UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE INFORMÁTICA

ProGenViZ: a novel interactive tool for prokaryotic genome visualization and comparison

Bruno Filipe Ribeiro Gonçalves

DISSERTAÇÃO
MESTRADO EM BIOINFORMÁTICA E BIOLOGIA COMPUTACIONAL

ESPECIALIZAÇÃO EM BIOINFORMÁTICA

2014

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE INFORMÁTICA

ProGenViZ: a novel interactive tool for prokaryotic genome visualization and comparison

Bruno Filipe Ribeiro Gonçalves

DISSERTAÇÃO
MESTRADO EM BIOINFORMÁTICA E BIOLOGIA COMPUTACIONAL

ESPECIALIZAÇÃO EM BIOINFORMÁTICA

Dissertação orientada por:

Professor Doutor João André Nogueira Custódio Carriço
Professor Doutor Octávio Fernando de Sousa Salgueiro Godinho Paulo

2014

Abstract

Everyday new sequencing data and draft microbial genomes are obtained by high-throughput sequencing (HTS) and made publicly available at NCBI Sequence Read Archive (www.ncbi.nlm.nih.gov/sra) and EBI European Nucleotide Archive (http://www.ebi.ac.uk/ena). It is now perceived that the limiting factor is not obtaining the sequence data but the current capacity of the existing analysis methods to extract relevant information from data. This procedure is still often dependent on the use of expensive software or open-source freely available software that commonly has a high level of complexity to operate. The combination of this factors are currently leading to large amounts of data in public databases, but its analysis are usually limited in nature.

The visual representation of data has a very important role in the perception of complex information. When used in combination with methods for comparison and querying of genomic data, different visualization methods can be used to facilitate and guide the identification of interesting features. In Microbiology, the ability to visualize and compare genomes can be applied in the development of genomic epidemiology studies, as well as to identify and characterize microorganisms by determining lineages associated to antibiotic resistance, pathogenicity and virulence. These methods can assist in the detection and prevention of infectious diseases. However, this is a recent area of research that is still missing visualization tools to compare prokaryotic genomes in terms of gene content variation that offer interactive ways to explore data. Here, we present ProGenViZ, a user-friendly web-application that gives options to visualize and explore several prokaryotic genomes and their annotations, also providing features to compare specific genomic regions. Moreover, it provides additional features such as the re-annotation of genes, ordering of draft genome sequences against a reference genome and subsequent annotation by annotation transfer from one or more references. ProGenViZ is available at http://darwin.phyloviz.net/ProGenViZ.

Keywords

Visual Analysis; Comparative Genomics; Prokaryotes; High-throughput sequencing; Sequence annotation

Resumo

Todos os dias, novos dados de genomas de vários organismos são obtidos através de sequenciação de alto débito (*high-throughput sequencing* ou HTS) e são tornados públicos no NCBI Sequence Read Archive (www.ncbi.nlm.nih.gov/sra) e no EBI European Nucleotide Archive (http://www.ebi.ac.uk/ena). Actualmente, o factor limitante não é a obtenção dos dados genómicos mas sim a capacidade actual dos métodos de análise para extrair informação relevante deles. Este processo é ainda muitas vezes dependente do uso de software de custo considerável ou, no caso de ser gratuito, apresenta um nível elevado de complexidade. A combinação destes factores estão a contribuir para a acumulação de dados nas bases de dados públicas mas que têm a sua capacidade de análise limitada.

A representação visual de dados complexos é bastante importante na percepção e apreenção de informação contida nos dados. Quando usada em combinação com métodos de comparação e exploração de dados genómicos, diferentes métodos de visualização podem ser usados para facilitar a identificação de caracteristicas relevantes em diversos estudos. Em Microbiologia, a capacidade de visualizar e comparar genomas pode ser aplicada em estudos epidemiológicos, bem como na identificação e caracterização de organismos através da determinação de linhagens associadas a resistência a antibióticos, patogenecidade e virulência, que podem assistir na detecção e prevenção de doenças infecciosas. No entanto, esta é ainda uma área de pesquisa recente onde faltam ferramentas de visualização que permitam comparar genomas de procariotas em termos de variação genómica em várias escalas e que ofereçam formas interactivas para explorar os dados.

Nesta tese foi desenvolvido o ProGenViZ, uma aplicação web que oferece opções para visualizar e explorar simultaneamente múltiplos genomas de procariotas e suas anotações, fornecendo também funcionalidades para comparar regiões genómicas específicas. Além disso, a aplicação fornece capacidades adicionais como a re-anotação de genes, ordenação de sequências de genomas parciais contra contra um genoma de referência e subsequente anotação por transferência de uma ou mais sequências de referência. ProGenViZ está disponível em http://darwin.phyloviz.net/ProGenViZ.

Para o desenvolvimento da estrutura básica da aplicação web foi utilizado o Bootstrap framework. A área de trabalho foi dividida em duas partes, uma com vários menus interactivos que permitem ao utilizador realizar várias análises aos dados carregados e outra com a representação visual das sequências genéticas e suas anotações. A aplicação aceita como input

ficheiros no formato GenBank/EMBL, General Feature Format (GFF) e FASTA, bem como ficheiros com sequências múltiplas (multi-FASTA), tipicamente provenientes de genomas parciais.

O ProGenViZ apresenta uma nova abordagem para conseguir visualizar vários genomas de procariotas numa única imagem. Utiliza uma representação abstracta onde as sequências genómicas são divididas de acordo com as suas anotações em *regiões* para reduzir a complexidade da visualização. As *regiões* são depois divididas em várias porções de 500 pares de bases de acordo com o seu tamanho e apresentadas numa de duas representações visuais baseada em grafos - hive plot ou numa representação linear – que foram desenvolvidas utilizando o a biblioteca de JavaScript D3. Foram também produzidas várias formas de interação entre as duas representação visuais e o utilizador através de zoom em *regiões* específicas, mas também através da disposição de informações sobre cada *região* e de menus que fornecem funcionalidades adicionais que permitem explorar e comparar os ficheiros carregados.

Foi também desenvolvido um sistema de pesquisas que o utilizador pode realizar aos dados. É possível aceder a informação global sobre os ficheiros ou fazer pesquisas sobre *regiões* específicas. No caso do acesso a informação global sobre os ficheiros, o utilizador pode aceder a dados como o tamanho total das sequências e a percentagem que está anotada, ou a estatísticas associadas com a distribuição do tamanho das diferentes *regiões* e dos seus produtos. As distribuições do tamanho e dos produtos das *regiões* são representados graficamente na forma de um gráfico de barras e de um gráfico circular interactivos, que dão a capacidade ao utilizador de filtrar os dados que são mostrados.

Procuras por regiões específicas e comparações podem também ser feitas através das anotações – por nome ou por produto - ou através do uso de sequências internas ou externas para determinar *regiões* com homologia de sequência utilizando BLAST. Os resultados de todas as procuras e relações entre *regiões* são apresentados numa tabela de resultados e através de modificações específicas na representação visual. Quando são estabelecidas relações entre *regiões*, essas relações são mostradas nas representações visuais através de ligações entre as *regiões* envolvidas, o que permite visualizar a sintenia entre as *regiões* de diferentes sequências genómicas.

Além dos resultados do BLAST serem mostrados em forma de texto na tabela de resultados e através de modificações na imagem, foi também criada uma forma de visualizar os alinhamentos ao nível da sequência nucleotídica. Adicionalmente são ainda detectados *single nucleotide polymorphisms* (SNPs) através da utilização de uma funcionalidade do software MUMmer que detecta os SNPs existentes entre duas sequências.

Como actualmente as tecnologias de HTS permitem obter rapidamente informação sobre genomas parciais, no ProGenViZ foi também incorporada a possibilidade de visualizar e analisar ficheiros com múltiplas sequências provenientes de sequenciação destes genomas (contigs). Além de ser possivel aceder tanto às informações globais como realizar qualquer uma das procuras referidas anteriormente, foi também desenvolvida uma funcionalidade para ordenar os contigs contra um genoma de referência, o que fornece uma perspetiva global de quais e de que forma os contigs estão distribuídos ao longo da sequência de referência. Além disso, como normalmente as sequências parciais após serem geradas não têm qualquer anotação, foi também criada uma abordagem para anotá-las através de transferência de anotações de um genoma anotado de referência através da combinação dos resultados dos software Prodigal e BLAST. O Prodigal, um software de previsão de genes em procariotas, é utilizado para prever *coding sites* (CDS) nos contigs enquanto que o BLAST é utilizado para determinar se alguma *região* do genoma de referência tem similaridade com o gene previsto pelo Prodigal.

Ao terem sido criadas maneiras de estabelecer relações entre *regiões* de diferentes ficheiros foi fornecida ao mesmo tempo uma forma de monitorizar a qualidade das anotações através de similaridade de sequência. Como algumas anotações pré-existentes podem estar erradas, foi desenvolvida uma funcionalidade para re-anotar o nome e o produto das diferentes *regiões*.

Criámos também uma série de funcionalidades para exportar dados da aplicação. Podem ser exportados os resultados apresentados nas tabelas, imagens, sequências genómicas específicas, bem como toda a informação existente de *regiões* e sequências genómicas associadas a cada um dos ficheiros carregados na aplicação.

Para demonstrar as diferentes capacidades da aplicação são também mostrados três casos de uso. No primeiro caso de uso são procurados os genes pertencentes ao esquema MLST de *Streptococcus pneumoniae* em dois genomas anotados para focar as capacidades do programa de realizar procuras por genes através do seu nome, produto e sequência. Esta análise demonstra os actuais problemas das anotações automáticas onde nem sequer os genes essenciais para manter funções básicas da célula estão bem anotados. Foi também possível determinar a existência de inversões na localização de dois dos genes após análise da representação visual.

No segundo caso de uso são procurados os genes regulatórios inseridos no *locus* da biosíntese da cápsula do serotipo 1 de *Streptococcus pneumoniae* num ficheiro de contigs para ilustrar as capacidades da aplicação para encontrar regiões de interesse em contigs. Nesta

análise é possível encontrar todos os genes regulatórios bem como outros pertencentes ao mesmo locus num único contig.

Finalmente, no último caso de uso, dois ficheiros com sequências parciais obtidas depois de sequenciar dois organismos da estirpe *Streptococcus pneumoniae OXC141* e um genoma anotado da mesma estirpe são utilizados para mostrar as capacidades do programa para ordenar e anotar todos os contigs de um ficheiro contra uma referência. Com esta abordagem de transferência de anotação por homologia foi possível transferir de uma media de 87% de anotações da referência para os ficheiros de contigs.

Palavras Chave

Análise Visual; Genómica Comparativa; Procariotas;Sequenciação de alto débito; Anotação de sequências

Acknowledgments

First of all, I would like to thank my supervisors João Carriço and Octávio Paulo for all the support they gave me through all the work done is this thesis. Especially to João, I am grateful for all the opportunities you gave me and for all the knowledge and motivation you were able to get into me. I must say that you are a true inspiration and model on how work should be done in these areas.

I would also like to acknowledge all my colleagues and researchers in the Microbiology and Infection Unit group of Instituto de Medicina Molecular who welcomed me with open arms for the last months. I am very thankful for all their help, suggestions and contributions to this thesis.

A big thanks to all my friends who were always there when I needed to refresh my mind.

Many thanks to all my family, especially to my grandparents, for all the support though my entire life. I would like to apologize to them for not receiving as many attention as they deserved for the last months.

To Sofia, a thanks with the size of the world, for all the love, support, optimism and comprehension. Sometimes it was not easy, but you were always there for me.

And finally, a big thanks to my parents and my brother, to whom I dedicate this work, for everything they have done for me, since ever. Without your full support I would never be here.

Contents

List of Figures

List of Tables

1

Introduction

1 Introduction

1.1 Context and motivation

In the recent years, high-throughput sequencing (HTS) has revolutionized the methods to obtain genomic data. Currently (November 2014) in NCBI there are 3249 prokaryotic genomes classified as complete and 13808 classified as draft genomes and each day more sequences are made publicly available. This ability to quickly get large volumes of information make HTS technologies able to be used in different types of studies, in particular the characterization of organisms and detection of sequence variation.

In Microbiology, HTS technologies provide the ability to obtain microbial draft genomes in a reduced period of time which can be used in the development of genomic epidemiology studies, as well as to identify and characterize microorganisms by determining lineages associated to antibiotic resistance, pathogenicity and virulence that can assist in the detection and prevention of infectious diseases[1]. However, despite the existence of the methods required to obtain genomic data, currently the limiting factor is the capacity of the existing analysis methods to extract relevant information from data since there are still missing visualization tools to compare prokaryotic genomes in terms of gene content variation that offer interactive ways to explore data. Many of the software available are expensive or open-source freely available software that commonly has a high level of complexity to operate, which limits the analysis of the data.

The creation of data driven images has been one of approaches to cope with the data from genomic sequences. The visual representation of data can join a single image information of one or more genomes and helps determine characteristics of the data that otherwise would not be possible such as genomic rearrangements through global representations of genomes or even sequence level variation when we have visual representations of nucleotide sequences. Also, when used in combination with methods for comparison and querying of genomic data, different visualization methods can be used to facilitate and guide the identification of interesting features.

Ideally, a tool to visualize and compare genomic data has to be able to import different available data formats, export the data and resulting visualizations, allow the comparison, display and exploration of unique sets of genes or entire genomes, as well as promote an interaction between the user and the image in order to assess the information at different levels[2]. However, the existing sequence comparison tools lacked some of these desired features. Some do not allow users to upload their own data and create difficult to interpret

representations or require installation of specific programs to be used. Others lack interaction between the user and the program, not giving control of the comparisons that are made. Also, a large majority of genome comparison applications focus only on a global associations, forgetting the relationships between specific sites in the genome.

Having all these concerns in mind, the aim of this thesis is to provide a public web-application to visualize, explore and compare prokaryotic genomes and draft genome data. We also want to offer other functionalities such as the re-annotation of genes, ordering of draft genome sequences against a reference genome and subsequent annotation transference from one or more references, allowing draft genome annotation of coding sequences.

1.2 Contributions

The contributions of this thesis to the area are:

- A new user-friendly interactive interface for viewing, explore and compare multiple genomes and draft genomes of prokaryotes, incorporating the BLAST[3], NUCmer[4] and Prodigal[5] software.
- The creation of a novel genome abstraction method that uses two visual representations based on graphs (Hive Plot and Linear) to display multiple genomes and draft genomes of prokaryotes and comparisons between them.
- A visual interface to reorder draft genomes data against a reference genome, using already established methodology.
- The development of a system to annotate coding-sites (CDS) in draft genomes by transferring annotations from one or several annotated reference sequences.
- The development of a system to re-annotate genes.

1.3 Thesis outline

This thesis is composed of three distinct parts: background, developed framework and use cases. In the first part, we present a historical overview since the discovery of DNA to the development of the different sequencing technologies, some of the methods used to align and annotate sequences, as well as some of its applications. Also, we show the theory to be considered when building a data driven visual representation and the different categories of software currently available to visualize and compare genomic data. In the second part, the

functionalities of a new tool to visualize, explore and compare multiple prokaryotic genomes and draft genome data is described. Finally, some of the applications of the developed tool are shown.

2

Background

2 Background

In this chapter, we provide an historical overview of events since the discovery of DNA to the creation of DNA sequencing technologies, some of its applications in Clinical Microbiology and some of the methods used to align and annotate sequences. Then, the theory to develop a good visualization system, as well as the existing methods to visualize, explore and compare genomic data are discussed.

2.1 DNA sequences

Since the early days of scientific research, researchers attempt to discover the reasons behind different characteristics exhibited by a population of individuals. Gregor Mendel, the proclaimed father of genetics, gave the first steps in 1865 on this subject and described the laws of genetic traits transfer through the study of crosses between peas[6]. Knowing that something had to be responsible for the transmission of these characteristics, Ernst Haeckel in 1866 suggested that these factors would be located in the nucleus[7] and Friedrich Miescher isolated the first DNA molecule from leukocytes in 1869[8]. Parallel to the findings associated with cytology, new ideas and concepts associated with heredity and evolution emerged in 1858 with the Theory of Evolution by Natural Selection by Charles Darwin, that were later published in the book “On the Origin of Species by Means of Natural Selection”[9].

It was necessary to wait until the next decade for the DNA to be thoughtful studied by Avery, MacLeod and McCarty who proposed a hypothesis where the DNA would function as genetic material, contrary to the common belief that proteins fulfilled this role[10]. This results would be confirmed in 1952 by Hershey and Chase that used bacteriophages T2 to demonstrate that phagic DNA enters the bacteria while viral proteins not[11].

Despite the certainty of DNA as genetic material and its role in passing traits between organisms, its structure remained to be discovered. It was then, in 1953, that Rosalind Franklin and Maurice Wilkins, using an X-ray analysis, obtained the first data on the repetitive helix structure of DNA. Its double-helix molecular structure would finally discovered by James Watson and Francis Crick in the same year[12].

At that time, one of the puzzles that was still unsolved was how the replication would proceed in cell division. Many were the models proposed but it is the semi-conservative model, where each original semi-helix molecule function as a template for the production of two identical ones, that is currently accepted[13].

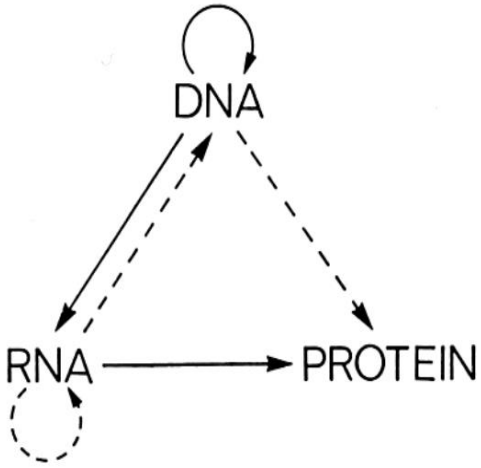


Figure 2.1: Central dogma of molecular biology proposed by Francis Crick in 1970, with some modifications to the one proposed by him in 1958. Reproduced from[14]

It was only after the proposal of the central dogma of molecular biology by Francis Crick in 1958[14], that would later be modified (Figure 2.1), that he and his colleagues would decipher the genetic code in 1961[15]. At the same time, new ways to manipulate sequences were being discovered that led to the possibility of using restriction enzymes to break DNA at specific sites and also to the production of the first pieces of recombinant DNA[16, 17].

Knowing that DNA was then linked to heredity and having now the ability to modify specific regions of sequences, it was necessary to begin developing tools to start characterizing the genomes of different organisms and to discover the differences between them.

2.2 DNA sequencing technologies

DNA sequencing techniques are very important in many areas of scientific research. A large number of scientific areas are taking advantage of these technologies, such as molecular biology, biotechnology, forensic science, genetics, ecology and environmental research[18–21]. However, a series of events were necessary to evolve from the sequencing of a very small number of nucleotides up to the megabases of information that are currently possible to obtain with the high–throughput technologies (HTS).

2.2.1 First DNA sequencing technologies

Early forms of DNA sequencing are quite time consuming, complex, and their laboratory work are very difficult and intense. The described sequence of the first 24 base pairs from the

lac operator was pioneer and it was a job made by Maxam and Gilbert with a method known as wandering spot-analysis[22]. From that point on, efforts where directed to reduce the complexity on how to obtain sequences. The first breakthrough was in 1977 when also Maxam and Gilbert described a method to reveal the nucleotides of a sequence through the cleavage of DNA sequences in specific sites by chemical degradation and consequent hybridization in electrophoresis gel[23]. In the same year, Frederick Sanger began to develop an efficient technique to sequence DNA, the chain-termination[24](Figure 2.2). The method consists in the use of a primer that anneals with the desired sequence, a DNA polymerase, and a series of dideoxynucleotide triphosphates (ddNTPs) that inhibit chain extension to create sequences of various lengths. By using this method, it is possible to determine which position belongs to each nucleotide in the sequence only by their disposition at the electrophoresis gel. Through the development of these techniques and for the pioneering work that led, as consequence, to the creation of improved methods for DNA sequencing, Sanger and Gilbert shared the Nobel Prize for Chemistry in 1980[25].

