State the Art of Microbial Genome Analysis

Md Jubayer Hossain

Submitted To: **Prof. Shamima Begum, Ph.D.**

Department of Microbiology
Jagannath University
Roll: B150605021
Session: 2015-16

August 12, 2020

Abstract

Genomics is the study of whole genomes of organisms, and incorporates elements from genetics. Genomics uses a combination of recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble, and analyse the structure and function of genomes. Genomics have become an inter-disciplinary(Computer Science, Statistics, Biology) science. The genomic data analysis steps typically include data collection, quality check and cleaning, processing, modeling, visualization and reporting.

Contents

List of Figures

List of Tables

1

Introduction

1.1 A Little Intro to Bioinformatics and Genomics

Genomics and Bioinformatics has become a buzzword in today's world of Science. About one or two decades ago, people saw biology and computer science as two different fields. One would learn about living beings and their functions whereas the other would learn about computers and underlying theories. However, at present, there seems to be a mere separation between two fields and this new-field, bioinformatics, has emerged as a combination of both Computer Science and Biology. And genomics is the study of genomes. Bioinformatics methods and tools are frequently used in genomics research, but genomics also makes use of many experimental methods[11].

1.2 Why Learn and Apply Genomics?

Bioinformatics and genomics have become an inter-disciplinary science and if you are a biologist, you will find that having knowledge in bioinformatics can benefit you immensely with your experiments and research.

A major application of bioinformatics can be found in the fields of **precision medicine** and **preventive medicine**. Precision medicine consists of health care techniques customized for individual patients, including treatments and practices. Rather than treating or curing diseases, precision medicine focuses of developing measures to prevent diseases. Some of the diseases being focused are **influenza**,**cancer**, **heart disease** and **diabetes** [8].

Researches are being carried out to identify genetic alterations in patients allowing scientists to come up with better treatments and even possible measures of prevention. Certain types of cancer, being caused by such genetic alterations can be identified beforehand and can be treated before the conditions get worse. `https://en.wikipedia.org/wiki/Genomics`

1.3 What is Genome Analysis?

Modern biology is undergoing an historical transformation, becoming among other things increasingly data driven. A combination of statistical, computational, and biological methods has become the norm in modern genomic research. Of course this is at odds with the standard organization of university curricula, which typically focus on only one of these three subjects. It is hard enough to provide a good synthesis of computer science and statistics, let alone to include molecular biology! Yet, the importance of the algorithms typical of this field can only be appreciated within their biological context, their results can only be interpreted within a statistical framework, and a basic knowledge of all three areas is a necessary condition for any research project [2].

Genome analysis, also known as genome mining or in silico analysis,currently constitutes an irreplaceable research tool for various aspects of microbiology[10]. In particular, the availability of genomes from virtually all bacterial human pathogens has opened

perspectives in the fields of diagnosis, epidemiology, pathophysiology and treatment. A major advantage of genome sequences over phenotypic methods is that data can rapidly be shared among scientists worldwide by being deposited in online databases and thus are easily comparable among laboratories.

1.4 The Main Ideas

The genomic sequencing era may be divided into two periods (Figure1.1). In the first decade, from 1995, when the sequencing of the *Haemophilus influenzae* genome was performed [10] to 2005, sequencing relied on the classic Sanger method, was time- and money-consuming and was reserved to a limited number of sequencing centers worldwide. Fewer than 300 bacterial genomes were sequenced during this period (Figure1.1). Since 2005, the development of new and high-throughput sequencing methods,2 together with a steep decrease of the sequencers' and reagents' cost enabling many laboratories to develop their own sequencing projects, led to a striking increase in the number of sequenced genomes, approaching 6000 for the year 2013 alone. The tremendous source of information provided by genome sequences revolutionized basic aspects of microbiology. In particular,genome sizes of bacteria range from 139 kb for Candidatus Tremblaya princeps to 14 782 kb for Sorangium cellulosum (`http://genomesonline.org/`) [10]

With more than 49 000 bacterial genome sequences currently available,including those from all significant human pathogens, genomics has a significant impact on clinical microbiology and infectious diseases by enabling the development of improved diagnostic, genotyping, taxo-nomic, antibiotic and virulence marker detection tools as well as development of new culture media or vaccines. This chapter summarizes the current achievements in bacterial genomics relevant to medical microbiology.

1.5 Genomic Approaches

The ultimate goal of genetic association studies is both to define the genetic architecture of complex traits and diseases and also to provide new insights into normal physiology and

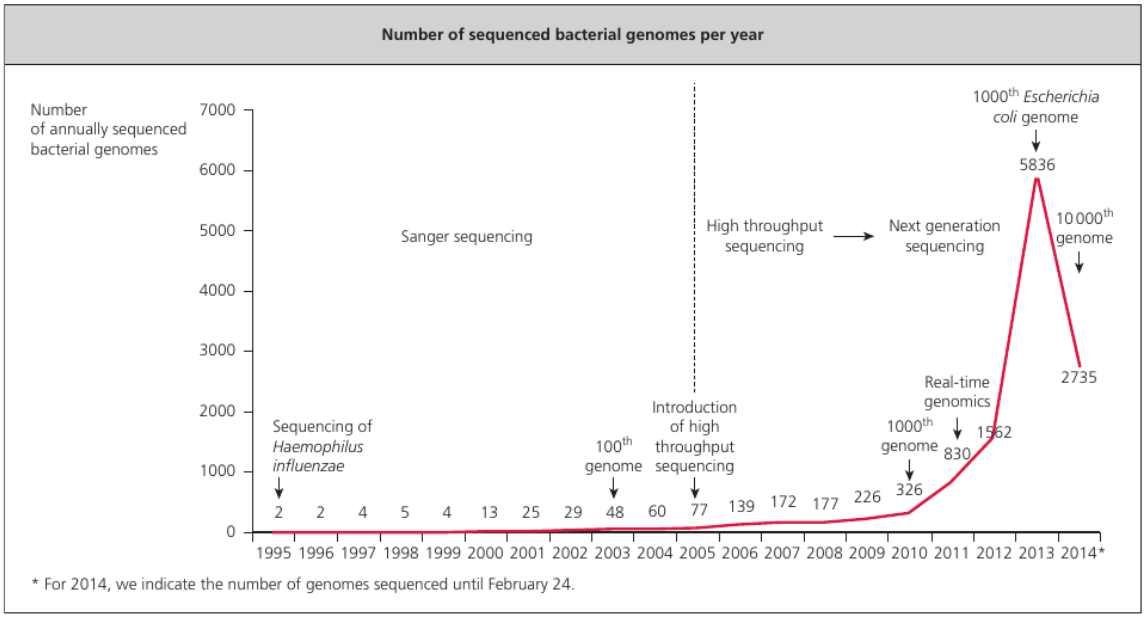

Figure 1.1: Number of sequenced bacterial genomes per year.

