FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

# Identificação Patogénica em Células Humanas Cancerígenas

**Nuno Miguel de Albuquerque Martinho**


Mestrado Integrado em Engenharia Informática e Computação

Orientador: Rui Camacho (FEUP)

Co-orientador: Pedro Ferreira (IPATIMUP/I3S)

25 de Julho de 2016

# Identificação Patogénica em Células Humanas Cancerígenas

**Nuno Miguel de Albuquerque Martinho**

Mestrado Integrado em Engenharia Informática e Computação

Aprovado em provas públicas pelo Júri:

Presidente: Eugénio Oliveira

Arguente: Sérgio Matos

Vogal: Rui Camacho

25 de Julho de 2016

# Resumo

Cerca de 15% a 20% dos cancros em humanos são devidos a infeções virais. Estas infeções, por vezes, têm a sua origem patogénica nas células humanas. A presença de vírus e bactérias nas células humanas, como o Vírus do Papiloma Humano, a Hepatite B, entre outros, aumenta o risco e a probabilidade de contrair cancro. Estas bactérias/vírus são formadas a partir da tradução, nos ribossomas, das sequências mARN, originando proteínas virais.

O aumento de investimentos e de esforços na área da Bioinformática, mais especificamente, *"Computational Transcriptomics"* e *"Sequencing and Genotyping Technologies"*, pode ajudar no estudo das infeções e agentes externos na formação de cancros.

Atualmente, existem diversas ferramentas que ajudam na identificação de sequências não humanas presentes no ARN. Estas ferramentas permitem, a partir do mapeamento do ARN, diferenciar as sequências do genoma humano das bacteriais/virais. Estas ferramentas têm diferentes graus de eficácia, dependendo da amostra e da finalidade da análise. Contudo, esta diferenciação, de maneira isolada, pouco contribui para o estudo e para a identificação patogénica de sequências não humanas. Todo o processo de execução destas ferramentas tende a ser complexo e complicado para investigadores com parcos conhecimentos na área de Informática.

Assim, como possível solução, surge a necessidade de fazer uma plataforma que, de forma automática, realize o mapeamento e alinhamento do ARN com *datasets* bacteriais/virais. A solução proposta é uma plataforma acessível da Web que, com ligação a um pipeline, analise as amostras fornecidas aplicando sucessivas ferramentas já existentes. Com a comparação das sequências nos vários *datasets* obtêm-se estatísticas visualizadas de forma gráfica para uma mais fácil compreensão sobre o resultado do mapeamento da amostra. É possível verificar regiões com um elevado grau de confiança e regiões do genoma onde há mais sequências com genes ativos. A solução tem como objetivo ser uma ferramenta útil no estudo das infeções e agentes externos na formação de cancros.

Nuno Martinho

# Abstract

About 15% to 20% of human cancers are due to viral infections. These infections sometimes have pathogenic origin. The presence of bacterias and viruses in human cells, such as human Papillomavirus, Hepatitis B, among others, increases the risk for developing cancer. These bacterias/viruses are formed from translation, at the ribosome, of the mRNA sequences, resulting in viral proteins.

The investment and efforts in the area of Bioinformatics, specifically "Computational Transcriptomics"and "Sequencing and Genotyping Technologies", can help to understand the role of infections and external agents in the formation of cancers.

Currently, there are several tools that help in the identification of non-human sequences present in the RNA. These tools allow, from RNA mapping, to differentiate the sequences between the human genome and bacterial/viral infections. These tools have different degrees of effectiveness depending on the sample and the purpose of the analysis. However, this differentiation, in isolation, does not have a big impact to the study and identification of pathogenic non human sequences. The whole process of implementation of these tools tend to be complex and difficult for expert researchers with a low level knowledge in the area of Informatics.

A possible solution is to make a platform that automatises, does mapping and alignment process of the RNA with bacterial/viral infections datasets. We propose an online platform to analyze the samples by applying successive existing tools on the user's sample. This mapping has a large computational weight and consume many resources, and the processing time is proportional to the number of reads to map. With the comparison of the sequences against the various datasets we expect to obtain pathoghen statistics of the samples mapping in a graphic way, to provide a usefull and easy comprehension of the analysis. You can check genes with a high RPKM (Reads Per kilobase transcript of per Million mapped reads) and check regions of the genome where are a high number of active genes. The solution aims to be a useful tool in the study of infections and external agents in the formation of cancers.

Nuno Martinho

# Agradecimentos

Sendo que, a presente dissertação, representa o final de mais um ciclo da minha vida, não podia deixar de agradecer a todos os que contribuíram para a realização deste trabalho, bem como, a todos que me ajudaram ao longo destes cinco anos de curso.

Em primeiro lugar, o meu agradecimento à minha família, nomeadamente os meus pais, por todo o esforço, apoio e confiança ao longo da minha formação académica, mesmo nos meus momentos de ausência, que foram bastantes.

Agradeço ao meu orientador, Professor Rui Camacho, pela orientação e ajuda facultada ao longo da dissertação.

Estou também grato, ao Doutor Pedro Ferreira, pela a ajuda prestada em matérias que não eram do meu domínio.

Um obrigado especial à Associação de Estudantes da FEUP, da qual fiz parte durante quatro anos, e a todos os seus membros com o qual tive o prazer de trabalhar e de formar amizade.

Agradeço aos meus amigos e colegas, que me ajudaram ao longo destes cinco anos, pelo companheirismo e por toda a companhia nas noites mal dormidas a trabalhar.

Para terminar, um agradecimento especial aos amigos de longa data de Guimarães, que apesar de ter falhado certos momentos e por vezes estar mais ausente que presente, sempre foram um grupo com o qual pude contar para tudo, seja para uma ajuda no estudo ou simplesmente para um fim-de-semana de descontração.

A todos vocês, muito obrigado, Nuno Martinho

*"My mission in life is not merely to survive, but to thrive;*
*and to do so with some passion, some compassion,*
*some humor, and some style."*

Maya Angelou

# Conteúdo

CONTEÚDO

# Lista de Figuras

# Lista de Tabelas

# Abreviaturas e Símbolos

| | |
|---|---|
| ADN | Ácido Desoxirribonucleico |
| ARN | Ácido Ribonucleico |
| ARNm | Ácido Ribonucleico mensageiro |
| bp | *based pairs* - pares de base |
| CSS | *Cascading Style Sheets* |
| CSV | *Comma-Separated Values* |
| ENCODE | *Encyclopedia of DNA Elements* |
| FEUP | Faculdade de Engenharia da Universidade do Porto |
| FTP | *File Transfer Protocol* |
| GB | Gigabyte |
| GFF | *General Feature Format* |
| GTF | *General Transfer Format* |
| VIH | *Human Immunodeficiency Virus* |
| HTML5 | *Hyper Text Markup Language 5* |
| I3S | Instituto de Investigação e Inovação da Universidade do Porto |
| ICGC | *International Cancer Genome Consortium* |
| IPATIMUP | Instituto de Patologia e Imunologia Molecular da Universidade do Porto |
| JSON | *JavaScript Object Notation* |
| MB | Megabyte |
| NCBI | *National Center for Biotechnology Information* |
| NGS | *Next Generation Sequencing* |
| REST | *Representational State Transfer* |
| RPKM | *Reads Per Kilobase of transcript per Million mapped reads* |
| SOAP | *Simple Object Access Protocol* |
| SQL | *Structured Query Language* |
| SSL | *Structured Query Language* |
| tARN | Ácido Ribonucleico transportador |
| UCSC | *University of California, Santa Cruz* |
| UML | *Unified Modeling Language* |
| URL | *Uniform Resource Locator* |
| XML | *Extensible Markup Language* |

# Capítulo 1

# Introdução

A Bioinformática tem um peso cada vez mais relevante em estudos relacionados com o genoma humano. A Bioinformática combina Ciência de Computadores, Estatística, Matemática e Engenharia Informática para interpretar dados biológicos [Bio]. O aumento de investimentos e de esforços nesta área, mais especificamente, *"Computational Genomics"* e *"Sequencing and Genotyping Technologies"*, pode ajudar no estudo da influência de agentes agentes patogénicos associados à formação de cancros e na descoberta e aplicação de novos medicamentos. Além destes métodos, também o processamento de imagem e de sinal estão em constante evolução. Com o aumento das bases de dados genómicas existentes e o aumento do poder computacional, é possível analisar uma grande quantidade de dados biológicos e extrair informação relevante e estatisticamente significativa para a área de Biologia Molecular e Genómica, em particular [SIL07]. Atualmente, já existe uma grande quantidade de projetos e instituições a desenvolver ferramentas que possibilitam a recolha, análise, tratamento e armazenamento de informação, por exemplo, o NCBI (*National Center of Biotechonology Information*), o ICGC e o ENCODE da UCSC (*University of California, Santa Cruz*).

## Contexto

Embora exista no genoma de cada um de nós uma pequena quantidade de código viral, cerca de 15% a 20% dos cancros em humanos são devidos a infeções virais. Estas infeções, por vezes, têm a sua origem patogénica nas células humanas. A presença de vírus e bactérias nas células humanas aumenta o risco e a probabilidade de contrair cancro [NG-].

Aliando o desenvolvimento de ferramentas de análise de sequências genómicas com o aumento de dados disponíveis, é possível obter informação muito importante relacionada com a expressão genómica. Por vezes, estas ferramentas são demasiado complexas e acarretam um elevado custo computacional. Estas ferramentas têm como principal objetivo serem úteis no diagnóstico e no tratamento de pacientes.

## Projeto

Desde o nascimento da espécie humana que vários agentes agentes patogénicos estão codificados no nosso código genético [NG-]. Assim, o estudo da expressão génica humana pode ser bastante útil na prevenção, no diagnóstico e no tratamento de doenças. A análise, a sequenciação e a visualização do nosso código genético são possíveis devido ao *RNA-sequencing* [1]. A tradução e a representação do nosso genoma para uma estrutura computacional são feitas usando esta tecnologia (RNA-seq). [2].

Com esta dissertação desenvolveu-se uma plataforma para ajudar os investigadores no estudo do impacto de agentes agentes patogénicos na origem de cancro. A plataforma integra várias ferramentas já existentes, de forma a obter uma análise adequada de amostras de sequências de ARN a partir do mapeamento e alinhamento com genomas de referência. Esta plataforma foi desenvolvida para ser de uso fácil para os investigadores. A plataforma tem um portal Web como interface, tornando-a ubíqua, e um *back-end* de recursos computacionais poderosos.

A solução proposta é constituída essencialmente por quatro módulos:

- ***Pipeline* de Mapeamento e Alinhamento**

  Responsável pelo alinhamento e mapeamento da amostra contra os genomas de referência (Genoma Humano, Viral e Bacterial). Também fornece a possibilidade de conversão de ficheiros (FASTA, FASTQ, SAM, BAM) e indexação dos mesmos.

- **FTP *Server***

  Transfere os ficheiros das amostras a analisar para o pipeline e, posteriormente, devolve os ficheiros com o resultado do alinhamento e mapeamento de volta para o utilizador da plataforma.

- **Interface Web**

  Permite ao utilizador a criação de projetos e a criação de pedidos de análise de amostras. Responsável pela exposição explícita e objetiva dos resultados da análise efetuada pelo pipeline.

- **Base de Dados**

  Guarda toda a informação necessária ao funcionamento da plataforma. Além de guardar dados dos genomas de referência e das amostras e os resultados da sua análise, também guarda informações sobre os utilizadores e os projetos que estes vão criando.

## Motivação e Objetivos

Aliando o desenvolvimento de ferramentas de análise de sequências genómicas com o aumento de dados disponíveis, é possível obter informação muito importante relacionada com a expressão

[1] Sequenciação de ARN

[2] Outras tecnologias existem, como *Micro-Arrays*, mas o RNA-Seq é, atualmente, a mais utilizada

genómica. Esta plataforma é uma contribuição para solucionar a necessidade de resolver um problema complexo e altamente relevante da Biologia Molecular e Médica. Normalmente, o processo de análise de sequências genómicas é repetitivo, moroso e a informação tende a ser bastante dispersa, tendo os investigadores que a reunir manualmente. A plataforma desenvolvida na sequência desta dissertação permite simplificar e acelerar todo este processo, diminuindo as ações a executar por parte dos investigadores.

## Estrutura da Dissertação

Além do capítulo de introdução, esta dissertação é constituída por mais quatro capítulos. O Capítulo 2 introduz conceitos importantes tanto no domínio da Biologia Molecular como no das tecnologias informáticas utilizadas neste trabalho de dissertação. Aborda essencialmente conceitos básicos de biologia, tipos de ficheiros biológicos, ferramentas que permitem manipular esses ficheiros e tecnologias para o desenvolvimento de portais web e armazenamento de informação. No Capítulo 3, é descrita a solução proposta para o problema, bem como, com mais detalhes técnicos, a arquitetura da solução proposta. O Capítulo 4 permite, a partir de *case studies*, avaliar a solução e a utilidade da mesma. Por fim, no Capítulo 5, são apresentadas as conclusões desta dissertação e o trabalho que pode ser desenvolvido no futuro. Depois, em anexo, apresenta-se o guia de utilização da plataforma e as suas interfaces.

# Capítulo 2

# Conceitos Básicos de Biologia e das Tecnologias Usadas

## 2.1 Conceitos Básicos de Biologia

### Expressão genética

É no núcleo das nossas células, mais propriamente no ADN, que está guardada toda a informação genética de um organismo, informação essa que vem sendo alterada desde o início da existência da espécie humana, seja devido à constante evolução, à ação do meio ambiente e aos erros genómicos ou devido a mutações ao longo da nossa história causadas por fatores externos. O ADN é uma sequência de bases ligadas em dupla-hélice. Estas bases podem ser de quatro tipos: Adenina (A), Timina (T), Citosina (C) e Guanina (G). As diferentes combinações das bases de ADN levam à diversidade humana. É no ADN funcional que temos os nossos genes, os quais contêm a informação essencial à formação de proteínas. Acontece que o ADN está "guardado"nos cromossomas no núcleo da célula e as proteínas são "construídas"nos ribossomas, que, por sua vez, estão fora do núcleo da célula, o que provoca todo um processo complexo de tradução da informação contida no ADN.

Numa descrição simplificada, esquematizada na figura 2.1, o ADN é transcrito para um tipo de ARN chamado ARN mensageiro (isto ainda no núcleo da célula). O ARN mensageiro viaja até aos ribossomas (no citoplasma), onde o mARN é traduzido e usado para sintetizar as proteínas.