During the next decade, Sanger's sequencing was widely adopted by the community and was improved, mainly by the automation of the process. The first automated sequencer used a modification of the Sanger's method that consisted in the use of specific fluorophores linked to each one of the nucleotides with its detection and interpretation being made computationally[26].

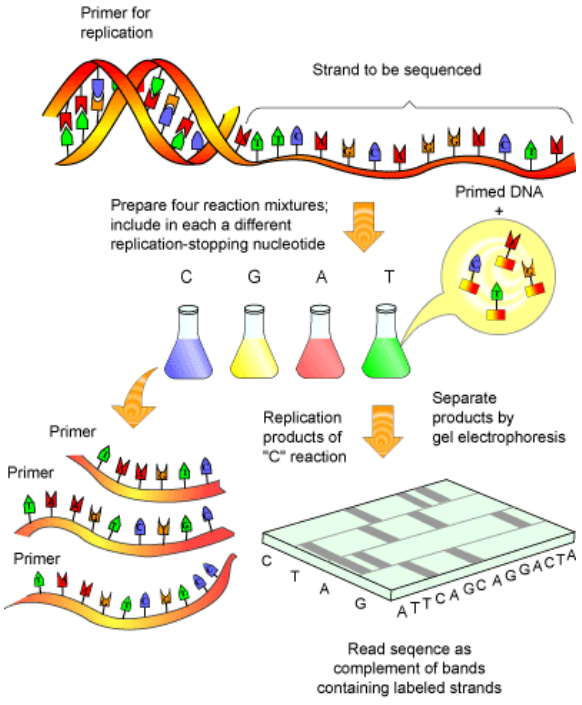


Figure 2.2: Representation of the Sanger's method to sequence DNA.

2.2.2 Human genome project

The Human Genome Project was a collaborative large-scale project which aimed at mapping and understanding all genes of the human being. The efforts to complete this project made it possible the development of new technologies, development of genomic maps of various organisms, as well as a well-designed sequence of the human genome[27].

The first serious thought to the possibility of sequencing the human genome were expressed in 1985 by the Director of University of California at Santa Cruz at the time, Robert Sinsheimer[28]. The idea would be considered a bit premature given the lack of resources to develop the project. However, in 1988, the U.S. National Research Council of the U.S. National Academy of Sciences proposed the beginning of the Human Genome Project (HGP), with a deadline of 15 years, despite the high cost of sequencing a nucleotide base at the time[27].

Since the early development of the project, one of the priorities was the creation of new methods to reduce the sequencing cost and increase the number of nucleotides sequenced per time unit[29]. Several steps were taken to reduce the necessary human intervention and thus make the process as automated as possible. Some factors were crucial to improve the sequencing processes, especially the emergence of commercial sequencing machines and the improvement of sequence assembly procedures[30].

With the emergence of capillary sequencing, which sequences DNA through a modification of the Sanger method and analyses several samples simultaneously[31], and also by the development of techniques that improved the data's quality and throughput, such as the shotgun sequencing procedure[32], improved fluorescent dyes[33, 34] and specific polymerases for sequencing[35], it was possible to give a boost to the success of the HGP. Due to the cooperation between various groups, it was also possible to develop the idea of "open" culture and information sharing between researchers, technologies and software[27]. This was the first time that software played a major role in the determination of sequence similarity and assisted in genome assemblies.

In 2001, the first drafts of the human genome were published, with about 90% of the full sequence[36, 37]. In the following years, studies were carried out to increase the coverage and quality of those results[38].

For over two decades, sequencing was dominated by Sanger's automated method. Although the evolution of techniques during this time led to the sequencing of the human genome, limitations associated with the expensive cost per base have shown that the creation of new technologies was needed to achieve the sequencing of various genomes in less time[39].

To make this possible, the priority was to develop tools capable of producing greater volumes of information, with greater coverage and lower financial costs, in order to produce whole-genome sequencing more quickly and with more quality. To achieve that goal, the HTS technologies were developed.

2.2.3 High-throughput sequencing (HTS) technologies

HTS technologies are a group of sequencing methodologies characterized by the production of a large amount of genomic data in a short period of time. They generate megabases of genomic information in the way of small DNA fragments – reads - which are then assembled into larger sequences – contigs - using specific assembly software. Currently, there are several HTS platforms available (Figure 2.3) that are organized into two major groups: those who need template amplification and those who use single molecule sequencing. Information about the technical aspects of the different technologies can be found described elsewhere[40, 41].

The differences between each sequencing technology are mostly associated with their monetary cost, read's size and sequencing quality. However, has to be said that the sequencing error probability of these platforms continues superior when compared to machines with Sanger technology[42].

Technologies that need template amplification – also called next-generation sequencing technologies (NGS) - differ primarily in 3 parameters: library creation, template amplification, and sequencing method. According to the amount of data that are capable of producing, platforms belonging to this category can range from the expensive high-end instruments that generate massive amounts of sequencing data such as the HiSeq instruments, Genome Analyzer IIx, SOLiD 5500 series and the 454 GS FLX+ system, to the most recent bench-top instruments with less throughput but ideal for rapid sequencing analysis and for microbial applications, such as MiSeq, 454 GS Junior and Ion Personal Genome Machine. On the other hand, single molecule sequencing platforms use a real time sequencing approach that as the advantage of eliminating the possible artefacts generated by sequence amplification and ends the need to make library preparations[43]. However, they have a high error rate and are extremely expensive. The Helicos BioSciences' HeliScope Single-Molecule Sequencer was the first single-molecule sequencing platform and more recently was developed the PacBio RS from Pacific Biosciences.

All these technologies revolutionized the sequencing methods and now we have the ability to produce genomic data thousands of times more cheaply than is possible with Sanger sequencing, something that would have been unthinkable a few years ago. These technologies

led to the current ability to obtain sequences of entire genomes and analyse them. In the next sections are described some of the whole genome sequencing (WGS) applications in Microbiology and also some of the existing methods to analyse DNA sequences such as sequence alignment methods and the strategies used to discover and infer gene locations in the sequences. Moreover, some of the existing software to visualize, explore and analyze sequencing data are described.

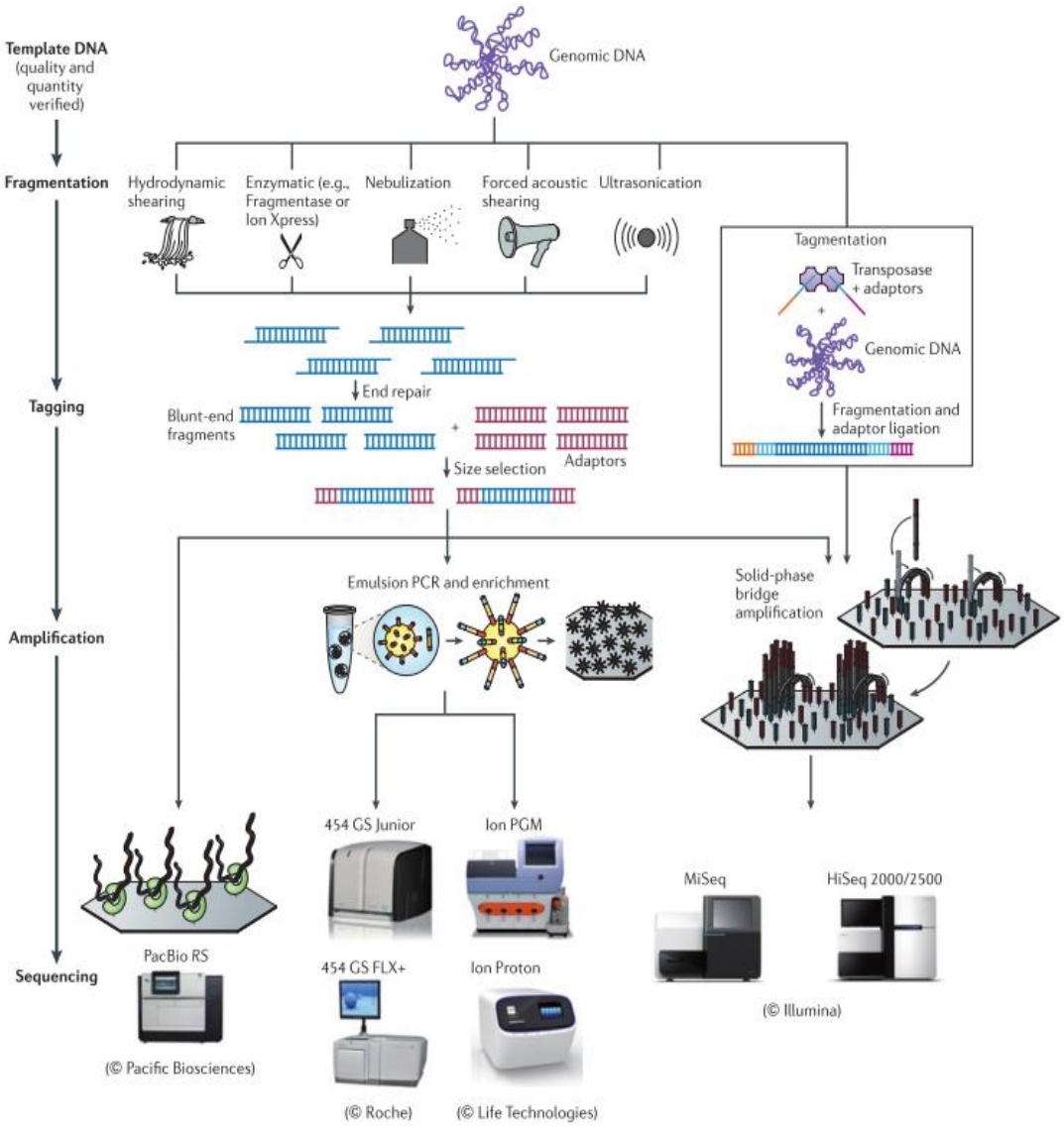


Figure 2.3: HTS platforms. Are described the different library preparation procedures and the methods used to amplify DNA. Reproduced from[41]

2.3 Whole genome sequencing (WGS) and Microbiology

HTS technologies enabled the development of whole genome sequencing (WGS) which leads to the generation of draft genomes that can be applied in different branches of Biological Sciences. Nevertheless, it is through the comparison of multiple genome sequences that relevant facts in the genome organization can be found. Comparative Genomics is a large-scale, holistic approach that compares two or more genomic sequences to discover the similarities and differences between them and to study the characteristics of the individual genomes[44]. Through this methodology, it is possible to study genomic rearrangements, find orthologs or paralogs, and compare gene content, among others.

One of the research areas that has benefited both from WGS as Comparative Genomics is Microbiology. The ability to sequence entire microbial genomes in a short time with bench-top sequencers and the identification of genomic regions of interest by comparing with reference sequences, is something that is currently within reach of all laboratories. Especially for bacteria, Comparative Genomics is being used for the fast identification of strains and to infer their evolutionary history, as well as to discover novel virulence factors and vaccine targets[45].

The potential of WGS for diagnosis and epidemiological studies was recognized in the last years[46, 47]. Although in the present WGS is still expensive to be used commonly in clinical microbial laboratories, its application in clinical samples in the future could reduce diagnostic times and thereby improve disease control and treatment. In the last few years, WGS has already been applied in outbreak investigations[48, 49]. Moreover, approaches to determine the efficacy of a comparative analysis to detect strain manipulation leading to increased virulence or antibiotic resistance in case of epidemic outbreak or a bioterrorist attack have been studied[50].

WGS is also revolutionizing molecular genotyping methodologies, mainly the sequence base typing methods[47]. Currently, methods like MultiLocus Sequence Typing (MLST) use only specific fragments of seven housekeeping genes to identify strains. This method, due to its limited number of target loci, as shown some lack of discriminatory power for resolving closely related strains, while providing an excellent tool for global population analyses. By analysing a broader range of loci on the scale of hundreds or thousands, WGS is proving to be a disruptive technology in the field, offering the highest discriminatory power available for epidemiological studies[47].

All these new perspectives that the combination of comparative genomics with WGS provide, not only for Clinical Microbiology but also for numerous other areas, will surely

revolutionize the genetic analyses that are made today. However, we are still on the early days of this novel technology and there is a long road ahead. We must develop better automatic annotation pipelines to deal with the continuous increase of genomic data and reduce the current persistent annotation errors. Moreover, we must create novel algorithms for quality control to guide the needed manual curation analysis. Also, visualization and genome mapping tools demand less complexity of use and a better representation of genome structures at different levels. The growth of Comparative Genomics analysis and its effective application in research and clinical settings will depend on how fast we can overcome these limitations.

2.4 Sequence alignment and annotation

A great variety of tools have been developed during the years to carry out studies using sequencing data since there is a need to analyse the information obtained from HTS. Sequence comparison, alignment, and annotation are some of the areas that need specific methods and software to produce the best results possible, more quickly and efficiently[2].

2.4.1 Sequence alignment algorithms

The main tools for sequence analysis and comparison are the sequence alignment algorithms. Sequence comparison allow users to obtain results that are biologically relevant such as search for orthologous genes[51], detection of variants[52], establish evolutionary relationships[53] and produce genome assemblies[54]. In these tools, most of the algorithms are grouped in one of three categories: algorithms based on hash tables, algorithms based on suffix trees and algorithms based on merge sorting[55]. This classification depends on the auxiliary data structure that is used in those algorithms.

One of the most popular algorithms to align sequences is BLAST[3] (Figure 2.4), an algorithm based on hash tables which allows comparisons between a query sequence and a database of sequences to find regions of local similarity. BLAST algorithm is available at the National Center for Biotechnology Information (NCBI) and may also be used locally. There are several programs based on BLAST that can be used, adapted to different types of sequences and operations. Some are used to compare nucleotides (BLASTN), others to make comparisons between protein and nucleotide sequences (BLASTX) or just between proteins sequences (BLASTP). The algorithm works in three steps[56] and many variables can be changed to adjust the sensitivity and speed of the comparisons between the query sequences and the database.

In the first step, there is a filtering of low complexity regions and then the query sequence is divided into sub-sequences. In a second stage, the previously generated sub-sequences are searched in the database and the results act as seeds for the determination of high-scoring segment pairs (HSPs). The search for these sub-sequences of defined size reduces the total number of comparisons that are required. Finally, there is a merge of seeds without gaps, followed by an extent on both sides using the Smith-Waterman algorithm[57] to find the best alignment possible. Only the alignments with a score higher than the cut-off score (S) determined are listed and returned as maximal scoring pairs (MSPs).

BLAST has been one of the most essential tool for research in Biological Sciences. Revolutionized the way of how to do analysis in various fields of research, answering several questions that could not be answered in laboratory and made the bioinformatics analysis accessible for researchers around the world. However, other alignment tools have emerged that have some useful characteristics for certain kinds of studies. Algorithms based on suffix trees have an approach which seek for the reduction of inexact matching by first identifying exact matches and only then build inexact alignments. One of those tools is MUMmer[58] package, a software that allows from alignment of entire genomes, and also alignment of contigs against a reference using NUCmer algorithm (nucleotide MUMmer)[4]. MUMmer has an approach that combines suffix trees, longest increasing subsequence (LIS), and the Smith-Waterman algorithm, to try to find regions that are exactly equal between two sequences (maximal unique matches - MUMs). Those regions function as starting points for the alignment. In the case of NUCmer (nucleotide MUMmer)[4], it uses MUMmer to map contigs against a reference sequence. Then, it uses a clustering algorithm on the MUMs to determine their location in the reference.

With the use of HTS and the ability to produce draft genomes in a short time, contigs can play an important role in the detection of genetic variations among organisms by detecting specific regions within them. However, there are few tools available that use contig files and allow their comparison with other sequences.

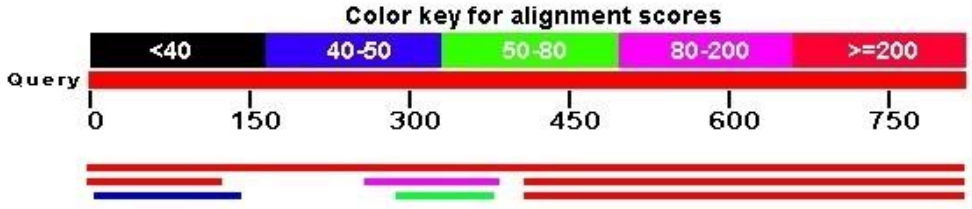


Figure 2.4: BLAST results visualization from the NCBI website (http://blast.ncbi.nlm.nih.gov/Blast.cgi). Different colours represent different scores.

Other alignment algorithms have been created to deal with specific sequencing data[59, 60]. Two of those alignment applications are inserted on the third category of alignment tools, the ones based on merge sorting. Slider[61] and Slider II[62] are programs that were developed specifically to improve the alignment and SNP detection of the Illumina's output. They use an auxiliary table with pre-defined sized fragments from the reference sequence, which are then sorted in a lexicographically form. After this, read alignments with exact matches and one-off matches are determined and the SNP prediction takes place.

2.4.2 Sequence annotation

Obtaining the DNA sequence is just the starting point. The main goal is to retrieve its information. Annotation can be defined as "a process by which structural or functional information is inferred for genes or proteins"[63] and it is essential for sequence interpretation. Because of that were developed pipelines to annotate genomic sequences (Figure 2.5).

Usually, sequence annotation relies on an automated annotation and a posterior manual curation[64]**.** However, at the rate that genetic information is produced, it is impossible to annotate all genes manually.

One of the critical steps to produce automatic annotations is the prediction of genes. There are several programs for this purpose but each of them has characteristics that make them more suited for the use in a particular species or type of gene that is intended to predict[65]. In the case of Prokaryotes, genes are organized in specific ways, having characteristic elements such as transcription promoter and terminator, operator, ribosome binding site (RBS), and start and stop codons between open reading frames (ORF). In order to predict the gene locations, the developed algorithms are based on the detection of these elements, mainly ORFs, in order to get an idea of how genes are organized and distributed throughout the genome. One of the programs that allows the

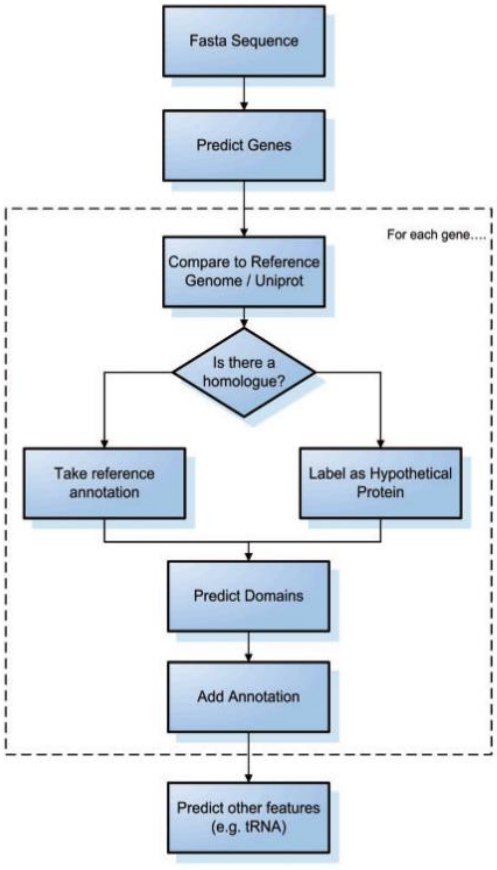


Figure 2.5: A generic process for genome annotation. Reproduced from[107]

identification of genes in prokaryotic sequences is Prodigal[5]. It uses a “trial and error" approach to search for genes, which begins with the search for all start and end codons included in the sequence. Then, a score is assigned to each possible gene which takes into account the bias of GC and the size of the open reading frame.

