cblB complementat<=
ion group. [provided by RefSeq]
==============================================
4524
5,10-methylenetetrahydrofolate reductase (NADPH)
Methylenetetrahydrofolate reductase (EC 1.5.1.20) catalyzes t<=
he conversion of 5,10-methylenetetrahydrofolate to 5-methylte<=
trahydrofolate, a cosubstrate for homocysteine remethylation <=
to methionine.[supplied by OMIM]
==============================================
```

Figure 3.20: Shows a practice example from creating an URL, parsing, retrieving, and storing info to create the desire output, gene information is sent to the user and display. Source: own work.

3.4 Performance evaluation

In this section, we will evaluate the performance of our model by focusing on execution time and key performances metrics, primary accuracy. This evaluation process includes generating histograms to visualize data distribution and using confusion matrices to assess model accuracy and other metrics.

We will start by explaining the prerequisites for creating histograms, including the necessary libraries and data acquisition methods. Following this, we will detail the construction of histograms for primer coverage and estimated coverage. The evaluation continues with the use of a confusion matrix to determine true positives, false negatives, and true negatives, from which we will derive performance metrics such as accuracy, precision, and recall. These metrics will provide a comprehensive understanding of the model´s effectiveness.

3.4.1 Previous requisites

This section describes how to obtain the histograms presented in the previous result. Below is a breakdown of each phase in this procedure:

- Linking required libraries:

  1. Libraries like Pandas were used to manage data structures such as data frames and dictionaries. Matplotlib. `pyplot` enables graphical representation of self-made histograms and diagrams. After this, to complete the preparation process we must incorporate required libraries for data handling and chart creation.

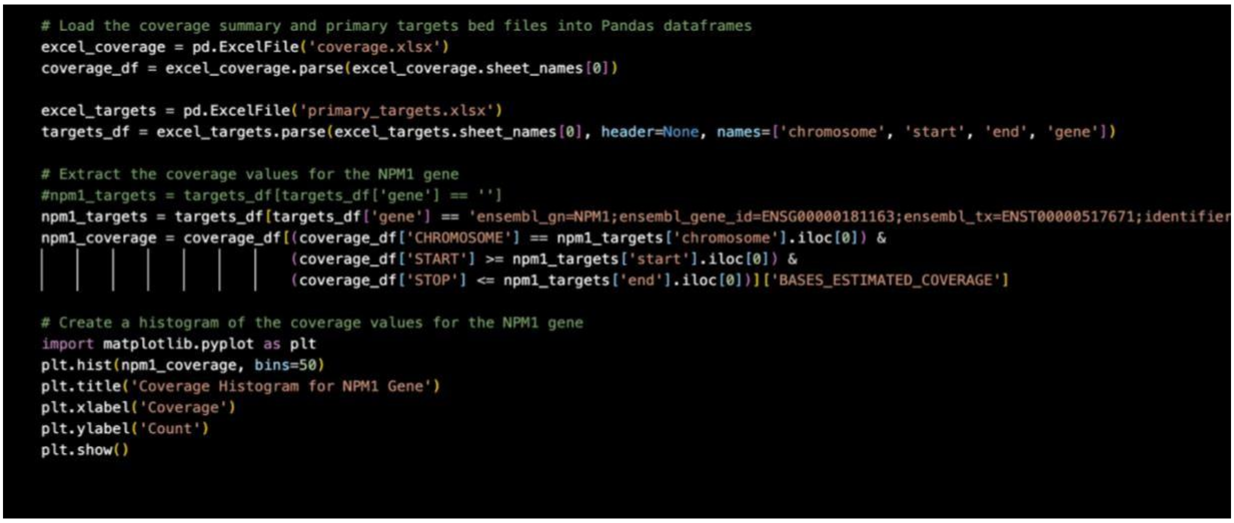
- Data Acquisition: The `read_excel ()` function in Pandas let us specify the path to files from which we can load the coverage information from an Excel document into a panda's data frame.

  1. Construction of Primer Coverage Histogram:

     - The `plt.figure()` function lets us set the size of the visual representation, to enhance adequate and legible representation.
     - Calling `sns.histplot()` from the Seaborn package to assemble the histogram.
     - Embellishing the histogram with a heading and axis titles by implementing `plt.title()`, `plt.xlabel()`, and `plt.ylabel()`.
     - Improve legibility by enabling the grid using the `plt.grid()` function.

  2. Establishing Estimated Coverage Histogram:

     - Replicate the mentioned steps for the histogram of estimated coverage, this occasion referencing the `'FRAC_ESTIMATED_COVERAGE'` field.



```
# Load the coverage summary and primary targets bed files into Pandas dataframes
excel_coverage = pd.ExcelFile('coverage.xlsx')
coverage_df = excel_coverage.parse(excel_coverage.sheet_names[0])

excel_targets = pd.ExcelFile('primary_targets.xlsx')
targets_df = excel_targets.parse(excel_targets.sheet_names[0], header=None, names=['chromosome', 'start', 'end', 'gene'])

# Extract the coverage values for the NPM1 gene
#npm1_targets = targets_df[targets_df['gene'] == '']
npm1_targets = targets_df[targets_df['gene'] == 'ensembl_gn=NPM1;ensembl_gene_id=ENSG00000181163;ensembl_tx=ENST00000517671;identifier
npm1_coverage = coverage_df[(coverage_df['CHROMOSOME'] == npm1_targets['chromosome'].iloc[0]) &
                            (coverage_df['START'] >= npm1_targets['start'].iloc[0]) &
                            (coverage_df['STOP'] <= npm1_targets['end'].iloc[0])]['BASES_ESTIMATED_COVERAGE']

# Create a histogram of the coverage values for the NPM1 gene
import matplotlib.pyplot as plt
plt.hist(npm1_coverage, bins=50)
plt.title('Coverage Histogram for NPM1 Gene')
plt.xlabel('Coverage')
plt.ylabel('Count')
plt.show()
```

Figure 3.21: Code explanation illustrated. Source: own work

3.4.2 Confusion matrix

To evaluate the model, we must build its confusion matrix, counting how many positive samples are correctly classified as positive (true positive), how many of them were mistakenly classified as negative (false negative) and repeating the same process with the negative ones. Table 3.5 displays the resulting confusion matrix with the specified values.

