Norwegian University of Life Sciences
Faculty of Chemistry, Biotechnology and Food Science

Philosophiae Doctor (PhD)
Thesis 2020:33

Back to basics: simplifying microbial communities to decrypt complex interactions

Tilbake til det grunnleggende: forenkling av mikrobielle samfunn for å tolke komplekse interaksjoner

Francesco Delogu

Back to basics:
simplifying microbial communities
to decrypt complex interactions

Tilbake til det grunnleggende:
Forenkling av mikrobielle samfunn for å tolke komplekse interaksjoner

Philosophiae Doctor (PhD) Thesis

Francesco Delogu

Norwegian University of Life Sciences
Faculty of Chemistry, Biotechnology and Food Science

Ås 2020

Thesis number 2020:33
ISSN 1894-6402
ISBN 978-82-575-1696-3

to Mitz

TABLE OF CONTENTS

ACKNOWLEDGMENTS

This dissertation is the result of the work carried out during the period 2016-2020 in the Protein Engineering and Proteomics (PEP) and the Microbial Ecology and Meta-Omics (MEMO) groups, Department of Chemistry, Biotechnology and Food Science (KBM) at The Norwegian University of Life Sciences (NMBU). This project was supported by the Research Council of Norway (RCN) with the FRIPRO program 250479 and the UM-NMBU with the Norwegian Centennial Chair Program (NOCC) 720500.

First of all I want to thank my supervisors for their patience, guidance and more in detail:

Phil Pope for his general advice, his endless endurance in correcting my manuscripts, giving me the chance to explore unorthodox paths in biology and sharing his knowledge about microbial ecology and metagenomics. **Torgeir Hvidsten** for his insight in transcriptomics, his help in coding, sharing his experience in eukaryotic data analysis and welcoming me in the Bioinformatics and Applied Statistics (BIAS) group for more than two years during the PhD. **Magnus Arntzen** for his guidance in metaproteomics, involving me in a new project and sanding a bioinformatician (me) back to the wet lab.

I also want to thank my hosts at the Eco-Systems Biology (ESB) group of the Luxembourg Center for Systems Biomedicine (LCSB):

Paul Wilmes for his scientific input on a faceted view of biological systems. **Patrick May** for teaching me several aspects of bioinformatics from his boundless knowledge of the field.

The five of you, alongside the microbes we studied, taught me the importance of complemention of knowledge to solve greater puzzles, because *in varietate concordia*. Moreover, you are five, like the Five Pieces of Exodia.

```
> to_thank=from_input()
> print(sort(unique(to_thank)))
> [Anto, Bastien, Ben, Eug, Garì, Giusi, Hilde, Javi, John, Lars, Leszek, Live, Malte, Mario,
Nikki, Raju, Sabina, Shaun, Simone, Sofi, Susana, Zarah]
```

And to you, the reader, if you find a typo, please do not report it to me.

F.D.

SUMMARY

Microbes are everywhere and contribute to many essential processes relevant for planet Earth, ranging from biogeochemical cycles to complex human behavior. The means to achieve these colossal tasks for such small and, at first glance, simple organisms rely on their ability to assemble in heterogeneous communities in which populations with different taxonomies and functions coexist and complement each other. Some microbes are of particular interest for human civilization and have long been used for everyday tasks, such as the production of bread and wine. More recently, large-scale industrial and civil projects have taken advantage of the transformative capabilities of microbial communities, with key examples being biogas reactors, mining and wastewater treatment.

Decades of classical microbiology, based on pure culture isolates and their physiological characterization, have built the foundations of modern microbial ecology. Molecular analysis of microbes and microbial communities has generated an understanding that for many microbial populations cultivation is hard to achieve and that breaking a community apart impacts its function. These limitations have driven the development of technical tools that bring us directly in contact with communities in their natural environment. In the mid 2000's the recently established "omics" techniques were quickly adapted to their "meta-omics" version, enabling direct analysis of the microbial samples without culture. Every class of molecules (DNA, RNA, protein, metabolite, etc.) can now theoretically be analyzed from the entire community within a given sample. Metagenomics uses community DNA to build the phylogenetic picture and the genetic potential, whereas metatranscriptomics and metaproteomics employ RNA and proteins respectively to inquire the gene expression of the community. Finally, meta-metabolomics can close the loop and describe the metabolic activity of the microbes.

Here, we combined the four aforementioned major meta-omics disciplines in a gene- and population-centric perspective to re-iterate the same Aristotelian question underlying microbial ecology: how is it possible that the whole is more than the sum of its parts? Along the detailed answers provided by the individual communities in various environments, we also tried to learn something about biology itself. We first addressed in a saccharolytic and

methane-producing minimalistic consortium (SEM1b), the strain-specific interplay engaged in (hemi)cellulose degradation, explaining the ubiquity of *Coprothermobacter proteolyticus* in biogas reactors. We showed through the genetic potential of the *C. proteolyticus*-affiliated COPR1 population, the putative acquisition via horizontal gene transfer of a gene cassette for hemicellulose degradation. Moreover, we showed how the gene expression of these COPR1 genes were both coherent with the release of hemicellulose by another population of the community (RCLO1) and synced with the gene expression of the orthologous genes of an already known hemicellulolytic population (CLOS1). Conclusively, we demonstrated how the same purified COPR1 protein (Glycosyl Hydrolases 16) showed endoglucanase activity on several hemicellulose substrates.

Secondly, we explored the combined application of absolute omics-based quantification of RNA and proteins using SEM1b as a benchmark community, due to its lower complexity (less than 12 populations) and relatively resolved biology. We subsequently demonstrated that the uncultured bacterial populations in SEM1b followed the expected protein-to-RNA ratio (10^2-10^4) of previously analyzed cultured bacteria in exponential phase. In contrast, an archaeon population from SEM1b showed values in the range 10^3-10^5, the same as what has been reported for eukaryotes (yeast and human) in the literature. In addition, we modeled the linearity (k) between genome-centric transcriptomes and proteomes over time and used it to predict the essential metabolic populations of the SEM1b community through converging and parallel k-trends, which was subsequently confirmed via classical pathway analysis. Finally, we estimated the translation and the protein degradation rates, coming to the conclusion that some of the processes in the cell that require a rapid tuning (e.g. metabolism and motility) are regulated (also) post-transcriptionally.

Thirdly we sought to apply our approach of collapsing complex datasets into simplistic metrics in order to identify underlying community trends, onto a more complex and "real-world" microbiome. To do this, we resolved more than one year of weekly sampling from a lipid-accumulating community (Shif-LAO) that inhabits a wastewater treatment in Shifflange (Luxembourg), and showed an extreme genetic redundancy and turnover in contrast to a more conservative trend in functions. Moreover, we demonstrated how the time patterns (e.g. seasonality) in both gene count and gene expression are linked with the physico-chemical parameters associated with the corresponding samples. Furthermore, we

built the static reaction network underlying the whole community over the complete dataset (51 temporal samples). From this, we characterized the sub-network for lipid accumulation, and showed that its more expressed nodes were defined by resource competition between different taxa (deduced via inverse taxonomic richness and gene expression over time). In contrast, the nitrogen metabolism sub-network instead exhibited a dominant taxon and a keystone ammonia oxidizing monooxygenase, the first enzyme of ammonia oxidation, which may lead to the production of nitrous gas (a powerful greenhouse gas).

Overall, our results presented in this thesis build a comprehensive repertoire of interactions in microbial communities ranging from a simplistic (10's of populations) consortium to a natural complex microbiome (100's of populations). These were ultimately uncovered using an array of techniques, including unsupervised gene expression clustering, pathway analysis, reaction networks, co-expression networks, eigengenes and linearity trends between transcriptome and proteome. Moreover, we learnt that to achieve a full understanding of microbial ecology and detailed interactions, we need to integrate all the meta-omics layers quantified with absolute measurements. However, when scaling these approaches to real-world communities the massive amounts of generated data brings new challenges and necessitates simplifying strategies to reduce complexity and extrapolate ecological trends.

SAMMENDRAG

Mikroorganismer er overalt og de bidrar til mange essensielle prosesser som er viktige for planeten vår, alt fra biokjemiske sykluser til kompleks menneskelig oppførsel. Midlene disse små, og ved første øyekast enkle organismene bruker for å oppnå så betydelige oppgaver på, ligger i deres evne til å forenes i et heterogent samfunn der ulike populasjoner med en forskjellig taksonomi og funksjoner sameksisterer og utfyller hverandre. Noen mikrobielle samfunn er av særlig interesse for oss mennesker, og har i lang tid blitt utnyttet i hverdagslige gjøremål, slik som produksjon av brød og vin. I senere tid har også stor-skala industri og kommunale anlegg, for eksempel biogass reaktorer og renseanlegg, dratt nytte av mikrobesamfunns evne til å transformere.

Tiår med klassisk mikrobiologi, basert på dyrking og fysiologisk karakterisering av renkulturer har bygget grunnlaget for moderne mikrobiell økologi. Molekylære analyser av mikrober og mikrobielle samfunn har resultert i forståelsen om at mange mikrobielle populasjoner er vanskelige å kultivere, og at en oppdeling av samfunnet vil påvirke dens funksjoner. Disse begrensningene har vært en drivkraft for utviklingen av tekniske verktøy som kan bringe oss i direkte kontakt med mikrobesamfunnet i deres naturlige miljø. I midten av 2000-talles ble de nylig etablerte «omikk»-teknikkene raskt adoptert til også å gjelde «meta-omikk», som muliggjør direkte analysering av mikrobielle samfunn uten kultivering. I dag kan i teorien hver molekylerære klasse (DNA, RNA, proteiner, metabolitter, osv.) bli analysert fra hele mikrobesamfunn i en bestemt prøve. I metagenomikk benyttes DNA-innholdet til å konstruere et fylogenetisk bilde av samfunnet og det genetiske potensiale, mens metatranskriptomikk og metaproteomikk bruker henholdsvis RNA og proteiner for å se på gen-uttrykket i samfunnet. Meta-metabolomikk kan slutte sirkelen ved å beskrive den metabolske aktiviteten til mikrobene.

I arbeidet som ligger til grunn for denne avhandlingen, kombinerte vi fire av de nevnte fagfeltene innen meta-omikk i et gen- og populasjons-orientert perspektiv for å gjenta det samme Aristoteliske spørsmålet bak mikrobiell økologi: hvordan er det mulig at helheten er større enn summen av enkeltdelene? Sammen med de detaljerte svarene som ble gitt av de enkelte mikrobesamfunnene i ulike miljøer, forsøkte vi også å lære noe om biologi i seg

selv. Først adresserte vi det stamme-spesifikke samspillet involvert i (hemi)cellulose degradering i et sakkarolytisk og metan-produserende minimalistisk konsortium (SEM1b), som belyser omfanget av *Coprothermobacter proteolyticus* i biogass reaktorer. Gjennom det genetiske potensiale til COPR1-populasjonen tilknyttet *C. proteolyticus*, viste vi den antatte ervervelsen, via horisontal gen-overføring, av en gen-kassett for nedbrytning av hemicellulose. Videre viste vi hvordan genuttrykket til disse COPR1-genene var i samsvar med frigivelsen av hemicellulose av en annen populasjon i samfunnet (RCLO1), og synkronisert med genuttrykket av de ortologe genene fra en allerede kjent hemicellulolytisk populasjon (CLOS1). Avslutningsvis demonstrerte vi hvordan det samme rensede COPR1-proteinet (glykosid-hydrolase 16) viste endoglukanase-aktivitet på flere hemicellulosesubstrater.

På grunn av lavere kompleksitet (færre enn 12 populasjoner) og en relativt kjent biologi, benytte vi SEM1b videre som et referansesamfunn for å utforske den kombinerte anvendelsen av absolutt omikk-basert kvantifisering av RNA og proteiner. Vi demonstrerte deretter at de ukultiverte bakterie-populasjonene i SEM1b fulgte en protein-til-RNA ratio (10^2-10^4) som var forventet basert på tidligere analyser av bakteriekulturer i eksponentiell fase. I kontrast til dette viste en arkeonpopulasjon fra SEM1b verdier i området mellom 10^3-10^5, som er det samme som tidligere rapportert i litteraturen for eukaryote (gjær og menneske). I tillegg modellerte vi lineariteten (k) mellom genom-orienterte transkriptomer og proteomer over tid, og brukte dette til å forutsi de essensielle metabolsk populasjon i SEM1b-samfunnet gjennom konvergerende og parallelle k-trender, som senere ble bekreftet via klassiske analyser av metabolske synteseveier. Til slutt estimerte vi frekvensen av translasjon og protein degradering, hvorpå vi konkluderte med at noen av prosessene i en celle som krever rask innstilling (som for eksempel metabolisme og bevegelse) er regulert (også) post- transkripsjonelt.

Til slutt ønsket vi å anvende vår tilnærming for å sette komplekse datasett inn i forenklede matriser for å identifisere underliggende trender i mikrosamfunnet, på et mer komplekst og virkelighetsnært mikrobiom. Til dette benyttet vi et mer enn ett år med ukentlige prøvetakninger fra en lipid-akkumulerende mikrobesamfunn (Shif-LAO) i et renseanlegg i Shifflange (Luxembourg), og avdekket en ekstrem genetisk redundans og turnover, i

motsetning til en mer konservativ trend i funksjoner. Videre demonstrerte vi hvordan tidsavhengige mønstre (som for eksempel sesongvariasjoner) i både antall gener og genuttrykk er knyttet til fysisk-kjemiske parameter assosiert med de tilsvarende prøvene. I tillegg rekonstruerte vi det underliggende statiske reaksjonsnettverket til mikrobesamfunnet over hele datasettet (51 prøver over tid). Basert på dette, karakteriserte vi sub-nettverk for lipid-akkumulering, og demonstrerte at mer uttrykte noder var definert av konkurransen om ressurser mellom ulike taksonomiske grupper (antatt via reversert taksonomisk diversitet og genuttrykk over tid). I motsetning til dette, viste nettverket for nitrogen-metabolismen i stedet et dominerende taxon og en keystone ammoniakk-oksiderende monooxygenase, det første enzymet i ammoniakk oksidasjon, som fører til produksjonen av lystgass (en svært sterk klimagass).

Resultatene presentert i denne doktorgradsavhandlingen bygger på et omfattende repertoar av interaksjoner i mikrobielle samfunn som spenner fra et forenklet konsortium (titalls populasjoner) til et naturlig komplekst mikrobiom (hundretalls populasjoner). Disse mikrobiomene ble til slutt kartlagt ved hjelp av en rekke teknikker, blant annet unsupervised gruppering av genutrykk, analyser av metabolisk synteseveier, nettverk av reaksjoner og co-uttrykte gener, eigengener og lineære trender mellom transkriptom og proteom. I tillegg erfarte vi at for å oppnå en full forståelse av mikrobiell økologi og detaljerte interaksjoner må vi integrere alle lagene av meta-omikk, kvantifisert med absolutte målinger. Når man oppskalering disse tilnærmingen til virkelige mikrobesamfunn, bringer imidlertid enorme mengder generert data til nye utfordringer som nødvendiggjør en forenkling av strategier for å redusere kompleksiteten og ekstrapolerer økologiske trender.

ABBREVIATIONS

16S rRNA	16S ribosomal ribonucleic acid
BWWTP	Biological wastewater treatment plant
CAZyme	Carbohydrate-active enzyme
cDNA	Complementary deoxyribonucleic acid
CKO	Collapsed KEGG orthology
ddNTP	Dideoxynucleotide triphosphate
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleotide triphosphate
EG	Eigengene
GH	Glycosyl hydrolase
LAO	Lipid-accumulating organism
KO	KEGG orthology
MAG	Metagenome-assembled genome
MG	Metagenomics
MT	Metatranscriptomics
MP	Metaproteomics
MB	Meta-metabolomics
NMR	Nuclear magnetic resonance
OMMC	Oleaginous mixed microbial community
ORF	Open reading frame
ORFG	Open reading frame group
PCR	Polymerase chain reactions
SCFA	Short chain fatty acid
WGCNA	Weighted correlation network analysis
WWTP	Wastewater treatment plant

LIST OF PAPERS

Paper I

From proteins to polysaccharides; lifestyle and genetic evolution of *Coprothermobacter proteoliticus.* B.J. Kunath*, F. Delogu*, A.E. Naas, M.Ø. Arntzen, V.G.H. Eijsink, B. Henrissat, T.R. Hvidsten, P.B. Pope (2019) ISME J. **13** 603–617

Paper II

Integration of absolute multi-omics reveals translational and metabolic interplay in mixed-kingdom microbiomes. F. Delogu, B.J. Kunath, P.N. Evans, M.Ø. Arntzen, T.R. Hvidsten, P.B. Pope (2020) Nat. Commun.. *In review.* bioRxiv doi: 10.1101/857599

Paper III

Functional dynamics of a microbial community from a wastewater treatment plant. F. Delogu, S. Martinez-Arbas, B.J. Kunath, M. Herold, J. Garcia, P.B. Pope, P. May, P. Wilmes (2020). *Manuscript*.

*** Authors contributed equally to the work**

Other publications by the author:

Self-domestication in Homo sapiens: Insights from comparative genomics. C. Theofanopoulou*, S. Gastaldon*, T. O'Rourke, B.D. Samuels, P.T. Martins, F. Delogu, S. Alamri, C. Boeckx (2017) PloS one **12**, 10

Reverse engineering directed gene regulatory networks from transcriptomics and proteomics data of biomining bacterial communities with approximate Bayesian computation and steady-state signalling simulations. A. Buetti-Dinh, M. Herold, S. Christel, M. El Hajjami, F. Delogu, O. Ilie, S. Bellenberg, P. Wilmes, A. Poetsch, W. Sand, M. Vera, I.V. Pivkin, R. Friedman, M. Dopson (2020) BMC Bioinformatics **21**, 23

Interspecies and intersubstrate comparison of the biomass-degrading enzyme repertoires of five filamentous fungi. M.Ø. Arntzen*, O. Bengtsson*, A. Várnai, F. Delogu, G. Mathiesen, V.G.H. Eijsink (2020) Sci. Rep.. *In review*.

*** Authors contributed equally to the work**

1 INTRODUCTION

1.1 Microbial ecology

The lineages of Microorganisms are dispersed everywhere on planet Earth[1] and make up to ~17% of the carbon biomass[2]. Among microbial ranks we count Bacteria, Archaea, Viruses, Protists and Fungi, with the last two belonging to the kingdom Eukarya. Not only are microbes ubiquitous, but they took part in the mastodontic biogeochemical processes that shaped our planet, such as the Great Oxidation Event in which the Cyanobacteria increased the oxygen level causing the largest extinction event so far[3]. Microbes still control important cycles today, such as the carbon and the nitrogen ones in soil[4,5]. Some of them live in inside other organisms, such as plants and mammals, augmenting their metabolic capabilities and generally contributing to their health[6,7].

1.1.1 A mechanistic view of microbial cells and populations

Biological cells are populated by a multitude of molecules that scientists seek to describe, both as their individual components and how they interact with one another. Moreover, how life itself is founded can be thought of via two central pillars: **the propagation of information** and **maintaining homeostasis**. Information is the set of directives on how an organism should be and function, and commonly takes the name of **genotype**. To propagate information is to propagate life. Homeostasis is the ability of an organism to maintain certain properties, in the context of their surrounding environment. The implementation of the biological information to propagate and maintain the homeostasis gives shape to the **phenotype** of the organism.

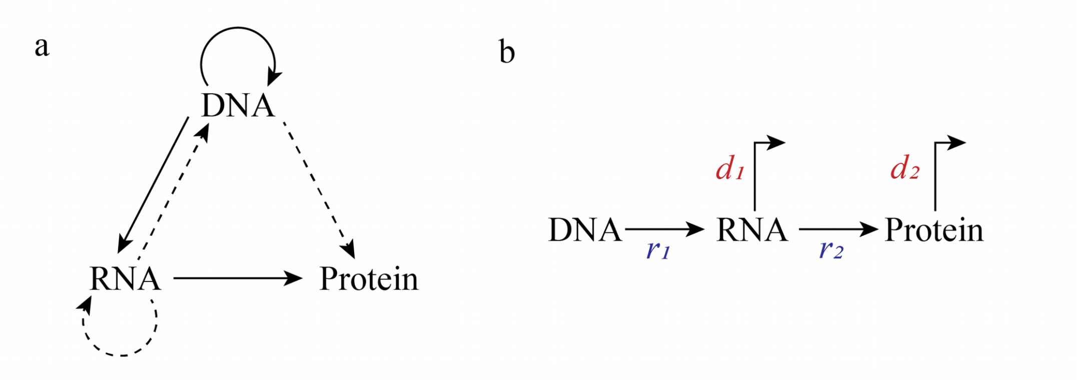


Fig1: a The Central Dogma of molecular biology from Crick 1970[8]. The solid arrows are the conventional paths of biological information: DNA replication, transcription and translation. Dashed arrows represent the unconventional paths: RNA replication, reverse transcription and direct translation from DNA. **b** Kinetics of the conventional paths of the biological information flow. DNA is transcribed with rate r_1 into RNA, which is then translated into protein with r_2. Both RNA and protein are subject to degradation, with rates d_1 and d_2 respectively. Measured degradation of those molecules was shown to be an "apparent measurement", comprising both molecular dilution through cell division and actual degradation. The fine tuning of these four rates allows the cell to reach and maintain the functions needed to express its phenotype.

The **central dogma** of molecular biology is the core rule of life as we know it. It states the direction of the information flow in the cell among DNA, RNA and proteins (**Fig. 1a**). The underlying dynamics have been studied for decades using absolute quantification of the molecules involved, characterization of the molecular machinery and inference of the regulatory network. The most schematic representation of the central dogma sees the information stored in the DNA **transcribed** into RNA that is then **translated** into proteins. However, during the formulation by Crick in 1958[9] and the restatement in 1970[8], it has been predicted that information could flow "backward" from RNA to DNA, which was proved real in the viral process of **retrotranscription**[10]. Also, the direct translation from DNA to protein had to be added among the possible paths[11,12]. Moreover, the dogma states that DNA and RNA can **replicate** themselves and both can be used as storage for biological information. While this statement seems obvious for DNA, it was proved true for RNA only decades after when RNA-based viruses were discovered[13].

Certain regions of DNA called genes can be used to produce gene products, i.e. RNAs and subsequently proteins, according to the central dogma[14]. These processes are often collectively referred to as **gene expression**. Gene expression is the dynamic tool that the cell uses to modify its internal status and interact with the environment. RNA levels are controlled by the regulation of transcription and degradation. In Bacteria the average half-life of RNA molecules is 2-10 minutes, which implies a quick recycle of nucleotides into new transcripts[15]. This makes the control of transcription the main target of regulation of gene expression in Bacteria[15]. Both in Prokaryotes and Eukaryotes the control of transcription can happen with a *cis* or a *trans* mechanism. *Cis* control is mediated by a DNA region placed in proximity of the gene which is targeted by **transcription factors**. More than one gene can be influenced by the same *cis* element. A common group of *cis* elements are the promoters that control one or more downstream genes. A group of genes under the regulation of the same promoter is called an **operon**. Translation can be also regulated at the RNA Polymerase level, in a phenomenon called *trans* regulation[16]. A classic example is the set Bacterial σ-factors which associate with the Polymerase and have different affinity for separate groups of promoters[17,18].

Similar to RNA, protein levels are controlled via regulation of translation, "control by dilution"[19] (dispersal of proteins via subsequent cell divisions) and rarely by protein degradation. A notable exception is the presence of pupylation, a mechanism to tag proteins for degradation in some bacterial taxa[20]. Like transcription control, translation can also be controlled by a dynamic pool of **translational factors**, such as initiation, elongation and ribosome components. Nevertheless, the control of transcription is believed to be the most important factor in the overall control of protein levels in Bacteria[21].

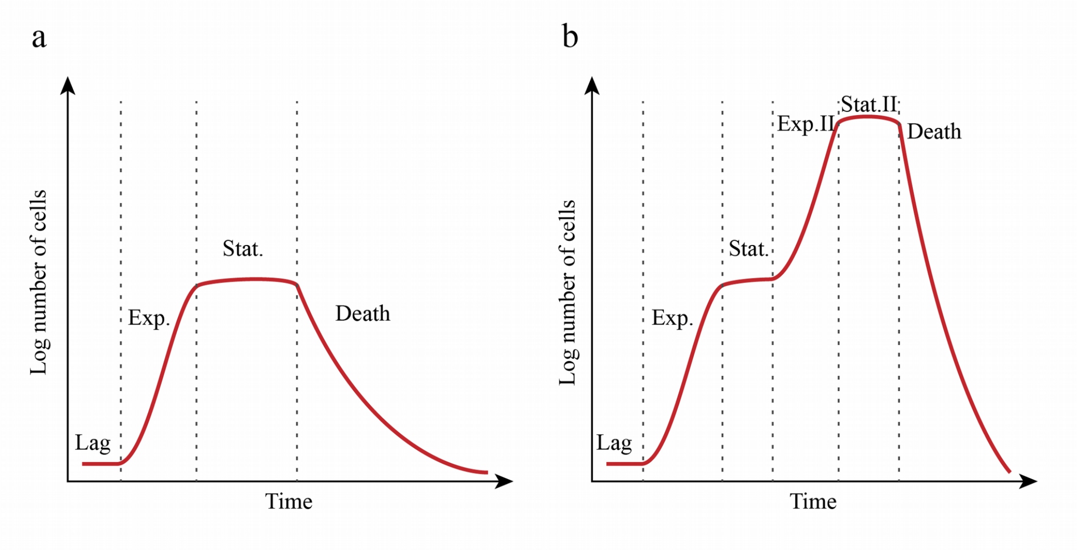


Fig 2. a. A typical bacterial colony growth curve. After inoculation the community strives to grow (lag), then, after adapting to the new environment its number grows exponentially (exp.) until resource depletion (stat.), finally it reduces (death). **b.** Double-growth colony curve. In case of injection of new substrate or shift to a secondary one, it is possible to ignite a second exponential growth of the community (exp.II), followed by another plateau when the resources are reduced again (Stat.II) until the community collapses at the end (death).

The complexity of the internal machinery of the microbial cell is however only one level of microbial life. Indeed, many microbial cells live in the same physical space and share a common ancestry (e.g. they are descendants of the same cell). In case of prokaryotic cells, when they also share the same taxonomy, this is commonly referred to as a **population**. Thus, in most cases in microbiology the unity of inquiry is the population rather than the individual cell, which also facilitates easier experimental approaches (as discussed later). When favorable conditions are present (e.g. temperature, pH, substrate, etc.) the cells will consume essential substrates to create new constituent molecules and replicate themselves. This process is simply called **growth** at the population level. The study of the **growth curve** is probably the most rudimentary element of every microbiology course. The parameters describing the curve may vary depending on the culturing conditions, but four main phases are usually present (**Fig. 2a**).

1. Lag phase: after the population is inoculated in a fresh space (e.g. plate or flask) with the appropriate medium, the number of cells drops as a result of the inoculation stress (**Fig 2a**: Lag.).
2. Log phase: the population activates the primary metabolism, starts to consume the necessary substrates to sustain their metabolism and replicates at the maximum speed allowed by the culturing conditions, which results in exponential growth (**Fig 2a**: Exp.).
3. Stationary phase: when the available substrate decreases significantly the number of cells remains constant (the number one of division is the same as the number of deaths) and the population activates the secondary metabolism (toxins, spores, etc.) to deal with resource scarcity (**Fig 2a**: Stat.).
4. Death phase: when the substrate is completely depleted the population collapses (**Fig 2a**: Death).

In general, the log phase is the most studied among the phases and it is described by the **growth rate** (also known as **doubling time**) which measures the replication time of a cell.

There are some main variations on the standard growth curve worth mentioning here. If more substrate is added to the medium during the stationary phase, the population will start growing again following a new log phase (**Fig 2b**: Log.II) and reaching another stationary phase (**Fig 2b**: Stat.II) before declining (**Fig 2b**: Death). In the case of two substrates being present from the beginning, but with a population with different affinities for them, the growth curve resembles the previous case. The most palatable substrate is consumed first and the population has to adjust its metabolism (first plateau) before switching to the second one (**Fig 2b**). In the latter case, if more substrate is constantly added and part of the population (and catabolites and toxins if produced) is removed, the population is maintained in log phase. This technique is commonly performed with a dedicated machine called a **chemostat**[22]. Following generations of cells in log phase, the growth rate fluctuates until it reaches a stable value. The use of the chemostat is the only way so far known to measure the true growth rate of a population with given growth conditions.

Biological systems are dynamic, which means that the parameters that describe them are changing over time. However, some of these systems possess special dynamic equilibria in which the variations of certain parameters are zero (or really close to it), a phenomenon known as **Steady State** (**SS**). In microbiology a population is considered to be in SS during the log phase of the growth curve, because its growth rate remains constant. This implies that both the flow of metabolites in the population and the reaction rates are constant. Since the reaching of a stable value of the growth rate requires the extensive use of a chemostat, we must remember that the assumption that log phase = SS is an approximation.

The phenotype implemented by a cell can be summarized as its **cell status**. A population is a group of cells of the same species, and when we take a measurement on it the result is an averaged value. It is however known that specific sub-groups of cells may have different cell statuses in the same population, forming two or more **subpopulations**[23]. A typical example is the coexistence of a subpopulation of actively replicating cells and one of sporulating/quiescent cells. Diversification of functions is a convenient way to overcome environmental challenges and increase the chances of survival for the population.

1.1.2 A society of microbes

Like larger organisms, microbes often live in the same physical space as other microbes. Sharing the same environment, they are brought to interact with each other. Such an ensemble of microbes is called a (microbial) **community** or **consortium** and has more recently also been termed a **"microbiome"**. The interactions that pairs of microbial populations may form follow the classical schemes of ecological **interactions**[24] (**Fig 3a**). The outcome of the interactions are "positive", "neutral" and "negative", which leads to 3^2 = 9 possible combinations (considering only pairs). A common example of such interactions in microbial communities is a form of metabolic commensalism in which the first microbe uses the substrate to grow and releases another compound used as substrate by the second microbe (**Fig 3b**). In this case the first microbe could thrive regardless the second one, but the latter would not be able to sustain itself without its partner. Sometimes the compound produced by the first microbe inhibits its growth. In this case a second microbe, whilst benefiting directly form consuming its substrate, allows the first one to grow, establish a mutualistic relationship (**Fig 3b**).

The piling-up of pairwise metabolic relationships in a microbial community can lead to the formation of a complex metabolic network. Exploiting such arrangements, microbial communities can perform wider **community functions** (or **community processes**) as multi-step substrate conversion and even shape their environment[25]. Consortia usually exhibit a certain level of tolerance to stress due to functional redundancy of the species involved. In case of a change in the physico-chemical parameters the decrease in relative abundance and/or gene expression in one or more populations is balanced by one or more other populations. When the conditions allow for it, we can expect the community to reach a SS and carry out its overarching function at capacity. Consortia that present high level of resilience and stability have been exploited for industrial purposes as cheap matter processors such as in biogas reactors or in wastewater treatment plants.

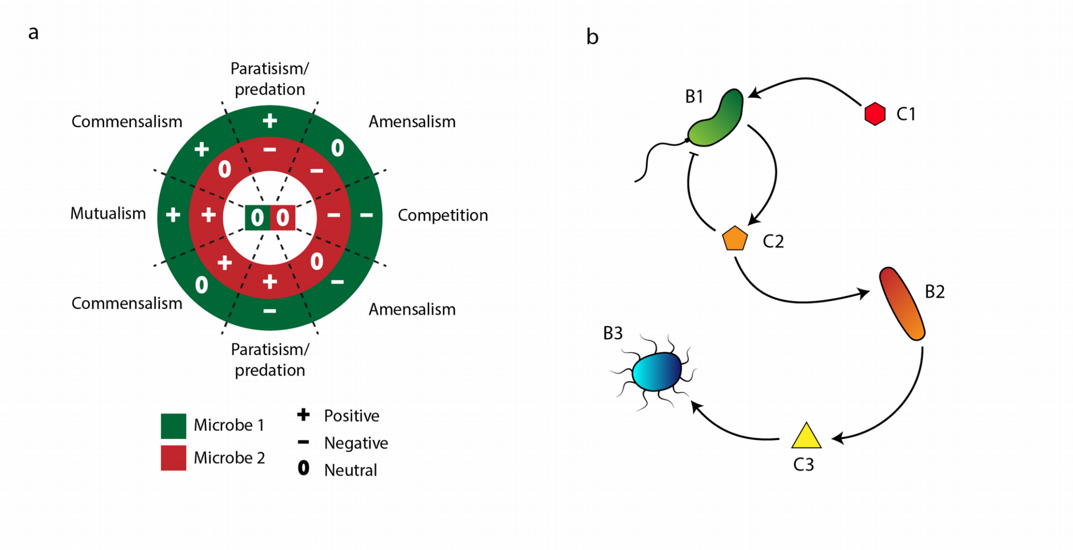


Fig 3. a. Wheel of pairwise ecological interactions from Faust et al.[24]. Two interacting microbes (green and red bands) and three one-sided outcomes (positive, negative and neutral) generate nine possible interaction types: parasitism/predation(x2), amensalism(x2), competition, commensalism(x2), mutualism and null interaction. **b.** Compounds (C1-3) are shared in a bacterial community (B1-3) outlying a metabolic network. B1 consumes C1 and produces C2, however it is inhibited by the latter. B2 consumes C2 and secretes C3, which is finally consumed by B3. In this case B1 and B2 have a mutualistic interaction because B1 is the only producer of C2 (its inhibitor) whilst (B2) is the only degrader of C2, allowing B1 to maintain its growth. B3 and B2 have a commensalistic interaction because B3 needs the activity of B2 to generate its substrate, whilst the activity of B3 is uninfluential to B2.

