Metabook

Leidy-Alejandra G. Molano

Update on January 10, 2025

Contents

List of Figures

List of Definitions

LIST OF DEFINITIONS

List of terms

bulge	Two short alternative directed paths between the same vertices of the de Bruijn graph. They are originated from sequencing errors, minor variations between sequences, or repetitive regions. 45
ecosystem	Community or group of organisms that live and interact with each other in a specific environment. 39
habitat	Refers to externalities such as the physical space and chemical environment that allow and organism to exist, including contributions from other members of the microbial community. 49
microbiota	The words "micro" and "biota" are also of Ancient Greek origin. It is a combination of "Micro" ($\mu\iota\kappa\rho o\zeta$, small), with the term "biota" ($\beta\iota o\tau\alpha$), which means the living organisms of an ecosystem or a particular area. 39
niche	Refers to the activity of an organism and the functional role that each member plays in the community. Interactions of the member both with one another and with the habitat drive the emergent organization of the community as a whole. 54

List of terms

Acronyms

CVD cardiovascular disease. 49

T2DM Type-2 Diabetes Mellitus. 49

Part I

Genomes

Chapter 1

Human Genome

1.1 Introduction

- 3.2 billion base pairs
- Haploid (n, gametes): 22 autosomal chromosomes + 1 sexual (X or Y).
- Diploid (2n, zygote): 46 autosomal chromosomes + 2 sexual (XX, XY). Every gene present in autosomes is present in two copies in the zygote. Then individuals contain two genomes, one maternal and one paternal, which get mixed in the cell nucleus after the first mitotic division.
- The total number of protein-coding genes distributed on the 23 chromosomes of the human genome is estimated to be 20,412, slightly less than 20,470 genes of the *Caenorhabditis elegans* [Pen21] and only the double of one strain of *Ktedonobacter racemifer*, with 11,453 protein-coding genes [HOH24].

1.2 Chromosomes

Chromosomes are the basic morphological division of the human genome.The number of genes on each human chromosome varies widely, from 2058 genes on chr1 to only 71 genes on Y chr. The density of genes on chromosomes also varies widely. For instance, chromosome 19 is smaller than chromosome 13, but contains almost four times more genes than the latter (chromosome 19 is the second in decreasing order of gene content, just behind the chromosome 1). The three autosomes with the fewest genes are chromosome 13 (327 genes), chr 18 (270 genes), and chr 21 (234 genes). It is thus no accident that the only autosomal human trisomies compatible with the survival of the fetus till birth are trisomies 13, 18 and 21!

There is apparently no specific reason why humans have 46 chromosomes in somatic cells. Our closest primate, the chimpanzee (*Pan troglodytes*) has 48 chromosomes. In the evolution of primates, two acrocentric chromosomes from the chimpanzee underwent centric fusion to form human chromosome 2, hence the reduction of chromosome number to 46. In contrast, the mouse (*Mus musculus*) has 56 chromosomes. The *Lysandra atlantica* butterfly has 446 chromosomes in diploid cells, while *Lysandra golga* has 268 and *Lysandra nivescens* has 82! In fact, there seems to be no correlation between the number of chromosomes or the size of the total genome or the biological complexity of the species. Both seem to vary at random. Thus, everything suggests that the chromosomes may be only physical frameworks that allow the realization of mitosis and meiosis in sexual species.

The chromosomes seem to behave functionally as "packages" of genes. In general, the functioning of individual genes is not affected by their chromosomal position. For instance, there are individuals with balanced chromosomal translocations, in which chromosomes have exchanged segments without loss or net gain of genetic material—such individuals do not present any clinical manifestation of translocation, except perhaps

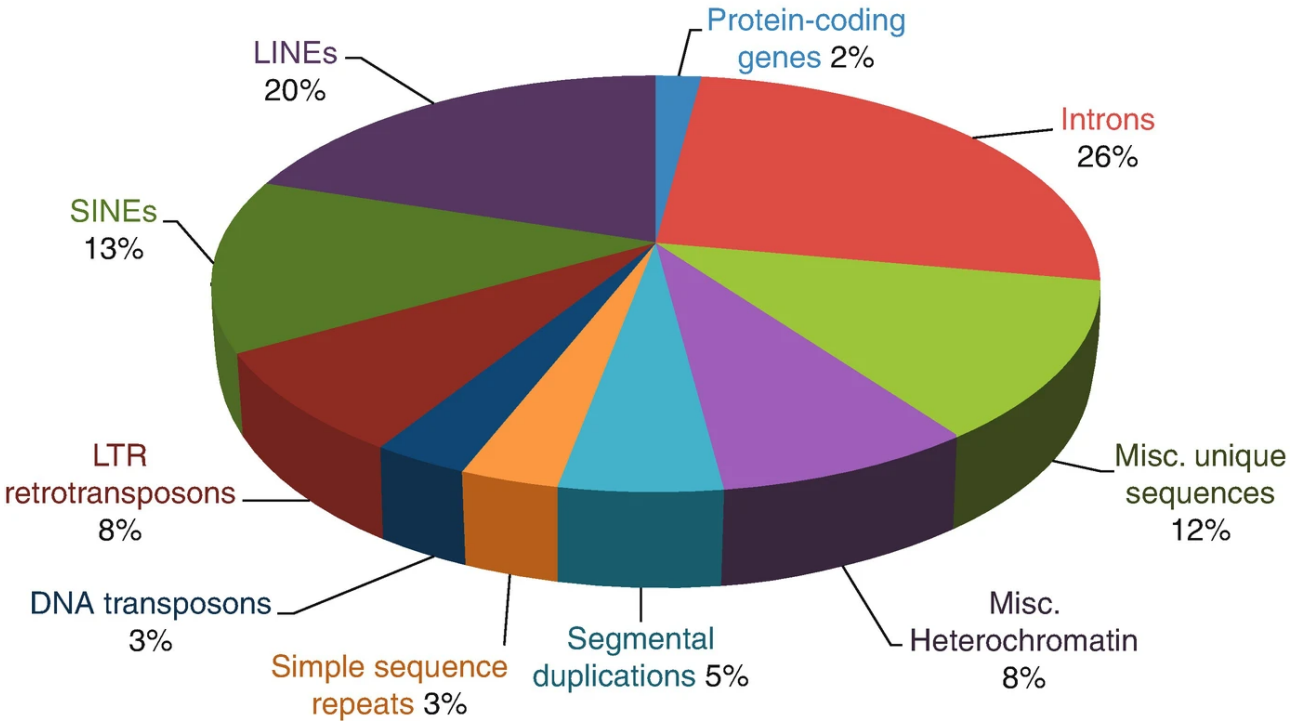

Figure 1.1: Composition of the human genome. Redrawn from a graph that was produced in 2014 by the NHS National Genetics and Genomics Education Centre. Borrowed from "An Overview of the Human Genome" [Pen21]

for reproductive difficulties, as some translocations may interfere with the production of gametes in meiosis, especially in the male.

Centromere. In chromosomes, DNA contains genes that are expressed according to the needs of the cell, but it also contains specialized sequences that are necessary for intrinsic functions of the chromosome itself. On one hand, chromosomes need to be properly aligned during cell division. This requires a centromere, a region where a pair of protein complexes, called kinetochores, binds just before the start of cell division. Microtubules are responsible initially for positioning the chromosomes correctly in the metaphase and then for pulling the individualized chromosomes to opposite poles of the mitotic spindle. The DNA sequences in the centromeres are very different in different organisms. In mammalian chromosomes, centromeric DNA is a heterochromatic region, with no informational content, dominated by repetitive DNA sequences that often extend monotonously by mega DNA bases.

Telomere. At the ends of chromosomes, there are specialized structures called telomeres, which are necessary for maintaining chromosomal integrity. If a telomere is lost after a break in a chromosome, the resulting chromosomal end is unstable and tends to merge with the broken ends of the other chromosomes, or even be degraded. In vertebrate telomeres the DNA consists of multiple copies in tandem of the oligonucleotide TTAGGG, sequence at which certain telomeric proteins bind. The repetitive units of the telomeres decrease in number with every division of the DNA. As the enzyme needed to regenerate telomeres (telomerase) is not available in normal somatic cells, telomeres are a kind of biological clock that records our age.

1.3 Coding and Non-coding DNA

The vast majority of genes are in the chromosomes of the nucleus; a few are also found in mitochondrial DNA.

Human vs Chimpanzee. Remarkable similarities of known human and chimpanzee protein sequences

initially led to the suggestion that significant differences might be primarily in gene and protein expression, rather than protein structure. Further analysis of alignable non-coding sequences affirmed this ∽1% difference. However, the subsequent identification of non-alignable sequences that were due to segmental deletions and duplications has shown that the overall difference between the two genomes is actually ∽4%.

Less than 2% of the human genome corresponds to protein-coding genes (Figure 1.1). The functional role of the remaining 98%, apart from repetitive sequences (constitutive heterochromatin) that appear to have a structural role in the chromosome, is a matter of controversy.

1.4 Retroposons, Retrotransposons, and Retrovirus

Transposable elements can be separated into two major classes:

- **DNA transposons**. Constitute approximately 3% of the human genome (Figure 1.1; Figure 1.2), can excise themselves from the genome, move as DNA and insert themselves into new genomic sites. Although DNA transposons are currently not mobile in the human genome, they were apparently active during early primate evolution.
- **Retroposition elements**. i.e. retroposons, retrotransposons and endogenous retroviruses, duplicate through RNA intermediates that are reverse transcribed and inserted at new genomic locations. Together, they constitute more than 40% of the human genome.

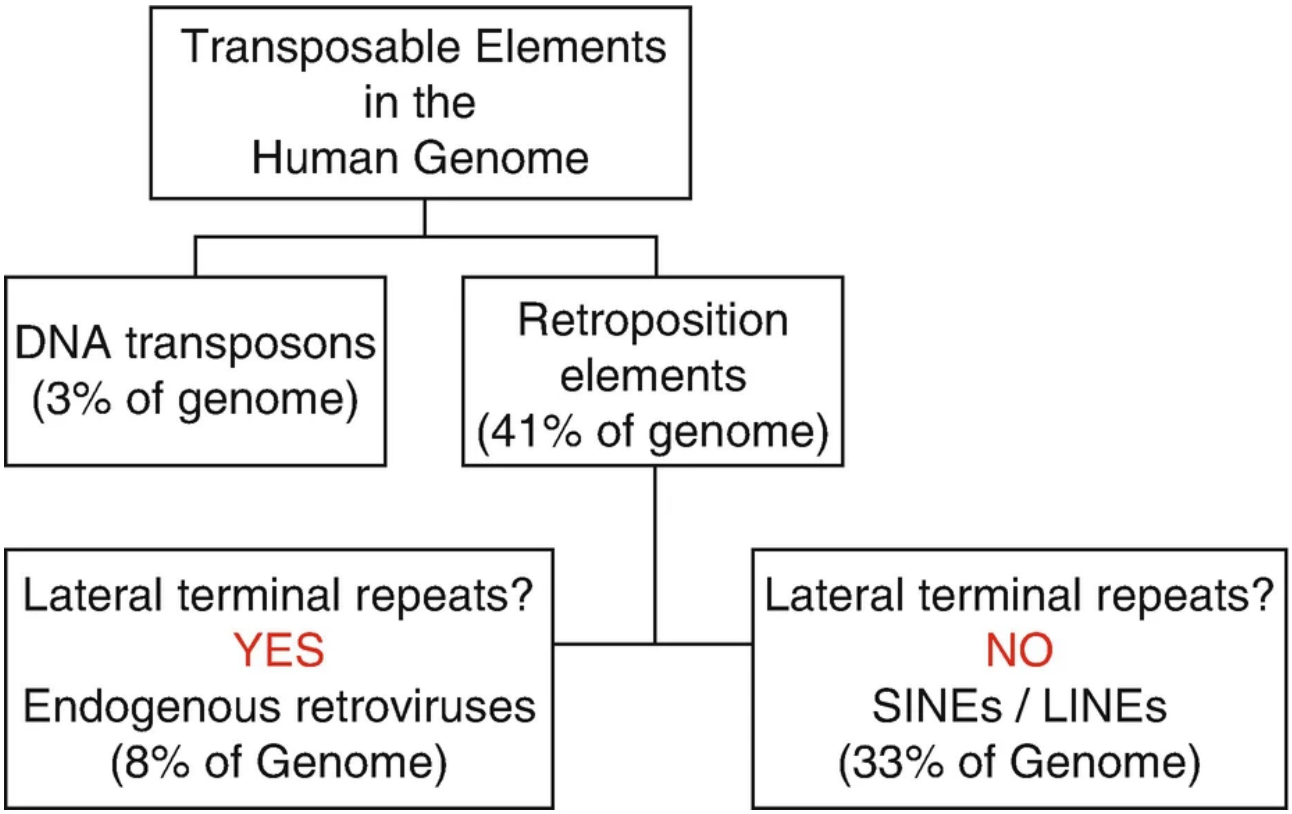

Figure 1.2: Classes of transposable elements in the human genome. Borrowed from "An Overview of the Human Genome" [Pen21]

Retroposons. Do not contain the gene for reverse transcriptase and thus are dependent on exogenous sources of the enzyme (mostly from Long Interspersed Nuclear Element s—LINEs) for retroposition. They

share similarity with genes transcribed by RNA polymerase III, the enzyme that transcribes genes into ribosomal RNA, tRNA and other small RNA molecules. An especially abundant group of retroposons in humans is the Alu family of SINEs (Short Interspersed Nuclear Elements), that basically represents a processed pseudogene of the Signal Recognition Particle (7SL) RNA. The Alu family of retroposons (thus called because they contain a site for digestion by the restriction enzyme AluI) makes up 13% of the human genome. Virtually all other mammalian SINEs differ from the human, being derived from tRNA genes.

