université PARIS-SACLAY

Nouvelles approches et concepts pour l'étude des communautés microbiennes complexes

New approaches and concepts to study complex microbial communities

Thèse de doctorat de l'université Paris-Saclay
École doctorale n°577
Spécialité de doctorat : Sciences de la vie et de la santé
Unité de recherche : Université Paris-Saclay, Univ Évry, CNRS, CEA, Génomique métabolique, 91057, Évry, France
Référent : Université d'Évry Val d'Essonne

Thèse présentée et soutenue à Évry-Courcouronnes, le 06 Octobre 2021, par

Chloé BAUM

Composition du Jury

Olga SOUTOURINA Professeur, I2BC, Université Paris-Saclay	Présidente
Chantal TARDIF Professeur, CNRS, Université Aix-Marseille	Rapporteur
Vincent THOMAS Directeur de recherches, BIAOSTER, Université Paris Descartes	Rapporteur
Bruno DUPUY Directeur de recherches, Institut Pasteur, Université Paris Descartes	Examinateur
Marc MONOT Chargé de recherches, Institut Pasteur, Université Paris Descartes	Examinateur

Direction de la thèse

Véronique DE BERARDINIS Directrice de recherches, CEA-Genoscope, Université Paris-Saclay	Directrice de thèse
Laurence ETTWILLER Chargée de recherches, New England Biolabs	Co-encadrante de thèse Invitée
Andrew TOLONEN Chargé de recherches, CEA-Genoscope, Université Paris-Saclay	Co-encadrant de thèse Invité

Thèse de doctorat
NNT : 2021UPASL050

ACKNOWLEDGEMENTS

The present thesis work could not have been completed without the support of great mentors, amazing colleagues and caring family members and friends.

First, I would like to thank my thesis directors for their help and support during these 3 years. I would like to thank Marcel Salanoubat, thank you for your support all along the PhD, for your precious help regarding the thesis manuscript and for helping me deal with the last-minute issues that come with the end of the PhD. I wish you all the best, enjoy your retirement. Thank you to Véronique De Berardinis for accepting to replace Marcel as my thesis director a few months before my defense.

Then, I would like to thank my thesis supervisors, Laurence Ettwiller and Andrew Tolonen. Laurence, I have learnt so much from you during this PhD, not only scientifically but also on a more personnel aspect. You taught me to think as a scientist, to ask the good questions but also to be resilient and to not give up during the hard times. I remember when you told me 'Learning the hard way is the best way'. Thank you for your trust, your support and always pushing me to give the best of myself during this entire PhD. Thank you for helping me to deal with the ups and downs of science and life. There is a lot more I could say but I will stop there, finally, thank you for being such a great mentor. Andy, thank you for your help, your kindness and support. I have learnt a lot about the microbiome thanks to you. It's too bad we didn't get the opportunity to meet more often in Boston, but I always enjoyed the time we spent discussing science and other topics together.

Also, I would like to thank my thesis committee members Ashlee Earl and Marc Monot for the great discussions that helped moving my project forward.

Thank you to Chantal Tardif, Olga Soutourina, Vincent Thomas, Bruno Dupuy and Marc Monot for accepting to be members of my thesis jury.

During this PhD, I had the chance to collaborate with great scientists. I would like to thank Tuval Ben Yehezkel from Loop Genomics for his help on developing the Loop-Cappable-seq project. Thanks to David Vallenet, David Roche and Stéphanie Fouteau from the Labgem team at the Genoscope for their help with the Microscope platform and data analysis. It has been a pleasure collaborating with you on the microbiome project, thanks for all your help. I also would like to thank Tanya Yatsunenko at Kaleido Biosciences for helpful discussions regarding the defined community.

Then, I would like to thank my amazing NEB colleagues, Bo Yan and Weiwei Yang for their help and support. Not only did I meet great co-workers but also friends. Bo, you have been an amazing colleague, I learnt so much from you, you are an excellent scientist and a caring person. You helped me go through the hard times and even if I know we will keep in touch, I definitely will miss our tea times. Weiwei, thanks for your support, I really enjoyed working with you and hope there will be more of our virtual social hours to come! I also had great times discussing with you about France, China and sharing our cultures. Thanks for making me discover the excellent Chinese cuisine (shabu-shabu). I also would like to address a great thank you to Yu-Cheng Lin for helping me with bioinformatics. Even if it was only for a short time, you taught me so much in Python, and always in a very kind way.

I also would like to thank other NEB researchers I had the chance to meet and work with. Thank you to Ira Schildkraut, probably the most knowledgeable scientist I have ever met, thanks for always being so nice whenever I stopped by for help or questions. Thanks to George Tzertzinis for his help, it was always a pleasure to stop by to discuss science and other topics. Thanks to Peter Weigele and Yan-Jiun (YJ) Lee, for sharing their expertise on phages and for providing Xp12 gDNA. Thank you to Elisabeth Raleigh for her counsel during my PhD, you always had amazing advice and ideas. Thank you to Richard Roberts, Alexey Fomenkov and Brian Anton for their help and contribution on the RIMS-seq paper. It was a pleasure working with you. Thank you to Luo Sun for helping me start with the Nanopore sequencing. I will miss our discussions about sequencing technologies. Also, a great thank you to Sean Maguire, an amazing scientist to work with but also a great friend, thanks for your help on the development of the splint polyA ligation. Thank you to the NEB sequencing core, Laurie Mazzola, Danielle Fuchs and Kristen Augewitz for all the Illumina sequencing.

Thank you to the IT/helpdesk team, especially Ching-Lun Lin, Tamas Vincze and Aaron Messelaar for their help, their comprehension (and their patience!).

I also would like to thank the amazing friends I had the chance to meet during my PhD at NEB. Laudine Petralia, Léa Chuzel, Julie Zaworski, Emilie Lefoulon, Katell Kunin, Augusto Garcia and Youseuf Suliman, thank you for being such great friends and for all the adventures (and Blue Moon) we shared together. Thank you to the postdoc team as well, especially Ece Alpaslan and Sean Maguire. These 2 years in Ipswich and this PhD experience would not have been the same without you.

Je voudrais maintenant remercier toutes les personnes que j'ai eu la chance de rencontrer et avec qui j'ai pu échanger au Genoscope en France. Un merci tout spécial à Magali Boutard pour son soutien tout au long de mes années de thèse. Merci à Pedro Oliveira d'avoir partagé son expertise sur la méthylation de l'ADN et pour les échanges enrichissants. Merci à Corinne Cruaud, Dominique Robert et Eric Mahieu pour m'avoir permis d'implémenter le RIMS-seq au Genoscope. Merci à l'équipe informatique du Genoscope, je voudrais remercier Claude Discala-Verdier, Claude Scarpelli, Eric Doutreleau et Franck Aniere pour leur aide précieuse, surtout concernant le disque dur.

Un grand merci à ma super collègue de bureau, Marion Schulz. Merci de m'avoir soutenue et remotivée dans les moments pas toujours évidents (appelons ça des galères) qui viennent avec la dernière année de thèse. Merci pour les soirées sport Sissy, les bobun, le Ground Control et tout le reste. Je te souhaite le meilleur pour la suite de ta thèse, aucun doute que tu vas réussir ! Merci également à Oriane Monet pour les bons moments passés ensemble, il y en aura plein d'autres j'en suis sûre. Merci également à Laurine et toute la team doctorants du Genoscope. Je remercie aussi chaleureusement la "coffee team" : Jean-Louis, Ivan, Isabelle, Peggy, Nadia, Sébastien, Agnès, Aurélie, Anne... pour leur gentillesse et pour avoir toujours réussi à faire des pauses café un moment fun. Merci aussi à Christophe Lechaplais pour sa gentillesse et sa bienveillance.

Enfin, je voudrais remercier mes amis et mes proches, Bérengère, Alice, Hélèna, Camille, Pauline, Ophélie et mes supers voisins Lucille et Tom pour leur soutien sans faille durant cette (longue) aventure qu'est la thèse. Merci d'avoir été là, dans les hauts comme dans les bas, et d'avoir su me redonner confiance dans les moments difficiles. Sans oublier mon chat Blue, merci de m'avoir tenu compagnie pendant les longs moments de rédaction en télétravail.

Last but not least, je voudrais remercier ma famille. Merci à mes grands-parents Christine et Daniel, pour leur soutien infaillible. Je tiens tout particulièrement à remercier mon frère Nicolas et mes parents Jacques et Christelle. Merci de m'avoir toujours soutenue et de croire en moi, quels que soient mes choix. Merci de m'avoir permis d'être arrivée là où j'en suis aujourd'hui, rien de tout cela n'aurait été possible sans vous.

TABLE OF CONTENTS

ABBREVIATIONS

ANI: Average Nucleotide Identity

BSL: Biosafety Level

cDNA: Complementary DNA

COG: Cluster of Orthologous Genes

C. phy. *Clostridium phytofermentans*

DEG: Differentially Expressed Gene

ddNTP: Dideoxynucleotide Triphosphate

dNTP: Deoxynucleotide Triphosphate

DTB: Desthiobiotin

DNA: Deoxyribonucleic Acid

ESKAPE group: *Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, *Enterobacter spp* group

HGT: Horizontal Gene Transfer

HMP: Human Microbiome Project

HR: Homologous Recombination

IBD: Inflammatory Bowel Disease

Indels: Insertions and deletions

MetaHIT: Metagenomics of the Human Intestinal Tract

MIC: Minimum Inhibitory Concentration

NEB: New England Biolabs

NER: Nucleotide Excision Repair

NGS: Next-Generation Sequencing

NIH: The United States National Institutes of Health

ONT: Oxford Nanopore Technologies

OTU: Operational Taxonomic Unit

PBS: Phosphate-Buffered Saline

PCR: Polymerase Chain Reaction

PFAM: Protein Families (database)

QRDR: Quinolone Resistance-Determining Region

R1: Read 1 (illumina sequencing)

R2: Read 2 (illumina sequencing)

RDP: Ribosomal Database Project

RIN: RNA Integrity Number

RM: Restriction-Modification systems

RNA: Ribonucleic acid

RPKM: Read Per Kilobase of Transcript

rRNA: Ribosomal RNA

RNAP: RNA Polymerase

RIMS-seq: Rapid Identification of Methylase Specificity

SBS: Sequencing By Synthesis

SD: Shine-Dalgarno

SLR: Synthetic Long Read

SMRT: Single Molecule, Real-Time

SNP: Single-Nucleotide Polymorphism

SOLiD: Sequencing by Oligonucleotide Ligation and Detection

T2T: Telomere-to-Telomere consortium

TEX: 5'Phosphate-dependent RNA exonuclease

tRNA: Transfer RNA

TSS: Transcription Start Site

TTS: Transcription Termination Site

VCE: Vaccinia Capping Enzyme

VRE: Vancomycin Resistant Enterococcus

ZMW: Zero Mode Waveguides

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

The first part of this introduction retraces history of DNA sequencing technologies and how the development of high-throughput sequencing revolutionized the study of complex bacterial communities (called microbiomes) in diverse environments, from the central oceans to the human intestine. The second part of the introduction describes the gut microbiome, its importance to human health, and the development of technologies to characterize its composition.

I. DNA-sequencing

A. A short history of DNA sequencing

Deoxyribonucleic Acid (DNA) consists of a succession of nucleotides (A, C, T or G) whose sequence contains the information for the hereditary and biochemical properties of all living organisms on earth. Therefore, the ability to determine and analyze such sequences is crucial for biological research. Over years, researchers have tried to address the problem of how to study and sequence DNA. From this, three generations of methodologies and sequencers have emerged and their history will be reviewed in this part.

Sequencing of nucleic acids emerged in the mid-1960s with the sequencing of low-molecular weight RNAs such as tRNA. At that time, sequencing was performed using base-specific cleavage by ribonucleases, combined with analytical techniques for separating and isolating nucleic acids, such as chromatography and electrophoresis. In 1965, Robert Holley *et al* determined the first whole nucleic acid sequence, the alanine tRNA from *Saccharomyces cerevisiae* (Holley *et al.*, 1965; Heather and Chain, 2016), for which he was awarded the Nobel Prize in 1986.

1. First generation of DNA sequencing

It was the development in the 70's of two different methods that could decode hundreds of bases that revolutionized the field and represent the first methods for the determination of nucleotide sequences in DNA.

In 1977, Maxam and Gilbert described a new DNA sequencing method called "chemical cleavage procedure" (Maxam and Gilbert, 1977). This method used double-stranded DNA, base-specific cleavage and radioactive labeling of the DNA. The fragments generated are processed by gel electrophoresis for size separation and exposed to an X-ray source for visualization. The DNA sequence can be inferred by reconstituting the order of cleavage. In parallel, Frederick Sanger and his team developed a second method called "chain termination procedure". After his Nobel prize in 1958 for the discovery of the 51 amino acids of the human insulin, Frederick Sanger shared a Nobel prize with Walter Gilbert and Paul Berg in 1980 for their contributions to the determination of base sequences in nucleic acids. The Sanger method is based on the partial incorporation by a DNA polymerase of dideoxynucleotides (ddNTPs) that interrupt elongation of DNA sequences. DNA strands are synthesized in the presence of natural deoxynucleotides (dNTPs) and radiolabeled ddNTPs that are used as non-reversible synthesis terminators. The DNA synthesis reaction is randomly terminated when a ddNTP is added to the growing chain, resulting in truncated products of varying lengths. By performing four parallel reactions containing each individual ddNTP base and running the results on four lanes of a polyacrylamide gel, it is possible to determine the nucleotide sequence (Sanger, Nicklen and Coulson, 1977). This method allowed the sequence determination of the first complete genome: the phiX174 bacteriophage genome (Sanger *et al.*, 1977; Heather and Chain, 2016).

The Sanger method rapidly became the gold-standard for sequencing. However, this method lacked automation and was time-consuming, which led to the development of the first-generation of automated capillary DNA sequencers. In fact, the Sanger method has been continuously improved in different ways: DNA fragments were labeled with fluorescent dyes instead of radioactive molecules, electrophoresis and fluorescence detection were automated using robotic platforms and central data repositories (such as GenBank) and search tools (such as BLAST (Altschul *et al.*, 1990) were created,

facilitating data analysis and sharing (Shendure *et al.*, 2017). First in 1987, Applied Biosystems (ABI) with its ABI Prism 310 and then GE healthcare with its MegaBACE 1000 were the first to release automated DNA sequencers containing capillaries. Capillaries coated with a polymer are used to perform DNA separation during electrophoresis, which allows simultaneous electrophoresis and processing of up to 96 samples independently (Swerdlow and Gesteland, 1990).

These automated Sanger sequencers are considered as first-generation sequencers. But such platforms were limited by the read length, as they produced reads no larger than one kilobase (kb) and provided low throughput. So in 1979, in order to increase the length of DNA fragments that could be analyzed, Staden (Staden, 1979) suggested shotgun sequencing. Shotgun sequencing consists in sequencing overlapping DNA fragments that are cloned, sequenced separately, and assembled into one long contiguous sequence (contig) *in silico*. Such strategies combined with polymerase chain reaction (PCR) and automatic Sanger sequencing were used in the Human Genome Project that was initiated in 1990, helping to produce the first draft in 2001 entitled "Initial sequencing and analysis of the human genome" (Lander *et al.*, 2001). It should be noticed that even nowadays, the sequencing of the human genome is still ongoing. New data are continually added, notably with the Telomere-to-Telomere (T2T) Consortium, getting everyday closer to filling the gaps (Reardon, 2021).

2. Second generation of DNA sequencing (NGS)

Since 2005, Next-generation sequencing (NGS) technologies, also known as second-generation sequencing, have entered the market and rapidly replaced Sanger sequencing, as these new technologies allow a much higher throughput for DNA and cDNA sequencing. Instead of one tube per reaction, a complex library of DNA templates is immobilized onto a two-dimensional surface. Bacterial cloning is replaced by *in vitro* amplification that generates copies of each template to be sequenced. Finally, instead of measuring fragment lengths, sequencing comprises cycles of biochemistry (such as polymerase-mediated incorporation of fluorescently labelled nucleotides) and imaging, also known as sequencing by synthesis (SBS). These techniques generate millions to billions

of DNA molecules that can be sequenced in parallel, providing massively parallel analysis from one or multiple samples at a much reduced cost (Shendure *et al.*, 2017).

Pyrosequencing (454 from Roche)

This method, called pyrosequencing marked the transition phase from first generation to next generation sequencing. It uses a luminescent reaction to measure pyrophosphate synthesis. Pyrosequencing enables the determination of DNA sequence by measuring the pyrophosphate release as one nucleotide is incorporated. More specifically, individual dNTPs are added in a predetermined order and when a base is incorporated into the DNA, pyrophosphate is converted by the ATP sulfurylase to ATP which is then used as the substrate by the luciferase enzyme, producing a visible light signal with an intensity proportional to the amount of pyrophosphate. By plotting the pattern of light intensity for each base, the sequence of the original piece of DNA can be decoded (Metzker, 2010).

From this emerged the 454 Genome Sequencing System commercialized by Roche in 2005. The 454 sequencers were the first high-throughput sequencing systems on the market. but were limited by a low throughput (no more than 1 000 000 reads per sequencing run) and are not commercialized anymore (Morey *et al.*, 2013).

Sequencing by ligation (SOLiD from Applied Biosystems)

In 2007, Applied Biosystems introduced the SOLiD technology (Sequencing by Oligonucleotide Ligation and Detection). Like Sanger sequencing, the system is based on the detection of fluorescence signals with the difference being that while in Sanger sequencing a fluorophore is used for each nucleotide, in SOLiD sequencing a fluorophore is used for a given combination of two nucleotides (two base encoding system). This methodology relies on a sequential ligation of fluorescent probes (16 possible combinations of nucleotides 2 by 2) and thanks to the known color-space technique, it is possible to determine which nucleotide occupies each position (Garrido-Cardenas *et al.*, 2017). However, the slow sequencing pace and the short read length (75bp) generated by the SOLiD platform limited its use (Hert, Fredlake and Barron, 2008).

Sequencing by synthesis (Solexa/illumina)

A third approach, which is still up to date and widely used, is the one developed by Solexa, a company founded by Balasubramanian and Klenerman in 1998. The polymerase-mediated Sequencing by synthesis (SBS) involves the incorporation of fluorescently labelled deoxynucleotides by an engineered polymerase. The key of this method was the development of reversible terminating fluorescent dNTPs such that each template incorporates a single dNTP on each cycle. After imaging to determine which of four colours was incorporated by each template on the surface, both blocking and fluorescent groups are removed to set up the next extension. With the Solexa method, the sequencing adapters are attached to DNA molecules, allowing the molecules to bind to a flowcell. Once bound to the flowcell, DNA templates will be clonally amplified by a solid phase PCR called "bridge amplification". During this amplification, single-stranded DNA containing terminator sequences complementary to oligonucleotides on the flowcell replicate in a confined area and then bend over to prime at neighboring sites, producing a local cluster of identical molecules. Clusters can be visualized by detecting fluorescent reversible-terminator nucleotides at the ends of each extension reaction, requiring cycle-by-cycle measurements and the removal of terminators (Garrido-Cardenas *et al.*, 2017; Shendure *et al.*, 2017; Sessegolo *et al.*, 2019).

In 2006, the Genome Analyzer (GA) was released, giving the power to sequence 1 gigabase (Gb) of data in a single run. Numerous bacterial, plant, human, and animal genomes were sequenced with this technology. In 2007, Solexa was acquired by Illumina. Advances in Illumina's technology over the years have largely set the pace for the tremendous gains in output and reductions in cost. Illumina is today the world's first company on the NGS market. Illumina sequencing technology uses four different fluorescent-labelled nucleotides to sequence the millions of clusters generated present on the flow cell surface through sequencing by synthesis. These nucleotides, specially designed to possess a reversible termination property, allow each cycle of the sequencing reaction to occur simultaneously in the presence of all four nucleotides (A, T, C, G). During each cycle, the polymerase can select the correct base to incorporate, with the natural competition between all four alternatives present in the reaction mix. Following scanning of the flow cell with a coupled-charge device (CCD) camera to determine the added nucleotide, the fluorescent moiety and the 3' block are removed, and the process is repeated (**Figure 1**).

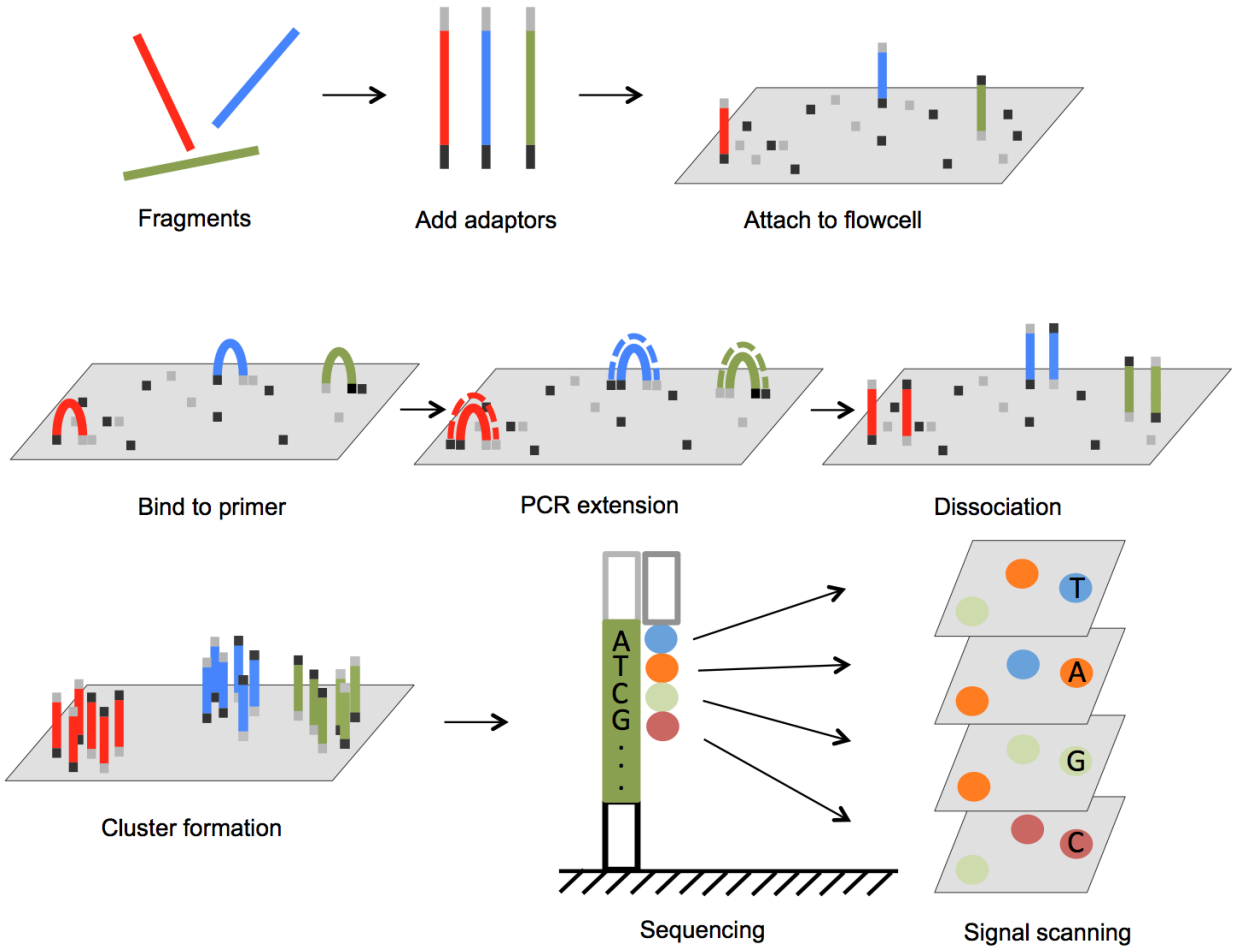


Figure 1: Principle of the illumina sequencing by synthesis (SBS) technology (Lu et al., 2016)

The addition of a unique barcode incorporated into the adapter added to the DNA fragment allows for multiple samples from different sources to be pooled and sequenced together, greatly enhancing the throughput of such techniques. Illumina currently produces a suite of sequencers (iSeq, MiniSeq, MiSeq, NextSeq, HiSeq and NovaSeq) developed for a variety of applications depending on the throughput and the read length needed, ranging from 4 million reads/run (1.2Gb) using the MiniSeq to 20 billion reads/run (6000Gb) using the NovaSeq6000.

Semiconductor sequencing (Ion Torrent from Life Technologies)

In 2010, Life Technologies launched the Ion Torrent, a fast and low-cost sequencer based on semiconductor technology. With the Ion Personal Genome Machine (PGM), nucleotide sequences are detected electronically by measuring the changes in the pH of the surrounding solution caused by the release of H+ protons during polymerization. The released H+ ions are proportional to the number of incorporated nucleotides. While the Ion PGM is designed for small genomes, the Ion Proton allows whole genome, transcriptome and exome sequencing. In 2015, Life Technologies released the Ion S5, designed for targeted sequencing workflows such as metagenomics analyses. But these sequencers are prone to insertion and deletion errors (indels) during sequencing and have troubles with homopolymer sequences (Glenn, 2011).

Synthetic long reads (SLRs)

The development of massively parallel short-read NGS sequencing permits the acquisition of high-throughput DNA sequences. However, it should be stressed that this incredible throughput is associated with a shortening of the read length which is at most 2x150bp for most Illumina platforms, with an exception for the MiSeq that has the ability to provide sequences up to 2x300bp (MiSeq V3 kits only). To circumvent this issue, diverse sample preparation protocols have emerged (Wu *et al.*, 2014; Stapleton *et al.*, 2016), enabling synthetic long reads (SLRs) to be constructed from short-reads. Those synthetic approaches rely on specific library preparations that use barcodes that will allow computational assembly of short fragments sharing the same barcode into a larger fragment (McCoy *et al.*, 2014).

Illumina first commercialized the TruSeq Synthetic Long-Read technology (also known as Moleculo), which allows construction of synthetic long reads from the short reads generated with the HiSeq platform (Li *et al.*, 2015). Another platform developed by 10X Genomics, called Chromium, relies on an emulsion-based system to partition and barcode the DNA fragments (Goodwin, McPherson and McCombie, 2016). Lastly, the LoopSeq platform developed by Loop Genomics, enables contiguous short read coverage of DNA molecules. Briefly, DNA molecules are barcoded with Unique Molecular Identifiers (UMIs) that are then intramolecularly distributed throughout the molecule. After fragmentation and sequencing, short reads that share the same UMI are assembled to reconstruct

the sequence of the full-length fragment (Callahan *et al.*, 2021; Liu *et al.*, 2021). However, one drawback of the SLRs technologies is their relatively high cost due to the high coverage required, compared to a typical short-read sequencing project.

3. Third generation of DNA sequencing

One major drawback of the second generation sequencing is the read length limitation. As a consequence, analysis of complex genomes, including repetitive regions, are difficult. Ideally, sequencing would be native, accurate and without read-length limitations.

Third-generation sequencing technologies address these issues as they endeavor to provide long-read sequencing. Single molecule sequencing technologies consider the sequencing of a single DNA molecule without the need for preliminary amplification. Consequently, biases, errors and information loss (such as loss of DNA methylation and modifications) that are related to DNA amplification are avoided (Kulski, 2016; Shendure *et al.*, 2017). In addition, longest read lengths, highest consensus accuracy, uniform coverage, real-time sequencing and single molecule resolution are possible. Long reads cDNA sequencing can also be useful for transcriptome analysis as they can span entire mRNA transcripts, allowing the identification of gene isoforms (Byrne *et al.*, 2019; Zhao *et al.*, 2019). Single-molecule real-time approaches differ from short read approaches as they do not rely on a clonal amplification of DNA fragments to generate a detectable signal (Goodwin, McPherson and McCombie, 2016). Currently, the leaders in the single-molecule real-time sequencing field are the technologies from Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT).

PacBio sequencing

In 2011, Pacific Biosciences launched the SMRT (Single-Molecule, Real-Time) DNA sequencing method. The principle is to optically observe polymerase-mediated synthesis in real time. The platform consists of nanostructures called Zero Mode Waveguides (ZMW) containing a single polymerase protein immobilized to the bottom (Levene *et al.*, 2003).

Template preparation involves ligation of single-stranded hairpin adapters onto the ends of DNA or cDNA molecules, generating a circular template (called SMRT-bell template). Once this circular DNA

is coupled with the DNA polymerase, fluorescently labelled nucleotides enter the ZMW and as each nucleotide is incorporated, the label is cleaved off and diffuses out of the ZMW. The ZMWs are continuously monitored using cameras and a series of pulses are converted into single molecular traces corresponding to the template sequence (**Figure 2**). By using a strand displacing polymerase, the original DNA molecule can be sequenced multiple times, increasing accuracy. Importantly, clonal amplification is avoided, allowing direct sequencing of native and potentially modified DNA.

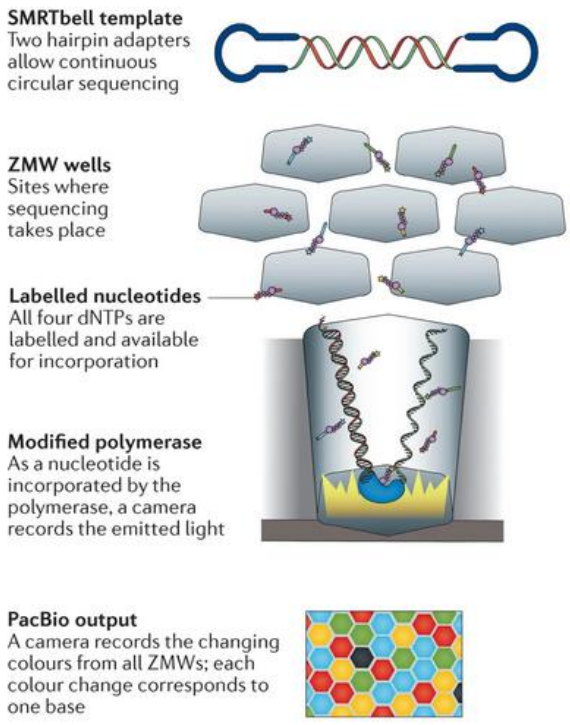


Figure 2: Principle of Single-molecule real-time (SMRT) sequencing from PacBio (Goodwin, McPherson and McCombie, 2016)

Different platforms have been developed through the years, the first platform being the PacBio RS II. This sequencer is now being replaced by the Sequel systems (Sequel II was launched in 2019) as they provide higher throughput, more scalability and lower sequencing project costs compared to the PacBio RS II System.

Oxford Nanopore sequencing

Nanopore-based sequencing is a strategy developed by Oxford Nanopore Technologies (ONT) and commercialized for the first time in 2014. The concept, which was first imagined in the 1980s, is based on the idea that the passage of a single-stranded DNA or RNA molecule through a nanopore channel subjected to a continuous current, will provoke specific current disruptions that can be used to detect the sequence composition (Deamer, Akeson and Branton, 2016). Whereas other sequencing platforms use a secondary signal derived from DNA synthesis (light, color or pH), nanopore sequencers detect electric signal fluctuation of a biopolymer that passes through the nanopore channel, opening up the door to very exciting and revolutionary applications, such as direct RNA-sequencing (Garalde *et al.*, 2018) or even protein sequencing (Ouldali *et al.*, 2020).

Oxford Nanopore's technology consists in a sequencing flow cell composed of hundreds independent micro-wells, each containing a synthetic bilayer perforated by nanopores. Library preparation is minimal, involving only fragmentation of DNA and ligation of adapters, PCR being optional. During sequencing, double stranded DNA gets denatured by a helicase enzyme that brings one DNA strand through one of the nanopores embedded in the synthetic membrane, across which a voltage is continuously applied (**Figure 3**). As the ssDNA passes through the nanopore, the different bases prevent ionic flow in a specific manner, allowing sequencing of the molecule by measuring characteristic changes in voltage at each channel, meaning sequencing happens in real-time (Clarke *et al.*, 2009; Reuter, Spacek and Snyder, 2015).

Multiple Nanopore sequencing devices have been developed through the years, starting with the MinION, the very first USB-powered portable sequencing device, slightly larger than a USB key, that provides up to 50Gb of data in 72h. Then, the GridION was developed, allowing to run up to five MinION flowcells at the same time. In 2019, a smaller flowcell, called Flongle, adaptable to the MinION and GridION was launched. The Flongle provides a smaller throughput than regular cells (2Gb), and is interesting for example for quick test experiments or sequencing of small genomes. Currently, the larger device of Oxford Nanopore is called PromethION and promises high throughput for large genomes sequencing (such as human), as well as highly multiplexed sequencing. Up to 48 flowcells can be run at the same time, theoretically providing up to 14Tb of data in 72h. The current

longest read record is held by Nanopore: 2.3Mb of continuous DNA molecule sequence generated with the MinION (Payne *et al.*, 2019).

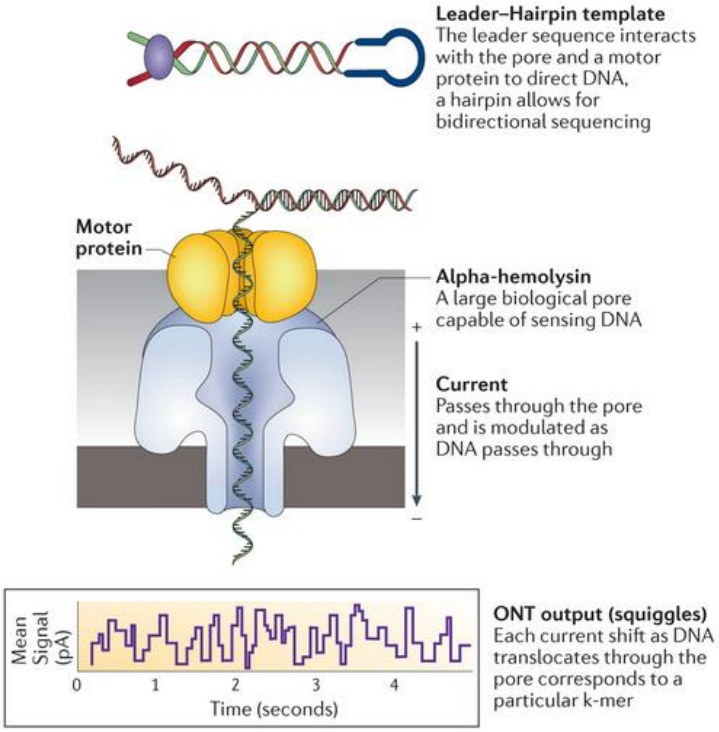


Figure 3: Principle of Nanopore sequencing developed by Oxford Nanopore Technologies (Goodwin, McPherson and McCombie, 2016).

Still, important challenges remain for long-read technologies. Although these platforms generate longer reads than the second generation sequencers (Illumina), PacBio and Oxford Nanopore sequencers are subjected to higher rates of sequencing error. Currently, PacBio has the ability to generate higher-quality data compared to Nanopore because the circular nature of the DNA in SMRT-bell library allows for multiple sequencing of the same starting molecule. But progress is rapid and in a few years only, basecalling accuracy of reads produced by both these technologies have drastically increased (Amarasinghe *et al.*, 2020). The raw base-called error rate is claimed to have been reduced to < 1% for PacBio sequencers (Wenger *et al.*, 2019) and < 5% for nanopore sequencers (M. Jain *et al.*, 2018). **Table 1** below resumes the features of the main sequencing platforms presented in the introduction.

	Platform	Sequencing chemistry	Max read length	Run time	Most frequent error type	Error rate (%)
1st generation	**Sanger capillary sequencing**	Size separation of specifically end-labeled DNA fragments	1000 bp	2h	substitutions	0.1
2nd generation	**454/Roche**	Sequencing by synthesis (Pyrosequencing)	1000 bp	10-24h	indels	1
	Ion Torrent	Sequencing by synthesis (Semiconductor sequencing)	400 bp	2-7h	indels	1
	SOLiD (Applied Biosystems)	Sequencing by ligation	75 bp	7d	substitutions	0.1-1
	Illumina (all platforms)	Sequencing by synthesis	300 bp	3d (average)	substitutions	0.1
3rd generation	**Pacific Biosciences (Sequel)**	Single Molecule, Real-Time (SMRT)	20 kb	0.5 - 30h	indels	<1
	Oxford Nanopore (MinION)	Nanopore sequencing	2.3 Mb	1min - 72h	indels	<5

Table 1: Comparison of the different generation of sequencing platforms presented in this thesis. Table adapted from multiple reviews (Shendure and Ji, 2008; Mestan et al., 2011; Fox and Reid-Bayliss, 2014; Reinert et al., 2015; Garrido-Cardenas et al., 2017).

Since 1977, DNA-sequencing technologies have evolved at an impressive pace and continue to progress rapidly. Although Illumina is still dominating the sequencing market, other technologies have emerged and expanded the scope of applications, for example PacBio used for *de novo* assembly of complex genomes and Nanopore bringing portable sequencing and revolutionary direct RNA sequencing. Next-generation DNA sequencing has the potential to accelerate biological and biomedical research, by enabling the comprehensive analysis of genomes and transcriptomes at continuously decreasing costs, enabling routine and widespread use of sequencing technologies. Together, these technologies bring huge research and applications potential, for clinical but also environmental research, with the possibility for real-time pathogen identification. Applied to environmental and microbial research, the possibility of in-field sequencing brings every day new knowledge on the microbial diversity surrounding us.

B. Deciphering the biology of complex bacterial communities

1. History and evolution of Microbiology

The term Microbiology was first introduced by Louis Pasteur around 1880. Microbiology can be defined as the study of microorganisms, all the living organisms that are too small to be visible with the naked eye. At the beginning of microbiology, the microscope was the main tool to study microorganisms and their interactions with the host. Later, the development of staining techniques such as Gram or Ziehl–Neelsen significantly improved their analysis and it was rapidly found that these microorganisms needed special conditions to grow. Robert Koch is credited for developing the first microbial isolation techniques. He is at the origin of the concept of bacterial colony and postulates that a colony forms from a single colonizing bacterium. His research allowed for the first pure bacterial culture experiments and the development of culture media adapted to the different types of bacteria. Koch grew the first bacterial colonies on thin potato slices, leading to the isolation of the etiological agent of anthrax, *Bacillus anthracis* in 1877. Ten years later, Robert Koch's assistant Julius Richard Petri, expanded on Koch's potato slices and invented what is now a basics in Microbiology: the Petri dish. From these breakthroughs and in the span of thirty years, emerged the "Golden age of Microbiology", during which the principal bacterial pathogens of human diseases were identified (**Table 2**) (Blevins and Bronze, 2010).

Koch's Legacy: The Discoverers of the Main Bacterial Pathogens

Year	Disease	Organism	Discoverer
1877	Anthrax	*Bacillus anthracis*	Koch, R.
1878	Suppuration	*Staphylococcus*	Koch, R.
1879	Gonorrhea	*Neisseria gonorrhoeae*	Neisser, A.L.S.
1880	Typhoid fever	*Salmonella typhi*	Eberth, C.J.
1881	Suppuration	*Streptococcus*	Ogston, A.
1882	Tuberculosis	*Mycobacterium tuberculosis*	Koch, R.
1883	Cholera	*Vibrio cholerae*	Koch, R.
1883	Diphtheria	*Corynebacterium diphtheriae*	Klebs, T.A.E., Loeffler, F.
1884	Tetanus	*Clostridium tetani*	Nicholaier, A.
1885	Diarrhea	*Escherichia coli*	Escherich, T.
1886	Pneumonia	*Streptococcus pneumoniae*	Fraenkel, A.
1887	Meningitis	*Neisseria meningitidis*	Weischselbaum, A.
1888	Food poisoning	*Salmonella enteritidis*	Gaertner, A.A.H.
1892	Gas gangrene	*Clostridium perfringens*	Welch, W.H.
1894	Plague	*Yersinia pestis*	Kitasato, S., Yersin, A.J.E. (independently)
1896	Botulism	*Clostridium botulinum*	van Ermengem, E.M.P.
1898	Dysentery	*Shigella dysenteriae*	Shiga, K.
1900	Paratyphoid	*Salmonella paratyphi*	Schottmüller, H.
1903	Syphilis	*Treponema pallidum*	Schaudinn, F.R., and Hoffmann, E.
1906	Whooping cough	*Bordtella pertussis*	Bordet, J., and Gengou, O.

Table 2: Main bacterial pathogens identified during the "Golden age of microbiology" (Blevins and Bronze, 2010).

So, for a long time, the study of microorganisms was based on morphology features, growth, and selection of some biochemical profiles (Maloy and Schaechter, 2006). Still, this approach provided a limited insight into the microbiological world, as the main focus was on bacterial pathogens and only cultivable bacteria could be studied. But in the late 1970s, Carl Woese proposed a revolutionary idea by suggesting the use of ribosomal RNA genes as a phylogenetic marker for bacterial classification (Woese *et al.*, 1985). Then, advances in molecular techniques, such as polymerase chain reaction (PCR), quantitative PCR (qPCR), cloning and sequencing, fluorescent in situ hybridization (FISH), restriction-fragment length polymorphism (RFLP), and terminal restriction-fragment length polymorphism (T-RFLP), revolutionized microbiology (Escobar-Zepeda, de León and Sanchez-Flores, 2015). These techniques opened considerable research paths, as it was now possible to characterize the "dark side of the microbiological world", the one of uncultivable bacteria.

In 1977, the 16S rRNA classification proposed by Carl Woese coupled to automated Sanger DNA sequencing revolutionized microbiology and the analysis of bacterial communities (Woese and Fox, 1977). Comparison of the 16S rRNA gene sequences has shown that this gene is highly

conserved within organisms of the same genus and species, but that they differ between organisms of other genera and species. Most prokaryotes contain 16S rRNA gene which is composed of 9 hypervariable regions flanked by conserved sequences (Yang, Wang and Qian, 2016). Thanks to the sequencing of both 16S rRNA gene from bacteria and 18S rRNA gene from eukaryotes, three domains of life: Archaea, Bacteria and Eukarya, were described (Woese, Kandler and Wheelis, 1990). In addition, the development of PCR and the design of primers that can be used to amplify almost the entire 16S rRNA gene lead to the discovery of numerous novel bacterial genus and species and more importantly, rendered the discovery, identification and classification of uncultivable bacteria possible (Handelsman, 2005; Woo *et al.*, 2008). The use of 16S gene sequencing will be further discussed in the part "Current techniques to characterize the gut microbiome composition".

2. The birth of Metagenomics and the exploration of bacterial diversity

As 16S rRNA gene sequencing studies expanded the understanding of microbial diversity and ecology and pushed microbiology towards the era of culture-independent studies, the microbial diversity seemed endless. But limits of amplicon sequencing began to arise as the technique was limited to phylogenetic applications and could not give any insight into microbial function. The idea of cloning DNA directly from environmental samples was first proposed by Pace (Pace *et al.*, 1986) and in 1996, Stein *et al* pushed the field forward with the first attempt of metagenomic sequencing in Hawaiian ocean water (Stein *et al.*, 1996). In this study, the authors attempted to clone large genomic DNA fragments isolated from ocean water into *E. coli* fosmid vectors. The clones were initially screened for archaeal DNA fragments by amplifying 16S rRNA genes content from the fragments and the selected clones were then sequenced and analyzed.

Yet, it is only a few years later that the term “metagenomics” was invented and defined for the first time by Handelsman *et al (Handelsman et al., 1998)*. This term refers to the culture-independent analysis of collective genomes from environmental samples. Metagenomic analysis consists in creating metagenomics libraries by: (1) isolating DNA from an environmental sample, (2)

cloning the DNA into a suitable vector, (3) transforming the clones into a host bacterium, and (4) screening the resulting clones. The clones can be screened for expression of specific traits, such as enzyme activity or antibiotic production (function-driven approach), for phylogenetic markers such as 16S rRNA, for conserved genes or can even be sequenced randomly (sequence-driven approach) (**Figure 4**). Metagenomics opened the possibility of discovering unknown sequences and functions from the environment without isolating or identifying individual organisms.

Together, these approaches provided new insights into bacterial diversity of various ecosystems by analyzing both cultivable and uncultivable bacteria (Handelsman, 2005; Srivastava, Ghosh and Pal, 2013), but the best is yet to come.

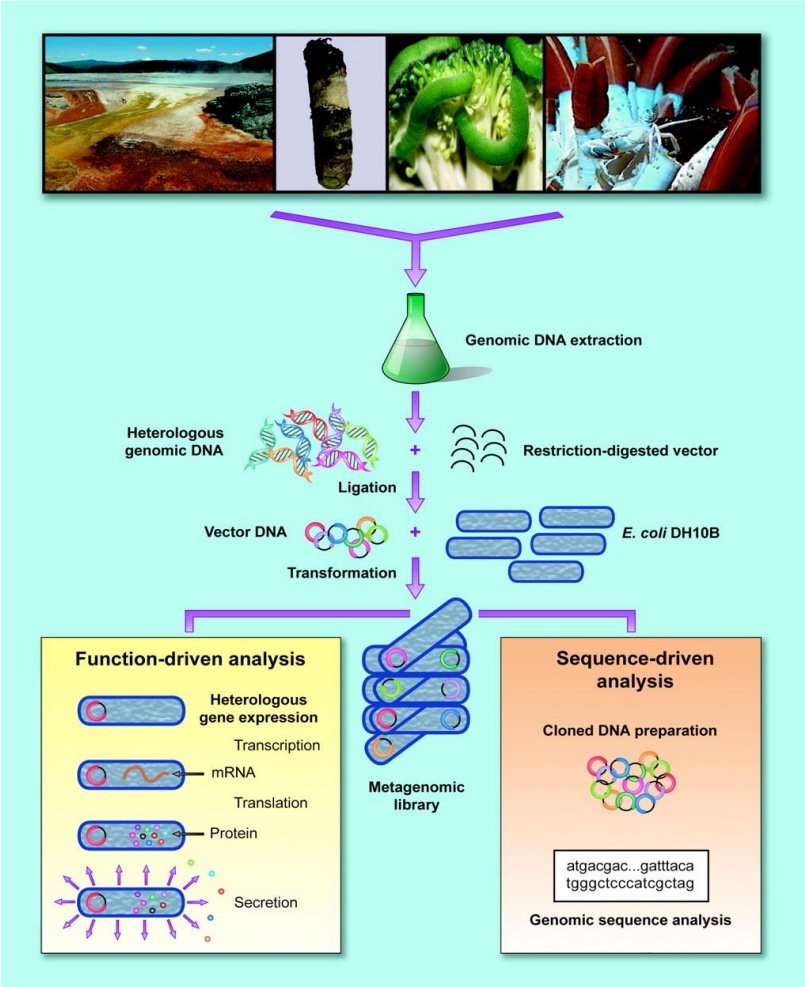


Figure 4: Schematic representation of construction of libraries from environmental samples (Pace et al., 1986; Handelsman, 2005).

3. Next-Generation Sequencing (NGS) and metagenomics

Over the years, technological advances have driven revolutions in microbiology. Since the first decade of the 2000s, the current revolution has been driven by novel DNA sequencing technologies called Next-Generation Sequencing (NGS). These new sequencing platforms provide high speed and high-throughput that can produce an enormous volume of data. The most important advantage provided by these platforms is their ability to determine the sequence from single DNA fragments of a library without the need for cloning. These techniques, accompanied by new bioinformatic approaches, brought a whole new level to metagenomics.

Two striking examples illustrate well the power of NGS to enrich our understanding of uncultured communities: the studies from Venter et al. on the Sargasso sea (Venter *et al.*, 2004) and from Tyson et al. on acid mine drainage (Tyson *et al.*, 2004). These studies have provided new linkages between phylogeny and function, shown the surprising abundance of certain types of genes, and reconstructed the genomes of organisms that couldn't be cultured.

The advance of high-throughput technologies also allowed the development of functional metagenomics, where the screening is based on enzyme activity and not sequence similarity to known enzymes. In such workflows, environmental DNA, cloned into vectors, can be screened for expression of a desired enzyme activity, using the appropriate substrate. This approach has become very popular to identify brand new enzyme activities and proteins from nature (Ngara and Zhang, 2018).

Together, such studies have shown the exciting potential of metagenomics to provide compositional and functional information community-wide from diverse environmental sample types. **Figure 5** below illustrates the key steps in the evolution of Microbiology to Metagenomics.

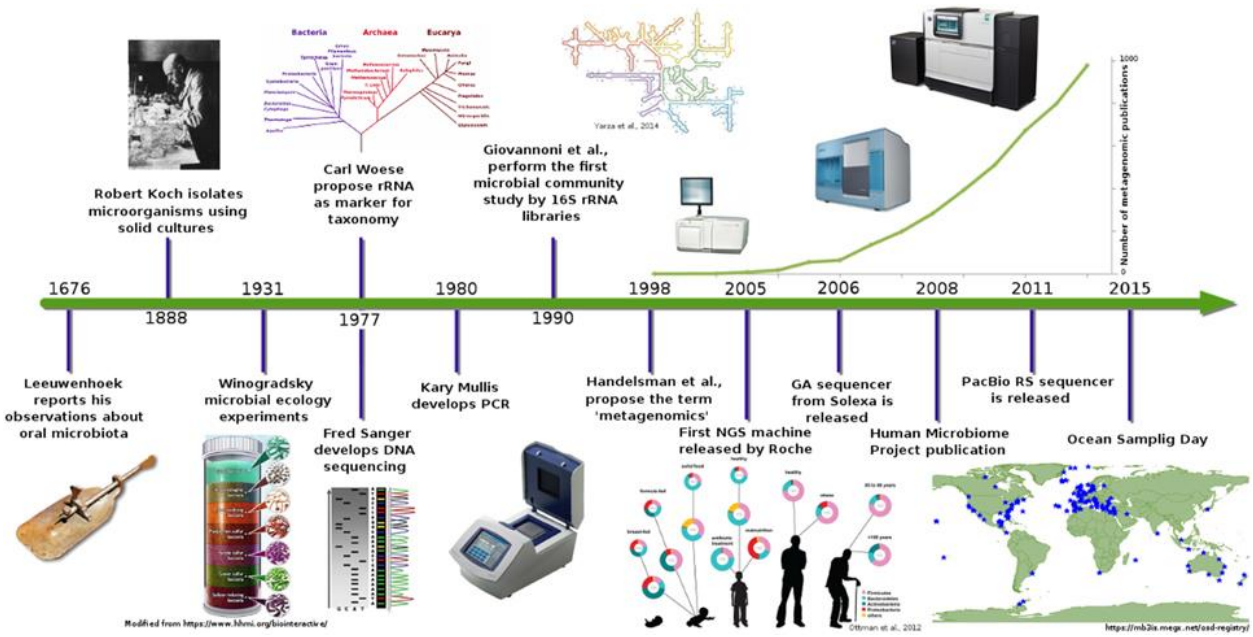


Figure 5: The road to Metagenomics (Escobar-Zepeda, de León and Sanchez-Flores, 2015).

II. The microbiome

A. The gut microbiome and human health

1. Definition

The fast evolution of sequencing techniques and the advent of metagenomics have led to the exploration of bacterial communities in different environments, from central oceans to the human gut. In the last decades, the microbiome field has exponentially expanded, bringing with it revolutionary discoveries. The term 'human microbiome' refers to the collective genomes of the microbes (bacteria, bacteriophage, fungi, protozoa and viruses) that live inside and on various sites of the human body (Consortium and The Human Microbiome Project Consortium, 2012b). Examples of occupied habitats include our oral cavity, genital organs, respiratory tract, skin, gastrointestinal system, and lungs (O'Dwyer, Dickson and Moore, 2016; Kho and Lal, 2018). The organ that contains most of the bacterial cells is the gastrointestinal tract, with an estimated $3.8x10^{13}$ of microbial cells. In a healthy individual, the mass of the gut microbiome is estimated to be 200 grams (Zhernakova *et al.*, 2016). The gut microbiome is predominantly composed of bacteria from three phyla: *Firmicutes,*

Bacteroidetes, and *Actinobacteria* (Tap *et al.*, 2009). This diverse and complex microbiome is considered another body organ and is estimated to harbor 150 fold more genes compared to the human host (Qin *et al.*, 2010). These extra genes add important functions not encoded by the host and play a critical role in host metabolism and physiology (Hooper and Gordon, 2001). Thus, the microbiome functions in tandem with the host, playing a pivotal role in critical processes such as aging, digestion, immunity, protection against pathogen colonization, and essential metabolic functions (**Figure 6**).

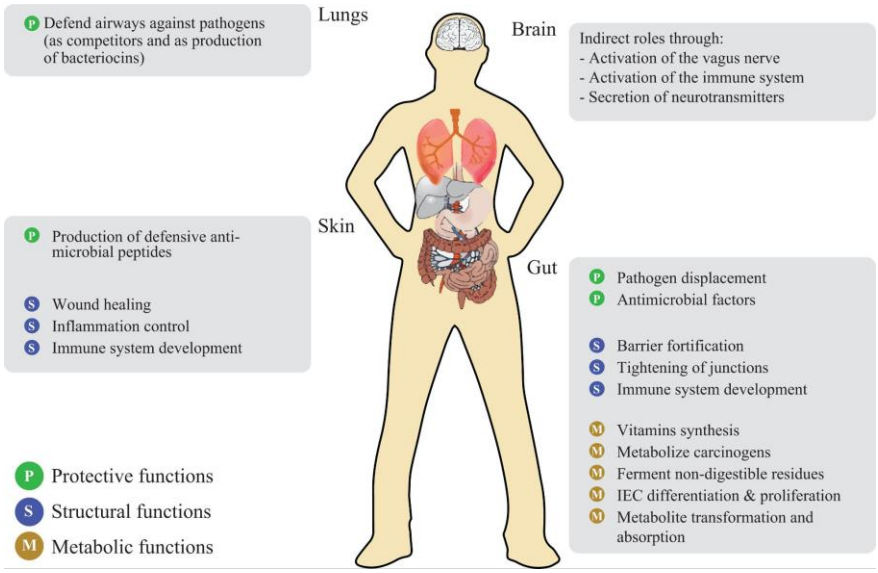


Figure 6: Main functions of bacteria in the human body. IEC: intestinal epithelial cell (Scotti et al., 2017).

While the role of the human microbiome is now considered essential, the composition of the human microbiome is far from being universal and highly varies within and between individuals depending on a variety of factors (**Figure 7**). The composition of the intestinal microbiome varies within an individual, depending on the anatomic site (stomach, small intestine, colon…), age (Bosco and Noti, 2021), sex (Kim *et al.*, 2020) or genetics (Goodrich *et al.*, 2014; Cahana and Iraqi, 2020). In addition, microbiome composition is strongly influenced by environmental factors (diet, geography, stress, medication), resulting in a high inter-individual variability. Both intra and inter-individual variability will be presented in the next section.

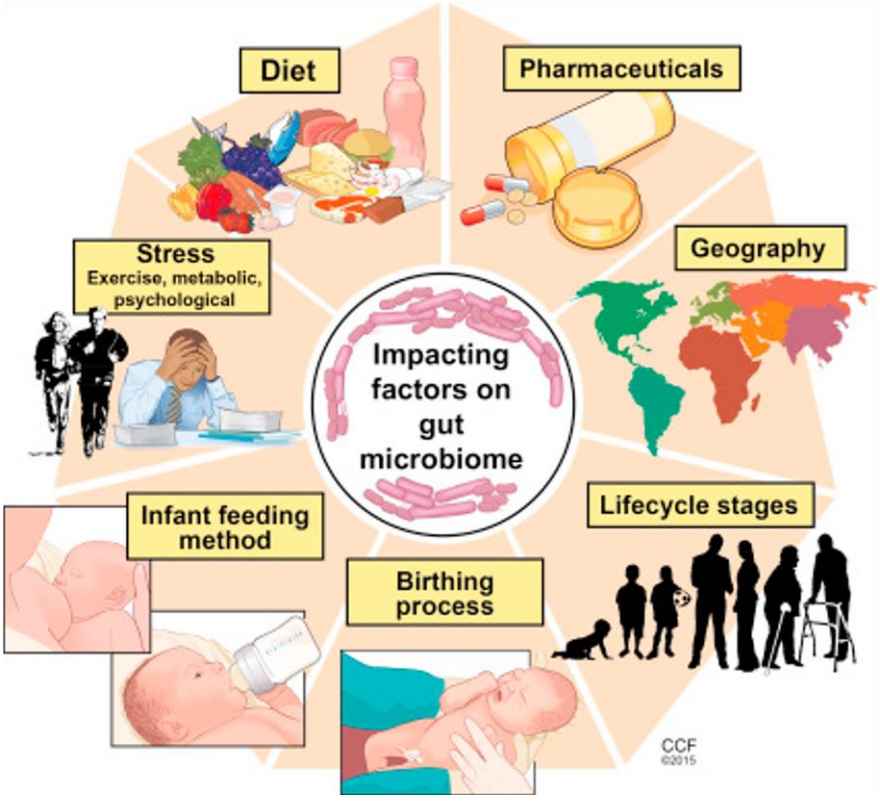


Figure 7: Factors impacting the human gut microbiome all along life ('Gut Microbiome', 2019).

2. Variability of the microbiome composition

While the fetus is considered sterile *in utero*, the microbial colonization of the newborn starts during birth and depends on the mode of delivery. Infants born by 'natural way' (vaginal birth) are colonized by the gut and vaginal microbiome of the mother, while the infants born through assisted delivery (C-section) are colonized by the skin microbiome of the mother (Rodríguez *et al.*, 2015). The difference in microbiome composition is especially marked among infants but tends to converge to more similar phyla later in life; it is considered to be similar to an adult microbiome by the age of 3 years (Palmer *et al.*, 2007; 'Gut Microbiome', 2019). Still, studies have demonstrated a great diversity of the gut microbiome composition between adults, which may depend on a large number of host and environmental factors such as age, sex, diet, medication, diseases and diet (Ley *et al.*, 2008; Healey *et al.*, 2017).

Below are some examples of factors inducing variability in the microbiome composition.

Age: compared to adults, the microbiome of the elderly is characterized by a decrease in bacterial diversity, with a decrease in Firmicutes coupled to an increase of the *Bacteroidetes.* These alterations in the gut microbiome during aging could provide a favorable environment for growth of pathogens, such as *Clostridium difficile* (Kumar *et al.*, 2016).

Spatial localization: The human digestive tract is composed of different organs, each one harboring different characteristic in terms of pH, oxygen concentration, and motility. The bacterial concentration in the stomach is relatively low (10^3 cell/mL) due to a very acidic pH (pH 2), while some acid tolerant species such as *Helicobacter pylori* can reside there. The microbial concentration progressively increases through the small intestine, with 10^4 cells/mL in the duodenum/jejunum and 10^8 cells/mL in the ileum, but the mucus present there is composed of antimicrobials that prevent a high colonization. The highest bacterial concentration is reached in the large intestine (colon) with 10^{11} cells/mL (see **Figure 8** for the structure of the colon). This can be explained by the fact that it takes up to 36h for the food residues to transit out of the colon, providing important energy sources for the development of micro-organisms (Canny and McCormick, 2008; Sender, Fuchs and Milo, 2016). In addition, an oxygen gradient is formed horizontally along the colon crypts, favoring obligate anaerobe growth in the anoxic intestinal lumen and facultative anaerobe growth closer to the intestinal epithelium. The presence of mucus, a permeable gel that lubricates and protects the colon epithelial cells against microbial invasion, adds another factor of selectivity, as it harbors a specific bacterial community composition (De Weirdt and Van de Wiele, 2015; Kastl *et al.*, 2020).

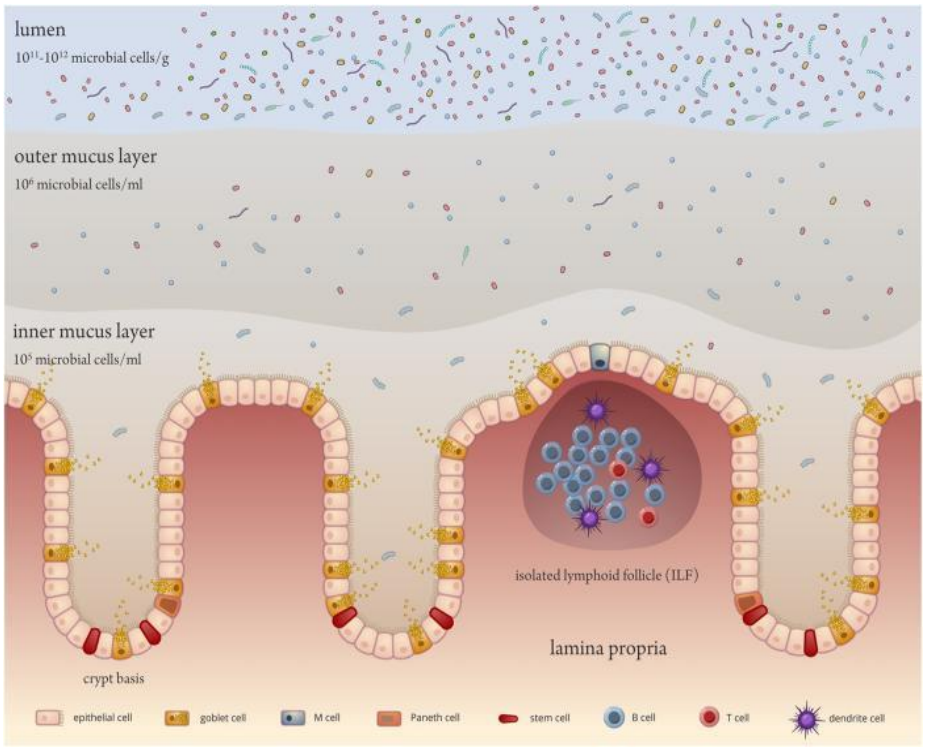


Figure 8: Structure of the colon and the differences in microbial colonisation (De Weirdt and Van de Wiele, 2015).

Diet: Numerous studies from humans and mouse models demonstrated that diet plays a major role in shaping the gut microbiome. For example, a plant-based diet has been associated with a higher abundance of *Bacteroidetes* and *Firmicutes* while an animal-based diet has been linked with a higher abundance of *Bacteroides (Ghaisas, Maher and Kanthasamy, 2016)*. Another study comparing the microbiome composition of children from Burkina Faso to Europeans, demonstrated that individuals from Burkina Faso consuming a high-fiber diet are enriched for *Prevotella* (*Bacteroidetes*) when compared with Europeans individuals consuming a Western diet (Filippo *et al.*, 2010).

3. Relationship between the gut microbiome and human health

Over the last two decades and thanks to the development of culture-independent methods like sequencing, there has been a growing interest in characterizing the microbiome in healthy individuals as well as in diseased states. Since 2007, large-scale sequence-based microbiome projects such as the Human Microbiome Project (HMP) consortium funded by The United States National Institutes of Health (NIH), and the MetaHIT (Metagenomics of the Human Intestinal Tract) consortium funded by the European Commission, have greatly improved research on the human microbiome. Both these large-scale projects aim at characterizing the human microbiome and its role in human health and diseases. Together, these consortia greatly helped setting up a framework for microbiome research (Consortium and The Human Microbiome Project Consortium, 2012a; Integrative HMP (iHMP) Research Network Consortium, 2014). Such studies have shown that all along our life, many factors shape our gut microbiome, altering its diversity and composition. Disruption of body homeostasis mediated by alteration of the microbiome is called dysbiosis. Dysbiosis is characterized by an imbalance in bacterial composition that can occur as a loss of beneficial bacteria (commensals), a decrease in the microbial diversity and richness, and an increase in potentially pathogenic strains (Mahnic *et al.*, 2020). Dysbiosis is likely to impair the normal functioning of gut microbiota in maintaining host well-being and has been associated with a wide-range of diseases and inflammatory disorders (**Figure 9**), including obesity, inflammatory bowel disease (IBD), allergies, diabetes, cardiovascular diseases and colorectal cancer, in both human and animal models (DeGruttola *et al.*, 2016; Kho and Lal, 2018).

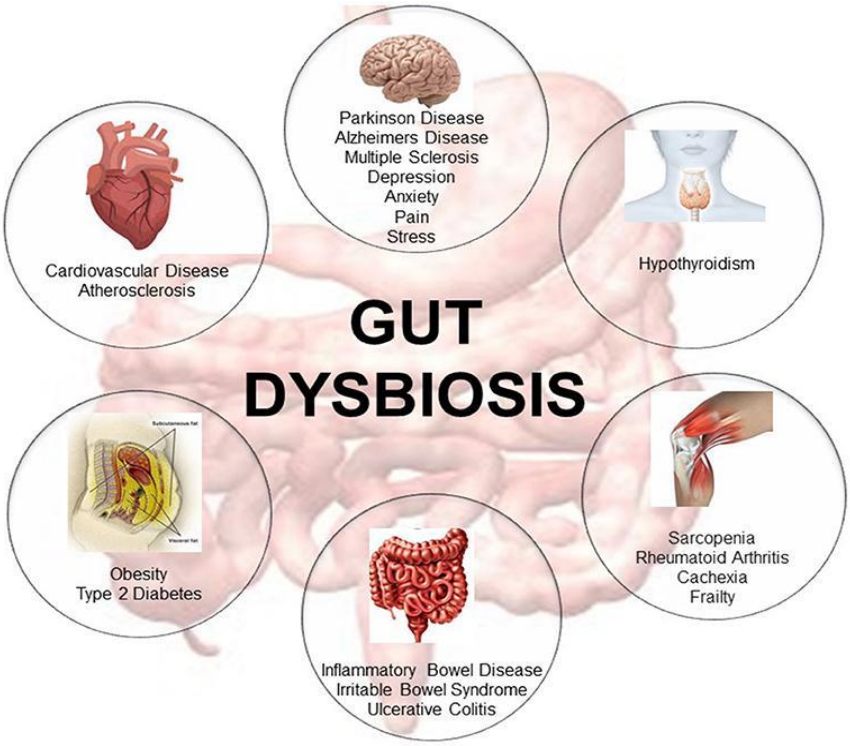


Figure 9: Description of diseases associated with gut dysbiosis (Baptista et al., 2020).

Several approaches have been developed to promote health by alleviating dysbiosis through changes to the composition of the gut microbiome. The use of 'nutritional tools' such as probiotics and prebiotics have been shown to alter the composition of the gut microbiota in favor of bacteria that improve gut barrier integrity (Madsen *et al.*, 2001; Leclercq *et al.*, 2014) and reduce inflammation (Delzenne *et al.*, 2011). In addition, Fecal Microbiota Transplantation (FMT), defined as the infusion of stool from a healthy individual to a patient (Brandt and Aroniadis, 2013)), is being actively developed. FMT aims at repopulating the gut with a healthy microbiome and has been successfully used in the treatment of dysbiosis-associated diseases such as *Clostridium difficile* infection (CDI), Inflammatory Bowel Disease (IBD), autoimmune disorders or even obesity (Choi and Cho, 2016; Kim and Gluck, 2019).

4. Antibiotics, dysbiosis of the gut microbiome and health consequences

In addition to combating pathogens, antibiotics can also inhibit growth of health-promoting commensals, thereby altering the taxonomic, genomic, and functional capacity of the human gut microbiota, with effects that are fast and sometimes persistent. Broad-spectrum antibiotics reduce bacterial diversity, select for resistant bacteria, increase opportunities for horizontal gene transfer (HGT), and open niches for intrusion of pathogenic organisms by removing commensals (Modi, Collins and Relman, 2014). For example, a study in adults found that five days of ciprofloxacin treatment impacted the abundance of about a third of the bacterial taxa in the gut and decreased taxonomic richness within days of exposure, leading to an overall decreased bacterial diversity. Many taxa returned to typical abundances in 4 weeks following the exposure, but some compositional changes lasted for six months (Dethlefsen *et al.*, 2008; Modi, Collins and Relman, 2014). Similar results have been observed in antibiotic-treated mice (Ubeda *et al.*, 2010). Increasing evidence suggests that antibiotics treatments in infants have a profound effect on the gut microbiome and can result in the later development of obesity (Trasande *et al.*, 2013), asthma (Patrick *et al.*, 2020), IBD and other disorders (Ledder, 2019).

5. Antibiotics and resistance

In addition to disrupting the human microbiome, extensive use of antibiotics has led to strong selective pressure for the emergence of pathogens able to resist antibiotic treatment. Antibiotic resistance is quickly becoming a worldwide significant health care problem as there is a real concern that existing antibiotics will become ineffective against these pathogens (Sommer *et al.*, 2017). Even the use of low or very low concentrations of antimicrobials (sub-inhibitory concentrations) can lead to selection of resistance in bacteria, increase the mutation rate, promote the movement of mobile genetic elements and thus, increase the ability of bacteria to acquire resistance (Blázquez *et al.*, 2012). Consequently, identifying the mechanisms of bacterial resistance is central to understanding its emergence.

Resistance can either be pre-existing (resistance genes already present in the bacteria) or acquired. Acquired resistance can arise spontaneously due to mutations in bacterial genomes (Bagel *et al.*, 1999), or it can emerge from the horizontal gene transfer (HGT) from one bacterial cell to another. This transfer of genetic material can happen through transformation, transposition, and conjugation (all part of the HGT mechanism). The bacteria can also acquire resistance through mutations to its own chromosomal DNA. This acquisition can be temporary or permanent, with plasmid-mediated transmission of resistance genes being the most common route for acquisition of outside genetic material (Reygaert *et al.*, 2018). There are four main types of antimicrobial resistance mechanisms: (1) limiting uptake of a drug, (2) modifying a drug target, (3) inactivating a drug and (4) active drug efflux. Because of differences in structure, the types of mechanisms used by Gram negative bacteria versus gram positive bacteria vary. Gram negative bacteria use of all four main mechanisms, whereas membrane-based resistance based on limited uptake and drug efflux are less common in Gram positive bacteria as they lack the LPS outer membrane (Tamaki, Sato and Matsuhashi, 1971), and don't have the capacity to use certain types of drug efflux mechanisms (Kapoor et al. 2017; Chancey et al. 2012).