Of note are communities that live in association with (or even in symbiosis with) multicellular organisms which function as a **host**. A common example is the root nodule microbiome in the legumes which is responsible for the assimilation of nitrogen by the plant. Another example is the human gut microbiome which was known to complement the metabolic needs of the human (e.g. through vitamin synthesis). Recently the human microbiome has been linked to central aspects of humans' lives such as their metabolism (e.g. predisposition to obesity[26] and diabetes[27]) and mental condition (e.g. pathological ones like depression[28] and schizophrenia[29]).

1.1.3 A brief history of microbial ecology

If microbiology started with the observation of "animalcules" (Bacteria) by Antoine van Leeuwenhoek at the microscope in the 17th century, it became a field of experimentation with Ferdinand Cohn, Louis Pasteur and Robert Koch during the 19th century. A student from Koch's lab, Julius Petri, standardized the cultivation of microbes on a medium solidified using agar: The Petri dish. The new technology led to the **isolation** of individual microbial populations from mixed colonies.

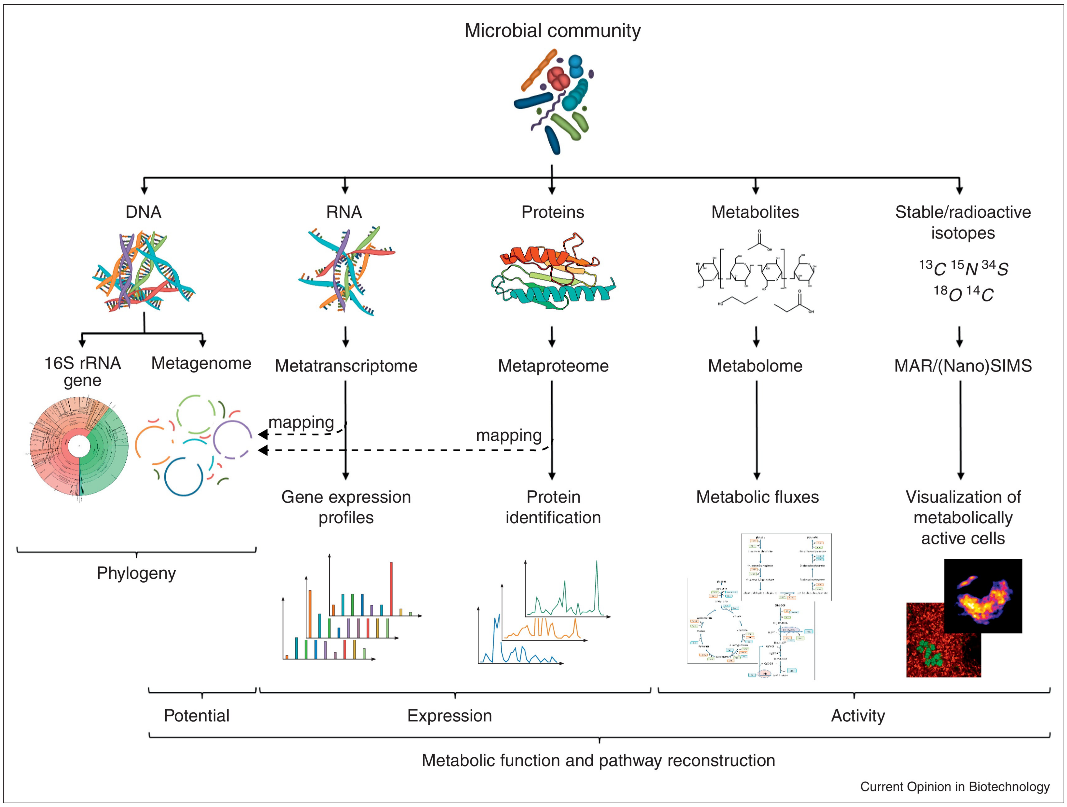
With the passing of time it became clear that something was missing from the greater picture of microbiology. An environmental sample observed directly under the microscope shows ~100 times more cells than colonies on the Petri dish inoculated with the same sample. This discrepancy was called "**the great plate anomaly**" and demonstrated that the vast majority of microbes could not be cultivated[30]. From the anomaly started the idea of culturing the microbes in conditions as close to their habitat as possible, if not accessing them directly from it.

With the advent of Polymerase Chain Reaction (PCR)[31] and the combination with DNA sequencing, microbes could be "massively" cataloged in their environment through **amplicon sequencing** techniques. The ideal DNA region to exploit has to be highly variable to easily identify different taxa, yet bordered by relatively conserved regions so that a single pair of PCR primers with degenerate bases can pair to all the Bacteria (and/or Archaea) in a given sample. This criteria led to 16S ribosomal gene becoming the keystone taxonomical marker gene, and subsequent collection of 16S rRNA gene sequences are constantly being generated and accumulated in huge databases. As a consequence of the new molecular approach to microbiology, the phylogeny of Prokaryotes was deeply altered and allowed microbial ecologists to efficiently barcode the microbial communities without cultivation.

The next leap in microbial ecology was brought to us in the *omics era*. Each different class of molecules in a single organism was interpreted as a single **-ome**. For example the collection of DNA material was referred as the *genome* of an organism. In this framework the first large-scale sequencing endeavors took place, such as the whole genome sequencing of many model organisms, human included (see section 1.1.3). The microbial populations were therefore characterized by their -omes, in technical approaches called **-omics** depending on their targeted -ome, such as genomics, proteomics, lipidomics, etc. Every class of molecules adds a new layer of information to the total picture of the microbe and its functions. The new paradigm of studying the microbial communities in their entirety pushed the adaptation of pre-existing technologies (such as sequencing) to be used on raw samples. The ensemble of the new approaches were labeled **meta-omics**, because they extended the previous techniques *beyond* the mere omics. With the two first metagenomics (MG) studies in 2004[32,33], the idea of fully sequencing the DNA from environmental samples was born and is still maturing today with the shift from Second generation sequencing to the Third (section 1.2.1). After the MG revolution, in time also metatranscriptomics (MT)[34], metaproteomics (MP)[35] and meta-metabolomics (MB)[36] were born. At this historical moment in microbial ecology we are facing the challenge to integrate the information from the different omics technologies in order to understand even more the microbial world, its components and their interactions.

1.2 Meta-omics methods

The two main experimental techniques to access the molecules contained in a microbial community are sequencing and mass spectrometry (MS). These techniques are profoundly different and require dedicated preparatory steps and subsequent analysis. Commonly the biological sample is separated into one batch for each omic layers that will be analyzed and is then processed with dedicated molecular extraction protocols (**Fig. 4**)[37,38]. However recently there has been a discussion about how much the splitting of the sample impacts the final reconstruction[39]. For instance, if the sample is divided in three to perform MG, MT and MP analysis, how can we know that each quantification performed in one third of the original sample is representative of the whole?



Current Opinion in Biotechnology

Fig 4. Schematics of the main meta-omics from Vanwonterghem et al.[40]. The microbial community is broken down into its constituent molecules (DNA, RNA, proteins and metabolites). From DNA it is possible to infer the phylogeny (markers genes) and the gene potential (metagenome). Mapping the metatranscriptome and

metaproteome on the metagenome-reconstructed community allows to quantify the gene expression. The metabolites can be used to assess the metabolic fluxes in the community to gain insight of the actual metabolic activity.

Another notable caveat of the omics technologies is their cost. The price of sequencing per kilobase has become cheaper over time while MS has remained expensive. Also, the price of extraction kits factors into the decision of how many and which omics to examine, with an RNA extraction kit being commonly more expensive (and the extraction itself more labor intensive) than the RNA-seq run itself.

1.2.1 Metagenomics

DNA constitutes the information template of the cell in the form of its genome. Usually Bacteria and Archaea have one single circular chromosome and they may possess extrachromosomal elements known as plasmids. The aim of MG is to reconstruct the DNA content of the microbial constituents in a community. Thereafter we can retrieve the taxonomy of the inherent microbes and their genetic potential (the genes encoded in the genome). Moreover, we can predict the collective functions of the community alongside hypothetical metabolic niches of individual populations.

There are three main steps to take into account when producing a MG dataset: DNA extraction, fragmentation and sequencing. Although DNA extraction is nowadays almost a trivial process, it is debated how much the differences/biases introduced in this initial step are influencing the final MG results and hindering the ability to compare data from different studies. It is good practice to adhere to the standards proposed for the microbial community under scrutiny and if not possible to add quality control samples[41]. These samples should come from either i) a complex sample (stool); ii) a chemostat culture of model community (see section 1.1.1); or iii) a mock community.

Fragmentation is the process of shredding the DNA into pieces of a chosen length. The length is chosen according to the sequencing technique and since the shredding process produces a distribution of length, the pieces are selected to be as close as possible to the selected length. The selected pieces are usually referred as **fragments** or **inserts**. The length of the fragments can be used downstream to help the data processing (e.g. during binning or read mapping).

DNA sequencing first started in 1977 with the Sanger and Maxam-Gilbert methods, in the so-called "**First generation of sequencing**". The Sanger method used a mix of normal and chemically modified nucleotides (dNTPs and ddNTPs) during the synthesis reaction of DNA. When integrated in the nascent strand, the ddNTPs blocked the synthesis reaction and generated a DNA fragment. Four different reactions were run with the four modified nucleotides and the resulting fragments were run in parallel lanes in a gel. In this way, the length of the fragment and the gel lane indicated which nucleotide was occupying which base and the DNA sequence could be read similarly to a punched card. This technique led to historical results such as the sequencing of the lambda bacteriophage in 1980, *Arabidopsis thaliana* in 2000 and *Homo sapiens* in 2001.

The "**Second generation of sequencing**" changed the paradigm introducing massive parallel reactions, increasing the throughput whilst slashing the costs and the time. Although Roche 454, based on pyrosequencing, was used to produce the first MG dataset[32,33], nowadays the standard technique is Illumina sequencing (originally developed by Solexa), which similarly to Sanger, is based on DNA synthesis. In Illumina sequencing, the DNA is firstly fragmented to the desired length, then the fragments are fixated (expectedly at a certain distance from one another) on a plate with a system of adapters. Here a "PCR bridge amplification" is performed to create a cluster of clonal copies of the original fragment. Finally, the synthesis occurs with cycles of reversibly-blocked nucleotides in which the blocker is also a fluorescent (one for each nucleotide type), florescence readings and washings.

Illumina HiSeq (the most commonly used machine for MG) produces up to 5×10^9 reads per run which are 150 nt long and with an error rate of 0.1% (one base out of 1000 is expected to be incorrect). Moreover, the Illumina technology allows the sequencing of both ends of the fragment, producing **paired-end reads**.

The "**Third generation of sequencing**" introduced the idea of "**single molecule sequencing**". The previous generations required amplification (PCR or a variant) to increase the signal and help the detection, which is time consuming, costly and prone to error. There are two main technologies in use at the time being: SMRT from PacBio and the Oxford Nanopore. The SMRT technology uses a sequencing by synthesis process in which the incorporated nucleotides are marked with a fluorescent. In this case however the synthesis occurs in a microwell in which the DNA polymerase is fixed and the diameter of the well is shorter than the wavelength of the emitted light, so that there is no backward propagation. The bottom of the well is a glass plate, and the light emission is read from there. The reads can be tens of kilobases in length but the error rate reaches up to 15%. The error rate can be decreased by using a circular consensus strategy in which the fragment is circularized and the sequencing continues on the target covering it several times. The post-sequencing processing reconstructs the original fragment with a reported error rate of 0.001%. With the Oxford Nanopore, the fragment is forced to pass through a protein nanopore fixated on a membrane while an electric current is applied. The steric hindrance of the nucleotide occupying the pore causes a change in the amount and placement of water molecules, resulting in a characteristic change in the current passing through the pore. The electric current is constantly registered and the changes over time are interpreted to decode the DNA sequence. A small device (minION) under 100g and connected to a mundane laptop can produce up to 30Gb of DNA sequences with a 5% error rate (nanopore R10.3).

A very common issue during sequencing is to estimate the amount of reads necessary to reconstruct the starting genome, commonly referred as **sequencing effort**. To compute this value, it is necessary to know the length of the genome, desired **coverage** (number of reads mapping at any given position of the genome) and length of the reads. However, in MG studies we have to deal with many populations (up to 100's-1000's in soil or sea samples) which probably possess even internal variability and with different population-specific abundance, hence the old-school approach is not a possibility. The obvious question being: is it possible to estimate something like genome length and population abundance before a MG study? It is not. The available options are to run a preliminary study to try to retrieve this information in an indirect way such as rarefaction curves[42] or to pool many different samples and combine short and long reads[43].

1.2.2 Metatranscriptomics

RNA is the second molecule in the canonical path of the Central Dogma (section 1.1.1). RNA is made of nucleotides like DNA, with the exception of Thymine swapped with Uracil, it is single stranded and usually has a short half-life. It can form secondary and tertiary structures and some RNA molecules have catalytic properties. One notable RNA is the ribosomal RNA (rRNA) which constitutes the core of the translation machine: the ribosome. The RNA species we are commonly interested in are the messenger RNA (mRNA), which is the product of gene transcription and awaits to be translated into a protein. Given the usually short half-life of RNAs (section 1.1.1), the metatranscriptomic profiling allows to record what are the current responses to stimuli (internal and external) of the community.

Studying the MT starts with the RNA extraction, which is more difficult than DNA extraction since RNA is single stranded and hence a less stable molecule. Moreover, like for DNA extraction, a sample-specific protocol is preferable. After this step, the unwanted RNA species should be removed. In the case of Eukaryota it is possible to use the characteristic poly-A tail of their mRNAs to select them. However, this is not possible in Prokaryota and usually the rRNA (which make up to 80% of the RNA in a sample[44]) is removed and the rest is kept for downstream analysis. The remaining RNA is retrotranscribed into complementary DNA (cDNA) using a genetically modified viral retrotranscriptase (hence bouncing back in one of the uncanonical paths of the Central Dogma). Finally, the MT can follow the MG procedure for sequencing. Sequencing in this case is commonly called RNA-seq. Other methods such as the microarrays have been the standard for RNA studies for decades and pushed the development of the computational methods associated with transcriptomics, but they have been outcompeted by RNA-seq over time. In RNA microarrays, short DNA sequences called probes were designed to be complementary to desired RNA sequences and attached (or directly synthesized) in clonal clusters, called spots, to a chip (usually a glass slide). After retrotranscription the cDNAs from the sample were labeled with a fluorescent marker, allowed to hybridize with the probes and then excess one washed away. The chip was therefore excited and “read” with the appropriate wavelengths recording the signal intensity per spot. The data were decoded linking each spot to the known probe sequence and normalizing the light signal.

A variant to the standard procedure for MT is the introduction of one or more **spike-in transcripts** (custom RNA molecules in known amount) at the beginning of RNA extraction. The addition of spike-ins allows for the absolute quantification of the transcripts in the sample with some dedicated steps in the data processing (section 1.3.2).

1.2.3 Metaproteomics

As with both DNA and RNA, the first step of MP is the extraction of the protein molecules of interest. However, conversely to the other omics technologies, MP present a unique problem: many proteins are encased in or associated to the cell membrane (also briefly addressed in **Paper II**). Therefore, the two caveats for MP extraction become: environment/sample specific method[45,46] and optimization of **membrane protein** yield[47].

In the most used approach to MP the proteins are digested (usually with the endoprotease trypsin) and the mass of the peptides is measured with mass spectrometry (MS). A useful subtype of bottom up MP called "shotgun" MP uses separation techniques such as high-performance liquid chromatography (HPLC) on the digested peptides before the MS measurements. Another technical improvement is the use of **tandem mass spectrometry** (**MS/MS**) in which two (orthogonal) mass analyzers are coupled within one mass spectrometer. The first mass analyzer (MS1) separates the peptides by their **mass-to-charge ratio** (m/z) and the peptides having a mass corresponding to a desired interval of this measure are selected for further fragmentation and measurement in the second mass analyzer (MS2). The output is a series of spectra recordings in three dimensions (m/z, time, relative abundance), which are subsequently matched against the previously computed database in order to find which peptides they originated from (section 1.3.2).

Obtaining absolute measurements in proteomics in general is relatively harder than in sequencing-based omics and several techniques require the use of isotopic labeling[48]. However, it is also possible to use an approach that does not require more instrumental effort than measuring the amount of protein after extraction. This method is called "Total protein approach"[49], which we adapted herein to a microbiome setting in **Paper II**.

1.2.4 Meta-metabolomics

While MG potentially predict the functions of the community, MT and MP quantification assess how much the community is actively implementing the functions they can perform. Both approaches can identify and quantify active functions. However, these omic-layers alone cannot be considered an ultimate proof and they do not allow to compute the biochemical rates at which these functions are operating. There are several reasons why functional omics-based analysis can build a misleading picture of the community: a fully folded enzyme may not be working or be inefficient, a competitive reaction may be using all the substrate, post translational modification, etc. In this context, even when we have MG, MT and MP data, we are still missing the last molecule class which stores this final piece of the puzzle: the metabolites. Meta-metabolomics (MB) is the large-scale study of the metabolic profile, usually compounds smaller than 1000 Dalton, in a given system. However, it does not exist as an overarching technique "to rule them all", and often a combination of technical assets is used. For instance, a chromatographic column can be loaded with different stationary phase components in order to bind to various metabolites of interest. Therefore, allowing the whole MB to be broken down into smaller experimental tasks.

The first technique to be used for metabolic profiling was nuclear magnetic resonance (NMR) during the 1940s, and its sensitivity improved over the decades to come, being applied in different biological systems[50,51]. More recently MS has been used more often and with better results[36,52]. As seen for MP, several specialized variants of MS' have been applied in MB, such as MS coupled with separation techniques[53,54]. Moreover, the highest quality of result is obtained through **targeted MB**, which applies sever techniques in parallel for different molecules, optimizing the individual outputs[55]. We used targeted MB in **Paper I** and **Paper II** to detect short chain fatty acids, monosaccharides and gas percentages.

1.3 Meta-omics analysis and integration

The two first aims to achieve with omics are to characterize and quantify. In order to characterize we need to reconstruct a qualitative picture of the molecules in the sample from which they originated. Sequencing-based data (reads) can be used to build a set of contiguous sequences (**contigs**) without the help of supplementary data. Contigs may represent fragments of DNA (MG) or RNA (MT) from which we can predict genes and assign taxonomic groups according to sequence feature and coverage. MP and MB data are usually not self-sufficient to characterize the molecule directly, but require a reference to be used in concert. However, some labor-intensive and time-consuming techniques exist for reference-free identification, such as peptide de novo sequencing. Building the reference, also called database, is a crucial step because the detection power of the dependent omic layer will be greatly affected (section 1.3.2). The quantification of the molecules always requires the use of a reference that it can be compared to. In the case of MP and MB, characterization and quantification are performed together since a reference is required from the beginning, whilst in MG and MT the quantification is a separate manual procedure.

1.3.1 Metagenomic assembly and binning

The common first step in multi-omic analyses of microbial communities is to reconstruct the least dynamic part of the sample, i.e. the MG. The reads from community DNA sequencing (section 1.2.1), when coming from the same genome are expected to hold redundant information, i.e. they are expected to overlap (the amount of overlap is usually linked to the coverage, section 1.2.1). The reconstruction of the community exploits this feature in a process called **assembly,** of whose final aim is to produce the longest continuous sequences possible according to the chosen assembly strategy and quality standard.

The first assembly method developed was the **string graph** and was based on sequence alignment (**Fig. 5**). In a very intuitive manner, all the pairwise alignments between reads were computed and charted in a graph structure (intro to graph theory in section 1.3.3). The paths in the graph that satisfy the user's criteria are selected and provided as contig in the output[56]. A string graph has an execution time that grows quadratically with the number of reads[57]. For this reason, string graphs fell out of fashion with the increase in the amount of reads produced by more modern technologies belonging to the second generation of sequencing. However, the string graph has seen a revival with the expansion of the extremely long yet less numerous reads produced in MG studies with the third generation of sequencing technologies[58].

The most used assembly strategy, the **deBruijn graph[59,60]**, boomed with the second generation of sequencing and elegantly manages and exploits the vast number of reads produced (**Fig 5a**). In this strategy the reads are broken into words of size *k* called **k-mers**. These words are used as vertices in a graph structure (intro to graph theory in section 1.3.3) and two of them are connected by an edge if the *k-1* prefix of one of them is the same as the *k-1* suffix of the other. In the end the graph is traversed to find the contigs. The main advantage of using the deBruijn graph is that the time and space complexity scale linearly with the number of k-mers in the reads. Moreover, in case of assembly of a single organism, the time is bound to the length of the genome to reconstruct because the number of different k-mers is approximately equal to the length of the genome. A typical problem for deBruijn graphs is the reconstruction of repetitive regions because any repetition of length $l<k$ will be collapsed in the same k-mer[61]. Usually a workaround technique is to iterate several steps of

assembly increasing k (limited by the length of the reads) and adding the contigs compute in the previous step to the reads[62].

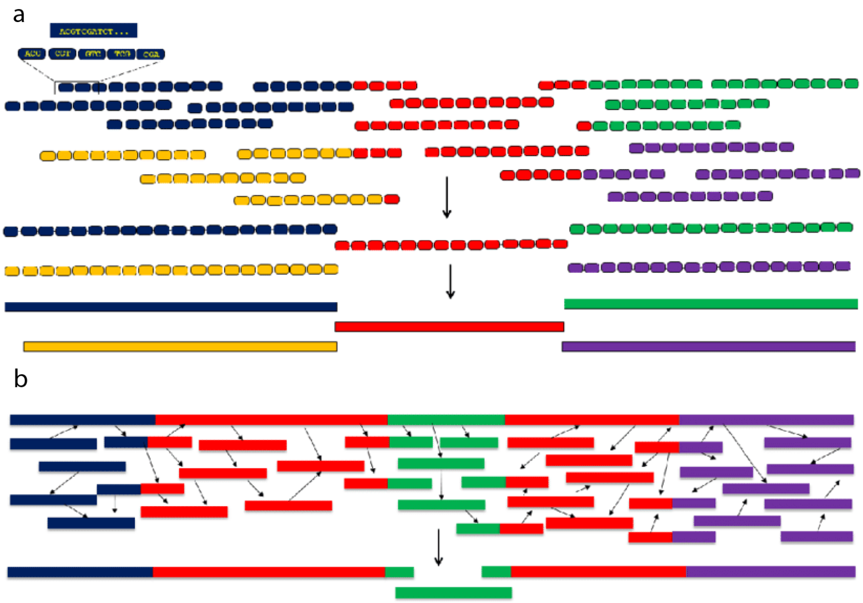


Fig 5. Two types of assembly, from Diniz & Canduri[63]. **a.** In the deBruijn graph the reads are in silico fragmented into k-mers and arranged on a graph structure in which they are the nodes and their k-1 suffixes/prefixes the edges. The deBruijn graph is then traversed to obtain the contigs. **b.** The overlap graph uses the whole reads, aligning them to each other and finding a path through the alignments in order to find the contigs.

Assemblies of metagenomes typically result in highly “hairy” graphs with bifurcating branches and some repeated regions that act as core for several arching contigs. Assemblers implement heuristics to traverse these complex structures and return linear paths (the contigs); however, sometimes it is better to visualize the assembly graph and make human-informed decisions. The standard tool to do so is bandage[64], which also allows plotting additional information on the edges of the graph such as length, coverage, BLAST match results, etc. Manual curation can therefore be performed easily with a graphical representation, for instance to recognize plasmids as circular elements with high coverage compared to the rest of the graph’s elements.

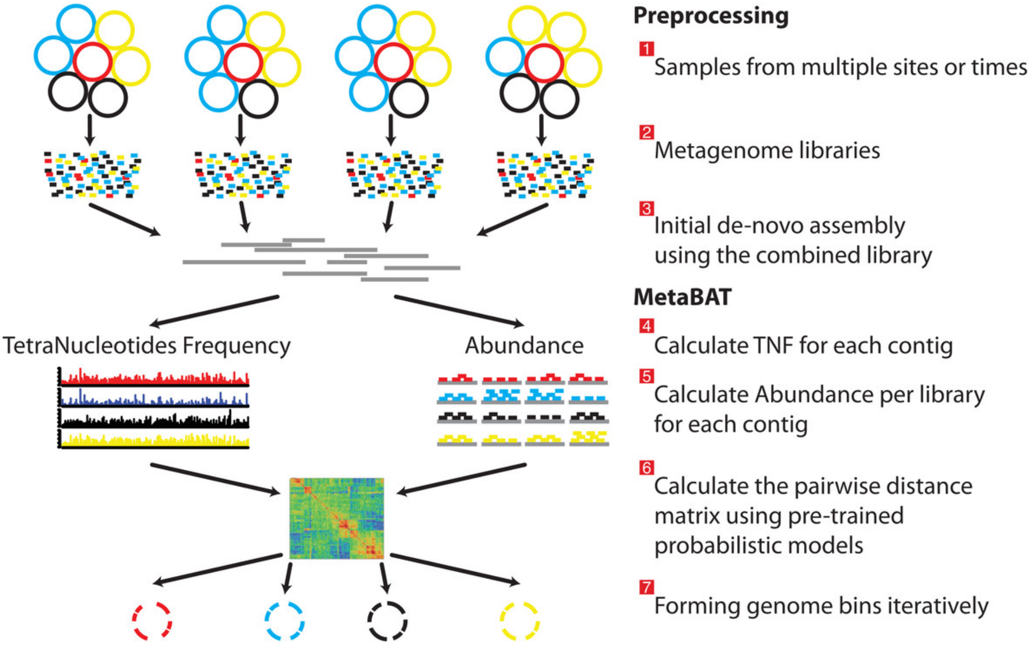

Preprocessing
1 Samples from multiple sites or times
2 Metagenome libraries
3 Initial de-novo assembly using the combined library
MetaBAT
4 Calculate TNF for each contig
5 Calculate Abundance per library for each contig
6 Calculate the pairwise distance matrix using pre-trained probabilistic models
7 Forming genome bins iteratively
TetraNucleotides Frequency
Abundance

Fig 6. An example of binning procedure as performed in MetaBAT from Kang et al.[65]. In the MG pipeline the genomes are sequenced and the resulting reads assembled into contigs. The tetra-nucelotide frequency is computed for each contigs, alongside the abundance per sample (mapping the original MG reads on the contigs). Finally, these data are used to separate the contigs into the bins.

The contigs represent segments of DNA from the microbes in the community, which ultimately we wish to piece together in order to reconstruct their genome(s). Unless the retrieved contigs span the entire length of the genomes in the community, which is unlikely unless using long reads, the contigs must be sorted to create coherent sets of contigs, ideally representative of individual populations, called "bins" (**Fig. 6**). Binning is a general problem in which data must be sorted into similarity buckets (the bins) in an unsupervised manner. For MG binning the contigs features such as k-mer composition, coverage and presence of marker genes are used in the clustering process. The most popular binning tools, such as MaxBin[66], MaxBin2[67] and CONCOCT[68] use the k-mer composition and the coverage across sample, whilst MetaBAT[65] and MetaBAT2[69] use the tetranucleotide (k-mer with k=4) composition, coverage and single-copy-genes. The quality of the bins are assessed according to a set of taxon-specific single copy marker genes by the software CheckM[70] which are used to estimate completeness, heterogeneity (similar concept to strain composition) and contamination (mix of bins). Bins with high quality are used as proxy for individually sequenced genomes and take the name of **Metagenome-Assembled Genomes** (**MAGs**). Additionally, it is possible to use more binning tools and combine the results with DAS Tool[71] to improve the quality of the bins. The recent increase in MG data produced per experiment and the observation that sample variability (when present) results in the most abundant genomes to be reconstructed better, lead to the development of dRep[72]. With dRep the samples are assembled and binned individually and the resulting bins are clustered and one representative per group is selected.

1.3.2 Functional omics quantification

As stated in the introduction to this section, the first concept to deal with during quantification is the set of objects (usually referred as **reference** or **database**) that we want to quantify. The most intuitive idea is to take inspiration from the Central Dogma (section 1.1.1) and consider the gene as the fundamental unity of the quantification inquiry. If the MAGs themselves have not been quantified or if we are interested in population-specific variants and/or unbinned material, the genes predicted from the MG assembly can be quantified using the MG reads. The quantification of the gene potential gives us an idea of the scale of the processes the community can handle. The MT reads can be used with the same gene dataset to quantify the amount of transcript produced from every gene. The procedure of quantifying MG or MT data uses the fundamental action of **aligning** the reads on the reference and find the best match in a process often referred as **mapping**. The reads are mapped using aligner tools, among which the most popular are BWA[73], Bowtie2[74], and the more recent kallisto[75], which is based on **pseudoalingment**. Pseudoalignment works on k-mer matching and is a probabilistic method which allows for considerably speed up of the mapping procedure without losing quality.

The quantification of the MP layer uses the mapping of the MP data on the selected protein database. Each protein within the database is first *in silico* digested into a set of peptides and theoretical fragmentation spectra are generated for each of these, which are subsequently used for matching with the experimentally acquired MS spectra. In the first MP experiment, and many to follow, the protein reference/database was a subset of publicly available repositories[35], curated to fit the expected organisms/environment. Nowadays it is becoming more common to couple the MP with MG or MT in order to create a sample-specific reference/database from the translation of the predicted genes. This approach has showed to identify a greater number of proteins compared to using large public repositories[76,77].

1.3.3 Biological networks

A common structure to put data in for further inquiry is that of a graph (a simple example in **Fig. 7a**). Since we used this concept before without much explanation, we here provide a quick definition. A graph G(V, E) contains two sets. The first, V, contains all the vertices whilst the second, E, contains all the edges between pairs of vertices. The vertices of a graph are the objects of the system we want to describe, e.g. the genes. The edges are the relationships between pairs of objects, e.g. temporal correlations, physical contact, etc. Some baseline assumptions settled in the scientific community about the topology of biological networks include that i) networks should have a certain degree distribution[78] (**Fig. 7b**), ii) the biological role/importance of a node is related to its degree[79] (**Fig. 7c**), and iii) nodes cluster into modules with defined biological meaning[80] (**Fig. 7d**). The use of these assumptions (and derived ones) has proven fruitful to better understand biology and predict its behavior[81,82], even if the first one has been widely discussed and not believed true in most biological systems[83].

One particular type of biological networks, in which V=genes and E=correlations, is a very common framework for biological knowledge inference from omics data: the **co-expression network**. In this network the gene expression (in form of quantified MT or MP) across samples (e.g. over time, space, conditions, etc.) is used to compute pairwise correlation measurements. Therefore, the edges take a value between -1 and 1 (or 0-1 if the absolute value is used instead). The general interpretation is that the gene expression of two genes is similar in shape (correlation measures are scale-invariant) for high values of the edge between them, is not similar at all for values around 0 and it is similar but mirrored the closer it gets to -1. In a co-expression network, the modules represent sets of genes that are expressed at the same time/condition (co-expressed), from which we can hypothesize that they may be involved in the same process or coordinated by the same regulators. One of the most common tools to build and analyze co-expression networks is the R package WGCNA (Weighted correlation network analysis)[84].

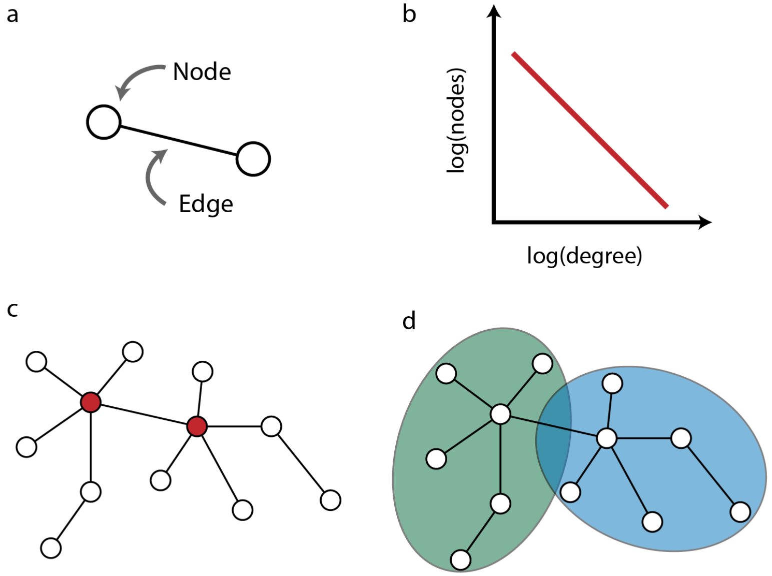


Fig. 7. a. A simple example of network where two nodes are connected by an edge. **b.** A particular case of degree distribution: the scale-free distribution. When the plotted in log-log scale the distribution is represented by a straight line. **c.** A simple network where the nodes with high degree are highlighted in red. **d.** A network can be partitioned in tightly connected regions called modules, here an example where the modules are highlighted in green and blue.