disease pathophysiology. Accomplishing that goal will require defining the causal variants that account for the observed associations, their mechanism of action, and their target genes. Success would have both near- and long-term benefits to health and science. In terms of health benefits, causal relationships between noncoding genetic variants and disease risk can be used to improve the prediction of disease onset and the design of prevention and early detection strategies. Subsequently determining the effects of causal variants on gene expression can prioritize downstream efforts to characterize causal genes and their role in disease etiology. That prioritization is particularly valuable when the target genes have an unknown function. This discovery pathway can ultimately lead to novel and potentially patient-specific therapeutic targets. In terms of scientific benefits, expanding the catalog of noncoding variants that are known to contribute to human traits is needed to determine general and transferrable principles about the genetic basis of complex human diseases. Recent conceptual and technical advances in genetics and genomics together have the potential to greatly improve our understanding of the noncoding genetic contributions to human traits. Although there are a wide variety of ways in which noncoding variants may affect phenotypes, we will focus specifically on variants that alter the activity of gene regulatory elements and, subsequently, the expression of target genes[8].

2

Genomic Data and Databases

2.1 What Are Genomic Data?

Genomic data refers to the genome and DNA data of an organism. They are used in bioinformatics for collecting, storing and processing the genomes of living things. Genomic data generally require a large amount of storage and purpose-built software to analyze.

2.2 Types of Genomic Data

Generally speaking, genomics data(Table 2.1) comes in four categories:

Table 2.1: Types of genomic data

Data Type	Unindexed format	Indexed formats
Sequence	FASTA	2bit
Annotations	BED, GTF2, GFF3, PSL	BigBed
Quantitative	bedGraph, wiggle	BigWig
Read alignments	bowtie, SAM, PSL	BAM

2.3 Genomic Databases

Genomic databases allow for the storing, sharing and comparison of data across research studies, across data types, across individuals and across organisms. These are not a new invention even before the popularisation of the modern internet, 'online' databases have been available in order to share data on key organisms, such as Escherichia coli (Blattner et al., 1997) and Saccharomyces cerevisiae (Cherry et al., 2012). Recent advances in both data sharing technology and genome sequencing technology have created an explosion of databases, based around particular organisms, as has been historically the case, as well as around particular data types, such as transcriptional data or short-read sequencing data [6].

2.4 Specific Organism Databases and the GMOD Project

It is possibly unsurprising that with the evolution of sequencing technology and the power to sequence the genome of most any organism, given a reasonable amount of time and a reasonable amount of research effort, individual databases have developed around the genomes of specific organisms [11]. In the past, this was mostly focussed around so-called 'model' organisms, or ones with large research bases, such as the mouse (Mus musculus) (Smith et al., 2018) and nematode (Caenorhabditis elegans) (Lee et al., 2018).

In many cases, these were created by their own research communities, to suit their own needs, both in terms of how the data could be accessed, as well as what tools were provided to dissect the data. As efforts continued, there have been moves to create some consistency between databases and the tools they offer, meaning new organism databases are not required to 're-invent the wheel' so to speak. In this regard, the Generic Model Organism Database (GMOD) project has served to provide a framework of tools and database

methods to allow new databases to be created. The 'users' of the GMOD project are no longer limited to 'model' organisms, and now consist of a variety of different species and databases. The GMOD project also has its own genome browser associated with it, GBrowse (as discussed further below), which can be integrated into the participating databases as a web-based genome browser [11].

2.5 Human Genome Databases

The breadth and depth of human genome databases is vast, as is to be expected when an organism attempts to study itself, and analyse its own biological problems. These databases are often structured around various data sources, such as transcriptional data,as is the case for the H-Invitational database (H-InvDB) (4). Particular study types have also given rise to specific databases:genome-wide association study databases such as GWASCentral (5), and structural variant study databases such as dbVar (6) and DGV (7). As is the case with other organisms, there are also some databases which seek to be more comprehensive in scope: DNA element databases such as ENCODE (8), and the 1000 Genomes project database, now hosted as the International Genome Sample Resource (IGSR) (9). Databases for even more specific purposes exist, such as a wealth of databases on cancer genomic data, and will need to be searched for on a case-by-case basis depending on need [7].

2.6 The Main Three Databases

The main three databases are the National Center for Biotechnology Information (NCBI, `www.ncbi.nlm.nih.gov/`), the DNA Data Bank of Japan (DDBJ, `www.ddbj.nig.ac.jp/`) and the European Bioinformatics Institute (EBI, `www.ebi.ac.uk/`). In addition to offering complete microbial genome sequences with links to corresponding publications, these databases provide online tools for analyzing genome sequences. As of February 24, 2014, 12 272 genome sequences from 2897 bacterial species are available online (www.genomesonline.org/, https://gold.jgi.doe.gov). For some species, several genomes have been sequenced. For 31 species, more than 50 genomes areavailable,including 16

species for which more than 100 genomes have been sequenced, the species holding the record being Escherichia coli, with 1261 currently available genomes. Sequenced genomes include the most significant human bacterial pathogens, covering all the phylogenetic domains of bacteria. In addition, more than 27 000 sequencing projects are ongoing (`www.genomesonline.org/`)[10]. Moreover, new sequencing technologies are making possible the sequencing of random community DNA and single cells of bacteria without the need for cloning or cultivation.