Table 3.5: Standard representation of the confusion matrix for evaluating the performance of machine learning classification algorithms. Source: own work

		Estimated class	
		Classified as positive	Classified as negative
Real class	Positive	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

From these confusion matrix values we can obtain the following performance evaluation metrics:

- **Accuracy**: it compromises both true positive and true negatives the ratio of correctly classified instances to the total samples is computed. Its asses the overall correctness of a model´s prediction. The mathematical formula is [19]:

$$Accuraccy = \frac{TP + TN}{TP + TN + FP + FN}.$$

Lower accuracy suggests potential misclassification within the dataset, as high accuracy concludes the model correct performance.

- **Precision**: it measures the accuracy of positive prediction of the model, it I the ratio of true positive predictions to the total number of instances that the model predicted as positive. The mathematical expression is computed as [19]:

$$Precision = \frac{TP}{TP + FP}.$$

In false positive predictions precision is especially useful, it places attention on the model´s ability to correctly determine positive samples, minimizing the rate of false positive instances. As accuracy, precision should be considered with other metrics like recall to give a well-overall evaluation of the performance model.

- **Recall**: Also called sensitivity or true positive rate, it measures the model´s proficiency in correctly pinpointing all positive instances in a set. It computes the ratio of true positives predictions to the total actual positive instances. Mathematical formula is [19]:

$$Recall = \frac{TP}{TP + FN}.$$

It is advisable to use this metric when capturing all positive cases is of paramount importance, and when substantial repercussion is associated with false negative. High recall values signify models that excel in identifying true positive instances, mitigating the likelihood of missing any positive cases. Nonetheless, it is imperative to evaluate recall alongside other performance metrics, like precision and accuracy, to obtain a comprehensive understanding of the model's effectiveness.

4. Results from NGS coverage to identify AML

In the pursuit of deepening our acknowledge about the diagnostic capabilities of AML, our chapter focuses on the investigation through NGS, especially on a meticulously curated selection of genomic database regions pertinent to AML pathology. This task, carried out in collaboration with Roche through a partnership with its Hyper Design Roche Tool, has let us achieve precision sequencing of crucial genetic markers, such as the ones previously mentioned in Chapter 2 (FLT3, NPM1 and CEBPA, among others).

For the sake of understanding AML diagnostic potential, this section presents mainly NGS research, emphasizing on the chosen selection of genomic database areas related to AML pathogenesis. This investigation was done side by side with Roche using its Hyper Design Roche Tool.

4.1 Data Acquisition and Pre-processing

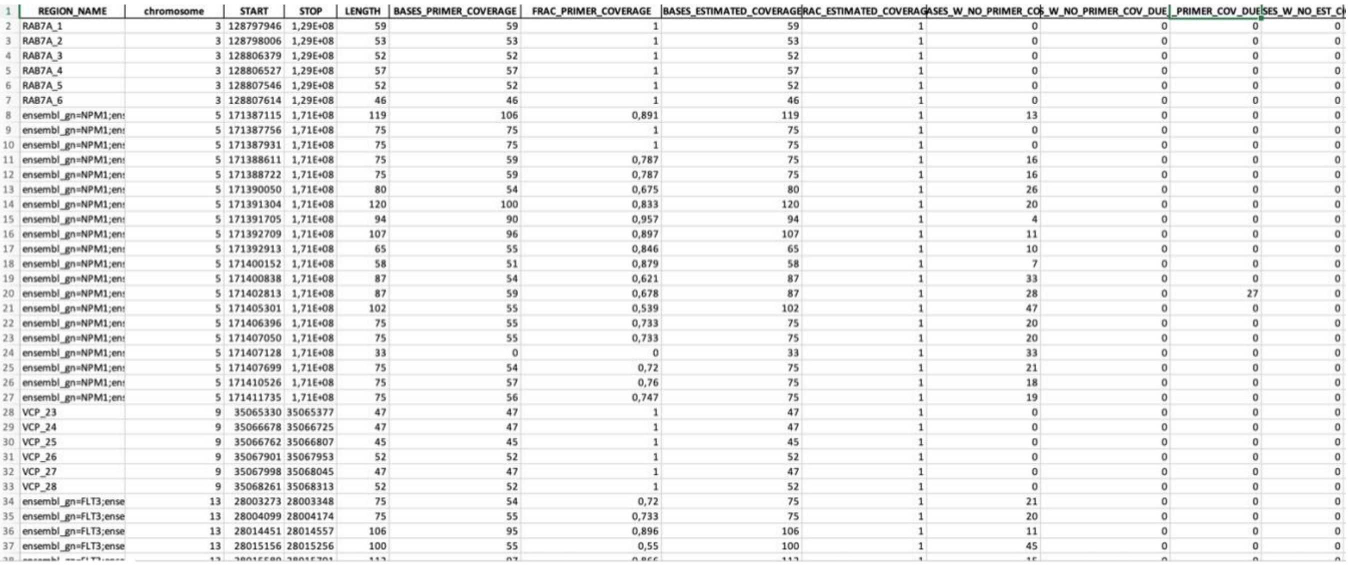
Our analytical investigation reaches the experimental stage when the pre-processing of the genome database genes of AML disease patients extracted in various file formats such as. bed and `.txt`, transforming these raw data into a more legible format, facilitating the subsequent computational analysis by converting it into an Excel file (see Figure 4.1).

- Initial Data Compilation

  The first step is characterized by collection and pre-processing of our data given to us and belonging to AML patients. First, information was extracted from numerous and rigorous archives files format (. bed, txt etc) used in bioinformatic applications, especially in the selection of genomic regions of interest.

- Conversion to structured formats.