B. Current techniques to characterize the gut microbiome composition

As we previously reviewed, an adequate balance of the gut microbiota is critical in maintaining the health status of the host and this fragile equilibrium can be impaired by various external factors, including antibiotics that have been described as a major cause of dysbiosis. Therefore, being able to identify the composition and the compositional changes following disruptions of the gut microbiome is crucial. Compositional changes and relative abundances in microbiomes have been well characterized thanks to two main techniques that will be presented in this section: 16S rRNA sequencing and shotgun metagenomics sequencing.

1. 16S rRNA gene sequencing

Ribosomal RNA genes are highly conserved, stable through evolution, and contain both conserved and hypervariable regions. Consequently, the 16S rRNA gene is commonly used as a marker for the characterization of microbial community diversity. More specifically, this 1500 bp gene contains nine hypervariable regions (ranging from V1 to V9) as well as highly conserved regions (Johnson *et al.*, 2019). The V3-V4 regions (Fadrosh *et al.*, 2014) and V4 region of the 16S rRNA gene have been recommended for profiling of human gut microbiomes (Qin *et al.*, 2010; Lozupone *et al.*, 2013). Sequencing of the 16S rRNA gene requires PCR amplification of a selected variable region using universal primers, followed by sequencing of the PCR amplicons. The 16S rRNA gene sequence has been determined for numerous bacterial strains, allowing comparison of gene sequences from strains of interest to databases of reference sequences. After sequencing, the hypervariable regions are used to discriminate bacteria between each other and to assign taxonomy. There are currently three main 16S databases used for taxonomic assignment: Greengenes (McDonald *et al.*, 2012), SILVA (Yilmaz *et al.*, 2014) and RDP (Wang *et al.*, 2007). The 16S bioinformatics pipelines were intensively developed during the last few years and today, pipelines such as QIIME allow straightforward 16S analysis, from the raw sequences to visualization (Caporaso *et al.*, 2010). Thus, 16S metagenomics provides a fast and cost-effective approach for bacterial taxonomic estimation (Osman *et al.*, 2018), even with a relatively small number of raw reads (as low as 18,000–20,000 reads per sample) (Kozich *et al.*, 2013). However, most 16S analyses are based on the Illumina platform (such as MiSeq), which produces paired-end reads only up to 2x300bp, limiting the 16S sequencing to a partial analysis. Consequently, short-read 16S analysis are unable to provide the taxonomic resolution achieved if the full 16S gene was sequenced. The development of long-read sequencing technologies, such as Nanopore or PacBio, recently provided a solution to sequence the full 16S gene, allowing for better taxonomic resolution (Wagner *et al.*, 2016; Nygaard *et al.*, 2020).

While 16S gene sequencing is a powerful tool for taxonomic analysis, the technique presents some biases. First, the choice of primers used to amplify the gene is critical, as it can lead to potential biases towards certain organisms, resulting in underrepresentation of some species or even whole groups (Klindworth *et al.*, 2013). In addition, the technique does not provide enough resolution to identify bacteria at the species/strain level, often only providing taxonomy at the genus level. Even when

sequencing the entire gene, the high error rate of long-read technologies can fail to provide accurate species resolution (Ardui *et al.*, 2018). Finally, the information obtained through 16S analysis is limited to taxonomic profiling and functional profiling (quantification of gene and metabolic pathway content) is not possible.

2. Shotgun metagenomics sequencing

In contrast to targeted 16S amplicon sequencing, shotgun metagenomics sequences all of the genomic DNA present in a sample. The library preparation workflow is similar to regular whole genome sequencing, including random fragmentation and adapter ligation. In addition to offering species level classification of bacteria, shotgun sequencing covers all genetic information in a sample. Thus, shotgun data can be used for additional analyses such as genome assembly, functional profiling, and antibiotic resistance gene profiling (Jovel *et al.*, 2016; de Abreu, Perdigão and Almeida, 2020). However, one drawback of shotgun metagenomics lies in the cost associated with high sequencing coverage required. The higher cost of shotgun compared to 16S sequencing can be a limiting factor for its use in routine analysis of microbiomes or large-scale projects (Rausch *et al.*, 2019). Table 3 below summarizes the main different features of each sequencing approach.

	16S sequencing	Shotgun sequencing
Taxonomy resolution	Genus-Species	Species-Strains
Host DNA interference	No	Yes
Functional profiling	No	Yes
Minimum DNA input	10 copies of 16S	1ng
Cost per sample	$80	$200

Table 3: Features of 16S sequencing and shotgun sequencing approaches. Table adapted from Zymo Research website (16S Sequencing vs Shotgun Metagenomic Sequencing, 2021).

3. The need for functional characterization

Thanks to the advent of sequencing technologies, it is now possible to identify the composition of microbiomes. However, metagenomics approaches are DNA-based and only answer the question "which bacteria and genes are present in the sample?". Thus, functional RNA-based approaches are necessary to provide functional characterization of microbiomes to complement our understanding of microbial communities' dynamics by answering the question "how are the bacteria responding and what are they doing?". Metatranscriptomics allows the investigation of functional activities of microbiomes by providing information on the genes that are expressed in complex communities. Such data allow the study of the microbiome-host interactions and derive metabolic pathways, enabling to explore the effect of different environments on bacterial activities and provide a better understanding of what can lead a healthy microbiome towards a dysbiosis or disease status (Bashiardes, Zilberman-Schapira and Elinav, 2016). As an example, it can take up to several days to observe the effect of a perturbation on the composition of the microbiome. In the case of antibiotic treatment, major changes in bacterial composition can be observed from 3 days following the treatment. The study from Abeles *et al* reported the majority of the diversity reductions in response to amoxicillin and azithromycin antibiotics occurred within the first 3 to 7 days of therapy (Abeles *et al.*, 2016). Conversely, it has been shown that transcriptional responses are among the earliest changes observed within minutes of antibiotic exposure, reflecting the ability of bacteria to rapidly acclimate to environmental perturbations. For example, global changes in gene expression (transcriptional reprogramming) were observed as soon as 5 min after injection of antibiotics in *E. coli* (Sangurdekar, Srienc and Khodursky, 2006).

Several studies have exploited transcriptome signatures and identified RNA markers enabling prediction of antibiotic susceptibility in pathogenic species such as *Neisseria gonorrhoeae (Khazaei et al., 2018)*, *Klebsiella pneumoniae* and *Acinetobacter baumanii* (Bhattacharyya *et al.*, 2017). In the case of *N. gonorrhoeae*, significant shifts in transcripts levels have been identified as soon as 10min after ciprofloxacin treatment. Such RNA signatures represent a promising approach to rapidly provide a phenotypic profiling for antibiotic susceptibility of pathogens. This could enable better and rapid adaptation of antibiotic treatment, according to the phenotype of the bacteria (Bhattacharyya *et al.*, 2017).

However, most of these studies have been done in mono-cultured bacteria and to our knowledge, studies on rapid transcriptional response (within minutes of treatment) have not been done in complex microbial communities. It is important to study the rapid transcriptional response to antibiotics in a complex community in order to identify those mRNA responses that best correlate with long term changes in community structure. One of the challenges of such a study lies in the ability to grow complex bacterial communities. To partly address this challenge, well-characterized synthetic communities have been developed, allowing better control and reproducibility. The reconstruction of synthetic gut microbial communities makes it easier to understand the structure and functional activities of more complex communities present in the human gut (Mabwi *et al.*, 2021). Prior knowledge of the genomic composition of the community also facilitates mapping-based identification of mRNA reads. Metagenomics and metatranscriptomics are thus complementary. Such multiomic approaches applied to defined synthetic communities offer the ability to link compositional changes to transcriptional response and to potentially identify early RNA marker that could predict the compositional outcome following external perturbations, such as antibiotic treatment.

OBJECTIVES OF THE PHD

As presented in the introduction, one of the major limitations of current microbiome studies has been the seldom integration of functional data such as transcriptomics to complement the interpretation of metagenomics (compositional) data. The research aim of my thesis is to develop new sequencing-based technologies and apply them to provide further insights into changes to the composition and activities of microbiomes. Specifically, Chapter One presents RIMS-seq (Rapid Identification of Methylase Specificity), a method to simultaneously obtain the DNA sequence and 5-methylcytosine (m5C) profile of bacterial genomes. Chapter Two introduces ONT-cappable-seq and Loop-Cappable-seq, two new techniques to reveal operon architecture through full-length transcript sequencing using Nanopore and LoopSeq sequencing, respectively. Finally, in Chapter Three, we applied a multi-omics approach using some of the tools developed in the previous chapters to study the dynamics of the response of a model human intestinal microbiome after treatment with ciprofloxacin, a widely used broad-spectrum antibiotic. We examined both the short and long-term transcriptional and genomic responses of the synthetic community and explored how the immediate transcriptomic response correlates and potentially predicts the later changes of the microbiome composition. We asked several questions: (1) can we identify an immediate transcriptional reprogramming in a complex community? (2) are bacteria from the same family responding the same way? Is there a phylum-specific response? (3) is there a specific response of the bacteria that will resist the treatment vs the susceptible ones? (4) and ultimately, can we identify some transcriptional markers (specific genes or pathways differentially expressed) that could be used to predict the outcome of the treatment?

RESULTS

Each of the 3 chapters will be organized and written as in a scientific publication format (Introduction, Material and methods, Results, Discussion) and contains one or more papers, published or in preparation.

Chapter I : Rapid Identification of Methylase Specificity (RIMS-seq) jointly identifies methylated motifs and generates shotgun sequencing of bacterial genomes

Author contribution

- Rapid Identification of Methylase Specificity (RIMS-seq) jointly identifies methylated motifs and generates shotgun sequencing of bacterial genomes.
  Baum C, Lin YC, Fomenkov A, Anton B, Chen L, Yan B, Evans TC, Roberts RJ, Tolonen AC, Ettwiller L. *Nucleic Acids Research, 2021, accepted*
  Preprint available in *bioRxiv https://doi.org/10.1101/2021.03.08.434449*

The manuscript has been published on August, 20th and is available in the **Appendix**.

This project is based on an idea of Laurence Ettwiller, Tom Evans and Lixin Chen. It was initiated by a postdoc Yu-Cheng Lin and I took off the project when he left the lab. I designed and performed the RIMS-seq, Bisulfite-seq and DNA-seq experiments. The EM-seq and MFRE-seq experiments were done by Alexey Fomenkov and Brian Anton and they also did some of the RIMS-seq experiments. Laurence Ettwiller wrote the analysis pipeline (published on Github https://github.com/Ettwiller/RIMS-seq) and I performed the RIMS-seq analysis using this pipeline. I also did all the Bisulfite-seq data analysis. Laurence and I wrote the manuscript and generated the figures together, I published the data on NCBI and participated in the publication of the RIMS-seq analysis pipeline on Github. When I moved back to the Genoscope in France, I successfully transferred the method and implemented RIMS-seq in the lab with two technicians. The sequencing lab of the Genoscope hopes to use this method routinely in the future. The manuscript has been published in *Nucleic Acids Research.*

Summary

DNA methylation is known to modulate gene expression in eukaryotes but is also widespread in prokaryotes, in which it confers viral resistance. Specifically, 5-methylcytosine (m5C) methylation has been described in genomes of various bacterial species as part of restriction-modification (RM) systems, each composed of a methyltransferase and cognate restriction enzyme. Methylases are site-specific and their target sequences vary across organisms. High-throughput methods, such as Bisulfite-sequencing (Bisulfite-seq) can identify m5C at base resolution but require specialized library preparations and genome assembly is not possible from these data. PacBio Single Molecule, Real-Time (SMRT) Sequencing is able to provide the DNA-sequence as well as the methylation information at the same time, but usually misses m5C. Here, we aimed at developing a new method that allows, similarly to PacBio, to get the DNA sequence and characterize m5C methylation of the genomes simultaneously, from one single library. In the following manuscript we present RIMS-seq (Rapid Identification of Methylase Specificity), a new method to simultaneously sequence bacterial genomes and determine m5C methylase specificities that is based on a simple and straightforward protocol that is very similar to the regular Illumina DNA-seq protocol.

Rapid Identification of Methylase Specificity (RIMS-seq) jointly identifies methylated motifs and generates shotgun sequencing of bacterial genomes

Chloé Baum[1,2], Yu-Cheng-Lin[1], Alexey Fomenkov[1], Brian P. Anton[1], Lixin Chen[1], Bo Yan[1], Thomas C. Evans Jr,[1] Richard J. Roberts[1], Andrew C. Tolonen[2], Laurence Ettwiller[1]

New England Biolabs, Inc. 240 County Road Ipswich, MA 01938, USA

Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91000 Évry, France

Abstract

DNA methylation is widespread amongst eukaryotes and prokaryotes to modulate gene expression and confer viral resistance. 5-methylcytosine (m5C) methylation has been described in genomes of a large fraction of bacterial species as part of restriction-modification systems, each composed of a methyltransferase and cognate restriction enzyme. Methylases are site-specific and target sequences vary across organisms. High-throughput methods, such as bisulfite-sequencing can identify m5C at base resolution but require specialized library preparations and Single Molecule, Real-Time (SMRT) Sequencing usually misses m5C. Here, we present a new method called RIMS-seq (Rapid Identification of Methylase Specificity) to simultaneously sequence bacterial genomes and determine m5C methylase specificities using a simple experimental protocol that closely resembles the DNA-seq protocol for Illumina. Importantly, the resulting sequencing quality is identical to DNA-seq, enabling RIMS-seq to substitute standard sequencing of bacterial genomes. Applied to bacteria and synthetic mixed communities, RIMS-seq reveals new methylase specificities, supporting routine study of m5C methylation while sequencing new genomes.

A. Introduction

DNA modifications catalyzed by DNA methyltransferases are considered to be the most abundant form of epigenetic modification in genomes of both prokaryotes and eukaryotes. In prokaryotes, DNA methylation has been mainly described as part of the sequence-specific restriction modification system (RM), a bacterial immune system to resist invasion of foreign DNA (1). As such, profiling methylation patterns gives insight into the selective pressures driving evolution of their genomes.

Around 90% of bacterial genomes contain at least one of the three common forms of DNA methylation: 5-methylcytosine (m5C), N4-methylcytosine (m4C), and N6-methyladenine (m6A))(2, 3). Contrary to eukaryotes where the position of the m5C methylation is variable and subject to epigenetic states, bacterial methylations tend to be constitutively present at specific sites across the genome. These sites are defined by the methylase specificity and, in the case of RM systems, tend to be fully methylated to avoid cuts by the cognate restriction enzyme. The methylase recognition specificities typically vary from 4 to 8 nucleotides and are often, but not always, palindromic (4).

PacBio Single Molecule, Real-Time (SMRT) sequencing has been instrumental in the identification of methylase specificity largely because, in addition to providing long read sequencing of bacterial genomes, m6A and m4C can easily be detected using the characteristic interpulse duration (IPD) of those modified bases (5). Thus, a single run on PacBio allows for both the sequencing and assembly of unknown bacterial genomes and the determination of m6A and m4C methylase specificities. However, because the signal associated with m5C bases is weaker than for m6A or m4C, the IPD ratio of m5C is very similar to the IPD of unmodified cytosine. Thus, PacBio sequencing misses the m5C methylases activities (2).

Instead, the identification of m5C requires specialized methods such as bisulfite sequencing or enzyme-based techniques such as EM-seq (8). Recently, MFRE-Seq has been developed to identify m5C methylase specificities in bacteria (10). MFRE-Seq uses a modification-dependent endonuclease that generates a double-stranded DNA break at methylated sites, allowing the identification of m5C

for the subset of sites. Unlike PacBio sequencing, these specialized methods do not provide the dual original sequence and methylation readouts from a single experiment.

Recently, m5C in the CpG context has been identified (11) and a signal for methylation can be observed at known methylated sites in bacteria using Nanopore sequencing (12)(13)(12). So far, no technique permits from a single experiment, the dual sequencing of genomes and the *de novo* determination of m5C methylase specificity for the non-CpG contexts typically found in bacteria.

Herein we describe a novel approach called RIMS-seq to simultaneously sequence bacterial genomes and globally profile m5C methylase specificity using a protocol that closely resembles the standard Illumina DNA-seq with a single, additional step. RIMS-seq shows comparable sequencing quality as DNA-seq and accurately identifies methylase specificities. Applied to characterized strains or novel isolates, RIMS-seq *de novo* identifies novel activities without the need for a reference genome and also permits the assembly of the bacterial genome at metrics comparable to standard shotgun sequencing.

B. Material and Methods

Samples and genomic DNA collection

Skin microbiome genomic DNA (ATCC® MSA-1005) and gut microbiome genomic DNA (ATCC® MSA-1006) were obtained from ATCC. *E. coli* BL21 genomic DNA was extracted from a culture of *E. coli* BL21 DE3 cells (C2527, New England Biolabs) using the DNEasy Blood and Tissue kit (69504, Qiagen). *E. coli* K12 MG1655 genomic DNA was extracted from a cell culture using the DNEasy Blood and Tissue kit (69504, Qiagen). All the other gDNA from the bacteria presented in **Table 1** were isolated using the Monarch genomic DNA purification kit (T3010S, New England Biolabs). Xp12 phage genomic DNA was obtained from Peter Weigele and Yian-Jiun Lee at New England Biolabs.

RIMS-seq library preparation

100ng of gDNA was sonicated in 1X TE buffer using the Covaris S2 (Covaris) with the standard protocol for 50µL and 200bp insert size.

The subsequent fragmented gDNA was used as the starting input for the NEBNext Ultra II library prep kit for Illumina (E7645, New England Biolabs) following the manufacturer's recommendations until the USER treatment step. The regular unmethylated loop-shaped adapter was used for ligation. After the USER treatment (step included), the samples were subjected to heat alkaline deamination: 1M NaOH pH 13 was added to a final concentration of 0.1M and the reactions were placed in a thermocycler at 60˚C for 3h. Then, the samples were immediately cooled down on ice and 1M of acetic acid was added to a final concentration of 0.1M in order to neutralize the reactions.

The neutralized reactions were cleaned up using the Zymo oligo clean and concentrator kit (D4060 Zymo Research) and the DNA was eluted in 20µL of 0.1X TE.

PCR amplification of the samples was done following NEBNext Ultra II library prep kit for Illumina protocol (ER7645, New England Biolabs) and the NEBNext® Multiplex Oligos for Illumina®(E7337A, New England Biolabs). The number of PCR cycles was tested and optimized for each sample following the standard procedure for library preparation. PCR reactions were cleaned up using 0.9X NEBNext Sample purification beads (E7137AA, New England Biolabs) and eluted in 25µL of 0.1X TE. All the libraries were evaluated on a TapeStation High sensitivity DNA1000 (Agilent Technologies) and paired-end sequenced on Illumina.

Bisulfite-seq library preparation

1% of lambda phage gDNA (D1221, Promega) was spiked-into 300ng gDNA to use as an unmethylated internal control. The samples were sonicated in 1X TE buffer using the Covaris S2 (Covaris) with the standard protocol for 50µL and 200bp insert size.

The subsequent fragmented gDNA was used as the starting input for the NEBNext Ultra II library prep kit for Illumina (E7645, New England Biolabs) following the manufacturer's recommendations until the USER treatment step. The methylated loop-shaped adapter was used for ligation. After USER, a 0.6X clean-up was performed using the NEBNext Sample purification beads (E7137AA, New England Biolabs) and eluted in 20µL of 0.1X TE. A TapeStation High Sensitivity DNA1000 was used to

assess the quality of the library before subsequent bisulfite treatment. The Zymo EZ DNA Methylation-Gold Kit (D5005, Zymo Research) was used for bisulfite treatment, following the manufacturer's suggestions.

PCR amplification of the samples was done following the suggestions from NEBNext Ultra II library prep kit for Illumina (ER7645, New England Biolabs), using the NEBNext® Multiplex Oligos for Illumina®(E7337A, New England Biolabs) and NEBNext® Q5U® Master Mix (M0597, New England Biolabs).

The number of PCR cycles was tested and optimized for each sample. The PCR reactions were cleaned up using 0.9X NEBNext Sample purification beads (E7137AA, New England Biolabs) and eluted in 25µL of 0.1X TE. All the libraries were screened on a TapeStation High sensitivity DNA1000 (Agilent Technologies) and paired-end sequenced on Illumina.

RIMS-seq data analysis

Paired-end reads were trimmed using Trim Galore 0.6.3 (option --trim1). The *Acinetobacter calcoaceticus* ATCC 49823 data have been trimmed using Trim Galore version 0.6.3 instead and downsampled to 1 million reads. Reads were mapped to the appropriate genome using BWA mem with the paired-end mode (version 0.7.5a-r418 and version 0.7.17-r1188 for the *Acinetobacter calcoaceticus*). When using an assembled genome directly from RIMS-seq data, trimmed RIMS-seq reads were assembled using SPAdes (SPAdes-3.13.0 (31)default parameters). Reads were split according to the read origin (Read 1 or Read 2) using samtools (version 1.8) with -f 64 (for Read 1) and -f 128 (for Read 2) and samtools mpileup (version 1.8) was run on the split read files with the following parameters: -O -s -q 10 -Q 0. For *Acinetobacter calcoaceticus*, the unmapped reads, reads without a mapped mate and the non-primary alignments were filtered out using the flags -F 12 and -F 256.

De-novo identification of motifs using RIMS-seq

Programs and a detailed manual for the *de-novo* identification of motifs in RIMS-seq are available on github (https://github.com/Ettwiller/RIMS-seq/). Using the mpileup files, positions and 14bp flanking regions in the genome for which a high quality (base quality score ≥ 35) C to T in R1 or a G

to A in R2 was observed were extracted for the foreground. Positions and 14bp flanking regions for which a high quality (base quality score ≥ 35) G to A in R1 or a C to T in R2 was observed were extracted for the background. C to T or G to A in the first position of reads were ignored. If the percentage of C to T or G to A are above 5% for at least 5 reads at any given position, the position was ignored (to avoid considering positions containing true variants). Motifs that are found significantly enriched (p-value < 1e-100) in the foreground sequences compared to background sequences were found using mosdi pipeline mosdi-discovery with the following parameters: '*mosdi-discovery -v discovery -q x -i -T 1e-100 -M 8,1,0,4 8 occ-count*' using the foreground sequences with x being the output of the following command: '*mosdi-utils count-qgrams -A "dna"* ' using the background sequences.

To identify additional motifs, the most significant motif found using *mosdi-discovery* is removed from the foreground and background sequences using the following parameters: '*mosdi-utils cut-out-motif -M X*' and the motif discovery process is repeated until no motif can be found.

Sequence logo generation

Using the mpileup files, positions in the genome for which a high quality (base quality score ≥ 35) C to T in R1 or a G to A in R2 was observed were extracted for the foreground using the get_motif_step1.pl program. Positions for which a high quality (base quality score ≥ 35) G to A in R1 or a C to T in R2 was observed were extracted for the background. The +/- 7bp regions flanking those positions were used to run Two sample logo (32). Parameters were set as t-test, p-value < 0.01.

Bisulfite-seq data analysis

Reads were trimmed using Trim Galore 0.6.3 and mapped to the bisulfite-converted concatenated reference genomes of each respective synthetic microbiome using bismark 0.22.2 with default parameters. PCR duplicates were removed using deduplicate_bismark and methylation information extracted using bismark_methylation_extractor using default parameters. For the microbiome, the bismark_methylation_extractor with --split_by_chromosome option was used to output one methylation report per bacterium. The motif identification was done as previously described in (10).

EM-seq

EM-seq was performed according to the standard protocol (NEB E7120S)

Motif identification was done as previously described in (10).

Analysis and abundance estimation in synthetic microbiomes

RIMS-seq, DNA-seq and Bisulfite-seq were performed on the synthetic gut and skin microbiome as described. Reads derived from RIMS-seq, DNA-seq and Bisulfite-seq were mapped as described to a 'meta-genome' composed of the reference genomes of all the bacteria included in the corresponding synthetic community (see Supplementary Table 3 for detailed compositions). Mapped reads were split according to each bacterium and RIMS-seq or bisulfite analysis pipelines were run on individual genomes as described above. Abundance was estimated using the number of mapped reads per bacteria and normalized to the total number of mapped reads. Normalized species abundances in RIMS-seq and Bisulfite-seq were compared to the normalized species abundances in DNA-seq.

Quality control of the data

The insert size for each downsampled filtered bam file was calculated using Picard version 2.20.8 using the default parameters and the option CollectInsertSizeMetrics ("Picard Toolkit." 2019. Broad Institute, GitHub Repository. http://broadinstitute.github.io/picard/; Broad Institute).

The GC bias for each downsampled filtered bam file was calculated and plotted using Picard version 2.20.8 using the default parameters and the option CollectGcBiasMetrics.

Xp12 genome assembly

Reads were downsampled to a 30X coverage using seqtk 1.3.106, trimmed using trimgalore 0.6.5 and assembled using Spades 3.14.1 with the --isolate option. Assembly quality was assessed using Quast 5.0.2. Reads used for assembly were then mapped back to the assembly using BWA mem 0.7.17 and mapping statistics were generated using samtools flagstat 1.10.2

Xp12 sequencing performance assessment

Reads were trimmed using trimgalore 0.6.5 and mapped to the Xp12 reference genome using BWA mem 0.7.17. Insert size and GC bias were assessed using Picard Toolkit and genome coverage using Qualimap 2.1.1.

C. Results

1. Principle of RIMS-seq

Spontaneous deamination of cytosine (C) leading to uracil (U) and of m5C leading to thymine (T) are examples of common damage found in DNA. *In-vitro*, this damage is often undesirable as it can interfere with sequencing. The type of interference during sequencing depends on whether the deamination occurs on C or m5C. U blocks the passage of high-fidelity polymerases typically used in library preparation protocols, preventing the amplification and sequencing of U-containing DNA fragments. Conversely, DNA harboring T derived from m5C deamination can be normally amplified, but results in C to T errors (14, 15). This distinction between blocking and mutagenic damage forms the basis of the RIMS-seq method, allowing the identification of methylase specificity based on an elevated number of reads containing C to T transitions specifically at methylated sites (**Figure 1A**). To increase the rate of m5C deamination, the DNA is subjected to a heat-alkaline treatment which has been previously demonstrated to elevate the rate of both C and m5C deamination with m5C having a 1.5–3 times higher deamination rate than for C (16). This treatment is aimed at inducing a level of deamination large enough to detect the m5C methylase specificity without affecting the sequencing quality. For this reason, the deamination levels typically obtained with RIMS-seq does not permit the quantitative measurement of methylation at each genomic site but rather provides a global methylation signal characteristic of the methylase specificity.

Illumina paired-end sequencing allows both ends of a DNA fragment to be sequenced, generating a forward read (R1) and reverse read (R2). Resulting from m5C deamination, R1 has the C to T read variants while R2 has the reverse-complement G to A variant. This difference leads to an overall imbalance of C to T variants between R1 and R2 (17) (see also **Supplementary Figure 1** for

explanation). Thus, sequence contexts for which the C to T read variants are imbalanced in R1 compared to R2 correspond to m5C methylase specificity(ies). Because of the limited deamination rate, RIMS-seq takes advantage of the collective signal at all sites to define methylase specificity. Because C to T imbalance can be observed at nucleotide resolution, RIMS-seq identifies at base resolution which of the cytosine within the motif is methylated.

The experimental steps for RIMS-seq essentially follow the standard library preparation for Illumina sequencing with an extra deamination step. Briefly, the bacterial genomic DNA is fragmented and adaptors are ligated to the ends of DNA fragments (**Figure 1B** and **Methods**). Between the ligation step and the amplification step, an alkaline heat treatment step is added to increase the rate of deamination. Only un-deaminated DNA or DNA containing deaminated m5C can be amplified and sequenced.

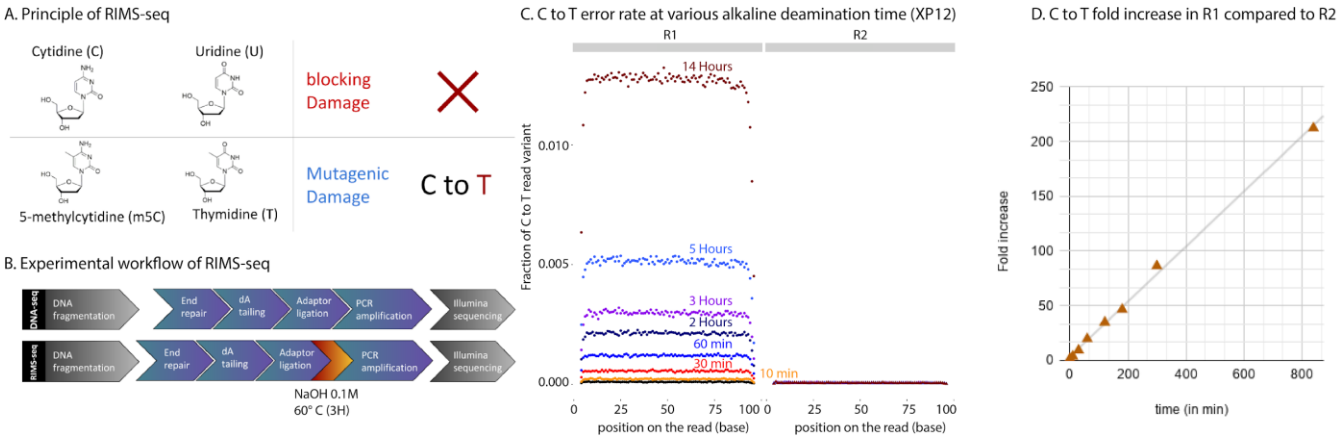


Figure 1: Principle of RIMS-seq. Deamination of cytidine leads to a blocking damage while deamination of m5C leads to a mutagenic C to T damage only present on the first read (R1) of paired-end reads in standard Illumina sequencing. Thus, an increase of C to T errors in R1 in specific contexts is indicative of m5C. **B.** The workflow of RIMS-seq is equivalent to a regular library preparation for Illumina DNA-seq with an extra step of limited alkaline deamination at 60°C. This step can be done immediately after adaptor ligation and does not require additional cleaning steps. **C.** Fraction of C to T variants in XP12 (m5C) at all positions in the reads for R1 and R2 after 0min (DNA-seq), 10min, 30min, 60min, 2h, 3h, 5h and 14h of heat-alkaline treatment. The C to T imbalance between R1 and R2 is indicative of deamination of m5C and increases with heat-alkaline treatment time. **D.** Correlation between the C to T fold increases in R1 compared to R2 according to time (r^2=0.998).

2. Validation of RIMS-seq

- *Optimization of the heat alkaline deamination step*

We first evaluated the conditions to maximize the deamination of m5C while minimizing other DNA damage. For this we used bacteriophage Xp12 genomic DNA that contains exclusively m5C instead of C (18) to measure the m5C deamination rates in various contexts.

To estimate the overall deamination rate of m5C, we quantified the imbalance of C to T read variants between R1 and R2 for 0, 10 and 30 minutes, 1h, 2h, 3h, 5h and 14h of heat alkaline treatment (**Figure 1C**). We observed an imbalance as early as 10 minutes with a 3.7-fold increase of C to T read variants in R1 compared to R2. The increase is linear with time with a maximum of 212-fold increase of C to T read variants in R1 compared to R2 after 14 hours of heat alkaline treatment (**Figure 1D**). Next, we quantified the deamination rate at all Nm5CN sequence contexts with N being A, T, C or G and show an increase of C to T variants in R1 in all contexts (**Supplementary Figure 2A)**. Together, these results show that a measurable deamination rate can be achieved in as soon as 10 minutes of heat alkaline deamination and that deamination efficiency is similar in all sequence contexts.

To estimate the non-specific damage to the DNA leading to unwanted sequencing errors, we quantified possible imbalances for other variant types (**Supplementary Figure 2B**). We found that G to T variants show imbalance in all the conditions investigated, likely the result of oxidative damage resulting from sonication, a common step in library preparation between RIMS-seq and DNA-seq (17). Slight elevation of G to C and T to C read variants can be observed in RIMS-seq compared to DNA-seq but this damage is of low frequency and therefore is not expected to notably affect the sequencing performance QC of RIMS-seq.

We performed QC metrics and assemblies of Xp12 for all the alkaline-heat treatment conditions, including a control DNA-seq. The overall sequencing performances were assessed in terms of insert size, GC bias, and genome coverage. Similar results were observed between RIMS-seq and the DNA-

seq control at all treatment times, indicating that the RIMS-seq heat-alkaline treatment does not affect the quality of the libraries (**Supplementary Figure 3**).

We also evaluated the quality of the assemblies compared to the Xp12 reference genome and found that all conditions lead to a single contig corresponding to essentially the entire genome with very few mismatches **(Supplementary Table 1)**. These results suggest that the heat-alkaline treatment does not affect the assembly quality, raising the possibility of using RIMS-seq for simultaneous *de novo* genome assembly and methylase specificity identification. We found that a 3-hour treatment provides a good compromise between the deamination rate (resulting in about ~ 0.3 % of m5C showing C to T transition) and duration of the experiment. We found that longer incubation times (up to 14h) increased the deamination rate by up to 1% and decided this is a slight sensitivity increase compared to the additional experimental time required.

- *RIMS-seq is able to distinguish methylated versus unmethylated motifs in E. coli*

To validate the application of RIMS-seq to bacterial genomes, we sequenced dcm+ (K12) and dcm- (BL21) *E. coli* strains. In K12, the DNA cytosine methyltransferase *dcm* methylates cytosine at CCWGG sites (C = m5C, W = A or T) and is responsible for all m5C methylation in this strain (19). *E. coli* BL21 has no known m5C methylation. Heat/alkaline treatments were performed at three time points (10min, 1h, and 3h). In addition, we performed a control experiment corresponding to the standard DNA-seq. Resulting libraries were paired-end sequenced using Illumina and mapped to their corresponding genomes (**Methods**).

For comparison, all datasets were downsampled to 5 million reads corresponding to 200X coverage of the *E. coli* genome and instances of high confidence C to T variants (Q score > 35) on either R1 or R2 were identified. As expected, control DNA-seq experiments show comparable numbers of C to T read variants between R1 and R2, indicating true C to T variants or errors during amplification and sequencing (**Figure 2A**). On the other hand, the overall number of C to T read variants in R1 is progressively elevated for 10 min, 1 hour and 3 hours of heat-alkaline treatment of the *E. coli* K12 samples with an overall 4-fold increase after 3 hours treatment compared to no treatment; heat-

alkaline treatments did not increase the rate of C to T read variants in R2 (**Figure 2A**). We anticipate that the elevation of the *E.coli* K12 C to T read variants in R1 is due to deamination of m5C. In this case, the elevation should be specifically found in Cs in the context of CCWGG (with the underlined C corresponding to the base under consideration). To demonstrate this, we calculated the fraction of C to T read variants in CCWGG compared to other contexts. We observed a large elevation of the C to T read variants in the CCAGG and CCTGG contexts for K12 (**Figure 2B**). As expected, the C to T read variants show no elevation at CCAGG and CCTGG contexts for the *E.coli* BL21 strain that is missing the *dcm* methylase gene (**Figure 2B**). Thus, this C to T read variant elevation is specific to the *E.coli* K12 strain subjected to heat-alkaline treatments, consistent with deamination detectable only on methylated sites. Taken together, these results indicate that the elevated rate of C to T variants observed in R1 from *E.coli* K12 is the result of m5C deamination in the CCWGG context.

Next, we assessed whether the difference in the C to T read variant context between R1 and R2 at the CCWGG motif provides a strong enough signal to be discernible over the background noise. For this, we calculated the fraction of C to T read variants in CCWGG and CCWGG compared to all the other NCNNN and CNNNN contexts, respectively. After 3 hours of heat-alkaline treatment, the fraction of C to T read variants in a CCWGG context increased, rising from only 1.9 % in regular DNA-seq to ~25% of all the C to T variants. This increase is only observable in R1 of the K12 strain (**Figure 2C**). Conversely, no increase can be observed in a CCWGG context for which the C to T variant rate at the first C is assessed (**Figure 2C**). Thus, RIMS-seq identified the second C as the one bearing the methylation, consistent with the well described dcm methylation of *E.coli* K12 (20) (19), highlighting the ability of RIMS-seq to identify m5C methylation at base resolution within the methylated motif.

Next, we calculated significant (p-value < 0.01) differences in position-specific nucleotide compositions around C to T variants in R1 compared to R2 using Two Sample Logo (21). We found a signal consistent with the dcm methylase specificity in K12 RIMS-seq samples at one and three hours of heat alkaline treatment (**Figure 2D**) demonstrating that it is possible to identify methylase specificities in genomic sequence subject to as little as 1h of alkaline treatment. These results support the application of RIMS-seq for the *de novo* identification of methylase specificity at base resolution.

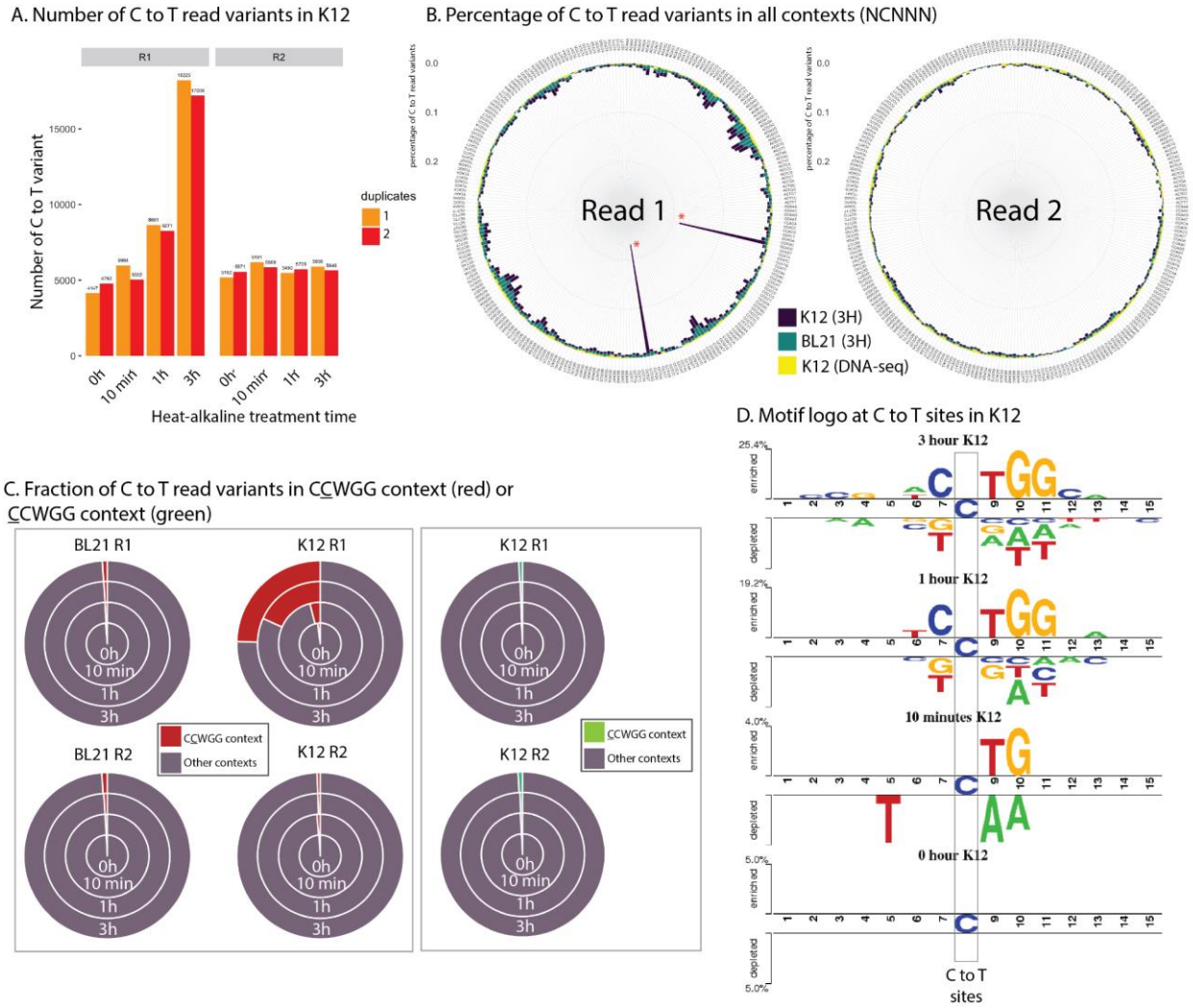


Figure 2: A. Bar plots representing the number of C to T read variants for K12 in R1 and R2 after different heat/alkaline treatment times. Colors represent duplicate experiments. **B.** Circular bar plots representing the percentage of C to T read variants in all NCNNN contexts (with N = A, T C or G) **C.** Proportion of C to T read variants in CCWGG (red) or CCWGG (green) contexts compared to other NCNNN or CNNNN contexts for R1 and R2 in K12 and BL21. The C to T read variants in CCWGG and CCWGG motifs represent less than 2% of all variants except in K12 (R1 only) after 10 minutes, 1hour and 3 hours treatments where the CCWGG motifs represent 4.1%, 22.5% and 32.6% of all C to T read variants respectively. The increase of C to T read variants in the CCWGG context is therefore specific to R1 in K12 strain. **D.** Visualization of the statistically significant differences in position-specific nucleotide compositions around C to T variants in R1 compared to R2 using Two Sample Logo (21)

for the K12 sample subjected to (from top to bottom) 3H, 1H, 10 min and 0 min heat alkaline treatment.

- RIMS-seq identifies the correct methylase specificity *de novo* in *E. coli* K12

In order to identify methylase specificities *de novo* in RIMS-seq sequencing data, we devised an analysis pipeline based on MoSDi (22) to extract sequence motif(s) with an over-representation of C to T transitions in R1 reads (**Figure 3A,** analysis pipeline available at https://github.com/Ettwiller/RIMS-seq). In brief, the pipeline extracts the sequence context at each C to T read variant in R1 (foreground) and R2 (background). MoSDi identifies the highest over-represented motif in the foreground sequences compared to the background sequences. To accommodate the presence of multiple methylases in the same host, the first motif is subsequently masked in both the foreground and background sequences and the pipeline is run again to find the second highest over-represented motif and so on until no significant motifs can be found (see **Methods** for details). Running the pipeline on strain K12 identifies one significant over-represented motif corresponding to the CCWGG motif (p-value = 9.71e−77, 4.25e−858 and 3.61e−4371 for 10min, 60min and 180min of alkaline treatment respectively) with the cytosine at position 2 being m5C. Summing up, we devised a novel sequencing strategy called RIMS-seq and its analysis pipeline to identify m5C methylase specificity *de novo*. When applied to *E. coli* K12, RIMS-seq identifies the dcm methylase specificity as CCWGG with the methylated site located on the second C, consistent with the reported dcm methylase specificity (**Table 1**).

- RIMS-seq identifies multiple methylase specificities *de novo* within a single microorganism

To assess whether RIMS-seq is able to identify methylase specificity in strains expressing multiple methylases, we repeated the same procedure on a strain of *Acinetobacter calcoaceticus* ATCC 49823 expressing two m5C methylases with known specificities (4). RIMS-seq identifies CGCG (p-value = 2.33e-174) and GATC (p-value = 3.02e-1308) (**Table 1**) both motifs have been confirmed by MFRE-seq (10). Thus, RIMS-seq is able to *de novo* identify methylase specificities in bacteria expressing multiple methylases.

- RIMS-seq can be applied for genome sequencing and m5C profiling in bacteria without a reference genome

We investigated whether RIMS-seq can be used to identify methylase specificities of uncharacterized bacteria for which a reference genome is unavailable. More specifically, we evaluated if the reads generated using RIMS-seq can be used for both identifying methylase specificities and generating an assembly of comparable quality to DNA-seq.

For this, we performed RIMS-seq on *A. calcoaceticus* ATCC 49823 genomic DNA as described above as well as a control DNA-seq experiment for which the alkaline treatment was replaced by 3 hours incubation in TE (DNA-seq(+3H)). We compared the *de novo* assembly obtained from the reads generated by the DNA-seq(+3H) and the *de novo* assembly obtained from the reads generated by RIMS-seq (see **Material and Methods**). In brief, the alkaline treatment did not alter the important metrics for assembly quality such as the rate of mismatches and N50 demonstrating that the elevated C to T variant rate at methylated sites is not high enough to cause assembly errors (**Figure 3B**).

We then proceeded to map the RIMS-seq reads to the assembly and motifs were identified using the RIMS-seq *de novo* motif discovery pipeline. As expected, the same motifs found when mapping to the reference genome are also found in the *A. calcoaceticus de novo* assembly with similar significance (GATC (p-value = 1.44e-1255) and CGCG (p-value = 8.6e-228) (**Figure 3C**). These motifs correspond to the methylase specificities expected in this strain indicating that RIMS-seq can be applied for genome sequencing and assembly of any bacterium without the need for a reference genome.

- RIMS-seq can be complemented with SMRT sequencing to obtain a comprehensive overview of methylase specificities

RIMS-seq performed in parallel with SMRT sequencing has the advantage of comprehensively identifying all methylase specificities (m5C, m4C and m6A methylations) and results in an assembly of higher quality than with short reads illumina data. We applied this hybrid approach to

Acinetobacter calcoaceticus ATCC 49823 for which a SMRT sequencing and assembly had been done previously (4). RIMS-seq was performed as described above and the reads were mapped to the genome assembly obtained from SMRT-sequencing. We again found the two m5C motifs : CGCG (p-value = 1.84e-1535) and GATC (p-value = 4.93e-6856) from the RIMS-seq data in addition to the 13 m6A motifs described previously using SMRT sequencing (4). This result demonstrates the advantage of such a hybrid approach in obtaining closed genomes with comprehensive epigenetic information.

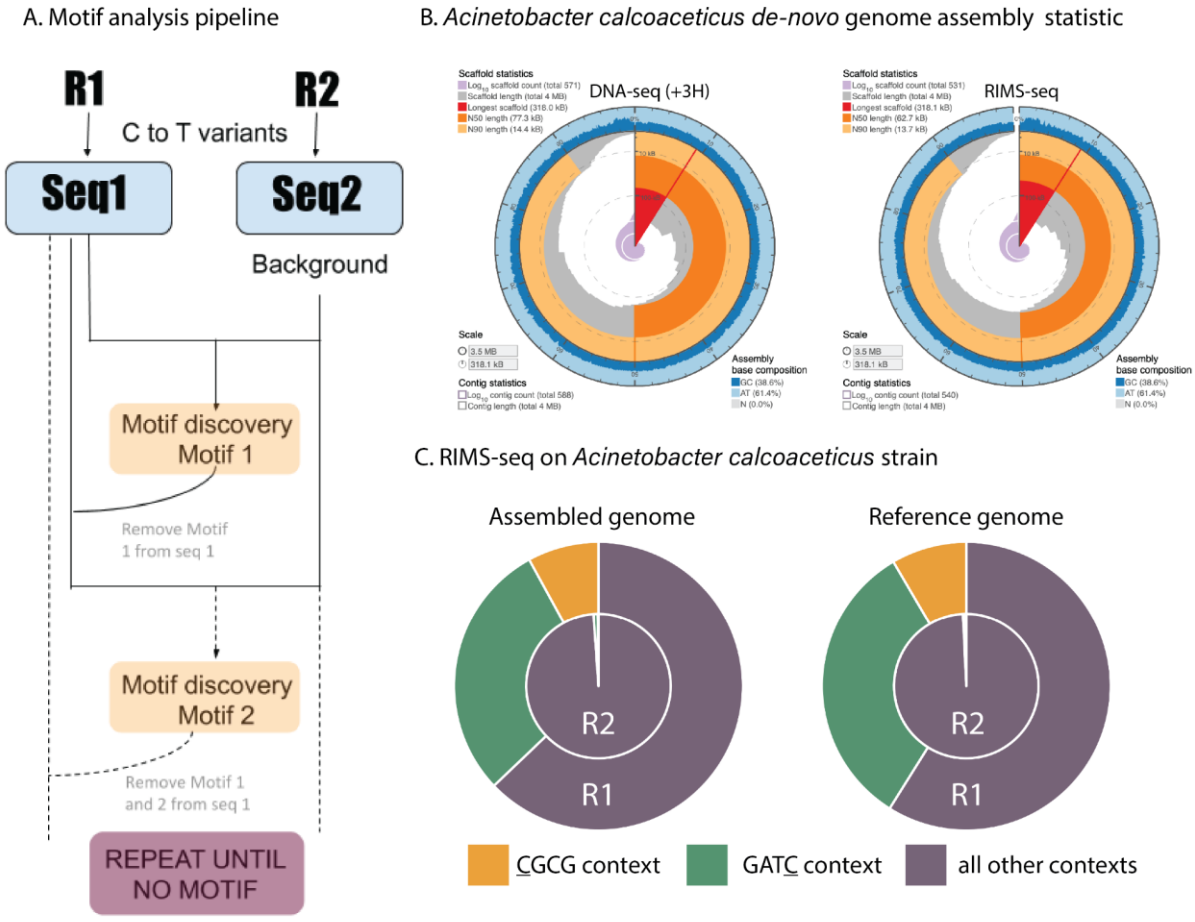


Figure 3: *De novo* discovery of methylase specificity using RIMS-seq. **A.** Description of the RIMS-seq motif analysis pipeline. First, C to T read variants are identified in both Read 1 and Read 2 separately. Then, the MosDI program searches for overrepresented motifs. Once a motif is found, the pipeline is repeated until no more motifs are found, enabling identification of multiple methylase specificities in an organism. **B.** Assembly statistics obtained using the sequence from the standard DNA-seq

(+3H, left) and RIMS-seq (right). Visualization using assembly-stats program (https://github.com/rjchallis/assembly-stats). The corresponding table with the statistical values is available in the supplementary material (**Supplementary Table 2**). **C.** Fractions of C to T read variants in CGCG (yellow) or GATC (green) contexts compared to other contexts for R1 and R2 in *Acinetobacter calcoaceticus* ATCC 49823 using the assembled or the reference genome. The increase of C to T read variants in these contexts are similar when using either the assembled or reference genomes.

3. RIMS-seq can be applied to a variety of RM systems

Methylases targets are usually palindromic sequences between 4-8nt, and a single bacterium often possesses several, distinct MTase activities (23). Next, we tested the general applicability of RIMS-seq and the *de novo* motif discovery pipeline using bacterial genomic DNA from our in-house collections of strains.

For some bacterial strains, the methylase recognition specificities have been previously experimentally characterized. In all of those strains, RIMS-seq confirms the specificities and identifies the methylated cytosine at base resolution (**Table 1**). We have tested the identification of 4-mers motifs such as GAT<u>C</u>, <u>C</u>GCG (*Acinetobacter calcoaceticus*) and G<u>C</u>GC (*Haemophilus parahaemolyticus*) up to 8-mers motifs such as AC<u>C</u>GCACT and AGTG<u>C</u>GGT (*Haemophilus influenzae*). Motifs can be palindromic or non-palindromic (**Table 1 and Supplementary Table 3**). In the latter case, RIMS-seq defines non-palindromic motifs at strand resolution. For example, RIMS-seq identifies methylation at two non-palindromic motifs A<u>C</u>CTGC as well as its reverse complement G<u>C</u>AGGT in the *Bacillus fusiformis* strain (**Table 1**).

A number of RM systems have been expressed in other hosts such as *E. coli* for biotechnological applications. For the methylase M.HhaI recognizing G<u>C</u>GC (4), we performed RIMS-seq and a control DNA-seq(+3H) on both the native strain (*Haemophilus parahaemolyticus* ATCC 10014) and in *E. coli* K12 expressing the recombinant version of M.HhaI. Interestingly, we found that the *de novo* RIMS-seq analysis algorithm identifies R<u>C</u>GC (with R being either A or G) for the recombinant strain and G<u>C</u>GC for the native strain (**Figure 4A**). Conversely, no notable elevation of C to T read variants are

observed for the native strain (**Figure 4B**), confirming the *de novo* motif discovery results from the analysis pipeline. Collectively, these results suggest that the recombinant methylase shows star activity, notably in the context of ACGC, that is not found in the native strain. We hypothesize that the star activity is the result of the over-expression of the methylase in *E. coli* K12. Interestingly, ACGC is not a palindrome motif and consequently the star activity results in hemi-methylation of the ACGC sites and not the GCGT motif.

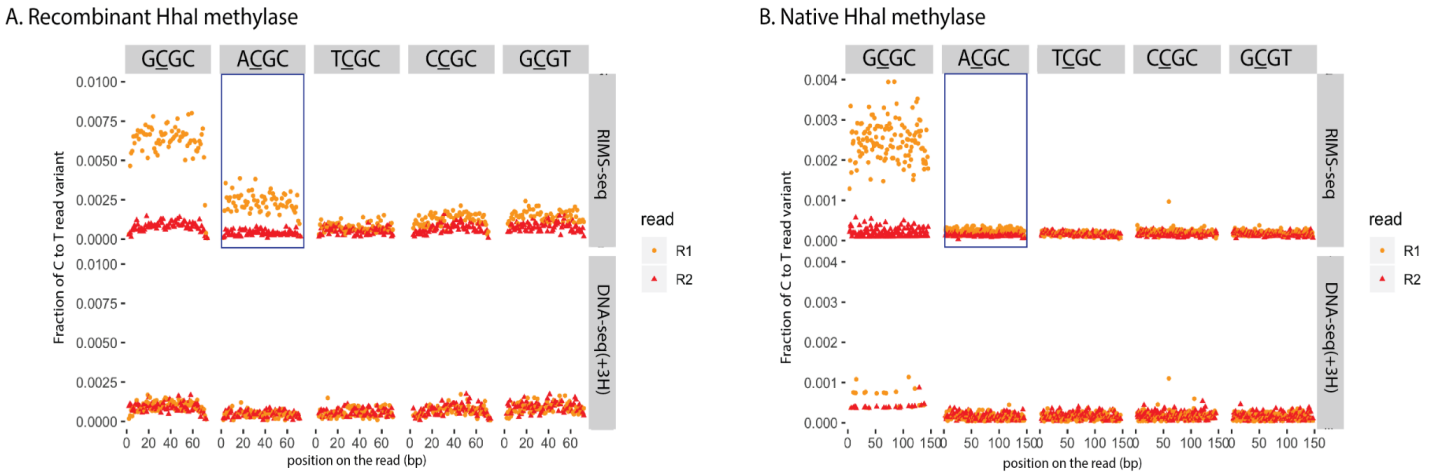


Figure 4: C to T error profile in GCGC (canonical recognition site), ACGC, TCGC, CCGC and GCGT. in R1 reads (orange) and R2 reads (red) for RIMS-seq (upper panel) and DNA-seq(+3H) (lower panel) **A.** Recombinant HhaI methylase expressed in *E. coli* **B.** Native HhaI methylase expressed in *Haemophilus parahaemolyticus*. Elevation of C to T in the R1 read variant can be observed in the context of GCGC for both the recombinant and native HhaI genomic DNA and in the context of ACGC only for DNA from the recombinant but not the native HhaI.

4. RIMS-seq can be applied to microbial communities

We assessed whether RIMS-seq can be applied to mixed microbial communities using synthetic gut and skin microbiomes from ATCC containing 12 and 6 bacterial species, respectively. We also complemented the RIMS-seq experiment with the control experiment DNA-seq(+3H) and a bisulfite treatment to validate the RIMS-seq findings. Reads were mapped to their respective microbiome reference genomes (**Methods**). For the gut microbiome we found a mapping rate (properly paired

only) of 95.79%, 95.77% and 66.2% for RIMS-seq, DNA-seq and bisulfite-seq respectively. Concerning the skin microbiome, 85.89%, 85.35% and 54.9% of reads were mapped for RIMS-seq, DNA-seq and bisulfite-seq respectively. The low mapping rate for bisulfite-seq is a known challenge as the reduction of the alphabet to A, G, T generates ambiguous mapping (24).

To use RIMS-seq as an equivalent to DNA-seq for mixed community applications, RIMS-seq should produce sequencing quality metrics that are similar to standard DNA-seq, especially on the estimation of species relative abundances. We therefore compared RIMS-seq sequencing performances with DNA-seq(+3H) and bisulfite sequencing. We found that bisulfite sequencing elevates abundances of AT-rich species such as *Clostridioides difficile* (71% AT), *Enterococcus faecalis* (63% AT) and *Fusobacterium nucleatum* (73% AT) (**Figure 5A, Supplementary Figure 4**). For example, bisulfite sequencing over-estimated the presence of *Clostridioides difficile* by a factor of 2.65 and *Staphylococcus epidermidis* by a factor of 3.9 relative to DNA-seq. This over-estimation of an AT rich genome by bisulfite is a known bias of bisulfite sequencing and relates to damage at cytosine bases (25). Conversely, we found that the species abundances are similar between DNA-seq(+3H) and RIMS-seq (abundance ratios between 0.8 and 1.2) indicating that RIMS-seq can be used to quantitatively estimate microbial composition.

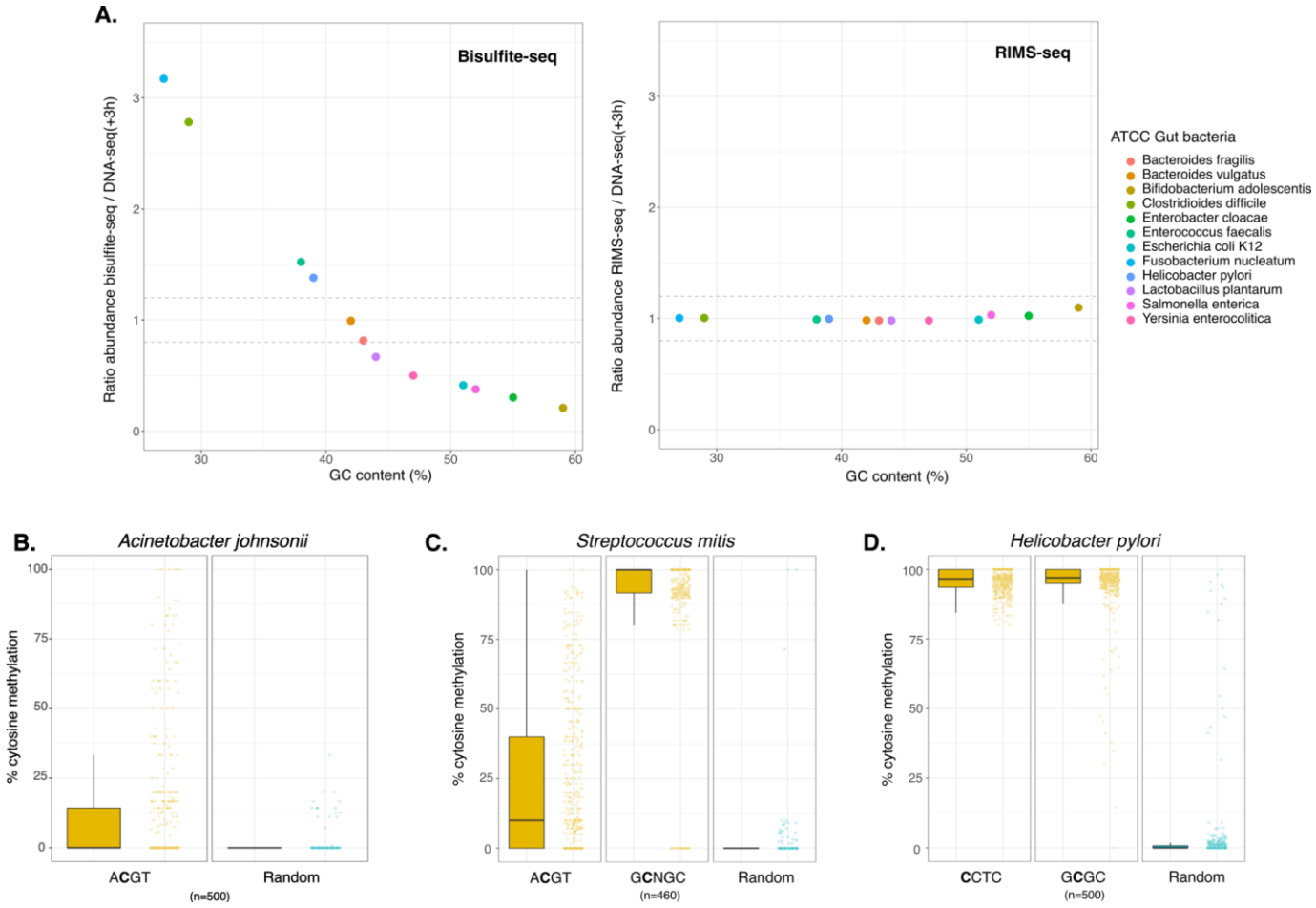


Figure 5: **A.** Bacterial abundance in the ATCC gut microbiome calculated from bisulfite-seq data (left) and RIMS-seq (Right) normalized to DNAseq(+3H). The normalized abundance is plotted relative to the GC content of each bacterium. **B.** Methylation levels in *Acinetobacter johnsonii* (ATCC skin microbiome). The methylation level was calculated for cytosine positions in the context of A<u>C</u>GT (yellow) and randomly selected positions in other contexts (blue). These bisulfite-seq data suggest some sites are methylated in the context of A<u>C</u>GT, but they are not fully methylated. **C.** Methylation level in Streptococcus mitis (ATCC skin microbiome) calculated from bisulfite-seq data. The methylation level was calculated for cytosine positions in the context of ACGT and GCNGC (yellow) as well as for randomly selected positions in other contexts (blue). **D.** Methylation level in Helicobacter pylori (ATCC gut microbiome) calculated from bisulfite-seq data. The methylation level was calculated for cytosine positions in the context of G<u>C</u>GC and <u>C</u>CTC (yellow) as well as for randomly selected positions in other contexts (blue).

- RIMS-seq identifies known and novel methylase specificities in synthetic microbial communities

Overall, we found motifs for 6 out of the 12 gut microbiome species and 5 out of the 6 skin microbiome species (**Supplementary Table 3**). The motifs range from 4 to 8 nucleotides long and 70% are palindromic. Interestingly, we found an unknown palindromic motif GGCSGCC (with S being either C or G) from *Micrococcus luteus* (NC_012803.1) in the skin community. To our knowledge, this is the first time this 7nt motif is identified, showing the potential of RIMS-seq to identify new methylase specificities.

We validated the results obtained with RIMS-seq using bisulfite sequencing. RIMS-seq identified 2 motifs in *Helicobacter pylori* from the ATCC synthetic gut microbiome: GCGC as well as an additional non-palindromic motif CCTC that was identified by the bisulfite analysis pipeline as CYTC with Y being either C or T. The CCTC motif is very common in *Helicobacter pylori*s species, it has been described to be modified at m5C on one strand, while modified at m6A on the other strand (4). In order to confirm the RIMS-seq motif, we investigated the bisulfite-seq data and compared the methylation level in cytosines present either in the CCTC context versus cytosines in any other context. We see a methylation level above 90% at the cytosines in the CCTC context confirming the existence of this methylated motif in *Helicobacter pylori* (**Figure 5D**). Interestingly, m4C methylation in *Helicobacter pylori* has been shown to also occur at TCTTC (26), resulting in the composite motif CYTC (TCTTC and NCCTC) found in the bisulfite data. Contrary to bisulfite, RIMS-seq does not identify m4C methylation (27).

Also, interestingly, bisulfite-seq results indicate that the ACGT motif in *Acinetobacter johnsonii* and *Streptococcus mitis* from the ATCC synthetic skin microbiome are not fully methylated (**Figure 5B**). Most of the sites in *Acinetobacter johnsonii* show a methylation of about 10% while in *Streptococcus mitis*, the average methylation per site is 23% (**Figure 5C**). These results highlight that despite the low methylation levels, RIMS-seq is able to detect the ACGT motif at high significance (p-value < 1e-100).

Organism	Accession numbers (biosample)	RIMS-seq motif(s)	Validated motif(s)
Escherichia coli K12	SAMN02604091	C<u>C</u>WGG	C<u>C</u>WGG (1,2,4)
Acinetobacter calcoaceticus ATCC 49823	SAMN14530202	GAT<u>C</u> <u>C</u>GCG	GAT<u>C</u> (4) <u>C</u>GCG (2,4)
Bacillus fusiformis 1083	SAMN17843035	A<u>C</u>CTGC G<u>C</u>AGGT	A<u>C</u>CTGC (2,3) G<u>C</u>AGGT (2,3)
Bacillus amyloliquefaciens H ATCC 49763	SAMN12284742	G<u>C</u>WGC	G<u>C</u>WGC (3)
Clostridium acetobutylicum ABKn8	SAMN17843114	G<u>C</u>NNGC	G<u>C</u>NNGC (3)
Aeromonas hydrophila NEB724	SAMN14533640	GC<u>C</u>GGC	GC<u>C</u>GGC (3)
Haemophilus influenzae Rd ATCC 51907	SAMN02603991	GR<u>C</u>GYC* AC<u>C</u>GCACT AGTG<u>C</u>GGT	GRCGYC (5)
Haemophilus parahaemoltyicus ATCC 10014	SAMN11345835	G<u>C</u>GC	G<u>C</u>GC (2)
M.HhaI clone (*E. coli*)	NA	R<u>C</u>GC C<u>C</u>WGG[(a)]	G<u>C</u>GC (4) C<u>C</u>WGG (1,2,4)[(a)]

Table 1: Methylases specificity obtained using RIMS-seq and validated using different methods. The method is indicated by a number next to the motif. Evidence for the validated motifs are (1) Bisulfite-seq (material and Methods), (2) REBASE (4), (3) EM-seq (material and method), (4) MFRE-seq (10), (5) mTet1-enhanced SMRT sequencing (6). (a) The E. coli strain used is Dcm+, resulting in the discovery of both the Dcm (C<u>C</u>WGG) and M.HhaI motifs (G<u>C</u>GC). RIMS-seq discovered R<u>C</u>GC instead of G<u>C</u>GC motif (see text for explanation).

D. Discussion

In this study, we developed RIMS-seq, a sequencing method to simultaneously obtain high quality genomic sequences and discover m5C methylase specificity(ies) in bacteria using a single library preparation. The simplicity of the procedure makes RIMS-seq a cost effective and time saving method with only an additional 3h sodium hydroxide incubation and an additional column-based cleaning step. Theoretically, the cleaning step can be avoided if a small volume of the library is used for the amplification step, but we have not tested this procedure.

Due to the limited deamination rate, RIMS-seq is equivalent to short read DNA-seq in terms of sequencing quality. Sequencing QC metrics such as coverage, GC content and mapping rate are similar for RIMS-seq and DNA-seq. Thus, RIMS-seq can be used for applications such as, but not limited to, shotgun sequencing, genome assembly and estimation of species composition of complex microbial communities. This dual aspect of RIMS-seq is analogous to SMRT sequencing for which methylation is inferred from the IPD ratio. We showed that both PacBio and RIMS-seq can be complementary with the ability to obtain a complete methylome: m6A and m4C methylase specificities can be obtained from SMRT sequencing while m5C methylase specificity can be obtained from RIMS-seq. Combining both sequencing technologies also allows for a hybrid assembly strategy resulting in closed reference genomes of high sequencing accuracy.

We applied RIMS-seq to several bacteria and identified a variety of methylation motifs, ranging from 4 to 8nt long, palindromic and non-palindromic. Some of these motifs were identified for the first time, demonstrating the potential of the technology to discover new methylase specificities, from known as well as from unknown genomes. We also validated that RIMS-seq can identify multiple methylase specificities from a synthetic microbial community and estimate species abundances. However, RIMS-seq has caveats similar to metagenomics sequencing when applied to study natural microbial communities. Closely related species are likely to co-exist and assigning the motif to the correct species can be challenging. Furthermore, single nucleotide polymorphisms may confound the identification of the C to T deamination, increasing the background noise for the detection of

motifs. Finally, species in microbiomes are unevenly represented which can cause RIMS-seq to identify motifs only in the most abundant species.

Because RIMS-seq is based on a limited deamination, it requires the combined signal over many reads to be large enough to effectively identify methylase specificity. For the vast majority of the methylases in RM systems, methylation is present at a sufficient number of sites across the genome for RIMS-seq to determine their specificities. Nonetheless, bacterial methylases can be involved in other processes such as, but not limited to, DNA mismatch repair (28), gene regulation (29) and sporulation (30) and the recognition sites may not necessarily be fully methylated. Partially methylated sites can be found using RIMS-seq but more analysis needs to be done to evaluate how pervasive methylation needs to be to provide a RIMS-seq signal. In other cases, methylated motifs are too specific or under purifying selection, resulting in just a handful of sites in the genome. In these cases, RIMS-seq signals can only be obtained with enough read coverage to compensate for the scarcity of those sites. While the methylase specificities are of great interest in bacteria due to their diversity in recognition sequences, applying RIMS-seq to humans would lead to the identification of the already well-described CpG context. In this case, other technologies such as EM-seq or bisulfite-seq are more appropriate as they enable the precise genomic location to be obtained.

In summary, RIMS-seq is a new technology allowing the simultaneous investigation of both the genomic sequence and the methylation in prokaryotes. Because this technique is easy to implement and shows similar sequencing metrics to DNA-seq, RIMS-seq has the potential to substitute DNA-seq for microbial studies.

Code and data availability

The data have been deposited with links to BioProject accession number PRJNA706563 in the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/).

Custom-built bioinformatics pipelines to analyse sequencing reads from RIMS-seq are available at https://github.com/Ettwiller/RIMS-seq/

Acknowledgments

We thank Peter Weigele and Yian-Jiun Lee from New England Biolabs for the Xp12 genomic DNA and genomic sequence. We thank Ivan Correa and Nan Dai for their assistance with LC-MS and Ira Schildkraut for his help with methylase specificities.

Competing interests

CB, YCL, AF, BPA, LC, TCE, RR and LE are or were employees of New England Biolabs Inc. a manufacturer of restriction enzymes and molecular reagents.

