Computational microbial genomics

Giacomo Fantoni
telegram: @GiacomoFantoni
Github: https://github.com/giacThePhantom/computationl-microbial-genomics

April 13, 2022

Contents

Chapter 1

Introduction

1.1 Microbes

Microbes are defined as whatever is not visible at the human eye: bacteria cells' dimensions are in the order of micrometre, while viruses in the order of nanometre. It is obvious how there is no visible part by eye. This is particularly true for viruses: their dimension make them almost impossible to perceive by any other method than genomics.

1.1.1 Prevalence of microbes

We are living in a microbial world: more than 90% of the biomass is composed of them and they are responsible for a great part of the biochemical cycle. Microbes can thrive in a variety of environment and according to some estimates they compose $10^{17}g$ of biomass. To put that in context the overall weight of the human species is three or four order of magnitude less. They also form the human microbiome, with important medical implication.

1.1.2 Difficulties in studying them

Of the predicted 30 million species to exist only thousands can be cultured in isolation in the lab. There is a need to create a way to directly study and characterized samples taken from the environment.

1.2 Genomics

Once the genetic material is isolated and sequenced a huge amount of information needs to be interpreted.

1.2.1 History of sequencing

- The first sequenced gene was one of a bacteriophage.
- Also the first complete genome wase one of a bacteriophage and is used by ILLUMINA as a control.
- The first bacterial genome was pub-

lished in 1995 and was that of Haemophilus influenzae. It has a dimension of $1.8Mb$ and sequencing took a year to complete.

- The first archea was sequenced in 1996.
- In 1996 the genome of S. cerevisiae has been sequenced and it was noticed that the genome shows a considerable amount of genetic redundancy.

After having sequenced the genomes there is a need to elucidate the biological functions of the genes contained in them.

1.2.2 Comparative genomics

Studying two different strains of the same organism allows to link difference in the genome to difference in phenotype.

1.2.3 Metagenomics

Metagenomics is the study of the DNA from all the genomes in an environment. By sampling all of the DNA from a given environment, it is possible to study the presence of bacterial ecosystems, independent of the ability to culture each bacterial strain in the laboratory. Large evolutionary radiation of bacterial lineages whose members are mostly uncultivated and only known through metagenomics and single cell sequencing have been described as nanobacterial. They have small genomes and lack several biosynthetic pathways and ribosomal proteins.

1.3 Leveraging computational power

Despite the advantages in DNA sequencing technology the sequencing of genomes has not progressed beyond clones on the order of the size of λ because of the lack of sufficient computational approaches that would enable the efficient assembly of a large number of independent, random sequences into a single assembly. When moving from low-throughput to high-throughput biology statistical power is needed: the genome of a bacterium must comprise all the DNA coding molecules present in the cell. With millions of reads from NGS of an environmental sample, it is possible to get a complete overview of any complex micro biome with thousands of species.

1.3.1 Comparing low-throughput and high-throughput pipeline

Let's consider the pipeline to find the pathogenic agent for a novel outbreak. In a low-throughput one a panel of reasonable putative causative agents is identified. Then one-by-one cultivation protocols to grow the agents from the infected tissue are performed. The pipeline ends when a pathogen grows in a sample. This is very time consuming. High-throughput instead sequences the full DNA repertoire of the sample and try to identify the pathogen by its genomic signature.

Chapter 2

Escherichia Coli general informations

2.1 *E. coli* genomics

Escherichia Coli is a Gram-negative, facultative anaerobic, rodshaped, coliform bacterium, it pertains to the phylum of proteobacteria and to the family of Enterobacteriaceae. It can be grown easily and inexpensively. Its genome has a length between 4.5-$4.7Mb$, including about 4000-5000 genes, and about seven ribosomal RNA operons. Only the 38% of the genes of K-12 (one of the most studied bacterial strains of *E. coli*) were experimentally identified, overall 40-50% of the genes are to date without a known function. The original *E. coli* strain K-12 was obtained from a stool sample of a diphtheria patient in Palo Alto, California in 1922.

2.1.1 *E. coli* long-term evolution experiment

The *E. coli* long-term evolution experiment led by Richard Lenski is one of the longest evolutionary experiments ever made. Starting from the 24th of Frbruary 1988 12 population of E. coli have been cultivated in parallel. After each day a portion of the population was introduced in a new flask, where it proliferated. Every 500 generations a sample from each flask is saved, so to track the evolutionary changes made. Today the $66000th$ generation have been reached. Long-term adaptation to a fixed environment can be characterized by a rich and dynamic set of population genetic processes, instead of the evolutionary desert expected near a fitness optima. In particular some bacteria developed the capacity to aerobically grow on citrate.

2.1.2 *E. coli* strains

E. coli could be found as commensal strains, pathogenic strains, or environmental strains. The pathogenic strains could pertain to these categories (which are not exclusive):

- enteropathogenic (EPEC),
- enteroinvasive (EIEC),
- enterotoxigenic (ETEC),
- diffusely adherent (DAEC),
- adherent invasive (AIEC),
- shiga-toxin producing (STEC),
- enteroaggregative(EAEC),
- extraintestinal pathogenic (ExPEC).

Resistances to antibiotics adds another layer of complexity in the categorization categorization of *E. coli*. Most of the genes are on plasmids, circular, additional to chromosome, and can be moved easily horizontally. Plasmids between different strains can be moved in enterobacteriacae, this doesn't happen normally in other families. Some *E. coli* strains are even capable of causing cancer in humans: for example, colibactin-positive *E. coli* can cause colon and rectal cancer, by creating mutations which are responsible of the of the cancer onset.

2.1.3 Stc-EAEC outbreak

In 2011 in Germany there was an outbreak of Stx-EAEC, An efficient counter-measure was found by sequencing the genome of those bacteria.

2.1.4 Shigella

Shigella is a strain of *E. coli* capable of producing the shiga toxin. It has been difficult to categorize for taxonomists. Several antigens can be used by taxonomists to categorize *E. coli* strains. In particular the O (171), H (56), K (80) antigens, respectively related to the somatic, the flagella and the capsule are widely used.

2.1.5 PanPhlAn - strain detection and characterization

Pangenome-based Phylogenomic Analysis (PanPhlAn) is a strain-level metagenomic profiling tool for identifying the gene composition and in-vivo transcriptional activity of individual strains in metagenomic samples. PanPhlAn's ability for strain-tracking and functional analysis of unknown pathogens makes it an efficient tool for culture-free infectious outbreak epidemiology and microbial population studies. This tool was for example used to study the strain responsible of an outbreak in Germany in 2011 and found a strain of shiga-toxigenic Escherichia coli (STEC). This method, due to its greater efficiency and accuracy has been used since instead of low-throughput, traditional pipelines.

2.2 Genomes of *E. coli*

2.2.1 Core and accessory genome

In the genome of *E. coli* strains, it is possible to distinguish between:

- Core genome: the set of all genes shared by all members of a bacterial species, it includes 1000 up to 3000 genes.
- Accessory or dispensable genome: the

set of genes present in some but not all genomes within the same bacterial species. It is found on a single strain or in a subset of strains.

2.2.2 Pangenome

The pangenome is the union of the core genome and the accessory genome. It is the set of all the set of all the genes that can be found in the species strains. The pangenome can be characterized regarding its size with respect to the number of genomes:

- Closed: the pangenome size tends to a maximum as number of genomes increases.
- Open: the pangenome size keeps increasing as you add new genomes.

Typically sequencing more organisms of the same species tends to lower the amount of genes in the core genome and increase the number of those in the pangenome, 2.1. Due to technical errors, the core genome should tend to a size of 0, but a more reasonable plateau can be predicted with a more accurate mathematical formulation.

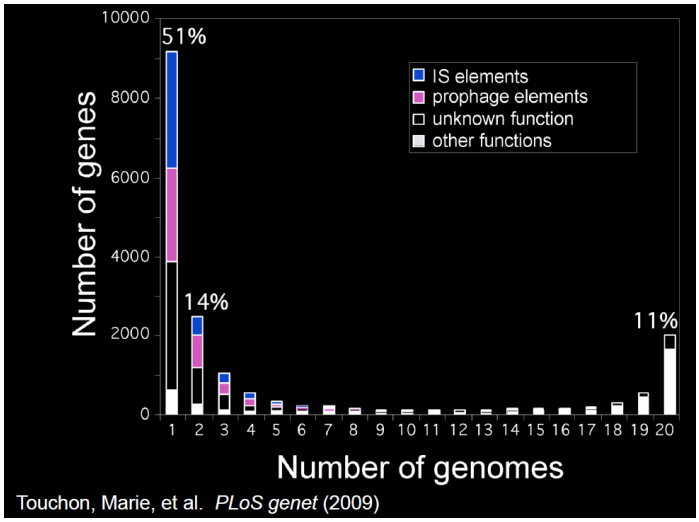


Figure 2.1: It can be seen that 51% of the genes are strain specific, and the other are shared between 2 to 20 strains of *E. coli*

Each *E. Coli* genome contains a balance genes of the core genome and of the pan genome, for a total amount of 4700 genes 2.2). Core genomes' genes are responsible of some basic cellular functionalities and utilities to survive the environment, while instead elements of the pangenome are quite usually specific to a single strain, like for example antibiotic resistance, and they are often not functionally well characterized.

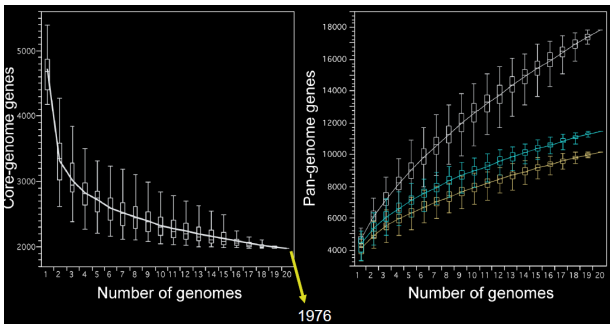


Figure 2.2: Balance between genes of the core- and of the pan-genome

Chapter 3

Next generation sequencing

3.1 Introduction

Next generation sequencing is the gold standard for sequencing nowadays. A series of discoveries allowed for the development of this technology:

- 1959: first homogeneous DNA purified.
- 1970: first discovery of type *II* restriction enzymes.
- 1972: first RNA gene sequence published.
- 1975: Sanger publishes his plus-minus method of sequencing, unable to distinguish homopolymers.
- 1977: Maxam and Gilbert publish their method that could distinguish homopolymers.
- 1977: Sanger publishes the dideoxy-sequencing method.

3.1.1 Progresses of sequencing

As can be seen in the graph 3.1 the cost of DNA sequencing is decreasing by a greater rate than the one predicted by Moore's law. This allows for greater number of samples and the sequencing of a different number of genomes.

3.1.2 Methods of sequencing

The methods of sequencing can be grouped in three groups:

- Chemical degradation of DNA: like the method of Maxam-Gilbert.
- Sequencing by synthesis ("SBS"): the most common approach and the first to be developed. It uses DNA polymerases in primer extension reactions. This technology is used by Illumina, Pacific Bioscences, Ion Torren and 454.
- Ligation-based: sequencing using short probes that hybridize to the template. This technology is used by SOLiD and Complete Genomics.
- Others like nanopores.

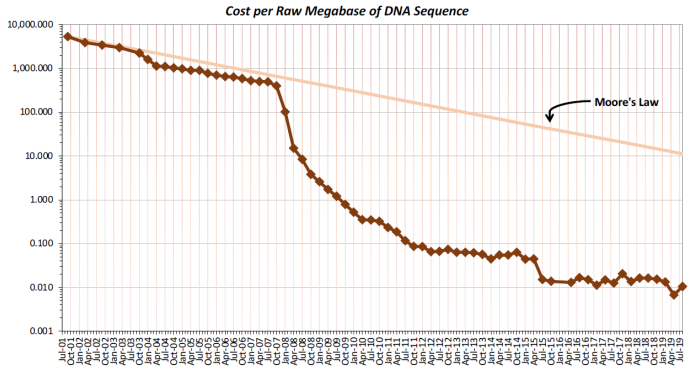
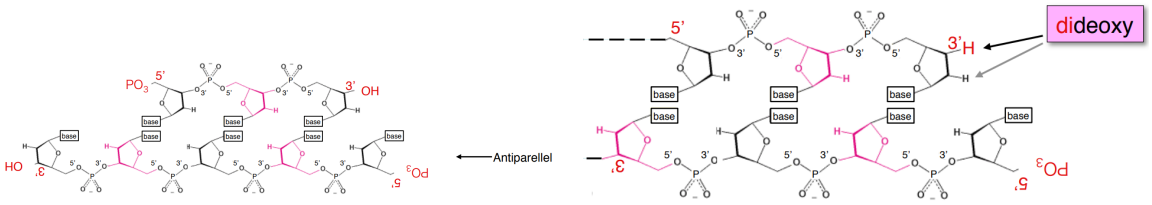


Figure 3.1

3.1.3 The Chain Terminators



(a) Normal DNA molecule, with oxydrilic group ligated to 3′-ends

(b) Figure representing the difference between a normal DNA chain and one with chain terminators

Figure 3.2: Normal DNA synthesys *vs* Chain terminators

Normally, the addition of new nucleotides to a generated molecule of DNA happens with the 3′-end of the nucleotide chain 3.2a. Chain terminators are dideoxy nucleotides, ddNTPs, that cannot be further extended. These nucleotides don't have the oxydrilic group at their 3′-end, so the DNA polymerase cannot further add other nucleotides to the chain 3.2b.

3.2 Sanger method

The first method ever used to sequence DNA was designed by Frederick Sanger. The Sanger manual sequencing system consists in an *in vitro* process described in figure 3.3. It is also named a primer extension method. It is performed over a single-filament DNA sample, and it uses chain terminators nucleotides, one for each type of nucleobase: ddATP, ddGTP, ddTTP, ddCTP. The reaction is done inside four different reactions tubes, each containing:

- The sample DNA to be reproduced.
- A DNA polymerase.
- The normal nucleotides.
- One of the four possible chain terminator marked with sulfur-35. In each tube, the corresponding dideoxynucleotide was used with a concentration 10 times lower than the other normal nucleotides.

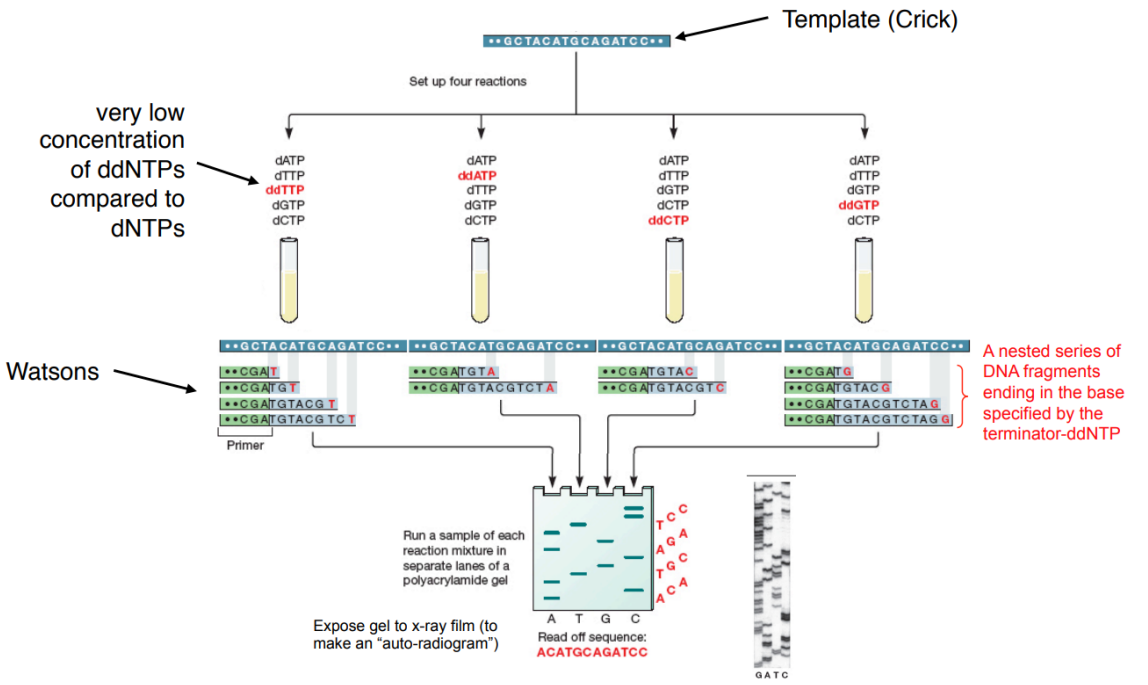


Figure 3.3: Sanger's method process

The polymerization reactions produces several molecules of DNA with different length: each replicative cycle is terminated after the addition of a chain terminator nucleotide. The initial DNA sequence is reconstructed using a long PAGE gel with high concentration of urea $(6-7M)$ to avoid the coiling of the DNA single-filaments. High voltages are required to achieve a highly risolutive run. This high resolution was needed as DNA's fragments are different only for a nucleotide. After having run it the bands were visualized through auto-radiography in order to evidentiate the phosphorescent signals.

The sequence is read starting from the shortest fragments at the end of the gel and going up along it, looking for the first presence of a band in one of the four runs.

3.2.1 Automatic sequencing

To automatize the Sanger methods fluorescent proteins substituted the radioactive signal. Several versions were developed:

1. Fluorescent primers marked with a single fluorochrome.
2. Four aliquotes of the same primer were used marked with four different fluorochromes, able to emit different fluorescences.
3. Four different fluorochromes were used to mark the single ddNTPs

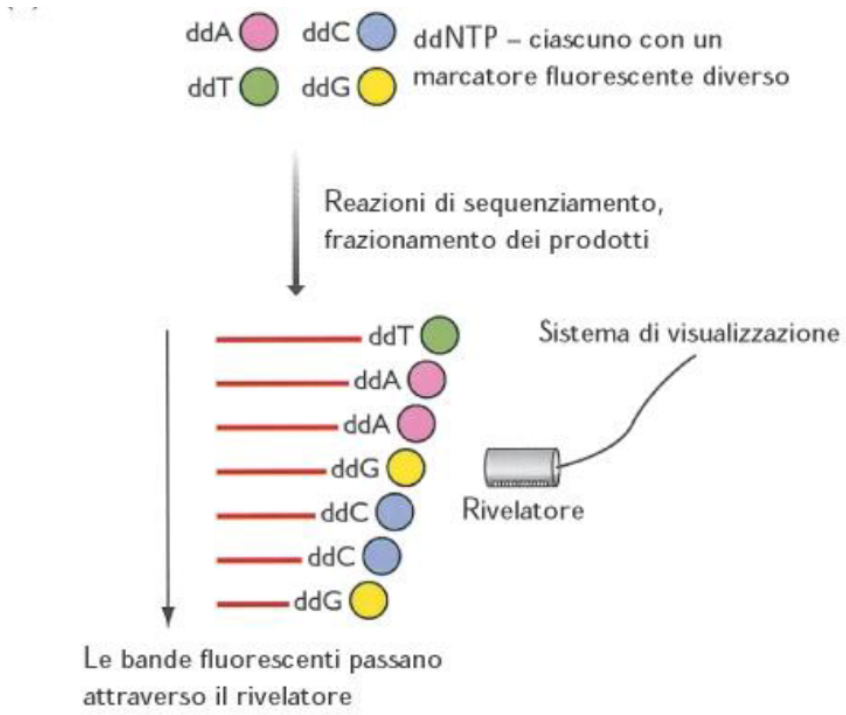
Thanks to the use of 4 different fluorochromes, it was possible to use a single electrophoretic lane to carry the sequencing reaction. Also a cyclic replicative reaction was

performed with this procedure using a thermal cycler:

1. Denaturation at 95°C of the DNA to be sequences
2. Annealing at $50 - 70$°C of the primer specific to one of the two filaments
3. Extension at 72°C by using a *Taq*-polymerase to avoid the formation of coiled structures in the DNA molecule to be sequenced.

The resulting molecules are run through a long PAGE gel and fluorescence is triggered irradiating the DNA molecules.

Figure 3.4



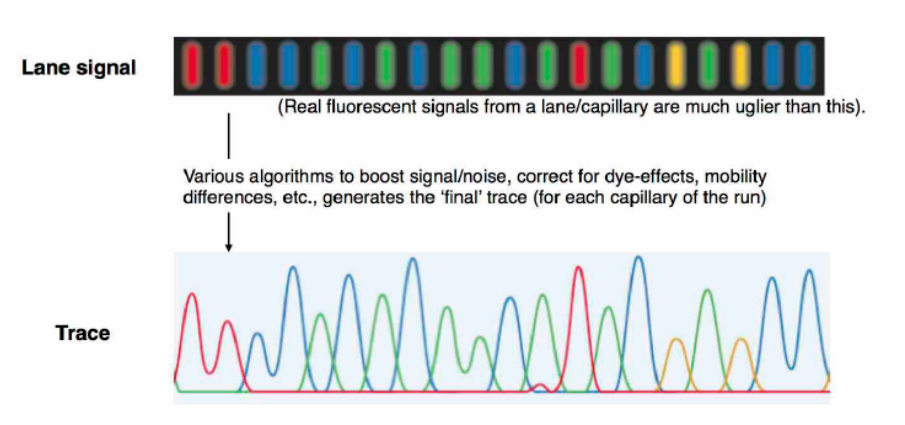
More usefully, this sequencing method is performed by using a capillar filled with a synthetic polymer with the same function of polyacrillamide. The analysis produces an electropherogram, with a color depicting the probability of each base being in each position. The electropherogram is refined through algorithms that can boost the signal to noise ration, correct the dye effect and reduce all systematic errors.

The Sanger automatic sequencing method was used extensively for the majority of the human project.

3.3 Development of Sequencing Machines

The method used for sequencing has to be chosen based on the wanted output, like the quality required and the type of input. Different technologies have different strengths and

Figure 3.5

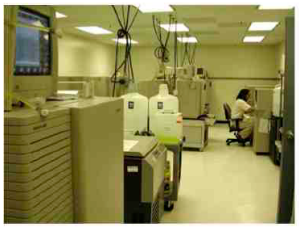


weaknesses:

- SOLID sequences reads long only 35-75 bases and it is not used anymore.
- Sanger sequencing, or the capillary, can read up to 1000 bases, but has a low throughput.
- MINION allowed to sequence an entire genome of *E. coli*.

A plethora of sequencing machines are available today. None of them is able to sequence DNA directly from a sample, requiring different preparation step. Today machines producing high-throughput output of short reads are preferred.