Retrotransposons. In contrast, do code for reverse transcriptase and hence are capable of autonomous retrotranscription. They also contain a promoter for RNA polymerase II, which allows it to insert itself into random positions. In humans, LINEs, which altogether make up 20% of the human genome, are the main class of retrotransposons. Although the vast majority of human LINE-1 sequences are inactive molecular fossils, an estimated 80-100 copies per individual still retain the ability to mobilize and expand in numbers within the human genome, by cycles of transcription, retrotranscription and retroposition. Some of these active LINEs constitute insertional polymorphisms in the human species. LINEs and SINEs continue growing in numbers in all mammalian genomes, and thus are "genomic parasites", the ultimate "selfish genes".

Endogenous retroviruses. The class of retrotransposons that contain lateral terminal repeats (LTRs), which are evolutionarily related to the exogenous retrovirus group of RNA virus and will be the focus of this section (Figure 1.2). They constitute around 8% of the human genome! This is ironic, considering that at the very moment that I am writing this chapter humanity is being held ransom by the RNA virus SARS-CoV-2 that causes the serious disease COVID-19. Thus, if not only for its timeliness, I think that today any discussion of the structure and function of the human genome should include a discussion of these endogenous retroviruses. In special I want to evaluate the evidence for a conceivable anti-viral effect of these mostly defective and dormant endogenous retroviruses, which eons ago were exogenous, infected germ cells, endogenized and multiplied to become 8% of the human genome.

An endogenous retrovirus is generally called ERV or EVE (endogenous viral element). Although not one of the thousands of retrovirus-related sequences found in the human genome contains a complete set of intact retroviral genes or can express infectious virus, these sequences are nonetheless referred to as Human Endogenous Retroviruses (HERVs). More info in the section of An Overview of the Human Genome.

Chapter 2

Prokaryotic Genomes

2.1 Introduction

To be done. Check “Chapter 26 - Bacterial whole-genome determination and applications” [HOH24]

Part II

Mobile Genetic Elements

Mobile genetic elements (MGEs) are selfish genetic entities that are unable to self-replicate and rely on host cells and cellular machinery to propagate. They can move around within a genome or be transferred across species.

Chapter 3

Plasmids

Contents

3.1 Introduction

Plasmid are DNA molecules located outside of the chromosomal DNA, i.e. extrachromosomal. Their topology is frequently circular although linear plasmid also exists. They have been extensively studied in Bacteria, even though Archaea and Eukaryota also carry them.

Plasmids are generally associated with a host range, which can be broad or narrow. Incompatibility groups used from plasmid classification. Seems that their host range is determined by their ability to escape host defenses and the use of host's machinery.

Mainly, plasmids mobility has been associated to its conjugation system, although other mechanisms have been described, such as membrane vesicles (castañeda 2024). Importantly, not all plasmids have transfer and mobility functions.

Due to their mobility, plasmids are included in the category of mobile genetic elements, along with Integrative Conjugative Elements (ICEs, which integrate into the host genome and carry a functional conjugation system for inter-cellular transfer, $\backsim$18-500 kb in length) and phages (forming viral particles that infect a prokaryotic cell, replicating within it and are transferred between the cells via transduction, $\backsim$11-500 kb in length) (Khedkar 2022).

Insertion sequences (IS, elements carry only a transposase gene, $\backsim$2.5 kb in length) and, transposons (elements that carry transposase and dispensable cargo genes, $\backsim$5 kb in length) and integrons (gene acquisition systems that are immobile without other MGEs, several kb in length) depend on other MGEs for inter-cellular transfer.

3.2 Structure

- **Backbone**. Consists in two differentiated parts.
  - Essential genes that ensure vertical inheritance (replication, copy number, partitioning, stability).
  - Inessential genes that code for horizontal gene transfers.

- **Genetic cargo**. Regions outside backbone that may contribute a phenotypic advantage to their hosts.

Plasmid stability depends on the balance between genetic burden and beneficial effect to the host (genetic cargo).

3.3 Replication

The fundamental characteristic that defines a plasmid is its ability to replicate autonomously. This independence allows the plasmid to present a copy number higher than the chromosome.

$$Plasmid\ copy\ number = \frac{\#\ plasmid}{\#\ chromosomal\ copies}$$

However, they also seem to replicate in step with the chromosome, doubling in number during the cell growth of their host, being vertically inherited from generation to generation. Plasmids use at least three distinct types of replication systems: rolling circle, theta, and linear replication. **Rolling circle** is generally confined to small and high copy number plasmids, whereas large and low copy number plasmids invariably use types of **theta** or linear replication systems.

Plasmids are replicons that transfer between cells via conjugation (6), up to 2.5 Mb in length. This independence from the chromosome defines them as genetic locus where genes may evolve faster than in the chromosome.

3.4 Toxin-Antitoxin systems

Many bacteria encode lethal proteins in their genome alongside antidotes that counteract their toxicity. These toxin-antitoxin (TA) systems are classified into different types according to the nature of the antitoxins and the mechanism of action of the toxins.

3.4.1 Hok/sok system

The hok/Sok system has been the most studied T1TA (RNA/RNA interacting systems). It was first discovered on *Escherichia coli* R1 plasmid where it acts by maintaining plasmid copies in a cell population through post-segregational killing of the plasmid-free cells.

- The Hok (host-killing) type I toxin is a small hydrophobic protein [52 amino acids (aa)] targeting the inner membrane and leading to cell death.
- The Sok (suppression of killing) antitoxin is an RNA that inhibits the production of Hok at the post-transcriptional level.
- The mok (modulation of killing), that overlaps with the hok coding sequence (CDS) and is required for hok translation. The translation of mok, rather than the Mok product, was shown to be important for proper hok regulation and expression. For simplicity, the mok_hok bicistronic mRNA will be referred to as the hok mRNA throughout the article "Profiling the intragenic toxicity determinants of toxin-antitoxin systems: revisiting hok/Sok regulation" [Le +22].

[TO DO]

- "Landscape of mobile genetic elements and their antibiotic resistance cargo in prokaryotic genomes" [Khe+22]
- "Toxins, Targets, and Triggers: An Overview of Toxin-Antitoxin Biology" [Har+18]

Part III

Transcriptomes

Chapter 4

Prokaryotic Transcriptomes

Contents

4.1 Introduction

Our perception of a bacterial transcriptome once used to be simple: mRNA, tRNA or rRNA genes are neatly arranged along the chromosome and expressed as distinct mono-cistronic or polycistronic transcripts. However, over the last two decades, new global methods have reported dense transcript patterns across the bacterial chromosome, and discovered a plethora of small regulatory RNAs (sRNAs) and antisense transcripts [SV14].

Prokaryotic mRNA are synthesized in the cytoplasm and do not require transport from the nucleus (Clark and Pazdernik, 2013). They also do not require processing and can begin translation immediately after the transcription is complete. Most mRNA contain a sequence at the 5' end of the mRNA prior to the AUG start codon, termed 5' untranslated region (5' UTR) (Meijer and Thomas, 2002) and a region following the stop codon, the 3' UTR. Most prokaryotic mRNA contain a sequence in the 5' UTR to position ribosomes for translation. This sequence is named after its discoverers as the Shine-Dalgarno sequence [GD16].

The Shine-Dalgarno sequence is present in nearly all prokaryotic mRNA; however, recent evidence has shown that there are prokaryotic mRNAs that lack a Shine-Dalgarno sequence in the 5' UTR (Londei, 2005) or lack a 5' UTR completely. These mRNAs appear to be more common in primitive prokaryotes, such as archaea, in which initiation of translation on leaderless transcripts is thought to be the evolutionary oldest mechanism. The mechanism of how these prokaryotes distinguish the start codon is not known [GD16].

In prokaryotic cells, a single mRNA may code for several proteins. Each message on the mRNA is contained in a single 'open reading frame' a sequence of codons bound by start and stop codons. There are no start or stop codons within the reading frame itself. The arrangement of messages in tandem along a single strand of mRNA allows the proteins (often called gene products) to be translated simultaneously; these gene products are often related in function. Because mRNAs are single stranded, some mRNA molecules are able to base-pair within themselves and can form secondary and tertiary three-dimensional structures. These structures can regulate the synthesis of polypeptides in the polycistronic mRNA. One example of this mechanism is MS2 bacteriophage (Kozak, 1983). The A protein is coded at the 5' end of the polycistronic message, but is needed in only small quantities. The 5' end of the mRNA is often blocked by tertiary folding of the mRNA allowing only limited translation of the A protein while allowing translation to occur at more accessible sites downstream from the A gene.

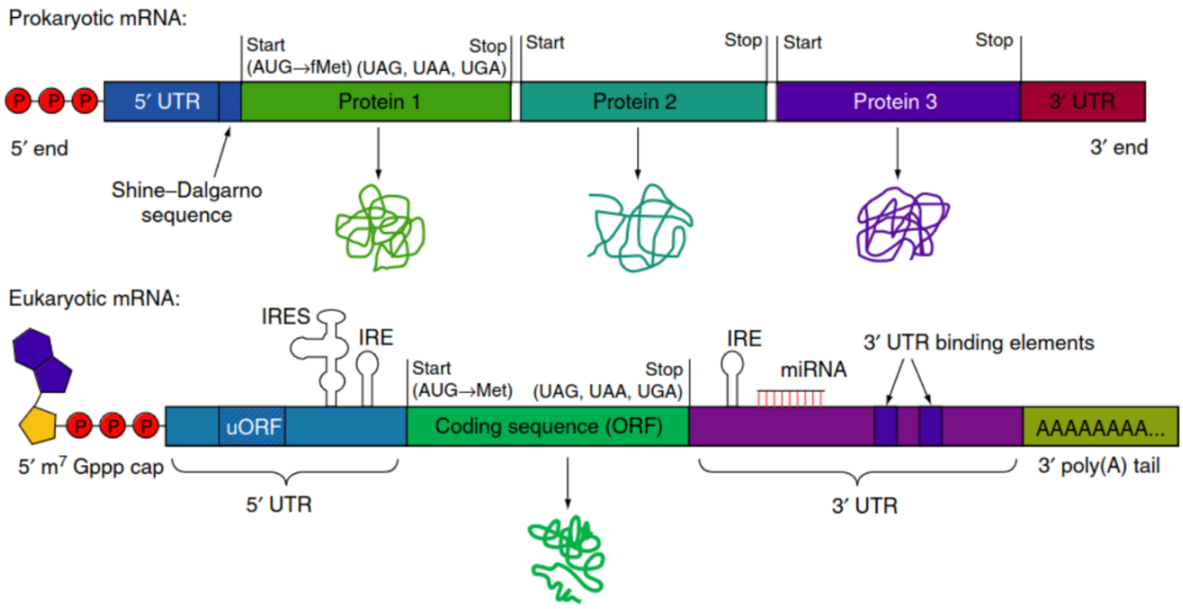

Figure 4.1: Schematic diagram of prokaryotic (top) and eukaryotic (bottom) mRNA. Bars indicate the relative length of the regions. Borrowed from "Messenger RNA (mRNA): The Link between DNA and Protein" [GD16].

