Metabook

Leidy-Alejandra G. Molano

Update on November 5, 2024

Contents

List of Figures

List of Theorems

List of Definitions

LIST OF DEFINITIONS

Part I

Genomes

Chapter 1

Human Genome

Contents

gls examples:

- Greatest Common Divisor (GCD); Greatest Common Divisor; GCD; Greatest Common Divisor (GCD)

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!

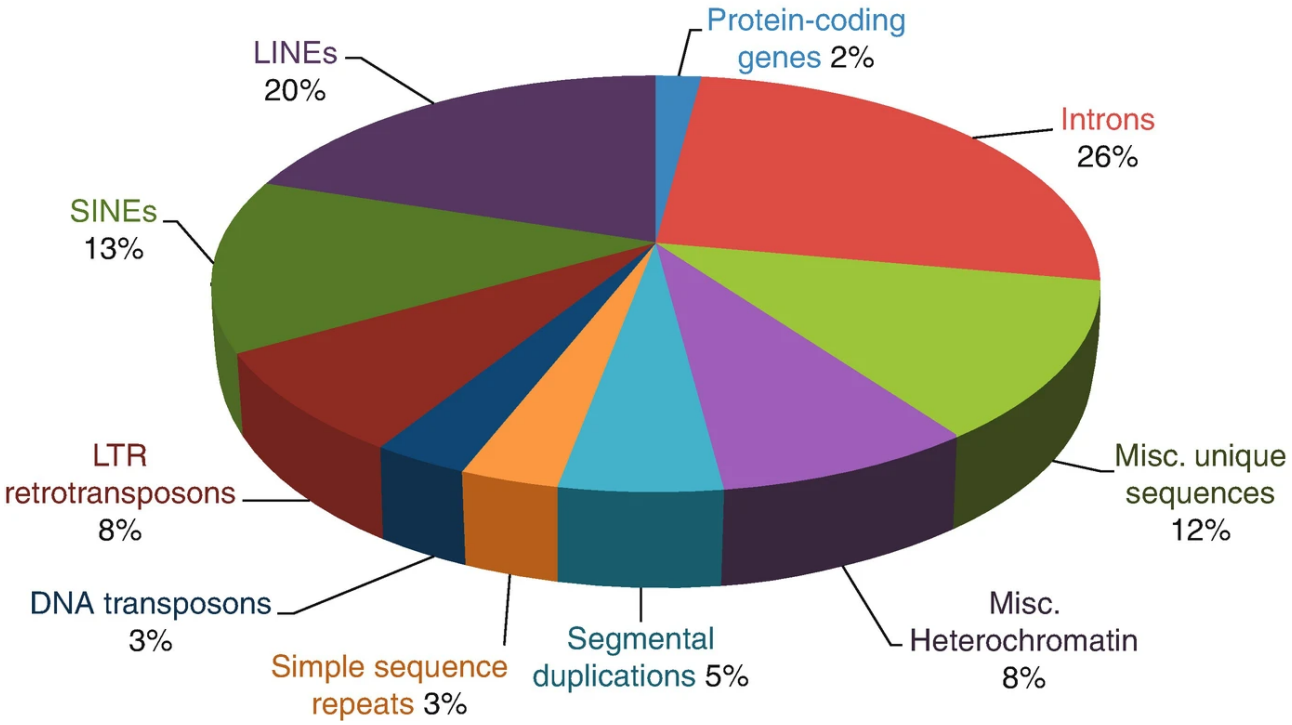

Figure 1.1: Composition of the human genome. Redrawn from a graph that was produced in 2014 by the NHS National Genetics and Genomics Education Centre. Borrowed from "An Overview of the Human Genome" [Pen21]

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 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 $\backsim$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 $\backsim$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.