Radioactive
polyacrylamide
slab gel
Low throughput,
labor intensive



AB slab gel sequencers
(370, 373, 377)
Fluorescent sequencing
1990-1999
6 runs/day
96 reads/run
500 bp/read
288,000 bp/day

AB capillary sequencers
(3700, 3730)
1998-now
24 runs/day
96 reads/run
550 – 1,000 bp/read
1-2 million bp/day

Figure 3.6: The implementation of capillary sequencing machines gave the possibility to make more runs than with the others. ~ 1000 fold productivity increase was allowed

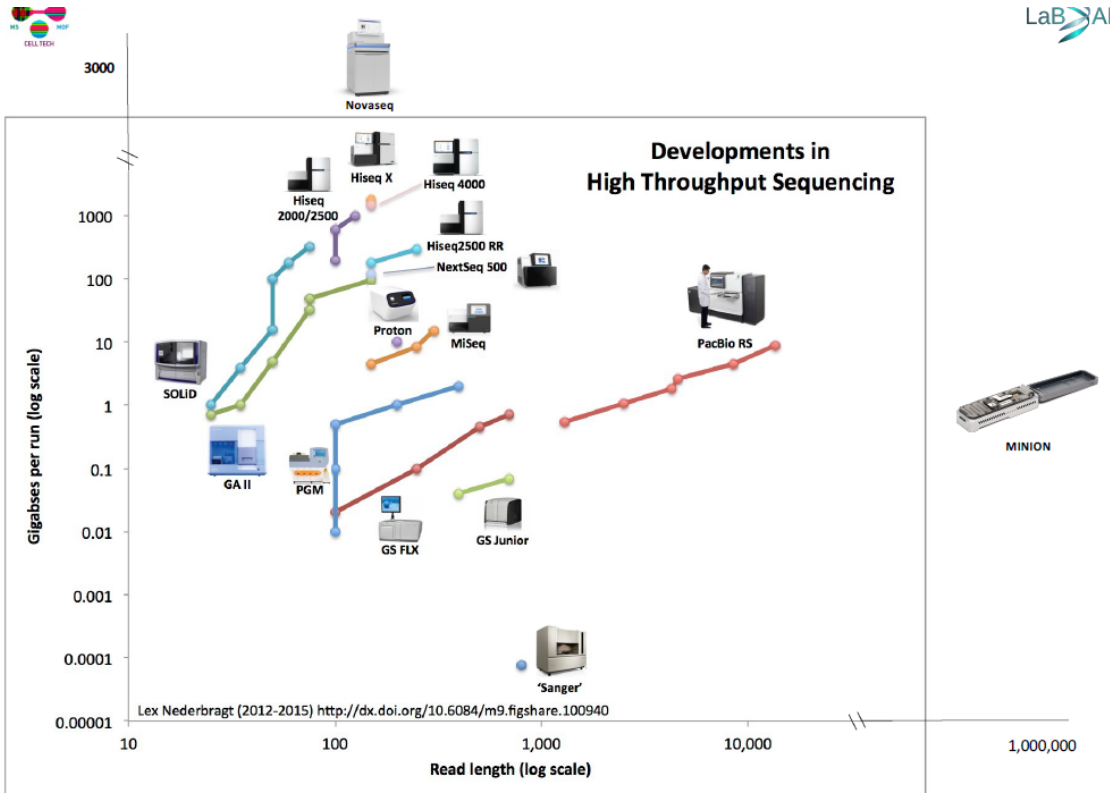


Figure 3.7: It can be noticed how recent developments had the scope of increasing the output data

The most wide-spread machines are from ILLUMINA like NovaSeq. They use sequencing

by synthesis and they amplify the signal through the use of clusters.

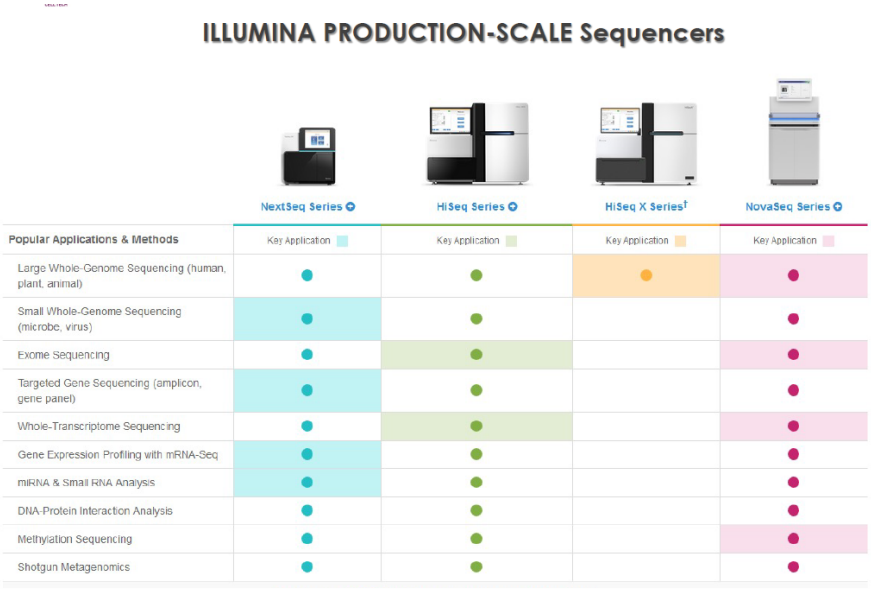


Figure 3.8

3.4 Next Generation Sequencing

The NGS protocol requires 3 steps:

1. Sample preparation: series of fragments added.
2. Clonal amplification: needed to replicate fragments attached to the solid surfaces, since machines are not sensible to single molecules.
3. Sequencing: ILLUMINA sequencing is one of the techniques used to obtain sequence data nowadays.

3^{rd} generation allow to read a molecule without replicating it.

3.4.1 ILLUMINA sequencing

3.4.1.1 Fragments and Library preparation

During fragmentation the DNA or RNA polymers that need to be sequenced are fragmented in short read sequences. This is because the machines are able to sequence only read with a length of a few hundred nucleotides. These fragments need to be tagmented: in this process one or two indexes, or barcodes are added to the fragment so that it has two sequencing primer binding sites and regions complementary to the oligonucleotides in the chamber. Indexes allow to distinguish between sample and, in the case of pair end sequencing, allow to distinguish between the forward and reverse read.

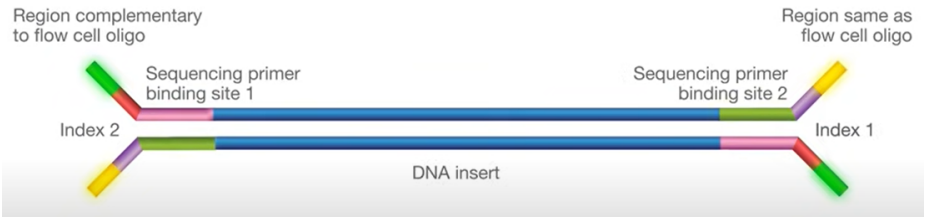


Figure 3.9: Figure representing the a good prepeared fragment, it has two indexes, two sequencing primer binding sites and regions complementary to the oligonucleotides present in the chamber

3.4.1.2 Clonal amplification and ILLUMINA sequencing procedure

Figure 3.10

Clonal amplification is necessary to amplify the signal from each fragment. ILLUMINA machines make use of clusters to sequence DNA: group of DNA strand positioned near each other that generate from a single fragment. Each cluter represents thousands of copies of the same DNA strand in a 1-2μm spot 3.10.

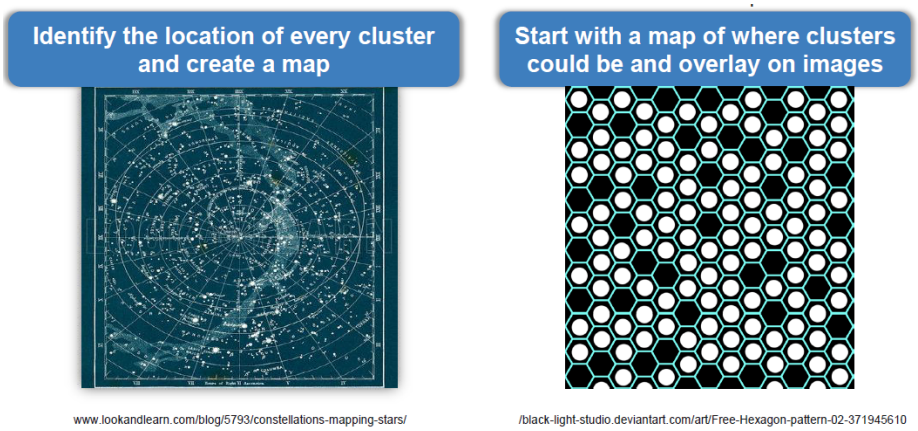


Figure 3.11

The clonal amplification process happens in flow cells, slides of glass in which fragments flow over channels. Changes in temperature allow for ligations and separations. The surface of the cells is functionalized with a series of oligos complementary to library adapters. These flow cell can be patterned (cluster in specific positions) or random (clus-

ter randomly positioned). In the former the location of the cluster can be known (rigid registration 3.11) or unknown.

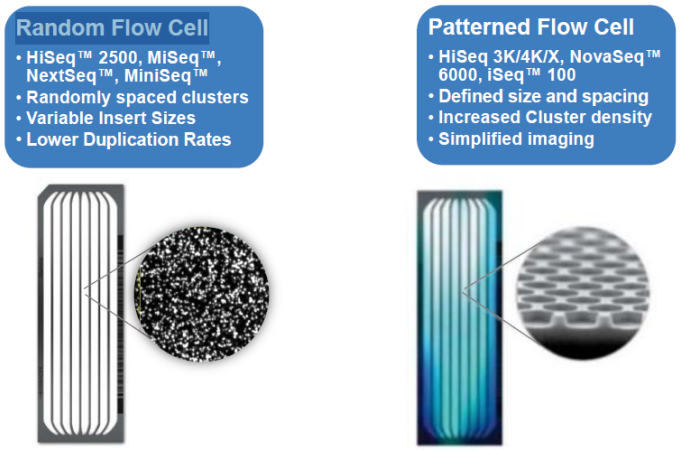
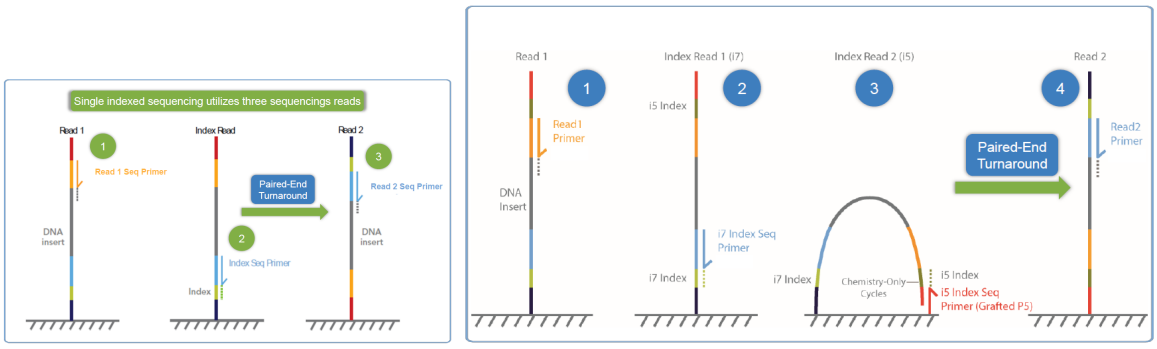


Figure 3.12

3.4.1.3 Fragment attachment to the clusters

Once the fragments start flowing over the chambers, they can bind only to one of the two oligos functionalizing the plate. Once they are attached to the surface the sequencing process is controlled through solvents and temperature. As shown in 3.13 the sequencing process differs between single end and paired end sequencing.

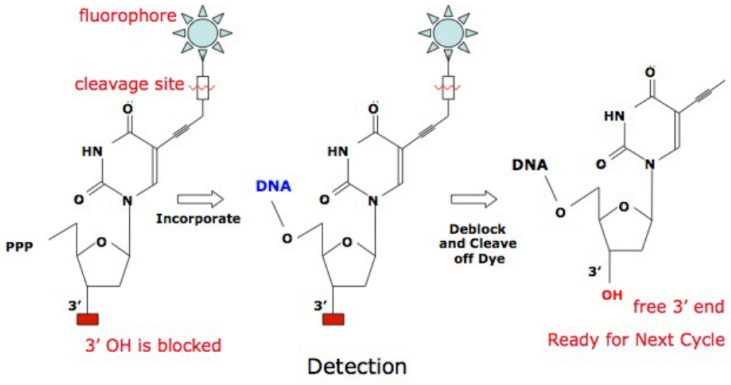


(a) single index **(b)** double index

Figure 3.13: Single/double index for ILLUMINA sequencing

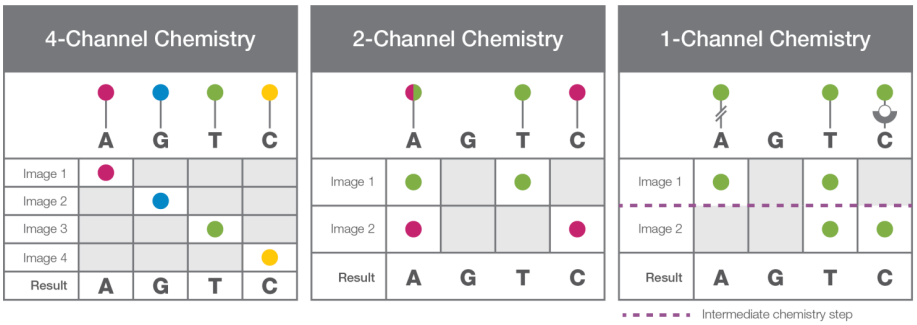
3.4.1.4 Sequencing

Figure 3.14



Reversible terminators 3.14 allow a real time analysis of the sequencing through the syntheses reaction. The fluorophore part of the terminator can be cleaved to eliminate the signal.

Figure 3.15



Depending on the number of fluorescent molecules used ILLUMINA sequencers are distinguished in the 4-channel, 2-channel or 1-channel type. In the case of the 4-channel 4 images are taken in each cycle and each cluster apperas in only one of the four images 3.15. The highest intensity base in a cluster is the called base for that cluster. In case no base is clearly related to a position the base calling returns N. This reading process can be done through single end reads on a single extreme of the fragment or through paired-ends reading, where each fragment is read in a forward and reverse way. This latter method gives structural information.

3.4.2 Pacific Bioscience

In the Pacific Bioscience PacBio sequencer the long DNA filament to be sequenced is attached to a polymerase, over the surface of a SMRT (Single Molecule Real Time) cell. This

cell is really small, and at each nucleation process a light signal is emitted. The produced light is not able to get out of the walls, and its duration is extremely restricted. The registration of the light signal correspond to a base calling. Its main advantage over ILLUMINA is the possibility of sequencing really long DNA molecules.

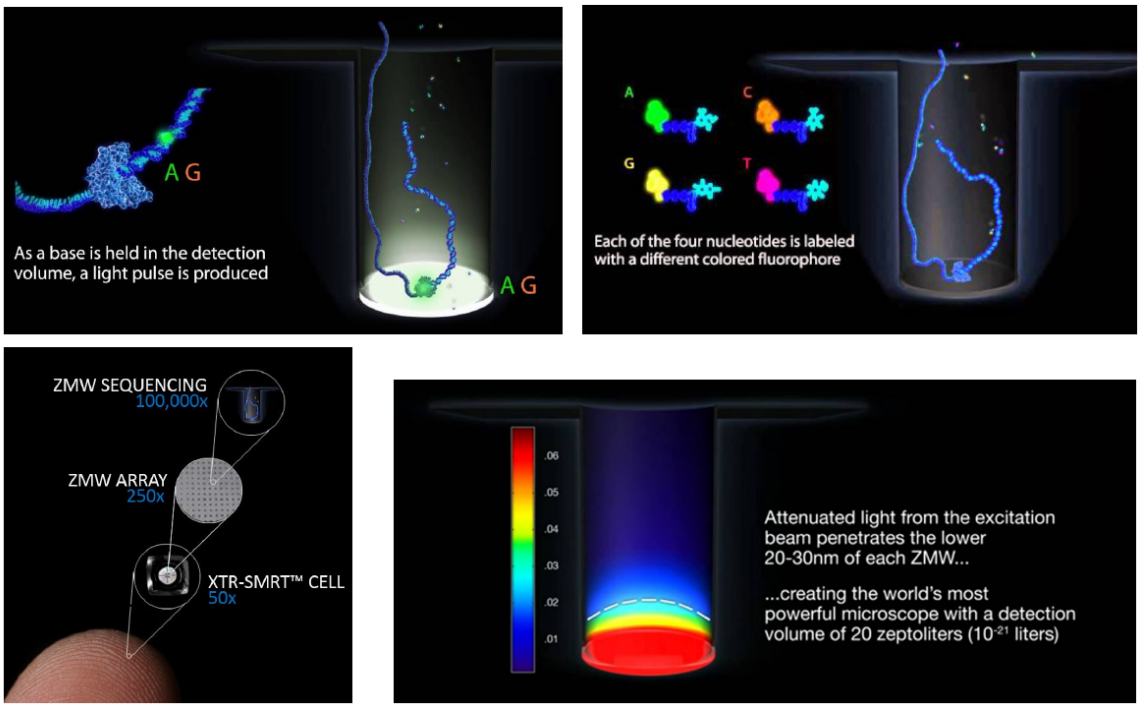


Figure 3.16

3.4.3 Nanopore sequencing

In nanopore sequencing the sequence is detected through the passage of DNA molecule into intramembrane protein. This produces a voltage changes that corresponds to a base calling. This technology is in constant development and is not widely used because of the, although improving, high error rate.

Chapter 4

Sequencing data

4.1 Choosing the optimal technology

When performing a genetics or genomics study it is best to be as hypothesis driven as possible and use already available data to guide the new analysis. Moreover to choose the optimal sequencer the parameters that need to be considered are:

- Throughput.
- Cost.
- Read lengths.
- Data output (reads per run).
- Coverage.
- Sequencing errors (indel, substitution, CG deletion, AT bias).
- Library preparation compatibility.
- Speed (run time).

4.1.1 Comparing different sequencing technologies

Illumina NovaSeq : optimal for sequencing a lot of DNA molecules at the same time like in the case of genomes or metagenomes. It can't go over $300bp$ readlines run, but it has the highest throughput so far. It is capable of multiplexing, differentiating the different samples with an unique barcode.

Illumina iSeq : optimal for sequencing shorter genomes.

NanoPore (minion) : it is a pocket-sized wet-lab free sequencer for DNA, RNA and (possibly) proteins, but the read lengths is smaller than Illumina's. The machine is cheap;, but the running flow is more expensive over time. It's a real-time sequencer.

PacBio : has very long reads but carries a high amount of error.

A solution to reduce the impact of error or weaknesses of one of the sequencers is to use more than one for the same project. Consider there is a need to sequence the genome of a bacterium: PachBio would give a lot of sequence errors, while Illumina wuild be unable to reconstruct the sample due to assembly ambiguities. In the end PacBio will construct the genome and Illumina will correct sequence errors. This solution doesn't work in complex

sample with more than one genome, because there is no way to a priori which reads are coming from one organism. Another widely adopted solution is to sequence multiple times one molecule so to reduce random errors. This is effective but does not resolve systematic error like the one of PacBio with homopolymers.

4.1.2 Sequencers' output

All sequencing platforms translate the physical read signal into files in FASTQ format. This files contain the sequencing reads and the quality of each base.

4.2 Base callers

A base caller is an algorithm that translates the analogical signal of the reading into numbers and nucleotides. The most popular algorithm is Phred. Phred tries to correct errors derived from the sequencing reaction and electrophoresis. It was tested on a huge dataset of gold standard sequences (finished human and C. elegans sequences generated by highly-redundant sequencing). Its results were compared with the traditional ABI base caller and Phred was considerably more accurate with 40-50% fewer errors. This algorithm need to be able to understand when it is impossible to recover an high quality sequencing and so it needs to be able to give up for low-quality reads. The confidence that the base caller has to call a certain nucleotide ATCG is annotated in the FASTQ file, allowing for quality control downstream.

4.2.1 Errors solved by Ilumina's base caller

Phasing noise ϕ : when a certain base is not seen frequently, the first time in which it will reappear there will be a spike in the graph, increasing the signal of the nearby bases. This will cause errors in estimating the real nucleotide that is occurring in the site. This problem can be solved by waiting more time between readings, but the sequencing will be less efficient and the throughput will be lower. The machine needs to find a trade-off between efficiency and clear reading.

Signal decay δ : after a while the sequencer has read the same base, or the same repeated couple of bases, the signal will go down and at some point will be indistinguishable. At some point you will need to cut the read since it will be not trustable after a while.

Mixed cluster μ : two different fragments can enter the same cluster and the sequencer will read the two signals simultaneously.

Boundary effects ω : in this case the machine needs to interpret an image and it cannot distinguish between the signal and the background.

Cross-talk $\mathcal{X}$

Fluophore accumulation τ

4.2.2 Density on the flow cell

The density of the flow cell is the number of cluster in it. Under clustering will reduce the sequencer throughput, while over clustering will cause errors due to the limited resolution

of the reading of the bases. There is a need to find the optimal number of cluster that maximises the throughput without introducing overlapping in the reading of clusters.

4.2.3 An ecology of base callers

Base callers need to find a satisfying trade-off between accuracy and computational efficiency. The quality is the estimation of the probability that the nucleotide in a certain position is correct. The PHRED score reported by the base caller and the PHRED score of mapping to a reference sequence are different. Base callers need to be calibrated with a standard to make sure that the estimation of the quality is accurate enough.

4.3 FASTQ format

4.3.1 Composition

The FASTQ format is composed by:

1. '@' followed by a sequence identifier. This identifier contains the unique instrument name, the flowcell and tile number, the x and y coordinates of the cluster within the tile, the index number for multiplexing and the pair number of paired-end sequencing. In Illumina MiSeq, each flowcell has 8 microfluidic channels (lanes), each lane contains three columns with 96 tiles, and can sequence up to 96 multiplexed samples.
2. The sequence. It could be mate paired for paired end sequencing.
3. '+', optionally followed by a sequence Identifier.
4. The quality scores. Quality is a number based on the estimated probability of error. p=probability of error, $Q = -10\dot{p}$. A base quality of at least 20 is needed to reach 1% of error. A base quality of 40 means the probability of error is 0.01%. The FASTQ quality score is the phred score +33, converted in CHAR code.

The FASTA file format is a FASTQ fomrat without the quality score reported and the seq ID is preceded by '>'.

4.3.2 Quality control: read length distribution

Quality scores are typically used to perform quality control and cleaning of the reads. This result in a FASTA output file to be used downstream. However there are algorithms that can use directly FASTQ files, performing autonomously the cleaning and quality control of the reads. The quality score is an indication of how well the sequencing run went. Typically the quality decreases when the read length increases due to the fact that sequencers have problem when are run in continuum. A solution to this problem is to cut the reads when the quality becomes too low. Another problem happens when the adapter is included into the read. This happens typically with short reads and a solution is to cut it and a part of the sequence. FastQC can be used to plot the quality distribution of the data. Another way to asses read quality is to consider the average quality of the entire read, discarding the low-quality ones.

4.3.3 Duplication artifacts

It is not frequent to see duplication, but it can be a problem especially when there is a need to quantify gene expression or copy number of genes in a bacterial genome. The distribution of the duplicates should be the same of the distribution of the reads. There should't be any bias along the length of the reads, but if there is it should be due to a repetition of sequencing of the same read or when the adapter and primer have been read.

4.3.4 GC content analysis

Each organism has a signature GC content, so when plotting it a normal distribution is expected. Multiple peaks are an indicator of the presence of two different organisms.

4.3.5 K-mers frequency plot

K-mer frequencies are a way to catch systematic sequencing error: when mapping a genome it is now the relative frequency of each K-mer. That can be compared with the K-mers frequencies to catch systematic errors in sequencing and to assess its quality. Frequent k-mers can be a signature, as the GC content. The expected coverage of a k-mer with reads of length L:

$$L_{cov} = \frac{L - k + 1}{L} \times Cov$$

Then, given a k-mer, it can be seen in how many reads it is present. For a typical K-mer its coverage should be around 40%. Coverage higher than 80% should indicate that this k-mer is located in more than one position. Other values arises from errors.