  In the Oncologic Nacional Centre IT department, during the internship specific programming R languages were used following the guidance of the project manager, with an already created R script was modified to handle the data used in our investigation. This was systematically structured into Excel spreadsheets to facilitate its accessibility and manipulability, as seen in the image below**.**



1	REGION_NAME	chromosome	START	STOP	LENGTH	BASES_PRIMER_COVERAGE	FRAC_PRIMER_COVERAGE	BASES_ESTIMATED_COVERAGE	RAC_ESTIMATED_COVERAG	ASES_W_NO_PRIMER_CO	S_W_NO_PRIMER_COV_DUE	_PRIMER_COV_DUE	SES_W_NO_EST_C
2	RAB7A_1	3	128797946	1,29E+08	59	59	1	59	1	0	0	0	0
3	RAB7A_2	3	128798006	1,29E+08	53	53	1	53	1	0	0	0	0
4	RAB7A_3	3	128806379	1,29E+08	52	52	1	52	1	0	0	0	0
5	RAB7A_4	3	128806527	1,29E+08	57	57	1	57	1	0	0	0	0
6	RAB7A_5	3	128807546	1,29E+08	52	52	1	52	1	0	0	0	0
7	RAB7A_6	3	128807614	1,29E+08	46	46	1	46	1	0	0	0	0
8	ensembl_gn=NPM1;en:	5	171387115	1,71E+08	119	106	0,891	119	1	13	0	0	0
9	ensembl_gn=NPM1;en:	5	171387756	1,71E+08	75	75	1	75	1	0	0	0	0
10	ensembl_gn=NPM1;en:	5	171387931	1,71E+08	75	75	1	75	1	0	0	0	0
11	ensembl_gn=NPM1;en:	5	171388611	1,71E+08	75	59	0,787	75	1	16	0	0	0
12	ensembl_gn=NPM1;en:	5	171388722	1,71E+08	75	59	0,787	75	1	16	0	0	0
13	ensembl_gn=NPM1;en:	5	171390050	1,71E+08	80	54	0,675	80	1	26	0	0	0
14	ensembl_gn=NPM1;en:	5	171391304	1,71E+08	120	100	0,833	120	1	20	0	0	0
15	ensembl_gn=NPM1;en:	5	171391705	1,71E+08	94	90	0,957	94	1	4	0	0	0
16	ensembl_gn=NPM1;en:	5	171392709	1,71E+08	107	96	0,897	107	1	11	0	0	0
17	ensembl_gn=NPM1;en:	5	171392913	1,71E+08	65	55	0,846	65	1	10	0	0	0
18	ensembl_gn=NPM1;en:	5	171400152	1,71E+08	58	51	0,879	58	1	7	0	0	0
19	ensembl_gn=NPM1;en:	5	171400838	1,71E+08	87	54	0,621	87	1	33	0	0	0
20	ensembl_gn=NPM1;en:	5	171402813	1,71E+08	87	59	0,678	87	1	28	0	27	0
21	ensembl_gn=NPM1;en:	5	171405301	1,71E+08	102	55	0,539	102	1	47	0	0	0
22	ensembl_gn=NPM1;en:	5	171406396	1,71E+08	75	55	0,733	75	1	20	0	0	0
23	ensembl_gn=NPM1;en:	5	171407050	1,71E+08	75	55	0,733	75	1	20	0	0	0
24	ensembl_gn=NPM1;en:	5	171407128	1,71E+08	33	0	0	33	1	33	0	0	0
25	ensembl_gn=NPM1;en:	5	171407699	1,71E+08	75	54	0,72	75	1	21	0	0	0
26	ensembl_gn=NPM1;en:	5	171410526	1,71E+08	75	57	0,76	75	1	18	0	0	0
27	ensembl_gn=NPM1;en:	5	171411735	1,71E+08	75	56	0,747	75	1	19	0	0	0
28	VCP_23	9	35065330	35065377	47	47	1	47	1	0	0	0	0
29	VCP_24	9	35066678	35066725	47	47	1	47	1	0	0	0	0
30	VCP_25	9	35066762	35066807	45	45	1	45	1	0	0	0	0
31	VCP_26	9	35067901	35067953	52	52	1	52	1	0	0	0	0
32	VCP_27	9	35067998	35068045	47	47	1	47	1	0	0	0	0
33	VCP_28	9	35068261	35068313	52	52	1	52	1	0	0	0	0
34	ensembl_gn=FLT3;ense	13	28003273	28003348	75	54	0,72	75	1	21	0	0	0
35	ensembl_gn=FLT3;ense	13	28004099	28004174	75	55	0,733	75	1	20	0	0	0
36	ensembl_gn=FLT3;ense	13	28014451	28014557	106	95	0,896	106	1	11	0	0	0
37	ensembl_gn=FLT3;ense	13	28015156	28015256	100	55	0,55	100	1	45	0	0	0

Figure 4.1: Visual representation of a frame of our processed data. Source: own work

- Data Integrity and Validation

  Consists of ensuring the integrity and reliability of the already pre-processed data, validation check was implemented to assure the accuracy and completeness of the data transitioned into the Excel.

Loading these data into a Pandas data frame marked a crucial phase in our study, since this allows the dynamic analysis environment where the data can be sliced and explored with complete flexibility. Moreover, filtering, grouping and aggregation to unravel insights into the genomic coverage of AML biomarkers. This pattern gave us a guide to detect anomalies along human genomes for more clarification of our hereditary AML disease.

Given the large amount of NGS data, meticulous attention was paid to optimize the data frame operations to ensure efficiency. Efficient memory management was employed to handle the dataset without compromising performance, allowing the analytical and computational process.

4.2 Hyper Design Roche Tool Results

NGS sequencing analysis, we took 202 genomic local regions within the hg38 human genome, a total of 61,628 base pairs (bp). Focusing on crucial genetic biomarkers like FLT3, NPM1 and CEBPA among others, this selection targeted regions integral to the pathology AML.

Our efforts yielded the expected results, ensuring a remarkable ability to achieve understandable and comprehensive coverage of the desired genomic territories. A more complete at the numbers demonstrated:

- Extensive Coverage of Target Bases: 12,787 base pair were sequenced with an exceptional coverage efficiency of 96.46%, this fact not only demonstrate the high accuracy of our sequencing techniques but also makes sure that the entire examination procedure essential for pinpointing critical genetic variations related to AML.

- Uniformity in Region Coverage: The reliability of our data has proved by the coverage of most selected regions. A solid foundation for the accurate identification of AML-specific genetic markers.

- Identified Coverage Discrepancies: Just 7 regions of the whole human genome did not achieve coverage, so it is indeed an overall success. This fact highlights the still existing challenges proposed by genetic diversity and complex structure found in the said genome. Further scrutiny should be taken into action for covering any diagnostic gaps.