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

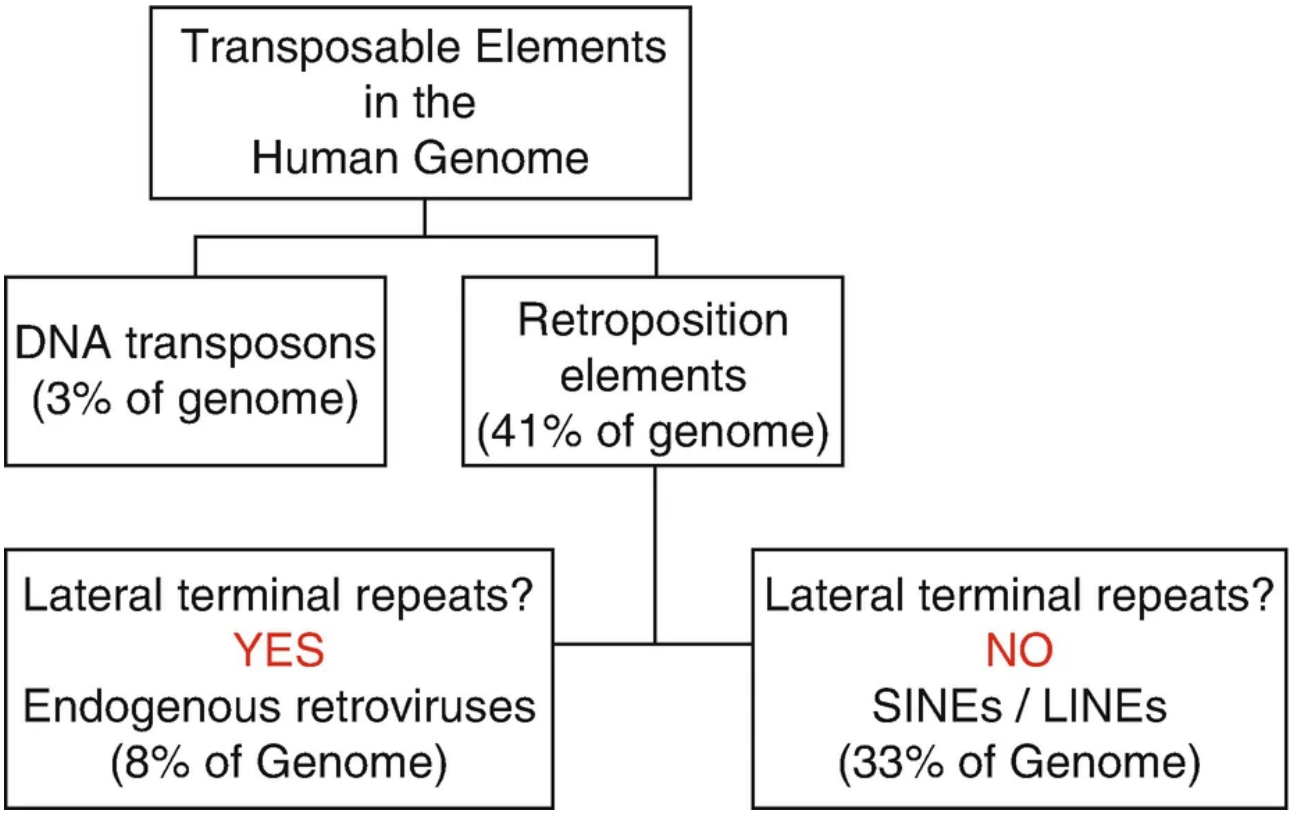

Figure 1.2: Classes of transposable elements in the human genome. Borrowed from "An Overview of the Human Genome" [Pen21]

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

Contents

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.