O ARN apresenta uma estrutura simples de cadeia única e é constituído pelas bases Adenina (A), Uracilo (U), Citosina (C) e Guanina (G). A transcrição de ADN em mARN (ARN mensageiro) é efetuada no núcleo das células. As cadeiras de ARN têm um tamanho bastante inferior às de ADN, pois já são cadeias maturadas. Esta diferença de tamanho deve-se ao fenómeno de *splicing* (remoção dos intrões e junção dos exões).[1] O ARN é a cadeia da qual irá ser feita a leitura de aminoácidos para posteriormente serem produzidas as proteínas expressadas e utilizadas pelo

[1] Os exões são a parte de ADN para codificar proteínas (regiões codificantes). Os intrões são regiões entre os exões e não são usados para codificar proteínas.

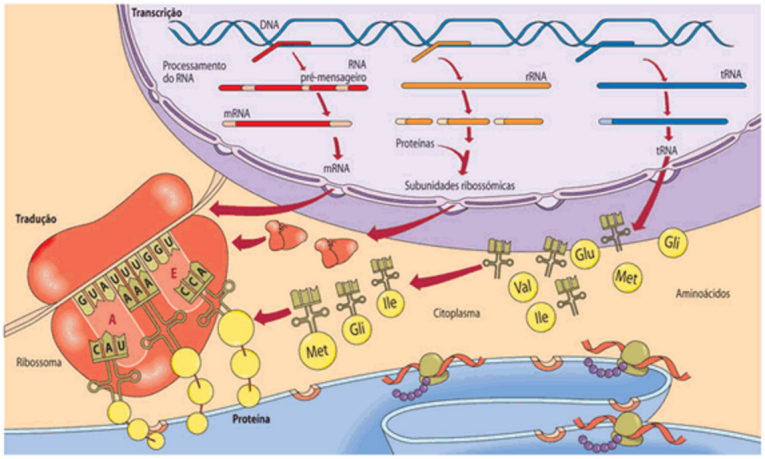


Figura 2.1: Transcrição de ADN e tradução de ARN em proteínas [DNA]

nosso organismo. Uma imagem do processo de "transcrição"e "tradução"pode ser visto na figura 2.1.

Esta fase de tradução ocorre nos ribossomas, para onde o tARN (também transcrito a partir do ARN) transporta e emparelha os aminoácidos com o seu respetivo par de base.

A fase de tradução, que ocorre nos ribossomas, é onde a cadeia de ARN é descodificada em aminoácidos, que, por sua vez, posteriormente, formam as proteínas. O mapeamento do código genético, que usa três pares de base para codificar um aminoácido, pode ser visto na figura 2.2

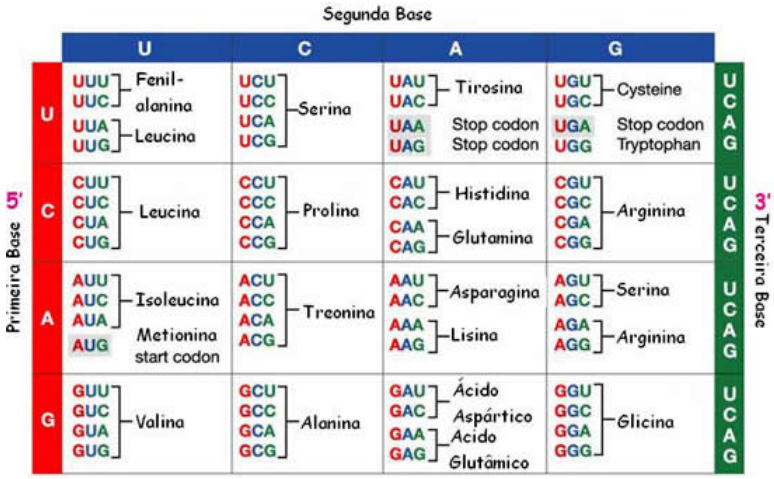


Figura 2.2: Código genético [Pro]

## ARN Viral

Todos os anos contraímos doenças a partir de vírus, sejam as mais comuns, como a gripe, ou as mais raras, como o Ébola. Geralmente, o nosso sistema imunitário combate-os do nosso corpo, mas, embora raras vezes, havendo mutações, estes vírus fundem-se com o nosso genoma, começando assim a ser parte do nosso legado genético ou permanecendo latentes nas células. Desde o

nascimento da espécie humana que temos ARN viral presente no nosso genoma [Vir]. Cerca de 8% do genoma humano é constituído por elementos patogénicos. São alguns destes elementos patogénicos presentes no nosso ADN que posteriormente, ao serem traduzidos e consequentemente expressos, podem ser agentes preponderantes na origem de cancro.

#### *RNA-sequencing*

As técnicas de sequenciação mais usadas atualmente são as técnicas NGS (*Next Generation Sequencing*). Estas técnicas são bastante populares em relação às anteriores, as técnicas de Sanger, pois conseguem produzir uma maior quantidade de dados em menos tempo, com menor custo e mais fiabilidade [Ill11]. A sequênciação de ARN tem como vantagem em relação aos *microarrays* de ADN não ter um limite máximo de quantificação [DDS+13], o que torna possível obter um maior número de sequências. Com a tecnologia *RNA-seq* podemos obter segmentos de ARN bastante precisos e exatos, havendo a possibilidade de obter segmentos bastante raros. Geralmente, estes segmentos são *short reads*, contendo entre 50 a 100 bp (pares de base).

#### *Assembly* de Segmentos de RNA-seq

Como referido, as técnicas de sequenciação NGS produzem *short-reads* - pequenos segmentos de ARN. Para se obter um genoma completo sem interrupções é necessário fazer a assemblagem dos *short-reads*, com o objetivo de ter a sequência completa de ADN. As duas principais técnicas de assemblagem são o *de-novo assembly* e o *genome reference assembly*. A principal diferença entre os dois é o uso de dados externos. O de *de-novo assembly*, ao contrário do *genome reference assembly*, não utiliza um genoma de referência, contudo é um método mais lento e com maior peso computacional. Ao não utilizar um genoma de referência, o *de-novo assembly* constrói a sequência final fazendo corresponder o fim de um *short-read* com o início do seguinte *short-read*, como demonstrado na figura 2.3. Em termos de resultado final, o *genome reference assembly* obtém, normalmente, sequências com mais informação disponível. Atualmente, em vários projetos, devido às vantagens/desvantagens de cada um, são usados processos híbridos, com os dois métodos a serem usados em simultâneo.

## 2.2 Formatos de dados Biológicos

### FASTA

O Formato FASTA representa, textualmente, sequências nucleotídicas ou peptídicas [FASb]. Atualmente, é um formato padrão no campo da Bioinformática. Cada sequência é representada em duas partes diferentes, como se pode verificar na tabela 2.1. A primeira linha (começa com o símbolo '>') é a descrição da sequência, contém o nome ou ID da sequência, comentários e a base de dados da sequência. As linhas seguintes representam, através de carateres únicos, bases

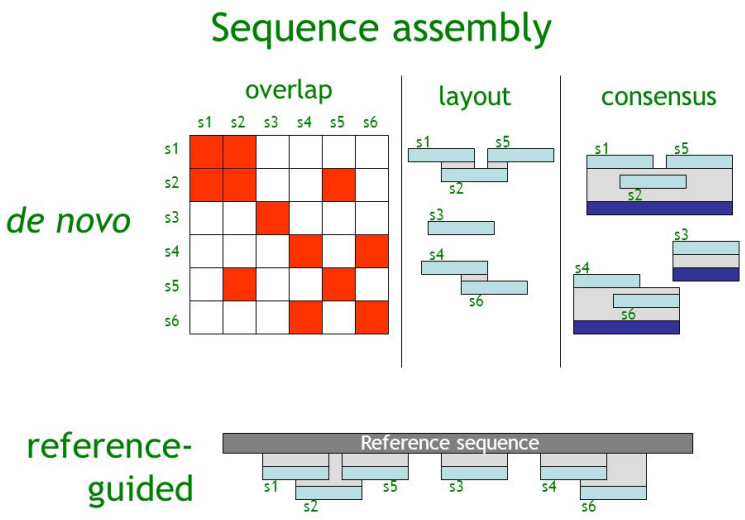


Figura 2.3: Diferença no método de *assembly* por *de-novo assembly* e *genome reference assembly [Seq]*

nucleotídicas ou peptídicas. A simplicidade da linguagem e da codificação permite facilmente a manipulação e a análise por diversas ferramentas [FASa].

| **Elemento** | **Exemplo** |
|---|---|
| Descrição e Anotação | >HSBGPG Human gene for bone gla protein (BGP) |
| Sequência | GGCAGATTCCCCCTAGACCCGCCCGCACCATGGTCAGGCA |

Tabela 2.1: Exemplo de formato FASTA

## FASTQ

O formato FASTQ é bastante semelhante ao FASTA, mas com mais informação disponível. Representa as sequências de um ficheiro FASTA e a qualidade das suas bases [CFG+09]. É usado em vez dos ficheiros em formato FASTA quando a qualidade das bases são importantes/relevantes, é o formato *standard* em NGS. A qualidade das bases pode ser calculada segundo várias fórmulas, PHRED ou SOLEXA. [Ill11]. Normalmente,as sequências são representadas por quatro linhas (ver tabela 2.2):

- Linha 1 - contém o identificador e a descrição da sequência;
- Linha 2 - representa as bases da sequência;
- Linha 3 - começa com o caráter '+' e, opcionalmente, tem o identificador da sequência;
- Linha 4 - codifica a qualidade (emparelhadamente) de cada um dos pares de base da sequência. Utiliza para a medida de qualidade os carateres ASCII, desde o ASCII-33 até ao ASCII-126 - sendo o '!' a pior qualidade e o "~"a melhor qualidade.

| Elemento | Exemplo |
|---|---|
| Identificação e Descrição | @EAS54_6_R1_2_1_413_324 |
| Sequência | CCCTTCTTGTCTTCAGCGTTTCTCC |
| Linha Separadora | + |
| Qualidade de cada par de base | ;;3;;;;;;;;;;;;7;;;;;;;88 |

Tabela 2.2: Exemplo de formato FASTQ com a descrição de cada linha

### SAM/BAM

O SAM [Li209] e o BAM (representação binária do SAM) guardam informações sobre o alinhamento ou mapeamento de sequências. Estes formatos suportam pequenas e longas sequências até 128Mbp. O formato SAM é essencialmente constituído por duas secções, a do cabeçalho e a de alinhamento, como se pode ver na Tabela 2.3. Cada linha de alinhamento é constituída por 11 elementos obrigatórios e outros elementos adicionais [SFW13]. O formato SAM foi criado para ser simples de trabalhar e flexível para ser manuseado por várias plataformas. O formato BAM, mais compacto que o SAM, permite indexação que, por sua vez, permite pesquisas mais rápidas.

Os dados que podemos obter de cada linha de um ficheiro SAM/BAM são os seguintes [2]:

- **Query Name (1)** - *read* em análise no alinhamento;
- **Flag (2)** - Flag com informação para o *parsing* do alinhamento;
- **Reference Name (3)** - genoma de referência usado para o alinhamento;
- **Position (4)** - posição do primeiro par de bases do alinhamento;
- **Mapping Quality (5)** - qualidade do alinhamento;
- **CIGAR (6)** - tipo de operação realizada;
- **Next Reference Name (7)** - nome do próximo segmento alinhado;
- **Next Reference Position (8)** - posição do próximo segmento alinhado;
- **Template Length (9)** - tamanho do segmento alinhado;
- **Segment Sequence (10)** - pares de base da sequência alinhada.

[2] a explicação detalhada de cada um dos campos pode ser analisada no manual *"Sequence Alignment/Map Format Specification"* [SFW13]

| Elemento | Exemplo |
|---|---|
| Header | @HD VN:1.5 SO:coordinate<br>@SQ SN:ref LN:45 |
| Alinhamento r001 | r001(1) 99(2) ref(3) 7(4) 30(5) 8M2I4M1(6) =(7) 37(8) 39(9) TTAGATAAAGG(10) * |
| Alinhamento r002 | r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA * |
| Alinhamento r003 | r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0; |
| Alinhamento r001 | r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1 |

Tabela 2.3: Exemplo Formato SAM

## GFF e GTF

O formato *Gene Transfer Format* (GTF) [GFFa] surgiu por refinamento do formato *General Feature Format* (GFF). Estes ficheiros, por norma, são usados para descrever genes e sequências de ARN e ADN. Em contexto biológico, o mais usado é o GTF [GFFb]. Cada linha deste tipo de ficheiro tem dez campos, oito deles iguais ao GFF. Este tipo de ficheiro permite saber a posição de cada gene no genoma da espécie em análise.

Cada linha de um ficheiro GTF, pode ser visto um exemplo na Tabela 2.4, tem o valor dos seguintes campos:

- **Seqname (1)** - nome da sequência em análise;
- **Source (2)** - nome do programa que gerou o ficheiro;
- **Feature (3)** - nome do tipo de *feature* em análise;
- **Start (4)** - posição de início da feature;
- **End (5)** - última posição da feature;
- **Gene ID (6)** - identificador único da origem do segmento genómico;
- **Transcript ID (7)** - identificador único do segmento genómico.

| |
|---|
| AB**(1)** Twinscan**(2)** CDS**(3)** 193817**(4)** 194022**(5)** . - 2 gene_id "AB.1"**(6)**; transcript_id "AB.1.2";**(7)** |
| AB Twinscan CDS 199645 199752 . - 2 gene_id "AB.1"; transcript_id "AB.1.2"; |
| AB Twinscan CDS 200369 200508 . - 1 gene_id "AB.1"; transcript_id "AB.1.2"; |

Tabela 2.4: Exemplo Formato GTF

## GenBank File

Os ficheiros disponíveis na base de dados do GenBank[3] contêm toda a informação sobre o genoma ou outras sequências de uma espécie. A extensão mais comum é ".gb"e são ficheiros

[3]http://www.ncbi.nlm.nih.gov/genbank/

de texto. No cabeçalho do ficheiro é possível obter informação geral sobre a espécie em análise: o nome, a versão do ficheiro, a origem, o organismo e publicações onde o gene foi citado [BKmL+07]. Após o cabeçalho, é possível obter informação detalhada sobre cada gene desse genoma, como: posição no genoma, nome e id do gene, sequência proteica que codifica. Nos ficheiros GenBank também é possível obter informações sobre os intrões e exões do genoma correspondente. Como campo opcional, após a definição de todos os genes, pode representar-se toda a sequência genómica em formato FASTA [Gen].