References

1. Loenen,W.A.M., Dryden,D.T.F., Raleigh,E.A., Wilson,G.G. and Murray,N.E. (2014) Highlights of the DNA cutters: a short history of the restriction enzymes. *Nucleic Acids Research*, **42**, 3–19.

2. Blow,M.J., Clark,T.A., Daum,C.G., Deutschbauer,A.M., Fomenkov,A., Fries,R., Froula,J., Kang,D.D., Malmstrom,R.R., Morgan,R.D., *et al.* (2016) The Epigenomic Landscape of Prokaryotes. *PLoS Genet.*, **12**, e1005854.

3. Beaulaurier,J., Schadt,E.E. and Fang,G. (2019) Deciphering bacterial epigenomes using modern sequencing technologies. *Nat. Rev. Genet.*, **20**, 157–172.

4. Roberts,R.J., Vincze,T., Posfai,J. and Macelis,D. (2015) REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Research*, **43**, D298–D299.

5. Flusberg,B.A., Webster,D.R., Lee,J.H., Travers,K.J., Olivares,E.C., Clark,T.A., Korlach,J. and Turner,S.W. (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods*, **7**, 461–465.

6. Clark,T.A., Lu,X., Luong,K., Dai,Q., Boitano,M., Turner,S.W., He,C. and Korlach,J. (2013) Enhanced 5-methylcytosine detection in single-molecule, real-time sequencing via Tet1 oxidation. *BMC Biol.*, **11**, 4.

7. Tse,O.Y.O., Jiang,P., Cheng,S.H., Peng,W., Shang,H., Wong,J., Chan,S.L., Poon,L.C.Y., Leung,T.Y., Chan,K.C.A., *et al.* (2021) Genome-wide detection of cytosine methylation by single molecule real-time sequencing. *Proc. Natl. Acad. Sci. U. S. A.*, **118**.

8. Sun,Z., Vaisvila,R., Hussong,L.-M., Yan,B., Baum,C., Saleh,L., Samaranayake,M., Guan,S., Dai,N., Corrêa,I.R.,Jr, *et al.* (2021) Nondestructive enzymatic deamination enables single-molecule long-read amplicon sequencing for the determination of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Genome Res.*, 10.1101/gr.265306.120.

9. Liu,Y., Siejka-Zielińska,P., Velikova,G., Bi,Y., Yuan,F., Tomkova,M., Bai,C., Chen,L., Schuster-Böckler,B. and Song,C.-X. (2019) Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. *Nat. Biotechnol.*, **37**, 424–429.

10. Anton,B.P., Fomenkov,A., Wu,V. and Roberts,R.J. (2021) Genome-Wide Identification of 5-Methylcytosine Sites in Bacterial Genomes By High-Throughput Sequencing of MspJI Restriction Fragments. *bioRxiv*, 10.1101/2021.02.10.430591.

11. Simpson,J.T., Workman,R.E., Zuzarte,P.C., David,M., Dursi,L.J. and Timp,W. (2017) Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods*, **14**, 407–410.

12. Rand,A.C., Jain,M., Eizenga,J.M., Musselman-Brown,A., Olsen,H.E., Akeson,M. and Paten,B. (2017) Mapping DNA methylation with high-throughput nanopore sequencing. *Nat. Methods*, **14**, 411–413.

13. Tourancheau,A., Mead,E.A., Zhang,X.-S. and Fang,G. (2021) Discovering multiple types of DNA methylation from bacteria and microbiome using nanopore sequencing. *Nat. Methods*, **18**, 491–498.

14. Fogg,M.J., Pearl,L.H. and Connolly,B.A. (2002) Structural basis for uracil recognition by archaeal family B DNA polymerases. *Nat. Struct. Biol.*, **9**, 922–927.

15. Duncan,B.K. and Miller,J.H. (1980) Mutagenic deamination of cytosine residues in DNA. *Nature*, **287**, 560–561.

16. Wang,R.Y., Kuo,K.C., Gehrke,C.W., Huang,L.H. and Ehrlich,M. (1982) Heat- and alkali-induced deamination of 5-methylcytosine and cytosine residues in DNA. *Biochim. Biophys. Acta*, **697**, 371–377.

17. Chen,L., Liu,P., Evans,T.C.,Jr and Ettwiller,L.M. (2017) DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science*, **355**, 752–756.

18. Kuo,T.T., Huang,T.C. and Teng,M.H. (1968) 5-Methylcytosine replacing cytosine in the deoxyribonucleic acid of a bacteriophage for Xanthomonas oryzae. *J. Mol. Biol.*, **34**, 373–375.

19. Marinus,M.G. and Morris,N.R. (1973) Isolation of deoxyribonucleic acid methylase mutants of Escherichia coli K-12. *J. Bacteriol.*, **114**, 1143–1150.

20. Palmer,B.R. and Marinus,M.G. (1994) The dam and dcm strains of Escherichia coli--a review. *Gene*, **143**, 1–12.

21. Crooks,G.E., Hon,G., Chandonia,J.-M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.

22. Marschall,T. and Rahmann,S. (2009) Efficient exact motif discovery. *Bioinformatics*, **25**, i356–64.

23. Vasu,K. and Nagaraja,V. (2013) Diverse functions of restriction-modification systems in addition to cellular defense. *Microbiol. Mol. Biol. Rev.*, **77**, 53–72.

24. Grehl,C., Wagner,M., Lemnian,I., Glaser,B. and Grosse,I. (2020) Performance of Mapping Approaches for Whole-Genome Bisulfite Sequencing Data in Crop Plants. *Front. Plant Sci.*, **11**, 176.

25. Olova,N., Krueger,F., Andrews,S., Oxley,D., Berrens,R.V., Branco,M.R. and Reik,W. (2018) Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. *Genome Biol.*, **19**, 33.

26. Vitkute,J., Stankevicius,K., Tamulaitiene,G., Maneliene,Z., Timinskas,A., Berg,D.E. and Janulaitis,A. (2001) Specificities of eleven different DNA methyltransferases of Helicobacter pylori strain 26695. *J. Bacteriol.*, **183**, 443–450.

27. Vilkaitis,G. and Klimasauskas,S. (1999) Bisulfite sequencing protocol displays both 5-methylcytosine and N4-methylcytosine. *Anal. Biochem.*, **271**, 116–119.

28. Modrich,P. and Lahue,R. (1996) Mismatch repair in replication fidelity, genetic recombination, and cancer biology. *Annu. Rev. Biochem.*, **65**, 101–133.

29. Casadesús,J. and Low,D. (2006) Epigenetic gene regulation in the bacterial world. *Microbiol. Mol. Biol. Rev.*, **70**, 830–856.

30. Oliveira,P.H., Ribis,J.W., Garrett,E.M., Trzilova,D., Kim,A., Sekulovic,O., Mead,E.A., Pak,T., Zhu,S., Deikus,G., *et al.* (2020) Epigenomic characterization of Clostridioides difficile finds a conserved DNA methyltransferase that mediates sporulation and pathogenesis. *Nature Microbiology*, **5**, 166–180.

31. Nurk,S., Bankevich,A., Antipov,D., Gurevich,A.A., Korobeynikov,A., Lapidus,A., Prjibelski,A.D., Pyshkin,A., Sirotkin,A., Sirotkin,Y., *et al.* (2013) Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J. Comput. Biol.*, **20**, 714–737.

32. Vacic,V., Iakoucheva,L.M. and Radivojac,P. (2006) Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics*, **22**, 1536–1537.

33. Emms,D.M. and Kelly,S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.*, **20**, 238.

34. Letunic,I. and Bork,P. (2021) Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.*, 10.1093/nar/gkab301.

Chapter II : Cappable-seq: a versatile method for the identification of transcriptional landmarks in bacteria

In this chapter, we present two new methods: ONT-Cappable-seq and Loop-Cappable-seq. We are planning to prepare a manuscript presenting the ONT-Cappable-seq method, and the Loop-Cappable-seq manuscript is already in preparation.

A. Introduction

1. Bacterial transcriptomics and Cappable-seq

a. RNA-sequencing (RNA-seq)

The transcriptome can be defined as: "the complete set of transcripts in a cell, and their quantity, for a specific developmental stage or physiological condition" (Wang, Gerstein and Snyder, 2009). Understanding the transcriptome is essential, notably to comprehend the functionality of a genome.

During the last decades, the throughput for quantifying gene expression has considerably increased, progressing from one or a few genes (using Northern blot, quantitative real-time polymerase chain reaction (RT-PCR) or hybridization of cDNA) to transcriptomics scale using RNA-seq (Moody, 2001). Before the advent of RNA-seq, hybridization-based approaches were routinely used to quantify transcripts. These approaches were mid- to high-throughput and relatively inexpensive but were limited by cross-hybridization artifacts, poor quantification of lowly and highly expressed genes, and the need to know the sequence of interest. In microbiology, it would mean a specific set of probes for each bacteria studied.

Because of these limitations, transcriptomics transitioned to NGS based sequencing methods and in the mid 2000's, RNA-seq revolutionized the field. This methodology, in which RNA is first fragmented, then complementary DNA is generated by reverse transcription, subjected to high-throughput sequencing and mapped to the genome, was developed and initially used to identify the transcriptional map of yeasts (Nagalakshmi *et al.*, 2008; Pinto *et al.*, 2011). This technique provides researchers with a revolutionary method that is high-throughput, high coverage, has a high sensitivity and ultimately, can be used to characterize the entire transcriptome of an organism. Although the RNA-seq pioneering studies were done on eukaryotes, their mRNAs with poly-A tails being easier to isolate, RNA-seq has also widely been applied to prokaryotes. In 2009, the study from Passalacqua *et al.* on *Bacillus anthracis* provided the first comprehensive, single-nucleotide resolution view of a bacterial transcriptome (Passalacqua *et al.*, 2009).

RNA-seq rapidly became a popular approach to study gene expression of diverse bacteria. While RNA-seq was developed primarily for transcript quantification, more specialized techniques based on RNA-seq were developed to target specific subsets of the transcriptome. While still based on high throughput sequencing of cDNA, these techniques have upstream treatments of the RNA that select for only certain types of transcripts or certain positions within the transcripts. For example, ribosomal profiling includes a set of techniques aiming at identifying the footprint of ribosomes to locate parts of the transcript that are in the process of being translated (Brar and Weissman, 2015). Here, the pre-treatment consists of eliminating the RNA fraction that is not bound to ribosomes before performing RNA-seq. The most relevant set of specialized techniques for this thesis work can be classified as transcript identification (as illustrated in Figure 10) for which pre-treatment of the RNA selects for either the 5′ end (Cappable-seq, dRNA-seq) or the 3′ end (term-seq) of the RNA before performing RNA-seq. These techniques are primarily aiming at annotating the architecture of transcripts and less so for transcript quantification. In addition to focusing on subsets of the transcriptome, RNA-seq has been applied to study host-pathogen interactions. One example is the dual RNA-seq, this technique relies on a parallel RNA-seq analysis of a bacterial pathogen and its eukaryotic host (Westermann, Barquist and Vogel, 2017), facilitating study of bacterial infection modes and the associated host responses.

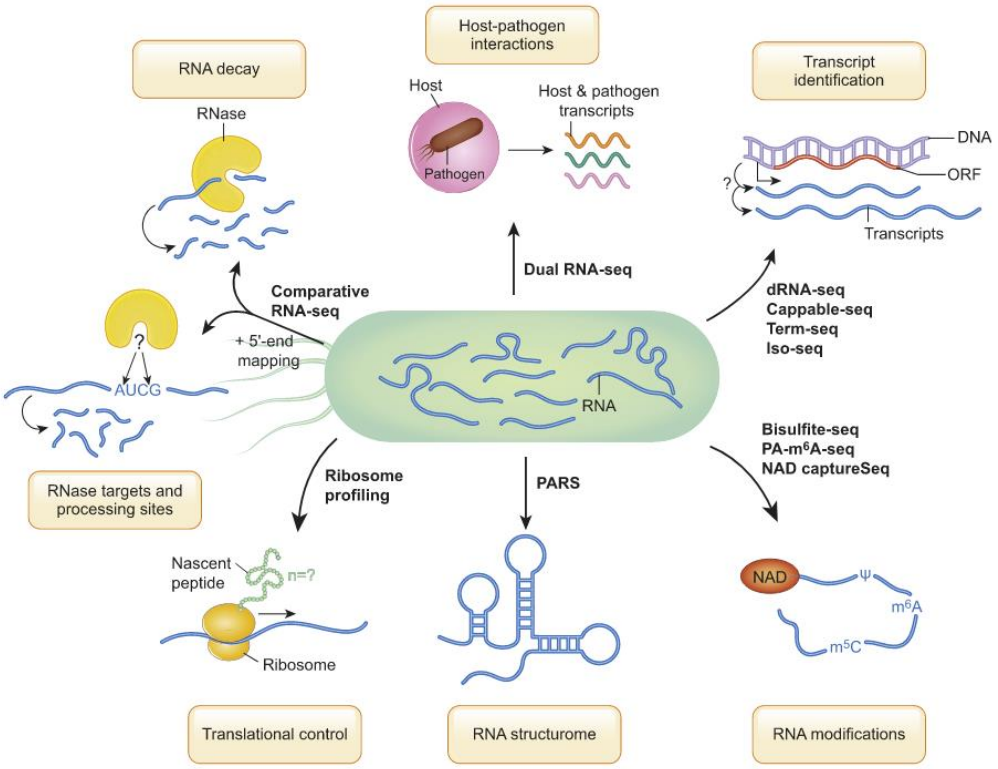


Figure 10: Technical variations and applications of RNA-seq using bacterial total RNA as starting material (Hör, Gorski and Vogel, 2018).

Strategies to identify transcripts architecture

Operons were first described in 1960 by Jacques Monod and Francois Jacob as means for bacteria to co-express functionally-related genes from a single promoter (Jacob and Monod, 1989). Studies have since shown that bacteria have complex transcriptional regulation mechanisms that give rise to condition-specific changes in operon structure (**Figure 11**).

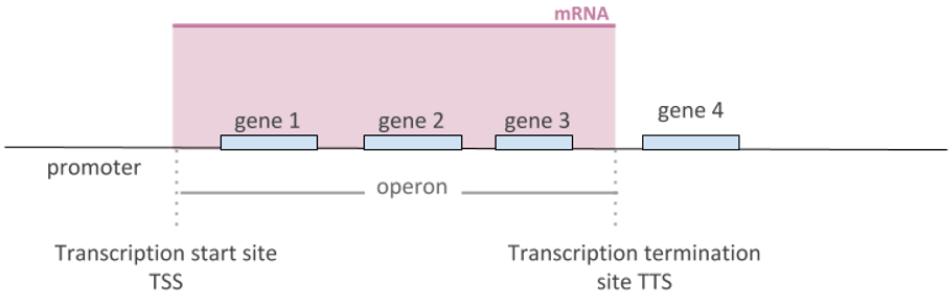


Figure 11: Structure of a prokaryotic operon. Operons are delimited by the Transcription Start Site on the 5'end (TSS) and the Transcription Termination Site (TTS) on the 3'end.

Because RNA-seq relies on a necessary RNA fragmentation step to be compatible with short-read sequencing, the information on the operon structure is lost, notably the position of the transcription start site (TSS) relative to the transcription termination site (TTS) for the same transcript. Furthermore, the standard strategies to reverse transcript RNA generates start and ends of cDNA that do not precisely match the original transcript start and end (**Figure 12**). In addition to these technical limitations, biological RNA processing confounds the identification of transcriptional landmarks and more generally, of primary transcripts, as processed RNA and rRNAs account for more than 95-99% of total prokaryotic RNA (Baracchini and Bremer, 1987).

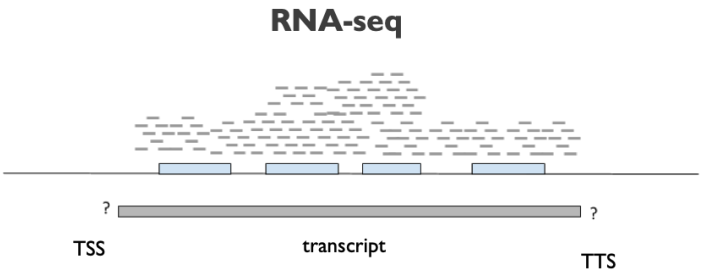


Figure 12: Limitations of RNA-seq for operon structure identification. Because the transcripts are fragmented and some of them are processed, it is very difficult to associate the start and end of specific transcripts and identify the number of transcript variants produced for a given operon.

Precise TSS mapping is important as it reveals the position of promoter elements such as transcription initiation factors and sigma factors binding sites. TSS marks the end of the 5′ untranslated region (5' UTR) of the transcripts, which often contains elements that regulate translation by forming secondary structures that can inhibit or promote translation of the downstream open reading frames (ORF). Moreover, 5'UTR can form riboswitches, regulatory molecules that sense the level of a chemical/physical signal, and can control the expression of downstream genes (Oliva, Sahr and Buchrieser, 2015).

Different strategies have been developed to identify the precise base position of TSS and they all exploit the fact that native RNAs are triphosphorylated at their 5′ end, while processed transcripts and ribosomal RNAs harbor a 5′ monophosphorylated end (Colgan, Cameron and Kröger, 2017). The most widely used method, differential RNA-seq (dRNA-seq), makes use of the Terminator 5′ Phosphate-dependent RNA exonuclease (TEX) that specifically degrades transcripts which exhibit a monophosphate at its 5′ end. The differential sequencing of two cDNA libraries: one generated from untreated RNA (TEX-) and another treated with TEX (TEX+) leads to an unenriched total RNA library and to a primary transcript enriched library. Comparison of the relative difference in sequencing depth between the two libraries at the 5'end of primary transcripts permits to annotate TSS (Sharma and Vogel, 2014). dRNA-seq notably allowed the first global identification of TSSs in the gastric pathogen *Helicobacter pylori* (Sharma *et al.*, 2010).

However, the exonuclease-dependent degradation of processed RNA fragments is not perfect and can be blocked by secondary structures, which could lead to incorrectly annotated TSS (Amman *et al.*, 2014; Wang, MacKenzie and White, 2015). It is based on these challenges of identifying the primary transcriptome and the transcription landmarks that Ettwiller *et al.* from New England Biolabs (NEB) developed Cappable-seq, a method that will be central to this thesis chapter and will be described in the next part.

2. Cappable-seq (Ettwiller *et al.*, 2016)

b. General principle of Cappable-seq

The first iteration of Cappable-seq aimed at identifying TSS of bacteria, genome-wide. Cappable-seq relies on the fact that only the primary transcripts contain a triphosphate present on the 5' nucleotide end. Indeed, the first step of most *in vivo* RNA degradation pathways in bacteria is believed to be the removal of the triphosphate, the maturation of RNA leaving a 5' OH or a 5' monophosphate (Schoenberg, 2007). Collectively these non-primary transcripts are referred to as *processed RNAs* and represent more than 95% of RNA in mass (Baracchini and Bremer, 1987). Thus, primary transcripts can be differentiated from *processed* RNAs based on their 5' end. Similarly to dRNA-seq that we previously described, this molecular distinction between *primary* and *processed* transcripts forms the basis of Cappable-seq. Cappable-seq differs from dRNA-seq as it involves a direct enrichment of primary transcripts rather than a depletion of 5' monophosphorylated transcripts. To enrich for the 5'end of primary transcripts that defines TSS, vaccinia capping enzyme (VCE) specifically caps triphosphorylated 5' end (5'PPP) of a primary transcript with a biotin-derived cap. These capped RNAs are fragmented and captured *via* a streptavidin bead system, allowing to isolate the 5'end of primary transcripts while removing uncapped RNAs. Thus, in addition to specifically capturing the primary transcripts of the bacteria, Cappable-seq also allows the removal of mature rRNA at the same time.

c. Method description and example of results

As mentioned earlier, the key step of Cappable-seq is the specific addition of a desthiobiotin-GTP onto the 5'PPP of primary transcripts using an enzyme called Vaccinia Capping Enzyme (VCE). The RNA is then fragmented prior to streptavidin enrichment, leading to the selection of the most 5' end fragment of the primary RNA. During this enrichment step, desthiobiotin-GTP capped transcripts specifically bind to the streptavidin beads, while the uncapped, and thus unbound, transcripts (processed RNAs, rRNAs and RNA in the process of being degraded) are washed away.
To obtain a cDNA library, the resulting cap structure of the selected primary transcripts is removed using RppH decapping enzyme leaving a ligatable 5'end. Finally, the Cappable-seq protocol

combined with a ligation-based library preparation protocol results in a cDNA library that can be sequenced on Illumina to identify TSS at nucleotide and strand resolution (**Figure 13**).

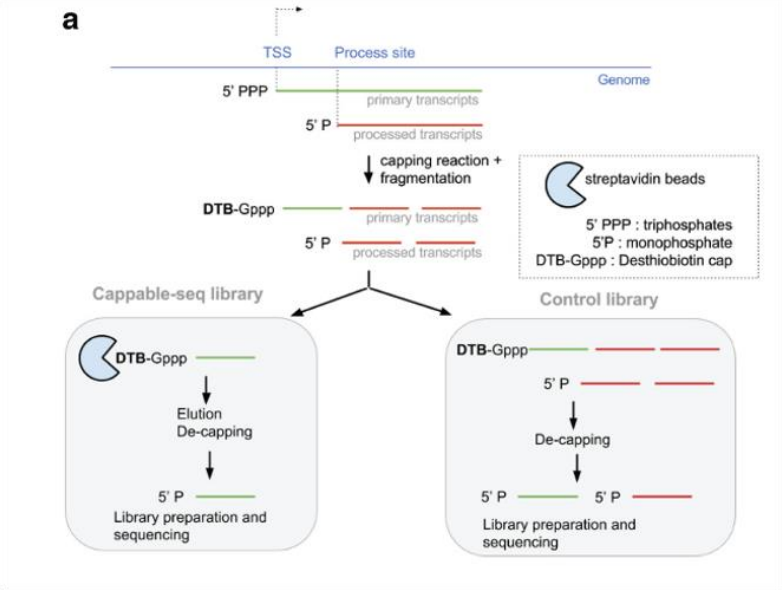


Figure 13: Principle of the Cappable seq protocol from (Ettwiller et al., 2016).

Applied to *E. coli,* Cappable-seq identified around 16,000 highly confident TSS clusters, detecting 76% of all *E. coli* genes, with only 4% of ribosomal RNA remaining. Cappable-seq libraries can be complemented with a control library for which the streptavidin enrichment step has been omitted. Similarly to differential RNA-seq (dRNA-seq), comparing the Cappable-seq with the control library offers the possibility to identify highly confident TSS. Still from the same study, Ettwiller *et al.* applied Cappable-seq to a mouse cecum microbiome sample, yielding to the first global TSS dataset of a gut microbiome. In addition, the method was able to uncover alternative modes of transcription such as leaderless transcription in *Akkermansia muciniphila*, an intestinal bacterium.

d. Short-reads limitations and the benefits of long-reads for Cappable-seq

Cappable-seq offers the possibility of sequencing at a high-throughput, resulting in a good quantification of the TSS usage. The identified TSS precisely locates the promoter, but the method is limited by the use of short reads. Indeed, identifying the full operon structure or transcripts isoforms

is impossible: as the RNA is fragmented into small pieces, the information on the transcription unit is lost. Short-read limitations also include bias in the data analysis, where ambiguous or multiple-mapping reads are difficult to handle (Stark, Grzelak and Hadfield, 2019). Multiple-mapping of short reads is particularly an issue in the case of a microbiome, where similar species or subspecies are often present. These reads are often discarded from the analysis, underestimating the presence of certain genes or organisms. Additionally, TSS reads cannot be used for *de novo* assembly of transcripts since they are all generated at the start of transcripts.

The emergence of long-read sequencing platforms (PacBio and Oxford Nanopore) offered new opportunities for transcriptome-wide analysis. The ambiguity in the mapping of sequence reads is reduced and full-length transcripts can be identified, which leads to a more complete capture of transcripts isoform diversity. Adapting the Cappable-seq technique to long-read sequencing offers the possibility to sequence full-length transcripts, delineate transcriptional landmarks (TSS and TTS) and ultimately, map the genome-wide operon structure of a bacteria.

3. SMRT-Cappable-seq (Yan *et al.*, 2018)

In 2018, the first long-read adaptation of Cappable-seq, SMRT-Cappable-seq, was published by Yan *et al.* This method combines the isolation of full-length primary transcripts with PacBio SMRT (Single Molecule Real-Time) sequencing. Similarly to the original Cappable-seq protocol, the 5’ triphosphorylated transcripts are first capped with a desthiobiotin GTP cap analog to capture the most 5'end. Then, a polyadenylation step (poly A-tailing) ensures the capture of the most 3'end before the transcripts are bound to the streptavidin beads to specifically capture primary transcripts. Finally, cDNA synthesis is performed through the priming of an anchored poly dT primer and a PacBio library for SMRT-sequencing can be prepared (**Figure 14**).

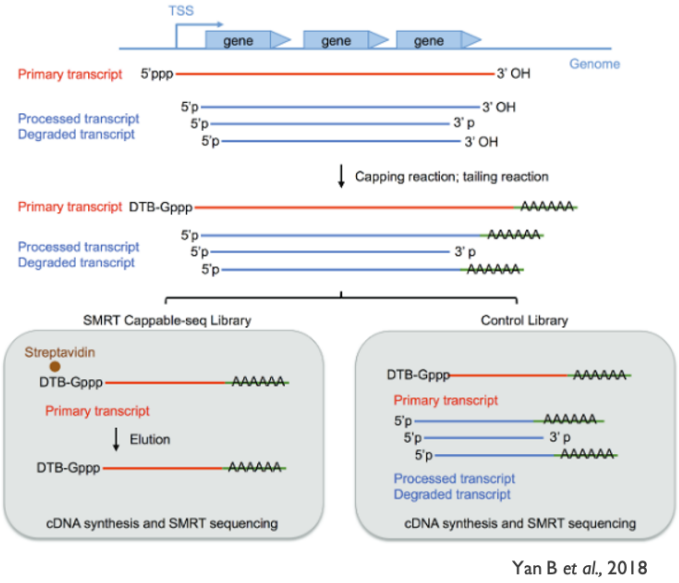


Figure 14: Principle of the SMRT-Cappable-seq protocol from (Yan et al., 2018).

The fact that RNA is not fragmented anymore and that full-length transcripts can be sequenced allows the determination of both TSS and TTS for each transcript. Applied to *E. coli*, this technology resulted in an accurate definition of the transcriptome with 34% of known operons from the RegulonDB being extended by at least one gene. In addition, this study showed that 40% of TTS have read-through at termination sites that alters the structure of the operons and that most of the bacterial genes are present in multiple operon variants. The phasing of TSS and TTS together along long distances reveals the complexity of operon structure, notably by identifying transcripts isoforms. Thus, the adaptation of Cappable-seq to long-read sequencing represents a valuable resource for the study of prokaryotic gene networks and regulation.

In addition, a new method called SEnd-seq was recently published. SEnd-seq is based on Cappable-seq and illumina circularized libraries, enabling to sequence both the TSS and the TTS from short-reads (Ju, Li and Liu, 2019). However, this technique requires a reference genome in order to phase the TSS and TTS and thus limits the technique's applications.

4. Adapting Cappable-seq to other long-read sequencing technologies

The association of the Cappable-seq technology with PacBio long-read sequencing permits the delimitation of transcription landmarks and the determination of the full operon structure in bacteria. This method published by Yan *et al* highlighted that Cappable-seq is a flexible technology that could be adapted to various sequencing platforms. Two other long-read platforms compete with PacBio on the long-read market, each one with its own advantages/disadvantages: Nanopore (ONT) and LoopSeq (Loop Genomics). The first platform provides throughput and affordability, while the second provides unprecedented accuracy.

In this section, we present the development of two new flavors of long-read Cappable-seq: ONT-Cappable-seq, adapted to the Nanopore MinION and Loop-Cappable-seq, adapted to the LoopSeq long-read sequencing platforms.

B. Development of ONT-Cappable-seq and comparison of different strategies to capture the 3'end

1. Introduction

The Nanopore sequencing technology from Oxford Nanopore Technologies (ONT) offers a high-throughput, affordable and easy to handle alternative to the PacBio platform. While Cappable-seq robustly captures the 5'end of the transcripts, defining prokaryotic TTS remains challenging because transcripts lack the 3’ poly-A tail used to ligate adapter sequences to eukaryotic transcripts. The common method to capture prokaryotic transcript's ends relies on the addition of a polyA tail to the 3'end of the transcripts using the polyA polymerase enzyme. This polyA tail will serve as an anchor for an oligod(T) primer used during the reverse transcription step to synthesize the first strand of cDNA. But it has been shown that polyA tailing can add biases. In the case of adenine-rich regions, the oligod(T) primer can prime internally to the adenine-rich region of the transcript and initiates reverse transcription from this region rather than the added polyA tail (Balázs *et al.*, 2019; Sessegolo

et al., 2019). This results in truncated cDNA molecules and errors in TTS identification, especially for bacteria containing AT-rich genomes. Accurate capture of the 3'end of the transcripts is critical as it not only alters the definition but also the quantification of transcriptional units.

So, in addition to the ONT-Cappable-seq development, we investigated different strategies to robustly capture the 3'end and developed a new strategy based on splint ligation. For this, we used *Escherichia coli* to first validate the method on a widely used model organism and we then used *Clostridium phytofermentans (C. phy)* as a model of AT-rich genome to investigate the effect of different strategies to capture the 3'end.

ONT-Cappable-seq is based on the previously developed SMRT-Capable-seq and was adapted to the MinION sequencing platform from Oxford Nanopore Technologies. The scheme below (**Figure 15**) illustrates the different steps of the protocol. The complete protocol is described in the material and methods part of this chapter. Briefly, the synthesized cDNA is used to prepare a nanopore library, using the genomic DNA by ligation kit (LSK-109, Oxford Nanopore Technologies). If samples need to be multiplexed on the same flowcell, the barcoding kit EXP-NBD104 (Oxford Nanopore Technologies) can be used prior and in combination to the LSK-109 kit.

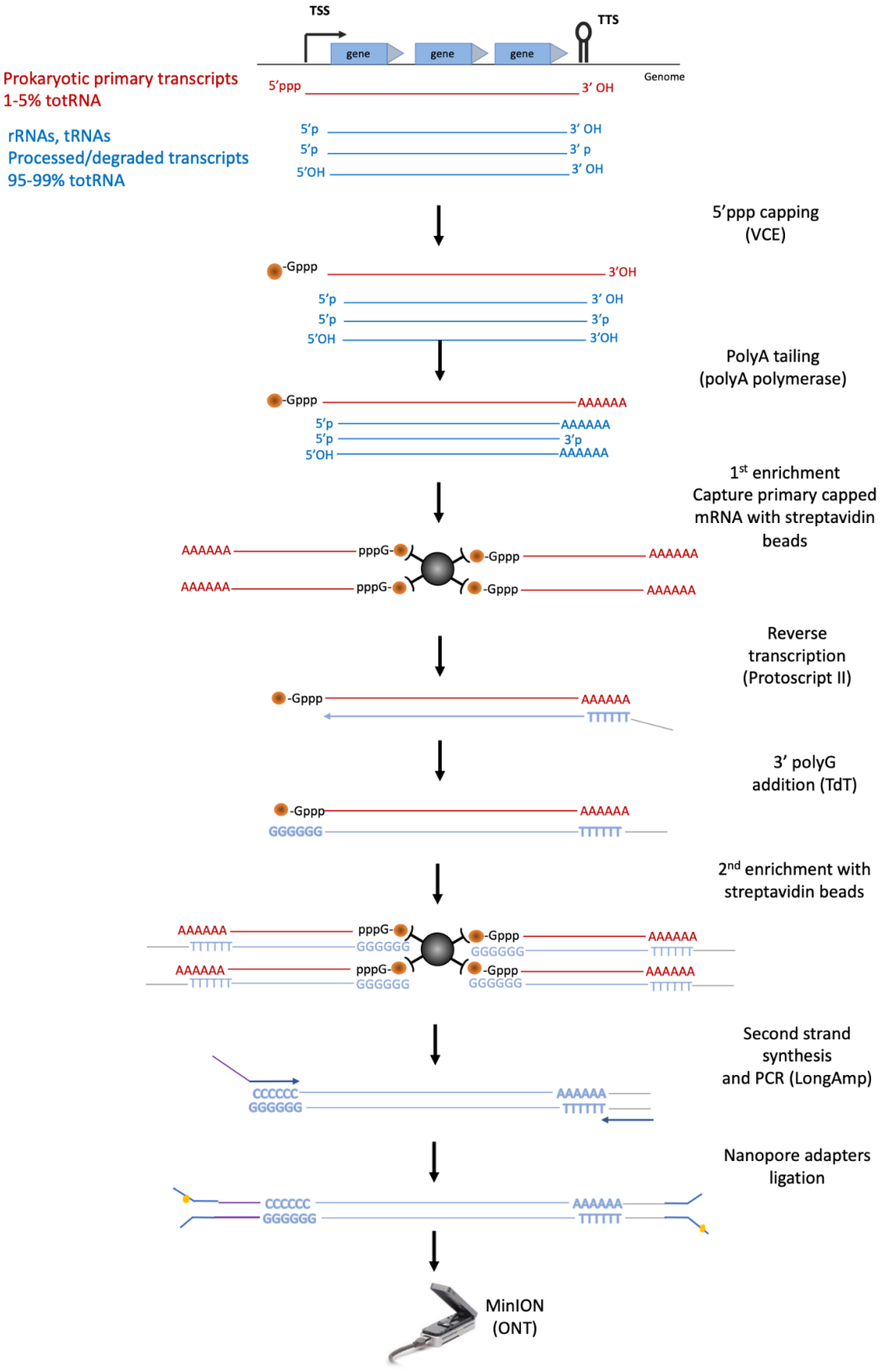


Figure 15: Detailed overview of the ONT-Cappable-seq method.

2. Material and Methods

Clostridium phytofermentans (*C. phy*) is a Gram positive, anaerobic bacterium that can ferment diverse plant substrates, such as cellulose, hemicellulose and pectin. Understanding gene regulation in such plant-fermenting bacteria has a significant potential for biotechnological applications, such as biofuel production, as these bacteria can serve as biocatalysts for industrial transformation of plant biomass.

A previous study using a technical variant of Cappable-seq (Capp-switch) investigated genome-wide patterns of *C. phy* transcription initiation on various plant substrates and demonstrated condition-dependent transcription regulation modifications. Amongst these changes, interesting regulatory mechanisms such as antisense transcription, leaderless transcription and non-coding RNA were identified (Boutard *et al.*, 2016).

We first developed and validated the ONT-Cappable-seq method using *Escherichia coli*. In a second time, using *Clostridium phytofermentans* as an AT-rich model bacterium, we investigated the effect of various strategies for *in vitro* tagging of 3' transcript ends: polyA tailing, polyU tailing, single strand ligation and splint ligation based on polyA tailing.

Culture of *Escherichia coli* K12 MG1655 and RNA extraction

E. coli total RNA was prepared and provided by Bo Yan, it is the same RNA used for the SMRT-Cappable-seq publication (Yan *et al.*, 2018). Briefly, *E. coli* K12 strain MG1655 was grown at 37°C in M9 minimal medium with 0.2% glucose. The culture was grown to late log phase (OD600 = 0.6). Two volumes of RNAlater (Life Technologies) were added to the culture and saved at 4°C overnight. The RNA was extracted using the RNeasy Midi kit (Qiagen). The isolated RNA had a RNA integrity number (RIN) above 9.0 as determined by Bioanalyzer (Agilent), and was used for SMRT-Cappable-seq.

Culture of *Clostridium phytofermentans* and RNA extraction

Clostridium phytofermentans ISDg ATCC 700394 (generously provided by Jeffrey L. Blanchard) was cultured at 30°C in GS2 medium (Johnson, Madia and Demain, 1981) containing 0.5% regenerated amorphous cellulose (RAC) from Avicel PH-101 (Sigma 11365). RAC was prepared by phosphoric acid

treatment (Hong *et al.*, 2008). The culture was incubated for 50h in anaerobic jars (260626, BD) containing one GasPak (260678, BD) to produce an anaerobic atmosphere, before harvesting the cells by centrifugation. Total RNA was extracted using the RNeasy Maxi kit (75162, Qiagen), using 5mg/mL lysozyme (L6876, Sigma) for lysis at 37°C during 30min. After extraction, the total RNA was treated with DNase I (M0303, New England Bioabs) for 1h at 37°C. RNA was purified using an Acid Phenol Chloroform extraction pH 4.5 (AM9720, Ambion). As we want to use this RNA for ONT-Cappable-seq and long read sequencing, the RNA quality and integrity is very important. The isolated RNA quality was checked using a Bioanalyzer (Agilent) and had a RNA integrity number (RIN) above 9.0.

ONT-Cappable-seq library preparation

Capping of full-length prokaryotic transcripts

The capping reaction was done using 9µg of *C. phy* total RNA. The RNA was incubated in the presence of 0.5 mM DTB-GTP (N0761, New England Biolabs) and 100 units of Vaccinia Capping Enzyme (M2080, New England Biolabs) and 0.25 units of *E. coli* pyrophosphatase (M0361, New England Biolabs) for 1h at 37°C in 50µl reaction volume. In order to measure the recovery of triphosphate transcripts, 1ng of in vitro synthesized Gluc (Gaussia Luciferase) transcripts can be mixed with the RNA in the capping and following reactions. The capped RNA was purified using the Zymo Clean and Concentrator 5G columns (R1013, Zymo) and eluted in 30µL of low TE buffer. Low TE buffer: 1mM Tris-HCl pH7.5; 0.1mM EDTA

Capture of the 3'end of full-length prokaryotic transcripts

The protocols differ depending on the 3'end method that has been investigated and a detailed version for each method is presented in the **Appendix**. Below, we present the polyA tailing version, which was the method developed in the first intention.

A polyA tail was added *in vitro* by incubating the capped RNA in 50µl reaction volume with 5 units of *E. coli* Poly(A) Polymerase (M0276, New England Biolabs) and 1 mM ATP for 15min at 37°C. The capped and tailed RNA was purified using the Zymo Clean and Concentrator 5G column kit and

eluted in 33µL low TE buffer. A volume of 3µL of the reaction (non-enriched RNA) can be put aside and used as control.

Enrichment of full-length prokaryotic transcripts

The capped RNA was enriched for a first round using hydrophilic streptavidin magnetic beads (S1421, New England Biolabs). A volume of 35µL of beads were prepared by washing 3 times with a Washing Buffer before being resuspended in 35µL of Binding Buffer (the difference lies in the NaCl concentrations, see above for composition). A volume of 30µL of capped and tailed RNA was incubated with 30µL of prepared streptavidin beads at room temperature for 30min on a hula mixer rotator. The beads were then washed thoroughly 3times with 60µL of Wash Buffer. To elute the RNA, the beads were resuspended in 30µL Elution buffer containing biotin (allowing to release the DTB-capped native transcripts from the streptavidin beads), and incubated at 37°C for 30min on a rotator. The 30µL biotin eluted enriched RNA were collected using a magnetic rack.

Binding Buffer: 10mM Tris-HCl pH7.5; 1mM EDTA; **1M NaCl**

Washing Buffer: 10mM Tris-HCl pH7.5; 1mM EDTA; **250mM NaCl**

Elution Buffer: **1mM Biotin**; 10mM Tris-HCl pH7.5; 0.1mM EDTA; **50mM NaCl**

First strand cDNA synthesis

A volume of 30µL of enriched RNA (and 3µL of non-enriched RNA if a control is needed) were used in 80µL first strand cDNA synthesis reaction. First, the RNA is incubated at 65°C for 2min with 5µM polydT oligo (RT_dTVN oligo), 1mM dNTP and then cooled down on ice. Next, 400 units of ProtoScript II Reverse Transcriptase (M0368, New England Biolabs) and murine RNAse inhibitor (M0314, New England Biolabs) were added to the reaction and incubated at 42 °C for 1h. After the first strand cDNA synthesis, 50 units RNase If (M0243, New England Biolabs) was added and incubated at 37°C for another 30min to remove single stranded RNA. The reactions were purified using 1.0X AMPure beads and eluted in 40µL Low TE, before being concentrated to 22µL using a Speed Vac for 15min.

A polyG was added to the 3’ end of cDNA for second-strand synthesis using TdT Terminal transferase (M0315, New England Biolabs). The purified cDNA/RNA duplex samples were incubated with 10 units of TdT enzyme and 3mM dGTP at 37°C for 30 min.

The RNA was enriched a second round using 30µl hydrophilic streptavidin magnetic beads as mentioned above, but was eluted in 30µL Low TE and not in the Elution buffer elution. The cDNA/RNA is on the streptavidin beads, so be careful not to discard the beads. If a control library is used, the polyG reaction products were purified using 1.0X AMPure beads and eluted in 30µL of Low TE.

At this point, if a control library is present, a qPCR can be performed to assess the quality of the libraries and of the enrichment by comparing the enrichment of the control vs enriched libraries. Gluc and rRNA gene primers can be used to respectively assess the enrichment of primary transcripts and the depletion of ribosomal/processed transcripts.

Second strand synthesis and PCR

Second-strand cDNA synthesis was performed on first strand cDNA/RNA duplex samples in 50µL reaction volume, using the LongAmp polymerase (M0533, New England Biolabs) a primer (Pac_oligodC20) containing several Cs allowing to prime to the added Gs and RNAse H (M0297, New England Biolabs). The reaction was first incubated at 37°C for 15min to allow the removal of residual RNA strands left after reverse transcription, then for 1 min at 94°C, next 4 cycles of 94°C for 1 min and 65°C for 15min, with a final extension time of 10min at 65°C. The tube was placed on a magnetic rack to keep the supernatant containing the double stranded cDNA. Finally, the reactions were purified using 1.0X AMPure beads and eluted in 50µL Low TE. Prior tests were done to determine the optimal number of PCR cycles required to amplify the cDNA library (10 cycles for the enriched library, but need to be adapted according to each sample). The PCR was performed using LongAmp polymerase, Pac_oligo_for_set2 and Pac_rev primers in a final volume of 50µL. The cycling was the following: 94°C for 1 min, then 10 cycles of 94°C for 30sec and 65°C for 8min, with a final extension time of 10min at 65°C. The PCR reaction was purified and size selected using 0.5X AMPure beads and eluted in 30µL Low TE.

Nanopore sequencing

The 1D Native barcoding genomic DNA kit (EXP-NBD104 and SQK-LSK109 kits, Oxford Nanopore Technologies) were used for library preparation. The barcode ligation was done using 500ng of PCR products, following Oxford Nanopore protocol (EXP-NB104), except that the incubation time was increased to 30 min. After clean-up using 1.0X AMPure beads and following the manufacturer's protocol, the samples were quantified using Qubit fluorometer (Thermo Fisher Scientific, Waltham MA, USA) and pooled in equimolar amounts to produce a final amount of 700 ng. The pool was ligated to the Nanopore adapter and purified according to the manufacturer's protocol. The library was quantified and 40-50fmol of this library was sequenced on a MinION or a GridION (Oxford Nanopore Technologies) using a FLO-MIN106 Rev D flow cell. Depending on when the data were generated, the Raw fast5 data were generated using MinKNOW and base called using Albacore or using the more recent Guppy base caller (in high accuracy mode).

Data analysis

Data preprocessing. For the *E. coli* experiment, reads were trimmed using Porechop (https://github.com/rrwick/Porechop) and mapped to the *Escherichia coli* str. K-12 substr. MG1655 (NC_00913.3) reference genome using Minimap2 version 2.10-r761 (Li, 2018). For the Clostridium phytofermentans experiment, reads were trimmed using Guppy 3.0.6 and mapped to *Clostridium phytofermentans* ISDg (CP000885.1) reference genome using Minimap2 version 2.10-r761.

Operon definition. The TSS and TTS were defined using custom scripts developed by Bo Yan from New England Biolabs (https://github.com/elitaone/SMRT-cappable-seq) that she adapted from the SMRT-Cappable-seq pipeline to the Nanopore data. For more details, please see the Material and Methods section from (Yan *et al.*, 2018)

Termination site predictions. Rho-independent transcription terminator positions were predicted using TransTermHP (Kingsford, Ayanbule and Salzberg, 2007). This algorithm, available at http://transterm.ccb.jhu.edu/, predicts rho-independent TTS from a genome (fasta file). The TTS positions that intersect within a -5/+5bp window of the predicted TTS positions were determined

using bedtools v2.27.1 (Quinlan and Hall, 2010) using slop (parameters -s -l 5 -r 5) and intersect (parameters -s -wa -wb).

Motif logo analysis at TTS and TSS. The sequence context of the −45bp to 5bp region around the defined TSS and the −30 bp to 10 bp region around the defined TTS were extracted for motif analysis. Motif logos were generated using the program weblogo3.6.0 (Crooks *et al.*, 2004).

Determination of intragenic/Intergenic TTS. The TTS positions located in a gene were determined using the annotation file and bedtools intersect v2.27.1 (parameters -s -u). Conversely, the TTS positions located outside a gene were determined using the annotation file and bedtools intersect (parameters -s -v).

Correlation analysis. Gene expression was determined using bedtools multicov version v2.27.1 (-s parameter) for the RNA-seq, SMRT-Cappable-seq and ONT-Cappable-seq datasets. After calculating the RPKM (Reads Per Kilobase of transcript), gene correlation was then plotted and the Pearson correlation coefficient was calculated using R (R Foundation for Statistical Computing, Vienna, Austria., 2020) and ggplot2 (Wilkinson, 2011).

3. Results

a. Validation of ONT-Cappable-seq using *E. coli*

As previously mentioned, we decided to develop and validate the new ONT-Cappable-seq method on a well-known model organism, *Escherichia coli*, grown in M9 minimal medium. We used RNA that has previously been used in the published paper presenting SMRT-Cappable-seq in order to directly compare both methods. We also benchmarked our methods to RNA-seq data obtained from the European Nucleotide Archive (SRR3132588) that used RNA from the same *E. coli* strain and grown in M9 medium as well.

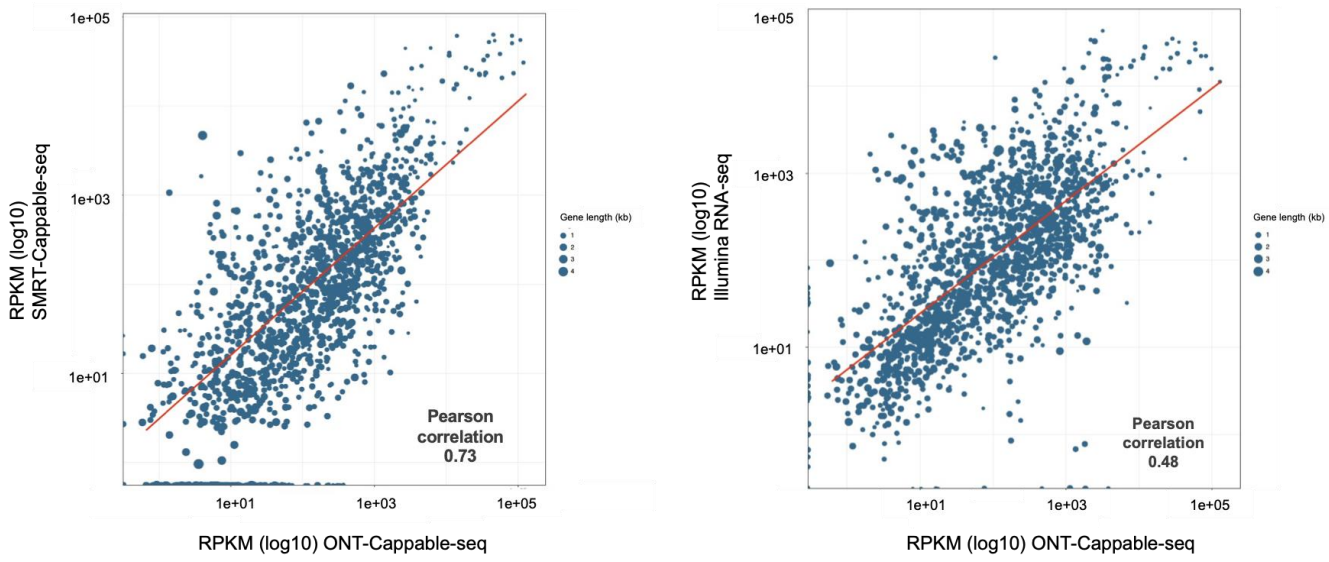


Figure 16: Gene expression correlation for SMRT-Cappable-seq vs ONT-Cappable-seq (left) and for ONT-Cappable-seq vs Illumina RNA-seq (right). The RPKM (Reads Per Kilobase of Transcript) and the Pearson correlation were calculated for all the data.

Figure 16 shows a decent correlation of gene expression between the two long-reads Cappable-seq technologies (Pearson correlation = 0.73). A lower correlation is observed when comparing ONT-Cappable-seq to RNA-seq (Pearson correlation = 0.48). This difference is due to the size selection performed with the long-reads; some short genes may not be present in the library while the short-reads RNA-seq on Illumina captures them. Overall, the ONT-Cappable-seq showed similar results to the SMRT-Cappable-seq and we validated the technology.

b. Investigation of different strategies to capture the 3'end using *C. phy*

Cappable-seq provides an accurate definition of the 5′end of transcripts and the current strategy to obtain the 3′end is based on polyA tailing. While this strategy proved to be successful in *E. coli,* polyA tailing may not be adapted to AT rich genome, where naturally occurring genomic poly-nucleotide stretches of A or T can confound the identification of the 3′end of transcripts. We therefore decided to apply ONT-Cappable-seq on an A-T rich genome microorganism, *Clostridium phytofermentans (C. phy)* cultured on cellulose (RAC), using four different strategies to capture the 3'end of transcripts with the aim to determine the best strategy that would give a robust 3'TTS

definition. One of such strategies has been especially developed for this purpose (splint ligation based on polyA tailing). The different methods and their principles are described below. The 4 methods were evaluated and compared based on different criteria, such as their ability to identify 5'TSS and 3'TTS sequences, the percentage of correct TTS identified and the correlation of the gene expression with RNA-seq data.

- **polyA tailing:** a polyA tail is added using the polyA polymerase enzyme (M0276, New England Biolabs). This is the standard way to capture the 3'end of the prokaryotic transcripts.
- **polyU tailing**: a polyU tail is added using the polyU polymerase (M0337, New England Biolabs). The principle is the same as for the polyA tailing, except that a polyU tail is added instead of a polyA tail.
- **single strand ligation:** a single-stranded DNA adapter is ligated directly onto the 3'end of the transcript. Because this strategy is ligation-based, in theory only the true 3'end is captured, and internal priming that can be observed with tailing is prevented. We used the thermostable 5´ App DNA/RNA Ligase (M0319, New England Biolabs) to perform the ligation. This enzyme requires a 5´ pre-adenylated adapter for ligation to the 3´OH end of either RNA, preventing the formation of undesired ligation products (concatemers). Also, this enzyme works at higher temperature (65°C), so it might reduce the constraints of RNA secondary structure that can prevent ligation. Because the ligation buffer contains 10mM MgCl2 and the reaction is performed at high temperature, we performed preliminary tests to optimize the magnesium concentration in the buffer and the ligation time in order to prevent degradation of the RNA. We found that a 1mM MgCl2 buffer incubated for 30min at 65°C would prevent most of the RNA degradation.
- **splint ligation based on polyA tailing:** this is a new strategy adapted from the technique published by Maguire *et al* (Maguire, Lohman and Guan, 2020). The original splint ligation strategy relies on the direct ligation of a double-stranded DNA adapter containing Ns on the 3'end of the bottom strand. We adapted this technique by replacing the Ns with polyTs so that it would specifically hybridize to the 3'end of the transcripts that are previously polyA tailed. The detailed principle is shown in **Figure 17** below. In brief, the adapter is composed

of a top strand containing a 5'P and a 3'end that is blocked (with an inverted dT modification), preventing elongation by the reverse transcriptase. The bottom strand contains, from the 5'end, a sequence that will be used for reverse transcription, followed by a deoxyuracil (dU) nucleotide that can be cleaved using the USER enzyme (M5508, New England Biolabs), followed by 18 Ts and an inverted dT blocking modification on the 3'end. The rows of Ts allow the hybridization of the adapter on the polyA tail previously added at the 3'OH of the transcript. The blocking modifications prevent truncated transcripts that can occur following hybridization to internal Adenine-rich regions. Once the ligation reaction is performed, the USER enzyme is used to cleave the adapter at the dU nucleotide. The remaining ligated sequence will serve as an RT oligo to perform reverse transcription and cDNA synthesis.

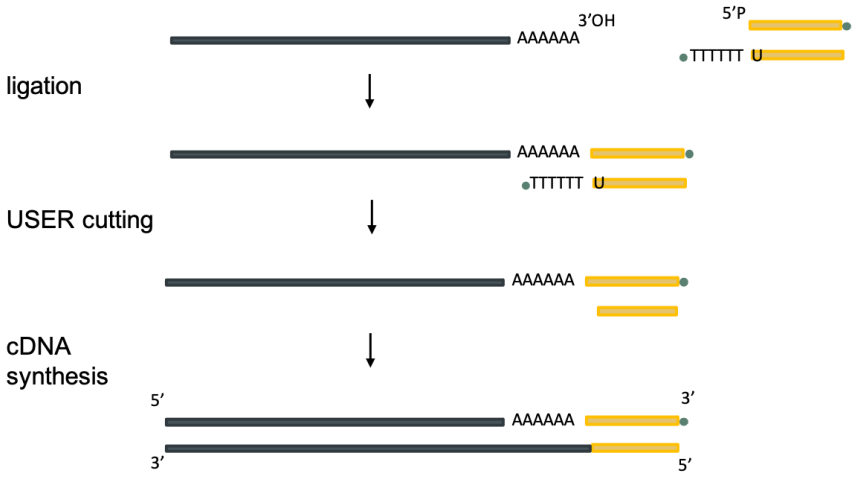


Figure 17: Principle of the splint polyA ligation. First, the double-stranded DNA adapter is ligated on the 3'OH of the RNA using splint ligation. Following adapter ligation, the bottom portion of the adapter is cleaved off by excising the deoxyuracil (U) using USER. Next, cDNA is synthesized using the remaining portion of the 3'bottom strand adapter that serves as a primer for the reverse transcription. The green dot on the 3'ends represents a blocking inverted dT modification.

3'end TTS sequence identification

Two different mechanisms of transcription termination have been described in bacteria: the rho-dependent termination and the rho-independent (or intrinsic) termination. The rho-dependent termination depends on the rho protein, which binds on the nascent mRNA, translocates up to the elongation complex and dissociates it. This interaction of the RNA polymerase with the rho protein provokes the mRNA release from the transcription complex and the transcription terminates (Jain, Gupta and Sen, 2019; Roberts, 2019). This type of transcription termination requires specific C-rich and G-low sequences, called *rut* (rho utilization) sites, to be present on the mRNA, (Ciampi and Sofia Ciampi, 2006). The computational prediction of rho-dependent terminators is difficult because the sequences required for the binding of the rho protein are complex, poorly defined and vary amongst bacterial species. As an example, even amongst several *E.coli* rho-dependent terminators that have been identified, no consensus sequence could be identified (Graham, 2004; Peters, Vangeloff and Landick, 2011; Grylak-Mielnicka *et al.*, 2016).

The rho-independent termination involves the formation of a secondary structure hairpin loop in the mRNA sequence upstream of the TTS. This type of terminator consists of a short GC-rich stem-loop, followed by a hairpin and by a polyU-rich region on the 3' end of the RNA (illustrated in **Figure 18**). The hairpin causes the RNA polymerase to stall and the transcription stops. Because of the stem-loop structure and the polyU-rich region, rho-independent TTS can be relatively easily identified using prediction programs, such as TransTermHP.

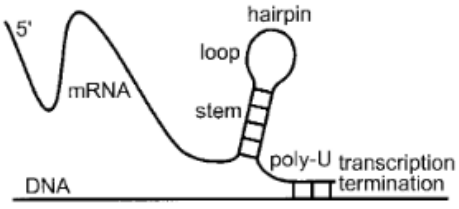


Figure 18: Model of a rho-independent transcription terminator (Ermolaeva et al., 2000).

Accordingly, we analyzed the sequence motif found within a window of +10nt and -30nt around the TTS positions predicted using the four strategies described above. The polyA tailing, the splint polyA

tailing and the polyU tailing methods identified a sequence composed of a stretch of Ts mostly upstream of the predicted TTS, consistent with a rho-independent terminator. However, the single stranded ligation did not identify any specific sequence upstream the predicted TTS (**Figure 19**).

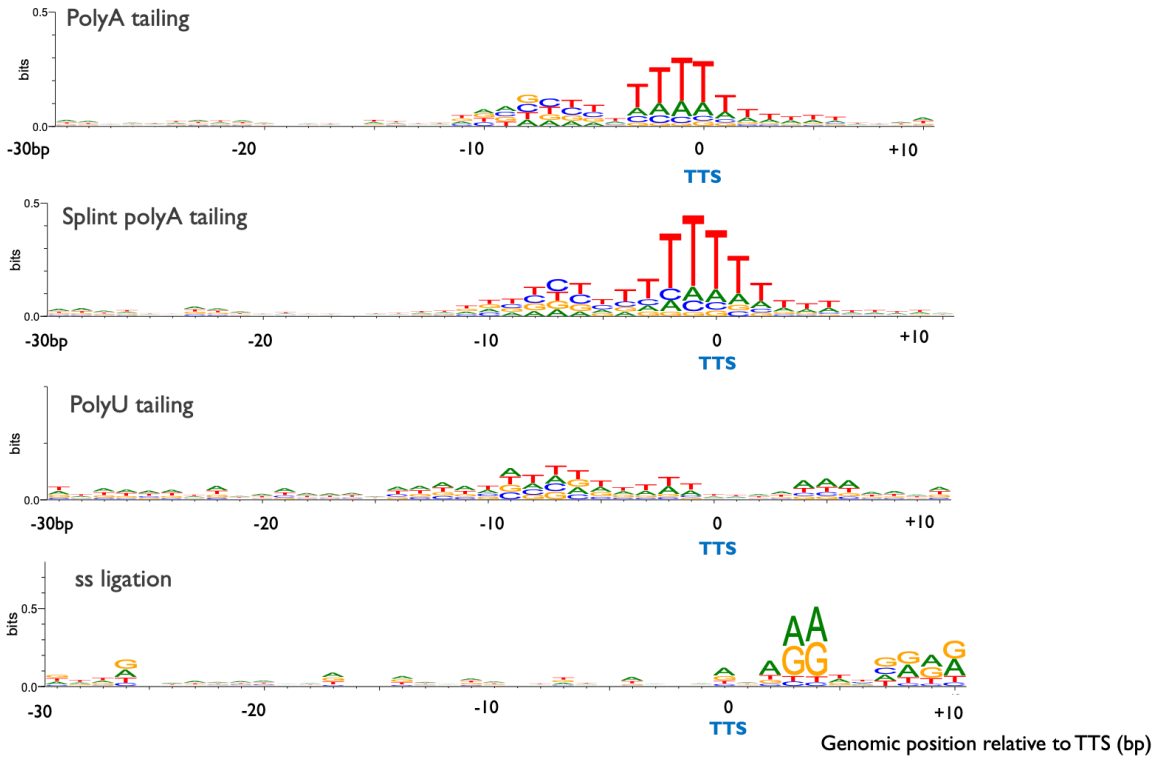


Figure 19: Transcription termination site (TTS) motifs determined by each method, found within a window of +10nt and -30nt around the TTS position (located at position '0'). Data obtained from C. phy grown on cellulose.

In order to validate the positions identified by the different methods, we used TransTermHP, an algorithm to predict the rho-independent terminator positions. We compared the experimental TTS positions with the computationally predicted ones, allowing a 10nt window around the predicted TTS position. PolyA tailing and splint ligation results correlate the best with the predicted rho-independent terminators, with > 60% of the TTS positions matching. However, the TTS identified with the single stranded ligation gave a very poor correlation (2% only) with the predictions, highlighting there is an issue with this method. In addition to this comparison, we calculated the proportion of TTS positions located inside and outside a gene, knowing that most of the TTS are

expected to be identified outside of a gene (**Table 4**). We found that more than 80% of the TTS were located outside a gene for the polyA, polyU tailing and splint ligation, while most of the TTS identified by the single stranded ligation appeared to be located inside a gene, highlighting again that this last method cannot be used to define robust TTS.

	% TTS positions inside a gene	% TTS positions outside a gene	% TTS positions corresponding to predicted Rho-indep. TTS (10bp window)
polyA tailing	17%	**83%**	62%
Splint ligation polyA	11%	**89%**	66%
polyU tailing	20%	80%	54%
Single strand ligation	89%	11%	**2%**

Table 4: Key statistics on TTS positions determined according to each 3'end strategy. The TTS positions determined experimentally were compared to predicted rho-independent TTS positions. The prediction was done using TransTermHP (Kingsford, Ayanbule and Salzberg, 2007).

5'end TSS sequence identification

We also analyzed the sequence motif of the TSS, for each method. Sequences upstream of TSS generally contain the consensus sequences −35 TTGACA and −10 TATAAT recognized by the sigma factor subunit of the RNA polymerase (RNAP). As the 4 methods only differ in the 3'end strategy that is used, the capture of the 5'end and thus, the TSS identification, should not be impacted. We observed similar sequence motifs for the -35 and -10 regions for all the methods. The TSS showed a nucleotide preference for A or G, observation that has previously been reported by Ettwiller *et al* and others (Hawley and McClure, 1983; Kim *et al.*, 2012; Ettwiller *et al.*, 2016) (**Figure 20**).

Overall, we found similar 5'end motifs in the promoter region, independently of the 3'end method. This result indicates that ONT-Cappable-seq is able to robustly identify TSS and promoter sequences in *C. phy*.

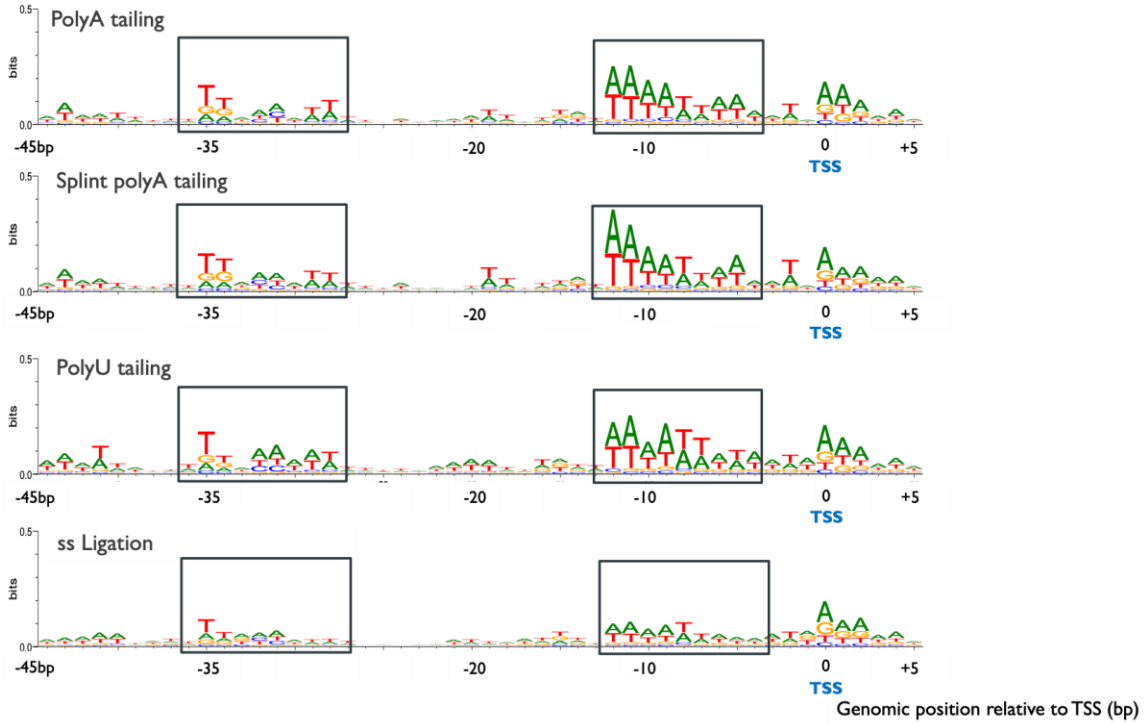


Figure 20: Consensus promoter motifs determined by each method. The -35 and -10 motifs are recognized by the RNA polymerase. The TSS is located at position '0'. These data obtained from C. phy grown on cellulose using different methods to capture the 3'end of transcripts.

Correlation with RNA-seq data

The last feature we used to evaluate the different strategies is the gene expression correlation of each method with RNA-seq. For this, we calculated levels of gene expression for each method and compared it to the gene expression results obtained from RNA-seq that was performed on the exact same RNA. The results are shown in **Figure 21**. PolyA tailing and splint ligation correlate the best with RNA-seq data with a Pearson correlation coefficient of 0.87 and 0.85, respectively. PolyU tailing has a slightly lower correlation but still correlates well, while single stranded ligation has the worst correlation (Pearson = 0.18).

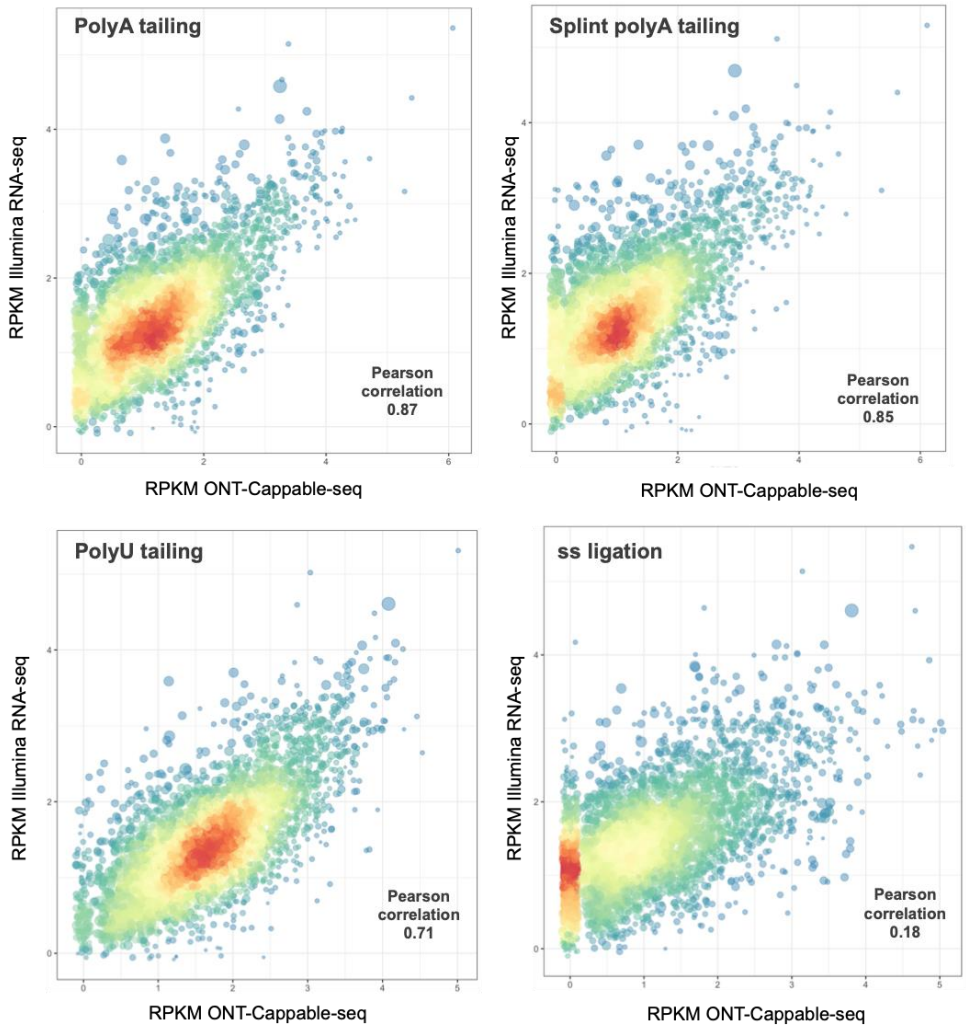


Figure 21: Gene expression correlation between RNA-seq and ONT-Cappable-seq data on C. phy grown on cellulose.

General results overview

Figure 22 below illustrates the results obtained using ONT-cappable-seq, showing the method's ability to define transcripts landmarks. Both methods polyA tailing and splint polyA ligation identified similar operon structures in *C. phy* grown on cellulose. The RPKM (Reads Per Kilobase of Transcript) and the Pearson correlation were calculated for all the data.

