Development of computational and experimental tools for the identification of small proteins in bacterial genomes

Samuel Miravet Verde

DOCTORAL THESIS UPF / YEAR 2020

Thesis supervisors

Prof. Luis Serrano Pubul
Dra. Maria Lluch Senar Pulmobiotics / UIC

THESIS DEPARTMENT

Department of Experimental and Health Sciences (UPF)

Design of Biological Systems Laboratory - Systems Biology Unit

Centre for Genomic Regulation (CRG)

A mis padres y a mis hermanas

Acknowledgements

I would like to start expressing how grateful I am to those great people whose feedback, ideas and support are also part of this project. First of all, I want to thank my supervisors, Luis and Maria, for accepting me as a PhD student and the guidance provided during these years. You have grounded my critical thinking capacity and my research vocation with an incommensurable support, giving me not only the opportunities but also your trust. I am a better version of myself thanks to sharing these great years, a time in which you have taught me to face any challenge, while enjoying every step with passion. Specially, thanks Maria for your positiveness, trusting me as a sidekick in bioinformatic analyses, and being such an inspiring scientist example of perseverance and success. Thanks Luis for your always on-point feedback, your great-crazy ideas, and your full availability despite your complicated calendar. Also, I would like to extend this acknowledgement to the support and fruitful discussion provided along four years by a great group of scientists belonging to my thesis advisory committee: James Sharpe, Juan Valcárcel, and Lucas Carey.

I also want to thank every former and current member of the Serrano lab for every discussion, beach volleyball game, retreat, video shooting, and parties, but specially for always being at my side during the ups and downs of the PhD. First, I would like to acknowledge the contribution of the people that have experimentally supported this thesis: Raul aka ‘uncle’, Carlos, Rocco, Tony, Alicia, Carolina, Eva García, Eva Yus, and Sira. I also want to thank the whole dry lab for those spontaneous and fruitful debates: Jae, Martin, Anas, Oscar, Miguel, and especially Javi, Marc and Leandro, with whom I have learned A LOT about computers, programming, data analysis, and life. Furthermore, I would also like to appreciate the feedback from the signaling team: Christina Violeta, and especially Claire, and Sarah. Finally, I would like to thank the other PhDs in the lab, with whom I have always found empathy and good moments. Vero, thanks for introducing me to Maria and Luis and sharing your knowledge on machine learning with me. Thanks Dan and Ari (la millor PhD i companya de batalles que ha donat la terreta) for your trust in the projects we have collaborated. Also Xavi, and Miquel, for interesting discussions on modeling and machine learning, I hope we can collaborate in the near future. I would like to thank specially Marie, Hannah, Ludo and Damiano, for being the most inspiring PhDs I have met. You have been friends, confidents, psychologists and overnight colleagues who have taught me that perseverance, positiveness, and living the present makes everything possible. In retrospective, I would change many moments from these years, however, every second with you I would repeat it again. Thanks to all of you for contributing to such a great atmosphere, it is a pleasure to work with you.

I would like to acknowledge the great months in New York learning about whole-cell modeling with Jonathan Karr, too. I enjoyed every scientific and social time with you, Arthur, Yin Hoon, John, Balazs, Roger, and Yosef. I would like to specially thank Yosef, extending it to Veronika, for these two years meeting great scientists around the world. Furthermore, thanks to all the Ramón y Cajal lab, specially Santiago, for trusting me in the task of analyzing transposon libraries from cancer cells. One of the things I have enjoyed the most while doing a PhD has been to meet amazing people, who are also scientists. Cris, thanks for all those valuable memories between coffee times, balcony discussions, and volleyball games. I would also like to acknowledge the starting PhD days with María Carla, Tobias, Artem, Aitor, Neus, Alejandra, Nieves, Iago, Beatrice, Ati, Manuel, Marcos, Sergi, Silvia, Mar, and Jackie. Also, Núria, Damjana and Imma, for their support and training opportunities, and Reyes, for being such a great logistic manager and always willing to help.

No puedo dejar de agradecer a las personas que me han acompañado, entendido e inspirado durante estos años. A mis amigos del mostacho: Carlos, Ernesto, Fran, Joaquín, Joan, Jose Luis, Jose Manuel y Victor. Habéis estado ahí siempre sin importar las circunstancias, celebrando mis visitas a Burriana con comida, bebida y conversaciones geniales, como si aún siguiéramos en el banco del cole. A Rubén y Álvaro por los años conviviendo, vuestros valores y vocaciones son referencia para mí. A Alberto, Guillem, Isa, Maria, Marta, Pau Jané, Pau Jurado, Paula y Jan, por haberme hecho sentir uno más del grupo. Ha sido un honor navegar con ustedes, en especial en Salit's con Paula y Jan; gracias por todos esos ratos de desconexión y ruido encima y abajo del escenario. Ratos comparables a los que he encontrado estos años con mis hermanos Fran, Carlos y Ernesto en Cheese and Onions y los chicos tragedia Álex, Edu, Mike, Iván y Alberto en The Taste of Tragedy. Gràcies també als actuals germans de batalla Saul, Manu i Fede, per sempre entendre les complicacions d'horari d'un PhD i donar-me, amb Astrial, la possibilitat de sonar més fort encara. También agradecer los trasnoches de entretenimiento junto a las leyendas Damiano, Joaquín, Joan, Jose Luis, Jose Manuel y Ernesto. Aunque también a aquellos que me han acompañado cuando ya nadie quedaba despierto: Coheed and Cambria, Press To Meco, Viva Belgrado, Mastodon, In Flames, Arcane Roots, la familia GTM, Santa Monica, Kojima Productions, From Software, Naughty Dog, Ryu ga Gotoku Studio y S. King.

A Mónica, porque tu apoyo y compañía en los momentos críticos han sacado la confianza que me faltaba en llevar adelante muchas de las ideas aquí representadas. Gracias por haber hecho este camino conmigo, especialmente este último año, ya que sin ti no lo habría conseguido. Te mereces todo y más.

Finalmente, quiero agradecer a mi familia, que a pesar de habernos distanciado físicamente en este período, yo los he sentido cada día más cerca. En primer lugar, a mis hermanas, Marta y Amanda, porque siempre me habéis dado más de lo que yo podía ofreceros, apoyándome y cuidándome. Cada recuerdo que con vosotras evidencia que no hace falta estudio científico para asegurar que sois las mejores hermanas del mundo. Gracias también a la más pequeña de la familia y mejor regalo de este año, Helena, que con la gracia de quien no es consciente, ha sido capaz de hacerme olvidar muchos problemas. Por último, a mis padres, Ramón y Tachi. Gracias por vuestro infinito apoyo académico y personal. No sólo me habéis dado todos los recursos que necesitaba para completar este proyecto, sino también la motivación para perseguir todos mis sueños. Todo lo que he conseguido y conseguiré, es mérito vuestro.

Abstract

Small proteins (SEPs; <100aa) are involved in essential processes such as cell homeostasis, signalling, or metabolism. However, they have been overlooked because of computational and experimental difficulties that prevent their annotation and rely their identification on serendipity. In this thesis, we present a series of tools to aid the characterization of bacterial SEPs. i) RanSEPs, the first bioinformatics tool to annotate SEPs, based on species-specific sequence features and random forest models. Running RanSEPs in 109 bacterial genomes reveals that SEPs could represent up to 20% of some species' proteomes. ii) FASTQINS and ANUBIS, two bioinformatics tools for the processing and analysis of transposon sequencing libraries to increase the accuracy in genome essentiality studies, including small genomic regions. iii) ProTInSeq, a novel transposon sequencing approach using mutated vectors to study bacterial proteomes, including SEPs, applied in *Mycoplasma pneumoniae*. Altogether, these tools aid the discovery of uncharacterized SEPs, including *quorum sensing* and antimicrobial SEPs, which functions could be exploited for the treatment of microbial diseases.

Keywords: small proteins, genome annotation, transposon sequencing, essentiality, proteomes

Resum

Les proteïnes petites de menys de 100 aminoàcids (SEPs) estan involucrades en processos essencials per a la cèl·lula com homeòstasis, senyalització o metabolisme. Nogensmenys han passat desapercebudes a causa de les limitacions computacionals i experimentals que impedeixen la seva identificació de SEPs son descobertes per serendipitat. En aquesta tesi presentem una sèrie d'eines per la caracterització de SEPs en bacteris. i) RanSEPs, la primera aplicació bioinformàtica destinada a l'anotació de SEPs, basada en propietats específiques de les seqüències de cada espècie y models de boscos aleatoris. A l'utilitzar RanSEPs en 109 espècies de bacteris observem que fins a un 20% de les proteïnes contingudes en un genoma podrien ser SEPs. ii) FASTQINS i ANUBIS, dues eines bioinformàtiques, per al processament i anàlisi de dades de seqüenciació d'elements genètics transposables per millorar la qualitat dels estudis d'essencialitat en genomes incloent petites regions genòmiques. iii) ProTInSeq, un nou protocol de seqüenciació d'elements genètics transposables utilitzant vectors mutats per estudiar proteomes en bacteris, inclòs SEPs, aplicat a *Mycoplasma pneumoniae*. En conjunt, aquestes eines assisteixen al descobriment de SEPs sense caracteritzar, incloent-hi SEPs de percepció de quòrum o antimicrobians, funcions les qual poden ser aplicades en el tractament de malalties microbianes.

Conceptes clau: proteïnas petites, anotació de genomes, seqüenciació d'elements genètics transposable, essencialitat, proteomes

Resumen

Las proteínas pequeñas de menos de 100 aminoácidos (SEPs) están involucradas en procesos esenciales para la célula como homeostasis, señalización o metabolismo. Sin embargo, han pasado desapercibidas debido a limitaciones computacionales y experimentales, haciendo que su identificación se base en serendipias. En esta tesis presentamos una serie de herramientas para la caracterización de SEPs en bacterias. i) RanSEPs, la primera aplicación bioinformática destinada a la anotación de SEPs, se basa en propiedades específicas de las secuencias de cada especie y modelos de bosques aleatorios. Al utilizar RanSEPs en 109 especies bacterianas se observa que hasta un 20% de las proteínas contenidas en un genoma podrían ser SEPs. ii) FASTQINS y ANUBIS, dos herramientas bioinformáticas para el procesamiento y análisis de datos de secuenciación de elementos genéticos transponibles para mejorar la calidad de los estudios de esencialidad en genomas, incluyendo pequeñas regiones genómicas. iii) ProTInSeq, un nuevo protocolo de secuenciación de elementos genéticos transponibles usando vectores mutados para estudiar proteínas en bacterias, incluidas SEPs, aplicado en *Mycoplasma pneumoniae*. En conjunto, estas herramientas asisten el descubrimiento de SEPs sin caracterizar, incluyendo SEPs de percepción de quórum o antimicrobianas, cuyas funciones podrían ser aplicadas en el tratamiento de enfermedades microbianas.

Conceptos clave: proteínas pequeñas, anotación de genomas, secuenciación de elementos genéticos transponibles, esencialidad, proteomas

List of publications

Miravet-Verde S; Lloréns-Rico V; Serrano L, 2017. *Alternative transcriptional regulation in genome-reduced bacteria.* Curr Opin Microbiol 39:89-95

Miravet-Verde S; Ferrar T; Espadas-García G; Mazzolini R; Gharrab A; Sabido E; Serrano L; Lluch-Senar M, 2019. *Unraveling the hidden universe of small proteins in bacterial genomes.* Mol Syst Biol 15(2):e8290

Montero-Blay A; **Miravet-Verde S**; Lluch-Senar M; Piñero-Lambea C; Serrano L, 2019. *SynMyco transposon: engineering transposon vectors for efficient transformation of minimal genomes.* DNA Res 26(4):327-339

Martín-Pardillos A; Valls Chiva Á; Bande Vargas G; Hurtado Blanco P; Piñeiro Cid R; Guijarro PJ; Hümmer S; Bejar Serrano E; Rodriguez-Casanova A; Diaz-Lagares Á; Castellvi J; **Miravet-Verde S**; Serrano L; Lluch-Senar M; Sebastian V; Bribian A; López-Mascaraque L; López-López R; Ramón Y Cajal S, 2019. *The role of clonal communication and heterogeneity in breast cancer.* BMC Cancer 19(1):666

Montero-Blay A; Piñero-Lambea C; **Miravet-Verde S**; Lluch-Senar M; Serrano L, 2020. *Inferring Active Metabolic Pathways from Proteomics and Essentiality Data.* Cell Rep 31(9):107722

Miravet-Verde S; Burgos R; Delgado J; Lluch-Senar M; Serrano L, 2020. *FASTQINS and ANUBIS: two bioinformatic tools to explore facts and artifacts in transposon sequencing and essentiality studies.* Nucleic Acids Res 48(17):e102

Shaw D; **Miravet-Verde S**; Piñero-Lambea C; Lluch-Senar, M; Serrano, L, 2020. *LoxTnSeq: Random Transposon insertions combined with cre/lox recombination and counterselection to generate large random genome reductions.*

Contents

List of figures

List of tables

Chapter 1. Introduction

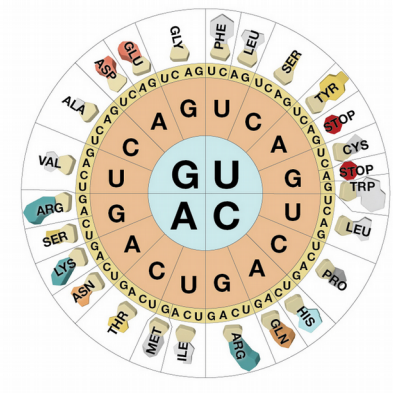
1.1. Information and function in life

1.1.1. From genes to genomes

The term ‘gene’ originally referred to ‘inheritance units’, the particular characteristics of an organism being inherited from its parentals [1]. At the molecular level, genes were an abstract concept that became physical with the evidence of DNA from pathogenic bacteria transforming nonvirulent strains [2–4]. Other DNA features, like the equimolar correspondence of the nucleobases: adenine (A) with thymine (T), and cytosine (C) with guanine (G), in addition to crystallographic observations [5,6], served as the founding ideas behind the first model of DNA by James D. Watson and Francis Crick in 1953 [7]. This model provided a molecular basis for heredity: two antiparallel DNA double-helix conformed by base-pairs (bp) forming hydrogen bonds (A=T; G≡C). It also fostered the definition of a “Central Dogma” in molecular biology: DNA as information storage system and self-preserved by replication; messenger RNA (mRNA; composed by ribonucleotide A, C, G, and uracil - U) synthesized from DNA in transcription; and proteins as functional units translated from RNA. Few additions highlighting RNA as more than intermediates were required to complete the information flow: synthesis of DNA from RNA (i.e. reverse transcription), and RNA replication, shown by viruses [8,9]; and genes encoding for functional non-translated RNA molecules: ribosomal and transfer RNAs (rRNA and tRNA, respectively), and other ‘non-coding’ RNA families that can have regulatory roles [10]. These last considerations made genes to be defined as the code in a nucleic acid (DNA or RNA) that gives rise to a functional product (RNA or protein) [11].

Less than a decade after the DNA model conception, researchers deciphered how the information was encoded in it and defined a genetic code (i.e. a ‘dictionary’ to translate DNA or RNA to proteins) [12,13]. This code is composed of non-overlapping combinations of three nucleotides called codons (4^3 = 64 codons), each of them encoding one amino acid out of 20, in a redundant or ‘degenerate’ manner [14]. The first codon of protein is methionine in eukaryotes and Archaea, and N-formylmethionine in bacteria (i.e. ‘start’ codon, generally AUG). In bacteria, alternative use of start codons such as GUG and UUG [15], while AUU has been shown to encode also for proteins in some species [16]. Alternatively, UAA, UAG, and UGA are ‘stop’ codons that do not encode for an amino acid but the signal of termination of protein synthesis (Figure 1.1). In the case of these codons the UGA can encode in bacteria of the Mollicute class for tryptophan. Thus depending on the species, attention should be paid to exceptions in the universal codon usage. As start and stop codons are conserved, the term Open Reading Frame (ORFs) is commonly used to refer to the span of the genome

between a start and a stop codon found in genomes. The term frame comes from considering the double-stranded DNA molecule as six possible ‘reading frames’, based on the triplet nature of codons [17]. As not all ORFs in a genome encode for a protein, the terms putative ORF and CDS (from coding sequence) are used to differentiate them [18].



Amino acid	Three letter symbol	One letter symbol
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Figure 1.1. The genetic code.
Left, in to out: first nucleotide in the triplet (blue), second (orange), list of possible third bases (yellow), encoded amino acid (white). The diverse amino acidic chemical structures and properties are represented with the 3D models. On the right, a table with the name and symbols assigned by IUPAC notation. * Figure adapted from kaiserscience (CC BY 4.0)

The genetic code was applicable to the first sequenced gene, encoding the coat protein of bacteriophage MS2 [19]. Later, the genome, or complete sequence of genetic material, of this virus was also the first to be sequenced (3,569 ribonucleotides) [20]; followed by the first DNA genome, from bacteriophage ϕX174 (5,386 bp) [21]. From there, sequencing and computational approaches facilitated large genomes sequencing of over 1 megabase (Mb = 1,000,000 bp): the bacteria *Haemophilus influenzae* (1,8 Mb), and the eukaryotic model *Saccharomyces cerevisiae* (12 Mb, ~6,000 genes) [22,23]. These events started the ‘Genomic Era’ that ended with the sequencing of the ~3.1 Gb (gigabases, 1 Gb = 10^6 bp), composing the first draft of the human genome containing ~20,000 CDS [24–26]. The European Nucleotide Archive (ENA) currently stores more than 30,000 completely assembled genomes, showing that current sequencing technologies have overcome limitations in resolution, cost and quality [26,27]. However, to estimate the number, location, and function of coding genes is still far from trivial and it requires iterative refinement and curation [28,29]. This task, known as genome annotation, is a key element in the study of function, structure, evolution, and editing of genomes, and their content; which ultimately provides a link between the genes of an organism (i.e. genotype) and the physiological features derived from them (i.e. phenotype) [30].

1.1.2. Protein functionality and biological complexity

Comparing sequenced genomes reveals a striking variety of sizes and numbers of CDS; with the eukaryotic domain presenting the highest values (Figure 1.2). This indicates that both genome sizes and CDS count play a role in increasing life complexity [31]. However, the rule does not apply at species level. For example, DNA content varies over 100-fold among herbaceous angiosperms DNA content varies well over 100-fold among diploid herbaceous angiosperms [32,33]. Also, the number of genes can be from just a few like in viruses [20], less than a thousand in bacterial endosymbionts or obligate pathogens [34,35], to a water flea reaching ~31,000 CDS [36]. Aside from gene number, complexity is consequence of the regulation of gene expression by different mechanisms ranging from the use of transcription factors to RNA molecules that do not encode for proteins (non-coding RNAS; ncRNAs), expressed from intergenic regions and overlapping with ORFs, that can have regulatory roles [10,37–39], or local changes in supercoiling [40]. In eukaryotes there is an added level of complexity due to alternative splicing [11]. However, proteins are the main biochemical and structural players in the cell so their identification in genomes is paramount to understand the range of biological functions an organism performs [41]. Essentiality classification has been the most direct way to understand how important a protein function is for a cell. Via gene knockout studies and/or random disruption with DNA elements [42,43], individual genes can be classified between those that are essential, thus required for the cell to survive; and non-essential, for those genes that in spite of being disrupted do not affect cell viability, and finally those that affect cell viability but still allows life (fitness). Of course the classification in these three categories can vary depending on the conditions of the experiment, except for some few genes that will be essential under any condition (e.g. RNA polymerase in prokaryotes). Studies comparing the intersection of these categories have highlighted biological functions which are intrinsic in every life domain [44,45].

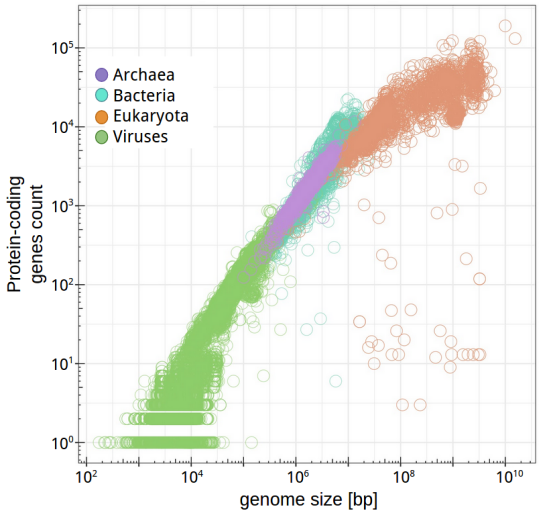


Figure 1.2. Log-log plot relating genome size with protein count
Scatter plot representing the genome size (X-axis) versus the number of protein-coding genes found annotated (Y-axis) for the complete National Center for Biotechnology Information genome database (NCBI). Different colors are used for Archaea (purple), Bacteria (blue), Eukaryota (orange), and viruses (green).
* Figure generated using saladi.shinyapps.io/Genome_size_vs_protein_count/

a) Replication and transcription

The first information process in the cell, and reason of inheritance, is DNA replication. This process is based on the DNA polymerase activity to copy the genome by complementary pairing of nucleotides present on each strand of the original DNA molecule [46]. Replication requires the formation of the DNA polymerase complex, also formed by enzymes to unwind the double-stranded DNA (e.g. helicases), relax the supercoiling (i.e. the property of DNA to wind over itself, controlled by topoisomerases), and ligate the DNA fragments (e.g. ligases) [47]. The second information step, transcription, is mediated by transcription factors and the activity of RNA polymerases (RNAP) to synthesize mRNA from genomic DNA. In bacteria, genes with related functions are commonly found, regulated, and expressed consecutively in genome regions so-called operons [48]. This mechanism produces polycistronic mRNA, which are single RNA molecules containing more than one gene, considering them as a transcriptional unit [49]. Transcription factors are in charge of controlling which operons are transcribed by specifically recognizing ‘promoter’ sequences in the DNA. For example, bacterial sigma factor 70 (σ^{70}) controls the expression of “housekeeping” genes, required to be expressed ubiquitously in the cell. Moreover, multiple variants can drive the expression of genes required for specific environmental conditions, such as heat or starvation stress [50,51].

The chromosome structure also plays a role in bacterial transcriptional regulation, in a dynamic process controlled by nucleoid-associated proteins (NAPs) that ensure the proper compaction of chromosomes in the cell [52]. For example, *Escherichia coli* presents its genome divided into 50 chromosomal interaction domains (CIDs), ranging from 40 to 100 kb, which are co-expressed [53]. This happens due to the proximity of those regions in three-dimensional space, not necessarily contiguous in the sequence, which make them accessible for the transcription machinery at the same time. Also, DNA supercoiling can regulate RNAP access in specific regions by packing the DNA and preventing its binding in a process mediated by topoisomerases [52]. Other factors described to affect transcription regulation include the initiating nucleotide (i.e. the first base being transcribed in the RNA). In this case, the availability of the complementary ribonucleotide in the cell can determine as well the expression of an RNA [54,55]. Also, metabolites can regulate the activation/inhibition of riboswitches, secondary structures in the RNA that can expose or hide different regulatory motifs depending on the concentration of a specific molecule [56]. The use of secondary structures in the RNA is also a regulatory mechanism in the process of transcription termination, where the formation of intrinsic terminators have been shown to impact transcription depending on environmental conditions [57]. Some RNA molecules need to be processed, or maturated, and at some point degraded. In this task, Ribonucleases (RNases) can catalyze the degradation or processing of RNA [58]. The recruitment of these RNases can be guided by non-coding or small RNAs (sRNAs) in bacteria. This family of RNAs have been reported to act as gene expression regulators at the transcriptional, post-transcriptional and translational level pairing with other RNAs in sense or antisense [59,60].

b) Translation

In the process of translation, mRNAs produced in transcription are used as templates to synthesize protein copies [61]. The machinery involved in this process includes ribosomes, which are rRNA-protein complexes composed of two subunits: the small subunit (the 30S in prokaryotes; 40S in eukaryotes), and the large subunit (the 50S in prokaryotes; 60S in eukaryotes) [62]. It also requires different tRNAs carrying the 20 amino acids in a “cognate” or specific pairing, known as aminoacyl-tRNA. Each aminoacyl-tRNA contains an anticodon, complementary to a codon, from which the genetic code is derived [63]. In the presence of an initiation factor (e.g. Translation initiation factor IF-1), the ribosomal small subunit complexes with a methionine-tRNA recognizing AUG codons in an mRNA found near RNA motifs that signal translation initiation (i.e. ribosome binding sites; RBS) [61]. Then, the large ribosomal subunit is recruited, unbinding IF-1. Elongation factors, such as EF-Tu in prokaryotes, bind the large subunit allowing the entrance of the required aminoacyl-tRNAs and proceed with

the polypeptide elongation [62]. During this step, the polypeptide chain is synthesized by 'reading' the mRNA three bases at a time by different sites that control the entry (A-site), peptide bond formation (P-site), and exit (E-site) of tRNAs in a cyclic manner (Figure 1.3) [61]. When a stop codon (UAA, UAG, and UGA) is found, the termination is mediated by release factors (e.g. bacterial TF-1), which are also codon-specific [64]. Also, it has been assessed the conservation of ribosome rescue factors, such as ribosome rescue factor A (arfA) or B (arfB) in bacteria. These proteins have the role of rescuing stalled ribosomes, blocked in the process of translation, due to the lack of a specific tRNA, collision against other ribosomes, or truncated RNAs [31,65]. This is highly important for the cell as protein synthesis is an energy-intensive event, due to the phosphate bonds expended for each peptide bond formed. Specifically, the delivery of each aminoacyl-tRNA by EF-Tu requires one Guanosine Triphosphate (GTP), and the translocation reaction consumes another [66]. Within the group of ribosomal proteins playing roles in this process, 34 are universally conserved across all domains of life [67].

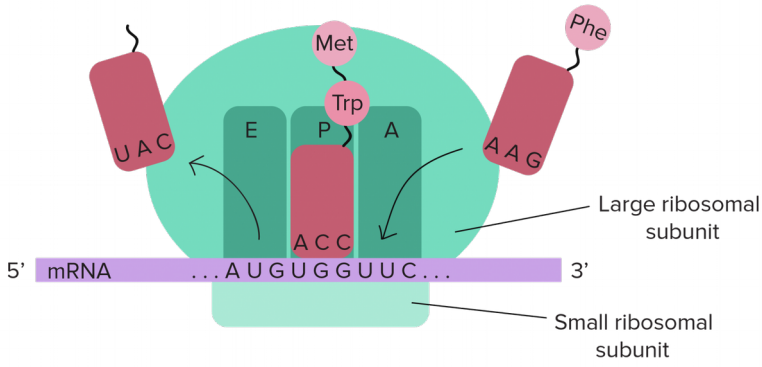


Figure 1.3. Schematic of the translation process
The ribosome begins translation at an AUG codon. Charged tRNAs (AAG-Phe in the figure) are loaded through the A-site while, in the P-site, the peptide bond between the nascent polypeptide and subsequent amino acids are formed (ACC-Trp+Met in the figure). The E-site serves as the exit point of uncharged tRNAs. *Figure adapted from "Translation: Figure 3," by OpenStax College, Biology (CC BY 4.0).

c) Homeostasis, metabolism, and structural roles

Once a polypeptide is synthesized, it generally requires to acquire a specific molecular structure to be biologically active [68]. Some proteins are able to fold in natural conditions but others require the action of chaperones, such as the groEL/groES system in bacteria [69]. These chaperone proteins guide the folding of newly translated polypeptides, identify miss-folded proteins, and prevent their aggregation. Thus, chaperone activity is considered an essential process in protein homeostasis [69,70]. Also, controlling the degradation of proteins is essential for cell homeostasis. In this task, proteases are responsible for processing and degrading proteins to peptides or amino acids with regulatory, recycling or homeostatic purposes [71]. For example, AAA+ proteases, ubiquitously found in every life domain, can degrade damaged and unneeded proteins, but also disaggregate and remodel them [72]. In general, proteolytic function in bacteria is highly redundant, and different proteins may serve for the same process and consequently being not essential for the cell. For instance, heat shock response proteases HslV and HslU, or proteases ClpP and ClpX are all homologs of the eukaryotic proteasome, which is the principal proteolytic machinery in this domain [73].This is not the case in genome-reduced bacteria like *Mycoplasma pneumoniae* where all proteases except one are essential delienating the minimal proteolytic machinery needed for life [74]. Some proteases may also be secreted to act as exotoxin, such as the exfoliative toxins in bacterial pathogens such as *Staphylococcus aureus*, which degrades extracellular structures of the host [75].

Metabolism, or the production of energy, is another requirement for the cell. For this task, enzymes (i.e. proteins with catalytic activity) interplay biochemical processes to synthesize and degrade biomolecules such as lipids, carbohydrates, coenzymes, amino acids, or nucleotides [76]. Generally, these reactions are mediated or have as aim the production either GTP or Adenosine Triphosphate (ATP), the principal bioenergetic molecules in the cell, via substrate-level phosphorylation and oxidative phosphorylation [77]. For example, in eukaryotes, the main amount of ATP is produced in the mitochondria, an organelle of endosymbiotic origin, by the ATP synthase using a proton gradient [78]. This protein, located within the inner membrane of mitochondria, represents one example of transmembrane protein [79]. These proteins perform functions such as granting direct communication with the environment, share molecules with neighboring cells, or provide fluidity to the cell membrane [79,80]. These proteins present hydrophobic amino acids which are stable in the lipidic membrane of cells [79]. Other proteins, like signalling peptides, are translocated to the membrane just to be processed and be delivered as a molecule with signalling or defense purposes [81].

Finally, certain protein sequences have been selected in evolution due to their structural properties, such as protein filaments forming the cytoskeleton, required to give shape and mechanical resistance to the cell [82]. Moreover, cytoskeleton is able to respond to different stimuli to make the cell contract, migrate or even form motility specialized structures such as flagella or cilia [82,83].

1.1.3. How small a protein can be?

The capacity of proteins to perform complex functions is implicitly related to their structure and the physicochemical properties of the sequence of amino acids composing them [45]. Traditionally it is assumed that for a protein to have function it needs to have a tertiary structure and it is known that intracellularly the smaller folded domains are around 60 aa in length (SH3, PDZ domains) with an exception like the WW domains (30 aa) [84]. For the known isolated proteins with tertiary structure (and not disulphide bridges) the majority are larger than 100 aa [85]. However, it was found that some very large proteins do not have a defined three-dimensional structure (e.g. Microtubule associated proteins), and that there could be very small peptides with functionality. For example, hormones and neuropeptides have an important signalling role that is performed with less than 100 amino acids, with very representative examples like insulin, the main anabolic hormone in mammals [86].

While these peptides derived from larger precursors by proteolytic processing, others can be directly encoded as small genes, of less than 300 bp or 100 aa (aa = amino acid length) [87]. For example, the majority of the *E. coli* proteins identified in the last 10 years have been those containing 50 or fewer amino acids, estimated to ~100 proteins in 2013 while now they are almost doubled, suggesting that this is the range where the most progress is needed [88]. The most extreme case of short protein is found in this same bacterium, with 7 aa, predicted to be membrane associated [88,89].

In this thesis project we will refer to genes encoding proteins less than 100 aa as small ORFs, or ‘smORFs’, and the abbreviation ‘SEP’ to refer to the protein encoded (from smORF-encoded protein). However, it has to be remarked the notation in the literature is inconsistent and alternative names such as ‘uORFs’, ‘sORFs’, or ‘alt-ORFs’ for small ORFs, and ‘miniprotein’, ‘micropeptide’ or ‘small protein’ for their products. As most of these smORFs can be found overlapping partially or completely with longer ORFs, the term ‘main-ORF’ will be used to designate these last. It is important to remark that SEPs are not the result of proteolytic processing and they are direct products of smORFs.

a) Regulatory smORFs

A common type of smORFs are found regulating transcription and translation of upstream main-ORFs. This is achieved by secondary structure changes in the mRNA induced by their translation, that hide or expose expression signals of the downstream main-ORF [90]. These smORFs, commonly referred as uORFs, can be found in up to 50% of mammalian transcripts [91]. Similar mechanisms have been described in bacteria, like the 'leader peptide' mechanism with examples in *E. coli* such as TrpL (14 aa), which regulates synthesis of tryptophan, or the pyr operon leader peptide (44 aa), regulating the pyrimidine biosynthesis [92]. In general, the translational product of these smORFs has no intrinsic function itself; however, exceptions have been described to play both regulatory and intrinsic roles in fruit flies development or in the cell cycle of plants [93,94]. Independent functions for leader peptides in bacteria have not been demonstrated [95]. Interestingly, a recent study has shown that mutation of 'minimal ORFs', composed of just two codons (start-stop), located out-of-frame of the main-ORF *yecJ* in *E. coli*, resulted in an increase in translation of YecJ. Thus, minimal ORFs could modulate translation as well [96].

b) Bacterial SEPs

Diverse smORFs have been experimentally demonstrated to encode for independent bioactive proteins (i.e. SEPs), commonly found by serendipity in gene screening analyses [97] (Table 1.1, Figure 1.4). One of the first SEP examples characterized in bacteria by loss-of-function mutations was SpoVM (26 aa), encoded by *spoVM*, a membrane associated protein that is essential for the sporulation process in bacteria like *Bacillus subtilis* [98]. In the same organism and process, CmpA (37 aa) was also found to participate in sporulation, but inhibiting it [99]. Other SEP examples include SgrS in *E. coli*, which is a 227-nucleotide RNA with two tasks in case of glucose toxicity: i) to hybridize to the *ptsG* mRNA inhibiting the translation of the glucose transporter PtsG; ii) to be translated to a SEP called SgrT (43 aa) able to block channels of active PtsG in the cell. This provides a robust mechanism to efficiently inhibit glucose influx [100]. The role of SEPs as membrane-associated proteins, anchoring, stabilizing and regulating membrane complexes, seems to be a common function of SEPs, participating in toxin/antitoxin systems (e.g. Hok, TisB, and PepA1) [101–103], transport (e.g. AcrZ, KdpF, and MntS) [104–106], cell division (e.g. MciZ and SidA) [107,108], and stress sensing and response (e.g. Prli, Brl, MgrB) [109–111]. Probably, the most paradigmatic organisms in this sense are cyanobacteria with photosynthetic complexes containing plenty of SEPs [112]. On the other hand, cytoplasmic SEPs have been also reported with functions such as chaperones in iron response (e.g. FbpB and FbpC) [113], being part of the ribosome complex [114], or attenuating ribosome-stalling by antibiotics (e.g. Prli53) [56].

Lastly, SEPs that are secreted have been associated with quorum signalling purposes [115] and antimicrobial activity (also known as antimicrobial peptides; AMPs). This last group includes SEPs secreted by pathogens to outcompete in specific biological niches like mammal microbiomes (i.e. symbiotic bacterial communities found in mouth, skin, gut, and respiratory, urinary and genital tracks) with role in infection [116–119], or protecting the host against pathogens [120] (Table 1.2, Figure 1.4). Interestingly, recent studies based on large-scale approaches suggest that this type of SEPs could be an important regulatory component in human microbiome homeostasis [121].

About the mechanisms of expression of SEPs little is known but it is considered that they are expressed by the same mechanisms observed in longer proteins [97,122,123]. On the other hand, cases of SEPs overlapping in sense to a main-ORF have been suggested to be produced by mechanisms such as ribosomal frameshifting, where the ribosome skips one or two ribonucleotides [124,125]. Another mechanism is the existence of internal ribosome binding sites in the transcript like is the case *Staphylococcus aureus*. In this case, the gene *sprG1* encodes for two versions of the same SEP but one is 13 amino acids shorter as it is expressed from an internal ribosome binding site and initiation codon (44 and 31 aa). While both are secreted to lyse human host erythrocytes, this task is more efficiently performed by the longer version [126].

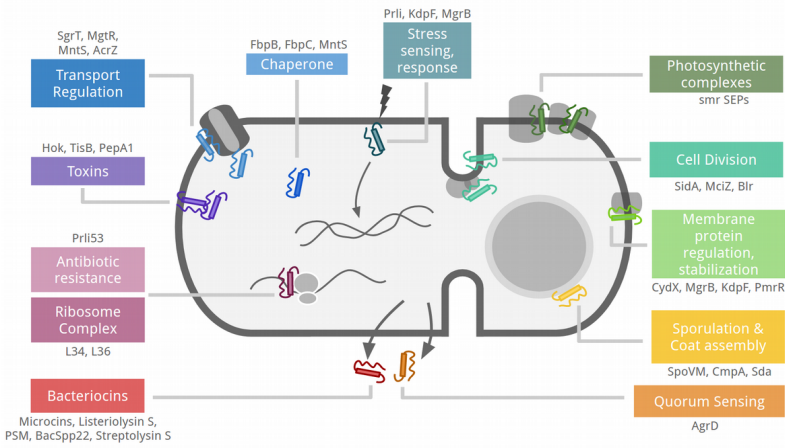


Figure 1.4. Location and function for SEPs characterized in Bacteria
In grey, the cell membrane is represented by a solid line, forespore as two grey circles (right), double-stranded DNA, and mRNA with ribosome bound (left). Transporters are represented as transmembrane proteins with a channel (top-left), transmembrane proteins as transparent grey capsules (top-right), and membrane associated as transparent circles. SEPs are represented as rectangles with tails in the same color as their categories. Below each category, examples of the SEPs are included (references in Table 1.1 and 1.2). Transmembrane SEPs are depicted traversing the membrane; if they are peripherally associated, only the tails traverse the membrane. Arrows within the cytoplasm represent signalling response, if they go out, secretion.

c) Eukaryotic SEPs

Relative to eukaryotes, studies in yeast show 247 SEPs, which loss led to lethality, slow growth, or temperature sensitive phenotypes [127]. In fruit flies, *tal* and *pri* genes encode for two SEPs required in development (11 aa, and 32 aa, respectively) [128]. In rodents, three SEPs (sarcolipin, phospholamban, and myoregulin) have been revealed as essential players in thermogenesis and muscle contraction by association to membrane transporters [129–131]. In humans, studies have reported 86 SEPs expressed from intergenic regions and ncRNAs, one of them demonstrated to start with an alternative ACG initiation codon [132]. The alternative start initiation in translation of SEPs has been also reported in leukemia cell lines and Hep3B cells [87,133]. Finally, not directly eukaryotic but still fundamental, mitochondria presents examples of SEPs such as humanin (24 aa), which protects cells from death and apoptosis through the inhibition of Bax, a pro-apoptotic factor [134]. The smORF encoding for humanin is contained in the mitochondrial 16S RNA and it has been related to the development of Alzheimer's disease [135]. Also localized in mitochondria, the *Boymaw* peptide is a SEP that can be found at high levels in post-mortem brains of patients suffering neuropsychiatric disorders such as inherited schizophrenia [136].

Bacteria	SEP	Function	Ref.
B. subtilis	SpoVM	Sporulation assembly point; SpoIVA recruitment; protease FtsH inhibition	[98,137]
	CmpA	Cortex assembly regulation in sporulation	[99]
	Sda	Inhibits kinases involved in regulation of sporulation	[138]
	FbpB, FbpC	Iron response; soluble chaperone	[113]
	MciZ	Inhibits FtsZ to prevent additional cell division	[107]
C. crescentus	SidA	Inhibits FtsW to prevent membrane constriction	[108]
E. coli	MntS	Manganese transporter MntP inhibition; soluble chaperone	[106]
	Blr	Antibiotic resistance and cell envelope stresses sensor	[110,139]
	AcrZ	Activates activity of AcrB-AcrA-tolC multidrug efflux pump	[104]
	SgrT	Inhibits glucose permease PtsG	[140]
	CydX	Activates the cytochrome bd oxidase	[141]
	MgrB	Represses PhoQ sensor kinase	[111]
	Hok	Toxin of type I toxin/antitoxin for plasmid maintenance	[101]
	TisB	Toxin of type I toxin/antitoxin involved in SOS-mediated response	[102]
	Ribosomal SEPs	SEPs in complex to 50S ribosome-EF-Tu	[114]
L. monocytogenes	Prli24, Prli42	Sensors and activators of stressosome in oxidative stress conditions	[109]
	Prli53	Controls antibiotic attenuation in response to lincomycin	[56]
M. bovis	KdpF	Stabilizes KdpABC; nitrosative stress resistance in replication inside macrophages	[105]
S. aureus	PepA1	Toxin of type I toxin/antitoxin	[103]
S. enterica	PmrR	Decreases lipopolysaccharides negative charge inhibiting LpxT	[142]
S. typhimurium	MgtR	Degradation of MgtC virulence factor by FtsH; Inhibits magnesium transporter MgtA	[143]
Synechocystis sp.	smr SEPs	Stabilize transmembrane complexes, Cytochrome b, NdhP, photosystem I, and II	[112]

Table 1.1. Bacterial SEPs, cytoplasm or membrane-located
Bacterial species, SEPs, function and reference where they have been characterized. Species abbreviations: *Bacillus subtilis, Caulobacter crescentus, Escherichia coli, Listeria monocytogenes, Mycobacterium bovis, Staphylococcus aureus, Salmonella enterica, Salmonella typhimurium, Staphylococcus pseudintermedius, and Streptococcus pyogenes.*

Bacteria	SEP	Function	Ref.
Gram+ bacteria	AgrD	Signalling peptides in quorum sensing	[115]
E. coli	Microcins	Bacteriocins produced to kill pathogens	[120]
L. monocytogenes	Listeriolysin S	Bacteriocin; involved in infection by killing host microbiota	[116]
S. pseudintermedius	BacSp222	Bacteriocin; involved in infection by killing host microbiota	[117]
Staphylococcus sp.	PSM	Phenol-soluble modulins; Toxin to kill skin pathogens	[118]
S. pyogenes	Streptolysin S	Bacterial cytotoxin	[119]

Table 1.2. Secreted Bacterial SEPs with characterized function
Bacterial species, SEPs, function and reference where they have been characterized. Species abbreviations: *Escherichia coli, Listeria monocytogenes, Staphylococcus pseudintermedius, and Streptococcus pyogenes.*

1.2. Gene identification: how to find the needle(s) in a haystack?

Despite SEPs being reported in different experiments, limitations in the approaches commonly used in genetic annotation and function studies have made researchers to generally ignore them, ultimately preventing their discovery and characterization [87]. In this section, we introduce the different approaches followed in gene identification: bioinformatic approaches (1.2.1-1.2.4); high-throughput analysis (1.2.5); and functional validation (1.2.6). These three independent elements can be considered sequential steps to follow in the task of gene identification. However once a gene is identified, the different approaches can feedback each other to ultimately support the characterization of a gene [87,144]. Due to the scope of this thesis project, the following sections will be mainly focused on prokaryotes. Despite most of the concepts, processes and limitations explained will be shared in eukaryotes, the splicing mechanism implies to consider signals and combinatorial approaches to represent the inherent complexity of an eukaryotic transcript which are not required in the study of prokaryotic genomes.

1.2.1. The ORF scanning concept

Bioinformatics, the branch of computational sciences oriented to solve biological problems, has been providing *Ab Initio* and *homology-based* methodologies for the annotation of genomes [145]. Their shared principle is the concept of ORF scanning, which consists in finding ORFs in genome sequences [146]. As assuming every ORF in a genome is coding for a protein would be ignoring that start and stop codons appear by chance in non-coding sequences, different assumptions are taken when defining the candidate ORFs to evaluate. The bioinformatic tools discussed are included in Table 1.3 at the end of Section 1.2.3.

Firstly, a minimal ORF length is required to not overpredict protein-coding genes. In a random DNA sequence (assuming 50% of GC-content), a stop codon is expected every 64 bp (21 codons). By cumulative probability, it can be estimated that probability to find a stop codon by chance under these conditions is maximum after 100 codons (Figure 1.5A). Thus, the probability of finding ORFs larger than 100 codons (100 aa) by chance is highly unlikely and they can be assumed to be CDS [146]. This concept has grounded the identification of ORFs in bacteria since the first bioinformatic tools, and in spite of being demonstrated as a good criteria in detecting bacterial proteins, this threshold is responsible for a drop at 100 codons when comparing ORFs versus ORF length [88].

Second assumption when creating the list of ORF candidates is that in case of multiple overlapping ORFs, in or out-of-frame in an mRNA, the longest ORF will

be the one considered to be translated [147]. This comes from the assumption that in a bacterial mRNA, the firsts RBS and start codon found by the ribosome will lead the translation process until the stop [148–150]. While true in most of the cases, proteins translated as a consequence of frameshifting and dual-coding genes exist in bacteria [124,126]. In addition, the partial overlap between adjacent genes is quite a common event in genome-reduced species such as *M. genitalium* and *M. pneumoniae* (Figure 1.5B) [151] .

Finally, the use of alternative start and stop codons, which number and frequency differs between species, has to be considered when generating ORF databases [146,152]. However, these codons are not selectively equivalent and this is complicated to include them in ORF scanning approaches [153]. For example, 80% of the ORFs in *E. coli* start with AUG, but also 15% and 5%, with GUG and UUG, respectively.

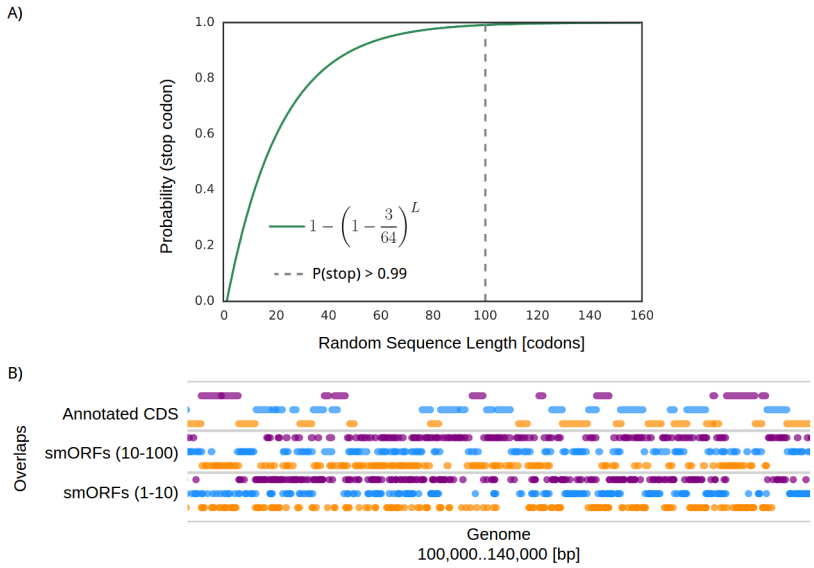


Figure 1.5. Stop codons by chance and overlap complexity in M. pneumoniae
(**A**) The probability to find a stop codon conditioned by length in a random sequence (solid green line) can be derived from a geometric distribution centered in the probability of finding a stop (3/64) and ORF length as exponent parameter (L; in codons). The probability to find a stop is total (>0.99) for random sequences longer than 100 codons (dashed grey line). (**B**) Example of overlap complexity in the genome-reduced model bacteria *M. pneumoniae*, only 5'3' orientation (i.e. positive strand) considering start codon AUG/GUG/UUG and stop codon UAA/UAG. Frame or 'phase' are represented by colors: phase 0 (frame in position n=0; orange), phase 1 (frame n+1; blue), phase 2 (frame n+2; purple). First track represents annotated CDS, second track represents not annotated smORFs (10-100 codons), and last track represents smORFs between 1-10 codons. Notice that even annotated presents overlaps.

1.2.2. *Ab Initio:* identifying genetic signals

Also referred to as '*de novo*', *Ab Initio* gene prediction relies on ORF scanning in combination to sequence inspection to detect gene signals and/or sequence content features; thus, they do not require any direct experimental input other than a sequence [154].

a) Sequence content

An ORF sequence itself can be used to explore functional and evolutionary determinants of coding potential [155]. From a functional perspective, RNA and proteins require specific conformation, derived from the biochemical features of their sequences, to perform specific functions [45,155]. From an evolutionary point of view, this implies that coding sequences will accumulate fewer mutations than non-coding regions, which should present random frequencies derived from the chemical environment of the DNA and the error characteristics of the DNA polymerase [156]. This is addressed in *Ab Initio* methods calculating the codon frequency of an ORF and comparing it to the frequencies observed in known genes or non-coding regions. Metrics like the Codon Adaptation Index (CAI), can be used to discriminate coding ORFs based on their resemblance to other genes in the organism, or bias from non-coding regions[157]. Implicitly, other features like GC-content of the ORF will also present different frequencies between coding and non-coding ORFs but sensitivity using the frequencies of 64 codons will be higher than using a four-letter code [148,158]. With the same in consideration, *Ab Initio* methods have extended the concept of codon frequency to dicodon frequencies, or higher orders, improving the identification power [159]. This is explained by the idea that evolution can select specific contiguous amino acids, or higher combinations, based on their physicochemical features and how they will interact in the protein structure conformation (i.e. protein domains) [160] .

Comparing putative ORF codon frequencies against non-coding sequences and/or known coding ORFs is a grounding principle shared by every *Ab Initio* bacterial gene predictor (Table 1.3). This criterion was used by pioneering tools such as *GeneMark* [148], or *Glimmer* [149], which are still kept updated [150,158]. Most of the tools using this approach rely on Markov models as statistical models, where the probability of each nucleotide/codon/k-mer depends on the state attained in the previous unit [149,158]. A limitation reported for these approaches include the incapability to evaluate short ORFs, which cannot reach a significant number of states in the model. This problem can be exacerbated by the effect of the GC-content: the higher this value, a higher threshold in size is required to not overpredict candidates deviating from the expected by chance [149]. For example, *Glimmer* indicates in their manuals that a high number of false positive predictions is expected for genomes with GC-content higher than 60% [149].

AMIGene, another tool from this family, exemplifies its size thresholds references as (GC-content between parentheses): *B. subtilis* (43%) - 500 bp; *E. coli* (51%) - 700 bp; *M. tuberculosis* (66%) - 900 bp [161]. Therefore, with the best conditions, this tool can evaluate ORFs larger than 160 codons, ignoring SEPs as a consequence.

b) Transcriptional and translational regulatory elements

Other coding signals can be derived from regions that are not part of ORFs themselves but their context. These regions of the ORF will not be represented in the product thus considered as untranslated regions (UTRs) [162]. For example, promoter sequences in bacteria are found preceding Transcriptional Start Sites (TSS), the point in genomic DNA where the transcript synthesis is initiated. Promoters are identified and bound by transcription factors, which also bind the RNA polymerase, to transcribe genomic DNA. For example, the aforementioned σ^{70} transcription factor recognizes two motifs: one 35 bases upstream the TSS, known as -35 box (TTGACA; *E. coli* consensus); and the other 10 bases upstream, known as -10 box or Pribnow (TATAAT; *E. coli* consensus) [163]. However, there are bacteria like *M. genitalium* or *M. pneumoniae* that do not have a -35 element at promoters [164].

On the other hand, transcription is found to terminate in Transcriptional Terminator Sites (TTS). TTS have been associated with two different termination processes: factor-dependent termination and intrinsic termination [165]. In the first case, bacteria rely on the Rho factor, a protein that disassembles the RNA polymerase complex recognizing C-rich sequences in the transcript [166]. In the second case, secondary structures formed in the RNA transcript that signals the termination of transcription in an intrinsic and protein-independent manner [167]. Interestingly, intrinsic terminators also have been shown to impact operon transcription responding to heat stress stimuli [57]. As TSS and TTS do not necessarily map to ORF coordinates, it is common to find UTRs in the 5’ and 3’ ends of mRNA (5’-UTR and 3’-UTR, respectively); in fact, it is in 5’-UTRs where Ribosome Binding Sites (RBS) to initiate translation can be found [168]. In bacteria, the most studied RBS is the Shine-Dalgarno motif, an RNA motif located around 8 bases upstream of the start codon AUG that pairs with the small subunit of rRNA [169]. In *E. coli* mRNA, the motif AGGAGGU was identified based on the paring with the 3’ end of its 16S rRNA subunit with sequence ACCUCCU, proving that this motif was the point of entrance for the ribosome to initiate translation of the mRNA to protein [168].

The most common applications of promoter and terminator mapping are applied to delimit operons in bacterial genomes or with vector design purposes despite some programs such as *FGENESB* using them as gene identification criteria

[170]. Their support is limited because TSS and TTS cannot provide direct information on the ORFs contained in a transcriptional unit and being expressed. On the other hand, RBS motifs were rapidly adapted into bioinformatics tools as a signal to consider, with tools such as *Prodigal*, *ORPHEUS*, *TICO*, or added to new versions of *Glimmer* and *GeneMark* (Table 1.3) [150,158,171,172]. These tools have been extensively tested and validated in *E. coli* and *B. subtilis*, which commonly have RBS motifs. However, reports show that RBS are not present in 20% of the genes (from a pool of 2,458 species), and another sequence(s) could be signaling translation initiation [173]. More extreme biases can be found when studying organisms such as *M. pneumoniae* or *Mycoplasma genitalium*, where the percentage drops below the 30% of genes including RBS motifs indicating that proteins can be expressed in a RBS-independent manner [174]. Thus, application of tools where the RBS is taken in strong consideration, such as *Glimmer*, results in a low prediction capacity in these species [150]. Concerning SEPs, a comparative with sets of SEPs and longer proteins from *E. coli* and *B. subtilis* has been done between *Glimmer*, *GeneMarkS*, and *ORPHEUS* showing constraints in the prediction of SEPs compared to longer proteins [158].

1.2.3. Homology-based gene identification

Homology-based gene prediction evaluates the shared ancestry between sequences by comparing them position by position (i.e. 'aligning') [175]. Two genes are homologs when they share the same or highly similar sequence. From an evolutionary point of view, this happens because of either a speciation event (orthologs) or a duplication event (paralogs) [145]. Consequently, if an ORF candidate shares sequence with one with annotated function, it can be assumed they are homologs and they will function similarly [175]. Mutations at the level of DNA can be synonymous, if they do not change the amino acid sequence, or non-synonymous mutations, otherwise. Thus, it is common to evaluate the ORFs converting them to amino acid sequences as the similarities or differences are more reliable; also because increasing it reduces the chances of spurious matches [145].

The most popular tool to evaluate homology against large databases is *BLAST* (Basic Local Alignment Search Tool) [176]. Given a 'query' sequence, it is evaluated by a sliding window, usually of 3 aa, comparing these subsequences to a database. If a 'hit' is returned (i.e. exact match), the position is extended evaluating contiguous windows and calculating an accumulated score that increases with the number of agreements. At the end of scanning the database, the results can be ranked statistically by an 'e-value', a metric that takes into consideration the accumulated score normalized by the probability to find the same match by chance in the database [176]. This directly limits the application with smORFs due to the high probability to match by chance with an unrelated

ORF segment. In these cases, factors like query and hit sharing the same sequence length have to be taken into consideration but reliability is still lower for SEPs than for larger proteins [121].

Commonly, homology-based approaches are not used independently but coupled to *Ab Initio* methodologies to support their predictions and provide an estimated function for each candidate [177]. For example, NCBI, the most important repository of genome sequences, provides the NCBI Prokaryotic Annotation Pipeline (PGAP), that can be run by request on genome sequence submissions [178]. The last version couples *Glimmer, GeneMarkS*, and homology searches against two databases curated to include only functionally annotated proteins: Clusters of Orthologous Groups or COGs, and NCBI Prokaryotic Clusters [179,180]. Other examples of annotation pipelines are *BASys* and *Prokka*, which use *Glimmer* and *Prodigal* for *Ab Initio* predictions, respectively, and *BLAST* to assign an e-value and function. Thus, these pipelines present the same aforementioned limitations predicting SEPs [181,182]. Furthermore, as the databases used in these approaches are derived from NCBI annotations, if an ORF candidate belongs to a not annotated family, it will not be identified.

Tool	Year	Type	Signals	Dependencies	Ref
GeneMark	1992	SC	-	-	[148]
GeneMark.hmm	1998	SC	-	-	[183]
Glimmer	1998	SC	-	-	[149]
ORPHEUS	1998	CM	RBS	DPS alignments	[171]
BLAST	1999	SH	-	-	[176]
COGs	2001	SH	-	-	[179]
AMIGene	2003	SC	-	-	[161]
GeneMarkS	2005	SC	5'-UTR motifs	-	[158]
TICO	2005	SC	RBS	Glimmer	[172]
EasyGene	2005	SC	-	-	[184]
BASys	2005	CM		Glimmer, BLAST	[181]
Glimmer3	2007	SC	RBS	-	[150]
ProtClustDB	2009	SH	-	BLAST	[180]
Prodigal	2010	SC	RBS	-	[185]
FGENESB	2011	SC	-	-	[186]
Prokka	2014	CM	RBS	Prodigal, BLAST	[182]
ZCURVE	2015	SC	RBS	-	[187]
PGAP	2016	CM	RBS	BLAST, COGs, ProtClustDB, Glimmer, GeneMarkS	[178]
CPC2	2017	CM	RBS	BLAST	[188]

Table 1.3. Bioinformatic tools for gene identification
Name of the tool, year of publication, gene identification criteria type: SC=Sequence Content *Ab Initio*; SM=Sequence Homology; CM = Combination of both approaches. If a tool considers genetic signals, the dependencies of some toolsuites and references are also included.

1.2.4. Machine Learning in the study of genetic signals

Artificial Intelligence (AI) advances can make computers to process and interpret biological data, learn from it, and take actions to maximize the chances of achieving a goal [189]. These algorithms are based on the definition of a mathematical model fitting patterns in the input data, to be used with predictive purposes like classification or regression [190] (Figure 1.6A). In the last decades, several AI models have been described with application in a vast range of disciplines. Between those, Machine Learning (ML) models have experienced the fastest adaptation in biology-related fields, from basic genetics to medicine, including the identification of genetic signals [190,191]. In this section we introduce the different types of ML: supervised, unsupervised, and reinforcement [192], and different examples of the type of information they can predict which can be used as evidence of existence or function of a coding gene.

Supervised learning relies on an input formed by input-output pairs (Figure 1.6B). In a computational biology context, this branch of ML is useful for both tasks such as prediction of categories (e.g. coding - non-coding) or quantities such as expression of a gene, respectively. Supervised learning algorithms like Random Forests (RF), which are based on averaging estimations from a pool of decision trees, have proven efficacy in diverse situations such as predicting protein interactions with other proteins, RNA, and/or DNA [193–195], distinguishing between productive and abortive promoters [164], or deciphering alternative splicing programs from -omics data [196]. Remarkably, RF algorithms can be used also to deconvolute the importance of the features based on how the estimations change including them or not [197]. Other models in this group are Support Vector Machines (SVM), which show good efficiencies in genomic studies like detecting transcriptional start sites in *Escherichia coli* [198]. SVM is the only ML model applied in the field of gene identification with CPC, which is used to distinguish between ncRNAs and mRNAs based on features extracted from homology searches, and ZCURVE [188].

When an input has not available or known labels, unsupervised learning models can cluster entries based on their features (Figure 1.6C). These algorithms excel in discovering hidden patterns, or data groups, resulting from similarities or differences in the information provided [199]. While algorithms like k-means, originally developed for signal processing, rely on vector quantization to separate clusters of comparable size within the population [200], others like Gaussian Mixture Models (GMM) try to deconvolute subpopulations following Gaussian distributions from a non-normal population, with no size dependency [201]. These methods are applied in different areas such as genomic population studies [202], disease subtyping [28], or discern when mitochondria were acquired by

eukaryotes [203]. Moreover, these approaches can be used to reduce the complexity of multivariable datasets, like those produced in -omics studies, with methods such as Principal Component Analysis (PCA) [204]. A k-means algorithm is fed with genetic signals and homology scores to cluster ORFs between coding and non-coding in the gene predictor program *AMIGene* [161].

The last group, relative to reinforcement learning algorithms, bases its power on the "exploration vs. exploitation" trade-off. In these models, 'agents' are designed to perform a series of actions with certain measurable variability and placed in a controlled environment. Then, interaction of thousands of those agents can be simulated and evaluated to use the best candidates performing a goal as 'parentals' of new generations. These iterations ultimately converge in an optimized set of features to perform the desired task [205]. These models have just started to be applied in the field of computational biology but they have already helped in the design of synthetic organisms with specific motility behaviors composed solely by cardiomyocytes and epidermal cells [206], the repurpose and discovery of antibiotics [207], and solving protein structures with unprecedented efficiency [208].