4.3.6 Low-complexity artefacts

Same nucleotide repeats (especially A) are in a lot of cases artefacts. This is due to systematic error and can be evidenced through quality control. To measure if these sequence are in fact artefacts parameters to take into consideration are:

- Low complexity.
- Low entropy.
- High compression (the artefacts increase the information inside the file).

But some low-complexity sequences are not artefacts:

- Hydrophobic transmembrane alpha-helical sequences in membrane proteins.
- CAG repeats in genes causing Huntington disease, spinal and bulbar muscular atrophy, dentatorubropalliodoluysian atrophy.
- Proline-rich regions in proteins.
- Poly-A tails in nucleotide sequence.
- Micro-satellites.

4.3.7 FASTQ quality control (QC)

FASTQ quality control is the first step in any NGS pipeline. It consists of:

Clipping/trimming : removing (low quality) parts of reads.

Masking : avoiding to consider parts of reads that can have low entropy for example.

Read removal : discard low quality reads or reads that are too short after clipping.

Additional features that can be exploited for QC are the GC content, clustering for contamination detection, TAG identification,ambiguous bases.

Chapter 5

Mapping

5.1 Introduction

Mapping and assembly are the operations that allow us to understand sequencing data producing assemblies.

- **Mapping** is a key step in a modern genomic analysis and consists in the process of aligning the reads on a reference genome in order to assign them to a specific location. Insights like the expression level of genes can be gained.
- **Assembly** is the process of aligning and merging overlapping sequences in longer consensus sequences in order to reconstruct the original sequence or genome.

A consensus sequence is the calculated order of most frequent residues, either nucleotide or amino acid, found at each position in a sequence alignment. In many cases, someone may have already assembled the genome or part of the genome (available reference sequences), so sequence assembly is not needed. Assembly will be needed however when studying new organisms. Mapping and assembly need to solve some problem related to the sequencing data:

- absence of DNA fragments covering the gaps, makes it difficult to order the contigs (since there is no connection)
- presence of DNA artefacts (those must be discriminated with Quality Control)
- repeated sequences

5.1.1 Coverage

The coverage, or read depth, is the average number of reads representing a given nucleotide in the reconstructed sequence. The coverage of a genome is defined as the average coverage of each single nucleotide across all nucleotides of the genome. The coverage can be represented with a coverage map and can also be defined theoretically as:

$$Cov = \frac{N \cdot L}{G} \sim \ln \frac{G}{L \cdot \varepsilon}$$

Where G is the length of the genome, N is the number of reads, L is the average read length and $1 - \varepsilon$ is the probability to cover the entire genome with coverage Cov. More reads obtained during sequencing the higher the coverage. In particular, the number N of reads needed to cover the entire genome with probability $1 - \varepsilon$ is:

$$N \sim \frac{G}{L} \ln \frac{G}{L \cdot \varepsilon}$$

5.1.2 Mapping process

In general, the sequence mapping process consists in performing comparisons between experimental sequence data with some reference information. The comparison can lead to obtaining new information, such as the presence of SNPs which can be linked to pathological conditions for example.

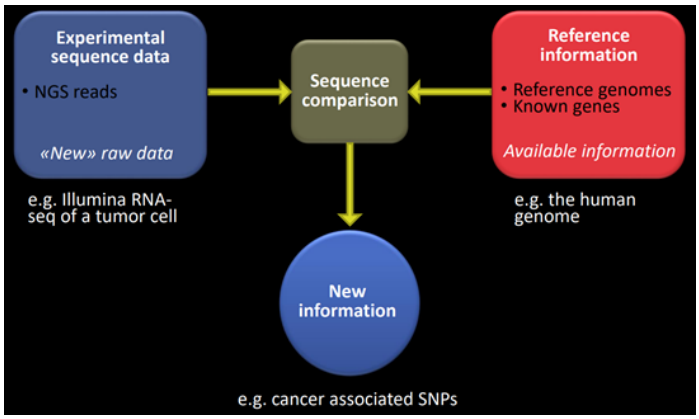


Figure 5.1

5.2 Mapping algorithms

Over time many different mapping algorithms were implemented. Ideally, the simplest aligning algorithm could consists in: a complete search: starting from the first position on the reference, the query sequence is controlled against it and the correct nucleotides are counted and summed forming a score. This is repeated for all positions until a perfect match is found or the position with the highest score is taken. This algorithm is very naive and does not take into account insertions and deletions.

5.2.1 Local vs Global alignment

Sequence alignment can follow two different approaches.

1. In global alignment an attempt is made to align completely the 2 sequences. It is referred as end to end alignment. It finds the best alignment across the whole two sequences. This approach is suitable for comparing closely related

sequences like homologous genes.

2. Local alignment, on the other hand, focuses on finding regions of similarity in parts of the sequences. It aligns subsequences of the query sequence to a subsequence of the target sequence. This approach is suitable for aligning more divergent sequences or distantly related sequences. It is used for example for finding out conserved patterns in DNA sequences of motifs in two proteins.

Sequence similarity is connected with evolutionary distance. Very high similarity implies a very low distance. But when the similarity goes down and reaches the twilight zone as in 5.2 it is more difficult to define the evolutionary distance and to give meaning to results.

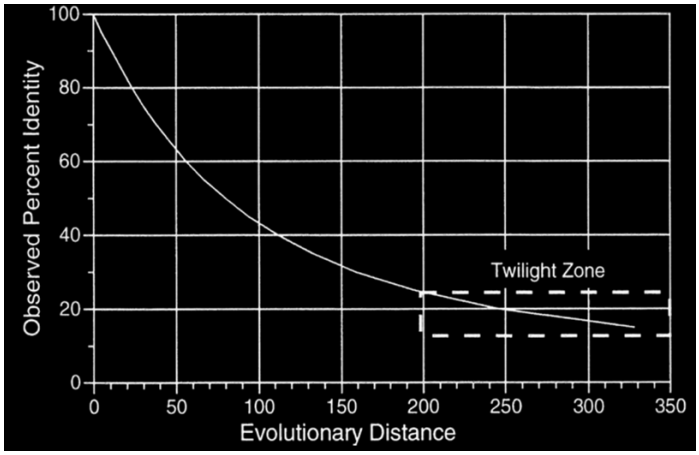


Figure 5.2

5.2.2 Smith-Waterman algorithm

The Smith-Waterman algorithm is a local-alignment algorithm based on dynamic programming, whose aim is to find the best match among all possible optimal local alignment with respect to the scoring system used. Consider two molecular sequences $A = a_1 a_2 \dots a_n$ and $B = b_1 b_2 \dots b_m$. Given a similarity $s(a, b)$ of elements of the sequence and W_k the weight of deletions of length k, to find pairs of segments with high degrees of similarity, a matrix H is set up such that:

$$H_{k0} = H_{0l} = 0 \qquad \forall 0 \leq k \leq n \wedge 0 \leq l \leq m$$

H_{ij} si the maximum similarity of two segments ending in a_i and b_j. H_{ij} is computed such that:

$$H_{ij} = \max(H_{i-1,j-1} + s(a_i, b_j), \max_{k \geq 1}(H_{i-k,j} - W_k), \max_{l \geq 1}(H_{i,j-l} - W_k), 0)$$

With $1 \leq i \leq n$ and $1 \leq j \leq m$. So H_{ij} is:

- $H_{i-1,j-1} + s(a_i, b_j)$ If a_i and b_j are associated.
- $H_{i-k,j} - w_k$ if a_i is at the end of a deletion of length k.
- $H_{i-k,j} - W_l$ if b_j is at the end of a deletion of length l.
- 0 is used to prevent calculated negative similarity, indicating no similarity up to a_i and b_j.

The pair of segments with maximum similarity is found first by locating the maximum element of H. The other elements are determined sequentially with a traceback procedure ending with an element of H equal to 0. This procedure other than identifying the elements produces their alignment. The parameters where:

$$s(a_i, b_j) = \begin{cases} 5 & a_i = b_j \\ -3 & a_i \neq b_j \end{cases}$$

And

$$W_k = \frac{1}{3}k$$

This algorithm in particular allows for the alignment of sequences that contained both mismatches and internal deletions.

5.2.3 Needleman-Wunsch algorithm

This global alignment algorithm was developed in 1970 and is also based on dynamic programming. The purpose of the algorithm is to find all possible alignments having the highest score. The idea is the same of the Smith-Waterman, but it has difference in initialization and computing of the matrix element, in particular:

$$H_{00} = 0$$

$$H_{ij} = \max(H_{i-1,j-1} + \alpha s(A_j, B_i), H_{i-i,j} + d, H_{i,j-1} + d)$$

Again a scoring function s needs to be defined. The fact that the max function contains no 0 anymore means that we are always comparing with the first nucleotide of the sequence.

5.2.4 Heuristic methods

New algorithms were implemented later. BWA and BowTie2 are the best ones available right now for short reads. Blast could also work too but it would take too much time.

This new algorithms are based on heuristic methods: the solution found is not the optimal, but it is the best approximation given some constrains. An heuristic is any approach to problem solving or self-discovery that employs a practical method that is not guaranteed to be optimal, perfect, or rational, but is nevertheless sufficient for reaching an immediate, short-term goal or approximation. Where finding an optimal solution is impossible or impractical, heuristic methods can be used to speed up the process of finding a satisfactory solution. For example BLAST return an alignment on a fixed level of

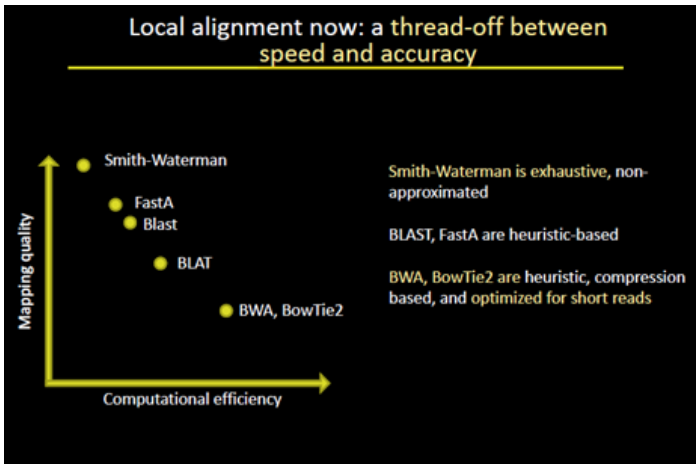


Figure 5.3

similarity and no result under that level of quality. BWA and BowTie2 are also compression based algorithms: they exploit data compression techniques in order to compress reference genome files to reduce the number of bits used to encode it.

5.2.5 BLAST (Basic Local Alignment Search Tool)

Despite its limitations, Blast is still widely used due to is low computational power. The Blast algorithm involves three steps:

1. Seeding: find perfect or almost exact k-mer matches. The idea is to look for identical short matches and try to expand from that.
2. Extension: extend the seeds at point one 1 with possibly some non-exact but high-score matches, that permit to obtain better alignments.
3. Evaluation: create alignments for the regions of high-scoring extended seeds. Every time the statistical significance of the match is evaluated with methods inspired on the NW and SW approaches.

If a k-mer length does not result in any matches its length can be decreased, increasing the computation time needed.

5.2.5.1 Blast flavours

Blast comes in many flavours, depending on what is the desired output and input:

- Blastp: compares an amino acid query sequence against a protein sequence database.
- Blastn: compares a nucleotide query sequence against a nucleotide sequence database.
- Blastx: compares a nucleotide query sequence translated in all reading frames against a protein sequence database. This can be done to find

potential translation products of unknown nucleotide sequences.

- Tblastn: compares a protein query sequence against a nucleotide sequence database dynamically translating in all reading frames.
- Tblastx: compares the six-frame translation of a nucleotide query sequence against the six-frame translation of a nucleotide sequence database.

5.2.5.2 Scoring matrices

Scoring matrices are the way in which blast score matches and mismatches. They are very important especially for amino acids and they are used to score alignments between protein sequences. Blosum 62 (BLOcks SUbstitution Matrix) is one of the most used. Non simple penalties in substitutions of different amino acids are used, based on the functional properties of the substitutions. A scoring matrix contains values proportional to the probability that amino acid i mutates into amino acid j for all pairs of amino acids. Such matrices are constructed by assembling a large and diverse sample of verified pairwise alignments of protein sequences.

5.2.5.3 Blast parameters

Other parameters can be set in BLAST:

- Max target sequences: number of reported sequences.
- Expected threshold: e-value.
- Words size: seed length.
- Gap cost: cost for adding multiple gaps. It can be linear or non linear.
- Filter low complexity regions to avoid getting stuck.

5.2.5.4 BLAST E-value

The *e-value* represents the number of distinct alignments with a score equivalent to or higher than S, that are expected to occur in a database search by chance. It is computed as:

$$Evalue = Kmne^{-\lambda S}$$

Where K and λ depend on the substitution matrix and on gab penalties, n is the query length, m is the length of the sequences of the database and S is the matching score. An e-value of 10 means that up to 10 alignments can be expected to be found just by chance, given the same size of a random database. E-value can be used as a first quality filter for the BLAST search result, to obtain only results equal to or better than the number given by the e-value option. Blast results are sorted by E-value by default (best hit in first line). The smaller the e-value, the better is the match. A small e-value means a low number of hits of high quality, whereas a high e-value indicates many hits, partly of low quality. There is a relationship between the E-value and the p-value.

- The E-value is the number of sequences that we would find by chance;
- The p-value measures the probability of finding by chance another sequence with an equal or better score.

In particular:

$$Pval = Ke^{-\lambda S} = Evalue/mn$$

5.2.6 Speed seed alignment

There are algorithms implemented 10/15 years ago that focus on seeds and are note used any more. Those methods cut both the reads and the reference sequence into small seeds. The reference seeds are then stored in an hash look-up table. The idea is to do that for all possible k-mers of the two sequences. Algorithms based on this approach are: Maq, SOAP, MOSAIK. This present problems in the presence of SNPs:

- 1 SNPs means that at most 1 seed is not-matching.
- X SNPs means that at most X seeds are not-matching.

The problem of these algorithms is that the lookup table is too big to be fully loaded in memory.

5.2.7 Burrow-Wheeler alignment

The Burrows-Wheeler algorithm is used in BowTie and Bowtie2. The Borrow-Wheeler algorithm is based on a very efficient way to store the reference genome, based on the BW transformation, that allows to have a reference hash table much smaller than the one of the speed seed approach. This particular method is successful because the compression is reversible and so the reference does not need to be decompressed. It is this reversible property in fact that guarantees that the reads can be found in the genome. The search is based on finding suffixes of the reads in the BW structure. With this approach the index of the human genome is around 2GB. The BW-based algorithms are the fastest currently available.

5.2.7.1 Reversibility of the Borrow-Wheeler compression

To compress a word:

1. Add a terminator to the word.
2. Form a matrix containing all the rotations of the word as rows.
3. Sort them based on lexicographic order.
4. Take the last column.

In the end this string is much more compressible because the same letters tend to be grouped together. In order to reverse the transformation

1. Repeat until the matrix is square:
   (a) the input sequence result of the compression is added as a column to a matrix.
   (b) Sort the rows of the matrix according to lexicographical order.
2. The row where the last element is the terminator is the untransformed input.

This reverse transformation can be made faster through the last-first property.

5.2.7.2 LF (Last-First) property

The LF property is then applied to the matrix of all rotations to get back the input sequence. This is possible because the ith occurrence of character X in the last column corresponds to the same text character as the ith occurrence of X in the first column. Following this property the word is reconstructed backward:

1. Look the first element c of the last column and add it to the end of the output.
2. Go to the corresponding ith occurrence of c in row j.
3. Add to the output element in the last column in row j.
4. Repeat with $c = MAT_{j,last}$ until the terminator is found.

5.2.7.3 Exact mapping using LF property

When performing a mapping operation sequence matching can be done using the compressed database. Backward exact mapping works by calculating the range of matrix rows beginning with successively longer suffixes of the query. Doing the backward matching we exploit the characteristics of the BWT transform. This same approach could be used to find a match between a query read and a reference genome which has been compressed using the BW alignment. The problem is that it does not take into accounts indels or mismatches.

5.2.7.4 Inexact mapping

A possible solution to this problem could be the inexact mapping. Perform exact mapping and, if the query is not found, go back and perform backtracking by hypothesising mismatches.

Chapter 6

Assembly

6.1 Introduction

Assembly is the process of aligning and merging overlapping sequences in longer consensus sequences to reconstruct an original sequence or genome. No reference database is needed.

6.1.1 Use cases for assembly

Assembly is needed to be performed when:

- Sequencing a genome for the first time.
- When the reference genome is not complete or very distant phylogenetically and hence not usable as a reference.
- In case of new genes, which cannot be discovered just by mapping.

This process is rarely needed in human genomics as the human genome is already available. By contrast it is particular important in microbial genomics to capture new genes of different strains not present in the reference genome.

6.1.2 General framework for assembly

In theory, all assembly algorithms could be based on a basic framework that:

1. Identifies overlaps between the reads obtained by sequencing.
2. Connects the reads through those overlaps.
3. Finds the Hamiltonian paths.
4. Finds the consensus sequence.

In practice this is not always possible, due to many issues that can occur, like:

- multiple overlapping of the same reads. This can be due to high coverage or location in the genomes that match at least partially the reads.
- Partial matches.

One possible solution would be to keep track of all possible matches and then try to reconstruct the sequence. One way to represent the possible overlaps between reads is by using a graph in which the nodes represent the reads and the edges represent the connections between the reads that are overlapping in some way. Then a mathematical algorithm is needed to find a solution in this very intricate network. Many assembly algorithms are available, and they are all based on this idea. For this reason, we always end up with multiple copies of genomes, not just one. Also, the output is almost never the full genome but pieces of the genome in the form of multiple contigs. By default, assembly is done in this way and gives a good representation of the genome, but not complete. Some quality control is also needed, especially to detect contamination.

6.1.3 General pipeline

The general pipeline for sequence assembly consists in a few steps:

1. Find overlapping reads.
2. Merge the good pairs of reads into longer contigs. A contig is a contiguous sequence formed by several overlapping reads with no gaps.
3. Link contigs to form scaffolds or supercontigs, which are ordered and oriented sets of contigs. Scaffold assembly usually exploits mate pairs which allows to link different contigs together, even if the information in the middle is not completely known. Mate pairs, also called long-insert paired-end reads (LIPERs), are a kind of read obtained from paired-end sequencing which can pair reads across great distances. Scaffolds do contain gaps, but at least the order of the contigs is known and helps to reconstruct the whole genome.
4. Derive consensus sequence. Sometimes there are multiple scaffolds and the consensus sequence is the set of all scaffolds representing the genome or the chromosome that is being reconstructed.

Finding the order of contigs is not easy and different approaches are used, hence scaffolding is usually a quite time-consuming operation. To close the genome additional experiments, like PCR to clone pieces between contigs could be needed. In microbial genomics, assembly usually finishes at the contig step. Contigs contained in the output files allow to define the genes present in the genome not defining their position in the genome, which is not needed and more difficult to find.

6.1.3.1 Mate pair library preparation process

Following DNA fragmentation, the DNA fragments are end-repaired with labeled dNTPs. The DNA fragments are circularized, and non-circularized DNA is removed by digestion. Circular DNA is fragmented, and the labeled fragments (corresponding to the ends of the original DNA ligated together) are affinity-purified. Purified fragments are end-repaired

and ligated to Illumina paired-end sequencing adapters. Additional sequences complementary to the flow cell oligonucleotides are added to the adapter sequence with tailed PCR primers. The final prepared libraries consist of short fragments made up of two DNA segments that were originally separated by several kilobases. These libraries are ready for paired-end cluster generation, followed by sequencing utilizing an Illumina next-generation sequencing system. Combining data generated from mate pair library sequencing with that from short-insert paired-end reads provides a powerful combination of read lengths for maximal sequencing coverage across the genome.

6.2 Sequence assembly algorithms

Doing a full assembly is a NP-hard computational problem. Because of this it cannot be done in reasonable time.

6.2.1 Quality parameters

Read length is the most important parameter when trying to perform assembly: without long enough reads it is impossible to assemble greater contigs. Also the coverage is important, but less so than read length. In assembly also helps to know the characteristics of the sequence reads, aiding in discriminating against articfacts.

6.2.2 Merging overlapping reads

When merging overlapping reads it must be considered that the strongest the similarity between the end of one read and the beginning of another the highest the likelihood the reads are coming from the same overlapping sequence region. So when overlapping reads the length of the overlap and accuracy must me maximised. Due to error all overlap can be not perfect and a score to each is assigned. A not perfect overlap can be due to the fact that:

- The two reads are not really contiguous but just similar.
- Sequencing error and noise.
- Diploid genome or multiple copy genomes.

6.2.3 Overlap graphs

Reads overlaps are represented by directed graphs in which:

- Nodes represents reads.
- Directed and weighted edges represents overlap: they connect the final part of the source node's sequence connect with the initial one of the target node's sequence. The weight is the score of the overlap.

Since most organisms' DNA is circular these graphs are ideally cyclic and the alignment is an Hamiltonian graph. However, sometimes cycles are just due to random overlapping or to the presence of repeats.

6.2.3.1 Repets

The problem of the repeats is a key aspect in bacterial genomics since bacterial genomes have many regions with repeats. Also, due to the short reads this repeats are difficult to find.

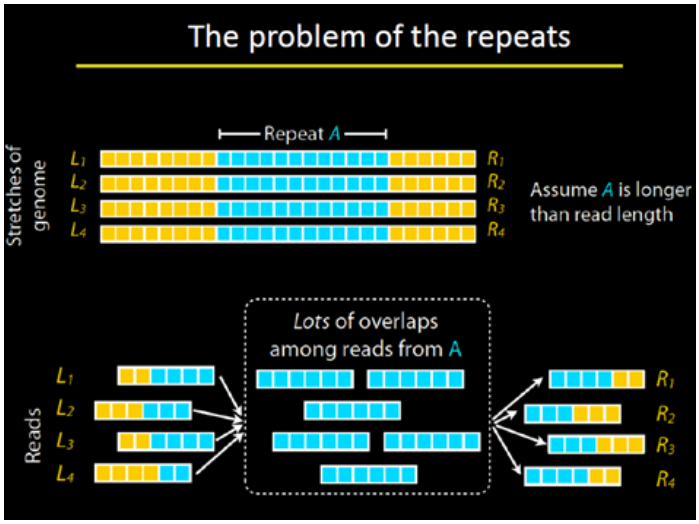


Figure 6.1: Example of genome with repeats

If the read length is not long at least the length of the repeat a solution cannot be found. It will be impossible to determine the length of these repeated regions. Also in the graph there will be a lot of cycles and connecting the two extreme of the repeat will not be possible. With overlaps graph in particular repeats can cause the production of wrong contigs that assume that two region are sequential when they are not. Hints of the presence of repeats are:

- Region with an abnormally higher coverage.
- Pairs over repeated regions have a shorter insert size.

6.2.3.1.1 Solve trough coverage This problem can be partially solved through coverage: assuming that the repeat has the same coverage of all other reads an expected length of the region could be computed. This is risky because repeated regions are problematic for the sequencers and a change in coverage for these region could happen.

6.2.3.1.2 Stopping the contig Stopping the contig before the repetitive region and starting it after it is the safest solution.