4.2 Prokaryotic vs Eukaryotic mRNA

4.2.1 Common structures

The structures of both eukaryotic and prokaryotic genes involve several nested sequence elements. Each element has a specific function in the multi-step process of gene expression. The sequences and lengths of these elements vary, but the same general functions are present in most genes. Although DNA is a double-stranded molecule, typically only one of the strands encodes information that the RNA polymerase reads to produce protein-coding mRNA or non-coding RNA. This 'sense' or 'coding' strand, runs in the 5' to 3' direction where the numbers refer to the carbon atoms of the backbone's ribose sugar. The open reading frame (ORF) of a gene is therefore usually represented as an arrow indicating the direction in which the sense strand is read.

Regulatory sequences are located at the extremities of genes. These sequence regions can be next to the transcribed region (the promoter) or separated by many kilobases (enhancers and silencers). The promoter is located at the 5' end of the gene and is composed of a core promoter sequence and a proximal promoter sequence. The core promoter marks the start site for transcription by binding RNA polymerase and other proteins necessary for copying DNA to RNA. The proximal promoter region binds transcription factors that modify the affinity of the core promoter for RNA polymerase. Genes may be regulated by multiple enhancer and silencer sequences that further modify the activity of promoters by binding activator or repressor proteins. Enhancers and silencers may be distantly located from the gene, many thousands of base pairs away. The binding of different transcription factors, therefore, regulates the rate of transcription initiation at different times and in different cells.

Regulatory elements can overlap one another, with a section of DNA able to interact with many competing activators and repressors as well as RNA polymerase. For example, some repressor proteins can bind to the core promoter to prevent polymerase binding. For genes with multiple regulatory sequences, the rate of transcription is the product of all of the elements combined. Binding of activators and repressors to multiple regulatory sequences has a cooperative effect on transcription initiation.

Although all organisms use both transcriptional activators and repressors, eukaryotic genes are said to be 'default off', whereas prokaryotic genes are 'default on'. The core promoter of eukaryotic genes typically requires additional activation by promoter elements for expression to occur. The core promoter of prokaryotic genes, conversely, is sufficient for strong expression and is regulated by repressors.

An additional layer of regulation occurs for protein coding genes after the mRNA has been processed to prepare it for translation to protein. Only the region between the start and stop codons encodes the final protein product. The flanking untranslated regions (UTRs) contain further regulatory sequences. The 3' UTR contains a terminator sequence, which marks the endpoint for transcription and releases the RNA polymerase. The 5' UTR binds the ribosome, which translates the protein-coding region into a string of amino acids that fold to form the final protein product. In the case of genes for non-coding RNAs the RNA is not translated but instead folds to be directly functional.

4.2.2 Eukaryotes

The structure of eukaryotic genes includes features not found in prokaryotes. Most of these relate to post-transcriptional modification of pre-mRNAs to produce mature mRNA ready for translation into protein. Eukaryotic genes typically have more regulatory elements to control gene expression compared to prokaryotes. This is particularly true in multicellular eukaryotes, including humans, where gene expression varies widely among different tissues.

A key feature of the structure of eukaryotic genes is that their transcripts are typically subdivided into exon and intron regions. Exon regions are retained in the final mature mRNA molecule, whereas intron regions are excised during post-transcriptional processing. Indeed, the intron regions of a gene can be considerably longer than the exon regions. Once spliced together, the exons form a single continuous protein-coding region, and the splice boundaries are not detectable. Eukaryotic post-transcriptional processing also adds a 5' cap to the start of the mRNA and a poly-adenosine tail to the end of the mRNA. These additions stabilise the mRNA and direct its transport from the nucleus to the cytoplasm, although neither of these features are directly encoded in the structure of a gene.

4.2.3 Prokaryotes

The overall organisation of prokaryotic genes is markedly different from that of the eukaryotes. The most obvious difference is that prokaryotic ORFs are often grouped into a polycistronic operon under the control of a shared set of regulatory sequences. These ORFs are all transcribed onto the same mRNA and so are co-regulated and often serve related functions. Each ORF typically has its own ribosome binding site (RBS) so that ribosomes simultaneously translate ORFs on the same mRNA. Some operons also display translational coupling, where the translation rates of multiple ORFs within an operon are linked. This can occur when the ribosome remains attached at the end of an ORF and simply translocates along to the next without the need for a new RBS. Translational coupling is also observed when translation of an ORF affects the accessibility of the next RBS through changes in RNA secondary structure. Having multiple ORFs on a single mRNA is only possible in prokaryotes because their transcription and translation take place at the same time and in the same subcellular location [SL17].

The operator sequence next to the promoter is the main regulatory element in prokaryotes. Repressor proteins bound to the operator sequence physically obstruct the RNA polymerase enzyme, preventing transcription. Riboswitches are other important regulatory sequences commonly present in prokaryotic UTRs. These sequences switch between alternative secondary structures in the RNA depending on the concentrations of key metabolites. The secondary structures then either block or reveal important sequence regions such as RBSs. Introns are extremely rare in prokaryotes and therefore do not play a significant role in prokaryotic gene regulation [SL17].

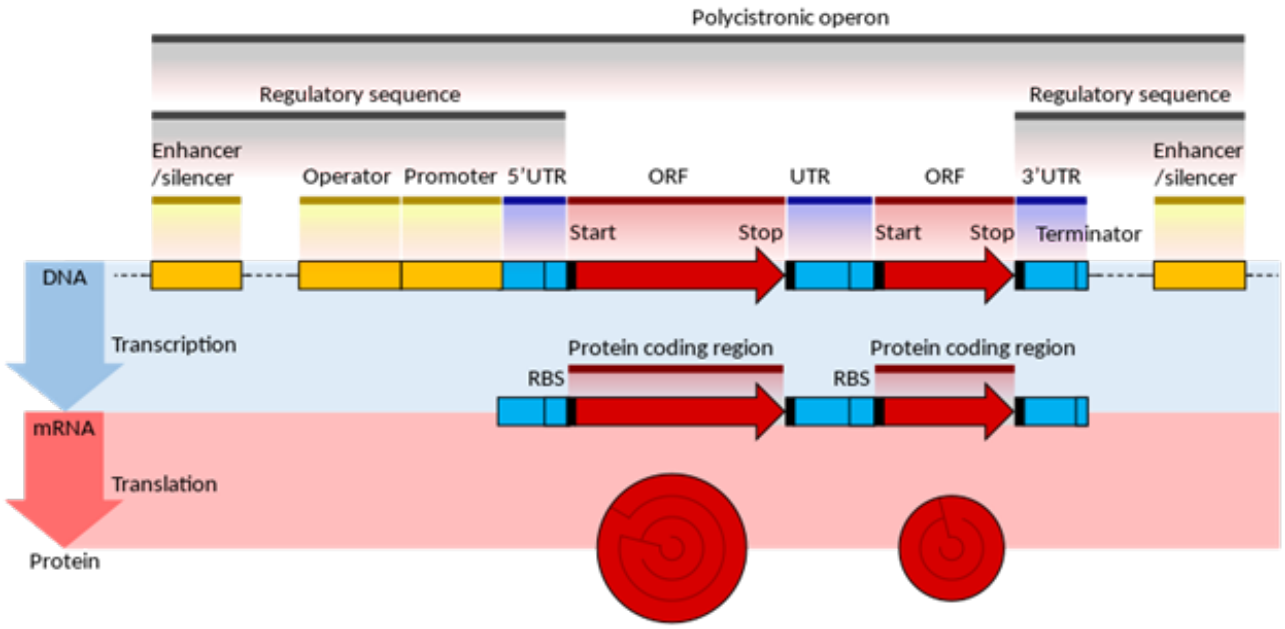

Figure 4.2: The structure of a eukaryotic protein-coding gene. Regulatory sequence controls when and where expression occurs for the protein coding region (red). Promoter and enhancer regions (yellow) regulate the transcription of the gene into a pre-mRNA which is modified to remove introns (light grey) and add a 5' cap and poly-A tail (dark grey). The mRNA 5' and 3' untranslated regions (blue) regulate translation into the final protein product. Borrowed from "Eukaryotic and prokaryotic gene structure" [SL17].

4.3 RNA-seq

While a major challenge for early bacterial RNA-seq experiments was the presence of highly abundant RNA species like rRNAs and tRNAs, which make up more than 95% of the RNA pool in a bacterial cell, this issue was overcome in eukaryotes by solely reverse-transcribing poly(A)-tailed mRNAs via oligo-d(T) priming during cDNA library preparation. Since poly(A)-tails represent a degradation signal in bacteria, several strategies for rRNA removal including oligonucleotide-based removal of rRNAs with magnetic beads or size fractionation using gel electrophoresis were employed [Bis+15].

In a typical RNA-seq experiment total RNA or a fraction thereof is first converted into cDNA in a reverse-transcription reaction, followed by PCR-based amplification of the library. Different library protocols are available, which are highly specific for the applied sequencing technique but can be subdivided into strand-specific and non-strand-specific protocols. Non-strand-specific protocols, for example, based on random hexamer priming and ligation of adapters to double-stranded cDNA have the drawback that they lose the information whether sequencing reads originate from the sense or the antisense strand. To overcome this problem, strand-specific protocols have been developed including direct sequencing of first strand cDNA, template switching PCR, RNA C to U conversion using bisulfite or second strand synthesis with dUTP followed by degradation after adapter ligation [Bis+15; SV14].

RNA-seq-based mapping of bacterial transcript boundaries enables a global elucidation of operon structures and facilitates annotation of untranslated regions (UTRs) of protein coding genes, which potentially contain gene regulatory elements. Additionally, it can improve genome annotation by providing extensive information on transcriptional start sites (TSS), untranslated regions (UTRs) of mRNA genes, and previously unknown open reading frames (ORFs) or sRNA genes [SV14].

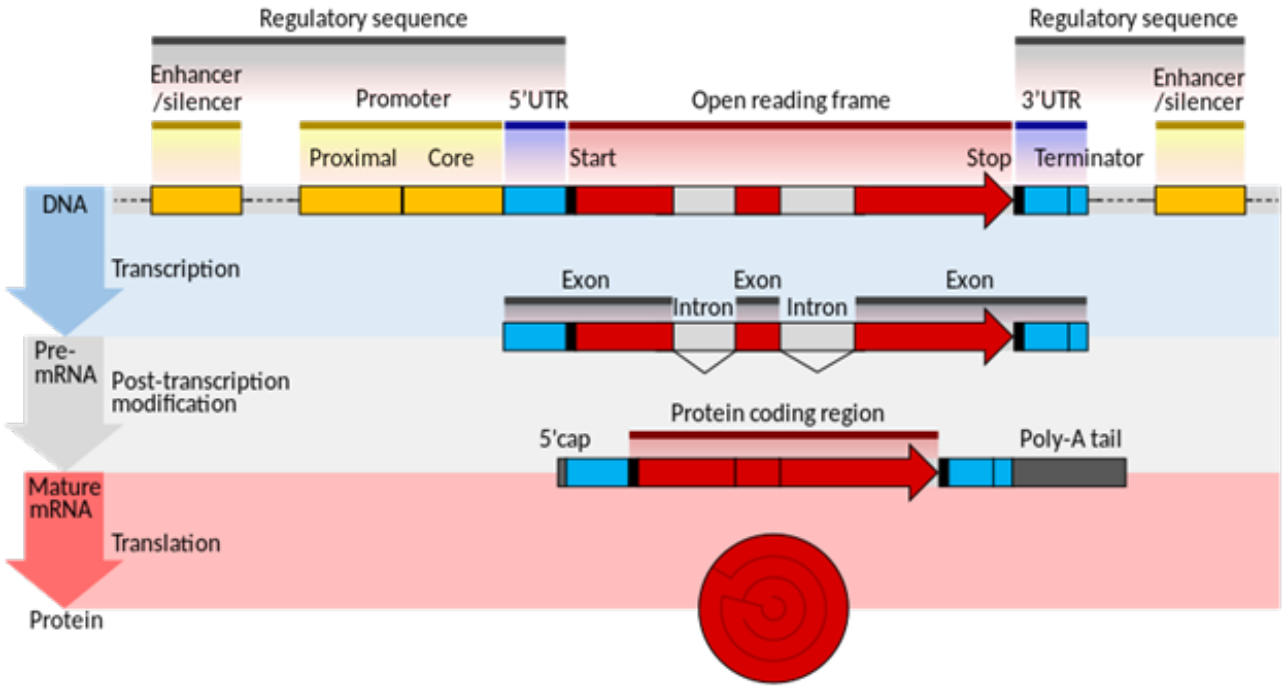

Figure 4.3: The structure of a prokaryotic operon of protein-coding genes. Regulatory sequence controls when expression occurs for the multiple protein coding regions (red). Promoter, operator and enhancer regions (yellow) regulate the transcription of the gene into an mRNA. The mRNA untranslated regions (blue) regulate translation into the final protein products. Borrowed from "Eukaryotic and prokaryotic gene structure" [SL17].