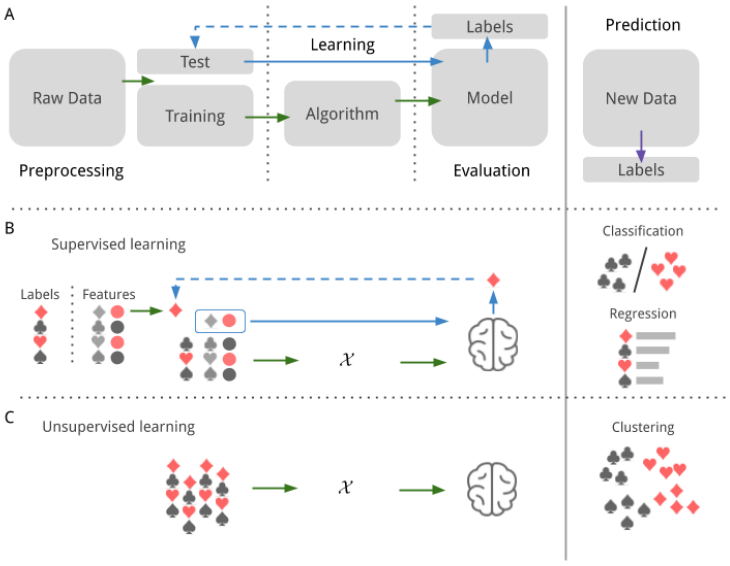


Figure 1.6. Machine learning approaches
(**A**) ML algorithms start with preprocessing raw data to satisfy modeling requirements (first column). Then, data is split into a training set, used by the model to 'learn' (green arrows), and a test set used to assess the model efficiency (blue arrows). Via iteration of these steps (blue dashed line), the model can be improved by training with different sets of data to make predictions from new data (right column, purple arrow). (**B**) In supervised learning, the user provides an input comprising features measured (exemplified as shape and color) for a set of entries for which the labels are known (card symbols). The model optimizes the task of predicting labels from the features after iterating splits of the data, ultimately returning a predictor able to perform classification and/or regression tasks. (**C**) Unsupervised learning can detect similarities and differences in the input data to define clusters grouping the data. These approaches do not implicitly require testing sets and they can be applied without preprocessing the data with clustering purposes.

1.2.5. High-throughput techniques to profile bacterial genomes

Instead of studying a single or reduced number of molecules, '-omics' techniques explore the whole set of molecular species (i.e. genes, transcripts, proteins, and metabolites) in a comprehensive manner, either independently or together. These approaches have as an advantage the consideration of the context of molecules in the experiment [209]. Genomics, transcriptomics, proteomics and metabolomics, have been the fields experiencing the biggest advances thanks to the appearance and optimization of high-throughput technologies.

a) DNA sequencing

Next-generation sequencing (NGS) refers to several standardized methods sharing features such as low-cost, large fragments sequencing capability, high-reliability, and massive parallelization of sequencing reactions (>10,000 kb/d). Therefore, all NGS methods are high-throughput techniques [27]. These methods rely on the random fragmentation and amplification of DNA ('shotgun'), and the application of bioinformatics assembly methods, to overcome the ~1 kb DNA chain length limitation of previous methods [210]. During the last decade, methods based on pyrosequencing [211], sequencing-by-synthesis [212], and sequencing-by-ligation [212], have significantly reduced the cost and increased the sequencing capacity to the order of gigabases per day, which makes these technologies to be referred as ultra-deep sequencing [27]. The last advances in the field belong to the *third generation* of sequencing methodologies which are able to sequence single molecules preventing biases derived from the amplification [213].

b) RNA sequencing and Ribosome profiling technologies

The transcriptome, or pool of RNA molecules in a cell, were initially studied in a high-throughput manner using microarrays, but the higher coverage and range of application provided by NGS technologies have made these last the current standard [214]. RNA sequencing (RNA-Seq) is supported by the same methodologies applied in DNA NGS technologies but with a previous step of RNA to cDNA by reverse transcription [215]. Current protocols allow the identification and quantification of RNA in a strand-specific manner and can be used to define TSS in a genome [216,217].

However, finding an RNA does not directly imply there is a protein being encoded in it. For this condition, ribosome sequencing (Ribo-Seq) has been

developed to capture mRNAs being translated (Ribo-Seq can be also referred to as ribosome profiling or ribosome foot-printing). By nucleolytic digestion, the RNA molecules in a sample are digested. In the case an mRNA is being translated, the regions with ribosomes will not be affected by the digestion [218]. Then, RNA-seq of those fragments provides a good estimation of the landscape of mRNAs. The data generated by Ribo-Seq, is analyzed in search of 3-nucleotide periodicity patterns to assess the synthesis of a protein. These codon patterns are well recognizable in eukaryotes, which have promoted its application in the detection of SEPs with enough resolution in fungi, plants, and mammals, including humans [219–222]. In contrast, codon resolution in bacteria is poor and highly variable and ribosome-stalling approaches are commonly required (e.g. using antibiotics such as tetracycline). By this approach, ribosome footprints have been detected for potential SEPs in *E. coli* including 312 smORFs [223], 120 smORFs translated in overlapping with known genes, 42 in-phase and 78 out-phase to main-ORFs [96], and another set of 41 smORFs, detected by Ribo-Seq and validated by chromosomal tagging, intergenic and in overlap with other genes [224]. The ribosome-stalling treatment has as a counterpart the detection of spurious ribosome binding to nonoptimal start codons which are unlikely to code for proteins [225]. Other limitations include the highly variable resolution between species studied, and low capacity in assessing overlapping ORFs [226,227]. Despite the limitations, Ribo-Seq methodologies can provide information even at population level, characterizing coding signals of SEPs in bacterial communities such as microbiomes [228].

c) *Transposon sequencing*

A common methodology used in bacteria to reveal the genes which encode for essential functions relies on coupling random mutagenesis to deep-sequencing technologies [43,229]. In transposon sequencing (Tn-Seq), a population of bacteria is transformed with transposon genetic elements. For example by using mini-transposons, which are suicide vectors that carry a transposase gene and a sequence to be inserted spanned between two inverted repeat (IR) sequence motifs [43]. Once the transposase is expressed, it will insert in a random position of the host genome the sequence delimited by IR and the plasmid vector in the cell will be inactivated [230]. Commonly, a bacterial population is transformed with this type of vector, cultured and selected by sequential passages to wash dead cells. Later, DNA sequencing, using specific steps enriching for sequences presenting the transposon, allows to determine the genomic point of insertion and how frequently it appears disrupted in the population [43]. If multiple insertions are found in a gene, it can be considered that it is not required for the cell viability, thus considered Non-Essential (NE). On the other hand, genes presenting clean profiles can be considered Essential (E) assuming that cells with insertions in them are non-viable and washed in passage selection steps [229].

For the proper analysis of this technique, is crucial to obtain the maximum number of initial insertions (i.e. coverage). The coverage is determined by the transformacion efficiency of the vector and the transposase used. Generally, two classes of transposases have been used for this kind of assays: Tc1/mariner-based, that only disrupt TA dinucleotides sites [231], and Tn5-based transposases, more recently applied, able to insert with no sequence constraints [232]. These differences are remarkable when considering the GC-content of the organism to be studied, as Tc1/mariner-based protocols will intuitively present less coverage for higher GC-contents. The coverage also determines the analyses performed in assessing the essentiality of genes. The higher the coverage, the higher resolution determining categories. Here it has to be considered the genome size as an additional factor, as larger genomes will require higher number of insertions to achieve significant coverages. This, together the differences derived from the transposase used, makes literature to report large deviations in terms of coverage. While in species like *E. coli*, a coverage of 10% of the genome can be achieved [233], in smaller genomes like *M. pneumoniae* a 41% can be recovered [234]. However, these are exceptions, as it is common to find libraries with small numbers of mutants even in closely-related species to those that present high insertion coverages.

For example, with the same technique in *Mycoplasma bovis*, only 319 insertions were determined (0.03% of coverage) [235]. While this number increased to 3,300 in *M. genitalium* (0.56% of coverage) [236]. Interestingly, in studies with high coverage such as *M. pneumoniae* example, or in Caulobacter bacterial species, a third essentiality category in-between E and NE can be observed. These genes are considered conditionally essential, or Fitness (F), and their disruption is viable for the cell but not as innocuous as NE [234,237,238]. In order to properly characterize these varieties of gene classes, the use of a ‘gold set’ of genes with known categories is commonly required, which prevents its direct application in genomes with little or no knowledge [239].

Additionally to the challenges derived from the aforementioned factors, Tn-Seq is still a recent development [237]. While standardized procedures have been established in DNA-Seq and RNA-Seq to preprocess the sequencing data and assess its quality, this does not happen in the field of Tn-Seq. This is also applicable at the level of essentiality estimation, where multiple models have been proposed but only reproducible in very specific conditions that prevent their general application out of conventional microbiology organisms models. For example, software tools to extract insertion profiles and posterior analyses from Tc1/mariner-based protocols exist [240–242], but these account for TA sites disruption sites so their application in Tn5-based transposon is inefficient and unsatisfactory. The lack of standardized protocols covering the different factors

related to the processing of Tn-Seq data, normalization, and essentiality estimation, prevent the general application of these types of studies and the development of new approaches using Tn-Seq.

d) Mass spectrometry

Mass spectrometry (MS) has been the main high-throughput technique in proteomics and metabolomic studies. This methodology measures the mass and charge of particles resulting from the ionization of a molecule, which can be used as a specific signal of that molecule in the sample with detection or quantification purposes [243]. For example, in 'shotgun' proteomics, a sample of proteins is digested with protocols such as tryptic digestion, which cleaves the C-terminal to lysine and arginine amino acids) producing tryptic peptides (TP) [244]. These peptides are ionized (e.g. bombarding them with electrons), which breaks them into smaller fragments that can be separated by their mass-to-charge ratio by accelerations and exposure to a magnetic field. A series of mass analyzers determine the spectra of the fragments which can be compared against a precalculated database of spectra and TPs with identification and relative quantification purposes [245]. Generally, only TPs that can be unambiguously assigned to a protein in the database are considered to validate the identification of a protein, these TP are known as unique tryptic peptides (UTPs) [246]. Thus, the larger the protein, the higher the probability to find associated UTPs. For shorter proteins, the number of possible UTPs is combinatorially reduced [247]. Also, Including all putative ORFs from a regular size genome in a MS search would result in reducing the potential number of UTPs for actual proteins to match the spectra, thus is not common to consider smORFs in these type of databases, which ultimately prevents the identification of new proteins and SEPs in conventional searches [244].

Moreover, MS is less sensitive in detecting hydrophobic TPs compared to hydrophilic peptides, a factor that also prevents the detection of transmembrane or membrane-associated proteins. If these membrane proteins are SEPs, their small size will make most of the SEP to be hydrophobic, thus less probable to be detected [248]. To overcome these limitations, protocolos like sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE) separation coupled to MS allow a higher recovery of hydrophobic peptides [249]. However, SEPs identification will still be limited by the number of UTPs in this condition.

The MS technique can be used also to detect and quantify a molecule with high sensitivity using labelling or targeted proteomic approaches [250,251]. These methodologies are based on growing cells or directly synthesizing peptides of interest with stable isotopes (e.g. ^{2}H, ^{13}C, ^{15}N and/or ^{18}O) which share the physicochemical, fragmentation and ionisation properties of their natural versions

[252]. Then, the presence and quantification of a specific molecule can be assessed if it shows identical mass-to-charge profiles but increased molecular when comparing non-labelled and labelled molecules (usually between 6 and 10 Daltons; Figure 1.7). The use of synthetic peptides has been proven to identify SEPs even in those cases where no peptides can be found in regular samples. The reason is that searches in targeted proteomics are heavily constrained to explore a specific spectra space, thus allowing a higher sensitivity detecting TP derived from the molecule of interest [253]. For example, Friedman and colleagues reported 17 SEPs of less than 50 aa, detected by isotope-labelling of synthetic peptides for the pathogenic bacteria *Helicobacter pylori*. Interestingly, all of them were encoded by previously considered non-coding RNAs [253]. As a disadvantage, identification of potential TP to synthesize and test by labelling is expensive in time and cost and each protein has to be assayed separately [132].

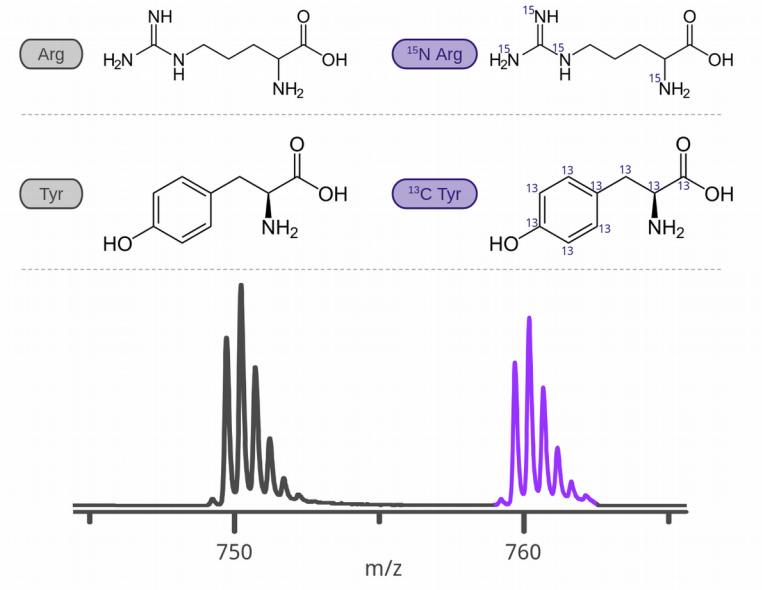


Figure 1.7. Identification of proteins by labelling peptide approaches
On top, two common examples of isotope-labelling with 15N and 13C. In this example labelled versions of Arginine and Tyrosine are presented (black for regular, purple for labelled). By synthetic peptide design different variants combining one or more isotopes can be applied. Bottom plot is a representation of a mass-to-charge (m/z) profile, showing a case where the spectra of a molecule matches between non-labelled (black) and labelled (purple).

1.2.6. Gene annotation validation

Technologies presented in the previous sections allow the validation of most of the predictions performed by bioinformatic approaches. However, multiple limitations have been discussed to prevent the experimental identification of proteins with specific features and also SEPs, and consequently, additional approaches can be used to validate these proteins [87]. For example, methods such as epitope tagging, consisting in fusing a 'epitope tag', such as fluorescent proteins or antigens, to a recombinant protein by means of genetic engineering. [89]. The epitope can be used later to validate the coding capabilities of a sequence considering that if the tag is detected, the protein is being expressed. One of the early examples of this technique was the A tag which is recognized by antibody immunoglobulin G, thus detectable by immunoprecipitation [254]. Later, multiple additional interaction-based tags were defined such as streptavidin or biotin [255,256]. For example, this last approach was used to identify the expression of 18 SEPs, one of them found in a dual-coding sequence, which are regulated by cell stress conditions, such as glucose starvation or heat stress in *E. coli* [100,257,258]. In addition, this methodologies can be used coupled to mass spectrometry approaches as it has been reported the fusion of epitopes can stabilize specific proteins that may be unstable or quickly degraded; however, they can also have the opposite effect and alter the function of the protein being tagged [87].

Despite the function for the protein encoded in a gene is commonly assigned by homology studies, a protein requires to be functionally validated as the last step in the characterization [87]. Generally, this process requires gene screening assays where the sequence of interest is over-expressed or inactivated to explore the phenotype produced in the population. Over-expression of a gene can be achieved by transforming the cells with a multicopy vector that can produce high levels of the protein encoded [259]. On the other hand, inactivation of genes can be achieved by different genetic engineering approaches consisting in disrupting the gene, either by techniques such as homologous recombination, or CRISPR/Cas9 [260,261]. However, these genetic tools cannot be used in a general manner and in cases like Mycoplasmas, alternative strategies such as Haystack mutagenesis. This protocol implies a laborious sequence of iterative PCR screening and ordered collection of pooled random transposon mutant libraries, until isolation of pure clones with the gene of interest disrupted are found [262]. Despite genetic screening having been the main source of functional SEPs discoveries presented in Section 1.1.3, it has to be remarked that expression of a protein could be dependent on very specific conditions, or be dispensable for the cell, thus complicating characterizing its function.

1.3. Applications in a genome reduced model

Understanding the set of proteins encoded in a bacterial genome can provide valuable insights on the range of biological functions that a species could perform. However, as introduced in previous sections, a hidden layer of complexity represented by SEPs has been ignored due to limitations in the computational and experimental methodologies. Thus, new methodologies are required in order to efficiently explore the uncharacterized SEPs a genome could contain [122,123]. While previous studies on SEPs have been performed in common bacterial and eukaryotic models, in this thesis we proposed *Mycoplasma pneumoniae* as a model organism to explore these family of proteins.

a) *Mycoplasma pneumoniae* biological features

Mycoplasma pneumoniae was isolated as a human pathogen proved to cause atypical pneumonia treatable with antibiotics [263,264]. This bacterium belongs to the Mollicutes class, group that includes other Mycoplasmas with parasitic or commensal lifestyles in a wide variety of species such as *Mycoplasma genitalium*, that infects human urinary and genital tracks [265]; *Mycoplasma gallisepticum* producing pneumonia in avian species [266]; or *Mycoplasma agalactiae* that infects ruminants [267]. Mycoplasmas are stained as gram-negative due to their lack of cell wall, however, phylogenetic studies showed they are more closely related to gram-positive Firmicutes [268]. Because of the cell wall being absent, Mycoplasmas are pleomorphic and different bacterial morphologies can be observed, ranging from coccus-shaped such as *Mycoplasma hyopneumoniae*, to flask-shaped morphologies like in the case *M. pneumoniae*, related to the structure of its attachment organelle (Figure 1.8) [269,270].

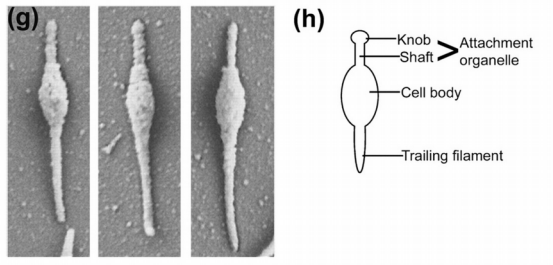


Figure 1.8. Scanning electron micrograph of M. pneumoniae
Left: M. pneumoniae cells grown on glass coverslips, approximate size 1-2μm long and 0.1-0.2μm wide. Right: schematic of the different cell parts. * Figure extracted from Figure 6 by Hatchel, J.M. and M.F. Balish (2008) [271].

Mycoplasmas share in common a genome downsizing process along evolution, termed degenerative evolution [272]. This is explained by the fact that most of the essential nutrients required by these bacteria are provided by the host cell during the infection process; as consequence, evolution has been permissive to the successive losses of genetic elements in this class [273]. Because of this, Mycoplasmas tend to present reduced gene numbers and genome sizes, with representative cases such as *M. genitalium* (580 kb; 525 CDS) or *M. pneumoniae* (816 kb; 689 CDS). At genome level, Mycoplasmas use a different genetic code compared to other bacteria. While the UGA codon generally encodes for a stop codon in prokaryotes, in Mycoplasmas it does for tryptophan [274]. Additionally, Mycoplasmas tend to present low GC-content genomes (e.g. 40% in *M. pneumoniae* - 50% in *E. coli*).

b) *M. pneumoniae* as Systems and Synthetic Biology model

The apparent simplicity in terms of number of genes compared to other bacteria, has made *Mycoplasma pneumoniae* an attractive Systems Biology model during the last decade. In the context of this branch of biology, which addresses the study of emergent properties derived from the interaction of the biological components in a system [275], genome reduced bacteria can be considered more approachable than other model organisms with higher number of genes; thus more predictable [276]. As it can be cultured in laboratory conditions, and a defined medium for growth of this bacterium is available [277], several studies have been performed to fully characterize the biology of this bacterium taking advantage of omics technologies. The transcriptome [278], the proteome [279,280], and the metabolome [277] were characterized ten years ago. The metabolome led to the definition of a flux balance analysis (FBA) model of the metabolism of this bacterium [281], while the proteome provided the basis to explore post-translational modifications in this bacterium [282]. Also, integration of proteomic and transcriptomic data, showed that majorly post-transcriptional, rather than post-translational mechanisms control the protein/mRNA ratios in *M. pneumoniae* [283,284].

In fact, this bacteria has been demonstrated to be a great model to study alternative transcriptional regulation. While regulation by transcription factors can explain most of the expression differences in bacterial models such as *E. coli*, the reduced number of transcription factors (i.e. nine), highlights the role of other mechanisms [57]. Recently, quantitative contribution of these mechanisms was characterized showing that only 20% of transcriptional regulation is mediated by canonical transcription factors and up to 70% of the total variance would be explained by different mechanisms such as supercoiling, metabolic control, RNA degradation, and chromosome topology, including transcriptional noise [285]. This is also supported by the genome three-dimensional structure, which has been

also characterized [40]. On the other hand, studies show that regulatory antisense non-coding RNAs in this bacterium are mainly products of transcriptional noise arising from spurious promoters [286]. Interestingly, using a random forest approach, productive and abortive promoters can be distinguished in *M. pneumoniae* [164].

Considering the lung pathogen nature of *M. pneumoniae*, the vast amount of biological information available, and its reduced genome, this bacterium has been proposed as 'chassis' for Synthetic Biology applications. This branch aims to the rational design of living systems to develop new biotechnological and biomedical applications such as the production of biomaterials or microbial therapies [287,288]. In this field, it is common to use the concept of chassis to refer to an organism with a genomic backbone, ideally depleted of irrelevant functions to its purpose (i.e. reduced genome), and additional modules to add novel functionalities [289]. Remarkably, the reduction point overlaps with a fundamental question in Systems Biology: which are the minimal number of genes required to sustain life?. In this question, transposon mutagenesis has been one of the most extended methodologies as it can provide the essentiality categories of genetic elements highlighting the genes which could be dispensable for the cell [43]. Transposon random mutagenesis presents a significantly higher resolution in *M. pneumoniae* compared to other organism models, reaching up to 4 bp for non-essential genes. With this high resolution coverage, comprehensive essentiality knowledge on the regulatory elements, UTRs, ncRNAs, ORFs and functional RNAs in this bacterium can be generated, in addition to characterize conditional essential genes [234]. These results are valuable in defining the fundamental genetic elements required for the system, which can be considered for the design of new chassis versions, even more considering the genome engineering tools developed for *M. pneumoniae* [290,291]. As example, genome reduction coupled to transposon random mutagenesis made possible to define the smallest self-replicating synthetic cell, derived from the close-relative to *M. pneumoniae*, *Mycoplasma mycoides*, with 473 genes, comprising 149 genes with unknown function, highlighting that genomic knowledge is still missing [276].

The vast amount of biological knowledge, available genetic tools, and potential capabilities in lung disease therapies presented by *M. pneumoniae*, makes it possible to envision a model-driven approach for the rational design of new versions of this bacterium. The comprehensive omic knowledge available can be integrated in a 'whole-cell model', a mathematical representation of the complete set of biological processes in the cell. This was achieved in *M. genitalium* with a multi-algorithmic model compiling 28 essential processes of the cell being able to predict the phenotype from single-knockout genotypes [292]. However, this model presented only 27.5% of the parameters extracted from *M. genitalium*, and despite the efforts of implementing the information known in *M. pneumoniae*,

technical advances, and better representation of specific processes, are still needed to define a final whole-cell model version of this bacterium [293]. Ideally, future advances could provide the capability to rationally design *M. pneumoniae* strains, in a computer-aided manner and supported by computational systems biology approaches, to develop desired functions in biotechnological or biomedical contexts with the help of genome engineering tools. Considering a 'whole-cell' model should inherently represent all the biology of an organism, the complete repertoire of SEPs in *M. pneumoniae* needs to be assessed in order to represent their roles in this type of models.

d) SEPs identification opportunities in *M. pneumoniae*

M. pneumoniae presents certain advantages for the study of SEPS compared to other bacterial species. First, complete ORF databases required in the study of small proteins are orders of magnitude smaller in genome reduced bacteria compared to other larger genomes, thus decreasing the space of sequences to evaluate [234]. Second, it is known that *M. pneumoniae* do not rely on Ribosome Binding Sites to initiate translation, despite the Shine-Dalgarno motif can be found in some genes located inside an operon [173]. Thus, theoretically, any ORF can be a potential coding gene. Third, the wide variety of omic datasets available for this bacterium provides an interesting benchmark to define the expression determinants for SEPs in each methodology, either known or non-annotated. Finally, the high coverage conditions observed when transforming *M. pneumoniae* with transposon random mutagenesis [234], allows the definition of up to 67 essential smORFs including examples such as MPN391a (30 aa), predicted to be involved in peroxide resistance [294], MPN347a (45 aa), as part of an anti-toxin pair [194], and MPN155a (90 aa) homologous to a putative RNA-binding protein, YlxR [295]. Remarkably, introducing these three smORFs to mass spectrometry databases results in peptides being recovered [234]. However, these advantageous coverage in transposon sequencing technologies are not commonly found in other bacteria [249,296,297], so further research on the standardization of these technologies is required to extend this application to other species.

In conclusion, *M. pneumoniae* is a good candidate for the definition of computational and experimental methodologies specifically prepared to assess the frequency and relevance for coding smORFs in bacterial genomes. These advances could increase and improve the quality of available genomic information in this model organism. Also, standardization of these tools could aid in the discovery of new small proteins in other bacterial species, which taking into consideration the functions SEPs can perform for the cell (e.g. *quorum sensing*, bacteriocins, chaperones, transport regulation, between others), could be exploited in new microbial therapies.

Chapter 2. Objectives

The objectives of this PhD thesis are: 1) To define new computational approaches for the efficient annotation of SEPs in bacterial genomes. 2) To critically assess the identification of small proteins by available technologies. 3) To standardize the bioinformatic analysis of transposon sequencing technologies, expanding their application in genome studies and improving essentiality studies. 4) To define a high-throughput experimental approach to identify expressed proteins, including SEPs.

In Chapter 3, we recapitulate the available knowledge on transcriptional regulation in bacteria and its relevance increasing the biological complexity of genome-reduced bacteria. Based on a bibliographical research, this chapter completes the introduction to gene expression regulation in minimal cells; highlighting the importance of regulatory elements other than transcription factors.

In Chapter 4, different ‘-omic' technologies, including mass spectrometry, RNA sequencing and ribosome profiling are evaluated to characterize their capability to recall SEPs. Limitations are evaluated and overcome by defining RanSEPs, a machine learning bioinformatic tool able to identify SEPs using species-specific sequence features, homology information and random forest models. We show that this approach predicts validated SEPs with an accuracy of 95%, outcompeting previous annotation algorithms. Moreover, running this tool in 109 bacterial genomes shows that the representation of SEPs in proteomes could increase from 10% to 25%. We also show that some annotated non-coding RNAs could encode for SEPs. A functional bioinformatic evaluation of the predicted SEPs highlights an enrichment in membrane, translation, metabolism, and nucleotide-binding categories; additionally, 9.7% of the SEPs included a N-terminus predicted signal peptide.

In Chapter 5, we present two different tools to aid the bioinformatic analysis of transposon sequencing (Tn-Seq) data. We present FASTQINS, a standardized pipeline to extract insertions profiles from raw data; and ANUBIS, a computational framework to cover every Tn-Seq data analysis step. Application of these tools under different sample conditions in *Mycoplasma pneumoniae* allow to recover unprecedented coverage levels (1.5 insertions per base resolution) which allow the characterization of specific artifacts. As a novelty in the field, we introduce a new model based on unsupervised clustering, to provide estimates without prior knowledge on the essentiality of the organism.

Finally, in Chapter 6 and 7 we present different applications resulting from the standardization of Tn-Seq approaches. First, in Chapter 6 we introduce ProTInSeq, a methodology to explore proteomes using ultra-deep sequencing and mutated transposon vectors where a resistance or marker is expressed only when inserted in-phase to an ORF. Preliminary results of this library indicate that it can be used to perform quantitative protein studies, reveal membrane topology features and also identify SEPs being expressed. Additionally to the ProTInSeq project, the approaches presented in Chapter 5 helped in different collaboration projects performed under this thesis project which are introduced in Chapter 7.

Miravet-Verde S; Lloréns-Rico V; Serrano L, 2017. Alternative transcriptional regulation in genome-reduced bacteria. Current Opinion in Microbiology 39:89-95

Chapter 3. Alternative Transcriptional Regulation in Genome-reduced Bacteria

3.1. Abstract

Transcription is a core process of bacterial physiology, and as such it must be tightly controlled, so that bacterial cells maintain steady levels of each RNA molecule in homeostasis and modify them in response to perturbations. The main regulators of transcription in bacteria (and in eukaryotes) are transcription factors. However, in genome-reduced bacteria, the limited number of these proteins is insufficient to explain the variety of responses shown upon changes in their environment. Thus, alternative regulation mechanisms could play a central role in orchestrating RNA levels in these microorganisms. These alternative mechanisms are dependent on cell metabolism (nucleotide levels), DNA topology and on intrinsic features within DNA and RNA molecules. This suggests they represent ancestral mechanisms shared among bacteria that have an increased relevance on transcriptional regulation in genome-reduced cells where the number of TFs is minimum. In this review, we summarize the alternative elements that can regulate transcript abundance in genome-reduced bacteria and how they contribute to the RNA homeostasis at different levels.

3.1.1. Highlights

- Genome-reduced bacteria have lost regulatory proteins acting at most regulatory levels.
- Minimal bacteria have retained sequence features to regulate transcription.
- Non-transcription factor regulation can occur at genome-wide, operon and transcript level.

Scan the QR code for full access to the original publication.

3.2. Introduction

Genome-reduced bacteria are of remarkable interest as model organisms to study basic aspects of bacterial physiology. Because of their inherent simplicity, they are attractive for systems biology studies, whose results can be generalized to larger, more complex bacteria. These organisms have encountered defined niches to colonize as endosymbionts or pathogens, and have adapted to their environments by eliminating genes that are not required for their development. For instance, they have usually lost metabolic pathways to synthesize elements present in their natural environment [277]. Also, this niche adaptation has affected how gene expression is regulated in these organisms. Transcription factors (TFs), which have been traditionally considered the major drivers of transcriptional regulation, are scarce in bacteria with small genomes. In bacterial models like *Escherichia coli* or *Bacillus subtilis*, TFs represent 5–6% their total number of genes. This number is reduced by half (2.5% on average) in the Mollicutes class, a bacterial group including multiple minimal bacteria, most of them Mycoplasmas [298] A comparative analysis of 50 Mollicutes genomes identified 1–5 global regulators and up to 15 TFs in the Mycoplasmas with larger genome sizes [298]. However, to the best of our knowledge, none of the putative global regulators has been characterized with the exception of the housekeeping sigma factor. Known transcription factors, including an additional sigma factor [299], only regulate a handful of genes [278].

Despite the tiny repertoire of TFs, these bacteria have not lost the ability to respond to a variety of external perturbations [278]. Therefore, it is possible that novel TFs remain undiscovered given the percentage of genes with unknown functions in these organisms, or that non-TF proteins with moonlighting functions act as TFs. Alternatively, different forms of regulating gene expression must exist, and may prevail, in these organisms. These alternative regulatory elements are probably not unique to genome-reduced bacteria, but they become more important as the process of genome reduction removes TFs to minimize the DNA content in these organisms. These alternative mechanisms of gene regulation are probably ancestral, as they are based in the chromosome structure and/or the intrinsic DNA or RNA sequences and not in proteins. The regulation they confer could have a smaller dynamical range and is more subtle than that by transcription factors, which makes it hard to observe in more complex bacteria. In this review, we focus on these other regulatory elements, from genome-wide to transcript-specific.

3.3. Results and discussion

3.3.1. Genome structure and DNA topology

First high-resolution 3D structure of a bacterial chromosome, obtained for *Caulobacter crescentus*, showed 23 interacting regions ranging from 30 to 400 kb bounded by highly transcribed genes, known as chromosomal interaction domains (CIDs) [300]. Lately, ~20 CIDs were defined in *Bacillus subtilis* with a size between 50 and 300 kb [301]. Disposition of these elements is regulated by DNA supercoiling, which is controlled by topoisomerases [302] and nucleoid-associated proteins (NAPs) [303] (Figure 3.1a). *B. subtilis* presents four DNA topoisomerases: two ATP-independent (I and III) and two ATP-dependent (II, known as DNA gyrase, and IV) [304]. Minimal cells commonly present no topoisomerase III and a significant reduction of NAPs [40,305]. With such a low number of DNA-binding proteins it was questionable whether small bacteria would preserve a chromosomal organization. A recent study in *Mycoplasma pneumoniae* found that small bacteria have enough components to maintain a defined chromosome structure and the presence of CIDs. In addition, this study provides the first evidence that genes inside CIDs tend to be co-regulated but the underlying mechanism to achieve this remains unknown. Interestingly, CIDs in *M. pneumoniae* are smaller (15–33 kb) but more frequent (44 CIDs) than *C. crescentus* and *B. subtilis* [40]. Additionally, promoters are sensitive to local superhelical states as it regulates the distance between the elements participating in the promoter [306]; even in small-genome bacteria with reduced number of topoisomerases [285,307]. Finally, ATP controls the ratio of ATP dependent/independent topoisomerases with direct effect on supercoiling and could imply a regulatory link between metabolism and genome topology and, consequently, expression [308].

3.3.2. Genome organization in operons

Genome organization in operons constitutes a first level of gene regulation in prokaryotes. As transcription and translation occur simultaneously in bacteria, positional effects exist, and expression levels of the individual proteins in an operon are inversely proportional to the distance to the transcription initiation site of the operon [308]. This represents a level of regulation that is used not only in small but in all bacteria. Traditionally, operons have been treated as static entities. However, recent research has shown that these structures are highly dynamic, being able to adapt in response to changing conditions, mainly thanks to termination, generating large transcripts or super-operons in some conditions, while producing short transcripts of sub-operons in others (Figure 3.1b) [309]. In M. pneumoniae, this condition-dependent transcriptional read-through can

explain a large part of how transcription is regulated [309]. This mechanism has been shown to occur also in larger bacteria such as E. coli and B. subtilis [310].

3.3.3. Bacterial promoters and transcription initiation

Promoter regions require certain features that make them recognizable by the RNA polymerase (RNAP) and the different TFs. Besides specific motifs binding sites for TFs, the most important sequence features are the boxes recognized by the RNAP complex and the different sigma factors. The housekeeping sigma factor binds two regions: the −10 box or Pribnow motif, and the −35 box. In genome-reduced bacteria, promoters have evolved towards the elimination of the −35 box, as this is non-existent or highly degenerated (Figure 3.1b) [164,303,311]. In *Buchnera aphidicola*, an aphid symbiont with a minimal genome, regions similar to the −10 box of *E. coli* have been found, while a −35 motif has been only found upstream the rRNA genes [312]. In Gram-positive bacteria like *B. subtilis*, absence of a −35 element has been shown to be compensated if the Pribnow motif is preceded by a 'TG' dinucleotide (the so-called extended −10 box), but this short motif is present in only a handful of promoters in *Mycoplasma gallisepticum* [311] and is not essential in determining promoters in *M. pneumoniae* [164]. This reduction in promoter complexity could be due to the scarcity of alternative sigma factors. This raises a question as to what makes promoters determine initiation of transcription and recognition by the RNAP complex. A recent study in *M. pneumoniae* points to the importance of the bases immediately surrounding the Pribnow motif, which tend to be A/T rich [313].

The structure of these regions is also important to trigger transcription. The double-stranded DNA should be less stable at the promoter region to unwind and accommodate the RNAP complex. Although the unwinding of the double helix is energetically favored at the promoters, the open complex formed between the promoter and the RNAP can be unstable. Unstable complexes require high concentrations of the initiating NTP (iNTP) to be stabilized so that RNA synthesis can be launched immediately. Otherwise, these complexes rapidly dissociate and transcription initiation is not produced. In contrast, very stable complexes require lower concentrations of the iNTP, as they will not easily dissociate [54]. Later, it was shown that the +2 nucleotide also modulates transcription initiation [314]. This mechanism establishes a link between cellular metabolism and transcriptional regulation and is not unique to genome reduced bacteria, but in the absence of major regulators this might be an elegant way to coordinate the expression of large groups of transcripts with identical +1 and +2 bases. An example of this nucleotide-based regulation includes the response to amino acid starvation (stringent response) in *B. subtilis*. In this scenario, concentration of ATP increases while GTP decreases as a consequence of the synthesis of

(p)ppGpp (Figure 3.1b) [315]. Upregulated genes in this condition have adenosine in the +1 position, while downregulated promoters have guanosine. This effect could also be present and play a major role in the absence of many TFs in minimal bacteria as a regulatory mechanism dependent only on sequence composition.

3.3.4. Termination

Transcription termination in bacteria can be accomplished by rho-dependent or intrinsic termination (IT). First type involves Rho protein moving through the nascent RNA and disassembling the transcription machinery [316]. IT depends on terminator sequences composed of a stem-loop hairpin followed by a poly-uridine (poly-U) tail. Poly-U induces the RNAP backtracking towards the nearest hairpin that disintegrates the elongation complex [317]. Rho is usually essential in Gram-negatives but Gram-positive model cells are viable without it [318,319]. Remarkably, few species directly lack this gene and any homolog: *Streptococcus pneumoniae*, *Streptococcus pyogenes*, *Mycoplasma genitalium*, *M. pneumoniae*, *Ureaplasma urealyticum*, and *Synechocystis* sp. PCC6803 [320]. Species in this group are all Gram-positive, present low GC contents and, except *Synechocystis* sp., have genome sizes between 0.5 and 2 Mb. Interestingly, low GC content has been presented as an impediment to form stable terminators. Analysis carried in several prokaryotes, including Rho-lack Mycoplasmas, showed that no free energy minimum to form hairpins is achieved close to stop codons although termination still occurs [321]. This could imply the existence of a third unknown mechanism that could be especially relevant in Rho-lack organisms, which are most of them genome-reduced bacteria.

IT regulation mainly relies on hairpin stability and poly-U length but three additional elements need to be considered. Firstly, low uridine triphosphate (UTP) concentration helps termination [322]. Secondly, elongation factors modify RNAP processivity and its sensitivity to terminators and they are reduced in minimal cells, like NusG or NusB, inexistent in most of them [305,323]. Finally and as mentioned above, IT can be condition-dependent as cases of readthrough and imperfect termination as response to different environmental stimuli have been observed [309,310,324] (Figure 3.1b).

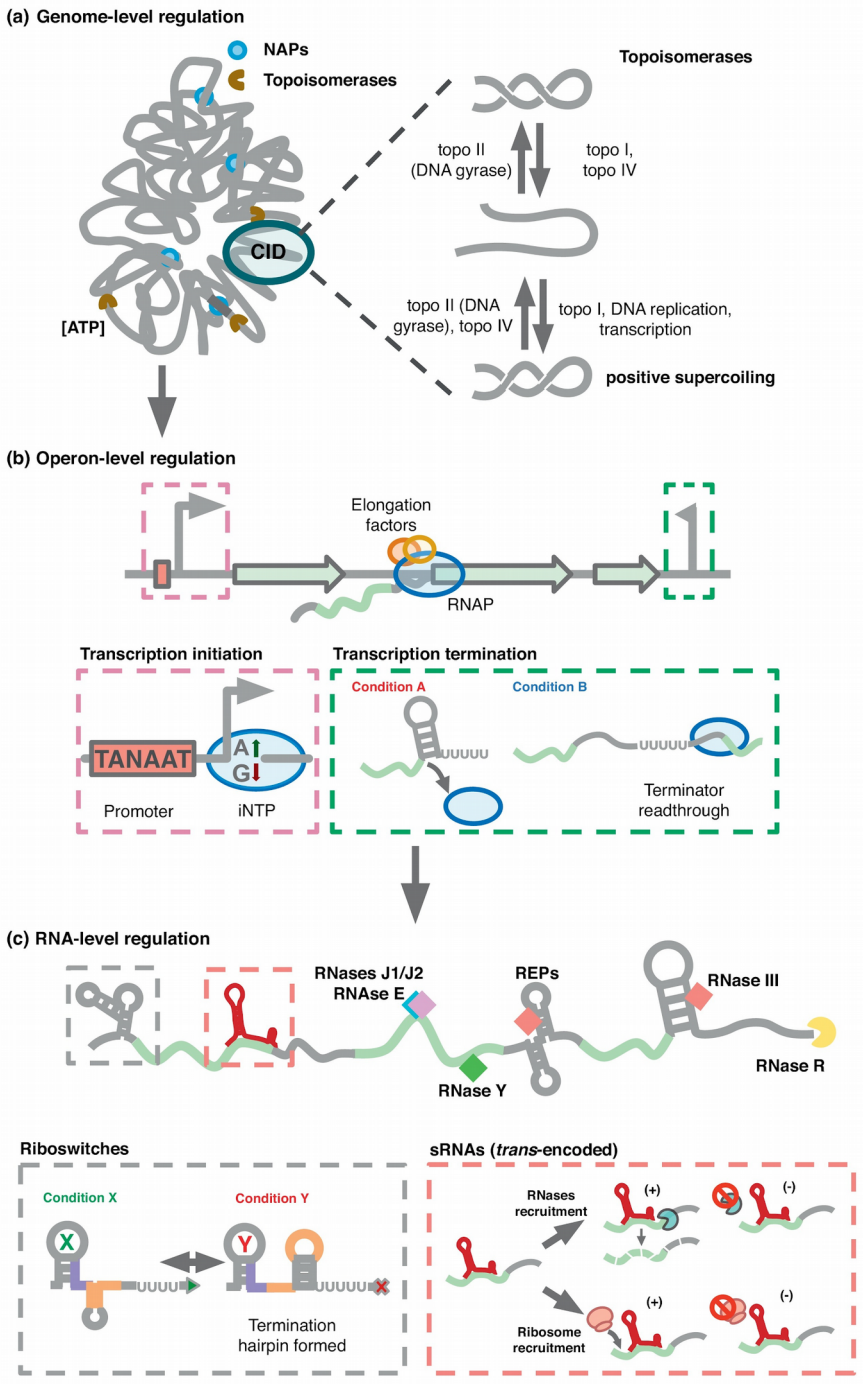

(a) Genome-level regulation
NAPs
Topoisomerases
Topoisomerases
topo II
(DNA gyrase)
topo I,
topo IV
CID
[ATP]
topo II (DNA
gyrase), topo IV
topo I, DNA replication,
transcription
positive supercoiling
(b) Operon-level regulation
Elongation
factors
RNAP
Transcription initiation
Transcription termination
Condition A
Condition B
TANAAT
A
G
Promoter
iNTP
Terminator
readthrough
(c) RNA-level regulation
RNases J1/J2
RNAse E
REPs
RNase III
RNase R
RNase Y
Riboswitches
sRNAs (trans-encoded)
Condition X
Condition Y
X
Y
Termination
hairpin formed
RNases
recruitment
(+)
(-)
Ribosome
recruitment
(+)
(-)

Figure 3.1. TF-independent regulation of transcription at three different levels observed in genome-reduced bacteria.

(**a**) Genome-wide level. At this level the principal actuators are the genome structure, organized in chromosomal interaction domains (CIDs) and maintained by nucleoid-associated proteins (NAPs); and the supercoiling, regulated by gyrases and topoisomerases (topo). ATP affects supercoiling through the regulation of the activity of ATP-dependent (in contrast to ATP-independent) topoisomerases. (b) Operon-level regulation. At this level, we consider how transcription is initiated (pink box) and terminated (green box), and the elements that regulate these processes including promoters, iNTP (initial nucleotide triphosphate, +1 position; A/G represent iNTPs that can control initiation rate) and intrinsic terminators that can be ignored due to specific environmental stimuli and in relation to different elongation factors (yellow and orange circles). (c) At the RNA or transcript-level regulation we encounter the effect of termination-related riboswitches (grey box) and sRNAs (red box). Riboswitches include those that promote or avoid premature termination in different conditions (X or Y) after being activated by presence or absence of a compound. sRNAs can control mRNA degradation by favoring or impeding recognition by different RNases, and can also regulate the recruitment of ribosomes that, besides affecting translation, may also affect mRNA stability.

3.3.5. Riboswitches

Riboswitches are segments within an mRNA that bind metabolites triggering a structural change that affects the encoded protein expression. This effect is a direct consequence of hiding or exposing terminators or ribosome binding sites [325]. Affecting transcription, only ribo-regulation based on termination has been defined with riboswitches usually located within 5′-UTR regions of metabolic genes and controlled by metabolite ligands appearing in the same pathways of the genes they regulate [326]. When active, they transform an anti-terminator in an intrinsic terminator, producing a premature termination, or vice versa producing readthrough (Figure 3.1c). Multiple cases of this ribo-regulation have been defined with strong importance in bacterial physiology and virulence [56]. Despite knowledge about ribo-regulation in small bacteria is still narrow, they are good alternatives to regulate genes in a TF-independent manner and, as occurs in termination, saving genomic space with a mechanism embedded into the sequence itself. As example, we know that multiple metabolic pathways in minimal cells are reduced to the core as they receive multiple resources from the host they parasite and some of these pathways commonly include regulation by riboswitches [327,328]. More interestingly, there are cases where ribo-regulation in small bacteria has evolved to high levels of complexity. One example includes multiple variants of guanine riboswitches found in the genome-reduced bacterium *Mesoplasma florum*, that are not seen in other organisms [329].

3.3.6. Small RNAs

Non-coding or small RNAs (sRNAs) in bacteria have traditionally been thought to act as gene expression regulators, either at the transcriptional, post-transcriptional or translational level [330–332]. The 6S RNA, which directly regulates the activity of the RNAP, is found only in Rickettsias, but not in other genera of genome-reduced bacteria such as *Buchnera* or *Mycoplasma* [333]. Despite the variety of possible mechanisms of action described for sRNAs (Figure 3.1c), only a minority of the discovered sRNAs have been characterized, most of which correspond to the *trans*-encoded sRNAs located in intergenic regions. In some *Mycoplasma* species, different intergenic sRNAs and their targets have been annotated using *in silico* approaches, and some of them have been found to be transcribed differently in various conditions [334]. In *Rickettsia conorii*, interaction between an intergenic sRNA and its targets could be experimentally validated [335].

However, genome compaction in small bacteria has caused intergenic regions to shrink substantially, therefore reducing the number of *trans*-sRNAs [286]. In contrast, genome-reduced bacteria are rich in *cis*-encoded antisense sRNAs. The functions of these have not been studied in depth, but a recent study provides evidence that in *M. pneumoniae*, most antisense RNAs could be the product of pervasive transcription arising at spurious promoters [286]. The low information content of promoters in these organisms, probably associated with the decrease of sigma factors, together with the higher probability of mutations from G/C to A/T in bacteria [336] could allow for a rapid formation of novel functional promoters, giving rise to these non-coding transcripts.

Indeed, the number of antisense transcripts in bacteria correlates with the AT content of their genomes [286]. Although this suggests that the individual antisense RNAs do not have a regulatory function, this pervasive transcription could have a role in generating variability in the bacterial population, and probably this phenomenon is not unique to genome-reduced bacteria. Other classes of small, non-coding RNAs are TSS-associated RNAs, that have been found in Mycoplasmas [217] and have been hypothesized to prevent transcription elongation until the correct RNAP complex has been assembled.

3.3.7. Post-transcriptional regulation

Maturation and degradation are essential events controlling RNA concentration catalyzed by enzymes with different specificities between stable (rRNA, tRNA) and messenger RNA (mRNA) including: exoribonucleases digesting from one end of the molecule and endoribonucleases cleaving the RNA internally [337].

Degradation and maturation start with an endoribonucleolytic primary cleavage as exoribonucleases cannot target newly produced RNA [338]. At this level, Gram-positive minimal cells do not show a clear reduction and they conserve the most important endoribonucleases found in *B. subtilis*: RNases III, H, J1, J2, P, Y and NrnA [339,340]. After a cleavage, stable and mRNA can be digested from 3′-end by PNPase, RNase R, YhaM, YhcR and YvaJ exoribonucleases in *B. subtilis* [340]. From 5′-end, RppH can remove the phosphate protection and trigger a rapid degradation by RNase J1 acting as 5′-exoribonuclease [340]. Unlike endoribonucleases, minimized bacteria show a strong reduction in 3′ exoribonucleases with only RNase R conserved. In addition, the lack of RppH and J1 acting as exoribonuclease being not proved yet in small bacteria, make unlikely their 5′-exoribonuclease activity [339–342] (Figure 3.1c).

In RNA maturation, participants in degradation, RNases III and YhaM, and specific RNases Bsn, M5 and PH, complete the task in *B. subtilis* [343]. No specific enzymes have been found in small Gram-positive bacteria and degradation-related RNases (III, P and R) participate in maturation of stable RNA [344]. In addition, ribonucleases can interact in a complex specialized in degradation and processing called degradosome. Gram-positive degradosome consists of three RNases (J1, J2 and Y), PNPase, CshA (RNA helicase) and two glycolytic enzymes (enolase and phosphofructokinase) [345]. Degradosome in small cells remains undefined and lack of PNPase is an important limitation; however, either as a remnant or because degradosome exists in small bacteria, same glycolytic enzymes than in *B. subtilis* interact in *M. pneumoniae* [346].

3.3.8. REP elements

Repetitive Extragenic Palindromic (REP) elements are species-specific conserved sequences that form RNA secondary structures. These elements were first found in *E. coli* representing close to 1% of its genome [347]. Lately, these have been characterized in minimal cells with multiple examples in *Mycoplasma* spp. [348].

Regulatory spectrum of REP sequences has not been fully explored yet but they seem to act at many levels during transcription. Firstly, REP elements preferentially bind gyrases so their effect could be extended to affect DNA supercoiling with its respective impact in transcription regulation [349]. Secondly, their recurrence within intercistronic regions has been associated with regulation of relative expression of genes within the same operon [350]. Finally, they can interact with the degradation machinery as potential target of RNaseIII and, due to their common presence upstream to terminators in the 3′-untranslated region (3′-UTR), it has been suggested REP elements could protect mRNA against 3′–5′ exoribonucleolytic activity [337,347].

3.4. Conclusion

Minimal bacteria arisen by degenerative evolution have in common a significant reduction of the number of proteins encoded within their genomes. This reduction implies a lack of multiple TFs resulting in an increased relevance of TF-independent transcriptional regulation at genome-wide, operon and transcript levels. At the genome-wide level, minimal cells have conserved a minimal set of proteins to maintain a structured genome and to control its superhelical state, both with direct effect on transcription. In addition, supercoiling could have an extended regulatory role in genome-reduced bacteria, as the high level of compaction makes it more likely to affect several operons with single local adjustment of the superhelical density. A second level based on operon organization comprises regulation of transcription initiation and termination. At this level we observe the impact of DNA sequence composition in transcriptional regulation with promoter motifs and the effect of the iNTP on initiation. We also find phenomena such as transcriptional read-through, with special relevance in minimal bacteria due to their high degree of compaction. Last type of TF-independent regulation occurs at the transcriptional level, where RNA structures that are part of the mRNA itself or additional interacting RNAs could critically impact the functional RNA concentration. For instance, riboswitches are a good alternative to regulate metabolic genes and operons based on the structure of the 5′ end of the mRNA. Transcript-level regulation also includes degradation of the mRNA where REP sequences (mRNA itself) and additional sRNAs interacting with it participate in the recruitment of ribonucleases, controlling the mRNA availability.

Throughout this review, we have noted that besides TFs, the specific protein machinery of these alternative regulatory elements has also been reduced (NAPs, elongation and termination factors and exoribonucleases have been lost in these organisms). Nevertheless, the core functionality remains, as cells are capable of displaying different transcriptional responses to perturbations. This functionality thus could rely on single proteins with moonlighting functions (e.g. RNase R) or more interestingly, on mechanisms that are implicit to the genome features or RNA molecules themselves without requiring any encoded protein. We believe that these could be ancestral mechanisms, that are not unique to genome-reduced bacteria, but that can be observed and studied in these organisms because the lack of TFs makes them more relevant. A question that remains open is to which extent each of the alternative mechanisms is responsible for the RNA regulation inside the cell, that is, how much of the RNA dynamics can be explained by each of these elements. Currently, there is no framework that allows to integrate the effect of the different mechanisms, but recent advances in modelling approaches, such as multi-scale models or even whole-cell models, could shed light on this question.

3.5. Author contributions

SMV and VLLR contributed equally to reviewing bibliography, writing and making figures. LS reviewed and supervised the work.

3.6. Acknowledgments

We acknowledge Dr. Eva Yus and Dr. Maria Lluch-Senar for providing valuable comments to the manuscript. We acknowledge support of the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program under agreement No 670216 (MYCOCHASSIS) and the Spanish Ministry of Economy and Competitiveness, 'Centro de Excelencia Severo Ochoa 2013-2017′.

3.7. Further research on alternative transcriptional regulation

Hypotheses and mechanisms presented in this work were later validated in the work '*Determination of the Gene Regulatory Network of a Genome-Reduced Bacterium Highlights Alternative Regulation Independent of Transcription Factors'* by Yus & Llorens-Rico, *et al.* (2019) [285]. In this work, authors quantitatively evaluate the importance of the different transcriptional mechanisms presented in this chapter in *M. pneumoniae*, by integration of multiple genetic and environmental perturbations. After testing 143 genes out of the 689 annotated proteins in this bacterium, they show that only 55% alter the phenotype, highlighting the robustness of the system. This study identifies nine transcription factors, their targets, and 16 proteins regulators, independently affecting transcription. Remarkably, only 20% of transcriptional regulation is mediated by canonical transcription factors. The contribution of different mechanisms such as supercoiling, metabolic control, RNA degradation, and chromosome topology to transcriptional changes are evaluated by using a Random Forest, explaining up to 70% of the total variance. These results highlight the importance of considering alternative transcriptional regulation when engineering bacteria.

Miravet-Verde S; Ferrar T; Espadas-García G; Mazzolini R; Gharrab A; Sabido E; Serrano L; Lluch-Senar M, 2019. Unraveling the hidden universe of small proteins in bacterial genomes. Molecular Systems Biology 15(2):e8290

Chapter 4. Unraveling the Hidden Universe of Small Proteins In Bacterial Genomes

4.1. Abstract

Identification of small open reading frames (smORFs) encoding small proteins (≤ 100 amino acids; SEPs) is a challenge in the fields of genome annotation and protein discovery. Here, by combining a novel bioinformatics tool (RanSEPs) with "-omics" approaches, we were able to describe 109 bacterial small ORFomes. Predictions were first validated by performing an exhaustive search of SEPs present in *Mycoplasma pneumoniae* proteome via mass spectrometry, which illustrated the limitations of shotgun approaches. Then, RanSEPs predictions were validated and compared with other tools using proteomic datasets from different bacterial species and SEPs from the literature. We found that up to 16 ± 9% of proteins in an organism could be classified as SEPs. Integration of RanSEPs predictions with transcriptomics data showed that some annotated non-coding RNAs could in fact encode for SEPs. A functional study of SEPs highlighted an enrichment in the membrane, translation, metabolism, and nucleotide-binding categories. Additionally, 9.7% of the SEPs included a N-terminus predicted signal peptide. We envision RanSEPs as a tool to unmask the hidden universe of small bacterial proteins.

4.1.1. Synopsis

RanSEPs is a random forest-based computational approach capable of predicting small encoded proteins in a species-specific context. Running this tool in 109 bacterial genomes indicated that up to 16 ± 9.5% of the proteins in a genome could be SEPs.

- Integration of transcriptomics and proteomics from 12 bacterial species showed that high-throughput experimental characterization of small proteins (SEPs) presents multiple limitations and false positive detections.
- RanSEPs is a computational approach that assigns coding potential scores to SEP candidates in a species-specific manner based on sequence features.
- After running RanSEPs in 109 bacterial genomes, we determined that between 6 and 25% of the proteins of a bacterial genome could be SEPs.
- Function prediction of RanSEPs-predicted SEPs revealed an enrichment in translation, metabolism and nucleotide-binding proteins.

4.1.2. Additional data access

Datasets covering RNA-seq, mass spectrometry, small gene annotation predictions and other integrative studies in several bacterial species can be accessed by scanning the QR code linked to the original publication. Datasets and supplementary figures will be numerically referred as “Dataset” and “Figure S”, respectively.

4.2. Introduction

Development of ultra-sequencing technologies has led to a considerable increase in the number of annotated bacterial genomes [351]. Classically, general genome annotation protocols only consider ORFs that encode for proteins larger than 100 amino acids [177,178]. This arbitrary cutoff was established to distinguish bona fide protein-coding ORFs from the numerous random in-frame arrangements of start and stop codons present in genomes [352]. However, recent studies have brought to light the importance of small open reading frame (smORF)-encoded proteins (SEPs; ≤ 100 amino acids) [353–355], such as the antimicrobial peptides (AMPs) secreted by insects, animals, plants, and humans in response to infection [356].

In bacteria, SEPs exhibit a wide range of functions that are essential for the cell. SEPs can be involved in cell division (Blr, MciZ, and SidA), transport (AcrZ, KdpF, and SgrT), and signal transduction (MgrB and Sda) or even act as chaperones (FbpB, FbpC, and MntS) [123]. They are also involved in protein complexes, stress responses, virulence, and sporulation [138,258,357,358]. Interestingly, these small proteins can also be used for communication between bacteria and phages, and as bacteriocins within niches like microbiota, thereby making them an important molecule to study when searching for new therapeutic protein candidates [97].

Identifying SEPs is both technically and computationally challenging. At the experimental level, techniques such as ribosome profiling (Ribo-Seq) [359] and mass spectroscopy (MS) [360] are typically used. However, as it is difficult to identify the translated frame in Ribo-Seq experiments, the identification of proteins encoded by overlapping ORFs is not feasible in most cases. Similarly, the absence of ribosome-binding sites (RBS, Shine–Dalgarno) in some bacterial genomes [361,362], and the existence of mRNA without UTRs, makes it difficult to discern smORFs [363]. The detection of SEPs with common tryptic-based bottom-up MS proteomics approaches is also difficult due to the mere fact that their small size correlates with a reduced number of tryptic peptides (TPs) [87,364]. Additionally, identification is further impeded by the fact that SEPs can be secreted, have relatively short half-lives, be present in low abundances, and exhibit tissue- and time-specific expression patterns [365,366].

Evolutionary pressure on genes leads to sequence conservation. As such, gene predictions by cross-species comparisons can be useful for predicting the existence of common proteins [367–369]. However, in such sequence conservation analyses, the probability of overprediction becomes higher for

shorter sequences [370]. Additionally, species-specific SEPs like the Sda protein of *Bacillus subtilis* (46 amino acids), which represses aberrant sporulation by inhibiting the activity of the KinA kinase, cannot be identified through comparative studies [138,357]. Furthermore, although computational methods based on the rate of synonymous and non-synonymous substitutions can differentiate between coding and non-coding regions, these alignment-based methods have two clear limitations. First, a closely related organism is required as a reference, and second, in order to avoid biases in the estimation, this type of method can only be applied to non-overlapping sequences [371]. Other approaches are based on machine learning (ML) algorithms like interpolated Markov models [149], support vector machine-based classifiers [372], logistic regression [372,373], and decompose–compose methods [374]. These methods analyze the coding potential of a genome in an alignment-free manner without the need for experimental information. However, as these approaches do not take into account the importance of species-specific coding features in the classification, they prove inadequate for analyzing the genomes of organisms that are not considered in the training process itself. Importantly, none of these computational methods are free of biases when classifying overlapping annotations, a situation that is common for SEPs [375].

Up until now, it has been difficult to determine the best method for comprehensively analyzing all putative SEPs. Here, by integrating more than 120 "-omics" datasets from *Mycoplasma pneumoniae*, we first assessed the experimental limitations of MS. Then, we developed RanSEPs, a random forest-based tool for the prediction of SEPs in any bacterial genome (Figure 4.1). We also validated the efficiency of RanSEPs by experimentally identifying SEPs in 12 bacterial species, including a set of 570 well-reported and experimentally characterized bacterial SEPs from different species [89,97,123,258,376–378]. We also performed the same efficiency test on other protein discovery software and found that RanSEPs stands out as the best predictor. The higher prediction accuracy of our method is explained by the iterative randomization of the training set, a technique that enables the capturing of additional protein-related information during training. In addition, as the training sets are biased to include more SEPs, they place a higher level of importance on the possible alternative features of these proteins in the classification (Figure 4.1).

By applying RanSEPs to 109 bacterial genomes, we showed that the average number of SEPs per organism could be much higher than previously thought, with SEPs accounting for up to 16 ± 9% of the total coding ORFs. This result suggests that a remarkable number of bacterial SEPs remain unexplored, as recently reported [89]. Additionally, even though most of the antisense non-coding RNAs (ncRNAs) are a product of transcriptional noise and dispensable for cell survival [286], some of them could encode for proteins. In fact, integration of

RanSEPs predictions with transcriptomics data from 11 bacteria species revealed that a fraction of ncRNAs (1%, mostly antisense and intergenic) could encode for SEPs. Finally, functional analysis of SEPs revealed an enrichment in functions related to the membrane, translation, metabolism, and nucleotide binding. As previously described [379,380], we observed a significant proportion of SEPs with N-terminus predicted signal peptide (9.7%) and transmembrane segments (15%). At a time when deep sequencing of microbiomes results in the identification of thousands of new bacterial species, our tool opens up the possibility to predict new SEPs that could modulate bacterial populations through quorum sensing or antimicrobial properties [97].