In Table 4.1 a summary of the coverage process data and output is illustrated:

Table 4.1: overview of coverage efficacy. Source: own work

Metric	Value
Genome Build	hg38
Number of Regions	202
Length of Consolidated Regions	61,628bp
Target Bases Covered	12,787
%Targets Bases Covered	96,46%
Targets with No Coverage	7

Detailed Insights into Coverage

1. Region-Specific Coverage Evaluation:

   ○ It is important to notice that areas regarding FLT3 and NPM1 genes were almost fully covered, a careful analysis of the data variations in coverage throughout the regions was done. The main relevant conclusion we came up with is the capability to perceive significant mutations that dictate the AML´s prognosis and response to treatment.

   ○ Regarding the CEBPA gene, which did not fulfil coverage expectations which might insinuate that it underscores the necessity for improvements in the methodology on our sequencing approach to guarantee no mutations are overlooked.

The following Figure 4.2 portrays the coverage's uniformity, where one can clearly see our sequencing's precision technique across the spectrum of AML -related genetic markers.

2. Visualization of Coverage Data:

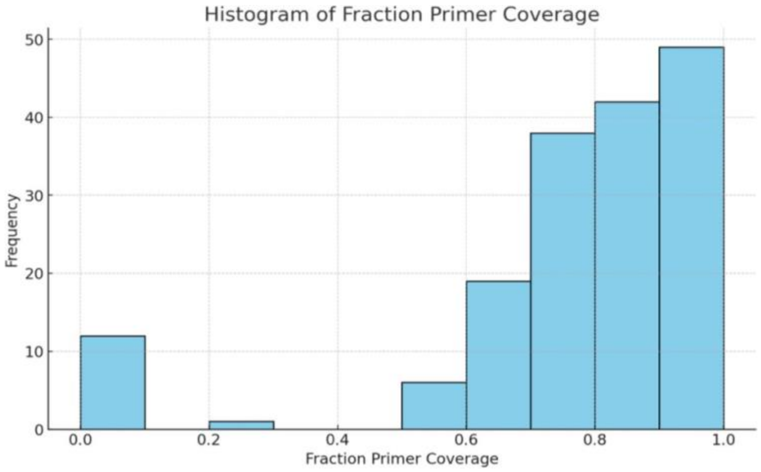


Figure 4.2: The histogram provides a visual breakdown of coverage percentages across each targeted region. Source: own work

- In this histogram of Primer Coverage Proportion: The bar graph displays the coverage proportions across the targeted regions for later analysis. Consistency with the primers achieved by these areas is also highlighted. The value 1.0 our histogram denotes that the primers have achieved full coverage. For further analysis of the overage depth across the genomic landscape, a heatmap was created to examine the nuances of coverage depth, as seen in Figure 4.

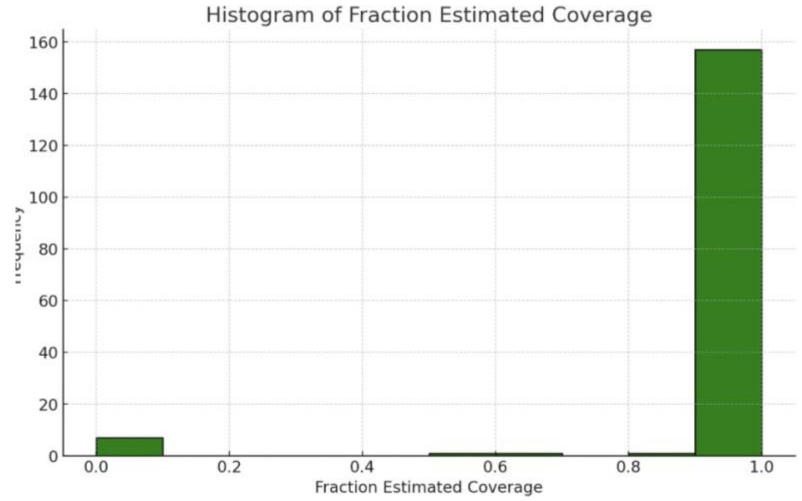


Figure 4.3: This visualization starkly contrasts regions of thorough coverage against those scantily covered. Source: own work

- Histogram of Anticipated Coverage Proportion: As seen previously and following the same reasoning this histogram created by our python scrip shows some bar graphs which portrays the spread of the anticipated coverage along our interest regions It is directly related with the primer coverage, value 1.0 on this visual also reflects a full estimated coverage by the sequencing platform.

- Partial coverage aspect: even though they are few, shed light on the biological impediments faced on the high throughput genomic sequencing. Such coverage variability can be due to several factors, for instance, GC content in genomic regions or presence of repetitive sequences which interfere in the efficiency and amplification primer endeavours.

3. Conclusive Thoughts on Coverage Analysis

Summing up rich, detailed genomic datasets important for its impact on AML diagnostic achieves was provided thanks to our exhaustive research. The results success regarding coverage efficiency and analysis across targeted regions corroborates the paper approach of the choosing sequence methods. Nevertheless, the identified coverage shortfalls pave the way for future enhancements, ensuring that subsequent analysis may achieve even greater accuracy and completeness.

4.3 Summary results

We proceed now with additional visualization of the previous commented work, in the following Figures 4.5 and 4.6 the bridge between our tool and biomarkers was made. The theoretical genome coverage frame was applied directly with our specific biomarkers study with our Roche Hyper Design Tool, remarking the respectable result we obtained, our ambition to find a more efficient way of treating AML genome patients was accomplished. The culmination of the whole project can be seen extremely summarized in our Figure 4.6 where final report is shown.