Prokaryotic vs Eukaryotic mRNA	
Prokaryotic mRNA is the RNA molecule which codes for prokaryotic proteins.	Eukaryotic mRNA is the RNA molecule which encodes for eukaryotic proteins.
Type	
Prokaryotic mRNA is polycistronic.	Eukaryotic mRNA is monocistronic.
Lifespan	
Prokaryotic mRNA has a shorter lifespan.	Eukaryotic mRNA has a comparatively a long lifespan.
Post Transcriptional Modifications	
Post transcriptional modifications are absent in Prokaryotic mRNA.	Post transcriptional modifications are present in eukaryotic mRNA

Figure 4.4: Table comparison. Borrowed from somewhereelse.

4.4 Differential RNA-seq

Differential RNA-seq (dRNA-seq) method allows for global annotation of all expressed transcriptional start sites (TSS) under the examined growth condition in an organism of interest in one sequencing experiment.

While it was originally developed to study the primary transcriptome of the major human pathogen Helicobacter pylori it has since been successfully applied for determination of TSS in a wide range of pro- and eukaryotic organisms. With ¿1900 unique TSS and at least one antisense TSS to 50% of all genes, the dRNA-seq approach revealed a very complex and compact transcriptional output from the small *H. pylori* genome and an unexpected number of $\geq$60 sRNA [SL17].

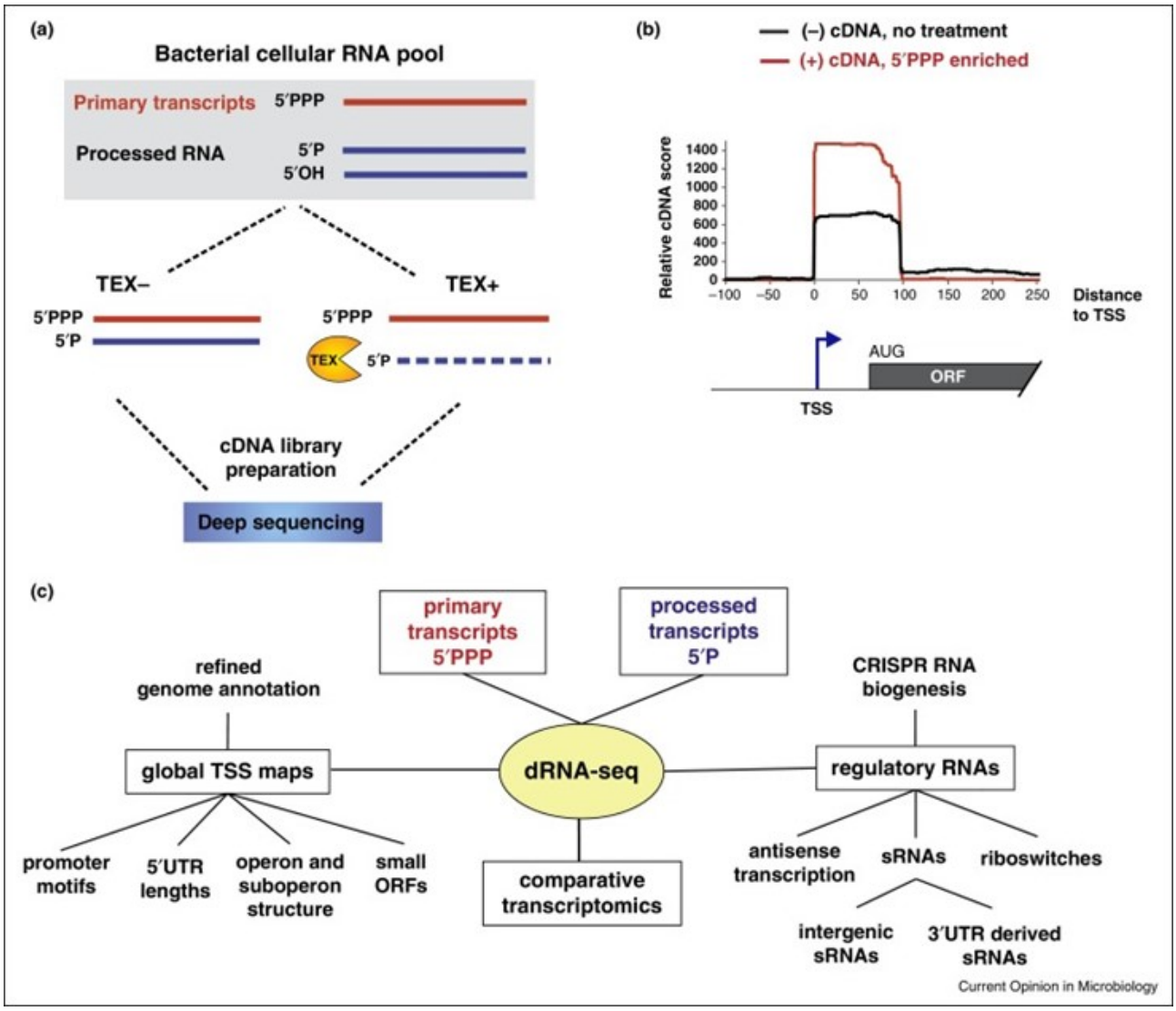

Figure 4.5: Rationale and output of the dRNA-seq approach. (a) Enrichment of primary transcripts using 5'-phosphate-dependent terminator exonuclease (TEX). The bacterial RNA pool consists of primary transcripts with a 5'PPP and processed RNAs with a 5'P or 5'OH. RNAs with a 5' OH group are not accessible for 5'-linker ligation during cDNA library constructions and, thus, will not be represented in the cDNA library. For the construction of dRNA-seq libraries, each RNA sample is split into two parts. One half remains untreated (TEX-), whereas the other half is treated with TEX which specifically degrades RNAs with a 5' P, and thereby enriches for primary transcripts with a 5'PPP in relative terms. Upon differential TEX treatment, both samples are converted into a cDNA library and analyzed by deep sequencing. (b) A dRNA-seq specific cDNA enrichment pattern can be observed at the primary 5' ends of genes. Treatment with TEX (red curve; (+) library) redistributes the cDNAs towards the nuclease-protected 5'-end, yielding a sawtooth-like profile with an elevated sharp 5' flank which can be used to annotate the TSS (blue arrow) of a gene of interest (grey bar). Note that dRNA-seq reads cluster towards a gene's 5' end if no fragmentation is used. (c) Schematic summary of information that can be gained from dRNA-seq to uncover transcriptome features and refine genome annotation. Borrowed from "Differential RNA-seq: the approach behind and the biological insight gained" [SV14].

Part IV

Metagenomics

Chapter 5

Metagenomics

Contents

5.1 Introduction to Microbiome

TO DO: Add One Health Concept Era and improve Figure 5.1.

Microbes are everywhere, almost every environment on earth is colonized by different types of organisms Figure 5.1.

So, microbes do not live isolated but instead live in different and complex ecosystem. We can distinguish three main categories of ecosystems: Host-Associated, Environmental, and Engineered.

- In the Host-Associated ecosystem the microbes live within a Host, that could be any animal that you think of but also humans, and they can play active roles in pathogen resistance, immune modulations, and food metabolism, among others.
- Also we have the environmental category, that includes the marine and soil ecosystems, being involved in carbon sequestration, pathogen resistance, nutrient absorption, and soil fertility.
- Finally, we can also classify the ecosystems as engineered all the artificial or human-created ecosystems where microbes can live. This includes the wastewater treatment in activated sludge…etc. They raise special interest in the biotechnology field.

Definition 5.1 (Microbiome)**.** The term microbiome includes both the microbiota (community of microorganisms) and their "theatre of activity", which includes structural elements such as proteins, lipids and theirs genomes, but also genetic mobile elements such as plasmid, the metabolites produced by the microbes, and the surrounding environmental conditions.

So, it's important to understand that, in contrast to the microbiota which can be studied separately, for example using a petri dish, the microbiome is not only composed of the microorganisms that live within, but also all this other elements ("Threatre of activity") Figure 5.2 that directly interact with each other and impact the function and type of microbiota present in this ecological niche.

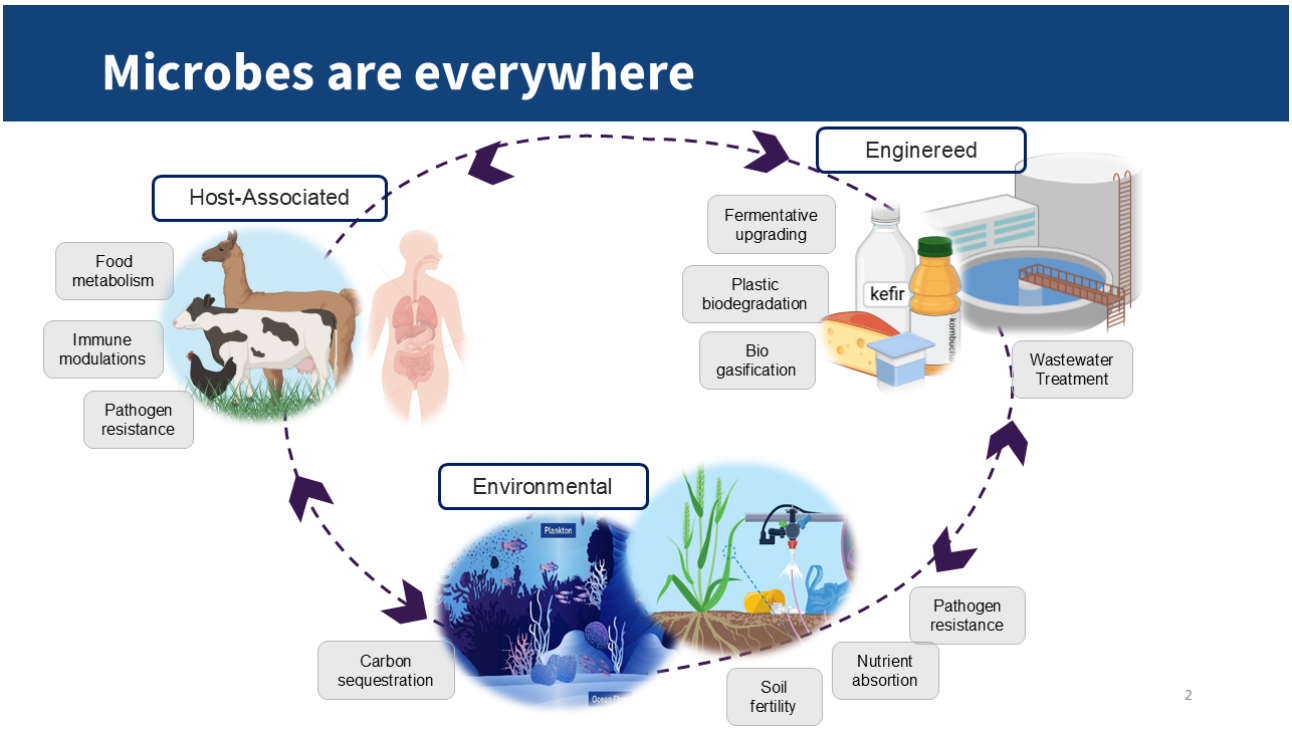

Figure 5.1: One Health concept, microbes are everywhere

You will see frequently that microbiome studies often focus on one or few specific groups of the microbiota, for example, only the bacteria part, using terms such as bacteriome. More unusual they can also focus on archaeome, or mycobiome, referring to archaeal and fungal communities, respectively. Lately some studies are also interested in the viral part of the microbiome, which can be referred to as virome. But also, they can focus on one of the elements of this "threatre of activity". When we focus on studying the genomes of the microbiome, which are called metagenomes, we are doing metagenomics. The same applies for the transcriptome, metatranscriptome - metratranscriptomics; metaproteome - metaproteomics; metabolome - metabolomics. Later on we will see some of these topics more in detail.

5.2 Taxonomical Classification

There is also a question of at what resolution each of the microbiome members should be studied. This resolution depends on the taxonomic level we are studying the microbiome. There are 9 major taxonomic levels or clades, with its corresponding abbreviation, ranging from very general (Domain) to specific (Strain), going to many other clades such as phylum, family or genus. There might also be intermediate clades, such as subphylum between phylum and class, and subspecies between species and strain.