Another particular graph is built using the reactions performed by the enzymes as nodes and the shared metabolites as edges. In this way we have a network we can interrogate to find patterns and topological features of interest in the metabolism under study. Moreover, not only we can use the presence/quantification for the nodes (e.g. via MT or MP), but we can directly use the MB data to validate the edges. For this thesis, we used a metabolic network in **Paper III** to summarize the shared metabolism of a lipid-accumulating community from a wastewater treatment plant.

1.4 Model environments

1.4.1 Biogas reactors

A widely studied process performed by a microbial community that has co-dependent functions is anaerobic digestion. During this process a multitude of different microbes collectively break down organic matter in a series of concatenated metabolic steps which culminates with the release of methane (CH_4) and carbon dioxide (CO_2). The whole conversion can be broken into four major steps (**Fig. 8a**).

1. **Hydrolysis**. The hydrolyzing populations colonize and attack the long-chain polymeric substrate and break it into medium chain molecules or monomers.
2. **Acidogenesis** or **fermentation**. The small molecules from hydrolysis are imported into the cells and fermented into short chain fatty acids (SCFA), alcohols and CO_2 + molecular hydrogen (H_2).
3. **Acetogenesis** or **anaerobic oxidation**. The SCFA produced during acidogenesis are used by specialized syntrophic microbes to be oxidized and produce CO_2 and H_2.
4. **Methanogenesis**. Archaeal populations of the community can take the Acetate and/or H_2/formate and produce CH_4 and CO_4.

Figure 8 illustrates some interesting cases that are explored herein. The conversion between Acetate and CO_2+H_2 (**Fig. 8a**, arrows 4 and 5) is performed by the same pathway, the Wood-Ljungdahl (carbon fixation) Pathway (WLP) which can be used in the reverse direction (oxidizing acetate) if the H_2 pressure is maintained low enough. The reaction is therefore usually coupled with hydrogenotrophic methanogenesis, which is the energy-yielding metabolism of methanogens and is unique to the Archaea (**Fig. 8a**, arrow 6). This is the case in **Paper II** where a syntrophic acetate-oxidizing bacterium performs the reactions on arrows 3 and 4 (**Fig. 8a**) in syntrophy with the community's Archaeon, which performs reaction 6. The composition of the community depends on several factors, such as the main substrate and the temperature, but a certain underlying structure and functional components is always present as addressed in **Paper I**.

Anaerobic fermentation is widely used in industrial process to harness the residual potential of other industries. For example, it is possible to build and maintain biological reactors which feed on substrates such as food waste, wood leftovers and agricultural dispose. These materials, instead of being disposed inefficiently and in a possibly un-sustainable fashion (e.g. burned directly), can be used for an efficient extraction of biochemical energy in the form of alcohol or methane. Methane-producing biogas reactors productivity depends on substrate and condition, but it can reach 88% of biomass conversion rate[85], however pushing this number requires a deeper knowledge of the mechanisms behind single populations' metabolism and their complementarity. In plant-based substrates the largest source of carbon is usually the cellulose, which is also the toughest to break down, and therefore we examined in **Paper I** the enzymatic dynamics over time needed to degrade it.

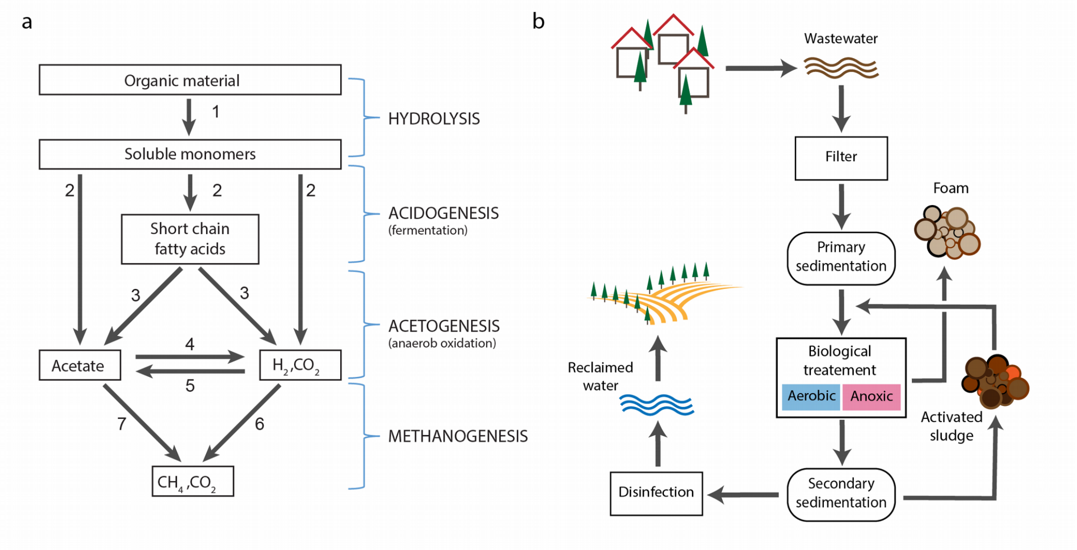


Fig 8. a. Anaerobic digestion diagram adapted from Hagen[86]. Organic material is broken down via (1) hydrolysis into soluble monomers, which in turn are used in (2) fermentation to produce Acetate, Short chain fatty acids (SCFAs) and H_2+CO_2. During acetogenesis the three previously produced compounds are interconverted in three process: (3) syntrophic oxidation of SCFAs, (4) syntrophic acetate oxidation and (5) hydrogen oxidation. Finally, acetate and H_2/formate are converted into CH_4+CO_2 during methanogenesis via (7) acetoclastic methanogenesis or (6) hydrogenotrophic methanogenesis (respectively), depending on the substrate respectively. b**.** Schematic wastewater cycle. Urban areas produce wastewater which is collected, filtered for solid matter and staged in a first sedimentation tank. The wastewater is then treated with the addition of the activated sludge. This step is usually composed of one or more tanks in one or more physicochemical conditions (e.g. O_2 saturation in the water). The foam sampled for **Paper III** is from an

anoxic tank from this step. A second sedimentation tank allows to collect and recycle the sludge. Finally, the water is disinfected and used for new civil purposes.

1.4.2 Wastewater treatment plants

Wastewater produced by human settlements still contain a high energetic chemical potential and a great amount of water, plus it cannot be simply disposed of for health and environmental concerns. Therefore, the solution is the use of a wastewater treatment plant that can separate the useful compounds, neutralize the potential harmful sources and return the clean water to the local system for other use. Depending on the amount of wastewater to be treated there are guidelines for the implant to be in place. In case of small to medium amounts, the biological process is a cheap and feasible option.

In a typical biological wastewater treatment plant (BWWTP), solid residual collected from the water before further processing. The main steps take place in large tanks where the water is subject to the work of bacterial communities in various oxygenation condition in order to perform changes in the profile of the chemicals in the water, such as nitrogen or phosphorus. The microbes have to coordinate their metabolism to perform multi-step process such as nitrification and denitrification. Moreover, the treatment needs may change depending on the starting conditions of the water, its source and destination, requiring a mix of physical, chemical and biological treatments to be reclaimed. The presence of lipids in the wastewater enables the growth of oleaginous mixed microbial communities (OMMCs), which accumulate on the surface of the tank resulting in a thick foam. The OMMC/foam moves the lipids from the water in a more accessible form, which may be exploited for biofuel production. On the other hand, the foam disturbs processes such aeration and moving the water throughout pipes, decreasing the efficiency of the BWWTP, therefore the study of the foam and its community becomes of great importance as we addressed in **Paper III**.

2 OUTLINE AND AIM OF THE THESIS

Microbial communities are often enigmatic entities. We can observe their general phenotype and manipulate them to identify their optima and boundaries (e.g. changing substrates or physico-chemical parameters), but their inner mechanisms remain mostly elusive to direct probing. This happens mostly because of the inability of many microbes to live outside their environment, i.e. without their microbial partners, and in most of the cases we have not yet elucidated what they require to stay alive (and functioning). To assist in these challenges, meta-omic approaches allow us to bypass cultivability and collect, characterize and quantify the main molecules that constitute a microbial community. We are now equipped with flexible tools to unlock an unprecedented amount of information; however, this creates new challenges such as how do we sort this data, and more importantly, what can we learn from them?

In **Paper I** we wanted to solve the duplicitous puzzle about bacterial complementarity in (hemi)cellulose degradation and the inexplicable ubiquity of *Coprothermobacter protelyticus* in biogas reactors. We therefore used the gene potential reconstructed from the metagenomes and isolated strain genomes to build a hypothesis on metabolic complementarity of the main populations in the community and the acquisition of hemicellulolytic enzymes by *C. proteolyticus*. Consequently, we sought to corroborate our hypothesis *in vivo*, using temporal metatranscriptomics paired with monosaccharide measurements. Finally, we proved that the newly acquired genes of *C.proteolyticus* are biochemically active *in vitro* using an enzymatic assay.

Microbial ecology based on multi-omics misses one big thing: absolute quantification. We therefore adapted our high-throughput approaches to obtain quantitative measurements of RNA and proteins in **Paper II**. In doing so, we showed the first Archaeal protein-to-RNA ratio, which matches Eukaryotic representatives in the literature. We then used the linearity between transcriptome and proteome to identify phenotypic complementarity and corroborated it with traditional pathway analysis. Hence, we propose that a fundamental biological feature such as the transcriptome-proteome linearity, can be used to highlight a different biological feature such as metabolism. Thanks to the absolute quantification we also aimed to estimate the impact of post-transcriptional regulation of the protein levels to identify the targeted functions.

The knowledge accumulated in the previous works, and the ones not included in this thesis, led to the benchmarking of a real-world complex microbiome. In **Paper III** we used metagenomes and metatranscriptomes that were generated over a one-year period from weekly sampling of a lipid accumulating community from a wastewater treatment plant. The fundamental aim was to understand community-wide taxonomy and function and interactions between constituent populations. Therefore, we used the eigengenes to reduce the complexity of our datasets, so that we could evaluate their time patterns and link them with the physico-chemical parameters of the system. Moreover, we analyzed two functions (lipid accumulation and nitrogen metabolism) comparing the gene expression over time and the taxonomic richness to discover competition for the substrate (lipid) and a keystone gene (ammonia oxidizing monooxygenase).

3 MAIN RESULTS AND DISCUSSION

Previous analysis on the Frevar biogas reactor (Fredrikstad, Norway) depicted the resident microbial community as dominated by heterogeneous strains of *C. proteolyticus*[87]. Samples taken from the original Frevar's reactor were further enriched for cellulose-degrading populations by serial dilution, resulting in SEM1b, which formed the community used for MG analysis in **Paper I** and **II**. The preliminary 16S rRNA amplicon analysis on SEM1b showed seven main populations, with strain variability (**Paper I**, Fig. 3). The consortium appeared co-dominated by *Clostridium (Ruminiclostridium) thermocellum* and strains of *C. proteolyticus*. Two individual strains of *C. proteolyticus* (BWF2A and SW3C) were isolated and sequenced independently and the MG assembly was helped by the subtraction of the reads mapping on the two strains. The co-assembly of two SEM1b MG samples produced 20,760 contigs (total 27 Mbp, longest 603 Kbp) which ultimately resulted in 11 MAGs, with taxonomic profile and abundances similar to the ones observed with 16S rRNA (**Paper I**, Fig. 2). The MAGs were analyzed using average nucleotide identity and Blastp, whereas the MAGs not complete enough were not taxonomically assigned. MAGs COPR1-3 were affiliated with *C. proteolyticus* (alongside the isolates BWF2A and SW3C), RCLO1 with *C. thermocellum*, CLOS1 matched a MAG from Frevar and *Clostridium stercorarium*. Moreover, TEPI1-2 were affiliated to *Tepidanaerobacter*, SYNG1-2 to *Synergistales*, TISS1 to *Tissierellales* and METH1 to *Methanothermobacter thermoautotrophicus*. RCLO1 and CLOS1, both associated with lineages known to degrade polysaccharides, encoded respectively 297 and 139 carbohydrate active enzymes (CAZymes), many of which were annotated as Glycosyl Hydrolases (GHs).

The two *C. proteolyticus* strains presented some genomic differences from the Type strain DSM 5265 obtained from the literature (**Paper I**, Fig. 3). Notably the species was reported as non-motile but DSM 5265 contained a set of flagellar genes, which were not present in the strains from SEM1b. In addition, BWF2A and SW3C showed acquisition of a CAZymes' cassette including GH16, GH3, and GH18-CBM35 (region-A, **Paper** I, Fig. 3). GH16 is an endo-β-1,3-1,4-glucanase and GH3 a β-glucosidase, suggesting the ability to degrade the hemicellulose beta-glucan, whilst GH18 encodes an endo-β-N-acetylglucosaminidase, conferring the ability to degrade the bacterial cell wall. Region-A had high similarity to a homologous region from Firmicutes (*Thermoanaerobacter*, *Clostridium cellulolyticum*, and *C. thermocellum*) and Thermotogae (*Thermosipho africanus*, *Fervidobacterium nodosum*, and *F. gondwanense*) as summarized in **Paper I** Fig. 4. The genomic sequence flanking region-A contains an incomplete prophage composed of a phage lysis holin and two downstream recombinases (**Paper I**, Fig. 3-4), which suggests phage-mediated horizontal gene transfer. Moreover, we proved biochemically the endoglucanase activity of the newly acquired GH16 on β-1,3 (pachyman, curdlan, laminarin) and β-1,3-1,4 (Barley) substrates (**Paper I**, Fig. S2A). We further showed that the activity of the GH16 showed a high production of glucose (**Paper I**, Fig. S2B), on which the *C. proteolyticus* strains can directly metabolize.

In order to understand the concerted work of SEM1b in degrading our (hemi)cellulose substrate, we analyzed the temporal 16S rRNA gene abundance and functional gene expression data from the substrate inoculum (T0) up to 42 hours of SEM1b growth, with samples being taken at 5-8 hour intervals. The 16S rRNA gene data showed a dynamic dominance of the community shared by *Coprothermobacter*- and *Clostridium*-affiliated populations (**Paper I**, Fig. 5a). For our metatranscriptome analysis of SEM1b, two or more CAZyme-annotated ORFs were grouped in "expression groups" if all the MT reads mapping on them were shared hits, therefore making explicit the taxonomic resolution limit of our dataset. This resulted in 274 singleton and 8 multi-ORF CAZymes. In case of close taxonomic relationships, such as for the *C. proteolyticus* strains, it was not possible to discern the exact origin of the expressed ORF (e.g. the GHs in the A-region).

The total gene expression data for the SEM1b CAZymes revealed six clusters (I-VI) and characteristic MAG/genome enrichments (**Paper I**, Fig. 5b). Clusters III and IV, accounting for 10 and 11 expression groups, had similar time patterns (Fig. 5c), with an initial increase (T2–3) followed by a later one (T6–8). Cluster III (alongside II and IV) was enriched in *C. proteolyticus*-affiliated populations. However, cluster II, containing 10 expressions groups, even if similar to III and IV, had a sharper upward trend in T2. These clusters contained N-acetylglucosamine- (CE9) and peptidoglycan-targeting (CE4, GH23, and GH73) CAZymes, both components of the bacterial cell wall. Therefore, we hypothesize they pertain to the recycling of the cell wall from dead cells in lag and late-stationary/death phases. After 13

hours (T2), the dominant phenotype of the community shifted toward cellulose degradation, which can be seen with the inversion in 16S rRNA gene profiles between *Clostridium*- and *C. proteolyticus*-associated populations, and returning toward the initial configuration over time. This trend pairs with the increased levels in clusters II, III and IV, associated with cell wall recycling by *C. protelyticus* (**Paper I**, Fig. 5b).
The two main (hemi)cellulolytic-associated clusters, V and VI (28 and 101 expression groups), contained CAZymes active on cellulose (e.g., GH5, GH9, GH44, GH48, CBM3) and hemicellulose (e.g., GH10, GH11, GH26, GH43, GH74) and were enriched in RCLO1 and CLOS1 (Table S5). The increase in exponential phase matches the surge in *Clostridium*-associated populations in the 16S rRNA gene analysis and the shift in community substrate.

With 121 expression groups, Cluster I is the largest in the analysis and contains ORFs for both hemicellulose degradation (e.g., GH3, GH10, GH29, GH31, GH43, and GH130) and carbohydrate deacetylation (e.g., CE4, CE7, CE8, CE9, CE12, and CE15) (**Paper I**, Table S5). Interestingly the newly acquired CAZymes GH16 and GH3 from the *C. protelyticus* strains were contained here, indicating that they were expressed with the same temporal pattern of their homologous genes. Moreover, clusters V and VI (**Paper I**, Fig. 5) preceded cluster I, indicating that the cellulolytic action was required to liberate embedded hemicellulose fibres from the substrate, before the hemicellulases could act upon it. As expected, we detected xylose (one of the products of hemicellulose hydrolysis) to increase from T5 to T7 (**Paper I**, Fig. 5a), indicated hemicellulose degradation. We can therefore summarize the life-cycle of SEM1b after substrate inoculum as: 1) lag phase and recycling of dead cells, 2) exponential phase and cellulose hydrolysis, which also liberates hemicellulose fibres 3) shortly after the previous step starts the hemicellulose hydrolysis and 4) the community goes back to recycling dead cells.

In **Paper II** we used the already reconstructed SEM1b community to characterize RNA/protein dynamics in a microbiome setting. Here we generalized the concept of "expression group" as a set of ORFs that are indistinguishable in MT and MP data, calling these sets ORF-groups (ORFGs), where a singleton ORFG is defined as a group with a single ORF, and thus a single gene. In this context, we obtained MT and MP data that identified 12552 MT- (96% singleton) and 3235 MP- (78% singletons) highly transcribed and translated ORFGs (respectively). Most of the ORFGs that contained multiple homologous ORFs were originating from strains of a single species. For instance, in the MT, 444 non-singleton ORFGs (88% of the total) contained ORFs from different strains of the same species, whilst this was the case for 294 ORFGs (32%) in the MP. Kegg Ontology (KO) codes were found for 19070 (49%) of the ORFs from SEM1b. The most abundant annotations included Membrane transport, Carbohydrate metabolism, Translation, Amino acid metabolism and Replication and repair (**Paper II**, Supplementary Fig. 1). These functional categories were also among the top five most abundant for the MT, and top six in the MP (plus Energy metabolism) The Membrane transport was poorly represented in the MP (2% of the terms), which reflects technical issues commonly encountered with transmembrane protein extraction. The abundance ranking of functional categories was more conserved between MG and MT (Kendall τ: 0.77, p<10-8) and MT and MP (τ 0.74, p<10-6) than between MG and MP (τ 0.68, p<10-5). This suggested that a more variable gene arsenal is present in the genomes than that expressed in the transcriptomes and the proteomes, which is the less variegate of the three. Collectively, this hinted to post-transcriptional regulation playing an important role in addition to transcriptional regulation in prokaryotes.

More importantly in **Paper II** we wanted to assess if microbial RNA/protein dynamics vary between ecological status (isolate vs community), metabolic states and/or taxonomic phylogeny. We therefore quantified and resolved the numbers of transcript and protein molecules per sample in our SEM1b community, which averaged 3.8×10^{12} (sd 3.0×10^{12}) and 2.2×10^{15}(sd 9.5×10^{14}), respectively. SEM1b approximated the exponential growth phase in t3 (18 hours), thus we used the protein-to-RNA ratio from this time point for comparison against estimates from the literature. The replicate-averaged protein-to-RNA ratio for the bacteria in SEM1b ranges from ~10^2 to 10^4 (median = 949, **Paper II** Fig. 1a), agreeing with the previously reported range for a pure culture of *Escherichia coli*[88]. We found a population-specific variation in the bacterial protein-to-RNA ratio (**Paper II**, Fig. 1a), with the median ratios at 18h ranging from 658 in CLOS1 to 1137 in RCLO1. In contrast to bacteria, the protein-to-RNA ratio for an Archaeal organism, which we report for the first time, was approximately 10x higher at 12035 protein molecules per detected RNA (**Paper II**, Fig. 1a: METH1). The values from literature for Eukaryotes are 4200-5600 in yeast and

2800-9800 in Homo sapiens; hence, bringing Archaeal and Eukaryotic translation dynamics in closer alignment.

Building on these initial observations, we modeled the relationship between proteome and transcriptome using a monomial function (**Paper II**, Eq. 1), which for our log10-transformed RNA and protein data can be fitted using a linear model. For the modeling we used the reconstructed MAGs with the highest quality (RCLO1, CLOS1, COPR1, TISS1, TEPI1, TEPI2 and METH1) (**Paper II**, Fig. 1d). The linearity parameter k can be interpreted as the rate of which a change in RNA level is reflected in a change in the corresponding protein level. With the exception of TEPI2, the linearity (k) between protein and RNA levels was observed to start at values between 0.6 and 0.8 at 13 hours (t2) (**Paper II**, Fig. 1d). The evolution of the MAGs' k values over time is then divided in three groups: one which is losing linearity rapidly (TISS1 and COPR1); one which is slowly declining (RCLO1, CLOS1 and METH1) and one which is staying constant if not increasing (TEPI1 and TEPI2) (**Paper II**, Fig. 1d). Notably CLOS1, METH1 and TEPI1 are converging towards the same linearity values, while RCLO1 has a parallel trend to them. If these trends can be used to retro-fit the steady state definition, we can hypothesize that these four populations possess a metabolic equilibrium and that this equilibrium is approximately reached within the 10 hour window between 33h and 43h (t6 and t7 respectively, **Paper II**, Fig. 1d).

To validate if the changing k-values could be extrapolated to greater interpretations of metabolic convergence and interlocking, we proceeded to investigate SEM1b with a more traditional pathway-guided analysis. We used the KO annotation of the ORFs to explore the metabolic modules' completeness for SEM1b MAGs (**Paper II**, Fig. 2) and reconstructed their temporal expression patterns (**Paper II**, Fig. 3). As previously shown in **Paper I**, SEM1b is able to convert (hemi)cellulose to methane via the combined metabolism of its seven major constituent populations (**Paper II**, Fig. 3a). Based on previous analysis that showed that RCLO1 is closely related to *R. thermocellum*, we predict that it senses its growth substrate (cellulose) and moves towards it (**Paper II**, Fig. 3d). RCLOS1 then invests in the production of the cellulosomal components, such as scaffoldins, dockerins and CAZymes, which assemble into a dynamic multi-proteins complex that degrades the substrate to smaller carbohydrates. Via the MG, we predicted that non-cellulosomal CAZymes were also employed by the *Clostridium*-affiliated CLOS1, which acted upon the hemicellulose fraction (mainly xylan) trapped in the spruce cellulose, which was supported by observed release of its main monomer xylose (**Paper II**, Fig. 3a). Sugars generated via the actions of RCLO1 and CLOS1 are subsequently consumed by RCLO1, CLOS1 and *Coprothermobacter*-affiliated populations (COPR1, BWF2A and SW3C), which were all observed to express sugar transporters (**Paper II**, Fig. 2). Interestingly, BWF2A and SW3C possess and express unique sugar transporters, likely gaining access to an undisputed pool of arabinogalactan or maltooligosaccharide. The transporter for pentamers ribose/xylose were the most common and possessed by RCLO1, *C. proteolyticus* populations and *Tepidanaerobacter*-affiliated populations (TEPI1 and TEPI2). Moreover from Fig. 2 in **Paper II**, it is clear that the proteins from the transporters are almost never found in the samples, even if the respective RNAs are abundant. This is likely due to the technical difficulties in extracting transmembrane proteins.

SEM1b activity of cellulose degradation lead to the formation short chain fatty acids (SCFAs) which are subsequently metabolized by the SCFA-oxidizing population TEPI1 (**Paper II**, Fig. 3a), which demonstrated a good linearity between protein and RNA levels that increased over time (**Paper II**, Fig. 1d). TEPI1 was also found to encode a complete Wood-Ljungdahl carbon fixation Pathway (WLP) that was detectable in both MT and MP (**Paper II**, Fig. 2). Interestingly the closely related MAG TEPI2 was observed to lack the WLP and to express ~10 times more transcripts for the ribose/xylose transporter than TEPI1; relegating it to the role of mere sugar degrader, and probably scavenger in the community. The TEPI1 MAG expresses the NAD+ (NADP+)-reducing hydrogenases complex, which reduces hydrogen ions to H_2 using NAD(P)H as the electron donor. The molecular hydrogen generated here would then be used by the syntrophic partner METH1 to form methane (**Paper II**, Fig. 3a). However, this reverse WLP-mediated acetate oxidation is thermodynamically unfavourable unless coupled with syntrophic hydrogenotrophs. Within SEM1b, the METH1 population is a hydrogenotrophic methanogen, thus we hypothesize that the molecular hydrogen generated by TEPI1 would then be used by the syntrophic partner METH1 to form methane (**Paper II**, Fig. 3a). Overall, our more classical pathway-wise exploration of the SEM1b populations supported that RCLO1, CLOS1, TEPI1 and

METH1 indeed share functional co-dependencies and supported our predictions via protein-RNA dynamics that they converge upon a dominant metabolic state.

The last part of **Paper II** explored the poorly understood aspect of microbiome protein-level regulation. The absolute quantification of transcripts and proteins in SEM1b were used to estimate the translation and protein degradation rates using PECA-R. The analysis found 305 significant changes in translation rate, accounting for 302 ORFs. Of the rate changes', 94% were downregulated and 71% of the ORF were functionally annotated. Among the main results. RCLO1 was found to downregulate 28 ORFs between 13h and 18h (t2-t3), mostly from complexes involved in chemotaxis, flagellum assembly and shape determination. Whilst in the following five hours, several systems concerning carbon fixation were affected. In the next five hours, RCLOS1 downregulates the translation of the cell division protein ZapA as well. The reduction protein production for chemotaxis, mobility and then cell division matches the idea that within 13h of the inoculation, RCLO1 sensed, reached and colonized the cellulose fibers. Contextually the release of medium length carbohydrates enables RCLO1 to engage in the more energetically favorable fermentation metabolism. TEPI1 downregulated 60 ORFs between 13h-18h, accounting for part of its carbohydrate metabolism, the amino acid transporters and the NADH dehydrogenase complex (HND). TEPI2 has 19 ORFs subject to downregulation in the 13h-18h interval, including carbohydrate metabolism and related transporters. In the last interval (33h-38h), RCLO1 upregulated the translation of 10 ORFs, including flagellar protein and shape determination; seemingly starting to restore the functions downregulated in the 13h-18h interval.

In **Paper III** we took what we had learned in the previous works and scaled it up to a real-world community (Schif-LAO) from a wastewater treatment plant from Schifflange (Luxembourg). We used the weekly sampling between 2011-03-21 and 2012-05-03 (51 samples) and combined the sample-wide MG analyses to produce a total of ~19.8×10^6 different ORFs (extended dataset). A KEGG Orthology group was assigned to 40.4% of the ORFs in the set, whilst taxonomic affiliations were designated to 38.5%. We quantified the gene count and their expression over time for the extended dataset using the MG and MT reads. The vast majority of the genes however were not found to be expressed over the entire dataset and were only detected in few samples alone, with as many as 16.8×10^6 in only one, hinting that the community relies on high gene redundancy. Subsequently we generated a more approachable dataset retaining only ORFs with a gene count or gene expression above 1 transcript per million (TPM) in at least one sample, obtaining 0.7×10^6 and 0.8×10^6 ORFs for the MG and MT respectively (core dataset).

In order to understand the temporal patterns underlying the core dataset, we reconstructed six (EG1-6) eigengenes associated with the MG and six (EG1-5, EG8) with the MT data (**Paper III**, Fig1a-b). The two sets of EGs however were highly correlated (**Paper III**, Fig 1c), and intuitively the curves described by the MG are smoother than by the MT data. Furthermore, we wanted to link these patterns with the environmental physio-chemical parameters, which we filtered for collinearity (the linear dependence of variables) to seven of them: conductivity, dry matter, ammonium (NH_4), nitrate (NO_3), oxygen, pH and temperature. The results show how the most relevant environmental factors are temperature and NO_3, linked with five EGs each, followed by dry matter with four EGs (**Paper III**, Fig 1d). Ammonium and pH were significant in explaining two EGs, whilst conductivity and oxygen contribute significantly to one EG each. One of the main processes happening in WWTPs is the conversion of Ammonium into Nitrate ($NH_4 \rightarrow NH_3 \rightarrow NO_2 \rightarrow NO_3$), therefore it is hard to establish the causal direction of the link between these two compounds and gene copy number/expression. The more intuitive causality is between temperature and the EGs, especially for EG2 (from MG and MT). Indeed, we fitted both the MG derived EG2 and the temperature with the sine function using a period (T) of 365 (days), giving perfect fits with F-statistics of 181 and 269 (p-values $< 10^{15}$), same phase and amplitude of opposite sign. These results point to a seasonal composition and behavior of the microbial community, with a set of genes whose presence in the Schif-LAO consortium depends on the temperature and is supposed to reach the same values at yearly intervals.

To reduce the complexity of our massive temporal omic datasets so that we could study the functional characteristic of Schif-LAO, we built a reaction network using the KO annotation from the extended dataset, in which every node (collapsed KO, hence CKO) represented a set of all the reactions using the same metabolites, and the edges the metabolites shared by these sets. We obtained a reaction network of 1,984 nodes and 13,350 edges. The number of

ORFs per CKO varied greatly, with a maximum of 77,474 and a median of 284. We speculated that the taxa contributing to a given function in the Schif-LAO community at any given point in the sampling time may change, thus we sought to taxonomically identify those ORFs that are crucial in the carrying of their function. To do this, we computed the normalized information entropy of the MT for every CKO at the Family level. Subsequently, we focused the analysis on the fatty acid biosynthesis, which is believed to be important in a LAO community, and nitrogen metabolism, which consequently is relevant in all wastewater treatment plants.

In Schif-LAO there are 17 CKOs associated with the KO term "Fatty Acid Biosynthesis", connected by 26 metabolites (edges) (**Paper III**, Fig. 2a), accounting for fatty acid initiation, elongation and termination. The most expressed reaction node is CKO1295, ranging between 695.3-1513.5 transcripts per million (TPM) and a median of 927.8 TPM per time point (**Paper III**, Fig. 2b). CKO1295 embeds the two opposite reactions that attach and detach the cofactor A (CoA) to the fatty acid chain. Interestingly the richness in taxa contributing to the node is inversely proportional to the gene expression (Spearman's ρ of -0.35, p=0.01) with a quadratic trend (**Paper III**, Fig. 2c). The second largest expressed node is CKO120 with a range of 539.6-1147.4 TPM and a median equal to 778.3 TPM per time point. Similar to the previous case, for CKO120 the gene expression is inversely proportional to the taxa richness (ρ=-0.38, p<0.01) but with a linear trend (**Paper III**, Fig. 2d). CKO1295 and CKO120 cover two fundamental aspects of FAS: activation/deactivation of the fatty acid and its extension; however, our data would suggest that different taxa enact a competitive takeover of these functions. The families detected at high MG abundance did not necessarily correspond to high MT activity, for example the widely abundant Comamonadaceae (24.8%) was observed to exert lower expression (13.2%) than Leptospiraceae (32%). In CKO120 Leptospiraceae (28.3%) is again the most active, whilst the most abundant (based on MG analysis) is the family Microthrixaceae (24.5%).

The Nitrogen-related metabolism of Schif-LAO includes 21 reaction nodes and 71 metabolic edges (**Paper III**, Fig. 2e). The entropy analysis points to CKO3145 and CKO3079 as potential keystone functions in the system (**Paper III**, Fig. 2f), having high expression and low taxonomical diversity. The first reaction node is overwhelmingly dominated by the ORFs from the family Nitrosomonadaceae (MG 97.8%, MT 99.1%) and contains the *amo* gene subunits A-B. The second node is dominated again by transcripts from Nitrosomonadaceae (MG 63.9%, MT 91%) and encodes the hydroxylamine dehydrogenases. Given the crucial importance of the presence of the gene *amo* in the environment to start the assimilation of ammonia, the main family producing transcripts from it, Nitrosomonadaceae, must be held carefully tuned to the optimal size to optimize the performance of Schif-LAO.