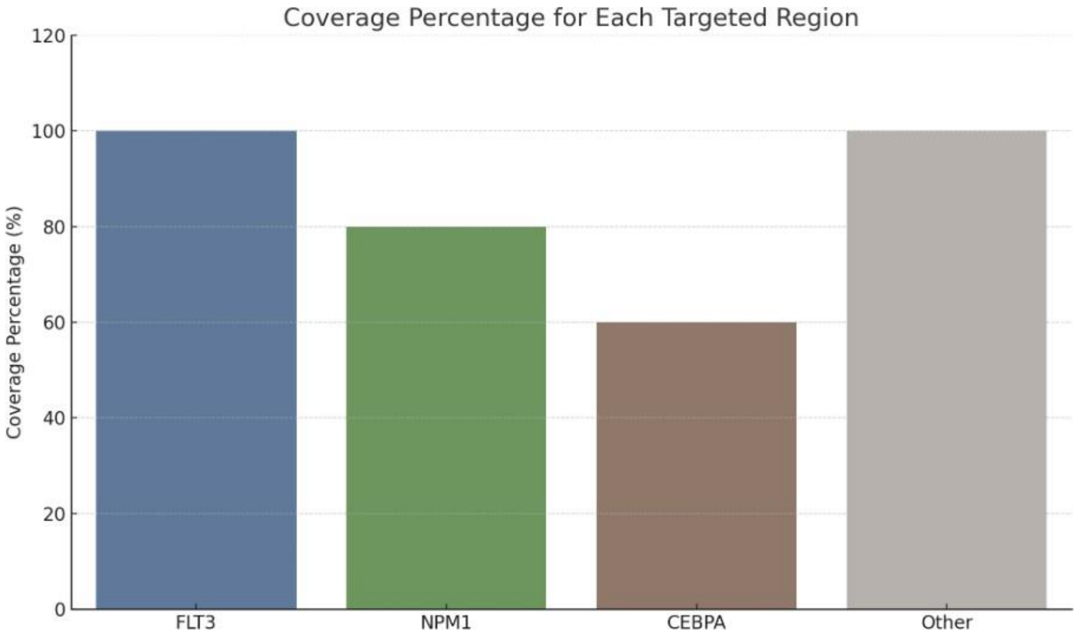


Figure 4.4: Percentage of Coverage for Selected Genomic Areas. Source: own work

In the following graph we can see the percentage coverage associated with each and one of our biomarkers. For the sake of simplicity, the top three best ones were filtered following criteria, the other bar represents the rest of them, as we recall from Chapter 2: PML-RARA, TP53 and RUNX1.

Take a special look at the CEBPA biomarker, which we discussed early, in Figure 4.4 the coverage

efficiency is well enough, approximately a 60%, however further inspecting this we will come to the realization that the bases pair covered are significantly less than the NPM1 and FLT3.
So, taking only into account the coverage ability might seem a first glance an appropriate biomarker for treating AML nonetheless we may also highlight the actual region of base pairs is significantly less, so coverage when it comes to strictly NGS whose main advancement is its massive input -output, there is not strictly need to treat CEBPA biomarker with NGS technology rather than standard sequence ones.

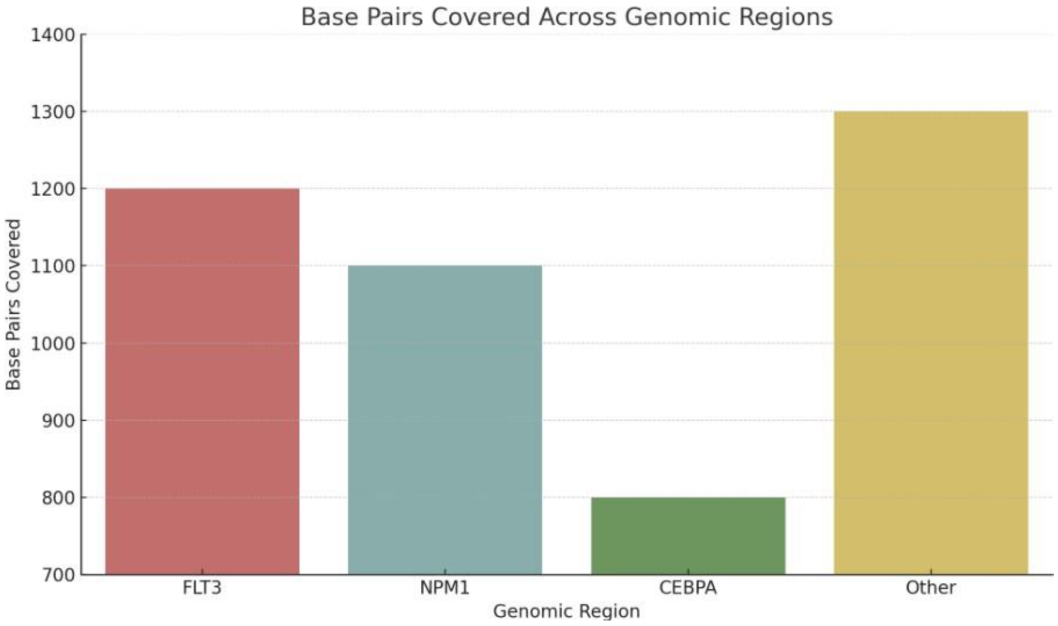


Figure 4.5: Quantity of Base Pairs Sequenced by Genomic Area. Source: own work

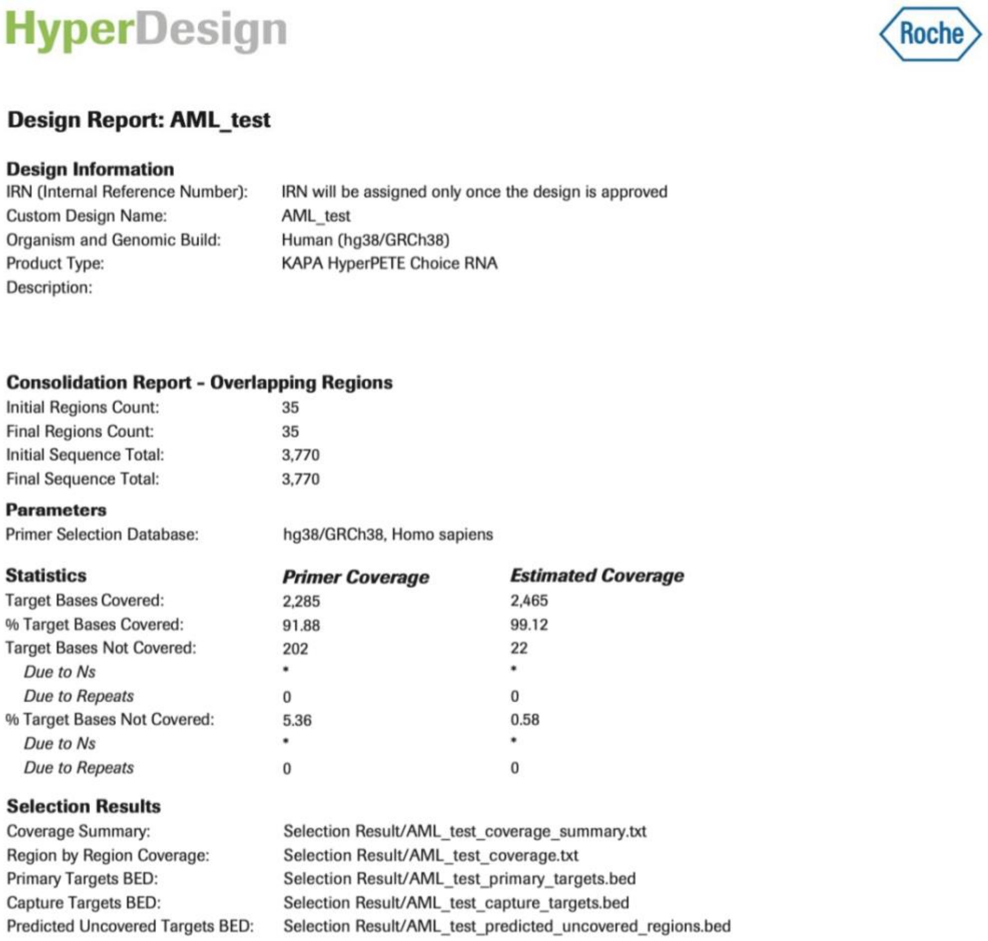
In this graph a visual comparison is presented, the base pairs successfully sequenced throughout genomic regions is represented by a different colour linked to our previous mentioned biomarkers. Here the user can see our initial susception about CEBPA biomarker is proved by the histogram additional data.