6.2.3.1.3 Long reads Sequencing with reads longer than the repeat is the only true solution.

6.2.4 Solving an overlap graph

The ideal way to solve the overlap graph is to find all Hamiltonian paths. These paths connect all nodes once and they provide all possible solutions. The best path is the one that maximizes the score of the overlaps. This problem however is NP-hard and not feasible. Another approach is to use a greedy approach:

1. Randomly select a starting node.
2. Select the connected node with maximum overlap as the next visitor.

This approach will find a local optimal solution which could not be the global optimum.

6.2.4.1 Graph simplification operations

To minimize the number of operation needed to solve the overlapping graph some simplification operations are needed. These can be:

- Merging consecutive nodes: if among more reads there is only one possible path they can be merged without loosing any information. Nodes that are sequentially connected only with each other are linked together.
- Remove dead ends: if some reads are branching from a path toward a dead-end, they can be removed. The removed path could be the right one and some information is lost.

6.2.4.2 Limitation of overlap graphs

Overlap graphs have some limitations:

- they are problematic when dealing with repeats: maximizing the overall weight will produce wrong assemblies.
- They are not tractable: finding the optimal solution is not computationally feasible.

Overlap graphs can work to some extent, but now other algorithms are used in assemblers, like the de Bruijn Graphs (DBG) and the Overlap Layout Consensus (OLC).

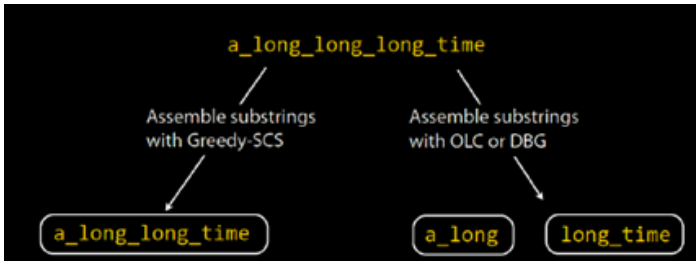


Figure 6.2

In the case of DBG for example, when facing repeats the contig is split into two: better two right contig than one wrong contig.

6.2.5 De Brujin graph assembly

In De Brujin graph assembly a graph is build such that:

- Nodes are all the unique k-mers present in the input reads. Where k is an hyperparameter chosen such that it is smaller than the reads lengths.
- Edges are the overlaps of the k-mers of length $k+1$ from 2 sequences.

The objective is to find the Eulerian path in the graph (path in which all edges are visited exactly once). This search is much more efficient than the one of the Hamiltonian path. Moreover the set of notes will not be too big compared with the number of reads. When choosing k a trade-off between number of nodes and edges needs to be done. The k-mers that appears only once with great coverage can be removed as they are result of sequencing error, pruning in this way the graph. Within the graph the paths are more linear with less possibility for wrong cycles to appear.

6.3 Post-assembly operations

6.3.1 Scaffolding

Scaffolding is an operation that tries to link together two contigs into a scaffold, consisting of sequences separated by gaps of known length. This can be done with paired-end sequencing or estimation of coverage.

6.3.2 Evaluating assemblies

N50 length is the sequence length of the shortest contig such that 50% of the sequence is contained in in it. Increasing the length of contigs increases in turn the N50, which can be used as a measure of the quality of the assembly process.

Chapter 7

16S-rRNA sequencing

7.1 Introduction to metagenomics

7.1.1 Definition of metagenomics

The term metagenomics refers to the study of uncultured microorganisms from the environment, which can include humans or other living hosts with focus on taxonomic and functional characteristics of the total collection of microorganisms within a community. The main way to analyse the entire microbial population of an environment is through high-throughput sequencing of nucleic acids isolated from the sample. Two approaches can be distinguished:

- 16S rRNA gene sequencing.
- Shotgun metagenomic.

7.1.2 Why studying the metagenome

Microbes are basically everywhere, in and outside of our bodies, in oceans, glaciers, hot springs and rocks. Given how widespread and abundant microbes are, studying the metagenome provides us plenty of information both on human and non-human microbiome and environment. For instance, it has been shown that the microbiome correlates to several diseases, therefore it can be used as a non-invasive biomarker. The list of activities microbes are involved in is constantly increasing.

7.1.3 Differences with older microbiome studies

The microbiome was discovered many years ago but there were no tools to analyze it properly: the only way was to culture and isolate each bacterium. This is an unfeasible approach to study the entire community, since only some bacteria can be grown in the laboratory and it would take an unreasonably long amount of time. The advent of high-throughput technologies is what made possible to study the microbiome of a sample, reducing significantly times, costs and increasing substantially the fraction of the microbiome that can be known.

7.1.4 Example: skin microbiome

Some studies were performed on skin microbiome (Segata et al, Nature Methods 2012, Truong et al, Nature Methods, 2015). Only about 60% of the contigs of various size were mapped to known microbes while 40% belonged to unknown species. When separating these sequences based on GC content and abundance, many clusters formed, some with higher abundance while others with lower abundance, probably due to the low GC content that makes more difficult for the machine to sequence them, therefore causing them to be underestimated. Studying this 40% of unknown sequences is one of the main tasks of metagenomics.

7.2 16S rRNA sequencing

16S rRNA sequencing is one of the first techniques developed to study the microbiome, since it does not require a huge amount of sequences nor excessive costs.

7.2.1 Simplified 16S rRNA analysis workflow

The general workflow for a 16S rRNA analysis is the following:

- DNA extraction from the entire community present in the sample. Some bacteria will be over-represented while other will be under-represented.
- Selective PCR amplification of 16S rRNA gene.
- High-throughput sequencing.
- Sequence mapping against genomes in databases. This allows to define which bacteria and which variants of those are present in the sample and to find new and unknown bacteria.

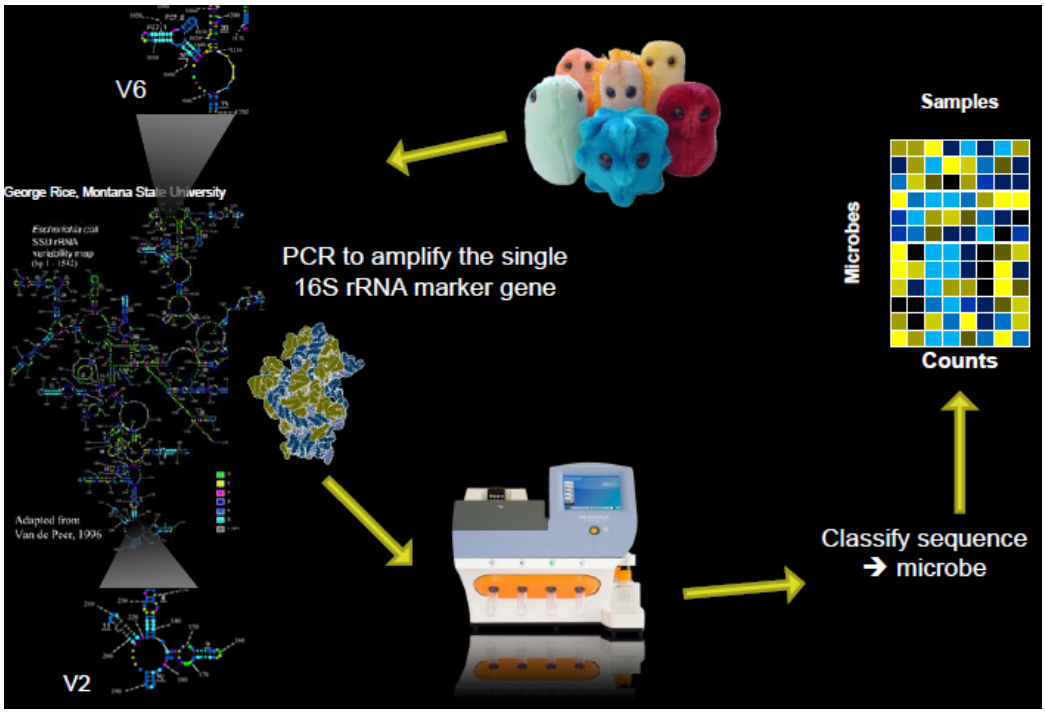


Figure 7.1: General 16S gene analysis workflow

7.2.2 16S rRNA gene

The ribosome is one of the most conserved, if not the most conserved, structure in all living organisms, making it one of the best phylogenetic markers. In prokaryotes, the ribosome is composed of several elements, both proteic and RNA based. Of the RNA based ones, 3 of them are ribosomial RNAs (rRNAs), namely 5S, 16S, 23S. Since these components are fundamental for any bacterium, all bacteria present the genes codifying for these rRNAs. Moreover most of the sequences are highly conserved but some regions have some species-specific variability which can be used as a barcode to find and classify species. The most conserved of the rRNAs is 23S but the one used for microbiome analysis is 16S (which corresponds to the human 18S). Its gene is a few thousands nucleotides long, most of which are highly conserved. The bulk of the differences among species is in the hypervariable regions named V1 to V9, the terminal loops of the structure. They are regions far away from the catalityc site. Despite the high degree of conservation, some variability can be found outside the hypervariable regions too. The annotation of which portions of the 16S rRNA gene are conserved has been performed using E. coli as a reference. For a few hundred organisms the gene has been compared to the reference one to define the degree of conservation of each stretch of nucleotides. Some totally conserved regions are present but they are not very big.

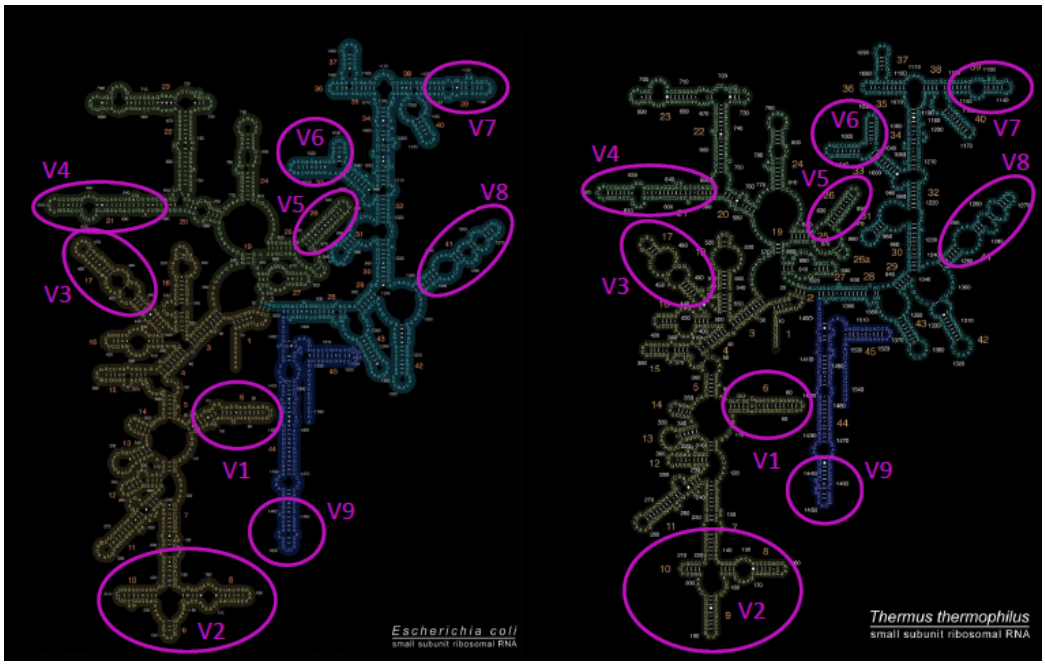


Figure 7.2: Structure of the 16S rRNA in *E. coli* and *T. termophilus*

7.2.3 Primer and high-throughput machine choice

One could sequence the entirety of the 16S rRNA gene, for example using NanoPore seq, but this would introduce many errors that could lead to mapping the sequence to the wrong organism. For this reason it is preferred to amplify only certain specific regions of the gene. To study the microbiome in a high-throughput way primers which can bind to all species are needed. Since the sequences conserved in all species are too short, you use primers that bind highly conserved regions. For this reason, regardless of which primers you choose there will be a bias in your results: some species will not be identifiable using those primers. This bias can be somewhat minimized using in silico primer validation, which means testing your primers against databases of 16S tRNA genes like silva and green genes, to test and decide the best pair of primers for your experiment.

Still, two experiments conducted with different primers will always have some differences. Moreover the binding regions must flank some variable region, in order to include it in the amplicon. Finally pair end amplification both primer back and forward is needed in order to have the complete amplicon to make the comparison easier. Given these characteristics there are multiple possible priming sites based on the sequence and on chemical properties of the primers. Moreover, primers can be used as forward or reverse to obtain different combinations and sequences.

As an example of the importance of the choice of primers, in some skin microbiome analyses, researchers could not find two bacteria always present on human skin due to the choice of primers. Moreover S. aureus seemed over-represented due to the non-amplification of other important species. Despite the biases, this technique is still extremely useful.

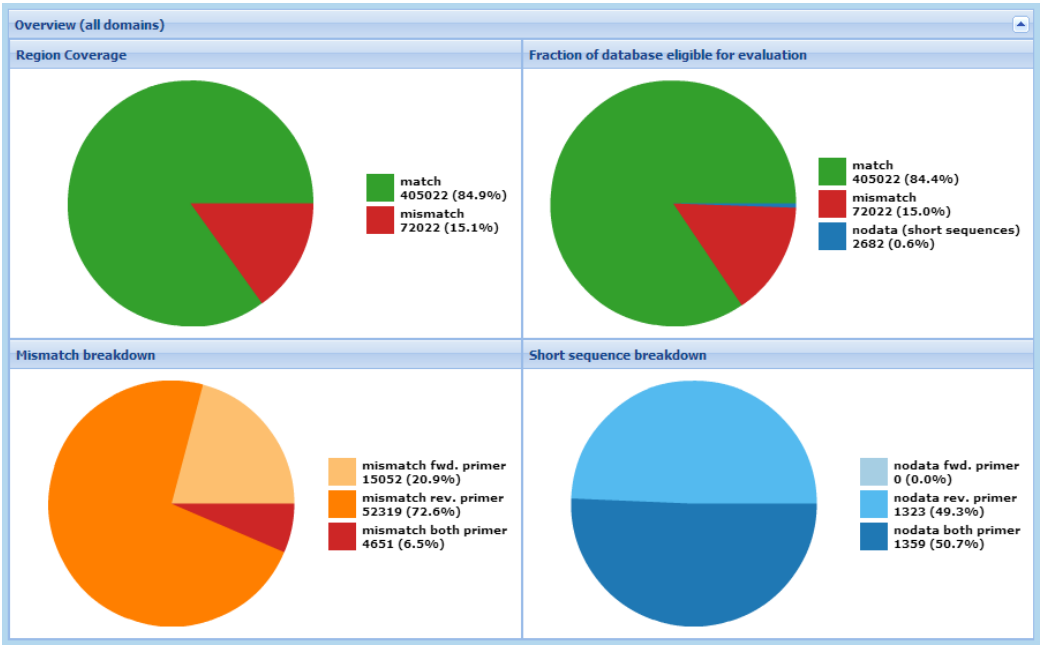


Figure 7.3: Example of *in silico* primer validation using silva; you can notice the different efficiency of the primers relative to different parameters.

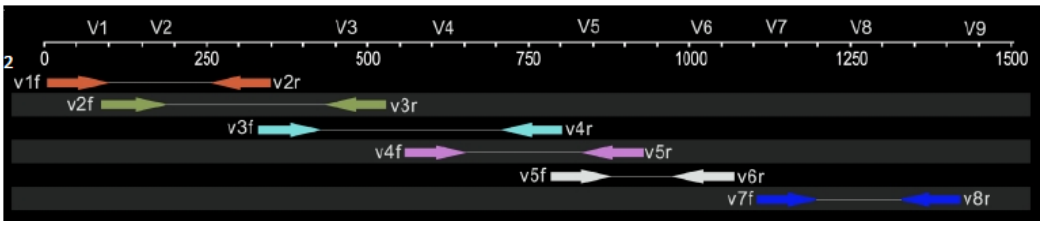


Figure 7.4: Examples of common primer placements relative to hypervariable regions

There are different protocols to target conserved regions also based on the machine used. In general:

- Sanger machines are not very good for this application since they have low throughput and they are more suited for longer sequencing tasks.
- Roche 454 machines have historically been well suited, since it was possible to sequence three hypervariable regions together using 400 nucleotides reads, providing a good cost and throughput trade-off.
- Illumina HiSeq is not the optimal choice since it has shorter reads and unnecessarily high throughput. Illumina MiSeq and IonTorrent can be a decent compromise.

7.2.4 In depth 16S rRNA analysis workflow

Adding detail, a more complete overview of the 16S rRNA analysis workflow is:

- DNA extraction from each of your samples
- Selective PCR amplification of 16S rRNA gene, introducing a barcode in the sequences using tagged primers.
- High-throughput sequencing of all the samples in a single run to reduce costs. The result is a set of amplicons belonging to different samples and with a barcode attached.
- Demultiplexing, which means removing the barcodes and assigning each sequence to the corresponding sample. Sequencing noise must be taken into account, therefore low quality reads must be removed.
- Multiple sequence alignment against reference sequences. Some reads will probably not map to any reference sequence.
- Group related sequences into OTUs (operational taxonomic units), which means grouping sequences that share some common variants. Since SNPs in the microbial genome are present, the similarity threshold between sequences cannot be too restrictive. OTUs can be used to define the relative abundance of each species in the sample, but in order to do so it is necessary to normalize for the copy number of the 16S gene sequence. This is very difficult since an accurate estimate can be made only if long read sequencing has been performed on the organism, which is almost never the case since for microbes that basically corresponds to full genome mapping.
- Build phylogenetic tree using one representative for each OTU.
- Annotate the OTUs using 16S gene databases.
- Downstream analysis is performed, such as clustering to visualize similarities among samples.

7.2.5 OTU clustering

Defining OTUs requires using multiple sequence alignment. Since this approach is a generalization of the mapping algorithm it is quite complex in terms of speed, but still feasible. Generally greedy algorithms which add the lowest possible amount of gaps are used to perform multiple sequence alignment. After the alignment, sequences are split into OTUs (operational taxonomic units), which are groups of 16S sequences very similar to each other. Generally a sequence is defined as the representative of the OTU, meaning that it has a certain threshold of identity with all other sequences in the OTU, usually 97% when considering species and that minimizes the differences of all other sequences of the OTU with itself. Some OTUs can be assigned univocally to a species, some others may be associated to more species, some others cannot be mapped to know species. The fact that a species may map to multiple OTUs is often an error but it may sometimes allow to find subspecies.

After sequence alignment, OTU clustering can be done through several supervised or unsupervised learning methods. The most common unsupervised clustering methods are:

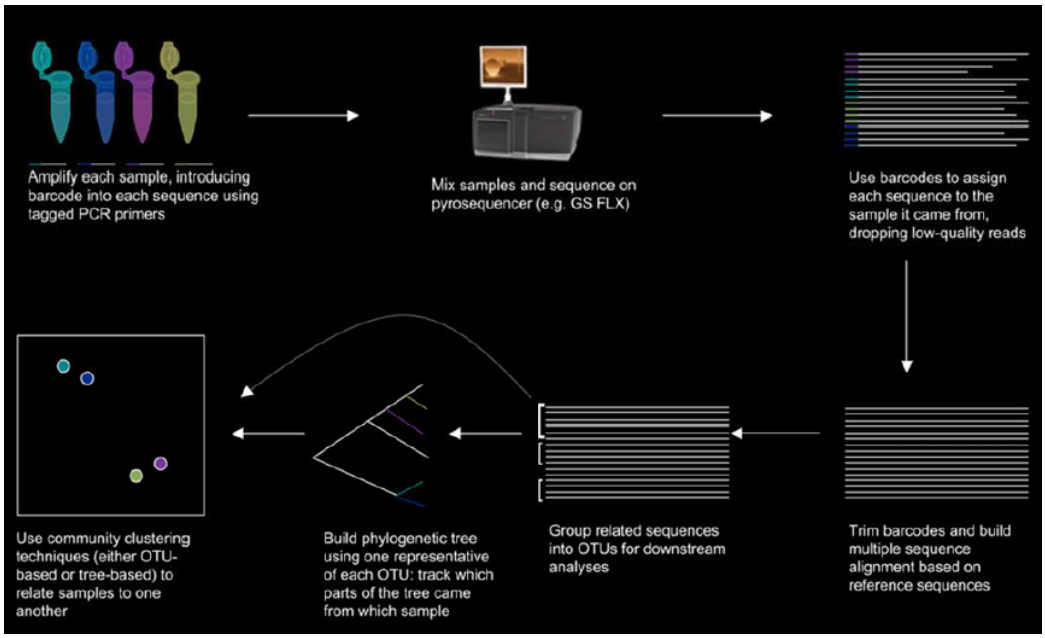


Figure 7.5: Expanded 16S gene analysis workflow

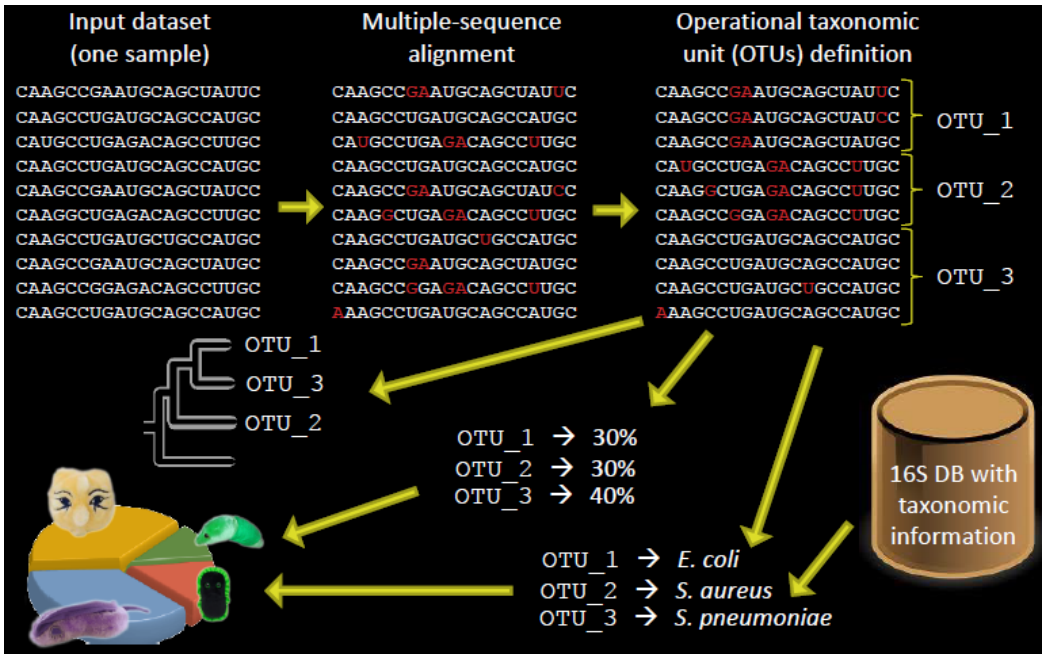


Figure 7.6: Zoom in on 16S gene analysis workflow

- Single linkage clustering nearest neighbour: assign the sequence to a cluster if that OTU already contains at least a sequence similar enough (97%).