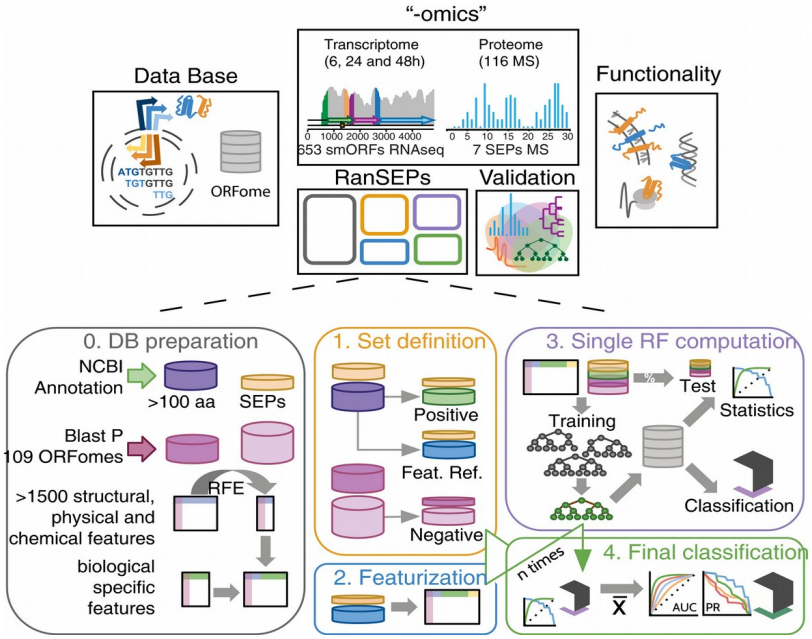


Figure 4.1. Graphical abstract.

First, we generated databases of all the putative ORFs encoded by the genomes of 109 different bacteria. The database of M. pneumoniae was used to perform the shotgun MS and RNA-Seq studies that were aimed at evaluating the coverage and performance of experimental approaches in the discovery of SEPs. In a parallel, experiment-independent manner, RanSEPs performed in silico predictions of potential novel proteins in the database. Results coming from both experimental and computational approaches are integrated in a validation step using a set of 570 SEPs characterized both in this work and in previous studies. Finally, RanSEPs predictions for the 109 bacterial genomes are combined together to assess the functional diversity and importance of predicted SEPs. The second part of the figure highlights how RanSEPs functions. In step 0 (gray box), RanSEPs detects annotated standard proteins (purple) and SEPs (yellow). By BLASTP, non-conserved standard and SEP proteins are detected (pink and light pink, respectively). In parallel, protein features are computed and filtered by Recursive Feature Elimination. These features are combined with general features of biological interest. In step 1 (yellow box), RanSEPs randomly subsets annotated standard and small proteins into a positive (green and yellow), a feature (blue and yellow), and a negative (pink and light pink) set from the bulk of non-conserved sequences. During step 2 (blue box), specific features that vary with each iteration are appended. In step 3 (purple box), the labeled positive and negative sets are divided into training and test sets. Step 4 (green box) consists of collecting the classifiers and classification task results, and computing the final statistics and scores for all the sequences. Step 0 is only run once, and then, it is out of the iteration process. Steps 1–3 are repeated as many times as iterations selected by the user. Step 4 is computed at the end to integrate the results of each iteration.

4.3. Results

4.3.1. Key factors and criteria for the experimental identification of SEPs

To experimentally identify all SEPs encoded by the minimal genome of *M. pneumoniae*, we integrated both proteomics (116 MS experiments) and transcriptomics (eight experiments: four samples of RNA-Seq at 6 h, two at 24 h, and two at 48 h) experiments (Figure 4.1; Datasets 1-3). Analysis of RNA-Seq and MS data was performed to identify possible new proteins having significant RNA expression and/or detected peptides. For this, we used a database including all putative proteins (length ≥ 10 amino acids) translated from the *M. pneumoniae* genome in all six frames (17,818 smORFs and 1,292 ORFs; Figure 4.1). A "decoy" protein dataset of comparable size (Table 4.1), base composition and codon adaptation index (CAI) to that of *M. pneumoniae*, was used as a negative control to detect possible MS artifacts (Dataset 3; see Materials and Methods).

	Type	N	Criteria		
			≥ 1 UTP	≥ 1 UTP; ≥ 1 NUTP	≥ 2 UTP
SEPs	Annotated	26	22	22	21
	Putative	17,792	42	29	7
	Decoy	20,1	19	0	0

Table 4.1. Detection of SEPs using MS in Mycoplasma pneumoniae.
The outcome of different results after MS searches using the decoy database (negative control) and translating all the possible ORFs in *M. pneumoniae*. When using the cutoff of at least 2 UTPs, the signal of every decoy protein was removed but the detection of putative SEPs consequently dropped, with one annotated SEP not being identified.

Using MS, we identified 42 potentially new SEPs in *M. pneumoniae* with ≥ 1 unique tryptic peptide (UTP) and RNA expression levels ≥ 4.5 log2(counts) (Figure 4.2A; Datasets 1,3). However, 19 "decoy" SEPs were also detected (Figure 4.2B). While we found that the number of novel SEPs identified with ≥ 1 UTP increased in proportion to the number of experiments being considered, this same trend was also observed for the "decoy" SEPs (Dataset 1 and Figure 4.2C). This trend suggested the existence of false positives in MS when considering no threshold for the number of identified UTPs. When we increased the number of detected UTPs to ≥ 2, we did not find any "decoy" protein but we did lose one NCBI-annotated SEP (Table 4.1 and Figure 4.2B) and the data quickly reached a plateau after four experiments (Figure 4.2C). The same happened using a threshold of one UTP and ≥ 1 non-unique tryptic peptide (NUTP). The number of

putative SEPs was reduced from 42 to 7 using the first threshold and from 42 to 29 using the more relaxed threshold (Table 4.1, Figure 4.2B). After filtering by ≥ 2 UTPs, 532 proteins remained: 521 annotated, four novel standard proteins, and seven novel SEPs (shortest presenting a length of 48 amino acids). To corroborate our ≥ 2 UTP threshold criteria, we performed targeted MS with C13C(6)15N(2)-labeled peptides of eight SEPs, four of which had ≥ 2 UTPs and four with one UTP (Dataset 4). All four of the novel SEPs detected with ≥ 2 UTPs were confirmed with the C13 peptides. In contrast, we only detected a signal for two of the SEPs identified with one UTP in targeted proteomics (Figures S1 and S2; Accession number of MS results: PXD008243). These results indicate that ≥ 2 UTPs should be considered as the threshold for protein discovery without false positives, but that true SEPs could be lost.

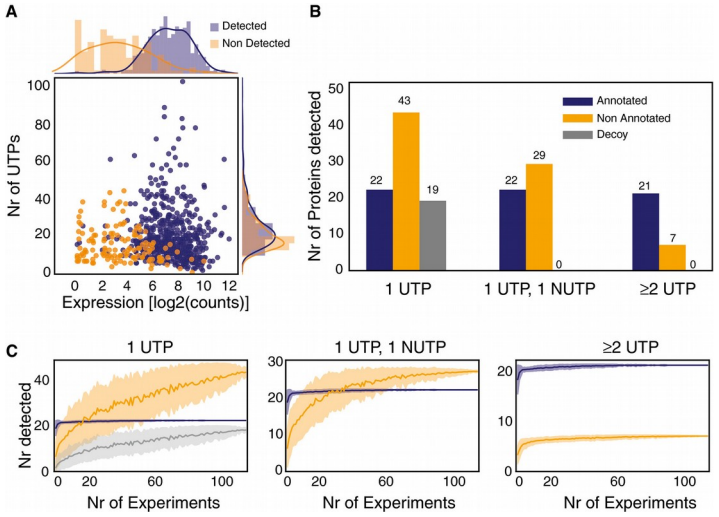


Figure 4.2. Assessment of the detection coverage by "-omics" approaches.
(**A**) Evaluation of expression by RNA-Seq and number of peptides required to detect an annotated protein by MS in *M. pneumoniae*. The plot represents the relationship between expression levels (average expression from RNA-Seq data) and number of possible unique tryptic peptides (UTPs) for two sets of studied proteins: detected (blue dots) and not detected (orange dots) by MS. (**B**) Evaluation of thresholds and artefactual signals in MS data. The histogram represents the total number of SEP proteins detected in 116 shotgun MS experiments with 1 UTP, 1 UTP and 1 NUTP, or ≥ 2 UTPs for three categories. Color code: annotated (blue bars), putative new (orange bars), and decoy set (gray bars). (**C**) Number of SEPs detected by increasing the number of experiments. Color code is the same as in panel (B). Each line represents the accumulated number of different SEPs detected (*y*-axis) when combining 1–116 MS datasets (*x*-axis) from *M. pneumoniae*. Each line has an associated error that is shaded and represents the standard deviation within combinations of datasets (e.g., $x = 80$ will present the average number of proteins detected taking every combination of datasets in groups of 80 samples).

Interestingly, 25% of the annotated proteins of *M. pneumoniae* were not identified by MS. By using PeptideSieve [381], we measured the responsiveness of the proteins to MS. We found that annotated proteins detected by MS had a significantly higher number of high-responsive UTPs (HR_UTPs) than undetected proteins (Mann–Whitney one-sided P-value = 0.03; Dataset 3 and Figure S3), revealing that not only the number of UTPs, but also their properties, could hamper protein detection by MS.

Analysis of the proteome (and its conservation) of five closely related Mycoplasmas revealed that 159 possible SEPs could be conserved in more than two species. Of these, we detected 48 by MS (Datasets 5-10), 30 with 1 UTP, and 18 with at least 2 UTPs. While these 18 SEPs were identified with ≥ 2 UTPs in some of the species, in others, they were detected with only 1 UTP. This reinforces the idea that some SEPs having only 1 UTP could in fact be real. Therefore, conservation analysis could be helpful in identifying new SEPs as long as it is performed in conjunction with MS experiments in multiple organisms. Nonetheless, this approach could be misleading in the case of overlapping genes (Figure S4).

To confirm that the ≥ 2 UTPs criteria enable us to identify true proteins, we studied the correlation between ribosome profiling and the number of UTPs. For this purpose, we used raw datasets of ribosome profiling that were recently published for *Escherichia coli* [382]. We then analyzed an *E. coli* extract enriched in SEPs by MS (see Materials and Methods, Dataset 11) and studied the correlation between both techniques. Ribosome profiling showed that the mRNA of SEPs detected with ≥ 2 UTPs presented significantly more bound ribosomes than both those detected with just 1 UTP (Mann–Whitney one-sided test P-value = 0.005) and those not detected by MS at all (Mann–Whitney one-sided test P-value = 0.001, Figure S5). Thus, ribosome profiling supports using a ≥ 2 UTP cutoff to extract potential positive SEPs by MS.

In conclusion, while true-positive SEPs can be identified by MS when filtering by ≥ 2 UTPs, SEPs with only 1 UTP or very low responsiveness cannot be experimentally assessed by label-free proteomics. Therefore, experimental validation of SEPs still remains a challenge, and development of computational prediction tools capable of identifying SEPs without compromising the false discovery rate is paramount.

4.3.2. RanSEPs: a novel random forest approach for the discovery of SEPs

Computational approaches are required not only to predict SEPs but also to reduce the required number of targeted validation experiments. For this purpose, we have developed RanSEPs, a variation of the random forest (RF) algorithm that iterates and randomizes training sets at the same time that it defines protein features (see Materials and Methods). These features are selected in a blind manner by their importance in test classifications (Figure 4.1). With this approach, positive and negative set selections are fully randomized in each iteration, thereby generating an individual classifier each time. The positive sets comprise subsets of annotated proteins from NCBI that are forced to include a minimum percentage of SEPs belonging to the target organism. For the negative set, RanSEPs creates random sets of smORFs that are located within intergenic regions (relative to annotated genes) and have no identified homologs in a database including the six translated reading frames of 109 different organisms. Conceptually, this set could include actual SEPs; however, the probability of maintaining a true SEP and biasing the prediction is virtually null (see Materials and Methods).

The output is a probability score for a specific protein belonging to the coding class. When assigning the coding class to SEPs of *M. pneumoniae*, we set the threshold to a score ≥ 0.5, while for standard proteins, it was set to a score ≥ 0.85 (95^{th} percentile for both distributions, Figure S6A). With the results of the previous prediction and using cross-validation, we obtained an average true-positive rate (TPR) of 96.3 and 90.3% for annotated SEPs and standard proteins, respectively, with a total area under the ROC curve (AUC) of 0.92 when considering both types of proteins (Figure S6B-C). Using these settings, RanSEPs predicted 756 ORFs for *M. pneumoniae*: 612 standard proteins (598 annotated and 14 new) and 144 SEPs (26 annotated and 118 new). All of the new SEPs detected by MS with ≥ 2 UTPs were classified by RanSEPs as coding (see Supplementary Methods and Figs S12–S14). Among the 23 SEPs detected with 1 UTP, RanSEPs predicted only five to be true, of which one was previously annotated with function while the other four were annotated by inference in closely related organisms. Interestingly, the other 18 putative smORFs with one UTP that were classified as non-coding by RanSEPs did not present homologous annotated candidates in closely related Mycoplasma species and their RNA levels were significantly lower compared with the five SEPs that had 1 UTP predicted as positive by RanSEPs (expression levels > 90^{th} percentile, Dataset 3). This agrees with what we found in the previous section and supports the application of RanSEPs as a tool for predicting those potential SEPs identified with 1 UTP.

Next, we determined the proportion of predicted SEPs that could be considered false positives: pseudogenes and highly repeated sequences. Within the complete smORFome of *M. pneumoniae*, we detected 44 smORFs that could be derived from the fragmentation of a larger protein found in *M. genitalium* and 242 with at least two homologous matches in the *M. pneumoniae* genome. RanSEPs classified eight of the 242 “repeated” annotations as coding and predicted the 44 fragments to be non-coding (Dataset 3). This homology information is integrated into every RanSEPs prediction to enable prioritization of results and provide more meaningful predictions.

4.3.3. RanSEPs validation and method comparative

To validate RanSEPs predictions and test its potential applicability in other bacterial genomes, we generated a positive small protein set (n = 570) including multiple sources. First, MS was used to identify SEPs from enriched protein extracts of *Escherichia coli*, *Pseudomonas aeruginosa*, and *Staphylococcus aureus* (see Materials and Methods, Datasets 11-13), as well as from total protein extracts of six Mycoplasma species (Datasets 6-10). Second, we re-analyzed publicly available MS datasets generated to detect SEPs and reported in the literature: *Lactococcus lactis* (PRD000266), *Synechocystis* sp. PCC6803 (PXD001246), and *Helicobacter pylori* (PXD000054; Datasets 14-16; see Materials and Methods). In total, 473 SEPs (25 potentially new SEPs; 11 corroborated also by targeted proteomics) were found with ≥ 2 UTPs in MS searches of these 12 bacterial species. Finally, 97 SEPs reported and validated in the literature were also added to this positive protein set (Dataset 18). We also defined a balanced negative protein set (n = 570), which included 13 smORFs tested by targeted proteomics with negative results and 536 putative smORFs expected to be true negatives. This 536 smORFs subset was extracted from a collection of 14,746 putative smORFs from the 12 bacterial species studied by MS (Dataset 18; see details in Materials and Methods). The criteria for selecting them were as follows: (i) They are not conserved in closely related species, and (ii) they have more than two high-responsive UTPs by PeptideSieve and are not detected by MS (Dataset 18).

For validation, we performed specific RanSEPs predictions for each species, ensuring that the SEPs included in the validation set were never used in any training step (details about species-specific parameters can be found in Supplementary Methods). The same test was replicated with commonly used annotation prediction tools: CPC2, GeneMarkS, BASys, Glimmer, and Prodigal [150,158,181,185,383]. One factor that makes RanSEPs different from other predictors is that it is able to compute and use species-specific feature weights to determine coding potential (see Materials and Methods). As specific features do

not necessarily share the same general importance across different organisms, this functionality allows unbiased searches to be carried out for any organism. For example, the Shine–Dalgarno sequence, which acts as an RBS and has an important role in translation, is not always present in bacterial species, including Mycoplasmas [173,384]. This can be observed when measuring the feature weights by RanSEPs, as RBSs, which are rarely found in *M. pneumoniae* genes [385], have a very low weight in this organism (Figure 4.3A).

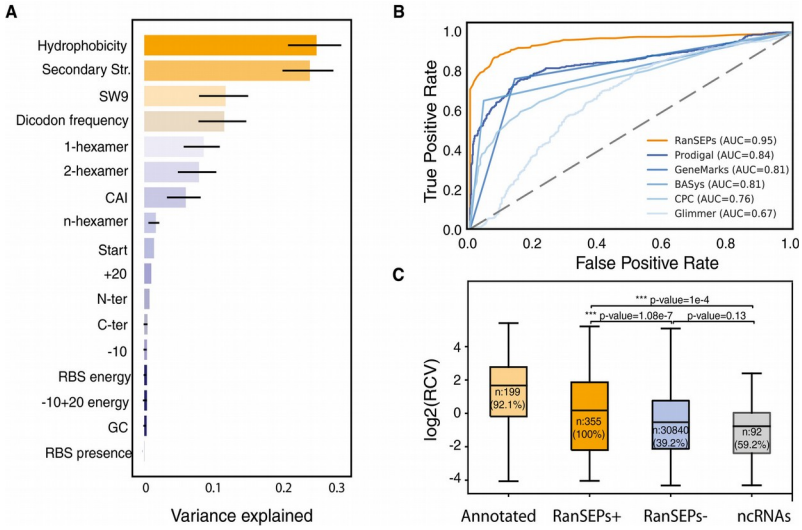


Figure 4.3. RanSEPs predictions.
(**A**) Feature weight prediction in *M. pneumoniae*. Weights of the different features considered in the classification by RanSEPs. Bars indicate the global averaged variance that each feature explains by itself along with its associated standard deviation (black line) (25 iterations to estimate the error). (**B**) Method accuracy comparative. Receiver operating characteristic curve for RanSEPs (orange) and five additional tools (blue gradient). The closer a curve to the left-hand border, the more accurate the tool. The area under the curve (AUC) associated with each method is presented, with values closer to 1 indicating a more accurate method. The dashed gray line represents a classifier that assigns the coding class randomly. (**C**) Boxplot representing the relationship between RanSEPs-positive ("RanSEPs^{+}", score ≥ 0.5) and RanSEPs-negative ("RanSEPs^{-}", score < 0.5) SEPs predictions and associated RCV (ribosome profiling ratio coverage, in $\log_2$) in *Escherichia coli*. Only annotations ≤ 300 nucleotides in length were included. As positive and negative controls, we considered annotated SEPs ("Annotated") and non-coding RNAs ("ncRNAs"), respectively. Annotations within RanSEPs^{+}, RanSEPs^{-}, and ncRNAs overlapping with known annotated genes were excluded. Annotations with RCV = 0.0 are filtered out, and the number within the box represents the percentage of values in that class that are kept in the comparative. Along the top, *P*-values computed by Mann–Whitney rank test are indicated.

We assessed and compared the quality of the predictions in terms of accuracy and AUC (see Materials and Methods), and also in terms of computational cost (Figure S8, see Supplementary Methods). RanSEPs was the best tool for predicting SEPs (AUC = 0.95; accuracy = 0.89) as none of the other tools had an AUC > 0.85 (Dataset 19, Figure 4.3B, and Figure S7). Remarkably, RanSEPs provided the best TPR (SEPs properly predicted as SEPs over total positives) for annotated proteins (86.8%), SEPs with ≥ 2 UTPs in MS (86.7%), and potential new SEPs (76%). It was also the only tool that predicted all the SEPs validated by targeted C13 proteomics without false positives (Dataset 19, Figure S15). In terms of false-positive rates (smORFs wrongly predicted as SEPs over total negatives, FPR), RanSEPs returned the third lowest value, coming after BASys and CPC. However, these two tools did not reach TPRs higher than 65%.

Finally, we further validated our prediction tool at the genome-wide level by studying the correlation between gene-expression-corrected Ribo-Seq coverage (RCV) and RanSEPs prediction in *E. coli* [382]. We found that SEPs predicted as positive showed significantly higher RCV levels compared with candidates predicted as negatives (Mann–Whitney one-sided test P-value=1×10^{-7}) and ncRNAs (Mann–Whitney one-sided test P-value=1×10^{-4}, Figure 4.3C). Additionally, while RanSEPs-positive predictions presented RCV values closer to the scores of annotated proteins, although still significantly lower (Mann–Whitney two-sided test, P-value = 1×10^{-10}), negative predictions were more similar to annotated ncRNAs (no significant differences by Mann–Whitney two-sided test, P-value = 0.13).

We next confirmed that the high success rate of RanSEPs was not due to an excess of positively scored annotations. In this analysis, we used the previously defined collection of 14,746 smORFs with low coding potential to search for false positives. Glimmer and CPC yielded the lowest FPRs but also had significantly limited TPRs. The rest of the tools presented comparable FPRs, with values of 5.1, 4.3, 3.6, and 3.9% for Prodigal, RanSEPs, BASys, and GeneMarkS, respectively. None of the false positives returned by RanSEPs presented a score higher than 0.65, indicating that a stricter score threshold would prevent the detection of false positives. However, this threshold led the average AUC falling to 0.88, indicating that we would miss valid SEPs.

Additionally, RanSEPs provides extra information associated with the scores for further prioritization of the predictions. This information includes aspects like the presence of an RBS and a preliminary classification of the predicted SEPs into one of the following groups: (i) conserved in closely related species but not annotated in the organism of interest or any other; (ii) conserved and annotated in other species with a known function; (iii) conserved and annotated in other species without a known function; (iv) highly repeated in the annotated reference genome; or (v) potential pseudogene (see Materials and Methods).

4.3.4. RanSEPs in a species-specific context and ncRNAs

To study the smORFomes in different bacterial genomes and to address the outstanding question regarding the percentage of coding annotations represented by SEPs, we applied RanSEPs to 109 bacterial genomes. RanSEPs was parameterized and ran independently for each genome (see details in Materials and Methods and Supplementary Methods), and we considered the two thresholds defined above: the one that maximizes true positives (RanSEPs score ≥ 0.5) and the one that minimizes false positives in *M. pneumoniae* (score ≥ 0.65). This resulted in an average TPR of 86 ± 7% for annotated SEPs (iteratively excluding them from the training sets) with the 0.5 score, and 67 ± 12% with the 0.65 score. On average, the number of annotated SEPs over the total number of annotated coding ORFs was 10 ± 5%, a value that reaches 16 ± 9.5% when adding SEPs predicted by RanSEPs with a score of ≥ 0.5 and 14 ± 7% when raising it to ≥ 0.65. On average, we determined that 1 ± 0.7% of the SEPs predicted by RanSEPs with a score ≥ 0.5 could be considered pseudogenes or "repeated" sequences when the SEP was a fragment of a larger protein in another organism or found several times in the reference genome. These values were reduced to 0.75 ± 0.1% when using the ≥ 0.65 threshold. Ultimately, this implies that between a minimum of 13 ± 7% and a maximum of 16 ± 9.5% of the proteins in each genome could be SEPs (Dataset 20). The prediction results for the 109 bacteria can be downloaded at www.ranseps.crg.es.

As in *M. pneumoniae*, secondary structure and hydrophobicity were the most important features for polypeptide classification in all bacteria (Figure 4.4). However, some features like the SW9 and the four dicodon frequencies (see Materials and Methods) showed weight differences that resulted in two clusters of bacterial species. The first cluster (higher weight for SW9 and lower values for dicodon frequencies) presented higher rates of encoded SEPs than the second cluster (low weight for SW9 and high weight for dicodon frequencies, unpaired *t*-test *P*-value = 0.04). In addition, we observed that organisms with higher percentages of SEPs (> 13.16%, $N = 55$) were associated with bacteria having low GC contents (38 ± 12%). In contrast, lower rates of SEPs (≤ 13.16%, $N = 54$) were predicted for bacterial species with higher GC contents (47 ± 10%, unpaired *t*-test *P*-value = 0.005, Dataset 20). These results agreed with previous studies, suggesting that a low GC content increases the number of stop codons and consequently results in an increased percentage of SEPs (see Supplementary Methods).

Over the past few years, it has been shown that sequences formerly described as ncRNAs could, in some cases, actually encode for proteins, with some of them being SEPs [253,382]. Thus, we combined RanSEPs predictions with the annotated ncRNAs of 11 bacterial transcriptomes [286] and found that 273 out of

8,056 ncRNAs could in fact encode for 289 proteins: 184 SEPs and 105 standard proteins (Dataset 21). Out of these 273 ncRNAs that could encode for proteins, 11 (4%) were overlapping in sense with genes, 185 (67.8%) in antisense, and 77 were located in intergenic regions (28.2%; Figure S9). The average length of the 184 SEPs encoded by these re-annotated RNAs is 96 amino acids. In contrast, standard proteins encoded by former ncRNAs had an average length of 132 amino acids.

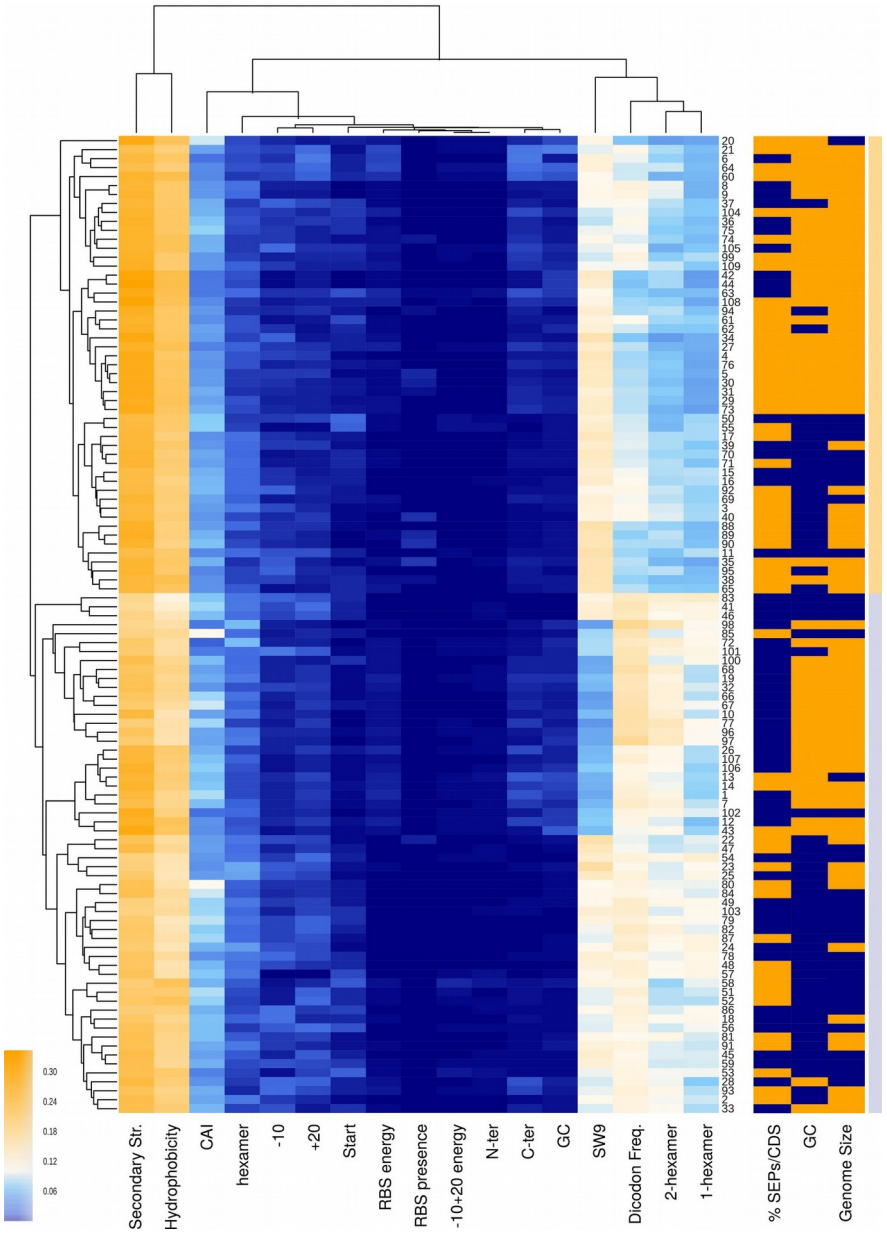

0.30
0.24
0.18
0.12
0.06
Secondary Str.
Hydrophobicity
CAI
hexamer
-10
+20
Start
RBS energy
RBS presence
-10+20 energy
N-ter
C-ter
GC
SW9
Dicodon Freq.
2-hexamer
1-hexamer
% SEPs/CDS
GC
Genome Size

Figure 4.4. A comparison of the feature weights used for the prediction of SEPs in 109 bacterial genomes

Clustered heat map using nearest point algorithm and representing the weights of different features in 109 bacterial genomes, and the clustering relations between features (top dendrogram) and species (side dendrogram). Rightmost light-orange and light-blue bars are included to differentiate the two main clusters. Numbers in the right vertical axis are short references representing the names of the bacterial genomes (Dataset 15). The right three columns represent biological features not used in the classification. The ratio of the percentage of SEPs compared to the median value is colored as blue and orange for ≤ 13.16% of SEPs and > 13.16%, respectively. Blue and orange colors in the %GC column represent genomes with ≤ 38 and > 38% GC content (median value = 38), respectively. Genome size column separates species into small-genome bacteria (≤ 1.5 Mb, blue) and large-genome bacteria (> 1.5 Mb, orange).

4.3.5. Functional assessment of novel SEPs

In total, 36,311 SEPs were collected, including annotated and predicted SEPs from the 109 genomes considered. Out of this group, while 25,229 were found annotated in their original genomes (231 ± 186 annotated SEPs per genome), the majority of them were annotated as hypothetical proteins or with unknown function. In fact, only 5,175 SEPs (20%) were associated with a function. The majority of the SEPs with assigned functions were involved in translation (mainly ribosomal proteins), metabolism, and DNA/RNA binding (Figure 4.5A; Dataset 22).

The total number of predicted SEPs not previously considered in their respective original reference genome was 14,773 using the ≥ 0.5 score criteria. To explore the possible functions of the proteins belonging to this group, we ran a BLAST search using the first group of SEPs with annotated functions as a database. Results indicated that, on average, a specific SEP with an undescribed function could be conserved in at least 15 different organisms (Figure S10). In addition, this analysis revealed that while 3,535 SEPs (24%) did not have annotated homologs, 11,238 (76%) were found annotated in other species: 5,038 (34%) with unknown function and 6,341 (42%) with different functions (Figure 4.5B). We repeated this search with the “decoy” protein dataset used for MS as the target, and found that no sequence passed the thresholds required to be considered homologous. As such, we would not expect to have false positives by chance. Although we have assigned functionality to most of the predicted SEPs in the 109 genomes, one needs to be cautious as sequence homology and functional annotation of small proteins is not always reliable.

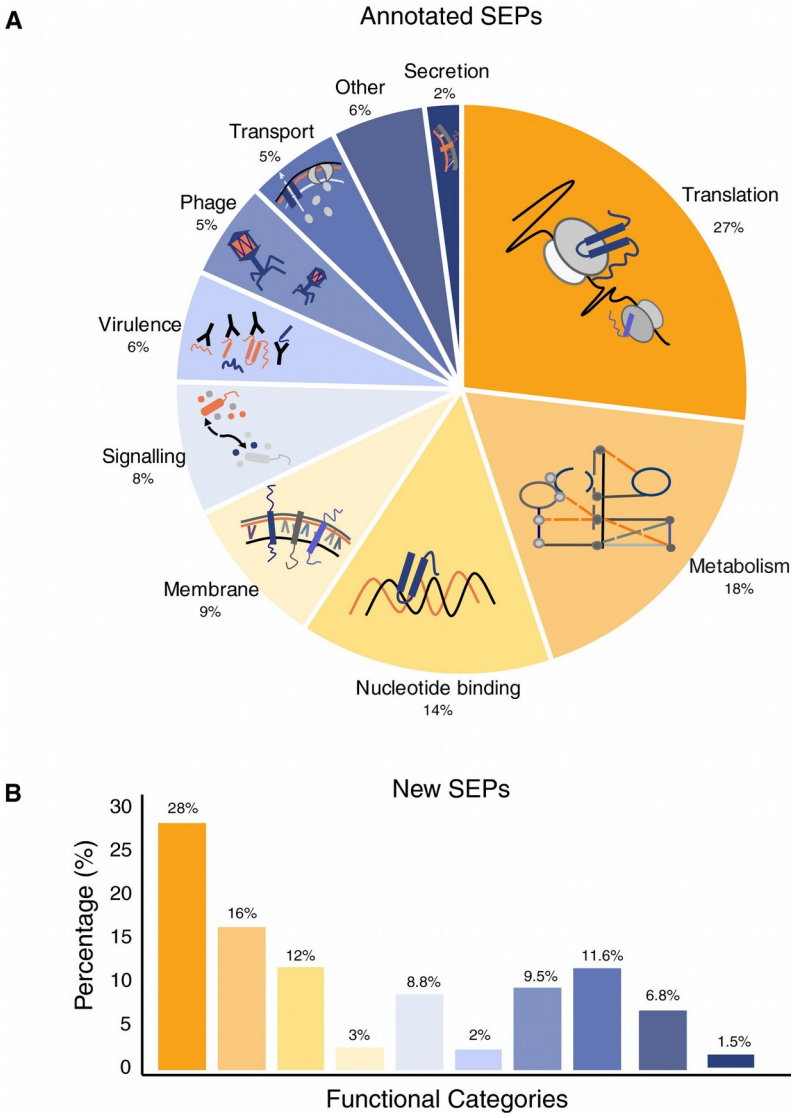


Figure 4.5. Functional assessment of RanSEPs results

(**A)** Landscape of the SEPs with functional annotations in NCBI considering 109 bacterial genomes (Number of SEPs = 25,229 SEPs). (**B)** Functional inference of the predicted SEPs (*N* = 11,238) as determined using BLASTP against NCBI-annotated SEPs having an associated function (*N* = 5,175). The color code associated with each category is the same as in panel (A).

Finally, in some bacteria, SEPs are known to be secreted and can play a role in communication or even act as toxins [97]. To determine whether some of the new SEPs we discovered could be secreted or be integrated into the membrane, we searched for signal peptide sequences as well as for transmembrane regions using Phobius [386]. We focused on the set of SEPs with unassigned function and found that 9.7% had a N-terminus predicted signal peptide sequence and 15% a transmembrane membrane region (Dataset 22). The percentage of SEPs with a signal peptide was higher than expected by chance when compared with the same "decoy" set of SEPs used in MS (9.7% for predicted SEPs, 1.2% for "decoy" SEPs, unpaired two-tailed t-test P-value = 0.018). Moreover, to confirm that the results obtained with Phobius were meaningful with regard to SEPs, and that protein size did not bias the analysis, we ran a test on a set of annotated standard proteins in which we sequentially shortened their C-terminus. The sensitivity of Phobius is higher than 80% for sequences over 30 amino acids. For sequences under 30 amino acids, however, we see values lower than 50%; this is expected when considering that Phobius specifically searches for a motif presented by the first 16–30 amino acids of the N-terminus of a protein. If the motif is located within these first amino acids and is short, Phobius will still detect the protein as positive (see Supplementary Methods and Figure S11).

4.4. Discussion

Genome annotations, which traditionally considered only standard proteins, ignored the existence of a layer of complexity represented by SEPs (i.e., the smORFome). After assessing the experimental limitations, we showed that the experimental detection and characterization of SEPs are challenging. On the one hand, as "decoy" protein sequences are detected by MS, proteins that do not exist can actually have spectra assigned (1 UTP or 1 UTP;1 NUTP). On the other hand, as these "decoy" proteins appeared across multiple experiments, discrimination criteria based on reproducibility are not feasible. This problem is solved by only accepting proteins detected with ≥ 2 UTPs. These criteria were corroborated by re-analyzing Ribo-Seq data from *E. coli*. The main drawback is that many SEPs have very few responsive UTPs and consequently they are discarded. Despite these constraints, however, we were still able to detect novel SEPs in *M. pneumoniae* by integrating 116 shotgun MS datasets. Thus, this represents the first comprehensive study of a bacterial proteome using MS without protein size thresholds. Label-free MS experiments on cell extracts and SDS gel extraction derived from 12 bacterial species identified 25 new SEPs not annotated in the reference genomes. Of course, the problem associated with only 1 UTP could be partly alleviated by doing targeted proteomics with labeled C13 peptides. However, taking into account the required number of experiments and the fact that many SEPs do not have high-responsive peptides, the extensive analysis of

SEPs encoded by a bacterial genome would be precluded. In addition, other factors could contribute to this problem like short protein half-lives, conditional gene expression, or special features in sequence associated with concrete functions (e.g., hydrophobicity).

Here, we developed RanSEPs to address the aforementioned limitations. Using *M. pneumoniae* as a reference, we developed RanSEPs as a predictor to define candidates given a specific genome and to score them by assigning a probability of being coding smORFs. Also, the assigned score provides meaningful information about features that can be important for the functional characterization of SEPs. Furthermore, we validated this application in other bacteria with SEPs that had been experimentally identified or described in the literature. Comparison of RanSEPs with five other tools showed that RanSEPs maximizes the correct prediction of true positives without increasing the false-positive rate. This could be attributed to its iterative method in which multiple classifications are averaged using different sets of annotated proteins in each iteration. This property permits the capture of a wide diversity of features presented by annotated genes, thereby resulting in more accurate predictions. Derived from this and considering that closely related species share sequence features, our scoring algorithm could also be modified to de novo annotate a genome of interest. In addition, the relationship between gene-expression-corrected ribosome profiling in *E. coli* and RanSEPs predictions showed that predicted SEPs generally have higher ratios than those predicted to be negative and resembling annotated ncRNAs.

Analysis of features that discriminate coding sequences in 109 bacterial genomes revealed that hydrophobicity and secondary structure are key factors. Also, we observed that the number of predicted SEPs encoded by a genome depends on the GC content. On the other hand, the importance of features governing coding potential is conserved across species. Strikingly, between a 13 ± 7 and 16 ± 9.5% of the genes (depending on the cutoff score used) in these 109 species encoded for SEPs, highlighting that the coding capacity of bacterial genomes has likely been underestimated. Noteworthy, genome annotations are critical for classifying a SEP as a new protein. In fact, for 76% of the SEPs predicted by RanSEPs, orthologous SEPs were identified by BLAST in closely related strains. This result indicates that reference genomes are still incomplete and not properly curated.

Possibly, some of the predicted SEPs could be pseudogenes or false positives. Identification via homology of mutations resulting in a premature stop codon can provide an estimation of the number of pseudogenes present in a genome of interest. With this approach, we estimated that 1 ± 0.7% of predicted SEPs could be pseudogenes. These genes can be excluded, however, by increasing the RanSEPs threshold, albeit at the cost of missing some true SEPs. Thus, our 13–

16% lower and upper estimates could still contain false positives but still represent a significant percentage.

Interestingly, some ncRNAs of multiple bacterial species could actually encode for proteins. While 63% of the proteins potentially encoded by ncRNAs were SEPs, 37% were standard proteins with an average amino acid length of 132 amino acids. This suggested that some ncRNAs could in fact be coding and that bacterial annotations could be missing not only SEPs but also longer proteins.

Functional analysis of the predicted and previously identified SEPs indicated that these proteins participate in basic processes of living systems such as transcription, translation, metabolism, signaling, quorum sensing, virulence, and pathogenicity. However, this analysis should be taken with caution as sequence homology and functional annotation of SEPs is challenging [89]. Interestingly, similar to what has been previously reported [379,380], we found a significant enrichment in SEPs presenting features indicative of being secreted (10%) or membrane localized (15%). This observation could have an impact not only on translational research but also on the study of the modulation of bacterial populations in microbiomes, thereby opening up a new line of research in the Systems Biology discipline [97].

With all our results in mind, we envision RanSEPs as a tool to help predict new SEPs, support detections, and discard artifactual proteins detected by MS that have low signals such as those detected with only one UTP and/or one NUTP. When no experimental information is available, RanSEPs can help guide the selection of potential new SEPs for validation and further characterization with the overall aim of uncovering their functions.

4.5. Material and Methods

4.5.1. ORFome database generation

We generated the *in silico* proteomes by translating all putative ORFs with sizes ≥ 10 amino acids from the six possible open reading frames of 109 bacteria. These bacteria included representative species of both gram types, and covered a wide spectrum of genome sizes (0.5–9 Mb), GC contents (20–70%), and generation times (0.48–12 h). Putative ORF databases were computed considering the codon translation table 11 (start codons: ATG, GTG, and TTG; stop codons: TAG, TAA, and TGA) for all cases except Mollicutes, which were based on translation table 4 (start codons: ATG, GTG, and TTG; stop codons: TAG and TAA). In all cases, only ORFs encoding theoretical proteins of at least 10 amino acids were accepted in the databases (www.ranseps.crg.es).

4.5.2. Decoy database generation

A “decoy” dataset to assess the presence of possible artifacts when searching SEPs in a specific organism was generated based on certain factors. First, we used a comparable number of SEPs and standard proteins to the number in the target organism as it is known that the database size can bias MS searches. Second, we forced the sequences to present a GC content and codon usage similar to those of the target organism. Lastly, we permitted only sequences that were not found in other organisms (BLASTP e-value > 0.1). In the end, the “decoy” dataset was composed of: (i) 2,433 translated stop-to-stop non-coding regions of *M. pneumoniae* without any start codon, (“in” prefix); (ii) 1,425 translated intergenic regions from the *M. pneumoniae* genome (without start codon, no overlap with any putative ORFs, “or” prefix); (iii) 8,740 pseudo-randomly generated peptides with a codon usage and GC content comparable to that of the *M. pneumoniae* genome, lengths between 20 and 100 amino acids, forced to have an average of three detectable UTPs, and comparable start and stop frequencies for start (ATG = 0.86, GTG = 0.073, TTG = 0.067) and stop: (TAA = 0.71, TAG = 0.28) codons (prefix “gc”); and (iv) 9,110 amino acid sequences obtained by translating the *in silico* random genome, preserving the GC content and codon usage of the *M. pneumoniae* genome (prefix “rd”). This genome is generated using frequencies and sizes of intergenic and coding regions similar to those of the annotated genome in NCBI. As GC content varies between coding and intergenic regions, we adjusted the “decoy” gene regions by codon adaptation index (CAI) and the intergenic ones by GC.

4.5.3. Bacterial strains and growth conditions

M. pneumoniae M129 was grown in 75-cm^2 tissue culture flasks with 50 ml of modified Hayflick medium at 37°C as previously described [217]. *M. genitalium* G-37 (wild-type) strain was grown in SP-4 medium [387] at 37°C under 5% CO_2 in tissue culture flasks (TPP). *M. gallisepticum* str. R (high), *M. hyopneumoniae* 232, *M. capricolum* subsp. capricolum ATCC 27343, and *M. mycoides* subsp. capri str. GM12 were all grown as suspension cultures in SP-4 medium at 37°C and 200 rpm. *E. coli*, *S. aureus*, and *P. aeruginosa* (strain PAO1) were grown overnight in 22 ml TSB medium, at 37°C, shaking at 180 rpm.

4.5.4. RNA extraction and library preparation for RNA-Seq

After growing *M. pneumoniae* for 6 h at 37°C, cells were washed twice with PBS and lysed with 700 µl of QIAzol buffer. RNA extractions were performed using the miRNeasy Mini Kit (Qiagen) following the instructions of the manufacturer. Libraries for RNA-Seq were prepared following directional RNA-Seq library preparation and sequencing. Briefly, 1 µg of total RNA was fragmented into ~100–150 nt using NEB Next Magnesium RNA Fragmentation Module (ref. E6150S, NEB). Treatments with Antarctic phosphatase (ref. M0289S, NEB) and PNK (ref. M0201S, NEB) were performed in order to make the 5′ and 3′ ends of the RNA available for adapter ligation. Samples were further processed using the TruSeq Small RNA Sample Prep Kit (ref. RS-200-0012, Illumina) according to the manufacturer's protocol. In summary, 3′ adapters and subsequently 5′ adapters were ligated to the RNA. cDNA was synthesized using reverse transcriptase (SuperScript II, ref. 18064-014, Invitrogen) and a specific primer (RNA RT Primer) complementary to the 3′ RNA adapter. cDNA was further amplified by PCR using indexed adapters supplied in the kit. Finally, size selection of the libraries was performed using 6% Novex® TBE Gels (ref. EC6265BOX, Life Technologies). Fragments with insert sizes of 100–130 bp were cut from the gel, and cDNA was precipitated and eluted in 10 µl of elution buffer. Double stranded templates were cluster-amplified and sequenced on an Illumina HiSeq 2000. The raw data of RNA-Seq were submitted to the ArrayExpress database (http://www.ebi.ac.uk/arrayexpress) and assigned the identifier: E-MTAB-6203.

For each experiment, both ends were treated as independent single-end reads in order to avoid the wrong assignment of read-pairs. Filtered reads were mapped to each reference genome using Maq mapping software. We mapped the reads containing 50 bp, allowing for one mismatch. The expression per ORF was computed based on:

$$Expression = \log_2\left(\frac{read\ count\ gene}{gene\ length}\right)$$

To define an ORF as “transcriptionally active”, its expression value had to pass a threshold established by the minimum expression value for all previously annotated genes of the organism of interest.

4.5.5. Prediction of possible and high-responsive UTPs

To determine the number of expected high-responsive UTPs, we used PeptideSieve [381] with the default properties file and selected results for “Page Electrospray: PAGE_ESI” with a probability score > 0.65. This threshold was selected as it provided the best correlation between predicted UTPs and those observed experimentally (0.61 correlation coefficient). A peptide was considered to be a UTP only when it was found to be associated with one protein and have a minimum length of 5 amino acids (Dataset 3).

4.5.6. Mass spectrometric analyses

a) Sample preparation

To generate new samples for MS analysis, 5 ml of the *P. aeruginosa*, *E. coli*, and *S. aureus* overnight cultures was centrifuged and resuspended in 500 µl lysis buffer (20 mM sodium phosphate, pH 7.4, 500 mM NaCl, 1% Triton, 2 mM DTT + protease inhibitors + lysozyme 50 µg/ml). Then, the lysates were incubated 20 min at RT, disrupted by sonication (15 min × hi 30’’ on/off on ice), and centrifuged for 30 min at 21,130 *g*. Twenty microliters of both the supernatant and the pellet was loaded on Novex 10–20% Tricine gels (Thermo Fisher # EC6625BOX) and run at 120 V for 30 min. Afterward, different portions of the gel were cut with a scalpel: one portion below the loading buffer line, and the other portion between the loading buffer line and the 10 kDa marker.

Data from 116 shotgun MS experiments corresponding to different mutants and conditions of *M. pneumoniae*, as shown in Datasets 1-3, were re-analyzed with the new database (see above) to re-annotate the *M. pneumoniae* genome (ID PRIDE: PXD008243).

Samples extracted with SDS were reduced with dithiothreitol (90 nmols, 30 min, 56°C), alkylated in the dark with iodoacetamide (180 nmols, 30 min, 25°C), and digested first with 3 µg LysC (Wako, cat # 129-02541) overnight at 37°C and then with 3 µg of trypsin (Promega, cat # V5113) for 8 h at 37°C following the fasp produce of Wiśniewski [388]. Samples extracted with urea were reduced with

dithiothreitol (90 nmols, 1 h, 37°C) and alkylated in the dark with iodoacetamide (180 nmol, 30 min, 25°C). The resulting protein extract was first diluted 1/3 with 200 mM NH_4HCO_3 and digested with 3 µg LysC (Wako, cat # 129-02541) overnight at 37°C, and then diluted 1/2 and digested with 3 µg of trypsin (Promega, cat # V5113) for 8 h at 37°C. After digestion, the peptide mix was acidified with formic acid and then desalted with a MicroSpin C18 column (The Nest Group, Inc) prior to LC-MS/MS analysis.

b) Sample acquisition

The peptide mixes were analyzed using a LTQ-Orbitrap Velos Pro mass spectrometer (Thermo Fisher Scientific, San Jose, CA, USA) coupled to an EasyLC [Thermo Fisher Scientific (Proxeon), Odense, Denmark]. Peptides were loaded onto the 2-cm Nano Trap column, which had an inner diameter of 100 µm and was packed with C18 particles of 5 µm (Thermo Fisher Scientific), and were separated by reversed-phase chromatography using a 25-cm column that had an inner diameter of 75 µm and was packed with 1.9-µm C18 particles (Nikkyo Technos Co., Ltd. Japan). Chromatographic gradients were started at 93% buffer A and 7% buffer B with a flow rate of 250 nl/min for 5 min and were then gradually increased to 65% buffer A and 35% buffer B over 60 or 120 min depending on the complexity of the sample. After each analysis, the column was washed for 15 min with 10% buffer A and 90% buffer B (buffer A: 0.1% formic acid in water; buffer B: 0.1% formic acid in acetonitrile).

The mass spectrometer was operated in DDA mode, and full MS scans with 1 micro scans at a resolution of 60,000 were used over a mass range of m/z 350–2,000 with detection in the Orbitrap. Auto gain control (AGC) was set to 1E6, dynamic exclusion to 60 s, and charge state filtering disqualifying singly charged peptides was activated. Following each survey scan of each cycle of the DDA analysis, the top twenty most intense ions with multiple charged ions above a threshold ion count of 5,000 were selected for fragmentation at a normalized collision energy of 35%. Fragment ion spectra produced via collision-induced dissociation (CID) were acquired in the Ion Trap, with an AGC of 5e4, an isolation window of 2.0 m/z, an activation time of 0.1 ms, and a maximum injection time of 100 ms. All data was acquired using Xcalibur software v2.2.

c) Database search

Proteome Discoverer software suite (v2.0, Thermo Fisher Scientific) and the Mascot search engine (v2.5, Matrix Science) were used for peptide identification and quantification [389]. Samples were searched against a customized database for each species as described in the corresponding section. Trypsin was chosen as the enzyme, and a maximum of three miscleavages were allowed.

Carbamidomethylation (C) was set as a fixed modification, whereas oxidation (M) and acetylation (N-terminal) were used as variable modifications. Searches were performed using a mass accuracy enforcement of 7 ppm, which goes accordingly with the accuracy of the Orbitrap mass analyzer, and a product-ion tolerance of 0.5 Da. Resulting data files were filtered for FDR < 1.

d) Targeted MS

MS1 Targeted Area Extraction was performed with Skyline v3.7.011317 and using RAW files acquired in the Orbitrap Velos Pro that contained heavy-labeled internal standards (Dataset 4).

4.5.7. Conservation analyses: detecting homology and potential pseudogenes

An ORF was considered as conserved when it was found in three or more species. Three different thresholds were taken into account to assess the presence of the annotation in different bacteria. These thresholds were applied to the results by running a BLASTP of the amino acid sequence of the ORF of interest against a protein database comprising a complete six-frame genome translation of 109 different bacterial species. Filter parameters included the e-value, the percentage of target sequence aligned, and the difference in length between the target and the hit. Thresholds for the three parameters were computed using the annotated proteins of the organism of interest as a reference. In the case of *M. pneumoniae*, 95% of the annotated proteins (with no size discrimination) have e-values smaller than 3×10^{-8}, more than 75% of their lengths aligned, and differ with the matched hit in < 20% of their length. We considered closely related species those sharing > 75% of their annotated proteins when applying the previously explained parameters.

Taking advantage of the conservation study, we implemented in RanSEPs an additional classification task to detect potential pseudogenes or highly repeated annotations that could be artifactually considered as coding. With this in mind, we classified every ORF into seven groups: 0 - no hits passed the thresholds defined; 1 - conserved with an annotated function; 2 - conserved as an annotated SEP in NCBI but no associated function; 3 - conserved in a different species but target and homologous sequence not found in NCBI; 4 - sequence is completely or partially (> 75%) repeated ≥ 3 times in the reference genome; 5 - potential pseudogene; and 6 - to depict those annotations that are found in the reference NCBI annotation file. Pseudogenes (type 5) are generally derived from a non-synonymous mutation that partially or totally truncates a protein. In these cases, the presence of an in-frame start codon downstream of the mutation can give rise

to a fragment of the original gene sharing its properties. To detect such cases, RanSEPs searches for cases where a SEP in the reference genome (gene A') was near a downstream or upstream gene (gene A) and these two together (gene A-A') were homologous to a single gene in any of the closely related species. In this case, gene A' would be labeled as a potential pseudogene.

4.5.8. RanSEPs methods

RanSEPs implementation is fully based on Python (version >2.7.x), using functions included in and tested in the scikit-learn package [390]. A fully functional version of RanSEPs is documented in and downloadable from GitHub and http://ranseps.crg.es/.

a) Set definition

In this step, it is important to define closely related organisms in the database to avoid an overestimation of conserved smORFs. This process is automatically performed by RanSEPs after evaluating the complete conservation database. The non-conserved smORFs are randomly and iteratively sampled with the selected set size. For the positive set, a minimum size of 100 true proteins is required. Although it is preferred that this set includes all the annotated SEPs of the organism, the user can define the specific percentage of SEPs that are included.

b) Protein feature computation

Complex featurization of sequences was performed using the Python package *propy* [391]. This package computes more than 1,500 features for each single sequence, covering protein attributes like amino acid composition, dipeptide and tripeptide composition, Moreau–Broto, Moran, and Geary autocorrelations, sequence-order-coupling number, and physicochemical properties. Importantly, as many of these features present a high correlation, including all of them could strongly over fit our training and test sets. To avoid this problem, we ran a Recursive Feature Elimination (RFE) to prune the least important features from the trees (i.e., features that do not efficiently separate positive and negative sequences). We applied this approach over the 109 organisms and selected the three best features by average: quasi-sequence-order-coupling numbers based on the Schneider–Wrede physicochemical distance matrix, hydrophobicity, and secondary structure.

RF classification enables features to be sorted by their importance and then compares these weights in a quantitative manner. Taking this into account, we added several sequence attributes of specific biological interest to the comparison of coding features between microorganisms. These are as follows:

- *Start codon:* The ATG start codon is prevalent over alternative start codons like GTG and TTG. To consider this effect in the classification, we assigned a binary classification where 0 represents annotations that do not have an ATG codon in their first 5 codons, and 1 represents otherwise.
- *GC content:* GC content is computed as the count of G+C divided by the length of the annotation. As described, GC content has a direct effect on the probability of finding start and stop codons.
- *Ribosome-binding site (RBS) stacking energy:* RBSs are important elements in translation regulation in some bacterial species. As motifs associated with this element can vary between organisms, we represented the stacking energy of the −15 to the start codon window. This value is close to −1.26 of free energy in the presence of the AGGAGG motif.
- *Ribosome presence:* Ribosome presence is included as a binary value where 1 indicates the presence of any of the possible Shine–Dalgarno sequences known to act as an RBS [392].
- *−10 + 20 stacking energy:* Multiple studies suggest that specific sequence requirements at the 5′ end of an mRNA impact translation efficiency. In the same way as for RBS, we computed the stacking energies for the 30 bases spanning the −10 to +20 region (with respect to the start codon).

Special features are measured in a relative manner using a “feature” set that is sampled from the positive set (same properties) but not used in the training process. Features extracted from this set are as follows:

- *−10 score and +20 score:* scores computed for the separate elements based on a position weight matrix (PWM) of those regions computed from annotated genes of the feature set.
- *Hexameric measures:* calculated by sliding a 6-base window along the sequence, starting in frame with the annotation (dicodon frequency), +1, +2, and the combination of all the possible hexamers (n hexamer measure). For each sequence, a single value per frame and in combination is extracted. This value is computed as the logarithmic odds ratio between the observed hexamer frequencies and the expected one computed from the feature set. Both sets of frequencies were normalized by the background frequencies based on the GC content.
- *Codon adaptation index (CAI):* a measure of the deviation of codon presence in a sequence from a background model that is extracted from the feature set of proteins. By implementing this measure in addition to the hexameric measures, we take into account synonymous codons [157].
- *2 amino acids, N and C terminal:* two features representing the importance of specific amino acids at the initiation and termination sites. The importance of these features depends on the species.

c) RF tuning calibration

After defining the types of sequences to include in each set, we exhaustively explored the parameter space to properly calibrate the single classifiers. RanSEPs presents two levels of complexity in its tuning, single classifiers and a global classifier, where the latter is the combination of single RFs. Tuning of single classifiers was performed in an exhaustive manner, iterating and testing every combination between: (i) 2, 4, 8, 16, 32, 64, 128, 256, 512, and 1,024 trees; (ii) 10- and 25-fold cross-validation; (iii) test sizes of 1, 5, 10, and 20%; (iv) positive, negative, and feature set sizes between 100, 200, and 300 sequences; (v) percentage of SEPs in each set between 0, 5, 10, 25, and 50%; (vi) maximum depth of the forest between 0, 10, and 20; and (vii) minimum samples per leaf from 1 to 20. For each combination in that parameter space, we combined 5, 10, 20, 30, and 50 single classifiers into the global classifier.

We then tested their accuracies based on the AUC of their ROC curves to find the best parameters using the same test size combinations of single RF in a global manner. In the end, we ended up with the default configuration shown in Table 4.2 for *M. pneumoniae*. This set of parameters worked properly in organisms with < 100 annotated SEPs and genome sizes < 1 kilobase. In the case of organisms with multiple SEPs (> 100) already annotated in NCBI and a bigger sized genome (> 1 kilobase), we observed more adjusted predictions (an equal TPR but a lower FPR) when increasing the negative set size to 2,000, and 85% of SEPs in the positive/feature set with size equal to 200. These rules are implemented in RanSEPs as automatic considerations. RanSEPs can run as a general random forest algorithm (1 classifier) with regular k-fold cross-validation procedures or generating multiple classifiers with randomized training sets to provide an averaged probability.

Table 4.2. RanSEPs default settings.
Parameter configuration used for the detection of proteins in *Mycoplasma pneumoniae*.

Parameter	Value
Positive set size	100
Negative set size	500
Feature set size	100
Percentage of SEPs in positive and feature set	25
Number of single classifiers per general classification	5
Number of trees	100
Maximum depth	0
Minimum samples per leaf	5

d) Feature weight estimation

An out-of-bag (OOB) approach was implemented to compute the importance of each feature in the classification task. This algorithm works by leaving a group of labeled points that will be classified out of the training set. For each classification, the algorithm permutes a feature while leaving the rest unchanged, and measures the error increase comparing the labels with the classes assigned.

e) RanSEPs output

RanSEPs output includes several files related to coding-potential features and classification stats, in addition to the classification task results (Dataset 23). An additional "parameters.txt" file is generated in order to keep track of the parameters used in each specific execution.

4.5.9. Validation set definition

The positive set (n = 570) comprises 307 SEPs detected by MS with ≥ 2 UTPs from the six Mycoplasma species considered, *Escherichia coli*, *Pseudomonas aeruginosa*, and *Staphylococcus aureus* (Datasets 2-3 and 6-13). Second, we performed searches in six additional public datasets for *Lactococcus lactis*, *Helicobacter pylori*, and *Synechocystis*, extracting a total of 166 SEPs (Datasets 14-16) [393–395]. These two sets together (n = 473) not only included multiple annotated proteins from the organism used as a reference (n = 335) or in a closely related one (n = 87), but also 25 potentially new SEPs that were not previously annotated in the corresponding reference genomes or other organisms of the RanSEPs database (Dataset 17). Third, six SEPs detected by targeted MS (MS1 Targeted Area Extraction) using C13C(6)15N(2)-labeled peptides from the *M. pneumoniae* proteome (Dataset 4). Fourth, 97 previously reported SEPs from six different bacteria, well-characterized in the literature and experimentally detected (Dataset 18) [97,123,258,376,396].

The negative set (n = 570) was extracted from two different sources. First, we randomly selected 556 SEPs from a collection of putative SEPs satisfying the following criteria: (i) ≥ 2 HR_UTPs by PeptideSieve; (ii) no NUTP/UTP signal by MS; and (iii) not conserved in any closely related bacteria (highest e-value > 0.01 by BLASTP). This set was balanced to be comparable with the positive set (the same average amino acid length (35 aa)) and to be representative of the 12 bacterial species considered (Dataset 18). Additionally, we included 14 SEPs detected with 1 UTP but not detected by C13 proteomics: 2 found in *M. pneumoniae* and 12 in *Helicobacter pylori* [253].