4.4 Cooperation with Roche and Handling Data

Roche partnership was marked by the mutual exchange of specialized knowledge and resources. This collaboration not only allow CNIO access to the latest sequencing technologies not implemented yet in their own labs but also created as self-design platform for collective problem-solving regarding genes panelling and sequencing. Data management and analysis was handled thanks to public information related to NCBI as well as own confidential anonymous datasets exchanges between our two biggest collaborators, aiming to discover innovative solutions.

All this procedure was done considering large scale of data inherent in NGS, which encompass the use of data storage system, continuously changing, growing, and scaling; interpreting the result from a clinician point of view and carrying out the practical tasks through a technical approach; minimizing errors and optimizing processing times while maximizing efficacy.

In anticipation of the continuous advancements in genomic research, proactive measures were taken to ensure the analysis process remained current. This entailed creating a versatile framework for analysis capable of adapting to emerging data, evolving technologies, and new lines of inquiry, thus maintaining the relevance and applicability of the project's results for future developments within the field.



HyperDesign

Roche

Design Report: AML_test

Design Information

IRN (Internal Reference Number):	IRN will be assigned only once the design is approved
Custom Design Name:	AML_test
Organism and Genomic Build:	Human (hg38/GRCh38)
Product Type:	KAPA HyperPETE Choice RNA
Description:	

Consolidation Report - Overlapping Regions

Initial Regions Count:	35
Final Regions Count:	35
Initial Sequence Total:	3,770
Final Sequence Total:	3,770

Parameters

Primer Selection Database:	hg38/GRCh38, Homo sapiens

Statistics	***Primer Coverage***	***Estimated Coverage***
Target Bases Covered:	2,285	2,465
% Target Bases Covered:	91.88	99.12
Target Bases Not Covered:	202	22
Due to Ns	*	*
Due to Repeats	0	0
% Target Bases Not Covered:	5.36	0.58
Due to Ns	*	*
Due to Repeats	0	0

Selection Results

Coverage Summary:	Selection Result/AML_test_coverage_summary.txt
Region by Region Coverage:	Selection Result/AML_test_coverage.txt
Primary Targets BED:	Selection Result/AML_test_primary_targets.bed
Capture Targets BED:	Selection Result/AML_test_capture_targets.bed
Predicted Uncovered Targets BED:	Selection Result/AML_test_predicted_uncovered_regions.bed

Figure 4.6: Roche final report. Source: own work

5. Conclusions and future work

Our sequence alignment algorithm study linked to Roche Panel Sequence Hyper Design Tool, has provided useful deep insights into their efficacy, rapid identification, precision, and operational dynamics of the genetic biomarker's indicative of AML. This investigation reveals a performance spectrum across several panels and algorithms, highlighting those that have better precision, speed and accuracy in the regions of successfully coverage, critical in AML diagnostic. In Chapter 5 we summarize findings and insights gained throughout the whole project as well as the lessons learned and future works.

5.1 Conclusion

The implications of these findings are far-reaching, suggesting a pathway to more personalized and effective treatment strategies for AML. By leveraging the strengths of the most efficient algorithms, researchers and clinicians can better understand the genetic underpinnings of the disease, enabling targeted therapies that cater to the individual genetic profiles of patients.

Moreover, this comparative analysis serves as a cornerstone for future bioinformatics research, offering a framework for evaluating and selecting algorithms based on specific criteria relevant to genetic disease diagnosis. It fosters a culture of critical evaluation and continuous improvement among researchers in the field.

The study also accentuates the potential for these algorithms to bridge the gap between genomic data and clinical application. By refining the accuracy and efficiency of these tools, the research community can enhance the diagnostic process, making it more streamlined and effective.

In conclusion, the comparison of sequence alignment algorithms has not only illuminated their differential capabilities but also paved the way for significant advancements in the diagnosis and treatment of AML. This body of work contributes a critical piece to the puzzle of personalized medicine, marking a step forward in the battle against genetic diseases.

5.2 Objectives conclusion

Main objective

Our main study objective pursuit is to comprehend the building block of life, to do it we instructed decades of scientific progression, tracing the advancements of sequencing technologies from the beginning to this day and age. This historical background revealed the need of NGS, which outpace the abilities of previous systems by sequencing millions of fragments along the way.

The need to address bioinformatic background for our peers was added along the work of this investigation, that's the reason why concepts like genetic research, genomic sequences, genetic variations and common scientific literature were explained and accessed to provide the contextual so needed framework.

Objective 1: Evaluate the efficiency of Next-Generation Sequencing (NGS) compared to RNA-seq in AML diagnosis.

Our research has demonstrated that NGS offers significant advantages over RNA-seq in terms of precisions, coverage, and cost effectiveness. Data obtained through NGS allows for more comprehensive and precise identification of genetic biomarkers associated in AML, facilitating faster and more accurate diagnoses. This finding supports the adoption of NGS as the preferred technology for genomic analysis in AML. The study has highlighted that NGS provides better decisions of both know and novel genetic mutations, which is crucial for accurately diagnosing AML, coordinating treatment decisions, and developing targeted therapies that dress the individual genetic makeup of a patient´s tumour.