## 2.3 Software de análise de dados Biológicos

Atualmente, existe uma enorme quantidade de ferramentas para análise de sequências de ARN[4] . Entre as existentes, as ferramentas que a seguir se apresentam foram as escolhidas para se usar no trabalho desta dissertação. Estas ferramentas podem servir para a execução de várias tarefas.

### Alinhamento e Mapeamento

#### Bowtie

Bowtie é uma ferramenta ultra-rápida de alinhamento de *short reads* com um genoma de referência, desenvolvida pela John Hopkins University [5]. É bastante eficaz quando utilizada em alinhamentos de genomas com grandes quantidades de dados. Para o seu alinhamento usa o algoritmo *Burrows-Wheeler transform*, que permite uma baixa taxa de utilização de memória. Este algoritmo usa transformações de strings, para as comprimir, facilitando assima a sua pesquisa. Devido às sequências biológicas serem representadas em texto, é recorrente ver este algoritmo ser utilizado pelas várias ferramentas que analisam amostras e segmentos genéticos.

Além do módulo de alinhamento do Bowtie, outra potencialidade desta ferramenta usada na dissertação foi o **bowtie-build indexer**. Este módulo usa como *input* um ficheiro FASTA e gera ficheiros indexadores. Estes ficheiros gerados servem como referência para alinhamentos efetuados pelo TopHat. O processo de indexação, necessário às ferramentas de alinhamento, é extremamente pesado computacionalmente, mas apenas necessita de ser efetuado uma vez, podendo os ficheiros resultantes serem utilizados nos vário alinhamentos.

O Bowtie é interoperável com o SAMTools e o TopHat, através de ficheiros BAM. Atualmente, já existe o Bowtie2, que está otimizado para alinhamentos mais longos, em comparação com o Bowtie original [LTPS09].

#### TopHat

[4] https://en.wikipedia.org/wiki/List_of_ARN-Seq_bioinformatics_tools
[5] http://bowtie-bio.sourceforge.net/index.shtml

| Program | Read Length (bp) | CPU Time | Memory Footprint (MB) | Bowtie Speed-Up | Reads Aligned (%) |
|---|---|---|---|---|---|
| **Bowtie** | **50** | 7 m 11 s | 1,310 | - | 67.5 |
| **Bowtie2** | | 5 m 32 s | 1,138 | - | 56.2 |
| **Maq** | | 2h39m56 s | 804 | 21.8x | 67.9 |
| **SOAP** | | 48h42m4 s | 13,619 | 691x | 56.2 |
| **Bowtie** | **76** | 18 m 58 s | 1,323 | - | 44.5 |
| **Bowtie2** | | 7 m 35 s | 1,138 | - | 44.9 |
| **Maq** | | 4h45m7s | 1,155 | 14.9x | 31.7 |

Tabela 2.5: Comparação de desempenho da leitura de *reads* usando como amostra dois milhões de *reads* [LTPS09]

A ferramenta TopHat [6] diferencia-se das outras ferramentas por ter bastante sucesso na identificação de *splice* [TPS09] em cadeias de ARN na análise dos resultados de mapeamento de sequências. É uma ferramenta de alinhamento de *short reads* que usa como base o Bowtie. Esta ferramenta utiliza, como base para a execução das suas tarefas, genomas de referência que são usados para mapear e alinhar a amostra a estudar.

A tipo de análise efetuado pelo TopHat pode ser especificado através dos seus parâmetros que podem ser personalizados:

- **-N/read mismatches** - número máximo de reads de um segmento não mapeados;
- **-o** - diretório de *output* dos ficheiros resultantes da análise;
- **-i/-I** - comprimento mínimo/máximo dos intrões no segmento a analisar;
- **-solexa qual** - escala Solexa para a qualidade dos ficheiros FASTQ;
- **-p/num threads** - número de *threads* a utilizar no alinhamento.

Como ficheiros de *input*, esta ferramenta requer ficheiros indexados pelo Bowtie para os genomas de referência e ficheiros FASTA e FASTQ como as amostras a analisar.

Para esta dissertação, os ficheiros de *output*, resultantes de uma análise efetuada pelo TopHat, que foram usados são:

- **align_summary.txt** - permite ter um *overview* de todo o processo. Este ficheiro indica-nos o número de *reads* da amostra e qual a percentagem de *reads* mapeados.
- **accepted_hits.bam** - Ficheiro BAM com informações sobre as sequências da amostra mapeadas no genoma de referência em utilização.
- **unmapped.bam** - Ficheiro BAM com informações sobre as sequências da amostra não mapeadas no genoma de referência em utilização.

[6] https://ccb.jhu.edu/software/tophat/index.shtml

Analisando a tabela 2.6 é possível constatar o melhor desempenho alcançado pela combinação da ferramenta TopHat2 juntamente com a ferramenta Bowtie, quando comparado com outras ferramentas adequadas para o mesmo tipo de análise. Apesar da rapidez de alinhamento da ferramenta STAR [DDS+13] em relação às restantes, o TopHat, com a elevada percentagem de *reads* corretamente mapeados [KPT+13] e a maior documentação existente, torna-se uma ferramenta fiável e confiável para o desenvolvimento desta dissertação, contudo a ferramenta STAR é agora bastante utilizada pelos investigadores na área da Bioinformática.

| **Program** | **No. of mapped reads** | **Correctly mapped reads, %** | **Incorrectly mapped reads, %** | **Unmapped reads, %** |
|---|---|---|---|---|
| **TopHat2 + Bowtie1** | 19,826,638 | 98.31 | 0.82 | 0.87 |
| **TopHat2 + Bowtie2** | 19,826,673 | 98.03 | 1.10 | 0.87 |
| **TopHat1.14** | 19,616,874 | 94.64 | 3.45 | 1.91 |
| **GSNAP** | 19,997,255 | 94.21 | 5.77 | 0.02 |
| **RUM** | 19,555,823 | 88.11 | 9.67 | 2.22 |
| **MapSplice** | 19,872,372 | 97.28 | 2.08 | 0.64 |
| **STAR** | 19,087,508 | 92.14 | 3.30 | 4.56 |

Tabela 2.6: *Benchmark* de ferramentas de alinhamento e mapeamento genómico para um *dataset* com 20 milhões de segmentos com 100 pares de base, segundo o artigo *"TopHat2 : accurate alignment of transcriptomes in the presence of insertions , deletions and gene fusions"* [KPT+13]

## Análise Taxonómica

### Megan5

Por vezes, um elemento patogénico sofre pequenas mutações. Estas mutações nos pares de bases de uma sequência de ARN são detetadas pelos programas de alinhamento e mapeamento, classificando-as como diferentes agentes patogénicos. Contudo, estas sequências, apesar de serem diferentes, continuam a representar a mesma espécie, mas diferentes estirpes. Por exemplo, o vírus Influenza é um vírus que de ano para ano sofre mutações. Neste caso, as ferramentas analisadas anteriormente detetam elementos diferentes. Por isso, para uma quantificação correta, a ferramenta Megan5[7] ajuda na procura destes elementos, agrupando-os e classificando-os como sendo o mesmo elemento patogénico [HM11]. Na figura 2.4 é possível verificar o agrupamento das várias estirpes do mesmo elemento patogénico.

## Conversão e visualização de ficheiros

Para a conversão, manipulação e visualização das sequências foram escolhidas duas ferramentas: **SamTools** e **Tablet**. O SamTools permite a conversão entre ficheiros BAM e SAM e, ao esconder tarefas de baixo nível, torna-se bastante útil e importante no manuseamento de sequências guardadas em ficheiros com este formato. A ferramenta SamTools permite a visualização

[7] http://ab.inf.uni-tuebingen.de/software/megan5/

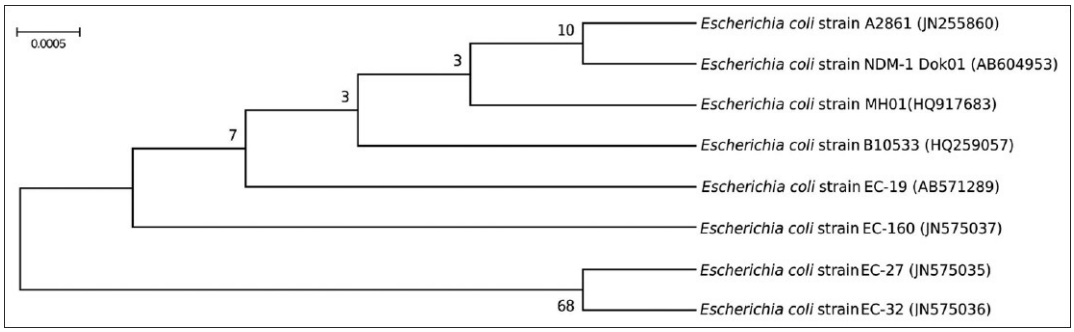


Figura 2.4: Exemplo de árvore taxonómica para a bactéria Eschrichia coli para o gene blaNDM-1 [BAH+13]. Os números nos ramos da árvore representam a percentagem de suporte de cada espécie em relação à amostra

textual de ficheiros de alinhamento e mapeamento SAM/BAM. As principais funções do SamTools usadas nesta dissertação foram:

- **samtools view** - conversão de SAM para BAM e vice-versa;
- **samtools index** - indexação de um ficheiro BAM.

**Tablet** [MSB+12] é um software de visualização gráfica que permite elevado desempenho na navegação e visualização de dados biológicos, como por exemplo a representação de um alinhamento de sequêñcias demonstrado na figura 2.5.Desenvolvido pelo The James Hutton Institute, o Tablet suporta todos os formatos analisados neste capítulo e permite a pesquisa por nome ou sequência dos dados em análise. Esta aplicação foi extremamente útil para visualizar os resultados das análises efetuadas neste trabalho de dissertação. A interface desta aplicação serviu como base da interface do utilizador na plataforma desenvolvida como consequência desta dissertação.

de

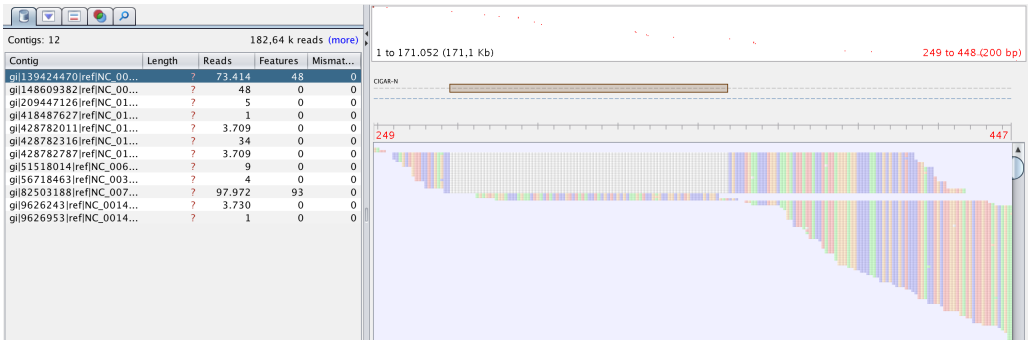


Figura 2.5: Exemplo da representação do alinhamento (ficheiro BAM) de uma amostra genómica (ficheiro FASTQ) com o genoma do *contig* "139424470"visualizado na ferramenta Tablet []

## 2.4 Dados Biológicos

Os dados escolhidos para a avaliação da plataforma foram amostras de células cancerígenas. Estas amostras foram recolhidas do ENCODE Project do USCS[8]. Estes dados foram usados para mapeamento e alinhamento contra genomas de referência. Assim, foi posssível a validação e verificação da plataforma recorrendo a dados reais. Os genomas de referência em uso são o genoma humano, o bacterial e o viral. Estes genomas foram recolhidos a partir da base de dados do National Center of Biotecnology Information (NCBI).

## 2.5 Tecnologias Informáticas

### 2.5.1 Web Services

Web Service é uma solução informática com o objetivo de fazer a comunicação e a ligação entre duas aplicações diferentes. Os *web services* trazem agilidade, eficácia e compatibilidade na interação entre plataformas diferentes[DB04]. Os Web Services permitem que os seus recursos estejam disponíveis em toda a rede para vários clientes, tornando-se interoperável [Web]. Os dois principais mecanismos de transporte em Web Services são SOAP e FTP. Geralmente, associado ao Web Service está uma interface para que o utilizador possa interagir com a base de dados e outras funcionalidades do Web Service, como a sua API. Um Web Service pode ser desenvolvido em qualquer linguagem de programação, mas ao usar formatos de ficheiros como intermediários, por exemplo, XML, CSV e JSON, torna-se universal.

### 2.5.2 Armazenamento de Informação

Atualmente, para armazenamento de informação, existem vários tipos de estruturas de base de dados, sendo as mais usadas SQL e no SQL. A principal diferença entre os dois tipos de base de dados está na estrutura e nos modelos de dados de cada uma. As base de dados SQL organizam-se por tabelas, estando os dados organizados de forma estrutural e rígida [SQL]. As bases de dados NoSQL são desenvolvidas em documentos, a estrutura dos seus dados é bastante dinâmica e mutável. Para esta dissertação foi escolhido o modelo SQL e, consequentemente, a linguagem MySQL, pois os dados da amostra, dos ficheiros de sequências de ARN e dos genomas de referência são bem conhecidos e estruturados.

## 2.6 Sumário

Neste capítulo, foram analisados alguns conhecimentos e software necessários ao desenvolvimento da dissertação. Os conceitos biológicos servem de base para a compreensão do problema da dissertação. Relativamente às ferramentas a usar, e aos ficheiros sobre os quais trabalham, é

[8] https://genome.ucsc.edu

bastante importante que sejam interoperáveis entre elas, usando o mesmo tipo de ficheiros como *input* e *output*.

# Capítulo 3

# Descrição da Solução

Neste capítulo está descrita a solução proposta como resposta ao problema encontrado.