Note that according to the International Code of Nomenclature of Prokaryotes, it is recommended to print scientific names by a different typeface, e.g., italic, or by some other device to distinguish them from the rest of the text. The name of a genus should be spelled without abbreviation the first time it is used. Later use of the name of the species previously cited usually has the name of the genus abbreviated, commonly to the first letter of the generic name. Example: Clostridioides difficile. -¿ C. difficile If, however, species are listed belonging to two or more genera which have the same initial letter, the generic name should be used in full, or initial two-letter or three-letter abbreviations should be used. Some subcommittees on taxonomy have recommended three-letter abbreviations to be used in such cases.

Here we have two samples, one the one side the taxonomic classification of C. difficile, bacteria known

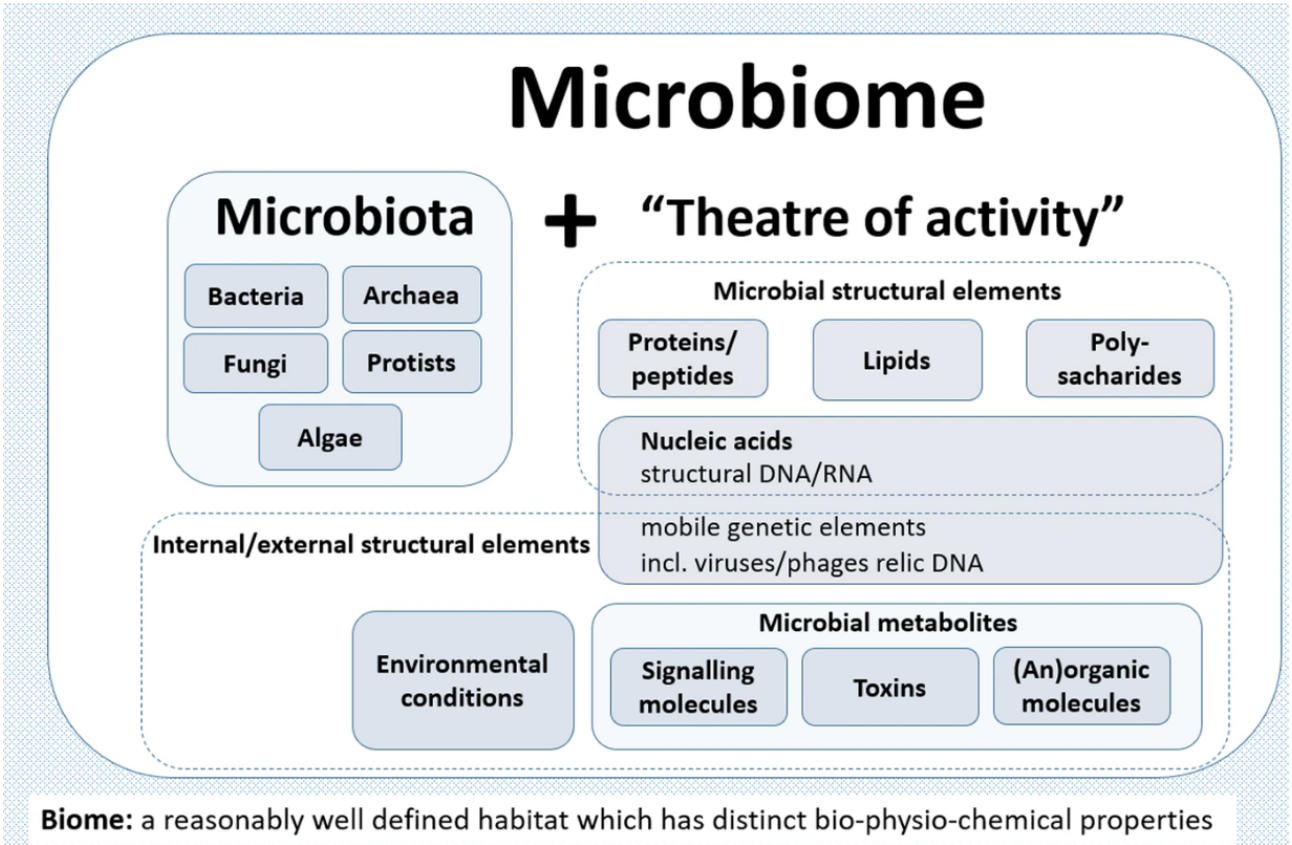

Figure 5.2: Borrowed from "Microbiome definition re-visited: old concepts and new challenges" [Ber+20]

to cause diarrheal infections, and on the other side L. crispatus, the taxonomic classification of one bacteria commonly found in the vaginal microbiome. You can see that for these organisms, all the taxonomic levels are complete, but that's the perfect scenario. The reality is that, in 2019, one study showed that 85/119 phyla have not had a single species described to date. So, you commonly will find microbes that are only taxonomic classified up to the genus or family level, having subsequent unknown clades. For this reason it is important to keep in mind that taxonomical classification is constantly changing.

Also, there might be discrepancies in the classification according to the methodology used to define each clade. For prokaryotes, two most widely used are NCBI taxonomy (which is the one actually used to these microbes) and GTDB (C. difficile is p__Bacillota_A). NCBI taxonomy is pretty useful because of its synonyms (example). Worth mentioning is the python package (ETE toolkit, v.4.) that allows access programmatically to both databases and retrieve the lineage of a desired microbe.

5.3 Quality Control

In this section, we will evaluate critical quality control metrics that are essential for ensuring the integrity and reliability of metagenomics datasets. These metrics help assess the overall quality of sequencing data and determine whether any technical artifacts or contamination could influence the downstream analysis. Below are the key metrics we will focus on:

igabase Pairs, Gbp) The number of reads per sample, expressed in Gbp, is a fundamental metric that indicates the amount of sequencing data generated for each sample. Sequencing depth is crucial for accurately detecting microbial species and other biological signals, as deeper sequencing provides a more comprehensive

representation of the microbial community. In the plot, we represent three stages of data processing, each with its own Gbp value:

Raw Data (Raw) This represents the total number of reads generated directly from sequencing, before any quality control or filtering.

After Filtering (Filt) This value reflects the number of reads retained after applying the tool fastp to remove low-quality sequences, adapters, and other artifacts. Filtering ensures that only high-quality reads are included in the analysis, improving accuracy.

contamination (Deconta) This shows the number of reads left after human DNA decontamination, where sequences that map to the human genome are removed. This step is particularly important in metagenomics studies to prevent human DNA contamination from masking microbial signals or skewing results.

ads to Human DNA This metric represents the proportion of reads that align to the human genome. A high percentage indicates significant contamination from human DNA, which is often unavoidable in samples like skin. However, too much human DNA can obscure microbial diversity, making it essential to identify and remove these sequences during the quality control process. Monitoring this metric helps ensure the decontamination step is effective.

ıtage of Duplication Duplication percentage refers to the proportion of reads that are duplicates of each other. High duplication rates can indicate over-sequencing or PCR amplification bias, which can distort microbial abundance estimates. Controlling for duplication ensures the data accurately reflects the diversity and relative abundance of species in the sample.

centage GC Content GC content is the percentage of guanine and cytosine nucleotides in the sequencing reads. This metric provides insight into the overall nucleotide composition of the dataset and can be used to identify potential biases or anomalies. For instance, extreme deviations in GC content might suggest contamination, poor sequencing quality, or species-specific biases in the dataset.

ge Adapter Content Adapter content refers to the proportion of sequences that still contain adapter sequences from the library preparation process. Adapter contamination can interfere with downstream analysis, so it is critical to ensure that adapters are effectively removed during the filtering process. A low percentage of adapter contamination indicates good-quality data.

omical profiling tool MetaPhlAn4 is a popular tool used to classify microbial taxa from metagenomic data. This metric shows the percentage of reads that could be assigned to specific taxa using MetaPhlAn4. A high percentage of assigned reads suggests successful identification of microbial species, while a low percentage may indicate poor sequencing quality, high contamination, or the presence of uncharacterized species in the dataset.

5.4 Microbial Diversity

The diversity definition according to the Cambridge dictionary is the fact of many different types of things or people being included in something. If we translate this term into the microbial context is the fact of many different microbes (or so called biological variability) included in something. Depending on this something, we will differentiate between alpha and beta diversity Figure 5.3.

5.4.1 Alpha diversity

Alpha diversity is a way to measure how many different types of microorganisms (species) are present in a sample and how evenly distributed they are. There are two main components:

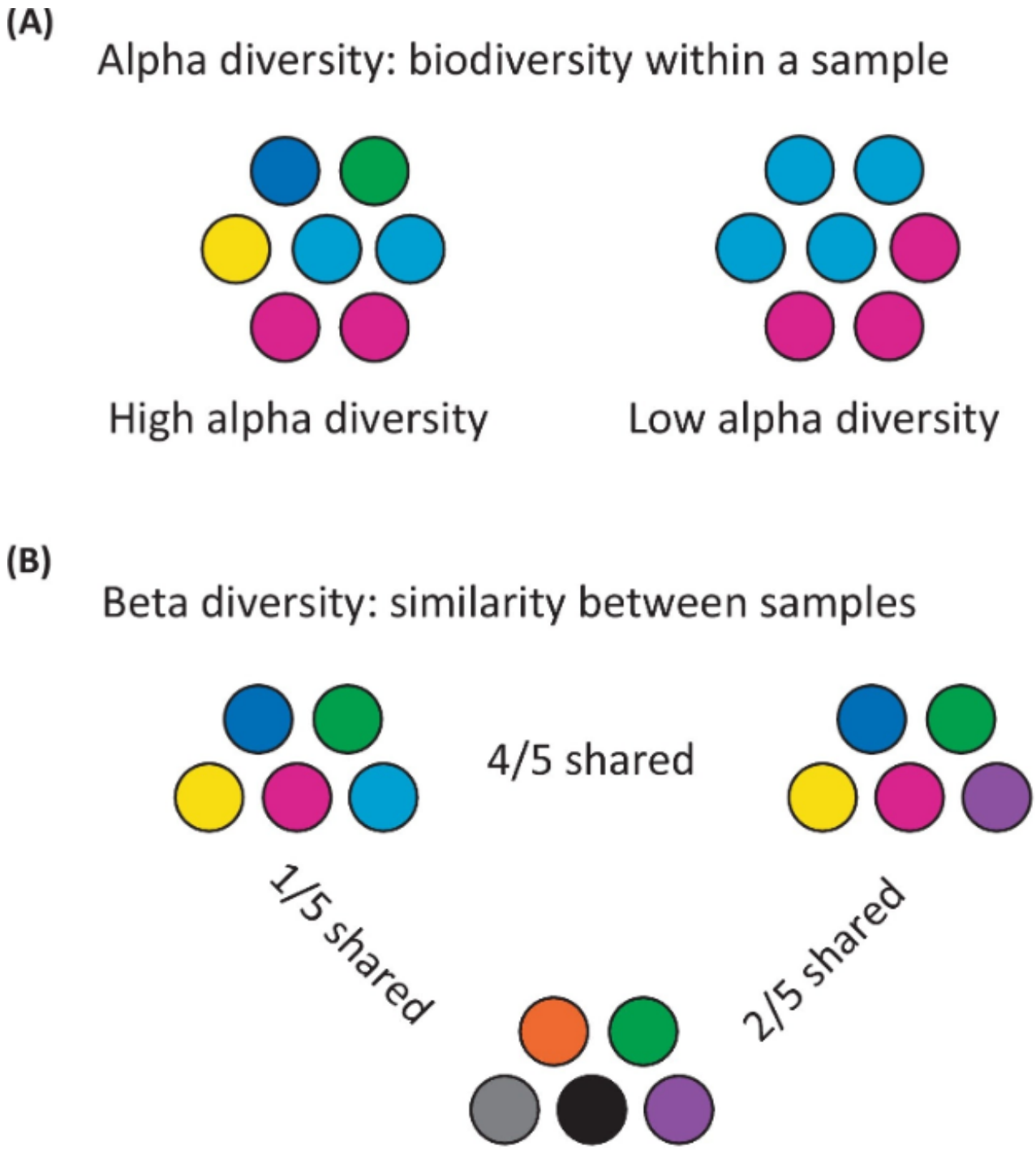

Figure 5.3: (A) Alpha diversity represents the biodiversity (species richness) within a specific community or individual sample. The sample on the left has high alpha diversity (five bacterial taxa), while the sample on the right has low alpha diversity (two bacterial taxa). (B) Beta diversity represents how similar one community or individual sample is to another. The samples on the left and right are similar to each other (4/5 shared bacterial taxa), while the sample on the left is not very similar to the sample in the middle (1/5 shared bacterial taxa). Borrowed from "Sex, Microbes, and Polycystic Ovary Syndrome" [Tha19].