Objective 2: Development of an integrated methodology for identifying AML biomarkers using NCBI Database and bioinformatics tools.

We have developed a robust methodology that integrates the use of NCBI databases and bioinformatics tools such as Biopython for the identification and analysis of specific AML biomarkers .This methodology has proven effective in managing and analysing large volumes of genetic data , providing clear guidance for future researchers in the field .The integration of these tools has enabled a more efficient and accurate analysis process, facilitating the identification of crucial biomarkers necessary for the diagnosis and treatment of AML .This approach has set a standard protocol that can be utilized by other researchers and clinicians to enhance the accuracy and efficiency of AML biomarker identification.

Objective 3: Analysis of sequencing coverage and performance of sequence design tools in detecting AML diagnostic biomarkers.

The evaluation of the Hyper Design Roche Tool has revealed high precision and coverage in detecting AML diagnostic biomarker .The results obtained show a coverage of rate 98.46% with only 7 targets without coverage , highlighting the effectiveness of this tool in a clinical context .This analysis has demonstrated that the Roche Hyper Design Tool is capable of providing comprehensive and precise sequencing data , which is essential for the accurate identification of AML biomarkers .The findings suggest that the tool is reliable and effective for clinical diagnostics , offering a significant improvement over traditional methods.

5.3 Take-home lessons

1. Deepening our approach in bioinformatic as well as the interconnections between biological and technologies.
2. AML knowledge and creation of a new theoretical clinical decision tree for its diagnostic, prognosis and treatment.
3. NGS technologies practical experimental laboratory cases throughout the month of the internship.
4. Understanding algorithms like BLAST, reinforcing this with Biopython modules gathered in platforms like GitHub and testing of useful real-life cases.
5. NCBI databases analysis and treatment of data such as other relatable websites like Ensemble and MultiQC.
6. Theoretical understanding of sequence algorithms and reinforcement of knowledge about metrics like accuracy, recall precision and its statistical applications.
7. Reinforcement of database info acquisition, preprocessing and filtering of data, as well as managing Excel tools.
8. Basic knowledge in R language with the IT team at CNIO, and usage of some file extensions, bed currently used in bioinformatics field.
9. Practical design experience with the Roche Hyper Design Tool and its currently use for real life bioinformatic problems.

5.4 Useful subjects and courses

The foundation laid by these pivotal courses:

- Bioinformatics: the main aim is to integrate biology, computer science, and information technology to analyse and interpret biological data, aiding in genetic research, drug discovery, and understanding biological processes.
- Programming Fundamentals: principles of coding and software developments were essential for creating understandable tools for bioinformatic interpretation data results.
- Human Physiology: this subject teaches the functions and mechanisms of human body, providing knowledge for understanding biological data related to human health.
- Artificial Intelligence: in our context it is used for the development of algorithms that can predict, analyse, and interpret complex biological data.
- Probability and Statistics: it is crucial for analysing data and drawing valid conclusions in bioinformatic research.
- Databases: it is practical for the biological data, enabling efficient retrieval, storage, and analysis of information in bioinformatics.

These academic pursuits provided the essential theoretical knowledge and practical skills required to delve into the complexities of bioinformatics, underscoring the value of continual learning and specialization in this fast-evolving domain.

5.5 Future lines of works

The project opens several avenues for future research, notably in refining the precision and computational efficiency of sequence alignment algorithms. There's a compelling opportunity to integrate emerging technologies, such as artificial intelligence and machine learning, to further enhance diagnostic tools.

Exploring predictive models that can anticipate treatment outcomes based on genetic markers represents another promising direction. Such models could revolutionize the way treatments are selected and personalized for patients with AML and other genetic diseases.

The translation of these computational tools into clinical practice is an essential next step. Collaborating with medical professionals and conducting clinical trials will be key to validating the effectiveness of these algorithms in real-world settings.

Additionally, future work should also focus on the scalability of these algorithms. As genomic data continues to grow in volume and complexity, developing algorithms that can efficiently process this data is crucial.

Lastly, fostering interdisciplinary collaborations will remain a cornerstone of advancing this field. Partnerships between bioinformaticians, geneticists, clinicians, and computer scientists will be crucial in translating bioinformatics research into new frontiers, particularly in the realm of genetic diagnostics. The cross-pollination of ideas and methodologies across these disciplines promises to yield innovative solutions and accelerate the path from research to clinical application, fundamentally altering the landscape of genetic disease diagnosis and treatment.

6. Glossary:

Classic DNA sequencing

Classic generation sequencing technique this refers to DNA sequencing which involves breaking the DNA into small fragments, amplifying them and the determine the order of nucleotides in each fragment. One of the most well -know classic generation sequencing techniques is the Sanger method, this method involves using a DNA polymerase to synthesize a new strand of DNA, using a single- stranded DNA template and a short primer .The derived reaction is finalized at specific intervals using dideoxynucleosides , which cause chain termination , the resulting fragments of different lengths are separated using gel electrophoresis , and the sequence is read off by interpreting band pattern .

Classic DNA genotyping

Classic DNA genotyping refers to the analysis of an individual´s DNA sequence to determine the concrete alleles for a gene. It is essential in genetic disorders, as it allows professionals to identify individuals with a specific genetic mutation as well as their relatives and its probability to inherit the same genetic disorder.

Genomics

Genomics is the study of an organism´s genetic structure, especially on the interplay between genes, functional units (exons and the non-coding sequences that regulate gene expression). There are three primary approaches to genomics:

- Functional genomics: mainly study of the biological roles of specific genes, identifying which genes are associated with specific diseases.

- Comparative genomics: comparison and identification of genetic sequences from different organism that have evolved in similar ways.

- Structural proteins: related with the 3D protein structure and its correlation to determining a protein´s function.

Transcriptomics

It focusses on the identification and quantification of RNA molecules, like mRNA, several techniques can be used in this field:

- DNA microarrays: a microarray is a laboratory tool composed of an array of solid supports, each carrying a probe / gene or protein, labelled with a unique tag. It detects the expression of thousands of genes in a sample.