Com essa finalidade, surge a plataforma **supRNA**, uma plataforma que procura otimizar todo o processo de alinhamento e mapeamento de amostras de ARN contra genomas de referência bem conhecidos. O grande objetivo desta plataforma é facilitar as ações a efetuar pelo investigador desde a introdução da amostra até à visualização dos resultados. O investigador tem o controlo de todo o fluxo de ficheiros nos projetos em que está inserido, tendo a possibilidade de os manusear e converter para aquilo que ache necessário. Os projetos poderão ser colaborativos, tendo vários investigadores acesso em simultâneo à análise de resultados. A plataforma supRNA, além de apresentar de forma gráfica e percetível, os resultados do alinhamento da amostra, oferece a possibilidade ao investigador de fazer *download* dos ficheiros resultantes da análise para poder usar outras ferramentas.

A plataforma supRNA tem como principais funcionalidades de destaque:

- Análise (*default* e personalizada) de amostras de células humanas
- Visualização gráfica dos resultados da análise
- Conversão de ficheiros com sequências genómicas
- Indexação de genomas de referência
- Criação de projetos colaborativos com outros invetigadores

## 3.1 Problema

Cerca de 8% do genoma humano são elementos patogénicos [GJ13] presentes no nosso código genético desde que nascemos. Atualmente, existe uma grande quantidade de software que permite estudos sobre o nosso genoma. Essas ferramentas são bastante complexas e especifícas quanto às suas funções, umas são mais eficazes a identificar *splicing*, outras, fusões e mutações. Além da eficácia, também o tempo computacional varia bastante entre elas.

Apesar de as ferramentas trabalharem o mesmo tipo de ficheiros, é necessário que estes sejam manuseados de aplicação para aplicação, o que pode levar a falhas e tornar-se tempo perdido durante a investigação. Sendo grande parte das ferramentas utilizadas através da linha de comandos, a falta de conhecimento do investigador na àrea de Informática pode comprometer a análise de sequências. Além de toda esta complexidade, por norma, o output final da análise em ficheiros que mostram o alinhamento de sequências é bastante complexo e de difícil compreensão para um investigador comum, tendo este que recorrer a programas de visualização de sequências e alinhamento genómico.

## 3.2 Solução

A solução proposta para solucionar o problema identificado passa por quatro módulos distintos, mas que se relacionam entre eles.

O primeiro módulo é o Web Service, que funcionará como um pipeline, onde as várias ferramentas anteriormente analisadas e descritas serão executadas de forma sequencial para analisar a amostra inserida pelo investigador.

O pipeline, além de analisar a amostra de sequências de ARN, irá funcionar como um filtro. O primeiro objetivo do pipeline é isolar os elementos bacteriais e virais, retirando da amostra as sequências humanas representativas. Isto será possível com o mapeamento e alinhamento da amostra contra o genoma humano de referência usando as ferramentas TopHat e Bowtie, como se pode ver na figura 3.1.

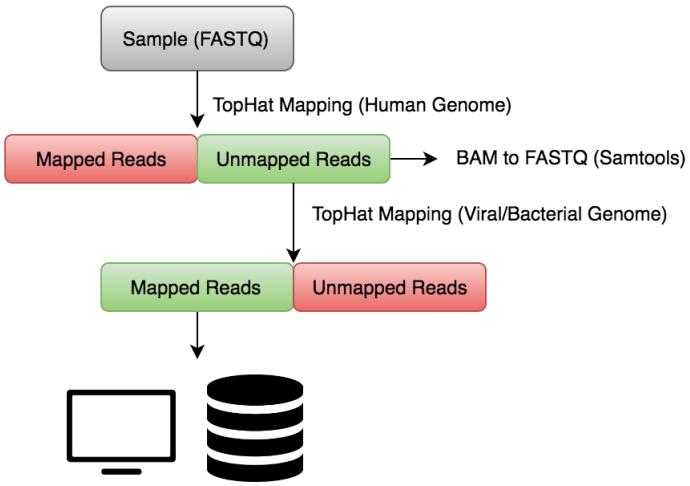


Figura 3.1: Diagrama representativo dos mapeamentos sucessivos efetuados pelo pipeline

Após o isolamento dos patogénicos, será feita a análise taxonómica e a quantificação de cada vírus ou bactéria. Esta quantificação é analisada, principalmente, segundo dois valores, o número de *reads* mapeados e o RPKM (*Reads Per Kilobase of transcript per Million mapped reads*) de

cada patogénico. O RPKM de cada patogénico é a relação entre os *reads* mapeados desse patogénico (C) multiplicado pela constante $10^9$ e o total de *reads* mapeados em toda a amostra (N) multiplicado pelo seu comprimento (total de *reads* - L). Estes dois valores permitem perceber a abundância de cada patogénico na amostra.

$$RPKM = \frac{10^9 \times C}{N \times L} \tag{3.1}$$

Os resultados da análise são enviados para a plataforma online para o investigador os poder analisar. Este pipeline também serve para conversão de ficheiros, que se poderão tornar úteis ao longo da investigação.

O segundo módulo é a plataforma online para uso remoto do pipeline. Esta plataforma é bastante simples e *user-friendly*, com o objetivo de o investigador visualizar os dados analisados e os resultados obtidos graficamente. O *website* faz a ligação entre o investigador e o *web service*, havendo troca de ficheiros entre eles. Nesta plataforma, o utilizador também tem a oportunidade de fazer o *download* dos ficheiros resultantes da análise para os poder utilizar com outras ferramentas.

O terceiro módulo é responsável pelo armazenamento de toda a informação necessária ao funcionamento de toda a plataforma. Utilizará uma Base de Dados para guardar os genomas de referência, a lista atualizada de patogénicos conhecidos, as amostras submetidas, os pedidos efetuados, os dados de experiências realizadas e os seus resultados.

O FTP Server é o quarto módulo desta solução. Este servidor gere as trocas de ficheiros entre o Web Service e a plataforma online, tanto num sentido como no outro.

Na figura 3.2 está representada a sequência de estados desde a inserção de uma amostra na plataforma até ao momento em que o investigador pode ver os resultados da análise. Os estados representados a azul são visíveis a partir da interface da plataforma e todas as ações são geridas pela Máquina ILP. Os estados a cinzento são ações que ocorrem no pipeline desenvolvido na Máquina Magalhaes03.GRID e não são estados visíveis ao utilizador. O processo descrito na figura 3.2 começa pelo *Login* na plataforma por parte do utilizador. Após o *Login*, o utilizador pode criar novos projetos ou aceder a projetos anteriores. Na página de projeto, além de poder converter ficheiros biológicos e genómicos, o utilizador pode fazer o *request* da análise de uma amostra de células humanas. Após a submissão da amostra, todo o processo é gerido na máquina Magalhaes03.GRID. É nesta fase que a amostra é filtrada e analisada no *pipeline*. Primeiro, são retirados todos os *reads* que mapeiam no genoma humano e, depois, em duas fases distintas, são agregados os *reads* que mapeiam no genoma viral e bacterial. Após esta sequência de filtros e mapeamentos executados pelas ferramentas SamTools, Bowtie e TopHat, os resultados são enviados de volta para a Máquina ILP, onde, a partir de gráficos e tabelas, a análise de resultados é exposta, podendo o investigador ver o mapeamento de cada patogénico segundo a quantidade de *reads* mapeada e o RPKM de cada patogénico.

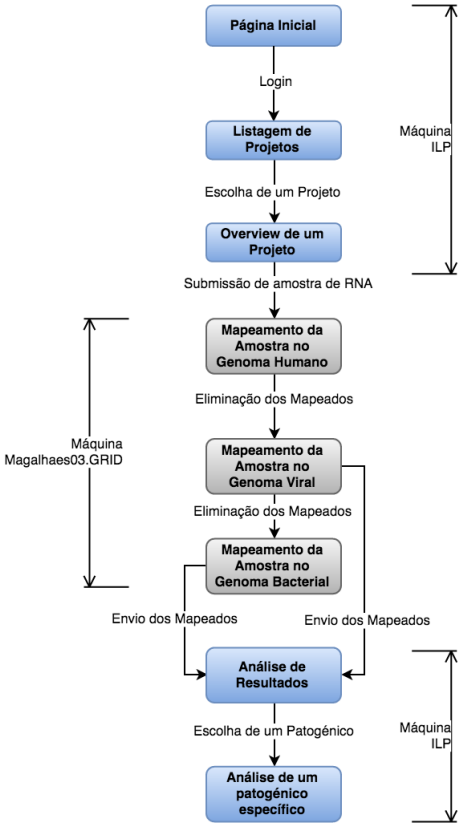


Figura 3.2: Workflow da solução para a análise de uma amostra (Distinção entre tarefas do servidor ILP e Magalhaes03.GRID)

## 3.3 Casos de Uso

A solução proposta apenas tem um ator a interagir com a plataforma, ou seja, o investigador que faz o pedido de análises de amostras de células humanas. Na figura 3.3 é possível ver as ações que o "Investigador"pode fazer na plataforma. Além das funcionalidades mais triviais, como fazer *Login*, *Logout*, registar-se e adicionar outros utilizadores a um projeto, o investigador pode converter ficheiros biológicos nos mais variados formatos e criar *Requests* para análise de ficheiros BAM/SAM ou FASTA/FASTQ. Após a análise dos ficheiros, também poderá ver os resultados da análise efetuada, seja ao nível geral dos patogénicos mapeados, como para cada patogénico em específico.

Na figura 3.4 estão representadas as trocas de mensagens do ator "Investigador"com o servidor e a troca de mensagens entre os dois servidores de modo a que o pedido do Investigador seja efetuado segundo o *workflow* representado na figura 3.2.

No anexo A está exemplificado um guião de utilizador de modo a analisar uma amostra genómica na plataforma desenvolvida. Este guião pretende mostrar todas as funcionalidades e todo o potencial da solução desenvolvida nesta dissertação.

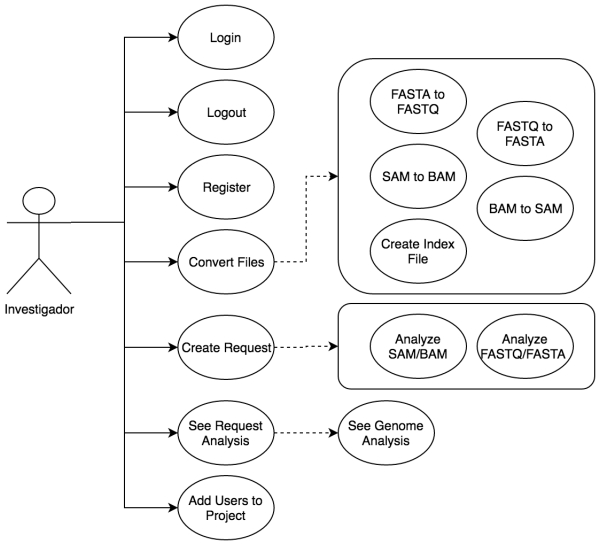


Figura 3.3: Diagrama de casos de uso para o ator Investigador

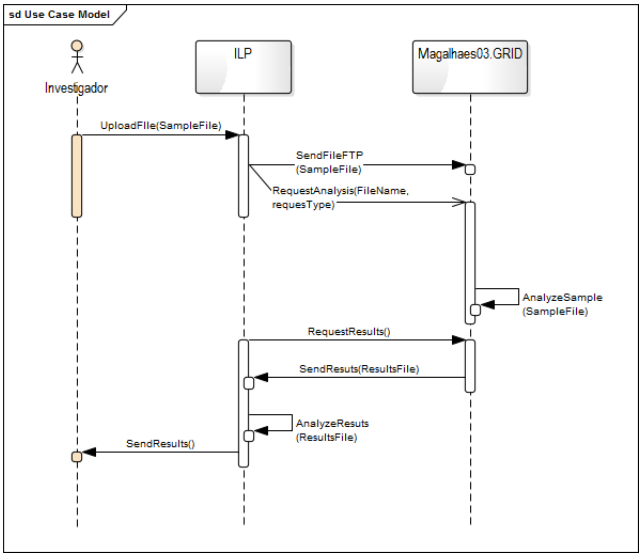


Figura 3.4: Diagrama de sequência das mensagens entre o investigador, ILP e Magalhaes03.GRID

## 3.4 Implementação

### Arquitetura da Solução

A solução é constituída por duas máquinas distintas, de modo a terem tarefas independentes e evitar problemas de excesso de memória em utilização . Cada uma das máquinas está desenvolvida para executar diferentes tarefas, de acordo com as suas características representadas na tabela 3.1. A Base de Dados e o Portal Web foram desenvolvidos na máquina ILP e o pipeline e o FTP Server na máquina Magalhaes03.GRID. Esta separação deve-se ao elevado consumo de memória das ferramentas biológicas, principalmente o TopHat e o Bowtie, que executam os algoritmos de alinhamento e mapeamento em milhões de sequências genómicas, além de carregarem os *index*

dos genomas em memória (cerca de 12GB) . A máquina Magalhaes03.GRID mantém todos os ficheiros usados na plataforma, sendo estes indexados a partir da Base de Dados na máquina ILP.

| Máquina | Processador | Memória RAM | Disco Rígido |
|---|---|---|---|
| **ILP** | Intel(R) Core(TM)2 Quad CPU Q9300, @ 2.50GHz | 8GB | 431GB |
| **Magalhaes03.GRID** | Intel(R) Xeon(R) Octa CPU E5430, @ 2.66GHz | 32GB | 11T |

Tabela 3.1: Características técnicas das máquinas ILP e Magalhaes03.GRID

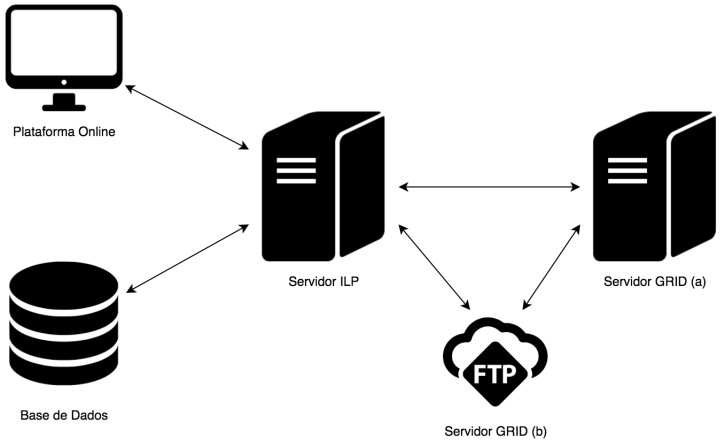


Figura 3.5: Diagrama de arquitetura da solução proposta

**Plataforma Online**

A plataforma online é a interface do utilizador com o pipeline e onde o utilizador interage com a aplicação. O seu desenvolvimento foi em linguagens web comuns como HTML5, CSS e Javascript, que permitem atingir os objetivos e as funcionalidades pretendidas.