After the identification of possible gene locations, the search for homologs is typically made by BLAST-based comparison of sequences[66, 67] and then the annotation from the best hits are transferred.

Annotated genomes can be used to detect differences between organisms. However, the accuracy of automated methods used to assign annotations have been questioned over time due to errors that have been accumulating in databases[68]. Errors can emerge at different stages of the annotation process: during sequencing, as a result of gene-calling procedures, and in the process of assigning gene functions[68]; and can lead to misleading results in different types of analyses. Therefore, in order to opt for an approach to detect and compare genomic regions by their annotations, these errors have to be diminished through methodologies to detect them and then re-annotate those genomic regions efficiently.

2.5 Genome data visualization

With the increasing ability to obtain whole-genome data, the need to develop tools to visualize, explore and analyse it as increased drastically. In the following sections, some design principles and theories to take into account when building a visualization software are described as well as some of the existing tools available to visualize, explore and compare sequencing data.

2.5.1 Visualization theory

Although many genomic data analysis can be carried out automatically, the large amount of information and complexity of the results make their understanding a difficult process and human judgment is often needed to interpret the results in the light of biological knowledge. The creation of visualization methods is one of the paths to overcome these problems.

Visualization has a very important role in human perception. The human mind can process very complex information through the use of vision. We managed to get more information by sight than by all other senses combined. The human visual system is great to look for patterns recognition, and the process is facilitated through the specific visualization techniques. Although we can create mental images, the thought process is facilitated when we have diagrams, maps,

information graphs or other way of data representation at our disposal that allow us to solve problems through visual thinking[69]. According to Ware, information visualization is “the use of interactive visual representations of abstract data to amplify cognition", promoting mental operations with rapid access to information derived from images[69].

One of the great advantages of using images to represent data is the ability to gather large amounts of data, being possible to understand the information given by thousands of objects together. It also promotes the perception of properties of the data that would otherwise be difficult to understand, like artefacts and other errors, thus also functioning as a way of quality control.

Visual Analytics is a field of study that promotes the connection between the human and the data through the use of visual interfaces in order to obtain information the easiest way possible[70]. It also tries to increase the humans’ capacity to understand and reason about complex data, revealing at the same time some relationships that can be unexpected[70]. It was defined by Thomas et.al as “the science of analytical reasoning facilitated by interactive visual interfaces”[71] and it as the following main focus areas of study:

- Creation of visual representations and interaction techniques to exploit the human eye’s broad bandwidth pathway to let users see, explore, and understand large amounts of information simultaneously.

- Use data representations and transformations that convert all types of conflicting and dynamic data in ways that support visualization and analysis.

To create efficient forms of visualization, it is necessary to know the characteristics of the data that will be used and the operations to be carried out. To help understand how a visual representation should be made in order to improve cognition, during the years were developed a group of representational principles (adapted from[70]):

- **Appropriateness Principle** – The visual representation should provide just the information that is needed for the task. Additional information may be distracting and makes the task more difficult.

- **Naturalness Principle** – Experiential cognition is most effective when the properties of the visual representation most closely match the information being represented.

- **Matching Principle** – Representations of information are most effective when they match the task to be performed by the user.

- **Principle of Congruence** – The structure and content of a visualization should correspond to the structure and content of the desired mental representation.

- **Principle of Apprehension** - The structure and content of a visualization should be readily and accurately perceived and comprehended.

An information display system needs to have two main components: representation and interaction[72]. The representation is associated with the way that the data is arranged and displayed, while interaction involves a "dialogue" between the user and the system in order to analyse the information. This is quite important because through interaction, the limits of a static representation may be exceeded, further enhancing cognition. In his book[69], Ware states that data visualization is divided into four phases: collection and storage of information, pre-processing to transform data into comprehensible information, image production, and finally the interaction with the perceptual and cognitive system of the human (Figure 2.5). Ware also says that there is an interaction between the different phases through feedback loops,

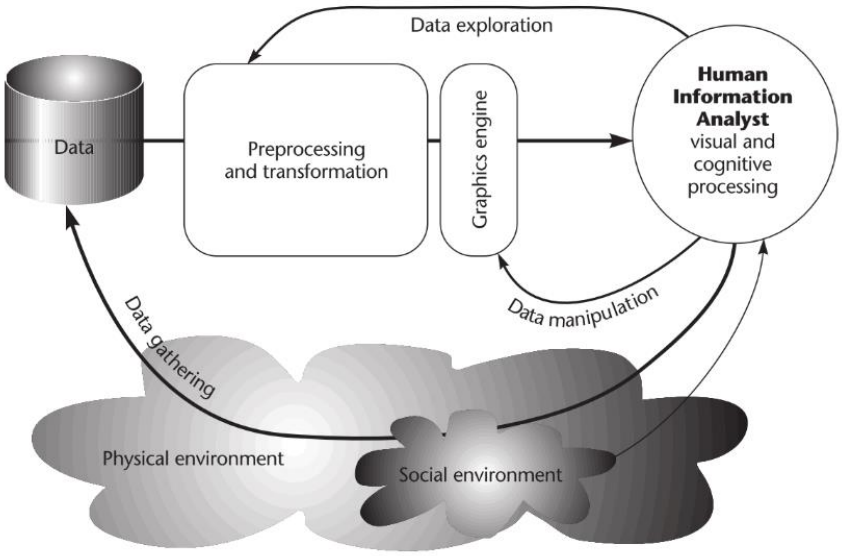


Figure 2.6: Ware's diagram of the visualization process. The human interact with the different data visualization stages through feedback loops. Reproduced from[69]

which correspond to possible interactions of the user with the viewing system leading to its modification.

There are a number of established techniques to provide an effective interaction interface. Dam et al. defined an interaction technique as the way of using a physical input or output to promote the realization of a task in a dialogue between human and computer[73]. Shneiderman[74], after summarizing what for him should be a framework to follow for the design of information visualization applications: "Overview first, zoom and filter, then details-on-demand", developed a taxonomic system to classify interaction techniques that help to understand how interaction can be added to a visual representation of data:

- **Overview**: Gain an overview of the entire collection.
- **Zoom**: Zoom in on items of interest.
- **Filter**: filter out uninteresting items.
- **Details-on-demand**: Select an item or group and get details when needed.
- **Relate**: View relationships among items.
- **History**: Keep a history of actions to support undo, replay, and progressive refinement.
- **Extract**: Allow extraction of sub-collections and of the query parameters.

Since then, several studies have proposed taxonomies with different levels of specificity that can be consulted elsewhere[72].

Visualization increases the ability to make sense of very complex data groups, such like the information generated by sequencing technologies. This makes the manual inspection of the data and the analysis of the results easier. It can also be complemented with the use of automatic methods of data analysis to successfully deal with large genomic datasets[2].

2.5.2 Genomic sequences and whole genome visualization

The study of DNA sequences and whole genomes of organisms became possible due to the great advances in sequencing technologies. However, to be able to make sense of all the information that is generated, it was necessary to create visualization methods to analyse the complex data that is produced. In recent years, several specific tools – alignment viewers, genome browsers, comparison software - were developed. There are a large number of applications that fall into one of these three categories and each of them has certain characteristics that makes them more capable of performing specific analysis. A summary of some of the tools that can be currently used can be found in the Table 2.1.

To specifically view reads alignment, assembly viewers such as EagleView[75] and MapView[76] were developed. This software category deals with large amounts of information and focus primarily on the ability to navigate, providing visual ways to test the alignment quality and to detect sequence variation. The representation of complete genome sequences remains a complicated task for these type of tools but they try to surpass their difficulties by an interactive navigation and through a division of the genomic sequence into a series of sections. This allows an optimization of the use of computer memory and increases the processing speed.

To facilitate the exploration and analysis of results after an assembly or to explore complete genomes, a series of tools called genome browsers were produced. These programs are characterized by disposing sequencing data or genomes and their annotations with the help of a graphical interface, and by enabling the analysis of specific regions of interest. One of those tools is the Integrative Genomics Viewer (IGV)[77], which enables the analysis of multiple genomic regions simultaneously and allows the visualization from complete genomes to specific sequences using different levels of complexity. There are also genome browsers operating as web applications, such as the JBrowse[78]. Both genome browsers and alignment viewers have problems related to the large amount of data that can be disposed and the maximum number of genomes that can be visualized at the same time. In the future is necessary to create ways to browse and filter through the information you want to view, and enable data and visual representation editing.

In the last years were also created a series of sequence comparison software to visualize relationships between genomic data - adapted or not to the comparison of microorganisms – that use different ways to represent comparisons. Some are more useful for global genome comparisons, while others to local ones. VISTA[79] is a web-based application that represent global comparison between two genomes, showing regions with peak identity. Software like Circos[80] and BLAST Ring Image Generator (BRIG)[81] display multiple whole genome comparisons and are characterized by a circular arrangement of information which are more suited to represent global comparisons between genomes. Others use linear representations of sequences which are more focused for local comparisons. In these cases the comparisons are represented by bands or lines. SynBrowse[82] is a web-based tool that shows global and microsynteny between two genomic sequences and allows to browse for annotations and specific comparisons. Artemis Comparison Tool (ACT)[83] and Genome Synteny Viewer (GSV)[84] allow the comparison of two or more genomes, showing sites of local similarity using bands in a horizontal linear layout. ACT compares the sequences through the use of BLAST[3] or parses files directly from other comparison tools like MUMmer[58], while GSV needs a synteny

Type	Name	Brief description
Alignment viewers		
Standalone	Hawkeye[108]	Visual analytics tool for genome assembly analysis and validation; identification of assembly errors.
Standalone	IGV[77]	Genome browser with alignment view support; supports a wide variety of data types, including array-based and next-generation sequence data, and genomic annotations.
Standalone	MapView[76]	Read alignment viewer; allows users to see the mismatches, base qualities and mapping qualities.
Web-based	LookSeq[109]	Supports multiple sequencing technologies and viewing modes; easy visualization of single nucleotide and structural variation
Standalone	Tablet[110]	High-performance graphical viewer for next generation sequence assemblies and alignments.
Genome browsers		
Standalone	CGView[111]	Java package to generate high quality, zoomable maps of circular genomes; Primary purpose of generate visual output for the web.
Web-based	JBrowse[78]	Fast, embeddable genome browser built with HTML5 and JavaScript.
Web-based	GBrowse[112]	Combination of database and interactive web pages to manipulate and display genome annotations.
Web-based	UCSC Genome Browser[113]	Comprehensive genome browser and database.
Comparison viewers		
Web-based	Cinteny[85]	Synteny identification and analysis of genome rearrangement; three-scale view of synteny calculated from user-specified markers.
Web-based	VISTA[79]	Comparative analysis of genomic sequences; conservation tracks connected to a variety of analysis tools.
Standalone	ACT[83]	Linked-track views; annotation track search; stacking of multiple genomes.
Standalone	Circos[80]	Circle-graph presentation of synteny; animations for increased dimensionality.
Standalone	Combo[86]	Dot-plot and linked-track views; integration of annotation in both views.
Standalone	SynBrowse[82]	Local synteny based on gene order, orthology or structure.
Both	SynTView[87]	Multiple data representations; genome visualization and comparison.
Standalone	BRIG[81]	Generate images that show multiple prokaryote genome comparisons as a set of concentric rings.
Web-based	GSV[84]	Presents two selected genomes in a single integrated track view for synteny visualization; requires a synteny file.

Table 2.1: Summary of some of the currently available alignment viewers, genome browsers and comparison viewers. Adapted from[2]

file to view the comparisons. Cinteny[85] is able to represent global and local synteny among multiple genomes on 3 levels of complexity, also providing reference genomes to visualize relationships between them. Another way to visualize comparisons is through dot plots. Combo[86] shows whole genome comparisons and provides two ways to view comparisons, by dot-plot or by horizontal linear layout. Also, it supports annotations that are arranged along the axes. Finally, SynTView[87] is a web-based/desktop application to visualize genomes and comparisons of microbial organisms, which offers several ways to represent data.

3

Developed Framework

3 Developed framework

3.1 Overview

ProGenViZ is an open-source freely available web tool to compare prokaryotic genomes and HTS contig data that provides an interactive way to explore genomic data and to visualize global and local relationships between genomic regions. Moreover, it provides additional features such as the re-annotation of genes, ordering of contigs against a reference and annotation of contigs by transfer from an annotated sequence.

Throughout the description of the developed framework, italic words will mark commands that can be accessed in the application or some new terms used to identify and explain certain features of the program.

Source code is available at https://github.com/B-UMMI/ProGenViZ and the tool is available at http://darwin.phyloviz.net/ProGenViZ.

3.2 Implementation

ProGenViZ was developed using a client-server approach. On the client-side we have the processing of visualization and user interaction through a web browser, while on the server-side we have all the operations leading to the creation of the basic data structures needed for visualization representation.

The Bootstrap framework[88] was used to develop the basic structure of the web application and D3 JavaScript framework[89] to carry out all the operations associated to the creation of visual representations and user interaction.

On the server-side, in order to process genomic data, we used Python[90] scripts to parse all input files and convert them to JavaScript Notation Format (JSON)[91], BLAST to search for genomic sequences, Prodigal to predict prokaryotic gene locations, and MUMmer to order contigs and find single nucleotide variations.

In the following sections we provide a more detailed description of several implementation aspects.

3.2.1 Input processing

ProGenViZ accepts three distinct file formats as input: the GenBank/EMBL format, which provides the genomic sequences and their features, the General Feature Format (GFF) which only provides information about features of genomic sequences, and the FASTA format which gives only the genomic sequence itself. More detailed information about the different file formats can be found elsewhere [92–94].

To process each of these formats, we use Python scripts to create two JSON files required to create the genomic data representation and perform other tasks. One of the JSON files as information about the genomic features, while the other as the genomic sequences itself if applicable.

Because the GFF format does not contain genomic sequences, we offer an additional option to upload GFF and FASTA files together. When this happens, we merge the information of the genomic sequences provided by the FASTA file with the features provided by the GFF.

We also developed a different kind of input processing for files that have more than one genomic sequence. In the case of FASTA files with multiple contig sequences, an additional step is taken in the input processing to add a specific attribute to the JSON file, which uniquely marks each sequence. This approach is essential to represent each individual sequence properly in the place reserved for the uploaded file in the visual representation.

3.2.2 Main work area

After the user uploads the first file, they are directed to the main work area. This area is divided into two parts: actions menu and visual representation area.

The actions menu gives access to a group of features that the user can use to explore and extract information of the uploaded files and to control some aspects of the visualization (Figure 3.1-a). The different functions of each action will be described throughout this chapter.

In the second part, the visual representation area, is where the representation of files will be displayed. The way of how genomic data is shown to users is described in the following sections.

3.2.3 Genomic data visualization

To be able to view several complete prokaryotic genomes in a single image we had to create an abstract representation of the genomic sequences and their annotations to reduce the complexity of the visualization (Figure 3.2). To do this we created two levels of abstraction.

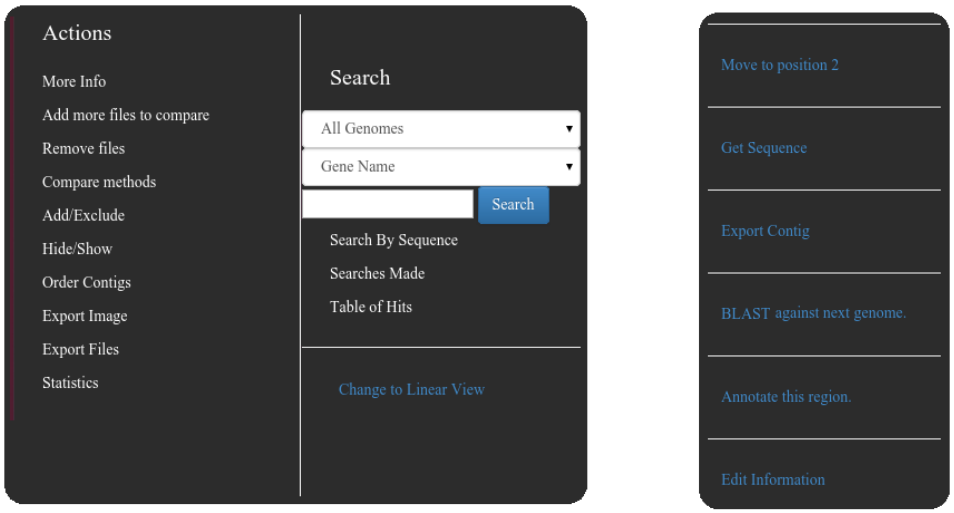


Figure 3.1: The two distinct menus of the application. (a) Actions menu. (b) Right-click menu that is activated when the user interacts with the visual representation.

First we used an approach where we divided all genomic sequences into *regions* according to their annotations. These annotations can be coding sites (CDS) and non-coding sequences that generate products such as tRNA and snRNA. Since not all *regions* of a genome are associated with an annotation, non-annotated *regions* are classified as undefined.

The second abstraction level was the division of all *regions* into intervals of 500 nucleotides, which we define as *nodes*. A *node* is thus the minimal size representation for a *region* in this tool. Therefore a *region* will be represented as many *nodes* as multiples of 500bp corresponding to its size. *Regions* with less than 500bp are still considered a single *node*. It is important to notice that what we achieve is an approximate representation of the length of the genome data and not a real one. *Nodes* are then represented as ellipses in all visual representations that are created.

We used the D3 JavaScript library as framework to develop the two ways to visualize genomic data. D3 allows to create powerful visualization components based on data and here it was used to transform all genomic data into an interactive representation.

In this tool, the representations of the genomic data and relationships between them are based on *graphs*. A *graph* is a representation of a set of objects, usually called *vertices* (singular *vertex*), where the relationships that exist between them are established by *edges* or *links*. In both visual representations developed each vertex has information about a single *node*.

The main visual representation in ProGenViZ is based on Hive Plots[95](Figure 3.3-a). Hive Plots are characterized by displaying vertices in a linear layout and by clustering different

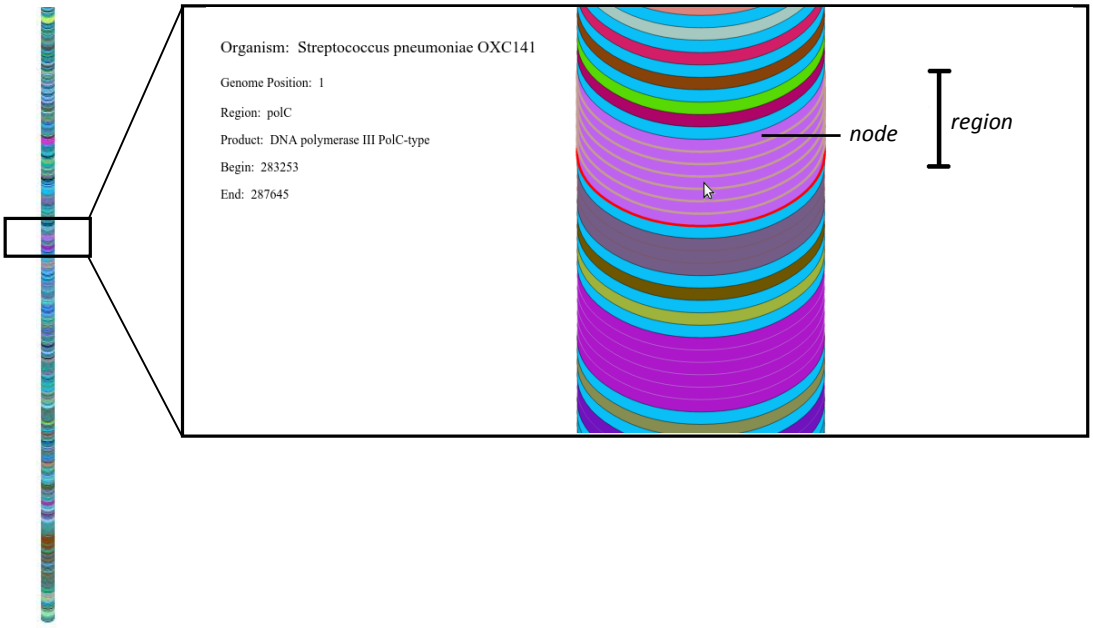


Figure 3.2: Representation of a part of the *Streptococcus pneumoniae OXC141* genome and definition of *node* and *region*. Different colours represent different products. In the case indicated in the image, the *region* corresponding to PolC gene consists of seven *nodes*, which means that the gene has a length between 3500-3999 base pairs. Mouse-over the region shows the information displayed on the left side of the visual representation.

vertices along a radial axis according to some property of the data[95]. We used the implementation in D3.js of Hive Plots by Mike Bostock's[96] as template for developing this visual representation.