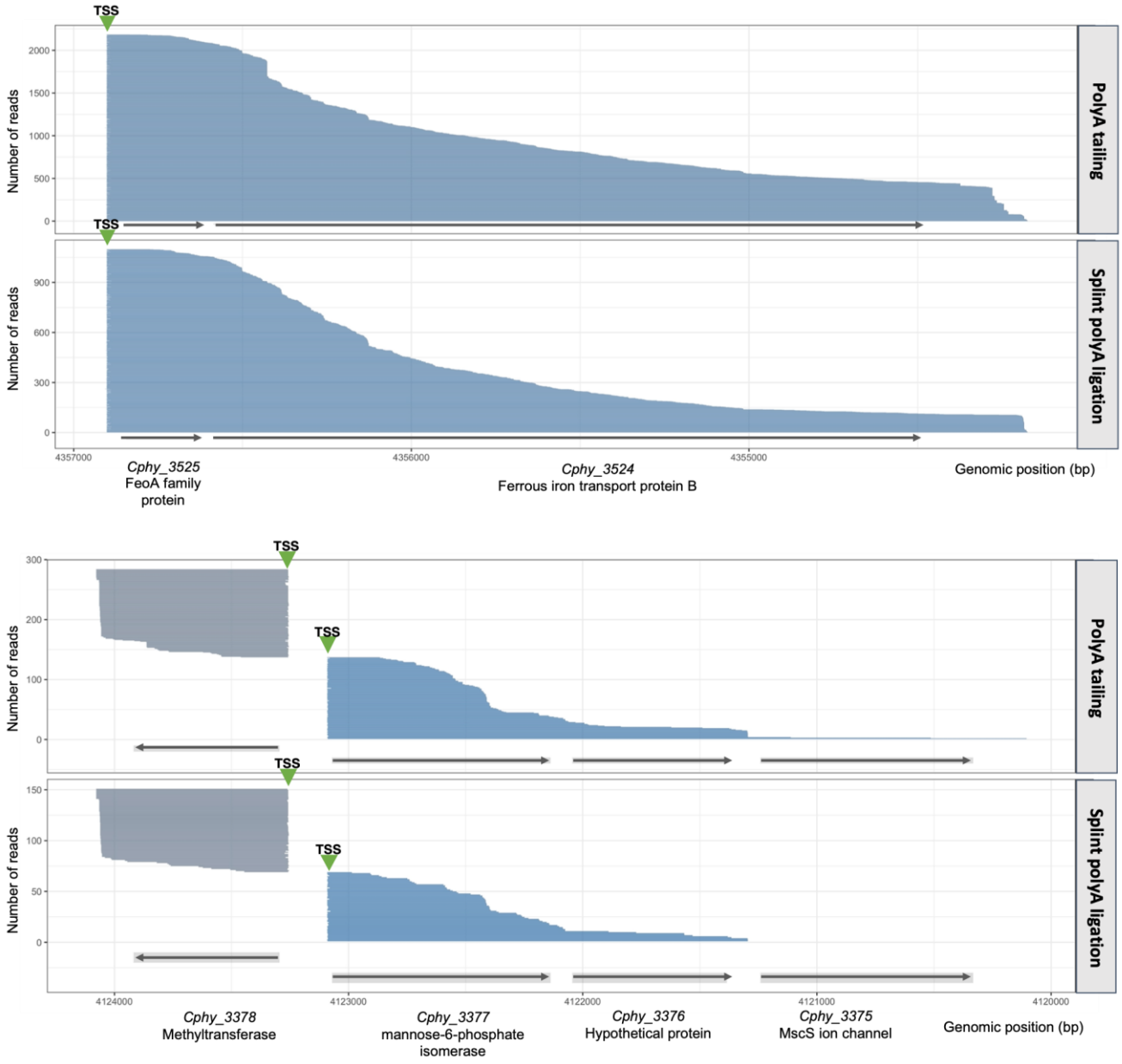


Figure 22: Example of operons identified in C. phy grown on cellulose substrate. The x-axis represents the position (in bp) on the reference genome (CP000885.1) and the y-axis represents individual mapped reads ordered by read size in ascending order. The TSS are indicated by a green arrow. The genes are indicated by a grey arrow and and are annotated. Reads going in the 'forward' direction of the genome are in blue, while the reads going in the 'reverse' direction are in grey.

4. Discussion and further outlooks

Overall, the polyA tailing and the splint ligation performed the best in terms of TTS identification, correlation with rho-independent predicted terminators and the gene expression obtained from these datasets correlated well with RNA-seq data. PolyU tailing gave satisfying results over these different parameters but the library yield was much lower compared to the other methods. We hypothesize this could be due to the polyU polymerase that is less processive than the polyA polymerase, leading to a lower proportion of transcripts tailed. As polyU tailing was used in the context of a comparison with other methods, we did not try to optimize the reaction. Conversely, single stranded ligation showed the worst results over all the features used in the comparison. Most of the TTS positions were identified within a gene and the data did not correlate with the predicted rho-independent terminators nor with RNA-seq. These results suggest that 3' truncated transcripts are being captured and sequenced leading to the false identification of TTS. We hypothesize this could be due to magnesium in the buffer, an essential component for ligation known to catalyze RNA degradation. Indeed, even if we optimized the ligation conditions that could have an impact on RNA integrity (temperature and time), it seems RNA degradation still happens.

While benchmarking the 3'end strategies, we realized later in the course of the project that the tube of polyA polymerase enzyme used in these experiments showed low activity. We hypothesized that the resulting polyA tails were shorter than expected, with some transcripts being not tailed at all. If very few or no A's were added onto the 3'end of the transcript, the RT primer could prime on internal A-rich regions, leading to the capture of truncated transcripts. This phenomenon would be amplified in A-T rich genomes such as *C. phy*. This technical problem delayed the ONT-Cappable-seq project and not all the experiments planned could be done, such as the investigation of the dynamics of operon structure changes according to the carbon source used by *C. phy.*

Another example of application for ONT-Cappable-seq is the use of the method on a complex community. The throughput and the long-reads provided by the Nanopore platform enables the identification of complete operon structure of the bacteria present in a sample. This notably offers the possibility to directly predict Open Reading Frames (ORFs) and get insights in the community

functionality, removing the need for a reference genome and annotation, which is often a limiting factor in microbiome studies.

In summary, both the polyA tailing and splint polyA ligation are valuable methods to capture the 3'end and provide similar results. The technical issues encountered with the polyA polymerase demonstrated that the choice of the method employed is crucial as it can have a profound impact not only on the TTS identification but also on transcripts identification and quantification. It shows the importance of having a robust strategy for capturing the 3' end in order to obtain an accurate transcriptome using long-read sequencing, especially when applied to complex communities where a diversity of genomes is present. Still, we developed a new strategy, the splint polyA ligation, that prevents internal priming that can be observed when using polyA tailing (especially in AT-rich bacterial genomes) and thus, prevents any false TTS identification. This technique, applied to a microbiome, would provide a robust and reliable capture of the transcriptome.

Another application for ONT-Cappable-seq moves away from the prokaryotic domain. A further outlook for this method would be to combine it with ReCappable-seq, developed by Yan *et al,* at New England Biolabs (Yan *et al.*, 2021). ReCappable-seq is the eukaryotic version of Cappable-seq that allows TSS identification from short-read sequencing. In eukaryotes, transcripts can be classified in 3 categories, depending on if they originate from RNA polymerase I (Pol I), polymerase II (Pol II) or polymerase III (Pol III) (Carter and Drouin, 2009). Pol I and Pol III transcripts harbor a 5'PPP and can be specifically capped with a DTB-GTP as described in the initial Cappable-seq protocol for prokaryotes. Pol II transcripts harbor a 7mGppp cap and therefore need to undergo a decapping step using the yDcpS enzyme, before being capped with a DTB-GTP cap. On their 3'end, Pol I and Pol II transcripts naturally harbor a polyA tail, but pol III transcripts are often omitted from transcriptome analysis as they do not have a polyA tail. Developing "ONT-ReCappable-seq", a hybrid version that combines the long-read ONT-Cappable-seq strategy with Recappable-seq would permit the capture of the complete and full-length transcriptome of eukaryotes.

C. Development of Loop-Cappable-seq

1. Introduction

Sharp improvements have been made thanks to the emergence of long-read sequencing technologies, such as PacBio and Oxford Nanopore. Despites these improvements, long-read sequencing technologies remain error-prone and the lack of accuracy can limit their use. The predominant errors in both PacBio and Nanopore sequencing technologies are insertions and deletions (indels). These errors can drastically confuse mapping algorithms and by introducing frameshifts and premature stop codons, critically affect the prediction of open reading frames directly from transcripts (Watson and Warr, 2019). The most common approach to overcome the high error rate limitation is to align the reads against a reference genome. Nonetheless, when no high-quality reference genomes are available (which is the case in most microbiome research) long-reads technologies are of limited use (Sahlin and Medvedev, 2021). Overall, such errors limit the scope of long-read technologies for complex community analysis. This motivated the development of several computational approaches to correct and reduce the number of errors in long-reads data. Two main strategies exist: (1) the hybrid-correction approach that uses short-read illumina data to correct long-reads and (2) the non-hybrid (self-correction) approach in which long-reads are self-corrected using the overlaps in high-coverage data (Magi *et al.*, 2018).

In this part, we present another version of Cappable seq. For this project, we collaborated with Loop Genomics (https://www.loopgenomics.com/) to adapt Cappable-seq to their LoopSeq platform, a new long-read sequencing technology based on Illumina sequencing. Because this technology is based on the Illumina platform, it provides affordability and sequencing accuracy. In addition, LoopSeq combined with Cappable-seq offers new possibilities, such as directly calling the ORF from the reads in complex microbial communities containing unknown species, as well as the ability to differentiate similar species between each other, reducing the multiple-mapping problem that is often faced in microbiome studies. Also, applications in microbiomes are particularly appealing since Loop-Cappable-seq would theoretically be able to uncover partial or complete metabolic pathways by phasing functionally related genes on the same sequencing reads. All the results that will be

presented in this following part are the result of the collaboration with Loop Genomics. In terms of organization, Bo Yan and I at NEB were in charge of the experiments and analysis, while Loop Genomics performed the LoopSeq sequencing and provided us the raw data for analysis.

First, we developed Loop-Cappable-seq on *E. coli* and evaluated the ability of different long-reads sequencing platforms to directly predict ORFs from the raw reads compared to mapped reads. In a second time, we created a synthetic mixed community composed of different *E. coli* subspecies and one *Bacillus* to demonstrate the ability of Loop-Cappable-seq to provide an accurate representation of the transcriptome of mixed communities, with the ability to distinguish species between each other even, at the subspecies level. However, this second part is still an ongoing project as the Covid-19 pandemic delayed the project. The experiments are currently ongoing but as a consequence, no results are available to be shown yet for this second part. A joint manuscript with Loop Genomics is in preparation with the aim to publish Loop-Cappable-seq.

2. Material and Methods

Loop-Cappable-seq library preparation

The protocol is very similar to the ONT-Cappable-seq presented in the previous part, except minor changes concerning the enrichment steps. Indeed, optimization has been done by Bo Yan reducing the number of enrichments rounds to a single one to remove most of the uncapped transcripts. The advantage of a single round of enrichment is that less material is lost and the RNA does not need to be eluted from the beads with biotin. The second strand cDNA synthesis can be directly done on the streptavidin beads, shortening the time for library preparation. Knowing from experience that the Cappable-seq protocol works well on *E. coli*, we decided to perform only one round of enrichment for the Loop-Cappable-seq library preparation.

Capping of full-length prokaryotic transcripts

The capping reaction was done using 4µg of *E. coli* total RNA grown in M9 medium. The RNA was incubated in the presence of 0.5mM DTB-GTP (N0761, New England Biolabs) and 100 units of Vaccinia Capping Enzyme (M2080, New England Biolabs) and 0.25 units of *E. coli* pyrophosphatase (M0361, New England Biolabs) for 1h at 37°C in 40µL reaction volume. The capped RNA was purified using the Zymo Clean and Concentrator 5G columns (R1013, Zymo) and eluted in 23µl of low TE buffer.

Capture of the 3'end of full-length prokaryotic transcripts

A polyA tail was added *in vitro* by incubating the capped RNA in 30µl reaction volume with 5 units of *E. coli* Poly(A) Polymerase (M0276, New England Biolabs) and 1 mM ATP for 15min at 37 °C. The capped and tailed RNA was purified using the Zymo Clean and Concentrator 5G column kit and eluted in 33µl of low TE buffer. A volume of 3µl of the reaction (non-enriched RNA) was put aside and used as control.

First strand cDNA synthesis

A volume of 30µL of RNA was used in a 40µl reaction for reverse transcription. First, the RNA was incubated at 65°C for 2min with 5µM of custom polydT oligo adapted for Loop sequencing containing a unique sample index (index loop RT primer), 1mM dNTP and then cooled down on ice. Next, 400 units of ProtoScript II Reverse Transcriptase (M0368, New England Biolabs), 1µL murine RNAse inhibitor (M0314, New England Biolabs) and 1µL of Actinomycin D (stock at 200ng/µL) were added to the reaction and incubated at 42 °C for 1h. After the first strand cDNA synthesis, 50 units RNase If (M0243, New England Biolabs) was added and incubated at 37 °C for another 30 min to remove single stranded RNA. The reactions were purified using 1.0X AMPure beads and eluted in 23µL Low TE.

A polyG was added to the 3’ end of cDNA for second-strand synthesis using TdT Terminal transferase (M0315, New England Biolabs). The purified cDNA/RNA duplex samples were incubated with 10 units of TdT enzyme and 3mM dGTP at 37 °C for 30 min.

Enrichment of full-length prokaryotic transcripts

The cDNA/RNA duplex was enriched for a unique round using hydrophilic streptavidin magnetic beads (S1421, New England Biolabs). A volume of 25µl of beads were prepared by washing 3 times with a Washing Buffer before being resuspended in 30µl of Binding Buffer (the difference lies in the NaCl concentrations, see below for composition). A volume of 30µl of cDNA/RNA from the previous step was incubated with 30µL of prepared streptavidin beads at room temperature for 30min on a hula mixer rotator. The beads were then washed thoroughly 3 times with 60µL of Binding buffer and then 3 times with 60µL of Wash Buffer. The beads were then resuspended in 15µL of Elution buffer (low-TE).

Binding Buffer: 10mM Tris-HCl pH7.5; 1mM EDTA; **2M NaCl**
Wash Buffer: 10mM Tris-HCl pH7.5; 1mM EDTA; **1mM NaCl**
Elution Buffer: 10mM Tris-HCl pH7.5; 0.1mM EDTA; **250mM NaCl**

Second strand synthesis and PCR

Second-strand cDNA synthesis was performed on first strand cDNA/RNA duplex samples in 50µL reaction volume, using the LongAmp polymerase (M0533, New England Biolabs) a custom primer adapted for Loop sequencing (loop_oligodC15 primer) containing Cs to prime to the added Gs and RNAse H (M0297, New England Biolabs). The reaction was first incubated at 37°C for 15min to allow the removal of residual RNA strands left after reverse transcription, then for 1min at 94°C, next 4 cycles of 94°C for 1min and 65°C for 15min, with a final extension time of 10min at 65°C. The tube was placed on a magnetic rack to keep the supernatant containing the double stranded cDNA. Finally, the reactions were purified using 1.0X AMPure beads and eluted in 40µL Low TE and concentrated to 20µL using a Speed Vac for 15 min. We then shipped the cDNA to Loop Genomics for sequencing.

LoopSeq Sequencing (Loop Genomics)

The LoopSeq platform generates synthetic long reads (SLRs). Briefly, a proprietary enzymatic barcoding technology distributes an intramolecular barcode, unique to each molecule. Thus, each short read contains the same barcode indicating which original molecule it came from. After barcoding, amplification and sequencing on an Illumina platform, short reads that share the same

barcode are *de novo* assembled together into a long-read sequence to reconstruct the full-length molecule (Callahan *et al.*, 2021; Liu *et al.*, 2021). The principle of LoopSeq is presented in **Figure 23** below.

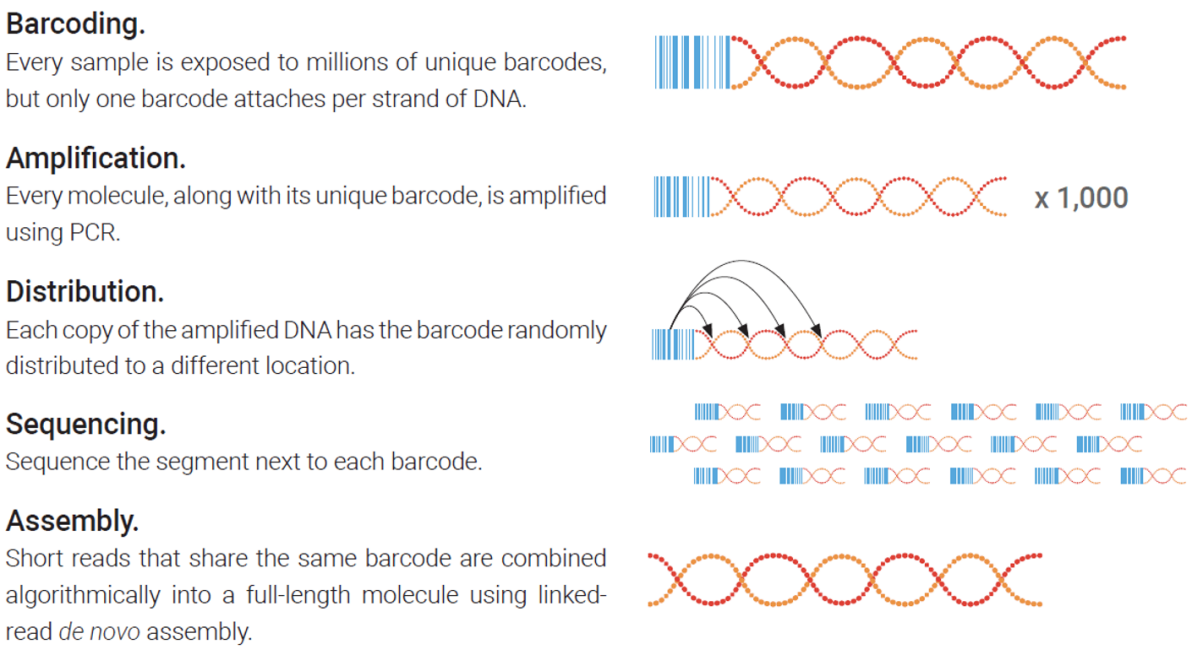


Figure 23: General principle of the LoopSeq Synthetic Long Reads (SLRs) (LoopGenomics — Overview, 2020).

Data analysis

Mapping of Loop-Cappable-seq data

Assembled and trimmed fastq files were provided by Loop Genomics. The reads were mapped to the *Escherichia coli* str. K-12 substr. MG1655 (NC_00913.3) reference genome using Minimap2 version 2.10-r761 (Li, 2018) using the parameters -ax asm20.

Canu correction

Two rounds of correction were performed for the ONT-Cappables-seq (nanopore) and SMRT-Cappable-seq (PacBio) datasets with Canu version 1.9 (Koren *et al.*, 2017)canu -correct option. The command lines are listed below.

ONT-Cappable-seq

canu -correct genomeSize=5M useGrid=false corOutCoverage=all minReadLength=100 minOverlapLength=100 -maxThreads=8 -maxMemory=8g corOverlapper=minimap -nanopore-raw $fastq stopOnLowCoverage=1

SMRT-Cappable-seq

canu -correct 1 genomeSize=5M useGrid=false corOutCoverage=all minReadLength=100 minOverlapLength=100 -maxThreads=8 -maxMemory=8g corOverlapper=minimap -pacbio-raw $fastq stopOnLowCoverage=1

The corrected fasta files were then downsampled to 20,000 reads. The ONT-Cappable-seq data were mapped to the *Escherichia coli* str. K-12 substr. MG1655 (NC_00913.3) reference genome using Minimap2 version 2.10-r761 with the parameters: minimap2 -ax map-ont --MD $refseq $fasta. The SMRT-Cappable-seq data were mapped to the same reference genome, using Minimap2 version 2.10-r761 with the parameters: minimap2 -ax map-pb --MD $refseq $fasta.

LoRDEC correction

LoRDEC version 0.6 (Salmela and Rivals, 2014)was used to perform a hybrid assembly with Illumina RNA-seq data downloaded from the European Nucleotide Archive SRR3132588. The following commands were used for both ONT-Cappable-seq and SMRT-Cappable-seq: lordec-correct -T 16 -k 19 -s 2 -i $longreads -2 $shortreads -o output.fasta

The corrected fasta files were then downsampled to 20,000 reads. The ONT-Cappable-seq data were mapped to the *Escherichia coli* str. K-12 substr. MG1655 (NC_00913.3) reference genome using Minimap2 version 2.10-r761 with the parameters: minimap2 -ax map-ont --MD $refseq $fasta. The SMRT-Cappable-seq data were mapped to the same reference genome, using Minimap2 version 2.10-r761 with the parameters: minimap2 -ax map-pb --MD $refseq $fasta.

ORF prediction and indels analysis

The ORFs length of each dataset was predicted from the bam files using a custom python script that uses Biopython (Cock *et al.*, 2009). The indels were counted from the cigar of the bam file, using pysam(https://github.com/pysam-developers/pysam).

3. Results

a. Loop-Cappable-seq development on *E. coli*

We evaluated the ability of each technology to predict ORFs directly from raw reads. First, SMRT-Cappable-seq (PacBio) and ONT-Cappable-seq (Nanopore) data were corrected either using a self-correction approach with Canu (Koren *et al.*, 2017) or a hybrid correction approach using illumina short reads with LoRDEC (Salmela and Rivals, 2014). Reads were then either:

(1) mapped to the reference genome, then the corresponding genomic sequence was extracted and the ORFs were predicted for the different datasets. Because the *E. coli* genome is of high quality, the sequences extracted from it are accurate. This corresponds to the 'mapped reads' represented in blue in **Figure 24** below.

(2) ORFs were directly predicted on the raw reads. This corresponds to the 'raw reads' represented in yellow in **Figure 24** below.

Then, the ratio of the predicted ORF length relative to the read length was calculated for each dataset. A ratio close to 1 represents an ORF that has been predicted along the whole read length, meaning the ORF is complete. The higher the ratio, the most accurate the prediction is. Even after correction, Nanopore data (ONT) is still error-prone, with the lowest ratio of all the datasets. PacBio data perform better, especially when corrected *in silico* but the low throughput can be a limited factor when applied to complex communities for which a high coverage is needed. LoopSeq performed the best and was able to perfectly call the ORF on *E. coli* unmapped data, without any data correction needed. Then, we looked at the indels for each dataset and calculated the ratio of the number of indels normalized to the read length (**Figure 25**). Again, even after correction, the number of indels per read is still very high in Nanopore data. Without any correction, PacBio shows a high number of indels, but correction can greatly improve the data.

Still, LoopSeq showed the lowest number of indels. Overall, Loop-Cappable-seq shows the best results. Because it is based on Illumina sequencing, very low indels are present in the data compared to other sequencing platforms. Therefore, Loop-Cappable-seq perfectly calls ORF on *E. coli* unmapped data and could therefore be used to annotate transcripts directly from the raw reads.

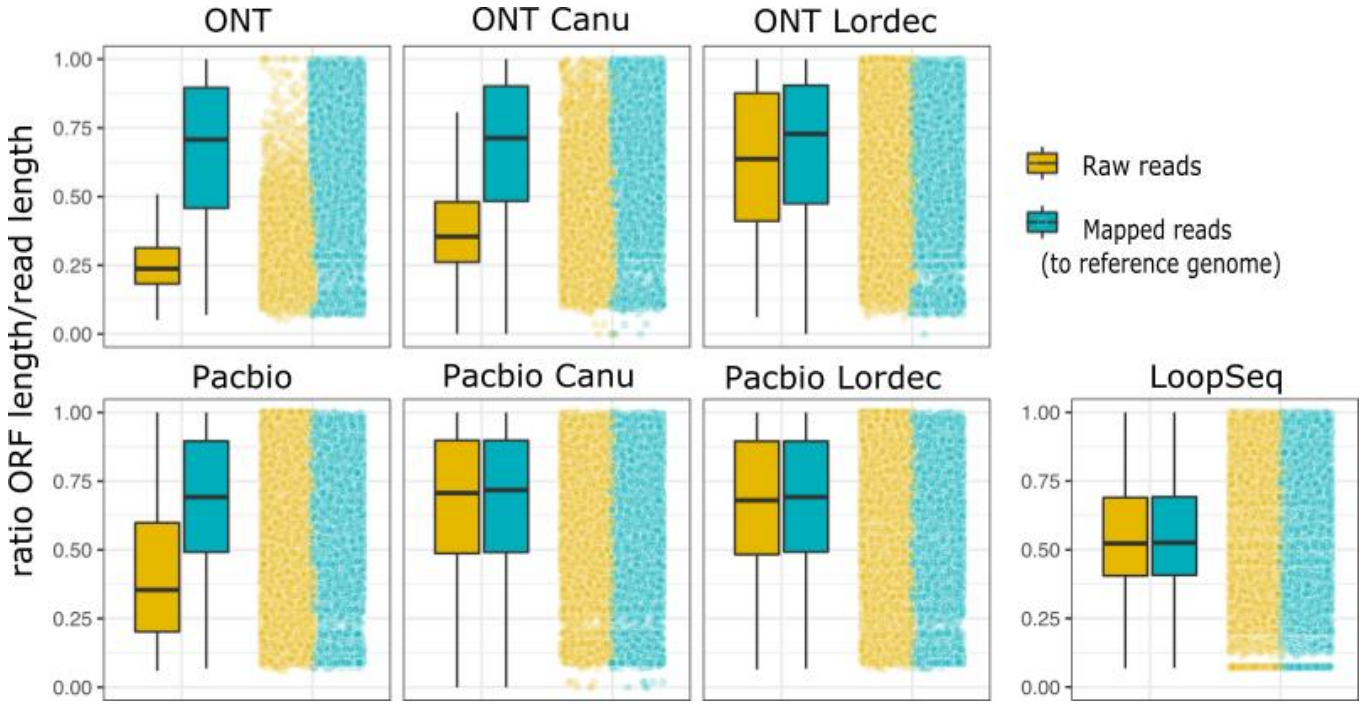


Figure 24: ORF prediction performed on raw reads directly (yellow) or after mapping the reads to the reference genome and extracting the correct sequence from the reference genome (blue), for the different datasets. ONT: ONT-Cappable-seq, Pacbio: SMRT-Cappable-seq, LoopSeq: Loop-Cappable-seq.

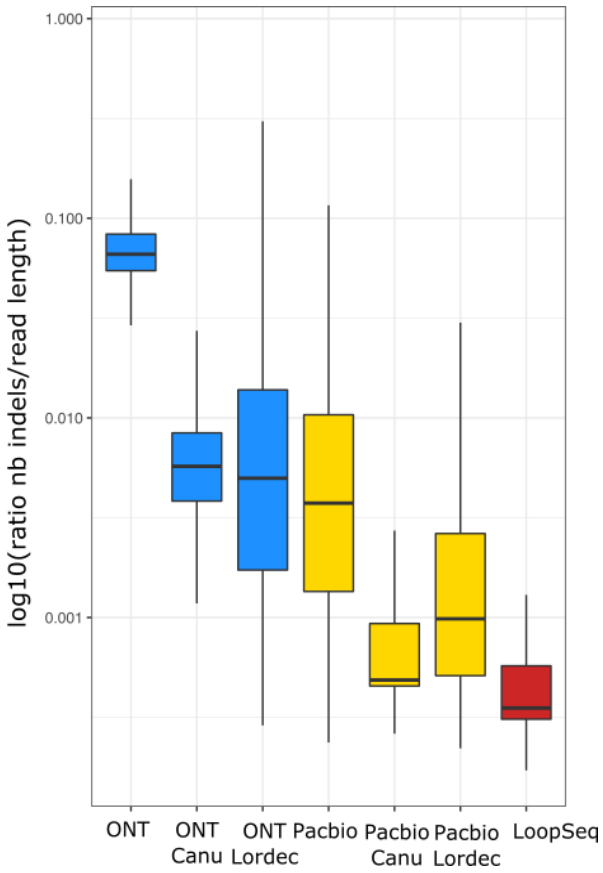


Figure 25: Insertions and deletions (indels) ratio calculated for each dataset. This number was normalized to the read length. ONT: ONT-Cappable-seq, PacBio: SMRT-Cappable-seq, LoopSeq: Loop-Cappable-seq.

b. Loop-Cappable-seq development applied to a mixed synthetic community

As mentioned in the introduction, unfortunately no results are available to be discussed in the thesis yet, but the experiments are ongoing. Still, the experiment's aim and design will be discussed in this part.

The goal here is to demonstrate the ability of Loop-Cappable-seq to provide an accurate transcriptome profiling of complex bacterial communities. Thanks to the accuracy provided by the synthetic long-reads of LoopSeq, combined to full operon structure obtained with Cappable-seq, this technique could be a new powerful tool for metatranscriptome analysis.

With this aim in mind, we designed a synthetic community composed of closely related *E. coli* strains and one *Bacillus* strain, more distant from *E. coli* in terms of genomic sequence. To decide which strains of *E. coli* would be appropriate, we considered the availability of the strain in the lab as well as strains with different degrees of similarity. The Average Nucleotide Identity (ANI) was calculated using an online calculator (Yoon *et al.*, 2017) and strains with different degrees of genomic similarity (99% ANI and 98% ANI) were picked. We added a *Bacillus* strain to add diversity to the synthetic community. **Table 5** shows the composition of the synthetic community and **Table 6** shows the Matrix of the ANI%.

Synthetic community	Reference	% ANI to *E. coli* MG1655	Gram	%GC
E. coli MG1655 (ER1506 at NEB)	ATCC 700926	/	-	51
E. coli BL21(DE3)	NEB C2527	99	-	51
E. coli ATCC 700728	ATCC 700728	98	-	51
Bacillus fusiformis 1226	NEB 1441	64	+	38

Table 5: Composition and properties of the synthetic community.

Matrix ANI %	*Bacillus fusiformis* 1226	*E. coli* BL21 (DE3)	*E. coli* MG1655 (ER1506 at NEB)	*E. coli* ATCC 700728
E. coli ATCC 700728	81	98	98	100
E. coli MG1655 (ER1506 at NEB)	81	99	100	98
E. coli BL21 (DE3)	81	100	99	98
Bacillus fusiformis 1226	100	64	64	65

Table 6: Matrix of the Percentage of Average Nucleotide Identity (ANI) calculated for all the strains in the synthetic community, compared two by two. ANI % was calculated using the 'ChunLab's online Average Nucleotide Identity (ANI) calculator'' (Yoon et al., 2017).

The bacteria were first grown individually in LB medium, except for *E. coli* K12 MG1655 that was grown in Rich medium. As the *E. coli* strains are closely related, the transcripts expressed are likely to be very similar in terms of sequence. So, in order to determine if Loop-Cappable-seq is able to detect, differentiate but also quantify accurately the transcripts expressed, we grow one of the strains in a different medium. The bacteria were grown individually up to OD=0.6, total RNA was extracted using the RNeasy Mini kit (Qiagen) and had a RIN above 9.5. The RNAs were then mixed in a 1:1 ratio

except for *E. coli* ATCC 700728, for which the RNA ratio was 1:10 in order to reflect the abundance diversity of complex bacterial communities and test Loop-Cappable-seq ability to quantify transcriptomes. 10µg of total RNA from *E. coli* MG1655, *E. coli* BL21, *Bacillus fusiformis* were mixed with 1µg of *E. coli* ATCC 700728.

RNA-seq and Loop-Cappable-seq are performed on the synthetic community in order to compare their ability to give an accurate transcriptome profiling. We expect that Loop-Cappable-seq data will allow a better read assignment than with short-read data, reducing the multiple-mapping issue. In addition, RNA-seq was performed on each individual strain, which will be used to assess the performance of RNA-seq and Loop-Cappable-seq on a metatranscriptome.

4. Discussion and further outlooks

Here, we adapted Cappable-seq to a new sequencing platform, LoopSeq, and developed Loop-Cappable-seq, a method that has the potential to be used to annotate transcript without the need for a reference genome. We evaluated the ability of different long-reads Cappable-seq versions (PacBio, Nanopore and LoopSeq platforms) to predict ORF directly from raw reads. Overall, PacBio (SMRT-Cappable-seq) and Nanopore (ONT-Cappable-seq) data are too error-prone to be used for direct ORF calling. Correction programs such as Canu and Lordec help correct the data and reduce the number of indels, but such programs also have major 'side-effects'. First, we should keep in mind these programs have been initially designed for DNA-seq data correction and might not be optimal to correct transcriptomic data. As an example, self-correction tools contain a step to generate consensus sequences using overlapping reads, which implies the need for a high sequencing coverage to get an effective error correction. In the case of complex communities, where similar species and similar transcripts are present, these are likely to be considered as a single transcript, removing information from the data (Lima *et al.*, 2020). Also, correction programs tend to produce shorter reads and reduce depths because they either discard uncorrected reads or trim the uncorrected regions. Such behaviors provoke data loss and may influence downstream analysis because information is lost if the reads are shortened, wrongly merged with others or even discarded (Zhang, Jain and Aluru, 2020).

On the other hand, Loop-Cappable-seq showed very good data quality, with the lowest indel rate and performed the best when predicting the ORF directly from raw reads. In contrast to other long-reads platforms, no error correction was needed to achieve a prediction that resembles the one we would have got with a perfect read. In other words, Loop-Cappable-seq has the potential to eliminate the need for a reference genome as entire genes and operons can be identified on a single raw read, allowing to predict gene's function and annotate unknown genes based on their neighboring genes on the operon. Also, because this method provides high quality data (accurate), it would reduce the mapping ambiguity when similar species are present (multiple-mapping problem). Reads would be assigned with a higher confidence to their correct genome. Applying the method to microbiome studies would then be particularly appealing.

The next step is to test whether the accurate long-reads obtained from Loop-Cappable-seq allow to distinguish very similar transcripts originating from several *E. coli* subspecies. We are currently performing the experiments and hope to analyze the data soon. The ultimate step to validate this method would be to apply it to a real microbiome sample and predict ORFs and PFAM domains from the data directly. This kind of analysis is very promising as it would give insights into the functionality of a microbiome even if the reference genomes are not available.

D. Conclusion

In this chapter, we presented different versions of Cappable-seq and developed two new versions based on long-reads sequencing: ONT-Cappable-seq and Loop-Cappable-seq. A variety of Cappable-seq 'flavors' are now available according to the sequencing platform used and to the application (summarized in **Table 7** below). The benefits and disadvantages of each version mainly depend on the sequencing platform and should be considered carefully depending on the application needed as this study showed the importance of choosing the right sequencing platform in order to obtain an accurate transcriptome.

	Cappable-seq	SMRT-cappable-seq	ONT-cappable-seq	Loop-cappable-seq	ReCappable-seq
Sequencing platform	Illumina	PacBio	Nanopore	LoopSeq	Illumina
Technology	Short-reads	Long-reads	Long-reads	Long-reads	Short-reads
Organism	Prokaryotes	Prokaryotes	Prokaryotes	Prokaryotes	Eukaryotes
Full operon structure	No (TSS only)	Yes	Yes	Yes	/
Benefits	Accuracy+ Throughput	Accuracy	Throughput	Accuracy+	Accuracy+

Table 7: Summary of the different Cappable-seq flavors presented in this thesis. The '+' sign represents a higher level of accuracy, as these technologies are based on the illumina platform.

Chapter III : Connecting transcriptional responses to compositional changes in a synthetic gut microbiome following antibiotic treatment

A. Introduction

The human microbiome is an exciting and rapidly expanding field of research that has gained significant interest over the past decade. Numerous studies have shown the biological relevance of the microbiome, especially the gut microbiome, for human health (Fan and Pedersen, 2021). The gut microbiome strongly influences host physiology, assisting in the bioconversion of nutrients and detoxification, supporting immunity, and protecting against pathogens. Perturbations to the composition of the microbiome (called dysbiosis) have been linked to the initiation and progression of numerous diseases and inflammatory disorders (Carding *et al.*, 2015; Scotti *et al.*, 2017). This fragile equilibrium between bacteria can be impaired by many factors, including antibiotics, which alter the bacterial population composition, enhance the spread of resistant strains, and may degrade the protective effect of microbiota against invasion by pathogens. Changes in the taxonomic composition of the gut microbiome are usually observed several days after the onset of treatment. However, it has been shown that bacteria can rapidly acclimate to environmental perturbations by transcriptional reprogramming (Sangurdekar, Srienc and Khodursky, 2006).

Here, we explore how the rapid transcriptomic response to antibiotics correlates, and potentially predicts, the later changes to microbiome structure. In this study, we examine the short and long-term responses of a phylogenetically diverse, defined community of gut bacteria to the widely used broad-spectrum antibiotic, ciprofloxacin. Following addition of ciprofloxacin to log phase cultures, samples were taken over a time course ranging from 5 minutes to 48 hours. We used a multiomic approach, using some of the methods developed and presented previously, in order to analyze transcriptional responses and community composition changes relative to minus-ciprofloxacin

controls. We performed RNA-seq and Cappable-seq to study the functional response as well as 16S and RIMS-seq (shotgun sequencing) to study the community-wide composition changes.

Ciprofloxacin: overview of a widely used antibiotic

In this study, we chose to subject the defined community to the synthetic antibiotic ciprofloxacin. Ciprofloxacin is a broad-spectrum antibiotic of the fluoroquinolones class, available in oral and intravenous formulations. The first fluoroquinolones were introduced in the late 1980s and were the only orally administered agents available for the treatment of serious infections at the time (King, Malone and Lilley, 2000). Therefore, ciprofloxacin rapidly became one of the most widely used antibiotics in the world because of its efficacy and relatively low cost. It is used for the treatment of a wide-range of infections, such as urinary tract, respiratory tract, skin, bones, gastrointestinal tract and gynaecological infections (Campoli-Richards *et al.*, 1988). Yet, because of its widespread use, resistance to this drug has emerged and rendered it less effective (Conley *et al.*, 2018).

Ciprofloxacin acts by inhibiting bacterial replication. The drug targets enzymes essential in DNA replication, namely the DNA gyrase (a type II DNA topoisomerase) and DNA topoisomerase IV, coded by the *gyrA*, *gyrB* and *parC*, *parE* genes, respectively (Drlica and Zhao, 1997). Topoisomerases bind to the DNA, cleave either one or both strands of the double helix, pass either the other strand of the same helix or another double strand through the break, and finally reseals the DNA backbone. By binding reversibly to the complexes of DNA with the gyrase and topoisomerase IV, ciprofloxacin inhibits the enzymes function by blocking the resealing of the DNA double-strand break and preventing the movement of the DNA replication fork (Wentzell and Maxwell, 2000), leading to double-strand DNA breaks and bacterial death (Drlica *et al.*, 2008). ciprofloxacin differentially targets DNA gyrase and topoisomerase IV in bacteria, with greater activity against DNA gyrase in Gram negative bacteria and greater activity against topoisomerase IV in Gram-positive bacteria (Hooper and Jacoby, 2016).

The SOS response pathway

By inducing DNA strand breaks, ciprofloxacin triggers the bacterial SOS response, a regulatory network found in most bacterial species (Drlica *et al.*, 2008). Beyond being a DNA repair process, the SOS induction leads to a very strong response to genotoxic stress, which promotes bacterial survival and adaptation to changing environments (**Figure 26**).

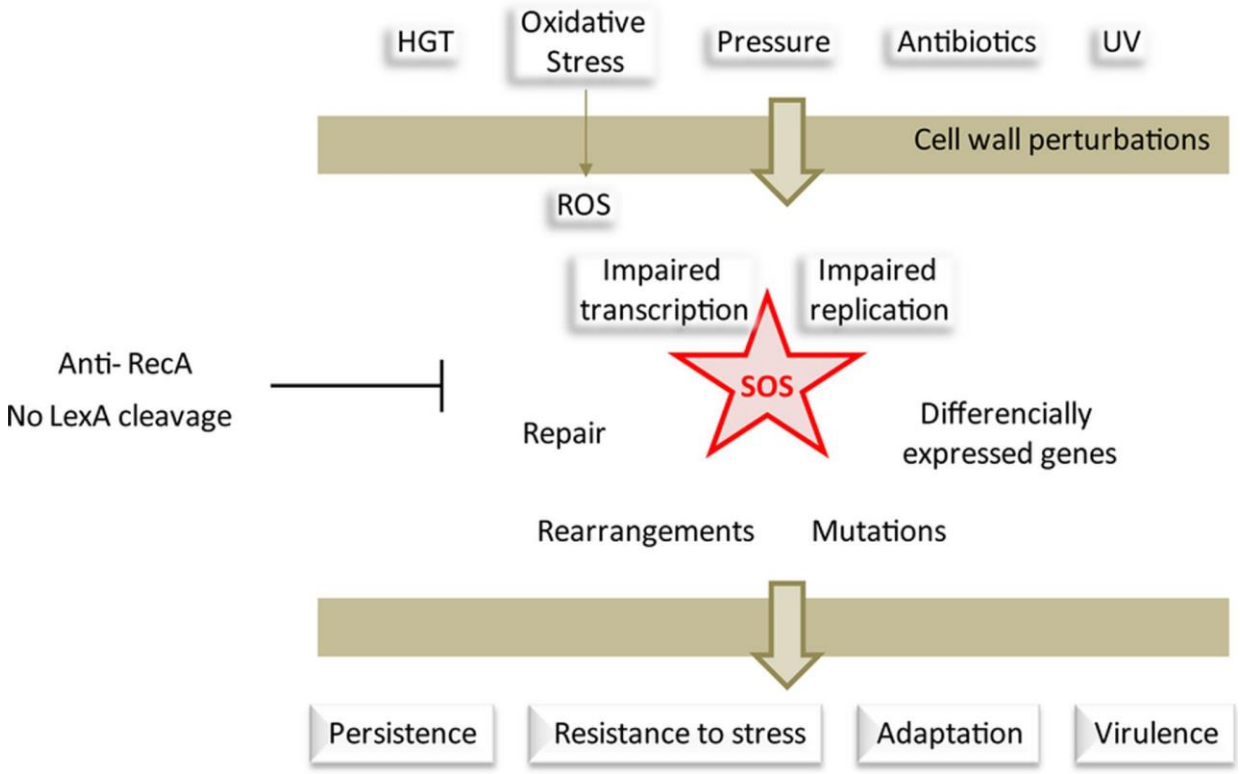


Figure 26: Perturbations leading to SOS response and mechanisms triggered by activation of the SOS response (Baharoglu and Mazel, 2014).

The SOS response is a well-characterized pathway and is controlled by 2 key regulators: RecA and LexA. Briefly, when a DNA damage causing double-strand DNA breaks is detected, RecA is recruited on the ssDNA and upon conformation changes, catalyzes the auto-cleavage of the LexA repressor, which in turns activates the transcription of the SOS genes. When in its dimer form, LexA represses the genes belonging to the SOS regulon by binding a LexA box sequence on the SOS genes promoter. LexA proteolysis thus leads to derepression of this regulon, comprising around 50 genes in *E. coli* (Courcelle *et al.*, 2001; Simmons *et al.*, 2008). The number and the type of genes found in the regulon vary among bacteria. For example, in *Bacillus subtilis*, the LexA regulon contains 33 genes

among which only eight are homologous to *E. coli* SOS genes (Au *et al.*, 2005). The SOS genes allow DNA repair and are also involved in DNA recombination, DNA replication and segregation of chromosomes during cell division (Cox, 1998). For example, the SOS gene *sulA* is induced to inhibit and delay cell division, leading to cell filamentation until DNA damage is fixed. Three main DNA repair pathways induced by the SOS response have been described in *E. coli* and other bacteria: homologous recombination (HR), nucleotide excision repair (NER), and translesion synthesis. In the homologous recombination (HR) pathway, RecA recruits other homologous recombination proteins such as RecBCD and RecFOR, which facilitate the repair of single-stranded lesions. In case of extensive and persistent damage, the DNA translesion synthesis pathway gets activated. This pathway involves several error-prone DNA polymerases: Pol II (*polB* gene), Pol IV (*dinB* gene) and Pol V (*umuC* and *umuD* genes), promoting mutations and genetic adaptation, including antibiotic resistance (Dallo and Weitao, 2010; Podlesek and Žgur Bertok, 2020). Thus, the induction of the SOS response by some antibiotics such as ciprofloxacin, can promote the emergence of antibiotic resistance.

Resistance to ciprofloxacin

Fluoroquinolone resistance has been attributed to point mutations in the bacterial genes *gyrA* and *parC*, which code for the target enzymes DNA gyrase and topoisomerase IV, respectively. The loci of these point mutations have been termed "quinolone resistance-determining region" (QRDR). Single target mutations produce eight- to 16-fold increases in fluoroquinolone resistance, while accumulating mutations in both target enzymes has been shown to cause increasing quinolone resistance (Hooper and Jacoby, 2016). In many species, high-level quinolone resistance is generally associated with mutations in both gyrase and topoisomerase IV (Schmitz *et al.*, 1998).

Other systems can also contribute to resistance. Indeed, quinolones must cross the bacterial envelope to interact with the gyrase and topoisomerase IV targets. Active quinolone efflux pumps can decrease cytoplasmic antibiotic concentrations and confer resistance. In Gram-positive bacteria, reduced diffusion across the cytoplasmic membrane has not been found to cause resistance, but active efflux transporters have been shown to cause resistance. In contrast, in Gram-negative bacteria, outer membrane porin diffusion channels reduce the diffusion of the antibiotic and can contribute to resistance. Other mutations can also occur in the genes that control the expression of outer

membrane proteins and efflux pumps (Hooper and Jacoby, 2016; Hamed *et al.*, 2018), enhancing antibiotic efflux and resistance.

Thus, treatment with ciprofloxacin is expected to trigger the SOS response in bacteria and resistance can occur by a diversity of mechanisms. In this study, we examine the short and long-term responses of a phylogenetically diverse, defined community of gut bacteria to the ciprofloxacin. Following addition of ciprofloxacin to log phase cultures, samples were taken over a time course between 5 minutes and 48 hours. We used a multiomic approach in order to analyze transcriptional responses and community composition changes relative to minus-ciprofloxacin controls. We performed RNA-seq and Cappable-seq to study the functional response as well as 16S and RIMS-seq (shotgun sequencing) to study the composition changes community-wide. We investigated several questions: (1) can we identify an immediate transcriptional reprogramming in a complex community? (2) are bacteria from the same family responding the same way? Is there a phylum-specific response? (3) is there a specific response of the bacteria that will resist the treatment vs the susceptible ones? (4) And ultimately, can we identify some transcriptional markers (specific genes or pathways differentially expressed) that could be used to predict the outcome of the treatment?

B. Material and Methods

This project was performed using a defined synthetic community (called DefCom) composed of 51 bacteria representatives of a human gut. The RNA-seq analysis was done in collaboration with the Labgem team from the Genoscope (David Vallenet, David Roche and Stéphanie Fouteau), using the MicroScope platform (Vallenet *et al.*, 2020).

Figure 27 below presents an overview of the experiment. Briefly, the DefCom community was grown in 3 biological replicates up to mid-log phase before being split into two cultures (6 cultures total) : one 'control culture' without ciprofloxacin and one 'treated culture' with 10µg/mL of ciprofloxacin added. The cultures were grown in parallel and cell pellets were collected at different time points over 48h. DNA and RNA were extracted from the collected cell pellets and used for subsequent library preparation.

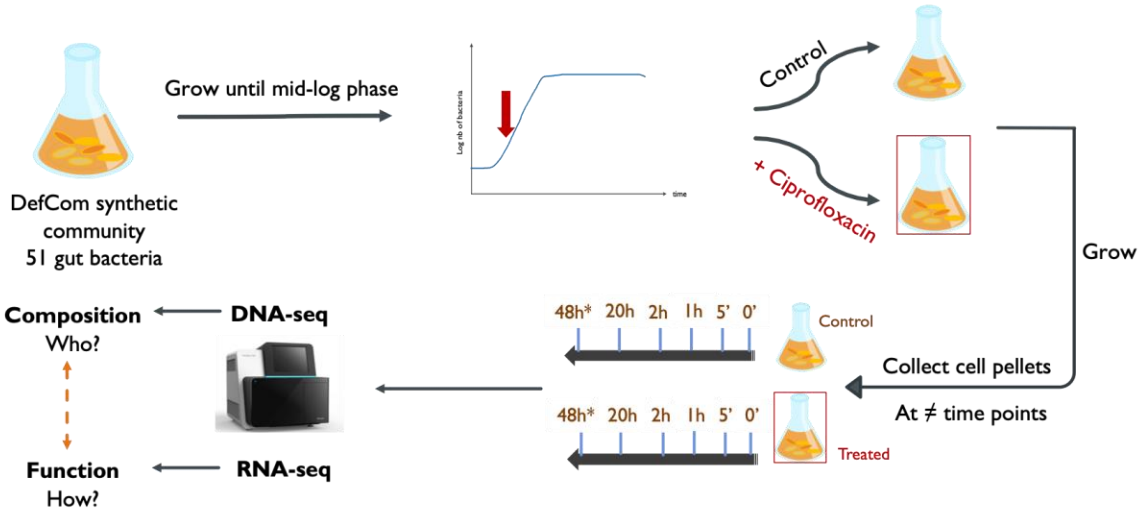


*Figure 27: Scheme presenting the design of the time course experiment on the DefCom community subjected to a ciprofloxacin treatment. Cultures were done in triplicates. *: cell pellets collected for DNA extraction only were sampled at 48h. The sampling for RNA was done until 20h.*

Description of the DefCom synthetic community

The DefCom synthetic community is composed of 51 bacteria encompassing the major phyla present in the human microbiome. This community is a consortium of dominant bacterial commensals that have been described in healthy human microbiomes (Hibberd *et al.*, 2017; Forster *et al.*, 2019) with BSL2 pathogens bacteria (see **Table 8** for a detailed list of the strains).

Bacteria	Phylum	NCBI taxid	Gram	BSL
Bifidobacterium adolescentis E298b (Variant c)	Actinobacteria	1680	+	1
Bifidobacterium angulatum DSM 20098	Actinobacteria	518635	+	1
Bifidobacterium catenulatum DSM 16992	Actinobacteria	566552	+	1
Bifidobacterium longum NCC2705	Actinobacteria	206672	+	1
Bifidobacterium longum sub. Infantalis ATCC 15697	Actinobacteria	391904	+	1
Bifidobacterium pseudocatenulatum DSM 20438	Actinobacteria	547043	+	1
Collinsella aerofaciens JCM 7790	Actinobacteria	74426	+	1
Akkermansia muciniphila ATCC BAA-835	Bacteroidetes	349741	-	1
Bacteroides caccae ATCC 43185	Bacteroidetes	411901	-	**2**
Bacteroides cellulosilyticus DSM 14838	Bacteroidetes	537012	-	1
Bacteroides cellulosilyticus WH2	Bacteroidetes	1268240	-	1
Bacteroides coprophilus DSM 18228	Bacteroidetes	547042	-	1
Bacteroides finegoldii CL09T03C10	Bacteroidetes	997888	-	1
Bacteroides thetaiotaomicron ATCC 29741	Bacteroidetes	818	-	**2**
Bacteroides thetaiotaomicron VPI-5482	Bacteroidetes	226186	-	**2**
Bacteroides uniformis ATCC 8492	Bacteroidetes	411479	-	**2**
Bacteroides dorei CL03T12C01	Bacteroidetes	997877	-	1
Bacteroides vulgatus ATCC 8482	Bacteroidetes	435590	-	**2**
Odoribacter splanchnicus DSM 20712	Bacteroidetes	709991	-	**2**
Parabacteroides distasonis ATCC 8503	Bacteroidetes	435591	-	**2**
Parabacteroides merdae CL09T00C40	Bacteroidetes	999421	-	1
Prevotella copri DSM 18205	Bacteroidetes	537011	-	1
Clostridium symbiosum WAL-14163	Firmicutes	742740	+	1
Anaerobutyricum hallii DSM 3353	Firmicutes	411469	+	1
Blautia coccoides YL58	Firmicutes	1532	+	1
Blautia hansenii DSM 20583	Firmicutes	537007	+	1
Blautia hydrogenotrophica ATCC BAA-2371	Firmicutes	53443	+	1
Blautia obeum ATCC 29174	Firmicutes	411459	+	1
Blautia producta ATCC 27340	Firmicutes	1121114	+	1
Clostridioides difficile 630	Firmicutes	272563	+	**2**
Enterocloster bolteae ATCC BAA-613	Firmicutes	411902	+	**2**
Clostridium celatum DSM 1785	Firmicutes	545697	+	1
Clostridium scindens ATCC 35704	Firmicutes	411468	+	1
Coprococcus catus VPI C6-61 [NCTC 11835]	Firmicutes	116085	+	1
Dorea formicigenerans ATCC 27755	Firmicutes	411461	+	1
Dorea longicatena DSM 13814	Firmicutes	411462	+	1
Enterococcus faecium ATCC 700221	Firmicutes	1352	+	**2**
Eubacterium eligens ATCC 27750	Firmicutes	515620	+	1
Eubacterium rectale ATCC 33656	Firmicutes	515619	+	1
Eubacterium ventriosum ATCC 27560	Firmicutes	411463	+	1
Faecalibacterium prausnitzii VPI C13-51	Firmicutes	853	+	1
Holdemanella biformis DSM 3989	Firmicutes	518637	+	1
Lachnospira multipara ATCC 19207	Firmicutes	1282887	+	1
Lactobacillus casei subsp casei ATCC 393	Firmicutes	1423732	+	1
Roseburia intestinalis DSM 14610	Firmicutes	166486	-	1
Ruminococcus gnavus AGR2154	Firmicutes	1384063	+	1
Tyzzerella nexilis DSM 1787	Firmicutes	500632	+	1
Escherichia coli ATCC BAA-97	Proteobacteria	562	-	1
Klebsiella pneumoniae ATCC 33259	Proteobacteria	72407	-	**2**
Salmonella enterica ATCC 27869	Proteobacteria	108619	-	**2**
Succinivibrio dextrinosolvens DSM 3072	Proteobacteria	1123324	-	1

Table 8: Composition of the DefCom synthetic community (51 bacteria). BSL: Biosafety Level classification. BSL2 bacteria are pathogenic.

Ciprofloxacin concentration tests

Preliminary growth tests were done before starting the experiment in order to determine a ciprofloxacin concentration that has a measured effect on the growth but does not completely stop the growth and wipe out the community. Based on the results shown on **Figure 28**, a concentration of 10µg/mL of ciprofloxacin was chosen.

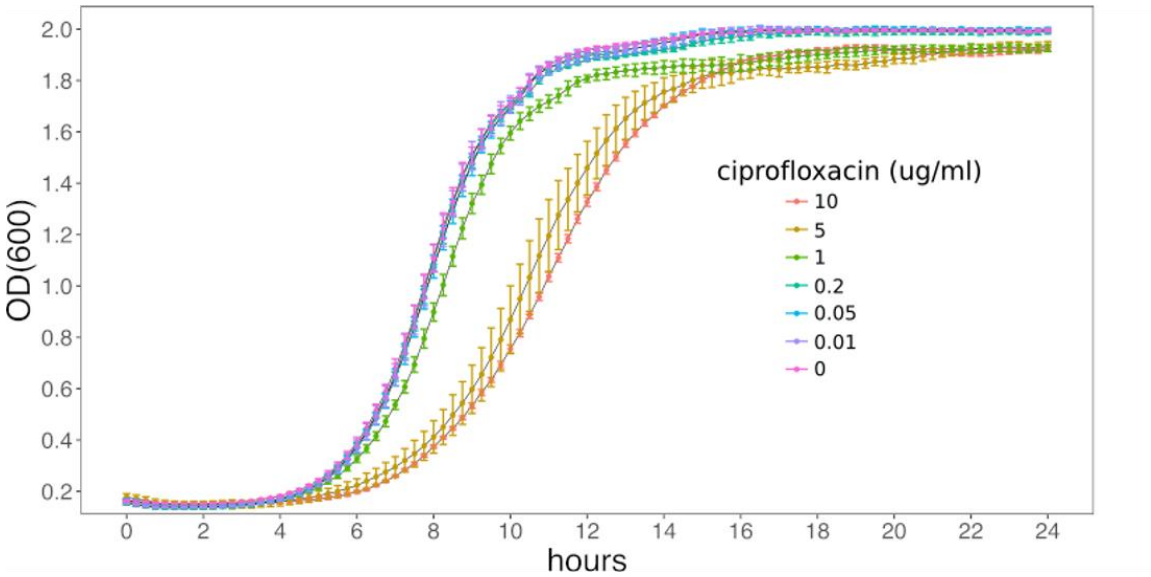


Figure 28: Growth curve of the DefCom community with different concentrations of ciprofloxacin added at the start of the culture. The OD600 was measured over 24h for each culture.

Culture of the DefCom community

All the following steps were done in a Coy anaerobic chamber (Coy Laboratory Products). The medium and consumables were stored in a Coy anaerobic chamber (5% H_2, 20% CO_2, and 75% N_2) at least 24 h prior to use. Glycerol stocks of the DefCom community were spun down and the pellet resuspended in 500 µL of PBS. For each biological triplicate, 3 glycerol stocks were combined together (1mL) and a total of 900 µL of DefCom community was inoculated in 100 mL of Mega Medium (see **Table 9** for the composition) and grown at 37°C for 48h. The cultures OD600 was monitored in parallel in a plate reader. After the cultures reached mi-log phase (after 5h), each triplicate was split in 2 bottles: one 'control culture' without ciprofloxacin and one 'treated culture' in which 10 µg/mL of ciprofloxacin was added (total of 6 cultures, 3 controls and 3 treated). The cultures were sampled at t=0min (before antibiotic addition), 5min, 20min, 1h, 2h, 20h and 48h. For each time

point, 5mL (for RNA extraction) and 1mL (for DNA extraction) of culture were sampled. The samples were centrifuged for 1 min at maximum speed, then the supernatant was discarded and the pellets freeze-thawed in a mix of dry-ice and ethanol before being stored at -80°C until nucleic acid extraction.

Component	Quantity/ L	Comments
Tryptone Peptone	10 g	
Yeast Extract	5 g	
D-glucose	2 g	
L-Cysteine HCl	0.5 g	
Potassium Phosphate Buffer	100 ml	1M stock solution, pH 7.2
Vitamin K_3 (menadione)	1 ml	1 mg/ml in 100% ethanol stock solution
$MgSO_4 \cdot 7\ H_2O$	0.02 g	
$NaHCO_3$	0.4 g	
NaCl	0.08 g	
$CaCl_2$	1 ml	0.8g/100 ml dH_20 stock solution
$FeSO_4 \cdot 7\ H_2O$	1 ml	40mg/100 ml dH_20 stock solution
Resazurin	4 ml	25mg resazurin/100 ml of dH_20 stock solution
Histidine Hematin	1 ml	1.2 mg hematin/ml in 0.2M histidine (pH 8.0) stock solution
Tween80	2ml	25% (vol/vol) dH_20 stock solution
Sodium Acetate	1 g	
Meat Extract	5 g	
ATCC Vitamin Mix	10 ml	
ATCC Trace Mineral Mix	10 ml	
Agar	15g	

Table 9: Composition of the Mega Medium, the composition is adapted from (Romano et al., 2015), except the medium contains 0.5% glucose.

DNA extraction

The DNA was extracted using the automated MagAttract PowerMicrobiome DNA kit (Qiagen). The DNA concentrations were quantified using PicoGreen on a plate reader.

RNA extraction

The RNA extraction method is critical when performing microbiome studies as a wide range of different bacteria are present (Gram+, Gram-). The lysis step is thus an important step that could represent a large source of bias depending on the protocol used. Studies have shown that the choice

of purification methods, and more importantly the choice of lysis procedures, has a large impact on the resulting microbiota composition and diversity (Yuan *et al.*, 2012; Knudsen *et al.*, 2016). Bead beating lysis has been shown to effectively lyse not only Gram- but also Gram+ bacteria that are harder to lyse due to their thick cell wall (Lim *et al.*, 2018).

We tested different kits for RNA extraction and picked the RNeasy mini kit (Qiagen) combined with a bead beating lysis on the FastPrep 120 disruptor (MP Biomedicals). An on-column DNAseI treatment was performed (Qiagen). The complete protocol can be found in **Appendix**. After extraction, the RNA quality was assessed using the Bioanalyzer RNA nano kit (Agilent) and the samples all had a RNA integrity number (RIN) above 9.0. The extracted RNA samples were quantified using the Qubit BR RNA (Thermofisher) kit and the Qubit HS DNA (Thermofisher) was used to assess the remaining DNA contaminants.

16S sequencing

16S libraries were prepared for all the samples (treated and control, all time points and triplicates) by amplifying the V4 region of the 16S rRNA gene using primers V4_515F and V4_806R (V4 primer 515F: GTGYCAGCMGCCGCGGTAA and V4 primer 806R: GGACTACHVGGGTWTCTAAT). 12.5ng of gDNA was used for the first round of PCR with NEBNext Q5 Ultra II Master mix (M0544, New England Biolabs), using 20 cycles of amplification. Amplicons were purified using 0.9X AMPure beads and eluted in 30µL of Low TE. 1µL of amplicon was used as input for a second PCR of 6 cycles, with the primers NEBNext single index NEBNext 96-well plate (E6609, New England Biolabs) containing p5, p7 and indexes. Amplicons were purified using 0.8X AMPure beads and eluted in 25µL of Low TE. All the libraries were evaluated on a TapeStation DNA1000 (Agilent) and paired-end sequenced on an Illumina MiSeq (2x250bp), using 10% of PhiX to add diversity to the libraries.

A total of 42 libraries were prepared and sequenced.

RIMS-seq

RIMS-seq libraries were prepared as described in the Material and methods of Chapter I, for treated and control samples, all the time points and in duplicates (B and C). 100ng of gDNA were used as

input for the RIMS-seq protocol. All the libraries were evaluated on a TapeStation High sensitivity DNA5000 (Agilent) and paired-end sequenced on Illumina (2x75bp).
A total of 28 libraries were prepared and sequenced.

RNA-seq

RNA-seq libraries were prepared using the NEBNext rRNA depletion kit for bacteria (E7850, New England Biolabs) and the NEBNExt Ultra II directional library prep kit for Illumina (E7760, New England Biolabs), following the manufacturer's protocol. 200ng of total RNA were used as input for the ribodepletion step. All the libraries were evaluated on a TapeStation High sensitivity DNA1000 (Agilent) and paired-end sequenced on Illumina (2x75bp).
A total of 24 libraries were prepared and sequenced.

Cappable-seq

The Cappable-seq libraries were prepared as described in the paper from Ettwiller *et al (Ettwiller et al., 2016)*. Libraries were made for the t=0 and t=5min, for the control and treated samples, in duplicates (B and C) and non-enriched controls were added (the RNA does not go through the Cappable-seq enrichment step, so the primary RNA are not enriched). Briefly, 5µg of total RNA was used as input for the capping reaction. 2 rounds of enrichment were used following the RNA fragmentation. After decapping with the RppH enzyme, the cDNA library was synthesized, amplified and prepared for Illumina sequencing using the NEBNext Small RNA Library Prep Set for Illumina (E7330, New England Biolabs). All the libraries were evaluated on a Bioanalyzer High sensitivity DNA (Agilent) and paired-end sequenced on Illumina (2x75bp).
A total of 16 libraries were prepared and sequenced.

All the different libraries performed on the DefCom community are indicated in the **Table 10** below.

Library type	Material	Starting material (ng)	Timepoints	Conditions	Replicates	total nb libraries
16S	gDNA	12.5ng	7	2	3	42
RIMS-seq (DNA-seq)	gDNA	100ng	7	2	2	28
RNA-seq	RNA	200ng	6	2	2	24
TSS cappable-seq	RNA	5000ng	2	2	2	16

Table 10: Summary of all the different libraries performed on the DefCom community. For the Cappable-seq library, a non-enriched control (no Cappable-seq enrichment) was performed for each sample (8x2 = 16 libraries).

Data analysis

Binning of highly similar genomes (with an ANI>95%)

As mentioned in Chapter 2, the correct attribution of multiple-mapping short reads from microbiome or complex synthetic communities containing highly similar species is a current challenge. Several sub-species are present in the DefCom community and in this case, it would be impossible to assign the corresponding reads with confidence to a particular genome. These reads would multiple-map on different genomes and bias the estimation of certain genes and organisms. Discarding these reads from the analysis would also underestimate the presence of genes or organisms. The Average Nucleotide Identity (ANI) measures the similarity between the coding regions (orthologous genes) of two genomes and can be used to define species boundaries (C. Jain *et al.*, 2018). Indeed, it has been suggested that species with an ANI > 95% can be considered as the same species (Konstantinos T. Konstantinidis, 2005; Goris *et al.*, 2007). So, in order to reduce the ambiguity in the mapping of sequence reads of highly similar genomes, we decided to 'bin' the genomes with an ANI > 95% among themselves and keep only one genome of reference. We performed a community-wide ANI analysis using the 'Genome clustering' option from the MicroScope platform (Vallenet *et al.*, 2020). **Figure 29** shows a tree representing the genomic similarity. Based on this genomic clustering, 8 species showed an ANI > 95% and could be considered as a single species (highlighted by a black box in the figure below). In order to pick the best reference genome for those species, statistics were calculated using seqkit stats option (Shen *et al.*, 2016). The genome length and completeness of the genome were taken into account for the decision. The species concerned by this binning and the

reference genome chosen are presented in the **Table 11** below. The 47 species community composition is presented in **Table 12**.

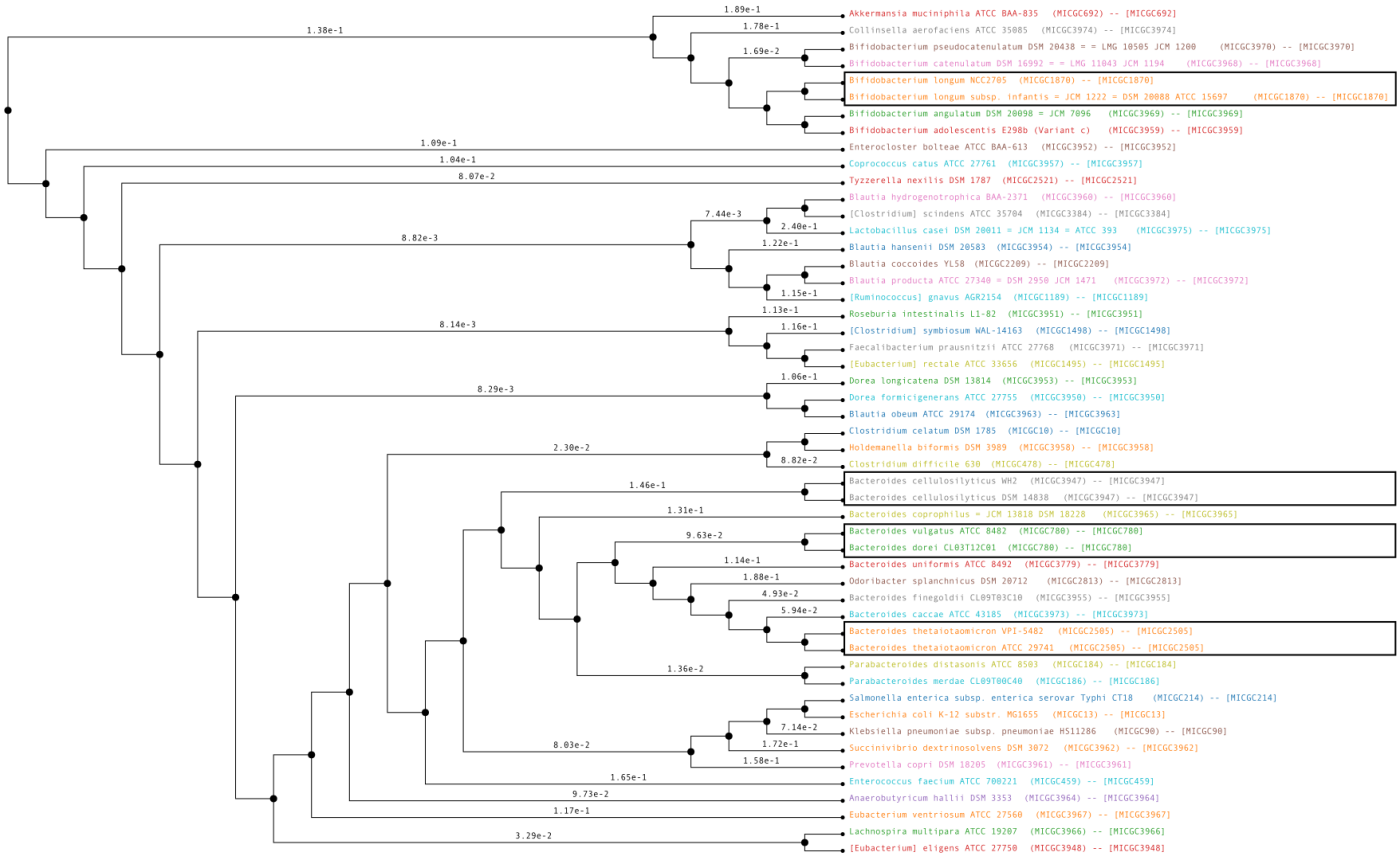


Figure 29: Tree representing the genomic clustering of the 51 genomes of the DefCom community. This clustering has been computed using the MicroScope platform and uses a 95% ANI that corresponds to the standard ANI used to define a species group. The species with an ANI > 95% are highlighted by a black box (8 species in total). MICGC: Microscope Genome Cluster.

Bacteria	taxid	MICGC number (Microscope)	Genome length (bp)	Number contigs	Keep as reference
Bifidobacterium longum NCC2705	206672	MICGC 1870	2,260,266	2	
Bifidobacterium longum sub. Infantalis ATCC 15697	391904	MICGC 1870	2,832,748	1	X
Bacteroides cellulosilyticus DSM 14838	537012	MICGC 3947	6,870,144	66	
Bacteroides cellulosilyticus WH2	1268240	MICGC 3947	7,084,828	1	X
Bacteroides vulgatus ATCC 8482	435590	MICGC 780	5,163,189	1	X (both are good)
Bacteroides dorei CL03T12C01	997877	MICGC 780	5,310,365	1	
Bacteroides thetaiotaomicron ATCC 29741	818	MICGC 2505	6,110,649	183	
Bacteroides thetaiotaomicron VPI-5482	226186	MICGC 2505	6,293,399	2	X

Table 11: Genomes with an ANI > 95 % and statistics calculated in border to keep one genome as reference for subsequent analysis.