Para uma interface *user-friendly*, responsiva e funcional foram usadas duas *frameworks*, Bootstrap3 e Highcharts. Estas duas frameworks são de uso grátis para produtos e aplicações não comercializados.

A *framework* Bootstrap3 utiliza HTML, CSS e Javascript e é usada para a interface com o utilizador, sendo responsável por componentes como formulários, botões, tabelas e listagens. A grande vantagem de usar frameworks como Bootstrap3 é que temos uma interface simples e responsiva, permitindo assim maior concentração e esforços no desenvolvimento e implementação de toda a lógica inerente à dissertação.

A *framework* Highcharts permite adicionar gráficos de análise de dados à interface online. Baseada em Javascript, esta ferramenta é compatível com os variados *Web Browsers* e as plataformas móveis existentes. Esta *framework* permite usar várias estruturas de dados nos gráficos, podendo

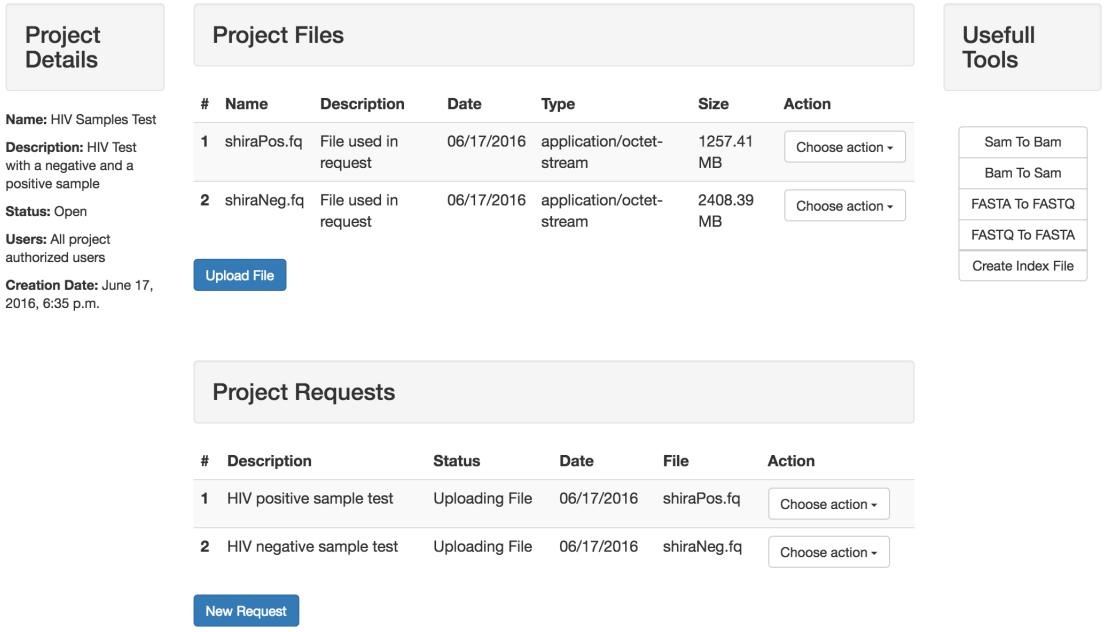


Figura 3.6: Exemplo de interface usando Bootstrap3

estes ser estáticos ou dinâmicos. Também permite a opção de *zoom*, bastante importante na implementação da solução, pois, por vezes, são mostrados gráficos com milhões de pontos, dando assim a opção ao utilizador de ver zonas específicas do mapeamento dos patogénicos.

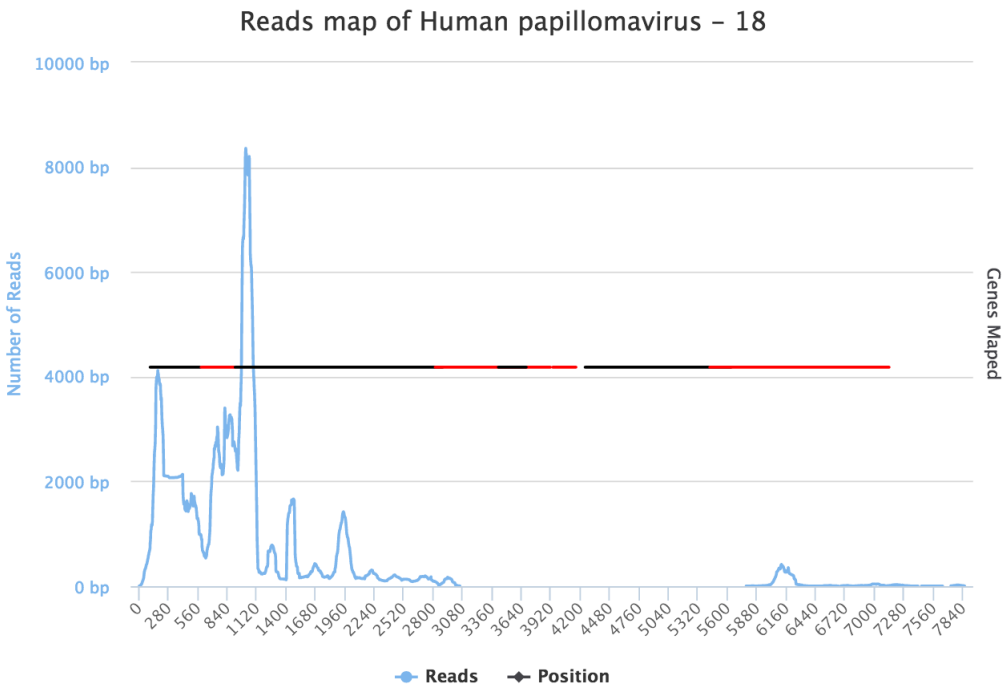


Figura 3.7: Exemplo de um gráfico usando Highcharts

**Servidor**

Tanto o servidor ILP como o Magalhaes03.GRID foram desenvolvidos em Python usando a *framework* Django. Python é uma linguagem mais direcionada e mais produtiva quando envolve *scripts* [Pyt]. Tendo em conta o software e as ferramentas utilizadas (analisadas no Capítulo 2), o uso de Python torna-as facilmente integráveis na plataforma. Além da ligação por FTP entre os servidores para a transferência de ficheiros, também são usados pedidos REST para transferência de informação e envio de mensagens.

Django [1] é uma *framework* direcionada para o desenvolvimento de *web services* usando a linguagem Python. É uma ferramenta grátis e *open-source* que permite o desenvolvimento rápido, seguro e escalável de plataformas online.

Além de Django, foram usadas várias bibliotecas externas desenvolvidas em Python:

- **BioPython**[2] - conjunto de módulos, classes e métodos para facilitar o uso de Python em Bioinformática. Inclui *parsers* para os vários tipos de ficheiros biológicos, possibilidade de aceder a recursos na web (ENCODE, NCBI,..) e também funções úteis para lidar com alinhamentos de sequências biológicas[CFG+09];

- **PySam**[3] - módulo desenvolvido em Python que usado para leitura, manipulação e escrita de *datasets* genómicos. Permite usar funcionalidades e comandos da ferramenta SamTools;

- **Requests**[4] - biblioteca usada para troca de HTTP *requests* entre os dois servidores. Além de pedidos GET/POST e da passagem de parâmetros no URL, também suporta diversos formatos (JSON, Binary,...) na resposta aos pedidos;

- **FTPlib** [5] - a biblioteca FTPlib implementa funções do lado do cliente para realizar pedidos FTP, seja para enviar ficheiros como para receber.

A transferência de dados será feita pelo protocolo FTP. Visto as amostras terem um tamanho considerável (normalmente entre os 3GB e os 6GB), estas não devem ser divididas em *chunks* nem serem enviadas em pacotes. O protocolo FTP (*File Transfer Protocol*) permite que os ficheiros das amostras sejam transferidos para o servidor de uma só vez (caso não ocorram falhas) e num espaço temporal menor [FTP]. Visto as amostras serem sequências genéticas, podem ser levantadas questões de segurança e privacidade. Mas, sendo o objetivo desta dissertação a investigação e não a comercialização de um produto, estas questões de segurança podem ser passadas para segundo plano, não sendo necessário implementar tecnologias como SSL (*Secure Sockets Layer*).

O servidor FTP foi desenvolvido na máquina Magalhaes03.GRID recorrendo à biblioteca django-ftpserver [6]. Esta biblioteca gere os acessos ao servidor e aos ficheiros nele existentes.

[1] https://www.djangoproject.com
[2] http://biopython.org/
[3] http://pysam.readthedocs.io/en/latest/
[4] http://docs.python-requests.org/en/master/
[5] https://docs.python.org/2.7/library/ftplib.html
[6] https://django-ftpserver.readthedocs.io

Permite criar diferentes grupos de acesso (*ftpusergroup*), cada um com o seu diretório, e juntar utilizadores a cada grupo FTP (*ftpuseraccount*).

```
1 ftp = FTP()
2 ftp.connect(host='magalhaes03.grid.fe.up.pt',port=9000)
3 ftp.login('admin','admin')
4 file = open(file_path,'rb')
5 ftp.storbinary('STOR '+fn, file)
6 file.close()
```

Listing 3.1: Exemplo de conexão FTP para envio de um ficheiro entre a máquina ILP e Magalhaes03.GRID

**Base de dados**

Essencialmente, a escolha da tecnologia a usar para a base de dados poderia recair entre uma base de dados SQL (MySQL ou PostgreSQL) ou NoSQL (MongoDB) Analisando o estudo efetuado na secção 2.5.2 no Capítulo 2 a base de dados foi desenvolvida em MySQL.

Para conectar a plataforma desenvolvida em Django e a base de dados presentes no servidor ILP, foi necessário recorrer à biblioteca PyMySQL, facilitando, assim, a conexão e envio de informação. [7]

Na figura 3.8 está representado o diagrama UML do modelo de classes e relações entre elas usado no desenvolvimento da base de dados da solução.

*Project* - A tabela *Project* guarda o nome, a descrição, a data de criação e o estado ("*open*"ou "*closed*") sobre cada projeto.

*User* - A tabela *User* guarda o primeiro e último nome, o email, a password, a data de registo, a instituição a que pertence e qual o grau académico dos utilizadores registados na plataforma. Os utilizadores têm associados a si vários projetos, que poderão ser partilhados e realizados em colaboração com outros utilizadores.

*File* - Cada ficheiro que é analisado ou convertido ficará indexado na base de dados com o nome, a descrição, a data de *upload*, o tipo e o tamanho. Os ficheiros estão disponíveis no painel de cada projeto.

*Request* - Um *Request* é um pedido de análise de uma amostra ou de conversão de ficheiro por parte do utilizador. A cada pedido estará associado um utilizador, um projeto e um ficheiro. Na base de dados ficará guardada informação sobre a descrição, o estado do *Request*, a data de criação e informação sobre o resultado final do mapeamento (tamanho da sequência, tamanho da amostra total e total de *reads* mapeados).

[7] https://github.com/PyMySQL/PyMySQL

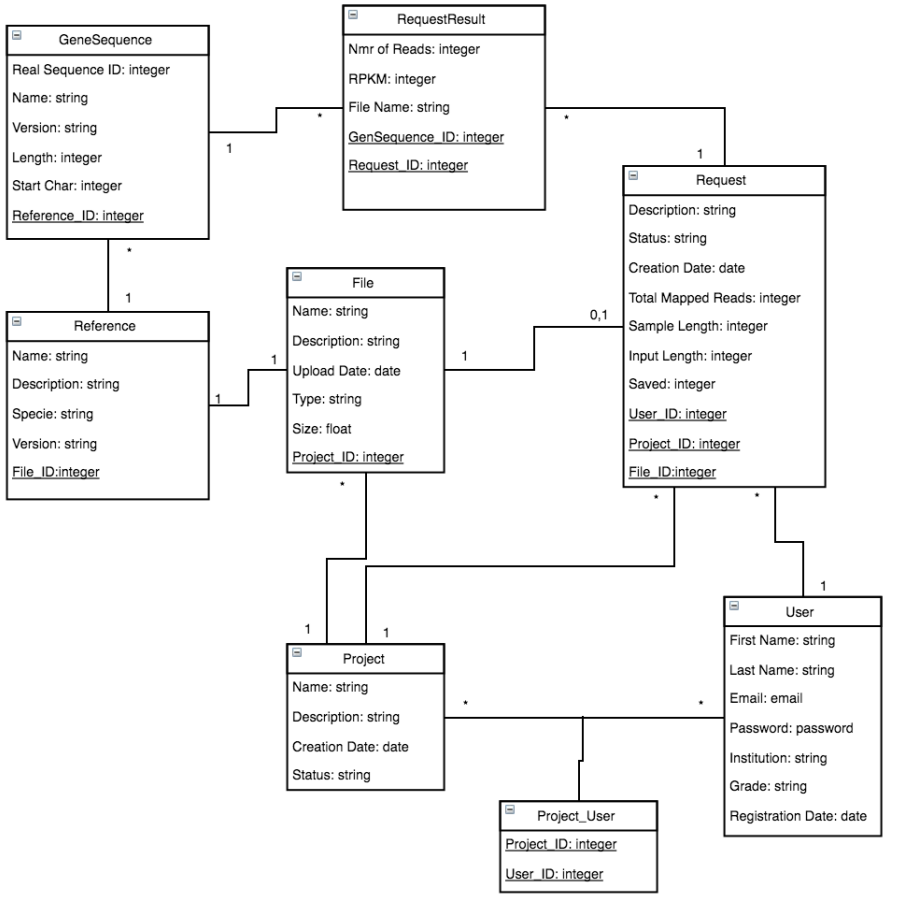


Figura 3.8: Diagrama UML de classes da solução

*Reference* - Cada *Reference* é um genoma de referência diferente. A cada um deles estará associada a espécie, a versão,a descrição e o respetivo ficheiro com as sequências biológicas.

*GeneSeq* - Cada um dos ficheiros dos genomas de referência contém diferentes patogénicos (Viral e Bacterial) e cromossomas (Humano). O modelo *GeneSeq* guarda a informação sobre cada patogénico/cromossoma: o ID na base de dados do NCBI, o nome, a versão, o tamanho da sequência e a posição (do caráter inicial) no ficheiro do respetivo genoma de referência.