To represent the different genomes in the Hive Plot we grouped the different *nodes* according to the file to which they belong. Depending on the number of files, the radial axis is divided into equal parts and the different linear layouts with all *nodes* belonging to a given uploaded file are disposed counter-clockwise. Near the centre of the display, we have *nodes* corresponding to the initial *regions* of the genome.

On the second visualization mode, the Linear representation (Figure 3.3-b), all *nodes* from different files are shown in a linear layout with a pre-set distance between them. *Nodes* are disposed from left to right and files added later are placed below the ones already displayed forming a stack.

On both visual representations, if a file is classified as having multi-FASTA sequences, all sequences are displayed in the same linear layout but separated from each other (Figure 3.2-c) in the same order provided in the file.

The interaction by the user with the visual representation is made through zoom and padding operations, and also by mouse-over interaction with the different *regions*. When the mouse pointer hovers a particular *region*, information about it pops-up in the upper left side of

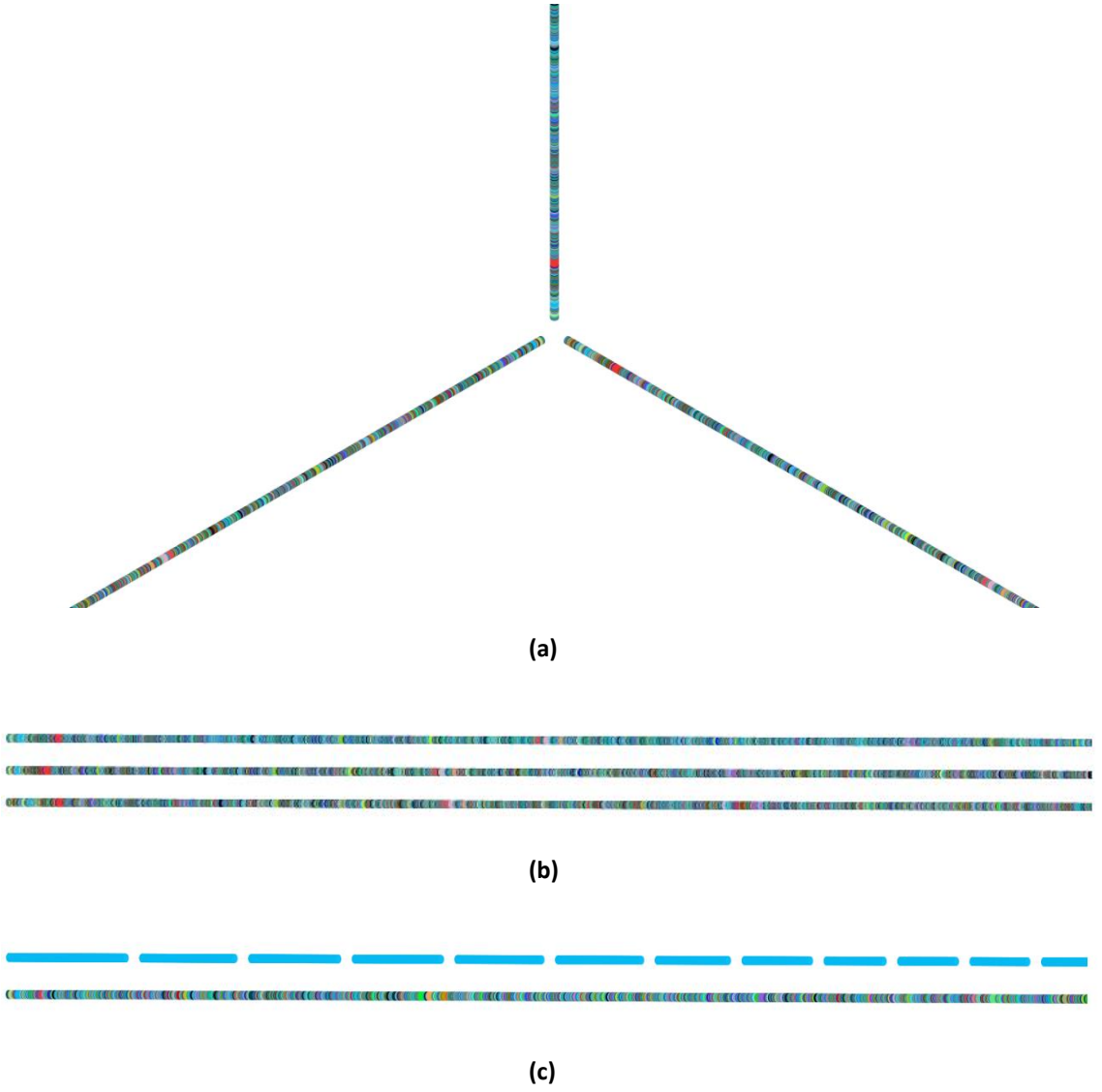


Figure 3.3: The two visual representations developed. (a) Part of the Hive Plot visual representation with 3 annotated *Streptococcus pneumoniae* genomes. (b) Part of the Linear representation with 3 annotated *Streptococcus pneumoniae* genomes. (c) Differences in the visualization of a file with contigs of *Streptococcus pneumoniae* (in blue) and an annotated genome.

the visualization area and all *regions* sharing the same annotation attribute product get highlighted (Figure 3.2). Also, by right clicking in any of the *regions*, a menu offers a series of operations than can be performed by the user (Figure 3.1-b).

Simple transitions can be made between the two visual representations by the actions menu's "*Change to Hive plot/Linear view*" option provided by the interface. Also, coupled with the ability to move between the two different views, the user has the possibility to reorder the files location in the representation by right clicking with the mouse in any of the represented files and by choosing the desired position. This feature is very important in this tool because it

is what enables the comparison of *regions* from one file with any other. Queries on different *regions* are discussed in the next section.

In order to distinguish between different *regions* we designed a colour based scheme to be assigned to them. The undefined *regions* are displayed in blue and different colours are assigned to annotated *regions* according to their products. These colours are randomly generated and their total number equals to the distinct products that exist in the analysis. Moreover, the colour scheme is updated whenever the number of products increases in an analysis, which is usually the case when a new file is added.

The application provides features to highlight or remove certain *regions* of the visual representation. Since a large portion of currently annotated genomes are genes classified as hypothetical proteins, the interface provides the option to highlight or remove from the analysis all genes with products classified as hypotheticals proteins. This is done by selecting the option "*Hide/Show hypothetical proteins*" on the actions menu of the interface to highlight those genes or by selecting the option "*Add/Remove hypothetical proteins*" to remove them from the visual representation. By choosing one of the options, a search is made in client-side for *regions* classified as hypothetical proteins. Those *regions* become red and with larger ellipses if the user choose the option to visualize hypothetical proteins, or are simply removed from the analysis leaving gaps in their location if the user choose the option to remove them.

The interface also offers an option to filter the *regions* that are displayed in the visual representation. The user can perform selections by mouse dragging through the *nodes* and by selecting the "*Use Selection*" option on the mouse right-click menu. Multiple selections can be made simultaneously by pressing the *ctrl k*ey. Selections can also be removed by the user when desired. This option was developed with the purpose of avoiding information overload in the visual representation when the user performs queries, by only displaying the selected *regions* while hiding the remaining sequence.

3.2.4 Querying on genomic data

We developed two distinct approaches to obtain information from the genomic data used for analysis: by providing global statistics, and by querying on specific *regions*.

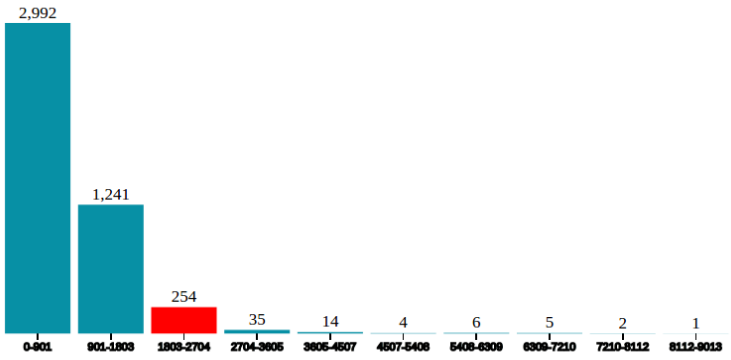
3.2.4.1 Global statistics

For every file uploaded, the user can access to general information regarding the files they uploaded (Figure 3.5-a). They can access to the total size of the genomic sequences and its percentage that is annotated by pressing the "*More Info*" button on the interface's action menu.

Is also presented information about specific genome features such as the number of transposases, the number of insertion sequences (IS), and the overall percentage of annotated genes corresponding to hypothetical proteins. Other statistics could be easily added to the interface as needed.

Other statistics can also be viewed through a representation of the distribution of the *regions'* sizes and products (Figure 3.4). The user can access to this information by selecting the "*Statistics*" button provided by the actions menu. The visual representations of these statistics are made by using a pie chart to show the distribution of *regions'* sizes and a bar chart to show the distribution of products. The colour corresponding to each product is the same as the one defined in the visual representation of genomic data. The pie and bar chart are also accompanied with a table that has information about the different products and their number of appearances in the genomic data.

The user interaction with the pie or bar chart leads to a filtration of data in the other representations. For example, when a range of sizes is chosen in the bar chart, only products in



(a) **(b)**

Colour	Function	Counts	Frequency
	conserved hypothetical protein	138	7.12%
	putative membrane protein	130	6.70%
	putative uncharacterized protein	60	3.09%
	hypothetical protein	42	2.17%
	ABC transporter ATP-binding protein	28	1.44%

(c)

Figure 3.4: Visual representations of the *Streptococcus pneumoniae 70585 regions*' product and size distribution. (a) Bar Chart with the *regions* size distribution. The size interval of 1808-2704 is selected. (b) Pie Chart with the products distribution in the selected size interval in the Bar Chart. Each colour represent a different product. (c) Table with the top counts of products in the selected size interval.

the selected size range are shown in the products table and in the pie chart. The reverse situation also occurs when a specific product is selected in the pie chart.

3.2.4.2 Querying on specific genomic *regions*

We developed two distinct approaches to perform queries on specific genomic *regions*: queries on annotations and sequence based queries. We also separated these queries into two different categories: *link queries*, when they establish relationships between *regions* of different files; and *basic queries* when they have only one target *region*.

The query system was developed so that the queries are cumulative in the sense that their results can be displayed simultaneously. All queries are also stored in a list and users can remove individual queries from it in order to tailor the final displayed targets to their needs.

Queries on annotations are performed by a client-side search in the annotation attributes name, product, or location. They are made by typing keywords on the search box field on the interface's actions menu and by selecting to option to search in one or in all files represented.

Queries on annotations are considered *link queries* when a comparison method is chosen to establish relationships between annotations of different files. These comparison methods can be by name or product and they are chosen by selecting the "*Comparison methods*" option on the actions menu. When comparing by name, there is a mapping between *regions* of different files with the same name, while the comparison by product finds *regions* of different files with the same product. It should be noted that are only created links for relationships found by a given query between *regions* that are in adjacent positions in the visual representation. This approach is used to avoid overloads when viewing these comparisons. If users want to visualize relationships between annotations of files that are in remote positions in the visual representation, they first have to use the option provided by the program to change the files position in the visual representation.

Sequence based queries are performed using an internal sequence from a *region* or by using an external sequence. Internal sequence queries are considered *link queries* while external sequences queries are considered *basic queries*. In all sequence based queries, all the sequence comparisons are made by BLAST and the user can control the minimum identity and minimum score parameters for a positive match.

Internal sequence queries are made by right clicking on a specific genomic *region* in the representation and by choosing the option "*BLAST against the next position sequences*" on the menu. Because BLAST only performs pairwise comparisons, the *region* chosen by the user is only

File Name	Genome Position	Name	Total Size	Annotated Portion	File Type	Number of Annotations	Number of Transposases	Number of IS	% of Hypothetical Proteins	Number of Contigs
14_contigs.fasta	1	Contigs File	2071551 bp	-	.fasta	-	-	-	-	124
FQ312030.gbk	2	Streptococcus pneumoniae INV104	2142122 bp	1800872 bp (84.07%)	.gbk	1939	65	42	9.95%	-

(a)

#	Type Search	Gene	Product	BLAST of Source	Number of Relations	Genome of Target	Gene Begin	Gene End
1	GeneProduct	INV104_01090 (See Position)	putative ABC transporter ATP-binding/permease (pseudogene)	BLAST against next genome.	0	2	129283	130901
2	GeneProduct	INV104_01210 (See Position)	ABC transporter, substrate binding protein	BLAST against next genome.	0	2	141302	142133
3	GeneProduct	INV104_01220 (See Position)	ABC transporter, substrate binding protein	BLAST against next genome.	0	2	142354	143185
4	GeneProduct	INV104_01250 (See Position)	putative ABC transporter ATP-binding protein	BLAST against next genome.	0	2	145658	146720

(b)

#	Type Search	Query Gene	Target Gene	Product	Genome of Target	Target Alignment Begin	Target Alignment End	Query Alignment Begin	Query Alignment End	Alignment Score	Strand of Query	Strand of Target
19	BLAST	NODE_1 (See Position)	aqpZ (See Position)	aquaporin Z	2	3	669	28193	28859	652.00	+	+
31	BLAST	NODE_1 (See Position)	exoA (See Position)	exodeoxyribonuclease	2	1	828	94117	94944	828.00	+	+
60	BLAST	NODE_1 (See Position)	fmt (See Position)	methionyl-tRNA formyltransferase	2	1	936	6065	7000	927.00	+	+
97	BLAST	NODE_1 (See Position)	galK (See Position)	galactokinase	2	1	1179	102781	103959	1179.00	+	+

(c)

Sequences Aligned

```
Query Sequence     T T C A A C T G T C T T A A C T T C A A G A G C T T C C C A G A T T T T T T C G T T G A A C
Matches            | | | | | | | | | | | |   | | | | | | | |   | | | | | | | | |   | | | | |   | | | | |   | | |
Target Sequence    T T C A A C T G T C T T C A C T T C A A G T G C T T C C C A A A T T T T C T C G T T A A A C
SNPs (Total: 56)   - - - - - - - - - - - - P - - - - - - - - P - - - - - - - - P - - - - - P - - - - - P - - -
```

(d)

Figure 3.5: The different query results tables. (a) More Info table. Has global information about the uploaded files. (b) *Hits Table* with results from queries by annotations. Has information about the results' name, product and position of the file in the visual representation as well as the location of the results in the sequence. (c) *Hits table* with sequence based query results. Shows the alignment location of the query sequence with the reference and its score. (d) HSP aligned table. Show the alignment of a high-similarity pair and their SNPs, at nucleotide level, of any results after a sequence based query.

compared with the genomic sequences that are immediately in the adjacent position in the representation.

Choosing the "*Search by Sequence*" option on the interface gives the option to perform external sequence queries. A file in the visual representation needs to be chosen to act as reference for a BLAST search and an external nucleotide sequence coupled with a name to identify it must be inserted to act as query.

In the next section we describe the developed ways to visualize the results of these queries on specific *regions*.

3.2.5 Visualizing results of queries on specific genomic *regions*: *Hits table* and representation modification

Results obtained after performing queries on specific *regions* are represented in a *Hits table* and by specific modifications in the visual representation of the *nodes*. However, there are some differences in the information that is shown when the user performs *basic* or *link queries*.

The *Hits table* provides text information about the queries results (Figure 3.5-b, c). It can be accessed after performing any query by choosing the "*Hits Table*" option on the actions menu. The table is fully customizable, being possible to organize all columns and filter the entries. This is achieved through the use of the DataTables[97] plugin for JQuery JavaScript library[98] which adds advanced interaction controls to HTML tables.

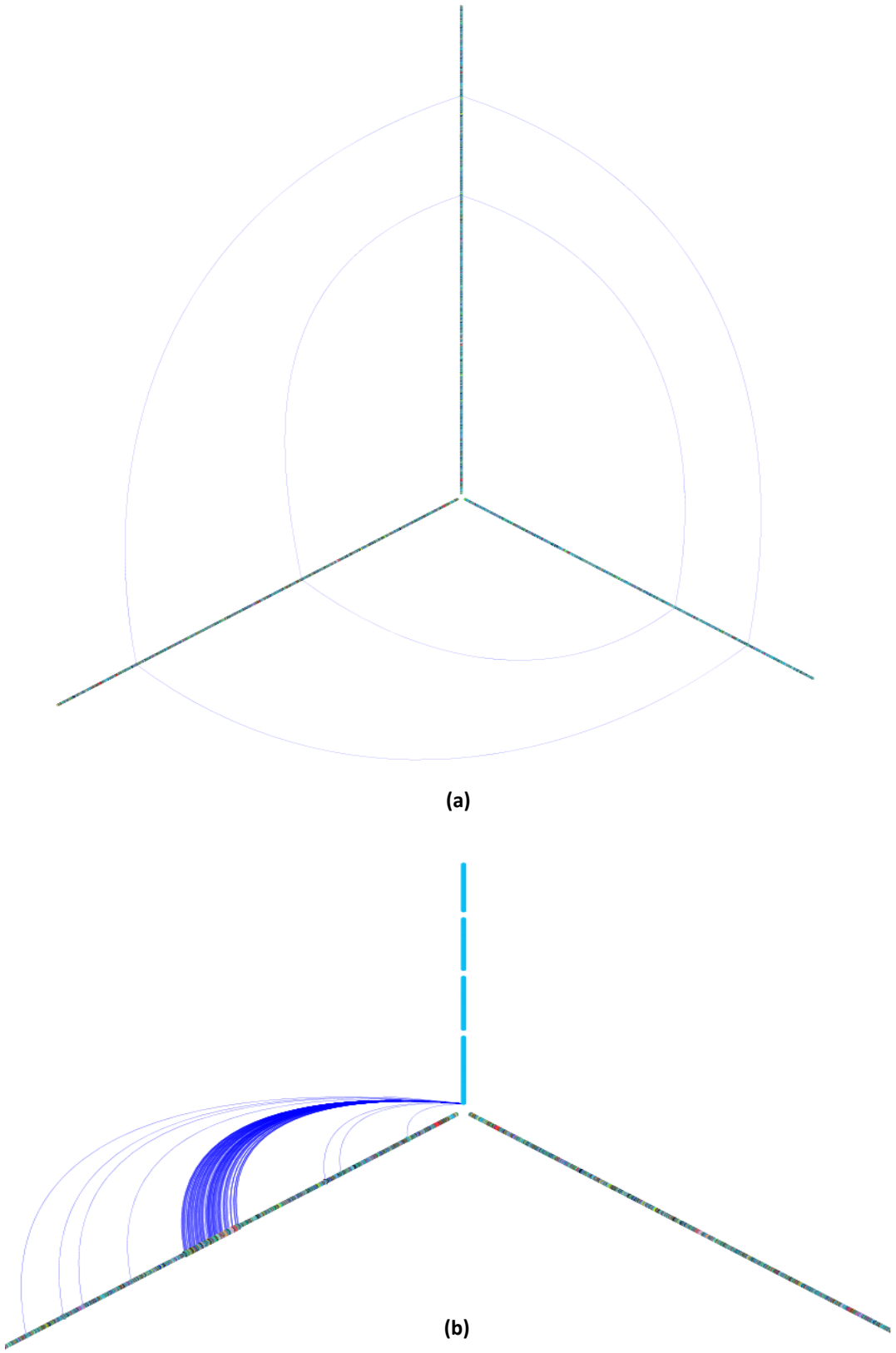
The *Hit table* also has a feature to allow user interaction between the table and the visual representation. This is achieved by using the "*See position*" field in the *Hits table*, which instantaneously directs the user to the location of the result's *region* in the visual representation. This operation is carried out by the use of the SVG coordinates to centre the image at the specific point, which leads to a precise focus at the desired *region* in the visualization.