Bacteria	Phylum	NCBI taxid	Gram	BSL
Bifidobacterium adolescentis E298b (Variant c)	Actinobacteria	1680	+	1
Bifidobacterium angulatum DSM 20098	Actinobacteria	518635	+	1
Bifidobacterium catenulatum DSM 16992	Actinobacteria	566552	+	1
Bifidobacterium longum sub. Infantalis ATCC 15697	Actinobacteria	391904	+	1
Bifidobacterium pseudocatenulatum DSM 20438	Actinobacteria	547043	+	1
Collinsella aerofaciens JCM 7790	Actinobacteria	74426	+	1
Akkermansia muciniphila ATCC BAA-835	Bacteroidetes	349741	-	1
Bacteroides cellulosilyticus WH2	Bacteroidetes	1268240	-	1
Bacteroides coprophilus DSM 18228	Bacteroidetes	547042	-	1
Bacteroides finegoldii CL09T03C10	Bacteroidetes	997888	-	1
Parabacteroides merdae CL09T00C40	Bacteroidetes	999421	-	1
Prevotella copri DSM 18205	Bacteroidetes	537011	-	1
Bacteroides caccae ATCC 43185	Bacteroidetes	411901	-	**2**
Bacteroides thetaiotaomicron VPI-5482	Bacteroidetes	226186	-	**2**
Bacteroides uniformis ATCC 8492	Bacteroidetes	411479	-	**2**
Bacteroides vulgatus ATCC 8482	Bacteroidetes	435590	-	**2**
Odoribacter splanchnicus DSM 20712	Bacteroidetes	709991	-	**2**
Parabacteroides distasonis ATCC 8503	Bacteroidetes	435591	-	**2**
Clostridium symbiosum WAL-14163	Firmicutes	742740	+	1
Anaerobutyricum hallii DSM 3353	Firmicutes	411469	+	1
Blautia coccoides YL58	Firmicutes	1532	+	1
Blautia hansenii DSM 20583	Firmicutes	537007	+	1
Blautia hydrogenotrophica ATCC BAA-2371	Firmicutes	53443	+	1
Blautia obeum ATCC 29174	Firmicutes	411459	+	1
Blautia producta ATCC 27340	Firmicutes	1121114	+	1
Clostridium celatum DSM 1785	Firmicutes	545697	+	1
Clostridium scindens ATCC 35704	Firmicutes	411468	+	1
Coprococcus catus VPI C6-61 [NCTC 11835]	Firmicutes	116085	+	1
Dorea formicigenerans ATCC 27755	Firmicutes	411461	+	1
Dorea longicatena DSM 13814	Firmicutes	411462	+	1
Eubacterium eligens ATCC 27750	Firmicutes	515620	+	1
Eubacterium rectale ATCC 33656	Firmicutes	515619	+	1
Eubacterium ventriosum ATCC 27560	Firmicutes	411463	+	1
Faecalibacterium prausnitzii VPI C13-51	Firmicutes	853	+	1
Holdemanella biformis DSM 3989	Firmicutes	518637	+	1
Lachnospira multipara ATCC 19207	Firmicutes	1282887	+	1
Lactobacillus casei subsp casei ATCC 393	Firmicutes	1423732	+	1
Roseburia intestinalis DSM 14610	Firmicutes	166486	-	1
Ruminococcus gnavus AGR2154	Firmicutes	1384063	+	1
Tyzzerella nexilis DSM 1787	Firmicutes	500632	+	1
Clostridioides difficile 630	Firmicutes	272563	+	**2**
Enterocloster bolteae ATCC BAA-613	Firmicutes	411902	+	**2**
Enterococcus faecium ATCC 700221	Firmicutes	1352	+	**2**
Escherichia coli ATCC BAA-97	Proteobacteria	562	-	1
Succinivibrio dextrinosolvens DSM 3072	Proteobacteria	1123324	-	1
Klebsiella pneumoniae ATCC 33259	Proteobacteria	72407	-	**2**
Salmonella enterica ATCC 27869	Proteobacteria	108619	-	**2**

Table 12: Composition of the DefCom synthetic community after binning the highly similar species with ANI > 95% (47 bacteria). BSL: Biosafety Level classification. BSL2 bacteria are pathogens.

16S data analysis

The analysis of the 16S sequences was performed using the QIIME2 package (Bolyen *et al.*, 2019) that contains several command lines. After the data were imported into a qiime2 compatible format using the command 'qiime tools import', the 16S primers and the illumina adapters were trimmed from the reads using the 'qiime cutadapt trim-paired' command. The DADA2 (Callahan *et al.*, 2016) option was used as an alternative to OTU (Operational Taxonomic Unit) clustering. 'qiime dada2 denoise-paired' was used to merge and denoise paired-end reads, with the complete command line: qiime dada2 denoise-paired --i-demultiplexed-seqs reads-trimmed.qza --p-trunc-len-f 220 --p-trunc-len-r 220 --p-n-threads 12 --output-dir dada2_output. The output data were exported for analysis outside qiime2 using the command 'qiime tools export'. In addition, qiime2 generates plots (.qzv) and artifacts (=data) (.qza) files that can be viewed in the qiime2 web browser (https://view.qiime2.org/).

RIMS-seq analysis

Data preprocessing

Custom python scripts were developed to automate the analysis. Paired-end reads were trimmed using Trim Galore 0.6.3 (option --paired). The 47 bacterial genomes were concatenated into one reference metagenome fasta file. Reads were mapped to the metagenome using BWA mem version 0.7.17-r1188 (Li, 2013), with the paired-end mode. The unmapped reads, reads without a mapped mate, the non-primary alignments and supplementary alignments were filtered out using samtools version 1.10 (Li *et al.*, 2009) and the flags -F 4 -f 2 -F 256 -F 2048.

Abundance analysis

Samtools idxstats was used to retrieve and print statistics from each indexed bam file. Then, the abundance of each bacterium in each sample was estimated using a custom python script that takes as input the idxstats file. The R packages phyloseq (McMurdie and Holmes, 2013) and DESEQ2 (Love, Huber and Anders, 2014) were used to perform a differential abundance analysis and visualize the abundance data. A simple DESEQ2 design formula was used to investigate the effect of the antibiotic treatment, for each time point individually, using duplicates: phyloseq_to_deseq2(phyloseq_object, ~treatment).

De novo m5C motif identification

Scripts and additional details for the *de novo* identification of motifs in RIMS-seq can be found on Github (https://github.com/Ettwiller/RIMS-seq/). The bam files were split by bacteria using a custom python script (x47 bam files per sample). In order to limit the number of files generated and to increase the amount of data for the lower abundant bacteria, the bam files were merged by replicate and treatment condition using a custom python script. This represents 47 bacteria x 2 replicates x 2 conditions = 188 files to analyze (instead of 1316 files if we performed the analysis on all the samples independently). The number of reads per bam file was calculated and the bam files with more than 5 million reads were downsampled to 5 million using a custom python script. The RIMS-seq pipeline was run on all the files using a custom python script that automates the RIMS-seq pipeline (split_mapped_reads.pl and get_motif_all.pl). The results for the binned genomes were not taken into account as the SNPs resulting from the difference between the genomes confuse the motif identification. High confidence motifs were identified for 10/47 bacteria.

RNA-seq analysis

Data preprocessing

Custom python scripts were developed to automate the analysis. Paired-end reads were trimmed using Trim Galore 0.6.3 (option --paired). The 47 bacterial genomes were concatenated into one reference metagenome fasta file. Reads were mapped to the metagenome using BWA mem version 0.7.17-r1188 (Li, 2013), with the paired-end mode. The unmapped reads, reads without a mapped mate, the non-primary alignments and supplementary alignments were filtered out using samtools version 1.10 (Li *et al.*, 2009) and the flags -f 2 -F 4 -F 256 -F 2048 -q 1.

Differential gene expressions analysis

The bam files were split by bacteria using a custom python script (x47 bam files per sample). The matrix of the number of mapped reads per gene was generated using featureCounts (Liao, Smyth and Shi, 2014)and served as input for the differential analysis performed using the R Package DESEQ2. A simple DESEQ2 design formula was used to investigate the effect of the antibiotic treatment, for each time point individually, using duplicates: DESeq(dds, ~treatment).

Analysis using the MicroScope platform

Following integration into the MicroScope platform, the 47 genomes were annotated, the Cluster of Orthologous Groups (COG) and metabolic pathways predicted for all the genes.

Cappable-seq analysis

Data preprocessing

Custom python scripts were developed to automate the analysis. Paired-end reads were trimmed using Trim Galore 0.6.3 (option --paired). The 47 bacterial genomes were concatenated into one reference metagenome fasta file. Reads were mapped to the metagenome using Bowtie2 version 2.3.4.3 (Langmead and Salzberg, 2012), with the --local and paired-end mode. The unmapped reads, reads without a mapped mate, the non-primary alignments and supplementary alignments were filtered out using samtools version 1.10 (Li *et al.*, 2009) and the flags -F 4 -f 2 -F 256 -F 2048. The bam files were split by bacteria using a custom python script.

TSS analysis

The TSS positions were defined for each bacteria using the script developed by Laurence Ettwiller. Scripts and additional details can be found on Github (https://github.com/Ettwiller/TSS).

Motif logo analysis at TSS

The sequence context of the −45bp to 5bp region around the defined TSS were extracted for motif analysis. Motif logos were generated for each bacteria using the program weblogo3.6.0 (Crooks *et al.*, 2004).

Leaderless analysis

The TSS positions were compared to the start of the annotated genes. If the position between the TSS and the start of the gene was equal to 0, the transcript was considered as leaderless.

Phylogeny of the DefCom community and visualization

The phylogenetic tree was done using OrthoFinder (Emms and Kelly, 2019) version 2.3.11 using the MSA workflow and MAFFT for the multiple sequence alignment program. The program options are

available at https://github.com/davidemms/OrthoFinder. The phylogenetic tree and abundance data obtained from RIMS-seq were visualized using iTOL (Letunic and Bork, 2021).

C. Results

1. Ciprofloxacin impacts the overall growth of the DefCom

The growth of the DefCom community was monitored over the experiment, before and after ciprofloxacin addition, for both 'control' and 'treated' cultures (**Figure 30**). Initially, the growth curves of the cultures were similar. After ciprofloxacin was added at mid-log phase (around 5h), a shift is observed for the treated cultures. Indeed, the antibiotic delays the overall growth of the community but does not completely stop it or wipes out the community.

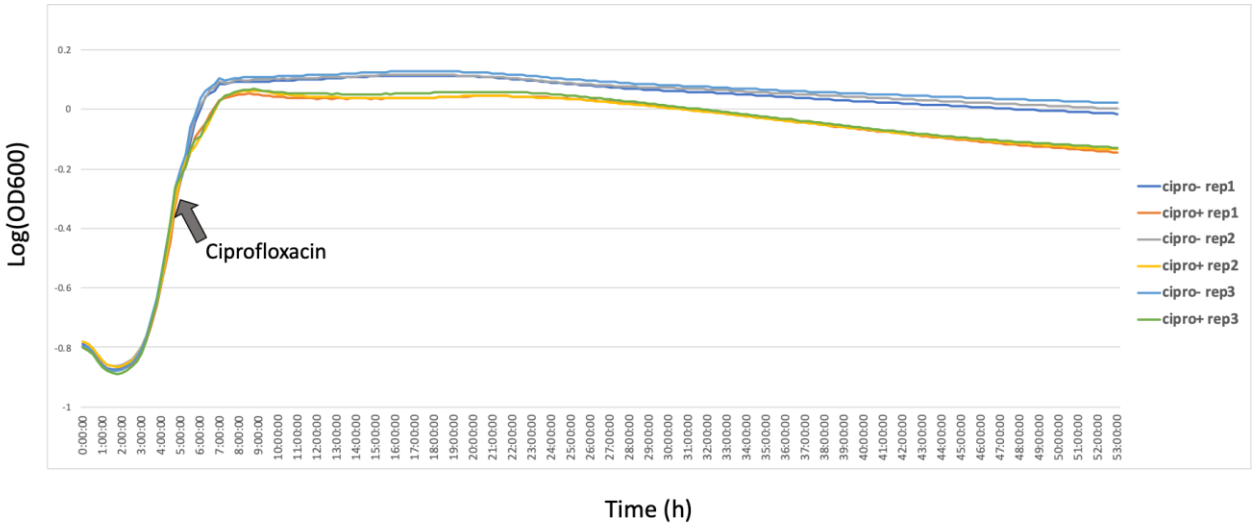


Figure 30: Growth curve of DefCom community grown in Mega medium supplemented with 0.5% glucose. 10 µg/mL ciprofloxacin was added to 'cipro+' cultures at t=5h (OD600=0.5).

2. Ciprofloxacin induces a shift in the DefCom composition

In order to identify the compositional changes induced by the ciprofloxacin addition, we performed 16S rRNA gene sequencing as well as shotgun metagenomics sequencing. For the shotgun sequencing, we used RIMS-seq, as the method allows to get the same information as regular DNA-seq, and provides additional information on the m5C status of the bacteria. In addition, this is a good opportunity to test and validate RIMS-seq on a complex microbial community.

Ciprofloxacin restructures the DefCom community

The *Firmicutes* to *Bacteroidetes* ratio is considered as having significant relevance to the composition of intestinal microbiota as *Firmicutes* and *Bacteroidetes* represent the predominant phyla in human and mice intestinal microbiota (Khan *et al.*, 2019). Overall, the *Firmicutes/Bacteroidetes* ratio was increased by ciprofloxacin (**Figure 31**), explained by a global increase in *Firmicutes* and a decrease in *Bacteroidetes*. Several studies have reported similar results in mice, with an increased *Firmicutes/Bacteroidetes* ratio (Zhang *et al.*, 2014) following ciprofloxacin treatment.

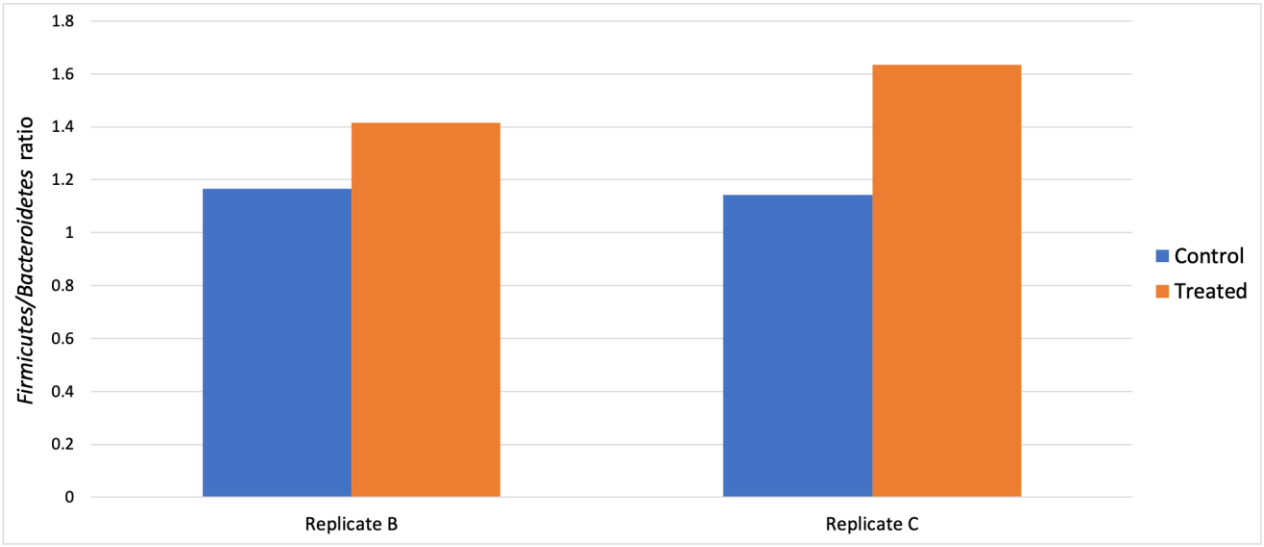


Figure 31: Barplot of the effect of ciprofloxacin on the ratio of total Firmicutes/Bacteroidetes species. The relative abundance of Firmicutes and Bacteroidetes was calculated from the RIMS-seq data over the whole treatment time.

Ciprofloxacin alters the relative abundance of bacterial phyla in the community

We analysed the community composition dynamics at the phylum level, for the control and treated samples, using 16S data (**Figure 32**) and RIMS-seq data (**Figure 33**). It should be noted that the 16S data contains 3 biological replicates (A, B and C), while the RIMS-seq data were generated for duplicates only (B and C). The reason is that 2 samples from replicate A were lost during RNA extraction (t=0 control and t=0 treated), therefore we couldn't use replicate A for RNA-seq and transcriptome analysis. We thus chose to rule out replicate A for the RIMS-seq analysis and performed the libraries only on duplicates. Still, we notice that the replicates are very similar intra-experimentally. Overall, the 16S and RIMS-seq data show similar results: following ciprofloxacin addition a decrease in the *Proteobacteria* and *Bacteroidetes* relative abundance is observed as well as an increase in the *Firmicutes* compared to the control sample. The *Actinobacteria* are present in a very low abundance and have very little coverage compared to the other phyla. This phylum of bacteria requires particular conditions to grow, such as the presence of gastric mucin (Ruas-Madiedo *et al.*, 2008). Taken together, these results also confirm that RIMS-seq can be used instead of regular shotgun sequencing to determine species abundance in complex microbial communities.

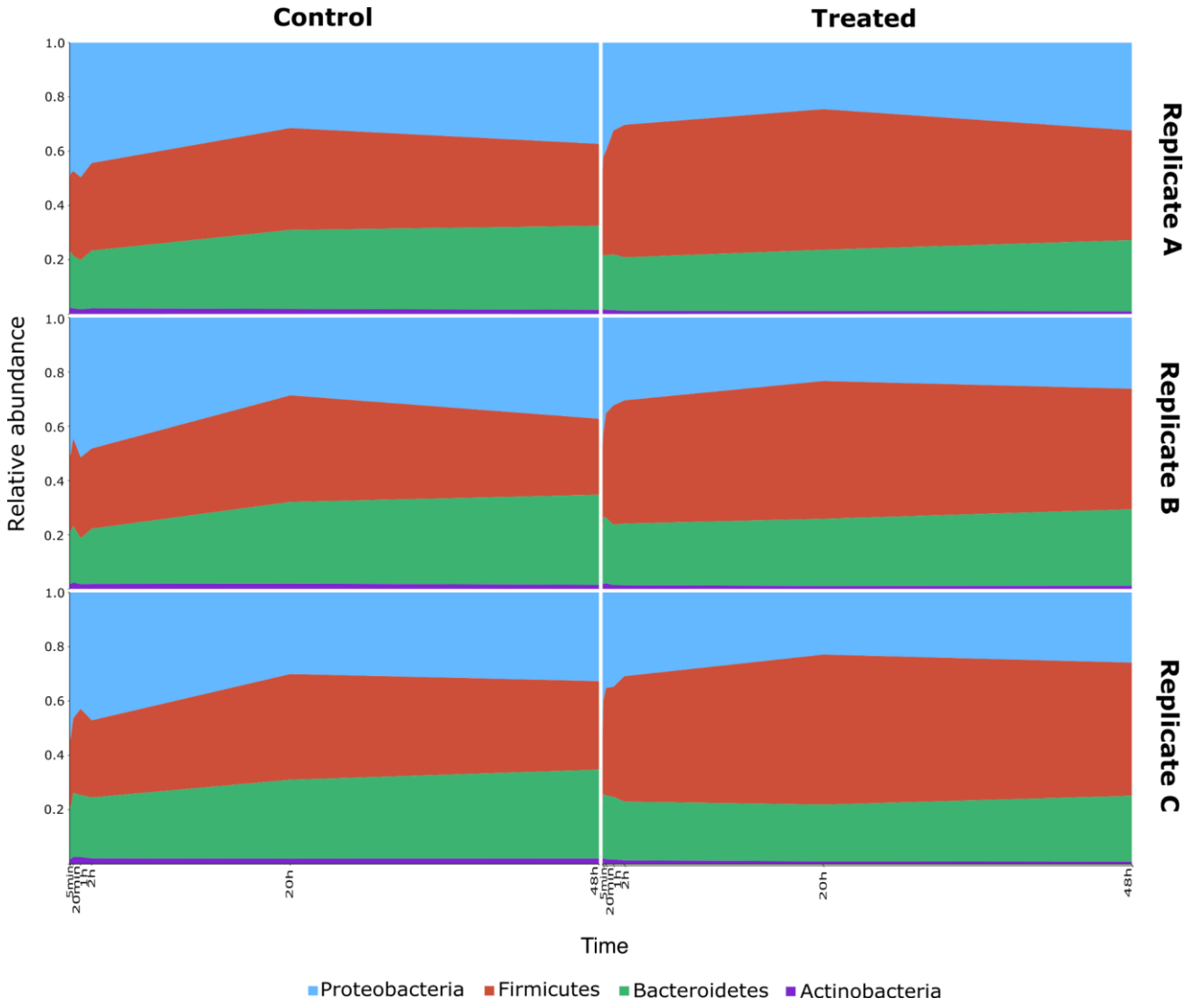


Figure 32: Community composition dynamics at the phylum level, over time for the control and the treated (ciprofloxacin treatment) replicates. The phylum relative abundance was determined using the 16S data.

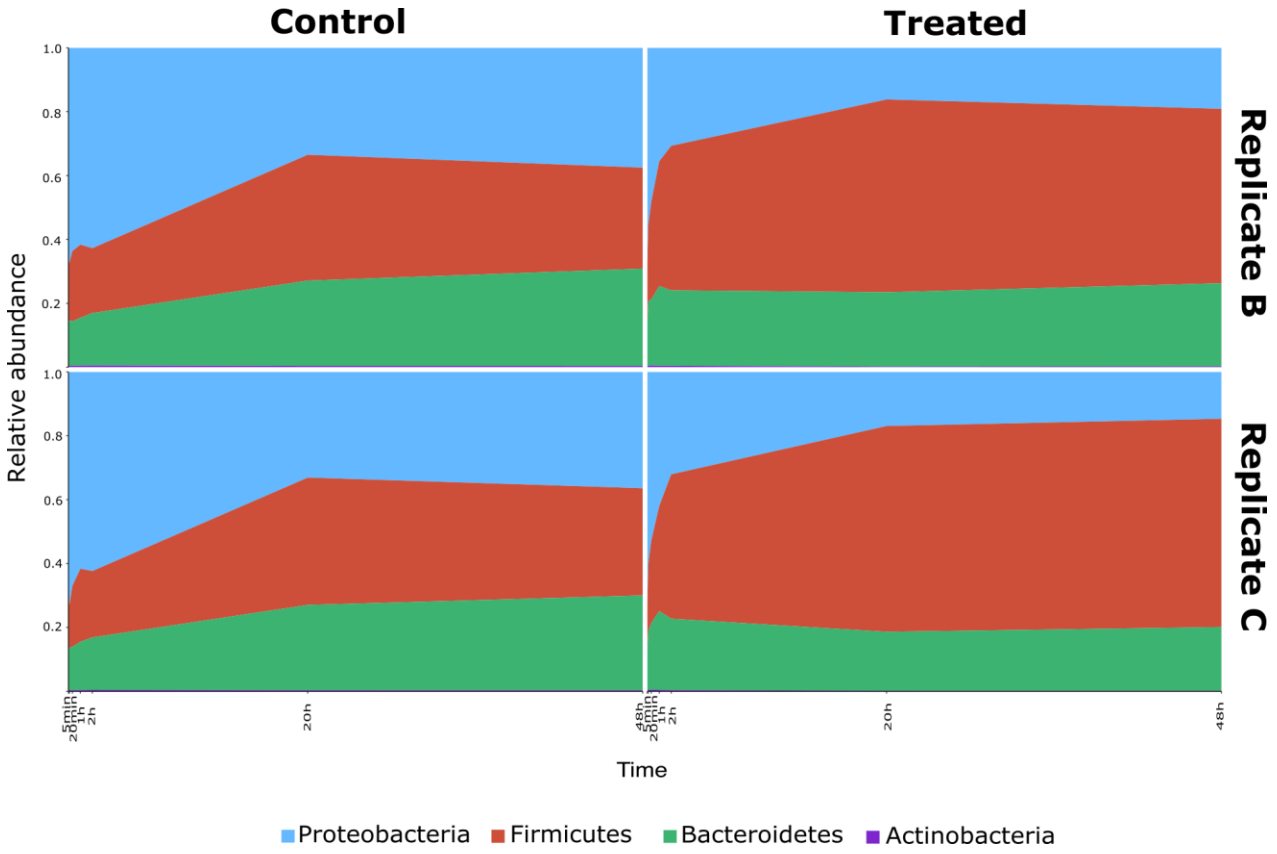


Figure 33: Community composition dynamics at the phylum level, over time for the control and the treated (ciprofloxacin treatment) replicates. The phylum relative abundance was determined using the RIMS-seq data.

Ciprofloxacin alters the relative abundance of various bacterial species in the community

In order to determine the bacteria significantly (p-value <0.05) impacted by ciprofloxacin after 48h of culture, we performed a differential abundance analysis between the control and treated samples using the R packages phyloseq (McMurdie and Holmes, 2013) and DESEQ2 (Love, Huber and Anders, 2014). Among the 47 bacteria present in the community, 31 have their abundance significantly modified by the antibiotic after 48h. 15/31 bacteria were significantly more abundant and 16/31 significantly less abundant in the treated sample compared to the control. After 48h, we can identify the 'winners' bacteria (abundance significantly increased) and the 'losers' bacteria (abundance significantly decreased). Overall, *Firmicutes* (*Ruminococcus gnavus, Tyzzerella nexilis, Enterocloster bolteae, Clostridium scindens, Dorea formicigenerans,* 4 *Blautia* strains) significantly increased in relative abundances following ciprofloxacin addition, while *Proteobacteria* (*E. coli,*

Salmonella, Klebsiella pneumoniae) and *Bacteroidetes* (*Odoribacter splanchnicus, Parabacteroides merdae, Bacteroides finegoldii, Prevotella copri*) decreased in relative abundance after ciprofloxacin addition (**Figure 34**).

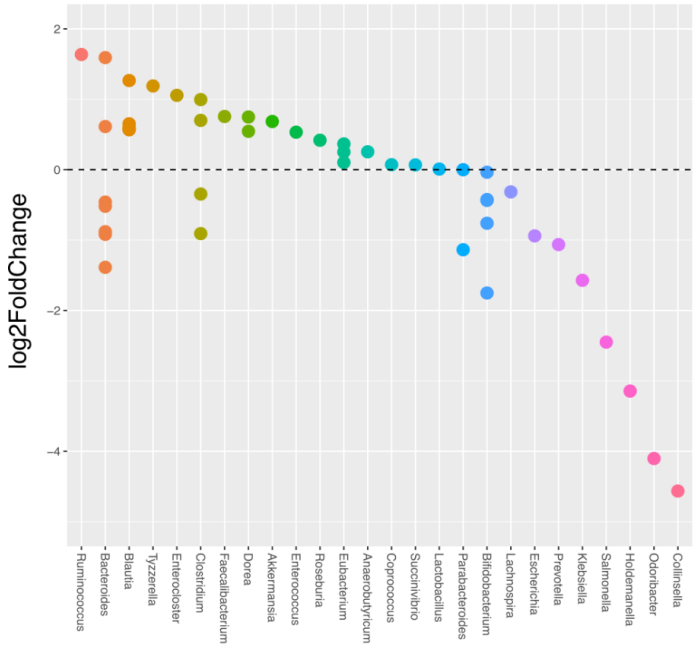


Figure 34: Log2FoldChange of the relative abundance after 48h of culture between the control and treated sample, for bacteria grouped at the genus level. The bacteria presented on these panels show a significantly different relative abundance (p-value <0.05) between the control and treated sample at 48h. The R packages phyloseq (McMurdie and Holmes, 2013) and DESEQ2 (Love, Huber and Anders, 2014) were used to perform a differential abundance analysis. Colors represent different bacterial genus.

We then focused on the ciprofloxacin effect on the community composition over time. We analyzed the dynamics of the community composition at the species level, using the RIMS-seq data (**Figure 35** below). Overall, the biological replicates show similar results, indicating the cultures are reproducible. Interestingly, we notice *Enterococcus faecium* blooms after 2h of ciprofloxacin addition, while the abundance of this bacteria was not significantly different from the control when we analysed the results at 48h. In replicate B, *E. faecium's* relative abundance progressively decreases to a level similar to the control at 48h, while in replicate C the relative abundance at 48h is superior to *E. faecium's abundance* in the control. Thus, we can't conclude on the *E. faecium* outcome after 48h of ciprofloxacin addition, however, these results indicate that *E. faecium* takes over the community

very rapidly after antibiotic addition. These results suggest that ciprofloxacin induced a shift in the community composition that favors *E. faecium* growth, a multidrug resistant opportunistic pathogen, part of the ESKAPE group of pathogens (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, *Enterobacter spp*) (Santajit and Indrawattana, 2016). More specifically, the strain in the DefCom community *E. faecium* ATCC700221 is a Vancomycin Resistant Enterococcus (VRE) known to be poorly sensitive to fluoroquinolones (Akpaka *et al.*, 2017), notably thanks to evolved mutations in the target genes of ciprofloxacin, DNA gyrase (*gyrA*) and topoisomerase (*parC*) (Leavis *et al.*, 2006). Using ResFinder 4.1 (Zankari *et al.*, 2012), an online website that allows to identify acquired antibiotic resistance genes and/or chromosomal point mutations from genomes, we identified 2 chromosomal point mutations in *E. faecium,* located in the *gyrA* (p.S83R) and in the *parC* (p.S801). Conversely, ciprofloxacin induces a decrease of the *Proteobacteria*, including the pathogens *Salmonella enterica* and *Klebsiella pneumoniae*. Taken together, ciprofloxacin addition resulted in a marked decrease of the *Proteobacteria* and *Bacteroidetes*, whereas it increased several bacterial species, mainly from the *Firmicutes* phylum. The antibiotic concentration used did not completely suppress bacteria from the community, but rather restructured the community's composition, notably with a shift observed toward the opportunistic pathogen *E. faecium*.

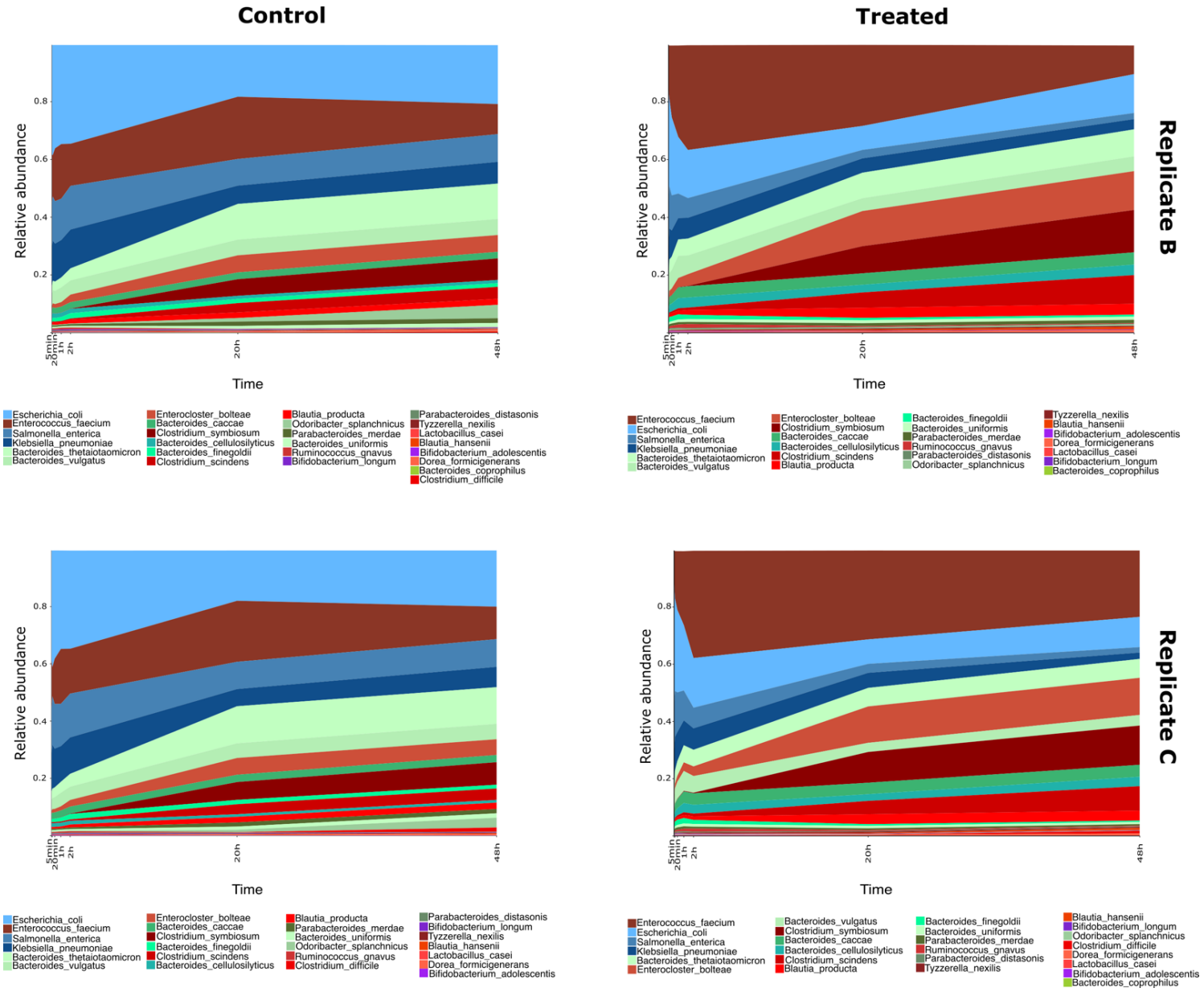


Figure 35: Community composition dynamics at the species level, over time (48h) for the control and the treated replicates B and C. The relative abundance was determined using the RIMS-seq data. Only the species with a total relative abundance greater than 0.5% were selected (B control: 27 species, C control: 27 species, B control: 25 species, B treated: 26 species). Each color represents a bacterium and the color gradient indicates in which phylum the bacteria belong. Red: Firmicutes, Blue: Proteobacteria, Green: Bacteroidetes, Purple: Actinobacteria.

3. *De novo* m5C motif identification in the DefCom

RIMS-seq allows to get both the DNA sequences (used previously for composition changes analysis) as well as the m5C motifs of the bacteria present in the community. Here, we performed a *de novo* analysis to identify the m5C specificities from the DefCom community. In order to limit analysis resources used and to increase the amount of data for the lower abundant bacteria, the data were merged by replicate and treatment condition. In addition to *de novo* determining the m5C specificities, we compare the motifs identified in the treatment *versus* control samples and investigate if condition-specific methylation could be identified (purely hypothetical). The results for the binned genomes were not taken into account as the SNPs resulting from the difference between the genomes confuse the motif identification. In addition, *Blautia* results need to be considered carefully, as 5 sub-species are present. In that case, the presence of several closely related *Blautia* strains can provoke ambiguous mapping and consequently, it would be difficult to assign an absolute motif to a bacterial genome with confidence.

The RIMS-seq pipeline for *de novo* m5C motif identification was run for all the 47 bacteria. It should be stressed that RIMS-seq requires between 1 to 4 million reads per bacteria to find fully and partially m5C methylated motifs *de novo*, respectively. Bacteria are present in very different proportions in the DefCom community and numerous species did not meet the required 1 million reads sequencing depth to ensure m5C methylation identification. So we consider a motif with high confidence if it is found in both replicates and has a high significance (p-value < 1e-100). High confidence motifs were identified for 10/47 bacteria (**Table 13**). These motifs vary in length from 4 to 6nt and are palindromic and have already been reported in the REBASE database (Roberts *et al.*, 2015), except one: interestingly, we identified a new m5C motif for *Odoribacter splanchnicus* (TC**C**GGA) that has not been reported in REBASE.

Bacteria	m5C specificity	p-value
Blautia producta	GAT**C**	1.1e-1269
Clostridium scindens	AG**C**T	2.4E-1268
Enterocloster bolteae	N**C**GGSNNN	9.1e-112
Enterococcus faecium	GAT**C**	1.1e-2934
Escherichia coli	C**C**WGG	5.2e-1125
Eubacterium ventriosum	C**C**GG	4.2E-61
Klebsiella pneumoniae	C**C**WGG	9.9e-487
Odoribacter splanchnicus	TC**C**GGA	8.3e-658
Parabacteroides distasonis	**C**GCG	4.30E-180
Salmonella enterica	C**C**WGG	8.2e-746

Table 13: High confidence m5C methylases specificities obtained using RIMS-seq. These motifs are present in both control and treated samples, with a significant p-value (p-value < 1e-100). All the motifs have been described in REBASE (Roberts et al., 2015) and can be validated, except a new motif that was identified for Odoribacter splanchnicus. The methylated cytosine within the motif is in bold and underlined.

Overall, we did not identify any condition-specific m5C motif. The identified m5C specificites were the same in both the control and treated conditions. Taken together, these results demonstrated that RIMS-seq can be applied to complex microbial communities for both compositional analysis and m5C specificities identification. This represents a nice additional validation of the method.

4. Preliminary analysis of the transcriptomic response of the DefCom after ciprofloxacin addition

In the first part, we analyzed the effect of ciprofloxacin on the composition of the DefCom community and identified significant changes in the overall structure of the community and in various bacterial species abundance. Now that we have identified which bacteria are impacted by the antibiotic, in this second part we focus on the transcriptomic data in order to understand how the bacteria are reacting. Adding the functional aspect will give important insights to understand what is happening to the bacteria in the community after ciprofloxain addition. Ideally, being able to link compositional to functional changes could give important keys that could help to potentially predict

the outcome of an antibiotic treatment. The RNA-seq analysis is done in collaboration with the Labgem team from the Genoscope (David Vallenet, David Roche and Stéphanie Fouteau), using the MicroScope platform (Vallenet *et al.*, 2020). This platform is of particular interest as it allows to annotate and perform comparative analysis of prokaryotic genomes (Médigue *et al.*, 2019), including comparative metabolic pathways analysis from KEGG (Kanehisa *et al.*, 2017) or MetaCyc (Caspi *et al.*, 2018). However, It should be stressed that the data and results that will be presented in this part are preliminary. These data are a gold mine but complex and thus, require time to be analysed entirely and carefully. In addition, the Covid-19 pandemic delayed the experiments and we are currently still analysing the transcriptomic data.

The transcriptomic response to ciprofloxacin is immediate

We calculated the number of differentially expressed genes (DEG) for the 47 bacteria of the DefCom observed between the control and treated samples at t=5min and t=20min and plotted the results with the differential abundance data obtained from RIMS-seq, for t=5min and t=48h (**Figure 36**). First, we observe that few compositional changes are observed after a short time of ciprofloxacin addition (5min), whereas significant changes in relative abundance are observed for all the bacteria after 48h. This confirms it takes time to observe compositional changes and that most of the changes are observed after hours/days of treatment. Conversely, the transcriptomic changes are fast, with about 1/4th of the genes of *Proteobacteria* being differentially expressed after 5min of ciprofloxacin addition. The transcriptomic response gets even stronger after 20min of ciprofloxacin, with important modifications observed in the transcriptome of *Actinobacteria.* Here, we highlight an immediate transcriptomic response following ciprofloxacin treatment. To our knowledge this is the first time such a fast transcriptomic response is observed in a complex community.

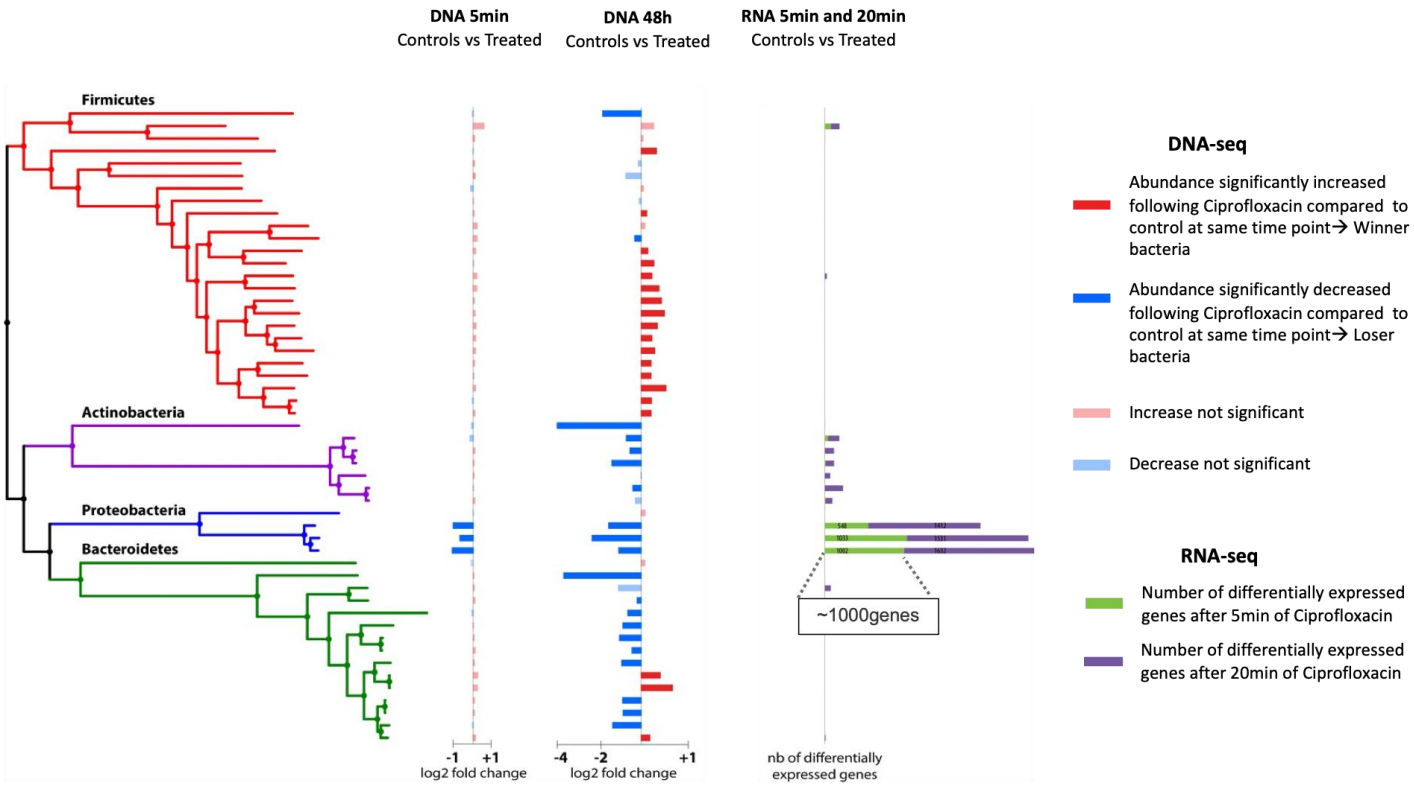


Figure 36: Phylogenetic tree of the community showing the differential abundance between the control and treated samples after 5min (first line) and 48h (second line) as well as the number of differentially expressed genes after 5min and 20min of ciprofloxacin (third line). The abundance log2foldChange for each bacterium was calculated and is represented on the histograms. A red bar represents an abundance increase for this bacterium compared to the control sample, while a blue bar represents an abundance decrease for this bacterium compared to the control sample. A transparent red bar is a non-significant increase, a transparent blue bar is a non-significant decrease. Conversely, an opaque red bar represents a significant abundance increase (p-value <0.05) and an opaque blue bar represents a significant abundance decrease in the treated sample compared to the control.

Miscellaneous observations on the transcriptomic response of the DefCom community (preliminary)

We previously highlighted an almost immediate transcriptomic response of the DefCom after ciprofloxacin addition. The next step is to characterize the functional response of the bacteria and investigate if a phylum-specific response can be identified, potentially enabling to predict the later compositional changes and ultimately the outcome of the ciprofloxacin treatment.

For a matter of time, we focused on a set of bacteria to analyze (**Table 14**). We chose bacteria to represent different phyla and that show different status after the ciprofloxacin addition (relative abundance increased, decreased or stable compared to the control).

Bacteria	Phylum	relative abundance after ciprofloxacin compared to control
Escherichia coli	*Proteobacteria*	decreased
Klebsiella pneuminiae	*Proteobacteria*	decreased
Salmonella enterica	*Proteobacteria*	decreased
Bacteroides cellulosilyticus	*Bacteroidetes*	increased
Bacteroides caccae	*Bacteroidetes*	increased
Bacteroides finegoldii	*Bacteroidetes*	decreased
Bacteroides thetaiotaomicron	*Bacteroidetes*	decreased
Blautia producta	*Firmicutes*	increased
Enterococcus faecium	*Firmicutes*	increased
Enterocloster bolteae	*Firmicutes*	increased

Table 14: List of bacteria selected for further transcriptomic analysis.

We found that *Proteobacteria* are among the first responder of the community and their relative abundance significantly decreased following ciprofloxacin compared to the control. Overall, we observed SOS response related genes being upregulated as soons as 5min and for several times following the antibiotic addition in *E. coli, Salmonella enterica* and *Klebsiella pneumoniae: recA, recN* (Homologous repair pathway), *uvrA* (Nucleotide Excision Repair pathway), *umuD, umuC, dinBDF* (DNA translesion pathway). Interestingly, we found phage-related genes being upregulated early in *E. coli,* with the *kilR* and *racR* genes from the Rac prophage. Similarly, we found phage-related genes overexpressed in *Salmonella* as soon as 20min after the antibiotic addition. The role of prophages in environmental adaptation has already been described in both bacteria, notably showing that the SOS response induced by fluoroquinolones may also induce the transduction of prophages (Bearson and Brunelle, 2015; Valat *et al.*, 2020). As prophages may carry virulence genes, their induction could

increase horizontal gene transfer and bacterial pathogenicity (Penadés *et al.*, 2015). The role of prophage in bacterial genomes is being studied and even considered as a potential target for antimicrobial drug development, as phages can help bacteria cope with various environmental perturbations (Wang and Wood, 2016).

Interestingly, we noticed ciprofloxacin had a different impact on different *Bacteroides* species, with *Bacteroides* species taking over the DefCom and some others being highly impacted after 48h of ciprofloxacin. We explored if a difference in the transcriptomic response could explain this resistance or sensitivity in the different *Bacteroides* species. For all the species we observed several genes coding for efflux pumps being upregulated, with a higher proportion of these genes detected in the species that increased in abundance after ciprofloxacin, *B. caccae* and *B. celulosilyticus* (*mexB, bepE,* MATE family efflux). Preliminary analysis did not provide further explanation on the differences in behaviors toward ciprofloxacin. It is likely the difference lies in the type, efficiency or number of efflux pumps expressed by the *Bacteroides*. We hope to investigate this question further with the determination and analysis of *Bacteroides* core-genome using the MicroScope platform.

Lastly, we focused on the *Firmicutes* phylum, which overall showed increased relative abundance following ciprofloxacin. Interestingly, we found *Blautia producta* upregulates, during the entire time course, a gene coding for a DNA topoisomerase III. In *E. coli,* this enzyme has been shown to perform a similar role as topoisomerase IV by assisting it during DNA replication (Lee *et al.*, 2019). A similar observation was found for *Enterocloster bolteae*, with the overexpression of the *traE* gene as soon as 20min after ciprofloxacin addition. The TraE protein has been shown to exhibit in *E. coli* a topoisomerase activity similar to that of topoisomerase III (Li *et al.*, 1997). As topoisomerase IV is one of the ciprofloxacin's targets, one hypothesis could be that those species of *Blautia* and *Enterocloster* use topoisomerase III to take over the function of the inhibited topoisomerase IV, maintaining DNA replication. We previously observed that *Enterococcus faecium* rapidly takes over the DefCom community following ciprofloxacin and that this strain carries chromosomal mutations in the genes coding for the targets of the antibiotic. Transcriptomic analysis of this bacteria revealed upregulation of various defense mechanisms, such as virulence factors (*agaS*) and numerous multidrug

transporters, with 14 drug transporters being upregulated after 1h of antibiotic addition. Efflux pump systems and mutations in target genes could explain the resistance of *Firmicutes* after ciprofloxacin addition.

5. TSS identification in the DefCom

In order to investigate ciprofloxacin treatment-dependent regulation mechanisms as well as determining the promoters structures, we prepared Cappable-seq libraries (Ettwiller *et al.*, 2016) on the control and treated samples, in duplicates (B and C), for the t=0min and t=5min. Our initial plan was to perform a differential analysis of TSS expression between t=0min and t=5min as TSS regulation occurs very rapidly and modifications in their expression should be observed even faster than the response observed from the RNA-seq data. Unfortunately, a crucial Cappable-seq library failed (replicate B, treated, t=5min) and the library couldn't be repeated. Without replicates, it is impossible to perform a differential TSS expression analysis. So here, we focus on the determination of promoter's structures.

Promoters are DNA sequences on which RNA polymerases bind to initiate transcription. The *E. coli* promoter structure has been well described and contains two consensus sequences: a -10 element (TATAAT box) and a -35 element (TTGACA), located at 10 and 35 bp upstream from the TSS, respectively. Even if promoters share common structural features, their sequences can vary extensively (Hawley and McClure, 1983). As an example, it has been shown that *Bacteroides fragilis* contains two conserved regions similar to *E. coli*, but the regions and sequences are different, with one element at -7 (TANNTTG) and one element at -33 (TTTG) (Mastropaolo, Thorson and Stevens, 2009).

We first identified the TSS positions for 35/47 species. There were not enough read coverage to call the TSS for the other 12 species. We determined the promoter structure of the bacteria in the DefCom community by analyzing the sequence around the defined TSS and identified a variety of promoter structures (**Figure 37**). We observed a preference for A and G at the TSS position, which was already reported (Hawley and McClure, 1983; Kim *et al.*, 2012; Ettwiller *et al.*, 2016). We found *Clostridium scindens* and *Enterocloster bolteae* have a promoter structure similar to *E. coli* with canonical -10 and -35 regions, whereas *Bacteroides thetaiotaomicron* and *Bifidobacterium catenulatum* have no

canonical promoter sequence. These results highlight the considerable diversity of promoter structures.

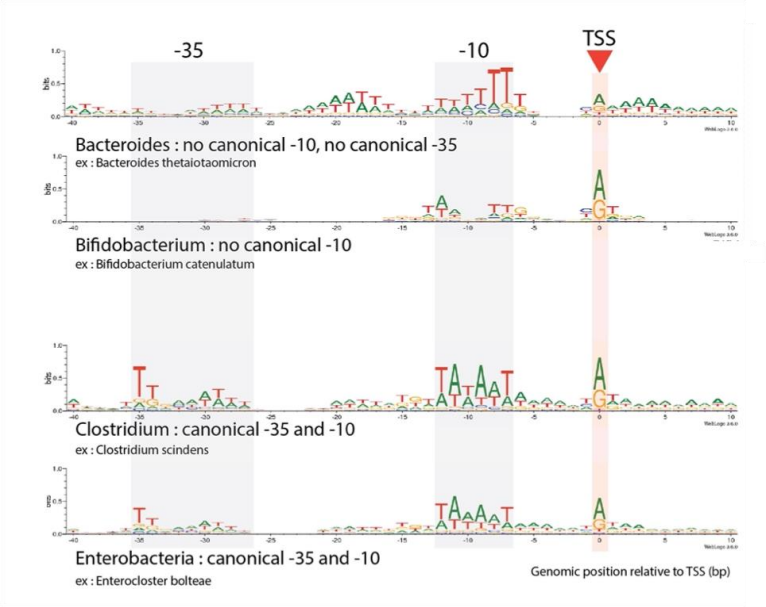


Figure 37: Examples of different promoter structures identified from the defined community. Bacteroides thetaiotaomicron and Bifidobacterium catenulatum have non canonical -10 and -35 promoter regions (top panel), while Clostridium scindens and Enterocloster bolteae (bottom panel) have a canonical -35 and -10 promoter structure. The red arrow indicates the TSS base position determined by Cappable-seq.

The Shine-Dalgarno signal has been described as the dominant translation initiation mechanism in prokaryotes. However, leaderless genes that lack the 5' Untranslated Transcribed Region (5'UTR) and the Shine-Dalgarno sequence on their transcripts, have been described in several bacteria (Zheng *et al.*, 2011; Schrader *et al.*, 2014) notably *Mycobacteria*, in which they are particularly abundant (14% of genes are leaderless) (Nguyen *et al.*, 2020). In the case of leaderless transcription, the ribosomes bind directly to the TSS, thus the AUG start codon itself serves as the signal for the translation initiation. Still, leaderless mRNAs seem to be more prevalent in Gram-positive bacteria and in archaea (Moll *et al.*, 2002) but remain poorly understood compared with canonical mRNA translation.

Using the Cappable-seq data, we calculated the proportion of leaderless transcripts in the bacteria from the DefCom (**Figure 38**). Interestingly, we found that *Akkermensia muciniphila* and the *Bifidobacterium* harbor a high proportion of leaderless transcripts, accounting for 6 to 15 % of

identified TSS. A previous Cappable-seq experiment in a mouse microbiome showed similar results, with abundant leaderless transcripts identified in *Akkermansia muciniphila* and *Bifidobacterium pseudolongum* (Ettwiller *et al.*, 2016). In addition, high occurrences of leaderless genes in *Actinobacteria* have previously been described (Zheng *et al.*, 2011). Conversely, leaderless transcripts are very low in *E. coli*, consistent with previous studies that suggest leaderless transcripts in *E. coli* are generally translated less efficiently (O'Donnell and Janssen, 2001).

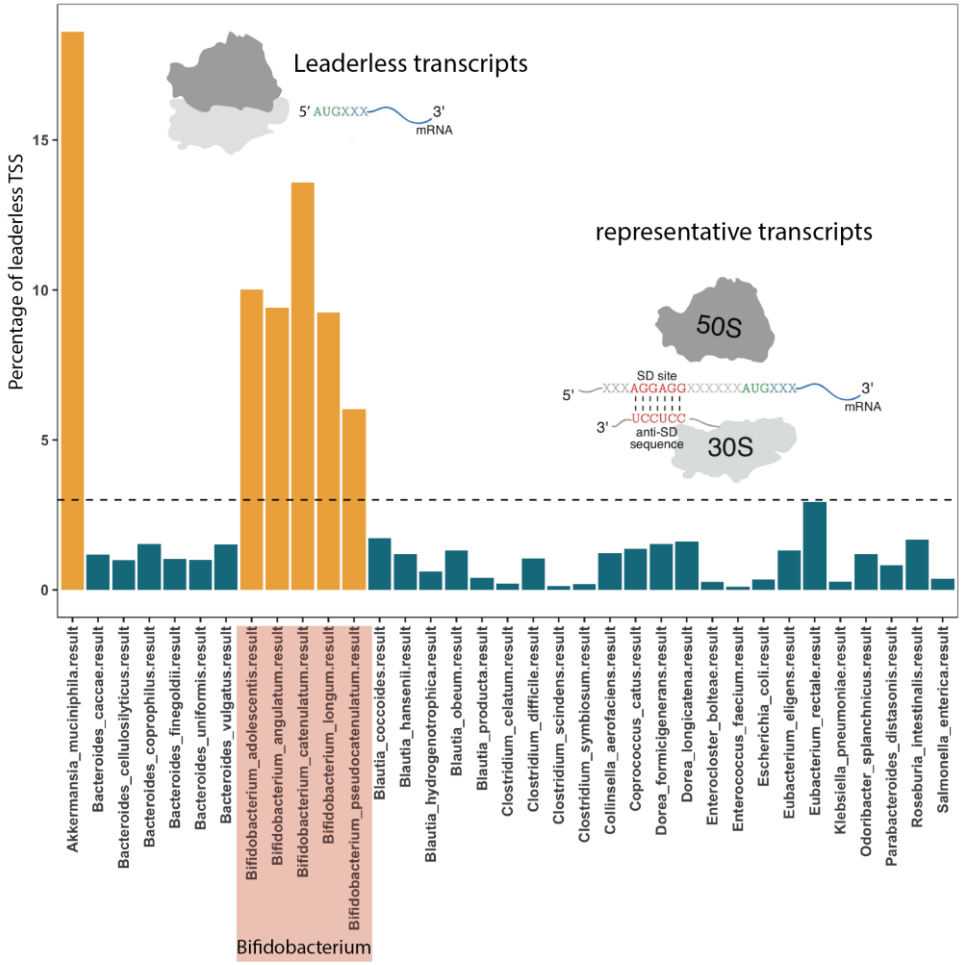


Figure 38: Percentage of leaderless transcripts calculated for 35/47 bacteria from the DefCom community. There were not enough reads to call the TSS for the other species.

To summarize this part, using Cappable-seq, we identified TSS from various bacteria in the DefCom community. The results showed a diversity of promoter structures and different translation initiation strategies, with abundant leaderless transcripts found in *Bifidobacterium*.

D. Conclusion and further perspectives

Here we used a defined synthetic community of 51 bacteria representative of the gut and investigated the impact of ciprofloxacin on both the community composition and functional response of the bacteria, over 48h. We explored multiple aspects of this community following antibiotic addition, including compositional changes, methylation characterization, TSS identification and regulation of gene expression. We identified significant shifts in the bacterial composition after several hours/days of antibiotic addition, with on one hand, species able to resist ciprofloxacin and taking over the community rapidly such as *Enterococcus faecium* and more generally *Firmicutes*, whereas some phyla of bacteria such as *Proteobacteria* and *Bacteroidetes* were significantly decreased by the antibiotic. We aim at correlating the community composition changes to the transcriptional reprogramming of bacteria that, as opposed to compositional changes, occurs very rapidly. Indeed, for the first time to our knowledge, we identified significant transcriptional reprogramming after only 5min of ciprofloxacin addition for several bacteria. More specifically, we identified the *Proteobacteria* as the first responders, those species immediately triggering the SOS response pathway, whereas the *Firmicutes* tend to set up defense mechanisms such as drug efflux pumps, suggesting there is a phylum-specific response. Overall, these preliminary observations are promising and correlate well with the expected response to ciprofloxacin that has been extensively described in literature. However, those results and the observations that emerged from them are preliminary and further analysis is needed to validate the hypothesis we presented in the above sections. The next step is to determine in which metabolic pathways the differentially expressed genes are involved. Currently, analysis of the transcriptomic response of the community on a pathway level is ongoing and could enable the identification of specific regulated pathways that could help predict the outcome of the antibiotic addition.

In addition to examining the short- and long-term responses of the community, we identified various promoter sequences as well as different transcription mechanisms such as leaderless transcription. Those results illustrate well the diversity of systems that bacteria possess to regulate transcription and adapt quickly to their environment. We also applied our new method RIMS-seq to characterize the m5C methylation community-wide. We identified a variety of m5C motifs and demonstrated the

ability of RIMS-seq to be used for both composition determination and m5C characterization in complex communities, showing the potential of RIMS-seq to replace regular standard DNA-seq for genome sequencing.

In addition to the further bioinformatic analysis, additional experiments are ongoing, notably with the determination of the Minimum Inhibitory Concentration (MIC) of ciprofloxacin for each bacterium in monoculture. This additional experiment will provide important information to better understand the expected and observed response of the bacteria following antibiotic addition. Indeed, it has been shown that bacterial tolerance to antibiotics differs between monoculture *versus* in a community. Higher than expected tolerance may occur if one or more species in a community excretes a compound which either degrades antibiotics which activates tolerance mechanisms such as efflux pump expression in other species. This phenomenon of community 'cross-protection or cross-feeding' could result in lower concentrations of antibiotics and more generally, alter the efficiency of antibiotic treatments (Yurtsev *et al.*, 2013; Adamowicz *et al.*, 2018). Additionally, we initially planned to perform ONT-Cappable-seq on the DefCom to explore the effect on the operon structure regulation following antibiotic treatment. This was unfortunately impossible due to the Covid-19 pandemic that delayed experiments. Another interesting path would be to apply Loop-Cappable-seq to the DefCom, enabling a better mapping resolution and thus enabling to distinguish subspecies between each other thanks to the LoopSeq technology accuracy.

More generally, microbiome analysis should be interpreted with caution as a variety of factors can have a profound impact on the conclusions. As an example, the culture medium is known to have an effect on antibiotic efficiency and to induce competition for nutrients between species (Adamowicz *et al.*, 2018; Maier *et al.*, 2020). In our case, we performed a batch culture of the DefCom (as opposed to a continuous culture), meaning the supply of nutrients is limited. The depletion of the medium will provoke a global acidification that is likely to add another selection pressure on the bacterial community. Another factor to take into account is the oxygen level of the culture. A gradient of oxygen exists *in vivo* in the digestive tract, varying from almost anoxic environment in the intestinal lumen with <1% O2 (0.1-1 mm Hg) to 5-20% O2 in the intestinal crypts (80 mm Hg) (Kim *et al.*, 2019).

Here we performed the culture in an anaerobic environment, which is likely non-optimal for certain species and adds an additional factor of stress, competition and selection, favoring anaerobic bacteria. Lastly, the growth rate and metabolic state of bacteria have been shown to impact the antibiotic efficiency (Eng *et al.*, 1991; Lopatkin *et al.*, 2019). Overall, this highlights the complexity of reproducing optimal *in vivo* growth conditions for microbiome and synthetic community studies. The microbiome is a vast ecosystem with constant interactions with the host and within bacterial individuals, in which bacteria compete but also cooperate with each other. Microbiome is an exciting and promising field of research that is not done revealing its secrets yet.

GENERAL CONCLUSION AND PERSPECTIVES OF THE THESIS

In this last part of the thesis, I would like to conclude on the work achieved but also on the rich personal experience these 3 (and a half) years have been.

In this thesis, I presented several new methods developed for bacterial and complex community characterization. We developed RIMS-seq, a new method based on an easy protocol that enables both sequencing of genomes and characterization of m5C methylation of bacterial genomes, demonstrating the potential of RIMS-seq to replace standard DNA-seq. We successfully validated the technique by applying it on a complex defined synthetic community and implemented the method from New England Biolabs to the Genoscope. The paper presenting RIMS-seq has been recently published in *Nucleic Acids Research*. We also developed ONT-Cappable-seq and Loop-Cappable-seq, two techniques enabling sequencing of full-length bacterial transcripts, based on Nanopore and LoopSeq sequencing, respectively and enabling to reveal the complexity of operon structure regulation. In the last part, we aimed at exploring the link between the long- and short-term response of a synthetic microbiome following an antibiotic perturbation using a multiomic approach. We performed various complex experiments and analyses with the aim to identify transcriptomic responses that best correlate with long term changes in community structure. The preliminary results are promising and revealed interesting paths to pursue, with a very fast transcriptional reprogramming identified as soon as 5min after ciprofloxacin addition. This ambitious project will require additional work and analysis to fully reveal the information from this huge amount of promising and exciting data.

On a more general view, this thesis has been a very rich scientific and personal experience. I have learned so much in so many aspects that it is hard to know where to begin. I had the chance to pursue my PhD in an international environment thanks to the collaboration between New England Biolabs (NEB) in the USA and the Genoscope in France. I started my PhD at NEB (for 2 years) and did my final year at the Genoscope in France. When I moved to the USA in Ipswich, it was the first time for me to move so far away from France for such a long period of time. This experience of moving abroad and setting my life for more than 2 years in a totally unknown environment, in a country with a different language and culture, has been a great personal challenge and is one of the best choices I have made in my life so far. I have met incredible people, scientists and friends and had the opportunity to work in an amazing environment. I also had the chance to use the most recent sequencing platforms on the market (I never thought I would see and even use a PacBio during my PhD!) and I even had the privilege to get a personal MinION Nanopore sequencer. Another big challenge was to get started with bioinformatics. Learning bioinformatics and Python was one of my goals during my PhD. Thanks to the help of great (and patient) bioinformaticians, a great dose of perseverance and a lot (really a lot) of google search, I am now able to perform my own data analysis. I never thought I would be able to write my own Python scripts or to not be scared of a Linux terminal. This represents a big personal achievement that will be helpful for my future career. During this PhD, I learned how to conduct a project from the very beginning with the design of the experiments to the data analysis and the finality of publishing. I also had the chance to be involved in projects with several collaborators and particularly enjoyed working in a team. Also, I had the opportunity to present my work in several conferences in the USA but also in Europe. Those conferences are always a great opportunity to share our research and I think sharing ideas and expertises is very important and crucial to move forward in research.

Transitioning back to France at the Genoscope in the middle of my PhD and above all in the middle of the Covid-19 pandemic, has also been one huge challenge. This pandemic impacted everyone in the world and I had to adapt my thesis project as it was impossible to perform experiments in the lab for several months. Also, I had to transfer all my sequencing data generated in the USA on a hard drive that happened to be broken once in France. Thanks to the help of the Genoscope and NEB we handled the situation and I could get all my data back.

Overall, Science and more generally Research requires us to constantly adapt to various situations. Every day is different, with its dose of pleasant discoveries and sometimes less nice surprises. This PhD experience taught me how to adapt and to find solutions to various situations. Finally, if I had to resume my PhD in one word, I think it would be 'resilience' and I am looking forward to pursuing my career in the microbiome and sequencing field areas that with no doubt, will be full of exciting surprises.

SCIENTIFIC COMMUNICATIONS

Publications

- Cappable-seq: A Versatile Toolkit for the Identification of Transcriptional Landmarks in Bacteria.

  Bo Yan, **Chloé Baum** and Laurence Ettwiller. *GEN News (Genetic Engineering & Biotechnology News), 2019*

  https://www.genengnews.com/resources/tutorial/cappable-seq-a-versatile-toolkit-for-the-identification-of-transcriptional-landmarks-in-bacteria/

- Non-destructive enzymatic deamination enables single molecule long read sequencing for the determination of 5-methylcytosine and 5-hydroxymethylcytosine at single base resolution.

  Sun Z, Vaisvila R, Hussong LM, Yan B, **Baum C**, Saleh L, Saranayake M, Guan S, Dai N, Correa I, Pradhan S, Davis T, Evan T, Ettwiller L. *Genome Research, 2021*

  https://doi.org/10.1101/gr.265306.120

- Rapid Identification of Methylase Specificity (RIMS-seq) jointly identifies methylated motifs and generates shotgun sequencing of bacterial genomes.

  Baum C, Lin YC, Fomenkov A, Anton B, Chen L, Yan B, Evans TC, Roberts RJ, Tolonen AC, Ettwiller L. *Nucleic Acids Research, 2021*

  https://doi.org/10.1093/nar/gkab705

Oral communications

- Flash talk at the Boston Bacterial Meeting (BBM 2020) (Boston, USA, online)

  "*Connecting rapid, transcriptional responses to compositional changes in the gut microbiome following antibiotic treatment*"

Posters

- Poster presentation at the Clostridia XV meeting 2018 (Munich, Germany). "*Comprehensive mapping of operon structure in Clostridium phytofermentans*"

- Poster presentation at Nanopore London Calling 2019 (London, UK) "*Distinct 3'end handling of bacterial transcripts for Nanopore sequencing leads to considerable disparity in the definition of full-length transcripts*"

- Poster presentation at the EMBL2021 Symposium: New approaches and concepts in Microbiology (Heidelberg, Germany, online)
  "RIMS-seq jointly identifies methylated motifs and generates shotgun sequencing of bacterial genomes"

Other

- Organization of the Boston Bacterial Meeting (BBM 2019), Registration team (Harvard Science Center, Boston, USA)

REFERENCES

16S Sequencing vs Shotgun Metagenomic Sequencing (2021). Available at: https://www.zymoresearch.com/blogs/blog/16s-sequencing-vs-shotgun-metagenomic-sequencing (Accessed: 15 July 2021).

Abeles, S. R. *et al.* (2016) 'Microbial diversity in individuals and their household contacts following typical antibiotic courses', *Microbiome*, 4(1), p. 39.

de Abreu, V. A. C., Perdigão, J. and Almeida, S. (2020) 'Metagenomic Approaches to Analyze Antimicrobial Resistance: An Overview', *Frontiers in genetics*, 11, p. 575592.

Adamowicz, E. M. *et al.* (2018) 'Cross-feeding modulates antibiotic tolerance in bacterial communities', *The ISME journal*, 12(11), pp. 2723–2735.

Akpaka, P. E. *et al.* (2017) 'Genetic characteristics and molecular epidemiology of vancomycin-resistant Enterococci isolates from Caribbean countries', *PloS one*, 12(10), p. e0185920.

Altschul, S. F. *et al.* (1990) 'Basic local alignment search tool', *Journal of molecular biology*, 215(3), pp. 403–410.

Amarasinghe, S. L. *et al.* (2020) 'Opportunities and challenges in long-read sequencing data analysis', *Genome biology*, 21(1), p. 30.

Amman, F. *et al.* (2014) 'TSSAR: TSS annotation regime for dRNA-seq data', *BMC bioinformatics*, 15, p. 89.

Ardui, S. *et al.* (2018) 'Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics', *Nucleic acids research*, 46(5), pp. 2159–2168.

Au, N. *et al.* (2005) 'Genetic composition of the Bacillus subtilis SOS system', *Journal of bacteriology*, 187(22), pp. 7655–7666.

Bagel, S. *et al.* (1999) 'Impact of gyrA and parC mutations on quinolone resistance, doubling time, and supercoiling degree of Escherichia coli', *Antimicrobial agents and chemotherapy*, 43(4), pp. 868–875.

Baharoglu, Z. and Mazel, D. (2014) 'SOS, the formidable strategy of bacteria against aggressions', *FEMS microbiology reviews*, 38(6), pp. 1126–1145.

Balázs, Z. *et al.* (2019) 'Template-switching artifacts resemble alternative polyadenylation', *BMC genomics*, 20(1), p. 824.

Baptista, L. C. *et al.* (2020) 'Crosstalk Between the Gut Microbiome and Bioactive Lipids: Therapeutic Targets in Cognitive Frailty', *Frontiers in nutrition*, 7, p. 17.