4 CONCLUDING REMARKS AND PERSPECTIVES

As the meta-omics field matures, new technologies are being constantly introduced creating larger and more resolute datasets, however we still have not reached any standardized way to study microbial communities and every study is a unique piece of research. At the same time, it is particularly challenging to work outside of the boundaries of model organisms to try and chart new microbial interactions when the microbes involved are uncharacterized. In our case we first covered a simplistic community with the greatest resolution possible (**Paper I**, **II**), retrieving a strain-level MG that included isolated genomes. In general, the combination of MG sequencing and isolation is a promising way to increase the resolution of the datasets, alongside the newly emerging strategies such as binning + dereplication of the MAGs. However, as we pointed out in **Paper I**, an increase in MG resolution, being virtually able to identify many strain-level populations' genomes (e.g. through third generation of sequencing), does not equate to being able to tell which are the strains contributing to a given function if their genes are highly similar. This phenomenon is tightly linked to phylogenetic affiliations, indeed, as addressed in **Paper II**, the distinction between gene products (transcript and protein) is inversely proportional to the phylogenetic distance of the microbes they come from. When adding the layers of annotations on the ORFs (KO, taxonomy, etc…), every comparison becomes challenging and a statistical and computation framework to address the task should be developed. In spite of its absence, we integrated and adapted methods developed for individual omics and eukaryotic data to render the highest resolution in molecular characterization and quantification.

In **paper II** in particular we addressed the need of absolute quantification in the meta-omics, showing that it is both achievable without increasing excessively the experimental work (nowadays commercial kits for RNA spike-in are widely available) and allows to answer questions concerning molecular level regulation and, more widely, making the samples (and different experiments) comparable. We introduced the novel idea of using the population-wide relationship (i.e. linearity) of the transcriptome and proteome as a proxy for the population activity. Moreover, when the relationships are compared among populations from the same community, the study of their trends (e.g. convergent, parallel, etc.) can identify metabolically intertwined microbes. This, of course, is still a hypothesis and it should be tested in bigger and more varied communities.

Another important aspect of this thesis is the use of bioinformaitcs and data analysis as a hypothesis generators to be used in other branches of biology. In **paper I** for instance we coupled the prediction from the MG with bacterial culturing and the hypothesis from the MT with enzymology. In **Paper III** we developed the reaction network with the precise intent to find keystone populations and genes for further investigation and potential exploitation. Most importantly, we learnt that to achieve a full understanding of microbial ecology, we need to integrate all the meta-omics layers quantified with absolute measurements and biochemical data from both the microbes and the environment. Moving forward, big technical and computational hurdles must be overcome to handle, interpret and visualize the massive amounts of generated data that will come from studying real-world communities at this proposed scale and resolution. However, we believe that it must be achieved if we are to truly and holistically appreciate naturally-occurring microbiomes at a fundamental level.

BIBLIOGRAPHY

1. Martiny, J. B. H. *et al.* Microbial biogeography: putting microorganisms on the map. *Nat. Rev. Microbiol.* **4**, 102–112 (2006).

2. Bar-On, Y. M., Phillips, R. & Milo, R. The biomass distribution on Earth. *Proc. Natl. Acad. Sci.* **115**, 6506–6511 (2018).

3. Hodgskiss, M. S. W., Crockford, P. W., Peng, Y., Wing, B. A. & Horner, T. J. A productivity collapse to end Earth's Great Oxidation. *Proc. Natl. Acad. Sci.* **116**, 17207–17212 (2019).

4. Gougoulias, C., Clark, J. M. & Shaw, L. J. The role of soil microbes in the global carbon cycle: tracking the below-ground microbial processing of plant-derived carbon for manipulating carbon dynamics in agricultural systems. *J. Sci. Food Agric.* **94**, 2362–2371 (2014).

5. Nelson, M. B., Martiny, A. C. & Martiny, J. B. H. Global biogeography of microbial nitrogen-cycling traits in soil. *Proc. Natl. Acad. Sci.* **113**, 8033–8040 (2016).

6. Vorholt, J. A. Microbial life in the phyllosphere. *Nat. Rev. Microbiol.* **10**, 828–840 (2012).

7. Groussin, M. *et al.* Unraveling the processes shaping mammalian gut microbiomes over evolutionary time. *Nat. Commun.* **8**, 14319 (2017).

8. CRICK, F. Central Dogma of Molecular Biology. *Nature* **227**, 561–563 (1970).

9. CRICK, F. H. On protein synthesis. *Symp. Soc. Exp. Biol.* **12**, 138–63 (1958).

10. Mizutani, S. & Temin, H. M. An RNA-Dependent DNA Polymerase in Virions of Rous Sarcoma Virus. *Cold Spring Harb. Symp. Quant. Biol.* **35**, 847–849 (1970).

11. McCarthy, B. J. & Holland, J. J. Denatured DNA as a direct template for in vitro protein synthesis. *Proc. Natl. Acad. Sci.* **54**, 880–886 (1965).

12. BRETSCHER, M. S. Direct Translation of a Circular Messenger DNA. *Nature* **220**, 1088–1091 (1968).

13. Ahlquist, P. RNA-Dependent RNA Polymerases, Viruses, and RNA Silencing. *Science (80-.).* **296**, 1270–1273 (2002).

14. Pearson, H. What is a gene? *Nature* **441**, 398–401 (2006).

15. Laalami, S., Zig, L. & Putzer, H. Initiation of mRNA decay in bacteria. *Cell. Mol. Life Sci.* **71**, 1799–1828 (2014).

16. Suzanne Clancy. RNA Transcription by RNA Polymerase: Prokaryotes vs Eukaryotes. *Nature Education* (2008).

17. Losick, R. & Stragier, P. Crisscross regulation of cell-type-specific gene expression during development in B. subtilis. *Nature* **355**, 601–604 (1992).

18. Browning, D. F. & Busby, S. J. W. Local and global regulation of transcription initiation in bacteria. *Nat. Rev. Microbiol.* **14**, 638–650 (2016).

19. Cai, L., Friedman, N. & Xie, X. S. Stochastic protein expression in individual cells at the single molecule level. *Nature* **440**, 358–362 (2006).

20. Barandun, J., Delley, C. L. & Weber-Ban, E. The pupylation pathway and its role in mycobacteria. *BMC Biol.* **10**, 95 (2012).

21. Kramer, G. *et al.* Proteome-wide Alterations in Escherichia coli Translation Rates upon Anaerobiosis. *Mol. Cell. Proteomics* **9**, 2508–2516 (2010).

22. Novick, A. & Szilard, L. Description of the Chemostat. *Science (80-.).* **112**, 715–716 (1950).

23. Rosenthal, A. Z. *et al.* Metabolic interactions between dynamic bacterial subpopulations. *Elife* **7**, (2018).

24. Faust, K. & Raes, J. Microbial interactions: from networks to models. *Nat. Rev. Microbiol.* **10**, 538–550 (2012).

25. Bier, R. L. *et al.* Linking microbial community structure and microbial processes: an empirical and conceptual overview. *FEMS Microbiol. Ecol.* **91**, fiv113 (2015).

26. Maruvada, P., Leone, V., Kaplan, L. M. & Chang, E. B. The Human Microbiome and Obesity: Moving beyond Associations. *Cell Host Microbe* **22**, 589–599 (2017).

27. Vatanen, T. *et al.* The human gut microbiome in early-onset type 1 diabetes from the TEDDY study. *Nature* **562**, 589–594 (2018).

28. Stevens, B. R. *et al.* Depression phenotype identified by using single nucleotide exact amplicon sequence variants of the human gut microbiome. *Mol. Psychiatry* (2020) doi:10.1038/s41380-020-0652-5.

29. Severance, E. G. & Yolken, R. H. Deciphering microbiome and neuroactive immune gene interactions in schizophrenia. *Neurobiol. Dis.* **135**, 104331 (2020).

30. Staley, J. T. & Konopka, A. Measurement of in Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats. *Annu. Rev. Microbiol.* **39**, 321–346 (1985).

31. Mullis, K. Process for amplifying, detecting, and/or cloning nucleic acid sequences. *Biotechnol. Adv.* **5**, 313 (1987).

32. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).

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

34. Poretsky, R. S. *et al.* Analysis of Microbial Gene Transcripts in Environmental Samples. *Appl. Environ. Microbiol.* **71**, 4121–4126 (2005).

35. Wilmes, P. & Bond, P. L. The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. *Environ. Microbiol.* **6**, 911–920 (2004).

36. Villas-Bôas, S. G., Mas, S., Åkesson, M., Smedsgaard, J. & Nielsen, J. Mass spectrometry in metabolome analysis. *Mass Spectrom. Rev.* **24**, 613–646 (2005).

37. Kunath, B. J., Bremges, A., Weimann, A., McHardy, A. C. & Pope, P. B. Metagenomics and CAZyme Discovery. in *Methods in Molecular Biology* 255–277 (2017). doi:10.1007/978-1-4939-6899-2_20.

38. Kunath, B. J. *et al.* Metaproteomics: Sample Preparation and Methodological Considerations. in *Advances in Experimental Medicine and Biology* 187–215 (2019). doi:10.1007/978-3-030-12298-0_8.

39. Roume, H. *et al.* A biomolecular isolation framework for eco-systems biology. *ISME J.* **7**, 110–121 (2013).

40. Vanwonterghem, I., Jensen, P. D., Ho, D. P., Batstone, D. J. & Tyson, G. W. Linking microbial community structure, interactions and function in anaerobic digesters using new molecular techniques. *Curr. Opin. Biotechnol.* **27**, 55–64 (2014).

41. Greathouse, K. L., Sinha, R. & Vogtmann, E. DNA extraction for human microbiome studies: the issue of standardization. *Genome Biol.* **20**, 212 (2019).

42. Rodriguez-R, L. M. & Konstantinidis, K. T. Estimating coverage in metagenomic data sets and why it matters. *ISME J.* **8**, 2349–2351 (2014).

43. Sanders, J. G. *et al.* Optimizing sequencing protocols for leaderboard metagenomics by combining long and short reads. *Genome Biol.* **20**, 226 (2019).

44. Lodish, H., Berk, A. & Zipursky, S. Processing of rRNA and tRNA. in *Molecular Cell Biology* (2000).

45. Wöhlbrand, L. *et al.* Impact of Extraction Methods on the Detectable Protein Complement of Metaproteomic Analyses of Marine Sediments. *Proteomics* **17**, 1700241 (2017).

46. Zhang, X. *et al.* Assessing the impact of protein extraction methods for human gut metaproteomics. *J. Proteomics* **180**, 120–127 (2018).

47. Leary, D. H., Hervey, W. J., Deschamps, J. R., Kusterbeck, A. W. & Vora, G. J. Which metaproteome? The impact of protein extraction bias on metaproteomic analyses. *Mol. Cell. Probes* **27**, 193–199 (2013).

48. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165**, 535–550 (2016).

49. Wiśniewski, J. R. & Rakus, D. Multi-enzyme digestion FASP and the 'Total Protein Approach'-based absolute quantification of the Escherichia coli proteome. *J. Proteomics* **109**, 322–331 (2014).

50. Bundy, J. G., Willey, T. L., Castell, R. S., Ellar, D. J. & Brindle, K. M. Discrimination of pathogenic clinical isolates and laboratory strains of Bacillus cereus by NMR-based metabolomic profiling. *FEMS Microbiol. Lett.* **242**, 127–136 (2005).

51. Nicholson, J. K. & Lindon, J. C. Metabonomics. *Nature* **455**, 1054–1056 (2008).

52. Nicholson, J. K., Holmes, E. & Wilson, I. D. Gut microorganisms, mammalian metabolism and personalized health care. *Nat. Rev. Microbiol.* **3**, 431–438 (2005).

53. Hollywood, K., Brison, D. R. & Goodacre, R. Metabolomics: Current technologies and future trends. *Proteomics* **6**, 4716–4723 (2006).

54. VerBerkmoes, N. C., Denef, V. J., Hettich, R. L. & Banfield, J. F. Functional analysis of natural microbial consortia using community proteomics. *Nat. Rev. Microbiol.* **7**, 196–205 (2009).

55. Roberts, L. D., Souza, A. L., Gerszten, R. E. & Clish, C. B. Targeted Metabolomics. *Curr. Protoc. Mol. Biol.* **98**, 30.2.1-30.2.24 (2012).

56. Myers, E. W. The fragment assembly string graph. *Bioinformatics* **21**, ii79–ii85 (2005).

57. Simpson, J. T. & Durbin, R. Efficient construction of an assembly string graph using the FM-index. *Bioinformatics* **26**, i367–i373 (2010).

58. Wee, Y. *et al.* The bioinformatics tools for the genome assembly and analysis based on third-generation sequencing. *Brief. Funct. Genomics* (2019) doi:10.1093/bfgp/ely037.

59. Pevzner, P. A., Tang, H. & Waterman, M. S. An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci.* **98**, 9748–9753 (2001).

60. Zerbino, D. R. & Birney, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).

61. Kingsford, C., Schatz, M. C. & Pop, M. Assembly complexity of prokaryotic genomes using short reads. *BMC Bioinformatics* **11**, 21 (2010).

62. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA – A Practical Iterative de Bruijn Graph De Novo Assembler. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 426–440 (2010). doi:10.1007/978-3-642-12683-3_28.

63. Diniz, W. J. S. & Canduri, F. REVIEW-ARTICLE Bioinformatics: an overview and its applications. *Genet. Mol. Res.* **16**, (2017).

64. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo genome assemblies: Fig. 1. *Bioinformatics* **31**, 3350–3352 (2015).

65. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).

66. Wu, Y.-W., Tang, Y.-H., Tringe, S. G., Simmons, B. A. & Singer, S. W. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* **2**, 26 (2014).

67. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).

68. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).

69. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).

70. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).

71. Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* **3**, 836–843 (2018).

72. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).

73. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

74. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

75. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

76. Ram, R. J. *et al.* Community proteomics of a natural microbial biofilm. *Science* **308**, 1915–20 (2005).

77. Tanca, A. *et al.* The impact of sequence database choice on metaproteomic results in gut microbiota studies. *Microbiome* **4**, 51 (2016).

78. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A.-L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).

79. Jeong, H., Mason, S. P., Barabási, A. L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).

80. Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabási, A. L. Hierarchical organization of modularity in metabolic networks. *Science (80-.).* **297**, 1551–1555 (2002).

81. Lee, T. I. Transcriptional Regulatory Networks in Saccharomyces cerevisiae. *Science (80-.).* **298**, 799–804 (2002).

82. Mazurie, A., Bonchev, D., Schwikowski, B. & Buck, G. A. Evolution of metabolic network organization. *BMC Syst. Biol.* **4**, 59 (2010).

83. Broido, A. D. & Clauset, A. Scale-free networks are rare. *Nat. Commun.* **10**, 1017 (2019).

84. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

85. Raposo, F. *et al.* Biochemical methane potential (BMP) of solid organic substrates: evaluation of anaerobic biodegradability using data from an international interlaboratory study. *J. Chem. Technol. Biotechnol.* **86**, 1088–1098 (2011).

86. Hagen, L. H. Quantitative metaproteomics highlight the metabolic contributions of uncultured phylotypes in a thermophilic anaerobic digester. *PhD thesis* (2016).

87. Hagen, L. H. *et al.* Quantitative Metaproteomics Highlight the Metabolic Contributions of Uncultured Phylotypes in a Thermophilic Anaerobic Digester. *Appl. Environ. Microbiol.* **83**, (2017).

88. Taniguchi, Y. *et al.* Quantifying E. coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells. *Science (80-.).* **329**, 533–538 (2010).

INCLUDED PAPERS

Paper I

The ISME Journal (2018) 13:603–617
https://doi.org/10.1038/s41396-018-0290-y

ARTICLE

From proteins to polysaccharides: lifestyle and genetic evolution of *Coprothermobacter proteolyticus*

Benoit J. Kunath[1] · Francesco Delogu[1] · Adrian E. Naas[1] · Magnus Ø. Arntzen[1] · Vincent G. H. Eijsink[1] · Bernard Henrissat[2] · Torgeir R. Hvidsten[1] · Phillip B. Pope[1]

Received: 13 March 2018 / Revised: 11 July 2018 / Accepted: 19 September 2018 / Published online: 12 October 2018
© The Author(s) 2018. This article is published with open access

Abstract

Microbial communities that degrade lignocellulosic biomass are typified by high levels of species- and strain-level complexity, as well as synergistic interactions between both cellulolytic and non-cellulolytic microorganisms. *Coprothermobacter proteolyticus* frequently dominates thermophilic, lignocellulose-degrading communities with wide geographical distribution, which is in contrast to reports that it ferments proteinaceous substrates and is incapable of polysaccharide hydrolysis. Here we deconvolute a highly efficient cellulose-degrading consortium (SEM1b) that is co-dominated by *Clostridium (Ruminiclostridium) thermocellum* and multiple heterogenic strains affiliated to *C. proteolyticus*. Metagenomic analysis of SEM1b recovered metagenome-assembled genomes (MAGs) for each constituent population, whereas in parallel two novel strains of *C. proteolyticus* were successfully isolated and sequenced. Annotation of all *C. proteolyticus* genotypes (two strains and one MAG) revealed their genetic acquisition of carbohydrate-active enzymes (CAZymes), presumably derived from horizontal gene transfer (HGT) events involving polysaccharide-degrading Firmicutes or Thermotogae-affiliated populations that are historically co-located. HGT material included a saccharolytic operon, from which a CAZyme was biochemically characterized and demonstrated hydrolysis of multiple hemicellulose polysaccharides. Finally, temporal genome-resolved metatranscriptomic analysis of SEM1b revealed expression of *C. proteolyticus* CAZymes at different SEM1b life stages as well as co-expression of CAZymes from multiple SEM1b populations, inferring deeper microbial interactions that are dedicated toward community degradation of cellulose and hemicellulose. We show that *C. proteolyticus*, a ubiquitous population, consists of closely related strains that have adapted via HGT to presumably degrade both oligo- and longer polysaccharides present in decaying plants and microbial cell walls, thus explaining its dominance in thermophilic anaerobic digesters on a global scale.

Introduction

The anaerobic digestion of plant biomass profoundly shapes innumerable ecosystems, ranging from the gastrointestinal tracts of humans and other mammals to those that drive industrial applications such as biofuel generation. Biogas reactors are one of the most commonly studied anaerobic systems, yet many keystone microbial populations and their metabolic processes are poorly understood due to a lack of cultured or genome sampled representatives. *Coprothermobacter* spp. are frequently observed in high abundance in thermophilic anaerobic systems, where they are believed to exert strong protease activity, while generating hydrogen and acetate, key intermediate metabolites for biogas production [1]. Molecular techniques have shown that their levels range from 10% to 90% of the total microbial community, irrespective of bioreactors being operated on lignocellulose- or protein-rich substrates (Fig. 1). Despite their promiscuous distribution, global abundance and key role in biogas production, only two species have been described: *Coprothermobacter platensis* [2] and *Coprothermobacter*

These authors contributed equally: B. J. Kunath, F. Delogu.

Electronic supplementary material The online version of this article (https://doi.org/10.1038/s41396-018-0290-y) contains Supplementary Material, which is available to authorized users.

✉ Phillip B. Pope
phil.pope@nmbu.no

[1] Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Ås 1432, Norway

[2] Architecture et Fonction des Macromolécules Biologiques, CNRS, Aix-Marseille Université, Marseille F-13288, France

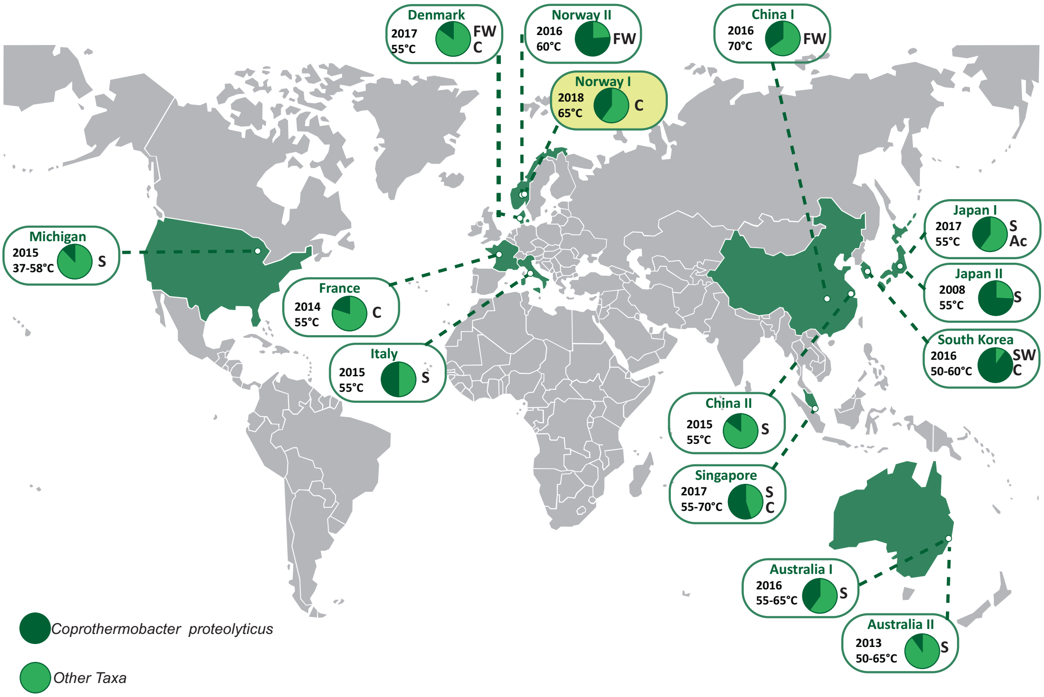


Fig. 1 Global distribution of *C. proteolyticus*-affiliated populations in anaerobic biogas reactors. Charts indicate relative 16S rRNA gene abundance of OTUs affiliated to *C. proteolyticus* (dark green), in comparison with the total community (light green). The year of publication, reactor temperature, and substrate (C cellulose, FW food waste, S sludge, SW Seaweed, Ac acetate) is indicated (details in Table S1). The SEM1b consortium analyzed in this study is highlighted in yellow

proteolyticus [3]. These two species and their inherent phenotypes have formed the predictive basis for the majority of *Coprothermobacter*-dominated systems described to date. Recent studies have illustrated that *C. proteolyticus* populations in anaerobic biogas reactors form cosmopolitan assemblages of closely related strains that are hitherto unresolved [4].

Frequently in nature, microbial populations are composed of multiple strains with genetic heterogeneity [5, 6]. Studies of strain-level populations have been predominately performed with the human microbiome and especially the gut microbiota [7, 8]. The reasons for strain diversification and their coexistence remain largely unknown [9]; however, several mechanisms have been hypothesized, such as microniche selection [5, 10], host selection [11], cross-feed interactions [12, 13], and phage selection [14]. Studies of axenic strains have shown that isolates can differ in a multitude of ways, including virulence and drug resistance [15–17], motility [18], and nutrient utilization [19]. Strain-level genomic variations typically consist of single-nucleotide variants, as well as acquisition/loss of genomic elements such as genes, operons, or plasmids via horizontal gene transfer (HGT) [20–22]. Variability in gene content caused by HGT is typically attributed to phage-related genes and other genes of unknown function [23], and can give rise to ecological adaptation, niche differentiation, and eventually, speciation [24–26]. Although differences in genomic features can be accurately characterized in isolated strains, it has been difficult to capture such information using culture-independent approaches such as metagenomics. Advances in bioinformatics have improved taxonomic profiling of microbial communities from phylum to species level but it remains difficult to profile similar strains from metagenomes and compare them with the same level of resolution obtained by comparison of isolate genomes [27]. As closely related strains can also differ in gene expression [28], being able to distinguish the expression profiles of individual strains in a broader ecological context is elemental to understanding the influence they exert towards the overall community function.

In this study, a novel population of *C. proteolyticus* that included multiple closely related strains was observed within a simplistic biogas-producing consortium enriched on cellulose (hereafter referred to as SEM1b). Using a combined metagenomic and culture-dependent approach, two strains and a metagenome-assembled genome (MAG) affiliated to *C. proteolyticus* were recovered and genetically compared with the only available type strain, *C. proteolyticus* DSM 5265 [29]. Notable genomic differences included the acquisition of an operon (region-A) encoding carbohydrate-active enzymes (CAZymes), which inferred that *C. proteolyticus* has adapted to take advantage of longer polysaccharides. Enzymology was used to further support our hypothesis that the CAZymes within region-A are functionally active. We further examined the saccharolytic potential of our recovered *C. proteolyticus* population in a broader community context, by examining genome-resolved temporal metatranscriptomic data generated from the SEM1b consortium. Collective analysis highlighted the time-specific polysaccharide-degrading activity that *C. proteolyticus* exerts in a cellulolytic microbial community.

Materials and methods

Generation of the SEM1b consortium

An inoculum (100 µl) was collected from a lab-scale biogas reactor (Reactor TD) fed with manure and food waste and run at 55 °C. The TD reactor originated itself from a thermophilic (60 °C) biogas plant (Frevar) fed with food waste and manure in Fredrikstad, Norway. Our research groups have previously studied the microbial communities in both the Frevar plant [4] and the TD bioreactor [30], which provided a detailed understanding of the original microbial community. The inoculum was transferred for serial dilution and enrichment to an anaerobic serum bottle and containing the rich ATCC medium 1943, with cellobiose substituted for 10 g/L of cellulose in the form of Borregaard Advanced Lignin technology (BALI^{TM})-treated Norway spruce [31]. Our enrichment was incubated at 65 °C with the lesser objective to study community biomass conversion at the upper temperature limits of methanogenesis. After an initial growth cycle, an aliquot was removed and used for a serial dilution to extinction experiment. Briefly, a 100 µl sample was transferred to a new 100 ml bottle containing 60 ml of anaerobic medium, mixed, and 100 µl was directly transferred again to a new one (six serial transfers in total). The consortium at maximum dilution that retained the cellulose-degrading capability (SEM1b) was retained for the present work and aliquots were stored at − 80 °C with glycerol (15% v/v). In parallel, continuous SEM1b cultures were maintained via regular transfers into fresh media (each recultivation incubated for ~2–3 days).

Metagenomic analysis

Two different samples (D1B and D2B) were taken from a continuous SEM1b culture and were used for shotgun metagenomic analysis. D2B was 15 recultivations older than D1B and was used to leverage improvements in metagenome assembly and binning. From 6 ml of culture, cells were pelleted by centrifugation at 14,000 × *g* for 5 min and were kept frozen at − 20 °C until processing. Non-invasive DNA extraction methods were used to extract high molecular weight DNA as previously described [32]. The DNA was quantified using a Qubit™ fluorimeter and the Quant-iT™ dsDNA BR Assay Kit (Invitrogen, USA), and the quality was assessed with a NanoDrop 2000 (Thermo Fisher Scientific, USA).

16S rRNA gene analysis was performed on both D1B and D2B samples. The V3–V4 hyper-variable regions of bacterial and archaeal 16S rRNA genes were amplified using the 341F/805R primer set: 5′-CCTACGGGNBGCASCAG-3′/5′-GACTACNVGGGTATCTAATCC-3′ [33]. The PCR was performed as previously described [30] and the sequencing library was prepared using Nextera XT Index kit according to Illumina's instructions for the MiSeq system (Illumina, Inc.). MiSeq sequencing (2 × 300 bp with paired ends) was conducted using the MiSeq Reagent Kit v3. The reads were quality filtered (Phred ≥ Q20) and USEARCH61 [34] was used for detection and removal of chimeric sequences. Resulting sequences were clustered at 97% similarity into operational taxonomic units (OTUs) and taxonomically annotated with the pick_closed_reference_otus.py script from the QIIME v1.8.0 toolkit [35] using the Greengenes database (gg_13_8). The resulting OTU table was corrected based on the predicted number of *rrs* operons for each taxon [36].

D1B and D2B were also subjected to metagenomic shotgun sequencing using the Illumina HiSeq 3000 platform (Illumina, Inc.) at the Norwegian Sequencing Center (NSC, Oslo, Norway). Samples were prepared with the TrueSeq DNA PCR-free preparation, and sequenced with paired ends (2 × 125 bp) on four lanes (two lanes per sample). Quality trimming of the raw reads was performed using cutadapt [37], removing all bases on the 3′-end with a Phred score lower than 20 (if any present) and excluding all reads shorter than 100 nt, followed by a quality filtering using the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). Reads with a minimum Phred score of 30 over 90% of the read length were retained. In addition, genomes from two isolated *C. proteolyticus* strains (see below) were used to decrease the data complexity and to improve the metagenomic assembly and binning. The quality-filtered metagenomic reads were mapped against the assembled strains using the Burrows-Wheeler Aligner with maximal exact matches (BWA-MEM) algorithm requiring 100% identity [38]. Reads that mapped the strains were removed from the metagenomic data and the remaining reads were co-assembled using MetaSpades v3.10.0 [39] with default parameters and k-mer sizes of 21, 33, 55, and 77. The subsequent contigs were binned with Metabat v0.26.3 [40] in "very sensitive mode", using the coverage information from D1B and D2B. The quality (completeness, contamination, and strain heterogeneity) of the bins (hereafter referred to as MAGs) was assessed by CheckM v1.0.7 [41] with default parameters.

Isolation of *C. proteolyticus* strains

Strains were isolated using the Hungate method [42]. In brief Hungate tubes were anaerobically prepared with the DSMZ medium 481 with and without agar (15 g/L). Directly after being autoclaved, Hungate tubes containing agar were cooled down to 65 °C and sodium sulfide nonahydrate was added. From the SEM1b culture used for D1B, 100 µl were transferred to a new tube and mixed. From this new tube, 100 µl was directly transferred to 10 ml of fresh medium, mixed, and transferred again (six transfers in total). Tubes were then cooled to 60 °C for the agar to solidify and then kept at the same temperature. After growth, single colonies were picked and transferred to liquid medium.

DNA was extracted using the aforementioned method for metagenomic DNA, with one amendment: extracted DNA was subsequently purified with DNeasy PowerClean Pro Cleanup Kit (Qiagen, USA) following manufacturer's instructions. To insure the purity of the *C. proteolyticus* colonies, visual confirmation was performed using light microscopy and long 16S rRNA genes were amplified using the primers pair 27F/1492R [43]: 5′-AGAGTTTGATCMTGGCTCAG-3′/5′-TACGGYTACCTTGTTACGACTT-3′ and sequenced using Sanger technology. The PCR consisted of an initial denaturation step at 94 °C for 5 min and 30 cycles of denaturation at 94 °C for 1 min, annealing at 55 °C for 1 min, and extension at 72 °C for 1 min, and a final elongation at 72 °C for 10 min. PCR products were purified using the NucleoSpin Gel and PCR Cleanup kit (Macherey-Nagel, Germany) and sent to GATC Biotech for Sanger sequencing.

The genomes of two isolated *C. proteolyticus* strains (hereafter referred to as *BWF2A* and *SW3C*) were sequenced at the NSC (Oslo, Norway). Samples were prepared with the TrueSeq DNA PCR-free preparation and sequenced using paired ends (2 × 300 bp) on a MiSeq system (Illumina, Inc). Quality trimming, filtering, and assembly were performed as described in the aforementioned metagenomic assembly section. The raw reads were additionally mapped on assembled contigs using bowtie2 (–very-sensitive -X 1000 -I 350) and the coverage was retrieved for every nucleotide with samtool depth –a. All the contigs with an average coverage higher than 100 were selected and individually inspected for coverage discontinuity. All the contigs selected with the average coverage criterion (BWF2A: 11, SW3C: 13) looked continuous in coverage and, together with the MAGs, they were submitted to the Integrated Microbial Genomes and Microbiomes system [44] for genomic feature prediction and annotation (pipeline version 4.15.1). Resulting annotated open reading frames (ORFs) were retrieved, further annotated for CAZymes using the CAZy annotation pipeline [45], and subsequently used as a reference database for the metatranscriptomics (with exception of glycosyltransferases). The genomes from both strains and MAGs corresponding to *C. proteolyticus* were compared with the reference genome from *C. proteolyticus* DSM 5265. Using the BRIG tool [46] for mapping and visualization, the different genomes were mapped against their pan genome generated using Roary [47].

Phylogenetic analysis

A concatenated ribosomal protein phylogeny was performed on the MAGs and the isolated strains using 16 ribosomal proteins chosen as single-copy phylogenetic marker genes (RpL2, 3, 4, 5, 6, 14, 15, 16, 18, 22, and 24, and RpS3, 8, 10, 17, and 19) [48]. The dataset was augmented with metagenomic sequences retrieved from our previous research on the original FREVAR reactor [4] and with sequences from reference genomes identified during the 16S rRNA analysis. Each gene set was individually aligned using MUSCLE v3.8.31 [49] and then manually curated to remove end gaps and ambiguously aligned terminal regions. The curated alignments were concatenated and a maximum likelihood phylogeny was obtained using MEGA7 [50] with 1000 bootstrap replicates. The radial tree was visualized using iTOL [51]. In addition, an average nucleotide identity (ANI) comparison was performed between each MAG and their closest relative using the ANI calculator [52].