Species Richness This refers to the number of different species present in a sample. For example, if you took a sample from the gut, species richness would tell you how many different species of bacteria are there.

Species Evenness This indicates how evenly the individuals are distributed among those species. A community where a few species dominate would have low evenness, whereas one where all species have similar abundance would have high evenness.

We will measure the two following indexes:

Chao1 index is used to estimate species richness, i.e., the total number of different species present in a sample. It is especially useful when the sample has many rare or unobserved species.

nnon-Weaver index) measures both species richness and evenness, i.e., how evenly the individuals are distributed among the different species in the sample. Chao1 is often used when there is concern about underestimating the true number of species due to sampling limitations, whereas Shannon is used to assess the overall diversity structure of a community, considering both richness and evenness cause is weighted by richness.

Simpson Distribution evenly taxa (evenness). Weighted by richness.

ACE

FaithPD phylogenetic relatedness between taxa.

5.4.2 Beta diversity

While alpha diversity measures the diversity within a single sample, beta diversity measures how different two or more samples are from each other. It gives an idea of how much the species composition varies between different environments or individuals. Beta diversity is crucial for identifying how the microbiome changes in response to factors like diet, disease, medication, or other environmental influences. For instance, researchers might compare the gut microbiome of healthy individuals with those who have a particular disease to identify differences.

We can measure beta-diversity using dissimilarity indexes, such as:

- Jaccard. Dissimilarity by presence/absence taxa. Range 0-1.
- Bray-Curtis. Dissimilarity weighted by taxa abundances. Range 0-1.
- Unifrac. Similarly based on phylogenetic relationships. Weighted/Unweighted.
- Aitchison. Euclidean distance between CLR-transformed data. Account for compositionality.

To analyse and visualize the differences in species composition across multiple samples we often use the Principal Coordinates Analysis (PCoA). PCoA is a technique used to visualize complex, multi-dimensional data (like Bray-Curtis dissimilarities) in a way that makes it easier to interpret. PCoA starts with a distance (or dissimilarity) matrix, which is a square matrix where each entry represents the dissimilarity between a pair of samples (like the Bray-Curtis dissimilarity). PCoA then finds a set of axes (called principal coordinates) that can best represent these distances in a lower-dimensional space, usually 2D or 3D. This is similar to how Principal Component Analysis (PCA) reduces data dimensionality but is specifically tailored to dissimilarity data.The first principal coordinate (PC1) explains the most variation in the data, the second principal coordinate (PC2) explains the next most, and so on.

Each point on the plot represents a sample, and the position of the point reflects its relationship with other samples based on their Bray-Curtis dissimilarity. Close Points: Samples that are close together on the plot have similar species compositions. Distant Points: Samples that are far apart have more different species compositions.

5.5 Metagenomes reconstruction

5.5.1 Assembly

Metagenomic assembly challenges that complicate the assembly and fragmenting of contigs:

- Highly non-uniform read coverage across different genomes. Widely different abundance levels of various species within a microbial sample.
- Shared conserved genomic regions across species, so call 'interspecies repeats'. They migh trigger intergenomic assembly errors.
- Closely related bacterial strains representing a bacterial species. Many bacterial species in a microbial sample are represented by *strain mitxures*, which are multiple related strain with varying abundances.
- High microbial diversity of certain enviroments. Comparison soil samples vs vaginal microbiome.

metaSPAdes

metaSPAdes is a state-of-the-art tool for assembling metagenomic sequencing data, addressing the challenges of mixed and complex microbial communities.

1. Sequence fragmentation into k-mers.
2. Error correction by k-mer abundance analysis.
3. Multisized de Bruijn graph construction with k-mers of different length.
4. Bulge corremoval by removing edged from bulger and recording its information for later use.
5. Detection and masking of strain variation. To respond the the microbial diversity challenge, metaSPAdes focuses on reconstructing a consensus backcone of a strain mixture, ignoring some strain-specific features corresponding to rare strains.
6. Iterative graph refinement.
7. Output Contigs by reconstruction paths in the assembly graph.

Sequence fragmentation. Input sequencing reads (e.g. paired-end reads of 150 bp, PE150) are split into smaller overlapping pieces called k-mers. These k-mers, substrings of fixed length k), will represent the building blocks for the subsequent graph construction. K-mers allow efficent handling of overlapping sequences, enabling the identification of how reads connect to from longer genomic fragments. For example: Read = ACGTAGC; k=3 | k-mers: ACG, CGT, GTA, TAG, AGC.

Error correction. Errors in sequencing (e.g., substitutions, insertions) create false k-mers that don't actually belong to the genome. MetaSPAdes uses the abundance of each k-mer to detect and correct these errors. High-abundance k-mers are more likely to be real, whereas low-abundant k-meres are considered likely errors and are corrected or removed. Removing errors reduces the graph's complexity, making the assembly more accurate and faster. If a sequencing error changes a k-mer ATCG to ATCC, and ATCC is detected in only one read, it can be corrected back to ATCG if supported by other reads.

Multisized de Bruijn graph construction. With our k-mers we will construct the alignment using a de Bruijn graph, where is node represents a k-mer and overlap between the k-mers are represented by an "arrow" (called directed edge). However, the choice of k affects the construction of the de Bruijn graph. Smaller values of k collapse more repeats together, making the graph more tangled. Larger values of k may fail

to detect overlaps between reads, particularly in low coverage regions, making the graph more fragmented. Since low coverage regions are typical for SCS data, the choice of k greatly affects the quality of single-cell assembly. Ideally, one should use smaller values of k in low-coverage regions (to reduce fragmentation) and larger values of k in high-coverage regions (to reduce repeat collapsing). Therefore metaSPAdes constructs multiple de Bruijn graphs using different k values, and combined them into a **multisized de Bruijn graph**, balacing resolution and connectivity. SPAdes is a universal A-Bruijn assembler in the sense that it uses k-mers only for building the initial de Bruijn graph and ”forgets” about them afterwards; on subsequent stages it only performs graph-theoretical operations on graphs that need not be labeled by k-mers. The operations are based on graph topology, coverage, and sequence lengths, but not the sequences themselves. At the last stage, the consensus DNA sequence is restored. Example: Bacteria A has many reads, so it forms a dense graph with many overlapping k-mers. Bacteria B and C have fewer reads, making their graphs sparser but still detectable due to combined k-mer sizes.

Bulge corremoval. Existing assembers often use two complementary approaches to deal with errors in reads: (1) error correction in reads and (2) bulge/bubble removal in de Bruijn graphs. Note the surprising contrast between these two approaches, both aimed at the same goal: the former approach corrects rather than removes erroneous k-mers in reads, while the latter approach removes rather than corrects erroneous k-mers in de Bruijn graphs. Removal (rather than correction) of bulges leads to deterioration of assemblies, since important information (particularly in the case of SCS) may be lost. SPAdes, unlike other NGS assemblers, records information about removed edges from bulges for later use before discarding them. We thus call this procedure ”bulge correction and removal” (or bulge corremoval). SPAdes maintains a data structure allowing one to backtrack all bulge corremovals. This is used in later stages of SPAdes to map reads to the assembly graph.

Decting and masking strain variation. Aiming at the consensus assembly of a strain mixture, metaSPAdes masks the majority of strain differences using a modification of the SPAdes procedures for masking sequencing errors. Genomic differences caused by sequecing errors often result in ”bulges” and ”tips” in the de Bruijn graphs. For example, an artifact often results in a bulge formed by two short alternative paths between the same vertices in the de Bruijn graph, a ”correct” path with high coverage and an ”erroneous” path with low coverage. Similarly, a substitution or a small indel in a rare strain (compared with an abundant strain) often results ina buldge formed by a high-coverage path corresponding to the abundant strain and an alternative low-coverage path corresponding to the rare strain.

Iterative graph refinement. The graph undergoes multiple rounds of refinement. (1) Ambigous regions are clarified, (2) Information from different k-mer sizes is combined, (3) remaining spurious connections are removed,

Output Contigs. Paths in the assembly graph are reconstructed into long sequences (contigs). This involves tracing through the graph to identify sequences that correspond to genomic fragments. Long, continuous sequences representing reconstructed portions of the genomes. Well-covered regions produce high-quality contigs, while low-coverage or ambiguous regions may remain fragmented.

MEGAHIT

5.5.2 Binning

Binning. Critical step required to establish a genome from a metagenomic assembly. This involves assignment of assembled fragments to a draft genome based on detection on any scaffold of some signal(s) that occur(s) locally within a genome and persists genome-wide [Che+20].

Genome curation [Hil+23; PP]. Filling scaffolding gaps and removal of local assembly errors. Gap filling strategies:

GapFiller Tool for filling the N’s gaps at scaffold joins. Often a few iterations are needed for gap closure. Using:

  - Unplaced pairs for reads adjacent to the gaps. When reads are mapped to genome fragments that compose a bin, a file of unplaced paired reads is generated for each fragment.
    * If due to low coverage gap filling is not achieve, potentially use of reads from other sample in which the sample population occurs.
    * Deeper sequencing of the same sample.
  - Placement of full metagenomic read data set to the new version of the scaffold.
  - Use of misplaced reads. This can be useful in cases where the necessary reads are misplaced, either elsewhere on that scaffold or on another scaffold in the bin. Misplaced read identification:
    * Read pileups with anomalously high frequencies of SNVs in a subset of reads.
    * Read pairs point outward (rather than toward each other, as expected).
    * Unusually long paired read distances.
  - Sometimes even with sufficient read depth, gap filling cannot be achieved due to complex repeats. Sometimes these repeat regions can be resolved careful read-by-read analysis, often requiring relocation of reads based on the placement of their pairs and sequence identity.

Local assembly errors (from more common to less):

- Error I:

  Identification Sequence in that region lacks perfect support, by even one read.

  Solution Consensus sequence should be replaced by Ns (gap), which can be further filled.

  Example https://genome.cshlp.org/content/suppl/2020/03/18/gr.258640.119.DC1/Supplemental_Fig_S3.pdf.

- Error II:

  Identification Ns have been inserted during scaffolding despite overlap between the flanking sequences.

  Solution Close the gap, eliminating the Ns and the duplicated sequence.

  Example https://genome.cshlp.org/content/suppl/2020/03/18/gr.258640.119.DC1/Supplemental_Fig_S4.pdf.

- Error III:

  Identification Incorrect number of repeats has been incorporated into the scaffold sequences. Anomalous read depth over that region.

- Error IV: Chimera sequences from two different organisms.

  Identification These joints typically lack paired read support and/or can be identified by very different coverage values and/or phylogenetic profiles on either side of the join.

- Error V: Artificial concatenation of an identical sequence.

  Identification Repeat finder.

5.6 Pangenomes

[TO DO] Differents genes within a population. Pangenome analysis.

5.7 Abundance estimation

[TO DO] How to quantify composition: markers genes and what makes good a marker gene (to be single and present in the core).

5.8 Sequencing depth

[TO DO] Huttenhower -¿ For strain analysis = ideally 10X; Gene-absence: ∽1X Discoveries and findings with microbelix.

Chapter 6

Oral Microbiome

Contents

6.1 Introduction

I didn't clean my teeth for three days and then took the material that had lodged in small amounts on the gums above my front teeth… I found a few living animalcules

- Antonie van Leeuwenhoek's letter to the Royal Society on observations made from his own dental plaque, translated by Clifford Dobell

Despite being considered by many as a relatively modern field of research, the first descriptions of human-associated microbiota date back to the 1670s-1680s, when Antonie van Leeuwenhoek started using his newly developed, handcrafted microscopes. In a letter written to the Royal Society of London in 1683, he described and illustrated five different kinds of bacteria (although he called them animalcules at the time) present in his own mouth and that of others, and subsequently also compared his own oral and faecal microbiota, determining that there are differences between body sites as well as between health and disease. Some of the first direct observations of bacteria were of human-associated microbiota.

The oral microbiome is the second large microbiome in the human body, after the gut, with more easy access. Presents spatially organized biofilms [MRB20; WMB20], which derives in the site-specialist hypothesis that predicts that most microbes in the human oral cavity have a primary habitat type within the mouth where they are most abundant.

Importance of the spatial organization in various aspects of the human microbial ecology [PR17].