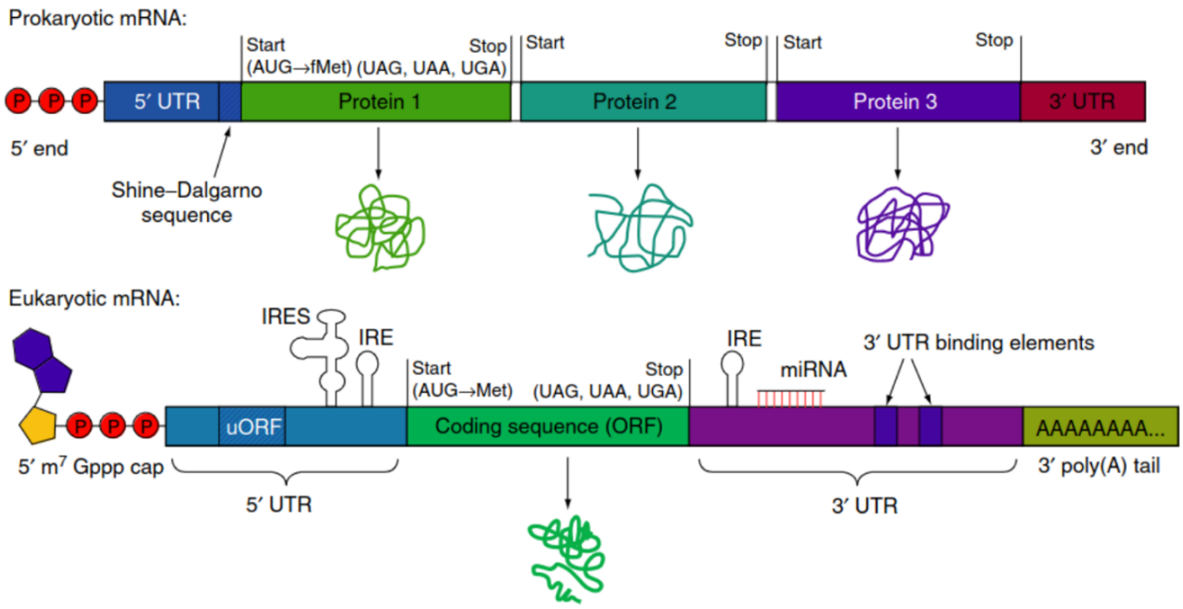

Figure 4.1: Schematic diagram of prokaryotic (top) and eukaryotic (bottom) mRNA. Bars indicate the relative length of the regions. Borrowed from "Messenger RNA (mRNA): The Link between DNA and Protein" [GD16].

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].

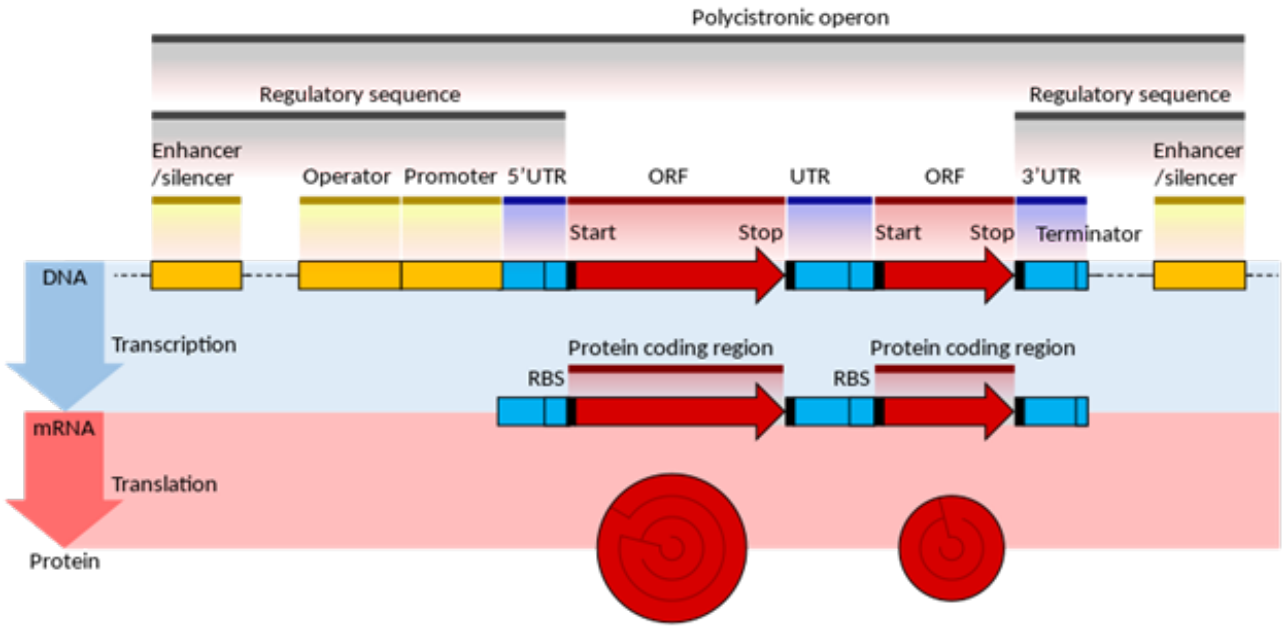

Figure 4.2: The structure of a eukaryotic protein-coding gene. Regulatory sequence controls when and where expression occurs for the protein coding region (red). Promoter and enhancer regions (yellow) regulate the transcription of the gene into a pre-mRNA which is modified to remove introns (light grey) 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].