Heterologous expression and purification of the GH16 enzyme

The *C. proteolyticus* BWF2A Ga0187557_1002 gene-sequence without predicted signal peptide [53] was cloned from isolated genomic DNA using the following primers; GH16_Fwd: 5′-TTAAGAAGGAGATATACTATGCTCGGCGTGAATGTGATG-AATATAAGTGA-3′; GH16_rev: 5′-AATGGTGGTGATGATGGTGCGCCTCATTTTCAAGCTTGTATA-CACGGACATAATC-3′, and cloned into the pNIC-CH plasmid in *Escherichia coli* TOP10 by ligation-independent cloning [54]. The transformant's sequence was verified by sequencing before transformation into OneShot® *E. coli* BL21 Star™ cells (Thermo Fischer Scientific, Waltham, MA, USA) for expression, where 200 ml Luria-broth containing 50 µg/ml kanamycin was inoculated with 2 ml overnight culture and incubated at 37 °C, 200 r.p.m. Expression was induced when the culture reached an OD600 of 0.6, by addition of isopropyl-β-D-1-thiogalactopyranoside. The culture was incubated at 22 °C, 200 r.p.m. for 16 h, before collection by centrifugation (5000 × *g*, 10 min) and storage of the pellet at − 80 °C. The frozen pellet was transferred to 20 mL buffer A (20 mM Tris-HCL pH 8.0, 200 mM NaCl, 5 mM imidazole) containing 1 × BugBuster (Merck Millipore, Berlington, MA, USA) and stirred for 20 min at room temperature to lyse the cells. Cell debris was removed by centrifugation (30,000 × *g*, 20 min) and the protein was purified by immobilized metal-ion chromatography using a 5 ml HisTrap FF column (GE-Healthcare, Little Chalfont, UK) pre-equilibriated with buffer A. The protein was eluted using a linear gradient to Buffer B (Buffer A with 500 mM imidazole). The purity of the eluted fractions were assessed by SDS-polyacrylamide gel electrophoresis and the imidazole was removed from the buffer by repeated concentration and dilution using a Vivaspin (Sartorius, Göttingen, Germany) concentrator with a 10 kDa cutoff. The protein concentration was determined by measured A280 and the calculated extinction coefficient.

Biochemical characterization of the GH16 enzyme

Assays were performed in triplicate in 96-well plates and contained 1 mg/ml substrate, 20 mM BisTris, pH 5.8 (50 °C), and 1 µM enzyme in a volume of 100 µl. The reactions were pre-heated to 50 °C before addition of enzyme and were sealed before incubation for 1 h in a Thermomixer C incubator with heated lid (Eppendorf, Hamburg, Germany). The substrates used were as follows: barley β-glucan, carboxymethyl-curdlan, carboxymethyl-pachyman, carob galactomannan, tamarind xyloglucan, wheat arabinoxylan, larch arabinogalactan (all from Megazyme, Bray, Co. Wicklow, Ireland), and laminarin from Laminaria digitate (Sigma-Aldrich, St. Louis, MO, USA). Reactions were stopped by addition of DNS reagent (100 µl, 10 g/l 3,5-dinitrosalicylic acid, 300 g/L potassium sodium tartrate, 10 g/L NaOH [55] for quantification, or NaOH to a final concentration of 0.1 M for product analysis. Reducing ends were quantified against a standard curve of glucose, where reactions with DNS reagent were incubated at 95 °C for 20 min before cooling on ice and the absorbance was measured at 540 nm. For product analysis, the reactions containing NaOH were further diluted 1:10 in water, before analysis by high-performance anion-exchange chromatography with pulsed amperometric detection (HPAEC-PAD), using a Dionex ICS3000 system with a CarboPac PA1 column (Sunnyvale, CA, USA). Oligosaccharides were eluted using a multi-step gradient, going from 0.1 M NaOH to 0.1 M NaOH–0.3 M sodium acetate (NaOAc) over 35 min, to 0.1 M NaOH–1.0 M NaOAc over 5 min, before going back to 0.1 M NaOH over 1 min, and reconditioning for 9 min at 0.1 M NaOH.

Temporal meta-omic analyses of SEM1b

A “meta-omic” time series analysis was conducted over the lifetime span of the SEM1b consortium (≈45 h). A collection of 27 replicate bottles containing ATCC medium 1943 with 10 g/L of cellulose (60 ml total volume) were inoculated from the same SEM1b culture and incubated at 65 °C in parallel. For each sample time point, three culture-containing bottles were removed from the collection and processed in triplicate. Sampling occurred over nine time points (at 0, 8, 13, 18, 23, 28, 33, 38, and 43 h) during the SEM1b life cycle and are hereafter referred as T0, T1, T2, T3, T4, T5, T6, T7, and T8, respectively. DNA for 16S rRNA gene analysis was extracted (as above) from T1 to T8 and kept at − 20 °C until amplification and sequencing, and the analysis was performed using the protocol described above. Due to low cell biomass at the initial growth stages, sampling for metatranscriptomics was performed from T2 to T8. Sample aliquots (6 ml) were treated with RNAprotect Bacteria Reagent (Qiagen, USA) following the manufacturer’s instructions and the treated cell pellets were kept at − 80 °C until RNA extraction.

In parallel, metadata measurements including cellulose degradation rate, monosaccharide production, and protein concentration were performed over all the nine time points (T0–T8). For monosaccharide detection, 2 ml samples were taken in triplicates, centrifuged at 16,000 × *g* for 5 min and the supernatants were filtered with 0.2 µm sterile filters and boiled for 15 min before being stored at − 20 °C until processing. Solubilized sugars released during microbial hydrolysis were identified and quantified by HPAEC with PAD. A Dionex ICS3000 system (Dionex, Sunnyvale, CA, USA) equipped with a CarboPac PA1 column (2 × 250 mm; Dionex, Sunnyvale, CA, USA) and connected to a guard of the same type (2 × 50 mm) was used. Separation of products was achieved using a flow rate of 0.25 mL/min in a 30 min isocratic run at 1 mM KOH at 30 °C. For quantification, peaks were compared with linear standard curves generated with known concentrations of selected monosaccharides (glucose, xylose, mannose, arabinose, and galactose) in the range of 0.001–0.1 g/L.

Total protein measurements were taken to estimate SEM1b growth rate. Proteins were extracted following a previously described method [4] with a few modifications. Briefly, 30 ml culture aliquots were centrifuged at 500 × *g* for 5 min to remove the substrate and the supernatant was centrifuged at 9000 × *g* for 15 min to pellet the cells. Cell lysis was performed by resuspending the cells in 1 ml of lysis buffer (50 mM Tris-HCl, 0.1% (v/v) Triton X-100, 200 mM NaCl, 1 mM dithiothreitol, 2 mM EDTA) and keeping them on ice for 30 min. Cells were disrupted in 3 × 60 s cycles using a FastPrep24 (MP Biomedicals, USA) and the debris were removed by centrifugation at 16,000 × *g* for 15 min. Supernatants containing proteins were transferred into low bind protein tubes and the proteins were quantified using Bradford's method [56].

As estimation of cellulose degradation requires analyzing the total content of a sample to be accurate, the measurements were performed on individual cultures that were prepared separately. A collection of 18 bottles (9 time points in duplicate) were prepared using the same inoculum described above and grown in parallel with the 27-bottle collection used for the meta-omic analyses. For each time point, the entire sample was recovered, centrifuged at 5000 × *g* for 5 min, and the supernatant was discarded. The resulting pellets were boiled under acidic conditions as previously described [57] and the dried weights, corresponding to the remaining cellulose, were measured.

mRNA extraction was performed in triplicate on time points T2–T8, using previously described methods [58] with the following modifications in the processing of the RNA. The extraction of the mRNA included the addition of an in vitro-transcribed RNA as an internal standard to estimate the number of transcripts in the natural sample compared with the number of transcripts sequenced. The standard was produced by the linearization of a pGem-3Z plasmid (Promega, USA) with ScaI (Roche, Germany). The linear plasmid was purified with a phenol/chloroform/isoamyl alcohol extraction and digestion of the plasmid was assessed by agarose gel electrophoresis. The DNA fragment was transcribed into a 994 nt-long RNA fragment with the Riboprobe in vitro Transcription System (Promega, USA) following the manufacturer's protocol. Residual DNA was removed using the Turbo DNA Free kit (Applied Biosystems, USA). The quantity and the size of the RNA standard was measured with a 2100 bioanalyzer instrument (Agilent).

Total RNA was extracted using enzymatic lysis and mechanical disruption of the cells and purified with the RNeasy mini kit following the manufacturer's protocol (Protocol 2, Qiagen, USA). The RNA standard (25 ng) was added at the beginning of the extraction in every sample. After purification, residual DNA was removed using the Turbo DNA Free kit, and free nucleotides and small RNAs such as tRNAs were cleaned off with a lithium chloride precipitation solution according to Thermo Fisher Scientific's recommendations. To reduce the amount of rRNAs, samples were treated to enrich for mRNAs using the MICROBExpress kit (Applied Biosystems, USA). Successful rRNA depletion was confirmed by analyzing both pre- and post-treated samples on a 2100 bioanalyzer instrument. Enriched mRNA was amplified with the MessageAmp II-Bacteria Kit (Applied Biosystems, USA) following manufacturer's instruction and sent for sequencing at the NSC (Oslo, Norway). Samples were subjected to the TruSeq stranded RNA sample preparation, which included the production of a cDNA library, and sequenced with paired-end technology (2 × 125 bp) on one lane of a HiSeq 3000 system.

RNA reads were assessed for overrepresented features (adapters/primers) using FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc/) and ends with detected features and/or a Phred score lower than 20 were trimmed using Trimmomatic v.0.36 [59]. Subsequently, a quality filtering was applied with an average Phred threshold of 30 over a 10 nt window and a minimum read length of 100 nt. rRNA and tRNA were removed using SortMeRNA v.2.1b [60]. SortMeRNA was also used to isolate the reads originating from the pGem-3Z plasmid. These reads were mapped against the specific portion of the plasmid containing the Ampr gene using Bowtie2 [61] with default parameters and the number of reads per transcript was quantified and scaled to match the length of the standard (x5.08). The remaining reads were pseudoaligned against the metagenomic dataset, augmented with the annotated strains, using Kallisto pseudo –pseudobam [62]. The resulting output was used to generate mapping files with bam2hits, which were used for expression quantification with mmseq [63], and the results were scaled to match the initial volume of the samples (x 10). Of the 40,046 ORFs identified from the assembled SEM1b metagenome and 2 *C. proteolyticus* strains, 17,598 (44%) were not found to be expressed, whereas 21,480 (54%) were expressed and could be reliably quantified due to unique hits (reads mapping unambiguously against one unique ORF) (Figure S1A). The remaining 968 ORFs (2%) were expressed but identified only with shared hits (reads mapping ambiguously against more than one ORF, resulting in an unreliable quantification of the expression of each ORF) (Figure S1B). As having unique hits improves the expression estimation accuracy, the ORFs were grouped using mmcollapse, in order to improve the precision of expression estimates, with only a small reduction in biological resolution [64]. The process first collapses ORFs into homologous groups if they have 100% sequence identity and then further collapses ORFs (or expression groups) if they acquire unique hits as a group (Figure S1C). This process generated 39,146 expression groups of which 38,428 (98%) were singletons (groups composed of single ORF) and 718 (2%) were groups containing more than one homologous ORF. From the initial 968 low-information ORFs, 661 (68%) became part of an expression group containing unique hits, 77 (8%) became part of ambiguous group (no unique hits), and 230 (24%) remained singletons (without unique hits). All expression groups without unique hits were then excluded from the subsequent analysis. A total of 21,480 singletons and 605

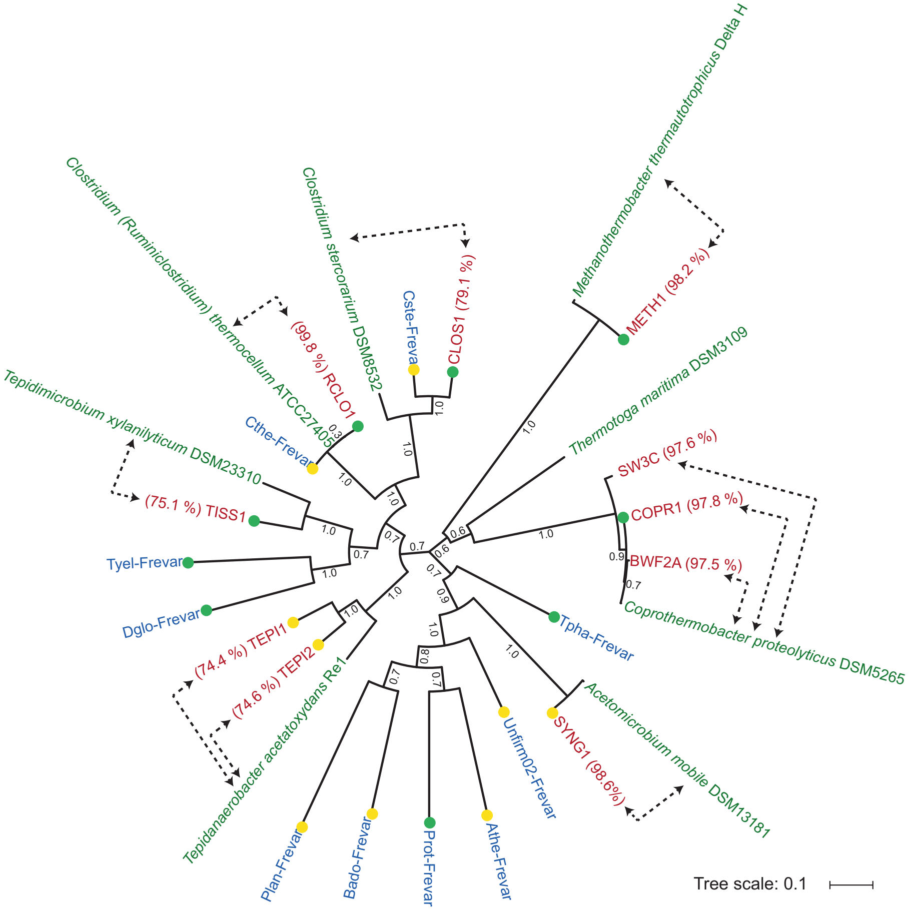

Methanothermobacter thermautotrophicus Delta H
METH1 (98.2 %)
Clostridium stercorarium DSM8532
Cste-Freva
CLOS1 (79.1 %)
Clostridium (Ruminiclostridium) thermocellum ATCC27405
(99.8 %) RCLO1
Cthe-Frevar
Tepidimicrobium xylanilyticum DSM23310
(75.1 %) TISS1
Tyel-Frevar
Dglo-Frevar
(74.4 %) TEPI1
(74.6 %) TEPI2
Tepidanaerobacter acetatoxydans Re1
Plan-Frevar
Bado-Frevar
Prot-Frevar
Athe-Frevar
Unfirm02-Frevar
SYNG1 (98.6%)
Acetomicrobium mobile DSM13181
Tpha-Frevar
Thermotoga maritima DSM3109
SW3C (97.6 %)
COPR1 (97.8 %)
BWF2A (97.5 %)
Coprothermobacter proteolyticus DSM5265
Tree scale: 0.1

Fig. 2 Phylogeny of *C. proteolyticus* strains and other MAGs recovered from the SEM1b consortium. Concatenated ribosomal protein tree of reference isolate genomes (green), MAGs from the previous Frevar study (blue [4]), and MAGs and isolate genomes recovered in this study (red). Average nucleotide identities (percentage indicated in parenthesis) were generated between SEM1b MAGs and their closest relative (indicated by dotted arrows). Bootstrap values are based on 1000 bootstrap replicates and the completeness of the MAGs are indicated by green (> 90 %) and yellow (> 80 %) colored dots

multiple homologous expression groups were reliably quantified between *BWF2A*, *SW3C*, and the SEM1b metatranscriptome (Figure S1C).

In order to normalize the expression estimates, sample sizes were calculated using added internal standards, as described previously [58]. The number of reads generated from the internal standard molecule were calculated to be 2.4 × 10^4 +/− 2.1 × 10^4 reads per sample out of 6.2 × 10^9 molecules added. Using this information, the estimated number of transcript molecules per sample was computed to be 1.0 × 10^{13} +/− 7.3 × 10^{12} transcripts. The resulting estimates for the sample sizes were used to scale the expression estimates from mmseq collapse and to obtain absolute expression values. During initial screening the sample T7C (time point T7, replicate C) was identified as an outlier using principle component analysis and removed from downstream analysis.

The expression groups were clustered using hierarchical clustering with Euclidean distance. Clusters were identified using the Dynamic Tree Cut algorithm [65] with hybrid mode, deepsplit = 1, and minClusterSize = 7. Eigengenes were computed for the clusters and clusters with a Pearson's correlation coefficient > 0.9 were merged. The MAG/strain enrichment of the clusters was assessed using the BiasedUrn R package. The *p*-values were corrected with the Benjamini–Hochberg procedure and the significance threshold was set to 0.05. Expression groups composed of multiple MAGs/strains were included in several enrichment tests.

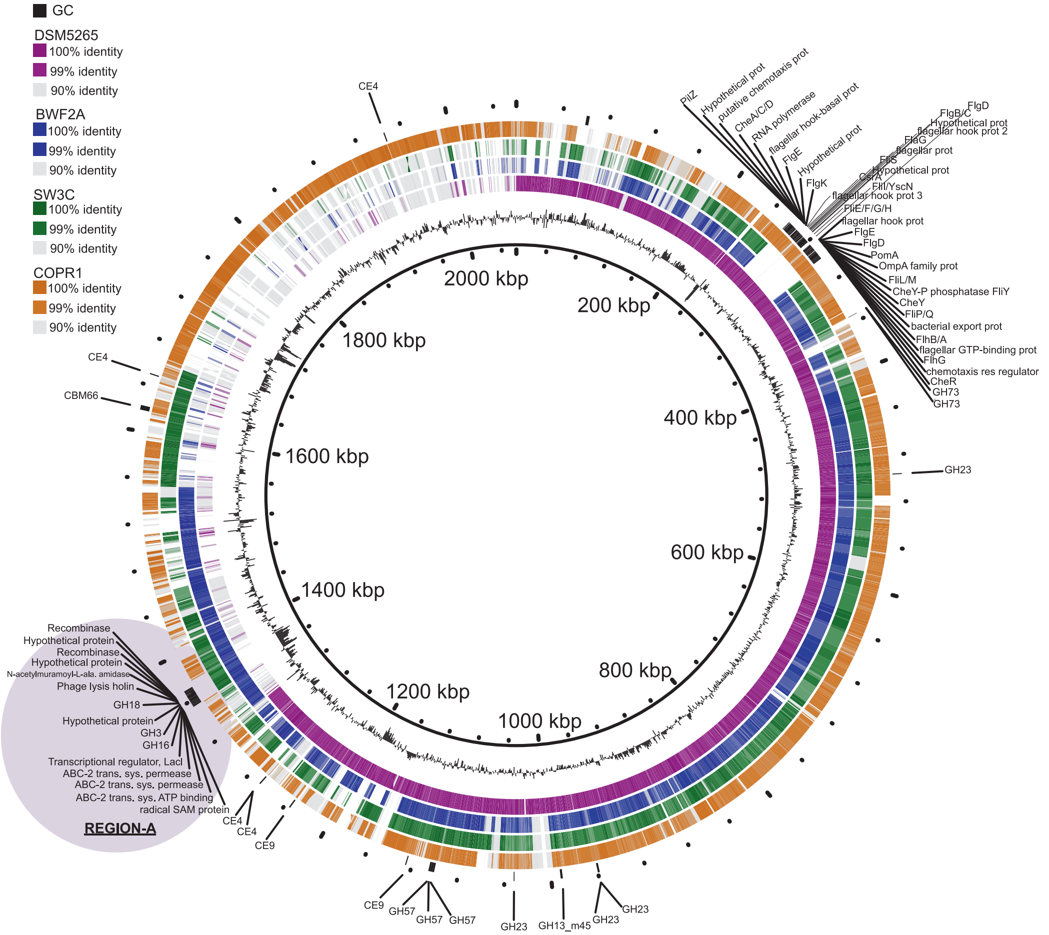


Fig. 3 Comparative genome content of *C. proteolyticus* representatives including isolated strains, a recovered MAG (COPR1), and the reference strain DSM 5265. The innermost ring corresponds to the pan genome of the three *C. proteolyticus* spp. genomes and one MAG as produced by Roary [47], and the second innermost ring represents the GC content. Outer rings represent the reference strain DSM 5265 (purple), the isolated strains *BWF2A* (blue) and *SW3C* (green), and the recovered COPR1 MAG (orange). Genes coding for carbohydrate-active enzymes (CAZymes) and flagellar proteins are indicted in black on the outermost ring. Genomic region-A is indicated by purple shading

Results and discussion

The SEM1b consortium is a simplistic community, co-dominated by *Clostridium (Ruminiclostridium) thermocellum* and heterogeneic *C. proteolyticus* strains

Molecular analysis of a reproducible, cellulose-degrading, and biogas-producing consortium (SEM1b) revealed a stable and simplistic population structure that contained approximately seven populations, several of which consisted of multiple strains (Fig. 2, Table S2–S3). 16S rRNA gene analysis showed that the SEM1b consortium was co-dominated by OTUs affiliated to the genera *Clostridium* (52%) and *Coprothermobacter* (41%), with closest representatives identified as *C. (Ruminiclostridium) thermocellum*, an uncharacterized *Clostridium spp.* and three *Coprothermobacter* phylotypes (Table S2). Previous meta-omic analysis on the parent Frevar reactor, revealed a multitude of numerically dominant *C. proteolyticus* strains, which created significant assembly and binning related issues [4]. In this study, multiple oligotypes of *C. proteolyticus* were also found (Table S2). We therefore sought to isolate and recover axenic representatives to complement our meta-omic approaches, and using traditional anaerobic isolation

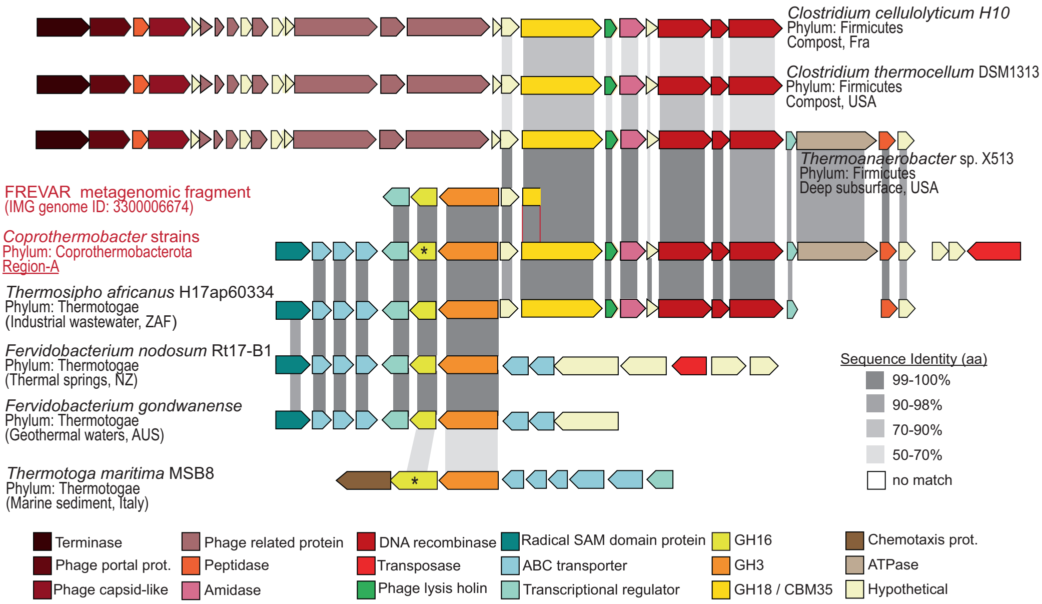


Fig. 4 Gene synteny of CAZymes within region-A encoded in *BWF2A* and *SW3C* genomes. The gene organization of CAZymes within region-A encoded in *BWF2A* and *SW3C* (see Fig. 3), as well as highly similar operons found in the original Frevar metagenome and isolated representatives from both phyla Firmicutes (*Thermoanaerobacter*, *C. cellulolyticum*, *C. thermocellum*) and Thermotogae (*T. africanus*, *F. nodosum*, *F. gondwanense*, and *Thermotoga maritima*). Grey shading between individual ORFs indicates amino acid sequence identity calculated between each query ORF (Frevar metagenome and isolates) and the reference ORF encoded in region-A from *BWF2A* and *SW3C* (identical in both strains). Asterisk denotes biochemically characterized GH16 enzymes, including the *C. proteolyticus* representative from this study and a laminarinase from *Thermotoa maritima* MSB8 that has previously been reported [79]

techniques, we were successful in recovering two novel axenic strains (hereafter referred to as *BWF2A* and *SW3C*). The genomes of *BWF2A* and *SW3C* were sequenced and assembled, and subsequently incorporated into our metagenomic and metatranscriptomic analysis below.

Shotgun metagenome sequencing of two SEM1b samples (D1B and D2B) generated 290 Gb (502 M paired-end reads) and 264 Gb (457 M paired-end reads) of data, respectively. Co-assembly of both datasets using strain-depleted reads with Metaspades produced 20,760 contigs totalizing 27 Mbp with a maximum contig length of 603 Kbp. Taxonomic binning revealed 11 MAGs and a community structure similar to the one observed by 16S analysis (Fig. 2, Table S3). A total of eight MAGs exhibited high completeness (> 80%) and a low level of contamination (< 10%). Three MAGs, COPR2, COPR3, and SYNG2, corresponded to small and incomplete MAGs, although Blastp analysis suggest COPR2 and COPR3 likely represent *Coprothermobacter*-affiliated strain elements.

All near-complete MAGs (> 80%), as well as *BWF2A* and *SW3C*, were phylogenetically compared against their closest relatives using ANIs and a phylogenomic tree was constructed via analysis of 16 concatenated ribosomal proteins (Fig. 2). One MAG was observed to cluster together with *C. proteolyticus* DSM 5265 and the two strains *BWF2A* and *SW3C*, and was defined as COPR1. Two MAGs (RCLO1-CLOS1) clustered together within the *Clostridium*; RCLO1 with the well-known *C. thermocellum*, whereas CLOS1 grouped together with another *Clostridium* MAG generated from the Frevar dataset and the isolate *C. stercorarium* (ANI: 79.1%). Both RCLO1 and CLOS1 encoded broad plant polysaccharide-degrading capabilities, containing 297 and 139 CAZymes, respectively (Table S4). RCLO1 in particular encoded cellulolytic (e.g., glycosyl hydrolase (GH) families GH5, GH9, and GH48) and cellulosomal features (dockerins and cohesins), whereas CLOS1 appears more specialized toward hemicellulose degradation (e.g., GH3, GH10, GH26, GH43, GH51, and GH130). Surprisingly, several CAZymes were also identified in COPR1 ($n = 65$), and both *BWF2A* (n ×= 37) and *SW3C* ($n = 34$) at levels higher than what has previously been observed in *C. proteolyticus* DSM 5265 ($n = 29$) (Table S4). Several MAGs were also affiliated with other known lineages associated with biogas processes, including *Tepidanaerobacter* (TEPI1-2), *Synergistales* (SYNG1-2), *Tissierellales* (TISS1), and *Methanothermobacter* (METH1).

Novel strains of *C. proteolyticus* reveal acquisition of CAZymes

Genome annotation of COPR1, *BWF2A*, and *SW3C* identified both insertions and deletions in comparison with the only available reference genome, sequenced from the type strain DSM 5265 (Fig. 3). Functional annotation showed that most of the genomic differences were sporadic and are predicted not to affect the metabolism of the strains. However, several notable differences were observed, which might represent a significant change in the lifestyle of the isolates. Both isolated strains lost the genes encoding flagellar proteins, although it is debatable that these genes originally conferred mobility in the type strain, as it has been previously reported as non-motile [3, 66]. Interestingly, both strains acquired extra CAZymes including a particular genomic region that encoded a cluster of three CAZymes: GH16, GH3, and GH18-CBM35 (region-A, Fig. 3). The putative function of these GHs suggests that both *BWF2A* and *SW3C* are capable of hydrolyzing various β-glucan linkages that are found in different hemicellulosic substrates (GH16: endo-β-1,3-1,4-glucanase; GH3: β-glucosidase). Regarding the putative GH18 encoded in both strains, it could have a role in bacterial cell wall recycling [67] as an endo-β-*N*-acetylglucosaminidase. Indeed, *C. proteolyticus* has previously been considered to be a scavenger of dead cells, even though this feature was mainly highlighted in term of proteolytic activities [68].

Taking a closer look, the region-A of CAZymes (GH16, GH3, and GH18-CBM35) in *BWF2A* and *SW3C* was located on the same chromosomal cassette but organized onto two different operons with opposite directions (Fig. 4). Comparison of the genes and their organization revealed a high percentage of gene similarity and synteny with genome representatives from both phyla Firmicutes (*Thermoanaerobacter*, *Clostridium cellulolyticum*, and *C. thermocellum*) and Thermotogae (*Thermosipho africanus*, *Fervidobacterium nodosum*, and *F. gondwanense*). Both *C. thermocellum* and *Fervidobacterium* populations were previously identified in the original Frevar reactor [4]. Moreover, a truncated contig from the Frevar metagenome (Scaffold Id:Ga0101770_1036339) exhibited 99.9 % nucleotide identity to the *BWF2A* and *SW3C* genomes spanning 4.7 Kb across the CAZymes and genomic sections from both phyla (Fig. 4), suggesting the acquirement of region-A preceded the SEM1b enrichment.

Examination of the flanking regions surrounding the CAZymes in region-A reveals the presence of an incomplete prophage composed of a phage lysis holin and two recombinases located downstream (Figs. 3, 4). Further comparisons revealed that only the Firmicutes lineages encoded the same prophage together with an additional terminase, phage-capsid-like proteins, and more phage-related components on the 5′-region (Fig. 4). Because of the high sequence homology and the presence of phage-genes in the surrounding, we hypothesized that the origin of region-A in *BWF2A* and *SW3C* is the result of phage-mediated HGT. Most likely, the operon from Firmicutes-affiliated lineages (e.g., *Thermoanaerobacter* and *C. thermocellum*) was transferred first due to the presence of its complete phage and generated a hotspot for further HGT for the GH16-GH3-encoding operon originating from Thermotogae-affiliated lineages (Fig. 4). Interestingly, *T. africanus* also encoded a syntenous region that covered Region-A in both *BWF2A* and *SW3C* almost in its entirety (Fig. 4), creating an alternative possibility that vertical gene transfer may also have had a role toward the evolution of this operon in *Coprothermobacter*. Gene transfer within anaerobic digesters has been reported for antibiotic resistance genes [69], whereas HGT of CAZymes have been detected previously among gut microbiota [70–72]. As many microbes express only a specific array of carbohydrate-degrading capabilities, bacteria that acquire CAZymes from gene transfer events may gain additional capacities and, consequently, a selective growth advantage [73].

In response to our discovery of *C. proteolyticus* CAZyme acquisition, we attempted to cultivate our axenic strains in minimal media containing only hemicellulosic substrates (pachyman, curdlan, barley β-glucan) as a sole carbon source. However, no growth was observed for either *BWF2A* or *SW3C* in polysaccharide-supplemented media that was without yeast extract. These results were consistent with the few available studies on type strain DSM 5265, which have shown weak and slow growth on proteins and monomeric sugars, and only in the presence of pluralistic organic compounds found in yeast extract and rumen fluid [3, 66]. Growth was observed in *BWF2A*/*SW3C* cultures with both yeast extract and polysaccharide substrates; however, we detected no increased levels of growth, indicating that in isolation our *C. proteolyticus* strains may require specific undefined cofactor(s) or collaborative microbial partners to support the activity encoded by their acquired CAZymes.

In lieu of axenic *C. proteolyticus* cultivation data to support a saccharolytic lifestyle, we biochemically interrogated the GH16 encoded in region-A (Fig. 4). The catalytic domain was synthesized and expressed in *E. coli*, followed by protein purification. As expected the GH16 demonstrated endoglucanase activity on β-1,3 (pachyman, curdlan, laminarin) and β-1,3-1,4 (Barley) substrates (Figure S2A), which supports our hypothesis that the CAZymes in region-A have transferred the ability of *BWF2A* or *SW3C* to degrade polysaccharides. Against all β-glucan substrates, GH16 hydrolysis generated a large fraction of glucose (Figure S2B), which has been shown to be readily fermented by *C. proteolyticus* [3, 66].