4.5.10. Annotation tool comparative

As quality metrics for the prediction, we used the accuracy (rate between true positives and true negatives over the total number of tested SEPs) and the AUC between true-positive and true-negative rates (the closer to 1 the better). AUC was measured by ROC curves, and accuracy was supported by precision–recall curves. All searches for validated SEPs using RanSEPs were performed excluding the target proteins of the training process. To run BASys predictions, we used their web service (basys.ca) and selected the arguments: gram-positive/negative and providing specific CDS nucleotide sequences of each target organism to perform a customized search. A CPC search was performed at their website (cpc2.cbi.pku.edu.cn) using the default general search, providing each target genome putative ORFs. GeneMarkS (exon.gatech.edu/Genemark) was run using the default search and selecting the TGA option as a Tryptophan codon for Mycoplasma species. To predict genes with Glimmer, we used the desktop version 3.0 downloaded at ccb.jhu.edu/software/glimmer/. In order to adjust the search for predicting small proteins in each organism, we specifically defined the use of start codons with custom probabilities based on their recurrence in annotated genes (e.g., *M. pneumoniae*: ATG = 0.86, GTG = 0.073, and TTG = 0.067). Additionally, we set a minimum size of 10, and trained the search with the annotated genes of each specific organism excluding the target proteins. To make the comparative meaningful, we standardized the metric provided by Glimmer to a probability scale of 0–1. The last software, Prodigal (github.com/hyattpd/Prodigal), was used as a desktop application forcing a full motif scan of Shine–Dalgarno subsequences, and using the annotated genes of each specific organism excluding the target proteins as a reference.

4.5.11. Functionality studies

Based on their described function in NCBI, the annotated SEPs from the 109 bacterial species were assigned to nine functional categories (Dataset 18). Functions assigned by homology inference were not taken into consideration. Annotated SEPs with known functions were used as the query database to assign functions by homology to the remaining putative SEPs with undefined functions. Homologous gene pairs were defined using the same e-value, aligned length, and shared size thresholds as in the other analyses. The desktop version of Phobius (http://phobius.sbc.su.se/) was used to predict any signal peptides and transmembrane segments in our predicted SEPs using default settings and only differentiating between gram positives and negatives.

4.6. Data and software availability

- Supplementary available at Molecular Systems Biology online.
- RNA-Seq datasets at ArrayExpress: E-MTAB-6203
- Proteomics datasets at PRIDE: PXD008243, PXD010490, PXD011038
- RanSEPs http://ranseps.crg.es/

4.7. Author contributions

SM-V performed computational and statistical analyses, methodology assessment, developed RanSEPs, and interpreted results. TF performed sample preparation for RNA-Seq and MS experiments, wrote the methodology, and corrected and improved the final version of the manuscript. GE-G processed the proteomics samples, wrote the methodology followed, and, together ES, provided valuable discussion about MS results interpretation. RM performed sample preparation for MS experiments and wrote methodology. AG created the webpage where RanSEPs program and results are located. LS provided direct supervision, interpreted results, and helped design this research. ML-S designed the experimental approach, interpreted results, and provided direct supervision of the research. SM-V and ML-S performed the functional study and created the figures and tables. SM-V, LS, and ML-S wrote the manuscript. All authors read and approved the final manuscript.

4.8. Acknowledgements

We thank Dr. Luca Cozzuto from the Bioinformatics Unit at CRG for providing valuable guidance for the conservation studies. Also, we would like to thank Dr. Carolina Gallo, Dr. Eva Yus, and Dr. Raul Burgos for providing MS data. Finally, we thank Dr. Marc Weber for his recommendation about how to analyze Ribo-Seq data. We acknowledge support of the Spanish Ministry of Economy, Industry and Competitiveness (MEIC) to the EMBL partnership, the Spanish Ministry of Economy and Competitiveness, “Centro de Excelencia Severo Ochoa”, the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program under agreement No 670216 (MYCOCHASSIS), the CERCA Programme/Generalitat de Catalunya, the European Regional Development Fund (ERDF) project from Instituto Carlos III (ISCIII, Acción Estratégica en Salud 2016; reference CP16/00094), and “Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya” (2014SGR678). The CRG/UPF Proteomics Unit is part of the “Plataforma de Recursos Biomoleculares y Bioinformáticos (ProteoRed)” supported by grant PT13/0001 of Instituto de Salud Carlos III from the Spanish Government.

Miravet-Verde S; Burgos R; Delgado J; Lluch-Senar M; Serrano L, 2020. FASTQINS and ANUBIS: two bioinformatic tools to explore facts and artifacts in transposon sequencing and essentiality studies. Nucleic Acids Research 48(17):e102

Chapter 5. FASTQINS and ANUBIS: two Bioinformatic Tools to Explore Facts and Artifacts in Transposon Sequencing and Essentiality Studies

5.1. Abstract

Transposon sequencing is commonly applied for identifying the minimal set of genes required for cellular life; a major challenge in fields such as evolutionary or synthetic biology. However, the scientific community has no standards at the level of processing, treatment, curation and analysis of this kind data. In addition, we lack knowledge about artifactual signals and the requirements a dataset has to satisfy to allow accurate prediction. Here, we have developed FASTQINS, a pipeline for the detection of transposon insertions, and ANUBIS, a library of functions to evaluate and correct deviating factors known and uncharacterized until now. ANUBIS implements previously defined essentiality estimate models in addition to new approaches with advantages like not requiring a training set of genes to predict general essentiality. To highlight the applicability of these tools, and provide a set of recommendations on how to analyze transposon sequencing data, we performed a comprehensive study on artifacts corrections and essentiality estimation at a 1.5-bp resolution, in the genome-reduced bacterium *Mycoplasma pneumoniae*. We envision FASTQINS and ANUBIS to aid in the analysis of Tn-seq procedures and lead to the development of accurate genome essentiality estimates to guide applications such as designing live vaccines or growth optimization.

5.1.1. Additional data access

Datasets covering Tn-seq in different selection conditions, gene essentiality predictions and other integrative studies can be accessed by scanning the QR code linked to the original publication. Datasets and supplementary figures and tables, will be numerically referred as “Daset”, “Table S” and “Figure S”, respectively.

5.2. Introduction

Synthetic biology aims to rationally design living systems for practical applications. Ideally, this requires a comprehensive understanding of the organism and a reduction of its genome by removing dispensable genes to create a so-called ‘chassis’ [397]. Transposon mutagenesis is one of the most informative methods for identifying non-essential genes and understanding what is the minimal set of genes required to sustain life. This technique relies on the random disruption of genes to discriminate between those genes that do not accept insertions and thus are required to sustain life (‘essential’; E), those that when inactivated decrease the fitness of the organism (‘fitness’; F), and those which are dispensable under the study conditions (‘non-essential’; NE) [398].

Disruption of genes by transposable elements is commonly driven by transposases [399]. Transposases are enzymes able to randomly insert genetic material into genome regions delimited by inverted repeats (IR) and they can be classified into two types depending on insertion site preferences: Tc1/mariner transposases, which are able to disrupt TA dinucleotide sites, and Tn-5 based transposases, which are assumed to insert without sequence composition restrictions [400]. After transforming the cells, the number of insertion sites in the population, or ‘coverage’, should ideally reach the maximum (i.e. every possible genome position disrupted at least once). Then, mutant cells are selected for by subsequent growth and serial passages. After several rounds of division, cells in which an E gene has been disrupted will disappear from the population and only NE genes will have insertions.

Remarkably, essentiality in an organism may vary between different genetic and/or environmental conditions like during infection [401]. Transposon insertion sites are commonly identified by ultra-deep sequencing in a technique known as Transposon sequencing (Tn-seq) [402,403]. Unfortunately, analysis of Tn-seq data to determine gene essentiality is not straightforward and both biological and technical factors can result in errors. In addition, essentiality is not Boolean (E or NE); there is also a third set of genes called fitness genes (F), in which the probability to find insertions depends on the capability of mutants carrying these mutations to compete with the culture population. Hence, F genes can be defined as NE or E depending on the rounds of passing [404] selection and experimental conditions [234]. In E genes, it is common to find insertions in the N- and C-terminal regions as these are not expected to disrupt the functional core of the encoded protein [405–408]. The presence of NE domains and high abundance and long protein half-lives are also factors to consider [234]. For example, cells with an insertion in an E gene that encodes a protein with a long half-life will still

survive until the corresponding protein is not depleted through dilution by cell division. Similarly, the gene of an essential metabolic enzyme could have insertions until the metabolite produced by the enzyme runs out. Finally, due to the high sensitivity of deep sequencing, it cannot be discarded that transposon insertions occurring in E regions (not viable) could still be detected if dead cells with those insertions remain in the sample.

At the technical level, increased read counts for an insertion position can be found because of PCR duplicates that are produced during the transposon sequence enrichment step [409] (Figure S1). Despite available software being able to count these duplicates as one [410,411], the effect of removing the duplicates on essentiality assignment is still unclear. Also, in Tn-seq the exact insertion position can be miss-mapped due to a high error rate when sequencing specific regions such as homopolymers [412]. Miss-mapped insertions can also arise due to chimeric sequences, which can be generated when combining chromosomal DNA with the inserted sequence, and that by chance, may match another genomic locus [413]. Furthermore, there can be issues regarding the transposon insertion itself because different transposases prefer different nucleotide compositions. For example, the Tc1/mariner transposase only disrupts TA dinucleotides sites and as such, it is necessary to correct for the GC content [231]. Even the Tn5-based transposases, which presumably do not present this bias [232], have been reported to favor AT-rich regions [234]. Some transposases also produce staggered cuts that result in target site duplications (TSD) [414,415]. The impact of these factors on the analyses and interpretation of Tn-seq data has not yet been addressed.

Finally, when running Tn-seq experiments it is also important to consider how essentiality is estimated. Multiple approaches have been proposed and include different metrics, normalizations [239,416], and methods based on different statistical models [405–408]. A complete Tn-seq analysis requires multiple parameters as well as the use of a training set of genes or ‘gold set’ that can introduce additional biases depending on the assumptions taken. For example, to define a NE gold set some models took genes not conserved in closely-related species [234] while others use non-coding regions [417]. This problem is especially important in organisms with little or no knowledge on their basic biology. In general, software tools to extract insertion profiles and posterior analyses of Tn-seq procedures are focused on Tc1/mariner-based protocols [241,418], and are not really applicable for Tn5-based Tn-seq as they only account for TA site disruption. Although a variety of methods have been proposed, there is still no in-depth study aimed at understanding how the combination of data treatments with different assumptions and approaches impacts the extraction of essentiality information.

To solve the above issues in an unbiased manner, we have developed two software packages: i) a pipeline for the detection of transposon insertions called FASTQINS, and ii) a framework for the ANalysis of UnBiased InSertions, or ANUBIS (Figure 5.1A). Together, these packages take into account the aforementioned issues that are ignored in currently available bioinformatic solutions (Figure 5.1B), and create a benchmark to facilitate comparison, analysis and assessment of genome essentiality. To test the methodology we generated a Tn-seq dataset by transforming the genome-reduced bacterium *Mycoplasma pneumoniae* with the mini-transposon pMTnCat_BDPr, which encodes the Tn5-like transposase Tn4001 (Figure 5.1C). This microorganism has a genome of ~860 Kbp, 40% GC-content, 689 protein-coding genes and is an excellent systems and synthetic biology model organism [277,278]. In addition, *M. pneumoniae* presents unprecedented high transposon transformation efficiency rates that ensure a high initial insertional coverage along the genome (1 insertion every ~3 bp in this study; 1 insertion every ~4 bp in a previous study [234]), preserved when only considering coding regions (2 insertions every ~7 bp). Using this model, we analyzed multiple rounds of passage selection and the associated essentiality estimates (Figure 5.1D). Using ANUBIS, we then compared different essentiality landscapes by passage, processing steps, and model estimates (Figure 5.1E).

In light of the increasing use and potential of Tn-seq, we envision that our new tools will further the development, implementation and understanding of this technique, and help pave the way toward new and improved applications. FASTQINS and ANUBIS will have a direct impact on concepts related to essentiality, like genome reduction, essentiality of genomic regulatory regions, and protein modularity. Moreover, with the current global need for new vaccines, accurate identification of virulence factors essential in the pathogenic process but not for the cell viability, by using a library of transposon mutants in animal models as inoculum, could make possible the design of effective attenuated vaccines.

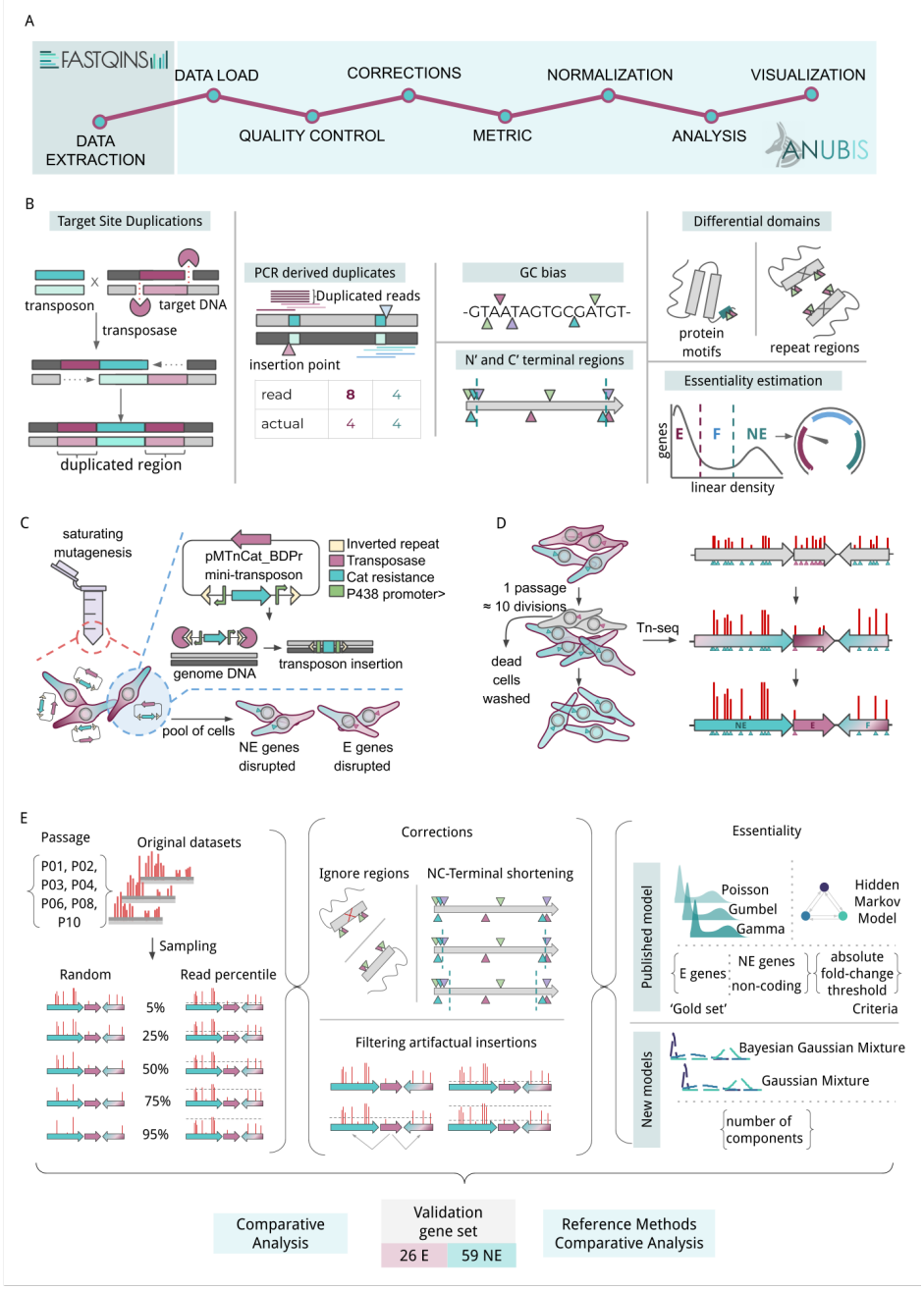

A
FASTQINS
DATA LOAD
CORRECTIONS
NORMALIZATION
VISUALIZATION
DATA EXTRACTION
QUALITY CONTROL
METRIC
ANALYSIS
ANUBIS
B
Target Site Duplications
transposon
target DNA
transposase
duplicated region
PCR derived duplicates
Duplicated reads
insertion point
read 8 4
actual 4 4
GC bias
-GTAATAGTGCGATGT-
N' and C' terminal regions
Differential domains
protein motifs
repeat regions
Essentiality estimation
genes
E
F
NE
linear density
C
saturating mutagenesis
pMTnCat_BDPr mini-transposon
Inverted repeat
Transposase
Cat resistance
P438 promoter>
genome DNA
transposon insertion
pool of cells
NE genes disrupted
E genes disrupted
D
1 passage
≈ 10 divisions
dead cells washed
Tn-seq
E
Passage
Original datasets
P01, P02, P03, P04, P06, P08, P10
Sampling
Random
Read percentile
5%
25%
50%
75%
95%
Corrections
Ignore regions
NC-Terminal shortening
Filtering artifactual insertions
Essentiality
Published model
Poisson
Gumbel
Gamma
Hidden Markov Model
E genes
NE genes
non-coding
absolute fold-change threshold
'Gold set'
Criteria
New models
Bayesian Gaussian Mixture
Gaussian Mixture
number of components
Comparative Analysis
Validation gene set
26 E
59 NE
Reference Methods Comparative Analysis

Figure 5.1. Graphical abstract.
(**A**) Proposed workflow using FASTQINS to process raw sequencing files into insertion profiles and ANUBIS to explore essentiality-related problems and provide estimates. (**B**) Graphical representation of the different issues that are not considered in previous essentiality studies. Target site duplications can double the signal of a transposition event (the transposon in blue is flanked by two different chromosome positions that are at a fixed distance equal to the duplication size). Reads derived from the PCR process can artifactually increase the signal of an insertion point (symbolized as triangles). GC content biases can occur when a transposase shows preference for TA sites. At the level of the protein, 5% of the N'- and C'-termini are arbitrarily not considered because they tend to accept insertions with no impact on essentiality. The differential essentiality of protein motifs and a lack of mapping due to repeated motifs should also be considered. Finally, essentiality can be estimated by different models and assumptions. (**C**) Saturating mutagenesis of *M. pneumoniae* with the mini-transposon pMTnCat_BDPr, which includes a Tn4001-derived transposase and a Cat resistance marker flanked by P438 promoters. With this approach, E and NE genes are expected to be disrupted in a random manner. (**D**) The library was selected along 10 serial selection passages (10 cell divisions each). (**E**) Information was collected from seven different passages (n = 2) and degenerated by two types of sampling. These samples were used to iterate and evaluate different combinations of corrections, essentiality models and criteria. Results were assessed by comparing the level of agreement between estimates and a validated set of 84 genes of known categories.

5.3. Material and Methods

5.3.1. Generation of sample datasets for transposon insertion sequencing analysis

Wildtype *M. pneumoniae* strain M129 (WT) was grown in modified Hayflick medium [277] at 37°C under 5% CO^2 in tissue culture flasks. To generate *M. pneumoniae* mutant libraries, 2 µg of mini-transposon plasmid DNA (pMTnCat_BDPr) was electroporated as previously described [419]. The resulting transformants were selected during 5 days in 5 ml of culture medium supplemented with 20 µg/ml of chloramphenicol, and then harvested in 1 ml of fresh medium. This cell stock was referred to as passage 0 (P0). To assess mutant fitness, transformants were serially cultured through ten consecutive passages as follows. Hayflick medium (5ml) supplemented with 20 µg/ml of chloramphenicol was inoculated with 25 µl of P0. After 4 days of culture (approximately 10 cell divisions), transformants were scraped off the flask in the culture medium, and 1 ml of cell culture (P1) was used for genomic DNA isolation using the MasterPureTMDNA Purification Kit (Epicentre, Cat. No. MCD85201).

In parallel, 25 µl of P1 was inoculated to obtain the next passage, and this procedure repeated until passage 10 (P10). Colony forming units (CFU) in the samples used for genomic DNA isolation ranged between 1×10^8-1×10^9 CFU/ml.

To account for any sampling batch effect, cell passaging and sample collection were performed in duplicate. The pMTnCat_BDPr plasmid used to obtain the transposon library is derived from the mini-transposon pMTnCat [420], which encodes a cat resistance marker. This mini-transposon was modified to include P438 promoters [421] at both ends of the cat resistance gene to minimize any polar transcriptional effects after transposon insertion. To perform these modifications, the cat gene was amplified using the Pr_cat_F and Pr_cat_R primers, and cloned by Gibson assembly into a pMTnCat vector opened by PCR using primers p_Pr_F and p_Pr _R (Table 5.1).

pMTnCm vector primers	
Pr_cat_F	ACTTTATTAATTCTAAATACTAGGGCCCCCCCTCGAGGTC
Pr_cat_R	ACTTTATTAATTCTAAATACTAGCGGCCGCTCTAGAACTA
p_Pr_F	TAGTATTTAGAATTAATAAAGTTTTTACACAATTATACGGACTTTATCAGCTA
p_Pr _R	TAGTATTTAGAATTAATAAAGTTTTTACACAATTATACGGACTTTATCTAGTC
PCR primers, the nested mix was composed of an equimolar mixture of:	
R2-Tn+1	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTVTTTTACACAATTATACGGAC
R2-Tn+2	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTVVTTTTACACAATTATACGGAC
R2-Tn+3	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTVVVTTTTACACAATTATACGGAC
R2-Tn+4	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTVVVVTTTTACACAATTATACGGAC
Illumina sequencing primers	
Tn-PA	TTTTACACAATTATACGG
R1-PA	ACACTCTTTCCCTACACGACGCTCTTC

Table 5.1. List of primers used in transformation and sequencing (5'– 3')

5.3.2. Library preparation

Between 10 ng and 500 ng of genomic DNA were fragmented to 200–300 bp using a Covaris S2 instrument (Figure S1). End repair and adaptor ligation was performed using the E7370L NEBNext Ultra DNA Library Prep kit for Illumina according to the manufacturer's instructions, except that the adaptor used contained only the read 1 adaptor sequence and not the standard Illumina Y-shaped adaptor containing read 1 and read 2 adaptor sequences (Figure S1). The adaptor ligated was amplified with NEBNext Q5 Hot Start HiFi PCR Master Mix in a 50-µl reaction with the R1 PA primer and Tn select PA primer (0.2 µM final concentration) using the following PCR program: 98 °C, 30 seconds; 8 cycles of

98 °C, 10 seconds and 65 °C 25 seconds; followed by a final extension of 5 minutes at 65 °C. The number of PCR cycles required for library amplification was estimated by preparing a 50-µl reaction of qPCR NEBNext Q5 Hot Start HiFi PCR Master Mix and adding SYBR Green I (10,000× in DMSO Sigma Aldrich) to at a final concentration of 0.1×.

PCR was performed in a Roche LightCycler LC480 for 30 cycles using the same conditions as for the first PCR reaction. The first PCR (1 µl) was used as template and the Universal PCR Primer (NEB) and R2 TN select nested primer mix were used at a final concentration of 0.2 µM. The remaining 49 µL of the first PCR were purified using 1.8 volumes of AMPure XP beads (Beckman Coulter) according to the manufacturer's protocol. The purified product was eluted in 48 µL of EB buffer (Qiagen). A second PCR was performed using 15 µl of the purified PCR product (Figure S1), with the number of cycles estimated from the previous qPCR (cycle number close to plateau minus 3 cycles due to the increased amount of template). PCR conditions were the same as in the qPCR using NEBNext Q5 Hot Start HiFi PCR Master Mix in a 50-µl reaction and the Universal PCR Primer (NEB) and R2 TN select nested primer mix were used at a final concentration of 0.2 µM, but with SYBR Green I omitted. The second PCR was purified using 1 volume of AMPure XP beads and eluted in 20 µl of EB buffer. To complete adaptor sequences and add sample barcodes, a third PCR was performed with NEBNext Q5 Hot Start HiFi PCR Master Mix in a 50-µl reaction using 19 µl of the second purified PCR as a template (Figure S1). The Universal PCR Primer and a suitable NEBNext Multiplex primer for Illumina at a final concentration of 0.6 uM (Table 5.1) were used. The PCR program used was: 98 °C, 30 seconds; 4 cycles of 98 °C, 10 seconds and 65 °C 75 seconds; followed by a final extension of 5 minutes at 65 °C. After the third PCR, libraries were purified using 1 volume of AMPure XP beads and eluted in 20 µl of EB buffer.

Final libraries were analyzed on a DNA High Sensitivity Bioanalyzer Chip (Agilent) and quantified using KAPA library quantification kit for Illumina (Roche). Libraries were sequenced on a HiSeq 2500 using HiSeq v4 sequencing chemistry and 2×125 bp paired-end reads (primers are shown in Table 5.1). The raw data was submitted to the ArrayExpress database and assigned the accession identifier E-MTAB-8918.

5.3.3. A standardized pipeline for transposon insertion mapping

We designed FASTQINS combining software tools generally used in nucleotide sequencing analysis to provide a standardized and reproducible pipeline to process, filter, and map insertions across a genome (Figure S1). FASTQINS

accepts randomly pooled transposon libraries generated using either Tc1/mariner or Tn5-based transposons and can analyze single-end or paired-end sequencing data. FASTQINS starts with an optional processing step where read duplicates are removed using Fastuniq [411]. The next step involves the trimming of specific IRs included in the raw reads (e.g. TTTTACACAATTATACGGACTTTATC, length=26) that are associated with a transposition event. This sequence, which must be provided by the user, is processed by FASTQINS to extract the shortest subsequence that is not present in the genome of interest (using the same previous example: TACGGACTTTATC, length=13). Trimming is required so that reads shorter than the original read that was covering the transposition event can be selected. The following step consists of mapping the reads to the reference genome selected using Bowtie2 [422]. Subsequently, FASTQINS filters the alignment with SAMtools to select paired reads mapped unambiguously with a minimum alignment quality [410]. If a user provides single-end reads or selects that option, the previous steps are identical except the condition of paired mapping is not considered and every mapped read is extracted. The final step of the process uses basic shell text processing tools (awk/grep/sed) paired to BEDTools [423] to subset those reads that are shorter than the original read length minus the shortest subsequence of the IR (expected read length after removing the IR). From these reads, the genomic base position contiguous to the previously removed IR is counted as the insertion point (Figure S1).

The final output includes a file detailing the list of positions where an insertion is found and the read counts associated with that position. Additionally, users can split the mapped insertions by forward and reverse orientation, which can be useful in cases like correcting TSD effects (see Results). Finally, a log file that details settings and messages from the application is generated. To expand the application of these tools, functionalities such as the control and recovery of intermediate processes and subtask parallelization have been added [424].

5.3.4. Insertion maps from transposon sequencing datasets

To generate the working dataset, we ran FASTQINS pipeline over 20 different samples covering 7 different cell passages (1, 2, 3, 4, 6, 8, and 10) with 2 biological replicates (replicate identifier 1 and 2) for each passage and two technical replicates for passages 2 to 4 (replicate identifier 3 and 4, related to replicates 1 and 2, respectively). We considered three different configurations: single-end, paired-end keeping read duplicates, and paired-end leaving out read duplicates (Table S1). As an output, we kept the log of the process with information like transposon recovery rate, and three insertions files: two considering each of the sequencing orientations and the merge. Finally, we also included the de-stranded versions ‘*fw*’ and ‘*rv*’ for forward- and reverse-mapped reads, respectively (Dataset 1).

5.3.5. ANUBIS: a Python framework to perform analyses of insertion profiles in an unbiased manner

We developed a Python framework -called ANUBIS (ANalysis of UnBiased InSertions) to cover from loading to analysis and visualization of data. ANUBIS is mainly supported by the *sample* object. Each *sample* includes specific functions to return basic statistics, parameters, and attributes, such as associated annotation, training gene sets, metadata like dilution, growth time, or passage. This information is used by different inner functions to perform the analyses required by the user (Figure 5.1D). The general flow of steps is as follows:

i) Data load and definition: data can be loaded as a single sample or as a collection. Files generated by FASTQINS, as well as those in WIG (wiggle) format, are automatically recognized as single samples. ANUBIS also accepts samples in bulk, using a tab-delimited file format that includes all the required information.

ii) Quality assessment: ANUBIS includes functions to explore the distribution of insertions, read coverage associated to each position, and correlation between replicates.

iii) Pre-processing: this step includes processes like checking sequencing and annotation biases. For example, the user can detect and apply a correction for positions prone to having artificial signals like those derived from GC biases at the level of the 4-mer, TSD, and mismatch-derived insertions. Also, at the level of annotation, N- and C- terminals, repeated regions (Dataset 2), and protein domains (either selected by the user or automatically predicted) can be corrected by using Change Point Detection algorithms from the Python module ruptures [425]. If CPD is asked, ANUBIS will use this tool to delimit regions with differential linear density using a penalized kernel change point detection as default.

iv) Custom read count filters: ANUBIS includes three filtering functions that can be applied or not depending on the needing of the user: 1) a read filter that accepts user-defined thresholds, useful to perform subsetting of insertion positions based on their read counts; 2) a filter to discard insertions with read counts in the tails of the read distributions based on the assumption that the right tail is composed by over-represented insertions due to sampling [241] and the left tail counts for poorly represented insertions usually associated to artifactual signals from dead cells and the mapping process [234]; and 3) a filter for positions with read values in the range of read counts mapped to E genes. This latter filter is based on the assumption that a list of known E genes should present a clean profile and any insertions within the genes would therefore come from dead cells and/or mapping process artifacts. In this filter, the 95^{th} percentile of read counts for insertions mapped to E genes in a gold standard set is calculated and later used as the minimum value required to trust an insertion. In ANUBIS, each of these filters can be applied with custom parameters defined by

exploration of the data or with a default based on their original reference (e.g. tail filter set to remove the insertion with read count below the 5^{th} and above the 95^{th} percentile of the read count distribution).

v) Metric calculation, standardization and normalization: in addition to general metrics (i.e. mean, standard deviation, median, minimum, and maximum) and common metrics in DNA/RNA sequencing (i.e. CPM or counts per million of reads and RPKM or reads per kilobase per million reads), ANUBIS also computes three specific metrics relative to a genomic region: transposon-inserted positions (I), read counts (R), and read counts per transposon-inserted position (RI). In a region from position n to m of the genome, I would be the count of disrupted positions from n to m, R would be the sum of reads from insertions found between n to m, and RI would result from the ratio between R and I (R/I). These values can be calculated for annotations provided by the user and/or sliding windows, either overlapping or not. When calculated for regions with a different annotation length (i.e. genes), these values are generally normalized by the length of the annotation. When I is normalized in this way, we obtain the metric known as linear density. Standardization methods such as min–max scaling and z-standardization can also be applied in ANUBIS.

vi. Sampling methods: these functions derive new datasets from previous samples. This process can be performed either randomly by removing a specific number of insertions sites or based on read count (Figure 5.1E).

vi. Analysis and visualization: ANUBIS provides multiple procedures to extract essentiality predictions with different methodologies (detailed below), perform differential insertion comparisons, and relate information such as protein domains, repeated regions, and structural information with Tn-seq profiles.

All these processes can be executed independently or in a combined and sequential manner through the protocol class. Furthermore, ANUBIS also include additional functions that can address issues during the design of a Tn-seq experiment, such as defining the most suitable IR for a specific genome, and defining the relationship between expected coverage, number of initial cells, and efficiency of transformation based on a probabilistic model of insertions (see Supplementary).

5.3.6. Gold standard and validation sets

Some of the methods required to predict essentiality categories rely on the definition of the center of each E and NE linear density distribution to later predict the probability of deviating from the center [426,427]. In these cases, a 'gold standard set' is required as a reference and usually includes a list of known E and NE genes for which an expected linear density for each category will be computed. Alternatively, the reference center for NE annotations can be calculated from non-coding regions [428] (although in this case, regulatory or important structural regions of the chromosome may be targeted). In this study, we used the same gold standard set as in previous studies using M. pneumoniae as a model [234] (Table S2). This list includes 27 known essential genes, and comprises ribosomal RNA, tRNA synthetases, DNA and RNA polymerases complexes, sigma 70 factor, and glycolytic enzymes required for ATP production. Also includes 29 genes not found in the very closely related species Mycoplasma genitalium as NE genes. Additionally, we defined a validation set for performing the accuracy assessment of each method. This validation set included the previously defined gold standard set plus 29 genes that were successfully knocked out or deleted [285] (n=85). For these 29 genes, we also had phenotypic growth information and information regarding transcriptional changes. This information enabled us to define a set of 6 genes that are potentially F genes because their deletion resulted in a 'slow' growth phenotype [285] (Table S2). Accordingly, we added the remaining 24 genes (no phenotypic changes) to the validation set of NE genes, leaving out the 6 genes that were likely to be F genes for specific observations. Alternatively, non-coding regions can be used as NE gold standard set (automatically defined as genome bp not located in known annotations), this is a common option when exploring essentiality based on linear density using Gamma and Gumbel distributions.

5.3.7. Essentiality estimate models

ANUBIS implements a collection of previously defined and novel methods (Figure 5.1E, Table 5.2). Firstly, we re-implemented as estimate models in the framework methods presented in previous studies based on Poisson [234], Gamma [427] and Gumbel [426] distributions (*italic* names will refer to a class object implemented in the framework). These methods rely on the definition of a gold standard set to estimate the centers of each gene population (E and NE/non-coding regions depending on the study; see previous section), and then classify each gene based on their probability of fitting the expected distributions. At this level, different criteria have been applied to assign essentiality classes. Poisson-based classification uses an 'absolute' criterion, assigning the labels E to genes with P (E)>0 and P (NE)=0, NE to genes satisfying P (E)=0 and P (NE)>0, and F

to any other cases [234]. On the other hand, Gamma- and Gumbel-based methods apply a ‘fold change’ approach and consider E genes to be those with $\log_2$(P (E) / P (NE)) > 2, NE to be those with a $\log_2$(P (E) / P (NE)) < -2, and F genes to be those which fall in between [426,427]. The final criterion that can be applied is a probability ‘threshold’ for trusting a probability or not, arbitrarily set to 0.01 in previous studies [429]. While all three methods were implemented in the ANUBIS framework so as to reproduce their original function, this was done in a more generalized manner to provide the user the option of separately selecting the criteria.

Secondly, we developed a new version of a prediction class based on Hidden Markov Models (HMM), taking into account principles from Tn-HMM such as read depth associated with each insertion [428]. This feature is interesting as it enables the detection of NE genes with minimal impact or even advantage on fitness. We defined a new version of Tn-HMM that maintains its basic functionality connected to functions of ANUBIS, but also adapted its application to Tn5-transposase studies and included additional parameterization options.

Thirdly, we implemented two novel methodologies based on Gaussian Mixture Models (GMM) and Bayesian Gaussian Mixture Models (BGMM). These two models share most of the principles with the exception of the algorithm used to fit the mixture-of-Gaussian models. While GMM relies on Expectation Maximisation (EM) to maximize data likelihood, BGMM extends that same EM algorithm to maximize model evidence, including priors, allowing the automatic estimation of components [430]. As an advantage, these methods do not rely on a gold standard set and consequently no prior knowledge about the expected essentiality of the organism is required. These methods enable evaluation by Akaike Information Criterion (AIC), which rewards goodness of fit, and Bayesian Information Criterion (BIC), which penalizes the number of parameters, to define the best fitting for number of categories and return the best model of essentiality [431]. For example, we could ask for three components as the three expected number of categories (e.g. 3 - E, F, NE; Figures 5.3B and 5.3C) and the model will determine the three best gaussian distributions that fit the observed data without requiring any gold standard set. Finally, if the user prefers to perform an essentiality estimate based on a visual exploration, ANUBIS includes a Mixture method that allows the combination of Poisson, Gamma, Gumbel, and lognormal [432] distributions to fit each subpopulation.

Model	Reference	Metric	Priors	Value	Crit
Poisson	Defining a minimal cell: essentiality of small ORFs and ncRNAs in a genome-reduced bacterium (2015).	linear density	goldset E and NE genes	P( )	absolute
Gamma	Defining the ABC of gene essentiality in streptococci (Amelia R. L. Charbonneau, 2017)	linear density	goldset E genes and intergenic NE	P( )	FC
Gumbel	Bayesian analysis of gene essentiality based on sequencing of transposon insertion libraries. Bioinformatics, 29(6):695-703	linear density	goldset E genes and intergenic NE	P( )	FC
HMM	A Hidden Markov Model for identifying essential and growth-defect regions in bacterial genomes from transposon insertion sequencing data	linear density and read values	goldset	States	3

Table 5.2. Previously published methods included in the comparative.

5.3.8. Method comparison

We ran essentiality estimates for all the samples in our dataset with five different model-based methods, testing corrections and parameterizations (Table 5.3, Dataset 3). For previously described methods (Poisson, Gamma, and Gumbel), each method was run under different parameters and class assignment criteria including the parameters associated with their original reference (Table S3; more details at the end of this section). For mixture models (GMM and BGMM), three different component numbers (number of components: 2, 3, and 4) were run. Each of these configurations were iterated with four different filter modes as well as different preprocessing parameters that included or excluded repeated regions and removed different percentages of N- and C-termin. The four filtering modes applied were: *i)* no filtering, *ii)* discarding insertions with a read count lower than 3 (assumes of 1 and 2 are background of the sequencing process), *iii)* filtering out insertions with a read count <95th percentile of reads mapping to E genes (assume E genes in the gold standard set should be clean of insertions), and *iv)* filtering out insertions with a read count below the 5th percentile or over the 95th percentile.

We also developed a sampling analysis that evaluates the robustness of a method and parameter set with the decay in coverage (Figure 5.1E). We reduced the coverage by two means: *i)* randomly and sequentially eliminating 5, 25, 50, 75, and 95% of insertions in each samples (4 replicates), and *ii)* with a gradual threshold to filter out 5, 25, 50, 75, and 95% of the insertions based on their rank in read counts. Each essentiality estimate task derived from one of the described

combinations of parameters was evaluated by two different accuracy values: accuracy and NE Accuracy. The first term is the total number of genes that were assigned to the same category in the method and the validation set (previous section), divided by the total number of genes in the validation set. The second term is computed in the same way but also counts as matches those cases where the model assigns an NE gene to the F class in the validation set. When referred to as ‘default’, we consider the conditions applied in the reference studies (Table 5.3). In all cases we performed basic data processing removing the 5% N’- and C’-termini regions of the genes and a >2 filter for read count positions.

Code	Description	Label	label description
I	Correct biases	0	no regions removed
		1	repeated regions removed, GC and TSD corrected
S	N' and C'-termini	0	No terminal sides removal
		10	CPD defined terminals
F	Filter of reads	0	No filter
		3	Filter out positions with read count <3
		E	Filter out < 95th read count percentile on E genes gold set
		T	Filter out < 5th and > 95th read count percentiles
M	Model	name	Poisson, Gamma, Gumbel, GMM and BGMM
C	Criteria	criterion	absolute, fold-change or threshold 0.01 (Poisson, Gamma, Gumbel)
	Components	2,3,4	number of component (GMM and BGMM)

Table 5.3. Processing and model estimate reference of conditions in the iterative study

5.4. Results

5.4.1. Extracting reproducible datasets from a high-coverage Tn-seq library with FASTQINS

We generated a library of *M. pneumoniae* pMTnCat_BDPr mutants (Figure 5.1C) for which ten passages had been performed (P0 to P10, each passage equivalent to approximately ten cell divisions, two biological replicas; see Material and Methods). Of these passages, we used seven in total: P01 to P04, P06, P08, and P10. Samples were processed using FASTQINS (Dataset 1) under three different processing conditions: i) single-end (U0_PE0, analogous to previously defined approaches [234]), ii) paired-end (U0_PE1), and iii) 'unique' paired-end removing read duplicates (U1_PE1; Material and Methods).

Different mapping modes were evaluated by means of: *i)* the recovery rate (percentage number of reads covering each insertion event), *ii)* the alignment rate of the mapping process (percentage of raw reads mapping unambiguously to the genome sequence), and *iii)* coverage (percentage of positions disrupted). Comparing the three different methods, paired-end processed samples (U0_PE1) showed improvement in all metrics (Table S1). Recovery rates, for example, were significantly higher (Figure 5.2A; Wilcoxon signed-rank test; P=0.005 when compared to U0_PE0), with improvements ranging from 3 ± 3% for P01 to 20 ± 10% for P10 when compared to U0_PE0. Similar improvements were seen with respect to alignment rates (Figure 5.2B; Wilcoxon signed-rank test; P=0.0004 when comparing U0_PE0 to U0_PE1). In terms of coverage, as expected, no difference was found between removing or not removing PCR-derived duplicates, but paired-end approaches performed better than single-end, with a 5 ± 2% increase per sample (Figure 5.2C; Wilcoxon signed-rank test; P=0.004; see Supplementary). These differences imply ~40,000 additional insertions; a meaningful difference when looking for specific disrupted positions. Based on these results, we used the U0_PE1 processed samples for further analyses.

Using the U0_PE1 samples as a reference (for this and the following Results sections), we first assessed the coverage of our library. We had an initial genome coverage of 37.5 ± 8%, which corresponds to 1 insertion every ~3 bp (2.8 ± 0.6 bp for P01, n=2; Table S1). When considering only coding genes in M. pneumoniae to measure saturation (size considered = 697,457 bp), we observed a similar coverage of 32.15 ± 7.8% (3.3 ± 0.8). These values increased to 70.5 ± 11%, which corresponds to 1 insertion every ~1.5 bp, when examining known NE genes from our validation set (1.45 ± 0.2 bp in P01, n=2; Table S2). We then explored the effect of cell passages at the gene level, comparing two metrics

typically used to estimate essentiality: linear density (number of insertions normalized by length) and read count per gene (considering Reads Per Kilobase Million, or RPKM, as a normalization method).

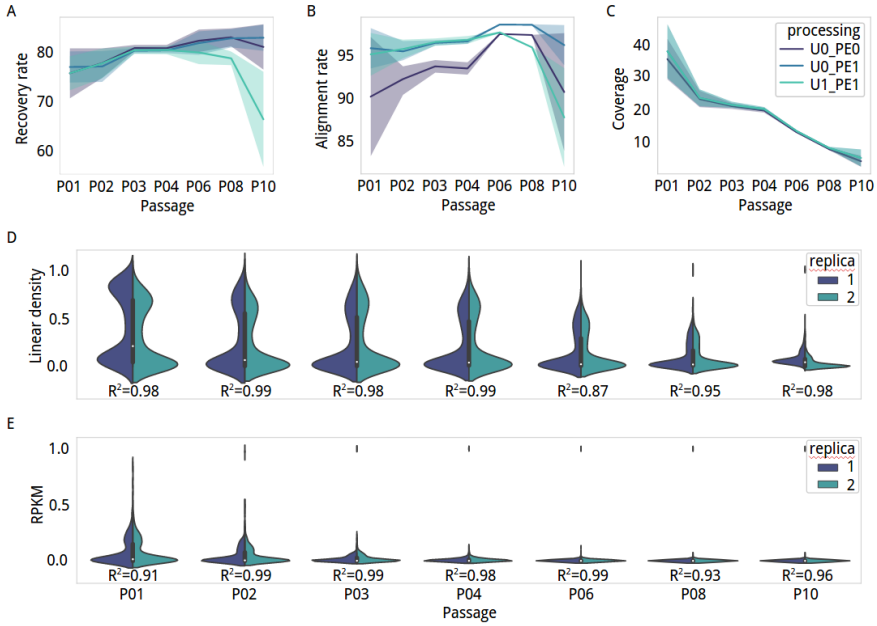


Figure 5.2. Variability of different FASTQINS modes and reproducibility of detection. A-C, Line plots of the (**A**) recovery rate, (**B**) alignment rate and (**C**) coverage (percentage of inserted positions in the genome; genome size: 816 394 bp) of FASTQINS modes run over seven points out of 10 cell passages. The solid lines represent the average values of each metric and the shadows represent variability U0_PE0 (purple) is for samples processed as single-end, and U0_PE1 (blue) and U1_PE1 (light blue) are for samples processed as paired-end, retaining PCR duplicates and filtering them out, respectively. D and E, Distribution of *linear density* (**D**) and *RPKM* (**E**) associated with the *M. pneumoniae* annotated genes (Table S3) by passage. Each side of the violin plot corresponds to one replica (purple for replica 1; blue for replica 2). The R^2 correlation factor between genes in replicas is shown at the bottom of each violin plot. To facilitate evaluation, both metrics were min–max scaled.

With respect to linear density, we observed a bimodal distribution separating E and NE genes even at P10 (Figure 5.2D). Read count distributions, on the other hand, presented a wider dynamic range, losing the bimodal distribution earlier (Figure 5.2E). This is important as a bimodal distribution is expected in essentiality estimate models. In terms of reproducibility, we observed that linear density was more reproducible than RPKM when comparing between replicas. These results indicate that linear density is a more convenient metric in conditions of high selection or with low coverage samples (Table S2 and Figure S2).

A decrease in the linear density associated with an E or F gene is expected with each passage, at least until selection and/or sampling leads to a reduced number of mutants with limited negative, no fitness effect or even positive fitness. Thus, genes with a high RPKM are expected to have a minimal fitness impact when disrupted, because cells with insertions in these genes are the most represented clones in the overall population after selection. For example, we detected that both P01 replicas shared the gene mpn358 (a hypothetical protein of 1,605 bp), with maximum percentage of bases disrupted and maximum read count (85 ± 7% and 9,923 ± 233 RPKM, respectively). This indicates that mpn358 could potentially be removed with no fitness impact or even provide an advantage in growth terms (Table S2). Supporting this, insertions in mpn358 were still overrepresented at P10.

5.4.2. Estimates of essentiality using different methods and default parameterization

We wanted to compare how gene essentiality changes when different published methods are used with their default parameters (see Material and Methods). We included models that statistically fit linear density distributions (number of transposon-inserted positions normalized by the length of the genome region of interest, see Materials and Methods), including Poisson [234], Gamma [426], and Gumbel [433]; as well as HMM [434], which also considers the read counts in the estimate. We also implemented and compared two new models, that do not require prior knowledge on the essentiality of the organism, based on linear density: Gaussian Mixture Models (GMM) and Bayesian Gaussian Mixture Models (BGMM; see Material and Methods) [435]. The only parameter required for these new models is a number of components, which we set to 3 (corresponding to E, F, and NE) to enable comparison with other estimates (supported below). To evaluate the accuracy of each method, we used essentiality information on knockouts and deletions of 29 genes [285]. These same genes are also used later as an NE validation dataset together with a gold standard set of E and NE genes (n=56) previously described [234] (n=85, Table S2; see Material and Methods).

We observed that accuracy (percentage of genes matching with the validation set) and NE accuracy (percentage of genes matching with the validation set considering F genes to be NE; see Material and Methods) gradually decreased with the number of passages due to NE genes being predicted as part of the F or E categories (Figure 5.3A, left panels; Table S3). This effect became more prominent in P08 and P10 indicating that at higher selection conditions only a subset of NE genes, those with minor fitness impact, will be detected as such. In terms of accuracy, Gumbel and the newly proposed methods of GMM and

BGMM, outperformed Poisson, Gamma, and HMM. The former models yielded accuracies of >75% up to P06, while Poisson returned a similar accuracy only for P01 and Gamma, at best, accurately assigned only 54% of the genes found in the validation set. When considering NE accuracy (considering F genes to be disruptible genes), all methods except for HMM performed at over 75% in every passage. HMM became unreliable after P03 (the point at which RPKM lost its bimodal distribution; Figure 5.2E) and did not perform accurately in one of the two replicates for P01.

We accounted for the number of genes that were assigned to each category along passages for each of the estimate models (Figure 5.3A, center panels). In general, we observed NE genes shifting to the F category, and consistency within models up to P06 in terms of the number of genes classified as E (Figure 5.3A, right panels; Figure S3 and Table S4). Interestingly, the best prediction in terms of accuracy and NE accuracy, (91 ± 6% and 97.6%, respectively; n=2) occurred for P01 when analyzed using GMM (Figure 5.3B). In the two P01 replicas, 644 of the genes were identically assigned: 232 E (33.6%), 165 F (23.9%), and 247 NE (35.8%). In contrast, there was a discrepancy for 45 of the genes (16 changed from E to F (2.3%) and 29 changed from F to NE (4.2%)). Additionally, the three components are supported by both the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC; see Material and Methods). Lower AIC and BIC values are associated with models that have a better trade-off between goodness-of-fit and model simplicity (penalizes number of parameters). We observed that with 3 components, AIC and BIC started to flatten (when the gradient stops decreasing there is no risk of overfitting or underfitting; Figure 5.3C and Figure S4).

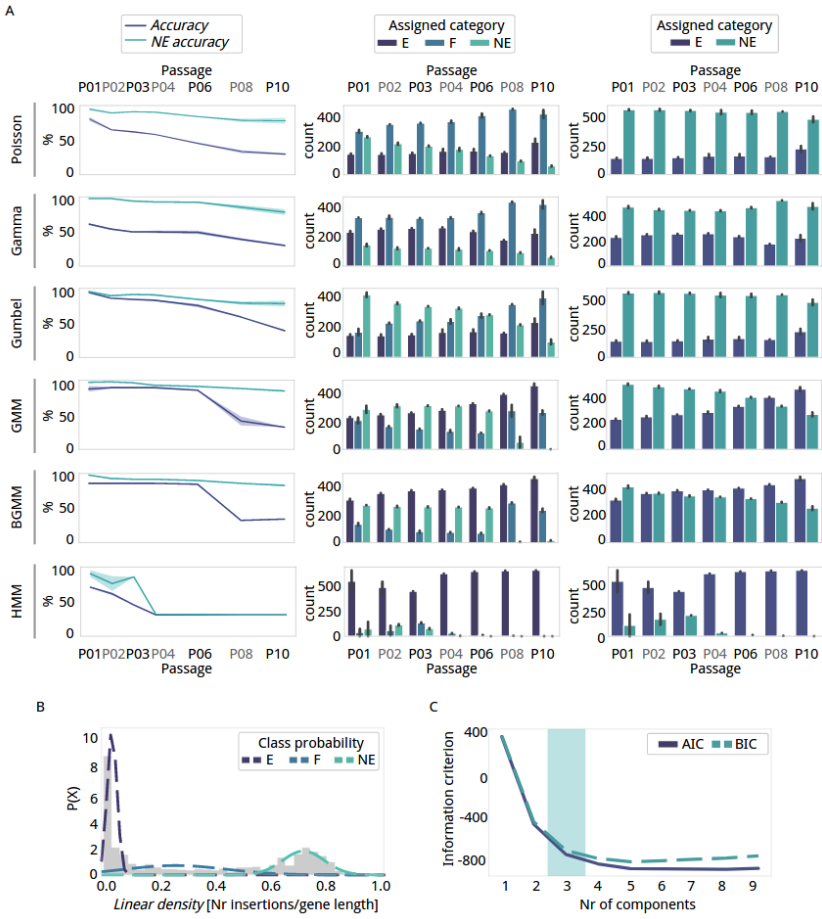


Figure 5.3. Comparison of accuracy and gene category assignment between reference and new essentiality estimate models.

The methods used are labeled on the left (GMM, Gaussian Mixture Model; BGMM, Bayesian Gaussian Mixture Model; and HMM, Hidden Markov Model). (**A**) left panel, *Accuracy* (purple) and *NE accuracy* (light blue) in percentage values for each method per passage. center panel, Number of genes classified as E (purple), F (blue), and NE (light blue). Error bars represent the standard deviation (n = 2). right panel, Number of genes classified E (purple) and NE (blue), with F and NE genes grouped together. Error bars represent the standard deviation (n = 2). (**B**) An example of an essentiality estimate using the Gaussian Mixture Model (GMM) with three components for P01, replica 1 (replica 2 in Figure S4). The gene *linear density* (grey histogram) has been properly fitted to the data using three Gaussian distributions (dashed lines: E (purple), F (blue) and NE (light blue)). (**C**) Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). The lower the AIC and BIC values, the better the balance between goodness-of-fit and model simplicity. The number of components (i.e. 3, blue shadowing) represents the elbow of the line where there is a good trade-off between fitting and the number of parameters.

5.4.3. Important factors to consider when estimating essentiality

We explored factors that could contribute to erroneous insertion signals, or artifacts. These factors were explored through filtering/correction, visualization, and statistical assessments using functions integrated into the ANUBIS framework. We used the U0_PE1 data subset to evaluate these factors at the level of the nucleotide base, the gene and/or essentiality estimate, to exemplify cases where a specific correction can be beneficial in terms of data reliability, accuracy and/or NE accuracy (see Material and Methods). For the sake of simplicity, we only describe the effects on a limited number of the default estimate models from the previous section (see Material and Methods and last section of Results).

a) PCR duplicates

We do not expect essentiality assignments based on linear densities to show differences when removing or not PCR duplicates because the positions inserted do not vary. However, essentiality estimates using models like HMM can be affected by PCR duplicates. We tested this using the P02 samples as a reference, and between replica 1 and 2, observed that 10 and 65 genes changed categories for the U0_PE1 and U1_PE1 mapping, respectively (Table S5). Accuracy did not change between mapping methods for either replica. However, when considering NE accuracy, we found that removing PCR duplicates was beneficial for replica 2 with the value increasing from 66% to 75%. This improvement was entirely due to the correct classification of seven validated NE genes (mpn307, mpn329, mpn346, mpn493, mpn495, mpn560, and mpn653) that were considered E in the U0_PE1 mapping mode.

Confidence detecting PCR duplicates in Tn-seq is problematic. This is because the probability of wrongly detecting reads coming from a clone that is highly represented in the population as PCR duplicates increases with the number of passages (Figure 5.2A). The use of barcodes can provide reliability when approaches like HMM are applied, as they allow for unique transposition events [43]. However, a general essentiality study based on linear density will not show advantages when using barcodes and removing PCR duplicates.

b) Sequence composition biases in Tn-seq

While insertions are only expected to occur at TA-sites with Tc1/mariner-based Tn-seq, when using Tn5 transposase, it is assumed that insertions are uniformly distributed along the genome with no significant biases [436,437]. However, we found some biases against GC sequences in our Tn5 dataset at the base level. As such, we explored the relationship between GC content of each available DNA 4-mer in non-coding *M. pneumoniae* regions and the probability that each gets

disrupted. We found a lower frequency of insertions in GC-rich 4-mers (≥3 G or Cs) as well as a preference for TA-rich 4-mers (4 A or Ts; Figure 5.4A). This effect was also observed when replicating the approach using NE genes from our validation set instead of non-coding regions, indicating that a GC bias also affects annotation (Pearson's R^2=0.92 and P=0.00, when correlating the frequency of 4-mer disruptions between validated NE and non-coding regions in *M. pneumoniae*). Consequently, in ANUBIS we have included a correction function to assess this bias and correct for the linear relationship between available and disrupted k-mers for each passage. We observed that the bias against GC was more prevalent for later passages, suggesting that sampling due to selection could increase this (Figure 5.4B). Finally, we concluded that the Tn4001 transposase (Tn5 family) prefers AT sites over GC ones despite being able to insert in GC-rich sites as well (Figure S5).

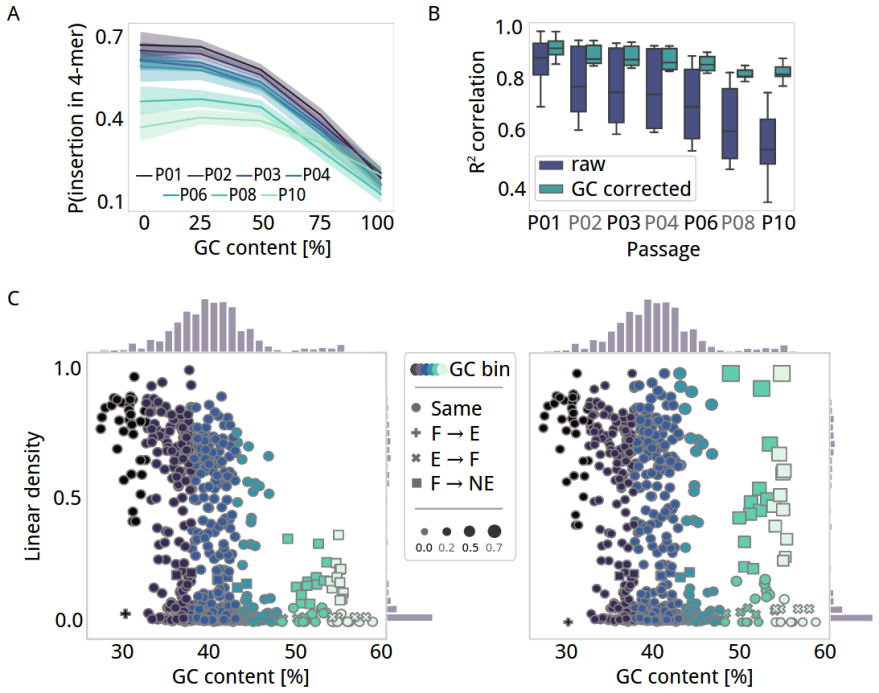


Figure 5.4. Corrections of GC content bias.

(**A**) Average frequency (line) and standard deviation (shadow) of each DNA 4-mer having a transposon insertion as a function of GC content (X-axis). Data is presented for each passage of the U0_PE1 dataset and shows that insertion probability is higher for 4-mers with lower GC content. (**B**) Boxplot representing the contribution of GC bias per passage, measured as the Pearson's R^2 correlation between available 4-mers and disrupted 4-mers. Raw profiles of insertions are shown in purple and ANUBIS-corrected profiles in blue. (**C**) Scatter plots (with histograms) of *linear density* as a function of percentage of GC content for each annotated gene in *M. pneumoniae*, before (left) and after (right) correcting for GC bias by Conditional Quantile Normalization. The legend is shared between the two panels, and a gradient from black to light green represents the following GC content (% units) bins: <32, 32–38, 38–43, 43–48, 48–53, >53 (minimum number of bins with >25 genes each). Changes between essentiality categories, as estimated by *GMM* with components, before and after correction are labeled with the following symbols: a dot for no change, a plus sign for F to E, a cross for E to F, and a square for F to NE. The symbol size represents the difference in terms of *linear density* between the corrected and uncorrected values.

We also evaluated the impact of sequence composition biases at the annotation level and on essentiality estimation using P02, replica 1, as an example. When relating linear density with GC content for each gene in i (genomic GC content of 40%; Figure 4C), we observed that almost all genes with a GC content ≤30% had more than 75% of their positions disrupted (28 out of 31 genes presented an average linear density of 85%). For genes with a GC content ≥50%, we observed significantly lower densities (average linear density of 27%; Wilcoxon signed-rank test; P=0.00). While in the first case we do not expect an impact on the essentiality estimation of AT-rich genes, we could be underestimating the number of NE genes with high GC content. In fact, when running a sliding window approach comparing gene local linear densities, we observed a clear anticorrelation with the percentage of GC (Figure S6).

ANUBIS implements a Conditional Quantile Normalization (CQN) method, validated to correct biases in sequencing processes [438]. This method corrects linear densities assuming full linear density for non-coding regions and using quantile normalization conditioned by GC content and linear regression correction. As changes between GC bias-corrected and non-corrected linear density were small (Pearson's R^2=0.95; P=0.002), we observed few differences in the predicted categories when estimating essentiality (Table S5). Looking at GMM with three components for example, only 45 genes presented different category estimations due to an increased linear density after correction. These genes have a high GC content (48%-54% and >54%). Fourteen genes were corrected from E to F and 31 from F to NE, indicating that their linear density values without correction could have been underestimated. No differences were observed in terms of accuracy or NE accuracy. GC content can be very different depending on the model organism and this kind of corrections could not be appropriate for those cases. However, this correction looks to ensure there are no unbalanced linear densities distributions by GC content and it should be generally effective in other models (see Supplementary).

c) Correlations at the base pair level: Target site duplications

Some transposases produce staggered cuts, and as a result, cause duplication of a fixed number of nucleotide bases during the repair process [439]. For a given insertion event, each of the flanking IR is followed by two different chromosome coordinates, and apparently for short read aligners, two different insertion positions. We evaluated biases at the nucleotide level by correlating read count values (i.e. a representation of a clone in the library) between insertion events and contiguous positions. The most noticeable correlation was between positions n+7 and n-7, a feature conserved in all passage conditions (Pearson's R^2 >0.5, Table S6; Figure 5.5A). This suggests that the Tn4001 transposase produces a 7-bp TSD.