2.7 Genome Browsers

Data access and quality means very little if no meaning can be gained from it. In a field with as complex and abstract data as genomics, methods for data visualisation and analysis are of even greater importance. These must be able to cope with vast amounts of data, in the order of gigabytes or terabytes, as well as be able to connect these to tangible, biological meaning in the form of genes and products. Genome browsers seek to fill this need by providing a pre-existing software basis to visualise and analyse genomic data. Due to the sheer variety of researchers, purposes, expectations, and goals involved in the field, a number of genome browsers are available. For the new user, there are three broad-class, easy-to-pick-up databases for generic uses that stand out at present: the UCSC Genome Browser, managed by the University of California, Santa Cruz (Casper et al., 2018); GBrowse, managed by the GMOD project (see “Relevant Website section”); and Ensembl, managed by EMBL-EBI and the Wellcome Trust Sanger Institute (Zerbino et al., 2018). This section will consider each of these browsers in turn, and then give an overview of the more specific browsers which have been created based on these three forerunners[11].

2.7.1 UCSC Genome Browser

UCSC Genome Browser is the one of the most widely regarded broad-class browser, and has been integrated into a number of major databases. Its initial conception in 2000 was to visualise the first working draft of the Human Genome Project, but has been adapted in

the following years to include a broad variety of organisms, and a vast suite of tools for visualising and analysing data.(`https://genome.ucsc.edu/`)

2.7.2 Gbrowse

Due to the 'generic' nature of the GMOD project (discussed above at 2.2), there was a need for a generic browser to accompany the suite of tools provided for new databases. GBrowse developed from this idea, and is therefore one of the more flexible genome browsers available. It has had a number of spin-off browsers created since its conception, tailored for particular purposes. As it is a part of the GMOD project, it is also available across many different databases [11]. (`http://gmod.org/wiki/GBrowse`)

2.7.3 Ensembl

The Ensembl genome browser created by EMBL-EBI and the Wellcome Trust Sanger Institute is the native genome browser for the Ensembl Genomes databases. Due to the broad nature of the databases it is used for, it contains a broad variety of tools for visualisation and analysis across a variety of kingdoms of organisms.(`http://ensemblgenomes.org/`)

2.7.4 Specialised Browsers

Browsers for more specialised purposes have been developed by particular groups, largely based on one of the primary three browsers. Due to its generic nature, the majority of 'subsidiary' browsers are based on GBrowse in particular. Lighter implementations have been created, such as JBrowse, as well as browsers more suited to collaboration and annotation, such as Apollo.

3

Genomics Application

3.1 Role of Genomics in Clinical Microbiology

There are multiple applications for genomics in clinical microbiology[10].

- Real-time genomics may be used to investigate infectious disease outbreaks
- Bacterial genomes may be used as target sources for molecular detection, identification or genotyping.
- The gene content, obtained by comparison to databases such as Clusters of Orthologous Groups (`www.ncbi.nlm.nih.gov/COG/`) or Kyoto Encyclopedia of Genes and Genomes (`www.genome.ad.jp/kegg/`), may be searched for specific phenotypic traits such as virulence or antibiotic resistance markers, or deficient metabolic