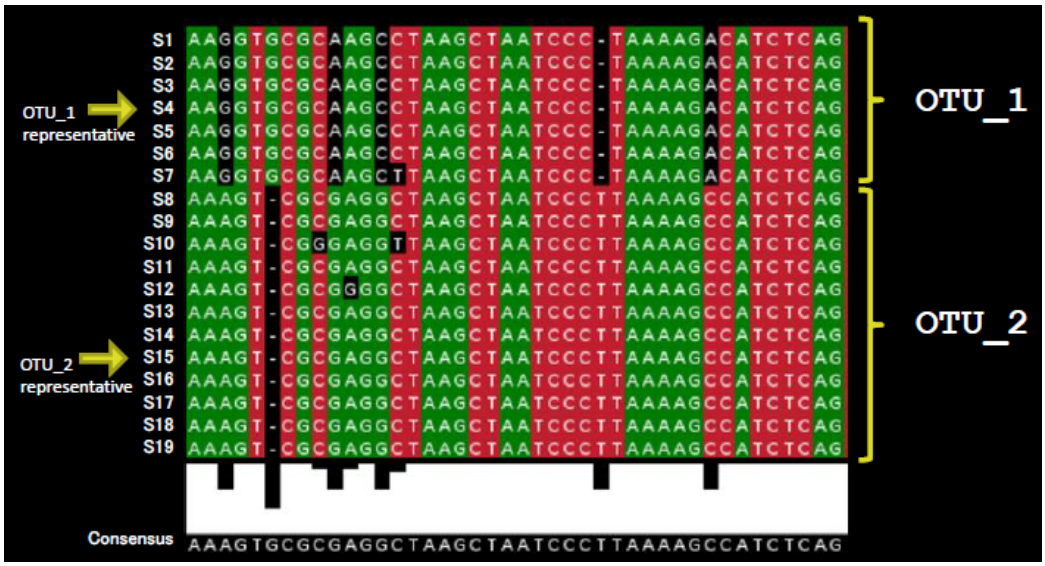


Figure 7.7: Example of multiple sequence alignment for OTUs

However two distant sequences in the OTU network could share a similarity which is way lower than 97%. This could result in underclustering.

- Complete linkage clustering furthest neighbour: assign the sequence to a cluster only if all the sequences of the OTU are similar enough (97%). However two sequences may be similar enough, yet belong to different OTUs, because the overall cluster width, or divergence, is at most 3%. This approach could then generate different solutions, based on the order the points are added in. Moreover, if the clustering conditions are too stringent, sequencing errors and SNPs in the microbial genome may result in overclustering (defining too many clusters).

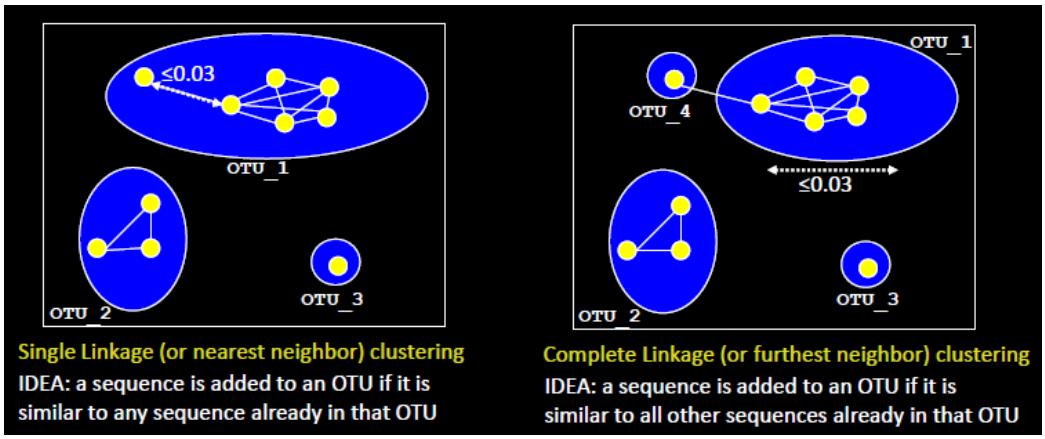


Figure 7.8: Visualization of single linkage analysis and complete linkage analysis

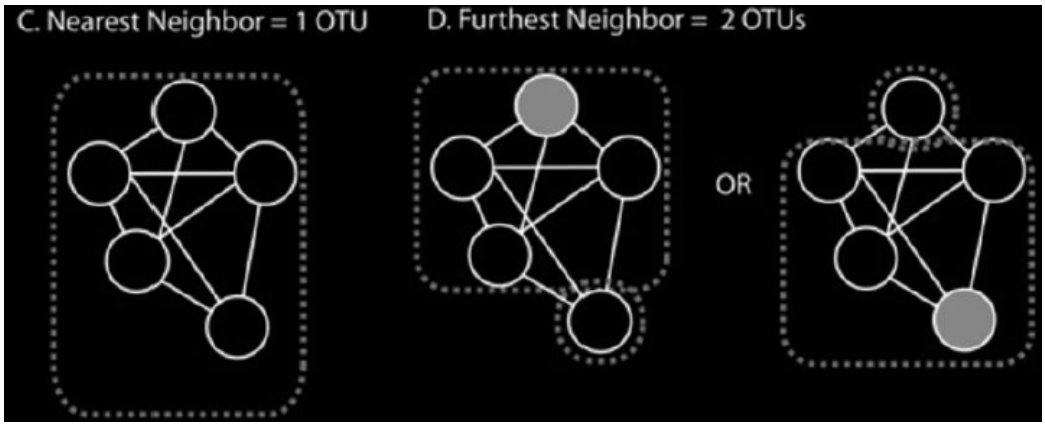


Figure 7.9: Example of overclustering and result multiplicity due to complete linkage analysis

7.2.6 OTU taxonomic annotation

Assigning a taxonomic annotation to an OTU cannot be done simply using BLAST to get the best matching sequence. This is because there is too much noise in the sequences and because it is difficult to classify new strains. A better way is using some other algorithm that assigns the terms of the taxonomic notation (since it is more than just one label) and provides some degree of confidence in the prediction. For instance the algorithm may be able to correctly assign the first taxonomical terms, up until Enterobacteriaceae, but then it provides a prediction of the OTU belonging to a list of species with confidence value for each one.

7.2.6.1 RDP classifier (Naive Bayes Model)

$P(S|G)$ is computed by RDP using a 8-mer strategy: comparing the 8-mers in S with the 8-mers in all the training sequences available for genus G. The confidence of each prediction is computed by bootstrapping:

1. Select a random subsequence of S, S'.
2. Compute the G that maximizes $P(S'|G)$.
3. Repeat the procedure a number of times.
4. The number of time G is selected by the bootstrapping procedure is the confidence

Leave one-out-cross validation:

- Take out one data point from the training set.
- Apply the classifier on the left-out point (without using it in the training set).
- Check the accuracy of the prediction.
- Repeat the procedure for each training data point.

The RDP classification accuracy is evaluated with the leave-one-out cross validation on the training set of hundreds thousands of 16S references:

- Accuracy is the percentage of correct classification (over the all leave-one-out runs).
- There are varying levels of taxonomic resolution (from phylum to genus).
- There are varying sequence lengths.
- In real applications accuracies are probably smaller due to sequencing errors or the presence of unknown bacteria.

7.3 Diversity analysis

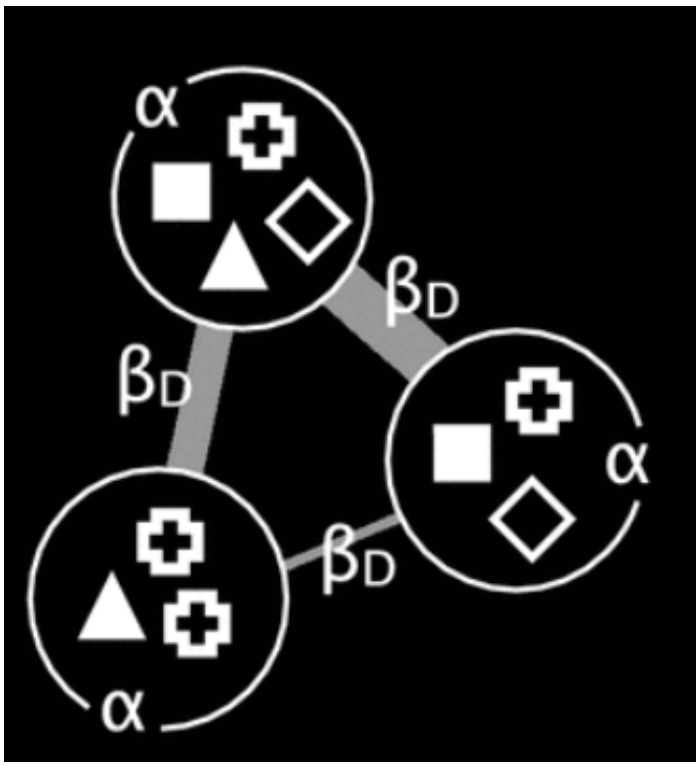


Figure 7.10

7.3.1 Alpha diversity analysis

Alpha diversity analysis is a measure of how diverse or complex a microbial community is. It measures within sample diversity. Species richness is a widely use alpha diversity index. All individuals considered have non-zero abundance, some will have high abundance ($\sim 99\%$) or low abundance (1%). High alpha diversity are usually associated with populations that are more robust and resilient to changes. For examples gut microbiome with a high

richness is usually associated with healthy state, instead of disease. Alpha diversity can be compared only between samples with the same sequencing depth. To do so between different samples usually a depth cut-off is chosen.

7.3.2 Beta diversity analysis

Beta diversity analysis is a measure of how different two microbial communities are. It measures between sample diversity. It is possible to measure the beta-diversity using the inverse of number of shared species. An example of beta-diversity is UniFrac. In UniFrac the distance is equal to the fraction of the total branch length that is unique to any particular environment. UniFrac can be also weighted in order to include abundances for each OTU.

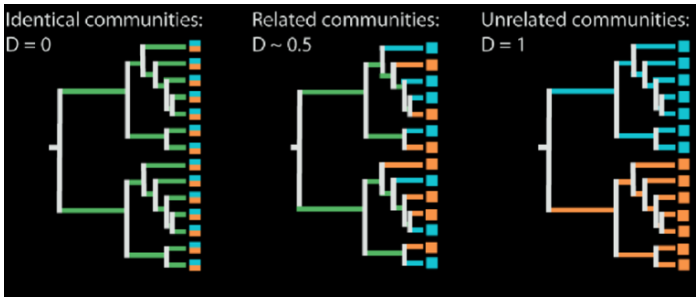


Figure 7.11: D=0. Blue and orange samples always have the same OTUs. Each 16S (each branch) is present in both samples. **D=∼0.5.**In reality we usually have a mix of the 2 situations. Some OTUs are present only in one of the samples and are either quite distant from the others or close (based on upstream the branch goes). **D=1.** Completely distinct OTUs . The difference is also in the upstream branches, which have different colors.

7.3.3 Principal Coordinate Analysis

PCoA is also known as multidimensional scaling. It is one of the most powerful approaches for exploratory analysis. The idea is to represent the multidimensional relationship between samples in a two or three dimensional space. It is possible to use any similarity function as Euclidean distance, UniFrac, bray-Curtis distance. We find frequently hierarchical clustering plots. It is mostly done to visualize the similarities and difference between species, identifying for example cluster of species.

Chapter 8

Shotgun Metagenomics

8.1 Introduction

8.1.1 Shotgun metagenomic analysis

A shotgun metagenomic analysis consists of different steps:

- Experimental pipeline: from sample collection to DNA sequencing.
- Preprocessing: decontamination and quality control.
- Mapping, assembling or both.
- Sequence analysis: identification of microbial species and identification of present pathways and functions.
- Post Processing: integrate data with other information coming from sample metadata.
- Validation: follow-up experiments with independent replicates.

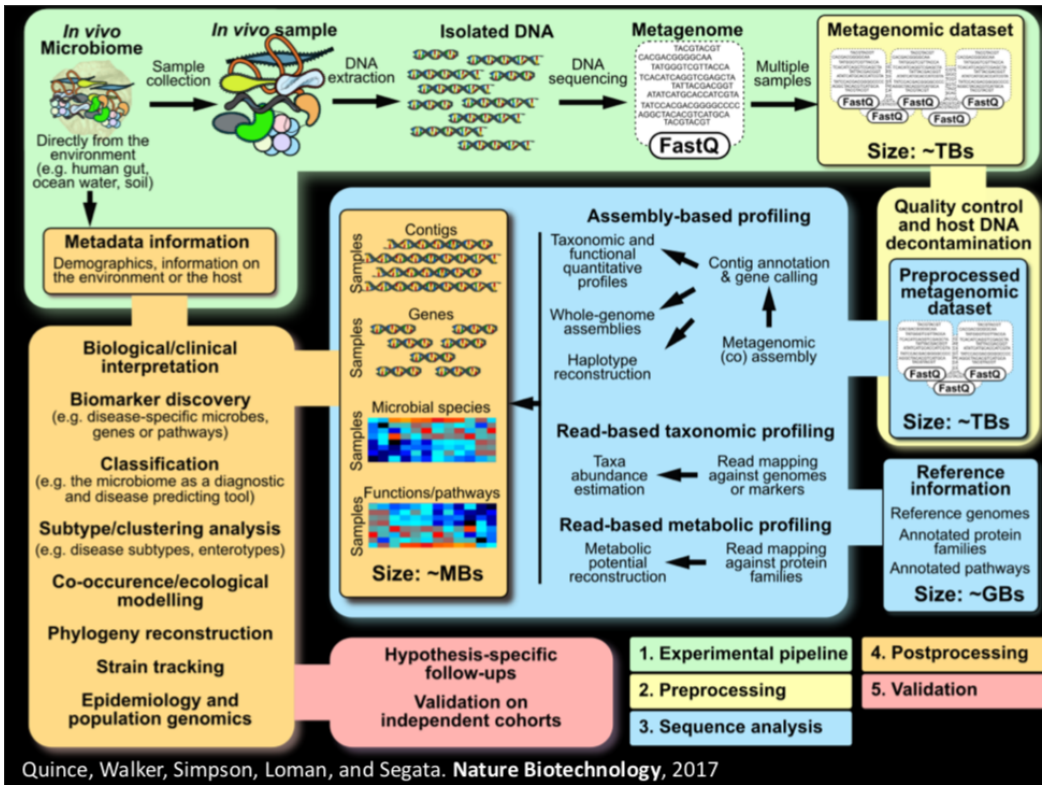


Figure 8.1: Shotgun metagenomics workflow

8.1.2 Comparison with the 16s sequencing

16S sequencing	
Pros	Cons
Cost-effective Avoids non-bacterial contamination Can catch low-abundance bacteria The output has reasonable size and complexity Mature software available to perform the computations	Non genome-wide Limited taxonomic resolution Not useful for pathogens profiling Does not detect viruses or eukaryotes Several biases Cross studies are difficult: the comparison is not possible due to biases

Shotgun sequencing	
Pros	Cons
Genome-wide: it is possible to retrieve information about all the genes present in the metagenome High taxonomic resolution Easy cross-study comparison thanks to the lack of biases All domains of life can potentially be observed in the same study	Expensive (but costs are decreasing: right now the cost for the sequencing of one sample is $\sim 100\$$) DNA contamination are hard to remove Low-abundance bacteria could be missed Large dataset as output, that can be difficult to process (TBs of data)

8.1.3 Latest technology

The cost of $100\$$ for the sequencing of one sample refers to a sample size of $5Gb$, that should be enough for shotgun metagenomics. More sequencing depth could be needed if microbes with very low abundances needs to be detected, if many new genomes need to be assembled or if there is a lot of contamination. Shotgun metagenomics is possible with Illumina Hiseq technology, but the latest and most used technology nowadays is Illumina NovaSeq.

8.2 Identification of microbes from Shotgun metagenomics data

The main challenges with respect to identifying microbes from metagenomics data are:

- How to obtain species-specific resolution.
- Computational feasibility.
- Being able to detect both bacteria and archaea. Phages are also a problem since there is little reference.
- Obtain relative abundances of organisms with different genome sizes.
- Consistent detection confidence for all clades.
- How to handle reads as short as $50nts$.
- Detect organisms without a sequenced genome or still unknown species.

8.2.1 MetaPhlAn: unique marker genes for taxonomic profiling

The main idea of MetaPhlAn is to find a marker gene that uniquely characterizes a species. This gene has to be present in all strains of a species and in no other species. This markers form taxonomic clades. ChocoPhlAn is the tool that generates the database necessary. From a number of genomes, ChocoPhlAn was used to create a database of marker genes that build the MetaPhlAn database. This database contains 200 markers per species and its reduced dimension with respect to the whole genomes database can align reads directly to the marker database.

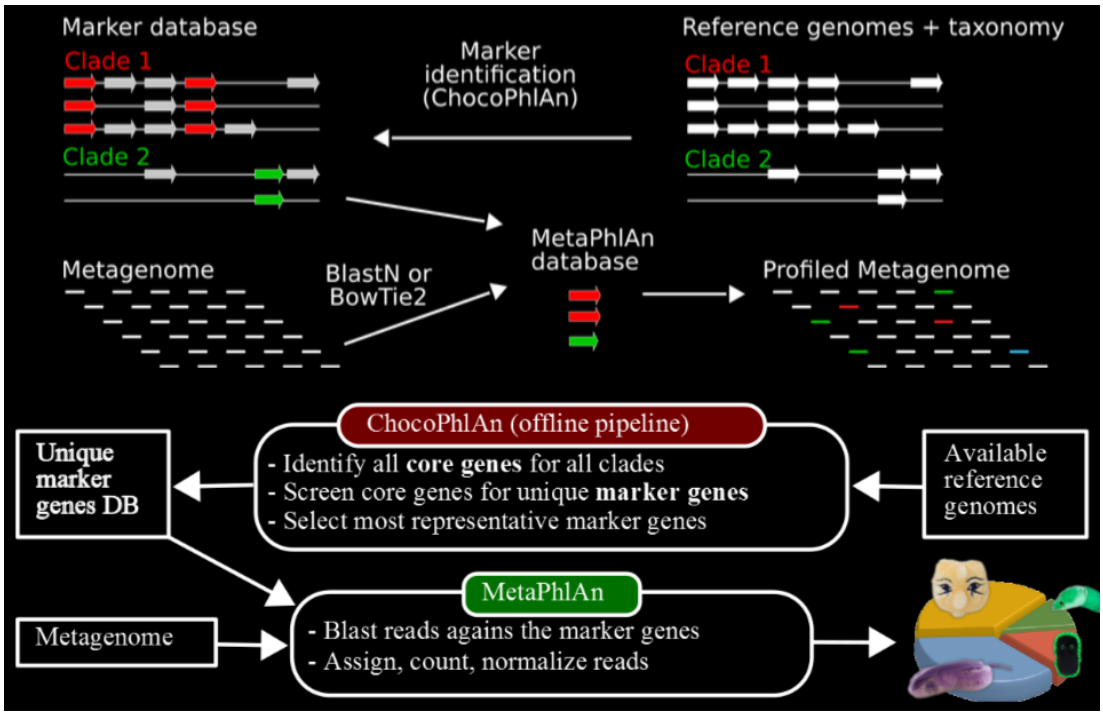


Figure 8.2: MetaPhlAn overview

8.2.1.1 Creation of the marker genes database

The marker database was created from 2887 genomes from 1222 species. The number of marker per species decreases with more sequenced genomes because the core genome tends to become smaller in size and noise is eliminated. In the current version 1 milion genomes are considered, with 200000 isolated genomes and 800000 from metagenomics assembled genomes or MAGs.

8.2.1.2 Efficiency and validation

MetaPhlAn can deal with 200000 reads per second and can profile thousands of microbiomes in a few hours. Validation can be done with synthetic metagenomes created with

errors or with biological methods: comparing the results with the ones obtained with 16S based abundance estimations.

8.2.1.3 The problem of the unknowns

Microbial species that were never observed and do not have a marker in the database cannot be detected with this method. A possible solution is to cluster together contigs obtained from the metagenome based on coverage, GC content, codon bias, and other possible features. Strict quality controls are then performed on these putative genomes: on the number of genes and the number of known single-copy genes in order to be sure that what has been found is not a mixture of genomes. High quality putative genomes are considered MAGs and version 4 of MetaPhlAn can include them in the creation of the marker database. This way, these species can be detected in metagenomes even though they do not have a name yet.

8.2.2 Other approaches

Other approaches include:

- Sequence-based clustering of contigs to create putative genomes. The output are unlabelled bins with relative abundances. The main problem is that high coverage is needed to obtain valid results.
- Machine learning algorithms that exploit GC content (and possibly other features) to give as output clades with relative abundances. This method is not completely reference-free as it uses reference genomes to extract the features used for the classification. The main problem is that many species could have really similar features.
- Read-to-genome sequence mapping. In this approach there is no processing of a reference genome.

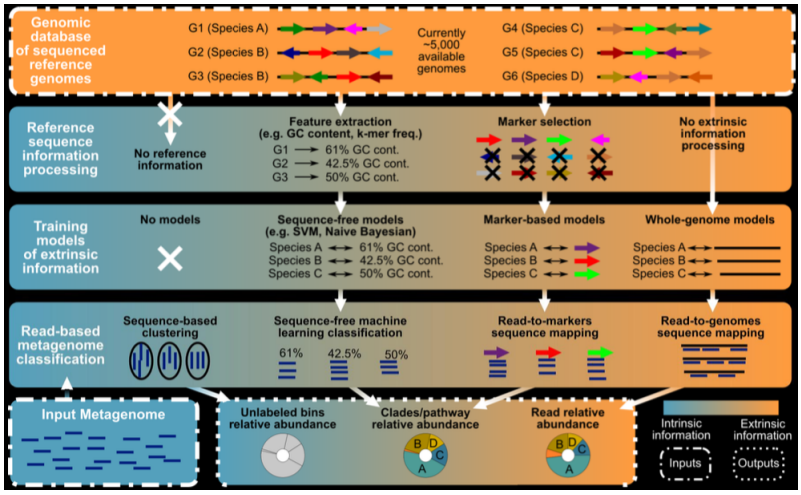


Figure 8.3: An overview of taxonomic profiling approaches

8.2.3 The curated MetagenomicData resource

Since raw metagenomic sequencing data can be quite difficult to deal with computationally, the curated metagenomic data database stores features obtained from raw metagenomic datasets uniformly processed (MetaPhlAn or HUMAnN2) and integrated with associated metadata obtained from NCBI, papers and authors. This database is accessible and can be exploited to perform various types of analysis.