Baracchini, E. and Bremer, H. (1987) 'Determination of synthesis rate and lifetime of bacterial mRNAs', *Analytical biochemistry*, 167(2), pp. 245–260.

Bashiardes, S., Zilberman-Schapira, G. and Elinav, E. (2016) 'Use of Metatranscriptomics in Microbiome Research', *Bioinformatics and Biology Insights*, p. BBI.S34610. doi: 10.4137/bbi.s34610.

Bearson, B. L. and Brunelle, B. W. (2015) 'Fluoroquinolone induction of phage-mediated gene transfer in multidrug-resistant Salmonella', *International journal of antimicrobial agents*, 46(2), pp. 201–204.

Bhattacharyya, R. *et al.* (2017) 'Rapid Phenotypic Antibiotic Susceptibility Testing Through RNA Detection', *Open Forum Infectious Diseases*, pp. S33–S33. doi: 10.1093/ofid/ofx162.082.

Blázquez, J. *et al.* (2012) 'Antimicrobials as promoters of genetic variation', *Current Opinion in Microbiology*, pp. 561–569. doi: 10.1016/j.mib.2012.07.007.

Blevins, S. M. and Bronze, M. S. (2010) 'Robert Koch and the "golden age" of bacteriology', *International Journal of Infectious Diseases*, pp. e744–e751. doi: 10.1016/j.ijid.2009.12.003.

Bolyen, E. *et al.* (2019) 'Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2', *Nature biotechnology*, 37(8), pp. 852–857.

Bosco, N. and Noti, M. (2021) 'The aging gut microbiome and its impact on host immunity', *Genes and immunity*. doi: 10.1038/s41435-021-00126-8.

Boutard, M. *et al.* (2016) 'Global repositioning of transcription start sites in a plant-fermenting bacterium', *Nature Communications*. doi: 10.1038/ncomms13783.

Brandt, L. J. and Aroniadis, O. C. (2013) 'An overview of fecal microbiota transplantation: techniques, indications, and outcomes', *Gastrointestinal endoscopy*, 78(2), pp. 240–249.

Brar, G. A. and Weissman, J. S. (2015) 'Ribosome profiling reveals the what, when, where and how of protein synthesis', *Nature reviews. Molecular cell biology*, 16(11), pp. 651–664.

Byrne, A. *et al.* (2019) 'Realizing the potential of full-length transcriptome sequencing', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 374(1786), p. 20190097.

Cahana, I. and Iraqi, F. A. (2020) 'Impact of host genetics on gut microbiome: Take-home lessons from human and mouse studies', *Animal models and experimental medicine*, 3(3), pp. 229–236.

Callahan, B. J. *et al.* (2016) 'DADA2: High-resolution sample inference from Illumina amplicon data', *Nature Methods*, pp. 581–583. doi: 10.1038/nmeth.3869.

Callahan, B. J. *et al.* (2021) 'Ultra-accurate microbial amplicon sequencing with synthetic long reads', *Microbiome*, 9(1), p. 130.

Campoli-Richards, D. M. *et al.* (1988) 'Ciprofloxacin. A review of its antibacterial activity, pharmacokinetic properties and therapeutic use', *Drugs*, 35(4), pp. 373–447.

Canny, G. O. and McCormick, B. A. (2008) 'Bacteria in the Intestine, Helpful Residents or Enemies from Within?', *Infection and Immunity*, pp. 3360–3373. doi: 10.1128/iai.00187-08.

Caporaso, J. G. *et al.* (2010) 'QIIME allows analysis of high-throughput community sequencing data', *Nature methods*, 7(5), pp. 335–336.

Carding, S. *et al.* (2015) 'Dysbiosis of the gut microbiota in disease', *Microbial Ecology in Health & Disease*. doi: 10.3402/mehd.v26.26191.

Carter, R. and Drouin, G. (2009) 'Structural differentiation of the three eukaryotic RNA polymerases', *Genomics*, 94(6), pp. 388–396.

Caspi, R. *et al.* (2018) 'The MetaCyc database of metabolic pathways and enzymes', *Nucleic acids research*, 46(D1), pp. D633–D639.

Choi, H. H. and Cho, Y.-S. (2016) 'Fecal Microbiota Transplantation: Current Applications, Effectiveness, and Future Perspectives', *Clinical endoscopy*, 49(3), pp. 257–265.

Ciampi, M. S. and Sofia Ciampi, M. (2006) 'Rho-dependent terminators and transcription termination', *Microbiology*, pp. 2515–2528. doi: 10.1099/mic.0.28982-0.

Clarke, J. *et al.* (2009) 'Continuous base identification for single-molecule nanopore DNA sequencing', *Nature Nanotechnology*, pp. 265–270. doi: 10.1038/nnano.2009.12.

Cock, P. J. A. *et al.* (2009) 'Biopython: freely available Python tools for computational molecular biology and bioinformatics', *Bioinformatics* , 25(11), pp. 1422–1423.

Colgan, A. M., Cameron, A. D. S. and Kröger, C. (2017) 'If it transcribes, we can sequence it: mining the complexities of host–pathogen–environment interactions using RNA-seq', *Current Opinion in Microbiology*, pp. 37–46. doi: 10.1016/j.mib.2017.01.010.

Conley, Z. C. *et al.* (2018) 'Wicked: The untold story of ciprofloxacin', *PLoS pathogens*, 14(3), p. e1006805.

Consortium, T. H. M. P. and The Human Microbiome Project Consortium (2012a) 'A framework for human microbiome research', *Nature*, pp. 215–221. doi: 10.1038/nature11209.

Consortium, T. H. M. P. and The Human Microbiome Project Consortium (2012b) 'Structure, function and diversity of the healthy human microbiome', *Nature*, pp. 207–214. doi: 10.1038/nature11234.

Courcelle, J. *et al.* (2001) 'Comparative gene expression profiles following UV exposure in wild-type and SOS-deficient Escherichia coli', *Genetics*, 158(1), pp. 41–64.

Cox, M. M. (1998) 'A broadening view of recombinational DNA repair in bacteria', *Genes to cells: devoted to molecular & cellular mechanisms*, 3(2), pp. 65–78.

Crooks, G. E. *et al.* (2004) 'WebLogo: a sequence logo generator', *Genome research*, 14(6), pp. 1188–1190.

Dallo, S. F. and Weitao, T. (2010) 'Bacteria under SOS evolve anticancer phenotypes', *Infectious agents and cancer*, 5(1), p. 3.

Deamer, D., Akeson, M. and Branton, D. (2016) 'Three decades of nanopore sequencing', *Nature Biotechnology*, pp. 518–524. doi: 10.1038/nbt.3423.

DeGruttola, A. K. *et al.* (2016) 'Current Understanding of Dysbiosis in Disease in Human and Animal Models', *Inflammatory bowel diseases*, 22(5), pp. 1137–1150.

Delzenne, N. M. *et al.* (2011) 'Targeting gut microbiota in obesity: effects of prebiotics and probiotics', *Nature reviews. Endocrinology*, 7(11), pp. 639–646.

Dethlefsen, L. *et al.* (2008) 'The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing', *PLoS biology*, 6(11), p. e280.

De Weirdt, R. and Van de Wiele, T. (2015) 'Micromanagement in the gut: microenvironmental factors govern colon mucosal biofilm structure and functionality', *NPJ biofilms and microbiomes*, 1, p. 15026.

Drlica, K. *et al.* (2008) 'Quinolone-mediated bacterial death', *Antimicrobial agents and chemotherapy*, 52(2), pp. 385–392.

Drlica, K. and Zhao, X. (1997) 'DNA gyrase, topoisomerase IV, and the 4-quinolones', *Microbiology and molecular biology reviews : MMBR*, pp. 377–392. doi: 10.1128/.61.3.377-392.1997.

Emms, D. M. and Kelly, S. (2019) 'OrthoFinder: phylogenetic orthology inference for comparative genomics', *Genome biology*, 20(1), p. 238.

Eng, R. H. *et al.* (1991) 'Bactericidal effects of antibiotics on slowly growing and nongrowing bacteria', *Antimicrobial agents and chemotherapy*, 35(9), pp. 1824–1828.

Ermolaeva, M. D. *et al.* (2000) 'Prediction of transcription terminators in bacterial genomes 1 1Edited by F. E. Cohen', *Journal of Molecular Biology*, pp. 27–33. doi: 10.1006/jmbi.2000.3836.

Escobar-Zepeda, A., de León, A. V.-P. and Sanchez-Flores, A. (2015) 'The Road to Metagenomics: From Microbiology to DNA Sequencing Technologies and Bioinformatics', *Frontiers in Genetics*. doi: 10.3389/fgene.2015.00348.

Ettwiller, L. *et al.* (2016) 'A novel enrichment strategy reveals unprecedented number of novel transcription start sites at single base resolution in a model prokaryote and the gut microbiome', *BMC genomics*, 17, p. 199.

Fadrosh, D. W. *et al.* (2014) 'An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform', *Microbiome*, 2(1), p. 6.

Fan, Y. and Pedersen, O. (2021) 'Gut microbiota in human metabolic health and disease', *Nature reviews. Microbiology*, 19(1), pp. 55–71.

Filippo, C. D. *et al.* (2010) 'Impact of diet in shaping gut microbiota revealed by a comparative study

in children from Europe and rural Africa', *Proceedings of the National Academy of Sciences*, pp. 14691–14696. doi: 10.1073/pnas.1005963107.

Forster, S. C. *et al.* (2019) 'A human gut bacterial genome and culture collection for improved metagenomic analyses', *Nature biotechnology*, 37(2), pp. 186–192.

Fox, E. J. and Reid-Bayliss, K. S. (2014) 'Accuracy of Next Generation Sequencing Platforms', *Journal of Next Generation Sequencing & Applications*. doi: 10.4172/2469-9853.1000106.

Garalde, D. R. *et al.* (2018) 'Highly parallel direct RNA sequencing on an array of nanopores', *Nature methods*, 15(3), pp. 201–206.

Garrido-Cardenas, J. A. *et al.* (2017) 'DNA Sequencing Sensors: An Overview', *Sensors* , 17(3). doi: 10.3390/s17030588.

Ghaisas, S., Maher, J. and Kanthasamy, A. (2016) 'Gut microbiome in health and disease: Linking the microbiome-gut-brain axis and environmental factors in the pathogenesis of systemic and neurodegenerative diseases', *Pharmacology & therapeutics*, 158, pp. 52–62.

Glenn, T. C. (2011) 'Field guide to next-generation DNA sequencers', *Molecular ecology resources*, 11(5), pp. 759–769.

Goodrich, J. K. *et al.* (2014) 'Human genetics shape the gut microbiome', *Cell*, 159(4), pp. 789–799.

Goodwin, S., McPherson, J. D. and McCombie, W. R. (2016) 'Coming of age: ten years of next-generation sequencing technologies', *Nature reviews. Genetics*, 17(6), pp. 333–351.

Goris, J. *et al.* (2007) 'DNA-DNA hybridization values and their relationship to whole-genome sequence similarities', *International journal of systematic and evolutionary microbiology*, 57(Pt 1), pp. 81–91.

Graham, J. E. (2004) 'Sequence-specific Rho-RNA interactions in transcription termination', *Nucleic acids research*, 32(10), pp. 3093–3100.

Grylak-Mielnicka, A. *et al.* (2016) 'Transcription termination factor Rho: a hub linking diverse physiological processes in bacteria', *Microbiology*, 162(3), pp. 433–447.

'Gut Microbiome' (2019) in *Adult Short Bowel Syndrome*. Academic Press, pp. 45–54.

Hamed, S. M. *et al.* (2018) 'Multiple mechanisms contributing to ciprofloxacin resistance among Gram negative bacteria causing infections to cancer patients', *Scientific reports*, 8(1), p. 12268.

Handelsman, J. *et al.* (1998) 'Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products', *Chemistry & biology*, 5(10), pp. R245–9.

Handelsman, J. (2005) 'Metagenomics: Application of Genomics to Uncultured Microorganisms', *Microbiology and Molecular Biology Reviews*, pp. 195–195. doi: 10.1128/mmbr.69.1.195.2005.

Hawley, D. K. and McClure, W. R. (1983) 'Compilation and analysis of Escherichia coli promoter DNA sequences', *Nucleic acids research*, 11(8), pp. 2237–2255.

Healey, G. R. *et al.* (2017) 'Interindividual variability in gut microbiota and host response to dietary interventions', *Nutrition reviews*, 75(12), pp. 1059–1080.

Heather, J. M. and Chain, B. (2016) 'The sequence of sequencers: The history of sequencing DNA', *Genomics*, 107(1), pp. 1–8.

Hert, D. G., Fredlake, C. P. and Barron, A. E. (2008) 'Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods', *Electrophoresis*, 29(23), pp. 4618–4626.

Hibberd, M. C. *et al.* (2017) 'The effects of micronutrient deficiencies on bacterial species from the human gut microbiota', *Science translational medicine*, 9(390). doi: 10.1126/scitranslmed.aal4069.

Holley, R. W. *et al.* (1965) 'STRUCTURE OF A RIBONUCLEIC ACID', *Science*, 147(3664), pp. 1462–1465.

Hong, J. *et al.* (2008) 'Bioseparation of recombinant cellulose-binding module-proteins by affinity adsorption on an ultra-high-capacity cellulosic adsorbent', *Analytica Chimica Acta*, pp. 193–199. doi: 10.1016/j.aca.2008.05.041.

Hooper, D. C. and Jacoby, G. A. (2016) 'Topoisomerase Inhibitors: Fluoroquinolone Mechanisms of Action and Resistance', *Cold Spring Harbor perspectives in medicine*, 6(9). doi: 10.1101/cshperspect.a025320.

Hooper, L. V. and Gordon, J. I. (2001) 'Commensal host-bacterial relationships in the gut', *Science*, 292(5519), pp. 1115–1118.

Hör, J., Gorski, S. A. and Vogel, J. (2018) 'Bacterial RNA Biology on a Genome Scale', *Molecular Cell*, pp. 785–799. doi: 10.1016/j.molcel.2017.12.023.

Integrative HMP (iHMP) Research Network Consortium (2014) 'The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease', *Cell host & microbe*, 16(3), pp. 276–289.

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

Jain, C. *et al.* (2018) 'High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries', *Nature communications*, 9(1), p. 5114.

Jain, M. *et al.* (2018) 'Nanopore sequencing and assembly of a human genome with ultra-long reads', *Nature biotechnology*, 36(4), pp. 338–345.

Jain, S., Gupta, R. and Sen, R. (2019) 'Rho-dependent transcription termination in bacteria recycles RNA polymerases stalled at DNA lesions', *Nature communications*, 10(1), p. 1207.

Johnson, E. A., Madia, A. and Demain, A. L. (1981) 'Chemically Defined Minimal Medium for Growth of the Anaerobic Cellulolytic Thermophile Clostridium thermocellum', *Applied and environmental microbiology*, 41(4), pp. 1060–1062.

Johnson, J. S. *et al.* (2019) 'Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis', *Nature communications*, 10(1), p. 5029.

Jovel, J. *et al.* (2016) 'Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics', *Frontiers in Microbiology*. doi: 10.3389/fmicb.2016.00459.

Ju, X., Li, D. and Liu, S. (2019) 'Full-length RNA profiling reveals pervasive bidirectional transcription terminators in bacteria', *Nature microbiology*, 4(11), pp. 1907–1918.

Kanehisa, M. *et al.* (2017) 'KEGG: new perspectives on genomes, pathways, diseases and drugs', *Nucleic acids research*, 45(D1), pp. D353–D361.

Kastl, A. J., Jr *et al.* (2020) 'The Structure and Function of the Human Small Intestinal Microbiota: Current Understanding and Future Directions', *Cellular and molecular gastroenterology and hepatology*, 9(1), pp. 33–45.

Khan, T. J. *et al.* (2019) 'Association of gut dysbiosis with intestinal metabolites in response to antibiotic treatment', *Human Microbiome Journal*, p. 100054. doi: 10.1016/j.humic.2018.11.004.

Khazaei, T. *et al.* (2018) 'RNA markers enable phenotypic test of antibiotic susceptibility in Neisseria gonorrhoeae after 10 minutes of ciprofloxacin exposure', *Scientific reports*, 8(1), p. 11606.

Kho, Z. Y. and Lal, S. K. (2018) 'The Human Gut Microbiome – A Potential Controller of Wellness and Disease', *Frontiers in Microbiology*. doi: 10.3389/fmicb.2018.01835.

Kim, D. *et al.* (2012) 'Comparative analysis of regulatory elements between Escherichia coli and Klebsiella pneumoniae by genome-wide transcription start site profiling', *PLoS genetics*, 8(8), p. e1002867.

Kim, K. O. and Gluck, M. (2019) 'Fecal Microbiota Transplantation: An Update on Clinical Practice', *Clinical endoscopy*, 52(2), pp. 137–143.

Kim, R. *et al.* (2019) 'An in vitro intestinal platform with a self-sustaining oxygen gradient to study the human gut/microbiome interface', *Biofabrication*, p. 015006. doi: 10.1088/1758-5090/ab446e.

Kim, Y. S. *et al.* (2020) 'Sex Differences in Gut Microbiota', *The world journal of men's health*, 38(1), pp. 48–60.

King, D. E., Malone, R. and Lilley, S. H. (2000) 'New classification and update on the quinolone antibiotics', *American family physician*, 61(9), pp. 2741–2748.

Kingsford, C. L., Ayanbule, K. and Salzberg, S. L. (2007) 'Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake', *Genome*

biology, 8(2), p. R22.

Klindworth, A. *et al.* (2013) 'Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies', *Nucleic acids research*, 41(1), p. e1.

Knudsen, B. E. *et al.* (2016) 'Impact of Sample Type and DNA Isolation Procedure on Genomic Inference of Microbiome Composition', *mSystems*, 1(5). doi: 10.1128/mSystems.00095-16.

Konstantinos T. Konstantinidis, J. M. T. (2005) 'Genomic insights that advance the species definition for prokaryotes', *Proceedings of the National Academy of Sciences of the United States of America*, 102(7), p. 2567.

Koren, S. *et al.* (2017) 'Canu: scalable and accurate long-read assembly via adaptive -mer weighting and repeat separation', *Genome research*, 27(5), pp. 722–736.

Kozich, J. J. *et al.* (2013) 'Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform', *Applied and environmental microbiology*, 79(17), pp. 5112–5120.

Kulski, J. K. (2016) 'Next-Generation Sequencing — An Overview of the History, Tools, and "Omic" Applications', *Next Generation Sequencing - Advances, Applications and Challenges*. doi: 10.5772/61964.

Kumar, M. *et al.* (2016) 'Human gut microbiota and healthy aging: Recent developments and future prospective', *Nutrition and healthy aging*, 4(1), pp. 3–16.

Lander, E. S. *et al.* (2001) 'Initial sequencing and analysis of the human genome', *Nature*, 409(6822), pp. 860–921.

Langmead, B. and Salzberg, S. L. (2012) 'Fast gapped-read alignment with Bowtie 2', *Nature methods*, 9(4), pp. 357–359.

Leavis, H. L. *et al.* (2006) 'High-level ciprofloxacin resistance from point mutations in gyrA and parC confined to global hospital-adapted clonal lineage CC17 of Enterococcus faecium', *Journal of clinical microbiology*, 44(3), pp. 1059–1064.

Leclercq, S. *et al.* (2014) 'Intestinal permeability, gut-bacterial dysbiosis, and behavioral markers of alcohol-dependence severity', *Proceedings of the National Academy of Sciences of the United States of America*, 111(42), pp. E4485–93.

Ledder, O. (2019) 'Antibiotics in inflammatory bowel diseases: do we know what we're doing?', *Translational Pediatrics*, pp. 42–55. doi: 10.21037/tp.2018.11.02.

Lee, C. M. *et al.* (2019) 'Topoisomerase III Acts at the Replication Fork To Remove Precatenanes', *Journal of Bacteriology*. doi: 10.1128/jb.00563-18.

Letunic, I. and Bork, P. (2021) 'Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree

display and annotation', *Nucleic acids research*, 49(W1), pp. W293–W296.

Levene, M. J. *et al.* (2003) 'Zero-mode waveguides for single-molecule analysis at high concentrations', *Science*, 299(5607), pp. 682–686.

Ley, R. E. *et al.* (2008) 'Worlds within worlds: evolution of the vertebrate gut microbiota', *Nature reviews. Microbiology*, 6(10), pp. 776–788.

Liao, Y., Smyth, G. K. and Shi, W. (2014) 'featureCounts: an efficient general purpose program for assigning sequence reads to genomic features', *Bioinformatics* , 30(7), pp. 923–930.

Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics* , 25(16), pp. 2078–2079.

Li, H. (2013) 'Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM', *arXiv*. Available at: https://arxiv.org/abs/1303.3997.

Li, H. (2018) 'Minimap2: pairwise alignment for nucleotide sequences', *Bioinformatics* , 34(18), pp. 3094–3100.

Lima, L. *et al.* (2020) 'Comparative assessment of long-read error correction software applied to Nanopore RNA-sequencing data', *Briefings in bioinformatics*, 21(4), pp. 1164–1181.

Lim, M. Y. *et al.* (2018) 'Comparison of DNA extraction methods for human gut microbial community profiling', *Systematic and Applied Microbiology*, pp. 151–157. doi: 10.1016/j.syapm.2017.11.008.

Li, R. *et al.* (2015) 'Illumina Synthetic Long Read Sequencing Allows Recovery of Missing Sequences even in the "Finished" C. elegans Genome', *Scientific Reports*. doi: 10.1038/srep10814.

Liu, S. *et al.* (2021) 'Targeted transcriptome analysis using synthetic long read sequencing uncovers isoform reprograming in the progression of colon cancer', *Communications biology*, 4(1), p. 506.

Li, Z. *et al.* (1997) 'The traE gene of plasmid RP4 encodes a homologue of Escherichia coli DNA topoisomerase III', *The Journal of biological chemistry*, 272(31), pp. 19582–19587.

LoopGenomics — Overview (2020). Available at: https://www.loopgenomics.com/how-it-works (Accessed: 19 July 2021).

Lopatkin, A. J. *et al.* (2019) 'Bacterial metabolic state more accurately predicts antibiotic lethality than growth rate', *Nature microbiology*, 4(12), pp. 2109–2117.

Love, M. I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome biology*, 15(12), p. 550.

Lozupone, C. A. *et al.* (2013) 'Meta-analyses of studies of the human microbiota', *Genome Research*, pp. 1704–1714. doi: 10.1101/gr.151803.112.

Lu, Y. *et al.* (2016) 'Next Generation Sequencing in Aquatic Models', *Next Generation Sequencing* -

Advances, Applications and Challenges. doi: 10.5772/61657.

Mabwi, H. A. *et al.* (2021) 'Synthetic gut microbiome: Advances and challenges', *Computational and structural biotechnology journal*, 19, pp. 363–371.

Madsen, K. *et al.* (2001) 'Probiotic bacteria enhance murine and human intestinal epithelial barrier function', *Gastroenterology*, 121(3), pp. 580–591.

Magi, A. *et al.* (2018) 'Nanopore sequencing data analysis: state of the art, applications and challenges', *Briefings in bioinformatics*, 19(6), pp. 1256–1272.

Maguire, S., Lohman, G. J. S. and Guan, S. (2020) 'A low-bias and sensitive small RNA library preparation method using randomized splint ligation', *Nucleic acids research*, 48(14), p. e80.

Mahnic, A. *et al.* (2020) 'Distinct Types of Gut Microbiota Dysbiosis in Hospitalized Gastroenterological Patients Are Disease Non-related and Characterized With the Predominance of Either Enterobacteriaceae or Enterococcus', *Frontiers in Microbiology*. doi: 10.3389/fmicb.2020.00120.

Maier, L. *et al.* (2020) 'Dissecting the collateral damage of antibiotics on gut microbes', *bioRxiv*. doi: 10.1101/2020.01.09.893560.

Maloy, S. and Schaechter, M. (2006) 'The era of microbiology: a golden phoenix', *International microbiology: the official journal of the Spanish Society for Microbiology*, 9(1), pp. 1–7.

Mastropaolo, M. D., Thorson, M. L. and Stevens, A. M. (2009) 'Comparison of Bacteroides thetaiotaomicron and Escherichia coli 16S rRNA gene expression signals', *Microbiology*, 155(Pt 8), pp. 2683–2693.

Maxam, A. M. and Gilbert, W. (1977) 'A new method for sequencing DNA', *Proceedings of the National Academy of Sciences*, pp. 560–564. doi: 10.1073/pnas.74.2.560.

McCoy, R. C. *et al.* (2014) 'Illumina TruSeq Synthetic Long-Reads Empower De Novo Assembly and Resolve Complex, Highly-Repetitive Transposable Elements', *PLoS ONE*, p. e106689. doi: 10.1371/journal.pone.0106689.

McDonald, D. *et al.* (2012) 'An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea', *The ISME journal*, 6(3), pp. 610–618.

McMurdie, P. J. and Holmes, S. (2013) 'phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data', *PloS one*, 8(4), p. e61217.

Médigue, C. *et al.* (2019) 'MicroScope—an integrated resource for community expertise of gene functions and comparative analysis of microbial genomic and metabolic data', *Briefings in Bioinformatics*, pp. 1071–1084. doi: 10.1093/bib/bbx113.

Mestan, K. K. *et al.* (2011) 'Genomic sequencing in clinical trials', *Journal of Translational Medicine*.

doi: 10.1186/1479-5876-9-222.

Metzker, M. L. (2010) 'Sequencing technologies - the next generation', *Nature reviews. Genetics*, 11(1), pp. 31–46.

Modi, S. R., Collins, J. J. and Relman, D. A. (2014) 'Antibiotics and the gut microbiota', *The Journal of clinical investigation*, 124(10), pp. 4212–4218.

Moll, I. *et al.* (2002) 'Leaderless mRNAs in bacteria: surprises in ribosomal recruitment and translational control', *Molecular microbiology*, 43(1), pp. 239–246.

Moody, D. E. (2001) 'Genomics techniques: An overview of methods for the study of gene expression', *Journal of Animal Science*, p. E128. doi: 10.2527/jas2001.79e-supple128x.

Morey, M. *et al.* (2013) 'A glimpse into past, present, and future DNA sequencing', *Molecular genetics and metabolism*, 110(1-2), pp. 3–24.

Nagalakshmi, U. *et al.* (2008) 'The transcriptional landscape of the yeast genome defined by RNA sequencing', *Science*, 320(5881), pp. 1344–1349.

Ngara, T. R. and Zhang, H. (2018) 'Recent Advances in Function-based Metagenomic Screening', *Genomics, proteomics & bioinformatics*, 16(6), pp. 405–415.

Nguyen, T. G. *et al.* (2020) 'The Impact of Leadered and Leaderless Gene Structures on Translation Efficiency, Transcript Stability, and Predicted Transcription Rates in Mycobacterium smegmatis', *Journal of Bacteriology*. doi: 10.1128/jb.00746-19.

Nygaard, A. B. *et al.* (2020) 'A preliminary study on the potential of Nanopore MinION and Illumina MiSeq 16S rRNA gene sequencing to characterize building-dust microbiomes', *Scientific Reports*. doi: 10.1038/s41598-020-59771-0.

O'Donnell, S. M. and Janssen, G. R. (2001) 'The initiation codon affects ribosome binding and translational efficiency in Escherichia coli of cI mRNA with or without the 5′ untranslated leader', *Journal of bacteriology*, 183(4), pp. 1277–1283.

O'Dwyer, D. N., Dickson, R. P. and Moore, B. B. (2016) 'The Lung Microbiome, Immunity, and the Pathogenesis of Chronic Lung Disease', *The Journal of Immunology*, pp. 4839–4847. doi: 10.4049/jimmunol.1600279.

Oliva, G., Sahr, T. and Buchrieser, C. (2015) 'Small RNAs, 5′ UTR elements and RNA-binding proteins in intracellular bacteria: impact on metabolism and virulence', *FEMS Microbiology Reviews*, pp. 331–349. doi: 10.1093/femsre/fuv022.

Osman, M.-A. *et al.* (2018) '16S rRNA Gene Sequencing for Deciphering the Colorectal Cancer Gut Microbiome: Current Protocols and Workflows', *Frontiers in microbiology*, 9, p. 767.

Ouldali, H. *et al.* (2020) 'Electrical recognition of the twenty proteinogenic amino acids using an

aerolysin nanopore', *Nature biotechnology*, 38(2), pp. 176–181.

Pace, N. R. *et al.* (1986) 'The Analysis of Natural Microbial Populations by Ribosomal RNA Sequences', *Advances in Microbial Ecology*, pp. 1–55. doi: 10.1007/978-1-4757-0611-6_1.

Palmer, C. *et al.* (2007) 'Development of the Human Infant Intestinal Microbiota', *PLoS Biology*, p. e177. doi: 10.1371/journal.pbio.0050177.

Passalacqua, K. D. *et al.* (2009) 'Structure and complexity of a bacterial transcriptome', *Journal of bacteriology*, 191(10), pp. 3203–3211.

Patrick, D. M. *et al.* (2020) 'Decreasing antibiotic use, the gut microbiota, and asthma incidence in children: evidence from population-based and prospective cohort studies', *The Lancet. Respiratory medicine*, 8(11), pp. 1094–1105.

Payne, A. *et al.* (2019) 'BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files', *Bioinformatics* , 35(13), pp. 2193–2198.

Penadés, J. R. *et al.* (2015) 'Bacteriophage-mediated spread of bacterial virulence genes', *Current opinion in microbiology*, 23, pp. 171–178.

Peters, J. M., Vangeloff, A. D. and Landick, R. (2011) 'Bacterial transcription terminators: the RNA 3'-end chronicles', *Journal of molecular biology*, 412(5), pp. 793–813.

Pinto, A. C. *et al.* (2011) 'Application of RNA-seq to reveal the transcript profile in bacteria', *Genetics and molecular research: GMR*, 10(3), pp. 1707–1718.

Podlesek, Z. and Žgur Bertok, D. (2020) 'The DNA Damage Inducible SOS Response Is a Key Player in the Generation of Bacterial Persister Cells and Population Wide Tolerance', *Frontiers in microbiology*, 11, p. 1785.

Qin, J. *et al.* (2010) 'A human gut microbial gene catalogue established by metagenomic sequencing', *Nature*, 464(7285), pp. 59–65.

Quinlan, A. R. and Hall, I. M. (2010) 'BEDTools: a flexible suite of utilities for comparing genomic features', *Bioinformatics* , 26(6), pp. 841–842.

Rausch, P. *et al.* (2019) 'Comparative analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of animal metaorganisms', *Microbiome*, 7(1), p. 133.

Reardon, S. (2021) 'A complete human genome sequence is close: how scientists filled in the gaps', *Nature*, 594(7862), pp. 158–159.

Reinert, K. *et al.* (2015) 'Alignment of Next-Generation Sequencing Reads', *Annual Review of Genomics and Human Genetics*, pp. 133–151. doi: 10.1146/annurev-genom-090413-025358.

Reuter, J. A., Spacek, D. V. and Snyder, M. P. (2015) 'High-throughput sequencing technologies', *Molecular cell*, 58(4), pp. 586–597.

Reygaert, W. C. *et al.* (2018) 'An overview of the antimicrobial resistance mechanisms of bacteria', *AIMS Microbiology*, pp. 482–501. doi: 10.3934/microbiol.2018.3.482.

R Foundation for Statistical Computing, Vienna, Austria. (2020) *R: A language and environment for statistical computing*. Available at: https://www.R-project.org (Accessed: 19 July 2021).

Roberts, J. W. (2019) 'Mechanisms of Bacterial Transcription Termination', *Journal of molecular biology*, 431(20), pp. 4030–4039.

Roberts, R. J. *et al.* (2015) 'REBASE--a database for DNA restriction and modification: enzymes, genes and genomes', *Nucleic acids research*, 43(Database issue), pp. D298–9.

Rodríguez, J. M. *et al.* (2015) 'The composition of the gut microbiota throughout life, with an emphasis on early life', *Microbial ecology in health and disease*, 26, p. 26050.

Romano, K. A. *et al.* (2015) 'Intestinal microbiota composition modulates choline bioavailability from diet and accumulation of the proatherogenic metabolite trimethylamine-N-oxide', *mBio*, 6(2), p. e02481.

Ruas-Madiedo, P. *et al.* (2008) 'Mucin degradation by Bifidobacterium strains isolated from the human intestinal microbiota', *Applied and environmental microbiology*, 74(6), pp. 1936–1940.

Sahlin, K. and Medvedev, P. (2021) 'Author Correction: Error correction enables use of Oxford Nanopore technology for reference-free transcriptome analysis', *Nature communications*, 12(1), p. 992.

Salmela, L. and Rivals, E. (2014) 'LoRDEC: accurate and efficient long read error correction', *Bioinformatics* , 30(24), pp. 3506–3514.

Sanger, F. *et al.* (1977) 'Nucleotide sequence of bacteriophage phi X174 DNA', *Nature*, 265(5596), pp. 687–695.

Sanger, F., Nicklen, S. and Coulson, A. R. (1977) 'DNA sequencing with chain-terminating inhibitors', *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), pp. 5463–5467.

Sangurdekar, D. P., Srienc, F. and Khodursky, A. B. (2006) 'A classification based framework for quantitative description of large-scale microarray data', *Genome biology*, 7(4), p. R32.

Santajit, S. and Indrawattana, N. (2016) 'Mechanisms of Antimicrobial Resistance in ESKAPE Pathogens', *BioMed Research International*, pp. 1–8. doi: 10.1155/2016/2475067.

Schmitz, F. J. *et al.* (1998) 'Characterization of grlA, grlB, gyrA, and gyrB mutations in 116 unrelated isolates of Staphylococcus aureus and effects of mutations on ciprofloxacin MIC', *Antimicrobial agents and chemotherapy*, 42(5), pp. 1249–1252.

Schoenberg, D. R. (2007) 'The end defines the means in bacterial mRNA decay', *Nature Chemical*

Biology, pp. 535–536. doi: 10.1038/nchembio0907-535.

Schrader, J. M. *et al.* (2014) 'The coding and noncoding architecture of the Caulobacter crescentus genome', *PLoS genetics*, 10(7), p. e1004463.

Scotti, E. *et al.* (2017) 'Exploring the microbiome in health and disease', *Toxicology Research and Application*, p. 239784731774188. doi: 10.1177/2397847317741884.

Sender, R., Fuchs, S. and Milo, R. (2016) 'Revised Estimates for the Number of Human and Bacteria Cells in the Body', *PLoS biology*, 14(8), p. e1002533.

Sessegolo, C. *et al.* (2019) 'Transcriptome profiling of mouse samples using nanopore sequencing of cDNA and RNA molecules', *Scientific reports*, 9(1), p. 14908.

Sharma, C. M. *et al.* (2010) 'The primary transcriptome of the major human pathogen Helicobacter pylori', *Nature*, 464(7286), pp. 250–255.

Sharma, C. M. and Vogel, J. (2014) 'Differential RNA-seq: the approach behind and the biological insight gained', *Current opinion in microbiology*, 19, pp. 97–105.

Shendure, J. *et al.* (2017) 'DNA sequencing at 40: past, present and future', *Nature*, pp. 345–353. doi: 10.1038/nature24286.

Shendure, J. and Ji, H. (2008) 'Next-generation DNA sequencing', *Nature biotechnology*, 26(10), pp. 1135–1145.

Shen, W. *et al.* (2016) 'SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation', *PloS one*, 11(10), p. e0163962.

Simmons, L. A. *et al.* (2008) 'The SOS Regulatory Network', *EcoSal Plus*, 2008. doi: 10.1128/ecosalplus.5.4.3.

Sommer, M. O. A. *et al.* (2017) 'Prediction of antibiotic resistance: time for a new preclinical paradigm?', *Nature reviews. Microbiology*, 15(11), pp. 689–696.

Srivastava, S., Ghosh, N. and Pal, G. (2013) 'Metagenomics: Mining Environmental Genomes', *Biotechnology for Environmental Management and Resource Recovery*, pp. 161–189. doi: 10.1007/978-81-322-0876-1_10.

Staden, R. (1979) 'A strategy of DNA sequencing employing computer programs', *Nucleic acids research*, 6(7), pp. 2601–2610.

Stapleton, J. A. *et al.* (2016) 'Haplotype-Phased Synthetic Long Reads from Short-Read Sequencing', *PloS one*, 11(1), p. e0147229.

Stark, R., Grzelak, M. and Hadfield, J. (2019) 'RNA sequencing: the teenage years', *Nature Reviews Genetics*, pp. 631–656. doi: 10.1038/s41576-019-0150-2.

Stein, J. L. *et al.* (1996) 'Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon', *Journal of bacteriology*, pp. 591–599. doi: 10.1128/jb.178.3.591-599.1996.

Swerdlow, H. and Gesteland, R. (1990) 'Capillary gel electrophoresis for rapid, high resolution DNA sequencing', *Nucleic Acids Research*, pp. 1415–1419. doi: 10.1093/nar/18.6.1415.

Tamaki, S., Sato, T. and Matsuhashi, M. (1971) 'Role of lipopolysaccharides in antibiotic resistance and bacteriophage adsorption of Escherichia coli K-12', *Journal of bacteriology*, 105(3), pp. 968–975.

Tap, J. *et al.* (2009) 'Towards the human intestinal microbiota phylogenetic core', *Environmental microbiology*, 11(10), pp. 2574–2584.

Trasande, L. *et al.* (2013) 'Infant antibiotic exposures and early-life body mass', *International journal of obesity* , 37(1), pp. 16–23.

Tyson, G. W. *et al.* (2004) 'Community structure and metabolism through reconstruction of microbial genomes from the environment', *Nature*, pp. 37–43. doi: 10.1038/nature02340.

Ubeda, C. *et al.* (2010) 'Vancomycin-resistant Enterococcus domination of intestinal microbiota is enabled by antibiotic treatment in mice and precedes bloodstream invasion in humans', *The Journal of clinical investigation*, 120(12), pp. 4332–4341.

Valat, C. *et al.* (2020) 'Overall changes in the transcriptome of Escherichia coli O26:H11 induced by a subinhibitory concentration of ciprofloxacin', *Journal of applied microbiology*, 129(6), pp. 1577–1588.

Vallenet, D. *et al.* (2020) 'MicroScope: an integrated platform for the annotation and exploration of microbial gene functions through genomic, pangenomic and metabolic comparative analysis', *Nucleic acids research*, 48(D1), pp. D579–D589.

Venter, J. C. *et al.* (2004) 'Environmental genome shotgun sequencing of the Sargasso Sea', *Science*, 304(5667), pp. 66–74.

Wagner, J. *et al.* (2016) 'Evaluation of PacBio sequencing for full-length bacterial 16S rRNA gene classification', *BMC microbiology*, 16(1), p. 274.

Wang, Q. *et al.* (2007) 'Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy', *Applied and Environmental Microbiology*, pp. 5261–5267. doi: 10.1128/aem.00062-07.

Wang, X. and Wood, T. K. (2016) 'Cryptic prophages as targets for drug development', *Drug resistance updates: reviews and commentaries in antimicrobial and anticancer chemotherapy*, 27, pp. 30–38.

Wang, Y., MacKenzie, K. D. and White, A. P. (2015) 'An empirical strategy to detect bacterial transcript structure from directional RNA-seq transcriptome data', *BMC genomics*, 16, p. 359.

Wang, Z., Gerstein, M. and Snyder, M. (2009) 'RNA-Seq: a revolutionary tool for transcriptomics',

Nature reviews. Genetics, 10(1), pp. 57–63.

Watson, M. and Warr, A. (2019) 'Errors in long-read assemblies can critically affect protein prediction', *Nature biotechnology*, pp. 124–126.

Wenger, A. M. *et al.* (2019) 'Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome', *Nature biotechnology*, 37(10), pp. 1155–1162.

Wentzell, L. M. and Maxwell, A. (2000) 'The complex of DNA gyrase and quinolone drugs on DNA forms a barrier to the T7 DNA polymerase replication complex', *Journal of molecular biology*, 304(5), pp. 779–791.

Westermann, A. J., Barquist, L. and Vogel, J. (2017) 'Resolving host–pathogen interactions by dual RNA-seq', *PLOS Pathogens*, p. e1006033. doi: 10.1371/journal.ppat.1006033.

Wilkinson, L. (2011) 'ggplot2: Elegant Graphics for Data Analysis by WICKHAM, H', *Biometrics*, pp. 678–679. doi: 10.1111/j.1541-0420.2011.01616.x.

Woese, C. R. *et al.* (1985) 'A phylogenetic definition of the major eubacterial taxa', *Systematic and applied microbiology*, 6, pp. 143–151.

Woese, C. R. and Fox, G. E. (1977) 'Phylogenetic structure of the prokaryotic domain: the primary kingdoms', *Proceedings of the National Academy of Sciences of the United States of America*, 74(11), pp. 5088–5090.

Woese, C. R., Kandler, O. and Wheelis, M. L. (1990) 'Towards a natural system of organisms: proposal for the domains Archaea, Bacteria and Eucarya', *Trends in Genetics*, 6, p. 281.

Woo, P. C. Y. *et al.* (2008) 'Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories', *Clinical microbiology and infection: the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, 14(10), pp. 908–934.

Wu, N. C. *et al.* (2014) 'HIV-1 quasispecies delineation by tag linkage deep sequencing', *PloS one*, 9(5), p. e97505.

Yan, B. *et al.* (2018) 'SMRT-Cappable-seq reveals complex operon variants in bacteria'. doi: 10.1101/262964.

Yan, B. *et al.* (2021) 'ReCappable Seq: Comprehensive Determination of Transcription Start Sites derived from all RNA polymerases', *bioRxiv*. doi: 10.1101/696559.

Yang, B., Wang, Y. and Qian, P.-Y. (2016) 'Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis', *BMC bioinformatics*, 17, p. 135.

Yilmaz, P. *et al.* (2014) 'The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks', *Nucleic Acids Research*, pp. D643–D648. doi: 10.1093/nar/gkt1209.

Yoon, S.-H. *et al.* (2017) 'A large-scale evaluation of algorithms to calculate average nucleotide identity', *Antonie van Leeuwenhoek*, 110(10), pp. 1281–1286.

Yuan, S. *et al.* (2012) 'Evaluation of Methods for the Extraction and Purification of DNA from the Human Microbiome', *PLoS ONE*, p. e33865. doi: 10.1371/journal.pone.0033865.

Yurtsev, E. A. *et al.* (2013) 'Bacterial cheating drives the population dynamics of cooperative antibiotic resistance plasmids', *Molecular systems biology*, 9, p. 683.

Zankari, E. *et al.* (2012) 'Identification of acquired antimicrobial resistance genes', *Journal of Antimicrobial Chemotherapy*, pp. 2640–2644. doi: 10.1093/jac/dks261.

Zhang, H., Jain, C. and Aluru, S. (2020) 'A comprehensive evaluation of long read error correction methods', *BMC Genomics*. doi: 10.1101/519330.

Zhang, Y. *et al.* (2014) 'Effect of various antibiotics on modulation of intestinal microbiota and bile acid profile in mice', *Toxicology and applied pharmacology*, 277(2), pp. 138–145.

Zhao, L. *et al.* (2019) 'Analysis of Transcriptome and Epitranscriptome in Plants Using PacBio Iso-Seq and Nanopore-Based Direct RNA Sequencing', *Frontiers in genetics*, 10, p. 253.

Zheng, X. *et al.* (2011) 'Leaderless genes in bacteria: clue to the evolution of translation initiation mechanisms in prokaryotes', *BMC genomics*, 12, p. 361.

Zhernakova, A. *et al.* (2016) 'Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity', *Science*, 352(6285), pp. 565–569.

APPENDIX

A. Appendix from Chapter I

Nucleic Acids Research, 2021 **1**
https://doi.org/10.1093/nar/gkab705

Rapid identification of methylase specificity (RIMS-seq) jointly identifies methylated motifs and generates shotgun sequencing of bacterial genomes

Chloé Baum[1,2], Yu-Cheng Lin[1], Alexey Fomenkov[1], Brian P. Anton[1], Lixin Chen[1], Bo Yan[1], Thomas C. Evans, Jr[1], Richard J. Roberts[1], Andrew C. Tolonen[2] and Laurence Ettwiller[1,*]

[1]New England Biolabs, Inc. 240 County Road Ipswich, MA 01938, USA and [2]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91000 Évry, France

Received April 08, 2021; Revised July 29, 2021; Editorial Decision July 29, 2021; Accepted August 16, 2021

ABSTRACT

DNA methylation is widespread amongst eukaryotes and prokaryotes to modulate gene expression and confer viral resistance. 5-Methylcytosine (m5C) methylation has been described in genomes of a large fraction of bacterial species as part of restriction-modification systems, each composed of a methyltransferase and cognate restriction enzyme. Methylases are site-specific and target sequences vary across organisms. High-throughput methods, such as bisulfite-sequencing can identify m5C at base resolution but require specialized library preparations and single molecule, real-time (SMRT) sequencing usually misses m5C. Here, we present a new method called RIMS-seq (rapid identification of methylase specificity) to simultaneously sequence bacterial genomes and determine m5C methylase specificities using a simple experimental protocol that closely resembles the DNA-seq protocol for Illumina. Importantly, the resulting sequencing quality is identical to DNA-seq, enabling RIMS-seq to substitute standard sequencing of bacterial genomes. Applied to bacteria and synthetic mixed communities, RIMS-seq reveals new methylase specificities, supporting routine study of m5C methylation while sequencing new genomes.

INTRODUCTION

DNA modifications catalysed by DNA methyltransferases are considered to be the most abundant form of epigenetic modification in genomes of both prokaryotes and eukaryotes. In prokaryotes, DNA methylation has been mainly described as part of the sequence-specific restriction modification system (RM), a bacterial immune system to resist invasion of foreign DNA (1). As such, profiling methylation patterns gives insight into the selective pressures driving evolution of their genomes.

Around 90% of bacterial genomes contain at least one of the three common forms of DNA methylation: 5-methylcytosine (m5C), N4-methylcytosine (m4C) and N6-methyladenine (m6A)) (2,3). Contrary to eukaryotes where the position of the m5C methylation is variable and subject to epigenetic states, bacterial methylations tend to be constitutively present at specific sites across the genome. These sites are defined by the methylase specificity and, in the case of RM systems, tend to be fully methylated to avoid cuts by the cognate restriction enzyme. The methylase recognition specificities typically vary from four to eight nucleotides and are often, but not always, palindromic (4).

PacBio single molecule, real-time (SMRT) sequencing has been instrumental in the identification of methylase specificity largely because, in addition to providing long read sequencing of bacterial genomes, m6A and m4C can easily be detected using the characteristic interpulse duration (IPD) of those modified bases (5). Thus, a single run on PacBio allows for both the sequencing and assembly of unknown bacterial genomes and the determination of m6A and m4C methylase specificities. However, because the signal associated with m5C bases is weaker than for m6A or m4C, the IPD ratio of m5C is very similar to the IPD of unmodified cytosine. Thus, PacBio sequencing misses the m5C methylase activities (2) unless the 5-methylcytosine detection is enhanced by treating the library with Ten-eleven translocation enzyme (6). A recent study uses a holistic kinetic model to identify m5C using PacBio reads (7). Nonetheless, methylation can only be identified in CpG context, restricting the use of this approach to organisms such as human, for which methylation is almost exclusively in CpG sites.

Consequently, the identification of m5C requires specialized methods such as bisulfite sequencing, enzyme-based

*To whom correspondence should be addressed. Tel: +1 978 998 7910; Fax: +1 978 921 1350; Email: ettwiller@neb.com

techniques such as EM-seq (8) or hybrid techniques such as TAPS-seq (9). Recently, MFRE-Seq has been developed to identify m5C methylase specificities in bacteria (10). MFRE-Seq uses a modification-dependent endonuclease that generates a double-stranded DNA break at methylated sites, allowing the identification of m5C for the subset of sites conforming to the recognition sites of the MFRE enzymes. Unlike PacBio sequencing, these specialized methods do not provide the dual original sequence and methylation readouts from a single experiment.

Recently, m5C in the CpG context has been identified (11) and a signal for methylation can be observed at known methylated sites in bacteria using Nanopore sequencing (12,13). So far no technique permits, from a single experiment, the dual sequencing of genomes and the *de novo* determination of m5C methylase specificity for the non-CpG contexts typically found in bacteria.

Herein, we describe a novel approach called RIMS-seq to simultaneously sequence bacterial genomes and globally profile m5C methylase specificity using a protocol that closely resembles the standard Illumina DNA-seq with a single, additional step. RIMS-seq shows comparable sequencing quality as DNA-seq and accurately identifies methylase specificities. Applied to characterized strains or novel isolates, RIMS-seq *de novo* identifies novel activities without the need for a reference genome and permits the assembly of the bacterial genome at metrics comparable to standard shotgun sequencing.

MATERIALS AND METHODS

Samples and genomic DNA collection

Skin microbiome genomic DNA (ATCC® MSA-1005) and gut microbiome genomic DNA (ATCC® MSA-1006) were obtained from ATCC. *Escherichia coli* BL21 genomic DNA was extracted from a culture of *E. coli* BL21 DE3 cells (C2527, New England Biolabs) using the DNEasy Blood and Tissue kit (69504, Qiagen). *Escherichia coli* K12 MG1655 genomic DNA was extracted from a cell culture using the DNEasy Blood and Tissue kit (69504, Qiagen). All the other gDNA from the bacteria presented in Table 1 were isolated using the Monarch genomic DNA purification kit (T3010S, New England Biolabs). Xp12 phage genomic DNA was obtained from Peter Weigele and Yian-Jiun Lee at New England Biolabs.

RIMS-seq library preparation

One hundred nanogram of gDNA was sonicated in 1× TE buffer using the Covaris S2 (Covaris) with the standard protocol for 50 μl and 200 bp insert size.

The subsequent fragmented gDNA was used as the starting input for the NEBNext Ultra II library prep kit for Illumina (E7645, New England Biolabs) following the manufacturer's recommendations until the USER treatment step. The regular unmethylated loop-shaped adapter was used for ligation. After the USER treatment (step included), the samples were subjected to heat alkaline deamination: 1 M NaOH pH 13 was added to a final concentration of 0.1 M and the reactions were placed in a thermocycler at 60°C for 3 h. Then, the samples were immediately cooled down on ice and 1 M of acetic acid was added to a final concentration of 0.1 M in order to neutralize the reactions. We also tested alkaline concentration of 0.5M and 1M NaOH, in these cases, equal amounts of acetic acid were added to the reaction to properly neutralize the PH. The neutralized reactions were cleaned up using the Zymo oligo clean and concentrator kit (D4060 Zymo Research) and the DNA was eluted in 20 μl of 0.1× TE.

PCR amplification of the samples was done following NEBNext Ultra II library prep kit for Illumina protocol (ER7645, New England Biolabs) and the NEBNext® Multiplex Oligos for Illumina® (E7337A, New England Biolabs). The number of PCR cycles was tested and optimized for each sample following the standard procedure for library preparation. PCR reactions were cleaned up using 0.9× NEBNext Sample purification beads (E7137AA, New England Biolabs) and eluted in 25 μl of 0.1× TE. All the libraries were evaluated on a TapeStation High sensitivity DNA1000 (Agilent Technologies) and paired-end sequenced on Illumina.

Bisulfite-seq library preparation

One percent of lambda phage gDNA (D1221, Promega) was spiked-into 300 ng gDNA to use as an unmethylated internal control. The samples were sonicated in 1× TE buffer using the Covaris S2 (Covaris) with the standard protocol for 50 μl and 200 bp insert size.

The subsequent fragmented gDNA was used as the starting input for the NEBNext Ultra II library prep kit for Illumina (E7645, New England Biolabs) following the manufacturer's recommendations until the USER treatment step. The methylated loop-shaped adapter was used for ligation. After USER, a 0.6× clean-up was performed using the NEBNext Sample purification beads (E7137AA, New England Biolabs) and eluted in 20 μl of 0.1× TE. A TapeStation High Sensitivity DNA1000 was used to assess the quality of the library before subsequent bisulfite treatment. The Zymo EZ DNA Methylation-Gold Kit (D5005, Zymo Research) was used for bisulfite treatment, following the manufacturer's protocol.

PCR amplification of the samples was done following the suggestions from NEBNext Ultra II library prep kit for Illumina (ER7645, New England Biolabs), using the NEBNext® Multiplex Oligos for Illumina® (E7337A, New England Biolabs) and NEBNext® Q5U® Master Mix (M0597, New England Biolabs).

The number of PCR cycles was tested and optimized for each sample. The PCR reactions were cleaned up using 0.9× NEBNext Sample purification beads (E7137AA, New England Biolabs) and eluted in 25 μl of 0.1× TE. All the libraries were screened on a TapeStation High sensitivity DNA1000 (Agilent Technologies) and paired-end sequenced on Illumina.

RIMS-seq data analysis

Paired-end reads were trimmed using Trim Galore 0.6.3 (option –trim1). The *Acinetobacter calcoaceticus* ATCC 49823 data have been trimmed using Trim Galore version 0.6.3 instead and downsampled to 1 million reads. Reads

were mapped to the appropriate genome using BWA mem with the paired-end mode (version 0.7.5a-r418 and version 0.7.17-r1188 for the *A. calcoaceticus*). When using an assembled genome directly from RIMS-seq data, trimmed RIMS-seq reads were assembled using SPAdes (SPAdes-3.13.0 (31) default parameters). Reads were split according to the read origin (Read 1 or Read 2) using samtools (version 1.8) with -f 64 (for Read 1) and -f 128 (for Read 2) and samtools mpileup (version 1.8) was run on the split read files with the following parameters: -O -s -q 10 -Q 0. For *Acinetobacter calcoaceticus*, the unmapped reads, reads without a mapped mate and the non-primary alignments were filtered out using the flags -F 12 and -F 256.

De-novo identification of motifs using RIMS-seq

Programs and a detailed manual for the *de-novo* identification of motifs in RIMS-seq are available on github (https://github.com/Ettwiller/RIMS-seq/). Using the mpileup files, positions and 14bp flanking genomic regions for which a high quality (base quality score $\geq$ 35) C to T in R1 or G to A in R2 was found, were extracted for the foreground. Positions and 14bp flanking regions for which a high quality (base quality score $\geq$ 35) G to A in R1 or C to T in R2 was found, were extracted for the background. C to T or G to A in the first position of reads were ignored. If the percentage of C to T or G to A are above 5% for at least 5 reads at any given position, the position was ignored (to avoid considering positions containing true variants). Motifs that are found significantly enriched (P-value $< 1e^{-100}$) in the foreground sequences compared to background sequences were found using mosdi pipeline mosdi-discovery with the following parameters: '*mosdi-discovery -v discovery -q x -i -T 1e-100 -M 8,1,0,4 8 occ-count'* using the foreground sequences with *x* being the output of the following command : '*mosdi-utils count-qgrams -A 'dna"* using the background sequences. To identify additional motifs, the most significant motif found using *mosdi-discovery* is removed from the foreground and background sequences using the following parameter: '*mosdi-utils cut-out-motif -M X'* and the motif discovery process is repeated until no significantly enriched motif can be found.

Sequence logo generation

Using the mpileup files, positions in the genome for which a high quality (base quality score $\geq$ 35) C to T in R1 or a G to A in R2 was observed were extracted for the foreground using the get_motif_step1.pl program. Positions for which a high quality (base quality score $\geq$ 35) G to A in R1 or a C to T in R2 was observed were extracted for the background. The $\pm$7 bp regions flanking those positions were used to run two sample logo (32). Parameters were set as *t*-test, pP-value <0.01.

Bisulfite-seq data analysis

Reads were trimmed using Trim Galore 0.6.3 and mapped to the bisulfite-converted concatenated reference genomes of each respective synthetic microbiome using bismark 0.22.2 with default parameters. PCR duplicates were removed using deduplicate_bismark and methylation information extracted using bismark_methylation_extractor using default parameters. For the microbiome, the bismark_methylation_extractor with –split_by_chromosome option was used to output one methylation report per bacterium. The motif identification was done as previously described in (10).

EM-seq

EM-seq was performed according to the standard protocol (NEB E7120S). Motif identification was done as previously described in (10).

Analysis and abundance estimation in synthetic microbiomes

RIMS-seq, DNA-seq and bisulfite-seq were performed on the synthetic gut and skin microbiome as described. Reads derived from RIMS-seq, DNA-seq and bisulfite-seq were mapped as described to a 'meta-genome' composed of the reference genomes of all the bacteria included in the corresponding synthetic community (see Supplementary Table S3 for detailed compositions). Mapped reads were split according to each bacterium and RIMS-seq or bisulfite analysis pipelines were run on individual genomes as described above. Abundance was estimated using the number of mapped reads per bacteria and normalized to the total number of mapped reads. Normalized species abundances in RIMS-seq and bisulfite-seq were compared to the normalized species abundances in DNA-seq.

Phylogeny of the ATCC synthetic microbiomes and visualization

The phylogenetic trees of both ATCC synthetic gut and skin microbiomes were done using OrthoFinder version 2.3.11 (33) using the MSA workflow and MAFFT for the multiple sequence alignment program. The program options are available at https://github.com/davidemms/OrthoFinder. The phylogenetic tree and abundance data obtained from DNA-seq, RIMS-seq and bisulfite-seq were visualized using iTOL (34) for each synthetic community (see Supplementary Figure S5).

Quality control of the data

The insert size for each downsampled filtered bam file was calculated using Picard version 2.20.8 using the default parameters and the option CollectInsertSizeMetrics ('Picard Toolkit.' 2019. Broad Institute, GitHub Repository. http://broadinstitute.github.io/picard/; Broad Institute).

The GC bias for each downsampled filtered bam file was calculated and plotted using Picard version 2.20.8 using the default parameters and the option CollectGcBiasMetrics.

Xp12 genome assembly

Reads were downsampled to a 30$\times$ coverage using seqtk 1.3.106, trimmed using trimgalore 0.6.5 and assembled using Spades 3.14.1 with the –isolate option. Assembly quality was assessed using Quast 5.0.2. Reads used for assembly

were then mapped back to the assembly using BWA mem 0.7.17 and mapping statistics were generated using samtools flagstat 1.10.2

Xp12 sequencing performance assessment

Reads were trimmed using trimgalore 0.6.5 and mapped to the Xp12 reference genome using BWA mem 0.7.17. Insert size and GC bias were assessed using Picard Toolkit and genome coverage using Qualimap 2.1.1.

Intact mass LC–MS

Intact mass analysis was performed by tandem liquid chromatography–mass spectrometry (LC–MS/MS) on an Vanquish Horizon UHPLC System equipped with a diode array detector and a Thermo Q-Exactive Plus mass spectrometer operating under negative electrospray ionization mode (–ESI). UHPLC was performed using a Thermo DNAPac™ RP Column (2.1 × 50 mm, 4 μm) at 70°C and 0.3 ml/min flow rate, with a gradient mobile phase consisting of hexafluoroisopropanol (HFIP)–*N*,*N*-diisopropylethylamine (DIEA) aqueous buffer and methanol. UV detection was performed at 260 nm. Intact mass analysis was performed under Full MS mode, and ESI-MS raw data was deconvoluted using Promass HR (Novatia Inc.).

RESULTS

Principle of RIMS-seq

Spontaneous deamination of cytosine (C) leading to uracil (U) and of m5C leading to thymine (T) are examples of common damage found in DNA. *In-vitro*, this damage is often undesirable as it can interfere with sequencing. The type of interference during sequencing depends on whether the deamination occurs on C or m5C. U blocks the passage of high-fidelity polymerases typically used in library preparation protocols, preventing the amplification and sequencing of U-containing DNA fragments. Conversely, DNA harboring T derived from m5C deamination can be normally amplified but results in C to T errors (14,15). This distinction between blocking and mutagenic damage forms the basis of the RIMS-seq method, allowing the identification of methylase specificity based on an elevated number of reads containing C to T transitions specifically at methylated sites (Figure 1A). To increase the rate of m5C deamination, the DNA is subjected to a heat-alkaline treatment which has been previously demonstrated to elevate the rate of both C and m5C deamination with m5C having a 1.5–3 times higher deamination rate than for C (16). This treatment is aimed at inducing a level of deamination large enough to detect the m5C methylase specificity without affecting the sequencing quality. For this reason, the deamination levels typically obtained with RIMS-seq does not permit the quantitative measurement of methylation at each genomic site but rather provides a global methylation signal characteristic of the methylase specificity.

Illumina paired-end sequencing allows both ends of a DNA fragment to be sequenced, generating a forward read (R1) and reverse read (R2). Resulting from m5C deamination, R1 has the C to T read variants while R2 has the reverse-complement G to A variant. This difference leads to an overall imbalance of C to T variants between R1 and R2 (17) (see also Supplementary Figure S1 for explanation). Thus, sequence contexts for which the C to T read variants are imbalanced in R1 compared to R2 correspond to m5C methylase specificity(ies). Because of the limited deamination rate, RIMS-seq takes advantage of the collective signal at all sites to define methylase specificity. Because C to T imbalance can be observed at nucleotide resolution, RIMS-seq identifies at base resolution which of the cytosine within the motif is methylated.

The experimental steps for RIMS-seq essentially follow the standard library preparation for Illumina sequencing with an extra deamination step. Briefly, the bacterial genomic DNA is fragmented, and adaptors are ligated to the ends of DNA fragments (Figure 1B and Materials and Methods). Between the ligation step and the amplification step, an alkaline heat treatment step is added to increase the rate of deamination. Only un-deaminated DNA or DNA containing deaminated m5C can be amplified and sequenced.

Validation of RIMS-seq

Optimization of the heat alkaline deamination step. We first evaluated the conditions to maximize the deamination of m5C while minimizing other DNA damage. For this we used bacteriophage Xp12 genomic DNA that contains exclusively m5C instead of C (18) to measure the m5C deamination rates in various contexts.

To estimate the overall deamination rate of m5C, we quantified the imbalance of C to T read variants between R1 and R2 for 0, 10 and 30 min, 1 h, 2 h, 3 h, 5 h and 14 h of heat alkaline treatment (Figure 1C). We observed an imbalance as early as 10 min with a 3.7-fold increase of C to T read variants in R1 compared to R2. The increase is linear with time with a maximum of 212-fold increase of C to T read variants in R1 compared to R2 after 14 h of heat alkaline treatment (Figure 1D). Next, we quantified the deamination rate at all Nm5CN sequence contexts with N being A, T, C or G and show an increase of C to T variants in R1 in all contexts (Supplementary Figure S2A). Together, these results show that a measurable deamination rate can be achieved in as soon as 10 min of heat alkaline deamination and that deamination efficiency is similar in all sequence contexts.

To estimate the non-specific damage to the DNA leading to unwanted sequencing errors, we quantified possible imbalances for other variant types (Supplementary Figure S2B). We found that G to T variants show imbalance in all the conditions investigated, likely the result of oxidative damage resulting from sonication, a common step in library preparation between RIMS-seq and DNA-seq (17). Interestingly, the imbalance is reduced in RIMS-seq, disappearing almost completely after 14 h of heat alkaline treatment (Supplementary Figure S2B). This result suggests that this treatment either converts 8-oxoG back to G or to another modification that ultimately blocks the polymerase

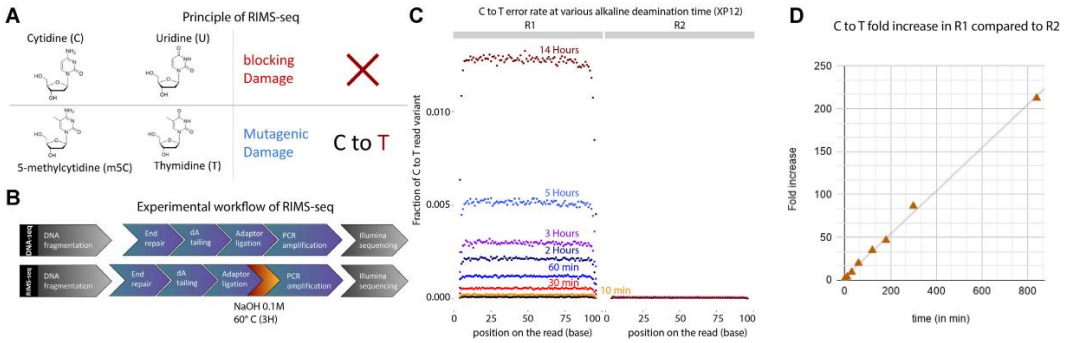


Figure 1. (**A**) Principle of RIMS-seq. Deamination of cytidine leads to a blocking damage while deamination of m5C leads to a mutagenic C to T damage only present on the first read (R1) of paired-end reads in standard Illumina sequencing. Thus, an increase of C to T errors in R1 in specific contexts is indicative of m5C. (**B**) The workflow of RIMS-seq is equivalent to a regular library preparation for Illumina DNA-seq with an extra step of limited alkaline deamination at 60°C. This step can be done immediately after adaptor ligation and does not require additional cleaning steps. (**C**) Fraction of C to T variants in XP12 (m5C) at all positions in the reads for R1 and R2 after 0min (DNA-seq), 10 min, 30 min, 60 min, 2 h, 3 h, 5 h and 14 h of heat-alkaline treatment. The C to T imbalance between R1 and R2 is indicative of deamination of m5C and increases with heat-alkaline treatment time. (**D**) Correlation between the C to T fold increases in R1 compared to R2 according to time ($r^2 = 0.998$).

from amplifying 8-oxoG-containing fragments. To properly address the disappearance of G to T variants due to oxidative damage in RIMS-seq, we designed an oligonucleotide containing a single 8-oxoG. Using LC–MS intact mass, we identified a strand break directly 5′ and 3′ of the 8-oxoG that is specific to oxidized G under heat alkaline treatments (Supplementary text 1 and Supplementary Figure S3). Thus, the heat-alkaline treatment performed in RIMS-seq induced strand breaks at oxidative damage sites, preventing the amplification of 8-oxoG-containing fragments and de-facto decreasing the frequency of G to T in the RIMS-seq libraries.

A slight elevation of G to C and T to C read variants can be observed in RIMS-seq compared to DNA-seq but this damage is of low frequency and therefore is not expected to notably affect the sequencing performance QC of RIMS-seq.

We performed QC metrics and assemblies of Xp12 for all the alkaline-heat treatment conditions, including a control DNA-seq. The overall sequencing performances were assessed in terms of insert size, GC bias and genome coverage. Similar results were observed between RIMS-seq and the DNA-seq control at all treatment times, indicating that the RIMS-seq heat-alkaline treatment does not affect the quality of the libraries (Supplementary Figure S4).

We also evaluated the quality of the assemblies compared to the Xp12 reference genome and found that all conditions lead to a single contig corresponding to essentially the entire genome with very few mismatches (Supplementary Table S1). These results suggest that the heat-alkaline treatment does not affect the assembly quality, raising the possibility of using RIMS-seq for simultaneous *de novo* genome assembly and methylase specificity identification. We found that a 3-h treatment provides a good compromise between the deamination rate (resulting in ∼0.3% of m5C showing C to T transition) and duration of the experiment. We found that longer incubation times (up to 14 h) increased the deamination rate by up to 1% and decided this is a slight sensitivity increase compared to the additional experimental time required.

RIMS-seq is able to distinguish methylated versus unmethylated motifs in E. coli. To validate the application of RIMS-seq to bacterial genomes, we sequenced dcm+ (K12) and dcm- (BL21) *E. coli* strains. In K12, the DNA cytosine methyltransferase *dcm* methylates cytosine at CCWGG sites (C = m5C, W = A or T) and is responsible for all m5C methylation in this strain (19). *E. coli* BL21 has no known m5C methylation. Heat/alkaline treatments were performed at three time points (10 min, 1 h and 3 h). In addition, we performed a control experiment corresponding to the standard DNA-seq. Resulting libraries were paired-end sequenced using Illumina and mapped to their corresponding genomes (Methods).