*Request Result* - Cada *Request* apresenta os patogénicos mapeados que estão guardados neste modelo de dados. Sobre cada patogénico mapeado é guardado o seu RPKM, total de *reads* mapeados e o ficheiro no qual os *reads* estão guardados. Os *reads* mapeados não são guardados na base de dados, mas, sim, num ficheiro associado a cada *Request Result*. Esta opção tem como objetivo não sobrecarregar a base de dados, pois por cada *Request Result* estão associados, geralmente, centenas de milhares de *reads*.

### *Pipeline* de Mapeamento e Alinhamento

A principal funcionalidade de toda a plataforma é a possibilidade dos utilizadores poderem fazer *requests* de análise de amostras de células humanas. Esta análise passa por três mapeamentos distintos consecutivos. Após o upload da amostra, esta é enviada por FTP para o servidor Magalhaes03.GRID. Juntamente com o envio do ficheiro é enviada uma mensagem HTTP para indicar o pedido efetuado e o nome do ficheiro enviado.

```
1 def handle_request_to_grid(f,fn,id):
2   handle_uploaded_file(f,fn) #transferencia de ficheiros por FTP
3   r = requests.get('http://magalhaes03.grid.fe.up.pt:9001/supRNA_analyzer?file='+fn
        +'&job=newreq'+'&id='+str(id))
```

Listing 3.2: Mensagem HTTP do ILP para Magalhaes03.GRID

Após a receção completa do ficheiro por parte da máquina Magalhaes03.GRID, o mapeamento da amostra é iniciado no pipeline. Primeiro, a amostra é mapeada contra o genoma humano, usando a ferramenta TopHat. Deste alinhamento resultam dois ficheiros BAM, **accepted_hits.bam**, que é descartado, e **unmapped.bam**, que é convertido para FASTQ (usando SamTools). O ficheiro **unmapped.fastq** contém os *reads* que não foram mapeados no genoma humano e é mapeado contra o genoma viral. Deste mapeamento volta a resultar dois ficheiros BAM, **viral_mapped.bam**, que neste caso é guardado, e **unmapped.bam**, que é convertido para FASTQ. Mais uma vez, o ficheiro **unmapped.fastq** é mapeado contra o genoma bacterial de onde resulta o ficheiro BAM **bacterial_mapped.bam** que contém os *reads* mapeados no genoma bacterial . Tanto o ficheiro **viral_mapped.bam** como o **bacterial_mapped.bam**, juntamente com um resumo do processo, são enviados de volta para a máquina ILP.

Na máquina ILP, estes ficheiros são analisados recorrendo à biblioteca ***pysam***, da qual resulta o conjunto de *reads* mapeados nos genomas de referência. Este conjunto é guardado num *hash-map* segundo a estrutura - *"hash-map['id do patogénico'] = total de reads."* Este *hash-map* é convertido em JSON para poder ser usado pela *framework* Highcharts.

```
1  hash_reads = {}
2  print("Start reading reads")
3  with open('supRNA/media/Request-analysis/'+rID+'/'+rID+"_reads.txt", "r") as file:
4    for line in file:
5      info = line.split("-")
6      if int(info[0]) == int(r_seq_id):
7        for i in range(int(info[2]),int(info[3])):
8          if str(i) in hash_reads:
9            hash_reads[str(i)]=hash_reads[str(i)]+1
10         else:
11           hash_reads[str(i)]=1
12 print("Stop reading reads")
13
14 gene_result = json.dumps(hash_reads, cls=DjangoJSONEncoder)
```

Listing 3.3: Leitura do ficheiro dos *reads* resultantes da análise para um *hash-map* e consequente conversão para JSON

Assim, após toda a execução das funções do pipeline, o utilizador pode analisar os resultados tendo como base os gráficos e tabelas produzidos na plataforma online. Para a obtenção de informações sobre cada gene é efetuado um pedido HTTP ao *web service* da NCBI usando a biblioteca de Python urllib[8]. Esta informação é recebida num ficheiro GenBank do patogénico em análise.

```
1 urllib.request.urlretrieve("http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.
    fcgi?db=nucleotide&id="+r_seq_id+"&rettype=gb", "supRNA/"+r_seq_id+".gb")
```

Listing 3.4: HTTP *request* para o *download* de um ficheiro GenBank de um gene

## 3.5 Sumário

Neste capítulo foram analisadas as tecnologias usadas para o desenvolvimento da solução proposta. Além das tecnologias, também se exemplificaram casos de uso da plataforma e quais os principais métodos que tornam possível a análise das amostras inseridas pelos investigadores.

[8]https://docs.python.org/2/library/urllib.html

# Capítulo 4

# Casos de Estudo

Neste capítulo são apresentados casos de estudo que permitem a validação e a verificação da solução.

Após a implementação da solução, foram escolhidas amostras reais provenientes de dois locais diferentes para a apresentação dos resultados. O objetivo da utilização das amostras é, após a conclusão da análise e da visualização dos resultados, ter informações reais para a poder validar com informação e resultados já existentes e verdadeiros.

## 4.1 Objetivos da Análise

Os objetivos da análise de amostras reais, além de provar as funcionalidades *core* da plataforma, consistem em validar os resultados do pipeline e de todo o mapeamento. Consequentemente, será possível ver a expressão genética dos patogénicos mapeados segundo o número de *reads* mapeados e o RPKM de cada patogénico. Estas experiências também permitem colocarmonos no "papel"do investigador e ter uma experiência de utilização da plataforma implementada.

## 4.2 Protocolo de Análise

Todas as amostras em análise foram submetidas às mesmas condições. Os mapeamentos contra os genomas de referência humano (grCh38), viral e bacterial foram efetuados pela ferramenta TopHat, com as mesmas configurações (4 *threads*), na máquina Magalhaes03.GRID. A homogeneidade de todo o processo garante o mesmo ambiente envolvente para todas as amostras a analisar. Estes *Case Studies* foram efetuados seguindo o Anexo A.

Nos resultados apresentados na secção 4.3 de Interpretação de Resultados apenas são apresentados os gráficos e tabelas de visualização dos patogénicos com mais *reads* mapeados para uma melhor perceção da análise efetuada pelo *pipeline* do *web service*.

O anexo A representa um possível guião de teste para a análise de resultados.

| Amostra | Tamanho (GB) | Mapeamento Genoma Humano (grCh38) (min) | Mapeamento Genoma Viral (min) | Mapeamento Genoma Bacterial (min) |
|---|---|---|---|---|
| **GM12878R1** | 4,7 | 123 | 33 | 80 |
| **HELA** | 2,9 | 94 | 18 | 66 |
| **HIVpos** | 1,3 | 77 | 2 | 31 |
| **HIVneg** | 2,5 | 111 | 2 | 60 |

Tabela 4.1: Tempo de mapeamento de cada amostra nos vários genomas

| Amostra | Total Reads (bp) | Reads Mapeados (bp) |
|---|---|---|
| **GM12878R1** | 24159698 | 89443 |
| **HELA** | 14830482 | 9677194 |
| **HIVpos** | 10891241 | 9429858 |
| **HIVneg** | 20229554 | 19522890 |

Tabela 4.2: Resultados totais da análise realizada

## 4.3 Interpretação de Resultados

### 4.3.1 Amostras do projeto ENCODE

Recorrendo ao repositório *Genome Browser* do projeto ENCODE foi possível obter amostras de teste de doenças recorrentes causadas por vírus e bactérias. O uso destas amostras procura mostrar quais os patogénicos presentes com maior abundância nestas células. Da tabela 4.1 é possível concluir que quanto maior o ficheiro da amostra mais tempo demora cada mapeamento, contudo o tempo dos diversos mapeamentos não é diretamente proporcional ao tamanho. Por exemplo, a amostra HIVneg e HELA têm sensivelmente o mesmo tamanho e o tempo de mapeamento do Genoma Viral é bastante maior no caso da amostra HELA. Essa diferença deve-se à quantidade de *reads* mapeados em cada um dos genomas e ao próprio tamanho médio das sequenças genómicas presentes em cada uma das amostras.

#### GM12878R1 - *Lymphoblastoid cell lines*

Amostra de célula humana infetada com o vírus Epstein-Barr, que origina células denominadas como *lymphoblastoid cell lines* [EP].

Como é possível ver na figura 4.1, o vírus *Human herpesvirus 4*, foi o patogénico mais mapeado na amostra. Este vírus, com um tamanho de 171823bp, apresenta 97972bp mapeados e um RPKM de 53357, valores bastante elevados, comprovando assim uma grande abundância deste patogénico na amostra GM12878R1.

#### HELA - Cancro cervical

Amostra de célula humana com cancro cervical (conhecido como cancro do colo do útero) causado pelo vírus do papiloma humano (HPV) [LPR+13].

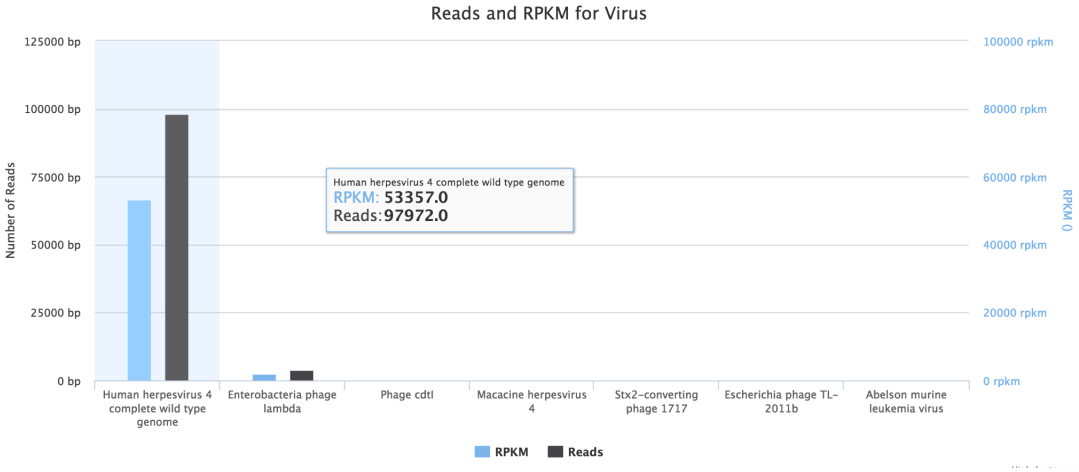


Figura 4.1: Gráfico representativo do mapeamento da amostra GM12878R1 no genoma viral

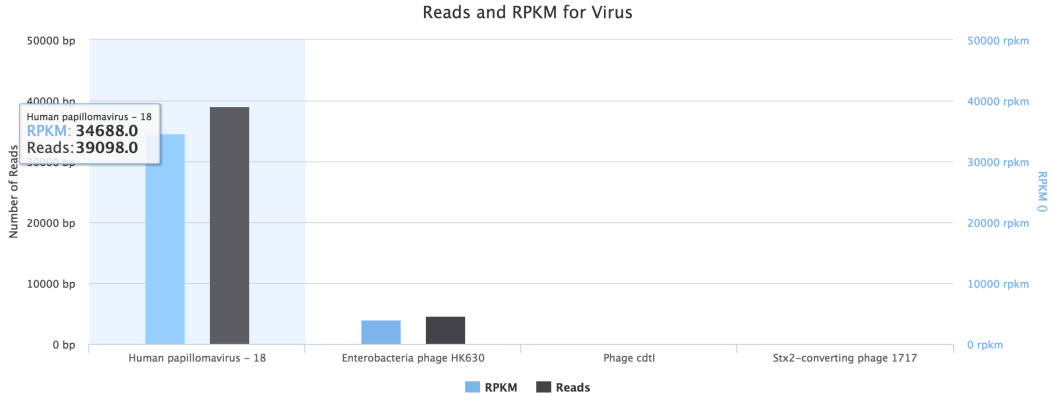


Figura 4.2: Gráfico representativo do mapeamento da amostra HELA no genoma viral

Na figura 4.2 é possível identificar claramente o patogénico com maior abundância na amostra. O vírus *Human papillomavirus - 18* apresenta um RPKM de 34688 e um mapeamento com a quantidade de 39098bp.

### 4.3.2 Amostras de HIV

Este teste foi efetuado com duas amostras de HIV, uma com um resultado positivo de infeção e outra com resultado negativo. As amostras foram obtidas após o contacto com Ofer Isakov, autor do artigo *Pathogen detection using short-RNA deep sequencing subtraction and assembly* [IMS11]. Assim, é possível confirmar os resultados obtidos na plataforma desenvolvida com esta dissertação comparando-os com os resultados obtidos descritos no artigo.

Os gráficos nas figuras 4.3 e 4.4 mostram o resultado do mapeamento, nos genomas de referência, da amostra de HIVpos com resultado positivo. Verifica-se grande abundância do vírus *Human immunodeficiency virus 1*, responsável da infeção HIV, com 5059 bp mapeados e um RPKM de 244470.