Considering the typical primer for PCR enrichment, which is designed to amplify the sequence from the IR to the contiguous genomic region (Material and Methods, Figure S1), we deducted forward-oriented (*fw*) mapped reads would always cover one side of the insertion while reverse-oriented (*rv*) reads would cover the other. In our case, insertions detected in rv reads corresponded to the same *fw* profile but were shifted by +7 (Pearson's R^2>0.8 for the position n+7; Figure 5.5A). This effect is related to the read count that is associated with each insertion because correlation with the +7 position became significant for those positions with a read count over the 90th percentile in the general read count distribution (Pearson's R^2>0.75; P<0.005 in all passages, Table S7; Figure 5.5B). This means that TSD are more probable to be detected when transposition occurs at an NE position (i.e. clones higher read count). In F regions, however, the read count will be lower and one of the two insertions could be missing and therefore only be counted one. Using the previous observations, we defined a correction that overlaps *fw* and *rv* insertion profiles, but shifting the *rv* positions by +7 if their read counts are over the 90th percentile (Figure 5.5A).

We applied the correction for TSD to sample P02, replica 1, and estimated essentiality using the GMM model with three components. We observed 19 genes changing categories after correction: 4 moving from F to E and 15 from NE to F (Table S5). Interestingly, despite not observing changes in terms of accuracy and NE accuracy, we could be improving the estimate of F genes. With no correction, GMM properly classified two out of six genes that could be considered as F in our validation set because deleting them confers a 'slow' growth phenotype to *M. pneumoniae* (see Material and Methods; Table S3). With the correction, all six genes were predicted as F.

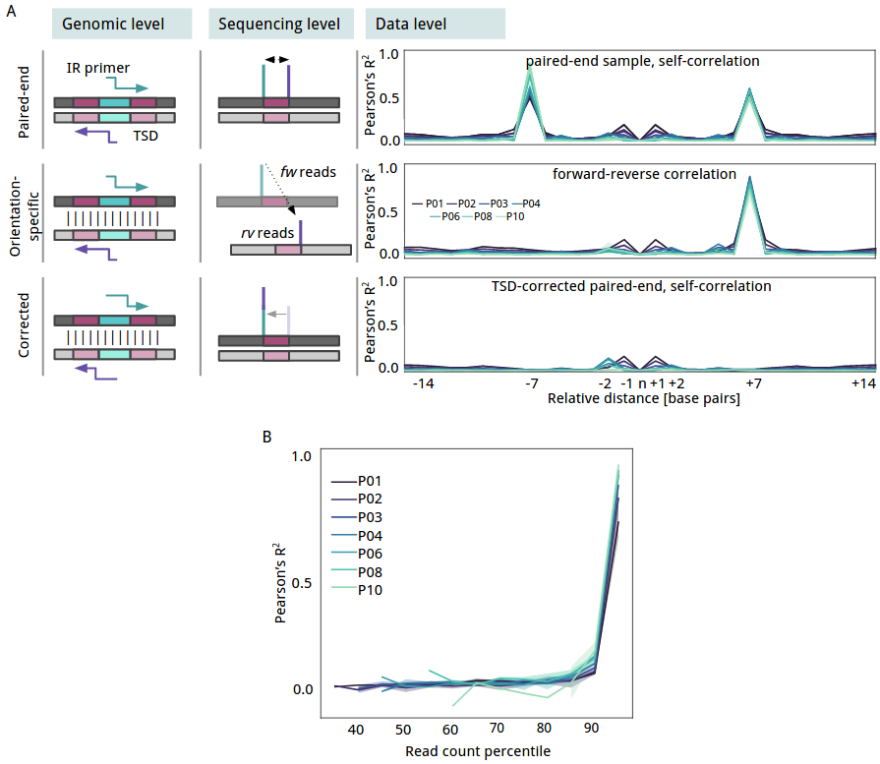


Figure 5.5. Read count correlation at the nucleotide level.
(A) Schema of how TSD causes read aligners to count for the same insertion twice. When we count for regular paired-end reads mapping (first row), at the genome level (left column), the IR sequencing primer (green and purple arrows) extend from two different positions. At the sequencing level (center column), aligners like Bowtie2 will assign the insertion to two positions with a distance that is equal to the size of duplication. At the data level (right column line plots), we show the average Pearson's R^2 correlation between relative positions for each passage (gradient of colors). The X-axes represent a relative insertion in the center and Y-axis correlation in R^2 values to contiguous up- and downstream positions. (B) Exploring the correlation at the level of read count percentile (X-axis) shows that Pearson's R^2 correlation (Y-axis) becomes relevant when the read count of insertions falls above the 90th percentile for each passage (gradient of colors).

d) Differential essentiality regions: N- and C-termini, repeated regions, and protein domains.

It is known that some coding genes can tolerate transposon insertions in the extreme N- and C-termini of their ORF because the insertions are not expected to disrupt the functional core of the encoded protein [405–408]. Previous studies have corrected for this by arbitrarily trimming 5% off each terminal region and considering only the inner 90% region [241]. More aggressive filters have been applied in some studies (e.g. removing 5% from the N-terminus and 20% from C-terminus [440]). These numbers are rather arbitrary and could impact essentiality estimates. We implemented a Change Point Detection (CPD) algorithm in ANUBIS that automatically analyzes the linear density of a gene by windows to detect significant changes [441]. This enables estimation of the best points (change points) delimiting the NE N- and C-termini regions of E genes that could have a different insertion profile to the rest of the gene. For example, taking the annotation of *M. pneumoniae* and all passages as input, we determined the average change points to be at 8% from the N-terminus and 10% from the C-terminus. In P10 for E genes, we detected the average change points at 3% and 4% for N- and C-terminal regions, respectively, indicating that they still conserve insertions at their terminal regions even after multiple selection passages.

In general, the extension of NE terminal regions for E and F genes becomes shorter with each cell passage. For example, *mpn116* is predicted to be a F gene (Poisson model, default) up to P06, at which point it starts to be classified as E using the arbitrary threshold of 5% from each termini. We analyzed this specific case and determined that, at P01, the first half of the protein is labeled as E while the second half is labeled as NE (Figure 5.6A). The differential NE region is maintained from P02 to P06, where it becomes reduced to the last 18% of the gene; being further reduced to 8% and 5% in P08 and P10, respectively. This effect was also observed for other genes, both in the N- and C-termini, indicating a progressive negative trend when insertions are further away from the N- and C-termini. Using P02 as a reference and the Poisson model as an example, we evaluated the effect of not filtering the terminal regions on predicting essentiality (arbitrary 5% cutoff and CPD methodology). Using different filters, we observed no difference in accuracy along passages. However, we did observe 61 genes changing categories when comparing the 5% termini removal versus the CPD approach. For example, genes like *mpn154*, *mpn214*, and *mpn339* were labeled as F when no filter or the arbitrary 5% cutoff filter was applied, but labeled as E when using CPD (Table S5, Figure 5.6B).

The CPD algorithm also enables the automatic detection of cases in which a protein comprises multiple differential essential domains. We hypothesized that E domains within apparently NE genes could either be the result of repeated loci in the genome preventing the mapping of insertions (ambiguously mapped reads are generally counted separately by aligners like Bowtie2) or a specific functional domain in the protein that, unlike the rest of the gene, is essential [234]. To test the first hypothesis, we generated a reference of repetitive DNA sequences in *M. pneumoniae* M129 and observed that mapping was efficient for repeated regions shorter than 100 bases, independent of the passage number (Figure S7, Dataset 2). Hence, repeated regions longer than 100 bases are ignored by ANUBIS when calculating metrics such as linear density.

For the latter hypothesis about protein domains, ANUBIS was designed to accept additional annotations such as HMMER protein domain predictions [442] and report differential essentiality assignments between those domains and the general gene. We tested the impact of these two types of regions on protein essentiality using the Poisson model along different passages. We observed minimal differences along passages, with only 10–15 genes changing category per passage. Despite most changes being between the F and NE categories, some interesting cases arose including *mpn141* and *mpn142* (Figure 5.6C). These genes were predicted as E in every passage condition when including repeated regions but predicted as F after correction (Table S5). In fact, spontaneous mutants for these cytadherence-related genes have been isolated, demonstrating they are dispensable for in vitro growth conditions [443]. Therefore, these results indicate that if repeated regions are considered, specific disruptable genes in an organism could be hidden. In addition, when looking for different essential HMMER domains, we found genes with apparent local differences in terms of linear density. However, all these cases could be explained by the protein having extended N' and C'-terminal NE regions, or E regions derived from repeated regions. Interestingly, while *mpn030* (168 amino acids), which has structural homology to NusB proteins [444], presented an enrichment in linear density from amino acid positions 13 to 53, the rest of the protein (corresponding to HMMER domain DUF1948) had no insertions. Interestingly, *mpn030* has an alternative start codon (GTG) after that specific NE region, suggesting this gene is essential with an NE N-terminal region not required for cell viability (Figure 5.6D). This is supported by the fact orthologs of mpn030 in other mycoplasmas do not present any extension. This could be an effect derived from the acquisition during evolution of an ATG start codon (preferred over GTG), which adds ~50 amino acids without affecting the original protein functionality.

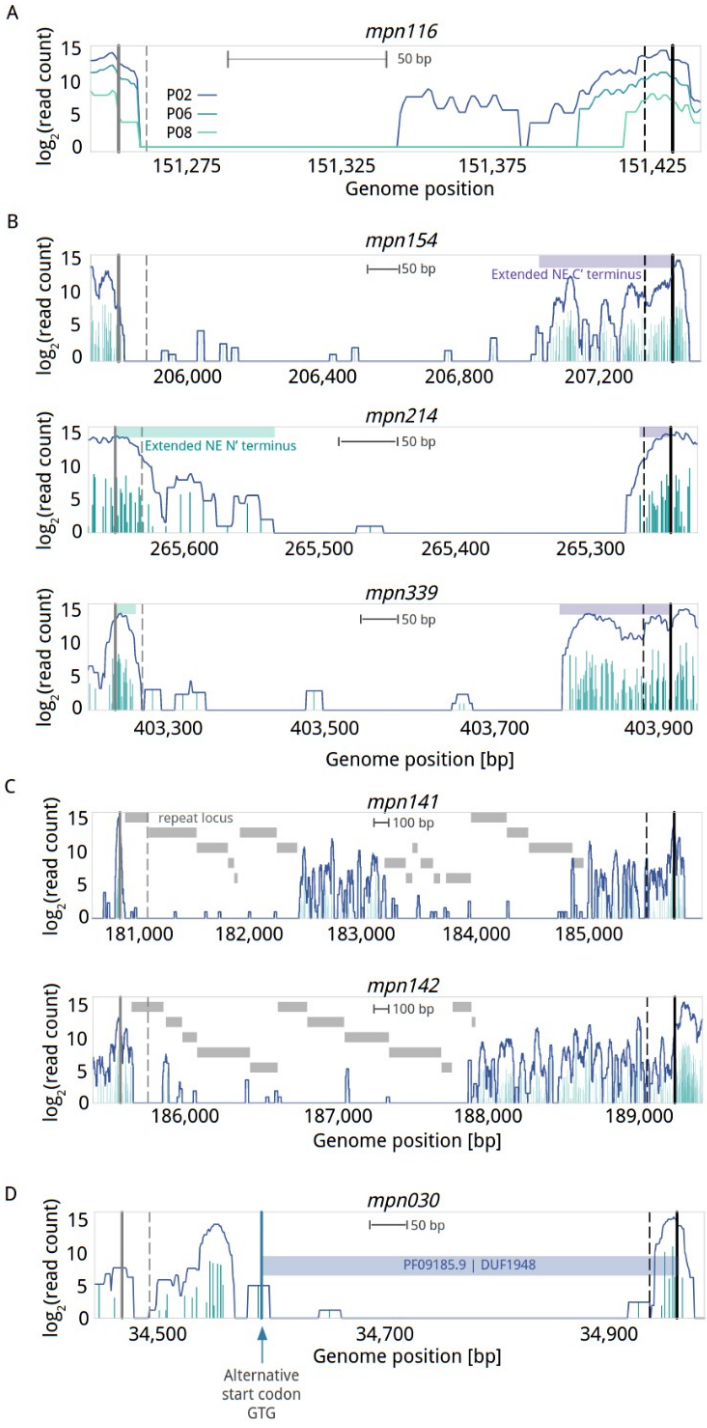

A
mpn116
50 bp
P02
P06
P08
log2(read count)
151,275
151,325
151,375
151,425
Genome position
B
mpn154
50 bp
Extended NE C' terminus
206,000
206,400
206,800
207,200
mpn214
Extended NE N' terminus
50 bp
265,600
265,500
265,400
265,300
mpn339
50 bp
403,300
403,500
403,700
403,900
Genome position [bp]
C
mpn141
repeat locus
100 bp
181,000
182,000
183,000
184,000
185,000
mpn142
100 bp
186,000
187,000
188,000
189,000
Genome position [bp]
D
mpn030
50 bp
PF09185.9 | DUF1948
34,500
34,700
34,900
Genome position [bp]
Alternative
start codon
GTG

Figure 5.6. Insertion profiles for different genes.
The genome coordinates of each gene are shown on the X-axis. Gene coordinates are delimited by start (solid vertical grey line) and stop (solid vertical black line) codon positions and their respective shifted position in the 5% N- and C-terminals (dashed lines). In every plot is shown the smoothed 20-bp distribution of read count per insertion (line) passed onto the CPD algorithm. Base-pair scales are shown below the gene name. (**A**) Gene *mpn116* at different passages, for passages 02, 06 and 08 (darker to lighter colors), presents and extended C-terminal of 50% (passage 1 and 2) that becomes shorter with selection (∼15% for P02–P06; for 5% P08 and P10). (**B**) E genes with extended NE N' and C'-termini at P02. Top profile represents *mpn154*, which presents insertions (solid blue vertical lines) in an extended C-terminal covering 23% of its length (purple box). In the middle, *mpn214* has an extended N-terminal region covering 13% of the protein (blue box) and a C-terminal region of 7% (purple box). The bottom profile represents *mpn339*, which has a shorter N-terminal region (3%) but a longer C-terminal region (18%). (**C**) *Genes with repeated regions* at P02. Examples of potential F/NE genes (*mpn141* and *mpn142*) that are predicted to be E when including repeated positions in the estimation (grey boxes). (**D**) *Mpn030* at P02. This gene is a NusB-like protein with a dispensable N'-terminal. Insertions before amino acid 58 still enable the expression of a functional, shorter version of the protein because of an internal start codon (labeled with blue arrow) that still expresses the domain of the protein found conserved by HMMER.

5.4.4. Effect of coverage, methodology, and corrections on predicting gene essentiality

We performed a general evaluation of models by examining how linear density is affected by transposon coverage and different estimate parameters (Figure 5.1E and Dataset 3; see Material and Methods). This analysis is important because in vivo essentiality studies, for example, result in a much lower transposon insertion density than in vitro studies due to stronger sampling and selection conditions. Also, we could have lower coverages when we analyze larger genomes like *Escherichia coli* (4,000 Kb) where transposon insertion saturation is harder to achieve. We first explored how accuracy is related to the coverage reduction that is produced by continuous selection (i.e. over passages). We observed that a genome coverage of at least 10% (10 insertions every 100 bp) is required to provide accurate estimates of both accuracy and NE accuracy. As described above, estimates made by Gumbel, GMM, and BGMM outperformed estimates made by Poisson and Gamma models independently to the filters and processing steps had been applied (Figure 5.7A and 5.7B).

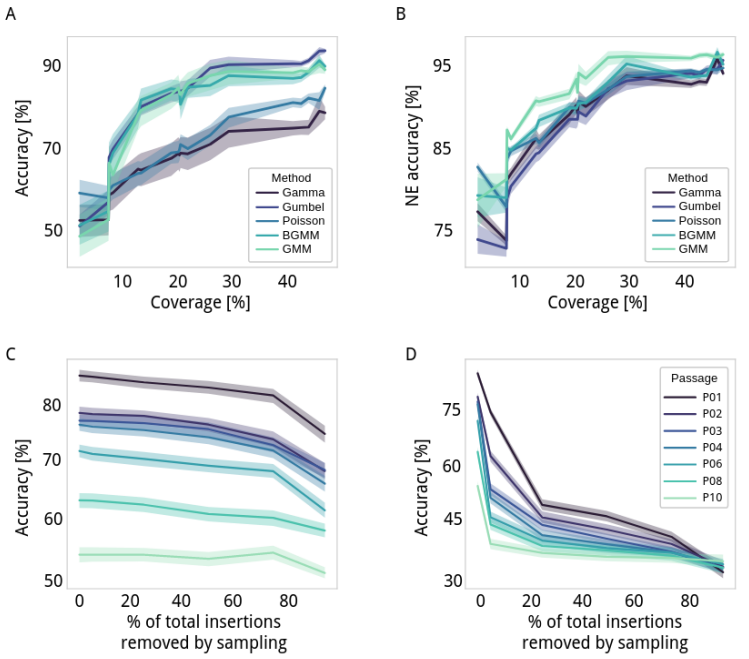


Figure 5.7. Comparison of essentiality estimates for different passages and different parameterizations.
(**A**) Line plots representing *accuracy* and (**B**), *NE accuracy* for each coverage found in our dataset). Solid lines represent the average accuracy of each model (different gradient colors) and shadows represent the expected variability as standard deviation. (**C**) Impact of randomly removing insertions on *accuracy*. The X-axis represents the sampling level, or the percentage of inserted positions in the sample that are randomly removed. Solid lines represent each of the samples (different gradient colors) and shadows represent the expected variability as standard deviation. (**D**) Same as panel c but with the sampling method based on *read count* values (e.g. at 75% we consider only those insertions with a *read count* >75th percentile of the total read distribution).

Secondly, we artificially produced a sampling effect by randomly removing insertions from a profile in a sequential and controlled manner (Figure 5.1E). We were able to randomly remove up to 75% of the insertions in a dataset without losing accuracy. This indicates that sampling effects that occur during passages (i.e. dilution of cell populations performed between each passage) do not account for large differences in essentiality estimates (Figure 5.7C) but it could affect specific annotations (e.g. short ones, see Supplementary). Additionally, we explored with a sampling method based on subsequently increasing a read count threshold (Figure 5.1E). We found that positions with a read count below the 5th percentile are required for proper estimation of essentiality based on our validation set. In each sample, the 5th percentiles corresponded to a read count of 3-4, indicating that most of these low read insertions are real despite the fact that they can be caused by artifactual factors such as the ones described above (Figure

5.7D). However, it is common to find insertions with a read count of ≤2 in E genes. Thus, we considered three different types of read filters in the comparative iterations: *i)* removing positions with a read count of ≤2, *ii)* trimming 5% of the read count distribution from the top and bottom (i.e. 'tails') [241], and *iii)* filtering out insertions with read values in the range of read counts mapped to know or validated E genes (i.e consider those insertions as 'noise' derived from dead cells or mismapped positions, see Material and Methods).

Lastly, we explored the variation in accuracy produced by each preprocessing mode, including models, the three different read threshold filters mentioned above, corrections for repeated, TSD, N- and C-terminal extended regions, criteria used for assigning essentiality categories, and definitions of expected NE linear density from the gold standard set or non-coding regions (Figure S8). As already mentioned, Gumbel, GMM, and BGMM models presented the best overall accuracy, with BGMM showing considerably less variability than other Material and Methods. With respect to filtering by read counts, we observed that removing insertions with a read count smaller than 3 was beneficial when estimating essentiality, improving estimation of E genes and F genes in the validation set, but at the cost of accuracy in detecting NE genes. The accuracy in detecting NE genes also decreased, albeit more aggressively, when applying filters based on E genes or removing tails (Figure S8). Correcting for repeated regions, TSD artifacts, and the use of a CPD-based definition of N- and C-termini did not improve overall accuracy (Figure S8). However, we already described how these corrections were beneficial for specific genes. Similar as when correcting for GC biases, these corrections should be specifically applied at the gene level. Finally, for Poisson, Gumbel and Gamma models we evaluated different gold standard sets, and class criteria definition (see the last two sections in Material and Methods). We found that the best criterion for estimating essentiality with these models is the fold change (FC) between E and NE probabilities, where $\log_2$FC<-2=NE and $\log_2$FC>2=E (Figure S8). For GMM and BGMM models, we found that two components provided the best accuracy, although this is at the cost of losing the F category. With respect to the gold standard set, estimating the expected linear density of NE genes from non-coding regions provided more accurate estimates than using a user-defined gold standard set (Figure S8).

Overall, estimation of essentiality is a complex task that requires multiple evaluation steps and consideration of factors that, despite not introducing dramatic changes in the general assessment of essentiality, can lead to the incorrect estimation of a specific set of genes. ANUBIS includes all the necessary functions to run Tn-seq data analyses from scratch so that the user can visually and analytically explore the impact of each of the introduced corrections.

5.5. Discussion

Here, we first presented FASTQINS, a pipeline able to extract transposon insertion profiles from sequencing data. FASTQINS considers available experimental and design conditions, accepting multiple input types to deliver results in a standardized format. We complemented it with ANUBIS, a Python standalone framework that helps to detect and correct factors that can cause deviations in essentiality estimates. ANUBIS combines, in a single tool, state-of-the-art Tn-seq analysis approaches, with new corrections for previously unconsidered factors, and novel models that do not require any previous knowledge on the essentiality of the organism considered.

We have discussed factors that greatly affect essentiality estimates, including TSD, PCR duplicates, GC bias, differential domains, and essentiality estimate models. We conclude that Tn-seq is a highly sensitive protocol that requires additional processing steps (compared to techniques such as DNA-seq and RNA-seq) and controlled supervision to retrieve accurate estimates. In this respect, ANUBIS provides routines and visualizations to guide along the best processing steps to use before predicting essentiality. Additionally, the user experimental design makes necessary specific considerations and correction/processing steps. For example, users can explore profiles at the level of insertion *read counts* (e.g. using *HMM*). If this is the case, we recommend performing minimal passages (≤30 cell divisions in our case) and PCR duplicates, GC content bias and TSD are highly recommended to be considered. When a more general perspective is desired (e.g. in gene essentiality studies), we found that to obtain good estimates a minimum genome transposon coverage of 10% is required, and that repeated regions and limits for NE N- or C-terminal regions should be properly assigned. ANUBIS also provides all the necessary tools to statistically and visually evaluate whether a gene can be removed from an organism or not, thereby aiding in the rational design of genome reductions.

Ultimately, ANUBIS collects functions to fit, predict, report, and visualize the estimation results using different models. It implements previously described estimators based on *Poisson*, *Gumbel*, *Gamma* and *HMM* models allowing the user to run previously described essentiality models. While these models have been proved to be useful in their original references, they present the limitation of depending on training sets, not always accessible for an organism of interest. This motivated us to implement unsupervised models based on mixture models such as *GMM* and *BGMM* that we believe can be useful in organisms with little knowledge about gene essentiality and/or gene function.

Altogether, we envision ANUBIS as a computational and customizable framework that can perform Tn-seq data treatments, benchmark essentiality studies or be integrated into larger analysis pipelines. Essentiality estimation is a complex task where multiple factors have to be taken into consideration and the requirements of the user can be very different. Thus, in ANUBIS all the corrections are optional and it is the user who decides which of them have to be applied, supported by visual and statistical exploration. However, it also includes specific procedures to automate these corrections based on statistical assumptions for those users with little background in essentiality studies. Both tools have been developed integrating available bioinformatic standards as well as general statistical assumptions that makes it possible to apply them to other organisms. This is important as factors presented here could present different impacts depending on the study species.

Nowadays, in the era of Synthetic Biology, a Tn-seq experiment processed by FASTQINS and explored and analyzed using ANUBIS, provides a perfect starting point to define the essential core machinery and elements that can be removed from a model organism in a sensitive and accurate manner. This, coupled together with targeted editing methodologies (e.g. CRISPR/Cas9 system), can represent a step forward in the rational design of genome-reduced organisms and biological chassis that have important biotechnological and/or biomedical applications.

5.6. Data and software availability

- Supplementary documents including supplementary figures and tables can be found in the “Supplementary Data” file available at Nucleic Acids Research online.
- The code and manuals for the two tools presented in this study can be downloaded as standalone applications or as Python packages from github.com/CRG-CNAG/fastqins and github.com/CRG-CNAG/anubis.
- Tn-seq raw data files have been deposited in the ArrayExpress database at EMBL-EBI, under accession number E-MTAB-8918, and are accessible from the following link: http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-8918.

5.7. Acknowledgements

We would like to thank Marc Weber for assistance and fruitful discussion that helped with the development of ANUBIS. We also thank Tony Ferrar for article revision and language editing.

5.8. Author contributions

S.M.V. performed computational and statistical analyses, developed FASTQINS and ANUBIS, interpreted results, created the figures and tables and wrote the manuscript. RB performed sample preparation for Tn-seq, wrote the methodology, and provided valuable discussion around interpreting the Tn-seq results. J.D. developed the first version of FASTQINS, whose principles have been applied in the version presented here. M.L.S. and L.S. provided direct supervision and were involved in the interpretation of results. All authors read and approved the final manuscript

5.9. Funding

ERASynBio 2nd Joint Call for Transnational Research Projects: ‘Building Synthetic Biology Capacity Through Innovative Translational Projects’, with funding from the corresponding ERASynBio National Funding Agencies; European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program [670216] (MYCOCHASSIS); CERCA Programme/Generalitat de Catalunya; Spanish Ministry of Science and Innovation to the EMBL partnership, ‘Centro de Excelencia Severo Ochoa 2013–2017’. Funding for open access charge: ERASynBio 2nd Joint Call for Transnational Research Projects: ‘Building Synthetic Biology Capacity Through Innovative Translational Projects’, with funding from the corresponding ERASynBio National Funding Agencies; European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program [670216] (MYCOCHASSIS); CERCA Programme/Generalitat de Catalunya; Spanish Ministry of Science and Innovation to the EMBL partnership, ‘Centro de Excelencia Severo Ochoa 2013–2017’

Chapter 6. ProTInSeq: Using ultra-deep sequencing to perform protein detection, quantification and functional studies

Miravet-Verde S; Mazzolini R; Segura C; Broto A; Serrano L; Lluch-Senar M. *ProTInSeq: Using ultra-deep sequencing to perform protein detection, quantification and functional studies.* In preparation.

6.1. Abstract

ProTInSeq is a novel "-omics" technique that allows the study of the proteome by DNA ultra-deep sequencing. The technique is based on transposons engineered to have a positive or negative protein selection marker that is only expressed when the transposon is inserted in-frame to a protein-coding gene. We have tested ProTInSeq in the genome-reduced bacterium *Mycoplasma pneumoniae*. We have identified 75% of all annotated expressed proteins of this bacterium, 5 new ORF (>100aa) and 153 small encoded proteins (<100 aa; SEPs). Being a quantitative and structural metric, it allowed a rough estimation of protein abundance, as well as the identification of special protein features like membrane topology (in 41 proteins). Fusion to unstable proteins (11 proteins) not detected by mass spectroscopy results in their stabilization. This and the detection of very smORFs (57 SEPs below 30aa) suggests this bacteria has a significant level of translational noise. ProTInSeq is a novel proteomics technique based on deep sequencing that can be applied to different genomes allowing a quantitative identification of ORFs, SEPs, and unstable proteins as well as studying membrane topology.

6.2. Introduction

A large number of spurious ORFs of shorter lengths could occur randomly within long non-coding RNAs. For this reason, the FANTOM genome annotation consortium initially relied on 100 and 50 amino acid cutoffs to distinguish protein-coding sequences in eukaryotes and prokaryotes, respectively [445–447]. However, analysis of genomes, transcriptomes, and proteomes revealed the existence of hundreds to thousands of translated, yet non-annotated, small open reading frames (smORFs) that encode for small proteins (<100 aa SEPs) with central roles in metabolism, apoptosis, and development [87]. The discovery of these bioactive SEPs emphasizes the functional potential of this unexplored class of biomolecules. There have been several computational and experimental approaches taken to systematically annotate SEPs in the genome [132,448]. The main challenge in computational approaches is to distinguish smORFs from start and stop codons in-frame by chance. Moreover, some smORFs [132], and ORFs in general [449], could use non-ATG start codons, which makes these assignments even more difficult. Nevertheless, several reports have attempted to computationally annotate smORFs [127,132,257,352,450–453]. In mammals, new algorithms have identified approximately 3000 candidate smORFs transcribed indicating that genomes may contain several thousand non-annotated smORFs [451]. Also, analysis of microbiome genomes reveals more than 4,000 conserved SEPs families, 30% predicted to be secreted or transmembrane. However, over 90% of them have no function associated and half are not represented in reference genomes [121]. Previous studies in the genome-reduced bacterium, *M. pneumoniae* have shown that 53% of its annotated smORFs are essential, while 11% affect fitness, indicating that SEPs play fundamental roles for the cell [234]. Recently, by combining a novel bioinformatics tool (RanSEPs) with “-omics” approaches, we were able to describe 109 bacterial small ORFomes (accuracy of the 94% calculated from 570 validated SEPs in 12 bacterial species). Strikingly, when running this tool along 109 bacterial species where SEPs represented 10% ± 5% of the annotations in the reference, we found that up to 16% ± 9% of proteins in an organism could be classified as SEPs based on species-specific protein features (Chapter 4).

As SEPs seem to be more frequent in genomes than previously thought, during recent years multiple high-throughput methodologies have tried to approach their detection. Techniques based on RNA-Seq like Ribo-Seq, which sequences fragments of transcripts bound by ribosomes, have been reported to be effective to detect short genes [223,450]. However, in bacteria, transcriptional units are polycistronic and a vast number of smORFs are spread in the genome overlapping with longer mRNAs. Consequently, the unambiguous identification of overlapping SEPs is not trivial. Despite this, of 80 smORFs, selected from a pool

of over 2,000 SEPs found downstream to a Ribosome Binding Site and tested by 3'- tagging and immunoblotting in *Escherichia coli*, 45% of them (n=36) resulted in encoded synthesized proteins [89]. Regarding proteomics, the main limitation comes from the complexity of the possible ORFs when small sizes are considered [454]. In eukaryotes, proteogenomics, the creation of proteomic databases from RNA-Seq data, is complicated because of the presence of splicing variants which increase the complexity of the database [455]. In addition, this technique presents limitations in the detection of SEPs where the number of unique peptides that will behave well in the machine is very small. We showed that at least two unique tryptic peptides (UTPs) are required to identify with confidence a protein by Mass Spectrometry (MS) and many SEPs have one or zero UTPs [448]. Thus, new experimental approaches should be developed to validate predicted SEPs in a high-throughput manner.

High-throughput transposon insertion tracking by ultra-sequencing (Tn-seq), the incarnation of transposon-based genomic analyses, enables genome-wide studies in a varied range of bacterial species, under a multitude of conditions, with unprecedented depth [229,234,456]. First, transposon mutagenesis is used to create a library in which ideally each mutant has a transposon inserted in one genomic locus. If the density of transposon insertions is high the disruptible (non-essential, NE) regions and the non-dispensable elements (essential, E) of all the genome can be identified. This library can then be grown under selective conditions, the prevalence of each mutant should be proportional to its fitness in each growth condition. Thus, attenuated mutants under a specific condition are outcompeted whilst mutants with increased growth and survival become overrepresented in the population. High-throughput sequencing is used to identify and quantify all transposon insertion sites in the population and recently we have developed a series of tools to accurately retrieve these profiles. Studies of transcription levels, protein location, and screenings of protein activity have been done by engineering the transposon vectors. Promoterless reporter vectors were engineered to characterize *in vivo* promoters activity in *M. genitalium* and *M. pneumoniae* [234,457]. Also, by random transposon-based GFP insertion, small libraries of full-length fluorescent fusion proteins were expressed at endogenous levels and were used for screenings of protein activity and protein location [458,459].

Here, we engineered a mini-transposon vector to identify experimentally all the ORFs of the genome-reduced bacterium *M. pneumoniae* genome being translated, including small ORFs encoding for SEPs. This human lung pathogen causes atypical pneumonia and it has been considered for more than a decade a model for Synthetic and Systems Biology [276]. This bacterium is a good model for SEPs identification as it has a reduced genome (n=816,394) so the sequence space of SEPs candidates is less imposing than in other bacterial species. In

addition, *M. pneumoniae* has been exhaustively explored by Tn-Seq technique reaching an average of 1 insertion every ~3 bp (Chapter 5) and the RanSEPs program suggested that the smORFome of this minimal cell could encode for 144 SEPs. Out of the 27 SEPs annotated in this bacterium, 21 can be detected by MS, but only 7 present at least 2 UTPs (Chapter 4).

In this study, we engineered the transposon Tn4001 inverted repeats (IR) to remove stop codons in a particular reading frame and fused the translated sequence to the following proteins: chloramphenicol acetyltransferase (*cat)*, erythromycin esterase *(ereA)*, and the RNAse *barnase (Barn)* [460–462], with no translation initiation codons at their N-terminus. The absence of a promoter and the fusion with the IR results in reporter expression only in the case it is transposed in-frame to a genome protein-coding sequence (in phase 0 of the ORF, Figure 6.1). The antibiotic resistance provides a positive selection marker while barnase is used as a negative selection marker. In the case of mutant libraries based on antibiotic resistance, cells were grown under different antibiotic concentrations thus selecting for expression levels of the fused protein to the resistance. Using this technique we could identify 518 annotated proteins (75.2% of the annotated *M. pneumoniae* proteome), including 18 annotated SEPs out of 27 described in this bacterium. A total of 158 new proteins were discovered of which 153 encoded for SEPs. A study of insertion *coverages* by ProTInSeq in relation to protein abundances obtained by MS revealed that this new technique makes possible both the identification and quantification of proteins. Another interesting feature of this technique is that it allows distinguishing the cytoplasmic and external regions of non-essential transmembrane proteins. Finally, our results also support that *M. pneumoniae* could present high levels of translational noise.

At a time when sequencing *coverage* is increasing and the cost of this technique is decreasing, ProTInSeq represents a new technology to study the proteome, including overlapping ORFs, at a resolution and cost not achieved by other "-omics" techniques.

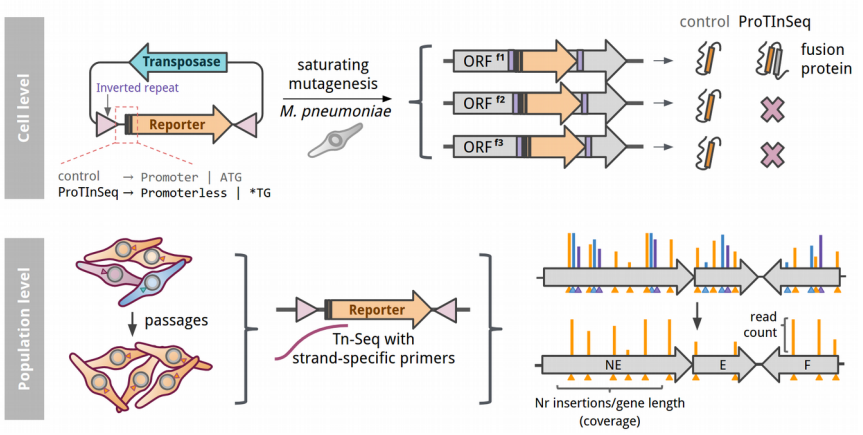


Figure 6.1. ProTInSeq rationale.
Cell level (top row), a schema of the vector used in this work. It is derived from the Tn4001 mini-transposon where the transposase (blue arrow) is expressed from the plasmid to randomly transpose the region flanked by the Inverted Repeats (IR, pink triangles). In a conventional Tn-Seq protocol, the reporter (e.g. antibiotic resistance; orange arrow) is expressed independently. In ProTInSeq, the reporter initiation codon and IR are mutated so the reporter is expressed only when it is inserted in-frame. The IR contains 2 stop codons in 2 out of the three possible open reading frames as a strategy to ensure that only one frame could produce the fusion protein. ***Population level (bottom row)***, Multiple individual transposition events occur in the population (orange cells have the insertion in-frame while blue and purple in frames 2 and 3, respectively). After growing in the presence of an antibiotic only the cells expressing the resistance reporter to a certain level will be viable. After sequencing and mapping insertions in these populations, profiles representing essentiality and protein abundance can be obtained. A similar approach can be done with a negative selection marker (barnase gene in the current study) and in this case, we should select frames 2 and 3, instead of 1.

6.3. Results

6.3.1. Mini-transposon engineering to obtain the ProTInSeq library

We have used two antibiotic resistance positive selection markers Cat, and Ery, and an RNase negative selection marker (barnase; Barn). It is important to mention that Barn is a very strong negative marker and it has been reported that just a few copies per cell are lethal [461,462]. In the Tn400 IRs sequences, there are three stop codons in the three putative open reading frames (ORFs) and a -10 Pribnow box. Different vectors removing one of the three stop codons of the inverted repeat (IR*), or the stop codon and a Pribnow box of the IR (IR**) in combination with mutated selection markers (Cat*; Ery*; Bar*; without promoter and start codon) in-frame, or with a promoter upstream of the start codon of the selection marker (Cat; Ery; Bar; positive controls for transformation and coverage) have been obtained (Figure 6.2). For the Cat and Bar libraries, the P438 promoter was used [460], while for Ery we used the Psyn promoter [463].

We generated different random libraries of Tn4001mini-transposon mutants after transforming *M. pneumoniae* cells with different vectors combining the above elements (TnCat, Figure 6.2A). Mutations in the IR* in the control vector *CmA* decreased the transformation efficiency (TnCatIR*; $1{,}14x10^{-3}$ %) with respect to the conventional mini-transposon vector (TnCatIR; $1{,}4 \times 10^{-2}$ %), while they were similar in the other control vector, *CmC* (TnCatIR**, $0.7x10^{-3}$ %). Transformations with vectors *CmB* (TnCat*IR*) and *CmD* (TnCat*IR**), where the resistance marker could only be expressed if it was in-frame with a translated gene, showed a significant decrease in the number of transformed cells when compared with controls *CmA* and *CmC*, respectively (T-test $P<0.05$ in both conditions; Figure 6.2B). Comparable results were observed in the *EryB* library (T-test $P<0.01$; Figure 6.2C), compared to the control version *EryA* [463]. In the *BarnB* library, as expected if selection by frame was working, we observed the opposite, higher transformation efficiency than in the control *BarnA* (Figure 6.2D; T-test $P<0.01$).

Previously to ultra-deep sequencing, six different individual colonies of the library, 5 from *CmB* and 1 from *CmD*, were picked and the transposon insertion sites were identified by Sanger sequencing (Table 6.1). Insertions were found in 2 NE, 2 F, and 1 E genes and they were in-frame in every case. We found insertions at the N- or C-terminal regions of *mpn165* (E) and *mpn624* (F) encoding for essential ribosomal proteins RplC and RmpB, respectively; thus indicating this technique can be used to study essential and fitness proteins as long as they occur at non-essential ends of the protein. One of the insertions was in a gene annotated as MPNs02 that we already showed that could encode a SEP of 12 aa (Chapter 4 and [234])

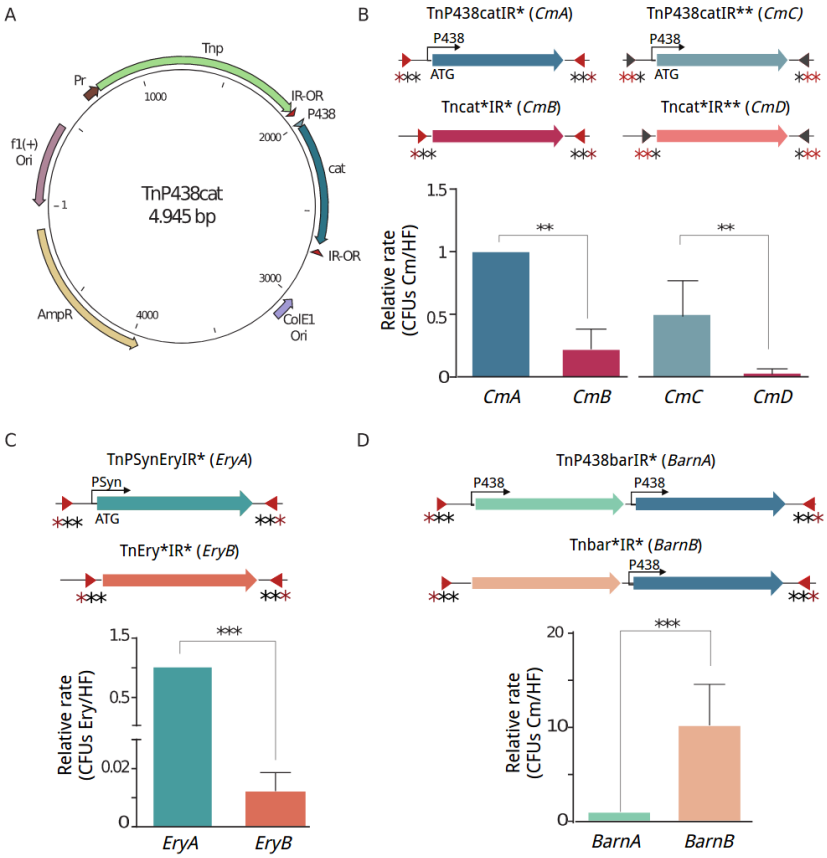


Figure 6.2. Efficiencies of transformation with different vectors.
(**A**) Template vector used in this work. When the transposase (Tnp, green) is expressed, it transposes the region between Inverted Repeats (IR-OR, red). Cat gene (blue) is inserted randomly in the genome. (**B**) Schematic representation of chloramphenicol libraries. The IR presents three asterisks representing the three stop codons in the three putative open reading frames. The asterisk labeled in red represents the A to T mutation (IR*) that leads to the change of stop codon to Leu amino acid in the fusion protein (red triangle; black triangle if not mutated). When the cat gene was mutated (*cat**, blue arrows) the P438 promoter (black arrow) and ATG codon were not present. The histogram shows the ratios of the efficiencies of transformation with different vectors normalized to sample TnP438*cat*IR*. Statistical comparison between pairs by T-test returned significant differences between *CmA* and *CmB* (p-value=0.018), and *CmC* and *CmD* (p-value=0.013 for *CmA*). (**C**) Schematic representation of vectors used to obtain different erythromycin libraries and the histogram with the efficiencies of transformation. In this case, the Psyn promoter was used. As observed in the case of *Cm* libraries, the *EryB* version presents a significant decrease in terms of the relative transformation rate compared to *EryA* (p-value<0.01). (**D**) Schematic representation of vectors to obtain *barnase* libraries and the histogram with the efficiencies of transformation. In this case, as it was a negative selection, *BarnA* version showed low transformation efficiencies as the barnase was expressed, when mutated, the number of recovered transformants also increased (p-value<0.01).

Insertion	ORF	Function	Category
469307	MPNs02	New SEP	NE
546063	MPN447	Hmw1 (adhesion)	NE
751308	MPN624	Ribosomal protein L28 (rmpB)	F
751308	MPN624	Ribosomal protein L28 (rmpB)	F
218775	MPN165	Ribosomal protein L3 (rplC)	E
751237	MPN624	Ribosomal protein L28 (rmpB)	F

Table 6.1. Insertions found by ProTInSeq and validated by Sanger sequencing

6.3.2. Generation of a transposon sequencing library to explore the coding genome of a genome-reduced bacteria

We sequenced 39 samples in total (including biological replicas), which comprise the 3 different selection reporters (*Cat*, *Ery*, and *Bar*), combined with the different IR sequences, the presence or absence of an internal promoter, and ATG sequence, and in the case of the antibiotics, different concentrations in the cell culture (0.5, 1, 2, 5, 10 and 15 µg/ml of chloramphenicol; 0,02 µg/ml for erythromycin). Transposon insertion sites were identified by using the FASTQINS pipeline paired-end and strand-specific mapping mode (Chapter 5), extending the sequence used to unambiguously identify the insert orientation and how many times this event is found by its *read count* (see Material and Methods).

To easily refer to specific samples, we defined an identification code with the conditions applied to each sample as *ReporterTypeConcentration* (e.g. *CmB15* corresponds to the mutated chloramphenicol transposon grown with 15 µg/ml of chloramphenicol). Each sample was explored in terms of *coverage* and *read counts*, distinguishing by types of positions in the *M. pneumoniae* genome: *annotated* for in-frame codon positions, considering annotated genes, *non-coding* for positions with no ORF and *putative* for all in-frame non-annotated ORFs of *M. pneumoniae* (n=29,424, see Material and Methods). In this way, we could easily evaluate the selection at the genome level comparing the different groups.

In terms of general genome coverage, we observed no significant differences for *CmB* and *CmD* at 0.5 and 1 µg/ml libraries (P>0.05) with respect to their references (*CmA* and *CmC*, respectively). When increasing the concentration of chloramphenicol (≥ 2 µg/ml), significantly lower *coverages* were obtained with respect to the *CmA* and *CmC* controls indicating a selection effect of the antibiotic (Figure 6.3A). We observed no significant differences in the controls *CmA, CmC, and EryA* when comparing *annotated*, *putative*, and *non-coding* groups in terms of coverage and read counts, indicating a random and homogenous rate of insertions with comparable values of reads between groups along the genome

(Figure 6.3A-C). Coverage of the genome in the case of control *CmA* increases with antibiotic resistance since cells with no insertion are progressively eliminated. The maximum coverage was recovered at 15 µg/ml (average of 27,9% ± 0.8%, 1 insertion every ~3 bp, which if we exclude essential genes represents 1 insertion every 2 bases).

For the *CmB* samples, we observed a significant increase toward in-frame insertions between *annotated* and *non-coding* positions in coverage and read counts in one of the biological replicates of *CmB2* and for all the biological replicas of *CmB5, CmB10, and CmB15* (Figure 6.3A). For the additional *CmD15* and *EryB1* libraries, we observed similar selection patterns (Figure 6.3B and C). In the case of the *BarnB* library, despite the fact that the general coverage of the samples was limited as expected by negative selection (coverage of 0.5 ± 0.4%, in 3 biological replicas), we observed a significantly reduced number of insertion in-frame for annotated positions respect to non-coding (Figure 6.3D). It is important to point out that this small number suggests the presence of a significant translational noise at a low level since in the best case, we should have expected coverage of around 2/3 of the 25% with *CmA15*.

Regarding the other metric (read counts) the differences between annotated, non-coding and putative were smaller for *Cm*, *Ery*, and *Barn* than those found for coverage, and in some cases we found high read count values for positions not expected to produce a protein fusion, suggesting they could be artifacts or the result of double-inserted cells in the population (one insertion in-frame providing resistance and the second off-frame in a NE region). This last hypothesis was supported by the fact that insertions off-frame with read counts over the median were all located in NE genes or intergenic regions, which are mostly NE. However, an average of 15% of the total signal in off-frame positions (non-coding) was found in E genes, with low read count values, that could come from dead cells (Mycoplasma cells tend to aggregate) or inherent artifacts of the technique.

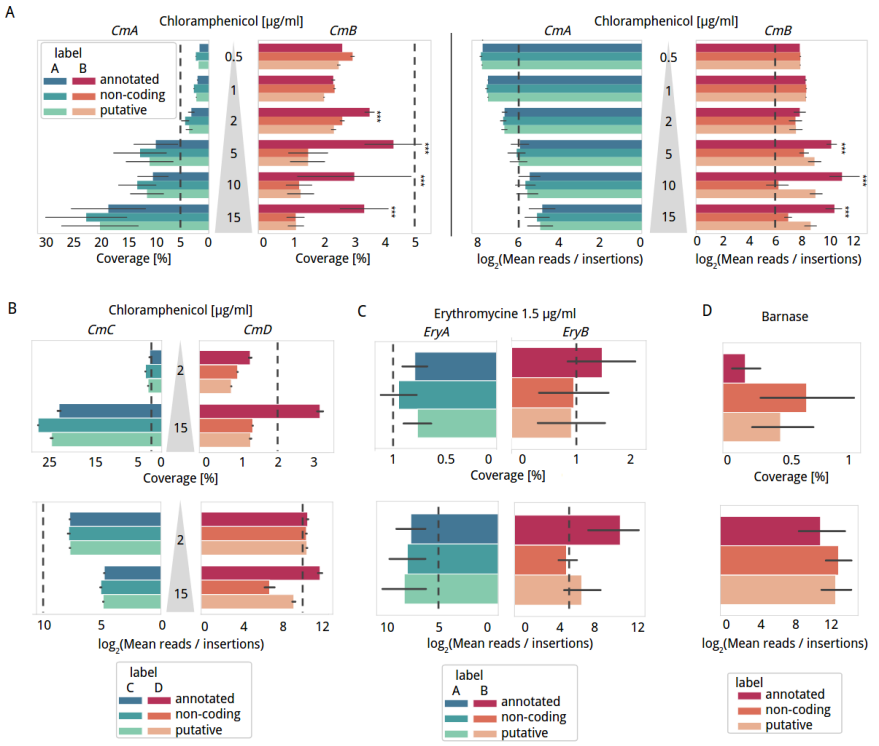


Figure 6.3. Transposon efficiencies of selection at the genome level

A general overview of the different samples in terms of coverage and read counts considering three different insertion positions (legend). Grey dashed lines are used to aid the direct comparison between horizontally related plots. **(A) left** - Cm coverage (X-axis) compared along 6 chloramphenicol concentrations (Y-axis) between control (CmA, blue to green) and mutated version (CmB, red to orange). It can be noticed that while insertions are maintained in the three positions types in *CmA* (significant decrease of *annotated* in *CmA15*, due to essential genes), the mutated version is inserted preferentially in in-frame positions of genes compared to *non-coding* positions with chloramphenicol concentrations higher than 2 µg/ml (one-tail Mann-Whitney-U p-value < 0.05 in every concentration after it). **(A) right** - same exploration but considering the $\log_2$(mean of reads/total insertions found) in the X-axis. A similar selection pattern can be observed in annotated positions but high read counts ($\log_2$ > 6) are still found in *non-coding positions*. (**B)** The same comparative between *CmC* control (left) and *CmD* (right) in terms of *coverage* (top) and read counts (right). Comparable results to *CmB15* were obtained with significant enrichment of in-frame insertions (one-tail Mann-Whitney-U P<0.001). (**C)** Coverage and read counts for *EryA* control (left) and *EryB* (right), both presented enrichment of *annotated* in-frame positions (P<0.05 in both metrics). (**D)** *BarnB* library, in this case, the selection is negative, if the barnase is expressed in fusion, the cell dies. This explains the significant reduction in coverage for *annotated* in-frame positions and the higher coverage in *non-coding* positions (one-tail Mann-Whitney-U P<0.001).

6.3.3. ProTInSeq selects in-frame insertions at the gene level

We evaluated the coverage of each gene (number of insertions normalized by gene length) associated with each of the 689 protein-coding sequences (CDS) of *M. pneumoniae*, distinguishing them by their essentiality (Essential n_E=299, Fitness n_F=59, Non-essential n_{NE}=331 [234]). The control libraries *CmA5* (three biological replicas), *CmA10* (two biological replicas), and *CmA15* (three biological replicas) showed comparable results to previous Tn-Seq experiments, with coverages following the expected E<F<NE distribution and no differences between the three ORFs in each CDS (Figure 6.4A). When comparing the corresponding *CmA* samples to *CmB5*, *CmB10*, and *CmB15*, we observed a significant reduction in coverage in off-frame positions of annotated ORFs compared in *CmB5*, *CmB10*, and *CmB15* samples (phases 1 or 2 of the ORF, corresponding to codon positions 2 and 3; P<0.05 in every compared condition evaluated by one-tail Mann-Whitney-U). Only when considering the in-frame position of the codon, coverage remains comparable to a regular Tn-Seq protocol (Figure 6.4B). Both *CmA* and *CmB* libraries were consistent with previous studies showing preferential insertions in the 5% of each N- and C- terminal gene regions of F and E encoded proteins (Chapter 5). Interestingly, we observed an enrichment in the rate of insertions found in the C-terminus for the *CmB* libraries when looking at E and F genes indicating that the *CmB* selection marker prefers to be fused at the end of a protein (P<0.05 in every compared condition evaluated by one-tail Mann-Whitney-U, Figure 6.4B).

For *EryB1* (two replicas, we observed the same kind of enrichment but in this case, the N-termini of NE genes were the preferent sites of insertion while in the C-terminus were rarely found (one-tailed Mann-Whitney-U P=0.001; Figure 6.5). When exploring the *BarnB* samples (three replicas), a specular image with no insertions in-frame indicates the negative selection is produced. Most of the insertions for *BarnB* were observed in F and NE genes in frames 2 and 3 where Barnase should not be translated so the fusion is not expected to happen, thus the cell can maintain the insertion (Figure 6.5).

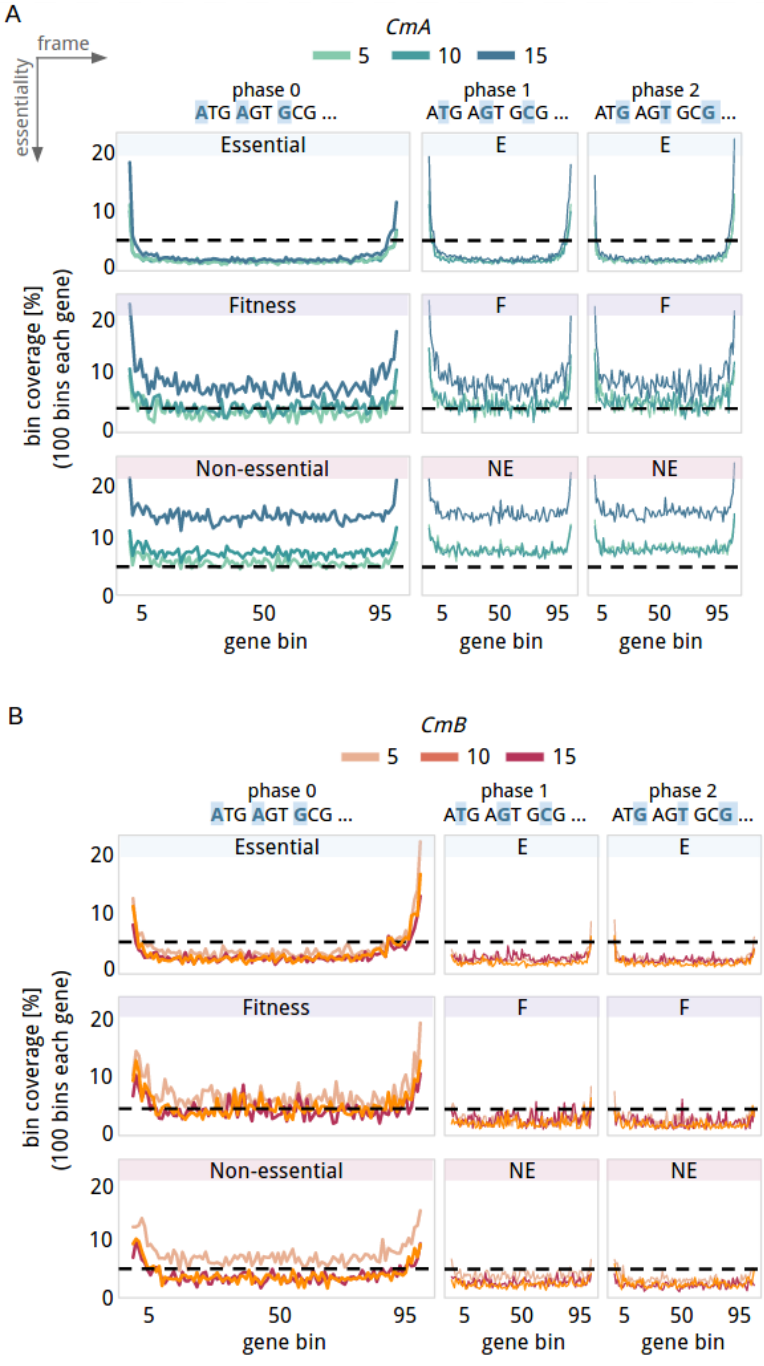

A
CmA
5
10
15
frame
essentiality
phase 0
ATG AGT GCG ...
phase 1
ATG AGT GCG ...
phase 2
ATG AGT GCG ...
Essential
E
E
Fitness
F
F
Non-essential
NE
NE
bin coverage [%]
(100 bins each gene)
20
10
0
5
50
95
gene bin
B
CmB
5
10
15
phase 0
ATG AGT GCG ...
phase 1
ATG AGT GCG ...
phase 2
ATG AGT GCG ...
Essential
E
E
Fitness
F
F
Non-essential
NE
NE
bin coverage [%]
(100 bins each gene)
20
10
0
5
50
95
gene bin

Figure 6.4. Metagene comparison between CmA and CmB libraries
Both panels represent the coverage (Y-axis) calculated for genes in *M. pneumoniae*, binning them in 100 non-overlapping regions with the same size within the same gene (X-axis, from N-terminus to C-terminus). We separated by 2 different variables: in each row, genes are grouped by their known essentiality category (in order: Essential, Fitness and Non-essential); in each column, the three possible frames of the genes are represented (in order: phase 0 (in-frame), phase 1 (position 2 of the codon) and phase 2 (position 3). To help the comparison, the grey dashed line is fixed to 5% coverage. (**A)** *CmA* library coverage along with genes, measured at concentrations 5, 10, 15 µg/ml of chloramphenicol (from light green to dark blue, *CmA5* and *Cm10* are overlapped). It can be observed that no differences are observed between frames and the order of E<F<NE in coverage is respected. (**B)** The same visualization in *CmB* library (from light orange to red for 5, 10, 15 µg/ml of chloramphenicol, respectively). It can be observed that phases 1 and 2 do not accumulate insertions while phase 0 resembles the coverages observed in the control. It can be noticed that there are preferential insertions at the C-terminus of E and F genes.

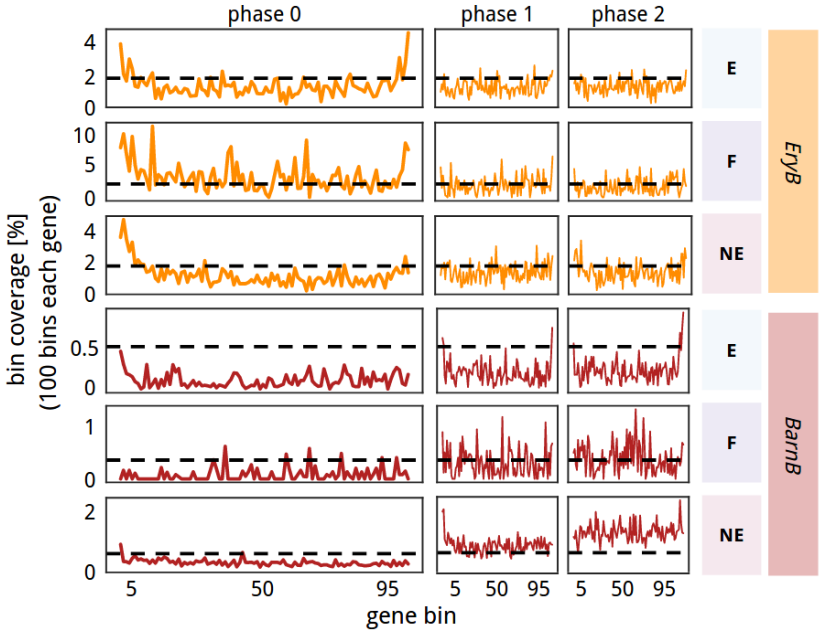


Figure 6.5. Metagene comparative for EryB and BarnB libraries
Coverage (Y-axis) calculated for genes in *M. pneumoniae*, binning them in 100 non-overlapping regions with the same size within the same gene (X-axis, from N-terminus to C-terminus). Library *EryB* (orange) presents the selection profiles observed. In this case F genes (second row) presented significantly higher coverages than NE (third row; one-tail Mann-Whitney-U p-value < 0.05). The grey-dashed line is set in coverage=2%. For the *barnase* library, the specular image can be observed despite the low coverage obtained (the grey-dashed line is at coverage= 0.5%). In this case, insertions in-frame are rarely found in the population while phases 1 and 2 accumulate insertions depending on their essentiality.

6.3.4. Identification of proteins with ProTInSeq

We evaluated if using the transposon insertion frame preferences we could identify coding sequences with special attention in the detection of SEPs. We analyzed the *CmB5*, *CmB10*, *CmB15*, *CmD15* and *EryB1* samples. We also included one additional replica for *CmB5*, *CmB10*, and *CmB15* where we did 3 additional cell passages to remove any possible dead cell that could produce the signal observed in off-frame insertions in the previous section. As a negative control, we used a list of strand-specific *non-coding* annotations derived from intergenic regions with low expression profile (log_2(CPM)<2; negative set, n=1700).