Studies have revelead that the microbes living in the oral cavity are a major contributor to the overall host health [HC21]. In periodontal disease, there is a low-grade systemic inflammatory state [1] that has been mechanistically linked to multiple chronic inflammatory diseases such as Type-2 Diabetes Mellitus (T2DM) and cardiovascular disease (CVD). The host immune response is tightly intertwined with metabolism, and dysfunction of this integrated system may lead to chronic metabolic inflammatory disorders.

[1] glslow-grade systemic inflammation

6.2 Previous studies

Most of the oral microbiome research has been based on 16S rRNA gene amplicon sequencing studies [Esc+18] Metagenomic studies:

- "Over 50,000 Metagenomically Assembled Draft Genomes for the Human Oral Microbiome Reveal New Taxa". Includes:
  - –

6.3 Major oral habitats

The mouth is an open system. Microbes are breathed with the air, ingested with the food, or acquired through close contact (animals, humans, or surroundings).

Although the millions of bacterial species on the planet discovered so far (and other millions to remain discovered), it is believed that only approximately 760[2] are primary residents, rather than transients in the mouth, according to the Human Oral Microbiome Database [Esc+18] .

- Supragingival plaque
- Subgingival plaque
- Keratinized gingiva
- Hard palate
- Buccal mucosa
- Throat
- Palatine tonsils
- Tongue dorsum

Each of these sites is not monolithic, rather sheltered or exposed to different environmental conditions:

- Crowns of teeth abundant of oxygen.
- Tooth surface in the gingival crevice anorexic environment bathed in gingival crevicular fluid, protein-rich exudate from the gingival tissues.
- Saliva film thinnest at the roof of the mouth in contrast with the saliva pools at the floor.
- Similarly, relative proximity to salivary glands influences the composition and rate of flow of saliva.

Although saliva is not a habitat per se, there are evidence that microbes found within the saliva are not abundant at any of the other sampled sites, suggesting additional unique micro-habitats elsewhere in the mouth.

[2] In my opinion this number is not informative, as don't take into account any prevalance among population, including rare and probably transient species

6.4 Selective force within the mouth

Flow and adhesion

- Salivary flow imposes a selective requirement for adherence: microbes can persist in exposed locations in the mouth only if they are adhered to an underlying substrate or to other microbes that are able to adhere to the substrate.
- In addition, salivary flow also requires closely proximity for microbial interactions, as microbial metabolites are constantly washed. Interestingly,
- In response to selective pressures, oral microbes developed highly specific adhesin-receptor interactions, which form the basis for cohesion or coaggregation phenomenon.

Shedding and colonization

- Dynamics of shedding of the underlying substrate and re-colonization back to the substrate.
- Overall thickness of the microbial biofilm is influenced by the rate of shedding. Exposed areas: enamel teeth surface, mucosal surfaces. Factors: oral hygiene, abrasion by chewing food.
- Colonization of fresh substates after shedding and abrasion. This colonization is dependent on both microbial and host sources, e.g. colonizing streptococci bind to cysteine repeat domains within glycoproteins or sialic acid of mucin in the enamel pellicle, whereas adherence of specific bacteria to the mucosa could be mediated in part by the secretory immunoglobulin A.

Host and microbe

- Saliva flow and immune surveillance are properties of the host that reduce the microbial load.
- Saliva is also a vehicle for positive selection of microbes cause mucins and nutrients such as lactate, bicarbonate, nitrate, and vitamins are actively secreted into saliva [Car20].

Testing figs (Figure 6.1,Figure 6.2)

Saliva

Saliva is formed by an active process of ion secretion into the lumen of the gland, creating an osmotic gradient which draws water through from the interstitial space.

- Most ions and metabolites are transported by specific channels into saliva.
- Proteins are synthesized in the glands and added mostly by a separate mechanism of storage granule release dependant on cyclic adenosine monophosphate (AMP) signaling:
  - Saliva directly from the duct: few serum proteins.
  - Whole mouth saliva: high amount of serum proteins derived from a serum transudate leaking around teeth (via gingival crevicular fluid).
- Urea concentrations parotid saliva whole mouth saliva/plasma → active transport of urea into parotid saliva + use by bacteria.
  - Urea is the most non-protein nutrient in saliva, used by *Streptococcus salivaris*, *Actinomyces naeslundii*, *Haemophilus* by their expression of urease (urea → ammonia + CO2 or urea → ammonium carbamate → Formate). Urease is not present in mammalian cells. Indeed,

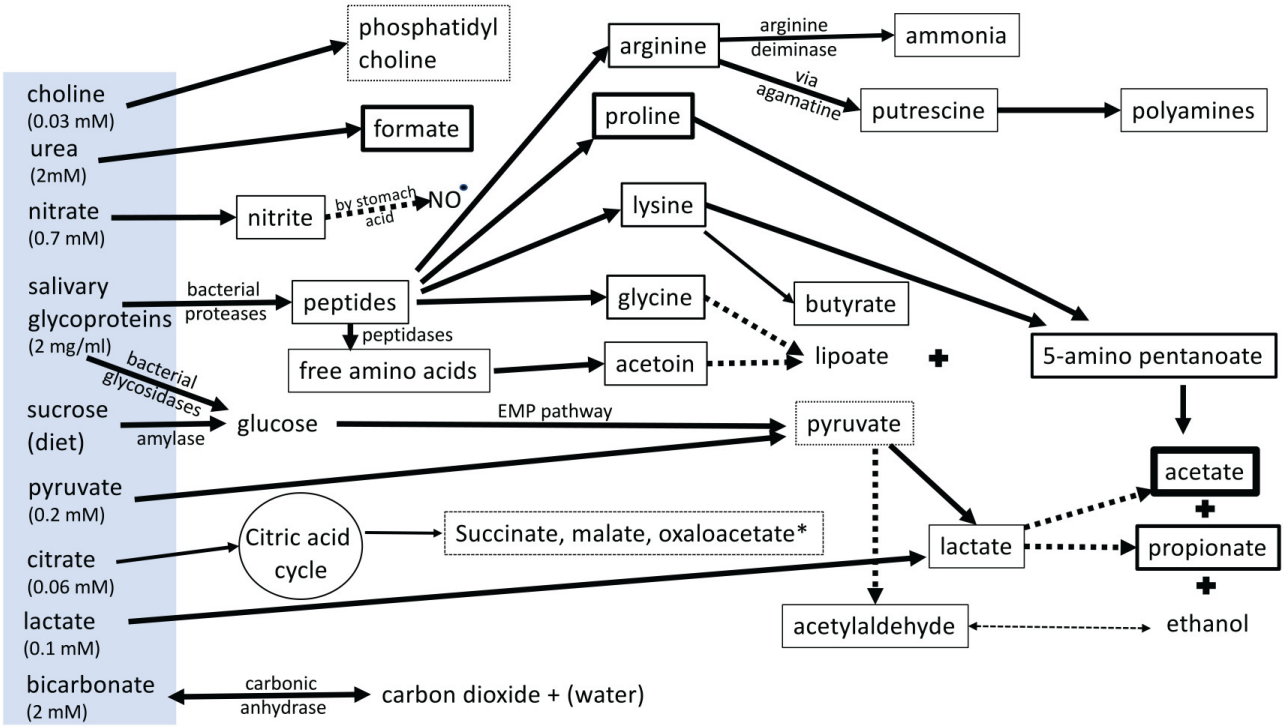

Figure 6.1: The main bacterial substrates (blue box) and detected metabolites (indicated by boxes) in whole mouth saliva. The thickness of arrows and boxes indicates relative abundance, dotted lines indicate possible connections. Under resting conditions between meals, the products of the citric acid cycle (indicated by *) are largely undetectable. Most metabolites indicate the breakdown of salivary glycoproteins as the main nutrient source, the amino acids yielding acetate and propionate, the N- and O-linked glycans leading to pyruvate via the Embden Meyerhof Parnas (EMP) pathway. Borrowed from "Salivary Factors that Maintain the Normal Oral Commensal Microflora" [Car20]

this reaction is so reliable that it is the basis of the urea breath test for *Helicobacter pylori* infections of the gut Figure 6.2.

- Low levels of sugars/carbohydrates in absence of food. Bacteria presumably rapidly utilize them via the Embden Meyerhof Parnas (EMP) pathway Figure 6.1.
  - Carbohydrates sources from food are still detectable after 20min, but usually clear in the mouth after 1h. → CH not may fuel source for commensal bacteria.
  - Proteins as main fuel source by proteolytic degradation of salivary proteins.
  - The Arginine Deiminase System (ADS) hydrolyses arginine to create citrulline and ammonia; the ammonia is beneficial to the host by neutralizing lactic acid in carious lesions. This pathway has become prominent as some dental products now contain arginine as an additive.
  - CH linked to proteins (glycoproteins) can also be used by sialidases action and other glycosidases (glycolytic EMP pathway = glucose → pyruvate). Here importance of bacteria cooperation in biofilms as no single bacterium contains all the necessary enzymes involved in the EMP pathway.
- Nitrate

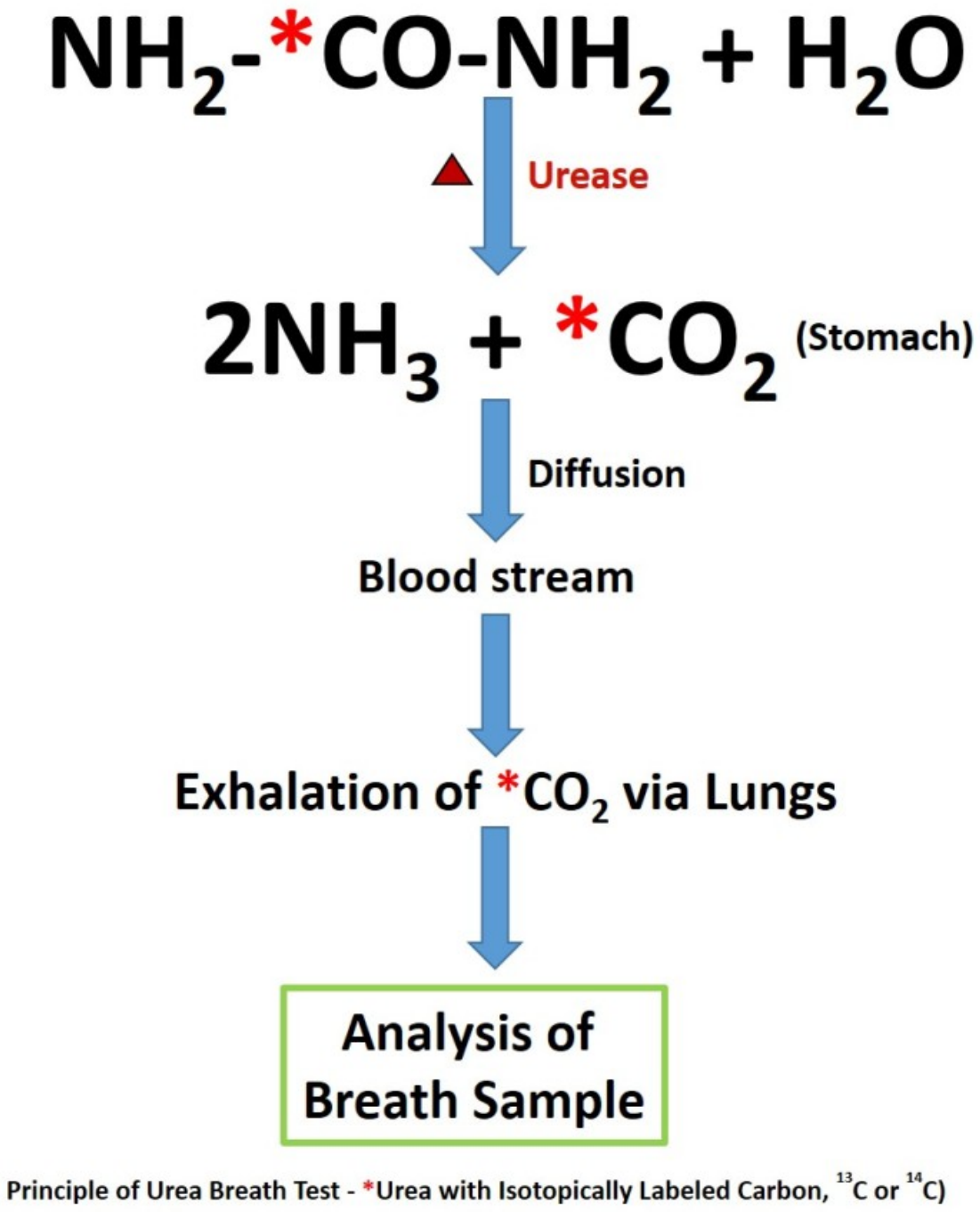

Figure 6.2: Urea breath test pathway. Borrowed from Sankararaman S, Moosavi L. Urea Breath Test. [Updated 2024 Feb 23]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK542286/

- Actively transported from the blood system by the salivary glands via the sialin transporter and delivered into the saliva.