Regarding the visual representation, a new colour is assigned to each query and all query results are represented by a size increment of the ellipses corresponding to the *regions* affected.

When the user performs a *link query*, there are some additional information displayed in both the *Hits table* as in the visual representation. These differences are described below.

3.2.5.1 Link querying: stablishing links and results table modification

When the user performs *link queries*, relationships between different *regions* are displayed through links in the visual representation (Figure 3.6). They are developed by assigning



(a)

(b)

Figure 3.6: Links between *regions* of different files. (a) Established links after choosing the comparison by gene name method and search by name for 2 genes (*aroE*, *xpt*) in 3 *Streptococcus pneumoniae* strains. (b) Part of the visual representation of an internal sequence based query of a contig of *Streptococcus pneumoniae* against an annotated genome. It is possible to verify the localization of the contig in the annotated genome.

the category of source and target to different *regions*, which are used as a start and end point to draw a path line between the two *regions* using D3.

The *Hits table* also as some additional fields that are specific from *link queries* (Figure 3.4-c). For each *link queries* results is shown the information about the name, product and location of the *regions* that functioned as source and target to create the link. In the case of *link queries* involving BLAST alignments, information about the sequence alignment locations and their respective scores is also shown.

3.2.6 Visualizing sequence alignment at nucleotide level

We have already shown two ways to visualize sequence alignment results: by the visual representation in the way of links and by the results in the *Hits table*. Another developed feature allows the representation of BLAST sequence alignments at the nucleotide level and highlights Single Nucleotide Polymorphisms (SNPs) between sequences (Figure 3.5-d).

After the user makes a query that produces results obtained by BLAST, he can choose to visualize the alignment at the nucleotide level by right clicking in one of the results of the *Hits table* and by selecting the option to "*view HSPs aligned*". By choosing this option, the high-similarity pairs (HSPs) obtained by BLAST are shown aligned with the representation of matches and gaps. The number of SNPs and their location is also shown after running the *show-snps* utility of the MUMmer software at the server-side.

3.2.7 Operations with contigs

ProGenViZ provides two different features which can be applied to contigs data: order contigs against a reference and single contig annotation. These two features are described in detail below.

3.2.7.1 Ordering contig data

Contigs prevenient of assembly of HTS data are usually available in multi-FASTA files. One feature that ProGenViZ allows users to do is to load these files and order them against any FASTA or GenBank file loaded on the interface.

To proceed with the ordering of contigs, the user must choose the "*Order contigs*" option in the actions menu and also a contigs file to act as query and other file to act as reference. After performing all these steps, two FASTA files are created. One FASTA file with the reference sequence and other with the contigs sequences. They are used as input for the alignment

software NUCmer inserted in the MUMmer software package, using the programs' default parameters. The NUCmer results are then filtered through the *delta-filter* algorithm which filters the alignment results of NUCmer, using as default parameters a minimum identity of 98% and a minimum alignment size of 300 nucleotides. These parameters used to filter the alignment results can be changed by the user.

After running the contig ordering option, the resulting order is displayed in the visual representation. It is important to notice that this alignment feature provided by the program aims only to show the relative position between contigs and also act as a filter to display only those contigs that align against the reference sequence under study.

3.2.7.2 Single contig annotation

Contig data is mainly available in multi-FASTA format and therefore, no annotation information is provided in this format. In order to allow contig annotations, we implemented an annotation process, which transfers features from a loaded annotated sequence to a target contig (Figure 3.7).

To perform the annotation of a given contig, we use a combination of BLAST to search for homologous genes in an annotated sequence and Prodigal to predict gene locations in the contigs. To annotate a contig the user must first perform an internal sequence query against the annotated reference by right clicking on the contig and choosing the "*BLAST against the next position sequences*" option on the menu. Then, to annotate it, it is necessary to right click on the contig again and choose the "*Annotate this Region*" option. The selection of the annotations that are transferred to the contig is made by combining the results obtained by BLAST and Prodigal. An annotation is transferred only if the start and end positions of an alignment of a *region* from the reference with the contig are the same or include the Prodigal's predicted start and end locations of a CDS in the contig (Figure 3.7-b). Because Prodigal only predicts genes, annotations that do not correspond to a CDS are not passed to the contigs. The annotation procedure can be repeated for each contig loaded in the interface.

3.2.8 Editing gene annotations

Designing ways to establish relationships between *regions* of different files also provides the possibility of monitoring the quality of annotations by sequence similarity. In order to be able to change the information of *regions* that are found to be poorly annotated, an option to perform changes in the pre-existing name and product of a *region* was developed. The user can edit an annotation by right clicking on the desired *region* and by selecting the "*Edit information*"

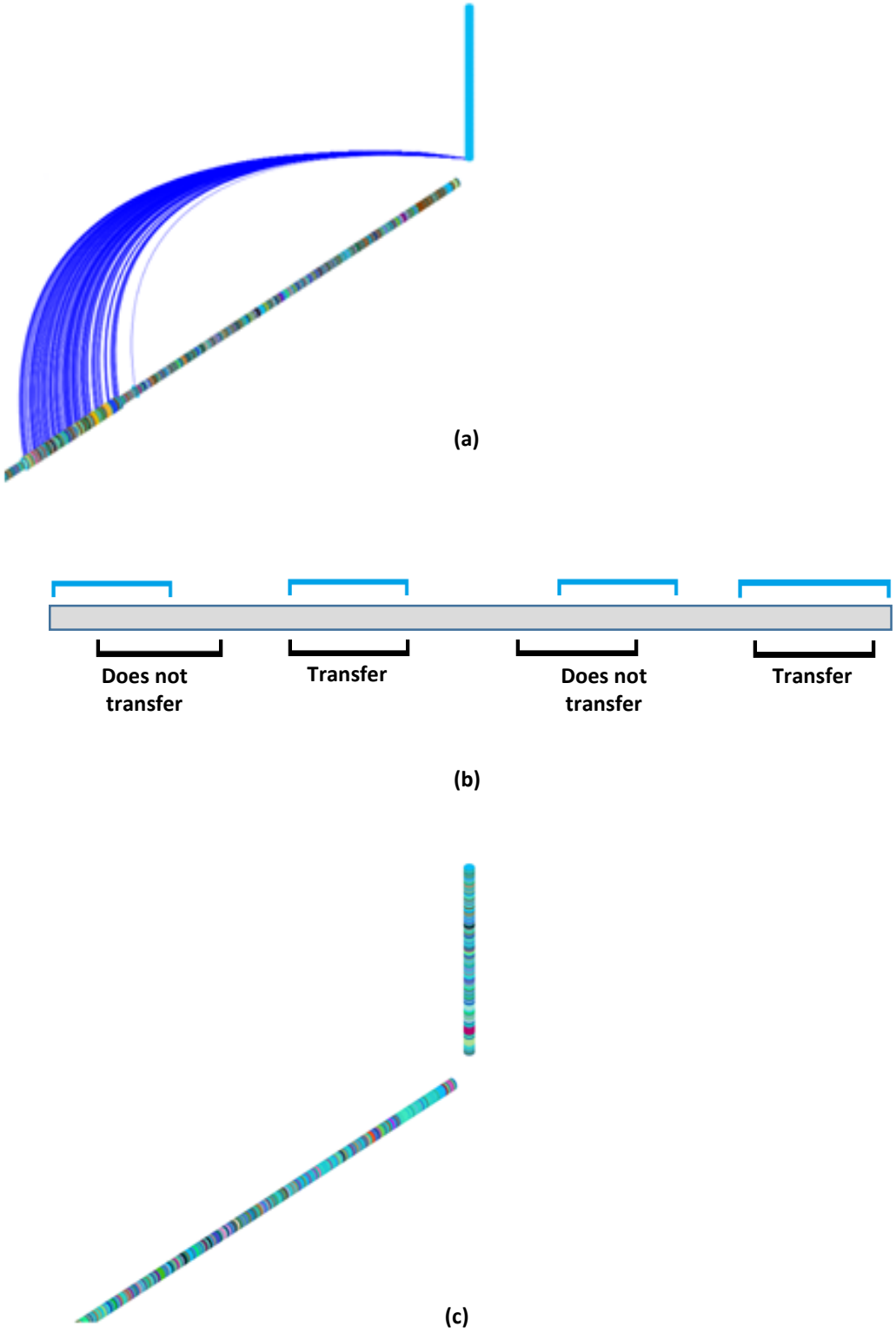


Figure 3.7: The contig annotation process. (a) Internal sequence query of a contig against a reference. (b) Transfer of annotation by combining the BLAST results and Prodigal. In grey is shown the representation of the contig's sequence. In blue is shown the BLAST alignment result of a *region* from the reference sequence with the contig. In black is represented the CDS predicted by Prodigal in the contig. (c) Annotated contig after choosing the option to "*Annotate this Region*" on the right-click menu.

option on the menu. The modification of pre-existing annotations is made by modifying the name or product of a *region* in the JSON file with the genomic features. These changes can then be exported using functionalities provided by ProGenViZ.

3.2.9 Exporting Data

There are several different data types that can be exported from ProGenViZ: results presented in tables, images, specific genomic sequences, specific contigs and whole files. The export functionality is very important in all data analysis software because it enables the use of results obtained from this tool in other software to allow further analysis.

The export of results from tables can be made through an option that is provided on their upper right corner when they are shown. The results can be exported in Comma Separated Values (CSV) and PDF format.

Images can also be exported as the current visual representation. This is done through the "*Export Image*" option of the actions menu. Images can be exported in PNG, high-resolution PNG and PDF formats.

Another option developed is the export of sequences associated with specific *regions*. This export option is made by right clicking on any of the *regions* in the visual representation and by choosing the "*Get Sequence*" option on the menu. Consequently, a FASTA file is generated and presented with the sequence of the chosen *region*.

Specific contigs can also be exported. This is made by right clicking on the desired contig and by choosing the "*Export contig*" option in the menu. In that time are generated two files at server-side, one FASTA file with the contig sequence and one GFF with its annotations, if any, which the user is given the option to download.

It is also possible to export all information of *regions* and genomic sequences for each file loaded in the interface. If changes occurred in the representation of a file of contigs that led to the annotation of some, the use of this option will only exports the annotated contigs. This type of export is done through the "*Export files*" option in the actions menu and by choosing the location of the file in the visual representation to export. After selecting a file to export we provide a combination of a FASTA file with the genomic sequences and a GFF file with the annotations.

4

Use Cases

4 Use Cases

The goal of this section is to illustrate the program's capacities through three use cases. In the first one we focus on the program's ability perform searches for genes via annotations and also through sequence similarity. In the second use case we show the tool capacities to discover locations of interest in HTS contigs and finally, in the third use case, we use the tool's features to order and annotate all contigs from an assembly against an annotated reference genome.

Use case 1 – Search for the MultiLocus Sequence Typing (MLST) scheme genes of Streptococcus pneumoniae

MLST is a technique used to characterize microbial strains by their DNA sequence variations in a set of housekeeping genes fragments, which are converted to allelic profiles by attributing an allele identifier to each unique DNA sequence of a fragment (typically spanning 400 to 700 nucleotides). The study of these variations is useful to characterize microorganisms at subspecies level, help understand the evolution of species and produce relevant data for epidemiological studies[99]. In this use case we show how to search for MLST genes by using information provided by annotated sequences and also by using external allele sequences taken from the *Streptococcus pneumoniae* MLST Database[100].

The *Streptococcus pneumoniae* MLST scheme uses internal fragments of the following seven housekeeping genes: *aroE* (shikimate dehydrogenase), *gdh* (glucose-6-phosphate dehydrogenase), *gki* (glucose kinase), *recP* (transketolase), *spi* (signal peptidase I), *xpt* (xanthine phosphoribosyltransferase), and *ddl* (D-alanine-D-alanine ligase). In this analysis, two GenBank files with genomes of different *Streptococcus pneumoniae* strains obtained from the National Center for Biotechnology Information (NCBI)[101], *Streptococcus pneumoniae 70575* (called SP70585) with the accession number CP000918 and *Streptococcus pneumoniae 670-6B* (called SP*670-6B*) with the accession number CP002176, were used to perform queries for the MLST genes and to analyse their synteny in both genomes.

Our first approach was to try to find the MLST genes by their name in the annotated genomes using the genes acronyms indicated previously (example: *gdh*). For this we performed *basic queries* by annotations using the interfaces' search area on the actions menu, by choosing

the option to perform the search by name in all genomes and by typing each of the seven acronyms of the MLST genes in the search box.

By analysing the *basic queries* by name results, matches were only found for some MLST genes (Table 4.1). The gene *aroE*, *ddl* and *xpt* were found in SP70585, while in SP670-6B only the genes *aroE* and *xpt* were found.

Gene	Product	Genome of Target	Gene Begin	Gene End
aroE	shikimate 5-dehydrogenase	1	1311640	1312495
ddl	D-alanyl-alanine synthetase A	1	1581501	1582545
xpt	xanthine phosphoribosyltransferase	1	1766318	1766900
aroE	shikimate 5-dehydrogenase	2	852702	853557
xpt	xanthine phosphoribosyltransferase	2	1790031	1790613

Table 4.1: Results after performing *basic queries* by name for all genes from the MLST scheme of *S. pneumoniae*. The Gene Begin and Gene End columns correspond to the location of the genes in the genome. Only the genes *aroE*, *ddl* and *xpt* had matches on SP70585 (genome of Target 1), while in SP670-6B (genome of Target 2) we had matches for *aroE* and *xpt*.

Since not all genes were found , our next approach was then to perform *basic queries* by annotations using the MLST gene products described in the MLST database instead of using their gene names (example: search for shikimate dehydrogenase product to find the *aroE* gene). This was made by selecting the option to search by product in all genomes and by typing each of the gene products of the MLST genes in the search area of the actions menu.

When analysing the results obtained from *basic queries* by product, as in *basic queries* by name, only some of the gene products were found in both genomes (Table 4.2). The xanthine phosphoribosyltransferase product was the only one that had matches in both genomes with the gene name equal to the expected (*xpt*). The glucose kinase product whose name should be *gki* only had a match for the gene *SP70585_0727* in the SP70585 genome. The glucose-6-phosphate dehydrogenase had also a match in the SP670-6B genome but with a gene name different than expected (*zwf* instead of *gdh*). Also, in both SP70585 and SP670-6B, the signal peptidase I was assigned to the *lepB* gene when is supposed to be the product of the *spi* gene.

The transketolase product was the only one that had more than one match in each genome due to the existence of different subunits. However, none of these matches had the name *recP*.

Gene	Product	Genome of Target	Gene Begin	Gene End
hpt	hypoxanthine phosphoribosyltransferase	1	11811	12354
lepB	signal peptidase I	1	426744	427359
SP70585_0727	glucokinase (Glucose kinase)	1	667833	668793
tkt1	transketolase	1	1528039	1530010
xpt	xanthine phosphoribosyltransferase	1	1766318	1766900
tkt2	transketolase	1	1939819	1941802
SP70585_2253	transketolase, C- subunit	1	2061389	2062322
SP70585_2254	transketolase, thiamine disphosphate-binding subunit	1	2062318	2063176
hpt	hypoxanthine phosphoribosyltransferase	2	27564	28107
lepB	signal peptidase I	2	433006	433507
zwf	glucose-6-phosphate dehydrogenase	2	959603	961091
lspA	signal peptidase II	2	1305039	1305501
SP670_1700	transketolase	2	1581315	1583286
xpt	xanthine phosphoribosyltransferase	2	1790031	1790613
SP670_2110	transketolase	2	1972206	1974183
SP670_2271	transketolase, C- subunit	2	2115199	2116132
SP670_2272	transketolase, thiamine disphosphate-binding subunit	2	2116128	2116986

Table 4.2: Results after performing *basic queries* by product for all genes from the MLST scheme of *S. pneumoniae*. Genome of Target correspond to the position of the genome in the visual representation. The Gene Begin and Gene End columns correspond to the location of the genes in the genome. Since the queries by annotations are case sensitive, there are some results that correspond to products with names similar to the ones used to perform the queries. The xanthine phosphoribosyltransferase was the only gene product that had matches from the expected gene name (xpt).

After these two approaches, it became clear that both the *basic queries* by name as by product did not show results for all genes of the MLST scheme. Two hypotheses can be put forward for the observations: the correct genomic sequences for those genes are present in the files but the genes have an alternative annotation, or the sequences encoding those genes did not exist in these genomes. To test both hypotheses, we performed external sequence *basic queries* using one allele of each gene from the MLST scheme taken from the MLST database to confirm if those genes were present in both genomes and also to see if they match to the ones

found previously through *basic queries* by annotations. The sequences of the alleles used can be found in the Appendix 1.

To carry out the external sequence *basic queries* we choose the *"Search by Sequence*" option on the actions menu for each of the seven gene fragments and we performed BLAST searches with those sequences against both genomes with a minimum e-value of 10^{-4} and minimum alignment length of 300 nucleotides. A unique identifier was assigned to each of the alleles with the name of the gene to which they belong and the allele number taken from the MLST database (example: *aroE1*).