We defined a method based on gene coverages where we evaluated if a gene presents an enrichment in the rate of insertion for in-frame positions with respect to what could be expected by chance in each sample. This method applies the same methodology previously used to assign essentiality classes assuming that insertion rate in a gene will follow a Poisson process conditioned by the gene length [464]. Applying this approach we accounted for a total of 7,388 ORFs with significant signal (24.5% of the possible ORFome). On average, 535 of the 689 known CDS in *M. pneumoniae* could be retrieved in this analysis (77.62% +9.5%), with the best sample identifying 598 (86.8% of the known proteome; *CmB5*), including 21 out of its 27 annotated SEPs (77.7%; Table 6.2). In a pool of 116 MS searches, only 561 proteins can be detected in this bacterium with at least 1 UTP, with 22 of them being annotated as SEPs, with the missing ones having few UTPs, or represent fragmented genes (pseudogenes) or duplicated proteins.

By using Receiver Operating Characteristic (ROC) curves in each sample, we obtained an average positive recall (i.e. percentage of annotated proteins retrieved) of 77.6% ± 8.9% and a negative recall of 0.65% ± 0.37%, which corresponded to intergenic sequences in the negative control being detected with a signal similar to genes (Table 6.2). When performing three additional passages, the negative recall decreases from 0.65% to 0.43% but also the number of identified proteins is reduced (from 60.4 to 34.8%; Table 6.2). The reduction in identified proteins is expected since by serial passages fitness genes become essential and some non-essential genes become fitness [234]. The putative false positives derived from short non-coding sequences (<90 bp) presenting up to 5 insertions in some samples which results in a high coverage. As filtering them out by coverage value would exclude a wide range of E genes from the study (they only have insertions at the N- and/or C-termini), we defined an additional filtering step to ensure we were retrieving a high-confidence set of identified proteins. For this, we set two conditions: a minimum threshold in the number of insertions required to ensure that negative control sequences are discarded; and to present a significant signal (P<0.05) in at least two biological replicas or in two

different selection conditions. Under these stringent criteria, we did not find false positives and reported a total of 518 CDS in *M. pneumoniae* (75.2% of the proteome, the average per sample of 60.4%), including 18 annotated SEPs (66%, n=27).

In total, 447 annotated proteins and 20 SEPs were found in the intersection of previously identified proteins by mass spectroscopy and the ones identified here, while 114 proteins and 2 SEPs (both are ribosomal proteins <50 aa, essential in *CmA*) were exclusively detected by MS. A total of 70 ORFs (69 NE and 1 F) were found with our approach but not by MS. Of these, 30 assigned hypothetical and the rest presenting diverse annotated functions in at least 2 closely related Mycoplasma species. Interestingly this group contained 11 NE proteins which could only be detected by MS when the Lon protease is knocked out, thus unstable and fastly degraded in the cell under normal conditions [465]. Considering that these proteins presented a coverage enrichment at the end of C'-termini regions with respect to the control (one-tail paired Mann-Whitney-U; $P<0.001$ in both comparisons for *CmB5*), it is possible that the fusion of the reporter stabilizes and protects them against Lon targeting as it has been described signal for degradation is located in the C'-termini of some of these genes (e.g. FtsZ and FtsA [74]). Finally, a total of 58 CDS were detected neither by MS nor in this approach, mostly adhesins (n=18) and hypothetical proteins (n=17) containing repeated sequences, a factor that limits both MS and Tn-Seq approaches as less number of unique peptides and unique reads can be detected, respectively.

		Rep	ROC curve			P < 0,05 (Coverage Poisson)					After filtering <thr of insertion + 2 independent samples						Min aa length
						Total			Recall			Total			Recall			
			TPR	FPR	AUC	Ann	New	Neg	Ann (n=689)	Neg (n=1700)	Thr	Ann	SEPs	New	Ann (n=689)	Ann SEPs (n=27)	
CmB	5	1	0,9	0,02	0,97	585	928	15	84,91	0,88	4	500	12	138	72,57	44,44	12
		2	0,92	0,01	0,99	598	825	6	86,79	0,35	5	518	13	132	75,18	48,15	9
		3	0,86	0,04	0,93	498	852	12	72,28	0,71	2	415	12	152	60,23	44,44	15
		4	0,88	0,04	0,93	453	678	8	65,75	0,47	2	331	4	118	48,04	14,81	25
	10	1	0,91	0,02	0,98	587	973	7	85,2	0,41	4	506	13	146	73,44	48,15	10
		3	0,9	0,07	0,94	513	1074	12	74,46	0,71	3	332	7	55	48,19	25,93	15
		4	0,81	0,04	0,9	388	490	10	56,31	0,59	3	176	2	27	25,54	7,41	32
	15	1	0,9	0,01	0,97	553	825	9	80,26	0,53	4	426	18	123	61,83	66,67	9
		2	0,92	0,06	0,95	520	876	12	75,47	0,71	3	327	10	92	47,46	37,04	13
		3	0,96	0,08	0,97	595	1586	30	86,36	1,76	3	465	13	122	67,49	48,15	9
		4	0,73	0,03	0,86	320	412	4	46,44	0,24	2	213	5	77	30,91	18,52	25
CmD	15	1	0,85	0,01	0,96	515	387	9	74,75	0,53	4	370	12	39	53,7	44,44	15
EryB	0,02	1	0,9	0,01	0,97	562	721	2	81,57	0,12	3	501	12	121	72,71	44,44	12
		2	0,75	0,04	0,87	357	416	8	51,81	0,47	2	218	5	81	31,64	18,52	33
Mean (rep 1-3)			**0,89**	**0,03**	**0,95**	**535**	**860**	**11**	**77,62**	**0,65**	**3**	**416**	**12**	**109**	**60,40**	**42,76**	**14**
Mean (rep 4)			**0,81**	**0,04**	**0,90**	**387**	**527**	**7**	**56,17**	**0,43**	**2**	**240**	**4**	**74**	**34,83**	**13,58**	**27**
Total unique						**598**	**6790**	**67**	**86,79**	**3,94**		**518**	**18**	**158**	**75,18**	**66,67**	**9**

Table 6.2. Results of the identification of ORFs
For different libraries, *CmB*, *CmD*, and *EryB*, including different concentrations, a ROC curve study is performed retrieving the True Positive Rate (TPR), False Positive Rate (FPR), and Area Under the Curve (AUC). The counts of ORFs estimated with this condition are expressed differentiating between Ann (annotated CDS in M. pneumoniae), New (putative ORFs), and Neg (negative control sequences). The recall (percentage of candidates in each group retrieved) is also included. The ROC values are used to define a sample-specific threshold (Thr), required to filter out negative control sequences and keep only those candidates that are significant with no negative control candidates. The following columns include the number of estimated proteins (also separating known SEPs in *M. pneumoniae*). The last column includes the length in aa of the shortest ORF identified. The last two rows include the mean values separating the replicas number 4 (the ones with extra passages), and the total unique ORFs identified in each category.

6.3.5. Exploration of smORFs identified as SEPs

Using the same criteria defined above, we identified 158 non-annotated ORFs (Table 6.2). This list included 5 ORFs (>300 nucleotides) with 3 of them being detected by MS, expected to encode for proteins of 104 aa (*mpneu10249*, 3 UTPs; note we use the *mpneu* prefix from the ORF database), 193 aa (*mpneu25274*; 5 UTPs, GTG start codon; predicted dihydroxyacetone kinase subunit L by BLASTP), and 252 aa (*mpneu06085*; 8 UTPs, also the largest ORF in this group; a predicted lipoprotein).

When considering the new 153 smORFs, we observed a normal distribution of sizes between 9 to 95 aa (40 ± 20 aa, median = 38). Out of this group, 32 of them were already reported in the list of 144 smORFs predicted to have SEP coding features (Chapter 4). Within this 32, 3 out of 6 potential SEPs were already validated by C_{13} labeled peptides: *mpneu00732* (MPN155a; 90 aa; 2 UTPs; YlxR, RNA binding protein). Another two candidates found out of this 32 were *mpneu14551* (MPN655b; 82 aa), and *mpneu14957* (MPN672a; 57 aa), both detected with 1 UTP by MS and predicted as SEPs with uncharacterized function.

We explored these 153 candidates in terms of conservation by BLASTP as done in our previous study, comparing the percentage of identity and alignment length (i.e. query and hit must share similar protein sizes), with the translated ORFomes from 109 bacterial species (Chapter 4). In total, 17 potential SEPs were reported as hypothetical in other bacterial species and 4 were reported with a predicted function: *mpneu12044* (type I restriction endonuclease-like protein; 69 aa), *mpneu00732* (K-like RNA binding protein; 90 aa), *mpneu24822* (thymidine kinase-like; 44 aa), and *mpneu14402* (ATP-binding protein-like protein; 33 aa). As an example, *mpneu14402* is located in between a NE region comprising two proteins not detected by MS: MPN634 (hypothetical) and MPN635 (hypothetical, potential pseudogene). This candidate, which was also predicted by RanSEPs, presented a significant signal in every *CmB* and *Ery* samples, and its AUG codon is located right downstream of a transcription start site (Figure 6.6A). The

genomic and transcriptomic context of these predicted SEPs was explored. We observed that within the list of 153 smORFs only 7 of them presented low RNA expression profiles in the range of what we considered as a negative control (log_2(reads)<2), while the remaining 146 showed an average expression of 8.68 ± 2.47 log_2 (reads).

High expression levels could be derived from the overlapping of the smORFs with other expressed genes. To address this, we defined 4 possible contexts: overlapping with an annotated gene, with a functional RNA (rRNA, tRNA, ncRNA), upstream to an annotated gene (10-30 bp between the predicted stop codon of the smORF and the start of the annotated gene), or intergenic (no ORF at 100 bp upstream and downstream). Out of the 153 candidates, 71 (46%) were located in transcribed intergenic regions with an average of 6.6 ± 2.64 log_2(reads) in expression (e.g. *mpneu14402* in Figure 6.6.A). The remaining 82 included 23 smORFs overlapping with annotated ncRNAs in *M. pneumoniae* (15%) [278,285]. Three were overlapping with the gene ncMPN037, including the aforementioned *mpneu12044* (69 aa), *mpneu02279* (14 aa), and *mpneu07215* (10 aa, Figure 6.6B). The smORF *mpneu07215* could be a potential regulatory smORFs (i.e. they regulate the expression of upstream genes by hiding/exposing genetic signals when being translated [90]). The function as regulatory smORF could be expected by location to 10 additional cases (6%). Finally, 38 smORFs were overlapping with annotated genes (24%), 9 did with an annotated ORF and an ncRNA at the same time (5%), and 2 candidates were found overlapping and annotated gene and being a potential regulatory smORF of upstream genes (1%). Two interesting cases of overlapping smORFs can be found exploring *mpn121*, a conserved hypothetical NE protein which is found in high abundance by MS (587 copies/cell), which has two overlapping significant smORFs: *mpneu00858* (10 aa) in the inner region, and *mpneu00861* (19 aa) in its C-terminus region. Also, the E gene *mpn120* partially overlaps with *mpn121* N'-terminus and it still has signals of translation (Figure 6.6C).

We also evaluated the presence of ribosome binding site (RBS) motifs that could be regulating the translation of these proteins. In total, we found 34 smORFs (22%) presenting a Shine-Dalgarno-like sequence, this is interesting as the ratio of inclusion of RBS in *M. pneumoniae* for annotated genes seems to be in line with this value (26.5%, n=689). Out of these 34 smORFs with RBS motifs, 5 were overlapping ncRNAs, 12 were overlapping ORFs, and 17 were intergenic smORFs.

Finally, a total of 39 candidates presented transmembrane predicted segments by TMHMM [386], 21 of them predicted to pass the cell membrane once while the remaining 18 were predicted as membrane-associated proteins with a transmembrane segment but not being exposed to the medium. Finally, 16 out of the 39 were predicted to present signal peptide features.

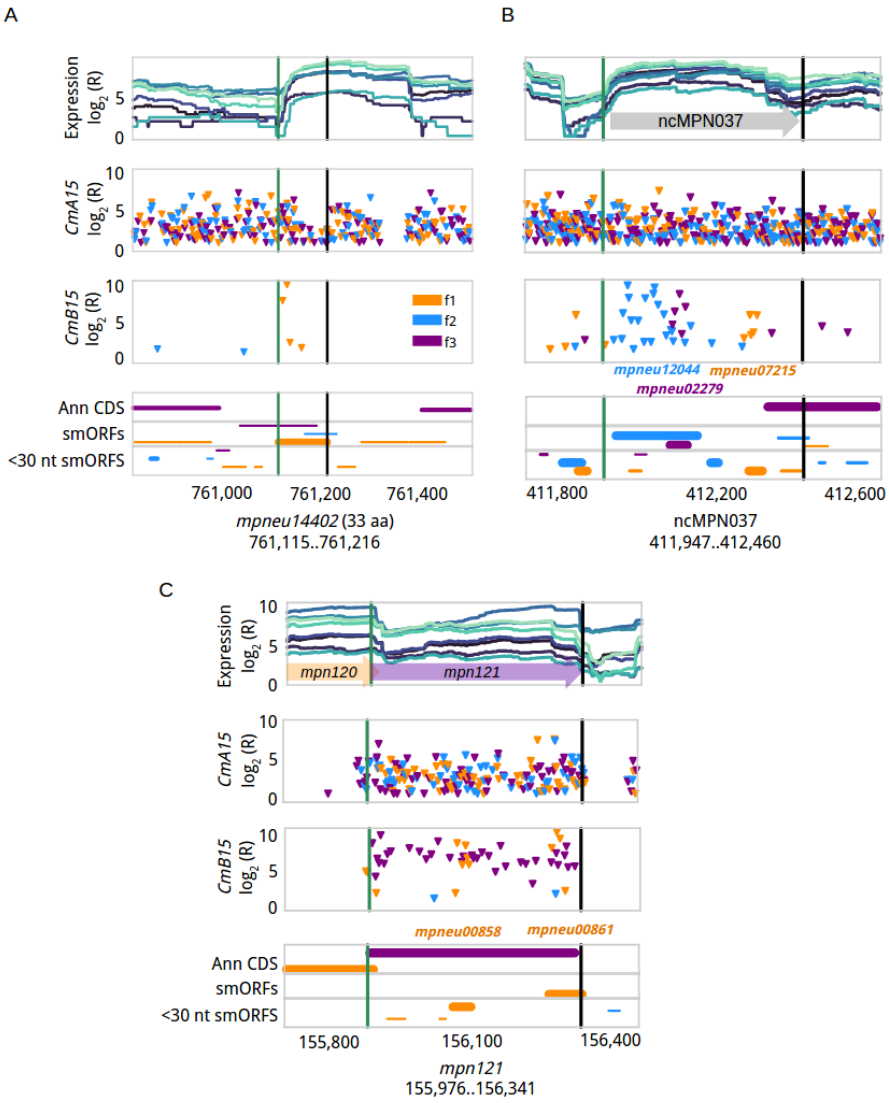

A
B
Expression log2 (R)
CmA15 log2 (R)
CmB15 log2 (R)
f1
f2
f3
ncMPN037
mpneu12044
mpneu07215
mpneu02279
Ann CDS
smORFs
<30 nt smORFS
761,000
761,200
761,400
411,800
412,200
412,600
mpneu14402 (33 aa)
761,115..761,216
ncMPN037
411,947..412,460
C
mpn120
mpn121
mpneu00858
mpneu00861
155,800
156,100
156,400
mpn121
155,976..156,341

Figure 6.6. Examples of profiles of smORFs detected with this approach
Profiles, first - RNA sequencing profile of the region measured as log_2(reads), **second** and **third** insertion profiles obtained by Tn-Seq from *CmA* and *CmB* are represented as log_2(reads) (Y-axis) along the genome (X-axis). Each inverted triangle represents an insertion found in at least two replicates, colors represent the 3 possible frames of the whole region: frame 1, 2, and 3 with colors orange, blue, and purple, respectively. The **bottom plot** shows the ORFs found, with the same frame color code as in the upper plots, distinguishing by annotated ('Ann CDS'), smORFs between 30-300 bp, and very small ORFs (3-30 bp). If an ORF is significant the size of the line is bigger. (**A) *mpneu14402*** (orange in smORFs tracks, start and stop delimited by green and black horizontal lines). This is an intergenic smORF, between *mpn634* (left purple) and *mpn635* (right purple), predicted as SEP with an ATP-binding domain by BLAST, scored positive by RanSEPs, and by this methodology. (**B)** profile of ncRNA ncMPN037 (grey arrow and delimiting horizontal lines) which has 3 overlapping smORF which are significantly enriched: *mpneu12044* (69 aa), *mpneu02279* (14 aa) and *mpneu07215* (10 aa), frame colors are maintained respect (A), we label the predicted new annotations with the same frame colors. (**C)** Example of translation signal overlapping with *mpn121* (delimited by horizontal lines and purple arrow on top plot), which has insertions in its N-terminus derived from *mpn120*, and two smORFs with translation signal (both in orange): *mpneu00858* (10 aa) and *mpneu00861* (19 aa).

6.3.6. Essentiality and protein abundances

Independently of the capacity of detection of this approach, we can still analyze the profiles as Tn-Seq conventional samples, evaluating the coverage variations depending on other factors. We explored the different biological factors that could play a role in the detection by Principal Component Analysis (PCA) of coverage in annotated genes and a series of biological features including localization, RNA expression, protein abundance, membrane topology, essentiality (measured in non-mutated versions), and conservation. We identified that up to 85% of the variability could be explained by the two first components (54% and 31%, respectively), mainly conformed by protein abundance, and essentiality in the first component, and the presence of transmembrane segments in the second. When exploring the essentiality of the genes detected as significant, as expected due to the nature of this methodology, E genes were the most missed (Figure 6.7A). When considering the set of 561 detected by MS and with protein abundance information available, we observed that the group of non-detected known proteins presented significant lower protein abundances with respect to the detected group in F and NE categories (one-tailed T-test $P<0.001$ in both comparisons, Figure 6.7B). This indicates that insertions in low expressed genes are not selected because the reporter, in frame, does not reach the required levels to provide resistance.

We evaluated this effect by exploring the relation of coverage in-frame, normalized by essentiality, with protein abundances (explored in 5 bins, Figure 6.7C). This revealed that while gene coverage in *CmA and EryA* samples remained comparable along with the abundance groups, *CmB* and *EryB* gene coverage increases with the protein abundance indicating that in addition to essentiality, these libraries selection is also dependent on the levels of protein expression. In a conventional Tn-Seq experiment, we would expect to see higher coverages for NE with respect to F proteins. However, as F proteins are expressed on average at higher levels than NE proteins (average copies/cell for F=304,9 and NE=212,3; one-tail Mann-Whitney-U P=0.003), this gene category present comparable coverages to NE genes in the *CmB10* and *CmB15* samples (Figure 6.4B).

In agreement with this, we observed in the *CmB5* samples higher *coverage* values for proteins with lower abundances compared to *CmB10* and *CmB15* samples (Figure 6.7C). In the *BarnB* library, E and F genes presented very residual *coverages*. For NE genes, those with higher abundances presented lower *coverages* compared to NE proteins with low expression (one-tail Mann-Whitney-U P=0.03), indicating that barnase can be inserted in genes with very little expression (Figure 6.7C). An example of the relation between essentiality and abundance can be observed in the genome region covering genes *mpn447* to *mpn452* which presented the highest in-frame *coverage* values in *CmB15* samples (Figure 6.8). Also, we already showed the example of MPN634 and MPN635, which are not found by MS and they are clearly NE, but we do not recover insertions in *CmB15* (Figure 6.6A).

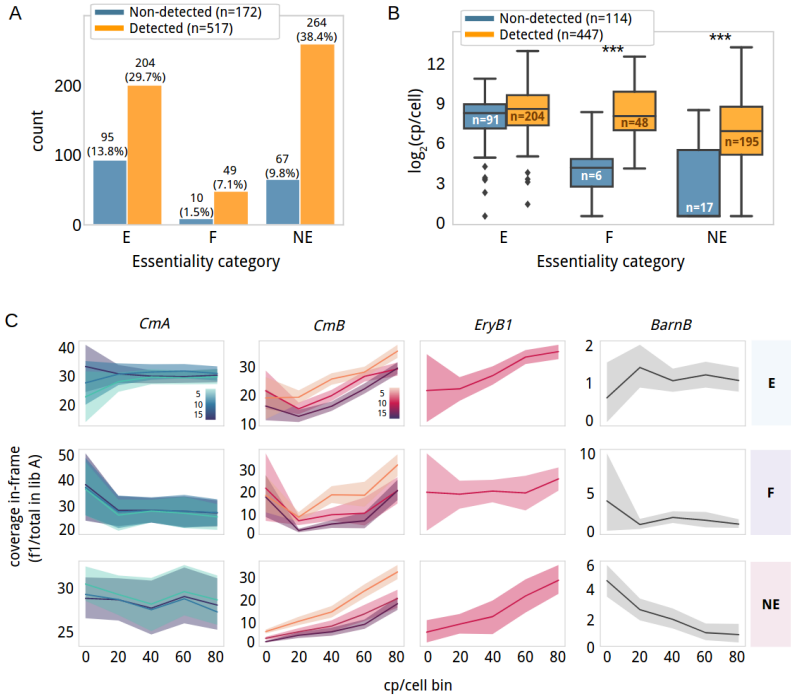


Figure 6.7. Essentiality and protein levels in relation to ProTInSeq

(**A**) Barplot counting the number of annotated proteins detected as significant (orange) and non-detected (blue) by essentiality categories in *M. pneumoniae*. Total counts and percentages over the total number of annotated genes are expressed on top of each bar. It can be noticed that most of the genes with no significant signal belong to the E category. (**B**) Boxplot comparing the protein abundances of 561 proteins detected by MS with those detected by our method. Non-detected proteins in the groups of F and NE are mostly low expressed proteins (one-tailed T-test P<0.001 in both comparisons, asterisks represent significance). In the case of E genes, differences are not significant. (**C**) Exploration of abundances (X-axis) in relation to essentiality categories E, F, and NE (rows) along with *CmA*, *CmB*, *eryB*, and *BarnB* libraries (columns). The X-axis represents 5 bins, each of them containing 112 proteins, and delimited by the following thresholds in protein copies per cell 0 (<1.82) , 20 (<50), 40 (<130), 60 (<305), and 80 (to 4033). The Y-axis represents the *coverage* in-frame normalized by the *coverage* measured along the whole gene in the respective antibiotic concentration of the *CmA* control libraries. Lines represent the average *coverage* for the genes in each abundance bin, while the shadow is the 95% confidence interval. If the shadows between two lines do not intersect (e.g. *CmB5* with *CmB10 and CmB15*- NE genes), the differences are significant. As expected in the control (first column) *coverage* in-frame tends to represent 1/3 of the total insertions found in each gene, independently of the essentiality category and the abundance bin. When exploring the *coverage* in-frame for *CmB* and *EryB* (second and third column), we observed that persistence of insertions in-frame is dependent on the abundances for the three essentiality categories: the highest the abundance, the higher the *coverage* found in in-frame positions. This is also shown by negative selection with the *BarnB* library, while E genes presented almost null *coverages* <1%, F and NE accepted *barnase* insertions only when they occur in low abundant proteins.

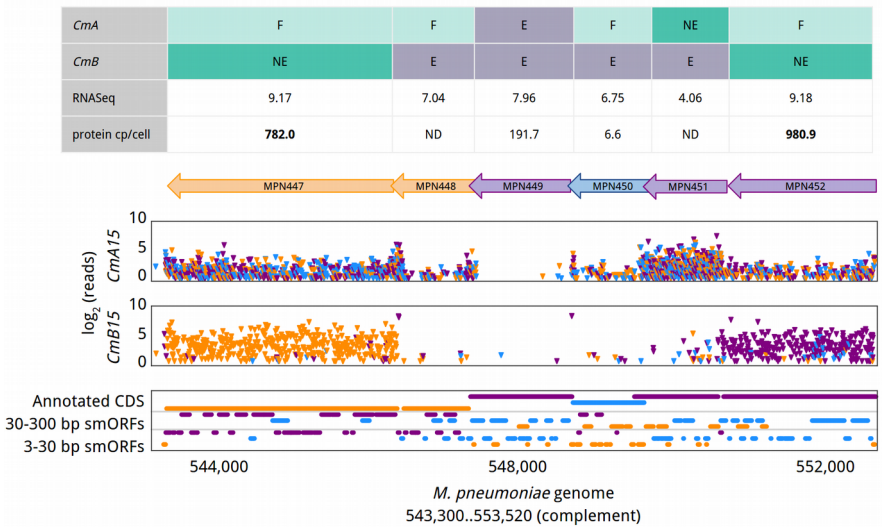


Figure 6.8. Example of the relation between essentiality and protein abundance explored with ProTInSeq

From top to bottom, **Table** with the essentiality category assigned in *CmA15* and *CmB15* conditions (average between read counts in inserted position if inserted in at least two replicas), followed by RNA-Seq measured as $\log_2$(reads/gene length) and protein abundance as copies per cell for 6 genes (arrows): *mpn447* (HMW1, attachment organelle protein), *mpn448*, *mpn449*, *mpn450* (three hypothetical proteins), *mpn451* (ComE, competence protein-like), and *mpn452* (HMW3, attachment organelle protein). The **bottom plot** shows the ORFs found, with the same frame color code as in the upper plots, distinguishing by annotated, smORFs between 30-300 bp, and very small ORFs (3-30 bp). We see that for NE genes in the control CmB with high protein copies per cell we find very clean profiles with in-frame insertions (orange and purple). For genes where we do not detect the protein by MS (MPN448 and MPN451) we do not see a clear preference for in-frame insertions. Finally, *mpn449*, an essential gene encoding for a conserved hypothetical protein, despite being present at 191.7 cps/cell, as it is essential only an insertion with a large number of reads is found in the N'-terminal (purple peak on the start of the gene).

6.3.7. Transmembrane topology explored by insertion coverage

In addition, we observed that the presence of one or more transmembrane segments was also determinant in the differences of coverage observed between control and mutated libraries. For example, *M. pneumoniae* has 41 annotated lipoproteins (all NE with the exception of 6 E), however, they were all estimated as E in every *CmB* sample. Lipoproteins are characterized by being fully exposed to the outside of the cell and anchored by an acyl group covalently attached to an N-terminal Cys residue. They are synthesized with an N-terminal region encompassing a transmembrane helical segment that is cleaved by a peptidase when acylating the Cys residue [466]. Thus, we would only expect insertions under antibiotic selection conditions at the N'-terminus, or in the case of barnase, in the external region (if 100% of the protein goes outside of the cell). While the control *CmA* library presented a homogenous *coverage* along with the NE lipoproteins (n=35), in the *CmB* samples we observed only insertions in the N-terminus (Figure 6.9A). For example, for the NE lipoprotein MPN648, we only found insertions for the first five aa in CmB15 (Figure 6.9B).

The effect was observed for other proteins with transmembrane segments. For example, when exploring 66 NE proteins with at least 2 predicted transmembrane segments by using TMHMM and measuring their *coverage* in-frame in cytoplasmic segments normalized by their total in-frame coverage in *CmB15*, an average of 81% ± 17% of the total insertions in the gene was found in cytoplasmic segments of the gene. Applying a Change Point Detection (CPD) algorithm, previously used to evaluate NE extensions in N' and C'-termini regions and to detect significant deviations in continuous data (Chapter 5), we could detect the transmembrane segments. An example is gene *mpn593* that also presents an overlapping SEP candidate, *mpneu15456* (30 aa) found in *CmB5*, *CmB10*, and *CmB15* (Figure 6.9C).

In some cases like *mpn359* gene, our results contradict the predictions made by TMHMM [34] (60% of in-frame insertions were located in outer-segment coding regions; Figure 6.9D). This could be a consequence of a wrong prediction by the software used. In fact, using the SPLIT server [467], we found a putative fourth transmembrane helix with a weak prediction which if real will indicate that the C-terminal region of the protein is internal as supported by our data. Remarkably, within the new SEPs predicted to have transmembrane segments (n=39), on average in *CmB15* samples, 63 % ± 27 % of the insertions were found in predicted cytoplasmic segments (lower than for known transmembrane NE genes; 81% ± 17%, one-tailed T-test P=0.12).

After running this algorithm in 101 NE known proteins (35 lipoproteins, 66 transmembrane), our results matched the TMHMM predictions, with an error of ± 10 aa, for 41 proteins, failed in one segment in 39 (31 predicted to be cytoplasmic which are exposed in TMHMM predictions; 8 predicted to be cytoplasmic in TMHMM but found clean of insertions), and for 21 we could not predict, 13 due to presenting repeated regions and the other 8 which presented at least three transmembrane segments and their in-frame coverage was considerably reduced (21% ± 8%, one-tailed T-test P=0.12); thus preventing the efficient application of the algorithm.

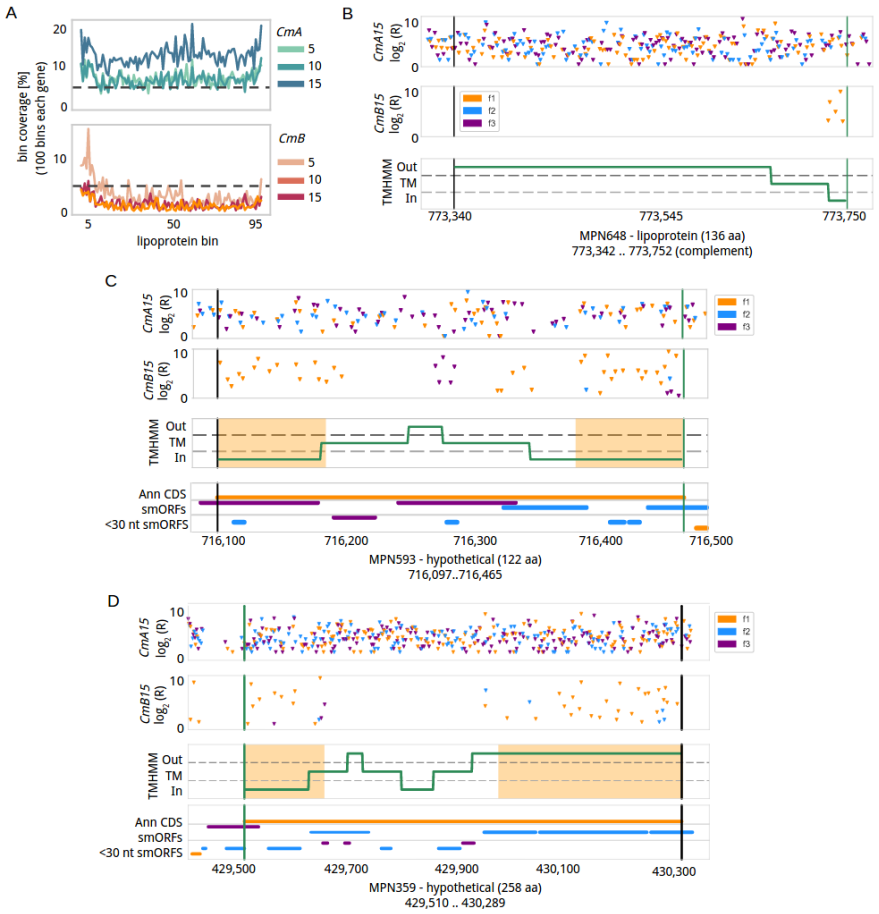


Figure 6.9. Transmembrane topology exploration using ProTInSeq

(A) Metagene representation of in-frame coverage of 35 NE lipoproteins comparing controls *CmA5*, *10* and *15* (top, light green to blue) and mutated versions (bottom, light orange to red). While insertions are homogeneously distributed in the control, insertions in the *CmB* are only recovered in the N'-terminus. Grey dashed lines are set in 5% of *coverage (***B)** ***Top two plots.*** Insertion profile of MPN648 lipoprotein (X-axis represents the position in the chromosome, green and black horizontal lines start and stop codons, respectively). Y-axis represents the $\log_2$ of average *read counts* ('R') of positions found inserted in at least two replicates. The top two plots are for CmA15 and CmB15 samples, respectively. ***Third plot.*** Prediction of transmembrane segments with TMHMM. If a predicted segment is located in the inner part (In, cytoplasmic segment), the line is below the grey-dashed horizontal line; when the amino acids are exposed (Out), the line is above the black-dashed horizontal line. Transmembrane segments (TM) are represented in between. ***Bottom plot.*** CDS and putative ORFs keep the same frame color than previous figures. (**C**) and (**D**) Same representation as in the B panel for the MPN593 and MPN359 proteins. In this case While in the case of MPN593 we see a perfect agreement between the predicted In regions and the in-frame insertions, this is not the case for the predicted second outer segment of MPN359.

6.4. Discussion

In conclusion, we confirmed the selection of in-frame insertions using a Tn-Seq approach with significant results in different conditions: two different versions of the chloramphenicol resistance transposon at 2, 5, 10, and 15 µg/ml antibiotic concentrations, and an erythromycin resistance version at 0.02 µg/ml. We also proved the anti-selection of in-frame insertion using a transposon encoding for barnase. When exploring these libraries at the genome and gene level, it can be observed that insertion occurs preferentially in in-frame positions for the antibiotic libraries and the opposite in the barnase library. Using common approaches applied in the field of genome essentiality, we define a statistical method for the detection of coverage-enriched ORFs that also recovers smORFs previously validated to encode for SEPs. In total and considering samples independently, more than 7,000 translation signals could be recovered in *M. pneumoniae.*

From a conservative perspective and considering the random nature of transposon sequencing protocols, the possibility of sequencing artifacts (base slippage) and of a small number of double transformations (around ~6% with the amount of plasmid DNA used), we define stringent criteria that recover 75.2% of *M. pneumoniae* proteome, including 66% of its annotated SEPs. We also identified 153 non-annotated smORFs, 5 ORFs, and 11 targets of Lon protease, which cannot be detected by MS as they are fastly degraded. Out of the 153 SEPs, 32 were predicted by our computational approach (n=144). Thus we should consider that RanSEPs will favor E or F SEPs considering the importance of homology in its searches, while ProTInSeq will favor NE genes.

Within the group of 158 newly identified proteins, we observe a wide variety of sizes, from a SEP of 9 aa to an ORF coding for a 252 aa protein. Within the 153 SEPs, we found intergenic SEPs, smORFs overlapping with other genes, and potential upstream regulatory smORFs, not described in *M. pneumoniae* but found in several bacterial species and predominant in eukaryotes [90]. We found 23 SEP candidates within annotated ncRNAs in this bacterium, with some complex translation contexts such as 3 proteins translated from ncRNA. Similarly, some NE genes like *mpn121* are shown to contain overlapping smORFs.

In line with the inclusion rate observed for annotated genes in *M. pneumoniae*, 22% out of the 153 smORFs identified as SEPs present a RBS defined as a Shine-Dalgarno sequence close to the translation initiation codon. The lack of Shine-Dalgarno motifs at the first gene of an operon suggests translational noise could be happening, in a similar manner to what has been described for antisense

ncRNAs due to the lack of a -35 element [286]. This is supported by the low coverage obtained in the *Barn* libraries and the high number of significant ORFs with enriched in-frame insertions (>7000). Thus we see a significant degree of low translation all over the genome of this bacterium. This is important since the low expression of ORFs paves the way for the evolution of new functionalities that when needed could be selected to increase their expression. It seems that the low translation of few protein copies in many regions of the genome is not so deleterious for the bacterium we have analyzed here.

As the expression of the resistance or anti-selection reporter is dependent on the expression of the protein with which is fused, we can roughly estimate protein abundance. We show that, within essentiality categories, those genes with higher coverages tend to be more abundant than those with lower coverage. This is exemplified by the fact F genes retain similar coverages to those found in NE genes, as F proteins are present at higher copies per cell than NE proteins in *M. pneumoniae*. With the barnase library, we detect insertion in-frame in NE genes but only when they are very lowly expressed.

The analysis of the libraries described hereunder different experimental conditions could be a useful tool in determining which proteins are being expressed and at which quantities. However, we should have a word of caution since we identified 70 F and NE genes which are not detected in free-label MS searches of which at least we know they can only be seen when preventing the expression of the Lon protease [74]. An interesting example mentioned in the results shows how we can detect in-frame insertions at the N-terminal region of a pseudogene not detected by MS and no insertions after the stop codon. Thus in some cases, protein abundance determined by ProtInSeq could be affected by protein stabilization due to the fusion to the antibiotic selection marker. This can be exemplified by looking at FtsA and FtsZ genes that have degrons at their C-terminal and are lowly expressed, which have preferential insertions at its C-termini and can be identified at a high concentration of chloramphenicol.

When exploring transmembrane and membrane-associated proteins, enrichment of insertions was observed in the predicted cytoplasmic segments of these proteins with respect to transmembrane and exposed segments. We hypothesize this happens because the fusions in the outer segment will expose the resistance and, consequently, will not confer resistance to the cell. This is well seen in the case of NE lipoproteins, which accumulate insertions only in their N'-termini regions, which is the only cytoplasmic segment presented by this family of proteins. Furthermore, for NE proteins with more complex membrane topology, we observed that most of the in-frame insertions found in these genes correspond to cytoplasmic segments predicted by TMHMM [34] (81% ± 17%).

We take advantage of the CPD algorithm, developed for the detection of differential essentiality domains in proteins, to predict the expected topology of 101 NE membrane related proteins (i.e. NE regions will be cytoplasmic, while E regions should correspond to exposed segments). We could predict 41 proteins with the same topology as TMHMM (± 10 aa), 21 could not be analyzed due to very small internal or external loops, and 39 were predicted properly except for one segment (cytoplasmic or exposed). This in principle incorrectly assigned segment could be because of the wrong assignment of TMHMM as exemplified by MP359 and supported by the SPLIT server [467]. Thus, this type of transposon library can be used to explore the topology of membrane proteins.

Altogether, this technique supports and complements information retrieved by proteomics using ultra-deep sequencing samples with positive and negative selection reporters. This methodology also allows the identification of annotated and new ORFs and smORFs in bacterial genomes, their relative quantification, and studying membrane topology features, a sequencing. As this technique can be applied generally in bacterial genomes, we envision ProTInSeq as a future standard in the experimental identification of SEPs. Ultimately, this tool could be used to identify bioactive SEPs in bacterial communities such as microbiomes, where SEPs play fundamental roles in the homeostasis of the population, with potential interest in microbial therapies.

6.5. Material and Methods

6.5.1. Experimental protocol to generate ProTInSeq libraries

a) Molecular cloning

Nine different vectors were obtained to define the three libraries used in the current study (chloramphenicol, erythromycin and barnase libraries). All the vectors were derived from the vector pMTnCat_BDPr [460]; a mini-transposon vector derived from the Tn4001 version (Table 2). They were obtained by using Gibson assembly (New England Biolabs) of three different fragments, following the instructions of the manufacturer.

b) Bacterial strains and growth conditions.

Escherichia coli strain Top10 (Thermo Fisher) cells were grown at 37ºC in 2YT broth or LB agar plates containing 75 µg/ml ampicillin when needed. *M. pneumoniae* M129 strain was grown in 75 cm2 tissue culture flasks with 50 mL of modified Hayflick medium (HF) at 37ºC and were transformed as previously described [217]. To select *M. pneumoniae* transformant cells, plates were supplemented with 20 µg/ml chloramphenicol or 0.02 µg/ml of erythromycin. Transformed cells were also grown in liquid cultures and testing different concentrations of antibiotics. First, *M. pneumoniae* M129 was grown in a 96 well plate format with 200 µl of HF and 5 µl of transformed cells. For chloramphenicol the tested concentrations were 0, 0.5, 1, 2, 5, 10, 15 and 20 µg/ml. In the case of erythromycin the tested concentrations were 0,002 and 0.02 µg/ml. Concentrations of 0.5, 1, 2, 5, 10, 15 µg/ml for chloramphenicol libraries and 0.02 µg/ml for erythromycin libraries were selected. To study the proteome of *M. pneumoniae*, transformed cells were grown in T75 flasks with different concentrations of antibiotic (0.5, 1, 2 and 15 µg/ml of chloramphenicol) to cover from low to highly expressed proteins. After 24 h cells were passed to a T300 cm^2 flask, cultures of cells grown with 0.5, 1, 2 and 15 µg/ml of antibiotic were confluent after 48h and cultures of cells grown in 15 µg/ml required three additional days.

c) Transformation of M. pneumoniae.

Transformations of *M. pneumoniae* were performed by electroporation as previously described [279] but with a slightly modified protocol. Briefly, cells grown in two T75 cm^2 flasks were recovered in a 2ml electroporation buffer and 80 µl of cells were electroporated with 2 pmol of different vectors (2 technical replicates per sample and two biological replicates). After electroporation cells

were resuspended in a final volume of 1 ml by adding 900 µl of HF. The 2 transformations of each vector were pooled (total volume of 2ml). Five hundred µl of cells were seed in 20ml of medium in a T75 flask with different concentrations of chloramphenicol (0.5, 1, 2 and 15 µg/ml) for 4 days at 37ºC in 5% CO_2. After one day of incubation each flask was resuspended in 1.5 ml of medium and cells were seeded in 150 ml of medium in a T300 flask. After 48 days of growth at 37ºC in 5% CO_2 DNA of samples treated with 0.5, 1, 2 µg/ml of chloramphenicol was extracted. The samples of 15 µg/ml were processed after 72 additional hours. This experiment was repeated twice and DNA samples were sequenced independently.

Also, in parallel, *M. pneumoniae* transformed cells were spread on Hayflick agar plates supplemented with 20µg ml^{-1} chloramphenicol and incubated at 37ºC in 5% CO_2. CFUs were accounted for after 1 week. Percentage of transformants was estimated by:

$$transformants\,[\%] = 100\ x\ \frac{CFUs\ in\ HF+Cm}{CFUs\ in\ HF}$$

d) Estimation of efficiencies of transformation in different libraries

As described above, the efficiencies of transformations shown in Figure 1 were measured by counting the colony forming units (CFUs) in plates with and without the antibiotic and doing the ratio. The analysis of the variance was done from four different transformations (n=4) and the different experiments were normalized versus one of the samples: TnP_{438}*cat*IR* for the libraries of the experiment of chloramphenicol selection, TnP_{Syn}eryAIR* for the libraries of the experiment of erythromycin selection and TnP438catIR* for the libraries of the barnase experiment.

e) DNA manipulations

Genomic DNAs of *M. pneumoniae* M129 were isolated with the Illustrabacteria genomic Kit (GE). The purification of PCR products and digested fragments from agarose gels were achieved using the QIAquick Gel Extraction Kit (Quiagen). Plasmid DNA was obtained by using the QIAprep Spin Miniprep Kit (Quiagen).

f) Sequencing of transposon libraries

Genomic DNA sequencing was performed in the Genomics facility at the Centre for Genomic Regulation in a HiSeq Sequencing v4 Chemistry controlled by Software HiSeq Control Software 2.2.58. Settings, 150 nucleotides in paired-end format. In the HiSeq Rapid Run sequencing technology from Illumina Genome Analyzer, the protocol starts with DNA fragmentation. Then, the fragmented

DNA is amplified using oligos specific for the *cat*, *ereA* or barnase genes that also add adapters to the glass flow cell. Later, the sequencing is performed by synthesis cycles, in which a single complementary base for each deoxynucleotide (dNTP) is incorporated using a fluorescently labeled dNTP. Finally, lasers excite the fluorophores while a camera captures images of the flow cell. In total, we sequenced 41 samples with 4 replicates for each *CmB5* and *CmB15* samples; 3 for each *Cm5A*, *Cm10A*, *Cm15A*, *Cm10B*, *Barnase*; and 2 replicates for the rest of conditions presented, including *EryB1*.

6.5.2. Database covering sequence features and measures

We used the *M. pneumoniae* M129 genome sequence, applying corrections from the latest in-house strain sequenced version, to define all putative ORFs , with translation product length ≥1 amino acids, from the six possible open reading frames (starts=ATG, TTG, GTG, stops=TAG, TAA). Considering *M. pneumoniae* do not require ribosome binding sites (RBS) motifs to start translation [463], we did not set any size threshold as, theoretically, the resistance of the mutated transposon could be expressed in fusion with any translated sequence independently of its size. In total, 30,113 sequences were defined, these included the 689 known annotated coding sequences of *M. pneumoniae*. For each sequence, all the available information was recapitulated including coordinates, protein localization and function. We also included transcription-related information as to whether the annotation belonged to an operon or not, average expression (as $\log_2$(gene read count/gene length) and estimated average RNA copies per cell considering 4 RNA sequencing samples covering different growth times (6, 24 and 48 hours, ArrayExpress identifier E-MTAB-6203). From previous studies, we considered the detection at protein and peptide level, available for 12,426 sequences that present an amino acid length ≥19 (from 116 mass spectrometry experiments, ID PRIDE: PXD008243), average protein copies per cell, estimated half-life, and homology with a database including 109 smORFomes [448]. Finally, we also included transmembrane segment predictions and signal peptide presence estimated using TMHMM [468] and Phobius [386], respectively.

For the Ribosome Binding Site inclusion rate calculation, 15 bp upstream start codons we look for any of the motifs reported to act as Shine-Dalgarno motif: GGA, GAG, AGG, AGGA, GGAG, GAGG, AGGAG, GGAGG, AGAAGG, AGCAGG, AGGAGG, AGTAGG, AGGCGG, AGGGGG and AGGTGG [173].

The topology prediction by TMHMM consists in assigning the label *i*=cytoplasmic, *m*=membrane, *o*=outer to represent the location of the segment with respect to the membrane. In order to perform the different analyses, we reduce this information to the percentage of aa with the *i* label respect the total aa length. In addition, and with the purpose of having a negative control in the

analyses, we defined a set of intergenic sequences with their coordinates extracted from all the genome spans between ORFs distinguishing between strands, and presenting low RNA expression profile with values log2(RNA read count/gene length) < 2 (n=1700). This set includes a total of 786 intergenic annotations extracted from the positive orientation (average sequence length = 25 ± 20 bp) and 914 from the negative orientation (24 ± 19 bp).

6.5.3. Identification of transposon insertion sites.

We used FASTQINS to retrieve the number of times, as read count, each base along *M. pneumoniae* genome (816,394 bp) was found next to a transposon insertion event. This tool selects for reads including specific sequences known as inverted repeats (IR) introduced during the transposition, trims the inserted segment and maps the remaining sequence to the reference genome reporting the base pair position next to the trimmed section that corresponds to the insertion point. We consider our settings strict as we only consider reads mapping in paired-end, unambiguously and with no mismatches. As we were interested in extracting the orientation of the transposon insert, the IR sequence used to select reads was extended to include the beginning of the resistance/marker and FASTQINS was run using the strand-specific mode in the way the results for the positive and negative strand will include only insertions with the resistance/marker oriented in the positive or negative sense, thus producing viable fusions, respectively. After running this procedure over our library including 39 samples, we obtained 78 profiles (one per each genome strand orientation) and the genome coverage (percentage of the genome that was found disrupted) and the total read count per sample (sum of read count for every position).

Taking as reference the 30,113 ORFs found in *M. pneumoniae* M129 and considering the design of our transposon where only insertions happening in the first position of a codon can produce viable fusions, we labeled each position in *M. pneumoniae* genome with the following excluding labels: the first label, *annotated*, is assigned to bases corresponding to the first positions of codons in annotated proteins, thus, an insertion found there would express that protein in fusion with our selection resistance/marker. We assigned this label to the 17.4% (n_{pos}=142,443) and 12.5% (n_{neg}=102,840) of the positions in the positive and negative strands, respectively, of *M. pneumoniae* (genome size = 816,394 bp). The second label, *putative*, considered the same as *annotated* but taking only non-annotated entries. This covered the 39.8% (n_{pos}=325,115 bp) and 44.1% (n_{neg}=360,302 bp) of the positive and negative genome strands. Finally, the *non-coding* label was assigned to the 42.7% (n_{pos}=348,836 bp) and 43.3% (n_{neg}=353,252 bp) of the positions, representing those cases where an insertion would be considered as inexplicable as no translation is expected. This last group includes for example second and third positions of codons in any annotation (if it does not present overlapping annotations) or any position located in-frame and

downstream to a stop codon (this last case will correspond to positions within the intergenic annotation defined in the previous section). Additionally, we also considered within the *annotated* two different subsets of positions corresponding to in-frame positions of a set of genes with known E and NE essentialities, described as essentiality 'gold set' in *M. pneumoniae* [234], including the following sizes nE_{pos}= 5,823 bp, nE_{neg} = 10,139 bp, nNE_{pos} = 4,258 bp and nNE_{neg} = 5,823 bp. For each of these positions types, we accounted for the coverage (percentage of positions found disrupted), total read count, mean, median and standard deviation of read distributions under 4 different filtering conditions: no filter (0), removing 0-reads positions (1), ≥16 reads positions (16) and filtering out reads below the 5th percentile and above the 95th percentile (90) as suggested by previous transposon sequencing studies [241]. Coverage and read count explorations were performed within the ANUBIS transposon sequencing exploration framework which includes automated functions to retrieve these values (Chapter 5).

6.5.4. Identification analysis

For each sample presenting a selective profile, we first filter out insertions with read values in the range of the tails of the *read counts* distribution and ignoring repeated regions where mapping is inefficient as done in previous studies [241,469]. Distinguishing by strand and replica, we model the background of the coverage distribution from *non-coding* positions with no RNA expression (log2(reads/bp)<2) in the *M. pneumoniae* genome and we calculate the probability of each ORF to fit that distribution. Then, we consider as 'identified' those ORFs presenting a significant increase of insertions (P<0.05), thus presenting a higher rate of in-frame insertions than expected by chance normalized by their expected gene length. These evaluations were performed with the *Poisson* prediction method implemented in ANUBIS.

In order to retrieve candidates with a higher number of insertion than expected by chance, we compared the annotated genes of *M. pneumoniae* (Positives, P; n=689) against the set of negative control sequences derived from intergenic regions (Negatives, N; n=1,700). Using a Receiver Operating Characteristic (ROC) we evaluated the relation between True Positive Rate (i.e. true positive, or TP, for annotated protein detected; and false negatives, or FN, for annotated proteins with no signal) and False Positive Rate (i.e. false positive, or FP, for intergenic annotations detected as ORF; and true negatives, or TN, for intergenic annotations with no signal).

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

And False Positive Rate:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

The Area Under the Curve (AUC) increases with high TPR and low FPR values; thus, it can be used to minimize the FPR and as threshold to ensure all the candidates present more insertions than what could be expected by chance. In addition to this, we set a second condition for the detection which requires an ORF to be reproducible in at least two samples.

In order to facilitate the analysis of these Tn-Seq mutated libraries, we have implemented new options to our previously published bioinformatic tools for essentiality studies. First, the pipeline of transposon insertions mapping (FASTQINS) includes a strand-specific mode to separate insertions by orientation. On the other hand, the set of essentiality assessment tools included in ANUBIS present new functions and subroutines to perform the different processing and estimation analyses distinguishing by frame, and visualize this data (Chapter 5).

Chapter 7. Additional Transposon Sequencing and Machine Learning applications in Diverse Biological Contexts

Standardization of Tn-Seq analyses with FASTQINS and ANUBIS, methods described in Chapter 5, promoted their application in different collaboration projects performed within the frame of this thesis project. In this chapter we introduce the main findings of these studies with special attention on the application of the developed tools.

7.1. Efficient transposon transformation in minimal genomes and application in metabolic studies

First example covers the evaluation of an optimized version of the Tn4001 mini-transposon vector with improved transformation efficiency in genome-reduced bacteria other than *M. pneumoniae*. In the work entitled *'SynMyco transposon: engineering transposon vectors for efficient transformation of minimal genomes'* by Montero-Blay, *et al.* (2019) [470], the use of vectors with additional regulatory regions, such as ribosome binding sites (RBS), is demonstrated to improve the transformation efficiency in *M. agalactiae* and *Mycoplasma feriruminatoris*. By inspection of alignments of the most expressed genes in 10 Mycoplasmas, different species-specific regulatory elements were characterized. Also, RBS-dependent expression is defined as one of the factors preventing the expression of transposases in Mycoplasmas other than *M. pneumoniae, M. genitalium* and *M. gallisepticum* via genome exploration of 20 bacterial species. The new vector version allowed the efficient recovery of Tn-Seq profiles in *M. agalactiae* 7784 and its genome essentiality estimation using ANUBIS from a coverage of ~23.3 insertions every 100 bp (Figure 7.1). The steps followed in the engineering of this vector can be used as reference to improve the application of transposon sequencing in other bacterial species where deep coverage was not reachable; such as *M. bovis* (coverage=0.03% of coverage) or *M. genitalium* (coverage=0.56%) [236]. In addition, this work presents the first complete genome sequence assembly and *de novo* annotation of *M. agalactiae* strain 7784, submitted as BioProject under accession PRJNA528179; and Tn-Seq datasets for this same species in ArrayExpress with accession number E-MTAB-7425.

The data generated in that first work was later analyzed from a metabolic perspective focusing on pathways connected to carbon metabolism in *M. pneumoniae* and *M. agalactiae*. In the work entitled *'Inferring Active Metabolic*

Pathways from Proteomics and Essentiality Data' by Montero-Blay, *et al.* (2020) [471], detailed maps of carbon metabolism in these Mycoplasma species are defined. The absence of key enzymes in *M. agalactiae* impacts the essentiality of the rest of the pathway. Assessing these differences combining homology studies, essentiality, and quantitative proteomics data, provide insights in function redundancy, essential components that intersect in different subpathways, fluxes and directionality of the pathways involved in glycolysis.

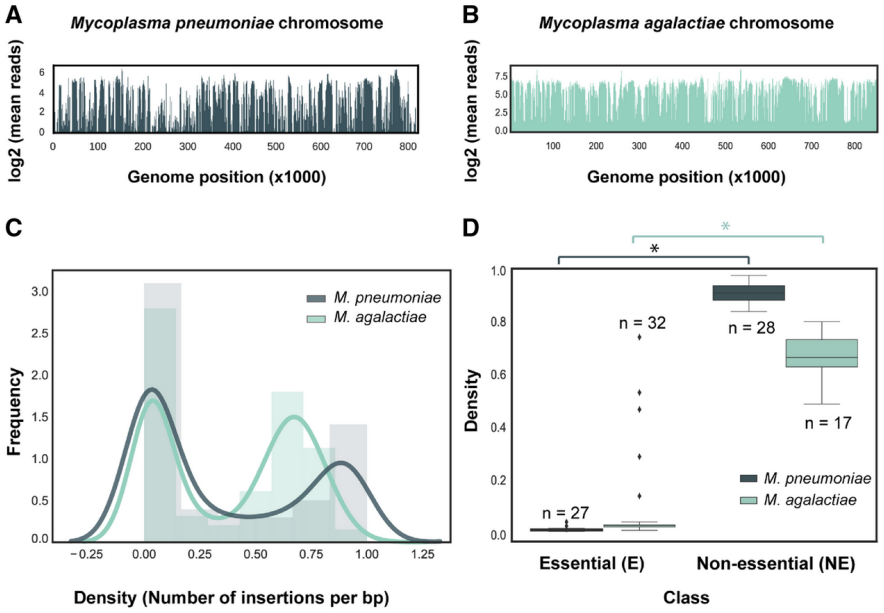


Figure 7.1. Essentiality study in M. agalactiae using the pMTnGm-SynMyco transposon and a comparison with previous studies in M. pneumoniae.

(**A**) Genome disruption profile for *M. pneumoniae*. The y-axis represents the logarithmic average of total reads covering a window of 1,000 bp (x-axis). (**B**) Genome disruption profile for *M. agalactiae* representing the same information as in the previous panel. (**C**) Insertion density by gene distribution in *M. pneumoniae* and *M. agalactiae* as indicated. The x-axis represents the percentage of bp in a gene that is disrupted and the y-axis the frequency of densities in the distribution. To better compare *M. pneumoniae* and *M. agalactiae* transposon insertion distributions, we standardized both distributions using min-max scaling. (**D**) Box-plot representing the statistical comparison of specific subsets of genes expected to be essential (E) and non-essential (NE) in *M. pneumoniae* and *M. agalactiae* as indicated. The asterisk represents *P*-value < 0.05 ($3.62\ e^{-41}$ and $1.20\ e^{-20}$) when comparing the density of insertions of reference E and NE genes in *M. pneumoniae* and *M. agalactiae*, respectively. * Figure extracted from Montero-Blay, *et al.* (2019).

7.2. Inducing random deletions in *M. pneumoniae* genome

In the work entitled *'LoxTnSeq: Random Transposon insertions combined with cre/lox recombination and counterselection to generate large random genome reductions'* by Shaw, *et al.* (2020) [472], a new methodology to randomly delete genome regions is presented. This technique is based on the sequential insertion of Lox66 and Lox71 sites using Tn4001 mini-transposons in the *M. pneumoniae* genome. By Cre recombinase induction, the span of the genome located in between these two sites will be deleted, leaving a Lox72 scar. In this context, FASTQINS was used first to assess the insertion coverage of the different transformation steps. From the first pool of mutants with Lox66, 355,319 unique insertion sites were recovered (43.5% of genome coverage). This represented an insertion every ~3 bases, an insertion frequency similar to what we describe in Chapter 5. The second pool derived from that first, and transformed with Lox71 sites, revealed that the second transposon was inserted in 187,814 unique positions (genome coverage of 23%; 1 insertion every ~4 bases). Combining the insertions of the two samples, a total of 387,962 unique insertions were recovered.

Pools of mutants carrying the two lox sites were induced by Cre recombinase and sequenced by ultra-deep technologies via a circularization protocol to identify both deletion sites in a single sequencing read. These sequencing samples represent reads with a lox72 site flanked by two genomic regions. A total of 1,291,712 reads were recovered using this protocol. However, due to the random nature of the circularization protocol, not all reads comprised the lox72 and enough genomic DNA to map the deletion points for each transposon. In this context, FASTQINS was used to map the flanking genomic regions located next to the inverted repeat contiguous to the lox72 scar. The different deletion candidates were prioritized by read count values and their genomic context using ANUBIS (i.e. if they were containing E genes, which should not be possible by principle). Then, using a read count as a threshold parameter, we performed a receiver operating characteristic (ROC) curve approach to defining 8 high-confidence deletions and 285 additional ones. The largest was 28.7 Kb, the smallest < 50 bp, both contained in the high-confidence set, and total mean size of 7,750 bp. Up to 147 genes could have been deleted across the pool, accounting for 171.2 Kb (21% of the genome), most of the presenting unknown functions. Remarkably, 139 genes were annotated as non-essential genes, with the remaining 8 classified as conditional fitness genes. By PCR, a selection of 4 of these cases were tested validating the deletion of regions containing: i) seven NE genes (*mpn096* to *mpn102*), ≈10 Kb in size; ii) four NE genes (*mpn397* to *mpn400*, ≈5 Kb); iii) 19 NE genes and 1 F gene (*mpn493* to *mpn512*), with a deleted area of ≈25 Kb; and a region comprising 6 NE genes (*mpn368* to *mpn373*, ≈8 Kb).

7.3. FASTQINS applied in a cancer context

Cancer is an evolving multifactorial disease where the different combinations of genetic and epigenetic alterations determine the aggressiveness and progression of the tumor cells [473]. This results in molecular and phenotypic heterogeneity within the tumor, the complexity of which is further amplified through specific interactions between cancer cells. In the study '*The role of clonal communication and heterogeneity in breast cancer*' by Martín-Pardillos, *et al*. (2019) [474], clonal cell lines are derived from the MDA-MB-231 breast cancer cell line, using the UbC-StarTrack PiggyBac transposon system, which allowed tracking by color: GFP C3, mKO E10, and Sapphire D7. Co-culture of these cells revealed genetic and epigenetic differences affecting growth rate, metabolic activity, morphology, and cytokine expression among the different cell lines. *In vivo*, all the clonal cell lines formed tumors. Interestingly, co-injection of an equal mix of the different labeled clones in mice showed that mKO E10 cells were unable to form lung metastases confirming that even in stable cell lines heterogeneity is present. Also, co-growth and co-injection of mKO E10 and GFP C3 clonal cell lines showed increased efficiency of invasion and migration, these findings support an interplay between clones determining the aggressiveness of a tumor. Further exploration of these results may allow the identification of cellular communication factors in clonal cooperation that could be targeted for preventing tumor progression.

In this context, FASTQINS was used to validate and characterize where the different fluorescent markers were inserted in MDA-MB-231 cells. Remarkably, the list of inserted positions required an additional prioritization step as the inverted repeat sequence used in the system was endogenously found in the human genome. To differentiate these artefactual cases from genuine transposition events we took advantage of one of the transposase properties in Chapter 5: they do not perform even cuts in both reverse/forward strands but a staggered cut, which generates a transposon duplication site, of 5 bases in the case of PiggyBac transposons [475]. While in a conventional study this could be considered as an artifact, this was taken as a criterion to distinguish actual from artifactual detections. Analysis of the genomic context of the different insertion points among cell lines resulted in the definition of a map of the chromosomes and genomic loci disrupted in each cell line which was taken into consideration in the final phenotypic assays performed experimentally. In conclusion, this study served as a proof-of-concept of the application of FASTQINS in a context other than bacteria, aiding the study of important biomedical questions.