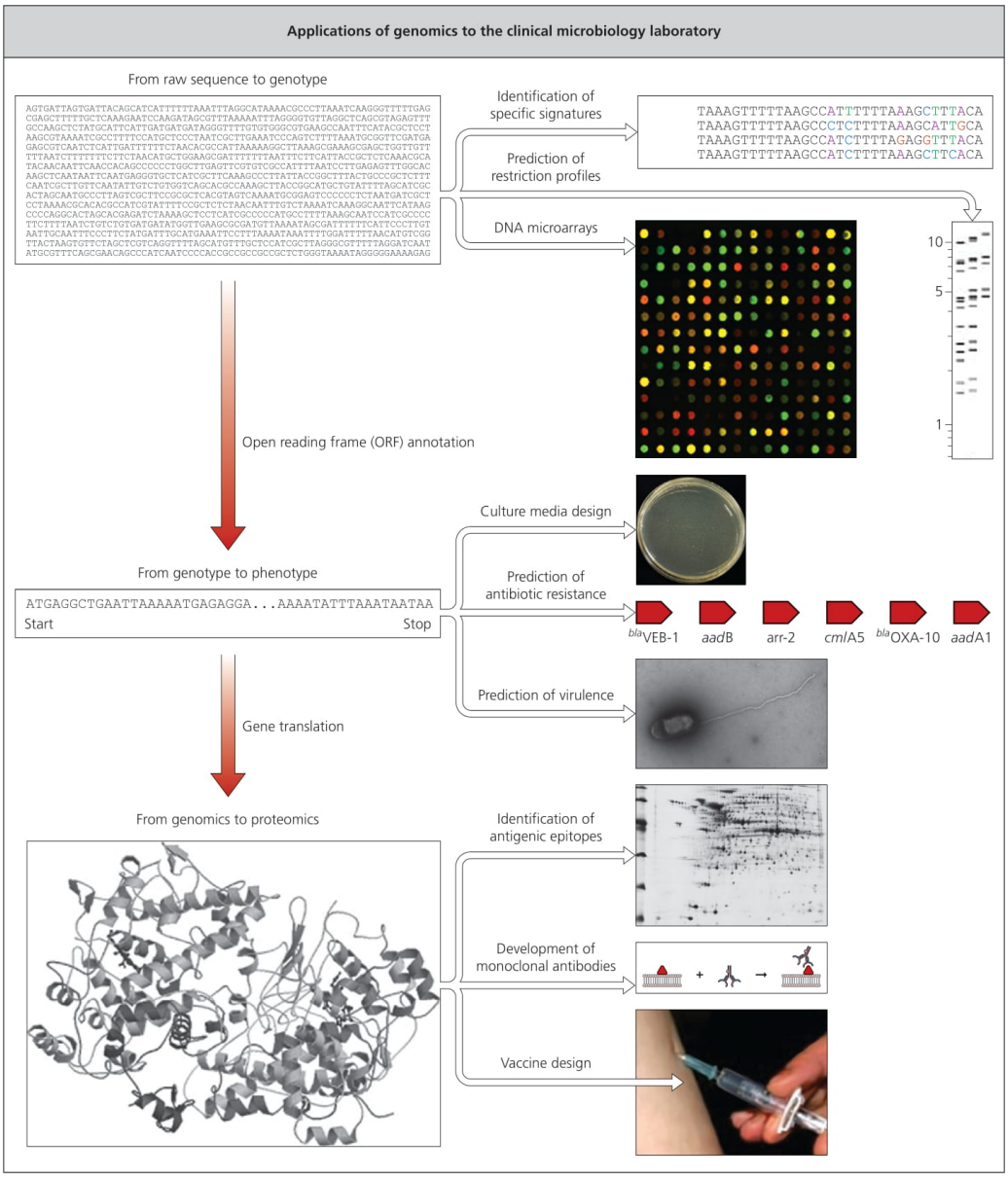

Figure 3.1: Applications of genomics to the clinical microbiology laboratory.

pathways enabling design of improved culture media.

- Antigenic epitopes detected in the deduced proteome may be used for serologic applications, development of monoclonal antibodies or development of vaccines (Figure 3.1).

- Taxonomic description of new bacterial species.

3.2 The Clinical Applications of Genomic Technologies

The clinical applications of genomic technologies are vast and offer opportunities to improve healthcare across the breadth of medical specialities. I will explore some of these applications in more depth this section.

3.2.1 Gene discovery and diagnosis of rare monogenic disorders

Genomic technologies can be used by clinicians from all specialities to diagnose their patients who have high-risk genetic errors causing disease. Researchers are using these techniques to identify new genes which cause genetic disease at an astonishing rate - over 4000 diseases now have a known single genetic cause, compared to around 50 in 1990.

3.2.2 Identification and diagnosis of genetic factors contributing to common disease

Genomic technologies are increasingly being used to understand the contribution of both rare and common genetic factors to the development of common diseases, such as high blood pressure, diabetes and cancer.

3.2.3 Pharmacogenetics and targeted therapy

Genetic information may be used to predict whether a person will respond to a particular drug, how well they will respond to that drug and whether they are likely to get any side effects from the use of a specific drug. This allows their treating team to make individualised decisions about the right drug treatment. In some cases, such as cancer, we can identify the genetic drivers of disease and then give drugs which specifically target that pathway. This is known as targeted therapy.

3.2.4 Prenatal diagnosis and testing

Genetic diseases are often devastating and may cause significant disability and even death in childhood. Prenatal diagnosis of genetic diseases allows parents to make decisions about whether to continue with the pregnancy or to allow early diagnosis and possible treatment in utero or at birth. Whilst previous approaches to prenatal diagnosis could put the pregnancy at risk, new methods using genomic technology can look directly at the DNA of the fetus from a maternal blood test, without increasing the risk of miscarriage - this is known as non-invasive prenatal testing. The use of NGS and array technology in prenatal samples is also on the increase to improve diagnostic yields in a pregnancy.

3.2.5 Infectious diseases

Sequencing the genomes of microorganisms which cause human infection can identify the exact organism causing symptoms, help to trace the cause of infectious outbreaks, and give information as to which antibiotics are most likely to be effective in treatment.

3.2.6 Personalised medicine

As the exact DNA sequence of the genome of each human is unique to them, we will all have unique disease susceptibilities and treatment responses. Personalised medicine describes the use of our genetic information to tailor health care intervention to our own individual need.

3.2.7 Gene therapy

Gene therapy involves the administration of DNA or RNA, in order to correct a genetic abnormality, or modify the expression of genes.

3.2.8 Genome editing

Genome editing uses molecular techniques to modify the genome - genome editing can add in, cut out, or replace sections of the DNA sequence.

3.2.9 Design of new antimicrobial agents and vaccines

One of the expected benefits of genome analysis of pathogenic bacteria is in the area of human health, particularly in the design of more rapid diagnostic reagents and the development of new vaccines and antimicrobial agents. These goals have become more urgent with the continuing spread of antibiotic resistance in important human pathogens. Moreover, results from the whole-genome analysis of human pathogens has suggested that there are mechanisms for generating antigenic variation in proteins expressed on the cell surface that are encoded within the genomes of these organisms[5]. These mechanisms include the following:

1. slipped-strand mispairing within DNA sequence repeats found in 58-intergenic regions and coding sequences as described for *H. influenzae2 Helicobacter pylori26* and *M. tuberculosis27*

2. recombination between homologous genes encoding outer-surface proteins as described for Mycoplasma genitalium28, Mycoplasma pneumoniae29 and Treponema pallidum30 (Figure 3.2)

3. clonal variability in surface-expressed proteins as described for Plasmodium falciparum31 and possibly Borrelia burgdorferi32. [5]