After we performed all external sequence *basic queries*, each allele sequence only got one high-similarity match on each genome although most of the gene names were not the ones of the *S. pneumoniae* MLST scheme (Table 4.3). Only the ddl1, aroE1 and xpt1 sequences matched to genes with the correct name in the SP70585 genome, confirming the previous findings. In the case of the SP670-6B genome, aroE1 and xpt1 were the only two fragments who matched to

Query Gene	Target Gene	Product	Genome of Target	Target Alignment Begin	Target Alignment End	Query Alignment Begin	Query Alignment End	Alignment Score
aroe1	aroE	shikimate 5-dehydrogenase	1	723	319	1	405	393.00
gdh1	zwf	glucose-6-phosphate 1-dehydrogenase	1	840	1299	1	460	445.00
gki1	SP70585_0727	glucokinase (Glucose kinase)	1	277	759	1	483	474.00
recP1	tkt2	transketolase	1	1841	1392	1	450	450.00
spi1	lepB	signal peptidase I	1	542	69	1	474	459.00
xpt1	xpt	xanthine phosphoribosyltransferase	1	46	530	1	485	470.00
ddl1	ddl	D-alanyl-alanine synthetase A	1	582	142	1	441	438.00
aroe1	aroE	shikimate 5-dehydrogenase	2	133	537	1	405	396.00
gdh1	zwf	glucose-6-phosphate dehydrogenase	2	649	190	1	460	448.00
gki1	SP670_0727	glucokinase	2	277	759	1	483	483.00
recP1	SP670_2110	transketolase	2	1835	1386	1	450	441.00
spi1	lepB	signal peptidase I	2	501	69	42	474	427.00
xpt1	xpt	xanthine phosphoribosyltransferase	2	46	531	1	486	483.00
ddl1	SP670_1758	D-alanine--D-alanine ligase	2	582	142	1	441	396.00

Table 4.3: Results after performing external sequence *basic queries* of alleles taken from the MLST database of all seven *S. pneumoniae* MLST scheme genes. The Query Gene column correspond to each of the external sequences used. Genome of Target correspond to the position of the genome in the visual representation. Target Alignment Begin and End indicate the beginning and end of the alignment of the reference genome with the external sequence. Query Alignment Begin and End indicate the portion of the external sequence that aligns with the reference. For each of the external sequences there was only one match in each of the genomes.

genes with the correct name of the MLST scheme. The gki1 fragment aligned to a gene with the name *SP670_0727*, the recP1 to *SP670_2110*, ddl1 to *SP670_1758*, spi1 to *lepB* and gdh1 to *zwf*.

With these results we managed to determine that the lack of results in *basic queries* by annotations were not due to the absence of genomic sequences for these genes but rather inconsistencies in the annotations of each of the genomes. As the vast majority of automatic annotation tools use an approach which searches for sequence homology and orthologous genes, some of the names that are assigned to certain genes are gene names from other organisms. This leads to the creation of erroneous annotations that can lead to incorrect conclusions by the user.

To eliminate the problem of having incorrect gene names for the *S. pneumoniae* MLST scheme genes and since we had very high alignment scores in the external sequence *basic queries,* we manually edited the gene name and gene product of each result through the option provided by ProGenViZ to edit annotations. For this we had to right-click in all *regions* in the

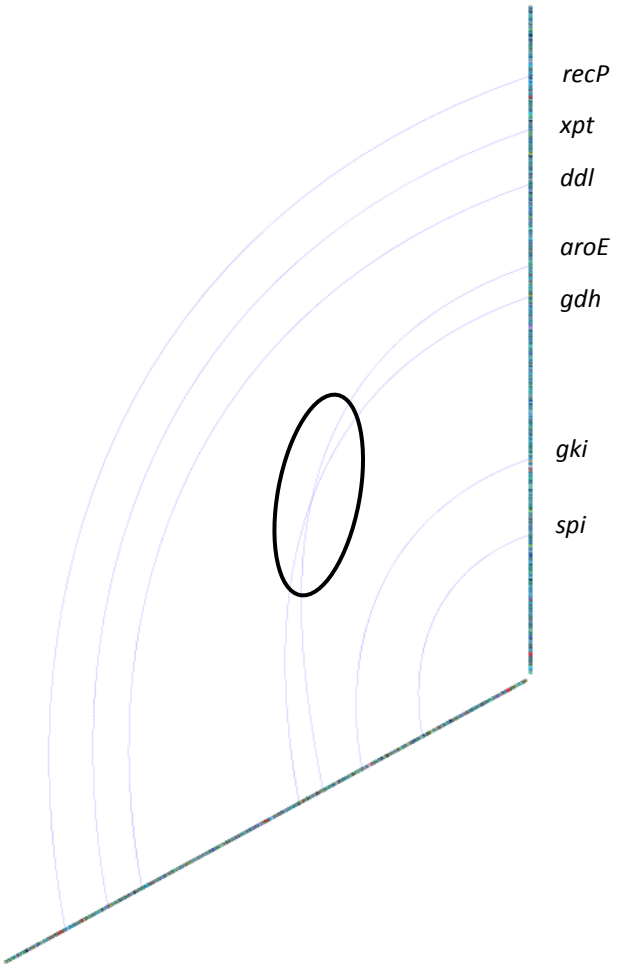


Figure 4.1: Global synteny of the MLST scheme genes in *Streptococcus pneumoniae 70575* and *Streptococcus pneumoniae 670-6B*. The black ellipse highlights the inversion of *aroE* and *gdh* genes between the two genomes. The remaining genes share the same position in the genome.

visual representation which corresponded to the external sequence *basic queries* results and choose the "*Edit Information*" option on the menu.

After having determined that all seven MLST genes were present in both genomes analysed, we wanted to analyze their arrangement in the two genomes and their synteny. For this we performed internal sequence *link queries* by right-clicking in the visual representation on each of the MLST genes of SP70585 and by choosing the "*BLAST against the next position sequences*" option on the menu to visualize their relationships with the SP670-6B genome. A minimum e-value of 10^{-4} and minimum alignment length of 300 nucleotides was used for the BLAST searches.

By analysing the links that were created between the two genomes, we managed to see that only five of the genes conserved the shared synteny in the two genomes (Figure 4.1). The *ddl* and *aroE* genes positions in the SP70585 genome are inverted in relation to the SP670-6B genome. This could be explained if there was a genome rearrangement of a larger fragment containing the two genes that could cause them to switch positions or this could be due to some genome assembly error. Further laboratory work would be needed to test these hypotheses.

The genomes with the changes in the annotations were exported at the end of the analysis in a combination of a FASTA file containing the sequences and GFF file with annotations. That was done through the "*Export files*" option offered by the interface and by selecting each genome position on the visual representation to export.

Use case 2 – Search for Capsule Biosynthesis locus (cps) genes in Streptococcus pneumoniae contigs

Several major invasive bacterial pathogens are encapsulated. Expression of a polysaccharide capsule is essential for their survival in the blood and it is associated with the virulence of the organism. Moreover, these same capsules are target for host antibodies and are often used as a basis for effective vaccines[102]. Serotyping is the identification of the capsule using antibodies, and was one of the first methodologies for the subtyping of *S. pneumoniae*[103].

Encapsulated species typically exhibit antigenic variation and express one of a number of immunochemically distinct capsular polysaccharides (CPSs) that define serotypes. In case of *Streptococcus pneumoniae*, with the exception of serotypes 3 and 37, CPSs are generally synthesised by the Wzx/Wzy-dependent pathway, whose genes for the latter pathway are located at the same chromosomal locus (*cps*)[104]. The *cps* locus consists of four regulatory

genes (*wzg*, *wzh*, *wzd* and *wze*) and a series of transferases, polymerases and flipases necessary for capsule biosynthesis.

In this case study we want to demonstrate the capabilities of the program to find locations of interest in contigs. We will try to find the regulatory genes belonging to the *cps* locus in a contigs file of *Streptococcus pneumoniae* (*S*erotype 1) isolated in Hospital de Santa Maria and provided by the Microbiology and Infection Unit of the Instituto de Medicina Molecular. For that we will use external sequences taken from the NCBI database of the genes *wzg*, *wzh*, *wzd* and *wze* (IDs 6216691, 6216690, 6216642 and 6217227 respectively). The contigs with information about the regulatory genes, if any, will consequently be annotated using a reference genome of *Streptococcus pneumoniae INV104* taken from the NCBI (accession number: FQ312030).

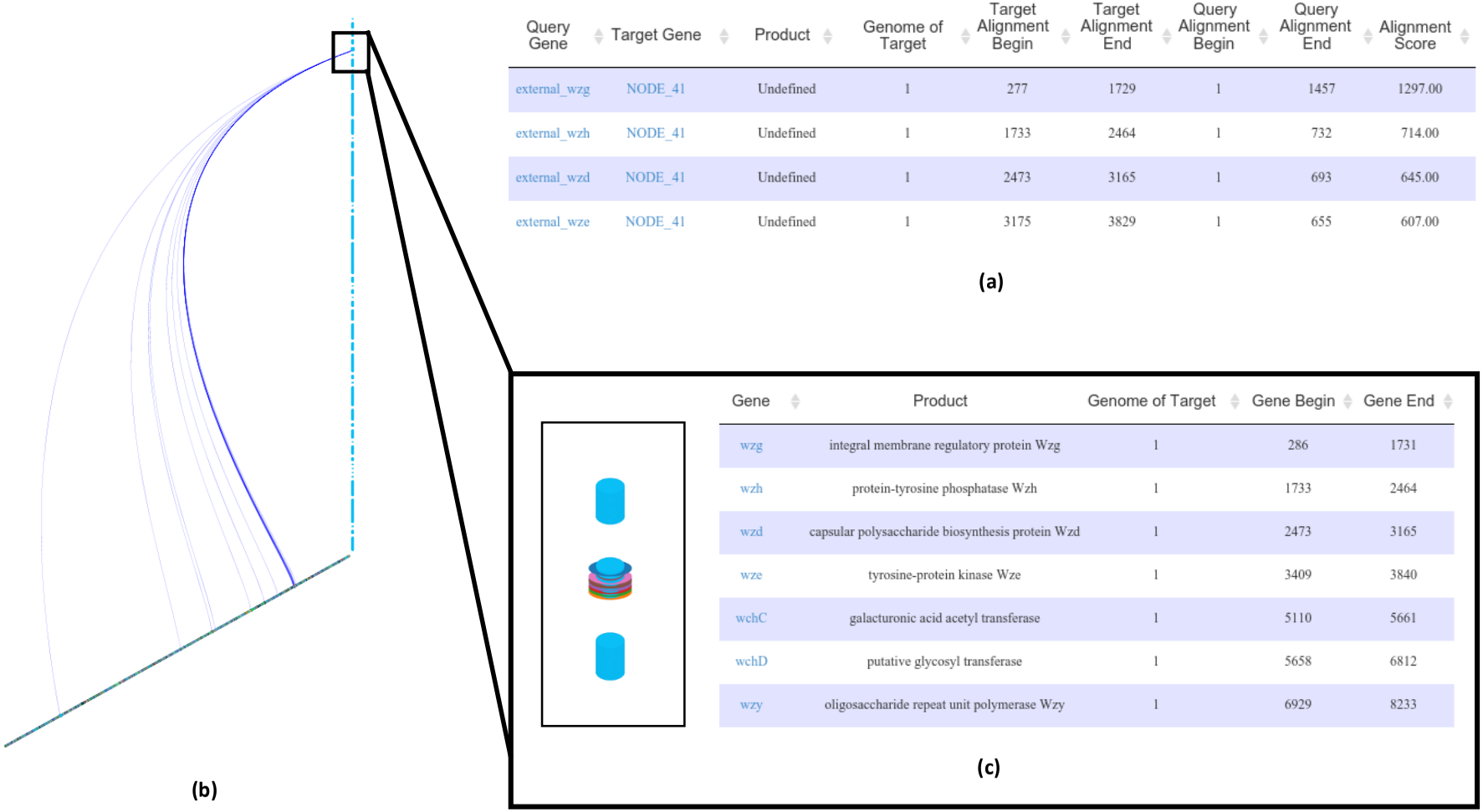
After uploading the file of contigs and the reference genome, we used the four external sequences of the regulatory genes taken from the NCBI database mentioned above to try to find matches in the contigs. This was done through external sequence *basics queries* for each gene using the "*Search by Sequence*" option of the actions menu with the default parameters and by choosing a unique identifier for each of them (example: *external_wzg*).

Matches were found for each of the genes in a single contig which had a length of 13145bp (Figure 4.2-a). The contig number 41 had matches for all of them. The *wzg* gene aligned from the 277bp of the contig until the 1729bp, the *wzh* from 1733bp to 2464bp, the *wzd* from 2473bp to 3165bp and the *wze* from 3175bp to 3829bp. Not only do these distances correspond to the total length of each inserted external sequences, as the arrangement of genes in the contig are those who would be expected to find in the *cps* locus.

As regulatory genes only covered a quarter of the total size of contig 41, we put the hypothesis for the existence of other genes of the *cps* locus in this contig. To verify this hypothesis we carried out an internal sequence query of contig 41 against the reference genome by right-clicking on contig 41 and choosing the "*BLAST against the next position sequences*" option on the menu.

When analysing the results of the internal sequence query, we confirmed that the contig 41 had not only matches for the regulatory genes of the *cps* locus but it also had matches to the genes *wzy*, *wzx*, *wchB*, *wchC* and *wchD* which are also part of the locus.

Knowing that contig 41 had information about several genes of the *cps* locus, the next step was to annotate it using the reference genome of the same strain. As we had already made the first step towards the annotation of contigs which is the internal sequence query, we only had to choose the option to annotate the contig by right-clicking on it and by choosing the "*Annotate region*" option on the menu.



Query Gene	Target Gene	Product	Genome of Target	Target Alignment Begin	Target Alignment End	Query Alignment Begin	Query Alignment End	Alignment Score
external_wzg	NODE_41	Undefined	1	277	1729	1	1457	1297.00
external_wzh	NODE_41	Undefined	1	1733	2464	1	732	714.00
external_wzd	NODE_41	Undefined	1	2473	3165	1	693	645.00
external_wze	NODE_41	Undefined	1	3175	3829	1	655	607.00

Gene	Product	Genome of Target	Gene Begin	Gene End
wzg	integral membrane regulatory protein Wzg	1	286	1731
wzh	protein-tyrosine phosphatase Wzh	1	1733	2464
wzd	capsular polysaccharide biosynthesis protein Wzd	1	2473	3165
wze	tyrosine-protein kinase Wze	1	3409	3840
wchC	galacturonic acid acetyl transferase	1	5110	5661
wchD	putative glycosyl transferase	1	5658	6812
wzy	oligosaccharide repeat unit polymerase Wzy	1	6929	8233

Figure 4.2: External sequence queries in contigs of *Streptococcus pneumoniae INV 104* and annotation against an annotated reference genome. (a) Results of an external sequence query for regulatory genes of the capsule biosynthesis. NODE_41 is referred in the text as contig 41. There were matches for the 4 regulatory genes in contig 41. (b) Hits from an internal sequence query of contig 41 against the reference genome. It as matches for the 4 regulatory genes in the reference and also for the genes *wzy*, *wzx*, *wchB*, *wchC* and *wchD* of the cps locus (links in dark blue). (c) Genes that were transferred to contig 41 after annotation transfer from the reference.

After the contig was annotated, we checked the annotation transference by performing *basic queries* by name for each of the results obtained earlier by the internal sequence query of contig 41 against the reference genome. To do this we used the search area of the actions menu and we choose the option to perform searches by name only in the file with contigs and we typed name each of the results of the internal sequence query in the search box.

The results obtained by the *basic queries* by name showed that not all the annotations were transferred to the contig (Figure 4.2-c). The annotations of regulatory genes were all transferred as well as those of the *wchC*, *wchD* and *wzy* genes. However, annotations from the genes *wzx* and *wchB* were not transferred. This may be due to discrepancies in the alignment location of these genes in the contig with the locations predicted by Prodigal, which prevents the annotations to be transferred.

After performing all the analyses, the annotated contig was exported using the "*Export files*" option on the actions menu. Only the annotated contigs were exported in a combination of a FASTA and a GFF file.

Use case 3 – Streptococcus pneumoniae OXC141 contigs annotation

The practical application of sequencing technologies in clinical diagnostic settings starts with the development of strategies to identify sequences obtained from HTS output. At first, contigs obtained by read assembly are annotated or ordered and it is often necessary to determine if specific genes are present. In this type of analysis, sequence alignment and annotation have an extremely important role.

The use case presented here provides features to order contigs using a reference genome and to perform annotations of those contigs by annotation transference from a reference.

For this analysis we used two files of contigs from two serotype 3 *Streptococcus pneumoniae OXC141* strains – named ContigsF1 and ContigsF2 for the purpose of this thesis – isolated in Hospital de Santa Maria and provided by the Microbiology and Infection Unit of the Instituto de Medicina Molecular. They were obtained by Illumina HiSeq sequencing resulting in 90bp Paired End reads with a genome coverage of approximately 100x. The reads were assembled using the Velvet[105] software. The parameters used for the assemblies can be found in the Appendix 2.

In order to align the contigs, an annotated genome for the strain OXC141, also serotype 3, obtained from the NCBI database with the accession number FQ312027 was selected.

After uploading both contigs and reference files, we first wanted to determine the global portion of the reference genome that was being covered by the two files of contigs. For this we used the "*More Info*" option from the actions menu to display the global information of the files (Figure 4.3-a). The ContigsF1 had a total of 100 contigs with a total length of 2016870bp, which corresponds to a coverage of 99% of the reference genome. In the case of ContigsF2, it had a total of 107 contigs and a total length of 2015251bp, which covers about 98.9% of the *S. pneumoniae OXC141* genome.

Next step was ordering the contigs against the reference to get a better sense of their spatial location in the genome. To do this we used the "*Order contigs*" option from the actions menu. The contigs file acted as query for the alignment and the annotated genome was used as reference using a minimum alignment of 500bp and minimum identity of 0.98. It should be noted that by using these parameters, we could be filtering out some contigs from the analysis due to the existence of contigs with size less than 500bp or due to the strictness of the minimum identity threshold that was selected. However, we are simultaneously reducing the possibility of having some overlapping contigs.

After ordering the two files against the reference, the number of contigs reduced by almost half (Figure 4.3-b). In ContigsF1 we now have 53 contigs and we were left with 47 contigs in ContigsF2. However, the contigs coverage of the reference did not follow the same path. There was only a slight drop in coverage from 99% to 97.8% in ContigsF1 and 98.9% to 98% in ContigsF2.

Now with the ordering and filtering of contigs made, we can move to their annotation. The contigs are annotated one at a time in two steps. The first step is to perform an internal sequence query of the contig against the reference to determine which *regions* from the reference match to the contigs' sequence. This is accomplished by right-clicking on the contig we want to annotate in the visual representation and by choosing the "*BLAST against the next position sequences*" option on the menu. After performing this step, the next one is to choose the option to annotate the contig, which is also made by right-clicking on the contig but this time we choose the "*Annotate Region*" option on the menu.

After the annotation of all contigs, which took about 20 minutes for each contigs file, it was possible to transfer annotations of 1544 *regions* of the reference to ContigsF1 and 1537 for ContigsF2 in a total of 1887 annotated *regions* in the reference, which corresponds to a rate of transfer of 81.8% and 81.4% respectively (Figure 4.3-c).

Files with the annotated contigs were exported after the analysis in a combination of GFF and FASTA files using the "*Export files*" option of the actions menu.

File Name	Genome Position	Name	Total Size	Annotated Portion	File Type	Number of Annotations	Number of Transposases	Number of IS	% of Hypothetical Proteins	Number of Contigs
ContigsF1.fasta	1	Contigs File	2016870 bp	-	.fasta	-	-	-	-	100
NC_017592.gbk	2	Streptococcus pneumoniae OXC141	2036867 bp	1667190 bp (81.85%)	.gbk	1887	9	9	23.95%	-
ContigsF2.fasta	3	Contigs File	2015251 bp	-	.fasta	-	-	-	-	107

(a)

File Name	Genome Position	Name	Total Size	Annotated Portion	File Type	Number of Annotations	Number of Transposases	Number of IS	% of Hypothetical Proteins	Number of Contigs
ContigsF1.fasta	1	Contigs File	2003527 bp	-	.fasta	-	-	-	-	47
NC_017592.gbk	2	Streptococcus pneumoniae OXC141	2036867 bp	1667190 bp (81.85%)	.gbk	1887	9	9	23.95%	-
ContigsF2.fasta	3	Contigs File	1999840 bp	-	.fasta	-	-	-	-	53

(b)

File Name	Genome Position	Name	Total Size	Annotated Portion	File Type	Number of Annotations	Number of Transposases	Number of IS	% of Hypothetical Proteins	Number of Contigs
ContigsF1.fasta	1	Contigs File	2002168 bp	1501477 bp (74.99%)	.fasta	1544	11	11	18.85%	47
NC_017592.gbk	2	Streptococcus pneumoniae OXC141	2036867 bp	1667190 bp (81.85%)	.gbk	1887	9	9	23.95%	-
ContigsF2.fasta	3	Contigs File	1993368 bp	1490531 bp (74.77%)	.fasta	1537	12	11	18.74%	53

(c)

Figure 4.3: Global information before / after ordering and annotation of contigs. (a) Global information after upload of two files with contigs (ContigsF1 and ContigsF2) and an annotated genome of *Streptococcus pneumoniae OXC141*. (b) Global information after ordering the contigs files against the reference genome. The number of contigs in question has been halved because some of them did not aligned with the reference with the parameters used. (c) Global information after all the contigs were annotated.