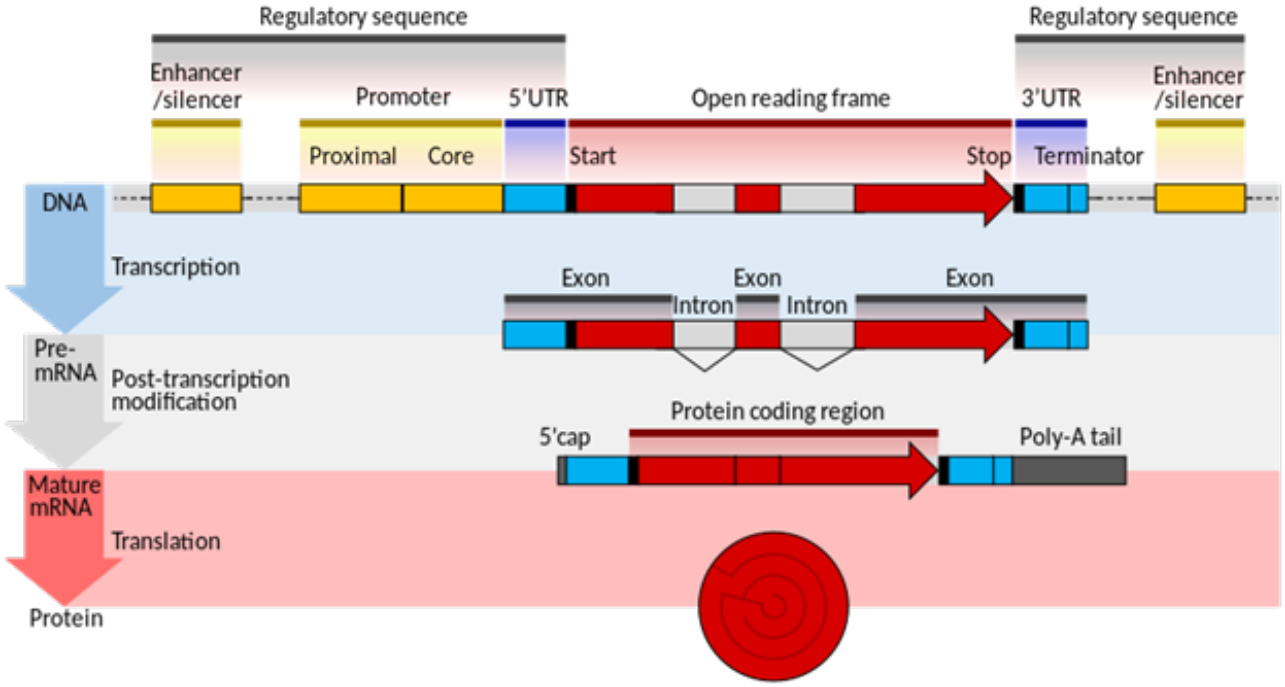

Figure 4.3: The structure of a prokaryotic operon of protein-coding genes. Regulatory sequence controls when expression occurs for the multiple protein coding regions (red). Promoter, operator and enhancer regions (yellow) regulate the transcription of the gene into an mRNA. The mRNA untranslated regions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