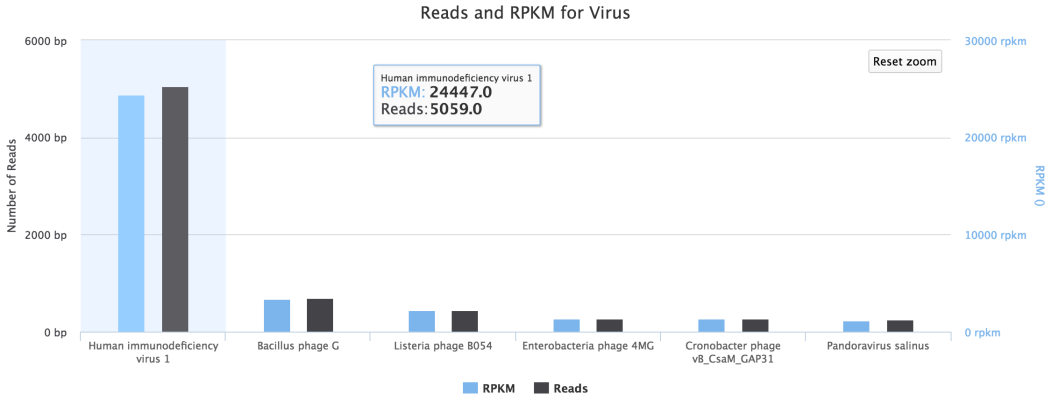


Figura 4.3: Gráfico representativo do mapeamento da amostra HIVpos no genoma viral

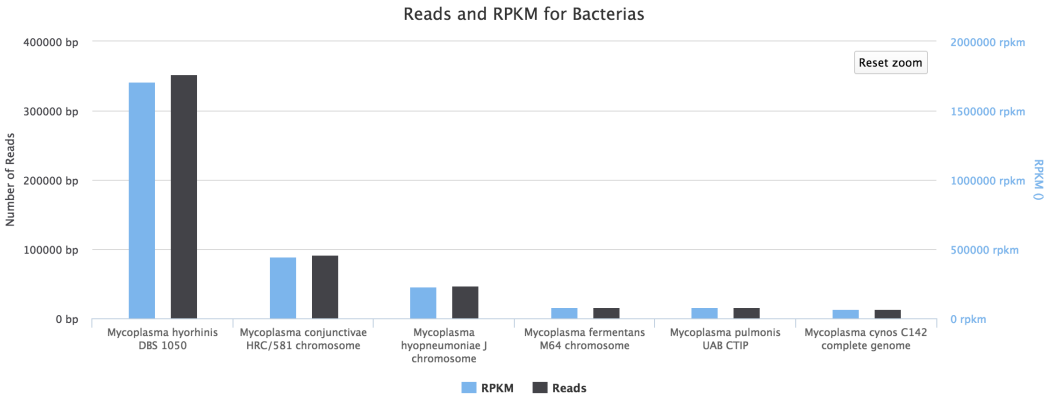


Figura 4.4: Gráfico representativo do mapeamento da amostra HIVpos no genoma bacterial

As bactérias encontradas com maiores índices de mapeamento são bactérias de origem não-humana, como *Mycoplasma hyorhinis* (cerca de 350000 bp mapeados), *Mycoplasma conjunctivae* (cerca de 90000bp mapeados) e Mycoplasma hyopneumoniae (cerca de 50000 bp mapeados).

Na figura 4.5, está representado, em tabela, o mapeamento da amostra HIVneg com resultado de HIV negativo. É possível observar, também, as bactérias mais mapeadas, contudo, a sua abundância é bastante menor em relação às encontradas na amostra HIVpos.

## 4.4 Sumário

A análise de resultados aqui efetuada não serve apenas de validação da solução, mas também procura exemplificar as potencialidades e funcionalidades da plataforma desenvolvida. Assim, com a análise dos resultados aqui demonstrados é possível validar a solução como uma resposta ao problema formulado e descrito nesta dissertação, pois com a demonstração gráfica implementada é bastante intuitiva a percepção e a identificação dos patogénicos presentes nas células humanas com e sem cancro.

| # | Pathogene Name | Type | Length | Total Reads | RPKM |
|---|---|---|---|---|---|
| 1 | Alteromonas macleodii str. 'Ionian Sea U8' | Bacteria | 4395035 | 19627 | 46200 |
| 2 | Achromobacter xylosoxidans NBRC 15126 = ATCC 27061 | Bacteria | 6683584 | 7671 | 18057 |
| 3 | Mycobacterium tuberculosis str. Beijing/NITR203 | Bacteria | 4411128 | 2893 | 6809 |
| 4 | Mycoplasma hyorhinis MCLD chromosome | Bacteria | 829709 | 828 | 1949 |
| 5 | Candidatus Accumulibacter phosphatis clade IIA str. UW-1 chromosome | Bacteria | 5058518 | 629 | 1480 |
| 6 | Herminiimonas arsenicoxydans chromosome | Bacteria | 3424307 | 543 | 1278 |
| 7 | Kitasatospora setae KM-6054 | Bacteria | 8783278 | 531 | 1249 |
| 8 | Glaciecola psychrophila 170 | Bacteria | 5413691 | 518 | 1219 |

Figura 4.5: Tabela representativa do mapeamento da amostra HIV-neg no genoma bacterial

# Capítulo 5

# Conclusões e Trabalho Futuro

## 5.1 Conclusão

Embora exista uma grande quantidade de software na área da Bioinformática tendo por objetivo ajudar estudos médicos, principalmente no cancro, a solução proposta nesta dissertação procura alguma inovação na área. Esta plataforma procura ser uma solução onde os investigadores apenas submetam a amostra a analisar e o resto da aplicação funcione como uma *black-box* onde, no fim de todo o processo, o investigador receba os resultados e os possa analisar. A plataforma apresenta uma interface *web* bastante simples e intuitiva, principalmente no modo de apresentar os resultados, para que, independentemente do conhecimento informático, os investigadores possam interpretar e utilizar a informação apresentada. Com o envolvimento dos investigadores, esta plataforma pode ser uma ferramenta importante no estudo de infeções e agentes internos como causas de cancro.

Com o desenvolvimento do *web service* projetado para análise de amostras de células humanas, na procura de patogénicos que possam levar a formações cancerígenas, e da plataforma online de visualização de resultados, os principais objetivos desta dissertação foram concretizados. Contudo, certos aspetos secundários, não necessários às funcionalidades principais da plataforma, ou não foram implementados ou poderão vir a ser melhorados no futuro. Também o tempo de resposta de certos serviços poderia ser reduzido, mas o fato de a plataforma lidar com uma grande quantidade de dados, por vezes, nem sempre é possível acelerar processos e analisar dados num curto espaço de tempo.

Este projeto teve como grande motivação não só o facto de poder ser mais uma ajuda e um esforço no estudo do impacto de patogénicos na origem de cancro, mas também, e sobretudo, a possibilidade e a oportunidade de criar algo que ajude e tenha impacto positivo na vida de milhões de pessoas, pelo contributo na prevenção, no diagnóstico e no tratamento das doenças oncológicas.

## 5.2 Trabalho Futuro

Como referido na secção 5.1, certamente que há aspetos e funcionalidades que podem ser melhoradas num futuro desenvolvimento da plataforma. A principal razão do não desenvolvimento destas pequenas funcionalidades nesta dissertação deve-se ao facto de terem sido definidas ou consideradas como requisitos de baixa prioridade na projeção da plataforma. São funcionalidades não necessárias ao uso da plataforma, mas que, caso fossem implementadas, poderiam aumentar o grau de satisfação na utilização da plataforma por parte do utilizador.

Por vezes, ao longo da experiência de utilização da plataforma não há mensagens por parte do sistema para o utilizador indicando o estado e o resultado das suas ações. Esta funcionalidade poderá ser implementada recorrendo ao uso da *framework* de mensagens do Django.

O uso de Data Mining, relacionando os vários resultados das análises em várias amostras e os dados retirados de projetos como o ENCODE, poderá ser um grande *upgrade* ao *web service* e à plataforma desenvolvida, oferecendo aos utilizadores mais e melhores funcionalidades.

Também a implementação de algoritmos de *Machine Learning*, poderá colocar a solução desenvolvida noutro nível de inovação e utilidade. Com algoritmos de *Machine Learning* e o constante uso da plataforma, com a criação de novos projetos e novas análises de amostras, a própria plataforma, usando modelos preditivos e de probabilidades, poderá sugerir e interpretar os resultados da análise de forma autónoma.

Caso o objetivo do projeto fosse a comercialização, deveria ser considerado não só o uso de protocolos de segurança na transferência de ficheiros (os dados biológicos são bastante sensíveis e contêm informações críticas e pessoais) como também a possibilidade de a plataforma se tornar escalável, podendo haver em simultâneo uma quantidade considerável de utilizadores a fazer pedidos de análises de amostras genómicas.

# Referências


[BAH+13] A Bora, G U Ahmed, N K Hazarika, K N Prasad, S K Shukla, V Randhawa e J B Sarma. Incidence of bla NDM - 1 gene in Escherichia coli isolates at a tertiary care referral hospital in Northeast India. 31:250–256, 2013. doi:10.4103/0255-0857.115628.

[Bio] Bioinformatics at ucsf. http://bioinformatics.ucsf.edu/about/program. Accessed: 2016-02-11.

[BKmL+07] Dennis A Benson, Ilene Karsch-mizrachi, David J Lipman, James Ostell e David L Wheeler. GenBank. 35(December 2006):21–25, 2007. doi:10.1093/nar/gkl986.

[CFG+09] Peter J A Cock, Christopher J. Fields, Naohisa Goto, Michael L. Heuer e Peter M. Rice. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38(6):1767–1771, 2009. doi:10.1093/nar/gkp1137.

[DB04] Francis McCabe David Booth, Hugo Haas. Web Services Architecture. (February), 2004.

[DDS+13] Alexander Dobin, Carrie a Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson e Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, 29(1):15–21, 2013. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3530905{&}tool=pmcentrez{&}rendertype=abstract, doi:10.1093/bioinformatics/bts635.

[DNA] Tradução de dna e transcrição de rna. http://www.sobiologia.com.br/conteudos/Citologia2/AcNucleico7.php. Accessed: 2016-02-10.

[EP] Encode-Project. GM12878 cell culture. d:3–5.

[FASa] Blast - query input and database selection. http://zhanglab.ccmb.med.umich.edu/FASTA/. Accessed: 2016-1-28.

[FASb] What is fasta format? http://blast.ncbi.nlm.nih.gov/blastcgihelp.shtml. Accessed: 2016-01-28.

[FTP] Ftps (ftp over ssl) vs. sftp (ssh file transfer protocol): what to choose. https://www.eldos.com/security/articles/4672.php?page=all. Accessed: 2016-02-08.

[Gen] Genbank flat file format. http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html#LocusB. Accessed: 2016-06-14.

# REFERÊNCIAS


[GFFa] Gff format. http://genome.ucsc.edu/FAQ/FAQformat.html#format3. Accessed: 2016-06-07.

[GFFb] Gff/gtf file format - definition and supported options. http://www.ensembl.org/info/website/upload/gff.html. Accessed: 2016-06-07.

[GJ13] Claudia Gonzaga-Jauregui. Human genome sequencing in health and disease. pages 35–61, 2013. doi:10.1146/annurev-med-051010-162644.Human.

[HM11] Daniel H. Huson e Suparna Mitra. Comparative Metagenome Analysis Using MEGAN. *Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches*, (February):341–352, 2011. doi:10.1002/9781118010518.ch39.

[Ill11] Illumina. Quality Scores for Next-Generation Sequencing. *Http://Res.Illumina.Com/Documents/Products/Technotes/Technote_Q-Scores.Pdf*, pages 1–2, 2011.

[IMS11] Ofer Isakov, Shira Modai e Noam Shomron. Pathogen detection using short-RNA deep sequencing subtraction and assembly. 27(15):2027–2030, 2011. doi:10.1093/bioinformatics/btr349.

[KPT+13] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley e Steven L Salzberg. TopHat2 : accurate alignment of transcriptomes in the presence of insertions , deletions and gene fusions. pages 1–13, 2013.

[Li209] The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009. URL: http://bioinformatics.oxfordjournals.org/cgi/content/full/25/16/2078.

[LPR+13] Jonathan J M Landry, Paul Theodor Pyl, Tobias Rausch, Thomas Zichner, Manu M Tekkedil, Adrian M Stütz, Anna Jauch, Raeka S Aiyar, Gregoire Pau, Nicolas Delhomme, Julien Gagneur, Jan O Korbel, Wolfgang Huber e Lars M Steinmetz. The Genomic and Transcriptomic Landscape of a HeLa Cell Line. 3:1213–1224, 2013. doi:10.1534/g3.113.005777.

[LTPS09] Ben Langmead, Cole Trapnell, Mihai Pop e Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. 10(3), 2009. doi:10.1186/gb-2009-10-3-r25.

[MSB+12] Iain Milne, Gordon Stephen, Micha Bayer, Peter J A Cock, Leighton Pritchard, Linda Cardle e Paul D Shaw. Using Tablet for visual exploration of second-generation sequencing data. 14(2), 2012. doi:10.1093/bib/bbs012.

[NG-] Our inner viruses: Forty million years in the making. http://phenomena.nationalgeographic.com/2015/02/01/our-inner-viruses-forty-million-years-in-the-making/. Accessed: 2016-02-05.

[Pro] Código genético de proteínas. http://www.sobiologia.com.br/conteudos/Citologia2/AcNucleico6.php. Accessed: 2016-06-12.

[Pyt] Python as web server. https://docs.python.org/2/howto/webservers.html. Accessed: 2016-02-04.

# REFERÊNCIAS


[Seq] Genome assembly: de novo versus mapping to a reference. https://era7bioinformatics.com/en/page.cfm?id=1500. Accessed: 2016-06-13.

[SFW13] The Sam, B a M Format e Specification Working. Sequence Alignment / Map Format Specification. (May):1–16, 2013.

[SIL07] Yvan Saeys, Iñaki Inza e Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007. doi:10.1093/bioinformatics/btm344.

[SQL] Sql vs nosql databases. http://www.sitepoint.com/sql-vs-nosql-differences/. Accessed: 2016-02-06.

[TPS09] Cole Trapnell, Lior Pachter e Steven L Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)*, 25(9):1105–11, 2009. URL: http://www.ncbi.nlm.nih.gov/pubmed/19289445, doi:10.1093/bioinformatics/btp120.

[Vir] The lurker: How a virus hide in our genome for six million years. http://phenomena.nationalgeographic.com/2013/05/10/the-lurker-how-a-virus-hid-in-our-genome-for-six-million-years/. Accessed: 2016-02-06.

[Web] Web service tutorial. http://www.tutorialspoint.com/webservices/. Accessed: 2016-02-02.

# REFERÊNCIAS

# Anexo A

# Guião de Utilizador

Este guião de utilizador procura exemplificar todos os procedimentos desde o registo na plataforma e a visualização de resultados da análise de uma amostra de ARN por parte do *web service*. Também, mostra todos os ecrãs de interface da plataforma online.

### A.0.1 *Homepage*

Na página inicial da plataforma (Figura A.1) existe uma pequena descrição das funcionalidades possíveis de serem realizadas na plataforma. A partir desta página é possível fazer o Registo e o Login para obter acesso ao painel de projetos e aí upload de amostras para serem analisadas pelo *pipeline*.