  - Bacteria including Rothia and Veillonella nitrate → nitrite (+ stomach acid) → NO.
  - Cor(Salivary nitrate, lowered caries risk).
  - Altered microbiome by long-term nitrate supplementation → utilization.
- Lactate
  - Connected to a high diversity in the mouth. Lactate consumers present in multi-species biolfilms → syntrophy.
  - Actively secreted by saliva.
- Bicarbonate
  - Actively secreted by salivary mucin-secreting sublingual and minor glands.
  - Consumers and producers such as Streptococcus anginosus and Porphyromonas gingivalis, respectively.
- Limitation of other nutrients availability
  - Chelation of iron by binding to iron-free lactoferrin.
  - Cobalamin (B12 vitamin). Not transport from serum to saliva + transcobalamin (vitamin-binding proteins) → Prevents use by bacteria such as P. gingivalis.

6.5 Site-specialist communities in the major oral habitats

Definition 6.1 (Site-specialist hypothesis)**.** Each microbe in the mouth is specialized for one habitat or another, so that the microbiota at one oral site is different from the microbiota at other oral sites not onli in overall composition and proportions of common taxa but also in specific membership. Each taxon had a primary "ecological niche" (e.g. teeth, tongue, or buccal mucosa) [MRB20].

Namely, it suggests that the most significant factor that determines the niche for a microbe is its local habitat, which includes its immediate neighbors. It implies that the co-evolution of microbes within the mouth has led to highly specific taxon-taxon interactions that results in most microbes being restricted to the habitat type in the mouth that is occupied by those neighbors.

Microbial communities at different sites in the mouth are organized similarly, in that they are composed predominantly of several dozen abundant and prevalent taxa from core genera that are represented in each of the site microbiomes. However, at the species level, the sites are distincts. This pattern of a common set of genera but different species from site to site, suggests that the different members of the microbial community are a adapted to one another and co-evolve as the community adapts from one oral site to the next [MRB20].

6.6 Microbial habitats and niches at the micron scale

In spatially organized ecosystems such as the oral microbiome, the physical proximity cell-to-cell plays an important role associations and steep gradients that are crucial in forming the habitat for each microbe.

Definition 6.2 (Hedgehog)**.** Outer shell of approximately 20-30 μm wide which is composed by aerobes and facultative anaerobes [MRB20]. Inside of this outer anaerobic shell lay a middle layer occupied by taxa that grow well in micro-aerobic conditions; filaments *Corynebacterium* spp. were densely packed at its base and extended through the middle layer to the outer shell. The deep core of the structure is rich in taxa that grow anaerobically. (include figure)

In addition to habitat zones and gradients, tight cell-to-cell associations between disparate taxa are characteristic of oral biolfims and are the micron-scale manifestation of the molecular-level coadhesion interactions among taxa.

Distinct differences in microbial composition occur in different regions of disease-associated biolfilms.

A direct connection between spatial organization and biofilm pathogenecity was demonstrated by Kim et al 2020 studying *Streptococcus mutants* in caries. They showed that early biofilms were thin, flat, and characterized by intermixing between *S. mutants* and the non-pathogenic *S. oralis*, whereas later biofilms formed a domed structure in which the two taxa segregated from one another. Production of glucans by *S. mutants* was required for the segregation, as well as for the formation of the domed structure which was associated with demineralization of enamel.

The site specialist hypothesis holds that each taxon is restricted to a single category of site within the mouth and therefore is retricted to a limited set of partners, yet both spatial nearest-neighbor arrangements of taxa in oral biofilms, and the molecular interactions that underlie them, indicate that many taxa have a range of potential partners within a site. Corncobs are an example of fairly specific taxon-taxon interactions but with some flexibility in membership.

Definition 6.3 (Corncobs)**.** Structures in the dental plaque with highly stereotypical arrangements, with a central filaments surrounded by a single or double layer of cocci. The participating cocci are non-specific (i.e. *Streptococcus*, *Porphyromonas*, *Haemophilus* genera). (include figure)

Fusobacterium nucleatum was widely thought to occupy a special position in the development and structure of the oral films, by being a central structural components of plaque and essential for plaque maduration and an increase in plaque diversity. However, recents studies suggest that *F. nucleatum* does not constitute a physical bridge between early and late-colonizing dental plaque organisms [MRB20]. Instead, now it is believe to be an opportunistic colonizer and indicator of the maduration of the plaque to the point where anoxic niches become available (*Fusobacterium* spp. are obligate anaerobes). In the dental plaque hedgehog structure with filaments of *Corynebacterium* spp. form the structural bridge, reaching from the base of the structure to the tip where they form the core of the corncob structure that fringe the hedgehog.

Given the importance of syntrophy to the oral microbiome, the question arises of how syntrophy is maintained during growth of the biofilm. Due to the fact that a microbe grows more efficiently when it can receive resources from disparate microbes a few microns away, its growth will slow or cease when it gets too far from those syntrophic partners, creating clonal clusters during the process. Current hypothesis are growth in vertical columns and in a filamentous morphology, but exception in large patches have been shown for members of the genus *Actinomyces*, *Rothia mucilaginosa*, and *Streptococcus salivarius*.

6.7 Short and large-range factors

Most likely the oral biofilms compositions are influence by the interaction of short and large-range factors:

- Short-range: direct adhesion, micron-scale strong gradients.
- Large-range: Saliva flow composition and velocity.

Changes in the velocity of salivary flow result in changes in the clearence rate of substances from the surface of the biolfim, which presumably stregthen or attenuate the micron-scale gradients within the biolfilm.

Proctor2018 sampled buccal and lingual aspects of teeth in 30 individuals and found that the most abundant taxa were consistent regardless of location, but less-abudant taxa showed a gradient in abundance from front to back of the mouth, particularly at lingual sites. This study and another previous one from Simon-Soro et al 2013 found evidence for shifts in the proportions of the dental plaque microbiota on different teeth or different aspects of teeth, but the details of the finding differed.

It is possible that the question of how dental plaque communities shift across the mouth has not yielded a straightforward answer because a mistmatch between the sampling method and the size and spatial organization of the communities under study. In short, DNA sequencing approaches can now tell us with great accuracy and completness what microbes are present in the samples we collected and homogenized, but the sampling technology for sequencing lacks the requisite spation resolution to investigate community structure. To address questions of how microbes are organized, how they interact, and how is the functional role of each microbe in the physiology of an oral biofilm at the sampled sites, we need analysis methods with a higher resolution and in which spatial organization remains as intact as possible.

Bibliography

[Ber+20] Gabriele Berg et al. *Microbiome definition re-visited: old concepts and new challenges*. Microbiome 8.1 (June 2020), p. 103 (cit. on p. 41).

[Bis+15] Thorsten Bischler et al. *Differential RNA-seq (dRNA-seq) for annotation of transcriptional start sites and small RNAs in Helicobacter pylori*. Methods 86 (2015). Bacterial and Archaeal Transcription, pp. 89–101 (cit. on p. 32).

[Car20] G.H. Carpenter. *Salivary Factors that Maintain the Normal Oral Commensal Microflora*. Journal of Dental Research 99.6 (2020). PMID: 32283990, pp. 644–649. eprint: https://doi.org/10.1177/0022034520915486 (cit. on pp. 51, 52).

[Che+20] Lin-Xing Chen et al. *Accurate and complete genomes from metagenomes*. Genome Research 30.3 (2020), pp. 315–333. eprint: http://genome.cshlp.org/content/30/3/315.full.pdf+html (cit. on p. 46).

[Esc+18] Isabel F. Escapa et al. *New Insights into Human Nostril Microbiome from the Expanded Human Oral Microbiome Database (eHOMD): a Resource for the Microbiome of the Human Aerodigestive Tract*. mSystems 3.6 (2018), 10.1128/msystems.00187–18. eprint: https://journals.asm.org/doi/pdf/10.1128/msystems.00187-18 (cit. on p. 50).

[GD16] D.J. Goss and A.V. Domashevskiy. *Messenger RNA (mRNA): The Link between DNA and Protein*. Encyclopedia of Cell Biology. Ed. by Ralph A. Bradshaw and Philip D. Stahl. Waltham: Academic Press, 2016, pp. 341–345 (cit. on pp. 29, 30).

[Har+18] Alexander Harms et al. *Toxins, Targets, and Triggers: An Overview of Toxin-Antitoxin Biology*. Molecular Cell 70.5 (June 2018), pp. 768–784 (cit. on p. 26).

[HC21] George Hajishengallis and Triantafyllos Chavakis. *Local and systemic mechanisms linking periodontal disease and inflammatory comorbidities*. Nature Reviews Immunology 21.7 (July 2021), pp. 426–440 (cit. on p. 49).

[Hil+23] Saskia Hiltemann et al. *Galaxy Training: A powerful framework for teaching!* PLoS Computational Biology 19.1 (Jan. 2023). Ed. by Francis Ouellette, e1010752 (cit. on p. 46).

[HOH24] Yongqun He, Edison Ong, and Anthony Huffman. *Chapter 26 - Bacterial whole-genome determination and applications*. Molecular Medical Microbiology Third Edition. Ed. by Yi-Wei Tang et al. Third Edition. Academic Press, 2024, pp. 511–525 (cit. on pp. 15, 19).

[Khe+22] Supriya Khedkar et al. *Landscape of mobile genetic elements and their antibiotic resistance cargo in prokaryotic genomes*. Nucleic Acids Research 50.6 (Mar. 2022), pp. 3155–3168. eprint: https://academic.oup.com/nar/article-pdf/50/6/3155/43246153/gkac163.pdf (cit. on p. 26).

[Le +22] Anaïs Le Rhun et al. *Profiling the intragenic toxicity determinants of toxin-antitoxin systems: revisiting hok/Sok regulation*. Nucleic Acids Research 51.1 (Oct. 2022), e4–e4. eprint: https://academic.oup.com/nar/article-pdf/51/1/e4/48723059/gkac940.pdf (cit. on p. 26).

[MRB20] Jessica L. Mark Welch, S. Tabita Ramírez-Puebla, and Gary G. Borisy. *Oral Microbiome Geography: Micron-Scale Habitat and Niche*. Cell Host and Microbe 28.2 (2020), pp. 160–168 (cit. on pp. 49, 54, 55).

[Pen21] Sergio D. J. Pena. "An Overview of the Human Genome". Human Genome Structure, Function and Clinical Considerations. Ed. by Luciana Amaral Haddad. Cham: Springer International Publishing, 2021, pp. 1–24 (cit. on pp. 15–17).

[PP] Nikos Pechlivanis and Fotis E. Psomopoulos. *Binning of metagenomic sequencing data (Galaxy Training Materials)*. [Online; accessed Wed Oct 30 2024] (cit. on p. 46).

[PR17] Diana M. Proctor and David A. Relman. *The Landscape Ecology and Microbiota of the Human Nose, Mouth, and Throat*. Cell Host and Microbe 21.4 (2017), pp. 421–432 (cit. on p. 49).

[SL17] Thomas Schafee and Rohan Lowe. *Eukaryotic and prokaryotic gene structure*. WikiJournal of Medicine 4.1 (Jan. 2017) (cit. on pp. 31–34).

[SV14] Cynthia M Sharma and Jörg Vogel. *Differential RNA-seq: the approach behind and the biological insight gained*. Current Opinion in Microbiology 19 (2014). Ecology and industrial microbiology • Special Section: Novel technologies in microbiology, pp. 97–105 (cit. on pp. 29, 32, 35).

[Tha19] Varykina G. Thackray. *Sex, Microbes, and Polycystic Ovary Syndrome*. Trends in Endocrinology and Metabolism 30.1 (2019), pp. 54–65 (cit. on p. 43).

[WMB20] Steven A. Wilbert, Jessica L. Mark Welch, and Gary G. Borisy. *Spatial Ecology of the Human Tongue Dorsum Microbiome*. Cell Reports 30.12 (2020), 4003–4015.e3 (cit. on p. 49).

[Zhu+22] Jie Zhu et al. *Over 50,000 Metagenomically Assembled Draft Genomes for the Human Oral Microbiome Reveal New Taxa*. Genomics, Proteomics & Bioinformatics 20.2 (2022), pp. 246–259 (cit. on p. 50).