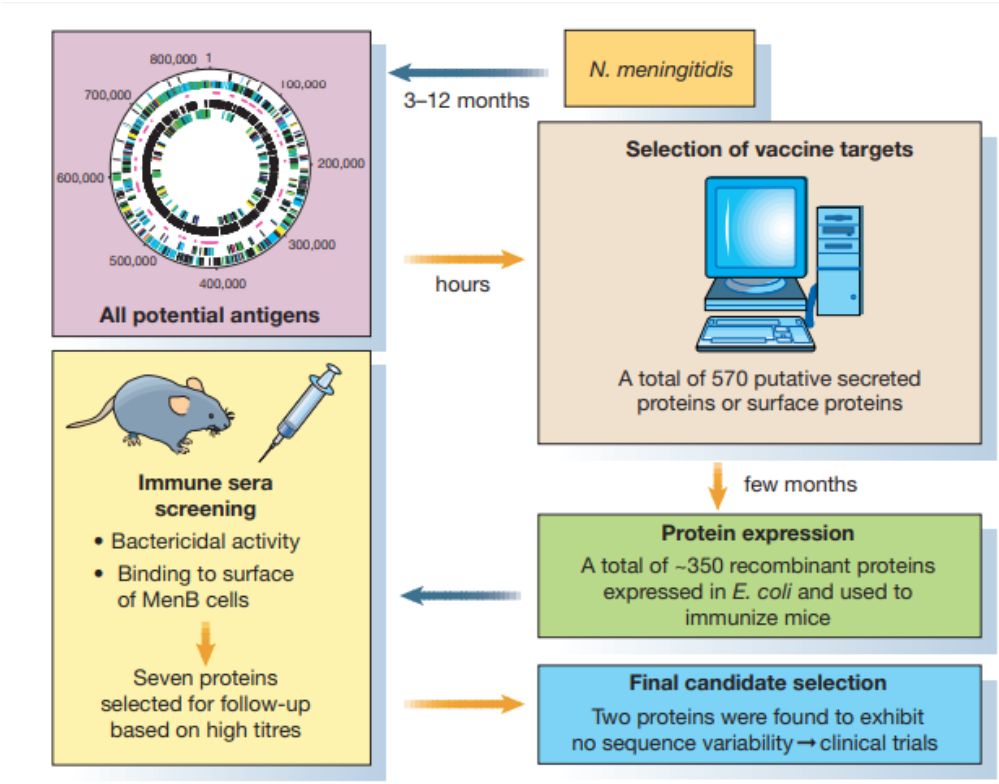

Figure 3.2: Diagram depicting how complete microbial genome sequence data can accelerate vaccine development.

4

Tools for Genomic Data Analysis

4.1 Basic Local Alignment Search Tool(BLAST)

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.(`https://blast.ncbi.nlm.nih.gov/Blast.cgi`) To overcome the limitations of the classic alignment algorithms, which are usually used for rather short sequences, new methods and heuristics were developed. By heuristic algorithm we imply a non-exact solution, which works faster than exact algorithms (sometimes much faster) and provides biologically meaningful results [1].

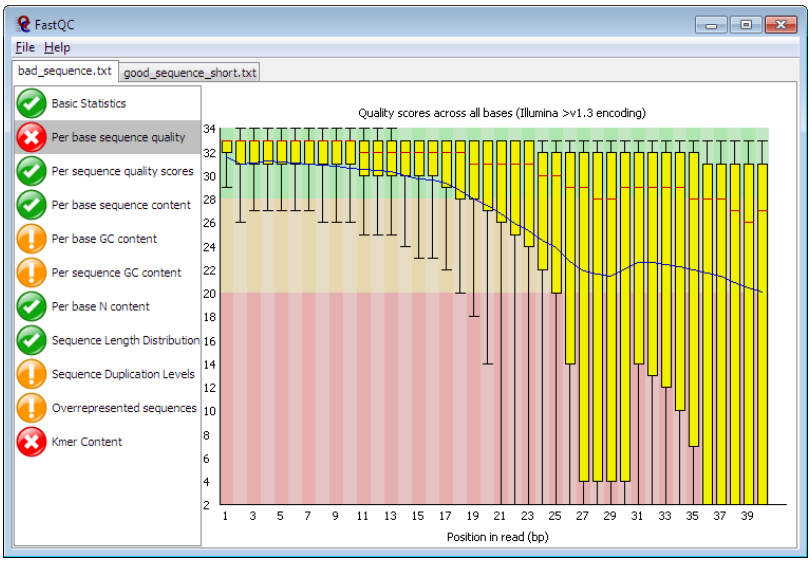

Figure 4.1: Example of FastQC Result

4.2 FastQC

FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis (Figure 6.1). (`https://www.bioinformatics.babraham.ac.uk/projects/fastqc/`)

The main functions of FastQC are:

1. Import of data from BAM, SAM or FastQ files (any variant)
2. Providing a quick overview to tell you in which areas there may be problems
3. Summary graphs and tables to quickly assess your data
4. Export of results to an HTML based permanent report
5. Offline operation to allow automated generation of reports without running the interactive application

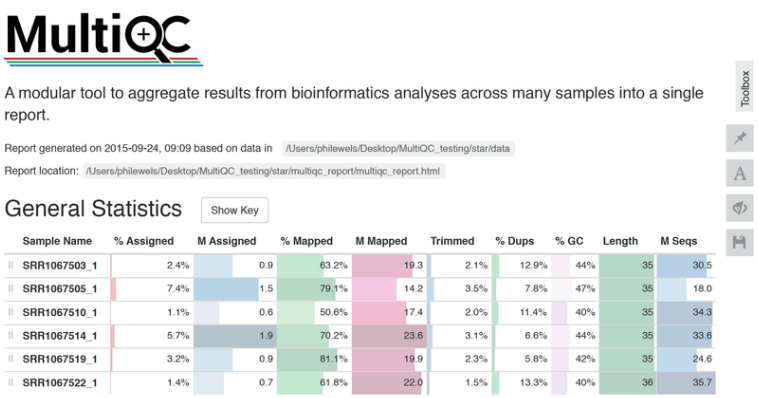

Figure 4.2: Example of MultiQC Report

4.3 MultiQC

MultiQC is a reporting tool that parses summary statistics from results and log files generated by other bioinformatics tools. MultiQC doesn't run other tools for you - it's designed to be placed at the end of analysis pipelines or to be run manually when you've finished running your tools. (Figure 6.1)

When you launch MultiQC, it recursively searches through any provided file paths and finds files that it recognises. It parses relevant information from these and generates a single stand-alone HTML report file. It also saves a directory of data files with all parsed data for further downstream use. (Figure 4.2) (`https://multiqc.info/`)

4.4 R Programming

The aim of computational genomics is to provide biological interpretation and insights from high dimensional genomics data. Generally speaking, it is similar to any other kind of data analysis endeavor but often times doing computational genomics will require domain specific knowledge and tools.