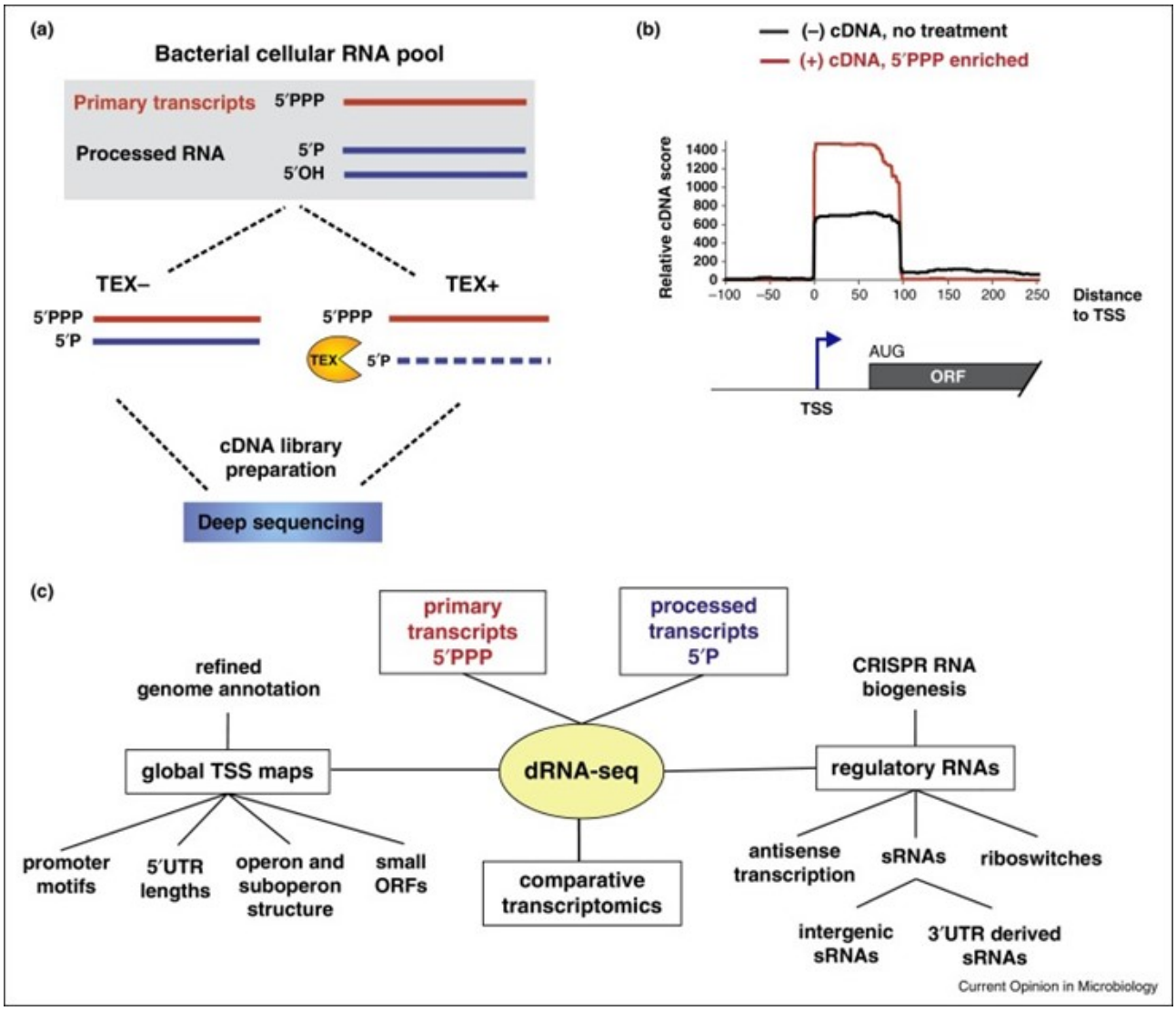

Figure 4.5: Rationale and output of the dRNA-seq approach. (a) Enrichment of primary transcripts using 5'-phosphate-dependent terminator exonuclease (TEX). The bacterial RNA pool consists of primary transcripts with a 5'PPP and processed RNAs with a 5'P or 5'OH. RNAs with a 5' OH group are not accessible 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

5.1 Metagenomes reconstruction

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.

5.2 Pangenomes

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

5.3 Microbila diversity

[TO DO] Beta diversity and its representation. Concept of dimensionality reduction techniques.

5.4 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).

5.5 Sequencing depth

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

Chapter 6

Oral Microbiome

6.1 Introduction

It 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.