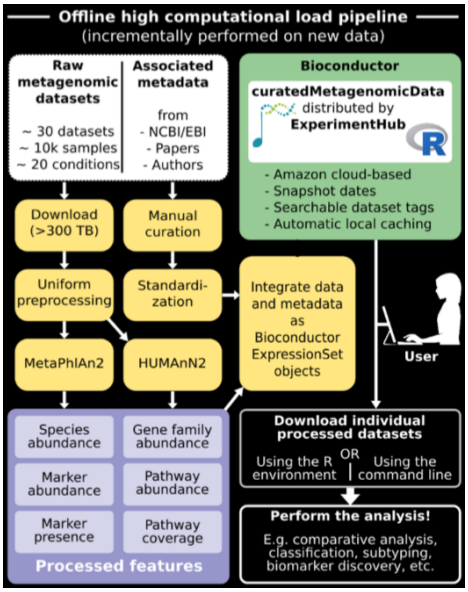
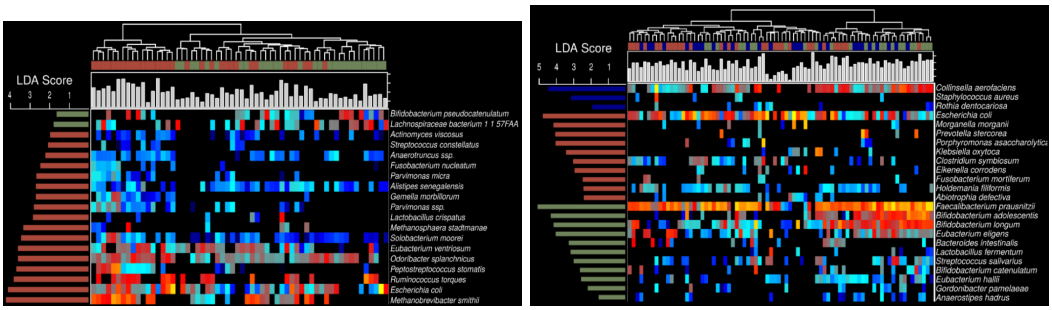


Figure 8.4: CuratedMetagenomicData pipeline

8.2.4 The link between the gut microbiome and colorectal cancer

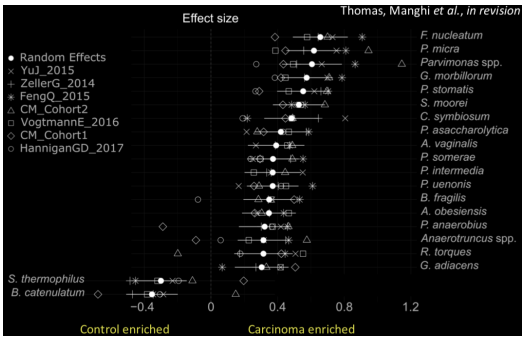
Colibactin is a genotoxic metabolite produced by E. coli: it causes damages to the DNA, possibly causing cancer onset. A fraction of human colorectal cancer (CRC) cases are caused by colibactin. A study published by Segata's group collected stool samples from people having a colonoscopy in Milan (8.5a) and Turin (8.5b). The samples were then categorized after the diagnosis provided by the colonoscopy. The aim was the identification of biomarkers associated with the CRC phenotype. Different results were obtained in the two cities.

Comparing this study with similar ones performed around the world (France, China, Austria, USA, Germany and Japan) some biomarkers appeared to be reproducible (8.6a). Moreover, an accuracy around 80% was observed when a machine learning approach (random forest) was applied on all the datasets combined and then the model was applied on a brand new one(8.6b). On the other hand, when each group tried the same approach on their data separately, completely different results were found for each dataset, showing that such a technique could be valid for some cases and completely unhelpful for others.

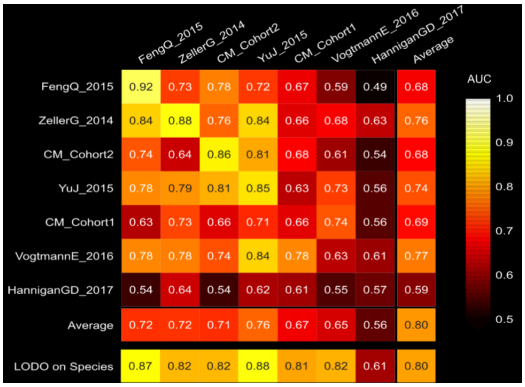


(a) Patients from Milan **(b)** Patients from Turin

Figure 8.5: Taxonomic profiling of gut microbiomes



(a) *F. nucleatum* found to be enriched in datasets with different origins



(b) Random forest approach on all the datasets shows an accuracy of 0.80 (AUC of 1 means always right prediction, 0.5 means no prediction ability)

Figure 8.6

8.2.4.1 Hypothesis-driven analysis

The cutC gene appears to be associated with the CRC phenotype (8.7). The problem is that it is present in several microbial species and it is unknown whether the function and the efficiency are maintained.

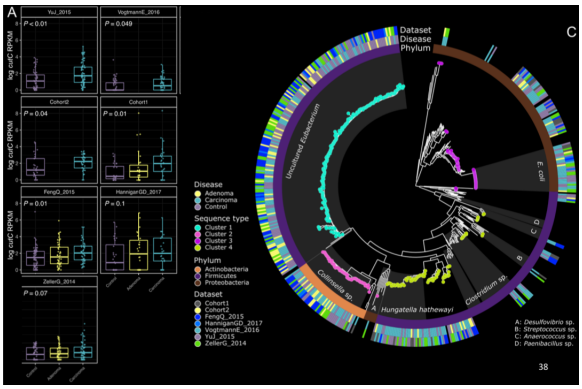


Figure 8.7

8.2.5 PanPhlAn: strain-level profiling

PanPhlAn is a tool for strain-level metagenomic profiling that allows to identificate gene composition of individual strains in metagenomic samples. The main difficulties to consider with this approach are:

- Identify the microbial strains present in the metagenome or check whether a specific one is present in the sample.
- Discover new strains and species.
- Characterize the metagenome genomically.
- Track across samples to find the same strain and eventually prove that transmission of bacterial strains occurred among them.

Let's consider an analysis regarding the *E. coli* pan genome contains about 20.000 gene-families. The goal is to find what strains are present in the metagenome and their abundances. First all genes are grouped in functional gene-families (8.8). Then genes from *E. coli* found in the metagenome are mapped on *E. coli* reference genomes with BowTie2. The coverage is computed for each gene and then they are grouped into gene-families (8.9). Then gene-families are ranked based on their coverage. Multi-copy gene families have really high coverage, while the plot shows a plateau of single-copy gene-families. These families correspond to the strains present in the metagenome and their abundance can be deducted thanks to the base coverage of single-copy genes.

8.2.5.1 Investigating population genomics thanks to PanPhlAn

8.2.5.1.1 *E. coli* population genomics with PanPhlAn Figure 8.11a shows the *E. coli* profiling of 1478 shotgun metagenomes carried out with PanPhlAn. Each column is either an *E. coli* strain obtained via shotgun metagenomics or a reference strain and the columns correspond to the gene-families that can be absent or present in each strain. The strains are then clustered (8.11b) based on which gene-families are present in order to study the population genomics: in this case we can see that the strains isolated from the German *E. coli* outbreak cluster together, while other strains are present in several different areas of the world.

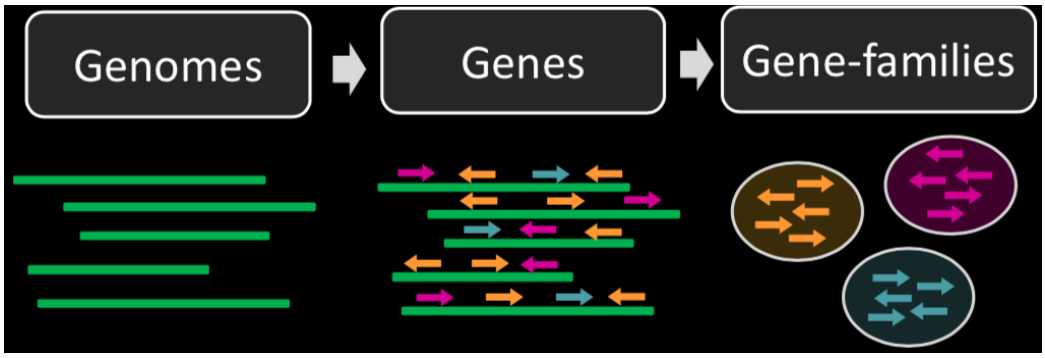


Figure 8.8: Genes are grouped in functional gene-families

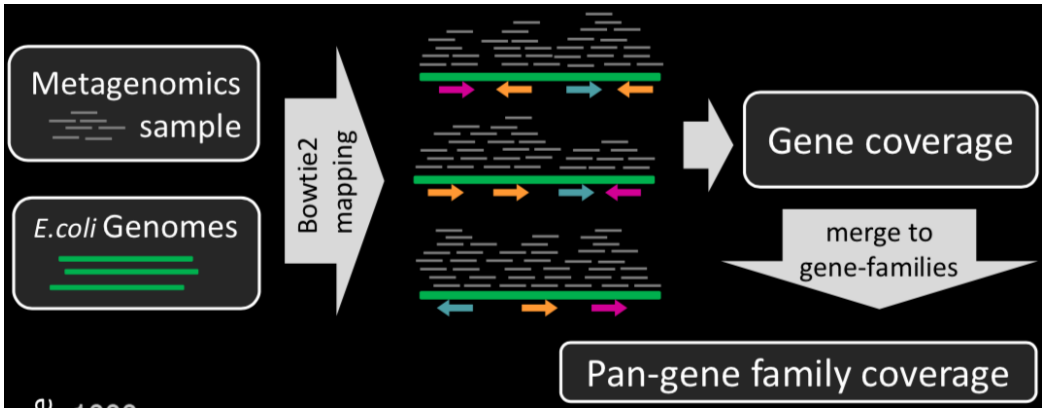


Figure 8.9: Mapping and subsequent coverage of gene-families

8.2.5.1.2 PanPhlAn on *Eubacterium rectale* Thanks to PanPhlAn, it was possible to identify many subtypes of *E.rectale* even though only one reference genome was available at the time (8.12).

8.2.5.1.3 The infant gut microbiome in disease Necrotizing Enterocolitis (NEC) is a devastating disease that affects mostly the intestines of premature infants. This study took samples from a cohort of 173 infants, 151 of them were preterm and 30 of them had NEC. They obtained 460 shotgun metagenomic samples and 284 shotgun metatranscriptomic samples, all of which were coupled with clinical data of the patients. The heatmap shows MetaPhlAn profiling of different bacterial species for all the samples: about 20% of the patients have a high *E. coli* predominance (8.13a: yellow circle). PanPhlAn was employed to investigate the *E. coli* strains found in these patients: of the 4 identified clades, only 2 are associated with NEC (8.13b).

8.2.6 StrainPhlAn

The idea in which StrainPhlAn is based on is to base the classification on the genetic variance of core genomes: look for unique combinations of SNPs in genes that are always

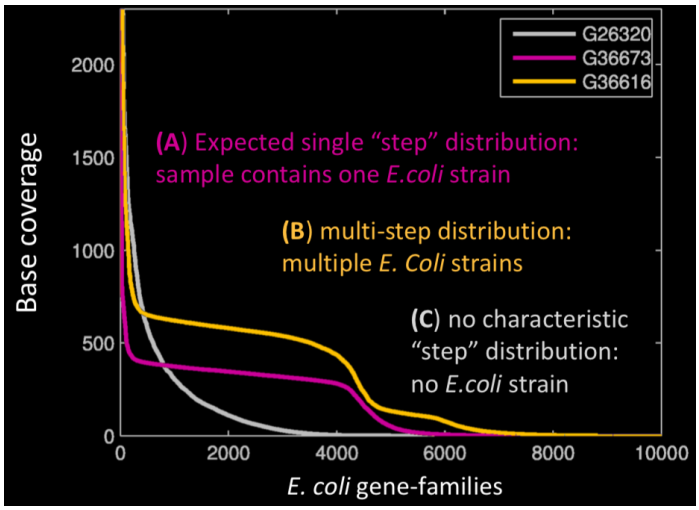
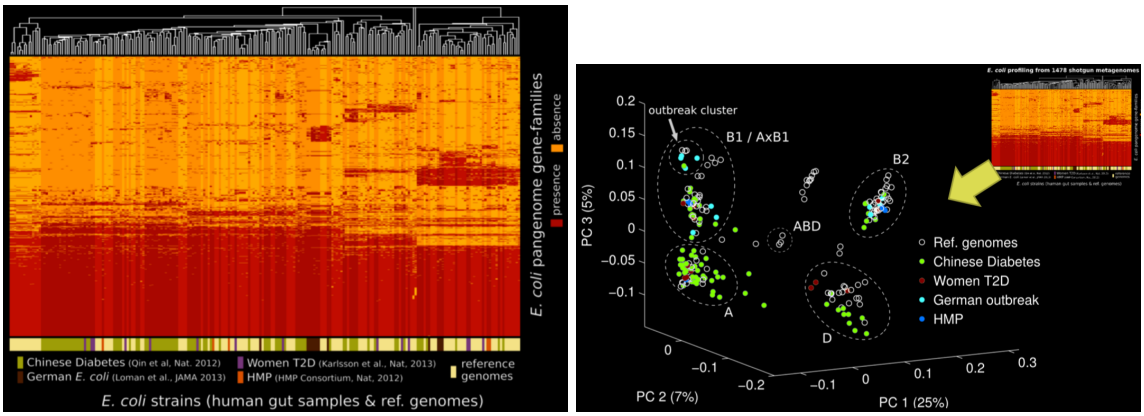


Figure 8.10: *E. coli* gene-family distribution: curve (A) shows the typical gene-families distribution: multi-copy genes with extremely high coverage, a plateau of single-copy genes and a tail of non-present gene-families. Curves like (C) should be discarded from the analysis because they indicate that no *E. coli* strain is detected in that sample.



(a) **(b)**

Figure 8.11: *E.coli* population genomics with PanPhlAn

present and analyse their variance to find some SNPs with a variance different from the others that could characterize a new strain. StrainPhlAn exploits MetaPhlAn to compute species-level abundances thanks to species-specific markers and then aligns the marker genes present in the samples to find the SNPs. The SNPs are then analysed to build a phylogenetic tree. This approach can be applied to many different species. For example studies concerning *E. rectale* seems to have a higher resolution compared to the previous approach.

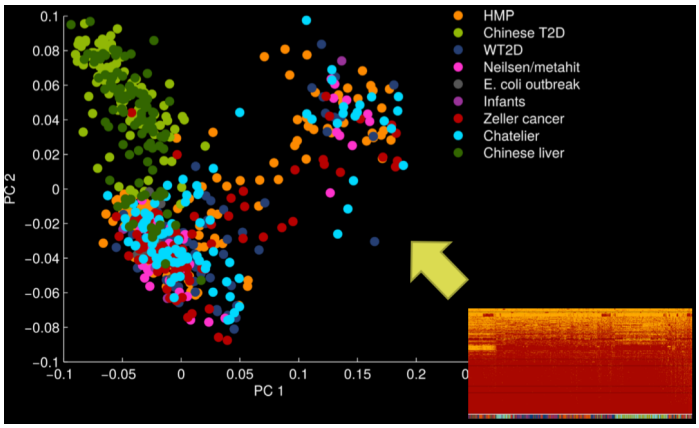
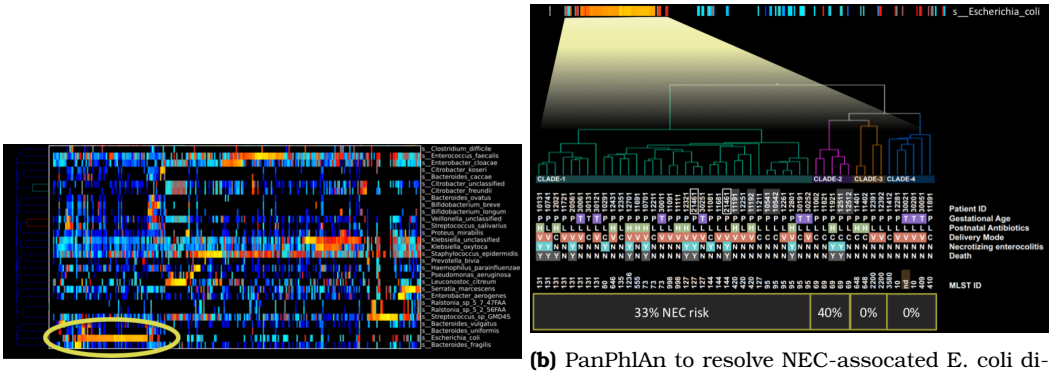


Figure 8.12: PanPhlAn on *Eubacterium rectale*



(a) MetaPhlAn2 profiling of the cohort of infants

(b) PanPhlAn to resolve NEC-assocated E. coli diversity

Figure 8.13: The infant gut microbiome in disease

8.2.6.1 StrainPhlAn applications

8.2.6.1.1 The stability of strains in the human gut This study analyses samples of the human gut microbiome obtained from different continents. The barplots (8.15) show the distances measured between results of SNPs analysis in different regions around the world. There are almost no pairs with zero or low distance and the results do not change if only subjects from the EU or US are considered. On the other hand, the red and blue barplots (8.15) show that comparing the SNPs of samples coming from the same individual but collected 6 months apart they show great similarity. This indicates that there is some stability in the human gut microbiome: usually the strains present in one's gut microbiome tend to remain almost the same. Nevertheless, changes of diet or other habits can bring variations in their abundances.

8.2.6.1.2 Identification of subspecies Some bacteria are strongly represented in the human population's microbiomes and their presence is detected in all continents. On the other hand subspecies tend to be highly region-specific. In Figure 8.16, each color

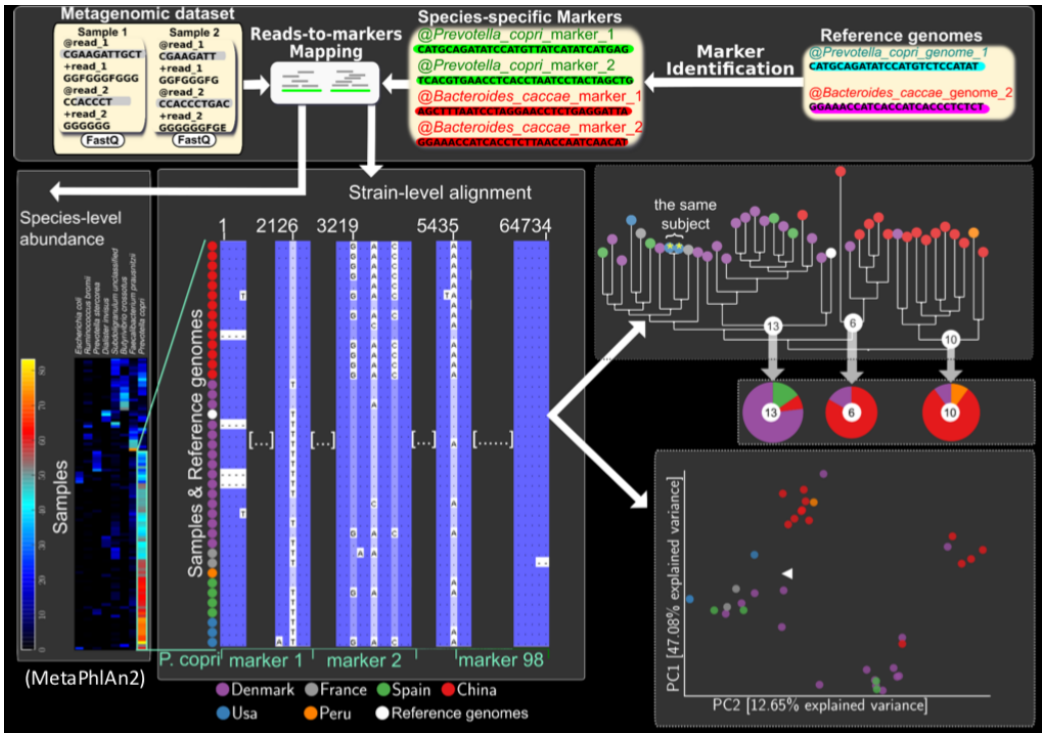


Figure 8.14: StrainPhlAn pipeline

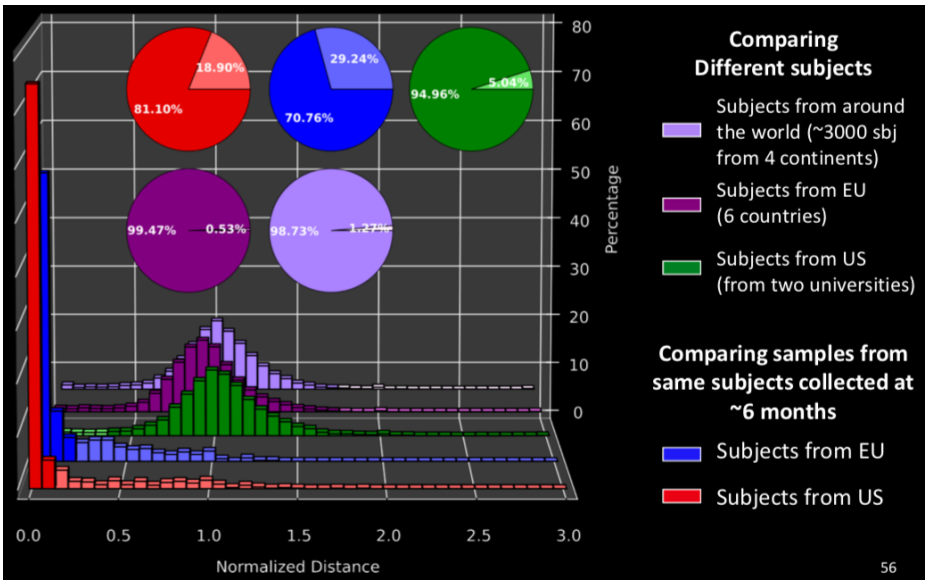


Figure 8.15: The stability of strains in the human gut

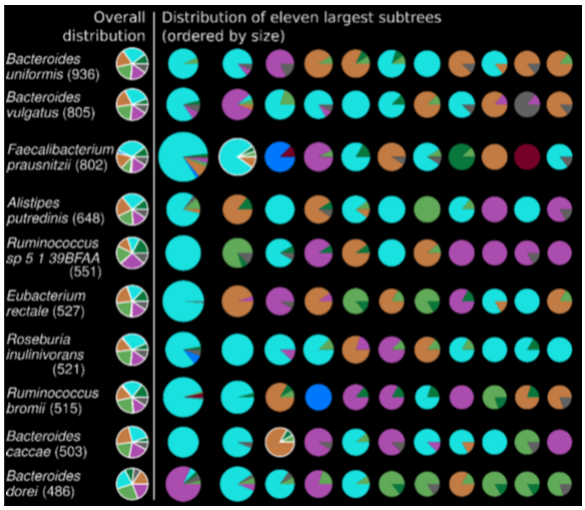


Figure 8.16: Association of sub-species structures with geography

indicates a different country: it is clear that when looking at subspecies of these common microbes, they are mostly associated with only one country or continent.

8.2.7 Uncharacterised species

A great amount of information about the human is still unknown:

- Functional unknowns: genes for which we still do not know the function because they do not match any functional database.
- Unknown species/strains: genes not matching any of the known reference genomes.
- Undetected unknowns: things we do not know and were not even observed yet.

8.2.8 Workflow for large scale metagenomic assembly

Westernized metagenomes are more characterized than non-westernized ones and this study performed a large scale metagenomic assembly on data from undercharacterized region of the world. They were able to reconstruct ~ 70.000 high quality genomes and ~ 85.000 medium-quality ones (with 50% completeness).

8.2.8.1 Project pipeline

Workflow (8.17):

- Assembly of the reads into contigs.
- Binning of contigs into putative genomes (MAGs = Metagenome-Assembled Genomes).
- Quality control.
- The MAGs are then clustered in species-level genome bins by measuring the distance scores between cou-

ples of putative genomes.

- SGB are divided into known, unknown and non-human.