As new high-throughput experimental techniques are on the rise, data analysis capabilities are sought-after features for researchers. The aim of this chapter is to first familiarize the readers with data analysis steps and then provide basics of R programming within the context of genomic data analysis. R is a free statistical programming language that is popu-

lar among researchers and data miners to build software and analyze data. Although basic R programming tutorials are easily accessible, we are aiming to introduce the subject with the genomic context in the background. The examples and narrative will always be from real-life situations when you try to analyze genomic data with R. We believe tailoring material to the context of genomics makes a difference when learning this programming language for sake of analyzing genomic data. (`https://www.r-project.org/about.html`)

4.5 Python

Python is a general-purpose programming language conceived in 1989 by Dutch programmer Guido van Rossum.Python is free and open source, with development coordinated through the Python Software Foundation.

Python has experienced rapid adoption in the last decade and is now one of the most commonly used programming languages. Python is also used widely in the field of bioinformatics, computational genomics, genomics and computational drug development. `https://www.python.org/`

4.6 Biopython

Biopython is a set of freely available tools for biological computation written in Python by an international team of developers. (`https://biopython.org/`)

4.7 Sciki-bio

scikit-bio™ is an open-source, BSD-licensed, python package providing data structures, algorithms, and educational resources for bioinformatics. (`http://scikit-bio.org/`)

4.8 Bioconductor

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, and an active user community. (`https://www.bioconductor.org/`)

5

Genomic Data Analysis Methods

5.1 Steps of Genomic Data Analysis

Regardless of the analysis type, the data analysis has a common pattern. I will discuss this general pattern and how it applies to genomics problems. The data analysis steps typically include data collection, quality check and cleaning, processing, modeling, visualization and reporting. Although, one expects to go through these steps in a linear fashion, it is normal to go back and repeat the steps with different parameters or tools. In practice, data analysis requires going through the same steps over and over again in order to be able to do a combination of the following:

- answering other related questions,
- dealing with data quality issues that are later realized, and,) including new data sets

to the analysis.

We will now go through a brief explanation of the steps within the context of genomic data analysis[1].

5.1.1 Data collection

Data collection refers to any source, experiment or survey that provides data for the data analysis question you have. In genomics, data collection is done by high-throughput assays. One can also use publicly available data sets and specialized databases. How much data and what type of data you should collect depends on the question you are trying to answer and the technical and biological variability of the system you are studying.

5.1.2 Data quality check and cleaning

In general, data analysis almost always deals with imperfect data. It is common to have missing values or measurements that are noisy. Data quality check and cleaning aims to identify any data quality issue and clean it from the dataset.

High-throughput genomics data is produced by technologies that could embed technical biases into the data. If we were to give an example from sequencing, the sequenced reads do not have the same quality of bases called. Towards the ends of the reads, you could have bases that might be called incorrectly. Identifying those low quality bases and removing them will improve read mapping step[4].

5.1.3 Data processing

This step refers to processing the data to a format that is suitable for exploratory analysis and modeling. Often times, the data will not come in ready to analyze format. You may need to convert it to other formats by transforming data points (such as log transforming, normalizing etc), or subset the data set with some arbitrary or pre-defined condition. In terms of genomics, processing includes multiple steps. Following the sequencing analysis example above, processing will include aligning reads to the genome and quantification

over genes or regions of interest. This is simply counting how many reads are covering your regions of interest. This quantity can give you ideas about how much a gene is expressed if your experimental protocol was RNA sequencing. This can be followed by some normalization to aid the next step.

5.1.4 Exploratory data analysis and modeling

This phase usually takes in the processed or semi-processed data and applies machine-learning or statistical methods to explore the data. Typically, one needs to see relationship between variables measured, relationship between samples based on the variables measured. At this point, we might be looking to see if the samples group as expected by the experimental design, are there outliers or any other anomalies ? After this step you might want to do additional clean up or re-processing to deal with anomalies.

Another related step is modeling. This generally refers to modeling your variable of interest based on other variables you measured. In the context of genomics, it could be that you are trying to predict disease status of the patients from expression of genes you measured from their tissue samples. Then your variable of interest is the disease status and . This is generally called predictive modeling and could be solved with regression based or any other machine-learning methods. This kind of approach is generally called “predictive modeling

Statistical modeling would also be a part of this modeling step, this can cover predictive modeling as well where we use statistical methods such as linear regression. Other analyses such as hypothesis testing, where we have an expectation and we are trying to confirm that expectation is also related to statistical modeling. A good example of this in genomics is the differential gene expression analysis. This can be formulated as comparing two data sets, in this case expression values from condition A and condition B, with the expectation that condition A and condition B has similar expression values[4].

5.1.5 Visualization and reporting

Visualization is necessary for all the previous steps more or less. But in the final phase, we need final figures, tables and text that describes the outcome of your analysis. This will be your report. In genomics, we use common data visualization methods as well as specific visualization methods developed or popularized by genomic data analysis.

6

Analyzing Genomic Sequences

6.1 Sequence Alignment

6.1.1 Classic alignment algorithms

Sequence alignment is the process of comparing and detecting similarities between biological sequences. What "similarities" are being detected will depend on the goals of the particular alignment process. Sequence alignment appears to be extremely useful in a number of bioinformatics applications [1].

6.1.2 Comparative genomics

Nucleic sequence alignment algorithms are widely used in comparative genomics and phylogenetic studies. Comparative geno- mics studies similarities between two or more genomes at the level of large rearrangement events, such as inversions, duplications, translocations, large insertions and deletions. Since the comparative genomics usually does not take into account small structural variations and single nucleotide polymorphisms (SNPs), it requires a specific kind of alignment software. These methods typically detect synteny blocks long sequences shared between genomes being compared. Indeed, those sequences may have differences at the nucleotide level, but are still highly similar overall. The genomes are then represented as a sequence of synteny blocks and rearrangements are detected (Hannenhalli and Pevzner, 1999). Phylogenetic studies use various multiple sequence alignment methods to detect the level of sequence dissimilarity. The distance between compared sequences is used to construct phylogenetic trees, in which the length of the branches typically correspond to the distance between analyzed sequences. To construct a biologically meaningful and realistic tree, various clustering methods can be used as well as different sequences may be provided as input (Felsenstein, 1981; Kumar et al., 1994). Phylogenetic studies can be done using whole genomes and rearrangement events, genes and proteins sequences, or even SNPs for closely related organisms[1].