Chapter 8. Discussion

Proteins are in charge of the main biochemical and structural functions in the cells. Thus, their annotation in genomes is a fundamental process in understanding the range of functions an organism can perform. In most genome projects, the first step after acquiring a genome sequence is predicting protein-encoding open reading frames (ORFs). However, these approaches largely underestimate a layer of proteins, encoded by short open reading frames (smORFs) encoding for small proteins (<100 amino acids; SEPs). In addition, high-throughput technologies are not always reliable in the detection of SEPs and most of the known SEPs have been found by chance [476]. From available studies, it can be seen that SEPs participate in a wide diversity of cellular processes with different mechanisms of action.

On the other hand, described bacterial SEPs can play roles in cell signaling, act as chaperons, antibiotics or toxins/anti-toxins, modify membrane properties, stabilize protein complexes, or serve as structural proteins [101–103,107,108,113]. The gap of information in currently available genomes where SEPs have not been annotated highlights the necessity of new computational and experimental approaches to study these proteins.

In this thesis project, we start defining the regulatory elements that govern transcription in a minimal cell and also increase genome complexity (Chapter 3). Then, we evaluated the detection by high-throughput technologies of SEPs in bacterial genomes and provided a gene identification tool to aid their discovery (Chapter 4). We then presented a couple of tools for the efficient analysis of transposon random mutagenesis coupled to ultra-deep sequencing (Tn-Seq, Chapter 5), required for the analyses presented in the last chapters. These tools could be used not only for the detection of SEPs but for any '-omics' study using transposons. In Chapter 6 we demonstrate that identification, quantification, and exploration of protein sequence features, can be performed from Tn-Seq libraries in addition to support the existence of SEPs predicted in Chapter 4. Furthermore, standardization of the tools presented propitiates additional applications such as metabolic network exploration, random deletion of genomic regions, or support cancer studies (Chapter 7).

8.1. SEPs identified by high-throughput and machine learning approaches

In Chapter 4 we present two lines of research: the first comprehensive study of a bacterial proteome using mass spectrometry without protein size thresholds, and a gene identification software for the annotation of SEPs in bacterial genomes.

8.1.1. Detection of SEPs by mass spectrometry

By integration of 116 MS, we provided the most extensive proteome study in *M. pneumoniae*. We showed that the detection and characterization of SEPs are challenging due to diverse factors. First, "decoy" protein sequences that do not exist in *M. pneumoniae* can have spectra assigned appearing across multiple experiments, making discrimination criteria based on reproducibility not feasible. This problem was solved by only accepting proteins detected with ≥ 2 UTP, a criterion that made even more complicated the proper identification of SEPs, which already have few UTPs due to their small size. Also, we evaluated the responsiveness of UTPs (i.e. present the features required to be detected by the mass spectrometer) showing that it also impacts the probability to identify a protein by MS.

Despite these constraints, however, we still identified 43 potential SEPs in *M. pneumoniae*, 7 of them passing the threshold of ≥ 2 UTP. These results were supported by C_{13} labeled peptides for 8 of these SEPs. The 4 SEPs in this group with 2 UTPs were validated, while only two out of 4 in the group of 1 UTP were real, supporting the idea that 1 UTP is not enough to validate the existence of a protein. Moreover, these criteria were corroborated by re-analyzing Ribo-Seq data from *E. coli*. Finally, we also applied these types of searches on cell extracts and SDS gel extraction derived from 12 additional bacterial species, identifying 25 new SEPs not annotated in their reference genomes.

Further research is required to completely assess the feasibility of proteomic studies in detecting SEPs. These thesis results were obtained by using Mascot searches, which are based on pre-computed databases, and a probability-based score system to assign the different spectra to the proteins considered [389]. However, other approaches not tested could identify SEPs previously missed. For example, by using SPIDER, a software for proteomic searches which do not rely on pre-computed databases and uses BLAST to match the peptides detected to the possible translated sequences in a genome [477]. Moreover, machine learning approaches have been shown to be able to recover proteins from Mascot searches that would be discarded in conventional searches as they do not present UTPs [478].

Also recently, protocols to enrich for SEPs by MS applied to human plasma samples have rescued more than 100 SEPs, including C5ORF46, a new SEP related to lipid homeostasis [479]. Evaluation of the performance of these approaches is required to ultimately assess the identification of SEPs by MS. In any case, taking into account the required number of experiments and the fact that many SEPs do not have high-responsive peptides, the extensive analysis of SEPs encoded by a genome would still be less than complete. In addition, other factors

could contribute to this problem like short protein half-lives, conditional gene expression, or special features in sequence associated with concrete functions (e.g. high hydrophobicity in transmembrane proteins). To overcome these limitations, this thesis project proposes novel solutions based on machine learning-based predictions and transposon sequencing protocols that allow the exploration of proteomes from ultra-deep sequencing technologies (Chapter 6).

8.1.2. Prediction of SEPs using machine learning

Using the proteomic knowledge acquired in *M. pneumoniae* as a reference, we develop RanSEPs to score smORFs candidates by their probability of being SEPs. This tool combines principles presented by the *Ab Initio* and homology gene identification algorithms in Chapter 1 but interpreted by a supervised machine learning algorithm able to prioritize genetic signals in a species-specific manner. As novelty compared to previous gene identification approaches, RanSEPs considers a wide range of nucleotide and amino acid sequence features (>3,000), together with regulatory genetic signals such as ribosome binding sites, promoters, and terminators, whose specificities in genome reduced bacteria are discussed in Chapter 3.

These features are used to train a random forest algorithm that prioritizes the most explanatory distinguishing between actual and non-coding sequences. In the end, this model predicts 756 ORFs in the *M. pneumoniae* genome: 612 ORFs (598 annotated, and 14 new) and 144 smORFs (26 annotated SEPs, and 118 new candidates). In the group of predicted new SEPs, we found the 4 SEPs with 2 UTPs, and the 2 with 1 UTP, validated by C_{13} labeled peptides. On the other hand, the 2 putative smORFs with 1 UTP that were not validated by C_{13} peptides were classified as negatives. This suggests that RanSEPs could be used to discriminate real proteins from noise in the group of candidates with only 1 detectable UTP.

Furthermore, we validate this application by running the prediction tool in a test set comprising 570 experimentally validated SEPs from 12 different bacterial species. The tool was compared to five other gene identification programs, including those used for the *de novo* annotation of genomes submitted to public databases (Chapter 1). Results suggest that RanSEPs maximize the correct identification of proteins with no increase in false positives, surpassing the rest of the tools. Considering that closely related species share sequence features, our scoring algorithm could also be adapted to *de novo* annotate a genome of interest.

Analysis of features that discriminate coding sequences in predictions of 109 bacterial small proteomes revealed that hydrophobicity and secondary structure are key factors in the predictions, conserved across species. Also, we observed that the number of predicted SEPs encoded by a genome depends on the GC

content. Strikingly, between 13 ± 7 and 16 ± 9.5% of the genes in the 109 species analyzed could encode for SEPs. Genome annotations are critical for classifying a SEP as a new protein. In fact, for 76% of the SEPs predicted by RanSEPs, orthologous SEPs were identified by BLAST in closely related strains. This result indicates that reference genomes are still incomplete and not properly curated.

Functional analysis of the predicted and previously identified SEPs corroborated their participation in essential processes such as transcription, translation, metabolism, signaling, *quorum sensing*, virulence, and pathogenicity. However, this analysis should be taken with caution as sequence homology and functional annotation of SEPs is challenging [89]. Interestingly, we found a significant enrichment in SEPs presenting features indicative of being secreted or membrane-localized (25%).

As a limitation of this tool, eukaryotic genomes cannot be analyzed due to the computational challenge of processing, featurization, and modeling of all the alternative splicing events that could synthesize a SEP. However, recent advances in the field of machine learning demonstrate that deep neural networks can handle this information [480]. This type of approach could be used coupled with RanSEPs to find those smORFs sharing features with known SEPs in eukaryotic organisms.

With all our results in mind, RanSEPs aids the computational annotation of SEPs, support their detection, and discard artifactual proteins detected by MS that have low signals, such as those detected with only one UTP. When no experimental information is available, RanSEPs can guide the selection of potential new SEPs for validation and further characterization.

8.2. Bioinformatic tools for the standardization of transposon sequencing technologies

One of the fundamental methodologies used in the field of Synthetic Biology is transposon random mutagenesis combined with deep sequencing (Tn-seq), which can be used to define the genes or regions in an organism that are required for sustain life, or that can be edited when rational engineering is required [234]. Despite this methodology being applied for more than a decade, we noticed that established guidelines and standards on how to analyze Tn-Seq data were missing [234,440,481–485]. Thus, to define a transposon methodology able to depict coding sequences in a genome, we were first required to evaluate the different proposed tools and define valid approaches for our purposes in Chapter 6. The evaluation of available tools in this field, and development of the required bioinformatics tools to fill the gap in Tn-Seq analyses platforms, propitiated the definition of FASTQINS and ANUBIS, which are presented in Chapter 5.

We developed FASTQINS to extract insertion profiles from Tn-seq data overcoming problems presented by other approaches proposed for the same task. First, most of the tools were designed to specifically analyze Tc1/Mariner transposon accounting only insertions in TA-sites [486], and consequently not being compatible with protocols using Tn5 transposases. On the other hand, tools that could analyze both types of libraries, either present licensed dependencies or they are web-based platforms, a factor that makes the tool more accessible but also prevents its scalation to bigger projects with more than a few samples [487,488]. Opposite to these tools, FASTQINS uses sequencing bioinformatics standards, which ultimately allows its application in a wide range of situations. For example, we showed in Chapter 7 that it can be applied to retrieve random deletions induced by a Cre/Lox system coupled to transposon sequencing. Moreover, it was also applied to analyze PiggyBac transposon libraries in breast cancer cell lines [474]; an example that proves our tool can process Tc1/mariner data (PiggyBac disrupts TTAA-sites), and it can be used in a eukaryotic cell context. In addition, the parallelization and task-recovery features implemented in FASTQINS have allowed its integration in an in-house web server ('DBSpipes'), which has already processed 148 Tn-Seq samples, together with the re-analysis of more than 350 samples previously generated by our group.

On the other hand, we present a bioinformatics framework, called ANUBIS, which covers data loading, quality control, preprocessing, metric extraction, normalizations, essentiality estimation, and visualization. This framework constitutes the most complete bioinformatic tool for Tn-Seq data analysis available, as previous ones were only focused on statistical evaluation [427,434], or only compatible with Tn5 transposase-based protocols [241]. To test this tool and FASTQINS, we generate a library of *M. pneumoniae* transposon mutants, sampling 7 representative timepoints out of 10 serial passage dilutions. This dataset achieves the highest insertion saturation obtained for a bacterial species (~1 insertion every bp). This enables us to explore potential artifacts and their impact on the essentiality estimates along with different selection, and consequently coverage, conditions.

With these datasets, we are able to fully characterize factors that could have been biasing previous essentiality estimates. Specifically, we highlight the importance of considering sequence composition biases, duplicated signals produced by transposases-derived duplications, repeated regions, and protein domains, either internal or in the N' and C'-termini domains of proteins. Despite some of these factors having been already described, like the accepted insertions in N' and C'-termini regions [236,481], none of these studies evaluated their impact in the estimates, neither proposed solutions to alleviate their effect. Also, for this study, it was crucial to understanding how the selection of an essentiality estimation model could affect the prediction. To evaluate this, we reimplement four

previously available approaches for essentiality estimation, in addition, to define a novel method based on Gaussian Mixture Models. As a novelty, this unsupervised machine learning algorithm is used to provide estimates without requiring a set of genes with known essentiality as parameter reference; thus reducing the biases induced by the selected sets by the user.

After i) testing, validating, and implementing corrections for the detected biases in ANUBIS; ii) benchmarking the different essentiality prediction models with randomized sets; iii) iterate the process along with different coverage conditions; we show that up to 125 genes could change categories depending on these factors. With these results, we highlight and provide tools to solve these considerations, which ultimately will be relevant not only for fundamental genomic essentiality studies but also when designing new organisms, where the proper categorization of genetic elements between dispensable or essential is determinant to succeed in the genome reduction of an organism [276].

In addition, in Chapter 7 our tools were used to analyze the results obtained with new versions of the Tn4001 mini-transposon designed to work efficiently in Mycoplasma species where standard transposons, used in *M. genitalium* and *M. pneumoniae*, were not working properly; and as a tool to induce random deletions based on a Cre/Lox system. Also, we performed the essentiality assessment in the infection process of *M. bovis*, results that are now under intellectual property protection). Taken as a whole, we believe these types of approaches, in combination with FASTQINS and ANUBIS, can propitiate the definition of new applications to rationally engineer cells. Moreover, the principle behind treatments against pathogenic bacteria or tumor cells is inhibiting their essential functions so the presented tools could aid in the development of new therapeutic targets.

8.3. A novel transposon sequencing approach to perform protein studies

Once FASTQINS and ANUBIS are available, we can envision novel approaches taking advantage of the high coverage found in *M. pneumoniae*. One of the relevant advances presented in this thesis project is a transposon-sequencing protocol to explore protein features from ultra-deep sequencing, and identify new SEPs.

8.3.1. Ultra-deep sequencing identification of SEPs

In Chapter 6 we present one of the novel applications derived from the standardization of transposon sequencing. With the aim of filling the gap in high-throughput technologies capable of detecting SEPs, we took advantage of the high resolution presented by transposon random mutagenesis in *M. pneumoniae* to define a new protocol that allows exploration of proteomes at a qualitative and quantitative level. In this work, we design a variation of the Tn4001 mini-transposon where the reporter gene is expressed only when inserted in-frame, thus producing a fusion, with a translated ORF.

We tested different antibiotic resistance genes (to chloramphenicol and erythromycin, respectively); and a third vector, carrying the barnase gene which encodes for a highly toxic ribonuclease, used as a negative selection marker [461,462]. By transposon random mutagenesis, *M. pneumoniae* populations are transformed and grow in different antibiotic concentrations to later be sequenced. Using the tools presented in Chapter 5, we extracted insertions profiles with clear enrichments of insertions, and read count associated, with in-frame positions compared to conventional transposon methodologies.

Based on essentiality estimation approaches, we define a statistical method for the identification of ORFs being translated. Being conservative, this approach can recover 75.2% of *M. pneumoniae* proteome (comprising 66% of annotated SEPs in this bacterium). On the other hand, proteomics can report up to 81% of the proteome. However, 70 proteins, including 11 unstable or protease targets, can be detected with this technique and not by MS. Moreover, this technique highlights 153 new SEPs candidates and 5 larger proteins. Remarkably, we found an intersection of 32 SEPs predicted in Chapter 4 (n=144) and this technique (n=153) in *M. pneumoniae*. This suggests that each methodology could be retrieving SEPs from different essentiality categories, which makes sense considering the importance of homology in RanSEPs searches (probably E genes), and the rationale of this approach where insertions will occur more often in F and NE genes. Similarly, E genes detected by MS are less likely to be detected with this approach, but ProTInSeq also detects 70 F and NE genes which are not detected in free-label MS searches.

Out of the 153 new SEPs (40 ± 20 aa), we found a striking variety of genomic contexts. While 46% of these proteins were located in intergenic regions, we also found examples of SEPs overlapping annotated genes (24%), non-coding RNAs (15%), and also potential upstream regulatory smORFs (6%), or a combination of these (9%). Also, 39 of these 153 SEPs (25%), presented transmembrane segments, membrane-associated features, or signal peptide predicted. This

percentage is the same as for the results predicted by RanSEPs (although only 10 SEPs are shared in these categories between the two approaches). This 25% is close to the results presented in recent large scale genomic comparative studies which report that up to 30% of the SEPs found in >4,000 conserved SEP families in 1,773 human-associated metagenomes are predicted to be secreted or transmembrane [121]. Also from this study, it is reported that 90% of the small protein families have no known domain and/or function. We saw similar results when evaluating the potential functions performed by the predicted SEPs, as we only found significant results by conservation studies for 21 SEPs (14%) while the remaining 132 (86%) did not retrieve any significant hit.

Remarkably, we also found 34 of these new SEPs (22%) presenting a Shine-Dalgarno-like sequence. This is in line with the RBS inclusion rate in *M. pneumoniae* for annotated genes (26.5%). This simplified translation regulation, together with the finding of very small coverage for the negative selection marker (Barnase), suggests that there could be many translation events all over the genome of an organism. These translational noise events would result in a few copies per cell of the translation product which prevents detecting them with a positive selection marker like an antibiotic. However, these products could evolve randomly with time and be a reservoir for new proteins and functions which once they are favorable will be selected for increased expression.

8.3.2. Determinants of ProTInSeq signal: protein abundances and transmembrane topology

Furthermore, we demonstrated that the coverage of a gene with ProTInSeq libraries is not only dependent on essentiality, but it also recapitulates protein abundance and membrane topology segments in proteins. Relative to quantification, as the mutated selection marker can only be expressed in fusion to the ORF it is inserted, we expect that its expression will be dependent on the protein abundance of that ORF. This is validated by showing that higher abundance proteins in *M. pneumoniae* maintain higher coverages along the different concentrations tested. This is explained by the fact they provide a fitness advantage to the cell as the resistance protein will be also more abundant. This is satisfied with the three essentiality categories, indicating that E genes can also be studied in this way.

Also, we saw that insertions of Barnase gene in NE genes were accepted by the cell but only when the NE protein product of those is not expressed or it is in very low abundance (<2 copies/cell). Further exploration of these libraries could provide more insights into the relation of essentiality with protein abundances, as in Chapter 7, where we show that the essentiality of metabolic enzymes is

conditioned by the directions of fluxes in the network comprising them. Also, as an interesting application of this approach, ProTInSeq libraries could be obtained in different media types and stress conditions exploring how the signal could change, as it has been reported SEPs can be expressed under very specific conditions [489,490].

Finally, we also demonstrate that in transmembrane or membrane-associated proteins, fusion occurs preferentially in cytoplasmic segments. We interpreted that this was due to the marker being exposed and thus not conferring resistance to the cell. We demonstrate this effect can be taken as an advantage for the characterization of transmembrane segments applying similar approaches to those used in defining extended NE N' and C'-termini regions of proteins in Chapter 5. Results of running this analysis in 101 NE membrane-related proteins, compared to predictions of transmembrane segments, shows an agreement in 41 proteins. This indicates that this approach could be improved to experimentally identify transmembrane domains. Moreover, when no agreement was found between the prediction and our estimate, we detected that in some cases our approach could be more feasible. For example, in the case of MPN359, where our approach identified a segment as cytoplasmic that was also supported by alternative topology prediction servers.

In conclusion, we provide an experimental alternative to ribosome profiling and proteomic studies for the identification of SEPs in bacterial genomes. Additionally, it allows exploring essentiality, protein abundance, membrane topology, and unstable or fastly degraded proteins.

8.4. Further perspectives

a) Translational noise, frameshifting, and overlapping

It is known that expression of antisense non-coding RNAs in *M. pneumoniae* is produced mainly by spurious transcription processes because of the lack of a -35 element at its promoters, and these being frequently found in low GC content genomes [286]. In a similar way, the low inclusion of ribosome binding sites in *M. pneumoniae* transcripts (mainly found in genes inside an operon and not at the first gene of a transcript), together with the higher chance to find start codons in low GC content genomes, suggest that significant translational noise could be happening in this bacterium as this process seems to be only dependent on the existence of an AUG codon at the 5'UTR for the first gene of a transcript [173,313]. This is supported by the fact that we detect a higher abundance of smORFs for species with lower GC content in Chapter 4, and also by the low coverage reported by the anti-selection with Barnase.

Regarding genes inside an operon, in many cases we find an overlap of the STOP codon and the initiation codon of the contiguous gene, indicating that the ribosome can reinitialize translation by going back one base. In other cases, we also see a more extensive overlap between contiguous genes which does not seem to prevent the expression of those proteins [151]. However, in some cases, we could detect a bona fide RBS at the right distance of the translation initiation codon for internal genes in an operon (the best example is the large ribosomal operon). This raises three questions: i) what determines an overlapping start codon to be recognized in an mRNA full of bound ribosomes translating the first ORF found?; ii) if translational noise occurs, what is the impact on the cell metabolism knowing the demanding energy process of protein synthesis, and degradation of spurious peptides?; iii) does this translational noise make *M. pneumoniae* a 'slow-growing' bacteria? Independently on the number of smORFs that could be coding for SEPs in *M. pneumoniae*, further study on the context of the signals found with ProTInSeq libraries could provide insights in these mechanisms; or on ribosomal frameshifting, as this process has been seen associated with expression of smORFs [124,125].

b) Regulatory upstream smORFs

Another open question refers to the frequency of upstream regulatory smORFs that could regulate expression of downstream ORFs but not encoding for functional SEPs [90], which have been described to be as short as just two codons in bacteria [96]. Although regulatory smORFs have not been reported in *M. pneumoniae*, we cannot discard that ProTInSeq is detecting some of these (they are translated in a conventional manner). As potential candidates to perform this role have been found in these libraries, validation of their regulatory role is required. This could be done by targeted deletion of these smORFs, which are mostly non-essential, and evaluating the expression of the downstream ORF compared to wild-type conditions. In the case some could be validated, the features of the mRNA sequence (e.g. secondary structures) could provide valuable insights in defining approaches for the computational identification of this family of ORFs. Remarkably, the study of these smORFs could have applications in understanding translational regulation control and, ultimately, providing a new way to regulate expression with synthetic biology interests. For example, in plants, pathogen-responsive smORFs have been shown to make plants resistant to diseases, or improve biosynthetic pathways [94].

c) Annotation-free and alternative analysis approaches

Moreover, as possible analysis alternatives, our searches consider ORFs starting with initiation codons found in *M. pneumoniae* (i.e. AUG, GUG, UUG). Annotation-free approaches could reveal enriched signals from considered off-frame positions that could be associated with SEPs encoded by alternative codons, as seen in other species [87,132,133,490]. Also, this could highlight examples of dual-coding genes (one mRNA; two or more encoded proteins in the same frame), which despite being infrequent in bacteria, representative SEP examples can be detected; such as the gene *sprG1*, in *Staphylococcus aureus*, which encodes for two versions of the same SEP (44 and 31 amino acids). While both versions of SprG1 are secreted to lyse human host erythrocytes, the longer version performs this task more efficiently [126]. Although it is not known how common this could be, this could also suggest that one SEP with antimicrobial and/or pathogenic roles could have different ranges of action depending on their size. From an evolutionary perspective, these cases are of great interest as they raise the question of which activity evolved first. From a synthetic biology perspective, it opens the possibility to imagine novel therapies, or bioremediation solutions, as the expression of SEPs with a different activity from the same transcript could be modulated depending on the requirements of an environment.

In addition, in this work, we determine multiple factors contributing to the coverage of a gene, including essentiality, abundance, membrane topology, in addition to those presented in Chapter 5. These could be integrated, together with experimental data, such as ribosome profiling or mass spectrometry, to perform more sophisticated identification studies considering all the factors in the estimation (e.g. a machine learning classifier).

d) Scaling these tools to eukaryotes

In theory, the tools presented could be adapted to eukaryotes by considering the possible alternative splicing forms in the estimations. Regarding the ProTInSeq technique, we have shown that we can efficiently recover insertions in cancer cells (Chapter 7); therefore, the application of this protocol in haploid cell lines could also highlight non-annotated proteins and exons. However, coverage is an important determinant, and it is intrinsically related to the genome size being studied. Interestingly, 102,960 unique insertions in genes can be retrieved with vectors based on the *Drosophila hydei* transposable element *Minos [491]*. This indicates that adaptation of the tools presented here could be feasible in larger genomes. To do this and since the coverage is smaller, it would be required to ensure unique transposition events as well as minimizing sequencing errors. This would be of great interest as SEPs in eukaryotes have been identified to regulate development, stress responses, muscle contraction, or be related to mental health disorders [127–131].

8.5. Concluding remarks

SEPs have been overlooked by computational and experimental approaches and their identification has mostly relied on serendipity. This implies a gap of knowledge in current reference genomes, which makes genome complexity to be underestimated. In this thesis project, we proposed novel bioinformatic and experimental solutions specifically designed for the identification of SEPs in bacterial genomes. In addition, standards required to define new methodologies based on transposon sequencing are developed to overcome the limitations of previous studies. As shown, this propitiated their application in a wide variety of contexts, from bacterial genome reduction to cancer studies.

From the perspective of understanding *M. pneumoniae* biology, the comprehensive study of its proteome is paramount in completing one of the most well-characterized models in Systems Biology. Thus, the proper annotation of all the proteins contained in its genome is a requirement to develop future rational engineering applications, such as drug delivery systems in the lung. Also, the proposed transposon sequencing analysis standards are envisioned to aid future essentiality studies, like those required to rationally design bacterial ‘chassis’.

Ultimately, the proposed computational and experimental methodologies lay the foundation of future studies in the search for bioactive small proteins in available and new genome projects. Considering the roles small proteins can play, from homeostasis regulation to antimicrobial capacities; we envision that future findings in the field of small proteins will be of great impact in areas such as rational genetic engineering, or microbial therapies.

References

1. Roll-Hansen N. The Genotype Theory of Wilhelm Johannsen and its Relation to Plant Breeding and the Study of Evolution. Centaurus. 1979. pp. 201–235. doi:10.1111/j.1600-0498.1979.tb00589.x

2. Griffith F. The Significance of Pneumococcal Types. Journal of Hygiene. 1928. pp. 113–159. doi:10.1017/s0022172400031879

3. Avery Ot, Macleod Cm, Mccarty M. Studies On The Chemical Nature Of The Substance Inducing Transformation Of Pneumococcal Types : Induction Of Transformation By A Desoxyribonucleic Acid Fraction Isolated From Pneumococcus Type Iii. J Exp Med. 1944;79: 137–158.

4. Hershey AD, Chase M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. J Gen Physiol. 1952;36: 39–56.

5. Franklin RE, Gosling RG. Molecular configuration in sodium thymonucleate. Nature. 1953;171: 740–741.

6. Hunter GK. Phoebus Levene and the Tetranucleotide Structure of Nucleic Acids. Ambix. 1999. pp. 73–103. doi:10.1179/amb.1999.46.2.73

7. Watson JD, Crick FH. Genetical implications of the structure of deoxyribonucleic acid. Nature. 1953;171: 964–967.

8. Baltimore D. Viral RNA-dependent DNA Polymerase: RNA-dependent DNA Polymerase in Virions of RNA Tumour Viruses. Nature. 1970. pp. 1209–1211. doi:10.1038/2261209a0

9. August JT, Shapiro L, Eoyang L. Replication of RNA viruses. Journal of Molecular Biology. 1965. pp. 257–271. doi:10.1016/s0022-2836(65)80056-0

10. Gaiti F, Calcino AD, Tanurdžić M, Degnan BM. Origin and evolution of the metazoan non-coding regulatory genome. Dev Biol. 2017;427: 193–202.

11. Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, et al. What is a gene, post-ENCODE? History and updated definition. Genome Res. 2007;17: 669–681.

12. Soll D, Ohtsuka E, Jones DS, Lohrmann R, Hayatsu H, Nishimura S, et al. Studies on polynucleotides, XLIX. Stimulation of the binding of aminoacyl-sRNA's to ribosomes by ribotrinucleotides and a survey of codon assignments for 20 amino acids. Proceedings of the National Academy of Sciences. 1965. pp. 1378–1385. doi:10.1073/pnas.54.5.1378

13. Nirenberg M, Leder P, Bernfield M, Brimacombe R, Trupin J, Rottman F, et al. RNA codewords and protein synthesis, VII. On the general nature of the RNA code. Proc Natl Acad Sci U S A. 1965;53: 1161–1168.

14. Lobanov AV, Turanov AA, Hatfield DL, Gladyshev VN. Dual functions of codons in the genetic code. Crit Rev Biochem Mol Biol. 2010;45: 257–265.

15. Castro-Chavez F. Most Used Codons per Amino Acid and per Genome in the Code of Man Compared to Other Organisms According to the Rotating Circular Genetic Code. NeuroQuantology. 2011. doi:10.14704/nq.2011.9.4.500

16. Sacerdot C, Fayat G, Dessen P, Springer M, Plumbridge JA, Grunberg-Manago M, et al. Sequence of a 1.26-kb DNA fragment containing the structural gene for E.coli initiation factor IF3: presence of an AUU initiator codon. EMBO J. 1982;1: 311–315.

17. Chen Y-PP. Bioinformatics Technologies. Springer Science & Business Media; 2005.

18. Snyder LAS. Bacterial Genetics and Genomics. Garland Science; 2020.

19. Fiers W, Contreras R, De Wachter R, Haegeman G, Merregaert J, Jou WM, et al. Recent progress in the sequence determination of bacteriophage MS2 RNA. Biochimie. 1971;53: 495–506.

20. Fiers W, Contreras R, Duerinck F, Haegeman G, Iserentant D, Merregaert J, et al. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature. 1976;260: 500–507.

21. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. 1977. Biotechnology. 1992;24: 104–108.

22. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science. 1995;269: 496–512.

23. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, et al. Life with 6000 genes. Science. 1996;274: 546, 563–7.

24. International Human Genome Sequencing Consortium Publishes Sequence and Analysis of the Human Genome. PsycEXTRA Dataset. 2001. doi:10.1037/e634052007-001

25. Pennisi E. The Human Genome. Science. 2001. pp. 1177–1180. doi:10.1126/science.291.5507.1177

26. Ezkurdia I, Juan D, Rodriguez JM, Frankish A, Diekhans M, Harrow J, et al. Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. Human Molecular Genetics. 2014. pp. 5866–5878. doi:10.1093/hmg/ddu309

27. Chiu KP. Next-Generation Sequencing and Sequence Data Analysis. Bentham Science Publishers; 2015.

28. Malcolm S. Genotype to phenotype: interpretation of the Human Genome Project. Genotype to Phenotype. 2003. pp. 1–12. doi:10.4324/9780203450420-1

29. Scalzitti N, Jeannin-Girardon A, Collet P, Poch O, Thompson JD. A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms. BMC Genomics. 2020;21: 293.

30. Atkinson P, Glasner P, Lock M. The Handbook of Genetics & Society: Mapping the New Genomic Era. Routledge; 2009.

31. Schad E, Tompa P, Hegyi H. The relationship between proteome size, structural disorder and organism complexity. Genome Biology. 2011. p. R120. doi:10.1186/gb-2011-12-12-r120

32. Pellicer J, Fay MF, Leitch IJ. The largest eukaryotic genome of them all? Botanical Journal of the Linnean Society. 2010. pp. 10–15. doi:10.1111/j.1095-8339.2010.01072.x

33. Price HJ, James Price H. DNA Content Variation among Higher Plants. Annals of the Missouri Botanical Garden. 1988. p. 1248. doi:10.2307/2399283

34. van Ham RCHJ, Kamerbeek J, Palacios C, Rausell C, Abascal F, Bastolla U, et al. Reductive genome evolution in Buchnera aphidicola. Proceedings of the National Academy of Sciences. 2003. pp. 581–586. doi:10.1073/pnas.0235981100

35. Fadiel A. Mycoplasma genomics: tailoring the genome for minimal life requirements through reductive evolution. Frontiers in Bioscience. 2007. p. 2020. doi:10.2741/2207

36. Cai M, Liu Z, Chen M, Zhang M, Jiao Y, Chen Q, et al. Comparative proteomic analysis of senescence in the freshwater cladoceran Daphnia pulex. Comp Biochem Physiol B Biochem Mol Biol. 2020;239: 110352.

37. Consortium TEP, The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science. 2004. pp. 636–640. doi:10.1126/science.1105136

38. Livny J, Waldor MK. Identification of small RNAs in diverse bacterial species. Curr Opin Microbiol. 2007;10: 96–101.

39. Thomason MK, Storz G. Bacterial antisense RNAs: how many are there, and what are they doing? Annu Rev Genet. 2010;44: 167–188.

40. Trussart M, Yus E, Martinez S, Baù D, Tahara YO, Pengo T, et al. Defined chromosome structure in the genome-reduced bacterium Mycoplasma pneumoniae. Nat Commun. 2017;8: 14665.

41. Pontarotti P. Evolutionary Biology: Biodiversification from Genotype to Phenotype. Springer; 2015.

42. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, et al. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol. 2006;2: 2006.0008.

43. Chao MC, Abel S, Davis BM, Waldor MK. The design and analysis of transposon insertion sequencing experiments. Nat Rev Microbiol. 2016;14: 119–128.

44. Weinhold N, Sander O, Domingues FS, Lengauer T, Sommer I. Local function conservation in sequence and structure space. PLoS Comput Biol. 2008;4: e1000105.

45. Charlebois RL, Doolittle WF. Computing prokaryotic gene ubiquity: rescuing the core from extinction. Genome Res. 2004;14: 2469–2477.

46. Badrinarayanan A, Le TBK, Laub MT. Bacterial chromosome organization and segregation. Annu Rev Cell Dev Biol. 2015;31: 171–199.

47. Yao NY, O'Donnell M. SnapShot: The replisome. Cell. 2010;141: 1088, 1088.e1.

48. Jacob F, Monod J. Genetic Regulatory Mechanisms in the Synthesis of Proteins. Molecular Biology. 1989. pp. 82–120. doi:10.1016/b978-0-12-131200-8.50010-1

49. Moreno-Hagelsieb G, Treviño V, Pérez-Rueda E, Smith TF, Collado-Vides J. Transcription unit conservation in the three domains of life: a perspective from Escherichia coli. Trends Genet. 2001;17: 175–177.

50. Grossman AD, Straus DB, Walter WA, Gross CA. Sigma 32 synthesis can regulate the synthesis of heat shock proteins in Escherichia coli. Genes Dev. 1987;1: 179–184.

51. McCann MP, Kidwell JP, Matin A. The putative sigma factor KatF has a central role in development of starvation-mediated general resistance in Escherichia coli. Journal of Bacteriology. 1991. pp. 4188–4194. doi:10.1128/jb.173.13.4188-4194.1991

52. Güell M, Yus E, Lluch-Senar M, Serrano L. Bacterial transcriptomics: what is beyond the RNA horiz-ome? Nat Rev Microbiol. 2011;9: 658–669.

53. Niki H, Yamaichi Y, Hiraga S. Dynamic organization of chromosomal DNA in Escherichia coli. Genes Dev. 2000;14: 212–223.

54. Gaal T, Bartlett MS, Ross W, Turnbough CL Jr, Gourse RL. Transcription regulation by initiating NTP concentration: rRNA synthesis in bacteria. Science. 1997;278: 2092–2097.

55. Sojka L, Kouba T, Barvík I, Sanderová H, Maderová Z, Jonák J, et al. Rapid changes in gene expression: DNA determinants of promoter regulation by the concentration of the transcription initiating NTP in Bacillus subtilis. Nucleic Acids Res. 2011;39: 4598–4611.

56. Dar D, Shamir M, Mellin JR, Koutero M, Stern-Ginossar N, Cossart P, et al. Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria. Science. 2016;352: aad9822.

57. Junier I, Unal EB, Yus E, Lloréns-Rico V, Serrano L. Insights into the Mechanisms of Basal Coordination of Transcription Using a Genome-Reduced Bacterium. Cell Syst. 2016;2: 391–401.

58. Aravind L, Koonin EV. A Natural Classification of Ribonucleases. Methods in Enzymology. 2001. pp. 3–28. doi:10.1016/s0076-6879(01)41142-6

59. Mandin P, Guillier M. Expanding control in bacteria: interplay between small RNAs and transcriptional regulators to control gene expression. Current Opinion in Microbiology. 2013. pp. 125–132. doi:10.1016/j.mib.2012.12.005

60. Kliemt J, Soppa J. Diverse Functions of Small RNAs (sRNAs) in Halophilic Archaea: From Non-coding Regulatory sRNAs to Microprotein-Encoding sRNAs. RNA Metabolism and Gene Expression in Archaea. 2017. pp. 225–242. doi:10.1007/978-3-319-65795-0_10

61. Kaeberlein M, Kennedy BK. Protein translation, 2007. Aging Cell. 2007. pp. 731–734. doi:10.1111/j.1474-9726.2007.00341.x

62. Spirin AS. Ribosomes. Springer Science & Business Media; 2006.

63. Peacock JR, Walvoord RR, Chang AY, Kozlowski MC, Gamper H, Hou Y-M. Amino acid-dependent stability of the acyl linkage in aminoacyl-tRNA. RNA. 2014. pp. 758–764. doi:10.1261/rna.044123.113

64. Scolnick E, Tompkins R, Caskey T, Nirenberg M. Release factors differing in specificity for terminator codons. Proc Natl Acad Sci U S A. 1968;61: 768–774.

65. Buskirk AR, Green R. Ribosome pausing, arrest and rescue in bacteria and eukaryotes. Philos Trans R Soc Lond B Biol Sci. 2017;372. doi:10.1098/rstb.2016.0183

66. Lynch M, Marinov GK. The bioenergetic costs of a gene. Proc Natl Acad Sci U S A. 2015;112: 15690–15695.

67. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. Nature Microbiology. 2016. doi:10.1038/nmicrobiol.2016.48

68. Pauling L, Corey RB, Branson HR. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci U S A. 1951;37: 205–211.

69. Mayhew M, da Silva AC, Martin J, Erdjument-Bromage H, Tempst P, Hartl FU. Protein folding in the central cavity of the GroEL-GroES chaperonin complex. Nature. 1996;379: 420–426.

70. Hayer-Hartl M, Bracher A, Ulrich Hartl F. The GroEL–GroES Chaperonin Machine: A Nano-Cage for Protein Folding. Trends in Biochemical Sciences. 2016. pp. 62–76. doi:10.1016/j.tibs.2015.07.009

71. Rawlings ND. MEROPS: the peptidase database. Nucleic Acids Research. 2000. pp. 323–325. doi:10.1093/nar/28.1.323

72. Olivares AO, Baker TA, Sauer RT. Mechanistic insights into bacterial AAA+ proteases and protein-remodelling machines. Nat Rev Microbiol. 2016;14: 33–44.

73. Gille C, Goede A, Schlöetelburg C, Preissner R, Kloetzel PM, Göbel UB, et al. A comprehensive view on proteasomal sequences: implications for the evolution of the proteasome. J Mol Biol. 2003;326: 1437–1448.

74. Burgos R, Weber M, Martinez S, Lluch-Senar M, Serrano L. Protein quality control and regulated proteolysis in the genome-reduced organism Mycoplasma pneumoniae. Mol Syst Biol. 2020;16: e9530.

75. Imanishi I, Nicolas A, Caetano A-CB, de Paula Castro TL, Tartaglia NR, Mariutti R, et al. Exfoliative toxin E, a new Staphylococcus aureus virulence factor with host-specific activity. Scientific Reports. 2019. doi:10.1038/s41598-019-52777-3

76. Schomburg I, Chang A, Placzek S, Söhngen C, Rother M, Lang M, et al. BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. Nucleic Acids Res. 2013;41: D764–72.

77. Chubukov V, Gerosa L, Kochanowski K, Sauer U. Coordination of microbial metabolism. Nat Rev Microbiol. 2014;12: 327–340.

78. Nath S, Villadsen J. Oxidative phosphorylation revisited. Biotechnol Bioeng. 2015;112: 429–437.

79. Elofsson A, von Heijne G. Membrane protein structure: prediction versus reality. Annu Rev Biochem. 2007;76: 125–140.

80. Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS. Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing. Cell. 2012. pp. 1607–1621. doi:10.1016/j.cell.2012.04.012

81. Maffei B, Francetic O, Subtil A. Tracking Proteins Secreted by Bacteria: What's in the Toolbox? Frontiers in Cellular and Infection Microbiology. 2017. doi:10.3389/fcimb.2017.00221

82. Herrmann H, Bär H, Kreplak L, Strelkov SV, Aebi U. Intermediate filaments: from cell architecture to nanomechanics. Nature Reviews Molecular Cell Biology. 2007. pp. 562–573. doi:10.1038/nrm2197

83. Sleigh MA. THE MOVEMENT OF CILIA AND FLAGELLA. The Biology of Cilia and Flagella. 1962. pp. 127–169. doi:10.1016/b978-1-4831-9772-2.50013-0

84. Sigrist CJA, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, et al. New and continuing developments at PROSITE. Nucleic Acids Res. 2013;41: D344–7.

85. Szabadka Z, Grolmusz V. Building a Structured PDB: The RS-PDB Database. 2006 International Conference of the IEEE Engineering in Medicine and Biology Society. 2006. doi:10.1109/iembs.2006.259331

86. Holst JJ. The Physiology of Glucagon-like Peptide 1. Physiological Reviews. 2007. pp. 1409–1439. doi:10.1152/physrev.00034.2006

87. Saghatelian A, Couso JP. Discovery and characterization of smORF-encoded bioactive polypeptides. Nat Chem Biol. 2015;11: 909–916.

88. Hemm MR, Weaver J, Storz G. Escherichia coli Small Proteome. EcoSal Plus. 2020. doi:10.1128/ecosalplus.esp-0031-2019

89. VanOrsdel CE, Kelly JP, Burke BN, Lein CD, Oufiero CE, Sanchez JF, et al. Identifying New Small Proteins in Escherichia coli. Proteomics. 2018;18: e1700064.

90. Occhi G, Regazzo D, Trivellin G, Boaretto F, Ciato D, Bobisse S, et al. A novel mutation in the upstream open reading frame of the CDKN1B gene causes a MEN4 phenotype. PLoS Genet. 2013;9: e1003350.

91. Lee S, Liu B, Lee S, Huang S-X, Shen B, Qian S-B. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. Proceedings of the National Academy of Sciences. 2012. pp. E2424–E2432. doi:10.1073/pnas.1207846109

92. Babitzke P, Gollnick P. Attenuation Of Transcription. Encyclopedia of Molecular Biology. 2002. doi:10.1002/047120918x.emb0118

93. Zhang H, Dou S, He F, Luo J, Wei L, Lu J. Genome-wide maps of ribosomal occupancy provide insights into adaptive evolution and regulatory roles of uORFs during Drosophila development. PLoS Biol. 2018;16: e2003903.

94. Lorenzo-Orts L, Witthoeft J, Deforges J, Martinez J, Loubéry S, Placzek A, et al. Concerted expression of a cell cycle regulator and a metabolic enzyme from a bicistronic transcript in plants. Nature Plants. 2019. pp. 184–193. doi:10.1038/s41477-019-0358-3

95. Orr MW, Mao Y, Storz G, Qian S-B. Alternative ORFs and small ORFs: shedding light on the dark proteome. Nucleic Acids Res. 2020;48: 1029–1042.

96. Meydan S, Marks J, Klepacki D, Sharma V, Baranov PV, Firth AE, et al. Retapamulin-Assisted Ribosome Profiling Reveals the Alternative Bacterial Proteome. Mol Cell. 2019;74: 481–493.e6.

97. Duval M, Cossart P. Small bacterial and phagic proteins: an updated view on a rapidly moving field. Curr Opin Microbiol. 2017;39: 81–88.

98. Sandman K, Losick R, Youngman P. Genetic analysis of Bacillus subtilis spo mutations generated by Tn917-mediated insertional mutagenesis. Genetics. 1987;117: 603–617.

99. Ebmeier SE, Tan IS, Clapham KR, Ramamurthi KS. Small proteins link coat and cortex assembly during sporulation in Bacillus subtilis. Mol Microbiol. 2012;84: 682–696.

100. Wadler CS, Vanderpool CK. A dual function for a bacterial small RNA: SgrS performs base pairing-dependent regulation and encodes a functional polypeptide. Proc Natl Acad Sci U S A. 2007;104: 20454–20459.

101. Chukwudi CU, Good L. The role of the hok/sok locus in bacterial response to stressful growth conditions. Microbial Pathogenesis. 2015. pp. 70–79. doi:10.1016/j.micpath.2015.01.009

102. Unoson C, Wagner EGH. A small SOS-induced toxin is targeted against the inner membrane in Escherichia coli. Mol Microbiol. 2008;70: 258–270.

103. Sayed N, Nonin-Lecomte S, Réty S, Felden B. Functional and structural insights of a Staphylococcus aureus apoptotic-like membrane peptide from a toxin-antitoxin module. J Biol Chem. 2012;287: 43454–43463.

104. Hobbs EC, Yin X, Paul BJ, Astarita JL, Storz G. Conserved small protein associates with the multidrug efflux pump AcrB and differentially affects antibiotic resistance. Proc Natl Acad Sci U S A. 2012;109: 16696–16701.

105. Gannoun-Zaki L, Alibaud L, Carrère-Kremer S, Kremer L, Blanc-Potard A-B. Overexpression of the KdpF membrane peptide in Mycobacterium bovis BCG results in reduced intramacrophage growth and altered cording morphology. PLoS One. 2013;8: e60379.

106. Martin JE, Waters LS, Storz G, Imlay JA. The Escherichia coli Small Protein MntS and Exporter MntP Optimize the Intracellular Concentration of Manganese. PLOS Genetics. 2015. p. e1004977. doi:10.1371/journal.pgen.1004977

107. Handler AA, Lim JE, Losick R. Peptide inhibitor of cytokinesis during sporulation in Bacillus subtilis. Mol Microbiol. 2008;68: 588–599.

108. Modell JW, Hopkins AC, Laub MT. A DNA damage checkpoint in Caulobacter crescentus inhibits cell division through a direct interaction with FtsW. Genes Dev. 2011;25: 1328–1343.

109. Impens F, Rolhion N, Radoshevich L, Bécavin C, Duval M, Mellin J, et al. N-terminomics identifies Prli42 as a membrane miniprotein conserved in Firmicutes and critical for stressosome activation in Listeria monocytogenes. Nat Microbiol. 2017;2: 17005.

110. Wong RS, McMurry LM, Levy SB. "Intergenic" blr gene in Escherichia coli encodes a 41-residue membrane protein affecting intrinsic susceptibility to certain inhibitors of peptidoglycan synthesis. Mol Microbiol. 2000;37: 364–370.

111. Salazar ME, Podgornaia AI, Laub MT. The small membrane protein MgrB regulates PhoQ bifunctionality to control PhoP target gene expression dynamics. Mol Microbiol. 2016;102: 430–445.

112. Baumgartner D, Kopf M, Klähn S, Steglich C, Hess WR. Small proteins in cyanobacteria provide a paradigm for the functional analysis of the bacterial micro-proteome. BMC Microbiol. 2016;16: 285.

113. Gaballa A, Antelmann H, Aguilar C, Khakh SK, Song K-B, Smaldone GT, et al. The Bacillus subtilis iron-sparing response is mediated by a Fur-regulated small RNA and three small, basic proteins. Proc Natl Acad Sci U S A. 2008;105: 11927–11932.

114. Fischer N, Neumann P, Konevega AL, Bock LV, Ficner R, Rodnina MV, et al. Structure of the E. coli ribosome–EF-Tu complex at <3 Å resolution by Cs-corrected cryo-EM. Nature. 2015. pp. 567–570. doi:10.1038/nature14275

115. Rutherford ST, Bassler BL. Bacterial quorum sensing: its role in virulence and possibilities for its control. Cold Spring Harb Perspect Med. 2012;2. doi:10.1101/cshperspect.a012427

116. Quereda JJ, Dussurget O, Nahori M-A, Ghozlane A, Volant S, Dillies M-A, et al. Bacteriocin from epidemic Listeria strains alters the host intestinal microbiota to favor infection. Proc Natl Acad Sci U S A. 2016;113: 5706–5711.

117. Wladyka B, Piejko M, Bzowska M, Pieta P, Krzysik M, Mazurek Ł, et al. A peptide factor secreted by Staphylococcus pseudintermedius exhibits properties of both bacteriocins and virulence factors. Sci Rep. 2015;5: 14569.

118. Cogen AL, Yamasaki K, Muto J, Sanchez KM, Alexander LC, Tanios J, et al. Staphylococcus epidermidis Antimicrobial δ-Toxin (Phenol-Soluble Modulin-γ) Cooperates with Host Antimicrobial Peptides to Kill Group A Streptococcus. PLoS ONE. 2010. p. e8557. doi:10.1371/journal.pone.0008557

119. Molloy EM, Cotter PD, Hill C, Mitchell DA, Ross RP. Streptolysin S-like virulence factors: the continuing sagA. Nat Rev Microbiol. 2011;9: 670–681.

120. Sassone-Corsi M, Nuccio S-P, Liu H, Hernandez D, Vu CT, Takahashi AA, et al. Microcins mediate competition among Enterobacteriaceae in the inflamed gut. Nature. 2016;540: 280–283.

121. Sberro H, Greenfield N, Pavlopoulos G, Kyrpides N, Bhatt AS. Large-scale analyses of human microbiomes reveal thousands of small, novel genes and their predicted functions. doi:10.1101/494179

122. Plaza S, Menschaert G, Payre F. In Search of Lost Small Peptides. Annu Rev Cell Dev Biol. 2017;33: 391–416.

123. Storz G, Wolf YI, Ramamurthi KS. Small proteins can no longer be ignored. Annu Rev Biochem. 2014;83: 753–777.

124. Fellner L, Simon S, Scherling C, Witting M, Schober S, Polte C, et al. Evidence for the recent origin of a bacterial protein-coding, overlapping orphan gene by evolutionary overprinting. BMC Evol Biol. 2015;15: 283.

125. Delaye L, DeLuna A, Lazcano A, Becerra A. The origin of a novel gene through overprinting in Escherichia coli. BMC Evol Biol. 2008;8: 1–10.

126. Pinel-Marie M-L, Brielle R, Felden B. Dual toxic-peptide-coding Staphylococcus aureus RNA under antisense regulation targets host cells and bacterial rivals unequally. Cell Rep. 2014;7: 424–435.

127. Kastenmayer JP. Functional genomics of genes with small open reading frames (sORFs) in S. cerevisiae. Genome Research. 2006. pp. 365–373. doi:10.1101/gr.4355406

128. Galindo MI, Pueyo JI, Fouix S, Bishop SA, Couso JP. Peptides encoded by short ORFs control development and define a new eukaryotic gene family. PLoS Biol. 2007;5: e106.

129. Bal NC, Maurya SK, Sopariwala DH, Sahoo SK, Gupta SC, Shaikh SA, et al. Sarcolipin is a newly identified regulator of muscle-based thermogenesis in mammals. Nat Med. 2012;18: 1575–1579.

130. Schmitt JP, Kamisago M, Asahi M, Li GH, Ahmad F, Mende U, et al. Dilated cardiomyopathy and heart failure caused by a mutation in phospholamban. Science. 2003;299: 1410–1413.

131. Anderson DM, Anderson KM, Chang C-L, Makarewich CA, Nelson BR, McAnally JR, et al. A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. Cell. 2015;160: 595–606.

132. Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, Levin JZ, et al. Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nat Chem Biol. 2013;9: 59–64.

133. Wang B, Hao J, Pan N, Wang Z, Chen Y, Wan C. Identification and analysis of small proteins and short open reading frame encoded peptides in Hep3B cell. J Proteomics. 2021;230: 103965.

134. Guo B, Zhai D, Cabezas E, Welsh K, Nouraini S, Satterthwait AC, et al. Humanin peptide suppresses apoptosis by interfering with Bax activation. Nature. 2003;423: 456–461.

135. Hashimoto Y, Niikura T, Tajima H, Yasukawa T, Sudo H, Ito Y, et al. A rescue factor abolishing neuronal cell death by a wide spectrum of familial Alzheimer's disease genes and Abeta. Proc Natl Acad Sci U S A. 2001;98: 6336–6341.

136. Ji B, Kim M, Higa KK, Zhou X. Boymaw, overexpressed in brains with major psychiatric disorders, may encode a small protein to inhibit mitochondrial function and protein translation. Am J Med Genet B Neuropsychiatr Genet. 2015;168B: 284–295.

137. Ramamurthi KS, Clapham KR, Losick R. Peptide anchoring spore coat assembly to the outer forespore membrane in Bacillus subtilis. Mol Microbiol. 2006;62: 1547–1557.

138. Burkholder WF, Kurtser I, Grossman AD. Replication initiation proteins regulate a developmental checkpoint in Bacillus subtilis. Cell. 2001;104: 269–279.

139. Karimova G, Davi M, Ladant D. The -Lactam Resistance Protein Blr, a Small Membrane Polypeptide, Is a Component of the Escherichia coli Cell Division Machinery. Journal of Bacteriology. 2012. pp. 5576–5588. doi:10.1128/jb.00774-12

140. Kosfeld A, Jahreis K. Characterization of the Interaction Between the Small Regulatory Peptide SgrT and the EIICBGlc of the Glucose-Phosphotransferase System of E. coli K-12. Metabolites. 2012;2: 756–774.

141. VanOrsdel CE, Bhatt S, Allen RJ, Brenner EP, Hobson JJ, Jamil A, et al. The Escherichia coli CydX protein is a member of the CydAB cytochrome bd oxidase complex and is required for cytochrome bd oxidase activity. J Bacteriol. 2013;195: 3640–3650.

142. Kato A, Chen HD, Latifi T, Groisman EA. Reciprocal control between a bacterium's regulatory system and the modification status of its lipopolysaccharide. Mol Cell. 2012;47: 897–908.

143. Choi E, Lee K-Y, Shin D. The MgtR regulatory peptide negatively controls expression of the MgtA Mg2 transporter in Salmonella enterica serovar Typhimurium. Biochemical and Biophysical Research Communications. 2012. pp. 318–323. doi:10.1016/j.bbrc.2011.11.107

144. Stodolsky M. Introduction: Validation methods for function genome annotation. BMC Genomics. 2011;12 Suppl 1: I1.

145. Abril JF, Castellano S. Genome Annotation. Encyclopedia of Bioinformatics and Computational Biology. 2019. pp. 195–209. doi:10.1016/b978-0-12-809633-8.20226-4

146. Furuno M. CDS Annotation in Full-Length cDNA Sequence. Genome Research. 2003. pp. 1478–1487. doi:10.1101/gr.1060303

147. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, et al. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol. 2007;3: 121.

148. Borodovsky M, McIninch J. GENMARK: Parallel gene recognition for both DNA strands. Computers & Chemistry. 1993. pp. 123–133. doi:10.1016/0097-8485(93)85004-v

149. Salzberg SL, Delcher AL, Kasif S, White O. Microbial gene identification using interpolated Markov models. Nucleic Acids Res. 1998;26: 544–548.

150. Delcher AL, Bratke KA, Powers EC, Salzberg SL. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics. 2007;23: 673–679.

151. Fukuda Y, Washio T, Tomita M. Comparative study of overlapping genes in the genomes of Mycoplasma genitalium and Mycoplasma pneumoniae. Nucleic Acids Res. 1999;27: 1847–1853.

152. Cao X, Slavoff SA. Non-AUG start codons: Expanding and regulating the small and alternative ORFeome. Exp Cell Res. 2020;391: 111973.

153. Povolotskaya IS, Kondrashov FA, Ledda A, Vlasov PK. Stop codons in bacteria are not selectively equivalent. Biology Direct. 2012. p. 30. doi:10.1186/1745-6150-7-30

154. Guigo R. Gene Prediction, Ab Initio (Intrinsic Gene Prediction, Template Gene Prediction). Dictionary of Bioinformatics and Computational Biology. 2004. doi:10.1002/0471650129.dob0274

155. Setubal JC, Almeida NF, Wattam AR. Comparative Genomics for Prokaryotes. Comparative Genomics. 2018. pp. 55–78. doi:10.1007/978-1-4939-7463-4_3

156. Gonzales MJ, Dugan JM, Shafer RW. Synonymous-non-synonymous mutation rates between sequences containing ambiguous nucleotides (Syn-SCAN). Bioinformatics. 2002. pp. 886–887. doi:10.1093/bioinformatics/18.6.886

157. Sharp PM, Li WH. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 1987;15: 1281–1295.

158. Besemer J, Lomsadze A, Borodovsky M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res. 2001;29: 2607–2618.

159. Jansen R. Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. Nucleic Acids Research. 2003. pp. 2242–2251. doi:10.1093/nar/gkg306

160. Orengo CA, Bateman A. Protein Families: Relating Protein Sequence, Structure, and Function. John Wiley & Sons; 2014.

161. Bocs S. AMIGene: Annotation of MIcrobial Genes. Nucleic Acids Research. 2003. pp. 3723–3726. doi:10.1093/nar/gkg590

162. Barrett LW, Fletcher S, Wilton SD. Untranslated Gene Regions and Other Non-coding Elements. Untranslated Gene Regions and Other Non-coding Elements. 2013. pp. 1–56. doi:10.1007/978-3-0348-0679-4_1

163. Jishage M, Ishihama A. Regulation of RNA polymerase sigma subunit synthesis in Escherichia coli: intracellular levels of sigma 70 and sigma 38. Journal of bacteriology. 1995. pp. 6832–6835. doi:10.1128/jb.177.23.6832-6835.1995

164. Lloréns-Rico V, Lluch-Senar M, Serrano L. Distinguishing between productive and abortive promoters using a random forest classifier in Mycoplasma pneumoniae. Nucleic Acids Res. 2015;43: 3442–3453.

165. Roberts JW. Mechanisms of Bacterial Transcription Termination. Journal of Molecular Biology. 2019. pp. 4030–4039. doi:10.1016/j.jmb.2019.04.003

166. Bossi L, Figueroa-Bossi N, Bouloc P, Boudvillain M. Regulatory interplay between small RNAs and transcription termination factor Rho. Biochim Biophys Acta Gene Regul Mech. 2020;1863: 194546.

167. Yakhnin AV, Babitzke P. Mechanism of NusG-stimulated pausing, hairpin-dependent pause site selection and intrinsic termination at overlapping pause and termination sites in the Bacillus subtilis trp leader. Molecular Microbiology. 2010. pp. 690–705. doi:10.1111/j.1365-2958.2010.07126.x

168. Gualerzi CO, Pon CL. Initiation of mRNA translation in bacteria: structural and dynamic aspects. Cell Mol Life Sci. 2015;72: 4341–4367.

169. Jin H, Zhao Q, de Valdivia EIG, Ardell DH, Stenstrom M, Isaksson LA. Influences on gene expression in vivo by a Shine-Dalgarno sequence. Molecular Microbiology. 2006. pp. 480–492. doi:10.1111/j.1365-2958.2006.05110.x

170. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature. 2004;428: 37–43.

171. Frishman D, Mironov A, Mewes HW, Gelfand M. Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. Nucleic Acids Res. 1998;26: 2941–2947.

172. Tech M, Pfeifer N, Morgenstern B, Meinicke P. TICO: a tool for improving predictions of prokaryotic translation initiation sites. Bioinformatics. 2005;21: 3568–3569.

173. Omotajo D, Tate T, Cho H, Choudhary M. Distribution and diversity of ribosome binding sites in prokaryotic genomes. BMC Genomics. 2015;16: 604.

174. Loechel S, Inamine JM, Hu PC. A novel translation initiation region from Mycoplasma genitalium that functions in Escherichia coli. Nucleic Acids Res. 1991;19: 6905–6911.

175. Hoff KJ, Stanke M. Current methods for automated annotation of protein-coding genes. Current Opinion in Insect Science. 2015. pp. 8–14. doi:10.1016/j.cois.2015.02.008

176. Basic Local Alignment Search Tool (BLAST). Bioinformatics and Functional Genomics. pp. 100–138. doi:10.1002/9780470451496.ch4

177. Angiuoli SV, Gussman A, Klimke W, Cochrane G, Field D, Garrity G, et al. Toward an online repository of Standard Operating Procedures (SOPs) for (meta)genomic annotation. OMICS. 2008;12: 137–141.

178. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, et al. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Research. 2016. pp. 6614–6624. doi:10.1093/nar/gkw569

179. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, et al. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. 2001;29: 22–28.

180. Klimke W, Agarwala R, Badretdin A, Chetvernin S, Ciufo S, Fedorov B, et al. The National Center for Biotechnology Information's Protein Clusters Database. Nucleic Acids Research. 2009. pp. D216–D223. doi:10.1093/nar/gkn734

181. Van Domselaar GH, Stothard P, Shrivastava S, Cruz JA, Guo A, Dong X, et al. BASys: a web server for automated bacterial genome annotation. Nucleic Acids Res. 2005;33: W455–9.

182. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30: 2068–2069.

183. Lukashin AV, Borodovsky M. GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res. 1998;26: 1107–1115.

184. Larsen TS, Krogh A. EasyGene--a prokaryotic gene finder that ranks ORFs by statistical significance. BMC Bioinformatics. 2003;4: 21.