Definition 6.1 (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.

Definition 6.2 (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.

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

6.2 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[1] 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

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

- 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.

6.3 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)

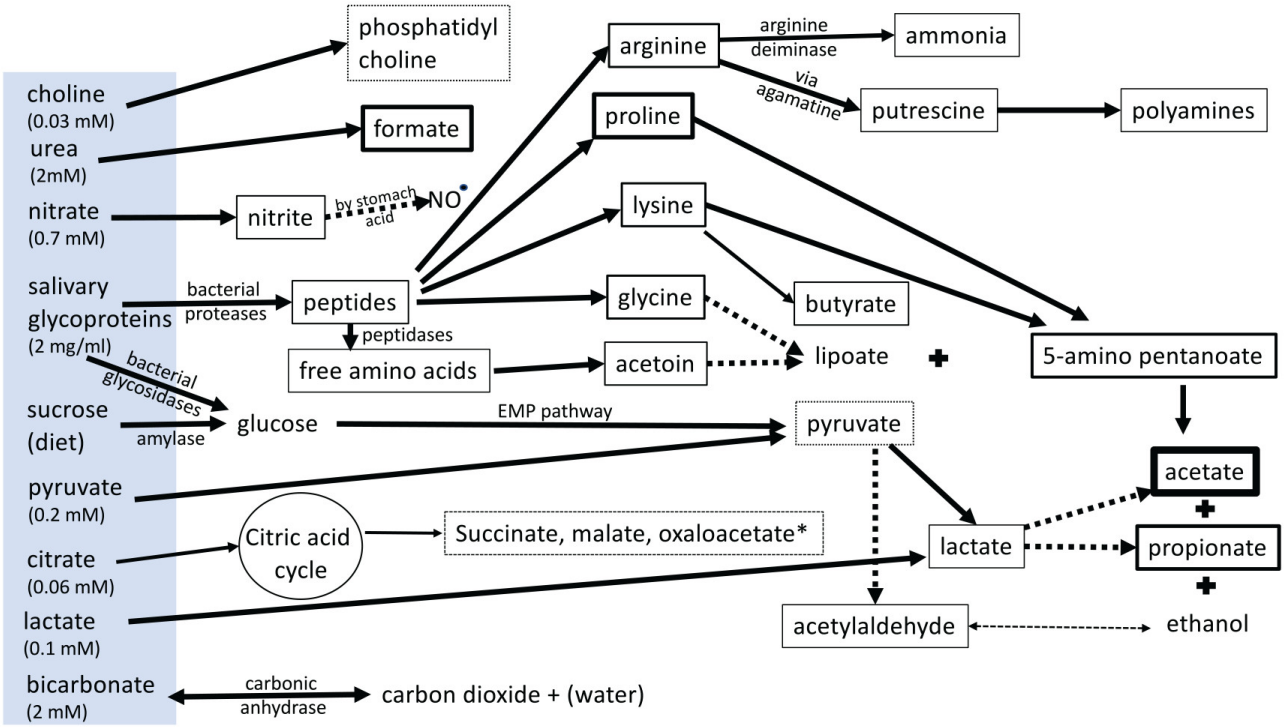

Figure 6.1: The main bacterial substrates (blue box) and detected metabolites (indicated by boxes) in whole mouth saliva. The thickness of arrows and boxes indicates relative abundance, dotted lines indicate possible connections. Under resting conditions between meals, the products of the citric acid 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]