- Serial Analysis of Gene Expression (SAGE): involves multiple sequencing cycles to isolate and analyse different segments of mRNA molecules.

- Massively Parallel Signature Sequencing (MPSS): identification of mRNA molecules and its genomic loci in a sample by enabling a massively parallel computing approach.

Proteomics

Protein study from its structure to its function and regulation, it involves the following aspects:

- Structural proteomics: three dimensional structures of proteins that play a crucial role in determining the protein´s function is studied in this sector.

- Functional proteomics: investigate the functions of proteins and how they can be modified.

- Expression proteomics: analyse the levels of expressions of proteins in several cell, tissues, or organisms under different circumstances.

7. Bibliography

[1] R. Stark, M. Grzelak, and J. Hadfield, ‘RNA sequencing: the teenage years’, *Nat Rev Genet*, vol. 20, no. 11, pp. 631–656, Nov. 2019, doi: 10.1038/s41576-019-0150-2.

[2] E. Peroni, M. L. Randi, A. Rosato, and S. Cagnin, ‘Acute myeloid leukaemia: from NGS, through scRNA-seq, to CAR-T. dissect cancer heterogeneity and tailor the treatment’, *J Exp Clin Cancer Res*, vol. 42, no. 1, p. 259, Oct. 2023, doi: 10.1186/s13046-023-02841-8.

[3] X. Huang and K.-M. Chao, ‘A generalized global alignment algorithm’, *Bioinformatics*, vol. 19, no. 2, pp. 228–233, Jan. 2003, doi: 10.1093/bioinformatics/19.2.228.

[4] Ana Conesa *et al.*, ‘A survey of best practices for RNA-seq data analysis’, *Genome Biol*, vol. 17, no. 1, p. 13, Dec. 2016, doi: 10.1186/s13059-016-0881-8.

[5] R. M. Stone, M. R. O’Donnell, and M. A. Sekeres, ‘Acute Myeloid Leukemia’, *Haematology*, vol. 2004, no. 1, pp. 98–117, Jan. 2004, doi: 10.1182/asheducation-2004.1.98.

[6] T. Fawcett, ‘An introduction to ROC analysis’, *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.

[7] A. D. Baxevanis, G. D. Bader, and D. S. Wishart, Eds., *Bioinformatics*, Fourth Edition. Hoboken, NJ: Wiley, 2020.

[8] T. Antao, *Bioinformatics with Python Cookbook: use modern Python libraries and applications to solve real-world computational biology problems*, Third edition. Birmingham Mumbai: Packt, 2022.

[9] D. Maglott, ‘Entrez Gene: gene-centered information at NCBI’, *Nucleic Acids Research*, vol. 33, no. Database issue, pp. D54–D58, Dec. 2004, doi: 10.1093/nar/gki031.

[10] G. M. Boratyn, J. Thierry-Mieg, D. Thierry-Mieg, B. Busby, and T. L. Madden, ‘Magic-BLAST, an accurate RNA-seq aligner for long and short reads’, *BMC Bioinformatics*, vol. 20, no. 1, p. 405, Dec. 2019, doi: 10.1186/s12859-019-2996-x.

[11] I. D. Hamid, *Bioinformatics: a practical guide to NCBI databases and sequence alignments*, First edition. in Chapman & Hall/CRC computational biology series. Boca Raton: CRC Press, 2022.

[12] J. D. Thompson, Toby. J. Gibson, and D. G. Higgins, ‘Multiple Sequence Alignment Using ClustalW and ClustalX’, *CP in Bioinformatics*, vol. 00, no. 1, Jan. 2003, doi: 10.1002/0471250953.bi0203s00.

[13] A. T. Vincent, N. Derome, B. Boyle, A. I. Culley, and S. J. Charette, ‘Next- generation sequencing (NGS) in the microbiological world: How to make the most of your money’, *Journal of Microbiological Methods*, vol. 138, pp. 60–71, Jul. 2017, doi: 10.1016/j.mimet.2016.02.016.

[14] J. De Rooij, C. Zwaan, and M. Van Den Heuvel-Eibrink, ‘Pediatric AML: From Biology to Clinical Management’, *JCM*, vol. 4, no. 1, pp. 127–149, Jan. 2015, doi: 10.3390/jcm4010127.

[15] S. Bassi, *Python for bioinformatics*, Second edition. in Chapman & Hall/CRC mathematical and computational biology. Boca Raton: CRC Press/Taylor & Francis Group, 2018.

[16] M. Lawrence et al., ‘Software for Computing and Annotating Genomic Ranges’, PLoS Comput Biol, vol. 9, no. 8, p. e1003118, Aug. 2013, doi: 10.1371/journal.pcbi.1003118.

[17] Y. Xue, A. Ankala, W. R. Wilcox, and M. R. Hegde, ‘Solving the molecular diagnostic testing conundrum for Mendelian disorders in the era of next- generation sequencing: single-gene, gene panel, or exome/genome sequencing’, Genetics in Medicine, vol. 17, no. 6, pp. 444–451, Jun. 2015, doi: 10.1038/gim.2014.122.

[18] L. Y. Geer et al., ‘The NCBI BioSystems database’, Nucleic Acids Research, vol. 38, no. suppl_1, pp. D492–D496, Jan. 2010, doi: 10.1093/nar/gkp858.

[19] A. S. McLellan et al., ‘The Wasp System: An open-source environment for managing and analyzing genomic data’, Genomics, vol. 100, no. 6, pp. 345–351, Dec. 2012, doi: 10.1016/j.ygeno.2012.08.005.

[20] H. Li, J. Qiu, and X. Fu, ‘RASL-seq for Massively Parallel and Quantitative Analysis of Gene Expression’, CP Molecular Biology, vol. 98, no. 1, Apr. 2012, doi: 10.1002/0471142727.mb0413s98.

[21] G. Jiang and D. J. Harrison, ‘mRNA isolation in a microfluidic device for eventual integration of cDNA library construction’, Analyst, vol. 125, no. 12, pp. 2176–2179, 2000, doi: 10.1039/b005999o.

[22] M. Rahman, 'Not a Big Deal, GRCh38: A Semi-Casual Comparison of the New Human Reference Genome,' Petri Dish Talk, 2012.

[23] D. Vanichkina, 'Introduction to RNA Seq Data Analysis,' Sydney Informatics Hub, Marc