6.2 Preprocessing Sequencing Data

Regardless of what technology, protocol or sample was used to generate sequencing data, quality control remains an integral part of every experiment. When performed correctly during the early stages of a project, quality control helps save time and thus, money. There have been many cases when false conclusions were made due to the poor quality of initial data, and, as a known saying states, “garbage in – garbage out”, meaning that great results do not come from low-quality data. FastQC is one of the most popular tools for basic quality control of different kinds of sequencing data (Andrews, 2010). FastQC does not require any additional data, such as a reference genome, and the QC is performed based on

just a FASTQ file with sequences and corresponding quality values. Below we will take a look at several important statistics produced by FastQC, how to interpret them and what the differences between high- and low- quality data are.

(https://galaxyproject.github.io/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html)

6.3 Quality control of sequencing data

During sequencing, the nucleotide bases in a DNA or RNA sample (library) are determined by the sequencer. For each fragment in the library, a short sequence is generated, also called a read, which is simply a succession of nucleotides.

Modern sequencing technologies can generate a massive number of sequence reads in a single experiment. However, no sequencing technology is perfect, and each instrument will generate different types and amount of errors, such as incorrect nucleotides being called. These wrongly called bases are due to the technical limitations of each sequencing platform.

Therefore, it is necessary to understand, identify and exclude error-types that may impact the interpretation of downstream analysis. Sequence quality control is therefore an essential first step in your analysis. (https://galaxyproject.github.io/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html)

6.4 NGS QC Toolkit

The quality of data may be affected by several factors regardless of the NGS platform. Although the commercial vendors for all the sequencing platforms provide a quality control (QC) pipeline for filtering of sequencing output, several sequence artifacts still remain in the dataset. Therefore, it is advisable to perform QC and filtering of high-quality (HQ) sequencing data at the end-user level. For example, we rejected about 8% of the sequence reads obtained after filtering through QC pipelines of sequencing platforms, in our QC analysis of Illumina and Roche 454 data. A few online/standalone software pack-

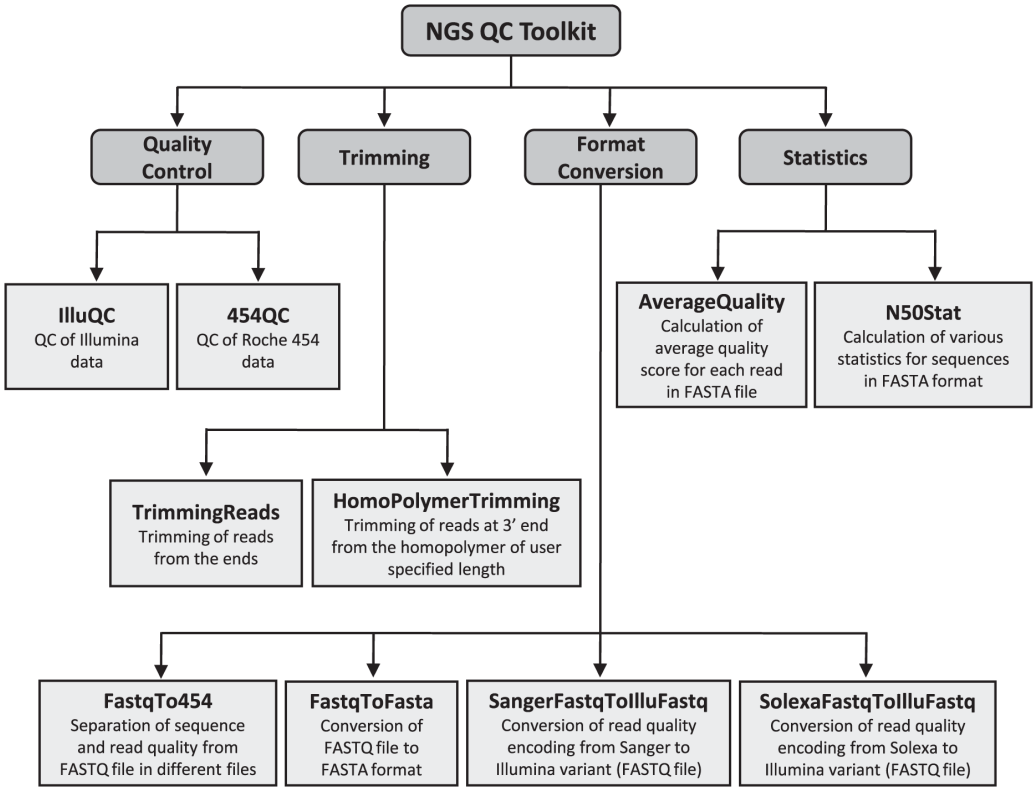

Figure 6.1: Flow chart showing various tools included in NGS QC Toolkit.

ages/pipelines with different features have been developed for QC of NGS data [5]–[9]. Many of these are specific for a particular sequencing platform and have one or the other limitation(s). Therefore, there is still a need for the development of better tools with additional/better features.

In this study, we have developed a NGS QC Toolkit, comprised of various easy-to-use standalone tools for quality check and filtering, trimming, generating statistics and conversion between different file formats/variants of NGS data from Illumina and Roche 454 platforms. The toolkit allows automatic and fast parallel processing of large amount of sequence data with user-friendly options. Given the importance of QC of NGS data, we anticipate that this toolkit will be very useful for the sequencing based biological research [9].