For comparison, all datasets were downsampled to 5 million reads corresponding to 200× coverage of the *E. coli* genome and instances of high confidence C to T variants (Q score > 35) on either R1 or R2 were identified. As expected, control DNA-seq experiments show comparable numbers of C to T read variants between R1 and R2, indicating true C to T variants or errors during amplification and sequencing (Figure 2A). On the other hand, the overall number of C to T read variants in R1 is progressively elevated for 10 min, 1 h and 3 h of heat-alkaline treatment of the *E. coli* K12 samples with an overall 4-fold increase after 3 htreatment compared to no treatment; heat-alkaline treatments did not increase the rate of C to T read variants in R2 (Figure 2A). We anticipate that the elevation of the *E. coli* K12 C to T read variants in R1 is due to deamination of m5C. In this case, the elevation should be specifically found in Cs in the context of CCWGG (with the underlined C corresponding to the base under consideration). To demonstrate this, we calculated the fraction of C to T read variants in CCWGG compared to other contexts. We observed a large elevation of the C to T read variants in the CCAGG and CCTGG contexts for K12 (Figure 2B). As expected, the

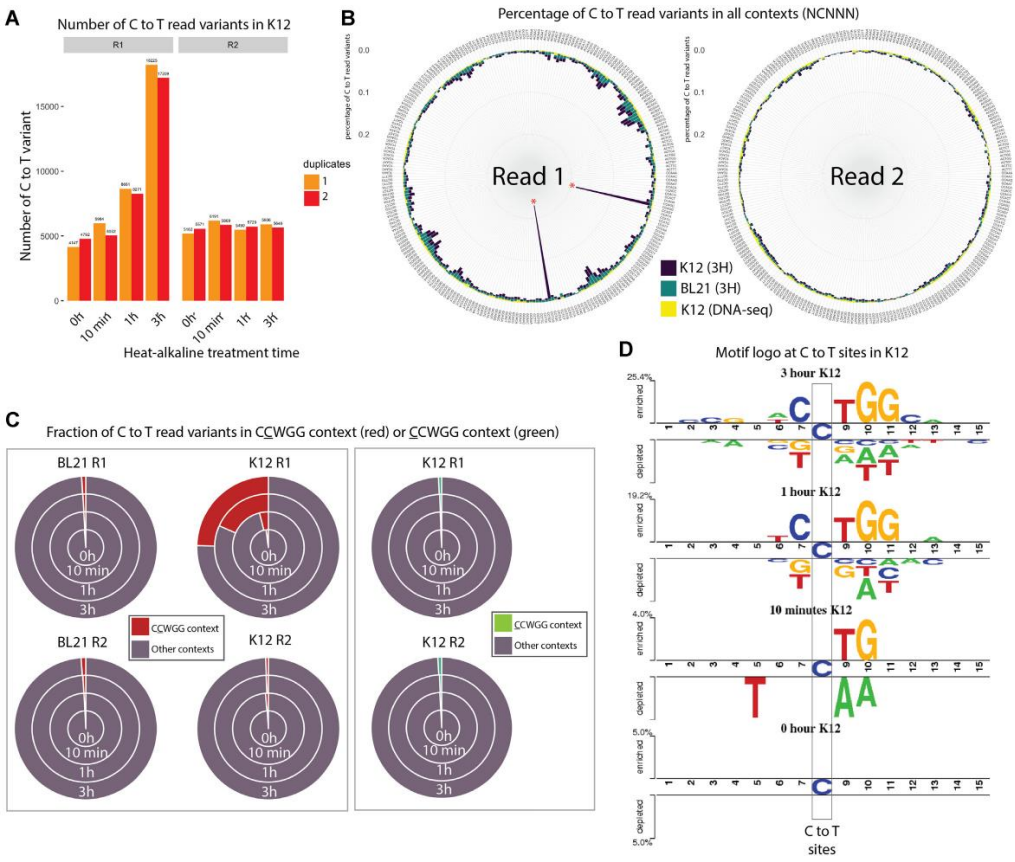


Figure 2. (**A**) Bar plots representing the number of C to T read variants for K12 in R1 and R2 after different heat/alkaline treatment times. Colors represent duplicate experiments. (**B**) Circular bar plots representing the percentage of C to T read variants in all NCNNN contexts (with N = A, T, C or G) for Read 1 (R1, left) and Read 2 (R2, right) in DNA-seq performed on K12 (yellow bars), RIMS-seq (3H) performed on BL21 (green) and RIMS-seq (3H) performed on K12 (dark blue). Red asterisks denote CCWGG contexts with W being either A or T. (**C**) Proportion of C to T read variants in CCWGG (red) or CCWGG (green) contexts compared to other NCNNN or CNNNN contexts for R1 and R2 in K12 and BL21. The C to T read variants in CCWGG and CCWGG motifs represent less than 2% of all variants except in K12 (R1 only) after 10 min, 1- and 3-h treatments where the CCWGG motifs represent 4.1%, 22.5% and 32.6% of all C to T read variants respectively. The increase of C to T read variants in the CCWGG context is therefore specific to R1 in K12 strain. (**D**) Visualization of the statistically significant differences in position-specific nucleotide compositions around C to T variants in R1 compared to R2 using Two Sample Logo (21) for the K12 sample subjected to (from top to bottom) 3 h, 1 h, 10 min and 0 min heat alkaline treatment.

C to T read variants show no elevation at CCAGG and CCTGG contexts for the *E. coli* BL21 strain that is missing the *dcm* methylase gene (Figure 2B). Thus, this C to T read variant elevation is specific to the *E. coli* K12 strain subjected to heat-alkaline treatments, consistent with deamination detectable only on methylated sites. Taken together, these results indicate that the elevated rate of C to T variants observed in R1 from *E. coli* K12 is the result of m5C deamination in the CCWGG context.

Next, we assessed whether the difference in the C to T read variant context between R1 and R2 at the CCWGG motif provides a strong enough signal to be discernible over the background noise. For this, we calculated the fraction of C to T read variants in CCWGG and CCWGG compared to all the other NCNNN and CNNNN contexts, respectively. After 3 h of heat-alkaline treatment, the fraction of C to T read variants in a CCWGG context increased, rising from only 1.9% in regular DNA-seq to ∼25% of all the C to T variants. This increase is only observable in R1 of the K12 strain (Figure 2C). Conversely, no increase can be observed in a CCWGG context for which the C to T variant rate at the first C is assessed (Figure 2C). Thus, RIMS-seq identified the second C as the one bearing the methylation, consistent with the well described dcm methylation of *E. coli* K12 (20) (19), highlighting the ability of RIMS-seq to identify m5C methylation at base resolution within the methylated motif.

Next, we calculated significant (P-value < 0.01) differences in position-specific nucleotide compositions around

C to T variants in R1 compared to R2 using Two Sample Logo (21). We found a signal consistent with the dcm methylase specificity in K12 RIMS-seq samples at 1 and 3 h of heat alkaline treatment (Figure 2D) demonstrating that it is possible to identify methylase specificities in genomic sequence subject to as little as 1 h of alkaline treatment. These results support the application of RIMS-seq for the *de novo* identification of methylase specificity at base resolution.

RIMS-seq identifies the correct methylase specificity de novo *in* E. coli *K12.* In order for RIMS-seq to identify methylase specificities *de novo*, we devised an analysis pipeline based on MoSDi (22) to find sequence motif(s) that are over-represented around C to T transitions in R1 reads (Figure 3A, analysis pipeline available at https://github.com/Ettwiller/RIMS-seq). In brief, the pipeline extracts the sequence context at each C to T read variant in R1 (foreground) and R2 (background). MoSDi identifies the highest over-represented motif in the foreground sequences compared to the background sequences. To accommodate the presence of multiple methylases in the same host, the first motif is subsequently masked in both the foreground and background sequences and the pipeline is run again to find the second highest over-represented motif and so on until no significant motifs can be found (see Materials and Methods for details). Running the pipeline using the K12 strain RIMS-seq data identifies one significant over-represented motif corresponding to the CCWGG motif (*P*-value = $9.71e^{-77}$, $4.25e^{-858}$ and $3.61e^{-4371}$ for 10, 60 and 180 min of alkaline treatment respectively) with the cytosine at position 2 being m5C.

Summing up, we devised a novel sequencing strategy called RIMS-seq and its analysis pipeline to identify m5C methylase specificity *de novo*. When applied to *E. coli* K12, RIMS-seq identifies the dcm methylase specificity as C$\underline{C}$WGG with the methylated site located on the second C, consistent with the reported dcm methylase specificity (Table 1).

RIMS-seq identifies multiple methylase specificities de novo *within a single microorganism.* To assess whether RIMS-seq can identify methylase specificity in strains expressing multiple methylases, we repeated the same procedure on a strain of *Acinetobacter calcoaceticus* ATCC 49823 expressing two m5C methylases with known specificities (4). RIMS-seq identifies C$\underline{C}$GCG (*P*-value = $2.33e^{-174}$) and GATC$\underline{C}$ (*P*-value = $3.02e^{-1308}$) (Table 1) both motifs have been confirmed by MFRE-seq (10). Thus, RIMS-seq is able to *de novo* identify methylase specificities in bacteria expressing multiple methylases.

RIMS-seq can be applied for genome sequencing and m5C profiling in bacteria without a reference genome. We investigated whether RIMS-seq can be used to identify methylase specificities of uncharacterized bacteria for which a reference genome is unavailable. More specifically, we evaluated if the reads generated using RIMS-seq can be used for both identifying methylase specificities and generating an assembly of comparable quality to DNA-seq.

For this, we performed RIMS-seq on *A. calcoaceticus* ATCC 49823 genomic DNA as described above as well as a control DNA-seq experiment for which the alkaline treatment was replaced by 3 h incubation in TE (DNA-seq(+3H)). We compared the *de novo* assembly obtained from the reads generated by the DNA-seq(+3H) and the *de novo* assembly obtained from the reads generated by RIMS-seq (see Materials and Methods). In brief, the alkaline treatment did not alter the important metrics for assembly quality such as the rate of mismatches and N50 demonstrating that the elevated C to T variant rate at methylated sites is not high enough to cause assembly errors (Figure 3B).

We then proceeded to map the RIMS-seq reads to the assembly and motifs were identified using the RIMS-seq *de novo* motif discovery pipeline. As expected, the same motifs found when mapping to the reference genome are also found in the *A. calcoaceticus de novo* assembly with similar significance (GATC$\underline{C}$ (*P*-value = $1.44e^{-1255}$) and C$\underline{C}$GCG (*P*-value = $8.6e^{-228}$) (Figure 3C). These motifs correspond to the methylase specificities expected in this strain indicating that RIMS-seq can be applied for genome sequencing and assembly of any bacterium without the need for a reference genome.

RIMS-seq can be complemented with SMRT sequencing to obtain a comprehensive overview of methylase specificities. RIMS-seq performed in parallel with SMRT sequencing has the advantage of comprehensively identifying all methylase specificities (m5C, m4C and m6A methylations) and results in an assembly of higher quality than with short reads illumina data. We applied this hybrid approach to *Acinetobacter calcoaceticus* ATCC 49823 for which a SMRT sequencing and assembly had been done previously (4). RIMS-seq was performed as described above and the reads were mapped to the genome assembly obtained from SMRT-sequencing. We again found the two m5C motifs: C$\underline{C}$GCG (*P*-value = $1.84e^{-1535}$) and GATC$\underline{C}$ (*P*-value = $4.93e^{-6856}$) from the RIMS-seq data in addition to the 13 m6A motifs described previously using SMRT sequencing (4). This result demonstrates the advantage of such a hybrid approach in obtaining closed genomes with comprehensive epigenetic information.

XP12 can be used as a spiked-in to measure the deamination rate. To ensure the correct level of heat-alkaline deamination rate, XP12 can be used as spiked-in to measure the deamination rate at m5C. To illustrate the practicality of such control, we subjected Haemophilus influenzae Rd ATCC 51907 (Table 1) spiked-in with XP12 DNA to various NaOH concentrations and treatment times. We observed deamination rates varying from 0.24% (0.1 M NaOH, 3 h) to 2.72% (0.5 M NaOH, 3 h) (Supplementary Figure S2C). We further investigated the error rates in both the bacteria and XP12 for substitutions other than C to T at various heat alkaline conditions (Supplementary Figure S2C) and found that all substitution rates are comparable to the rates obtained using standard DNA-seq. Taken together, these results indicate that the heat alkaline treatments in the measured ranges are not expected to notably affect the sequencing performance QC in bacteria.

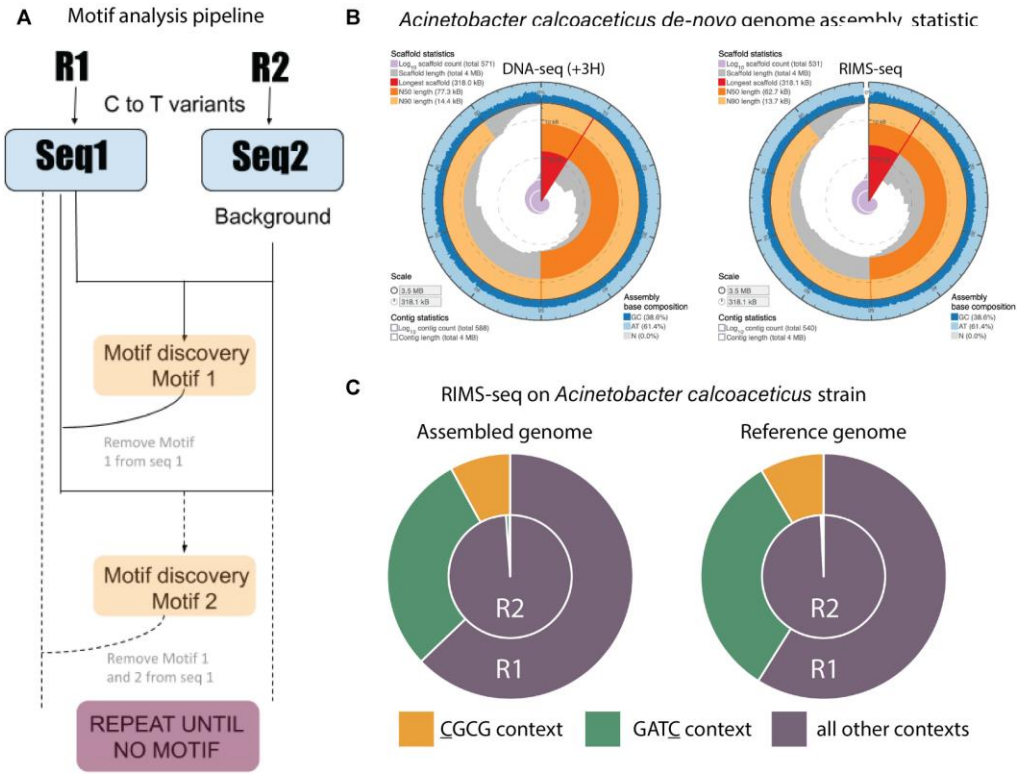


Figure 3. *De novo* discovery of methylase specificity using RIMS-seq. (**A**) Description of the RIMS-seq motif analysis pipeline. First, C to T read variants are identified in both Read 1 and Read 2 separately. Then, the MosDI program searches for overrepresented motifs. Once a motif is found, the pipeline is repeated until no more motifs are found, enabling identification of multiple methylase specificities in an organism. (**B**) Assembly statistics obtained using the sequence from the standard DNA-seq (+3H, left) and RIMS-seq (right). Visualization using assembly-stats program (https://github.com/rjchallis/assembly-stats). The corresponding table with the statistical values is available in the supplementary material (Supplementary Table S2). (**C**) Fractions of C to T read variants in CGCG (yellow) or GATC (green) contexts compared to other contexts for R1 and R2 in Acinetobacter calcoaceticus ATCC 49823 using the assembled or the reference genome. The increase of C to T read variants in these contexts are similar when using either the assembled or reference genomes

Table 1. Methylases specificity obtained using RIMS-seq and validated using different methods. The method is indicated by a number next to the motif. : Evidence for the validated motifs are (1) bisulfite-seq (Materials and Methods), (2) REBASE (4), (3) EM-seq (material and method), (4) MFRE-seq (10), (5) mTet1-enhanced SMRT sequencing (6)

Organism	Accession numbers (biosample)	RIMS-seq motif(s)	Validated motif(s)
Escherichia coli K12	SAMN02604091	CCWGG	CCWGG (1,2,4)
Acinetobacter calcoaceticus ATCC 49823	SAMN14530202	GATC	GATC (4)
		CGCG	CGCG (2,4)
Bacillus fusiformis 1083	SAMN17843035	ACCTGC	ACCTGC (2,3)
		GCAGGT	GCAGGT (2,3)
Bacillus amyloliquefaciens H ATCC 49763	SAMN12284742	GCWGC	GCWGC (3)
Clostridium acetobutylicum ABKn8	SAMN17843114	GCNNGC	GCNNGC (3)
Aeromonas hydrophila NEB724	SAMN14533640	GCCGGC	GCCGGC (3)
Haemophilus influenzae Rd ATCC 51907	SAMN02603991	GRCGYC*	GRCGYC (5)
		ACCGCACT	
		AGTGCGGT	
Haemophilus parahaemoltyicus ATCC 10014	SAMN11345835	GCGC	GCGC (2)
M.HhaI clone (*E. coli*)	NA	RCGC	GCGC (4)
		CCWGG[a]	CCWGG (1,2,4)[a]

(a) The *E. coli* strain used is Dcm+, resulting in the discovery of both the Dcm (CCWGG) and M.HhaI motifs (GCGC). RIMS-seq discovered RCGC instead of GCGC motif (see text for explanation). * *P*-value = $1.0e^{-91}$ (standard detection threshold of $<1.0e^{-100}$ would miss this motif).

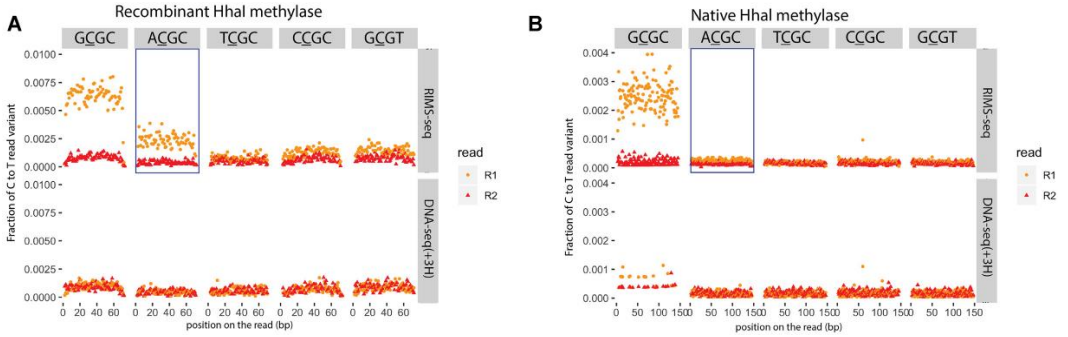


Figure 4. C to T error profile in GCGC (canonical recognition site), ACGC, TCGC, CCGC and GCGT. in R1 reads (orange) and R2 reads (red) for RIMS-seq (upper panel) and DNA-seq(+3H) (lower panel) *A.* Recombinant HhaI methylase expressed in *E. coli B.* Native HhaI methylase expressed in *Haemophilus parahaemolyticus*. Elevation of C to T in the R1 read variant can be observed in the context of GCGC for both the recombinant and native HhaI genomic DNA and in the context of ACGC only for DNA from the recombinant but not the native HhaI.

RIMS-seq can be applied to a variety of RM systems

Methylases targets are usually palindromic sequences between 4 nt and 8 nt, and a single bacterium often possesses several, distinct MTase activities (23). Next, we tested the general applicability of RIMS-seq and the *de novo* motif discovery pipeline using bacterial genomic DNA from our in-house collections of strains.

For some bacterial strains, the methylase recognition specificities have been previously experimentally characterized. In all of those strains, RIMS-seq confirms the specificities and identifies the methylated cytosine at base resolution (Table 1). We have tested the identification of 4-mers motifs such as GATC, CGCG (*Acinetobacter calcoaceticus*) and GCGC (*Haemophilus parahaemolyticus*) up to 8-mers motifs such as ACCGCACT and AGTGCGGT (*Haemophilus influenzae*). Motifs can be palindromic or non-palindromic (Table 1 and Supplementary Table S3). In the latter case, RIMS-seq defines non-palindromic motifs at strand resolution. For example, RIMS-seq identifies methylation at two non-palindromic motifs ACCTGC as well as its reverse complement GCAGGT in the *Bacillus fusiformis* strain (Table 1).

A number of RM systems have been expressed in other hosts such as *E. coli* for biotechnological applications. For the methylase M.HhaI recognizing GCGC (4), we performed RIMS-seq and a control DNA-seq(+3H) on both the native strain (*Haemophilus parahaemolyticus* ATCC 10014) and in *E. coli* K12 expressing the recombinant version of M.HhaI. Interestingly, we found that the *de novo* RIMS-seq analysis algorithm identifies RCGC (with R being either A or G) for the recombinant strain and GCGC for the native strain (Figure 4A). Conversely, no notable elevation of C to T read variants are observed at ACGC for the native strain (Figure 4B), confirming the *de novo* motif discovery results from the analysis pipeline. Collectively, these results suggest that the recombinant methylase shows star activity, notably in the context of ACGC, that is not found in the native strain. We hypothesize that the star activity is the result of the over-expression of the methylase in *E. coli* K12. Interestingly, ACGC is not a palindrome motif and consequently the star activity results in hemi-methylation of the ACGC sites and not the GCGT motif.

RIMS-seq can be applied to microbial communities

We assessed whether RIMS-seq can be applied to mixed microbial communities using synthetic gut and skin microbiomes from ATCC containing 12 and 6 bacterial species, respectively. We also complemented the RIMS-seq experiment with the control experiment DNA-seq(+3H) and a bisulfite treatment to validate the RIMS-seq findings. Reads were mapped to their respective microbiome reference genomes (Materials and Methods). For the gut microbiome we found a mapping rate (properly paired only) of 95.79%, 95.77% and 66.2% for RIMS-seq, DNA-seq and bisulfite-seq respectively. Concerning the skin microbiome, 85.89%, 85.35% and 54.9% of reads were mapped for RIMS-seq, DNA-seq and bisulfite-seq respectively. The low mapping rate for bisulfite-seq is a known challenge as the reduction of the alphabet to A, G, T generates ambiguous mapping (24).

To use RIMS-seq as an equivalent to DNA-seq for mixed community applications, RIMS-seq should produce sequencing quality metrics that are similar to standard DNA-seq, especially on the estimation of species relative abundances. We therefore compared RIMS-seq sequencing performances with DNA-seq(+3H) and bisulfite sequencing. We found that bisulfite sequencing elevates abundances of AT-rich species such as *Clostridioides difficile* (71% AT), *Enterococcus faecalis* (63% AT) and *Fusobacterium nucleatum* (73% AT) (Figure 5A, Supplementary Figure S5). For example, bisulfite sequencing over-estimated the presence of *Clostridioides difficile* by a factor of 2.65 and *Staphylococcus epidermidis* by a factor of 3.9 relative to DNA-seq. This over-estimation of an AT rich genome by bisulfite is a known bias of bisulfite sequencing and relates to damage at cytosine bases (25). Conversely, we found that the species abundances are similar between DNA-seq(+3H) and RIMS-seq (abundance ratios between 0.8 and 1.2) indicating that RIMS-seq can be used to quantitatively estimate microbial composition.

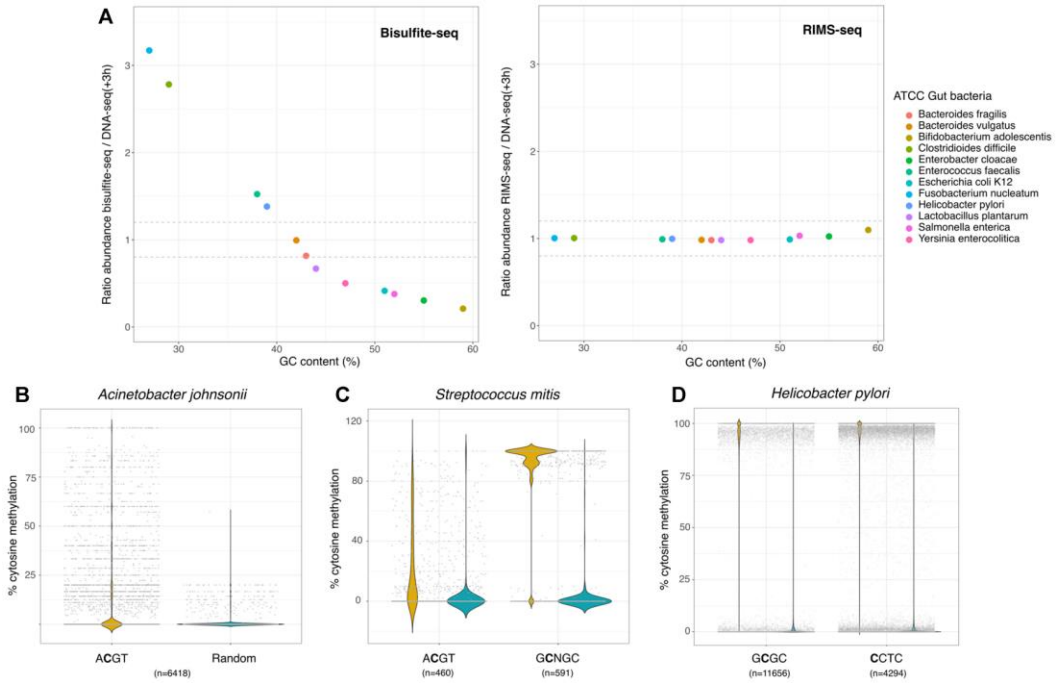


Figure 5. (**A**) Bacterial abundance in the ATCC gut microbiome calculated from bisulfite-seq data (left) and RIMS-seq (Right) normalized to DNAseq(+3H). The normalized abundance is plotted relative to the GC content of each bacterium. (**B**) Methylation levels in *Acinetobacter johnsonii* (ATCC skin microbiome).The methylation level was calculated for cytosine positions in the context of ACGT (yellow) and randomly selected positions in other contexts (blue). These bisulfite-seq data suggest some sites are methylated in the context of ACGT, but they are not fully methylated. (**C**) Methylation level in Streptococcus mitis (ATCC skin microbiome) calculated from bisulfite-seq data. The methylation level was calculated for cytosine positions in the context of ACGT and GCNGC (yellow) as well as for randomly selected positions in other contexts (blue). (**D**) Methylation level in Helicobacter pylori (ATCC gut microbiome) calculated from bisulfite-seq data. The methylation level was calculated for cytosine positions in the context of GCGC and CCTC (yellow) as well as for randomly selected positions in other contexts (blue).

RIMS-seq identifies known and novel methylase specificities in synthetic microbial communities. Overall, we found motifs for 6 out of the 12 gut microbiome species and five out of the six skin microbiome species (Supplementary Table S3). The motifs range from four to eight nucleotides long and 70% are palindromic. Interestingly, we found an unknown palindromic motif GGCSGCC (with S being either C or G) from *Micrococcus luteus* (NC_012803.1) in the skin community. To our knowledge, this is the first time this 7nt motif is identified, showing the potential of RIMS-seq to identify new methylase specificities. Results obtained with RIMS-seq were also validated using bisulfite sequencing. RIMS-seq identified two motifs in *Helicobacter pylori* from the ATCC synthetic gut microbiome: GCGC as well as an additional non-palindromic motif CCTC that was identified by the bisulfite analysis pipeline as CYTC with Y being either C or T. The CCTC motif is very common in *Helicobacter pylori*s species, it has been described to be modified at m5C on one strand, while modified at m6A on the other strand (4). In order to confirm the RIMS-seq motif, we investigated the bisulfite-seq data and compared the methylation level in cytosines present in the CCTC context versus cytosines in any other context. We see a methylation level above 90% at the cytosines in the CCTC context confirming the existence of this methylated motif in *Helicobacter pylori* (Figure 5D). Interestingly, m4C methylation in *Helicobacter pylori* has been shown to also occur at TCTTC (26), resulting in the composite motif CYTC (TCTTC and NCCTC) found in the bisulfite data. Contrary to bisulfite, RIMS-seq does not identify m4C methylation (27), hence the identification of the CCTC motif instead.

Also, interestingly, bisulfite-seq results indicate that the ACGT motif in *Acinetobacter johnsonii* and *Streptococcus mitis* from the ATCC synthetic skin microbiome are not fully methylated (Figure 5B). Most of the sites in *Acinetobacter johnsonii* show a methylation of about 10% while in *Streptococcus mitis*, the average methylation per site is 23% (Figure 5C). These results highlight that despite the low methylation levels, RIMS-seq is able to detect the ACGT motif at high significance (P-value $< 1e^{-100}$). We took advantage of the fact that Streptococcus mitis has two methylated motifs, ACGT and GCNGC with an average methylation per site at 23% and 91% respectively (Figure 5C) to evaluate the sequencing depth required for RIMS-seq to

de-novo identify both motifs. As expected, the fully methylated GCNGC motif is found using 4 times fewer sequencing reads than the ACGT motif, with a required 1 million and 4 million mapped reads respectively (Supplementary Figure S6A and B).

DISCUSSION

In this study, we developed RIMS-seq, a sequencing method to simultaneously obtain high quality genomic sequences and discover m5C methylase specificity(ies) in bacteria using a single library preparation. The simplicity of the procedure makes RIMS-seq a cost effective and time saving method with only an additional 3 h sodium hydroxide incubation and an additional column-based cleaning step. Theoretically, the cleaning step can be avoided if a small volume of the library is used for the amplification step, but we have not tested this procedure. By increasing the sodium hydroxide concentration to 0.5M or even 1M, the incubation time can be reduced to 30 min.

Due to the limited deamination rate, RIMS-seq is equivalent to short read DNA-seq in terms of sequencing quality. Sequencing QC metrics such as coverage, GC content and mapping rate are similar for RIMS-seq and DNA-seq. Thus, RIMS-seq can be used for applications such as, but not limited to, shotgun sequencing, genome assembly and estimation of species composition of complex microbial communities. This dual aspect of RIMS-seq is analogous to SMRT sequencing for which methylation is inferred from the IPD ratio. We showed that both PacBio and RIMS-seq can be complementary with the ability to obtain a complete methylome: m6A and m4C methylase specificities can be obtained from SMRT sequencing while m5C methylase specificity can be obtained from RIMS-seq. Combining both sequencing technologies also allows for a hybrid assembly strategy resulting in closed reference genomes of high sequencing accuracy.

We applied RIMS-seq to several bacteria and identified a variety of methylation motifs, ranging from 4 nt to 8 nt long, palindromic and non-palindromic. Some of these motifs were identified for the first time, demonstrating the potential of the technology to discover new methylase specificities, from known as well as from unknown genomes. We also validated that RIMS-seq can identify multiple methylase specificities from a synthetic microbial community and estimate species abundances. However, RIMS-seq has caveats similar to metagenomics sequencing when applied to study natural microbial communities. Closely related species are likely to co-exist and assigning the motif to the correct species can be challenging. Furthermore, single nucleotide polymorphisms found in microbial communities may confound the identification of the C to T deamination, increasing the background noise for the detection of motifs. Finally, species in microbiomes are unevenly represented which can cause RIMS-seq to identify motifs only in the most abundant species.

Because RIMS-seq is based on a limited deamination, it requires the combined signal over many reads to be large enough to effectively identify methylase specificity. For the vast majority of the methylases in RM systems, methylation is present at enough sites across the genome for RIMS-seq to determine their specificities. Nonetheless, bacterial methylases can be involved in other processes such as, but not limited to, DNA mismatch repair (28), gene regulation (29) and sporulation (30) and the recognition sites may not necessarily be fully methylated. Partially methylated sites can be found using RIMS-seq but more analysis needs to be done to evaluate how pervasive methylation needs to be to provide a RIMS-seq signal. In other cases, methylated motifs are too specific or under purifying selection, resulting in just a handful of sites in the genome. In these cases, RIMS-seq signals can only be obtained with enough read coverage to compensate for the scarcity of those sites. While the methylase specificities are of great interest in bacteria due to their diversity in recognition sequences, applying RIMS-seq to humans would lead to the identification of the already well-described CpG context. In this case, other technologies such as EM-seq or bisulfite-seq are more appropriate as they enable the precise genomic location to be obtained.

In summary, RIMS-seq is a new technology allowing the simultaneous investigation of both the genomic sequence and the methylation in prokaryotes. Because this technique is easy to implement and shows similar sequencing metrics to DNA-seq, RIMS-seq has the potential to substitute DNA-seq for microbial studies.

DATA AVAILABILITY

The data have been deposited with links to BioProject accession number PRJNA706563 in the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/).

Custom-built bioinformatics pipelines to analyse sequencing reads from RIMS-seq are available at https://github.com/Ettwiller/RIMS-seq/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Peter Weigele and Yian-Jiun Lee from New England Biolabs for the Xp12 genomic DNA and genomic sequence. We thank Ivan Correa and Nan Dai for their assistance with LC–MS and Ira Schildkraut for his help with methylase specificities.

FUNDING

Funding for open access charge: New England Biolabs.

Conflict of interest statement. C.B., Y.C.L., A.F., B.P.A., L.C., T.C.E., R.R. and L.E. are or were employees of New England Biolabs Inc. a manufacturer of restriction enzymes and molecular reagents.

REFERENCES

1. Loenen,W.A.M., Dryden,D.T.F., Raleigh,E.A., Wilson,G.G. and Murray,N.E. (2014) Highlights of the DNA cutters: a short history of the restriction enzymes. *Nucleic Acids Res.*, **42**, 3–19.
2. Blow,M.J., Clark,T.A., Daum,C.G., Deutschbauer,A.M., Fomenkov,A., Fries,R., Froula,J., Kang,D.D., Malmstrom,R.R., Morgan,R.D. *et al.* (2016) The epigenomic landscape of prokaryotes. *PLoS Genet.*, **12**, e1005854.

3. Beaulaurier,J., Schadt,E.E. and Fang,G. (2019) Deciphering bacterial epigenomes using modern sequencing technologies. *Nat. Rev. Genet.*, **20**, 157–172.
4. Roberts,R.J., Vincze,T., Posfai,J. and Macelis,D. (2015) REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.*, **43**, D298–D299.
5. Flusberg,B.A., Webster,D.R., Lee,J.H., Travers,K.J., Olivares,E.C., Clark,T.A., Korlach,J. and Turner,S.W. (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods*, **7**, 461–465.
6. Clark,T.A., Lu,X., Luong,K., Dai,Q., Boitano,M., Turner,S.W., He,C. and Korlach,J. (2013) Enhanced 5-methylcytosine detection in single-molecule, real-time sequencing via Tet1 oxidation. *BMC Biol.*, **11**, 4.
7. Tse,O.Y.O., Jiang,P., Cheng,S.H., Peng,W., Shang,H., Wong,J., Chan,S.L., Poon,L.C.Y., Leung,T.Y., Chan,K.C.A. *et al.* (2021) Genome-wide detection of cytosine methylation by single molecule real-time sequencing. *Proc. Natl. Acad. Sci. U.S.A.*, **118**, e2019768118.
8. Sun,Z., Vaisvila,R., Hussong,L.-M., Yan,B., Baum,C., Saleh,L., Samaranayake,M., Guan,S., Dai,N., Corrêa,I.R. Jr *et al.* (2021) Nondestructive enzymatic deamination enables single-molecule long-read amplicon sequencing for the determination of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Genome Res.*, **31**, 291–300.
9. Liu,Y., Siejka-Zielińska,P., Velikova,G., Bi,Y., Yuan,F., Tomkova,M., Bai,C., Chen,L., Schuster-Böckler,B. and Song,C.-X. (2019) Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. *Nat. Biotechnol.*, **37**, 424–429.
10. Anton,B.P., Fomenkov,A., Wu,V. and Roberts,R.J. (2021) Genome-wide identification of 5-methylcytosine sites in bacterial genomes by high-throughput sequencing of MspJI restriction fragments. *PLoS One*, **16**, e0247541.
11. Simpson,J.T., Workman,R.E., Zuzarte,P.C., David,M., Dursi,L.J. and Timp,W. (2017) Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods*, **14**, 407–410.
12. Rand,A.C., Jain,M., Eizenga,J.M., Musselman-Brown,A., Olsen,H.E., Akeson,M. and Paten,B. (2017) Mapping DNA methylation with high-throughput nanopore sequencing. *Nat. Methods*, **14**, 411–413.
13. Tourancheau,A., Mead,E.A., Zhang,X.-S. and Fang,G. (2021) Discovering multiple types of DNA methylation from bacteria and microbiome using nanopore sequencing. *Nat. Methods*, **18**, 491–498.
14. Fogg,M.J., Pearl,L.H. and Connolly,B.A. (2002) Structural basis for uracil recognition by archaeal family B DNA polymerases. *Nat. Struct. Biol.*, **9**, 922–927.
15. Duncan,B.K. and Miller,J.H. (1980) Mutagenic deamination of cytosine residues in DNA. *Nature*, **287**, 560–561.
16. Wang,R.Y., Kuo,K.C., Gehrke,C.W., Huang,L.H. and Ehrlich,M. (1982) Heat- and alkali-induced deamination of 5-methylcytosine and cytosine residues in DNA. *Biochim. Biophys. Acta*, **697**, 371–377.
17. Chen,L., Liu,P., Evans,T.C. and Ettwiller,L.M. (2017) DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science*, **355**, 752–756.
18. Kuo,T.T., Huang,T.C. and Teng,M.H. (1968) 5-Methylcytosine replacing cytosine in the deoxyribonucleic acid of a bacteriophage for Xanthomonas oryzae. *J. Mol. Biol.*, **34**, 373–375.
19. Marinus,M.G. and Morris,N.R. (1973) Isolation of deoxyribonucleic acid methylase mutants of Escherichia coli K-12. *J. Bacteriol.*, **114**, 1143–1150.
20. Palmer,B.R. and Marinus,M.G. (1994) The dam and dcm strains of Escherichia coli–a review. *Gene*, **143**, 1–12.
21. Crooks,G.E., Hon,G., Chandonia,J.-M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
22. Marschall,T. and Rahmann,S. (2009) Efficient exact motif discovery. *Bioinformatics*, **25**, i356–i364.
23. Vasu,K. and Nagaraja,V. (2013) Diverse functions of restriction-modification systems in addition to cellular defense. *Microbiol. Mol. Biol. Rev.*, **77**, 53–72.
24. Grehl,C., Wagner,M., Lemnian,I., Glaser,B. and Grosse,I. (2020) Performance of mapping approaches for whole-genome bisulfite sequencing data in crop plants. *Front. Plant Sci.*, **11**, 176.
25. Olova,N., Krueger,F., Andrews,S., Oxley,D., Berrens,R.V., Branco,M.R. and Reik,W. (2018) Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. *Genome Biol.*, **19**, 33.
26. Vitkute,J., Stankevicius,K., Tamulaitiene,G., Maneliene,Z., Timinskas,A., Berg,D.E. and Janulaitis,A. (2001) Specificities of eleven different DNA methyltransferases of *Helicobacter pylori* strain 26695. *J. Bacteriol.*, **183**, 443–450.
27. Vilkaitis,G. and Klimasauskas,S. (1999) Bisulfite sequencing protocol displays both 5-methylcytosine and N4-methylcytosine. *Anal. Biochem.*, **271**, 116–119.
28. Modrich,P. and Lahue,R. (1996) Mismatch repair in replication fidelity, genetic recombination, and cancer biology. *Annu. Rev. Biochem.*, **65**, 101–133.
29. Casadesús,J. and Low,D. (2006) Epigenetic gene regulation in the bacterial world. *Microbiol. Mol. Biol. Rev.*, **70**, 830–856.
30. Oliveira,P.H., Ribis,J.W., Garrett,E.M., Trzilova,D., Kim,A., Sekulovic,O., Mead,E.A., Pak,T., Zhu,S., Deikus,G. *et al.* (2020) Epigenomic characterization of Clostridioides difficile finds a conserved DNA methyltransferase that mediates sporulation and pathogenesis. *Nat. Microbiol.*, **5**, 166–180.
31. Nurk,S., Bankevich,A., Antipov,D., Gurevich,A.A., Korobeynikov,A., Lapidus,A., Prjibelski,A.D., Pyshkin,A., Sirotkin,A., Sirotkin,Y. *et al.* (2013) Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J. Comput. Biol.*, **20**, 714–737.
32. Vacic,V., Iakoucheva,L.M. and Radivojac,P. (2006) Two sample logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics*, **22**, 1536–1537.
33. Emms,D.M. and Kelly,S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.*, **20**, 238.
34. Letunic,I. and Bork,P. (2021) Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.*, **49**, W293–W296.

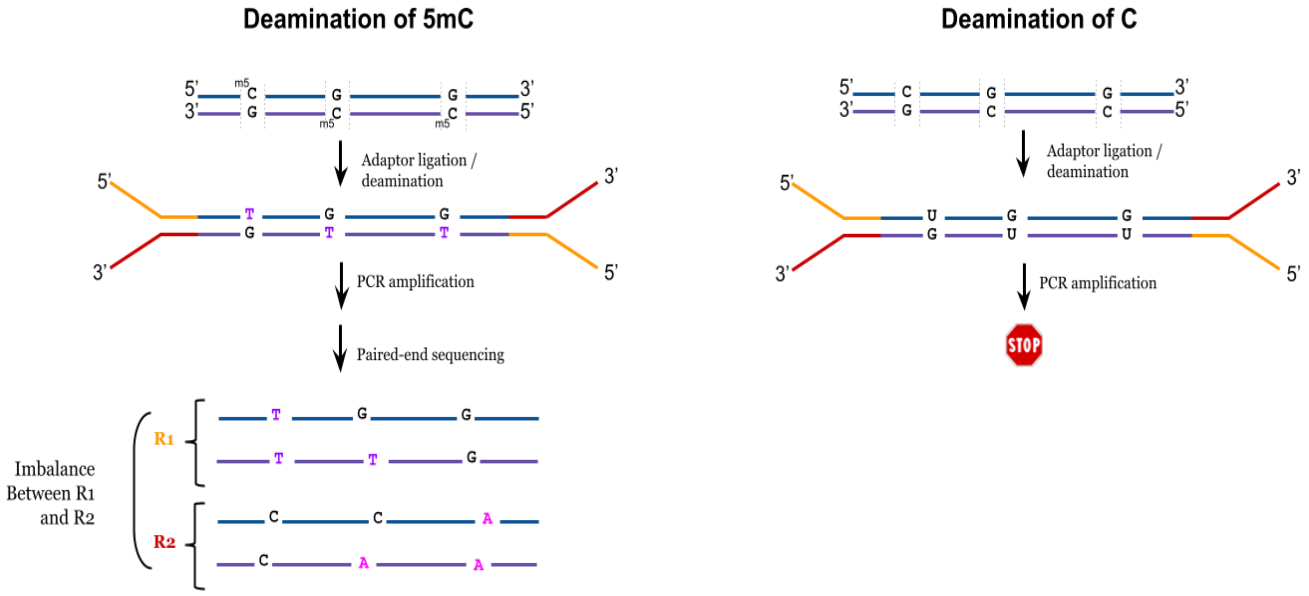
Supplementary material: Rapid Identification of Methylase Specificity (RIMS-seq) jointly identifies methylated motifs and generates shotgun sequencing of bacterial genomes

Chloé Baum, Yu-Cheng-Lin, Alexey Fomenkov, Brian P. Anton, Lixin Chen, Bo Yan, Thomas C. Evans Jr, Richard J Roberts, Andrew C Tolonen, Laurence Ettwiller

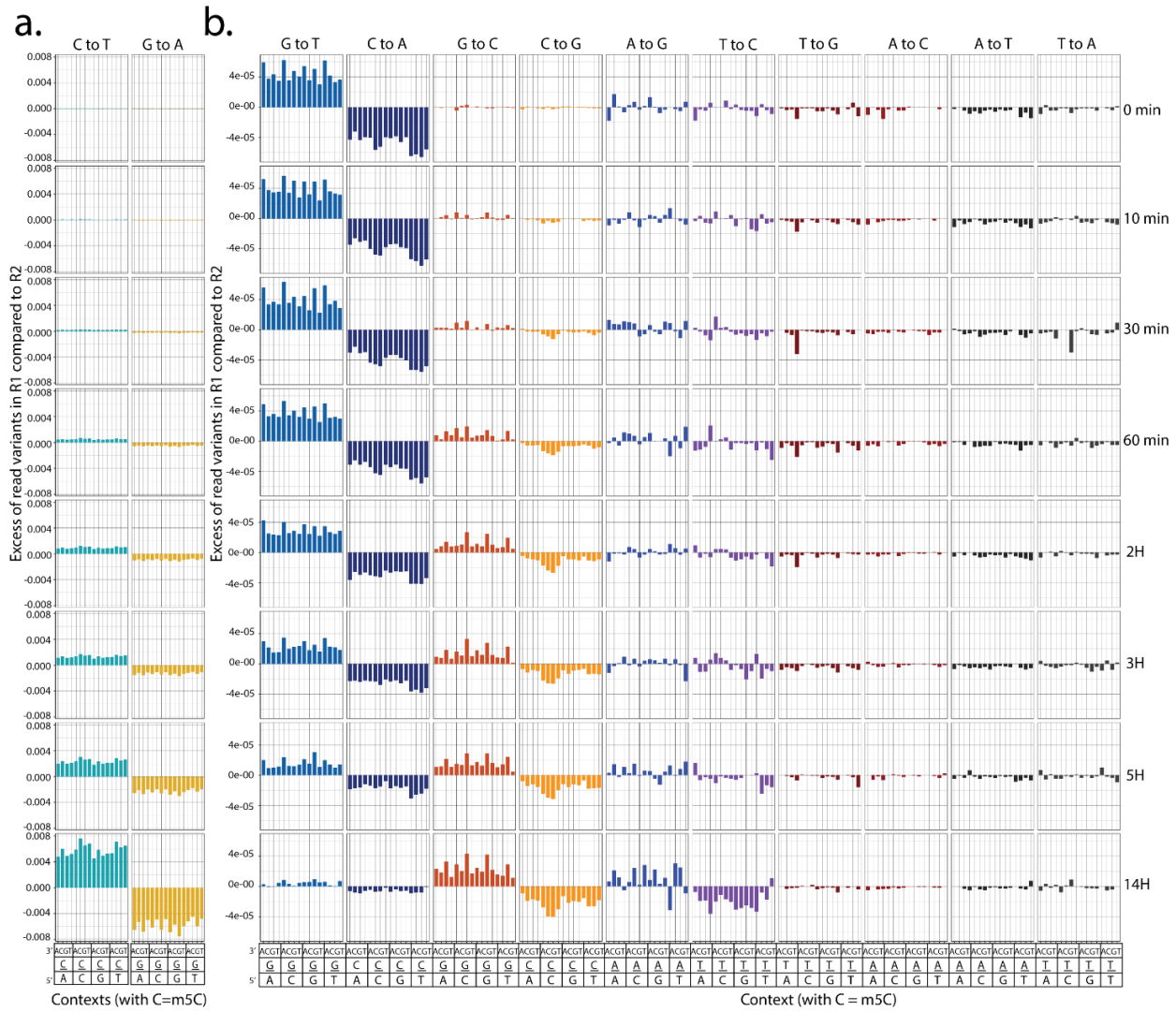
Supplementary text 1:

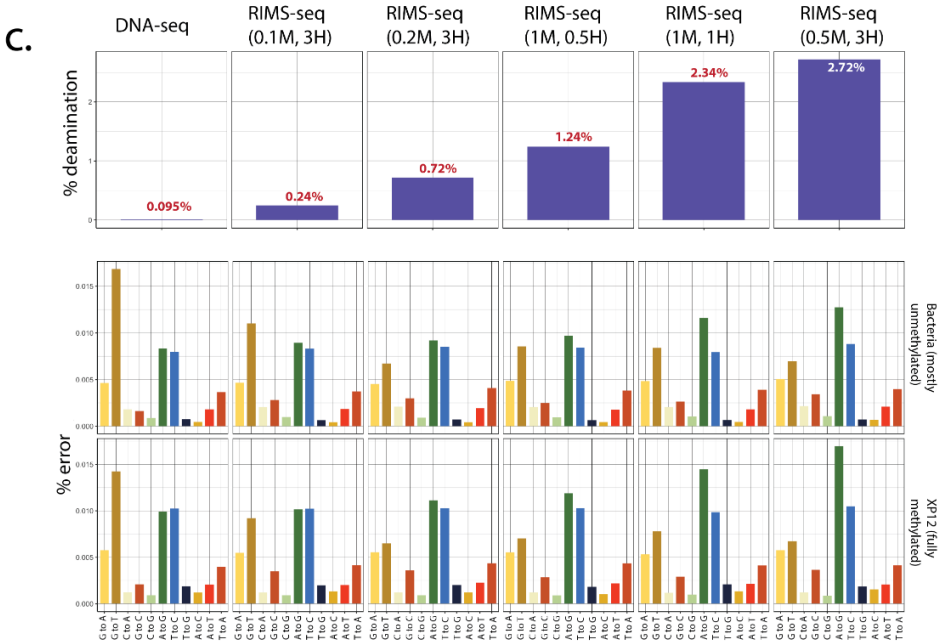
We designed an oligonucleotide containing a single 8-oxo-dG (/5FAM/TGGAGATTTTGATCACGGTAACC/i8oxodG/ATCAGAATGACAACAAGCCCGAATTCACCCAGGAGG/3Rox_N/). 50 uM of this 8-oxo-dG containing oligonucleotide was treated with 0.1M NaOH at 60C for 3 hours or 16 hours, respectively. Following alkaline and heat treatment, all reactions were neutralized with acetic acid and a clean-up with Monarch PCR & DNA cleanup kit (NEB, Ipswich). All cleanup DNA eluted with 20 ul of water were subjected for LC-MS analysis. Untreated 8-oxo-dG oligonucleotide was also subjected to LC-MS in the same run (**Material and Methods**).

In the two alkaline and heat treatment conditions tested, LC-MS confirms two DNA fragments which are the strand-breaks products at 8-oxo-dG: 1) 7692.309 Dalton, which is the mass of 5'FAM-DNA fragment with a 3'-phosphate end before 8-oxo-dG; 2) 11867.1510 Dalton, which are the mass of 3' ROX-DNA fragment with a 5'-phosphate end after 8-oxo-dG. A 11139.888 Dalton peak was also detected, which is the 3'-DNA fragment without 3'Rox. Conversely, the untreated 8-oxo-dG oligonucleotide LC-MS shows a single peak corresponding to the intact oligonucleotide (**Supplementary Figure 3**).

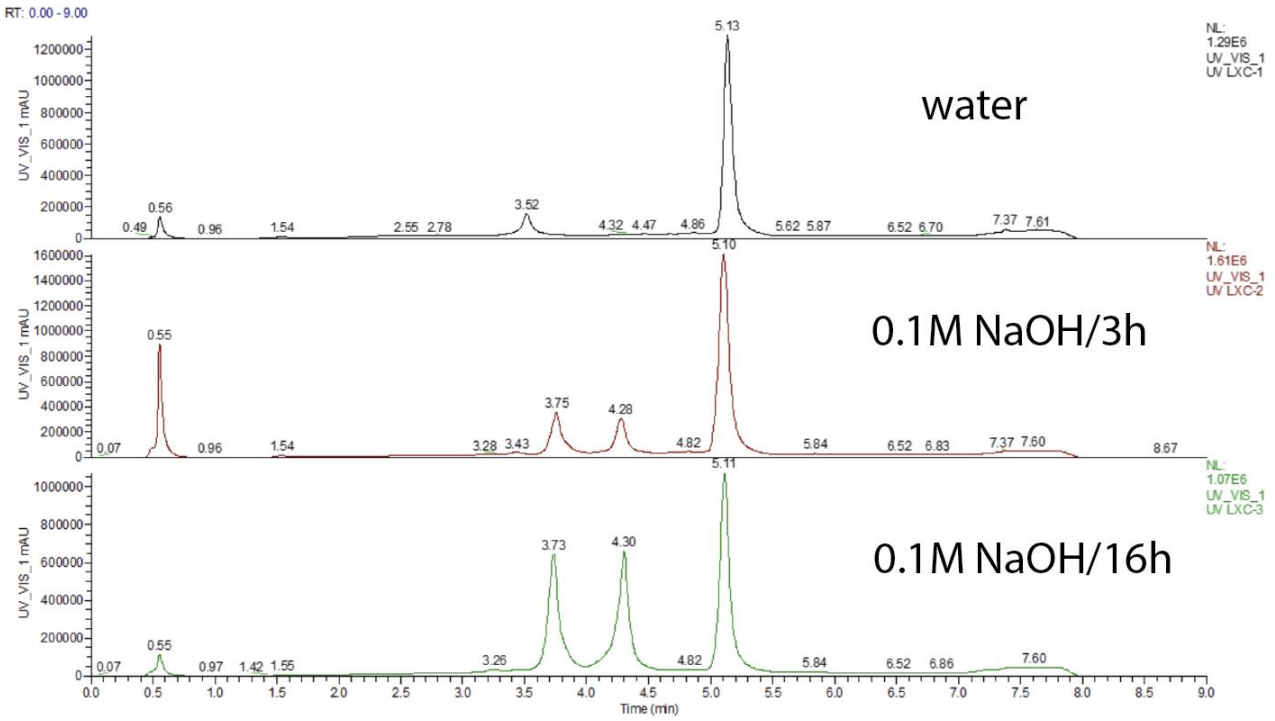


Supplementary Figure 1: Schema describing the sequencing of 5mC and C containing DNA fragments. After DNA fragmentation and adaptor ligation (see methods), DNA fragments are subjected to a limited deamination under heat alkaline conditions. Deamination converts m5C to T (left schema) and C to U (right schema). U is a blocking damage for the polymerase, thus fragments containing U will not be amplified and sequenced. Conversely, fragments containing deaminated 5mC (T) are amplified and sequenced leading to C to T variants. Because of the sequencing directionality of Illumina library, sequencing from the forward adaptor (Read 1) corresponds to the original strand; while sequencing from the reverse adaptor (Read2), corresponds to the reverse complement of the original strand. Thus, Read 1 shows an excess of C to T read variants compared to Read 2 (that shows an excess of G to A variant instead) leading to an imbalance of C to T/G to A variants. This imbalance is directly proportional to the deamination rate. In fragments containing a mixture of C and 5mC, the C to T/G to A imbalance can only be observed at methylated sites.

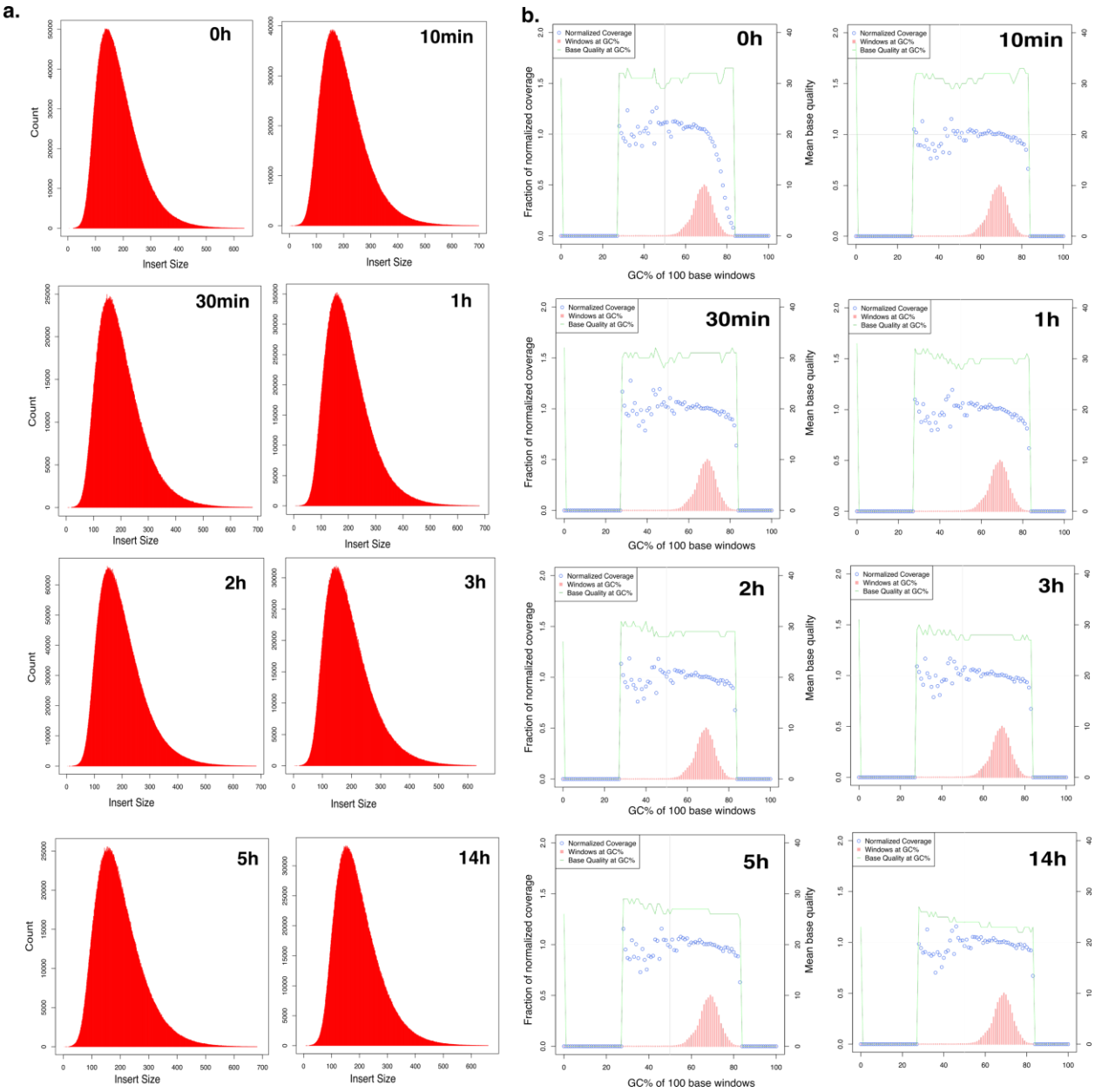

a.
b.
C to T
G to A
G to T
C to A
G to C
C to G
A to G
T to C
T to G
A to C
A to T
T to A
0 min
10 min
30 min
60 min
2H
3H
5H
14H
Excess of read variants in R1 compared to R2
Contexts (with C=m5C)
Context (with C = m5C)


c.
DNA-seq
RIMS-seq (0.1M, 3H)
RIMS-seq (0.2M, 3H)
RIMS-seq (1M, 0.5H)
RIMS-seq (1M, 1H)
RIMS-seq (0.5M, 3H)
% deamination
0.095%
0.24%
0.72%
1.24%
2.34%
2.72%
% error
Bacteria (mostly unmethylated)
XP12 (fully methylated)

Supplementary Figure 2 Imbalance indicative of damage between R1 and R2 in a fully methylated genome (XP12) after RIMS-seq. **a.** Excess of read variants in R1 compare to R2 for control (t=0) and various heat alkaline treatment times (t= 10, 30 minutes, 1,2,3,5 and 14 hours) for C to T and G to A. G to A values are mirroring C to T because of the imbalance. **b.** Same as **a.** for all the other substitutions. Note that the Y-axis scale between **a.** and **b.** is different and the C to T excess in R1 is up to a 100-fold greater than for the other substitutions. Time scale represents the heat-alkaline incubation time (in minutes or hours). X-axis represents the different genomic context NNN context with N being A, T, m5C or G. c. Deamination rates (calculated as the % of C to T in excess in R1 compared to R2 in XP12) for 0.1, 0.2, 0.5 and 1M NaOH 60 degree C at various times (ranging from 0.5 to 3 hours, top panels) and the respective error rate for each substitution (calculated as the % of read variants, bottom panels) for bacteria (*Haemophilus influenzae* Rd ATCC 51907, mostly unmethylated) and XP12 (fully methylated).



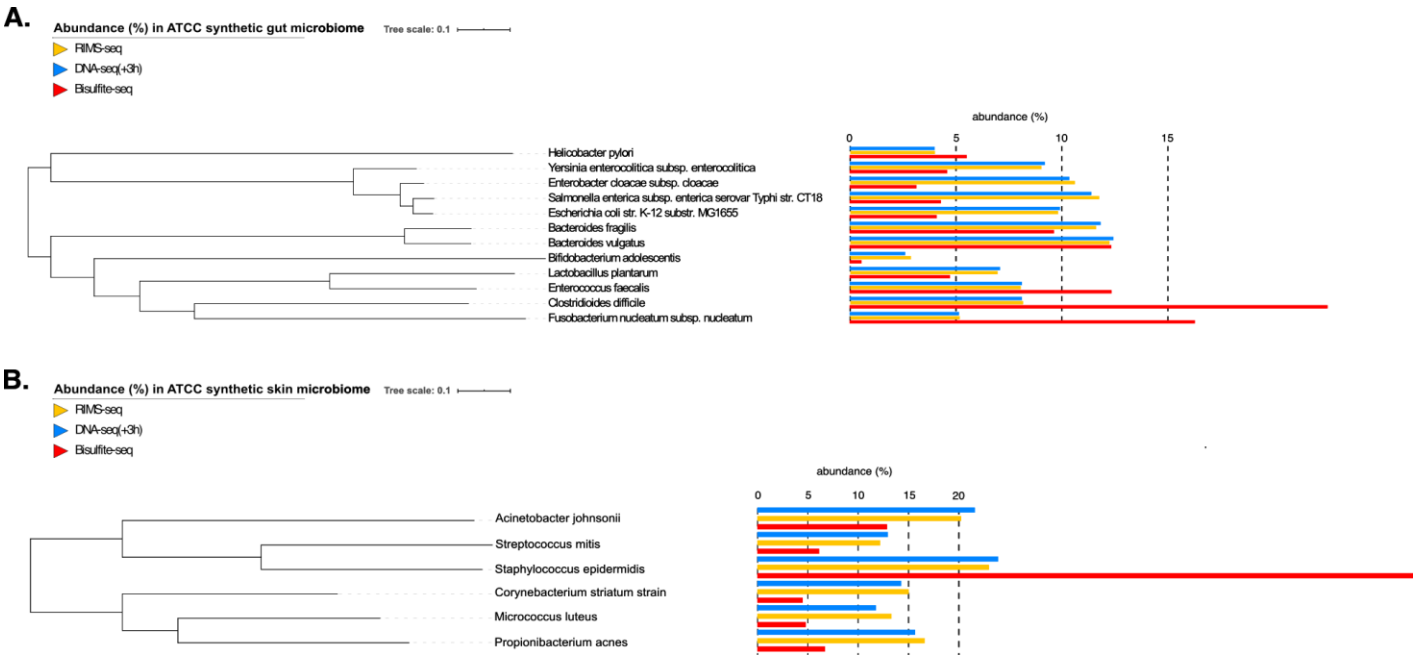
Supplementary Figure 3: LC-MS result



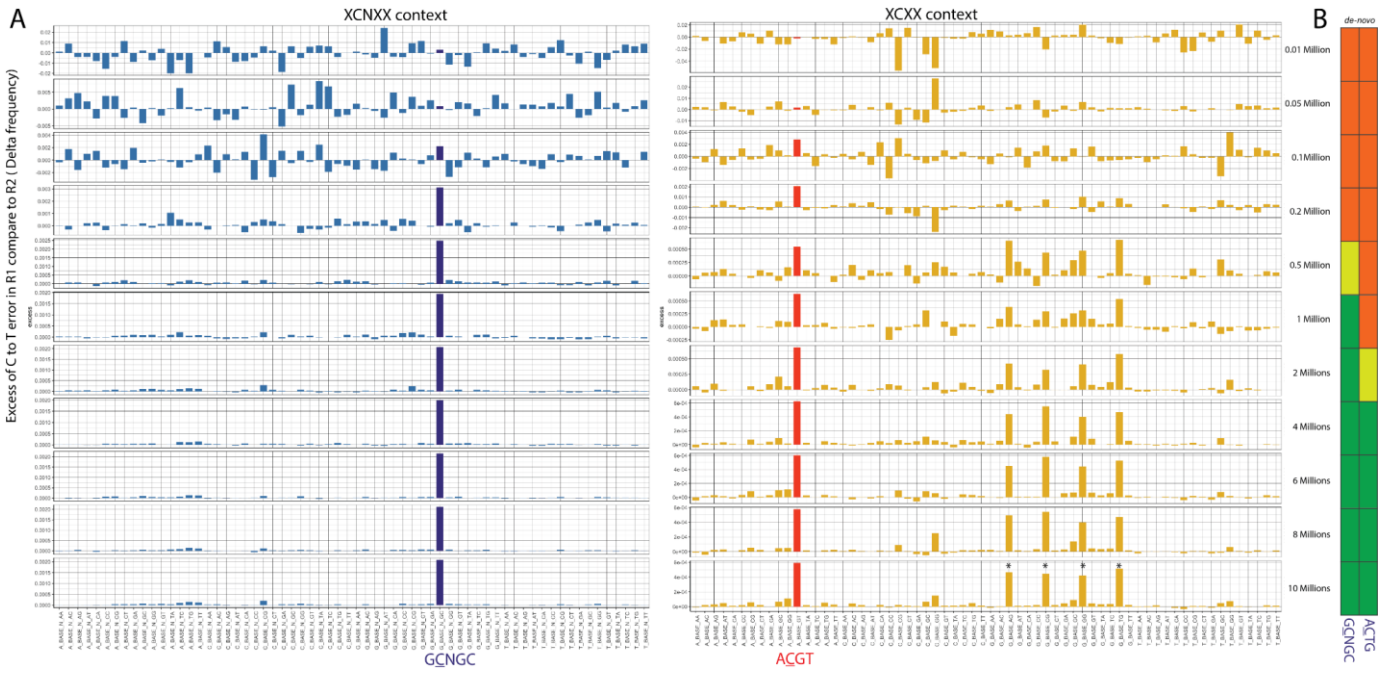
Supplementary Figure 4: Quality control of the sequencing performances for Xp12 gDNA

a. Insert size distribution (bp) for the control (0h) and different heat-alkaline treatment times. Input genomic DNA was sheared by Covaris treatment (200 bp target) and no further size selection was applied.

b. GC bias for the control (0h) and different heat-alkaline treatment times.



Supplementary Figure 5: Phylogenetic tree and barplot showing the relative abundance of each species in the ATCC synthetic (A.) gut and (B.) skin microbiome for RIMS-seq (yellow), DNA-seq(+3H, blue) and bisulfite sequencing (red). The relative abundance is calculated using the number of reads mapping to each species normalized to the total number of mapped reads.



Supplementary Figure 6: A. Barplots showing the excess of C to T error in R1 compared to R2 (y-axis) in XCXX (left) and XCNXX (right) contexts (with X=either A,T,C or G and N being any nucleotides). The dataset has been downsampled to various numbers of read mapping to Streptococcus mitis ranging from 0.01 (top) to 10 million (bottom) reads. The ACGT and GCNGC context is highlighted in red and dark blue respectively. Asterix (*) denote the GCAGN, GCTGN,GCCGN,GCGGN motifs which correspond to a subset of the GCNGC motif. **B.** De-novo identification of the ACGT and GCNGC motifs. Red denotes motif not found, yellow denotes the correct motif was found with p-values between 1e-50 and 1e-100 and green denotes the correct motif was found with p.values < 1e-100.

Alakline-heat treatment time	0h	10min	30min	1h	2h	3h	5h	14h
Statistics without reference genome								
nb contigs	1	1	1	1	1	1	1	1
largest contig	63783	63882	64181	63774	63873	63839	63873	63774
Statistics with reference genome								
largest alignment	63783	63782	63774	63774	63285	63839	63773	63774
total aligned length	63783	63782	63774	63774	63285	63839	63773	63774
GC %	68.17	68.17	68.17	68.17	68.17	68.18	68.17	68.17
N50	63783	63882	63774	63774	63873	63839	63873	63774
Genome fraction (%)	99.239	99.238	99.225	99.225	99.224	99.241	99.224	99.225
% reads mapping back to assembly	99.75	99.73	99.8	99.76	99.82	99.78	99.75	99.84
Misassemblies								
nb misassemblies	0	0	0	0	0	0	0	0
misassembled contig length	0	0	0	0	0	0	0	0
local misassemblies	1	1	1	1	1	1	1	1
Mismatches								
N's per 100kbp	0	156.14	0	0	156.56	0	156.56	0
nb mismatches per 100kbp	3.14	3.14	1.57	1.57	3.14	4.7	1.57	3.14
nb indels per 100kbp	0	0	0	0	0	0	0	0

Supplementary Table 1: Xp12 assembly statistics for various heat/alkaline treatment times (see Material and Methods)

	dnaseq 3h	RIMS	reference
span (bp)	3,540,609	3,515,857	3,543,981
N (%)	0.02	0.01	0.00
GC (%)	38.62	38.59	38.31
AT (%)	61.38	61.41	61.69
scaffold count	571	531	1
longest scaffold (bp)	317,976	318,149	3,543,981
scaffold N50 length (bp)	77,278	62,717	3,543,981
scaffold N50 count	14	15	1
scaffold N90 length (bp)	14,355	13,679	3,543,981
scaffold N90 count	52	59	1
contig count	588	540	1
contig N50 length (bp)	62,601	58,046	3,543,981
contig N50 count	15	16	1
contig N90 length (bp)	13,997	13,455	3,543,981
contig N90 count	57	63	1

Supplementary Table 2: Assembly statistics for *Acinetobacter calcoaceticus* ATCC 49823 assemblies obtained using the sequences from the standard DNA-seq (+3H) and RIMS-seq, compared to the reference genome.

Skin microbiome	Chromosome accession	GC content (%)	Bisulfite-seq motif	RIMS-seq motif	p-value RIMS-seq
Micrococcus luteus	NC_012803.1	73	GGCSGCC	GGCSGCC	7.965455e-572
				CGSNNW	1.74E-154
Propionibacterium acnes	NC_006085.1	60	NA	CSNNNNCG	2.34E-132
Corynebacterium striatum	NZ_CP021252.1	59	GCCGGC	GCCGGC	1.37E-268
				CNNYRNNG	4.72E-135
Acinetobacter johnsonii	NZ_CP010350.1	41	ACGT	ACGT	1.19E-105
				ATCNNGRC	3.12E-123
Streptococcus mitis	NC_013853.1	40	GCNGC	GCNGC	1.540920e-603
			ACGT*	ACGT	6.07E-108
Staphylococcus epidermidis	NC_004461.1	32	NA	NA	NA
Gut microbiome	**Chromosome accession**	**GC content (%)**	**Bisulfite-seq motif**	**RIMS-seq motif**	**p-value RIMS-seq**
Bifidobacterium adolescentis	NC_008618.1	59	GATC	GATC	7.609400e-718
			CCNGG	CCNGG	3.970975e-364
Enterobacter cloacae	NC_014121.1	55	CCWGG	CCWGG	3.129282e-1371
Salmonella enterica	NC_003198.1	52	CCWGG	CCWGG	1.071714e-1331
Escherichia coli K12	NC_000913.3	51	CCWGG	CCWGG	2.277395e-1220
Yersinia enterocolitica	NC_008800.1	47	CCWGG	CCWGG	2.314551e-649
Lactobacillus plantarum	NC_004567.2	44	NA	NA	NA
Bacteroides fragilis	NC_006347.1	43	NA	NA	NA
Bacteroides vulgatus	NC_009614.1	42	NA	NA	NA
Helicobacter pylori	NC_000915.1	39	GCGC	GCGC	2.291291e-516
			CYTC*	CCTC	2.387073e-318
Enterococcus faecalis	NC_004668.1	38	NA	NA	NA
Clostridioides difficile	NC_009089.1	29	NA	NA	NA
Fusobacterium nucleatum	NC_003454.1	27	NA	NA	NA

Supplementary Table 3: Methylases specificity of the synthetic ATCC microbiomes. Motifs followed by an asterisk (*) are motifs for which the bisulfite analysis pipeline had to be adapted to find the motif. For *Streptococcus mitis*, the ACGT sites were found to be poorly methylated genome-wide (15% were fully or hemi-methylated). For Helicobacter pylori, literature describes 2 motifs CCTC and TCTTC https://www.ncbi.nlm.nih.gov/pmc/articles/PMC94898/ with an m5C and m4C, respectively. Bisulfite cannot distinguish between m5C and m4C, so it is likely the CYTC motif picked up in the bisulfite data shown in the table results from a composite between the m5C and m4C motifs of H. pylori.