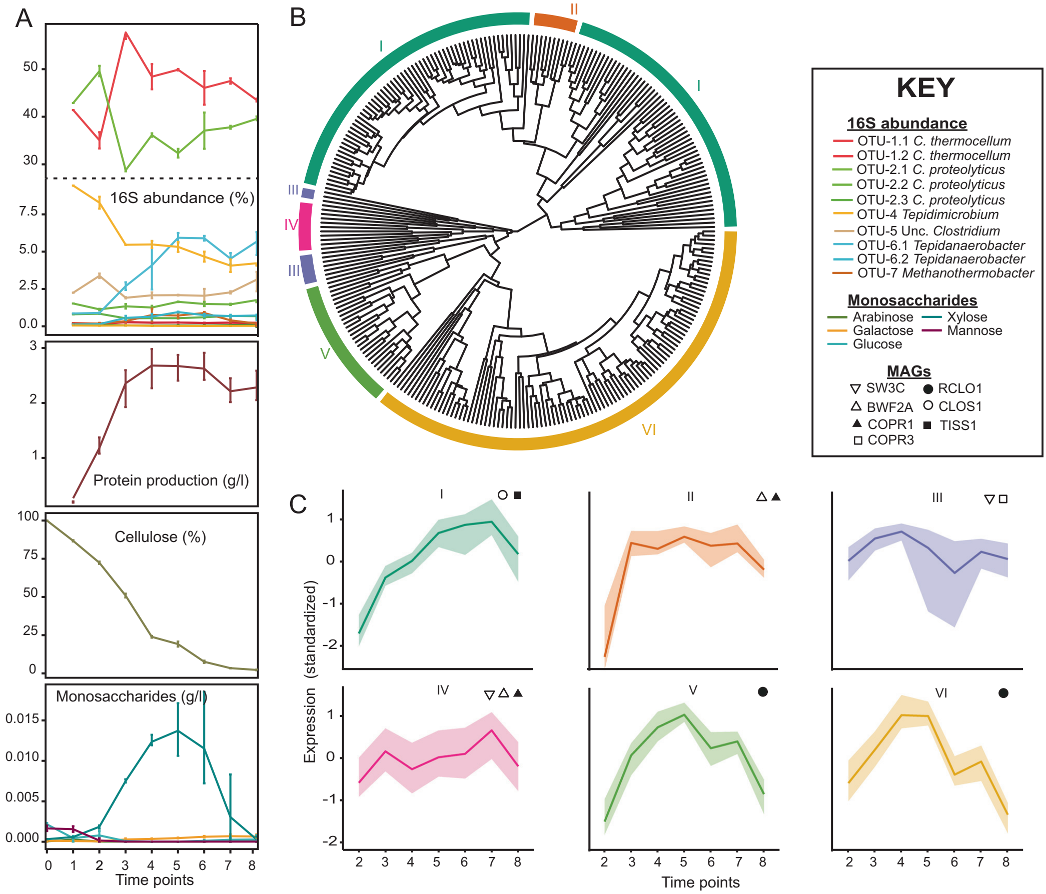


Fig. 5 Temporal meta-analysis of the SEM1b consortium. **a** 16S rRNA gene amplicon and metadata analysis was performed over a 43 h period, which was segmented into nine time points. OTU IDs are detailed in Table S2. Cellulose degradation rate, monosaccharide accumulation, and growth rate (estimated by total protein concentration) are presented. **b** Gene expression dendrogram and clustering of CAZymes from *BWF2A*, *SW3C*, and MAGs: RCLO1, CLOS1, COPR1-3, and TISS1. Six expression clusters (I–VI) are displayed in different colors on the outer ring. **c** Clusters I–VI show characteristic behaviors over time summarized by the median (solid line) and the shaded area between the first and third quartile of the standardized expression. Bacteria that are statistically enriched (p-value < 0.05) in the clusters are displayed in the subpanels

C. proteolyticus expresses CAZymes and is implicit in collaborative polysaccharide degradation within the SEM1b consortium

Although we confirmed that the acquired *C. proteolyticus* GH16 is functionally active, we also sought to better understand the role(s) had by it and other *C. proteolyticus* CAZymes in a saccharolytic consortium, by analyzing the temporal metatranscriptome of SEM1b over a complete life cycle. 16S rRNA gene analysis of eight time points (T1–8) over a 43 h period reaffirmed that *C. thermocellum*- and *C. proteolyticus*-affiliated populations dominate SEM1b over time (Fig. 5a). Highly similar genes from different MAGs/genomes were grouped together, in order to obtain "expression groups" with discernable expression profiles (see Methods and Figure S1A/B). A total of 274 singleton CAZyme expression groups and 8 multiple ORF groups were collectively detected in the two *C. proteolyticus* strains and MAGs suspected of contributing to polysaccharide degradation (RCLO1, CLOS1, COPR1-3, and TISS1, Figure S1D, Table S5). In several instances, expressed CAZymes from *BWF2A* and *SW3C* could not be resolved between the two strains and/or the COPR1 MAG. For example, all GHs within region-A could be identified as expressed by at least one of the isolated strains but could not be resolved further between the strains.

From the CAZymes subset of expression groups, a cluster analysis was performed to reveal six expression clusters (I–VI, Fig. 5b). Clusters II, III, and IV were enriched with *C. proteolyticus*-affiliated MAGs and isolated strains. Clusters III and IV comprised 10 and 11 expression groups, respectively, and followed a similar profile over time (Fig. 5c), increasing at earlier stages (T2–3) and again at later stationary/death stages (T6–8). Cluster II (10 expression groups) was slightly variant and increased more rapidly at T2 and sustained high levels over the course of SEM1b. All three clusters consisted of CAZymes targeting linkages associated with *N*-acetylglucosamine (CE9) and peptidoglycan (CE4, GH23, and GH73), suggesting a role in bacterial cell wall hydrolysis (Table S5). This hypothesis was supported by 16S rRNA gene data, which illustrated that *C. proteolyticus*-affiliated populations (OTU2), were high at initial stages of the SEM1b life cycle when cell debris was likely present in the inoculum that was sourced from the preceding culture at stationary phase (Fig. 5a). At T2, the abundance of *C. thermocellum*-affiliated populations (OTU-1) was observed to outrank *C. proteolyticus* as the community predictably shifted to cellulose utilization. However, toward stationary phase (T6–8) when dead cell debris is expected to be increasing, expression levels in clusters II, III, and IV were maintained at high levels (Fig. 5b), which was consistent with high *C. proteolyticus* 16S rRNA gene abundance at the same time points.

Clusters V and VI comprised 28 and 101 expression groups (respectively), and were enriched with the RCLO1 MAG that was closely related to *C. thermocellum*. As expected, numerous expressed genes in cluster V and VI were inferred in cellulosome assembly (via dockerin domains) as well as cellulose (e.g., GH5, GH9, GH44, GH48, CBM3) and hemicellulose (e.g., GH10, GH11, GH26, GH43, GH74) hydrolysis (Table S5). Both clusters increased throughout the consortium's exponential phase (time points T1–4, Fig. 5a), whereas 16S rRNA data also shows *C. thermocellum*-affiliated populations at high levels during the same stages (Fig. 5a).

Cluster I was determined as the largest with 121 expression groups and was particularity enriched with CLOS1, which expressed many genes involved in hemicellulose deconstruction (e.g., GH3, GH10, GH29, GH31, GH43, and GH130) and carbohydrate deacetylation (e.g., CE4, CE7, CE8, CE9, CE12, and CE15) (Table S5). Genes encoding CAZymes from both *BWF2A* and *SW3C* were also expressed in cluster I including the functionally active GH16- and GH3-encoding ORFs from region-A, which reaffirms our earlier predictions that certain *C. proteolyticus* populations in SEM1b are capable of degrading hemicellulosic substrates. The expression profile of cluster I over time was observed to slightly lag after cluster V and VI (Fig. 5), suggesting that genes encoding hemicellulases in cluster I are expressed once the hydrolytic effects of the RCLO1 cellulosome (expressed in cluster V and VI) have liberated hemicellulosic substrates [74]. Although *C. thermocellum* cannot readily utilize other carbohydrates besides glucose and longer glucans [75], the cellulosome is composed of a number of hemicellulolytic enzymes such as GH10 and GH11 endoxylanases, GH26 mannanases, GH74 xyloglucanases, and GH43 arabinanases/xylosidases [76], which are involved in the deconstruction of the underlying cellulose–hemicellulose matrix [74]. Interestingly, RCLO1 representatives of GH10, GH11, GH5, GH9, GH16, and GH43 were all expressed in the additional RCLO1-enriched cluster V and are presumably acting on the hemicellulose fraction present in the spruce-derived cellulose [77]. Furthermore, detection of hydrolysis products (Fig. 5a) revealed that xylose increased significantly at T5–7, indicating that hemicellulosic polymers containing β-1-4-xylan were likely available at these stages. Cluster V exhibited a similar profile to the other RCLO1-enriched cluster (Cluster VI), however its high expression levels were extended to T7, consistent with our observed levels of xylose release (Fig. 5c).

An additional GH16 from RCLO1 was also expressed in SEM1b cluster V, which has 99.5% amino acid sequence identity to Lic16A, a biochemically characterized endoglucanase that exerts specific β-1,3 activity similar to the *BWF2A*/*SW3C* GH16 that we report here. Notably, Lic16A is a cell wall anchored, non-cellulosomal CAZyme that is believed to enable *C. thermocellum* to grow exclusively on β-1,3-glucans [78]. All in all, the SEM1b expression data shows sequential community progression that co-ordinates putative hydrolysis of cellulose and hemicellulosic substrates as well as carbohydrates that are found in the microbial cell wall. In particular, *C. proteolyticus* populations in SEM1b were suspected to have key roles degrading microbial cell wall carbohydrates and hemicellulosic substrates, possibly in cooperation or in parallel to other clostridium populations at the later stages of the SEM1b growth cycle.

Conclusions

Unraveling the interactions occurring in a complex microbial community composed of closely related species or strains is an arduous task. Here we have leveraged culturing techniques, metagenomics, time-resolved metatranscriptomics, and enzymology to describe a novel *C. proteolyticus* population that comprised closely related strains that have acquired CAZymes via HGT and putatively evolved to incorporate a saccharolytic lifestyle. The co-expression patterns of *C. proteolyticus* CAZymes in clusters II, III, and IV supports the adaptable role of this bacterium as a scavenger that is able to hydrolyze cell wall polysaccharides during initial phases of growth and in the stationary/death phase, when available sugars are low. Moreover, the acquisition of biochemically verified hemicellulases by *C. proteolyticus* and their co-expression in cluster I at time points when hemicellulose is available further enhances its metabolic versatility and provides substantial evidence as to why this population dominates thermophilic reactors on a global scale, even when substrates are poor in protein.

Data availability

All sequencing reads have been deposited in the sequence read archive (SRP134228), with specific numbers listed in Table S6. All microbial genomes are publicly available on JGI under the analysis project numbers listed in Table S6. The code used to perform the computational analysis is available at: https://github.com/fdelogu/SEM1b-CAZymes.git

Acknowledgements We are grateful for support from The Research Council of Norway (FRIPRO program, P.B.P.: 250479/NorZymeD, V.G.H.E.: 221568), as well as the European Research Commission Starting Grant Fellowship (awarded to P.B.P.; 336355—MicroDE). The sequencing service was provided by the Norwegian Sequencing Centre (www.sequencing.uio.no), a national technology platform hosted by the University of Oslo and supported by the "Functional Genomics" and "Infrastructure" programs of the Research Council of Norway and the Southeastern Regional Health Authorities.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, which permits any non-commercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. If you remix, transform, or build upon this article or a part thereof, you must distribute your contributions under the same license as the original. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/4.0/.

References

1. Tandishabo K, Nakamura K, Umetsu K, Takamizawa K. Distribution and role of Coprothermobacter spp. in anaerobic digesters. J Biosci Bioeng. 2012;114:518–20.
2. Etchebehere C, Pavan ME, Zorzópulos J, Soubes M, Muxí L. *Coprothermobacter platensis* sp. nov., a new anaerobic proteolytic thermophilic bacterium isolated from an anaerobic mesophilic sludge. Int J Syst Bacteriol. 1998;48:1297–304.
3. Ollivier BM, Mah RA, Ferguson TJ, Boone DR, Garcia JL, Robinson R. Emendation of the Genus *Thermobacteroides*: *Thermobacteroides proteolyticus* sp. nov., a proteolytic acetogen from a methanogenic enrichment. Int J Syst Bacteriol. 1985;35:425–8.
4. Hagen LH, Frank JA, Zamanzadeh M, Eijsink VGH, Pope PB, Horn SJ, et al. Quantitative metaproteomics highlight the metabolic contributions of uncultured phylotypes in a thermophilic anaerobic digester. Appl Environ Microbiol. 2016;83:pii: e01955–16.
5. Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, et al. Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. Science (New Y, NY). 2014;344:416–20.
6. Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, et al. Genomic variation landscape of the human gut microbiome. Nature. 2013;493:45–50.
7. Spanogiannopoulos P, Bess EN, Carmody RN, Turnbaugh PJ. The microbial pharmacists within us: a metagenomic view of xenobiotic metabolism. Nat Rev Microbiol. 2016;14:273–87.
8. Bron PA, Van Baarlen P, Kleerebezem M. Emerging molecular insights into the interaction between probiotics and the host intestinal mucosa. Nat Rev Microbiol. 2012;10:66–78.
9. Ellegaard KM, Engel P. Beyond 16S rRNA community profiling: Intra-species diversity in the gut microbiota. Front Microbiol. 2016;7:1–16.
10. Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ, Polz MF. Resource partitioning and sympatric differentiation among closely related bacterioplankton. Science. 2008;320:1081–5.
11. McLoughlin K, Schluter J, Rakoff-Nahoum S, Smith AL, Foster KR. Host selection of microbiota via differential adhesion. Cell Host Microbe. 2016;19:550–9.
12. Rosenzweig RF, Sharp RR, Treves DS, Adams J. Microbial evolution in a simple unstructured environment: genetic differentiation in *Escherichia coli*. Genetics. 1994;137:903–17.
13. Zelezniak A, Andrejev S, Ponomarova O, Mende DR, Bork P, Patil KR. Metabolic dependencies drive species co-occurrence in diverse microbial communities. Proc Natl Acad Sci USA. 2015;112:6449–54.
14. Rodriguez-Valera F, Martin-Cuadrado A-B, Rodriguez-Brito B, Pašić L, Thingstad TF, Rohwer F, et al. Explaining microbial population genomics through phage predation. Nat Rev Microbiol. 2009;7:828–828.
15. Gill SR, Fouts DE, Archer GL, Mongodin EF, Deboy RT, Ravel J, et al. Insights on evolution of virulence and resistance from the complete genome analysis of an early methicillin-resistant *Staphylococcus aureus* strain and a biofilm-producing methicillin-resistant *Staphylococcus epidermidis* strain. J Bacteriol. 2005;187:2426–38.
16. Sharon I, Morowitz MJ, Thomas BC, Costello EK, Relman DA, Banfield JF. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. Genome Res. 2013;23:111–20.
17. Solheim M, Aakra Å, Snipen LG, Brede DA, Nes IF. Comparative genomics of *Enterococcus faecalis* from healthy Norwegian infants. BMC Genom. 2009;10:1–11.
18. Zunino P, Piccini C, Legnani-Fajardo C. Flagellate and non-flagellate *Proteus mirabilis* in the development of experimental urinary tract infection. Microb Pathog. 1994;16:379–85.
19. Siezen RJ, Tzeneva VA, Castioni A, Wels M, Phan HTK, Rademaker JLW, et al. Phenotypic and genomic diversity of

Lactobacillus plantarum strains isolated from various environmental niches. Environ Microbiol. 2010;12:758–73.

20. Koskella B, Vos M. Adaptation in natural microbial populations. Annu Rev Ecol, Evol, Syst. 2015;46:503–22.
21. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial “pan-genome”. Proc Natl Acad Sci USA. 2005;102:13950–55.
22. Treangen TJ, Rocha EPC. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genet. 2011;7:e1001284.
23. Ochman H, Lawrence JG, Grolsman EA. Lateral gene transfer and the nature of bacterial innovation. Nature. 2000;405:299–304.
24. Bendall ML, Stevens SLR, Chan LK, Malfatti S, Schwientek P, Tremblay J, et al. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. ISME J. 2016;10:1589–601.
25. Biller SJ, Berube PM, Lindell D, Chisholm SW. Prochlorococcus: The structure and function of collective diversity. Nat Rev Microbiol. 2015;13:13–27.
26. Shapiro BJ, Timberlake SC, Szabó G, Polz MF, Alm EJ. Population genomics of early differentiation of bacteria. Science. 2012;336:48–51.
27. Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. Microbial strain-level population structure & genetic diversity from metagenomes. Genome Res. 2017;27:626–38.
28. González-Torres P, Pryszcz LP, Santos F, Martínez-García M, Gabaldón T, Antón J. Interactions between closely related bacterial strains are revealed by deep transcriptome Sequencing. Appl Environ Microbiol. 2015;81:8445–56.
29. Alexiev A, Coil DA, Badger JH, Enticknap J, Ward N, Robb FT, et al. Complete genome sequence of *Coprothermobacter proteolyticus* DSM 5265. Genome Announ. 2014;2: pii: e00470–14.
30. Zamanzadeh M, Hagen LH, Svensson K, Linjordet R, Horn SJ. Anaerobic digestion of food waste - effect of recirculation and temperature on performance and microbiology. Water Res. 2016;96:246–54.
31. Rødsrud G, Lersch M, Sjöde A. History and future of world’s most advanced biorefinery in operation. Biomass Bioenergy. 2012;46:46–59.
32. Kunath BJ, Bremges A, Weimann A, McHardy AC, Pope PB. Metagenomics and CAZyme Discovery. In: Abbott DW, Lammerts van Bueren A, editors. Protein-carbohydrate interactions: methods and protocols. New York, NY: Springer New York; 2017. p. 255–77.
33. Takahashi S, Tomita J, Nishioka K, Hisada T, Nishijima M. Development of a prokaryotic universal primer for simultaneous analysis of Bacteria and Archaea using next-generation sequencing. PLoS ONE. 2014;9:e105592.
34. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26:2460–1.
35. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7:335–6.
36. Stoddard SF, Smith BJ, Hein R, Roller BRK, Schmidt TM. rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. Nucleic Acids Res. 2015;43:D593–8.
37. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal. 2011;17:10–10.
38. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arxiv. 2013;00:1–3.
39. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. MetaSPAdes: a new versatile metagenomic assembler. Genome Res. 2017;27:824–34.
40. Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ. 2015;3:e1165–e1165.
41. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25:1043–55.
42. Hungate RE (1969). Chapter IV A roll tube method for cultivation of strict anaerobes. In: Norris JR, Ribbons DWBTMiM, editors. Methods in microbiology. Chapter IV. Academic Press. p. 117–32.
43. Schumann P. Nucleic acid techniques in bacterial systematics (modern microbiological methods). J Basic Microbiol. 1991;31:479–80.
44. Chen IMA, Markowitz VM, Chu K, Palaniappan K, Szeto E, Pillay M, et al. IMG/M: Integrated genome and metagenome comparative data analysis system. Nucleic Acids Res. 2017;45: D507–16.
45. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. 2014;42:D490–5.
46. Alikhan NF, Petty NK, Ben Zakour NL, Beatson SA. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics. 2011;12:402.
47. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: Rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31:3691–3.
48. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. Nat Microbiol. 2016;1:1–6.
49. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.
50. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol. 2016;33:1870–4.
51. Letunic I, Bork P. Interactive tree of life (iTOL)v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 2016;44:W242–5.
52. Rodriguez-R LM, Konstantinidis KT. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. PeerJ Prepr. 2016;4:e1900–1.
53. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 2011;8:785.
54. Aslanidis C, de Jong PJ. Ligation-independent cloning of PCR products (LIC-PCR). Nucleic Acids Res. 1990;18:6069–74.
55. Miller GL. Use of dinitrosalicylic acid reagent for determination of reducing sugar. Anal Chem. 1959;31:426–8.
56. Bradford MM. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem. 1976;72:248–54.
57. Zhou Y, Pope PB, Li S, Wen B, Tan F, Cheng S, et al. Omics-based interpretation of synergism in a soil-derived cellulose-degrading microbial community. Sci Rep. 2014;4:1–6.
58. Gifford SM, Sharma S, Rinta-Kanto JM, Moran MA. Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. ISME J. 2011;5:461–72.
59. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
60. Kopylova E, Noé L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics. 2012;28:3211–7.
61. Langmead. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.

62. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7.
63. Turro E, Su SY, Gonçalves Â, Coin LJM, Richardson S, Lewin A. Haplotype and isoform specific expression estimation using multimapping RNA-seq reads. Genome Biol. 2011;12:1–15.
64. Turro E, Astle WJ, Tavaré S. Flexible analysis of RNA-seq data using mixed effects models. Bioinformatics. 2014;30:180–8.
65. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics. 2008;24:719–20.
66. Kersters I, Maestrojuan GM, Torck U, Vancanneyt M, Kersters K, Verstraete W. Isolation of *Coprothermobacter proteolyticus* from an anaerobic digest and further characterization of the species. Syst Appl Microbiol. 1994;17:289–95.
67. Johnson JW, Fisher JF, Mobashery S. Bacterial cell wall recycling. Ann New Y Acad. 2013;1277:54–75.
68. Lü F, Bize A, Guillot A, Monnet V, Madigou C, Chapleur O, et al. Metaproteomics of cellulose methanisation under thermophilic conditions reveals a surprisingly high proteolytic activity. ISME J. 2014;8:88–102.
69. Miller JH, Novak JT, Knocke WR, Pruden A. Survival of antibiotic resistant bacteria and horizontal gene transfer control antibiotic resistance gene content in anaerobic digesters. Front Microbiol. 2016;7:1–11.
70. Hehemann JH, Correc G, Barbeyron T, Helbert W, Czjzek M, Michel G. Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. Nature. 2010;464:908–12.
71. Ricard G, McEwan NR, Dutilh BE, Jouany JP, Macheboeuf D, Mitsumori M, et al. Horizontal gene transfer from bacteria to rumen ciliates indicates adaptation to their anaerobic, carbohydrates-rich environment. BMC Genomics. 2006;7:1–13.
72. Song T, Xu H, Wei C, Jiang T, Qin S, Zhang W, et al. Horizontal transfer of a novel soil agarase gene from marine bacteria to soil bacteria via human microbiota. Sci Rep. 2016;6:1–10.
73. Modi SR, Lee HH, Spina CS, Collins JJ. Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. Nature. 2013;499:219–22.
74. Zverlov VV, Schantz N, Schmitt-Kopplin P, Schwarz WH. Two new major subunits in the cellusome of *Clostridium thermocellum*: xyloglucanase Xgh74A and endoxylanase Xyn10D. Microbiology. 2005b;151:3395–401.
75. Demain AL, Newcomb M, Wu JHD, Demain AL, Newcomb M, Wu JHD. Cellulase, clostridia, and ethanol. Microbiol Mol Biol Rev. 2005;69:124–54.
76. Zverlov VV, Kellermann J, Schwarz WH. Functional subgenomics of *Clostridium thermocellum* cellulosomal genes: identification of the major catalytic components in the extracellular complex and detection of three new enzymes. Proteomics. 2005a;5:3646–53.
77. Chylenski P, Petrović DM, Müller G, Dahlström M, Bengtsson O, Lersch M, et al. Enzymatic degradation of sulfite-pulped softwoods and the role of LPMOs. Biotechnol Biofuels. 2017; 10:1–13.
78. Fuchs K-P, Zverlov VV, Velikodvorskaya GA, Lottspeich F, Schwarz WH. Lic16A of *Clostridium thermocellum*, a non-cellulosomal, highly complex endo-β-1,3-glucanase bound to the outer cell surface. Microbiology. 2003;149:1021–31.
79. Jeng W-Y, Wang N-C, Lin C-T, Shyur L-F, Wang AHJ. Crystal structures of the laminarinase catalytic domain from *Thermotoga maritima* MSB8 in complex with inhibitors: essential residues for β-1,3 and β-1,4 glucan selection. J Biol Chem. 2011;286: 45030–40.

Paper II

Integration of absolute multi-omics reveals translational and metabolic interplay in mixed-kingdom microbiomes.

F. Delogu[1], B.J. Kunath[1,2], P.N. Evans[3], M.Ø. Arntzen[1], T.R. Hvidsten[1], P.B. Pope[1,4]

[1] Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, 1432 Ås, Norway

[2] Luxembourg Centre for Systems Biomedicine, Université du Luxembourg, L-4362 Esch-sur-Alzette, Luxembourg

[3] Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, University of Queensland, St Lucia, Queensland, Australia

[4] Faculty of Biosciences, Norwegian University of Life Sciences, 1432 Ås, Norway

Corresponding authors:

Phillip B. Pope phil.pope@nmbu.no

Francesco Delogu delogu.francesco@nmbu.no

Abstract

Microbiology is founded on the study of well-known model organisms. For example, the majority of our fundamental knowledge regarding the quantitative levels of DNA, RNA, and protein backdates to keystone culture-based studies. Nowadays, meta-omic approaches allow us to directly access the molecules that constitute microorganisms and microbial communities, however due to a lack of absolute measurements, many original culture-derived "microbiology statutes" have not been updated or adapted to more complex microbiome settings. Within a cellulose-degrading and methanogenic consortium, we temporally measured genome-centric absolute RNA and protein levels per gene and obtained a protein-to-RNA ratio of 10^2-10^4 for bacterial populations. In contrast, Archaeal RNA/protein dynamics (10^3-10^5: *Methanothermobacter thermoautotrophicus*) were more comparable to Eukaryotic representatives humans and yeast. The linearity between transcriptome and proteome had a population-specific change over time, highlighting a

minimal subset of four functional guilds (cellulose degrader, fermenters, syntrophic acetate-oxidizer and methanogen) that coordinated their respective metabolisms, cumulating in the overarching community phenotype of converting polysaccharides to methane. Our findings show that upgrading multi-omic toolkits with traditional absolute measurements unlocks the scaling of core biological questions to dynamic and complex microbiomes, creating a deeper insight into inter-organismal relationships that drive the greater community function.

Introduction

The foundations of microbiology have been built within the constrained framework of pure culture studies of model organisms that are grown under controlled steady state conditions. However, we are constantly told that microorganisms grown in single culture behave in a different manner to those in mixed natural communities. For example, when *Escherichia coli* is grown axenically in steady state*,* we can expect that each RNA molecule results in 10^2 to 10^4 of the corresponding protein (protein-to-RNA ratio) and the variation in the level of cellular RNA explains ~29% of the variation in the amount of detectable protein[1]. Yet does this notion hold true when a given bacterial population is part of a larger community and subject to transitions from one state of equilibrium to another due to limiting and/or confronting environmental factors? In this context, the exploration of temporal interplay between populations with different lifestyles (comprising metabolism, motility, sporulation, etc.) becomes of primary importance to interpret the changes in fundamental quantities in a microbial community, such as the protein-to-RNA ratio that ultimately impacts the overarching community phenotype(s). In order to perform studies of such design and test if previously defined quantitative data about the functioning of microorganisms (i.e. protein-to-RNA ratio) is applicable to real world consortia, we must first sample microbial communities across transition events and employ quantification techniques that are absolute.

Meta-omics techniques, such as metagenomics (MG)[2,3], metatranscriptomics (MT)[4] and metaproteomics (MP)[5] are routinely used to assess prokaryotes in the natural world, where they are part of communities that are frequently dominated by as-yet uncultivated populations[6]. The quantities retrieved from the meta-omics are usually expressed in relative terms, which makes comparison between samples and between omic layers inaccurate[7,8]. Moreover, within dynamic data measurements, such as the MT or MP, the notion of steady state becomes relevant as it is extremely rare that parameters (e.g. bacterial growth rate and nutrient availability) are stable over time[8].

Here, we present an absolute temporal multi-omic analysis of a minimalistic cellulose-degrading and methane-producing consortium (SEM1b), which was resolved at the strain level and augmented with two strain isolates[9]. We combined both a RNA-spike-in for

MT[10,11] and the *total protein approach* for MP[12] for the absolute quantification of high-throughput data. We not only demonstrate that temporal SEM1b samples were comparable within the same omic layer, but also between the MT and MP. Indeed, the protein-to-RNA ratio per sample of the bacterial populations matched previous calculations for the existing example from axenically cultured *E.coli*[1]. For the first time, we present protein-to-RNA ratios for an archaeon (*Methanothermobacter thermoautotrophicus*), which are similar to those reported for the Eukarya, and support crystallography and homology studies that suggest the translation system of archaea more closely resembles eukaryotes[13]. Our approach enabled us to explore the linearity of the protein-to-RNA ratio and if it is influenced by changes in community state and/or specific population lifestyle. Finally, we estimated the translation and protein degradation rates, showing that a downregulation of the former marks main lifestyle changes (e.g. motility/chemotaxis and metabolism) during the community development.

Results and Discussion

Taxonomic and functional resolution of the omics

In order to characterize RNA/protein dynamics in a microbiome setting, we first needed to reconstruct our test community over time at the molecular level. Previous analysis of the simplistic SEM1b community genomically reconstructed and resolved 11 metagenome assembled genomes (MAGs) as well as two isolate genomes[9], covering the taxonomic and functional niches that are required to convert cellulosic material to methane/CO_2 in an anaerobic biogas reactor[14]. Taxonomic analysis of SEM1b inferred population-level affiliations to *Rumini(Clostridium) thermocellum* (RCLO1), *Clostridium sp.* (CLOS1), *Coprothermobacter proteolyticus* (COPR1, BWF2A, SW3C), *Tepidanaerobacter* (TEPI1-2), *Synergistales* (SYNG1-2), *Tissierellales* (TISS1), and the methanogen *Methanothermobacter* (METH1)[9]. Herein we estimated that the total genomic potential of SEM1b includes 39144 Open Reading Frames (ORFs) (Supplementary Dataset 1). Since ORFs with very high sequence similarity may produce RNAs and proteins that are indistinguishable in MT and MP data, we instead gathered all ORFs into ORF-groups (ORFGs), where a singleton ORFG is defined as a group with a single ORF, and thus a single gene. Using this approach, our MT and MP data identified 12552 (96% singleton) and

3235 (78% singletons) highly transcribed and translated ORFGs, respectively. The discrepancy between the singleton percentages was as expected, due to the fact that variations in the DNA/RNA sequences are expected to be greater than in the protein since different codons can code for the same amino acid (codon degeneracy). Degeneracy implies that the chance to distinguish between homologous genes using MT is greater than using MP. Previous MG analyses using assembly algorithms has shown that problematic genomic regions in a given environmental contig can harbor variants from multiple, closely-related strains, which can be further linked to normal strain-level variability within a population and speciation[15–17]. Within SEM1b, the ORFGs that contained multiple homologous ORFs predominantly originated from several strains of a single species. For example, in the MT, 444 non-singleton ORFGs (88% of the total) contained ORFs from different strains of the same species, whilst this was the case for 294 ORFGs (32%) in the MP.

All ORFs were annotated using Kegg Ontology (KO), and at least one term was found for 19070 (49%) representatives from our complete dataset (Supplementary Dataset 2). The predominant ORF annotations included *Membrane transport*, *Carbohydrate metabolism*, *Translation*, *Amino acid metabolism* and *Replication and repair* (Supplementary Fig. 1). As expected, these functional categories were also among the top five most abundant for the MT, and top six in MP (plus *Energy metabolism*), although in a different order. The *Membrane transport* category is extremely poorly represented in the MP (2% of the terms), which is likely explained by well-known technical issues that limit the extraction of transmembrane proteins[18]. The most abundant annotation categories mentioned above are all in line with the community function of cellulose degradation. The abundance ranking of the KO categories changes slightly from MG to MT (Kendall τ: 0.77, $p<10^{-8}$) and from MT to MP (τ 0.74, $p<10^{-6}$) whilst moderately from MG to MP (τ 0.68, $p<10^{-5}$), which means that the functional potential observed in the genomes is more preserved in the diversity of produced transcripts than the one of proteins and thus hints to post-transcriptional regulation playing an important role in addition to transcriptional regulation in prokaryotes.

Absolute quantification extends expectation from *E.coli* RNA/protein dynamics and positions Archaea alongside the Eukarya

To determine whether or not microbial RNA/protein dynamics vary between ecological status (isolate vs community), metabolic states and/or taxonomic phylogeny, we quantified and resolved the numbers of transcript and protein molecules per sample in our SEM1b community, which averaged 3.8×10^{12} (sd 3.0×10^{12}) and 2.2×10^{15}(sd 9.5×10^{14}), respectively (Supplementary Datasets 3-4). Microbial cell volume and its transcriptome size has been shown to change in yeast according to cell status (proliferation vs. quiescence), whilst the proteome is merely reshaped in its composition between these states[19]. In our case, the number of total transcripts per sample increased more than three-fold during the first 15 hours (from ~1.2×10^{12} in t1 to ~4.0×10^{13} in t4) in the SEM1b consortium's life cycle and then decreased sharply, whereas the number of proteins per sample reached a plateau after 18 hours post-inoculation at ~2.7×10^{15} molecules. SEM1b approximated the exponential growth phase in t3 (18 hours), therefore we used the protein-to-RNA ratio from this time point for comparison against previously reported axenic estimates[1,20–23]. The replicate-averaged protein-to-RNA ratio for the bacteria in SEM1b ranges from ~10^2 to 10^4 (median = 949, Fig. 1a), which fits the estimated range reported for *E.coli*[1]. This means that for every RNA molecule one can expect from 100 to 10000 protein molecules with a value of 949 being the most likely. Our results showed a population-specific variation in the protein-to-RNA ratio within bacteria (Fig. 1a), with the median ratios for the bacteria in SEM1b at 18h ranging from 658 in CLOS1 to 1137 in RCLO1. While the limited number of published studies and data that enable estimation of protein-to-RNA ratios prevented our assessment of higher-resolution taxon-specific distributions within Bacteria, clear patterns were observed at a broader Domain level and are presented below (Fig. 1a).