6.5 Ten steps to get started in Genome Assembly and Annotation

The advice here presented is based on a need seen while working in the ELIXIR-EXCELERATE task "Capacity Building in Genome Assembly and Annotation". In this capacity we have held courses and workshops in several European countries and have encountered many users in need of a document to support them when they plan and execute their projects. With these 10 steps we aim to fill this need[3].

In a de novo genome assembly and annotation project, the nucleotide sequence of a genome is first assembled, as completely as possible, and then annotated. The annotation process infers the structure and function of the assembled sequences. Protein-coding genes are often annotated first, but other features, such as non-coding RNAs or presence of regulatory or repetitive sequences, can also be annotated (Figure 6.2).

A checklist of things to keep in mind when starting a genome project:

- For the DNA extraction, select an individual which is a good representative of the species, and able to provide enough DNA.
- Extract more DNA than you think you need, or save tissue to use for DNA extraction later. If you need to produce more data later, it is critical to be able to use the same DNA to make sure the data assembles together
- Remember to extract RNA and order RNA-sequencing if you want to use assembled transcripts in your annotation (which is strongly recommended). If possible, extract RNA from the same individual as used in the DNA extraction to make sure that the RNA-seq reads will map well to your assembly.
- Decide early on which sequencing technology you will be using, and also consider which assembly tools you want to try. These two choices will greatly influence what kind of compute resources you will need, and you do not want to end in a situation where you have data that you cannot analyze anywhere. Plan compute resources accordingly.

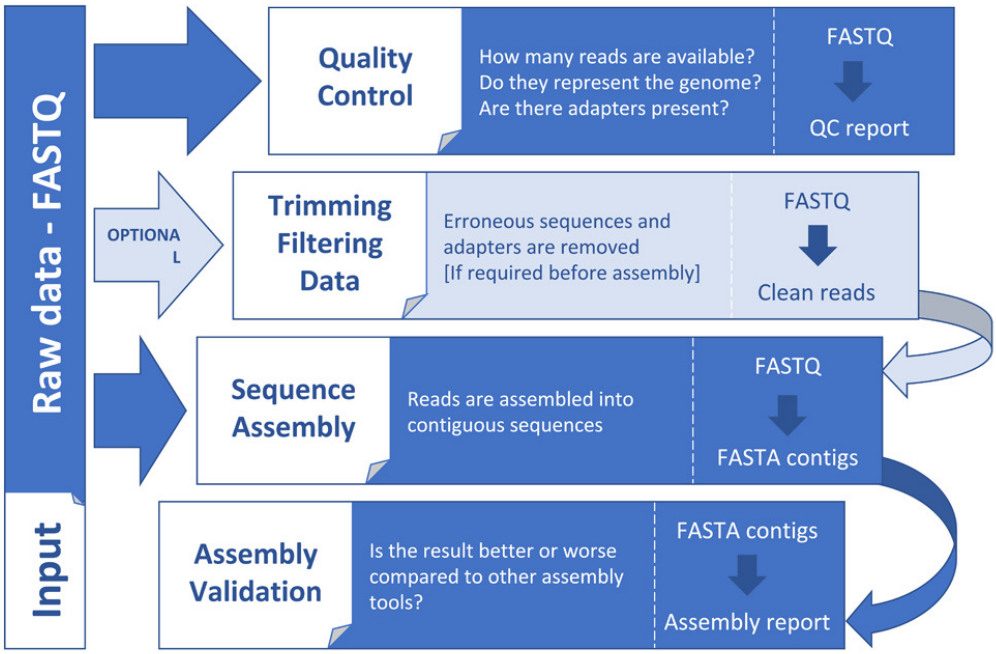

Figure 6.2: teps to get started in Genome Assembly and Annotation.

Refrences

[1] Anton I Korobeynikov Andrey D Prjibelski and Alla L Lapidus. *Sequence Analysis*. Elsevier, 2018.

[2] Nello Cristianini and Matthew W. Hahn. *Introduction to Computational Genomics*. 1st edition, 2006.

[3] Victoria Dominguez Del Angel, Erik Hjerde, Lieven Sterck, Salvadors Capella-Gutierrez, Cederic Notredame, Olga Vinnere Pettersson, Joelle Amselem, Laurent Bouri, Stephanie Bocs, Christophe Klopp, et al. Ten steps to get started in genome assembly and annotation. *F1000Research*, 7, 2018.

[4] David J Edwards and Kathryn E Holt. Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. *Microbial informatics and experimentation*, 3(1):2, 2013.

[5] Claire M Fraser, Jonathan A Eisen, and Steven L Salzberg. Microbial genome sequencing. *Nature*, 406(6797):799–803, 2000.

[6] Ana Gutiérrez-Preciado, Philippe Deschamps, Tom Delmont, Claudia Chica, Nathan AM Chrismas, and Ricardo Rodriguez de la Vega. Genome sequence databases: Types of data and bioinformatic tools, 2019.

[7] Neil C Jones, Pavel A Pevzner, and Pavel Pevzner. *An introduction to bioinformatics algorithms*. MIT press, 2004.

[8] William L Lowe and Timothy E Reddy. Genomic approaches for understanding the genetics of complex disease. *Genome research*, 25(10):1432–1441, 2015.

[9] Ravi K Patel and Mukesh Jain. Ngs qc toolkit: a toolkit for quality control of next generation sequencing data. *PloS one*, 7(2):e30619, 2012.

[10] Didier Raoult Pierre-Edouard Fournier. *Infectious Diseases*. 4th edition, 2017.

[11] Shoba Ranganathan, Kenta Nakai, and Christian Schonbach. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*. Elsevier, 2018.

Typesetting Credits

This report is created byJubayer Hossain using LaTeX. LaTeXis a high-quality typesetting system; it includes features designed for the production of technical and scientific documentation. LaTeXis the de facto standard for the communication and publication of scientific documents. I would like to mention some sources of learning LaTeXfor the rising writers.

1. https://www.latex-project.org/
2. https://www.overleaf.com/learn
3. https://www.overleaf.com/learn/latex/How_to_Write_a_Thesis_in_LaTeX
4. https://www.latextemplates.com/