5

Discussion & Final Remarks

5 Discussion & Final Remarks

5.1 Discussion

In this thesis we described the development of ProGenViZ, a prokaryotic genome comparison tool that provides an interactive way to visualize local and global associations between genomes by their sequence or by their annotations. Another key feature it the ability to interact with HTS contig data. Moreover, it also offers additional features such as re-annotation of genes and annotation of contig data. ProGenViZ can be accessed online (http://darwin.phyloviz.net/ProGenViZ) but the source code can also be obtained (https://github.com/B-UMMI/ProGenViZ) to be installed locally.

The capabilities that have been developed in this program enable its use in several comparative genomics studies. The search for specific genes, the detection of sequence variations, as well as the rapid visualization of pairwise comparisons between genomes and HTS contigs data, allow a quick analysis of the genomic information in terms of gene content variation. This can be applied directly in Clinical Microbiology and research settings to detect factors associated with disease and for the characterization of microorganisms, as well as in other areas of study that rely on visualization of comparisons between sequences.

The visual framework developed proved to be able to represent multiple genomes, with most of the computation requirements being done at server-side. On the client-side the image creation is the more computational demanding step. No limit was imposed for the number of genomes to be analysed at the same time however, the tool interaction fluidity depends on the processing speed of the personal computer used, mainly due to changes made when the image is manipulated. That said, the representation of 3 or 4 genomes is advisable to keep the all-round performance of the application.

The primary method chosen to represent and visualize complete genomes, the hive plot, proved efficient in the distinction of local comparisons, as well as to show global and local synteny. By comparing with other layouts such as the circle or the linear, the hive plot has the advantage of showing relationships between *regions* in multiple genomes in a more straightforward fashion, avoiding cluttering of multiple links. Also, the associations between genes far apart from each other are easier to visualize, providing a “birds-eye” view of several genome relationships. Nevertheless, the visual perception of differences in the overall size of distinct genomes is easily accomplished with the linear representation. By allowing to exchange

between the hive plot and the linear representation, we aimed to provide the users with the benefits of either view.

The approach of dividing the genome's total sequence into *regions* according to their annotations was chosen because it lowers the complexity of the representation. Moreover, it facilitates the comparison between annotated areas. However, it has some disadvantages. Currently, using this tool, users are limited to use the entire *region*'s sequence when making BLAST searches, not giving the user full control to subdivide them. Also, despite the existence of proportionality between the genomes' representation and their actual sizes, in some cases can lead to discrepancies in the overall size. The division of the genomic sequences according to their annotations into *regions* and the consequent division of those *regions* into multiple *nodes* of 500bp does not allow the distinction of *regions* with less than 500bp apart, representing them all the same way. If we have a *region* that represents a gene with 800bp and one that represents other with 500bp, they will be shown the same way in the visual representation. It should be noted that our objective was not to create an exact and real representation of genomes but to develop an intuitive and fast way to explore them.

The creation of tables that show general information about the files and the results of different queries are an asset in this tool. They complement the visual representation and facilitate the access to more specific information such as the *regions*' positions and their products, as well as access to information about the sequence alignments performed. The ability to quickly access the location in the image of the *regions* involved in query results through the table serves as a connecting bridge and allows interaction between the two elements. Moreover, the possibility to visualize alignments at sequence level and detect single nucleotide variation, coupled with the ability to represent the complete genomes and their relationships, turns the application capable of performing analyses at different levels of genome organization.

We also present ways to visualize statistics associated with *regions*' sizes and products. With them it is possible to get an overview of the distribution of products in the genome and the *regions*' size distribution. However, the visualization methods were proved less sensitive to small variations when there are a large number of different products in the analysis. This problem can be overcome by analysing each gene sizes sections separately, reducing the number of products in analysis.

Unlike other tools to visualize whole genome comparisons, ProGenViZ does not show all the existing relationships directly. The displayed comparisons depend on the user choices. This approach was chosen because information overload frequently complicate the analyses and often the detection of locations of interest does not necessarily depends on whole genome

comparisons, only being necessary to compare specific areas of interest. In the case of users having to perform analysis where a global visualization of comparisons is more suited for their data, other available software like BRIG or Circos should be used.

When we performed queries for genes belonging to the MLST scheme of *Streptococcus pneumoniae*, we showed the characteristics of the tool to search for genes by their annotations and sequences. The analysis highlighted one of the major problems of genome annotation in public available genomes, where not even genes that are essential to maintain the basic functions of the cell were classified with the correct name. When *basic queries* by annotations returned no results, the ability to use external sequences by ProGenViZ to perform queries proved essential and more reliable, revealing matches when annotations did not. This analysis also demonstrated that is not yet possible to conduct queries using annotations and at the same time have full confidence in the results. Manual curation usually mitigates this problem but the use of sequences to perform queries for specific *regions* is therefore more reliable.

We also demonstrated the tool's features to explore and analyse contigs data. Incorporating contigs data in the program reduces the number of steps to be taken from the raw data obtained by HTS technologies until the sequence analysis because only the assembly of reads is necessary for the data be used as input and conduct searches by sequence homology and find patterns of interest. The search for the *cps* locus genes in contigs and their consequent annotation using an annotated reference genome demonstrated the ability of the program to perform rapid analysis and find patterns of interest that can assist in characterization of organisms and that offers a direct way to discover the contigs content using direct comparison with a reference.

The homology-based approach to annotate all contigs from a file allowed the transfer of an average of 87% of the annotations from the reference to the contigs. One of the reasons so that the numbers were not higher was due to the alignment parameters used for the contigs ordering that, despite eliminate any overlapping contigs, dismissed some of the smaller ones. Moreover, some genomic regions may not had enough coverage depth when sequencing was carried out, which leads to unrepresented regions on the resulting contigs. Another factor that may have contributed so that the percentage of transferred genes were not greater was the possible discrepancies between the alignments and the predictions made by Prodigal. Also, since Prodigal only predicts CDSs, annotated *regions* from the reference that are not genes could not be transferred.

When comparing ProGenViZ with other sequence comparison tools available, it not only provides the essential capabilities that characterize them, as innovates by incorporating new

	ProGenViZ	Circos	Cinteny	SynBrowse	Vista	Combo	ACT	BRIG	SynTView	GSV
Standalone tool/web-application	-/x	x/-	-/x	x/-	-/x	x/-	x/-	x/-	x/x	-/x
Shows an interactive representation	x	-	x	x	x	x	x	-	x	-
Support contig data	x	x	-	-	-	-	-	x	-	-
Built in comparison software	x	-	-	-	x	-	-	x	-	-
Requires a file with comparisons	-	x	x	x	-	x	x	-	x	x
SNP determination and visualization	x	-	-	-	-	-	-	-	x	-
Load annotations from existing files (e. g. Genbank, EMBL)	x	-	x	x	-	x	x	-	x	x
Query on existing annotations	x	-	-	x	-	-	-	-	x	-
Represents multiple/pairwise Comparisons	-/x	x/x	-/x	-/x	x/x	x/x	-/x	x/x	x/x	-/x
Provides percentage identity and e-value filtering for BLAST	x	-	-	-	-	-	x	x	x	-
Order contigs against a reference	x	-	-	-	-	-	-	-	-	-
Annotate/re-annotate genomic regions	x/x	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-

Table 5.1: Differences between ProGenViZ and other sequence comparison tools available. The X means that the program has the feature described.

features and combine functionalities available in multiple programs (Table 5.1). The target user for ProGenViZ, are non-bioinformaticians that need a user-friendly interface for basic genome comparison operations. By being a web application, there is no installation needed to use ProGenViZ, and all the ancillary software it uses (BLAST, NUCmer and Prodigal). Unlike the vast majority of comparison viewers that use pre-made alignment files, comparison files or customized files to visualize the comparisons, the developed application gives ability to control the comparisons that are made and the parameters that are used to carry out BLAST searches, representing them in real-time. Moreover, it allows the use of contigs data like Circus or BRIG, but adding also the ability to interact with the visual representation, instead of providing a static representation of the genome comparison like the former.

Although ProGenViz does not allow a simultaneous representation of multiple comparisons of sequences from different files against the same reference, multiple pairwise comparisons can be performed and viewed simultaneously in a stack, as in ACT, or through the Hive plot representation. Moreover, each uploaded genome can be set by the user as a reference or as a query depending on its position in the visual representation.

ProGenViZ allows data exploration from a whole genome perspective through annotations such as SynBrowse or SynTView, adding new ways of interaction between the image and the Hits table. Tools like VISTA, Combo, Cinteny and other global comparison viewers focus exclusively on the global representation of genomes, not giving emphasis to viewing nucleotide sequences. We offer not only an overview of the various annotated genomes and sequences, as BLAST alignments can be shown at the nucleotide level accompanied by the detection of SNPs.

Some additional features were also implemented that are not found in other tools. Options to order contigs against a reference and capabilities to annotate contigs based on sequence homology were developed. Moreover, the ability to re-annotate *regions* is also made available, allowing the user to redefine erroneous annotations by comparison with well-annotated genomes.

We believe that it is the combination of presented features and the fact it is a freely available web based tool that makes ProGenViZ a useful and powerful tool for every day analysis of prokaryotic genomes and draft genomes.

5.2 Final Remarks & Future Work

A visual representation of complex data has a very important role in the perception of information. Its main advantage is the ability to summarize large amounts of data, making possible to understand the information provided by hundreds of thousands of objects simultaneously. In the case of representation of genomic sequences, an image promotes the perception of data properties that otherwise would not be possible as the determination of synteny between genomes and detection of genomic regions of interest.

With each passing day, more sequencing data and draft genomes are being added to public databases, mainly due to the recent year's advances in high-throughput sequencing technologies. However, despite the continuous development of faster and more efficient methods for obtaining genomic data, we must not forget that it is also necessary to focus on developing tools to analyze it because the limiting step is still the data analysis of this ever increasing amount of information.

ProGenViZ puts together visual representations of genomic sequences and the basic exploratory functionalities needed in comparative genomic studies in a user-friendly interface, removing the need of programming knowledge typically required to perform such analyses in freely available software. The guiding principle of ProGenViz development was not the analysis and visualization of large studies of tens to thousands of strains, but to empower the everyday user in the visualization and querying the continuous flow of complete and draft prokaryotic genome data. The user friendly interface, the search for specific genes, the detection of sequence variations, the annotation of sequences, as well as the rapid visualization of pairwise comparisons between genomic sequences and draft genomes, offer efficient ways to analyze genetic information in terms of gene content variation by any individual with background in Biology. Moreover, the simple approach chosen to represent multiple genomes and to visualize comparisons between them facilitates the understanding by the user of the large amount of genomic data obtained from HTS. These features can be directly applied in research to detect factors associated with disease and for the characterization of microorganisms, as well as in other areas of study that rely on comparison and search for specific genomic regions and sequences. However, there are some aspects for the tool that can be improved.

On the usability aspect, our objective is to develop a user management system that allows users to create and save different data analysis sessions facilitating future access to previous analysis. By providing users with data persistence capacity, another goal would be to allow users to create their own private databases of sequence that could be reused in any of their analysis. With the existence of an area for each user, other methods of sequence analysis that are more

time consuming could be incorporated in the application. One would be to create an approach to provide a more thorough capability to annotate sequences by integrating the software Prokka[106] . This would add to ProGenViZ the ability to annotate genomes of prokaryotes without the need to perform annotation transfer from a reference sequence. Also, because currently ProGenViZ only annotates CDSs, Prokka would cause annotations to be more complete with the annotation of other elements such as tRNA, snRNA and rRNA.

ProGenViZ can also be upgraded through the incorporation of more analysis methods for extracting additional information about the data, both globally, like the GC content of a sequence or part of the sequence, and by the creation of more precise methods to compare specific genomic sequences, such as the possibility to perform BLAST searches with a sub-division of the sequence from a *region* defined in ProGenViZ. We also aim to allow the manual annotation of specific regions of sequences after having matches from external sequence queries defined by the user. In terms of visualization, new layouts can be created to represent the genomes at different levels of complexity, such as a circular layout to easily get a “birds eye view” of synteny between similar genomes and also develop a better nucleotide level visualization for all regions and not only for the HSPs aligned obtained from BLAST.

Methodologies for visual representation of genomic data and comparison have evolved over the years with different approaches, from global representations of single or multiple genomes and comparisons between them, to representations of genomic sequences up to the nucleotide level. However, the visual methods that exist today still have a long way to go before they can cope with the current demands of research as visualization of comparisons between hundreds of genomes. Despite the ProGenViZ represent and compare multiple draft genomes with only one image, is necessary to continue the development of new visual representations where the user can visualize data of hundreds of genomes simultaneously but still retrieving the relevant information.

6

Bibliography

6 Bibliography

1. Pallen MJ, Loman NJ, Penn CW: **High-throughput sequencing and clinical microbiology: progress, opportunities and challenges.** *Curr Opin Microbiol* 2010, **13**:625–31.

2. Nielsen CB, Cantor M, Dubchak I, Gordon D, Wang T: **Visualizing genomes: techniques and challenges.** *Nat Methods* 2010, **7**(3 Suppl):S5–S15.

3. Stephen F. AltschuP, Warren Gish, Webb Miller EWM and DJL: **Basic Local Alignment Search Tool**. *J Mol Biol* 1990:403–410.

4. Delcher AL, Phillippy A, Carlton J, Salzberg SL: **Fast algorithms for large-scale genome alignment and comparison.** *Nucleic Acids Res* 2002, **30**:2478–83.

5. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ: **Prodigal: prokaryotic gene recognition and translation initiation site identification.** *BMC Bioinformatics* 2010, **11**:119.

6. Gregor Mendel: **Versuche über Plflanzenhybriden**. 1865.

7. Moore J: **Heredity and development.** *Hered Dev* 1963.

8. Dahm R: **Friedrich Miescher and the discovery of DNA.** *Dev Biol* 2005, **278**:274–88.

9. Darwin C: **On the origins of species by means of natural selection**. *London: Murray* 1859:1–313.

10. McCarty M, Avery O: **STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES II. EFFECT OF**. *J Exp Med* 1944, **79(2)**:137–158.

11. HERSHEY AD, CHASE M: **Independent functions of viral protein and nucleic acid in growth of bacteriophage.** *J Gen Physiol* 1952, **36**:39–56.

12. Watson JD: **Molecular structure of nucleic acids. A structure for deoxyribose nucleic acid.** *JAMA J Am Med Assoc* 1993, **269**:1966–1967.

13. Leslie A. Pray: **Semi-conservative DNA replication: Meselson and Stahl**. *Nat Educ* 2008, **1(1)**:98.

14. CRICK F: **Central Dogma of Molecular Biology**. *Nature* 1970, **227**:561–563.

15. CRICK FHC, BARNETT L, BRENNER S, WATTS-TOBIN RJ: **General Nature of the Genetic Code for Proteins**. *Nature* 1961, **192**:1227–1232.

16. MESELSON M, YUAN R: **DNA Restriction Enzyme from E. coli**. *Nature* 1968, **217**:1110–1114.

17. Jackson DA, Symons RH, Berg P: **Biochemical method for inserting new genetic information into DNA of Simian Virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of Escherichia coli.** *Proc Natl Acad Sci U S A* 1972, **69**:2904–9.

18. Maclean D, Jones JDG, Studholme DJ: **Application of "next-generation" sequencing technologies to microbial genetics**. *Nat Rev Microbiol* 2009, **7**:96–97.

19. Varshney RK, Nayak SN, May GD, Jackson S a: **Next-generation sequencing technologies and their implications for crop genetics and breeding.** *Trends Biotechnol* 2009, **27**:522–30.

20. Berglund EC, Kiialainen A, Syvänen A-C: **Next-generation sequencing technologies and applications for human genetic history and forensics.** *Investig Genet* 2011, **2**:23.

21. Shokralla S, Spall JL, Gibson JF, Hajibabaei M: **Next-generation sequencing technologies for environmental DNA research.** *Mol Ecol* 2012, **21**:1794–805.

22. Gilbert W, Maxam a: **The nucleotide sequence of the lac operator.** *Proc Natl Acad Sci U S A* 1973, **70**:3581–4.

23. Maxam a M, Gilbert W: **A new method for sequencing DNA. 1977.** *Biotechnology* 1992, **24**:99–103.

24. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors.** *Proc Natl Acad Sci U S A* 1977, **74**:5463–7.

25. Kolata G: **The 1980 Nobel Prize in Chemistry**. *Science (80-)* 1980, **210**:887–889.

26. Lloyd M. Smith, Jane Z. Sanders, Robert J. Kaiser, Peter Hughes, Chris Dodd, Charles R. Connell, Cheryl Heiner SBHK& LEH: **Fluorescence detection in automated DNA sequence analysis**. *Nature* 1986, **321**:674–679.

27. Collins FS, Morgan M, Patrinos A: **The Human Genome Project: lessons from large-scale biology.** *Science* 2003, **300**:286–90.

28. Sinsheimer R: **The Santa Cruz Workshop—May 1985**. *Genomics* 1989, **5**:954–956.

29. Watson J: **The human genome project: past, present, and future**. *Science (80-)* 1990, **248**:44–49.

30. Shendure J, Mitra RD, Varma C, Church GM: **Advanced sequencing technologies: methods and goals.** *Nat Rev Genet* 2004, **5**:335–44.

31. Swerdlow H, Wu SL, Harke H, Dovichi NJ: **Capillary gel electrophoresis for DNA sequencing. Laser-induced fluorescence detection with the sheath flow cuvette.** *J Chromatogr* 1990, **516**:61–7.

32. Venter JC: **GENOMICS: Shotgun Sequencing of the Human Genome**. *Science (80-)* 1998, **280**:1540–1542.

33. Metzker ML, Lu J, Gibbs R a: **Electrophoretically uniform fluorescent dyes for automated DNA sequencing.** *Science* 1996, **271**:1420–2.

34. Ju J, Ruan C, Fuller CW, Glazer a N, Mathies R a: **Fluorescence energy transfer dye-labeled primers for DNA sequencing and analysis.** *Proc Natl Acad Sci U S A* 1995, **92**:4347–51.

35. Reeve MA, Fuller CW: **A novel thermostable polymerase for DNA sequencing.** *Nature* 1995, **376**:796–7.

36. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, et al.: **The sequence of the human genome.** *Science* 2001, **291**:1304–51.

37. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford a, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan a, Sougnez C, et al.: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**:860–921.

38. Hattori M: **Finishing the euchromatic sequence of the human genome.** *Nature* 2004, **431**:931–45.

39. Metzker ML: **Sequencing technologies - the next generation.** *Nat Rev Genet* 2010, **11**:31–46.

40. Schuster S: **Next-generation sequencing transforms today's biology**. *Nature* 2007, **5**:16–18.

41. Loman NJ, Constantinidou C, Chan JZM, Halachev M, Sergeant M, Penn CW, Robinson ER, Pallen MJ: **High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity.** *Nat Rev Microbiol* 2012, **10**:599–606.