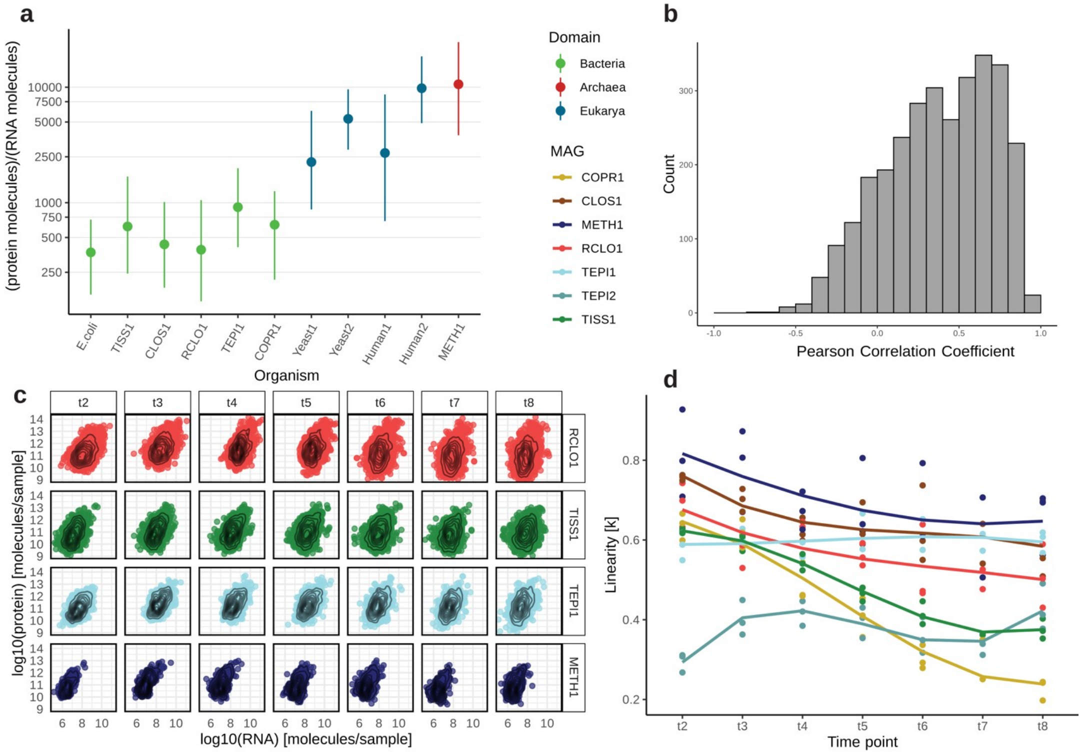


Figure 1. Protein-to-RNA ratio distributions of as-yet uncultured bacterial and archaeal populations within a microbial community. a. Comparison of protein-to-RNA ratio distributions of selected MAGs reconstructed from the SEM1b community as well as those previously reported in the literature. The dots represent the median values and the bars span from the first to the third quartiles (Bacteria: green, Archaea: red, Eukarya: blue). The protein-to-RNA ratios for *E.coli* was retrieved from Taniguchi et al.[1], Yeast1 from Ghaemmaghami et al.[20], Yeast2 from Lu et al.[21], Human1 from Schwanhausser et al.[22] and Human2 from Li et al.[23]. **b.** The distribution of the Pearson Correlation Coefficients (PCC) between transcripts and their corresponding proteins computed across the time points. With a median PCC of 0.41, the change in the amount of a given transcript over time seemingly does not translate into a change in the amount of the corresponding protein. **c.** Per-time-point scatterplots of the absolute protein and transcript levels for ORFs that produced both detectable transcript and protein in SEM1b datasets. For simplicity, only four representative MAGs are shown, with all MAGs depicted in Supplementary Fig. 2. **d.** The plot shows how the linearity parameter *k* between RNA and protein changes over time for the different MAGs. The linearity represents how a change in RNA level is reflected in a change in the corresponding protein level. The parameter ranges from 0 to 1, and increasingly smaller values translate in fewer proteins being expected for the same level of RNAs. The populations CLOS1, METH1 and TEPI1 are converging towards the same values, while RCLO1 has a parallel trend. Hinting to the existence, and the reaching of an equilibrium among them.

In contrast to bacterial protein-to-RNA ratios that were relatively comparable to one another, the median protein-to-RNA ratio for an Archaeal organism, which we report herein for the first time, was approximately 10x higher at 12035 protein molecules per detected

RNA (Fig. 1a: METH1). The reported values for Eukaryotes are 4200-5600 for yeast[20,21] and 2800-9800 for *Homo sapiens*[22,23]; therefore, we find that Archaeal translation dynamics are closer to that observed within the Eukaryotic kingdom than that of Bacteria. Structurally, the translation system of archaea more closely resembles eukaryotes[13]. Moreover, the RNA of Eukarya and Archaea have been shown to exhibit longer half-lives than Bacteria[24,25], with Archaea found to contain a novel triphosphate structure at the 5' end of the RNA molecule that is involved with mRNA stability[26]. Also, like Eukaryotes, it has been shown that archaeal RNA is regulated by post-translational modification of the RNA molecule in order to up-regulate protein expression[27,28]. Findings that show transcripts are present in archaeal cells for longer than bacterial cells can be used to hypothesize that this feature could play a greater role in optimizing efficient production of protein molecules. In a microbiome-setting, the greater turnover of RNA molecules and lower protein-RNA ratio in bacteria could potentially facilitate their faster adaption to changes in metabolic state and substrate availabilities in their environment, at higher rates than their archaeal counterparts. However, in many complex microbiome's archaea occupy highly specialized niches such as the biological production of methane via methanogenesis, which is the energy-yielding metabolism of methanogens and is unique to the Archaea. In this context, proteins involved with hydrogenotrophic methanogenesis have been shown to be the most highly detectable in methanogens grown in co-culture with syntrophic acetate oxidizing (SAO) bacteria, when compared to the same methanogen grown in axenic culture with higher concentrations of supplemented H_2[29]. This discrepancy between H_2 supply and protein levels suggests there is a requirement for methanogens to maintain highly active protein expression levels in order to keep H_2 at levels that are low enough to keep SAO energetically favourable[30]. We therefore speculate that methanogens, via their molecular mechanisms of maintaining high protein levels, are at an advantage to stably and efficiently maintain low H_2-levels, a process that is critical to the metabolic equilibrium of many microbial ecosystems[31].

In axenic culture, a microorganism is considered to be in steady state during the log phase of its growth cycle[8,32,33], specifically when the changes in proteome size are believed to be mainly dictated by a change in the transcriptome[34]. In contrary to these assumptions, comparisons of RNA and protein levels between single cells of *E. coli* grown at steady state have not been shown to correlate, however patterns do emerge when the cells are considered at the population level[1]. In SEM1b, we wanted to see if correlations between RNA and protein levels exist in a larger microbial community, and if they are affected by changes in time and life stages. We calculated gene-wise Pearson Correlation Coefficients (PCCs) of protein and transcripts over time for all SEM1b populations and showed that the PCC value varied greatly (Fig. 1b) with a median of 0.41, suggesting that no direct correlations between RNA and proteins levels exist at any stage in a microbiome and that it is nearly impossible to predict the level of the given protein based on the level of the corresponding transcript.

Looking at relationships between proteome and transcriptome for individual populations within SEM1b (examples form four populations in Fig. 1c) was observed to follow a more predicable relationship, which can be described by the monomial function:

$$protein = a \cdot RNA^{k} \text{(eq.1)}$$

The formula for log10-transformed RNA and protein levels takes the form of a linear model (see methods) that was fitted to protein and RNA distributions per time point from MAGs with the highest quality (RCLO1, CLOS1, COPR1, TISS1, TEPI1, TEPI2 and METH1) (Fig. 1d). The linearity parameter k can be interpreted as the rate of which a change in RNA level is reflected in a change in the corresponding protein level. For example, if k=1, a doubling in RNA level means a doubling in protein level, whereas if k=0.5 a doubling in RNA level means a ~40% increase in protein level. Ranging from 0 to 1, it implies that, in the "perfect" condition where k=1, the number of proteins is linked to the number of RNAs by the scalar constant *a*, whilst if k approaches 0, there will be much lower expected protein levels for the same number of RNAs. With the exception of TEPI2, the linearity (*k*) between protein and RNA levels was observed to start at values between 0.6 and 0.8 at 13 hours (t2) (Fig. 1d). The evolution of the MAGs' *k* values over time is then divided in three groups: one which is losing linearity rapidly (TISS1 and COPR1); one which is slowly declining (RCLO1, CLOS1 and METH1) and one which is staying constant if not increasing (TEPI1 and TEPI2) (Fig. 1d). Notably CLOS1, METH1 and TEPI1 are converging towards the same linearity values, while RCLO1 has a parallel trend to them. If these trends can be used to retro-fit the steady state definition, we can hypothesize that these four populations possess a metabolic equilibrium and that this equilibrium is approximately reached within the 10 hour window between 33h and 43h (t6 and t7 respectively, Fig. 1d).

Interpretation of functional specialization in the light of RNA-protein dynamics

Using multi-omic data and the above described RNA-protein dynamics, we were able to visualize that at least four populations within SEM1b converge upon a dominant metabolic state that we speculate to strongly shape the overall SEM1b community phenotype and suggest a functional co-dependence between the individual populations. To determine if this was the case, we annotated the genes and metabolic pathways for SEM1b MAGs (Fig. 2) and reconstructed their temporal expression patterns (Fig. 3). The SEM1b consortium is able to convert cellulose (and hemicellulose) to methane via the combined metabolism of its seven major constituent populations (Fig. 3a). Based on previous analysis that showed that RCLO1 is closely related to *R. thermocellum*[9], we predict that it senses[35] its growth substrate (cellulose) and moves towards it (Fig. 3d). RCLOS1 then transcribes, translates and secretes the components of the cellulosome, such as scaffoldins, dockerins and carbohydrate-active enzymes (CAZymes)[36], which assemble into a dynamic multi-proteins complex that degrades the substrate to smaller carbohydrates. Via the MG, we predicted that non-cellulosomal CAZymes were also employed by the *Clostridium*-affiliated CLOS1, which acted upon the hemicellulose fraction (mainly xylan) trapped in the spruce cellulose, which was supported by observed release of its main monomer xylose (Fig. 3a). Sugars generated via the actions of RCLO1 and CLOS1 are subsequently consumed by RCLO1, CLOS1 and *Coprothermobacter*-affiliated populations (COPR1, BWF2A and SW3C), which were all observed to express sugar transporters (Fig. 2). Notably CLOS1 has the most diversified transporters, making it a flexible consumer, and for the most part demonstrated highest levels of hydrolytic and fermentative gene expression after RCLO1, which again is likely tied to xylose release at later stages of the SEM1b lifecycle (Fig. 3a). However, some of the transporters, such as the one for oligogalacturonide, raffinose/stachyose/melibiose and rhamnose, were not expressed, likely due to the absence of their substrates in the largely cellulose and xylan dominated spruce wood used in this study. CLOS1 was also the only population to possess the aldouronate transporter with 20 copies of gene lplA, 20 of lplB and 16 of lplC (20/20/16) and expressing 0.4/0.7/3.8×10^{10} and 92.8/3.5/7.0/×10^{11} combined median transcripts and proteins per sample; making it one of the few transporters detectable at the protein level. Similarly, the *C. proteolyticus* strains (BWF2A and SW3C) possess and express unique sugar transporters, likely gaining access to an undisputed pool of arabinogalactan or maltooligosaccharide. The transporter for pentamers ribose/xylose were the most common and possessed by RCLO1, *C. proteolyticus* populations and *Tepidanaerobacter* populations (TEPI1 and TEPI2). Notably from Fig. 2, it is clear that the proteins from the transporters are almost never found in the samples, even if the respective RNAs are abundant. This is likely due to the difficulties in extracting transmembrane proteins[18].

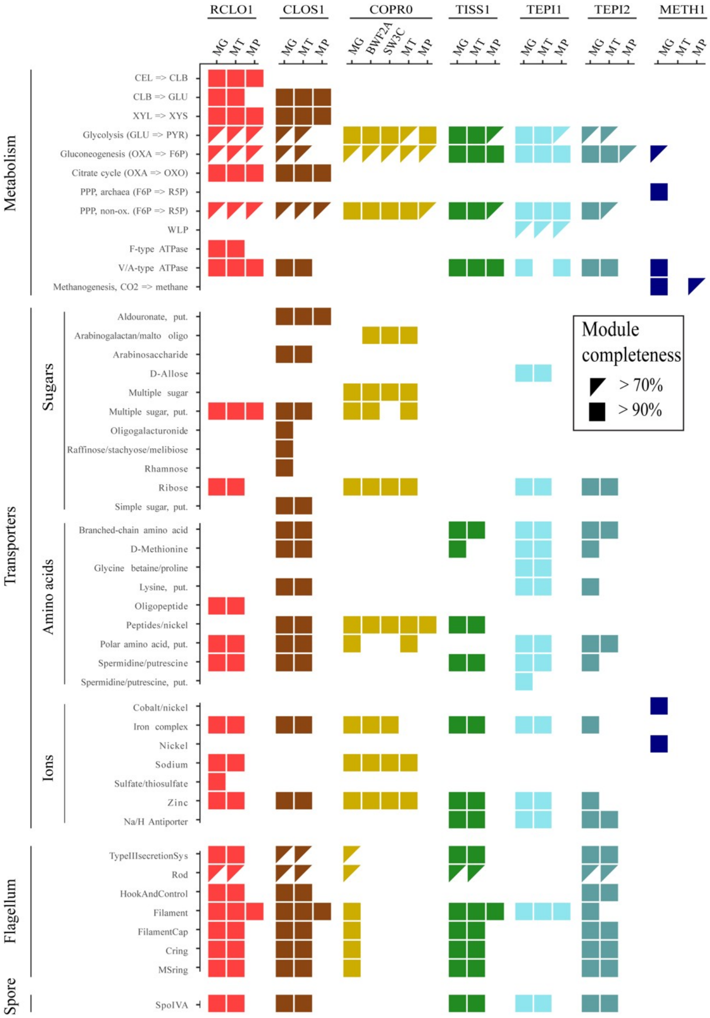


Figure 2. Overview of the genetic potential and expressed modules in the seven populations of SEM1b. Module completeness denotes the level of detected RNA and proteins mapped to major genes/metabolic pathways that are critical to the SEM1b lifecycle. Only MAGs with the highest quality reconstruction (RCLO1, CLOS1, COPR1, TISS1, TEPI1, TEPI2 and METH1) are included as well as two isolated and genome-sequenced *Coprothermobacter* strains, for which the transcriptome and the proteome were considered as the species level.

The process of degrading cellulose and simple saccharides via hydrolysis and fermentation ultimately results in the production of short chain fatty acids (SCFAs) such as proprionate, butyrate and acetate, which are subsequently metabolized by the SCFA-oxidizing populations in SEM1b (TISS1, TEPI1, TEPI2) (Fig. 3a). The only metabolically-active SCFA-oxidizing population in SEM1b was predicted to be TEPI1, as it demonstrated good linearity between protein and RNA levels that increased over time (Fig. 1d) and harbored a complete Wood-Ljungdahl carbon fixation Pathway (WLP) that was detectable in both MT and MP (Fig. 2). It has been shown that oxidizers can improve oxidization of SCFAs (up to double speed) when superior NADPH and ATP generators (e.g. glucose) are consumed in small amounts to complement the stoichiometry through the Pentose Phosphate Pathway (PPP) without triggering the shift of the entire cells metabolism toward another substrate[37]. In this context, it is interesting to note that TEPI1 was the only MAG that encoded and expressed a hexose (allose) transporter (Fig. 2). Aldohexoses (such as D-allose, D-glucose, D-mannose, etc.) are imported and transformed into fructose-6P in two reactions (both expressed in TEPI1), which can then be fed into both the PPP or the Glycolysis pathways. Xylose, is a product of the degradation of hemicellulose present in our system (Fig. 3a) and can be converted to ribulose-5P and fed to the PPP in three reactions. This data, in combination with a highly expressed and detectable WLP over time (Fig. 3a), points to the establishment of TEPI1 as the only SAO bacteria in the SEM1b consortium. We speculate that TEPI1's SAO-metabolism is helped by the other SEM1b populations that generate acetate as a fermentation end-product and the supplement of the sugars released by the cellulosomal complex such as glucose and xylose. Interestingly the closely related MAG TEPI2 was observed to lack the WLP and to express ~10 times more transcripts for the ribose/xylose transporter than TEPI1; relegating it to the role of mere sugar degrader, and probably scavenger in the community.

While TISS1 seems mostly to phase out of the community and lose linearity in its protein to transcript relationship (Fig. 1d), TEPI2 implements an exit strategy in the form of sporulation. All the gram-positive populations from the SEM1b consortium (RCLO1, CLOS1, TISS1, TEPI1 and TEPI2) were able to produce spores and express the spore marker *spoIV*, an ATPase associated to the surface of the neospore that promotes the assembly of the coating and is common to all the spore forming bacteria[38] (Fig. 3b). TEPI2 however increased the level of transcripts for spoIV by 1000 times within the 13h and the 18h time points, reaching the maximum at 23h, and having a production 10 times higher than the phylogenetically related TEPI1. All SEM1b populations, except the *C. proteolyticus* isolates and TEPI1, have the genetic potential for flagellar synthesis but the respective transcripts were only observed for RCLO1, CLOS1, TISS1 and TEPI2. The filament protein of RCLO1 is by far the most abundant protein in the samples with an average of 2.8×10^{13} molecules per sample, which matches the idea of RCLO1 investing in motility to reach the cellulose fibers and starting with the highest level of marker flgD in the community (Fig. 3d).

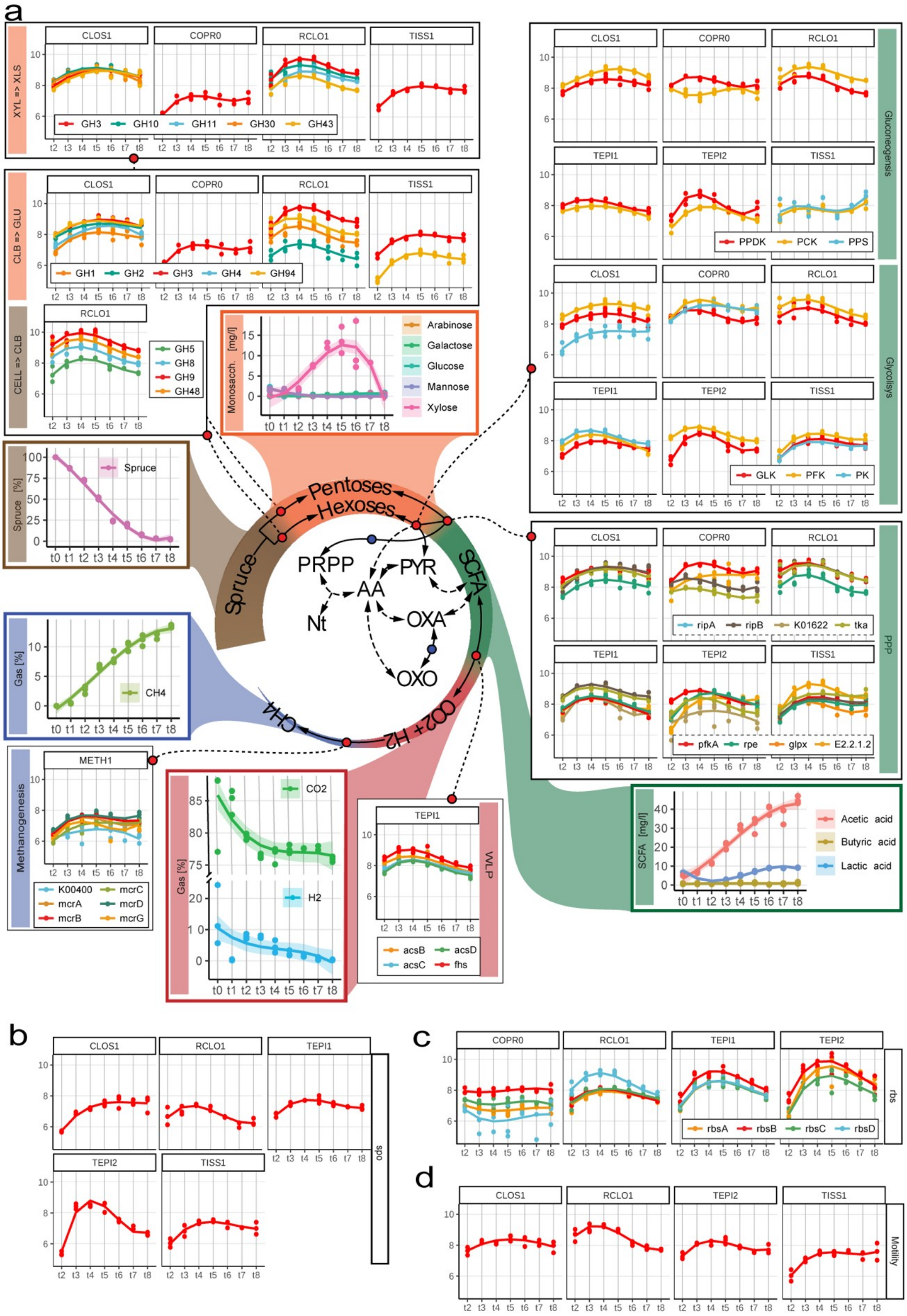

a
XYL => XLS
CLB => GLU
CELL => CLB
CLOS1
COPR0
RCLO1
TISS1
TEPI1
TEPI2
METH1
GH3
GH10
GH11
GH30
GH43
GH1
GH2
GH4
GH94
GH5
GH8
GH9
GH48
Monosacch. [mg/l]
Arabinose
Galactose
Glucose
Mannose
Xylose
Spruce [%]
Spruce
Gas [%]
CH4
CO2
H2
Methanogenesis
K00400
mcrA
mcrB
mcrC
mcrD
mcrG
WLP
acsB
acsC
acsD
fhs
Pentoses
Hexoses
Spruce
SCFA
CO2+ H2
CH4
PRPP
PYR
AA
Nt
OXA
OXO
Gluconeogenesis
PPDK
PCK
PPS
Glycolisys
GLK
PFK
PK
PPP
ripA
ripB
K01622
tka
pfkA
rpe
glpx
E2.2.1.2
SCFA [mg/l]
Acetic acid
Butyric acid
Lactic acid
t0 t1 t2 t3 t4 t5 t6 t7 t8
b
spo
c
rbs
rbsA
rbsB
rbsC
rbsD
d
Motility

Figure 3. Schematic representation of the temporal and co-dependent metabolism of SEM1b that converts spruce-derived cellulose to methane. **a.** Within SEM1b, four major metabolic stages are required: Spruce → Hexoses/Pentoses, Hexoses/Pentoses → SCFAs, SCFAs → CO_2+H_2 and CO_2+H_2 → methane. Metabolites (spruce, sugars, SCFAs, CO_2+H_2 and methane) involved in these processes were measured and the temporal analysis of the metabolic pathways involved in their interconversion is depicted for the major SEM1b populations. Other metabolites (for which abbreviations are: Nt=Nucleotides, PRPP=Phosphoribosyl pyrophosphate, AA=Amino acids, PYR=Pyruvate, OXA=Oxaloacetate and OXO=Oxo-glutarate) are shown to highlight the essential metabolism of the microbes. In the central metabolic network the metabolites are linked by solid arrows if the interconversion requires one step or the link between them is addressed more in detail (blue dot if in Fig. 2, red dot if in a pathway plot herein). Metabolic pathways are quantified via marker genes (selection in methods section) in the scale of log10-transformed transcript molecules per sample whilst the solid lines in the plots represent the qubic fitting of the data points. More metabolites' abbreviations are CELL=Cellulose, CLB=Cellobiose, GLU=Glucose, XYL=Xylan, XLB=Xylobiose and pathways' abbreviations are WLP="Wood-Ljungdahl Pathway", PPP="Pentose Phosphate Pathway". **b.** Sporulation is common to all Gram positive bacteria of the community and it is quantified with the marker *spoIVA*. Notably TEPI2 is investing greatly in spore formation until 28h after the inoculum (t4). **c.** The genes for the Ribose and xylose transporter (*rbs*) are expressed in four populations. Notably TEPI2 produces more rbs transcripts than the closely related MAG TEPI1; indeed, the first has been predicted to be a mere fermenter whilst the latter bases its metabolism on the WLP pathway (Fig. 3a). **d.** Microbial motility is represented by the marker gene *flgD*. RCLO1 is the most active bacterium, producing less and less flagella over time after t4. It starts ahead of the others at t2, presumably finishing the colonization of the substrate (Spruce-derived cellulose).

In microbial ecosystems, acetate is oxidized by secondary fermenters to CO_2/H_2 or formate, a process that is mediated by the WLP in reverse. The oxidation of acetate associated with the reverse WLP is coupled with the transition between NADH/NAD$^+$, and translocates Na$^+$ to create an electrochemical gradient, which is then used by the type-V ATPase to synthesize ATP[39]. Indeed the NAD$^+$-Fd$_{red}$-dependent Na$^+$ translocation system *rnf* is expressed in both the fermenting and SAO bacteria of SEM1b, while type-V ATPase, which produce energy by exploiting the Na$^+$/H$^+$ gradient, were detected by all the SEM1b populations aside from METH1 and *C. proteolyticus*-affiliated populations (**Fig. 2**). Moreover, the TEPI1 MAG expresses the NAD$^+$ (NADP$^+$)-reducing hydrogenases complex, which reduces hydrogen ions to H_2 using NAD(P)H as the electron donor. The molecular hydrogen generated here would then be used by the syntrophic partner METH1 to form methane (**Fig. 3a**). However, this reverse WLP-mediated acetate oxidation is thermodynamically unfavourable unless coupled with syntrophic hydrogenotrophs. Within SEM1b, the METH1 population is a hydrogenotrophic methanogen and the methanogenesis pathway, which is observed in the METH1 MG and MP, is the largest pathway in SEM1b according to the number of genes involved (n=112). In METH1, we also observed transporters for nickel, the metal ion found in the F_{430} prosthetic group in the methyl-coenzyme M reductase complex (McrABG), which is responsible for the terminal step in anaerobic methanogenesis[40]. Transporters for another key metal, cobalt, which is utilized by cobalamide-requiring enzymes such as the energy conserving methyl-H_4MPT:CoM-SH methyltransferase complex (MtrABCDEFGH), were also detected in the MG and MP of METH1. Within hydrogenotrophic methanogens, electrochemical gradients generated Na^+ ion exclusion by the Mtr complex allows for the inflow of Na^+ through ATP synthases to generate energy. Surprisingly, no H^+/Na^+ *nha* ion transporter which are commonly observed in methanogens were observed in the METH1 MG, MT or MP. Only in the populations TEPI1, TEPI2 and TISS1 were the Na^+/H^+ antiporter *nha* encoded and expressed (**Fig. 2**), which does point to an important role of these ions in the bacterial component of the SEM1b consortium. Overall, our more classical pathway-wise exploration of the SEM1b populations supported that RCLO1, CLOS1, TEPI1 and METH1 indeed share functional co-dependencies and supported our predictions via protein-RNA dynamics that they converge upon a dominant metabolic state.

Translation control drives changes in cell status and source utilization

In addition to RNA/proteins ratio assessments, our collection of absolute multi-omic data allowed us to explore the crucial aspect of protein-level regulation, which is poorly understood in microbiomes. The control of protein levels in bacteria is believed to occur predominately via transcription control, “control by dilution”[41] (dispersal of proteins via subsequent cell divisions) and rarely by protein degradation[42]. Similar to transcription control, translation can also be controlled by a dynamic pool of translational factors, such as initiation, elongation and ribosome components[43]. The processes targeted by these systems require a rapid change in the number of proteins in the cell that cannot wait for a change in RNA levels or a dilution effect. The absolute quantification of transcripts in SEM1b and proteins was used to estimate the translation and protein degradation rates using PECA-R[44] (Supplementary Dataset 5). The analysis found 305 significant changes in translation rate, accounting for 302 ORFs. Of the rate changes’, 94% were downregulated and 71% of the ORF were functionally annotated. RCLO1 has 28 downregulated ORFs between 13h and 18h (t2-t3), mostly from complexes involved in chemotaxis (*cheY*, *cheW*, *mcp*), flagellum assembly (*flgG*, *flgK*, *fliD*) and shape determination (*mreB*). In the following five hours, several systems concerning carbon fixation are affected, such as phosphoglycerate kinase (PGK), triosephosphate isomerase (TPI), phosphate acetyltransferase (EC 2.3.1.8), isocitrate dehydrogenase (IDH1) and *pyruvate* orthophosphate dikinase (PPDK). In the next five hours it downregulates the translation of the cell division protein ZapA as well. The reduction protein production for chemotaxis, mobility and then cell division matches the idea that within 13h of the inoculation, RCLO1 sensed, reached and colonized the cellulose fibers. Contextually the release of medium length carbohydrates enables RCLO1 to engage in the more energetically favorable fermentation metabolism. TISS1 has a decrease in translation rates of ORFs related to metabolic processes between 13h and 18h, mostly involving cofactors (*fhs*, *folC*, *folD*, *lplA*, *metH*, *pdu0* and *nadE*) and amino acids (*aorQ*, *hutI*, LDH, *metH*, *mtaD* and *pip*). TEPI1 down-regulated 60 ORFs, accounting for part of its carbohydrate metabolism (e.g. PGK, TPI), the amino acid transporters and the NADH dehydrogenase complex (HND). TEPI2 has 19 ORFs subject to downregulation in the 13h-18h interval, such as Pyruvate ferrodoxin odidoreductase (PFOR), GK, fructose-bisphosphate aldolase (FBA), tansaldolase EC 2.2.1.2 and the ribose/xylose transporter subunit *rbsB*. In the last interval (33h-38h), RCLO1 upregulated the translation of 10 ORFs, among which the flagellar FlbD and shape determination MreB; seemingly starting to restore the functions downregulated in the 13h-18h interval.

Conclusions

We present the reconstruction of a microbiome from a model environment and quantified the number of RNAs and proteins over time in absolute terms. This approach enabled us to assess and report, for the first time, the protein-to-RNA ratio of multiple microbial populations simultaneously, which individually engage in distinct, yet integrative metabolic pathways that ultimately cumulate into the community's principal phenotype of converting cellulose to methane. We extended the results from Taniguchi et al.[1], showing that our populations had a varying protein-to-RNA ratio in the predicted interval of 10^2-10^4 while presenting for the first time the same quantity for an Archaeal population (METH1): 10^3-10^5, which resembled the previously measured values for Eukaryotes[20–23]. The greater ecological significance of the seeming Archaeal capacity to generate higher protein levels at a lower "RNA-cost" is of interest, as many Archaeal populations in mixed-kingdom microbiomes are known to occupy essential ecological niches and exert strong functional influence, despite their cell concentrations being orders of magnitude lower than their bacterial counterparts (i.e. methanogens in the rumen microbiome[45]).

Moreover, we assessed the linearity between transcriptome and proteome for each population over time (Eq.1), finding that three major populations of the community, a fermenter (CLOS1), a SAO bacteria (TEPI1) and a methanogen (METH1), were converging on the same values in parallel with the primary cellulose degrader (RCLO1) (Fig. 1d). The highlight of their seemingly intertwined protein/RNA dynamics matches with their metabolic complementarity, starting from RCLO1 degrading cellulose to sugars and SCFAs, CLOS1 fermenting sugars to SCFA, TEPI1 oxidizing SCFAs to H_2 and METH1 converting CO_2 and H_2 to methane. Closer examination revealed even more intricate relationships involving Na^+ and H^+ ions as well as secondary sugars (i.e. xylose) reiterating that each population needs the metabolic activity and subsequent byproducts of the previous one to provide a supply of growing metabolites (Fig. 3a). Moreover, the estimation of translation and protein degradation rates pointed at a translational negative control for several ORFs involved in chemotaxis/motility and central metabolism, marking important changes in the community status. In conclusion, our data highlights that simple modifications to multi-omics toolkits can reveal much deeper functional-related trends and integrative co-dependent metabolisms that drive the overall phenotype of microbial communities, with potential to be expanded to more-complex and less-characterized microbial ecosystems.

Data availability

All sequencing reads have been deposited in the sequence read archive (SRP134228), with specific numbers listed in Supplementary Table 6 in Kunath et al.[9]. All microbial genomes are publicly available on JGI under the analysis project numbers listed in Supplementary Table 6 in Kunath et al.[9]. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE[46] partner repository with the dataset identifier PXD016242. The code used to perform the computational analysis is available at: https://github.com/fdelogu/SEM1b-Multiomics.