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

186. Noguchi H. MetaGene: Prediction of Prokaryotic and Phage Genes in Metagenomic Sequences. Handbook of Molecular Microbial Ecology I. 2011. pp. 433–439. doi:10.1002/9781118010518.ch50

187. Hua Z-G, Lin Y, Yuan Y-Z, Yang D-C, Wei W, Guo F-B. ZCURVE 3.0: identify prokaryotic genes with higher accuracy as well as automatically and accurately select essential genes. Nucleic Acids Res. 2015;43: W85–90.

188. Kang Y-J, Yang D-C, Kong L, Hou M, Meng Y-Q, Wei L, et al. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Res. 2017;45: W12–W16.

189. Kaplan A, Haenlein M. Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Business Horizons. 2019. pp. 15–25. doi:10.1016/j.bushor.2018.08.004

190. Michalski RS, Carbonell JG, Mitchell TM. Machine Learning: An Artificial Intelligence Approach. Elsevier; 2014.

191. Moses A. Statistical Modeling and Machine Learning for Molecular Biology. 2017. doi:10.1201/9781315372266

192. Trappenberg TP. Machine learning with sklearn. Fundamentals of Machine Learning. 2019. pp. 38–65. doi:10.1093/oso/9780198828044.003.0003

193. Lin N, Wu B, Jansen R, Gerstein M, Zhao H. Information assessment on predicting protein-protein interactions. BMC Bioinformatics. 2004;5: 154.

194. Wu J, Liu H, Duan X, Ding Y, Wu H, Bai Y, et al. Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. Bioinformatics. 2009;25: 30–35.

195. Liu Z-P, Wu L-Y, Wang Y, Zhang X-S, Chen L. Prediction of protein–RNA binding sites by a random forest method with combined features. Bioinformatics. 2010. pp. 1616–1622. doi:10.1093/bioinformatics/btq253

196. Hu Q, Greene CS, Heller EA. Specific histone modifications associate with alternative exon selection during mammalian development. Nucleic Acids Res. 2020;48: 4709–4724.

197. Supervised Learning. Neural Smithing. 1999. doi:10.7551/mitpress/4937.003.0003

198. Gordon JJ, Towsey MW, Hogan JM, Mathews SA, Timms P. Improved prediction of bacterial transcription start sites. Bioinformatics. 2006;22: 142–148.

199. Berry MW, Mohamed A, Yap BW. Supervised and Unsupervised Learning for Data Science. Springer Nature; 2019.

200. Hartigan JA, Wong MA. Algorithm AS 136: A K-Means Clustering Algorithm. Applied Statistics. 1979. p. 100. doi:10.2307/2346830

201. Gaussian Mixture Models. SpringerReference. doi:10.1007/springerreference_70943

202. Wong K-C, Li Y, Zhang Z. Unsupervised Learning in Genome Informatics. Unsupervised Learning Algorithms. 2016. pp. 405–448. doi:10.1007/978-3-319-24211-8_15

203. Pittis AA, Gabaldón T. Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. Nature. 2016;531: 101–104.

204. Jolliffe IT. Principal Component Analysis. Springer Science & Business Media; 2013.

205. Chiu C-C, Soo V-W. Subgoal Identifications in Reinforcement Learning: A Survey. Advances in Reinforcement Learning. 2011. doi:10.5772/13214

206. Kriegman S, Blackiston D, Levin M, Bongard J. A scalable pipeline for designing reconfigurable organisms. Proc Natl Acad Sci U S A. 2020;117: 1853–1859.

207. Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, et al. A Deep Learning Approach to Antibiotic Discovery. Cell. 2020;181: 475–483.

208. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Improved protein structure prediction using potentials from deep learning. Nature. 2020;577: 706–710.

209. Fritsche-Neto R, Borém A. Omics: Opening up the “Black Box” of the Phenotype. Omics in Plant Breeding. 2014. pp. 1–11. doi:10.1002/9781118820971.ch1

210. Datta S, Nettleton D. Statistical Analysis of Next Generation Sequencing Data. Springer; 2014.

211. Ronaghi M, Uhlén M, Nyrén P. A sequencing method based on real-time pyrophosphate. Science. 1998;281: 363, 365.

212. Ju J, Kim DH, Bi L, Meng Q, Bai X, Li Z, et al. Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. Proc Natl Acad Sci U S A. 2006;103: 19635–19640.

213. Branton D, Deamer D. The Development of Nanopore Sequencing. Nanopore Sequencing. 2019. pp. 1–16. doi:10.1142/9789813270619_0001

214. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10: 57–63.

215. Chu Y, Corey DR. RNA sequencing: platform selection, experimental design, and data interpretation. Nucleic Acid Ther. 2012;22: 271–274.

216. Zhao S, Zhang Y, Gordon W, Quan J, Xi H, Du S, et al. Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. BMC Genomics. 2015;16: 675.

217. Yus E, Güell M, Vivancos AP, Chen W, Lluch-Senar M, Delgado J, et al. Transcription start site associated RNAs in bacteria. Molecular Systems Biology. 2012. p. 585. doi:10.1038/msb.2012.16

218. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009;324: 218–223.

219. Johnstone TG, Bazzini AA, Giraldez AJ. Upstream ORFs are prevalent translational repressors in vertebrates. EMBO J. 2016;35: 706–723.

220. Spealman P, Naik AW, May GE, Kuersten S, Freeberg L, Murphy RF, et al. Conserved non-AUG uORFs revealed by a novel regression analysis of ribosome profiling data. Genome Res. 2018;28: 214–222.

221. Hsu PY, Calviello L, Wu H-YL, Li F-W, Rothfels CJ, Ohler U, et al. Super-resolution ribosome profiling reveals unannotated translation events in. Proc Natl Acad Sci U S A. 2016;113: E7126–E7135.

222. Fritsch C, Herrmann A, Nothnagel M, Szafranski K, Huse K, Schumann F, et al. Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. Genome Res. 2012;22: 2208–2218.

223. Nakahigashi K, Takai Y, Kimura M, Abe N, Nakayashiki T, Shiwa Y, et al. Comprehensive identification of translation start sites by tetracycline-inhibited ribosome profiling. DNA Res. 2016;23: 193–201.

224. Weaver J, Mohammad F, Buskirk AR, Storz G. Identifying Small Proteins by Ribosome Profiling with Stalled Initiation Complexes. MBio. 2019;10. doi:10.1128/mBio.02819-18

225. Santos DA, Shi L, Tu BP, Weissman JS. Cycloheximide can distort measurements of mRNA levels and translation efficiency. Nucleic Acids Res. 2019;47: 4974–4985.

226. Ribosome profiling: a Hi-Def monitor for protein synthesis at the genome-wide scale. Wiley Interdiscip Rev RNA. 2017;8. doi:10.1002/wrna.1434

227. Glaub A, Huptas C, Neuhaus K, Ardern Z. Recommendations for bacterial ribosome profiling experiments based on bioinformatic evaluation of published data. J Biol Chem. 2020;295: 8999–9011.

228. Fremin BJ, Sberro H, Bhatt AS. MetaRibo-Seq measures translation in microbiomes. Nat Commun. 2020;11: 3268.

229. van Opijnen T, Camilli A. Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. Nat Rev Microbiol. 2013;11: 435–442.

230. Mhatre E, Snyder DJ, Sileo E, Turner CB, Buskirk SW, Fernandez NL, et al. One gene, multiple ecological strategies: A biofilm regulator is a capacitor for sustainable diversity. Proc Natl Acad Sci U S A. 2020;117: 21647–21657.

231. Plasterk RH, Izsvák Z, Ivics Z. Resident aliens: the Tc1/mariner superfamily of transposable elements. Trends Genet. 1999;15: 326–332.

232. Goryshin IY, Miller JA, Kil YV, Lanzov VA, Reznikoff WS. Tn5/IS50 target recognition. Proc Natl Acad Sci U S A. 1998;95: 10716–10721.

233. McCarthy AJ, Stabler RA, Taylor PW. Genome-Wide Identification by Transposon Insertion Sequencing of Escherichia coli K1 Genes Essential for Growth, Gastrointestinal Colonizing Capacity, and Survival in Serum. J Bacteriol. 2018;200. doi:10.1128/JB.00698-17

234. Lluch-Senar M, Delgado J, Chen W-H, Lloréns-Rico V, O'Reilly FJ, Wodke JA, et al. Defining a minimal cell: essentiality of small ORFs and ncRNAs in a genome-reduced bacterium. Mol Syst Biol. 2015;11: 780.

235. Sharma S, Markham PF, Browning GF. Genes found essential in other mycoplasmas are dispensable in Mycoplasma bovis. PLoS One. 2014;9: e97100.

236. Glass JI, Assad-Garcia N, Alperovich N, Yooseph S, Lewis MR, Maruf M, et al. Essential genes of a minimal bacterium. Proceedings of the National Academy of Sciences. 2006. pp. 425–430. doi:10.1073/pnas.0510013103

237. Cain AK, Barquist L, Goodman AL, Paulsen IT, Parkhill J, van Opijnen T. A decade of advances in transposon-insertion sequencing. Nat Rev Genet. 2020;21: 526–540.

238. Hentchel KL, Reyes Ruiz LM, Curtis PD, Fiebig A, Coleman ML, Crosson S. Genome-scale fitness profile of Caulobacter crescentus grown in natural freshwater. ISME J. 2019;13: 523–536.

239. DeJesus MA, Ioerger TR. Normalization of transposon-mutant library sequencing datasets to improve identification of conditionally essential genes. J Bioinform Comput Biol. 2016;14: 1642004.

240. McCoy KM, Antonio ML, van Opijnen T. MAGenTA: a Galaxy implemented tool for complete Tn-Seq analysis and data visualization. Bioinformatics. 2017. pp. 2781–2783. doi:10.1093/bioinformatics/btx320

241. DeJesus MA, Ambadipudi C, Baker R, Sassetti C, Ioerger TR. TRANSIT--A Software Tool for Himar1 TnSeq Analysis. PLoS Comput Biol. 2015;11: e1004401.

242. Veeranagouda Y, Didier M. Transposon Insertion Site Sequencing (TIS-Seq): An Efficient and High-Throughput Method for Determining Transposon Insertion Site(s) and Their Relative Abundances in a PiggyBac Transposon Mutant Pool by Next-Generation Sequencing. Current Protocols in Molecular Biology. 2017. pp. 21.35.1–21.35.11.

243. Maher S, Jjunju FPM, Taylor S. Colloquium: 100 years of mass spectrometry: Perspectives and future trends. Reviews of Modern Physics. 2015. pp. 113–135. doi:10.1103/revmodphys.87.113

244. Alves P, Arnold RJ, Clemmer DE, Li Y, Reilly JP, Sheng Q, et al. Fast and accurate identification of semi-tryptic peptides in shotgun proteomics. Bioinformatics. 2008;24: 102–109.

245. Alves P, Arnold RJ, Novotny MV, Radivojac P, Reilly JP, Tang H. Advancement in protein inference from shotgun proteomics using peptide detectability. Pac Symp Biocomput. 2007; 409–420.

246. Rajoria S, Sabna S, Babele P, Kumar RB, Kamboj DV, Kumar S, et al. Elucidation of protein biomarkers for verification of selected biological warfare agents using tandem mass spectrometry. Sci Rep. 2020;10: 2205.

247. Mesuere B, Van der Jeugt F, Devreese B, Vandamme P, Dawyndt P. The unique peptidome: Taxon-specific tryptic peptides as biomarkers for targeted metaproteomics. Proteomics. 2016;16: 2313–2318.

248. Müller SA, Kohajda T, Findeiss S, Stadler PF, Washietl S, Kellis M, et al. Optimization of parameters for coverage of low molecular weight proteins. Anal Bioanal Chem. 2010;398: 2867–2881.

249. DI Girolamo F, Ponzi M, Crescenzi M, Alessandroni J, Guadagni F. A simple and effective method to analyze membrane proteins by SDS-PAGE and MALDI mass spectrometry. Anticancer Res. 2010;30: 1121–1129.

250. Ong S-E, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, et al. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteomics. 2002;1: 376–386.

251. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol. 1999;17: 994–999.

252. Gerber SA, Rush J, Stemman O, Kirschner MW, Gygi SP. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. Proc Natl Acad Sci U S A. 2003;100: 6940–6945.

253. Friedman RC, Kalkhof S, Doppelt-Azeroual O, Mueller SA, Chovancová M, von Bergen M, et al. Common and phylogenetically widespread coding for peptides by bacterial small RNAs. BMC Genomics. 2017;18: 553.

254. Uhlén M, Nilsson B, Guss B, Lindberg M, Gatenbeck S, Philipson L. Gene fusion vectors based on the gene for staphylococcal protein A. Gene. 1983. pp. 369–378. doi:10.1016/0378-1119(83)90025-2

255. Schmidt TGM, Skerra A. The Strep-tag system for one-step purification and high-affinity detection or capturing of proteins. Nature Protocols. 2007. pp. 1528–1535. doi:10.1038/nprot.2007.209

256. Grisshammer R, Averbeck P, Sohal AK. Improved purification of a rat neurotensin receptor expressed in Escherichia coli. Biochem Soc Trans. 1999;27: 899–903.

257. Hemm MR, Paul BJ, Schneider TD, Storz G, Rudd KE. Small membrane proteins found by comparative genomics and ribosome binding site models. Mol Microbiol. 2008;70: 1487–1501.

258. Hemm MR, Paul BJ, Miranda-Ríos J, Zhang A, Soltanzad N, Storz G. Small stress response proteins in Escherichia coli: proteins missed by classical proteomic studies. J Bacteriol. 2010;192: 46–58.

259. Prelich G. Gene overexpression: uses, mechanisms, and interpretation. Genetics. 2012;190: 841–854.

260. Sawitzke JA, Thomason LC, Bubunenko M, Li X, Costantino N, Court DL. Recombineering: using drug cassettes to knock out genes in vivo. Methods Enzymol. 2013;533: 79–102.

261. Cobb RE, Wang Y, Zhao H. High-efficiency multiplex genome editing of Streptomyces species using an engineered CRISPR/Cas system. ACS Synth Biol. 2015;4: 723–728.

262. Halbedel S, Stulke J. Tools for the genetic analysis of Mycoplasma. International Journal of Medical Microbiology. 2007. pp. 37–44. doi:10.1016/j.ijmm.2006.11.001

263. Eaton MD, Meiklejohn G, van Herick W, Corey M. STUDIES ON THE ETIOLOGY OF PRIMARY ATYPICAL PNEUMONIA. Journal of Experimental Medicine. 1945. pp. 317–328. doi:10.1084/jem.82.5.317

264. Chanock RM, Dienes L, Eaton MD, ff. Edward DG, Freundt EA, Hayflick L, et al. Mycoplasma pneumoniae: Proposed Nomenclature for Atypical Pneumonia Organism (Eaton Agent). Science. 1963. pp. 662–662. doi:10.1126/science.140.3567.662

265. Weinstein SA, Stiles BG. Recent perspectives in the diagnosis and evidence-based treatment of Mycoplasma genitalium. Expert Rev Anti Infect Ther. 2012;10: 487–499.

266. Evans RD, Hafez YS. Evaluation of a Mycoplasma gallisepticum Strain Exhibiting Reduced Virulence for Prevention and Control of Poultry Mycoplasmosis. Avian Diseases. 1992. p. 197. doi:10.2307/1591490

267. Kumar A, Rahal A, Chakraborty S, Verma AK, Dhama K. Mycoplasma agalactiae, an Etiological Agent of Contagious Agalactia in Small Ruminants: A Review. Vet Med Int. 2014;2014: 286752.

268. Browning G, Citti C. Mollicutes: Molecular Biology and Pathogenesis. Horizon Scientific Press; 2014.

269. Mare CJ, Switzer WP. NEW SPECIES: MYCOPLASMA HYOPNEUMONIAE; A CAUSATIVE AGENT OF VIRUS PIG PNEUMONIA. Vet Med Small Anim Clin. 1965;60: 841–846.

270. Wilson MH, Collier AM. Ultrastructural study of Mycoplasma pneumoniae in organ culture. Journal of Bacteriology. 1976. pp. 332–339. doi:10.1128/jb.125.1.332-339.1976

271. Hatchel JM, Balish MF. Attachment organelle ultrastructure correlates with phylogeny, not gliding motility properties, in Mycoplasma pneumoniae relatives. Microbiology. 2008. pp. 286–295. doi:10.1099/mic.0.2007/012765-0

272. Woese CR, Maniloff J, Zablen LB. Phylogenetic analysis of the mycoplasmas. Proceedings of the National Academy of Sciences. 1980. pp. 494–498. doi:10.1073/pnas.77.1.494

273. Kandavelmani A, Piramanayagam S. Comparative genomics of Mycoplasma: Insights on genome reduction and identification of potential antibacterial targets. Biomedical and Biotechnology Research Journal (BBRJ). 2019. p. 9. doi:10.4103/bbrj.bbrj_142_18

274. Inamine JM, Ho KC, Loechel S, Hu PC. Evidence that UGA is read as a tryptophan codon rather than as a stop codon by Mycoplasma pneumoniae, Mycoplasma genitalium, and Mycoplasma gallisepticum. J Bacteriol. 1990;172: 504–506.

275. Voit EO. 1. What is systems biology all about? Systems Biology: A Very Short Introduction. 2020. pp. 1–8. doi:10.1093/actrade/9780198828372.003.0001

276. Hutchison CA 3rd, Chuang R-Y, Noskov VN, Assad-Garcia N, Deerinck TJ, Ellisman MH, et al. Design and synthesis of a minimal bacterial genome. Science. 2016;351: aad6253.

277. Yus E, Maier T, Michalodimitrakis K, van Noort V, Yamada T, Chen W-H, et al. Impact of Genome Reduction on Bacterial Metabolism and Its Regulation. Science. 2009;326: 1263–1268.

278. Guell M, van Noort V, Yus E, Chen W-H, Leigh-Bell J, Michalodimitrakis K, et al. Transcriptome Complexity in a Genome-Reduced Bacterium. Science. 2009;326: 1268–1271.

279. Kühner S, van Noort V, Betts MJ, Leo-Macias A, Batisse C, Rode M, et al. Proteome Organization in a Genome-Reduced Bacterium. Science. 2009. pp. 1235–1240. doi:10.1126/science.1176343

280. Lluch-Senar M, Mancuso FM, Climente-González H, Peña-Paz MI, Sabido E, Serrano L. Rescuing discarded spectra: Full comprehensive analysis of a minimal proteome. Proteomics. 2016;16: 554–563.

281. Wodke JAH, Puchałka J, Lluch-Senar M, Marcos J, Yus E, Godinho M, et al. Dissecting the energy metabolism in Mycoplasma pneumoniae through genome-scale metabolic modeling. Mol Syst Biol. 2013;9: 653.

282. van Noort V, Seebacher J, Bader S, Mohammed S, Vonkova I, Betts MJ, et al. Cross-talk between phosphorylation and lysine acetylation in a genome-reduced bacterium. Molecular Systems Biology. 2012. p. 571. doi:10.1038/msb.2012.4

283. Maier T, Schmidt A, Güell M, Kühner S, Gavin A, Aebersold R, et al. Quantification of mRNA and protein and integration with protein turnover in a bacterium. Molecular Systems Biology. 2011. p. 511. doi:10.1038/msb.2011.38

284. Chen W-H, van Noort V, Lluch-Senar M, Hennrich ML, Wodke JAH, Yus E, et al. Integration of multi-omics data of a genome-reduced bacterium: Prevalence of post-transcriptional regulation and its correlation with protein abundances. Nucleic Acids Res. 2016;44: 1192–1202.

285. Yus E, Lloréns-Rico V, Martínez S, Gallo C, Eilers H, Blötz C, et al. Determination of the Gene Regulatory Network of a Genome-Reduced Bacterium Highlights Alternative Regulation Independent of Transcription Factors. Cell Syst. 2019;9: 143–158.e13.

286. Lloréns-Rico V, Cano J, Kamminga T, Gil R, Latorre A, Chen W-H, et al. Bacterial antisense RNAs are mainly the product of transcriptional noise. Sci Adv. 2016;2: e1501363.

287. Hara KY, Araki M, Okai N, Wakai S, Hasunuma T, Kondo A. Development of bio-based fine chemical production through synthetic bioengineering. Microb Cell Fact. 2014;13: 173.

288. Piñero-Lambea C, Ruano-Gallego D, Fernández LÁ. Engineered bacteria as therapeutic agents. Current Opinion in Biotechnology. 2015. pp. 94–102. doi:10.1016/j.copbio.2015.05.004

289. de Lorenzo V, Krasnogor N, Schmidt M. For the sake of the Bioeconomy: define what a Synthetic Biology Chassis is! N Biotechnol. 2021;60: 44–51.

290. Piñero-Lambea C, Garcia-Ramallo E, Martinez S, Delgado J, Serrano L, Lluch-Senar M. Mycoplasma pneumoniae Genome Editing Based on Oligo Recombineering and Cas9-Mediated Counterselection. ACS Synthetic Biology. 2020. pp. 1693–1704. doi:10.1021/acssynbio.0c00022

291. Garcia-Morales L, Ruiz E, Gourgues G, Rideau F, Piñero-Lambea C, Lluch-Senar M, et al. A RAGE Based Strategy for the Genome Engineering of the Human Respiratory Pathogen Mycoplasma pneumoniae. ACS Synthetic Biology. 2020. pp. 2737–2748. doi:10.1021/acssynbio.0c00263

292. Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival B Jr, et al. A whole-cell computational model predicts phenotype from genotype. Cell. 2012;150: 389–401.

293. Marucci L, Barberis M, Karr J, Ray O, Race PR, de Souza Andrade M, et al. Computer-Aided Whole-Cell Design: Taking a Holistic Approach by Integrating Synthetic With Systems Biology. Front Bioeng Biotechnol. 2020;8: 942.

294. Zimmerman C-U, Herrmann R. Synthesis of a small, cysteine-rich, 29 amino acids long peptide in Mycoplasma pneumoniae. FEMS Microbiol Lett. 2005;253: 315–321.

295. Osipiuk J, Maltseva N, Dementieva I, Clancy S, Collart F, Joachimiak A. Structure of YidB protein from Shigella flexneri shows a new fold with homeodomain motif. Proteins: Structure, Function, and Bioinformatics. 2006. pp. 509–513. doi:10.1002/prot.21054

296. Wilson KJ, Sessitsch A, Corbo JC, Giller KE, Akkermans AD, Jefferson RA. beta-Glucuronidase (GUS) transposons for ecological and genetic studies of rhizobia and other gram-negative bacteria. Microbiology. 1995;141 (Pt 7): 1691–1705.

297. Cebolla A, Guzmán C, de Lorenzo V. Nondisruptive detection of activity of catabolic promoters of Pseudomonas putida with an antigenic surface reporter system. Applied and environmental microbiology. 1996. pp. 214–220. doi:10.1128/aem.62.1.214-220.1996

298. Fisunov GY, Garanina IA, Evsyutina DV, Semashko TA, Nikitina AS, Govorun VM. Reconstruction of Transcription Control Networks in Mollicutes by High-Throughput Identification of Promoters. Front Microbiol. 2016;7: 1977.

299. Torres-Puig S, Broto A, Querol E, Piñol J, Pich OQ. A novel sigma factor reveals a unique regulon controlling cell-specific recombination in Mycoplasma genitalium. Nucleic Acids Research. 2015. pp. 4923–4936. doi:10.1093/nar/gkv422

300. Le TBK, Imakaev MV, Mirny LA, Laub MT. High-resolution mapping of the spatial organization of a bacterial chromosome. Science. 2013;342: 731–734.

301. Marbouty M, Le Gall A, Cattoni DI, Cournac A, Koh A, Fiche J-B, et al. Condensin- and Replication-Mediated Bacterial Chromosome Folding and Origin Condensation Revealed by Hi-C and Super-resolution Imaging. Mol Cell. 2015;59: 588–602.

302. Staczek P, Higgins NP. Gyrase and Topo IV modulate chromosome domain size in vivo. Mol Microbiol. 1998;29: 1435–1448.

303. Fritsche M, Li S, Heermann DW, Wiggins PA. A model for Escherichia coli chromosome packaging supports transcription factor-induced DNA domain formation. Nucleic Acids Research. 2012. pp. 972–980. doi:10.1093/nar/gkr779

304. Tadesse S, Graumann PL. Differential and dynamic localization of topoisomerases in Bacillus subtilis. J Bacteriol. 2006;188: 3002–3011.

305. Herrmann R, Reiner B. Mycoplasma pneumoniae and Mycoplasma genitalium: a comparison of two closely related bacterial species. Curr Opin Microbiol. 1998;1: 572–579.

306. Zhi X, Leng F. Dependence of transcription-coupled DNA supercoiling on promoter strength in Escherichia coli topoisomerase I deficient strains. Gene. 2013;514: 82–90.

307. Dorman CJ. Regulation of transcription by DNA supercoiling in Mycoplasma genitalium: global control in the smallest known self-replicating genome. Molecular microbiology. 2011. pp. 302–304.

308. Lim HN, Lee Y, Hussein R. Fundamental relationship between operon organization and gene expression. Proc Natl Acad Sci U S A. 2011;108: 10626–10631.

309. Junier I, Unal EB, Yus E, Lloréns-Rico V, Serrano L. Insights into the Mechanisms of Basal Coordination of Transcription Using a Genome-Reduced Bacterium. Cell Syst. 2018;7: 227–229.

310. Junier I, Rivoire O. Conserved Units of Co-Expression in Bacterial Genomes: An Evolutionary Insight into Transcriptional Regulation. PLoS One. 2016;11: e0155740.

311. Mazin PV, Fisunov GY, Gorbachev AY, Kapitskaya KY, Altukhov IA, Semashko TA, et al. Transcriptome analysis reveals novel regulatory mechanisms in a genome-reduced bacterium. Nucleic Acids Research. 2014. pp. 13254–13268. doi:10.1093/nar/gku976

312. Clark MA, Baumann L, Baumann P. Sequence analysis of a 34.7-kb DNA segment from the genome of Buchnera aphidicola (endosymbiont of aphids) containing groEL, dnaA, the atp operon, gidA, and rho. Curr Microbiol. 1998;36: 158–163.

313. Yus E, Yang J-S, Sogues A, Serrano L. A reporter system coupled with high-throughput sequencing unveils key bacterial transcription and translation determinants. Nat Commun. 2017;8: 368.

314. Lew CM, Gralla JD. Mechanism of stimulation of ribosomal promoters by binding of the +1 and +2 nucleotides. J Biol Chem. 2004;279: 19481–19485.

315. Krásný L, Tiserová H, Jonák J, Rejman D, Sanderová H. The identity of the transcription +1 position is crucial for changes in gene expression in response to amino acid starvation in Bacillus subtilis. Mol Microbiol. 2008;69: 42–54.

316. Ciampi MS, Sofia Ciampi M. Rho-dependent terminators and transcription termination. Microbiology. 2006. pp. 2515–2528. doi:10.1099/mic.0.28982-0

317. Peters JM, Vangeloff AD, Landick R. Bacterial Transcription Terminators: The RNA 3′-End Chronicles. Journal of Molecular Biology. 2011. pp. 793–813. doi:10.1016/j.jmb.2011.03.036

318. Quirk PG, Dunkley EA Jr, Lee P, Krulwich TA. Identification of a putative Bacillus subtilis rho gene. J Bacteriol. 1993;175: 8053.

319. Washburn RS, Marra A, Bryant AP, Rosenberg M, Gentry DR. rho Is Not Essential for Viability or Virulence inStaphylococcus aureus. Antimicrobial Agents and Chemotherapy. 2001. pp. 1099–1103. doi:10.1128/aac.45.4.1099-1103.2001

320. Opperman T, Richardson JP. Phylogenetic analysis of sequences from diverse bacteria with homology to the Escherichia coli rho gene. J Bacteriol. 1994;176: 5033–5043.

321. Washio T, Sasayama J, Tomita M. Analysis of complete genomes suggests that many prokaryotes do not rely on hairpin formation in transcription termination. Nucleic Acids Research. 1998. pp. 5456–5463. doi:10.1093/nar/26.23.5456

322. Farnham PJ, Greenblatt J, Platt T. Effects of NusA protein on transcription termination in the tryptophan operon of Escherichia coli. Cell. 1982;29: 945–951.

323. Gusarov I, Nudler E. Control of Intrinsic Transcription Termination by N and NusA. Cell. 2001. pp. 437–449. doi:10.1016/s0092-8674(01)00582-7

324. Nicolas P, Mäder U, Dervyn E, Rochat T, Leduc A, Pigeonneau N, et al. Condition-dependent transcriptome reveals high-level regulatory architecture in Bacillus subtilis. Science. 2012;335: 1103–1106.

325. Coppins RL, Hall KB, Groisman EA. The intricate world of riboswitches. Curr Opin Microbiol. 2007;10: 176–181.

326. Serganov A, Nudler E. A Decade of Riboswitches. Cell. 2013. pp. 17–24. doi:10.1016/j.cell.2012.12.024

327. Barrick JE, Breaker RR. The distributions, mechanisms, and structures of metabolite-binding riboswitches. Genome Biol. 2007;8: R239.

328. Bocobza SE, Aharoni A. Small molecules that interact with RNA: riboswitch-based gene control and its involvement in metabolic regulation in plants and algae. Plant J. 2014;79: 693–703.

329. Kim JN, Roth A, Breaker RR. Guanine riboswitch variants from Mesoplasma florum selectively recognize 2'-deoxyguanosine. Proc Natl Acad Sci U S A. 2007;104: 16092–16097.

330. Brantl S, Wagner EGH. An antisense RNA-mediated transcriptional attenuation mechanism functions in Escherichia coli. J Bacteriol. 2002;184: 2740–2747.

331. Urban JH, Papenfort K, Thomsen J, Schmitz RA, Vogel J. A Conserved Small RNA Promotes Discoordinate Expression of the glmUS Operon mRNA to Activate GlmS Synthesis. Journal of Molecular Biology. 2007. pp. 521–528. doi:10.1016/j.jmb.2007.07.035

332. Opdyke JA, Kang J-G, Storz G. GadY, a Small-RNA Regulator of Acid Response Genes in Escherichia coli. Journal of Bacteriology. 2004. pp. 6698–6705. doi:10.1128/jb.186.20.6698-6705.2004

333. Wehner S, Damm K, Hartmann RK, Marz M. Dissemination of 6S RNA among bacteria. RNA Biol. 2014;11: 1467–1478.

334. Siqueira FM, de Morais GL, Higashi S, Beier LS, Breyer GM, de Sá Godinho CP, et al. Mycoplasma non-coding RNA: identification of small RNAs and targets. BMC Genomics. 2016;17: 743.

335. Narra HP, Schroeder CLC, Sahni A, Rojas M, Khanipov K, Fofanov Y, et al. Small Regulatory RNAs of Rickettsia conorii. Sci Rep. 2016;6: 36728.

336. Hershberg R, Petrov DA. Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet. 2010;6: e1001115.

337. Deutscher MP, M. P. Deutscher. Degradation of RNA in bacteria: comparison of mRNA and stable RNA. Nucleic Acids Research. 2006. pp. 659–666. doi:10.1093/nar/gkj472

338. Hui MP, Foley PL, Belasco JG. Messenger RNA Degradation in Bacterial Cells. Annual Review of Genetics. 2014. pp. 537–559. doi:10.1146/annurev-genet-120213-092340

339. Kaberdin VR, Singh D, Lin-Chao S. Composition and conservation of the mRNA-degrading machinery in bacteria. J Biomed Sci. 2011;18: 23.

340. Bechhofer DH, Oussenko IA, Deikus G, Yao S, Mathy N, Condon C. Chapter 14 Analysis of mRNA Decay in Bacillus subtilis. RNA Turnover in Bacteria, Archaea and Organelles. 2008. pp. 259–276. doi:10.1016/s0076-6879(08)02214-3

341. Hsieh P-K, Richards J, Liu Q, Belasco JG. Specificity of RppH-dependent RNA degradation in Bacillus subtilis. Proc Natl Acad Sci U S A. 2013;110: 8864–8869.

342. Zuo Y, Deutscher MP. Exoribonuclease superfamilies: structural analysis and phylogenetic distribution. Nucleic Acids Res. 2001;29: 1017–1026.

343. Condon C. RNA Processing and Degradation in Bacillus subtilis. Microbiology and Molecular Biology Reviews. 2003. pp. 157–174. doi:10.1128/mmbr.67.2.157-174.2003

344. Lalonde MS, Zuo Y, Zhang J, Gong X, Wu S, Malhotra A, et al. Exoribonuclease R in Mycoplasma genitalium can carry out both RNA processing and degradative functions and is sensitive to RNA ribose methylation. RNA. 2007;13: 1957–1968.

345. Cho KH. The Structure and Function of the Gram-Positive Bacterial RNA Degradosome. Front Microbiol. 2017;8: 154.

346. Dutow P, Schmidl SR, Ridderbusch M, Stülke J. Interactions between glycolytic enzymes of Mycoplasma pneumoniae. J Mol Microbiol Biotechnol. 2010;19: 134–139.

347. Stern MJ, Ames GF-L, Smith NH, Clare Robinson E, Higgins CF. Repetitive extragenic palindromic sequences: A major component of the bacterial genome. Cell. 1984. pp. 1015–1026. doi:10.1016/0092-8674(84)90436-7

348. Mrázek J. Analysis of distribution indicates diverse functions of simple sequence repeats in Mycoplasma genomes. Mol Biol Evol. 2006;23: 1370–1385.

349. Yang Y, Ames GF. DNA gyrase binds to the family of prokaryotic repetitive extragenic palindromic sequences. Proc Natl Acad Sci U S A. 1988;85: 8850–8854.

350. Cattani AM, Siqueira FM, Guedes RLM, Schrank IS. Repetitive Elements in Mycoplasma hyopneumoniae Transcriptional Regulation. PLoS One. 2016;11: e0168626.

351. Kim Y, Koh I, Rho M. Deciphering the human microbiome using next-generation sequencing data and bioinformatics approaches. Methods. 2015;79-80: 52–59.

352. Crappé J, Van Criekinge W, Trooskens G, Hayakawa E, Luyten W, Baggerman G, et al. Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs. BMC Genomics. 2013. p. 648. doi:10.1186/1471-2164-14-648

353. Ota T, Suzuki Y, Nishikawa T, Otsuki T, Sugiyama T, Irie R, et al. Complete sequencing and characterization of 21,243 full-length human cDNAs. Nat Genet. 2004;36: 40–45.

354. Savard J, Marques-Souza H, Aranda M, Tautz D. A segmentation gene in tribolium produces a polycistronic mRNA that codes for multiple conserved peptides. Cell. 2006;126: 559–569.

355. Makarewich CA, Olson EN. Mining for Micropeptides. Trends Cell Biol. 2017;27: 685–696.

356. Avila EE. Functions of Antimicrobial Peptides in Vertebrates. Current Protein & Peptide Science. 2017. doi:10.2174/1389203717666160813162629

357. Rowland SL, Burkholder WF, Cunningham KA, Maciejewski MW, Grossman AD, King GF. Structure and mechanism of action of Sda, an inhibitor of the histidine kinases that regulate initiation of sporulation in Bacillus subtilis. Mol Cell. 2004;13: 689–701.

358. Alix E, Blanc-Potard A-B. Peptide-assisted degradation of the Salmonella MgtC virulence factor. EMBO J. 2008;27: 546–557.

359. Mumtaz MAS, Couso JP. Ribosomal profiling adds new coding sequences to the proteome. Biochem Soc Trans. 2015;43: 1271–1276.

360. D'Lima NG, Khitun A, Rosenbloom AD, Yuan P, Gassaway BM, Barber KW, et al. Comparative Proteomics Enables Identification of Nonannotated Cold Shock Proteins in E. coli. J Proteome Res. 2017;16: 3722–3731.

361. Dandekar T, Huynen M, Regula JT, Ueberle B, Zimmermann CU, Andrade MA, et al. Re-annotating the Mycoplasma pneumoniae genome sequence: adding value, function and reading frames. Nucleic Acids Res. 2000;28: 3278–3288.

362. Lluch-Senar M, Vallmitjana M, Querol E, Piñol J. A new promoterless reporter vector reveals antisense transcription in Mycoplasma genitalium. Microbiology. 2007;153: 2743–2752.

363. Goyal A, Belardinelli R, Rodnina MV. Non-canonical Binding Site for Bacterial Initiation Factor 3 on the Large Ribosomal Subunit. Cell Rep. 2017;20: 3113–3122.

364. Yang X, Tschaplinski TJ, Hurst GB, Jawdy S, Abraham PE, Lankford PK, et al. Discovery and annotation of small proteins using genomics, proteomics, and computational approaches. Genome Res. 2011;21: 634–641.

365. Wang R, Braughton KR, Kretschmer D, Bach T-HL, Queck SY, Li M, et al. Identification of novel cytolytic peptides as key virulence determinants for community-associated MRSA. Nature Medicine. 2007. pp. 1510–1514. doi:10.1038/nm1656

366. Goldberg AL. Correlation between rates of degradation of bacterial proteins in vivo and their sensitivity to proteases. Proc Natl Acad Sci U S A. 1972;69: 2640–2644.

367. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 1980;16: 111–120.

368. Ina Y. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. J Mol Evol. 1995;40: 190–226.

369. Makalowski W, Boguski MS. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. Proc Natl Acad Sci U S A. 1998;95: 9407–9412.

370. Ochman H. Distinguishing the ORFs from the ELFs: short bacterial genes and the annotation of genomes. Trends Genet. 2002;18: 335–337.

371. Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics. 2011;27: i275–82.

372. Kong L, Zhang Y, Ye Z-Q, Liu X-Q, Zhao S-Q, Wei L, et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res. 2007;35: W345–9.

373. Zhao J, Song X, Wang K. lncScore: alignment-free identification of long noncoding RNA from assembled novel transcripts. Sci Rep. 2016;6: 34838.

374. Hu L, Xu Z, Hu B, Lu ZJ. COME: a robust coding potential calculation tool for lncRNA identification and characterization based on multiple features. Nucleic Acids Research. 2017. pp. e2–e2. doi:10.1093/nar/gkw798

375. Mount DW. Using the Basic Local Alignment Search Tool (BLAST). Cold Spring Harbor Protocols. 2007. p. db.top17. doi:10.1101/pdb.top17

376. Kodama T, Matsubayashi T, Yanagihara T, Komoto H, Ara K, Ozaki K, et al. A novel small protein of Bacillus subtilis involved in spore germination and spore coat assembly. Biosci Biotechnol Biochem. 2011;75: 1119–1128.

377. Baumgartner D, Kopf M, Klähn S, Steglich C, Hess WR. Small proteins in cyanobacteria provide a paradigm for the functional analysis of the bacterial micro-proteome. BMC Microbiology. 2016. doi:10.1186/s12866-016-0896-z

378. Impens F, Rolhion N, Radoshevich L, Bécavin C, Duval M, Mellin J, et al. Author Correction: N-terminomics identifies Prli42 as a membrane miniprotein conserved in Firmicutes and critical for stressosome activation in Listeria monocytogenes. Nature Microbiology. 2018. pp. 962–962. doi:10.1038/s41564-018-0197-4

379. Kemp G, Cymer F. Small membrane proteins - elucidating the function of the needle in the haystack. Biol Chem. 2014;395: 1365–1377.

380. Sheng H, Stauffer WT, Hussein R, Lin C, Lim HN. Nucleoid and cytoplasmic localization of small RNAs in Escherichia coli. Nucleic Acids Res. 2017;45: 2919–2934.

381. Mallick P, Schirle M, Chen SS, Flory MR, Lee H, Martin D, et al. Computational prediction of proteotypic peptides for quantitative proteomics. Nat Biotechnol. 2007;25: 125–131.

382. Hücker SM, Ardern Z, Goldberg T, Schafferhans A, Bernhofer M, Vestergaard G, et al. Discovery of numerous novel small genes in the intergenic regions of the Escherichia coli O157:H7 Sakai genome. PLoS One. 2017;12: e0184119.

383. Kang Y-J, Yang D-C, Kong L, Hou M, Meng Y-Q, Wei L, et al. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Research. 2017. pp. W12–W16. doi:10.1093/nar/gkx428

384. Fusaro VA, Mani DR, Mesirov JP, Carr SA. Prediction of high-responding peptides for targeted protein assays by mass spectrometry. Nat Biotechnol. 2009;27: 190–198.

385. Weiner J 3rd, Herrmann R, Browning GF. Transcription in Mycoplasma pneumoniae. Nucleic Acids Res. 2000;28: 4488–4496.

386. Käll L, Krogh A, Sonnhammer ELL. A combined transmembrane topology and signal peptide prediction method. J Mol Biol. 2004;338: 1027–1036.

387. Tully JG, Rose DL, Whitcomb RF, Wenzel RP. Enhanced isolation of Mycoplasma pneumoniae from throat washings with a newly-modified culture medium. J Infect Dis. 1979;139: 478–482.

388. Wiśniewski JR. Quantitative Evaluation of Filter Aided Sample Preparation (FASP) and Multienzyme Digestion FASP Protocols. Anal Chem. 2016;88: 5438–5443.

389. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis. 1999;20: 3551–3567.

390. Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, et al. Machine learning for neuroimaging with scikit-learn. Front Neuroinform. 2014;8: 14.

391. Cao D-S, Xu Q-S, Liang Y-Z. propy: a tool to generate various modes of Chou's PseAAC. Bioinformatics. 2013. pp. 960–962. doi:10.1093/bioinformatics/btt072

392. Mir K, Neuhaus K, Scherer S, Bossert M, Schober S. Predicting statistical properties of open reading frames in bacterial genomes. PLoS One. 2012;7: e45103.

393. Lahtvee P-J, Adamberg K, Arike L, Nahku R, Aller K, Vilu R. Multi-omics approach to study the growth efficiency and amino acid metabolism in Lactococcus lactis at various specific growth rates. Microbial Cell Factories. 2011. p. 12. doi:10.1186/1475-2859-10-12

394. Müller SA, Findeiß S, Pernitzsch SR, Wissenbach DK, Stadler PF, Hofacker IL, et al. Identification of new protein coding sequences and signal peptidase cleavage sites of Helicobacter pylori strain 26695 by proteogenomics. J Proteomics. 2013;86: 27–42.

395. Gao L, Ge H, Huang X, Liu K, Zhang Y, Xu W, et al. Systematically ranking the tightness of membrane association for peripheral membrane proteins (PMPs). Mol Cell Proteomics. 2015;14: 340–353.

396. Impens F, Rolhion N, Radoshevich L, Bécavin C, Duval M, Mellin J, et al. N-terminomics identifies Prli42 as a membrane miniprotein conserved in Firmicutes and critical for stressosome activation in Listeria monocytogenes. Nature Microbiology. 2017. doi:10.1038/nmicrobiol.2017.5

397. Chi H, Wang X, Shao Y, Qin Y, Deng Z, Wang L, et al. Engineering and modification of microbial chassis for systems and synthetic biology. Synth Syst Biotechnol. 2019;4: 25–33.

398. Salama NR, Shepherd B, Falkow S. Global transposon mutagenesis and essential gene analysis of Helicobacter pylori. J Bacteriol. 2004;186: 7926–7935.

399. Wong SMS, Gawronski JD, Lapointe D, Akerley BJ. High-throughput insertion tracking by deep sequencing for the analysis of bacterial pathogens. Methods Mol Biol. 2011;733: 209–222.

400. Green B, Bouchier C, Fairhead C, Craig NL, Cormack BP. Insertion site preference of Mu, Tn5, and Tn7 transposons. Mob DNA. 2012;3: 3.

401. Arenas J, Zomer A, Harders-Westerveen J, Bootsma HJ, De Jonge MI, Stockhofe-Zurwieden N, et al. Identification of conditionally essential genes for Streptococcus suis infection in pigs. Virulence. 2020. pp. 446–464. doi:10.1080/21505594.2020.1764173

402. van Opijnen T, Bodi KL, Camilli A. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. Nature Methods. 2009. pp. 767–772. doi:10.1038/nmeth.1377

403. Barquist L, Mayho M, Cummins C, Cain AK, Boinett CJ, Page AJ, et al. The TraDIS toolkit: sequencing and analysis for dense transposon mutant libraries. Bioinformatics. 2016. pp. 1109–1111. doi:10.1093/bioinformatics/btw022

404. Arenas J, Zomer A, Harders-Westerveen J, Bootsma HJ, De Jonge MI, Stockhofe-Zurwieden N, et al. Identification of conditionally essential genes for Streptococcus suis infection in pigs. Virulence. 2020. pp. 446–464. doi:10.1080/21505594.2020.1764173

405. Langridge GC, Phan M-D, Turner DJ, Perkins TT, Parts L, Haase J, et al. Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. Genome Res. 2009;19: 2308–2316.

406. Glass JI, Assad-Garcia N, Alperovich N, Yooseph S, Lewis MR, Maruf M, et al. Essential genes of a minimal bacterium. Proceedings of the National Academy of Sciences. 2006. pp. 425–430. doi:10.1073/pnas.0510013103

407. Akerley BJ, Rubin EJ, Camilli A, Lampe DJ, Robertson HM, Mekalanos JJ. Systematic identification of essential genes by in vitro mariner mutagenesis. Proc Natl Acad Sci U S A. 1998;95: 8927–8932.

408. Iii CAH, Hutchison CA III. Global Transposon Mutagenesis and a Minimal Mycoplasma Genome. Science. 1999. pp. 2165–2169. doi:10.1126/science.286.5447.2165

409. Parekh S, Ziegenhain C, Vieth B, Enard W, Hellmann I. The impact of amplification on differential expression analyses by RNA-seq. Sci Rep. 2016;6: 25533.

410. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079.

411. Xu H, Luo X, Qian J, Pang X, Song J, Qian G, et al. FastUniq: A Fast De Novo Duplicates Removal Tool for Paired Short Reads. PLoS One. 2012;7: e52249.

412. Laehnemann D, Borkhardt A, McHardy AC. Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. Brief Bioinform. 2016;17: 154–179.

413. Wetmore KM, Price MN, Waters RJ, Lamson JS, He J, Hoover CA, et al. Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons. MBio. 2015;6: e00306–15.

414. Levin HL, Moran JV. Dynamic interactions between transposable elements and their hosts. Nature Reviews Genetics. 2011. pp. 615–627. doi:10.1038/nrg3030

415. Han M-J, Xu H-E, Zhang H-H, Feschotte C, Zhang Z. Spy: a new group of eukaryotic DNA transposons without target site duplications. Genome Biol Evol. 2014;6: 1748–1757.

416. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003;19: 185–193.

417. Osterman AL, Gerdes SY. Microbial Gene Essentiality: Protocols and Bioinformatics. Humana Press; 2010.

418. McCoy KM, Antonio ML, van Opijnen T. MAGenTA: a Galaxy implemented tool for complete Tn-Seq analysis and data visualization. Bioinformatics. 2017. pp. 2781–2783. doi:10.1093/bioinformatics/btx320

419. Weber M, Burgos R, Yus E, Yang J-S, Lluch-Senar M, Serrano L. Impact of C-terminal amino acid composition on protein expression in bacteria. doi:10.1101/751305

420. Burgos R, Totten PA. Characterization of the operon encoding the Holliday junction helicase RuvAB from Mycoplasma genitalium and its role in mgpB and mgpC gene variation. J Bacteriol. 2014;196: 1608–1618.

421. Pich OQ, Burgos R, Planell R, Querol E, Piñol J. Comparative analysis of antibiotic resistance gene markers in Mycoplasma genitalium: application to studies of the minimal gene complement. Microbiology. 2006;152: 519–527.

422. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9: 357–359.

423. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010. pp. 841–842. doi:10.1093/bioinformatics/btq033

424. Goodstadt L. Ruffus: a lightweight Python library for computational pipelines. Bioinformatics. 2010;26: 2778–2779.

425. Truong C, Oudre L, Vayatis N. Selective review of offline change point detection methods. Signal Processing. 2020. p. 107299. doi:10.1016/j.sigpro.2019.107299

426. DeJesus MA, Zhang YJ, Sassetti CM, Rubin EJ, Sacchettini JC, Ioerger TR. Bayesian analysis of gene essentiality based on sequencing of transposon insertion libraries. Bioinformatics. 2013. pp. 695–703. doi:10.1093/bioinformatics/btt043

427. Charbonneau ARL, Forman OP, Cain AK, Newland G, Robinson C, Boursnell M, et al. Defining the ABC of gene essentiality in streptococci. BMC Genomics. 2017;18: 426.

428. DeJesus MA, Ioerger TR. A Hidden Markov Model for identifying essential and growth-defect regions in bacterial genomes from transposon insertion sequencing data. BMC Bioinformatics. 2013;14: 303.

429. Osterman AL, Gerdes SY. Microbial Gene Essentiality: Protocols and Bioinformatics. Humana Press; 2010.

430. Garreta R, Moncecchi G. Learning scikit-learn: Machine Learning in Python. Packt Publishing Ltd; 2013.

431. Akaike H. Information Theory and an Extension of the Maximum Likelihood Principle. Springer Series in Statistics. 1992. pp. 610–624. doi:10.1007/978-1-4612-0919-5_38

432. Poulsen BE, Yang R, Clatworthy AE, White T, Osmulski SJ, Li L, et al. Defining the core essential genome of. Proc Natl Acad Sci U S A. 2019;116: 10072–10080.

433. Charbonneau ARL, Forman OP, Cain AK, Newland G, Robinson C, Boursnell M, et al. Defining the ABC of gene essentiality in streptococci. BMC Genomics. 2017;18: 426.

434. DeJesus MA, Ioerger TR. A Hidden Markov Model for identifying essential and growth-defect regions in bacterial genomes from transposon insertion sequencing data. BMC Bioinformatics. 2013;14: 303.

435. Garreta R, Moncecchi G. Learning scikit-learn: Machine Learning in Python. Packt Publishing Ltd; 2013.

436. Kumar A, Seringhaus M, Biery MC, Sarnovsky RJ, Umansky L, Piccirillo S, et al. Large-scale mutagenesis of the yeast genome using a Tn7-derived multipurpose transposon. Genome Res. 2004;14: 1975–1986.

437. Green B, Bouchier C, Fairhead C, Craig NL, Cormack BP. Insertion site preference of Mu, Tn5, and Tn7 transposons. Mob DNA. 2012;3: 3.

438. Hansen KD, Irizarry RA, Wu Z. Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics. 2012;13: 204–216.

439. Levin HL, Moran JV. Dynamic interactions between transposable elements and their hosts. Nature Reviews Genetics. 2011. pp. 615–627. doi:10.1038/nrg3030

440. Gawronski JD, Wong SMS, Giannoukos G, Ward DV, Akerley BJ. Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for Haemophilus genes required in the lung. Proceedings of the National Academy of Sciences. 2009. pp. 16422–16427. doi:10.1073/pnas.0906627106

441. Truong C, Oudre L, Vayatis N. Selective review of offline change point detection methods. Signal Processing. 2020. p. 107299. doi:10.1016/j.sigpro.2019.107299

442. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011;39: W29–37.

443. Balish MF, Ross SM, Fisseha M, Krause DC. Deletion analysis identifies key functional domains of the cytadherence-associated protein HMW2 of Mycoplasma pneumoniae. Mol Microbiol. 2003;50: 1507–1516.

444. Liu J, Yokota H, Kim R, Kim S-H. A conserved hypothetical protein from Mycoplasma genitalium shows structural homology to nusb proteins. Proteins. 2004;55: 1082–1086.

445. Fickett JW. ORFs and Genes: How Strong a Connection? Journal of Computational Biology. 1995. pp. 117–123. doi:10.1089/cmb.1995.2.117

446. Dinger ME, Pang KC, Mercer TR, Mattick JS. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. PLoS Comput Biol. 2008;4: e1000176.

447. Storz G, Wolf YI, Ramamurthi KS. Small proteins can no longer be ignored. Annu Rev Biochem. 2014;83: 753–777.

448. Miravet-Verde S, Ferrar T, Espadas-García G, Mazzolini R, Gharrab A, Sabido E, et al. Unraveling the hidden universe of small proteins in bacterial genomes. Molecular Systems Biology. 2019. doi:10.15252/msb.20188290

449. Ingolia NT, Lareau LF, Weissman JS. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell. 2011;147: 789–802.

450. Aspden JL, Eyre-Walker YC, Phillips RJ, Amin U, Mumtaz MAS, Brocard M, et al. Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. Elife. 2014;3: e03528.

451. Frith MC, Forrest AR, Nourbakhsh E, Pang KC, Kai C, Kawai J, et al. The Abundance of Short Proteins in the Mammalian Proteome. PLoS Genetics. 2006. p. e52. doi:10.1371/journal.pgen.0020052

452. Ladoukakis E, Pereira V, Magny EG, Eyre-Walker A, Couso JP. Hundreds of putatively functional small open reading frames in Drosophila. Genome Biol. 2011;12: R118.

453. Vanderperre B, Lucier J-F, Roucou X. HAltORF: a database of predicted out-of-frame alternative open reading frames in human. Database . 2012;2012: bas025.

454. Kumar D, Yadav AK, Dash D. Choosing an Optimal Database for Protein Identification from Tandem Mass Spectrometry Data. Methods Mol Biol. 2017;1549: 17–29.

455. Oyama M, Kozuka-Hata H, Suzuki Y, Semba K, Yamamoto T, Sugano S. Diversity of translation start sites may define increased complexity of the human short ORFeome. Mol Cell Proteomics. 2007;6: 1000–1006.

456. Barquist L, Boinett CJ, Cain AK. Approaches to querying bacterial genomes with transposon-insertion sequencing. RNA Biol. 2013;10: 1161–1169.

457. Halbedel S, Stülke J. Probing in vivo promoter activities in Mycoplasma pneumoniae: a system for generation of single-copy reporter constructs. Appl Environ Microbiol. 2006;72: 1696–1699.

458. Mealer R, Butler H, Hughes T. Functional fusion proteins by random transposon-based GFP insertion. Methods Cell Biol. 2008;85: 23–44.

459. Bednarz H, Niehaus K. Using transposition to introduce eGFP fusions in Sinorhizobium meliloti : A tool to analyze protein localization patterns in bacteria. Journal of Biotechnology. 2017. pp. 139–149. doi:10.1016/j.jbiotec.2016.12.013

460. Weber M, Burgos R, Yus E, Yang J-S, Lluch-Senar M, Serrano L. Impact of C-terminal amino acid composition on protein expression in bacteria. Mol Syst Biol. 2020;16: e9208.

461. Matsuoka M, Sasaki T. Inactivation of macrolides by producers and pathogens. Curr Drug Targets Infect Disord. 2004;4: 217–240.

462. Hartley RW. Barnase and barstar: two small proteins to fold and fit together. Trends Biochem Sci. 1989;14: 450–454.

463. Montero-Blay A, Miravet-Verde S, Lluch-Senar M, Piñero-Lambea C, Serrano L. SynMyco transposon: engineering transposon vectors for efficient transformation of minimal genomes. DNA Res. 2019;26: 327–339.

464. Osterman AL, Gerdes SY. Microbial Gene Essentiality: Protocols and Bioinformatics. Humana Press; 2010.

465. Burgos R, Weber M, Martinez S, Lluch-Senar M, Serrano L. Protein quality control and regulated proteolysis in the genome-reduced organism Mycoplasma pneumoniae. Mol Syst Biol. 2020;16: e9530.

466. Wiktor M, Weichert D, Howe N, Huang C-Y, Olieric V, Boland C, et al. Structural insights into the mechanism of the membrane integral N-acyltransferase step in bacterial lipoprotein synthesis. Nat Commun. 2017;8: 15952.

467. Juretić D, Zoranić L, Zucić D. Basic charge clusters and predictions of membrane protein topology. J Chem Inf Comput Sci. 2002;42: 620–632.

468. Sonnhammer EL, von Heijne G, Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol. 1998;6: 175–182.

469. Subramaniyam S, Zaveri A, DeJesus MA, Smith C, Baker RE, Ehrt S, et al. Statistical Analysis of Variability in TnSeq Data Across Conditions Using Zero-Inflated Negative Binomial Regression. doi:10.1101/590281

470. Montero-Blay A, Miravet-Verde S, Lluch-Senar M, Piñero-Lambea C, Serrano L. SynMyco transposon: engineering transposon vectors for efficient transformation of minimal genomes. DNA Res. 2019;26: 327–339.

471. Montero-Blay A, Piñero-Lambea C, Miravet-Verde S, Lluch-Senar M, Serrano L. Inferring Active Metabolic Pathways from Proteomics and Essentiality Data. Cell Rep. 2020;31: 107722.

472. Shaw D, Miravet-Verde S, Piñero-Lambea C, Serrano L, Lluch-Senar M. LoxTnSeq: random transposon insertions combined with cre/lox recombination and counterselection to generate large random genome reductions. Microb Biotechnol. 2020. doi:10.1111/1751-7915.13714

473. Wu S, Zhu W, Thompson P, Hannun YA. Evaluating intrinsic and non-intrinsic cancer risk factors. Nat Commun. 2018;9: 3490.

474. Martín-Pardillos A, Valls Chiva Á, Bande Vargas G, Hurtado Blanco P, Piñeiro Cid R, Guijarro PJ, et al. The role of clonal communication and heterogeneity in breast cancer. BMC Cancer. 2019;19: 666.

475. Sarkar A, Sim C, Hong YS, Hogan JR, Fraser MJ, Robertson HM, et al. Molecular evolutionary analysis of the widespread piggyBac transposon family and related “domesticated” sequences. Mol Genet Genomics. 2003;270: 173–180.

476. Boekhorst J, Wilson G, Siezen RJ. Searching in microbial genomes for encoded small proteins. Microb Biotechnol. 2011;4: 308–313.

477. Han Y, Ma B, Zhang K. SPIDER: software for protein identification from sequence tags with de novo sequencing error. Proceedings. 2004 IEEE Computational Systems Bioinformatics Conference, 2004. CSB 2004. doi:10.1109/csb.2004.1332434

478. Brosch M, Yu L, Hubbard T, Choudhary J. Accurate and sensitive peptide identification with Mascot Percolator. J Proteome Res. 2009;8: 3176–3181.

479. Harney DJ, Hutchison AT, Su Z, Hatchwell L, Heilbronn LK, Hocking S, et al. Small-protein Enrichment Assay Enables the Rapid, Unbiased Analysis of Over 100 Low Abundance Factors from Human Plasma. Molecular & Cellular Proteomics. 2019. pp. 1899–1915. doi:10.1074/mcp.tir119.001562

480. Louadi Z, Oubounyt M, Tayara H, Chong KT. Deep Splicing Code: Classifying Alternative Splicing Events Using Deep Learning. Genes . 2019;10. doi:10.3390/genes10080587

481. Langridge GC, Phan M-D, Turner DJ, Perkins TT, Parts L, Haase J, et al. Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. Genome Res. 2009;19: 2308–2316.

482. Goodman AL, McNulty NP, Zhao Y, Leip D, Mitra RD, Lozupone CA, et al. Identifying genetic determinants needed to establish a human gut symbiont in its habitat. Cell Host Microbe. 2009;6: 279–289.

483. Kakkanat A, Phan M-D, Lo AW, Beatson SA, Schembri MA. Novel genes associated with enhanced motility of Escherichia coli ST131. PLoS One. 2017;12: e0176290.

484. Dorman MJ, Feltwell T, Goulding DA, Parkhill J, Short FL. The Capsule Regulatory Network of Defined by density-TraDISort. MBio. 2018;9. doi:10.1128/mBio.01863-18

485. Matern WM, Jenquin RL, Bader JS, Karakousis PC. Identifying the essential genes of Mycobacterium avium subsp. hominissuis with Tn-Seq using a rank-based filter procedure. Sci Rep. 2020;10: 1095.

486. Liu F, Wang C, Wu Z, Zhang Q, Liu P. A zero-inflated Poisson model for insertion tolerance analysis of genes based on Tn-seq data. Bioinformatics. 2016;32: 1701–1708.

487. Zomer A, Burghout P, Bootsma HJ, Hermans PWM, van Hijum SAFT. ESSENTIALS: software for rapid analysis of high throughput transposon insertion sequencing data. PLoS One. 2012;7: e43012.

488. Solaimanpour S, Sarmiento F, Mrázek J. Tn-seq explorer: a tool for analysis of high-throughput sequencing data of transposon mutant libraries. PLoS One. 2015;10: e0126070.

489. Koonin EV. How many genes can make a cell: the minimal-gene-set concept. Annu Rev Genomics Hum Genet. 2000;1: 99–116.

490. Khitun A, Ness TJ, Slavoff SA. Small open reading frames and cellular stress responses. Mol Omics. 2019;15: 108–116.

491. Klinakis AG, Zagoraiou L, Vassilatis DK, Savakis C. Genome-wide insertional mutagenesis in human cells by the Drosophila mobile element Minos. EMBO Rep. 2000;1: 416–421.