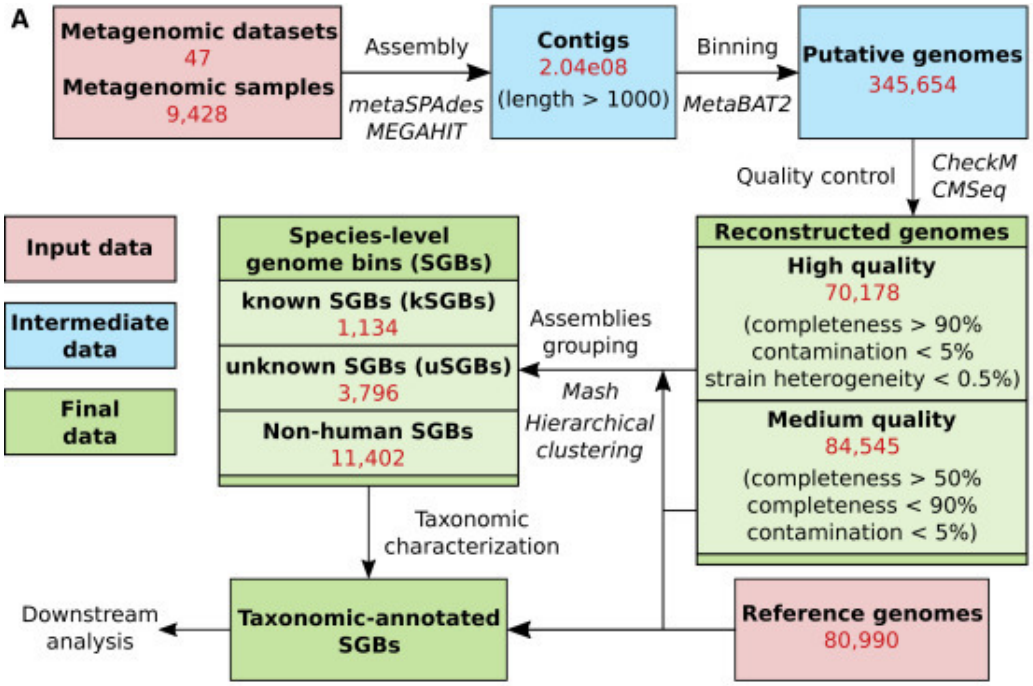


Figure 8.17: Workflow of large-scale metagenomic assembly

This study allow to characterize the new cibiobacter. In particular Madagascar associated strains of cibiobacter uniquely possess the trp operon for tryptophan metabolism.

8.3 Applications of strain-level metagenomic profiling

8.3.1 *E.rectale* refined population genomics

Thanks to the advancements brought by strain-level metagenomic profiling, a new subtype of *E. rectale* was discovered. Subtype 3 lacks the operon coding for motility and other genes that become useless for the bacteria if it is not motile. On the other hand, they present more copies of the CAZy genes: these are involved in metabolism and are required by non-motile bacteria in order to be more efficient in the exploitation of carbon sources since they cannot move to reach them.

8.3.2 *Prevotella copri* is strongly lifestyle-associated

P. copri is a frequent bacterium in the gut microbiome and it tends to be an on/off one: if it is present, it is the most abundant in the body. 4 *P. copri* clades with the ability of degrading complex fibers were found. They are called clades just because it is not

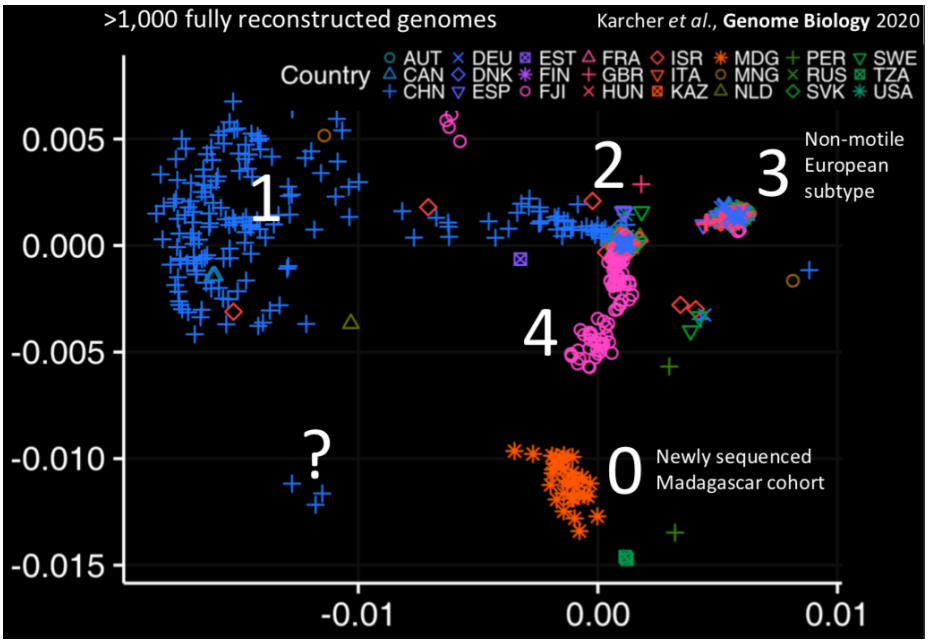


Figure 8.18: *E.rectale* refined population genomics

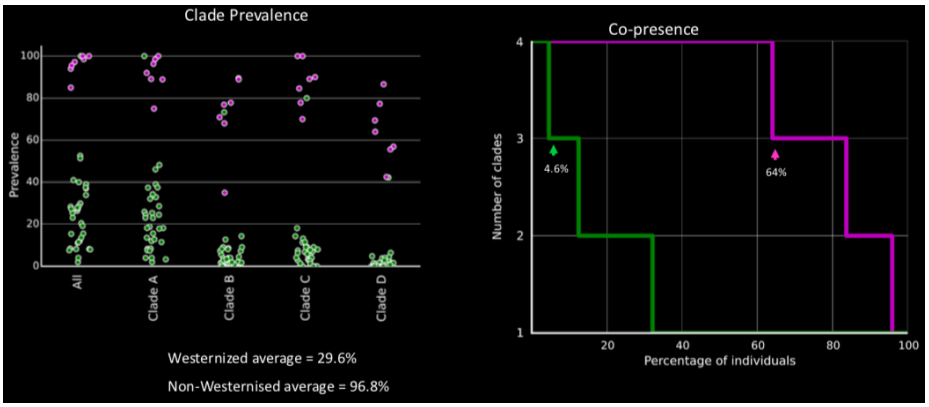


Figure 8.19: *Prevotella copri* is strongly (Western) life-style associated

yet confirmed that they can be considered as new strains. The interesting part is that westernized populations are associated with a lower prevalence of these clades and with a lower probability of presenting all the 4 clades together (8.19). This is probably caused by our diet: westernized populations tend to eat less complex fibers than non-westernized ones. This was confirmed by the analysis of the microbiome found in Ötzi (3.300 BC) and some ancient Mexican coprolites (600-1300 AD). The *P. copri* clades were found in these samples, indicating that we are possibly losing *P. copri* in the westernized populations.

8.3.3 Identification of *Akkermansia* candidate subspecies

Only two subspecies of *Akkermansia* have been described and characterized so far: *A. muciniphila* and *A. glycaniphila*. In this paper, they used PhyloPhlAn3 to identify 4 MAGs that are candidate *Akkermansia* species. Moreover, they observed that these candidate *Akkermansia* species are mutually exclusive for what concerns hosts: they were rarely found coexisting in the same sample. They also appear to have different associations in respect to *A. muciniphila*: one is associated with decreased host body mass index (BMI) but the others are not (8.20).

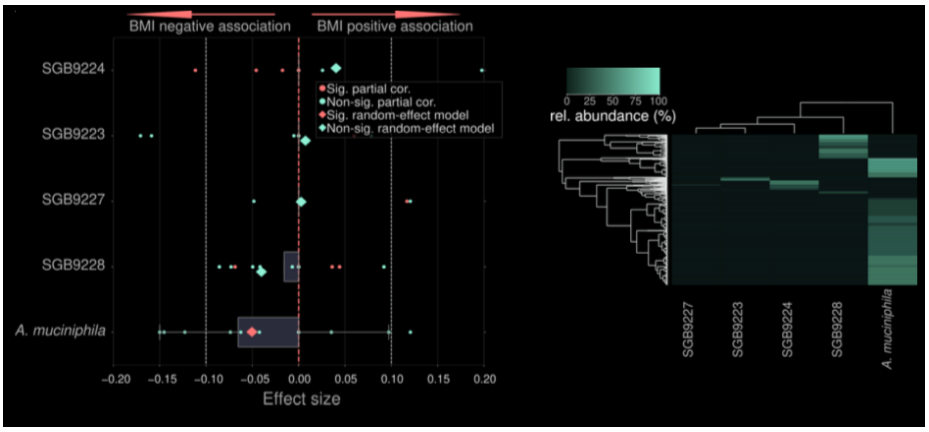


Figure 8.20: Distinct associations and co-exclusion for the *Akkermansia* (candidate) species

They also identified putative bacteriophages with spacer hits from *Akkermansia* candidate species and found that viral detectability correlates strongly with the relative abundance of the *Akkermansia* candidate species, suggesting an intimate ecological interplay. Analysing *A. muciniphila* subspecies, they determined that these are host-specific: some are found only in mice and some only in humans.

8.3.4 An example of eukaryotic microorganism: *Blastocystis*

In this work they developed a pipeline to detect *Blastocystis* subtypes and applied it on 12 large datasets composed of 1689 subjects of different geographic origin, disease status and lifestyle. They confirmed that *Blastocystis* is a component of the healthy gut microbiome and found a higher prevalence in non-westernized individuals. Moreover, they were able to construct and functionally profile 43 new *Blastocystis* genomes. A strong association of specific microbial communities with *Blastocystis* was confirmed by the high predictability of the microorganism colonization based on the species-level composition of the microbiome.

8.3.5 Bacteriophages profiling

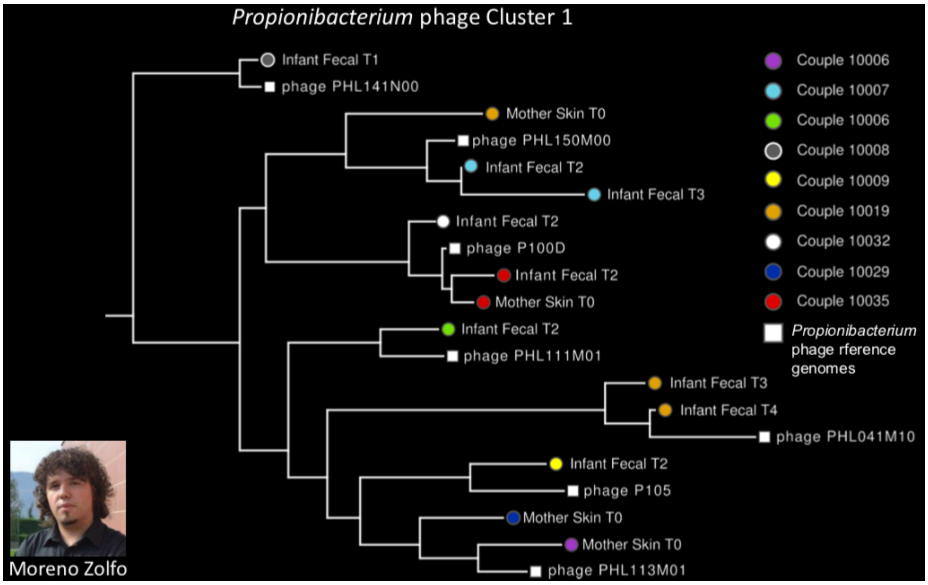


Figure 8.21: *Propionibacterium* phage Cluster 1

8.3.6 HUMAnN2: Functional profiling

The task of mapping nucleotide sequences to proteins that can be done by blastX on a smaller scale is important but computationally challenging. Multiple bacteria can be responsible for the same function in the microbiome. Thanks to this redundancy, functions are more conserved than bacterial prevalence: the abundance of a bacteria can vary but its function remains efficient because another microbe supplies it. For functional profiling, the idea is to reduce the reference dataset by mapping the genes only across proteins that we know are present in the bacteria found in the sample. Then the remaining unclassified reads can be mapped to a comprehensive protein database 8.22.

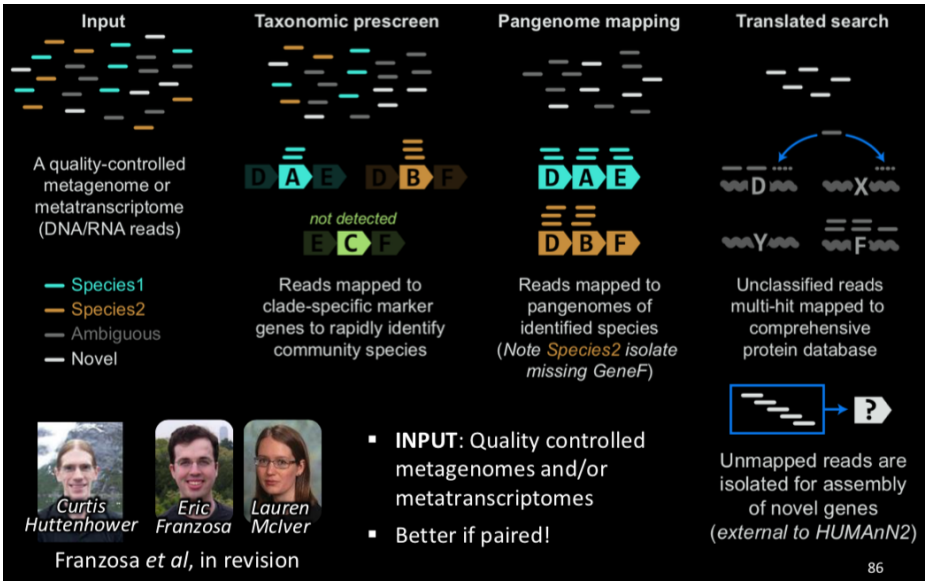


Figure 8.22: HUMAnN2 workflow

Chapter 9

Staphylococcus aureus

9.1 Introduction

Staphylococcus aureus is a gram positive (it has a peptidoglican layer into the cell wall) and it is a facultative anaerobe bacterium. This is very for its epidemiology because it is able to colonize nostrils where there is oxygen, but it is also able to colonize organs that are inside the body. It is one of the main players in common food poisoning. It is also involved in the menstrual toxic shock syndrome. It plays a key role, also, in other serious disease, like osteomyelitis (infection of the bones) or sepsis (systemic infections). It is a common skin colonizer and for this reason 25% of the people probably have it, but it is also the cause of very bad skin infection. The main reason for *S. aureus* to be tricky to treat because of his immune evasion strategies.

9.1.1 Immune evasion strategies

S. aureus has two different main strategies that used in order to stop the immune system of the host from getting rid of it.

9.1.1.1 Prevention of the engagement of the host immune system

It prevents the engagement of the host immune system, so it is not be recognized by the host immune system. That is done by for different classes proteins that are present on the surface of *S. aureus* able to hide it from the host immune system:

- Adhesins bind complement factors to inhibit complement activation cascade.
- Leukocidins are instead a number of toxins that are selectively killing the adapting immune cells, so killing those immune cells that would be able to kill *S. aureus*.
- Immunoglobulin binding proteins bind and immobilize IgGs, so they cannot start the cascade of activation of the immune systems.
- Proteases that cleave the immunoglobulin that are responsible for the activation of the host immune system.

9.1.1.2 Overactivation of the non-specific immune system

It is able to trigger a lot of inflammation to cytokine release, a non-specific reaction of the body, and also to facilitate invasion of the so-called non-professional phagocytes, the neutrophiles. There is the production of autholysins, that facilitate invasion of non-professional phagocytes, and super-antigens, that activate T cells and trigger the cytokine release. This could be an advantage to *S. aureus* because when a neutrophiles phagocytes a bacterium or any pathogen can happen that the microbiome is uptaken by the neutrophile, is killed through degranulation and ROS production. The neutrophile then undergoes apoptosis and it is removed by macrophages and so there is a resolution of the infection. However, *S. aureus* stops the apoptosis of the neutrophile that can uptake it and is able to divide inside it and be screened by the immune system of the host and then, with the leukocidins, it can cause some holes in the neutrophile and in turn the release of its content outside causing an extra inflammation.

9.2 Antibiotic resistance in *S. aureus*

After the discovery in the $1940s$ of penicilin S. Aureus developed quickly the protein penicilinase, becoming resistant to that antibiotic. Also a few years after the discovery of methicillin, a semisynthetic penicillinase-resistance β-lactam antibiotic some cases of S. aureus resistant to that antibiotic or MRSA were found in an hospital in the UK and then quickly spread in many countries.

9.2.1 Methicillin-resistant *S. aureus* (MRSA)

S. aureus is resistant to *beta*-lactam and is also able to acquire other resistances, event to last resource antibiotics, like vancomycin, linezolid, daptomycin. Because *S. aureus* is a gram-positive bacterium, the peptidoglycan layer of the cell wall is extremely important for the correct assembly of the cell membrane and the β-lactam antibiotics have the ability to act as an analog substrate causing an impaired transpeptidation of the peptidoglycan and the creation of a defective cell wall during cell division.

1. In absence of β-lactam antibiotics, there is the normal cell-wall biosynthesis.
2. In presence of β-lactam antibiotics, the binding of the antibiotic to the PBP active site causes an inability to transpeptidate the peptidoglycan. Because of this the peptidoglycan layer of the cell wall cannot be produced and the bacterium dies during cell division.
3. In presence of the β-lactam antibiotics, but with mutated PBP (PBP2a, or MecA), β-lactam is not able to bind the modified PBP, the peptidoglycan can be normally transpeptidated and *S. aureus* can produce the cell wall and proliferate.

9.2.2 Coding of methicillin resistance

The resistance to methicillin, but more in general to β-lactam is encoded on the mobile genetic element staphylococcal chromosome cassette *mec* (SCC*mec*). SCC*mec* are mobile

genetic elements that are wide spread across staphylococci genome. They commonly carries genes that might confer some increased fitness for specific environments. This type of mobile genetic elements can be easily integrated into the genome and also can easily excises from it. That means that is really easy for a MSSA to integrate the mobile genetic element in case of a strong selective pressure that might be given by the presence of antibiotic. When the antibiotic is not more present, it is easy for MSSA to excise the mobile genetic element and return to the basic state of methicillin asset of the *S. aureus*. This genetic element is not maintained inside the cell but it is integrated into the it and it is excised and released outside when is no longer needed.

9.2.3 Methicillin-resistant *S. aureus* (MRSA)

S. aureus is not well recognized for the problems that it causes. There are 80 thousands new patients per year in US that have invasive infections and the mortality rate is at 20%. Hospitalized patients, immune compromised patients or patients with conditions like cystic fibrosis are very exposed to *S. aureus*. This is why, in 2017, the World Health Organization insert *S. aureus* resistant to meticillin and partially resistant to vancomycin or completely as the fifth highest priority bacterium for the research and the development of new antibiotics. An article of three years ago estimated about 5 millions deaths associated with bacterial antibiotic resistances (not *S. aureus* only) in 2019. The World Health Organization has estimated that by 2025 there will be 10 millions deaths per years because of antibiotic resistance.

9.2.4 *S.aureus* worldwide

There are lot of studies that focused on MRSA or *S. aureus* infections, but the problem is that there are quite biases toward specifically lineages. Lineages are specific strains or a group of strains that are known to be particularly hyper-virulent or resistant or affecting a specific population. Because of this research is restricted to only a part of the pool of infections that S. Aureus can cause. There is a great variability in MRSA that can change epidemiology:

- 60s-70s. There were lots of hospitals associated infections, so people go to the hospital, get a surgery and get *S. aureus*. Also, nowadays *S. aureus* is a key player in post-surgery infections.
- 80s-90s. The community start talking about community-associated *S. aureus* infections and methicillin resistance of *S. aureus* because the study starting focusing on the dissemination, also in healthy people.
- 2000s. There was great studies on livestock-associated MRSA. Zoonotic infections are pretty relevant, but treatment with antibiotics cause resistances in that community. Livestock diseases and resistances are serious consequences.

9.3 Whole genome epidemiology, characterization, and phylogenetic reconstruction of S. *aureus* strains in a pediatric hospitals

The work presented here represent a full pipeline to perform a survey of the general population of S. aureus in a specific place.

Figure 9.1

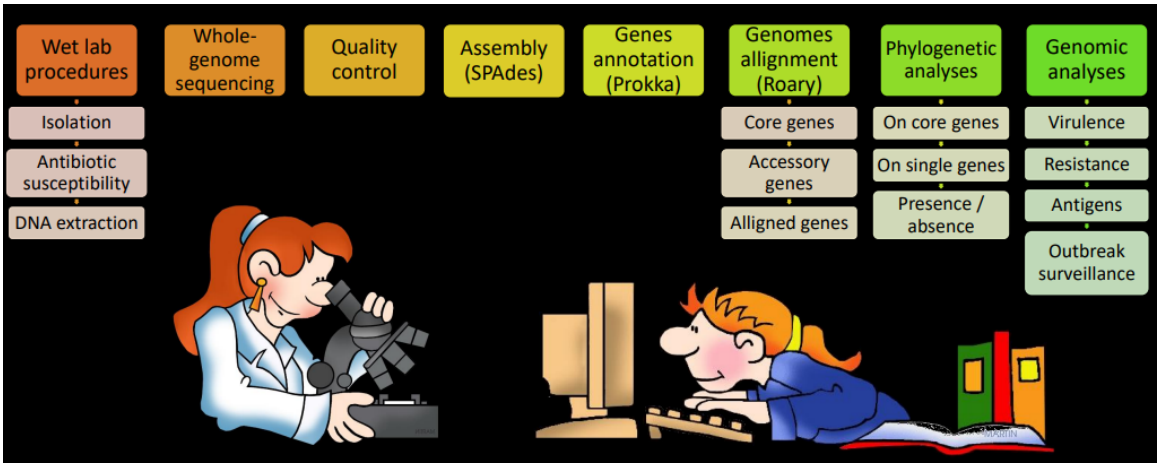


Figure 9.2: Wet and dry lab workflow

9.3.1 Methods

The authors worked with 11 operative units isolated from each other and they started with 234 S. aureus isolates performing antibiotic susceptibility test in vitro and whole-genome sequencing. After that, they selected 184 of the isolates with $N50 > 50k$ to consider only high-quality genomes. Patients were not selected on the bases of their status and came from different department of the hospital. 135 single patients were selected so that all single patients were isolated. The genome considered had:

- Average nr of contigs = 51 (12-138)
- Average N50 = 206k (50k-970k)

Also the samples derived from different tissues like sputum, nasal, pharyngeal and lesion swabs among the others.

9.3.2 Typing methods

Usually the typing of S. *aureus* is based on four different typing methods created for wet-lab work:

1. Multilocus sequence typing (MLST).
2. *S. aureus* proteins A (spa). It is one of the major determinant of virulence on *S. aureus*.
3. Staphylococcal cassette chromosome *mec* (SSC*mec*). Looked at the presence or absence of it.
4. Proton-Valentine Leukocydin (PVL). Looked at the presence or absence of it.

9.3.2.1 MLST

Multilocus sequence typing consisted in:

1. Characterizing isolates by sequencing fragments of house-keeping genes.
2. Each isolate is characterized by its allelic profile at the house keeping loci determining its sequence type (ST).
3. Based on multilocus enzyme electrophoresis (MLEE). This step is not usually done because of its inefficiency and the high number of PCR required. Allelic profiling is usually much faster since only sequencing and comparison are required.

MLST take fragments of house-keeping genes and look at their sequence to assign an allelic profile. Looking at each of the 7 house-keeping genes of S. aureus a sequence of numbers created by the composition of its sequence is compared to a database to determine the lineage of the sample.