42. Kircher M, Kelso J: **High-throughput DNA sequencing--concepts and limitations.** *Bioessays* 2010, **32**:524–36.

43. Pareek CS, Smoczynski R, Tretyn A: **Sequencing technologies and genome sequencing.** *J Appl Genet* 2011, **52**:413–35.

44. Wei L, Liu Y, Dubchak I, Shon J, Park J: **Comparative genomics approaches to study organism similarities and differences.** *J Biomed Inform* 2002, **35**:142–50.

45. Ali A: **Microbial Comparative Genomics: An Overview of Tools and Insights Into The Genus Corynebacterium**. *J Bacteriol Parasitol* 2013, **04**.

46. Fricke WF, Rasko D a: **Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions.** *Nat Rev Genet* 2014, **15**:49–55.

47. Köser CU, Ellington MJ, Cartwright EJP, Gillespie SH, Brown NM, Farrington M, Holden MTG, Dougan G, Bentley SD, Parkhill J, Peacock SJ: **Routine use of microbial whole genome sequencing in diagnostic and public health microbiology.** *PLoS Pathog* 2012, **8**:e1002824.

48. Underwood AP, Dallman T, Thomson NR, Williams M, Harker K, Perry N, Adak B, Willshaw G, Cheasty T, Green J, Dougan G, Parkhill J, Wain J: **Public health value of next-generation DNA sequencing of enterohemorrhagic Escherichia coli isolates from an outbreak.** *J Clin Microbiol* 2013, **51**:232–7.

49. Snitkin ES, Zelazny AM, Thomas PJ, Stock F, Henderson DK, Palmore TN, Segre J a: **Tracking a hospital outbreak of carbapenem-resistant Klebsiella pneumoniae with whole-genome sequencing.** *Sci Transl Med* 2012, **4**:148ra116.

50. La Scola B, Elkarkouri K, Li W, Wahab T, Fournous G, Rolain J-M, Biswas S, Drancourt M, Robert C, Audic S, Löfdahl S, Raoult D: **Rapid comparative genomic analysis for clinical microbiology: the Francisella tularensis paradigm.** *Genome Res* 2008, **18**:742–50.

51. Wu F, Mueller L a, Crouzillat D, Pétiard V, Tanksley SD: **Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the euasterid plant clade.** *Genetics* 2006, **174**:1407–20.

52. Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, Krabichler B, Speicher MR, Zschocke J, Trajanoski Z: **A survey of tools for variant analysis of next-generation genome sequencing data.** *Brief Bioinform* 2014, **15**:256–78.

53. Lemmon EM, Lemmon AR: **High-Throughput Genomic Data in Systematics and Phylogenetics**. *Annu Rev Ecol Evol Syst* 2013, **44**:99–121.

54. Schatz M, Delcher A, Salzberg S: **Assembly of large genomes using second-generation sequencing**. *Genome Res* 2010:1165–1173.

55. Li H, Homer N: **A survey of sequence alignment algorithms for next-generation sequencing.** *Brief Bioinform* 2010, **11**:473–83.

56. Mount D: **Steps used by the BLAST algorithm**. *Cold Spring Harb Protoc* 2007(May):2014.

57. Smith T, Waterman M: **Identification of common molecular subsequences**. *J Mol Biol* 1981:195–197.

58. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: **Versatile and open software for comparing large genomes.** *Genome Biol* 2004, **5**:R12.

59. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10**:R25.

60. Smith A, Chung W, Hodges E: **Updates to the RMAP short-read mapping software**. *Bioinformatics* 2009, **25**:2841–2.

61. Malhis N, Butterfield YSN, Ester M, Jones SJM: **Slider--maximum use of probability information for alignment of short sequence reads and SNP detection.** *Bioinformatics* 2009, **25**:6–13.

62. Malhis N, Jones SJM: **High quality SNP calling using Illumina data at shallow coverage.** *Bioinformatics* 2010, **26**:1029–35.

63. Ouzounis C a, Karp PD: **The past, present and future of genome-wide re-annotation.** *Genome Biol* 2002, **3**:COMMENT2001.

64. Stothard P, Wishart DS: **Automated bacterial genome analysis and annotation.** *Curr Opin Microbiol* 2006, **9**:505–10.

65. Do JH, Choi D-K: **Computational approaches to gene prediction.** *J Microbiol* 2006, **44**:137–44.

66. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M: **Pfam: the protein families database.** *Nucleic Acids Res* 2014, **42**(Database issue):D222–30.

67. Wheeler TJ, Clements J, Eddy SR, Hubley R, Jones T a, Jurka J, Smit AF a, Finn RD: **Dfam: a database of repetitive DNA based on profile hidden Markov models.** *Nucleic Acids Res* 2013, **41**(Database issue):D70–82.

68. Poptsova MS, Gogarten JP: **Using comparative genome analysis to identify problems in annotated microbial genomes.** *Microbiology* 2010, **156**(Pt 7):1909–17.

69. Ware C: *Information Visualization: Perception for Design*. 2013.

70. Thomas J, Cook K: *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. 2005.

71. Thomas J, Cook K: **A visual analytics agenda**. *IEEE Comput Graph Appl* 2006, **26**:10–13.

72. Yi JS, Kang YA, Stasko J, Jacko J: **Toward a deeper understanding of the role of interaction in information visualization.** *IEEE Trans Vis Comput Graph* 2007, **13**:1224–31.

73. Dam A Van, Feiner S: *Computer Graphics: Principles and Practice*. 2014.

74. Shneiderman B: **The eyes have it: a task by data type taxonomy for information visualizations**. In *Proc 1996 IEEE Symp Vis Lang*. IEEE Comput. Soc. Press; 1996:336–343.

75. Huang W, Marth G: **EagleView: a genome assembly viewer for next-generation sequencing technologies.** *Genome Res* 2008, **18**:1538–43.

76. Bao H, Guo H, Wang J, Zhou R, Lu X, Shi S: **MapView: visualization of short reads alignment on a desktop computer.** *Bioinformatics* 2009, **25**:1554–5.

77. Thorvaldsdóttir H, Robinson JT, Mesirov JP: **Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.** *Brief Bioinform* 2013, **14**:178–92.

78. Skinner ME, Uzilov A V, Stein LD, Mungall CJ, Holmes IH: **JBrowse: a next-generation genome browser.** *Genome Res* 2009, **19**:1630–8.

79. Mayor C, Brudno M, Schwartz JR, Poliakov a, Rubin EM, Frazer K a, Pachter LS, Dubchak I: **VISTA : visualizing global DNA sequence alignments of arbitrary length.** *Bioinformatics* 2000, **16**:1046–7.

80. Krzywinski M, Schein J, Birol İ: **Circos: an information aesthetic for comparative genomics**. *Genome …* 2009:1639–1645.

81. Alikhan N-F, Petty NK, Ben Zakour NL, Beatson S a: **BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons.** *BMC Genomics* 2011, **12**:402.

82. Pan X, Stein L, Brendel V: **SynBrowse: a synteny browser for comparative sequence analysis.** *Bioinformatics* 2005, **21**:3461–8.

83. Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J: **ACT: The Artemis comparison tool**. *Bioinformatics* 2005, **21**:3422–3423.

84. Revanna K V, Chiu C-C, Bierschank E, Dong Q: **GSV: a web-based genome synteny viewer for customized data.** *BMC Bioinformatics* 2011, **12**:316.

85. Sinha AU, Meller J: **Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms.** *BMC Bioinformatics* 2007, **8**:82.

86. Engels R, Yu T, Burge C, Mesirov JP, DeCaprio D, Galagan JE: **Combo: a whole genome comparative browser.** *Bioinformatics* 2006, **22**:1782–3.

87. Lechat P, Souche E, Moszer I: **SynTView - an interactive multi-view genome browser for next-generation comparative microorganism genomics.** *BMC Bioinformatics* 2013, **14**:277.

88. **Bootstrap** [http://getbootstrap.com/]

89. **D3.js** [http://d3js.org/]

90. **Python** [https://www.python.org/]

91. **JSON format** [http://www.json.org/]

92. **GFF file format** [http://www.ensembl.org/info/website/upload/gff.html]

93. **GenBank/EMBL file format** [http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html]

94. **FASTA file format** [http://www.bioinformatics.nl/tools/crab_fasta.html]

95. Krzywinski M, Birol I, Jones SJM, Marra MA: **Hive plots--rational approach to visualizing networks.** *Brief Bioinform* 2012, **13**:627–44.

96. **Hive Plot implementation in D3.js** [http://bost.ocks.org/mike/hive/]

97. **DataTables JQuery plug-in** [http://www.datatables.net/]

98. **JQuery JavaScript library** [http://jquery.com/]

99. Pérez-Losada M, Cabezas P, Castro-Nallar E, Crandall K a: **Pathogen typing in the genomics era: MLST and the future of molecular epidemiology.** *Infect Genet Evol* 2013, **16**:38–53.

100. **Streptococcus pneumoniae MLST Database** [http://spneumoniae.mlst.net/]

101. **National Center for Biotechnology Information (NCBI)** [http://www.ncbi.nlm.nih.gov/]

102. Bentley SD, Aanensen DM, Mavroidi A, Saunders D, Rabbinowitsch E, Collins M, Donohoe K, Harris D, Murphy L, Quail M a, Samuel G, Skovsted IC, Kaltoft MS, Barrell B, Reeves PR, Parkhill J, Spratt BG: **Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes.** *PLoS Genet* 2006, **2**:e31.

103. Serrano I, Melo-Cristino J, Carriço JA, Ramirez M: **Characterization of the genetic lineages responsible for pneumococcal invasive disease in Portugal.** *J Clin Microbiol* 2005, **43**:1706–15.

104. Mavroidi A, Aanensen DM, Godoy D, Skovsted IC, Kaltoft MS, Reeves PR, Bentley SD, Spratt BG: **Genetic relatedness of the Streptococcus pneumoniae capsular biosynthetic loci.** *J Bacteriol* 2007, **189**:7841–55.

105. Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Res* 2008, **18**:821–9.

106. Seemann T: **Prokka: rapid prokaryotic genome annotation.** *Bioinformatics* 2014, **30**:2068–9.

107. Richardson EJ, Watson M: **The automatic annotation of bacterial genomes.** *Brief Bioinform* 2013, **14**:1–12.

108. Schatz MC, Phillippy AM, Shneiderman B, Salzberg SL: **Hawkeye: an interactive visual analytics tool for genome assemblies.** *Genome Biol* 2007, **8**:R34.

109. Manske HM, Kwiatkowski DP: **LookSeq: a browser-based viewer for deep sequencing data.** *Genome Res* 2009, **19**:2125–32.

110. Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, Marshall D: **Tablet--next generation sequence assembly visualization.** *Bioinformatics* 2010, **26**:401–2.

111. Stothard P, Wishart DS: **Circular genome visualization and exploration using CGView.** *Bioinformatics* 2005, **21**:537–9.

112. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S: **The generic genome browser: a building block for a model organism system database.** *Genome Res* 2002, **12**:1599–610.

113. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler a. D: **The Human Genome Browser at UCSC**. *Genome Res* 2002, **12**:996–1006.

7

Appendices

7 Appendices

Appendix-1 – Allele sequences of the MLST scheme genes of S. pneumoniae used in use case 1

The following sequences correspond to the alleles extracted from the MLST database[100] and used to search for genes belonging to the MLST scheme for *S. pneumoniae* in the use case 1:

>aroe1
GAAGCGAGTGACTTGGCAGAAACAGTGGCCAATATTCGTCGCTACCAGATGTTTGGCATCAATCTGTCCATGCCCT
ATAAGGAGCAGGTGATTCCTTATTTGGATGAGCTAAGCGATGAAGCGCGCTTGATTGGTGCGGTTAATACGGTTG
TCAATGAGAATGGCAATTTAATTGGATATAATACAGATGGCAAGGGATTTTTTAAGTGCTTGCCTTCTTTTACAATT
TCAGGTAAAAAGATGACCCTGCTGGGTGCAGGTGGTGCGGCTAAATCAATCTTGGCACAGGCTATTTTGGATGGC
GTCAGTCAGATTTCGGTCTTTGTTCGTTCCGTTTCTATGGAAAAAACAAGACCTTACCTAGACAAGTTACAGGAGC
AGACAGGCTTTAAAGTGGATTTGTGT

>gdh1
AGAACACTTTATCCGTGGGCAATACCGCTCTGGTAAGATTGATGGCATGAAATACATCTCTTATCGTAGCGAGCCA
AATGTGAATCCAGAATCAACAACTGAAACCTTTACATCAGGTGCCTTCTTTGTAGACAGCGATCGATTCCGTGGTG
TTCCTTTCTTTTTCCGTACAGGTAAACGACTGACTGAAAAAGGAACCCATGTCAACATCGTCTTTAAACAAATGGAT
TCTATCTTTGGAGAACCACTTGCTCCAAATATTTTGACCATCTATATTCAACCAACAGAAGGCTTCTCTCTTAGCCTA
AATGGGAAGCAAGTAGGAGAAGAATTTAACTTGGCTCCTAACTCACTTGATTACCGTACAGATGCGACTGCAACT
GGTGCTTCTCCAGAACCATACGAAAAATTGATTTATGATGTCCTAAATAACAACTCAACTAACTTTAGCCACTGGGA
T

>gki1
ACCCTTCAACCAATCAAACAAAAGATTGAAAAAGCTTTGGGCATTCCATTTTTCATCGATAATGATGCCAACGTAGC
AGCTCTTGGTGAGCGCTGGATGGGTGCTGGAGATAACCAACCAGACGTTGTCTTTATGACACTCGGTACTGGTGT
TGGTGGCGGTATCGTCGCAGAAGGCAAATTGCTTCACGGTGTTGCTGGTGCAGCAGGTGAGCTTGGTCACATCAC
TGTTGACTTTGACCAGCCAATCTCATGTACTTGTGGTAAAAAAGGCTGCCTTGAGACAGTTGCTTCAGCAACAGGG
ATTGTCAACTTGACTCGTCGCTATGCCGATGAATACGAAGGCGATGCAGCCTTGAAACGCTTGATTGATAACGGA
GAAGAAGTAACTGCTAAGACTGTCTTTGATCTCGCAAAAGAAGGAGACGACCTTGCTTTGATTGTTTACCGTAACT
TCTCACGTTACTTGGGAATCGCTTGTGCT

>recP1
CTCAACCAAACTGGATTAACCGCGACCGCTTTATTCTTTCAGCAGGTCATGGTTCAATGCTCCTTTATGCTCTTCTTC
ACCTTTCTGGTTTTGAAGATGTCAGCATGGATGAGATTAAGAGTTTCCGTCAATGGGGTTCAAAAACACCAGGTCA
CCCAGAATTTGGTCATACGGCAGGGATTGATGCTACGACAGGTCCTCTAGGGCAAGGGATTTCAACTGCTACTGG
TTTTGCCCAAGCAGAACGTTTCTTGGCAGCCAAATATAACCGTGAAGGCTACAATATCTTTGACCACTATACTTACG
TTATCTGTGGAGACGGAGACTTGATGGAAGGTGTCTCAAGCGAGGCAGCTTCATACGCAGGTTTGCAAAAACTTG
ATAAGTTGGTTGTTCTTTATGATTCAAATGATATCAACTTGGATGGTGAGACAAAGGATTCCTTTACAG

>spi1
GTATCTTTTTTTGGAGCAATGTTCGCGTAGAAGGACATTCCATGGATCCGACCCTAGCGGATGGCGAAATTCTCTT
CGTTGTAAAACACCTTCCTATTGACCGTTTTGATATCGTGGTTGCCCATGAGGAAGATGGCAATAAGGACATCGTC
AAGCGCGTGATTGGAATGCCTGGCGACACCATTCGTTACGAAAATGATAAACTTTACATCAATGACAAAGAAACG
GACGAGCCTTATCTAGCAGACTATATCAAACGCTTCAAGGATGACAAACTCCAAAGCACTTACTCAGGCAAGGGCT
TTGAAGGAAATAAAGGAACTTTCTTTAGAAGTATCGCTCAAAAAGCCCAAGCCTTCACAGTTGATGTCAACTACAA
CACCAACTTTAGCTTTACTGTTCCAGAAGGAGAATACCTTCTCCTCGGAGATGACCGCTTGGTTTCGAGCGACAGC
CGCCACGTAGGTACCTTCA

>xpt1
GGTGATAACATCCTCAAGGTAGATTCCTTTTTAACCCACCAAGTTGACTTTAGCTTGATGCGAGAGATTGGTAAGG
TTTTTGCGGAAAAATTTGCTGCTACTGGCATTACCAAGGTCGTAACCATTGAAGCGTCGGGTATTGCCCCAGCCGT
TTTTACAGCTGAAGCCTTAAACGTTCCCATGATTTTCGCCAAAAAAGCTAAGAACATCACCATGAACGAAGGCATC
TTAACTGCTCAAGTCTACTCCTTTACCAAGCAGGTGACCAGCACTGTTTCTATCGCTGGAAAATTCCTCTCACCAGA
GGACAAGGTTTTGATTATCGACGATTTCCTTGCTAATGGCCAAGCTGCTAAAGGCTTGATTCAAATCATCGAACAG
GCCGGTGCCACAGTCCAAGCTATCGGTATCGTGATTGAGAAATCCTTCCAAGATGGTCGTGATTTGCTTGAAAAAG
CAGGCTACCCTGTCCTATCACTTGCTCGC

>ddl1
GCTAAAATAGCTGAAGTGGAAGAAAAATTGGCTTATCCAGTCTTCACTAAGCCGTCAAACATGGGGTCTAGTGTC
GGTATTTCTAAGTCTGAAAACCAAGAAGAACTCCGTCAAGCCTTAAAACTTGCCTTCCGATATGACAGCCGTGTCTT
GGTTGAGCAAGGAGTGAATGCCCGTGAAATTGAGGTTGGCCTCTTGGGTAACTACGATGTCAAGAGCACGCTACC
TGGAGAAGTTGTCAAGGACGTTGCCTTTTATGACTACGATGCCAAGTATATTGATAACAAGGTTACTATGGATATT
CCTGCCAAAATCAGTGATGATGTGGTGGCTGTCATGCGTCAAAATGCAGAAACAGCCTTCCGTGCCATTGGTGGC
CTTGGTCTATCTCGTTGCGATTTCTTCTATACAGATAAGGGAGAGATTTTTCTCAACGAGCTC

Appendix-2 – Assembly of HTS data from use case 3

The raw Illumina HiSeq 90bp Paired End read files where first filtered for quality using in-house scripts. The filtering criteria was the selection the largest substring with nucleotide quality above Q20 (99.99% certainty of base call) for each read. The minimum read size after filtering was set at 50bp.

The reads where then assembled using Velvet[105] assembler, with VelvetOptimiser script in order to optimize the k-mer size for N50. The range of k-mers size to test was estimated using the *velvetk.pl* script of the *velvetoptimiser* package. The minimum contig length was set to 200 nucleotides and the scaffolding option was allowed.

The command lines used were the following:

ContigF1:

VelvetOptimiser.pl –d /velvet_assembly –f ‘ –fastq.gz –short /Strain1.cleandata_500_1.p.fq.gz /Strain1.cleandata_500_2.p.fq.gz’ –t 8 –s 45 –e 57 –o ‘-min_contig_lgth 200 – scaffolding yes’

ContigF2:

VelvetOptimiser.pl –d /velvet_assembly –f ‘ –fastq.gz –short /Strain2.cleandata_500_1.p.fq.gz /Strain2.cleandata_500_2.p.fq.gz’ –t 8 –s 43 –e 55 –o ‘-min_contig_lgth 200 – scaffolding yes’