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,

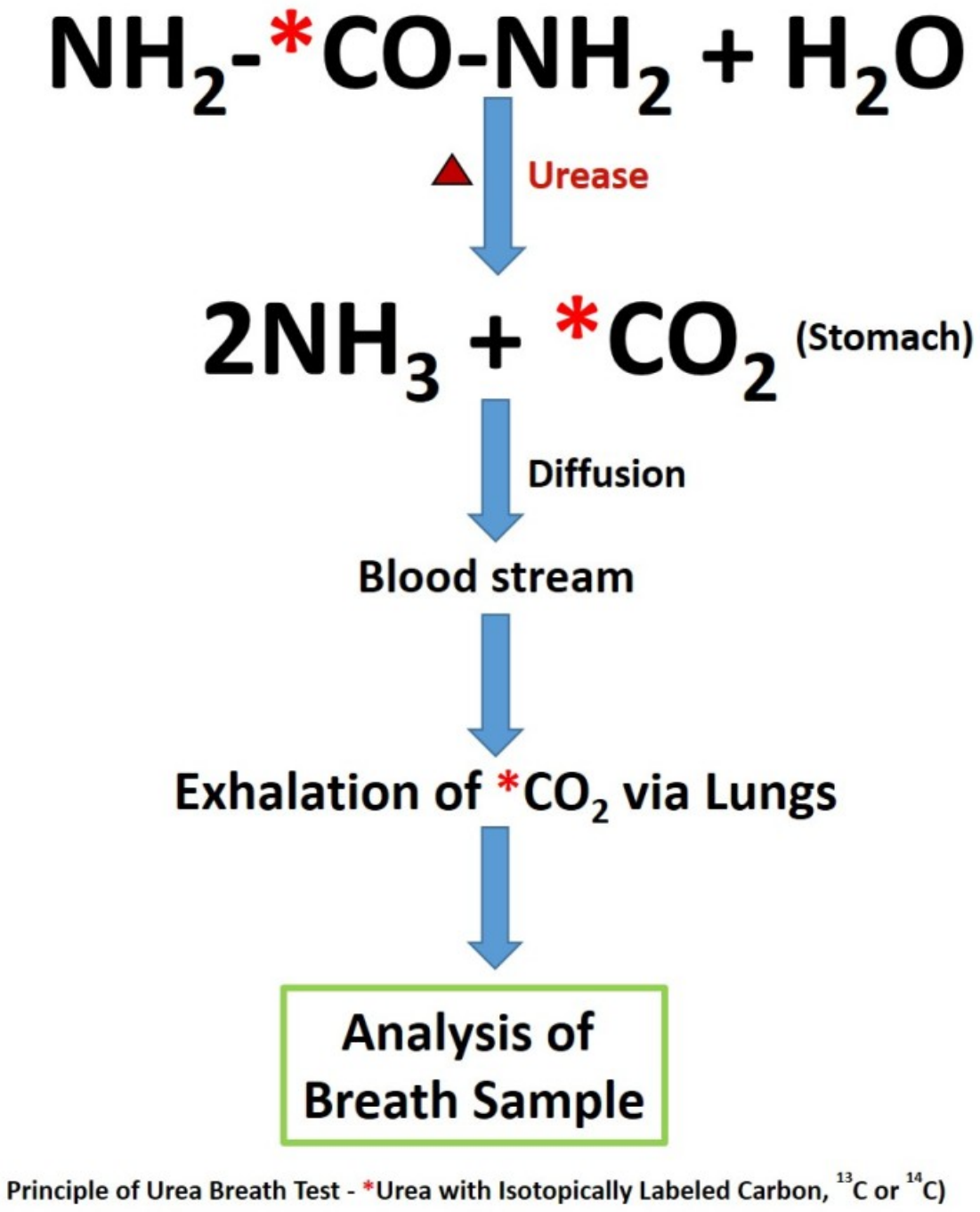

Figure 6.2: Urea breath test pathway. Borrowed from Sankararaman S, Moosavi L. Urea Breath Test. [Updated 2024 Feb 23]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK542286/

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
  - 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.

Bibliography

[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. 28).

[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. 38, 39).

[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. 35).

[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. 37).

[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. 25, 26).

[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. 22).

[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. 35).

[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. 11, 15).

[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. 22).

[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. 22).

[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 p. 37).

[Pen21] Sérgio 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. 11, 12, 14).

[PP] Nikos Pechlivanis and Fotis E. Psomopoulos. *Binning of metagenomic sequencing data (Galaxy Training Materials)*. [Online; accessed Wed Oct 30 2024] (cit. on p. 35).

[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. 37).

[SL17] Thomas Schafee and Rohan Lowe. *Eukaryotic and prokaryotic gene structure*. WikiJournal of Medicine 4.1 (Jan. 2017) (cit. on pp. 27–30).

[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. 25, 28, 31).

[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. 37).