9.3.2.2 Spa-typing

Spa-typing consists in:

1. Single locus DNA-sequencing of the repeated region of the Staphylococcus protein gene (spa).
2. This in turn is used to further discriminate STs.
3. Repeats are assigned a numerical code and the spa-type is deduced from the order of repeats.

The Staphylococcus protein A typing looks at the differences in the repeat sequence internal in the Staphylococcus protein A gene. This gene has a part of repeats that can be in different positions. The order of the repeats is the determinant of the spa-type. To perform this analysis 135 PCRs were necessary.

9.3.2.3 SCC*mec*

During the sequencing of the region containing *mecA* it was found a distinct mobile genetic element named the staphylococcal chromosome cassette *mec* (SCC*mec*).

The mec gene complex is formed by mecA, mecI and mecR. The first is responsible for antibiotic resistance, the second is the inhibitor of the first and the third is the inhibitor of the second. While mecA is always present, the other two contribute to the typing of the cassette. Another part important for the typing is the cassette chromosome recombinase ccr. Moreover there are 3 joining regions responsible for other resistance that can be used for subtyping. There are specific inverted and directed repeats containing the insertion site

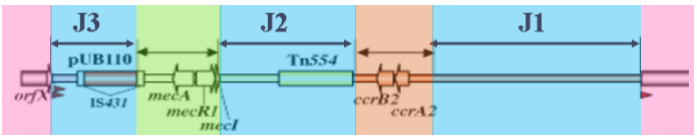


Figure 9.3

recognized by the ccr-encoded recombinase. There are 11 types of cassette with different sized and gene content.. SCCmec perform typing on the mec and ccr gene complez and subtyping on the J region. To perform this analysis 675 PCRs were performed.

9.3.2.4 PVL

PVL is a regarded as an important indicator of *S. aureus* virulence. The PVL factor is encoded in a prophage that secretes two toxins LukS-PV and LukF-PV. It is a good indicator of how invasive an infections can be. PVL:

- Assemble in the membrane of host white blood cells, monocytes, and macrophages.
- Form a ring with a central pore through which immune-cell contents leak.
- The ring also acts as a superantigen and we have the suppression of adaptive immune response.

To characterize this gene 135 PCRs were made. In total to get the typing of the community 1890 PCRs have been performed.

9.3.3 The cohort

1464 core genes that are present in at least 99% of the strains and some trees that are quite specific, highly observed through MLST typing. A lot of information about sample type, operative unit, PVL presence, SCC*mec* type and presence of virulence genes about the samples. To catch virulence genes a list of genes known in literature to cause virulence in S. aureus was compared with the analyzed genoms. In the cohort they found:

- 28 STs (14 CCs)
- 41 *spa*-types
- 4 SCC*mec* types
- 27.4% PVL+

There are 8373 genes in total that are present in at least 1 isolates.

9.3.4 Co-presence of local, global, animal-associated and hypervirulent clones

1. Highly virulent STs:
    - USA300 ST8-SCC*mec*IV PVL+.
    - ST239 HA-MRSA had high transmissibility and quickly develops into bacteria.

Lineage	Epidemic clone	# of isolates	Notes
ST228-I	South German / Italian	16	Regional
ST22-IV	E-MRSA-15	13	
ST1-IV	USA400 / CA-MRSA-7	11 (5 var)	
ST5-IV	USA800	10	Paediatric
ST8-IV (PVL-)	USA500 / E-MRSA-2/6	4	
ST8-IV (PVL+)	USA300	2	Highly virulent
ST45-IV	Berlin / USA600	2	Highly virulent (bacteremia)
ST152-V	Balkan clone	2	Highly virulent
ST5-II	USA100 / New York / Japan	1	
ST247-I	Iberian / E-MRSA-5 clone	1	
ST5-I	E-MRSA-3	1	

Figure 9.4

- ST45-SCC*mec*IV causes bacteremia and was an MSSA isolated from infectious diseases.
- ST121 MSSA obtained from lesion swabs.
- ST152-SCC*mec*V that caused severe infections.

2. Livestock-associated MRSA (LA-MRSA), like ST398 and ST97 that causes mastitis, but also found in children not exposed to livestock. Which were increased in non at-risk individuals.
3. They found also the ST395 lineage that is very peculiar and it is usually not found in humans. It was find in a child that was not at risk. It is particularly interesting because he can exchange DNA with the coagulase negative *S. aureus* (CoNS) because it has modified wall teichoic (WTA).

9.3.5 Genomic signature of chronic versus acute *S. aureus* infections

Correlation of specific departments with virulent STs and PVL+:

- CF and intensive care correlated with PVL + ST121
- first aid and diseases correlated with PVL
- infectious diseases correlated with ST45

Correlation of sample tipes with virulent STs and PVL+:

- lesion swabs correlated with MSSA, ST121 and PVL. Here they found virulent and not resistant infections, and that make sense because it is an acute infections.
- Lung isolates (bronchoaspiration material + sputum) correlated with ST128, PVL and MSSA.

They observed that chronic infections are usually less virulent, while normally acute infections are more virulent.

9.3.6 Variability in SSC*mec* cassettes

They took cassettes and they performed genomic analysis to specifically check the genes that are present. When focusing on a specific part of the genome, presence or absence of genes and their functions can be analysed. They found two cassettes harboring extra genes that were resistant at antibiotics (kanamycin, bleomycin and trimethoprim).

9.3.7 Diversity of virulence factors and antigens

Specific class of genes can be considered, like immune evasion genes. Some immune invasion genes are present in almost all of the isolates. The resistant to vancomycin is never present. But is present the resistant to penicillin. There is only one sample (first aid) positive for Edin (epidermal cell differentiation inhibitor) that causes translocation into the bloodstream. One USA300 isolate positive for the arginine catabolic mobile element (ACME) and that increased the pathogenicity of the strain. There was an higher prevalence of:

- resistance genes in chronic infections.
- virulence genes in acute infections.

Toxins primarily responsible for *S. aureus* skin manifestations (Eta and Etb) were strongly associated with ST121 and lesions. Staphylococcal enterotoxins are present in infections department, but are not also present in CF and intensive care departments.

9.3.8 Virulence factors with available vaccines targets

There are no vaccine approved now for *S. aureus*. They took a list of genes that encode for the target of these vaccines and they checked for the prevalence in the community of isolates and also the presence of SNPs or deletions. There are different strategies for the development of vaccines:

1. highly prevalent genes, but with an high degree of variability/indels.
2. most virulent or lethal infections.
3. non-virulence genes, more prevalent and conserved than virulence ones.

They mentioned antigens are part of the formulation of putative vaccines tested in published clinical trial, with only a few of them getting favourable results and no approved vaccine to date.

9.3.9 Phylogenetic of specific STs highlights the aggressive spread of a novel independently acquired ST1 clone

They investigated the hypothesis that some of the prevalent STs could be hospital-associated clones:

1. isolates sharing the same ST, SCC*mec*, and *spa* types, were not monophyletic subtrees when considering external genomes for the same STs. There is and independent acquisition of the clones and there is no evidence of

transmission in the hospital.

2. two ST121 MSSA isolates from two patients in the same time window were found to be almost identical.
3. all but two isolates belonged to the same sub-lineage, typed as SCC*mec*IV t127 PVL-.

It was used a Bayesian phylogenetic modelling approach integrating in the analysis all the ST1 reference genomes publicly available and the two ST1 SCC*mec*V:

1. Meyer's clone emerged 6 to 28 years ago as a specific branch of the ST1 tree (26-160 years old).
2. An isolate obtained in a recent study investigating the spread of a ST1 SCC*mec*IV t127 clone in Irish hospitals and carrying a virulence and resistance profile very close to the one of our cohort (differences in gene presence: 2/79 and 0/18 respectively) is phylogenetically rooted inside the Meyer's cluster.

ST1 SCC*mec*IV t127 is not specific of the Meyer's hospital, but might represent a newly arising community clone that is now spreading in the nosocomial environment of different countries.

9.3.10 Conclusions

With a whole genome sequencing approach we can:

1. Type and phylogenetically reconstruct a large *S. aureus* cohort.
   - Observe emerging / unexpected clones.
   - Spot potential outbreaks.
2. Test for antibiotic resistances and virulence factors.
3. Discover variants in genes of interest or of unknown relevant genes.
4. Track strain transmission among patients.

Chapter 10

Ancient DNA

10.1 Iceman's history

Thirty years ago Erika and Helmut Simon, while hiking on the Italian - Austrian border, stumbled over a skeleton half sticking in the ice. The hut owner called the police and the dead body was recovered (believing it was a recent death, no archeological retrieval). Luckily, many photos were taking during the excavation, allowing to keep track of more details. After some days, an archeologist was called and established that the body had to be at least 4000 years old. The mummy is now conserved in Bolzano under specific conditions; the body, which is around $50kg$, is losing weight over time, so spraying and temperature/humidity modifications are required to ensure a correct conservation.

10.1.1 Iceman's equipment

The Iceman's equipment is composed by tools and clothes. 61 tattoos were identified all over the body with the help of multispectral photography, mainly lines and crosses. Most likely this was a treatment for pain, stitch tattoos. It was initially thought that the Iceman died for the cold. Interestingly, in the stomach traces of hop hornbeam pollen were retrieved. This particular pollen is only present when hop hornbeam is in bloom, so Ötzi died in late spring. In 2001, by examining radiological analyses, it was possible to discover an arrowhead: the Iceman was murdered, a new 3D reconstruction proved that the arrow injury was lethal. The Iceman tried to pull out the arrow, but the arrowhead remained in the shoulder – maybe he wanted to hide that he was injured.

10.1.2 Ötzi life

Ötzi lived approx. 3.300 BC and died at 40-50 years old. He suffered from mild osteoarthritis of new joints (probably due to mountain hiking), intestinal parasites and arteriosclerosis. The isotopic analysis confirmed that he grew and lived in Northern Italy. A lot of results came back in the last years, but many samples were lost. In 2007 the Institute for Mummies and the Iceman was built, focus on molecular analysis (especially DNA) and preservation.

10.2 Ancient DNA analysis

The first study ancient DNA study was performed on a museum quagga specimen, an extinct member of the horse family. Svante Pääbo is famous for his research on Neanderthal genome: by analyzing genomes retrieved from mummies through histology and DNA analysis he was able to perform mitochondrial DNA extraction and cloning. It was later seen that a long mitochondrial DNA stretch was a modern human DNA contamination, bringing awareness in the field of the risk of contamination.

10.2.1 Impact of NGS

Thanks to NGS and PCR, it is now possible to exactly discriminate ancient DNA from modern DNA (by paying attention to potential overlaps). After an organism dies, post-mortem degradation begins. We have differences in pH, enzyme degradation, obtaining highly fragmented biomolecules. If the conditions are right e.g. permafrost, we can still use 7000 years old DNA. Modern DNA can also degrade, we can find small fragments. Through new techniques we can perform shotgun sequencing (all DNA) in silico and distinguish modern from ancient on the basis of damage patterns. Cytosines tend to deaminate over time, the C to T change can be effectively measured. In ancient DNA, we have a higher substitution frequency (figure 10.1).

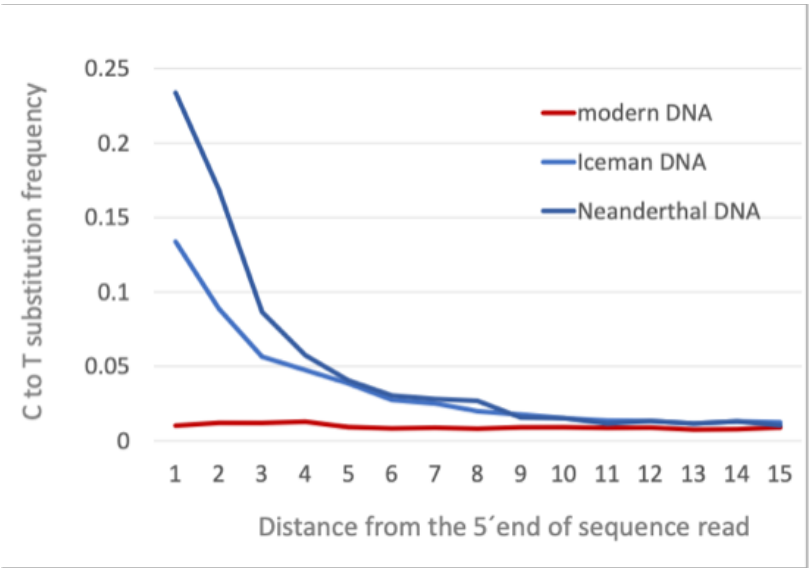


Figure 10.1: C to T substitution frequency in modern, Iceman and Neanderthal DNA

Most of the ancient DNA is highly contaminated, and much depends on the conservation environment (we should protect samples from external DNA). For ancient DNA analysis, we extract the sample, build a DNA library and either perform shotgun sequencing or enrich for a specific site of interest.

10.3 Iceman's genome

While performing a metagenomic analysis of the Iceman's shoelace leather, it was possible to identify the particular cattle species used for producing the leather, gaining insights into the animal sources of the Iceman's Copper Age clothing. The genetic analysis of the Iceman human genome started in 1994.

10.3.1 Genome analysis

After some unsuccessful attempts, it was possible to sequence the entire mitochondrial genome and nuclear genome. The Iceman genome project managed to over most of the genome ($> 85\%$) with average coverage of $7x$ [SOLiD 4 sequencer, old machine]. It was able to confirm authenticity and had a low level of contamination. Thanks to the project, it was possible to obtain data on pigmentation, SNP association indicates a high probability for brown eyes, light skin color and brown/blond hair, and clinical diseases: Ötzi was lactose intolerant, increased risk factor for cardiovascular disease (high common in other mummies, old lifestyle led to a higher risk).

10.3.2 Population genomics

Population genomics: project comparing a group of Early European farmers to modern European. Most of the skeleton samples clustered close to the Sardinian. This is probably due to an island effect.

10.4 Iceman's metagenome

The Iceman's metagenome was compared to 16s databases, finding a high abundance of Clostridia and Pseudomonas. The third genus was Treponema, where most reads were assigned to T. denticola, belonging to a group of pathogens responsible for periodontitis (chronic inflammatory disease of the periodontium). These oral flora bacteria are also involved in plaque formation.

10.4.1 Plaque analysis

BY applying PCR-based detection of red complex bacteria in Iceman's mouth samples, it was possible to identify Treponema denticolsa and Porphyromonas gingivalis. Ancient dental calculus, present on top of the teeth, is an interesting plaque formation, which could be useful for many archeological studies (e.g. food traces, pathogens, human proteins). A diacronic sample set (different location and time origin) was used to study the dental calculus. A high abundance of the archea Methanovrevibacter was previously linked to ancient calculus. The study aimed at analyzing the diversity of archea in ancient calculus over time. All calculus samples contained reads assigned to the genus Methanovrevibacter and some had a very high abundance. Only one strain was linked to the known Methanovrevibacter, while the others unravelled new branches in the phylogeny – one linked to copper age, others to early middle age. There was a clear tendency in time for the diverse presence of different strains. Through functional analysis, it was seen that all the strains are involved in methanogenesis. To summarize, this study was able to perform de novo

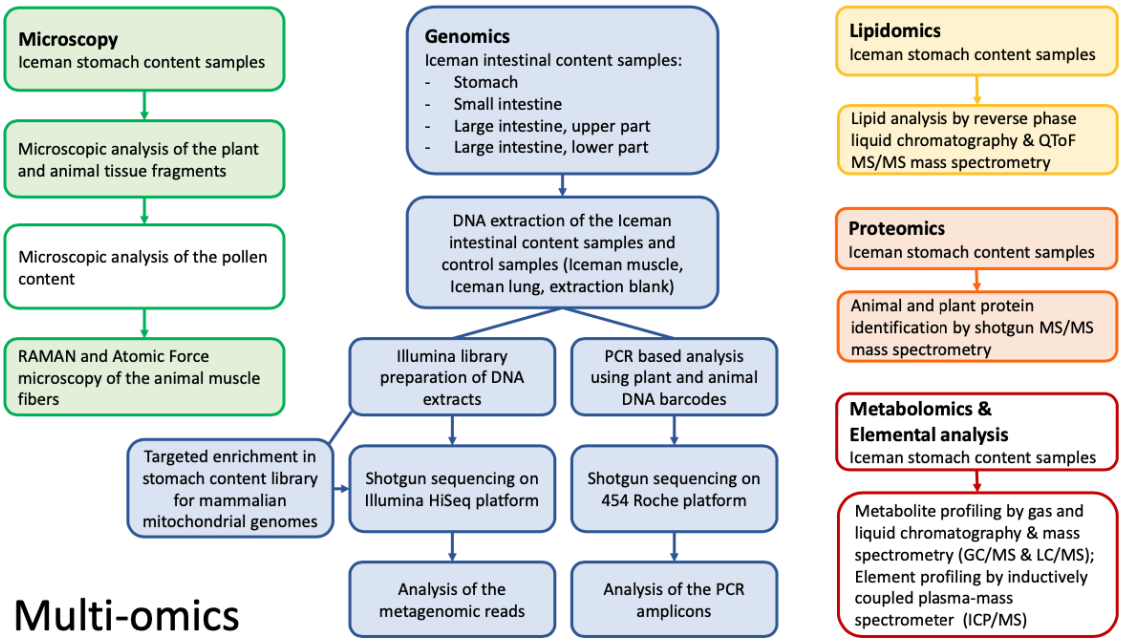


Figure 10.2: Multi-omics analysis of the Iceman samples

assembly on ancient dental calculus. There is a possible decline in Methanovrevibacter over time, but we lack modern calculus for comparison to fully assess it.

10.4.2 Iceman's stomach

The completely filled stomach was not in its original anatomical position, it moved to the lung region. By sampling the stomach, it was possible to reconstruct diet and perform Helicobacter pylori diagnostics (figure 10.2.

The Iceman had an omnivorous diet: plants, muscle fibers with different length and breadth. The DNA-based analysis showed a high presence of background proteobacteria and firmicutes. Eukaryotic reads linked to diet are only 0.7% to 0.2%. Through them, it was possible to trace capra ibex reads and cerbus elaphus reads (last meals for the Iceman). Thanks to protein-based analysis, an animal diet was confirmed. For what concerns the plant diet, chloroplast genomes were traced – Poaceae reads traced to a monococcum wheat (cultivated) and Dennstaedtiaceae to Pteridium aquilinum, toxic herb (also consumed in Asia). The high hydrophobic property of the stomach content were studied in order to discriminate if fat derives from animals or plants, or dairy products. Initial histology highlighted the presence of adipose tissue, meat related. A lipid analysis was performed with HPLC-Chip/QTOF-MS, which discriminates fat types according to chain length; Ötzi TAG profile points towards ruminant meat and/or dairy products. By performing a comparative TAG profiling for meat and fat from ibex and red deer, milk and cheese from sheep and goat, it was seen that the Iceman had a high level of saturated fatty acids.

10.4.2.1 Summary

Ötzi had an omnivorous diet including wild meat, cereals and bracken. DNA- and Protein-based analysis indicate that the meal contained tissue of two different wild animals (red deer, ibex). High amount of lipids: TAG, DAG, cholesterol esters, phosphatidyl-cholines, sphingomyelins TAG profile matches that of mammalian tissues, especially fat from wild ruminants. No indication of dairy products consumption - Iceman could have been a shepherd or a migratory herder. RAMAN and AFM analysis of meat fibers reveal no indications for meat preparation with fire (cooked, fried/ roasted). Meat was most presumably air dried or smoked over the fire (under 60°C).

10.5 Helicobacter pylori

Helicobacter pylori is a gram-negative bacterium, which infects $> 50\%$ of humans globally. It causes chronic gastritis, ulcers, adenocarcinoma. It is transmitted predominantly vertically (intra-familial) and is usually acquired early in childhood causing a lifelong infection. It is used as a marker for tracing human demographic events and recombines frequently.

10.5.1 Presence in the Iceman

H.pylori in the Iceman's stomach: usually H. pylori is studied through the histological analysis of the stomach mucosa (staining). Through PCR-based diagnostics, $\frac{2}{20}$ samples resulted to be PCR positive, but seemed to be a false positive. Through metagenomics, a decent number of unambiguous H. pylori DNA were retrieved in all gastrointestinal tract contents. The reads display a damage pattern typical of the ancient DNA. The distribution pattern of the DNA is coherent with the location, as most reads mapped to the stomach corpus. In order to overcome the background noise, RNA-based enrichment for H. pylori was performed. In this way we reduce diversity e.g. specific Iceman features, but we obtain a high enrichment effect (216-fold enrichment). The Iceman H. pylori reads map on 90% of the reference genome (only one strain, 26695) with a 20-fold average coverage. Virulence factors indicate how severe an infection can be. CagA (cytotoxine-associated gene A) and VacA (vacuolating cytotoxin A) virulence factors were present in the Iceman, they provide a high probability of developing a severe disease form. To check whether the Iceman suffered from the disease, we would need to perform an histological exam (infeasible). Some proteins linked to infection were upregulated, suggesting the presence of the infection.

10.5.2 Genetic diversity of H. pylori

Studies on genetic variation showed that H. pylori diversity is highly structured. Looking at SNPs and flanking sites, we can sense the differences between European, Asian, African H. pylori strains. We can see a geographic pattern. Iceman's strain belonged to a prehistoric European branch of hpAsia2 that is different from the modern hpAsia2 population from northern India. The distribution of hpAsia2 suggests it was widespread across Eurasia during the paleolithic and subsequently replaced by hpEurope. Despite the constant Middle Eastern immigration starting 10,000 BP (LGM) into Europe 5300 years ago, the

replacement of ancient European hpAsia2 strains (like Ötzi's) was not yet complete. These migrations were continuous, and occurred much more recently than previously thought.

10.5.3 Taxonomic profiling of the Iceman's microbiome

The mummification was not fully finished in Ötzi's corpse, the body was dehydrated and frozen and enzymes could not work. This is the main cause of high presence of Pseudomona and Clostridiae, involved in post-mortem processes (proteolytic breakdown of proteins, hydrolysis of fat and carbohydrates). The taxonomic profiling plus filtering indicates the presence of endogenous ancient gut bacteria.

10.5.4 Coprolite samples analysis

Ancient coprolite and gastric content samples were collected to get insights into the community composition of gut bacteria. A study focused on Mexican coprolite samples for caves, pre-colombian (660 to 1430 AD), established that H. pylori is less present in these samples with respect to the Iceman. The human DNA was linked to different sex and to genetic variation. The taxonomic composition shows high similarity to modern gut communities.

10.6 Prevotella copri

Prevotella copri is a frequently observed inhabitant of the human gut. It aids in glucose tolerance but is responsible for rheumatoid arthritis and is abundant in HIV patients. To extend the knowledge on its population structure and genetic diversity, a huge collection of modern gut samples from different sources was collected and analyzed. A novel mapping based on a contig binning approach was applied along with a strict quality control. Four different clades were identified with some country-specific sub-types. A high prevalence and co-presence of multiple P. copri complex clades is typical in individuals from non-Westernized populations. There is a considerably larger P. copri functional potential in non-Westernized populations e.g. distinct carbohydrate metabolism repertoires in the four clades. P. copri diversity in ancient human gut contents resembles that of non-Westernised populations. Clocking-based approach and dating methods can be used to see when the splits of certain clades happend e.g. humankind in Africa, then migration events carried Prevotella clades. All this links to an interesting observation: there are indications for a loss of diversity throughout time, our ancestors had more diversity.