B. Appendix from Chapter II

ONT-Cappable-seq protocol – splint polyA ligation

1. Capping

total RNA	10µg
Vaccinia Capping enzyme (M2080)	5µl
10X Vaccinia Capping buffer (M2080)	5µl
5mM Destiobiotin-GTP (N0761)	5µl
E.Coli IPP (M0361)	5µl
H2O	qsp
Total	50µl

37°C for 1h
Purify with a Zymo Clean and Concentrator 5G column (standard protocol total RNA).
Wash 4 times with wash buffer: 2 times 700µL, 1min + 2 times 400µL, 1min. Centrifuge one more time 2min to dry. All centrifugations @12000G
Elute with 40µl low-TE (in 2 times 20µL). Wait 2min before centrifugation for 1min.

2. Addition of a polyA tail to the 3'end of transcript

Capped Ecoli RNA	40µl
Ecoli polyA polymerase (M0276)	1µl
10X polyA polymerase buffer (M0276)	5µl
10mM ATP (M0276)	5µl
Total	50µl

Incubate at 37°C for 15min.

Purify the reaction with 1.0X Ampure beads (=50µl) → Capped-tailed RNA
Incubate sample with beads for 5min with 300 rpm agitation
Wash 2 times with 200µL EthOH 80%
Spin down to remove EtOH remaining and dry for ~ 10min
Elute in 34 µL low TE for 5min with 300 rpm agitation

- Use 30µl for enrichment
- Save 1µl as Control RNA (Enrich:Control 10:1).
  NB: save the control in a PCR tube with the volume of H2O required for the later RT reaction (10µL).
  This is to avoid evaporation.

3. Enrichment

2X Binding Buffer	Elution buffer	Wash buffer
250mM NaCl	50mM NaCl	60 mM NaCl
10mM Tris-HCl pH 7.5	10mM Tris-HCl pH 7.5	10 mM Tris-HCl pH 7.5
1mM EDTA	0.1mM EDTA	1mM EDTA
	1mM biotin	

Prepare 35µL of streptavidin beads:

- Wash 2 times with 400 µl of Wash buffer
- Wash 2 times with 400 µl of 2XBinding buffer
- Resuspend in 35uL 2X Binding buffer

Add 30ul Capped-tailed RNA to 30µL prepared beads.
Incubate at room temperature for 30min on a hula mixer.
Wash the beads 3 times with 400µL **Wash Buffer.**
Elute RNA from beads by incubating the beads with 25µl **Elution Buffer** at 37°C for 30min – 1h.
Collect the biotin-eluted RNA by placing the tube on the magnetic rack -> **get 25µL of enriched RNA**

4. Splint Ligation to 3'end

Prepare the adapter
Dilute top and bottom strands of the adapters in annealing buffer (10mM Tris, 50mM NaCl, 0.1 mM EDTA). Anneal with 10µM of the top strand and 20uM of the bottom strand.

Heat the mixed DNA adapters at 95°C for 5 min followed by 70°C for 1minute. Start cycling at 95°C until 25°C and decrease the temperature by 1°C each cycle. After ~70 min the temperature will go down to 25°, the adapters are annealed. Adapters can be stored at -20°C and used indefinitely, don't need to anneal a new batch each time. If in a hurry for annealing, just heat the adapters to 95°C and then do a slow ramp (.1 degree per second) to 4°C. This takes about 15 min and also seems to work just fine. 15µM final.

Ligation protocol

	Ci	Cf	Vol (uL)
Enriched RNA	/	/	25
Adapter	15µM	1µM	4
T4 RNAl2 buffer			5
PEG	50%	10%	10
T4 RNAl2 enzyme	/	/	3
H2O Vf=50uL	/	/	3

	Ci	Cf	Vol (uL)
Control RNA	/	/	11 (1µL + 10µL H2O)
Adapter	15µM	1µM	4
T4 RNAl2 buffer			5
PEG	50%	10%	10
T4 RNAl2 enzyme	/	/	3
H2O Vf=50uL	/	/	17

Incubate 1h at 25°C in thermomix

5. USER treatment
Use 3.4µL endoIV + 1.6µL UDG for 50uL. (2µL endoIV + 1µL UDG for one reaction of 30µL)
Incubate 1h at 37°C

6. Zymo oligo column clean-up (D4060)
Elute in 2x15µL RNase free H2O (30µL final)

7. Reverse transcription, first strand cDNA synthesis
The bottom 3'end adapter serves as the RT oligo.

USER cleaned reaction	30 µL
dNTP	8 µL
5X Protoscript II buffer	10 µL
0.1M DTT	5 µL
Protoscript II (M0368)	2 µL
Murine Rnase inhibitor (M0314)	1 µL
H1O Vf=60uL	4 µL

Incubate at 42℃ for 1h (lid at 50℃).

For Enrich group, add 2µl RNaseIf (M0243) and incubate at 37°C for 30min to 1h;
then purify the reaction with 1.0X Ampure beads (=60µL beads), elute with 23µl Low-TE.

For Control group, DO NOT do the RNaseIf. Directly purify the reaction with 1.0X Ampure beads(=60µL beads), elute with 23µl Low-TE (Vf=22µL)

8. Addition of a polyG linker to the 3'end of synthesized cDNA

cDNA	22µl
Terminal transferase (M0315)	1µl
10X TdT buffer	3µl
100mM dGTP	1µl
2.5mM CoCl2	3µl
Total	30µL

Incubate at 37°C for 30min.

9. 2nd enrichment of the sample & AMPure purification of the control
Enrich sample:
Add 10µL water to the enriched sample to qsp 40µL.
Add 30µl prepared streptavidin beads to the enriched sample
Incubate at room temperature for 30min on a hula mixer
wash the beads 3 times with 400µL **1X Binding buffer**
Elute in 40µL elution buffer -> get enrich cDNA

Control sample:
Purify the TdT reaction with 1.0X AMPure beads. Elute in 40µl low-TE -> get control cDNA

10. Second cDNA strand synthesis

cDNA	20µL	→ 1/2 of the cDNA for the reaction
10µM Pac_oligodc20_for_set2	2µL	
RnaseH (M0297)	2µL	
LongAmp ReadyMix (M0533)	20µL	
Total	44µL	

Cycling:

37ºC	15min	
Flick the tube (beads)		
94ºC	1min	
94ºC	1min	⎫
65ºC	15min	⎭ 4 cycles
65ºC	10min	

11. Amplification of the cDNA

Set up reaction as follow:

	For 1 reaction
ds cDNA	10µL
10µM Pac_fw_set2	2.5µL
10µM Pac_rev	2.5µL
LongAmp ReadyMix (M0533)	25µL
H2O	10µL
Total	50µL

Cycling:

94ºC	1min	
94ºC	30sec	⎫
65ºC	8min	⎭ 4 - 10 cycles
65ºC	10min	

12. AMPure purification 0.6X

13. Second round amplification of the cDNA

According to the previous Qubit results, determine the number of cycles required for this additional round of PCR.

→ use all the 40µL cDNA reaction from first round

Set up reaction as follow:

ds cDNA	20µL
10µM Pac_fw_set2	2.5µL
10µM Pac_rev	2.5µL
LongAmp ReadyMix (M0533)	25µL
H2O	16µL
Total	50µL

Cycling:

94ºC	1min	
94ºC	30sec	1-10 cycles
65ºC	8min	
65ºC	10min	

14. AMPure purification 0.9X

ONT-Cappable-seq protocol – polyU tailing

1. Capping

total RNA	10µg
Vaccina Capping enzyme (M2080)	5µl
10X Vaccina Capping buffer (M2080)	5µl
5mM Destiobiotin-GTP (N0761)	5µl
E.Coli IPP (M0361)	5µl
H2O	qsp
Total	50µl

37°C for 1h
Purify with a Zymo Clean and Concentrator 5G column (standard protocol total RNA).
Wash 4 times with wash buffer: 2 times 700µL, 1min + 2 times 400µL, 1min. Centrifuge one more time 2min to dry. All centrifugations @12000G
Elute with 40ul low-TE (in 2 times 20µL). Wait 2min before centrifugation for 1min.

2. Purify the reaction with 1.0X AMPure beads (50µL beads, elute in 40µL low TE).

3. Addition of a polyU tail to the 3'end of transcript

Capped RNA	39µl	
polyU polymerase (M0337)	5µl	
10X polyA polymerase buffer (B7002)	5µl	10U final
10mM UTP (N0453A)	2.5µl	0.5mM final
Total	50µl	

Incubate at 37°C for 20min.

Purify the reaction with 1.0X Ampure beads (=50µl) → Capped-tailed RNA
Incubate sample with beads for 5min with 300 rpm agitation
Wash 2 times with 200µL EthOH 80%
Spin down to remove EtOH remaining and dry for ~ 10min
Elute in 34µL low TE for 5min with 300 rpm agitation. **Split the sample in two:**

- Use 30µl for enrichment → Adjust volume to 40µL.
- Save 3µl as a control (Enrich:Control 10:1).

4. Enrichment

2X Binding Buffer	Elution buffer	Wash buffer
250mM NaCl	50mM NaCl	60 mM NaCl
10mM Tris-HCl pH 7.5	10mM Tris-HCl pH 7.5	10 mM Tris-HCl pH 7.5
1mM EDTA	0.1mM EDTA	1mM EDTA
	1mM biotin	

Prepare 40uL of streptavidin beads:

- Wash 2 times with 400µl of Wash buffer
- Wash 2 times with 400µl of 2XBinding buffer
- Resuspend in 40µL 2X Binding buffer

Add 40ul Capped-tailed RNA to the prepared beads.
Incubate at room temperature for 30min on a hula mixer.
Wash the beads 3 times with 400µL **Wash Buffer.**
Elute RNA from beads by incubating the beads with 25ul **Elution Buffer** at 37°C for 30min.
Collect the biotin-eluted RNA by placing the tube on the magnetic rack -> get 25µL of enriched RNA

5. Reverse transcription, first strand cDNA synthesis

10mM dNTP	8µl
100uM RT_dABN_UID primer	4µl
RNA template	25µl (Enrich) or 3µl (Control)
Add H2O to total	51µl (14µL Enrich) (36µL Control)

Incubate at 65°C for 2min, cool down at room temperature.

Add:	
5X ProtoscriptII buffer	16µl
0.1M DTT	8µl
ProtoscriptII (M0368)	4µl
Murine RNase Inhibitor (M0314)	1µl
Total	80µl

Incubate at 42°C for 1h.

For Enrich group, add 2ul RNaseIf (M0243) and incubate at 37°C for 30min to 1h;
then purify the reaction with 1.0X Ampure beads (=80µL beads), elute with 23µl Low-TE.

For Control group, DO NOT do the RNaseIf. Directly purify the reaction with 1.0X Ampure beads (=80µL beads), elute with 23µl Low-TE.

6. Addition of a polyG linker to the 3'end of synthesized cDNA

cDNA	22µl
Terminal transferase (M0315)	1µl
10X TdT buffer	3µl
100mM dGTP	1µl
2.5mM CoCl2	3µl
Total	30µL

Incubate at 37°C for 30min.

7. 2nd enrichment of the sample & AMPure purification of the control

Enrich sample:
Add 10µL water to the enriched sample to qsp 40µL.
Add 40µl prepared streptavidin beads to the enriched sample

Incubate at room temperature for 30min on a hula mixer
wash the beads 3 times with 400µL **1X Binding buffer**
resuspend beads in 40µl low-TE, do not elute with biotin -> get enrich cDNA (+ strepta beads)

Do not use more than 15µl enrich cDNA (containing the streptavidin beads) in one 50µl PCR reaction, since the beads might inhibit the PCR.

Control sample:
Purify the TdT reaction with 1.0X AMPure beads. Elute in 40µl low-TE -> get control cDNA

8. Second cDNA strand synthesis

cDNA	10µL	→ ¼ of the cDNA-RNA reaction (30uL left)
10µM Pac_oligodc20_for_set2	4µL	
RnaseH (M0297)	2µL	
LongAmp ReadyMix (M0533)	20µL	
H2O	4µL	
Total	40µL	

Cycling:

37ºC	15min	
Flick the tube (beads)		
94ºC	1min	
94ºC	1min	⎫
65ºC	15min	⎭ 4 cycles
65ºC	10min	

9. Amplification of the cDNA

→ Use half of the cDNA (20uL left in PCR tube at -20ºC)
Set up the reactions as follow:

	For 1 reaction	
ds cDNA		4µL
10µM Pac_fw_set2		2.5µL
10µM Pac_rev		2.5µL
LongAmp ReadyMix (M0533)	25uL	150µL
H2O		16µL
Total		50µL

Cycling:

94ºC	1min	
94ºC	30sec	⎫
65ºC	8min	⎭ 4-10 cycles
65ºC	10min	

10. AMPure purification 0.6X

11. Second round amplification of the cDNA

According to the previous Qubit results, determine the number of cycles required for this additional round of PCR.
→ use all the 40µL cDNA reaction from first round

Set up 10 reactions as follow:

ds cDNA	4µL
10µM Pac_fw_set2	2.5µL
10µM Pac_rev	2.5µL
LongAmp ReadyMix (M0533)	25µL
H2O	16µL
Total	50µL

Cycling:

94ºC	1min	
94ºC	30sec	1-10 cycles
65ºC	8min	
65ºC	10min	

12. AMPure purification 0.9X

ONT-Cappable-seq protocol – single strand ligation

1. Capping

total RNA	10µg
Vaccinia Capping enzyme (M2080)	5µl
10X Vaccinia Capping buffer (M2080)	5µl
5mM Destiobiotin-GTP (N0761)	5µl
E.Coli IPP (M0361)	5µl
H2O	qsp
Total	50µl (23µL)

37°C for 1h
Purify with a Zymo Clean and Concentrator 5G column (standard protocol total RNA).
Wash 4 times with wash buffer: 2 times 700µL, 1min + 2 times 400µL, 1min. Centrifuge one more time 2min to dry. All centrifugations @12000G
Elute with 32µl low-TE (in 2 times 16µL). Wait 2min before centrifugation for 1min.

- Use 30µl for enrichment
- Save 1µl as Control RNA (Enrich:Control 30:1).
  NB: save the control in a PCR tube with the volume of H2O required for the later RT reaction (10µL).
  This is to avoid evaporation.

2. Enrichment

2X Binding Buffer	Elution buffer	Wash buffer
250mM NaCl	50mM NaCl	60 mM NaCl
10mM Tris-HCl pH 7.5	10mM Tris-HCl pH 7.5	10 mM Tris-HCl pH 7.5
1mM EDTA	0.1mM EDTA	1mM EDTA
	1mM biotin	

Prepare 35uL of streptavidin beads:

- Wash 2 times with 400µl of Wash buffer
- Wash 2 times with 400µl of 2XBinding buffer
- Resuspend in 35µL 2X Binding buffer

Add 30µl Capped-tailed RNA to the prepared beads.
Incubate at room temperature for 30min on a hula mixer.
Wash the beads 3 times with 400µL **Wash Buffer.**
Elute RNA from beads by incubating the beads with 25µl **Elution Buffer** at 37°C for 30min to 1h.
Collect the biotin-eluted RNA by placing the tube on the magnetic rack -> **get 15µL of enriched RNA**

3. Speed vac to evaporate and reduce the volume (for further ligation)

Dry to 10µL final

3. Ligation to 3'end with 5'App thermostable ligase (M0319)

	Ci	Cf	**ENRICHED** Vol (µL)	**CONTROL** Vol (µL)
Capped RNA	10ng/µL	30ng (0.08pmol)	10	1
5'App	~5pmol/µL	1.5uM (1.5pmol/µL)	6	6
NEB1 homemade buffer	10X	1X	2	2
Thermo ligase M0319	20pmol/µL	40pmol	2	2
MgCl2	25mM	1mM	0.8	0.8
H2O	/	/	/	7
Total	/	/	22.8µL	22µL

Incubate at 65℃ for 30min.

4. AMPure beads clean-up and size selection 1.0X

To remove the spike-in DNA.
Add 22µL beads. Elute in 12µL (10µL final).

5. Reverse transcription, first strand cDNA synthesis

	Volume (µL)
Template RNA from ligation	10
RT oligo (100uM)	1 (5µM final)
10mM dNTP	1
5X Protoscript II buffer	4
0.1M DTT	2
Murine RNase inhibitor	0.2
Protoscript II RT	1
H2O	Total 20µL (0.8µL H2O)

Incubate at 42℃ for 1h (lid at 50℃).

For Enrich group, add 2µl RNaseIf (M0243) and incubate at 37°C for 30min to 1h;
then purify the reaction with 1.0X Ampure beads (=60µL beads), elute with 23µl Low-TE.

For Control group, DO NOT do the RNaseIf. Directly purify the reaction with 1.0X Ampure beads (=60µL beads), elute with 23µl Low-TE.

6. Addition of a polyG linker to the 3'end of synthesized cDNA

cDNA	22µl
Terminal transferase (M0315)	1µl
10X TdT buffer	3µl
100mM dGTP	1µl
2.5mM CoCl2	3µl
Total	30µl

Incubate at 37°C for 30min.

7. 2nd enrichment of the sample & AMPure purification of the control

Enrich sample:

Add 10µL water to the enriched sample to qsp 40µL.
Add 30µl prepared streptavidin beads to the enriched sample
Incubate at room temperature for 30min on a hula mixer
wash the beads 3 times with 400µL **1X Binding buffer**
elute with 40µL biotin at 37°C for 30min.

Control sample:

Purify the TdT reaction with 1.0X AMPure beads. Elute in 40µl low-TE -> get control cDNA

8. Second cDNA strand synthesis

cDNA	20µL	→ half of reaction (keep the other 20µL product)
10µM Pac_oligodc20_for_set2	2µL	
RnaseH (M0297)	2µL	
LongAmp ReadyMix (M0533)	20µL	
Total	44µL	

Cycling:

37ºC	15min	
Flick the tube (beads)		
94ºC	1min	
94ºC	1min	4 cycles
65ºC	15min	
65ºC	10min	

9. Amplification of the cDNA

→ Use half of the cDNA (20µL left in PCR tube at -20ºC)

Set up reaction as follow:

	For 1 reaction
ds cDNA	10µL
10µM Pac_fw_set2	2.5µL
10µM Pac_rev	2.5µL
LongAmp ReadyMix (M0533)	25µL
H2O	10µL
Total	50µL

Cycling:

94ºC	1min	
94ºC	30sec	4 - 10 cycles
65ºC	8min	
65ºC	10min	

10. AMPure purification 0.6X

11. Second round amplification of the cDNA

According to the previous Qubit results, determine the number of cycles required for this additional round of PCR.

→ use all the 40uL cDNA reaction from first round

Set up reaction as follow:

ds cDNA	20µL
10µM Pac_fw_set2	2.5µL
10µM Pac_rev	2.5µL
LongAmp ReadyMix (M0533)	25µL
H2O	16µL
Total	50µL

Cycling:

94ºC	1min	
94ºC	30sec	1-10 cycles
65ºC	8min	
65ºC	10min	

12. AMPure purification 0.9X

C. Appendix from Chapter III

RNA extraction protocol for the DefCom community

Hybrid protocol combining: Trizol + RNeasy mini kit (Qiagen) + bead beating using Fastprep 120

- Add 1mL Trizol Reagent to 1mL of cell pellet
- Incubate 5min at room temperature
- Transfer the 1mL to a lysing matrix B tube (blue cap, from MP Biomedicals)
- Lyse using a bead beating machine: Fastprep 120 (MP Biomedicals): 40sec at 6.0m/s speed
- Immediately transfer the samples on ice after the bead beating
- Add 200µL of Chloroform
- Incubate 3min at room temperature
- Centrifuge at 12000G for 15min at 4°C
- Carefully remove the upper aqueous phase (~500µL)
- Add an equivalent volume of pure 100% Ethanol (~500µL)
- Follow the procedure of the RNEasy mini kit (Qiagen), start by loading the sample onto a column
- Perform the on-column DNAseI treatment following Qiagen's recommendations.

Elute in 40µL of nuclease-free water

Résumé de la thèse en français

Nouvelles approches et concepts pour l'étude des communautés microbiennes complexes

Introduction

L'acide désoxyribonucléique (ADN) est constitué d'une succession de nucléotides (A, C, T ou G) dont la séquence contient les informations sur les propriétés héréditaires et biochimiques de tous les organismes vivants sur terre. Il est donc crucial pour les chercheurs en biologie de pouvoir analyser de telles séquences. Au fil des années, les chercheurs ont ainsi tenté de développer des méthodes et techniques permettant le séquençage de l'ADN. De ces recherches ont découlé trois générations de méthodologies et de séquenceurs.

Le développement dans les années 70 de deux différentes méthodes capables de décoder des centaines de bases a révolutionné le domaine de la biologie. Ces méthodes sont considérées comme les premières méthodes permettant de déterminer les séquences nucléotidiques de l'ADN. La première méthode repose sur une procédure de clivage chimique (Maxam et Gilbert, 1977) et a été développé en 1977 par Maxam et Gilbert. En parallèle, Frederick Sanger et son équipe ont développé une seconde méthode, appelée la méthode Sanger, qui est rapidement devenue la référence en matière de séquençage. Cependant, cette méthode manquait d'automatisation et prenait beaucoup de temps, ce qui a conduit au développement de la première génération de séquenceurs ADN capillaires automatisés.

Depuis 2005, les technologies de séquençage de nouvelle génération (NGS), également connues sous le nom de séquençage de deuxième génération, sont entrées sur le marché et ont rapidement remplacé le séquençage de Sanger, ces nouvelles technologies permettant un débit beaucoup plus élevé pour le séquençage de l'ADN et de l'ADN complémentaire. Au lieu d'analyser un tube par réaction, une banque complexe de matrices d'ADN est immobilisée sur une surface bidimensionnelle et amplifié *in vitro*, générant des copies de chaque matrice à séquencer. Au lieu de mesurer la longueur des fragments, le séquençage comprend des cycles biochimiques (tels que l'incorporation par une polymérase de nucléotides marqués par fluorescence) et d'imagerie (méthode également connu sous le nom de séquençage par synthèse). Ces techniques génèrent des millions à des milliards de molécules d'ADN qui peuvent être séquencées en parallèle, permettant une analyse massive à partir d'un ou plusieurs échantillons, à un coût très réduit (Shendure et al., 2017). Parmi les différentes technologies de séquençage de seconde génération, on peut citer : le pyroséquencage (platerforme 454 de Roche), le séquençage par ligation (SOLiD from Applied Biosystems), le séquençage ion torrent (Ion Torrent de Life Technologies), le séquencage par synthèse (Solexa/illumina) et les longues lectures synthétiques. Un inconvénient majeur du séquençage de deuxième génération est la limite au niveau de la longueur de lecture qui reste relativement courte (300bp max). Par conséquence, l'analyse de génomes complexes et des régions répétitives est difficile. Idéalement, le séquençage se ferait sur les molécules d'ADN (ou ARN) natives, serait précis et sans limitation de longueur de lecture.

Les technologies de séquençage de troisième génération tentent de palier à ces problèmes en s'efforçant de fournir de longues lectures, en temps réel. Ces technologies permettent de séquencer l'ADN sans amplification préalable, avec la résolution d'une molécule à la fois. Par conséquent, les biais, les erreurs et les pertes d'informations (telles que la perte de la méthylation et de modifications de l'ADN) liés à l'étape d'amplification sont évités (Kulski, 2016 ; Shendure et al., 2017). De plus, les tailles de lecture plus longues, la couverture uniforme, le séquençage en temps réel et la résolution d'une seule molécule à la fois sont possibles. Le séquençage d'ADN complémentaire avec de tels techniques apporte un vrai avantage pour l'analyse de transcriptome, car il permet de séquencer des transcrits d'ARNm entiers, permettant d'identifier des isoformes de gènes (Byrne et al., 2019; Zhao et al., 2019). Ces approches à longues lectures diffèrent des approches à lecture courte car elles ne reposent pas sur une amplification clonale de fragments d'ADN pour générer un signal détectable (Goodwin, McPherson et McCombie, 2016). Actuellement, les leaders dans le domaine du séquençage de troisième génération sont les technologies de Pacific Biosciences (PacBio) et d'Oxford Nanopore Technologies (ONT).
Pourtant, des défis importants subsistent pour les technologies à lecture longue. Bien que ces plateformes génèrent des lectures plus longues que les séquenceurs de deuxième génération (Illumina), les séquenceurs PacBio et Oxford Nanopore ont des taux d'erreur de séquençage plus élevés. Mais les progrès sont rapides et en quelques années seulement, la précision des lectures produites par ces deux technologies a considérablement augmenté (Amarasinghe et al., 2020). Le taux d'erreur a été réduit à < 1% pour les séquenceurs PacBio (Wenger et al., 2019) et < 5% pour les séquenceurs Nanopore (M. Jain et al., 2018).

Depuis 1977, les technologies de séquençage de l'ADN ont évolué à un rythme impressionnant et continuent de progresser rapidement. Bien qu'Illumina domine toujours le marché du séquençage, d'autres technologies ont émergé et ont élargi les champs d'applications, par exemple PacBio est utilisé pour l'assemblage *de novo* de génomes complexes et Nanopore a permi le développement d'approches révolutionnaires telles que le séquençage portable et le séquençage direct de l'ARN. Le séquençage nouvelle génération a le potentiel d'accélérer la recherche biologique et biomédicale, en permettant l'analyse complète des génomes et des transcriptomes à des coûts continuellement réduits, permettant une utilisation systématique et généralisée des technologies de séquençage. Ces technologies apportent avec elles un énorme potentiel de recherche et d'applications, pour la recherche clinique avec la possibilité d'identifier les agents pathogènes en temps réel mais aussi environnementale et microbienne, avec la possibilité de séquencer en temps réel sur le terrain, ce qui apporte chaque jour de nouvelles connaissances sur la diversité microbienne qui nous entoure.

Ainsi, l'évolution rapide des techniques de séquençage et l'avènement de la métagénomique ont conduit à l'exploration de communautés bactériennes dans différents environnements, des océans centraux à l'intestin humain. Au cours des dernières décennies, le domaine du microbiote s'est étendu de façon exponentielle, apportant avec lui des découvertes révolutionnaires. Le terme « microbiote humain » fait référence aux génomes collectifs des microbes (bactéries, bactériophages, champignons, protozoaires et virus) qui vivent à l'intérieur et sur divers sites du corps humain (Consortium et The Human Microbiome Project Consortium, 2012). Des exemples d'habitats occupés comprennent notre cavité buccale, nos organes génitaux, nos voies respiratoires, notre peau, notre système gastro-intestinal et nos poumons (O'Dwyer, Dickson et Moore, 2016; Kho et Lal, 2018).

L'organe contenant la plupart des cellules bactériennes est le tractus gastro-intestinal, avec environ $3{,}8x10^{13}$ de cellules microbiennes. Le microbiote intestinal est principalement composé de bactéries de trois phylums : *Firmicutes, Bacteroidetes* et *Actinobacteria* (Tap et al., 2009). Ce microbiote diverse et complexe est considéré comme un organe additionnel et on estime qu'il abrite 150 fois plus de gènes que l'hôte humain (Qin et al., 2010). Ces gènes supplémentaires apportent des fonctions importantes non codées par l'hôte et jouent un rôle essentiel dans le métabolisme et la physiologie de l'hôte (Hooper et Gordon, 2001). Ainsi, le microbiote fonctionne en tandem avec l'hôte, jouant un rôle central dans des processus critiques tels que le vieillissement, la digestion, l'immunité, la protection contre la colonisation par des agents pathogènes et les fonctions métaboliques essentielles. Alors que le rôle du microbiote humain est désormais considéré comme un « organe » essentiel, sa composition est loin d'être universelle et varie fortement au sein et entre les individus en fonction d'une multitude de facteurs.

Des projets d'étude du microbiote à grande échelle, tels que le consortium Human Microbiome Project (HMP) et le consortium MetaHIT ont grandement aidé à mettre en place un cadre pour la recherche sur le microbiote humain. De telles études ont montré que tout au long de notre vie, de nombreux facteurs façonnent notre microbiote intestinal, modifiant sa diversité et sa composition. Un déséquilibre au niveau de la composition bactérienne est appelé dysbiose. Cet état de dysbiose se caractérise par un déséquilibre qui peut se manifester par une perte de bactéries bénéfiques (commensales), une diminution de la diversité et de la richesse microbienne, ainsi qu'une augmentation des souches pathogènes (Mahnic et al., 2020). La dysbiose est susceptible d'altérer le fonctionnement normal du microbiote intestinal dans le maintien du bien-être de l'hôte et a été associée à un large éventail de maladies et de troubles inflammatoires, notamment l'obésité, les maladies inflammatoires de l'intestin (MICI), les allergies, le diabète, les maladies cardiovasculaires et le cancer colorectal, dans des modèles humains et animaux (DeGruttola et al., 2016 ; Kho et Lal, 2018). Parmi les facteurs pouvant perturber l'équilibre du microbiote intestinal, on retrouve les antibiotiques. Ces médicaments permettent de lutter contre les agents pathogènes mais peuvent également inhiber la croissance de bactéries bénéfiques pour la santé, altérant ainsi la capacité fonctionnelle du microbiote intestinal humain, induisant des effets rapides et qui peuvent persister dans le temps. Les antibiotiques à large spectre réduisent la diversité bactérienne, sélectionnent les bactéries résistantes, augmentent les opportunités de transfert horizontal de gènes (HGT) et ouvrent des niches pour l'intrusion d'organismes pathogènes en éliminant les commensaux (Modi, Collins et Relman, 2014).

Un équilibre adéquat du microbiote intestinal est donc essentiel au maintien de l'état de santé de l'hôte, cet équilibre fragile pouvant être altéré par divers facteurs externes. Par conséquent, il est crucial d'être capable d'identifier la composition et les changements de composition suite à des perturbations du microbiote. Les changements de composition et les abondances relatives dans les microbiotes ont été bien caractérisés grâce à deux techniques majeures : le séquençage de l'ARN ribosomal16S et le séquençage métagénomique shotgun. Cependant, ces approches métagénomiques sont basées sur l'étude de l'ADN et répondent uniquement à la question « quelles bactéries et quels gènes sont présents dans l'échantillon ? ». Des approches fonctionnelles basées sur l'ARN sont nécessaires pour fournir une caractérisation au niveau fonctionnel des microbiotes afin de compléter notre compréhension de la dynamique des communautés microbiennes en

répondant à la question « comment les bactéries réagissent-elles et que font-elles ? ». La métatranscriptomique permet justement d'étudier les activités fonctionnelles des microbiotes en fournissant des informations sur les gènes exprimés dans des communautés complexes. De telles données permettent d'étudier les interactions microbiote-hôte et de dériver des voies métaboliques, permettant d'explorer l'effet de différents environnements sur les activités bactériennes et de mieux comprendre ce qui peut conduire un microbiote sain vers une dysbiose ou un état pathologique (Bashiardes, Zilberman- Schapira et Elinav, 2016). A titre d'exemple, cela peut prendre jusqu'à plusieurs jours pour observer l'effet d'une perturbation sur la composition du microbiote. Dans le cas d'un traitement antibiotique, des modifications importantes de la composition bactérienne peuvent être observées 3 jours suivant le traitement (Abeles et al., 2016). Inversement, il a été démontré que les réponses transcriptionnelles sont parmi les premiers changements observés dans les premières minutes suivant l'exposition aux antibiotiques, reflétant la capacité des bactéries à s'acclimater très rapidement aux perturbations environnementales. Par exemple, des changements globaux dans l'expression des gènes (reprogrammation transcriptionnelle) ont été observés dès 5 min après l'injection d'antibiotiques dans *E. coli* (Sangurdekar, Srienc et Khodursky, 2006). Plusieurs études ont ainsi exploité les signatures du transcriptome et identifié des marqueurs ARN qui permettent de prédire la sensibilité aux antibiotiques de souches pathogènes telles que *Neisseria* gonorrhoeae (Khazaei et al., 2018). De telles signatures d'ARN représentent une approche prometteuse pour fournir rapidement un profil phénotypique de la sensibilité aux antibiotiques des agents pathogènes. Cela pourrait permettre une adaptation meilleure et rapide du traitement antibiotique, en fonction du phénotype de la bactérie (Bhattacharyya et al., 2017).

Cependant, la plupart de ces études ont été réalisées sur des bactéries en monoculture et, à notre connaissance, aucune étude sur la réponse transcriptionnelle rapide (après quelques minutes de traitement) n'a été réalisée dans des communautés microbiennes complexes. Il est important d'étudier la réponse transcriptionnelle rapide aux antibiotiques dans une communauté complexe afin d'identifier les réponses d'ARNm qui correspondent le mieux aux changements à long terme de la structure de la communauté. L'un des défis d'une telle étude réside dans la capacité à développer des communautés bactériennes définies et complexes. Pour relever en partie ce défi, des communautés synthétiques bien caractérisées ont été développées, permettant un meilleur contrôle et une meilleure reproductibilité. Utiliser la métagénomique et la métatranscriptomique de manière complémentaires, appliquées à des communautés synthétiques définies offrent la possibilité de lier les changements de composition à la réponse transcriptionnelle et d'identifier un potentiel marqueur ARN qui pourrait prédire de manière précoce l'impact sur la composition du microbiote suite à des perturbations externes, telles qu'un traitement antibiotique.

Objectifs de la thèse

Comme présenté dans l'introduction, l'une des principales limites des études actuelles sur le microbiote a été l'intégration rare de données fonctionnelles telles que la transcriptomique pour compléter l'interprétation des données métagénomiques (compositionnelles). L'objectif de recherche de ma thèse est de développer de nouvelles technologies basées sur le séquençage et de les appliquer pour fournir des informations supplémentaires sur les changements dans la composition et les activités des microbiotes. Plus précisément, le premier chapitre présente RIMS-seq (Rapid Identification of Methylase Specificity), une méthode pour obtenir simultanément la séquence d'ADN et le profil de 5-méthylcytosine (m5C) des génomes bactériens. Le chapitre deux présente ONT-cappable-seq et Loop-Cappable-seq, deux nouvelles techniques pour révéler l'architecture des opérons *via* le séquençage de transcrits complets utilisant respectivement le séquençage Nanopore et LoopSeq. Enfin, dans le chapitre trois, nous avons appliqué une approche multi-omique en utilisant certains des outils développés dans les chapitres précédents pour étudier la dynamique de la réponse d'un modèle de microbiote intestinal humain après traitement par la ciprofloxacine, un antibiotique à large spectre largement utilisé. Nous avons examiné les réponses transcriptionnelles et génomiques à court et à long terme de la communauté synthétique et avons exploré comment la réponse transcriptomique immédiate est corrélée et peut potentiellement prédire les changements ultérieurs de composition du microbiote. Nous nous sommes posé plusieurs questions : (1) peut-on identifier une reprogrammation transcriptionnelle immédiate dans une communauté complexe ? (2) les bactéries de la même famille réagissent-elles de la même manière ? Existe-t-il une réponse spécifique au phylum ? (3) y a-t-il une réponse spécifique des bactéries qui résisteront au traitement par rapport aux bactéries sensibles ? (4) et finalement, pouvons-nous identifier des marqueurs transcriptomiques (gènes ou voies métaboliques spécifiques exprimés de manière différentielle) qui pourraient être utilisés pour prédire l'issue du traitement ?

Résultats

I. RIMS-seq

Le manuscrit est publié dans *Nucleic Acid Research* :
Rapid Identification of Methylase Specificity (RIMS-seq) jointly identifies methylated motifs and generates shotgun sequencing of bacterial genomes.
Baum C, Lin YC, Fomenkov A, Anton B, Chen L, Yan B, Evans TC, Roberts RJ, Tolonen AC, Ettwiller L.
Nucleic Acids Research, 2021 https://doi.org/10.1093/nar/gkab705

La méthylation de l'ADN est connue pour moduler l'expression des gènes chez les eucaryotes, mais elle est également répandue chez les procaryotes, auxquels elle confère une résistance virale. Plus précisément, la méthylation de la 5-méthylcytosine (m5C) a été décrite dans les génomes de diverses espèces bactériennes dans le cadre de systèmes de restriction-modification (« RM systems »), chacun composé d'une méthyltransférase et d'une enzyme de restriction apparentée. Environ 90 % des génomes bactériens contiennent au moins l'une des trois formes courantes de méthylation de l'ADN : la 5-méthylcytosine (m5C), la N4-méthylcytosine (m4C) et la N6-méthyladénine (m6A). Contrairement aux eucaryotes où la position de la méthylation m5C est variable et sujette à des états épigénétiques, les méthylations bactériennes ont tendance à être présentes à des sites spécifiques à travers le génome. Ces sites sont définis par la spécificité de la méthylase et, dans le cas des systèmes

RM, ont tendance à être entièrement méthylés pour protéger l'hôte des digestions par l'enzyme de restriction associée. Les méthylases sont donc spécifiques d'un site et leurs séquences cibles varient selon les organismes.
Les méthodes à haut débit, telles que le séquençage au bisulfite (Bisulfite-seq), peuvent identifier le m5C à une résolution à la base près, mais nécessitent des préparations de banque spécialisées et l'assemblage du génome n'est pas possible à partir de ces données. Le séquencage PacBio a joué un rôle déterminant dans l'identification de la spécificité des méthylases, en grande partie parce qu'en plus de fournir un séquençage à longue lecture des génomes bactériens, les modifications m6A et m4C peuvent facilement être détectés. Ainsi, une seule analyse sur PacBio permet à la fois le séquençage et l'assemblage de génomes bactériens ainsi que la caractérisation des modifications et m4C et m6A. Mais le signal associé aux bases m5C est plus faible que pour les m6A ou m4C, le séquençage PacBio ne peut généralement pas identifier les méthylations m5C.

Jusqu'à présent, aucune technique ne permet donc le séquençage simultané de génomes et la caractérisation de la spécificité des méthylations m5C chez les bactéries. Nous avons développé une nouvelle méthode appelée RIMS-seq pour séquencer simultanément les génomes bactériens et caractériser la spécificité des méthylases m5C en utilisant un protocole simple, rapide et qui ressemble étroitement au protocole standard d'Illumina.
Nous avons appliqué RIMS-seq à plusieurs bactéries et identifié une variété de motifs de méthylation, allant de 4 à 8 nt de long, palindromiques et non palindromiques. Certains de ces motifs ont été identifiés pour la première fois, démontrant le potentiel de la technologie pour découvrir de nouvelles spécificités de méthylase, à partir de génomes connus comme inconnus. Appliqué à des souches caractérisées ou à de nouveaux isolats, RIMS-seq permets d'identifier *de novo* de nouvelles activités sans avoir besoin d'un génome de référence et permet également l'assemblage du génome bactérien à une qualité comparable à un séquençage standard.

Nous avons également validé que RIMS-seq peut identifier plusieurs spécificités de méthylases à partir d'une communauté microbienne synthétique et estimer l'abondance des espèces. Cependant, les espèces dans les microbiotes sont inégalement représentées, ce qui peut amener RIMS-seq à identifier des motifs uniquement dans les espèces les plus abondantes. Parce que RIMS-seq est basé sur une déamination limitée, il est nécessaire que le signal soit suffisamment grand pour identifier efficacement la spécificité de la méthylase. Pour la grande majorité des méthylases dans les systèmes RM, la méthylation est présente sur un nombre suffisant de sites à travers le génome pour que RIMS-seq détermine leurs spécificités. Néanmoins, les méthylases bactériennes peuvent être impliquées dans d'autres processus tels que la réparation des mésappariements de l'ADN, la régulation des gènes et la sporulation et il est possible que les sites de reconnaissance ne soeint pas entièrement méthylés. De tels sites partiellement méthylés peuvent être trouvés à l'aide de RIMS-seq, mais une analyse plus approfondie doit être effectuée pour évaluer à quel point la méthylation doit être omniprésente pour fournir un signal RIMS-seq. Dans d'autres cas, les motifs méthylés sont trop spécifiques ou sous sélection purificatrice, entraînant seulement une poignée de sites dans le génome. Dans ces cas, les signaux RIMS-seq ne peuvent être obtenus qu'avec une couverture de lecture suffisante pour compenser la rareté de ces sites. Alors que les spécificités de la méthylase sont d'un grand intérêt chez les bactéries en raison de leur diversité dans les séquences de reconnaissance, l'application de RIMS-seq à l'homme conduirait à l'identification du contexte CpG

déjà bien décrit. Dans ce cas, d'autres technologies telles que EM-seq ou bisulfite-seq sont plus appropriées car elles permettent d'obtenir la localisation génomique précise. En résumé, RIMS-seq est une nouvelle technologie permettant l'étude simultanée de la séquence génomique et de la méthylation chez les procaryotes. Étant donné que cette technique est facile à mettre en place et présente une qualité de séquençage similaire à celles de l'ADN-seq, RIMS-seq a le potentiel de remplacer le DNA-seq standard pour les études microbiennes.

II. Développement de ONT-Cappable-seq et Loop-Cappable-seq

Les opérons ont été décrits pour la première fois en 1960 comme un moyen pour les bactéries de co-exprimer des gènes fonctionnellement liés à partir d'un seul promoteur. Depuis, les chercheurs ont montré la complexité de la transcription bactérienne et l'implication de multiples mécanismes pour la contrôler. Les méthodes les plus largement utilisées pour étudier les transcriptomes reposent sur le séquençage à lecture courte (par exemple le RNA-seq). Mais ces techniques à lecture courte nécessitent de fragmenter les transcrits au préalable, provoquant la perte de nombreuses informations contenues dans les transcrits entiers et limitant les analyses. Il n'est par exemple pas possible de phaser les sites de démarrage de la transcription (TSS) avec les sites terminateurs (TTS). Cappable-seq est une méthode développée et publiée en 2016 par l'équipe de Laurence Ettwiller à New England Biolabs (Ettwiller et al., 2016). Cette méthode permet de capture spécifiquement les transcrits primaires, élimine les transcrits ribosomaux et dégradés et permet d'identifier les TSS à une résolution à la base près.

L'association de la technologie Cappable-seq avec le séquençage longue lecture comme PacBio a démontré qu'il est possible de délimiter les sites de début et de fin de la transcription ainsi que de déterminer la structure opéronique complète chez les bactéries. Cette méthode publiée par Yan et al (Yan et al., 2018) a mis en évidence que Cappable-seq est une technologie flexible qui pourrait être adaptée à diverses plateformes de séquençage. Deux autres plateformes rivalisent avec PacBio sur le marché des lectures longues, chacune avec ses propres avantages/inconvénients : Nanopore (ONT) et LoopSeq (Loop Genomics). La première plate-forme offre un haut débit, une facilité d'utilisation et un prix abordable, tandis que la seconde offre une précision sans précédent. Dans ce second chapitre nous présentons le développement de deux nouvelles versions de Cappable-seq longue lecture : ONT-Cappable-seq, adapté au MinION de Oxford Nanopore et Loop-Cappable-seq, adapté à la plateforme LoopSeq développée par Loop Genomics.

ONT-cappable-seq

La technologie de séquençage Nanopore d'Oxford Nanopore Technologies (ONT) offre un séquençage longue lecture à haut débit, abordable et facile à manipuler. Alors que la méthode Cappable-seq capture de manière robuste l'extrémité 5' des transcrits, la définition du TTS procaryote reste difficile car les transcrits n'ont pas de queue poly-A en 3' utilisée pour liguer spécifiquement les adaptateurs comme pour les transcrits eucaryotes. La méthode courante pour capturer les extrémités des transcrits procaryotes repose sur l'ajout d'une queue polyA à l'extrémité 3' des transcrits en utilisant l'enzyme polyA polymérase. Cette queue polyA servira d'ancrage pour une amorce oligod(T) utilisée lors de l'étape de transcription inverse pour synthétiser le premier brin d'ADNc. Mais il a été montré que la queue polyA peut ajouter des biais. Dans le cas des régions riches en adénine, l'amorce oligod(T) peut s'hybrider en interne sur des régions du transcrit riches en adénine et initier la transcription inverse à partir de cette région plutôt que de la queue polyA ajoutée (Balázs et al., 2019 ;

Sessegolo et al., 2019). Cela se traduit par des molécules d'ADNc tronquées et des erreurs dans l'identification du TTS, en particulier pour les bactéries contenant des génomes riches en AT. La capture précise de l'extrémité 3' des transcrits est essentielle car elle modifie non seulement la définition mais également la quantification des unités de transcription. Ainsi, en plus du développement de ONT-Cappable-seq, nous avons étudié différentes stratégies pour capturer de manière robuste l'extrémité 3' et avons développé une nouvelle stratégie basée sur la 'splint ligation'. Pour cela, nous avons utilisé *Escherichia coli* pour d'abord valider la méthode sur un organisme modèle et nous avons ensuite utilisé *Clostridium phytofermentans* (*C. phy*) comme modèle de génome riche en AT pour étudier l'effet de différentes stratégies pour capturer l'extrémité 3'. ONT-Cappable-seq est basé sur la stratégie de la méthode SMRT-Capable-seq développé précédemment et a été adapté à la plateforme de séquençage MinION d'Oxford Nanopore Technologies.

Dans l'ensemble, la queue polyA et la 'splint ligation' ont donné les meilleurs résultats en termes d'identification TTS, de corrélation avec les terminateurs prédits rho-indépendants et l'expression génique obtenue à partir de ces ensembles de données était bien corrélée avec les données de RNA-seq. La queue PolyU a donné des résultats satisfaisants sur ces différents paramètres mais le rendement de la bibliothèque était beaucoup plus faible par rapport aux autres méthodes. Nous émettons l'hypothèse que cela pourrait être dû à la polyU polymérase qui est moins processive que la polyA polymérase, conduisant à une plus faible proportion de transcrits contenant une queue polyU. Le 'polyU tailing' ayant été utilisé dans le cadre d'une comparaison avec d'autres méthodes, nous n'avons pas cherché à optimiser la réaction. A l'inverse, la ligation simple brin a montré les moins bons résultats sur toutes les caractéristiques utilisées dans la comparaison. La plupart des positions TTS ont été identifiées au milieu d'un gène et les données n'étaient pas corrélées avec les terminateurs rho-indépendants prédits ni avec les données RNA-seq. Ces résultats suggèrent que des transcrits tronqués en 3' sont capturés et séquencés, ce qui conduit à une fausse identification du TTS. Nous émettons l'hypothèse que cela pourrait être dû au magnésium dans le tampon, un composant essentiel pour la ligation mais qui est connu pour catalyser la dégradation de l'ARN. En effet, même si nous avons optimisé en amont les conditions de ligation qui pourraient avoir un impact sur l'intégrité de l'ARN (température et temps), l'ARN subit une dégradation qui provoque des transcrits tronqués.

En résumé, à la fois la queue polyA et la 'splint ligation' polyA sont des méthodes efficaces pour capturer l'extrémité 3' et ont fourni des résultats similaires. Les problèmes techniques rencontrés avec la polyA polymérase ont démontré que le choix de la méthode employée est crucial car elle peut avoir un impact profond non seulement sur l'identification TTS mais aussi sur l'identification et la quantification des transcrits. Ce travail montre l'importance d'avoir une stratégie robuste pour capturer l'extrémité 3' afin d'obtenir un transcriptome précis en utilisant le séquençage à lecture longue, en particulier lorsqu'il est appliqué à des communautés complexes où une diversité de génomes est présente. Néanmoins, nous avons développé une nouvelle stratégie, la 'splint ligation', qui empêche l'amorçage interne qui peut être observé lors de l'utilisation de la queue polyA (en particulier dans les génomes bactériens riches en AT) et empêche ainsi toute fausse identification de TTS. Cette technique, appliquée à un microbiote, permettrait une capture robuste et fiable du transcriptome.

Loop-Cappable-seq

De nettes améliorations ont été apportées grâce à l'émergence de technologies de séquençage à lecture longue, telles que PacBio et Oxford Nanopore. Malgré ces améliorations, les technologies de séquençage à lecture longue restent sujettes aux erreurs et le manque de précision peut limiter leur utilisation. Les erreurs prédominantes dans les technologies de séquençage PacBio et Nanopore sont les insertions et les deletions (indels). Ces erreurs peuvent considérablement dérouter les algorithmes de « mapping » et, en introduisant des décalages de lecture et des codons stop prématurés, affecter de manière critique la prédiction des cadres de lecture ouverts directement à partir des transcrits (Watson et Warr, 2019). L'approche la plus courante pour surmonter ces taux d'erreur élevé consiste à aligner les lectures sur un génome de référence. Néanmoins, lorsque aucun génome de référence de haute qualité n'est disponible (ce qui est le cas dans la plupart des recherches sur le microbiote), les technologies de lecture longue sont d'une utilité limitée (Sahlin et Medvedev, 2021). Dans l'ensemble, de telles erreurs limitent la portée des technologies de lecture longue pour une analyse communautaire complexe. Cela a motivé le développement de plusieurs approches informatiques pour corriger et réduire le nombre d'erreurs dans les données de lecture longue. Il existe deux stratégies principales : (1) l'approche de correction hybride qui utilise des données illumina à lecture courte pour corriger les lectures longues et (2) l'approche non hybride (autocorrection) dans laquelle les lectures longues sont autocorrigées à l'aide du chevauchement des données de séquencage à couverture élevée (Magi et al., 2018). Dans cette partie, nous présentons une autre version de Cappable seq. Pour ce projet, nous avons collaboré avec Loop Genomics pour adapter Cappable-seq à leur plateforme LoopSeq, une nouvelle technologie de séquençage à lecture longue basée sur le séquençage Illumina. Parce que cette technologie est basée sur la plate-forme Illumina, elle offre un prix abordable et une haute précision de séquençage. De plus, LoopSeq combiné à Cappable-seq offre de nouvelles possibilités, telles que la prédiction des ORFs à partir des lectures brutes dans des communautés microbiennes complexes contenant des espèces inconnues, ainsi que la possibilité de différencier des espèces similaires entre elles, réduisant ainsi le problème de mapping multiple qui est souvent rencontré dans les études sur le microbiote. De plus, les applications dans les microbiotes sont particulièrement attrayantes puisque Loop-Cappable-seq serait théoriquement capable de découvrir des voies métaboliques partielles ou complètes en phasant des gènes fonctionnellement liés sur les mêmes lectures de séquençage. Tout d'abord, nous avons développé Loop-Cappable-seq sur *E. coli* et évalué la capacité de différentes plates-formes de séquençage de lectures longues à prédire directement les ORF à partir des lectures brutes par rapport aux lectures mappées. Dans un second temps, nous avons créé une communauté mixte synthétique composée de différentes sous-espèces d'*E. coli* et d'un *Bacillus* pour démontrer la capacité de Loop-Cappable-seq à fournir une représentation précise du transcriptome de communautés mixtes, avec la possibilité de distinguer les espèces entre elles et à une résolution au niveau de la sous-espèce. Cependant, cette deuxième partie est toujours un projet en cours car la pandémie de Covid-19 a retardé le projet. Les expériences sont actuellement en cours mais par conséquence, aucun résultat n'est encore disponible pour être montré dans cette deuxième partie. Un manuscrit conjoint avec Loop Genomics est en préparation dans le but de publier Loop-Cappable-seq.

Ici, nous avons adapté Cappable-seq à LoopSeq, une nouvelle plate-forme de séquençage, et développé Loop-Cappable-seq, une méthode qui a le potentiel d'être utilisée pour annoter les transcrits sans avoir besoin d'un génome de référence. Nous avons évalué la capacité de différentes

versions Cappable-seq à lecture longue (plateformes PacBio, Nanopore et LoopSeq) à prédire les ORFs directement à partir de lectures longues brutes. Dans l'ensemble, les données PacBio (SMRT-Cappable-seq) et Nanopore (ONT-Cappable-seq) sont encore trop sujettes aux erreurs pour être utilisées pour ce type d'analyse. Les programmes de correction tels que Canu et Lordec aident à corriger les données et à réduire le nombre d'erreurs, mais ces programmes ont également des « effets secondaires » majeurs. Tout d'abord, nous devons garder à l'esprit que ces programmes ont été initialement conçus pour la correction des données de DNA-seq et pourraient ne pas être optimaux pour corriger les données transcriptomiques. À titre d'exemple, les outils d'autocorrection contiennent une étape pour générer des séquences de consensus à l'aide de lectures qui se chevauchent, ce qui implique la nécessité d'une couverture de séquençage élevée pour obtenir une correction efficace. Dans le cas de communautés complexes, où des espèces similaires et des transcrits similaires sont présents, ceux-ci sont susceptibles d'être combinés en un seul transcrit, supprimant des informations (Lima et al., 2020). De plus, les programmes de correction ont tendance à produire des lectures plus courtes et à réduire la profondeur de séquençage, car ils éliminent les lectures non corrigées ou rognent les régions non corrigées. De tels comportements provoquent une perte de données et peuvent influencer l'analyse en aval car des informations sont perdues si les lectures sont raccourcies, fusionnées à tort avec d'autres ou même supprimées (Zhang, Jain et Aluru, 2020). D'un autre côté, Loop-Cappable-seq a montré une très bonne qualité de données, avec le taux d'indels le plus bas et a donné les meilleurs résultats lors de la prédiction des ORFs directement à partir des lectures brutes. Contrairement aux autres plates-formes de lecture longue, aucune correction d'erreur n'a été nécessaire pour obtenir une prédiction qui ressemble à celle que nous aurions obtenue avec une lecture parfaite sans erreurs. En d'autres termes, Loop-Cappable-seq a le potentiel d'éliminer le besoin d'un génome de référence car des gènes et des opérons entiers peuvent être identifiés sur une seule lecture brute, ce qui permet de prédire la fonction du gène et d'annoter les gènes inconnus en fonction de leurs gènes voisins sur l'opéron. De plus, parce que cette méthode fournit des données de haute qualité (précises), elle réduirait l'ambiguïté du mapping lorsque des espèces similaires sont présentes (problème de mapping multiple). Les lectures seraient attribuées avec une plus grande confiance à leur génome correspondant. L'application de la méthode aux études du microbiote serait alors particulièrement intéressante. Pour conforter cette hypothèse, l'étape suivante consiste à tester si les lectures longues et précises obtenues à partir de Loop-Cappable-seq permettent de distinguer des transcrits très similaires provenant de plusieurs sous-espèces d'*E. coli*. Nous effectuons actuellement les expériences et espérons analyser les données bientôt. L'étape finale pour valider cette méthode serait de l'appliquer à un échantillon de microbiote réel et de prédire les domaines ORF et PFAM directement à partir des données. Ce type d'analyse est très prometteur car il donnerait un aperçu de la fonctionnalité d'un microbiote même si les génomes de référence ne sont pas disponibles.

III. Relier les réponses transcriptionnelles aux changements de composition dans un microbiote intestinal synthétique soumis à un traitement antibiotique

Depuis la dernière décennie, le microbiote humain a suscité un intérêt considérable et fait l'objet de recherches passionnantes. De nombreuses études ont montré l'importance du microbiote, en particulier du microbiote intestinal, pour la santé humaine (Fan et Pedersen, 2021). Le microbiote intestinal influence fortement la physiologie de l'hôte, en aidant par exemple, à la bioconversion des nutriments et à la détoxification, en soutenant l'immunité et en protégeant l'hôte contre les agents

pathogènes. Des perturbations au niveau de la composition du microbiote ont été reliées à l'initiation et à la progression de nombreuses maladies et troubles inflammatoires (Carding et al., 2015 ; Scotti et al., 2017). Cet équilibre fragile entre les bactéries du microbiote peut être altéré par de nombreux facteurs, dont les antibiotiques, qui modifient la composition de la population bactérienne, favorisent la propagation des souches résistantes et peuvent dégrader l'effet protecteur du microbiote contre l'invasion par des agents pathogènes. Cependant, il a été démontré que les bactéries peuvent s'acclimater rapidement aux perturbations environnementales par reprogrammation transcriptionnelle (Sangurdekar, Srienc et Khodursky, 2006). Ici, nous explorons comment la réponse transcriptomique rapide aux antibiotiques est corrélée et prédit potentiellement les changements ultérieurs de la structure du microbiome. Dans cette étude, nous examinons les réponses à court et à long terme d'une communauté phylogénétiquement diverse et définie de bactéries intestinales à l'antibiotique à large spectre largement utilisé, la ciprofloxacine. Après l'ajout de ciprofloxacine aux cultures en phase logarithmique, des échantillons ont été prélevés sur une durée allant de 5 minutes à 48 heures. Nous avons utilisé une approche multiomique, en utilisant certaines des méthodes développées et présentées précédemment, afin d'analyser les réponses transcriptionnelles et les changements de composition de la communauté par rapport aux contrôles sans ciprofloxacine. Nous avons effectué RNA-seq et Cappable-seq pour étudier la réponse fonctionnelle ainsi que 16S et RIMS-seq (séquençage shotgun) pour étudier les changements de composition à l'échelle de la communauté.

Dans cette étude, nous avons étudié une communauté synthétique définie de 51 bactéries représentatives de l'intestin et étudié l'impact de la ciprofloxacine sur la composition de la communauté et la réponse fonctionnelle des bactéries, sur 48h. Nous avons exploré plusieurs aspects de cette communauté suite à l'ajout d'antibiotiques, y compris les changements de composition, la caractérisation de la méthylation, l'identification des TSS et la régulation de l'expression des gènes. Nous avons identifié des changements significatifs dans la composition bactérienne après plusieurs heures/jours d'ajout d'antibiotiques, avec d'une part, des espèces capables de résister à la ciprofloxacine et prenant rapidement le dessus sur la communauté comme *Enterococcus faecium* et plus généralement les *Firmicutes*, alors que certains phylums de bactéries comme les *Proteobacteria* et *Bacteroidetes* ont été significativement diminués par l'antibiotique. Le but est de corréler les changements de composition de la communauté avec la reprogrammation transcriptionnelle des bactéries qui, contrairement aux changements de composition, se produit très rapidement.

Pour la première fois à notre connaissance, nous avons identifié une reprogrammation transcriptionnelle significative chez plusieurs bactéries seulement 5min après ajout de la ciprofloxacine. Plus précisément, nous avons identifié les *Proteobacteria* comme les premiers répondeurs, ces espèces déclenchant immédiatement la voie de réponse SOS, tandis que les *Firmicutes* ont tendance à mettre en place des mécanismes de défense tels que des pompes à efflux, suggérant qu'il existe une réponse spécifique au phylum. Dans l'ensemble, ces observations préliminaires sont prometteuses et sont bien corrélées avec la réponse attendue suite à un traitement avec la ciprofloxacine qui a été largement décrite dans la littérature.

Cependant, ces résultats et les observations qui ont émergé de ces analyses sont préliminaires et une analyse plus approfondie est nécessaire pour valider l'hypothèse que nous avons présentée dans les

sections ci-dessus. L'étape suivante consiste à déterminer dans quelles voies métaboliques les gènes différentiellement exprimés sont impliqués. Actuellement, l'analyse de la réponse transcriptomique de la communauté au niveau des voies métaboliques est en cours et pourrait permettre l'identification de voies spécifiques qui pourraient aider à prédire l'impact du traitement antibiotique sur la composition du microbiote. En plus d'examiner les réponses à court et à long termes de la communauté, nous avons identifié diverses séquences de promoteurs ainsi que différents mécanismes de transcription tels que la transcription « leaderless ». Ces résultats illustrent bien la diversité des systèmes que possèdent les bactéries pour réguler la transcription et s'adapter rapidement à leur environnement. Nous avons également appliqué notre nouvelle méthode RIMS-seq pour caractériser la méthylation m5C à l'échelle de la communauté entière. Nous avons identifié une variété de motifs m5C et démontré la capacité de RIMS-seq à être utilisé à la fois pour la détermination de la composition et la caractérisation des méthylations m5C dans des communautés complexes, démontrant le potentiel de RIMS-seq pour remplacer le DNA-seq standard pour le séquençage de génomes bactériens.
En plus de la poursuite des analyses bioinformatiques, des expériences complémentaires sont en cours, notamment la détermination de la Concentration Minimale Inhibitrice (CMI) de la ciprofloxacine pour chaque bactérie en monoculture. Cette expérience supplémentaire fournira des informations importantes pour mieux comprendre la réponse attendue et observée des bactéries suite à l'ajout de l'antibiotique. En effet, il a été montré que la tolérance bactérienne aux antibiotiques diffère en monoculture comparé à en communauté. Une tolérance plus élevée que prévu peut se produire si une ou plusieurs espèces d'une communauté secrètent un composé qui dégrade les antibiotiques, ce qui peut activer des mécanismes de tolérance tels que l'expression de pompes à efflux chez d'autres espèces. Ce phénomène de « protection croisée ou d'alimentation croisée » communautaire pourrait se traduire par des concentrations plus faibles d'antibiotiques et plus généralement, altérer l'efficacité des traitements antibiotiques (Yurtsev et al., 2013 ; Adamowicz et al., 2018). De plus, nous avions initialement prévu d'utiliser notre méthode ONT-Cappable-seq sur la communauté afin d'explorer l'effet de la ciprofloxacine sur la régulation de la structure opéronique. Cela a malheureusement été impossible en raison de la pandémie de Covid-19 qui a retardé les expériences. Une autre piste intéressante à explorer serait d'appliquer notre autre méthode Loop-Cappable-seq, permettant une meilleure résolution de mapping et permettant ainsi de distinguer les sous-espèces entre elles grâce à la précision de la technologie LoopSeq.

Plus généralement, l'analyse du microbiote doit être interprétée avec prudence car divers facteurs peuvent avoir un profond impact sur les conclusions. À titre d'exemple, le milieu de culture est connu pour avoir un effet sur l'efficacité des antibiotiques et pour induire une compétition pour les nutriments entre les espèces (Adamowicz et al., 2018 ; Maier et al., 2020). Dans notre cas, nous avons effectué une culture en « batch » de la communauté (par opposition à une culture continue), ce qui signifie que l'apport en nutriments est limité. L'épuisement du milieu va provoquer une acidification globale susceptible d'ajouter une pression de sélection supplémentaire sur la communauté bactérienne. Un autre facteur à prendre en compte est le niveau d'oxygène de la culture. Un gradient d'oxygène existe *in vivo* dans le tube digestif, variant d'un environnement presque anoxique dans la lumière intestinale avec <1% O2 (0,1-1 mm Hg) à 5-20% O2 dans les cryptes intestinales (80 mm Hg) (Kim et al., 2019). Ici, nous avons effectué la culture dans un environnement anaérobie, ce qui n'est probablement pas optimal pour certaines espèces et ajoute un facteur supplémentaire de stress, de

compétition et de sélection, favorisant les bactéries anaérobies. Enfin, il a été démontré que le taux de croissance et l'état métabolique des bactéries ont un impact sur l'efficacité des antibiotiques (Eng et al., 1991; Lopatkin et al., 2019). Fianlement, tout cela met en évidence à quel point il est complexe de reproduire des conditions de croissance *in vivo* optimales pour les études sur le microbiote et les communautés synthétiques. Le microbiote est un vaste écosystème avec des interactions constantes avec l'hôte et au sein des individus bactériens, dans lequel les bactéries sont en compétition mais aussi coopèrent entre elles. Le microbiote est un domaine de recherche passionnant et prometteur qui n'a pas encore fini de révéler ses secrets.

Conclusion

Dans cette dernière partie de la thèse, je voudrais conclure sur le travail accompli mais aussi sur la riche expérience personnelle que ces 3 (et demi) années ont été. Dans cette thèse, j'ai présenté plusieurs nouvelles méthodes développées pour la caractérisation de communautés bactériennes complexes. Nous avons développé RIMS-seq, une nouvelle méthode basée sur un protocole simple qui permet à la fois le séquençage des génomes et la caractérisation de la méthylation m5C des génomes bactériens, démontrant le potentiel de RIMS-seq pour remplacer le DNA-seq standard. Nous avons validé avec succès la technique en l'appliquant sur une communauté synthétique définie complexe et transféré avec succès la méthode de New England Biolabs au Genoscope. L'article présentant RIMS-seq a été récemment publié dans *Nucleic Acids Research*. Nous avons également développé ONT-Cappable-seq et Loop-Cappable-seq, deux techniques permettant le séquençage de transcrits bactériens pleine longueur, basées respectivement sur le séquençage Nanopore et LoopSeq et permettant de révéler la complexité de la régulation de la structure des opérons. Dans la dernière partie, nous avons cherché à explorer le lien entre la réponse à long et à court terme d'un microbiote synthétique complexe suite à une perturbation liée à un traitement antibiotique en utilisant une approche multiomique. Nous avons effectué diverses expériences et analyses complexes dans le but d'identifier les réponses transcriptomiques qui correspondent le mieux aux changements à long terme de la structure de la communauté. Les résultats préliminaires sont prometteurs et ont révélé des pistes intéressantes à poursuivre, avec notamment l’identification d’une reprogrammation transcriptionnelle très rapide observée dès 5min après l'ajout de ciprofloxacine. Ce projet ambitieux nécessitera des expériences et des analyses supplémentaires pour révéler pleinement les informations contenues dans cette énorme quantité de données prometteuses et passionnantes.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Agriculture, alimentation, biologie, environnement, santé (ABIES)

Titre : Nouvelles approches et concepts pour l'étude des communautés microbiennes complexes
Mots clés : Séquençage haut débit, Méthylation de l'ADN, Transcriptomique, Microbiome, Antibiotiques

Résumé : Le développement du séquençage à haut débit a révolutionné l'étude des communautés microbiennes complexes, appelées "microbiomes", dans divers environnements, des océans centraux à l'intestin humain. L'objectif de la recherche menée dans le cadre de cette thèse est de développer de nouvelles technologies basées sur le séquençage haut débit et de les appliquer à l'étude des changements de compositions et d'activités des microbiomes. Le Chapitre Un présente RIMS-seq (Rapid Identification of Methylase Specificity), une méthode permettant d'obtenir simultanément la séquence ADN ainsi que le profil de méthylation 5-methylcytosine (m5C) des génomes bactériens. Le Chapitre Deux introduit ONT-cappable-seq et Loop-Cappable-seq, deux nouvelles techniques révélant la structure opéronique des transcrits via le séquençage de transcrits pleine longueur, en utilisant le séquençage Nanopore et LoopSeq, respectivement. Dans le Chapitre Trois nous avons utilisé une approche multiomique en appliquant certains des outils que nous avons développés dans les précédents chapitres, afin d'étudier la dynamique de réponse d'un microbiome synthétique représentatif d'un intestin humain après traitement avec la ciprofloxacine, un antibiotique à large spectre très largement utilisé. Les antibiotiques sont indispensables pour traiter les infections par des bactéries pathogènes mais ils éliminent également des bactéries commensales qui ont un rôle important pour la santé, entraînant le développement de souches résistantes et réduisant l'effet protecteur du microbiote contre l'invasion par des pathogènes. Il est donc crucial de pouvoir caractériser l'impact des traitements antibiotiques sur le microbiote, à la fois sur le plan compositionnel mais aussi sur le plan fonctionnel. Nous avons examiné à la fois la réponse au niveau transcriptomique et génomique, à court et à long terme, de la communauté synthétique. Nous avons exploré comment la réponse transcriptomique immédiate peut corréler et potentiellement prédire les changements de composition du microbiote, généralement observés bien plus tard après le traitement antibiotique. Le but est d'essayer d'identifier un marqueur apparaissant après quelques minutes/heures de traitement et qui pourrait être utilisé pour potentiellement prédire l'impact de ce traitement antibiotique afin de se diriger vers une médecine plus personnalisée.

Title: New approaches and concepts to study complex microbial communities
Keywords: Next-Generation Sequencing, DNA methylation, Transcriptomics, Microbiome, Antibiotics

Abstract: The development of high-throughput DNA sequencing revolutionized the study of complex bacterial communities called "microbiomes'', in diverse environments, from the central oceans to the human intestine. The research aim of this thesis is to develop new sequencing-based technologies and apply them to provide further insights into changes to the composition and activities of microbiomes. Specifically, Chapter One presents RIMS-seq (Rapid Identification of Methylase Specificity), a method to simultaneously obtain the DNA sequence and 5-methylcytosine (m5C) profile of bacterial genomes. Modification by m5C has been described in the genomes of many bacterial species to modulate gene expression and protect from viral infection. Chapter Two introduces ONT-cappable-seq and Loop-Cappable-seq, two new techniques to reveal operon architecture through full-length transcript sequencing, using Nanopore and LoopSeq sequencing, respectively. In Chapter Three, we applied a multiomics approach using some of the tools developed in the previous chapters to study the dynamics of the response of a model human intestinal microbiome after treatment with ciprofloxacin, a widely used broad-spectrum antibiotic. Antibiotics are critical treatments to prevent pathogenic infections, but they also kill commensal species that promote health, enhance the spread of resistant strains, and may degrade the protective effect of microbiota against invasion by pathogens. Therefore, it is crucial to be able to characterize both the composition but also the functional response of a microbial community to antibiotic treatment. We examined both the short and long-term transcriptional and genomic responses of the synthetic community and explored how the immediate transcriptomic response correlates and potentially predicts the later changes of the microbiome composition. The goal is to try to identify a marker appearing a few minutes/hours after the treatment that could be used to potentially predict the outcome of an antibiotic treatment, opening up the path to a more personalized medicine.

Maison du doctorat de l'Université
2ème étage aile ouest, Ecole normale supérieure paris-Saclay
4 avenue des Sciences
91190 Gif sur Yvette