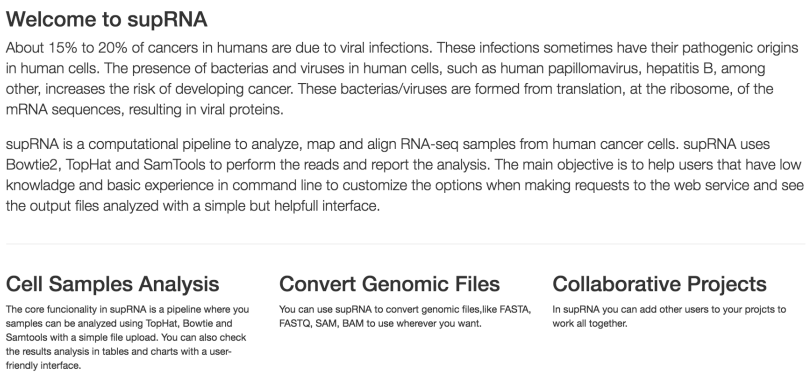


Figura A.1: *Homepage* da plataforma online

#### A.0.1.1 Registo

Para efetuar o registo, é necessário carregar no link *"Sign up"* no canto superior direito da página e preencher o formulário (Figura A.2) com os seguintes campos:

- **First Name** - primeiro nome do utilizador;
- **Last Name** - último nome do utilizador;

- **Email** - email do utilizador, necessário para o *Login* na aplicação;
- **Password** - *password*, necessária para o *Login* na aplicação;
- **Insitution** - Instituição de ensino ou de investigação a que o utilizador pertence;
- **Grade** - grau académico do utilizador.


Figura A.2: Formulário de registo na plataforma

#### A.0.1.2 *Login*

Para efetuar o *login* na plataforma, o utilizador deve inserir, no formulário de *Login* (Figura A.3), o email e a password gravados aquando o registo na plataforma:

- **Email** - email do utilizador registado;
- **Password** - password do utilizador registado:


Figura A.3: Formulário de *login* na plataforma

Após o *Login* o utilizador fica com acesso livre à plataforma podendo usufruir de todas as funcionalidades.

### A.0.2 Painel de Projetos

Neste painel (Figura A.4) é possível ver os projetos em que o utilizador colabora. Os projetos com cor verde são projetos em aberto ("*Open*") e os de cor vermelha são projetados dados como terminado ("*Closed*"). Nesta página também é possível adicionar novos projetos.

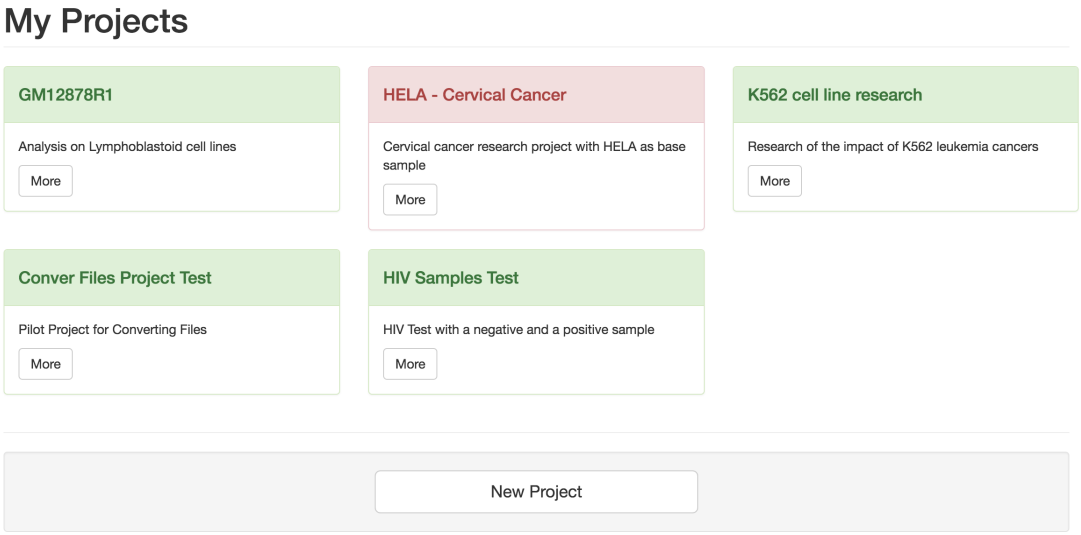


Figura A.4: Vista geral dos projetos, nos quais o utilizador colabora

#### A.0.2.1 Novo Projeto

Para a criação de um projeto é necessário preencher o formulário (Figura A.5) com os seguintes campos:

- **Name** - nome atribuído ao projeto;
- **Description** - descrição do projeto;
- **Collaborators** - email dos utilizadores que serão adicionados como colaboradores ao projeto.

### A.0.3 Projeto

Na página de cada projeto (Figura A.6) é possível ter uma vista geral do projeto. Este painel inclui os detalhes dos projetos, os ficheiros convertidos (com a possibilidade de *download*) e os amostras em análise ou já analisadas. Também, é possível, através de formulários, converter novos ficheiros, analisar amostras e adicionar colaboradores ao projeto.

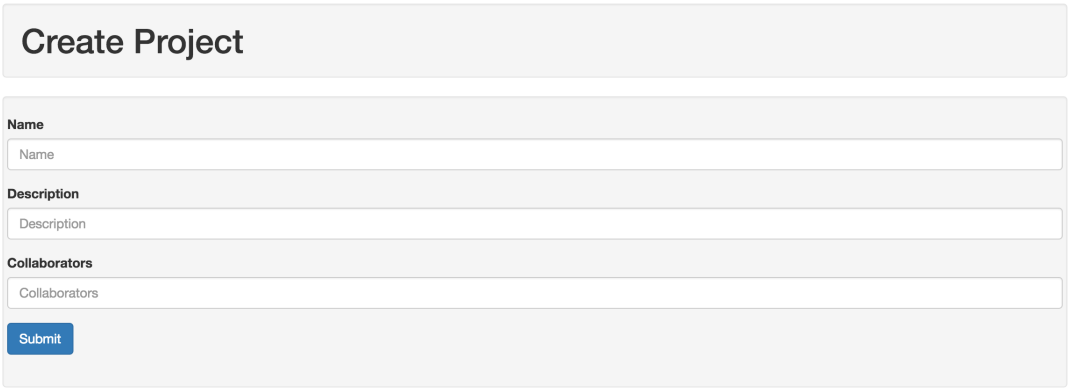


Figura A.5: Formulário para a criação de um novo projeto

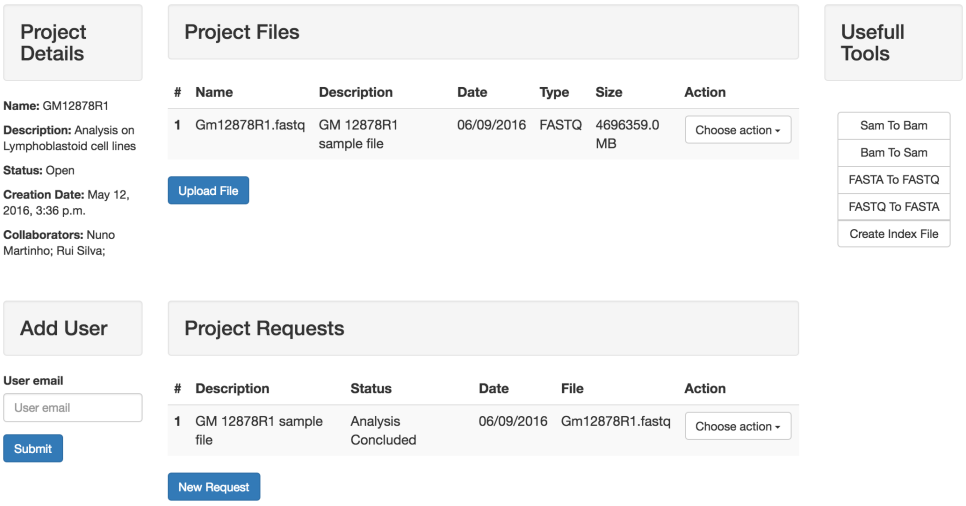


Figura A.6: *Overview* de um projeto

### A.0.3.1 Novo Pedido

Para efetuar um novo pedido de análise de uma amostra, é necessário carregar no botão *"New Request"* (Figura A.6). De seguida aparecerá um formulário para inserir a amostra e informações sobre o *request*:

- **Description** - descrição da amostra;
- **File** - ficheiro que contém a amostra, pode ser FASTQ, FASTA, BAM, SAM;
- **Type of Request** - o tipo do *request* pode ser *default* (usando os parâmetros definidos pela plataforma) ou *personalized* (usando os parâmetros definidos pelo utilizador)

### A.0.3.2 Conversão de Ficheiros

Para converter um ficheiro, é necessário escolher o tipo de conversão na secção *"Usefull Tools"* (Figura A.6). Após a escolha, aparecerá um formulário onde, apenas, tem que inserir o ficheiro,

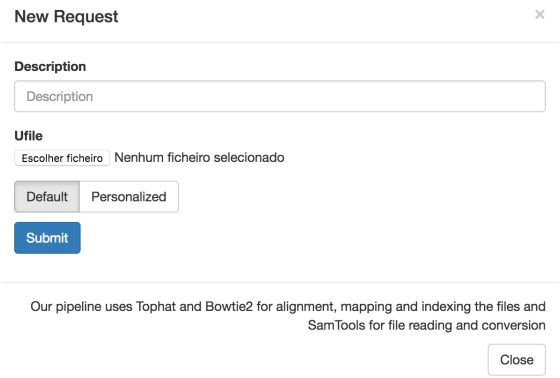


Figura A.7: Formulário de inserção de nova amostra para análise

no formato especificado, a converter.

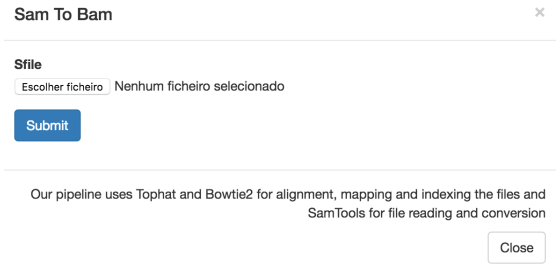


Figura A.8: Formulário de inserção de ficheiro a converter, neste caso, de SAM para BAM

#### A.0.3.3 Adicionar Colaboradores

Para adicionar um novo utilizador, como colaborador, a um projeto é necessário introduzir o email do utilizador na secção *"Add User"*.

### A.0.4 Visualização dos resultados

Após o a análise da amostra inserida estar concluída, o utilizador para ver os resultados da análise deve pressionar o botão *"Action"* da linha da amostra a analisar na tabela *"Requests"* no painel do projeto (Figura A.6). Existem duas maneiras diferentes de analisar os resultados. A mais geral, permite ver a quantidade de *reads* mapeados e o RPKM de cada agente patogénico. A mais específica, permite a visualização das zonas de mapeamento e quais os genes com mais *reads* mapeados em cada agente patogénico específico.

#### A.0.4.1 Visualização do mapeamento no genoma viral e bacterial

Na página de resultados de cada *"Request"* (Figura A.9) é possível na barra lateral esquerda, ver um sumário do processo de mapeamento. Por baixo dos *"Request Details"* é possível filtrar os resultados segundo o tipo de patogénico e ordenar os resultados segundo os parâmetros de cada

coluna da tabela com os patogénicos mapeados. Além destas funcionalidades também é possível procurar os patogénicos pelo nome. Os gráficos apresentados nesta página, por exemplo os da secção 4.3, representam graficamente a informação contida na tabela, sendo que há a possibilidade de aproximar zonas específicas dos gráficos para melhor compreensão. Ao percorrer os gráficos com o rato é possível obter informações de cada patogénico.

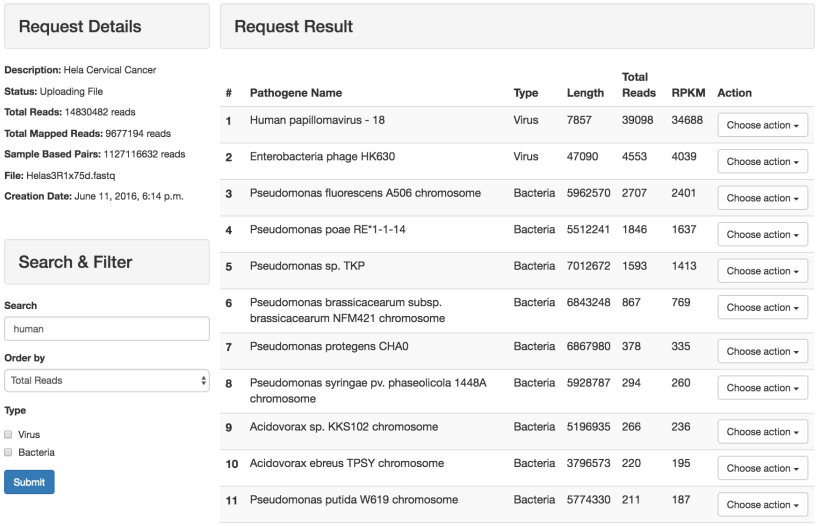


Figura A.9: Página de visualização dos resultados da análise de amostras

### A.0.4.2 Visualização do mapeamento por patogénico

Caso o utilizador pretenda ver em específico o resultado do mapeamento de um patogénico, terá que carregar no botão *"Choose Action"* do menu escolher a opção *"See chart analysis*. Assim, será redirecionado para uma página (Figura A.10) contendo o gráfico dos reads mapeados ao longo do genoma do patogénico (linha azul). Este gráfico também com a opção de *zoom*, tem a cor vermelha e preta a posição de cada gene, sendo que ao fazer *over* com o rato é possível ver as informações sobre cada gene do patogénico em análise.

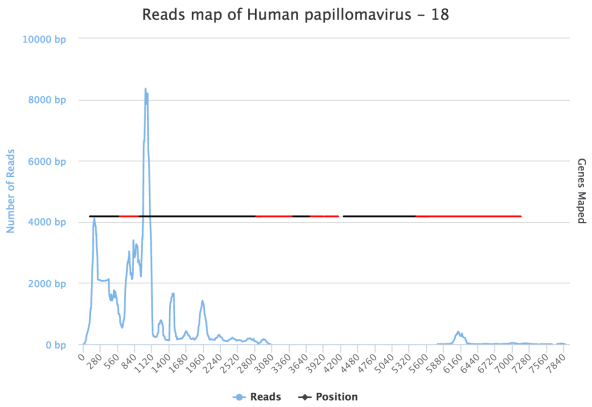


Figura A.10: Gráfico do resultado do mapeamento da amostra para cada patogénico