Acknowledgements

We are grateful for support from The Research Council of Norway (FRIPRO program, PBP: 250479), as well as the European Research Commission Starting Grant Fellowship (awarded to PBP; 336355 - MicroDE). The sequencing service was provided by the Norwegian Sequencing Centre (www.sequencing.uio.no), a national technology platform hosted by the University of Oslo and supported by the “Functional Genomics” and “Infrastructure” programs of the Research Council of Norway and the Southeastern Regional Health Authorities. P.N.E. is supported by an Australian Research Council Discovery Early Career Researcher Award (1700100428)

Author contributions

F.D., T.R.H. and P.B.P. conceived the study, performed the primary analysis of the data and wrote the paper (with input from all authors). B.J.K., and M.Ø.A. generated the data and contributed to the data analyses. P.N.E and G.W.T contributed to metabolic reconstructions for syntrophic acetate oxidizers and methanogens.

Competing interests

The authors declare there are no competing financial interests in relation to the work described.

Materials and Methods

Multiomics data acquisition

Background The full experimental setup and the methods concerning the retrieval of biological samples and data preprocessing were performed during a previous study[9] and can be summarized as follows: a microbial consortium called SEM1b was obtained from a biogas reactor using serial dilution and enrichment methods on spruce cellulose. A metagenomic analysis was initially performed on the SEM1b community using two different generations that had consistent population structure, and was used as a supporting database for a subsequent SEM1b time series experiment. The time series analyses consisted of metabolomics, metaproteomics and metatranscriptomics over nine time points (at t0, 8, 13, 18, 23, 28, 33, 38 and 43 hours) in triplicate (A, B and C), spanning the consortium life-cycle.

Metagenomics For generation of metagenomic data, 6ml samples of SEM1b culture were taken and cells were pelleted prior to storage at -20°C. Non-invasive DNA extraction methods were used to extract high molecular weight DNA as previously described in Kunath et al.[47]. DNA samples were prepared with the TrueSeq DNA PCR-free preparation, and sequenced with paired-ends (2×125bp) on one lane of an Illumina HiSeq3000 platform (Illumina Inc) at the Norwegian Sequencing Center (NSC, Oslo, Norway). Metagenomic analyses comprising quality trimming and filtering, reads assembly, binning and annotations were performed as previously described[9]. Resulting annotated open reading frames (ORFs) were retrieved and used as a reference database for the metatranscriptomic and metaproteomic analysis.

Metatranscriptomics mRNA extraction was performed in triplicate on time points t2 to t8, using previously described methods[11]. The extraction of the mRNA included the addition of an in vitro transcribed RNA as an internal standard to estimate the number of transcripts in the natural sample compared with the number of transcripts sequenced. For further normalization, total RNA was extracted using enzymatic lysis and mechanical disruption of the cells and purified with the RNeasy mini kit (Protocol 2, Qiagen, USA). The RNA standard (25ng) was added at the beginning of the extraction in every sample. After purification, residual DNA, free nucleotides and small RNAs were removed. Samples were treated to enrich for mRNAs and then amplified before being sent for sequencing at the Norwegian Sequencing Center (NSC, Oslo, Norway). Samples were subjected to the TruSeq stranded RNA sample preparation, which included the production of a cDNA library, and sequenced with paired-end technology (2×125bp) on one lane of a HiSeq 3000 system.

The resulting sequences were filtered and rRNA and tRNA reads were removed as performed in Kunath et al.[9]. The reads mapping on the internal standard pGEM-3Z were extracted using SortMeRNA[48] v2.1b and their counts used as I_R in the "Functional omics absolute quantification" section of the material and Methods, whilst the not mapping reads (the transcriptome in the sample) were used as $\sum T_R$. The retained reads were mapped against the predicted genes dataset using Kallisto pseudo -pseudobam[49] and the mapping files were produced with bam2hits. Transcripts were quantified with mmseq[50] and collapsed using mmcollapse[51].

Metaproteomics Proteins were extracted from t1 to t8 in triplicate following a previously described method[52] with a few modifications. Briefly, 30ml of cultures containing cells and substrate were centrifuged at 500x g for 5 minutes to pellet the substrate. The supernatant was centrifuged at 9000 x g for 15 minutes to collect the cells. Cell lysis was performed by resuspending the cells in 1ml lysis buffer (50 mM Tris-HCl, 0.1% (v/v) Triton X-100, 200 mM NaCl, 1 mM DTT, 2mM EDTA) and keeping them on ice for 30 minutes. Cells were disrupted in 3×60 seconds cycles using a FastPrep24 (MP Biomedicals, USA). Debris were removed by centrifugation at 16000 x g for 15 minutes. The supernatants containing the proteins were kept at -20°C until further processing. Extracted proteins were quantified using the Bradford's method. 50μg of each sample were denatured using SDS sample buffer and loaded on an Any-kD Mini-PROTEAN gel (Bio-Rad Laboratories, USA) and separated by SDS-PAGE for 20 minutes at 270V. Each gel lane was cut into 16 slices and the reduction, alkylation and tryptic digestion of the proteins into peptides were performed in-gel. The tryptic peptides were extracted from the gel and desalted prior to mass spectrometry analysis. Peptides were analyzed using a nanoLC-MS/MS system connected to a Q-Exactive hybrid quadrupole-orbitrap mass spectrometer (Thermo Scientific, Germany) equipped with a nano-electrospray ion source. The Q-Exactive mass spectrometer was operated in data-dependent mode and the 10 most intense peptide precursors ions were selected for fragmentation and MS/MS acquisition. The selected precursor ions were then excluded for repeated fragmentation for 20 seconds. The resolution was set to R=70,000 and R=35,000 for MS and MS/MS, respectively.

A total of 384 raw MS files (8 samples × 3 biological replicates × 16 fractions) were analyzed using MaxQuant[53] version 1.4.1.2 and proteins were identified and quantified using the MaxLFQ algorithm[54]. The data was searched against the generated MG dataset from

Kunath et al.[9] supplemented with common contaminants such as human keratin and bovine serum albumin. In addition, reversed sequences of all protein entries were concatenated to the database for estimation of false discovery rates. The tolerance levels for matching to the database was 6 ppm for MS and 20 ppm for MS/MS. Trypsin was used as digestion enzyme, and two missed cleavages were allowed. Carbamidomethylation of cysteine residues was set as a fixed modification and protein N-terminal acetylation, oxidation of methionines, deamidation of asparagines and glutamines and formation of pyro-glutamic acid at N-terminal glutamines were allowed as variable modifications. The 'match between runs' feature of MaxQuant[54] was applied. All identifications were filtered in order to achieve a protein false discovery rate (FDR) of 1%. Quantitative information was retrieved using the LFQ intensities of each proteins.

Metabolomics For monosaccharide detection, 2 ml samples were taken in triplicates, filtered and sterilized with 0.2µm sterile filters and 15 minutes boiling. Soluble sugars were identified and quantified by high-performance anion exchange chromatography (HPAEC) with pulsed amperiometric detection (PAD). For quantification, peaks were compared to linear standard curves generated with known concentrations of selected monosaccharides (glucose, xylose, mannose, arabinose and galactose) in the range of 0.001-0.1 g/L.

For the short chain fatty acids (SCFAs), 1ml was taken in triplicate from each time point, they were centrifuged at 16000x g for 5 minutes and the supernatants were filtered with 0.2µm sterile filters. 5µL of Sulfuric Acid 72% were added to the filtrates and let at rest for 2 minutes before being centrifuged again at 16000 x g for 5 minutes, transferred in a new tube and stored at -20°C until processing. SCFAs were then analyzed using a Dionex 3000 HPLC as described in Estevez et al.[55].

Functional omics absolute quantification

Metatranscriptomics The absolute quantification of transcripts was taken from Mortazavi et al.[10] using the internal standard from Gifford et al.[11] as reference to estimate the length of the initial transcriptome. The number of reads produced in a given sample is proportional to the total amount (in Nt) of starting material. With the addition of an internal standard we have the following proportion between the starting material for transcripts (T_{Nt}) and the internal standard (I_{Nt}) and the reads they produce (T_R and I_R respectively):

$$\frac{\sum T_{Nt}}{\sum T_R} = \frac{\sum I_{Nt}}{I_R},$$

in which the sums are taken over a single sample. The formula can be rearranged as:

$$\sum T_{Nt} = \sum I_{Nt} \times \frac{\sum T_R}{I_R}.$$

Since we know the number of molecules of internal standard added (I_M) and its length (I_{Nt}), we can substitute them in the equation as:

$$\sum T_{Nt} = I_M \times I_{Nt} \times \frac{\sum T_R}{I_R}.$$

We can now use the estimation of the starting length of the transcriptome and the TPMK transcript measurements in the formula from Mortazavi et al.[10]:

$$T_M = \frac{T_{RPMK}}{10^9} \times \sum T_{Nt},$$

which becomes:

$$T_M = \frac{T_{RPMK}}{10^9} \times I_M \times I_{Nt} \times \frac{\sum T_R}{I_R}.$$

<u>Metaproteomics</u> The “Total protein approach” method from Wiśniewski et al.[12] relies on the use of the protein mass per sample, the computed Molecular Weight (MW) of the detected proteins to transform the LFQ values into absolute ones. Here we omitted the per-cell quantification since SEM1b is a heterogeneous community and MG measurements were not taken for the time series.

We computed the *Total protein$_i$* as:

$$Total\ protein_i = \frac{LFQ\ intensity_i}{\sum LFQintensity}$$

Then the *Protein concentration$_i$* was obtained from the previous with:

$$Protein\,concentration_i = \frac{Total\ protein_i}{MW_i}$$

The method was developed on the assumption that the reference proteome is complete and that the total mass of the peptides detected is equal to the total mass of peptides processed by the machine. This is not necessarily valid in a microbiome for which the reference cannot be completely reliable. Thus we computed the fraction of identified mass using the raw MP files with the following formula:

$$Detectedproteinmass = \frac{Totalproteinmass \times \sum_{i=1}^{Pep_{id.}} Base\ peak\ intensity_i \times Mass_i}{\sum_{j=1}^{Pep_{tot}} Base\ peak\ intensity_j \times Mass_j}$$

Finally the copy number of proteins per sample was computed using the Avogadro's Number (N_A) as:

$$Copy\ number_i = Protein\ concentration_i \times Detectedproteinmass \times N_A$$

Multiomics dataset integration

Data preprocessing The MT and MP datasets estimate absolute abundance of ORFGs over time. An expression group is defined in this study as a set of ORFs which cannot be further resolved using the available data. When the analysis required the direct comparison of ORFs (e.g. transcript-protein correlation) only the singleton subset of the ORFGs was considered. The reliability of the expression estimation is linked to the number of unique hits (reads or peptides) available for a given ORF, therefore all the entries with 0 unique hits were filtered out. The datasets were then log10-transformed with a pseudocount equal to one. After expression density plotting, the minimum expression thresholds of 5 and 9 were selected for MT and MP, respectively, and the data was filtered accordingly. Principal component analysis was used to screen the samples and t7C (time point 7, replicate C) was identified as an outlier and removed before downstream analysis.

MP/MT linear fit We took the intersection of ORFs present in the MT and MP layers of the dataset for each of the selected MAGs (COPR1, CLOS1, COPR1, METH1, RCLO1, TEPI1, TEPI2, TISS1), and, for each sample, we performed a regression analysis in R. The values span several orders of magnitudes, thus we decided to fit the monomial functional:

$$protein = a \cdot RNA^k$$

which can be rewritten as:

$$\log(protein) = a + k \cdot \log(RNA)$$

to be more easily fitted as a linear model. The previously log10 transformed protein levels were used as *y* while the log10-transformed RNA was used as *x* in a linear model using the lm function. The slopes of the models were then used to fit a third grade polynomial function to obtain the linearity change profile in Fig. 1d.

Functional annotation and module completeness The KEGG Orthology (KO) numbers were assigned to the ORFs as a part of the annotation pipeline from IMG[56]. The ORF-wise annotation was then translated into the MT/MP-ORFGs assigning to each ORFG a non-redundant set of all the terms assigned to all the ORFs in the group. We used the KO numbers to estimate the KEGG module completeness using the R package MetaQy[57] v.1.1.0. The Glycosyl Hydrolases annotation was retrieved from Kunath et al.[9].

Metabolic marker genes selection The metabolic marker genes for Fig. 2 were selected with the following criterion. Glycolysis/Gluconeogenesis: enzyme with irreversible reactions. PPP: genes involved in the main interconversion loop between Ribose-5 Phosphate and Fructose-6 Phosphate. WLP: marker genes from Can et al.[58]. Methanogenesis: markers genes from Scheller et al.[40]. The Glycosyl Hydrolases were manually curated to assemble a set able to perform the substrate conversion.

PECA analysis We ran PECA-R[44] to estimate translation and protein degradation rates using the absolute quantification tables for transcripts and proteins with default parameters. The rates are estimated between two consecutive time points, therefore the sample from 8h was not included because it is missing the corresponding MT data. We filtered the results to identify the changing point using a score threshold of 0.9 and a FDR equal to 0.05.

Figure legends

Figure 1. Protein-to-RNA ratio distributions of as-yet uncultured bacterial and archaeal populations within a microbial community. a. Comparison of protein-to-RNA ratio distributions of selected MAGs reconstructed from the SEM1b community as well as

those previously reported in the literature. The dots represent the median values and the bars span from the first to the third quartiles (Bacteria: green, Archaea: red, Eukarya: blue). The protein-to-RNA ratios for *E.coli* was retrieved from Taniguchi et al.[1], Yeast1 from Ghaemmaghami et al.[20], Yeast2 from Lu et al.[21], Human1 from Schwanhausser et al.[22] and Human2 from Li et al.[23]. **b.** The distribution of the Pearson Correlation Coefficients (PCC) between transcripts and their corresponding proteins computed across the time points. With a median PCC of 0.41, the change in the amount of a given transcript over time seemingly does not translate into a change in the amount of the corresponding protein. **c.** Per-time-point scatterplots of the absolute protein and transcript levels for ORFs that produced both detectable transcript and protein in SEM1b datasets. For simplicity, only four representative MAGs are shown, with all MAGs depicted in Supplementary Fig. 2. **d.** The plot shows how the linearity parameter *k* between RNA and protein changes over time for the different MAGs. The linearity represents how a change in RNA level is reflected in a change in the corresponding protein level. The parameter ranges from 0 to 1, and increasingly smaller values translate in fewer proteins being expected for the same level of RNAs. The populations CLOS1, METH1 and TEPI1 are converging towards the same values, while RCLO1 has a parallel trend. Hinting to the existence, and the reaching of an equilibrium among them.

Figure 2. Overview of the genetic potential and expressed modules in the seven populations of SEM1b. Module completeness denotes the level of detected RNA and proteins mapped to major genes/metabolic pathways that are critical to the SEM1b lifecycle. Only MAGs with the highest quality reconstruction (RCLO1, CLOS1, COPR1, TISS1, TEPI1, TEPI2 and METH1) are included as well as two isolated and genome-sequenced *Coprothermobacter* strains, for which the transcriptome and the proteome were considered as the species level.

Figure 3. Schematic representation of the temporal and co-dependent metabolism of SEM1b that converts spruce-derived cellulose to methane. **a.** Within SEM1b, four major metabolic stages are required: Spruce → Hexoses/Pentoses, Hexoses/Pentoses → SCFAs, SCFAs → CO_2+H_2 and CO_2+H_2 → methane. Metabolites (spruce, sugars, SCFAs, CO_2+H_2

and methane) involved in these processes were measured and the temporal analysis of the metabolic pathways involved in their interconversion is depicted for the major SEM1b populations. Other metabolites (for which abbreviations are: Nt=Nucleotides, PRPP=Phosphoribosyl pyrophosphate, AA=Amino acids, PYR=Pyruvate, OXA=Oxaloacetate and OXO=Oxo-glutarate) are shown to highlight the essential metabolism of the microbes. In the central metabolic network the metabolites are linked by solid arrows if the interconversion requires one step or the link between them is addressed more in detail (blue dot if in Fig. 2, red dot if in a pathway plot herein). Metabolic pathways are quantified via marker genes (selection in methods section) in the scale of log10-transformed transcript molecules per sample whilst the solid lines in the plots represent the qubic fitting of the data points. More metabolites' abbreviations are CELL=Cellulose, CLB=Cellobiose, GLU=Glucose, XYL=Xylan, XLB=Xylobiose and pathways' abbreviations are WLP="Wood-Ljungdahl Pathway", PPP="Pentose Phosphate Pathway". **b.** Sporulation is common to all Gram positive bacteria of the community and it is quantified with the marker *spoIVA*. Notably TEPI2 is investing greatly in spore formation until 28h after the inoculum (t4). **c.** The genes for the Ribose and xylose transporter (*rbs*) are expressed in four populations. Notably TEPI2 produces more rbs transcripts than the closely related MAG TEPI1; indeed, the first has been predicted to be a mere fermenter whilst the latter bases its metabolism on the WLP pathway (Fig. 3a). **d.** Microbial motility is represented by the marker gene *flgD*. RCLO1 is the most active bacterium, producing less and less flagella over time after t4. It starts ahead of the others at t2, presumably finishing the colonization of the substrate (Spruce-derived cellulose).

References

1. Taniguchi, Y. *et al.* Quantifying E. coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells. *Science (80-.).* **329**, 533–538 (2010).

2. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).

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

4. Shi, Y., Tyson, G. W. & DeLong, E. F. Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. *Nature* **459**, 266–269 (2009).

5. Wilmes, P. & Bond, P. L. Metaproteomics: studying functional gene expression in microbial ecosystems. *Trends Microbiol.* **14**, 92–97 (2006).

6. Dewi Puspita, I., Kamagata, Y., Tanaka, M., Asano, K. & Nakatsu, C. H. Are Uncultivated Bacteria Really Uncultivable? *Microbes Environ.* **27**, 356–366 (2012).

7. Lovell, D., Pawlowsky-Glahn, V., Egozcue, J. J., Marguerat, S. & Bähler, J. Proportionality: A Valid Alternative to Correlation for Relative Data. *PLOS Comput. Biol.* **11**, e1004075 (2015).

8. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* **13**, 227–232 (2012).

9. Kunath, B. J. *et al.* From proteins to polysaccharides: lifestyle and genetic evolution of Coprothermobacter proteolyticus. *ISME J.* **13**, 603–617 (2019).

10. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).

11. Gifford, S. M., Sharma, S., Rinta-Kanto, J. M. & Moran, M. A. Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. *ISME J.* **5**, 461–472 (2011).

12. Wiśniewski, J. R. & Rakus, D. Multi-enzyme digestion FASP and the 'Total Protein Approach'-based absolute quantification of the Escherichia coli proteome. *J. Proteomics* **109**, 322–31 (2014).

13. Benelli, D., La Teana, A. & Londei, P. Translation Regulation: The Archaea-Eukaryal Connection. in *RNA Metabolism and Gene Expression in Archaea* 71–88 (Springer, 2017).

14. Achinas, S., Achinas, V. & Euverink, G. J. W. A Technological Overview of Biogas Production from Biowaste. *Engineering* **3**, 299–307 (2017).

15. Caro-Quintero, A. & Konstantinidis, K. T. Bacterial species may exist, metagenomics reveal. *Environ. Microbiol.* **14**, 347–355 (2012).

16. Sharon, I. *et al.* Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res.* **23**, 111–20 (2013).

17. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* **31**, 533–538 (2013).

18. Leary, D. H., Hervey, W. J., Deschamps, J. R., Kusterbeck, A. W. & Vora, G. J. Which metaproteome? The impact of protein extraction bias on metaproteomic analyses. *Mol. Cell. Probes* **27**, 193–199 (2013).

19. Marguerat, S. *et al.* Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. *Cell* **151**, 671–83 (2012).

20. Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425**, 737–41 (2003).

21. Lu, P., Vogel, C., Wang, R., Yao, X. & Marcotte, E. M. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.* **25**, 117–124 (2007).

22. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337–42 (2011).

23. Li, J. J., Bickel, P. J. & Biggin, M. D. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ* **2**, e270 (2014).

24. Hennigan, A. N. & Reeve, J. N. mRNAs in the methanogenic archaeon Methanococcus vannielii: numbers, half-lives and processing. *Mol. Microbiol.* **11**, 655–670 (1994).

25. BINI, E., DIKSHIT, V., DIRKSEN, K., DROZDA, M. & BLUM, P. Stability of mRNA in the hyperthermophilic archaeon Sulfolobus solfataricus. *RNA* **8**, S1355838202021052 (2002).

26. Hasenohrl, D., Lombo, T., Kaberdin, V., Londei, P. & Blasi, U. Translation initiation factor a/eIF2(-) counteracts 5' to 3' mRNA decay in the archaeon Sulfolobus solfataricus. *Proc. Natl. Acad. Sci.* **105**, 2146–2150 (2008).

27. Jäger, D. *et al.* An archaeal sRNA targeting cis - and trans -encoded mRNAs via two distinct domains. *Nucleic Acids Res.* **40**, 10964–10979 (2012).

28. Li, J. *et al.* Global mapping transcriptional start sites revealed both transcriptional and post-transcriptional regulation of cold adaptation in the methanogenic archaeon Methanolobus psychrophilus. *Sci. Rep.* **5**, 9209 (2015).

29. Luo, H.-W., Zhang, H., Suzuki, T., Hattori, S. & Kamagata, Y. Differential Expression of Methanogenesis Genes of Methanothermobacter thermoautotrophicus (Formerly Methanobacterium thermoautotrophicum) in Pure Culture and in Cocultures with Fatty Acid-Oxidizing Syntrophs. *Appl. Environ. Microbiol.* **68**, 1173–1179 (2002).

30. Thauer, R. K., Kaster, A.-K., Seedorf, H., Buckel, W. & Hedderich, R. Methanogenic archaea: ecologically relevant differences in energy conservation. *Nat. Rev. Microbiol.* **6**, 579–591 (2008).

31. Jackson, B. E. & McInerney, M. J. Anaerobic microbial metabolism can proceed close to thermodynamic limits. *Nature* **415**, 454–456 (2002).

32. Shachrai, I., Zaslaver, A., Alon, U. & Dekel, E. Cost of Unneeded Proteins in E. coli Is Reduced after Several Generations in Exponential Growth. *Mol. Cell* **38**, 758–767 (2010).

33. Wang, P. *et al.* Robust growth of Escherichia coli. *Curr. Biol.* **20**, 1099–103 (2010).

34. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165**, 535–50 (2016).

35. Dumitrache, A. D., Wolfaardt, G., Allen, G., Liss, S. N. & Lynd, L. R. Form and function of Clostridium thermocellum biofilms. *Appl. Environ. Microbiol.* **79**, 231–239 (2013).

36. Artzi, L., Bayer, E. A. & Moraïs, S. Cellulosomes: bacterial nanomachines for dismantling plant polysaccharides. *Nat. Rev. Microbiol.* **15**, 83–95 (2017).

37. Park, J. O. *et al.* Synergistic substrate cofeeding stimulates reductive metabolism. *Nat. Metab.* **1**, 643–651 (2019).

38. Galperin, M. Y. *et al.* Genomic determinants of sporulation in Bacilli and Clostridia : towards the minimal set of sporulation-specific genes. *Environ. Microbiol.* **14**, 2870–2890 (2012).

39. Bertsch, J., Öppinger, C., Hess, V., Langer, J. D. & Müller, V. Heterotrimeric NADH-Oxidizing Methylenetetrahydrofolate Reductase from the Acetogenic Bacterium Acetobacterium woodii. *J. Bacteriol.* **197**, 1681–1689 (2015).

40. Scheller, S., Goenrich, M., Boecher, R., Thauer, R. K. & Jaun, B. The key nickel enzyme of methanogenesis catalyses the anaerobic oxidation of methane. *Nature* **465**, 606–608 (2010).

41. Cai, L., Friedman, N. & Xie, X. S. Stochastic protein expression in individual cells at the single molecule level. *Nature* **440**, 358–62 (2006).

42. Maurizi, M. R. Proteases and protein degradation inEscherichia coli. *Experientia* **48**, 178–201 (1992).

43. Rodnina, M. V. Translation in Prokaryotes. *Cold Spring Harb. Perspect. Biol.* **10**, a032664 (2018).

44. Teo, G., Vogel, C., Ghosh, D., Kim, S. & Choi, H. PECA: a novel statistical tool for deconvoluting time-dependent gene expression regulation. *J. Proteome Res.* **13**, 29–37 (2014).

45. Janssen, P. H. & Kirs, M. Structure of the Archaeal Community of the Rumen. *Appl. Environ. Microbiol.* **74**, 3619–3625 (2008).

46. Perez-Riverol, Y. *et al.* The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D450 (2019).

47. Kunath, B. J., Bremges, A., Weimann, A., McHardy, A. C. & Pope, P. B. Metagenomics and CAZyme Discovery. in *Methods in Molecular Biology* 255–277 (Humana Press, 2017).

48. Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012).

49. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

50. Turro, E. *et al.* Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biol.* **12**, R13 (2011).

51. Turro, E., Astle, W. J. & Tavaré, S. Flexible analysis of RNA-seq data using mixed effects models. *Bioinformatics* **30**, 180–188 (2014).

52. Hagen, L. H. *et al.* Quantitative Metaproteomics Highlight the Metabolic Contributions of Uncultured Phylotypes in a Thermophilic Anaerobic Digester. *Appl. Environ. Microbiol.* **83**, (2017).

53. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).

54. Cox, J. *et al.* Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. *Mol. Cell. Proteomics* **13**, 2513–2526 (2014).

55. Estevez, M. M., Sapci, Z., Linjordet, R. & Morken, J. Incorporation of fish by-product into the semi-continuous anaerobic co-digestion of pre-treated lignocellulose and cow manure, with recovery of digestate's nutrients. *Renew. Energy* **66**, 550–558 (2014).

56. Chen, I.-M. A. *et al.* IMG/M: integrated genome and metagenome comparative data analysis system. *Nucleic Acids Res.* **45**, D507–D516 (2017).

57. Martinez-Vernon, A. S., Farrell, F. & Soyer, O. S. MetQy—an R package to query metabolic functions of genes and genomes. *Bioinformatics* **34**, 4134–4137 (2018).

58. Can, M., Armstrong, F. A. & Ragsdale, S. W. Structure, Function, and Mechanism of the Nickel Metalloenzymes, CO Dehydrogenase, and Acetyl-CoA Synthase. *Chem. Rev.* **114**, 4149–4174 (2014).

Paper III

Functional dynamics of a microbial community from a wastewater treatment plant.

F. Delogu[1], S.A. Martinez[2], B.J. Kunath[2], M. Herold[2], J. Garcia[2], P.B. Pope[1,3], P. May[2], P. Wilmes[2]

[1] Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, 1432 Ås, Norway

[2] Luxembourg Centre for Systems Biomedicine, Université du Luxembourg, L-4362 Esch-sur-Alzette, Luxembourg

[3] Faculty of Biosciences, Norwegian University of Life Sciences, 1432 Ås, Norway

Abstract

Biological wastewater treatment plants exploit microbial consortia to perform important chemical transformations (e.g. nitrification and denitrification) that are needed before water reclamation. In particular, the lipid fraction in wastewater remains an untapped energy source, although it is also the perfect substrate for lipid-feeding bacteria that form an inhibitive bulky foam. To characterize the community from such foam and to undercover the temperature-driven seasonality and other physicochemical influence on the system, we analyzed temporal meta-omics data that was generated over a one-year period (weekly samples). Our analysis rendered the overarching biochemical reaction network deployed by the community and retrieved the critical components of fatty acid synthesis and nitrogen metabolism. The gene expression of key components from fatty acid synthesis were found to be inversely proportional to taxa richness, suggesting inter-species competition for substrates. Nitrogen metabolism instead was dominated by a single family: Nitrosomonadaceae, which is linked to greenhouse emission (nitrous oxide) and therefore should be controlled. Our findings suggest a cyclical and dynamic interaction of the taxa and genetic pool in the community maintains certain metabolic functions, highlighting critical nodes in the reaction network that should be considered when devising improvements and/or direct exploitation of these (or similar) processes for lipid harvesting in wastewater treatment plants.

Introduction

Microbes are ubiquitous on planet Earth1 and make up to 17% of its carbon biomass[2]. The pervasiveness of microbes is the foundation for their ability to form complex communities of heterogeneous taxonomy and function. Different microbial lineages are continuously evolving to fit the most diverse ecological niches[1], which ultimately gives rise to communal living that thrives on its metabolic complementarity. Humans have learned how to exploit microbial communities to perform complex tasks such as baking and brewing, and more recently to process waste[3]. The latter activity has been coupled with the idea of reclamation of resources and harnessing of residual chemical energy, making systems such as Biological Wastewater Treatment Plants (BWWTPs) de facto model environments on which several interests converge. It is known that Lipid Accumulating Organisms (LAOs) accumulate lipids from the environment or synthesize them themselves[4], and that solid fraction of domestic wastewater can contain more than 40% of lipids[5]. Moreover, the recovery of LAO populations grown on wastewater to produce biofuel has been estimated to be profitable[6]. Processes such as nitrification/denitrification are required to reclaim the water[7], thus they must be carried out alongside the lipid accumulation. Therefore, in order to exploit and improve the potential of BWWTP grown LAOs we must understand their communities and concerted metabolism within their natural environment and range of physio/chemical factors they are subject to.

Modern microbial ecology has been updated with newly developed meta-omics techniques that enable direct access to the main biological molecules that constitute a microbial community in its native environment. Briefly, metagenomics (MG) charts the taxonomic composition and genetic potential of the community, hence predicts its metabolism and lifestyle[8,9]; metatranscriptomics (MT) assesses the functions in which the microbes are investing via gene expression[10]. Here we present a temporal reconstruction of the LAO surface community (Schif-LAO) from an anaerobic tank at the BWWTP in Schifflange (Luxembourg). The sampling spans more than one year with 51 samples from which we analyzed the MG, the MT and the physio/chemical factors measured at the site. We reconstructed the MG structure of the community, alongside its taxonomy, genetic potential and gene expression, from which we extracted the time patterns (e.g. cyclicity). The patterns were linked with the environmental parameters to build an explanatory model (e.g. seasonality). Moreover, the entire metabolism of the community was represented as a single network. We extracted two functions relevant for LAO ecosystem in particular and wastewater treatment plants in general: lipid biosynthesis and nitrogen metabolism. We showed that in both metabolic subsystems there are some reactions that were performed mainly by a single taxon or genes from unknown taxa. The reactions (and the genes enabling them) highlighted by our analysis are candidates to be "keystone" units that are irreplaceable to the entire community. We therefore integrated the exploration of the operational boundaries of the system and its putative keystone components.

Results and Discussion

Functional composition and time patterns

The Schif-LAO community was firstly sampled in 2010-10-04 and 2011-01-25 and led to an estimate of approximately 600 resident operational taxonomic units (determined via 16S rRNA gene analysis)[11] and accounted for a total of 23,317 open reading frames (ORFs) from four samples (with four biological replicates each). From the following sampling between 2011-03-21 and 2012-05-03 we obtained 51 weekly samples, that we analyzed individually to obtain 51 MGs and sets of ORFs. In order to form a coherent ORF set spanning the whole time-series, we clustered them according to their sequence (see methods), which lead to a total sum of ~19.8×10^6 different ORFs (extended dataset). A KEGG Orthology group was assigned to 40.4% of the ORFs in the set, whilst taxonomic affilaitions were designated to 38.5%. The number of ORFs' copies as well as their detected gene expression were estimated over the extended dataset (see method). The vast majority of the genes however were not found to be expressed over the entire dataset and were only detected in few samples alone, with as many as 16.8×10^6 in only one sample. This indicates that a large share of the gene pool in Schif-LAO is not specifically required for the enduring well-being of the community but rather their cumulative functional effort may be compartmentalized, fitting the previous results from Roume et al.[12].

To reduce the complexity of the dataset and resolve the gene pool with the strongest signal in the community, we first filtered for the "core set", which contains all the genes with at least 1 transcript per million TPM) in at least one sample. The core gene and transcript sets comprised 0.7×10^6 and 0.8×10^6 ORFs for the MG and MT respectively, and were subsequently used to compute eigengenes; a set of vectors that are associated to a gene count/expression matrix and summarize their pattern over a set of given samples[13]. In a time-series context, we reconstructed six (EG1-6) eigengenes associated with the MG and six (EG1-5, EG8) with the MT data (**Fig. 1a-b**). When the set of eigengenes were compared they formed five pairs with an absolute Pearson Correlation Coefficient (PCC) ≥ 0.7 and one pair with an absolute PCC = 0.5 (**Fig. 1c**). The correlated eigengenes pairs indicate that the pattern of change in gene number and gene expression in Schif-LAO are similar and intuitively the curves described by the MG are smoother than by the MT data. The phenomenon is perhaps the reflection of the more transient characteristic of RNA molecules, if not also an inferior extraction yield. The environmental parameters were measured often in multiple different ways (e.g. online measurement form the WWTP, manual, air, etc.), so we selected them in order to reduce the co-linearity, resulting in a shrink from 15 to 7 parameters, which include: conductivity, dry matter, ammonium (NH_4), nitrate (NO_3), oxygen, pH and temperature.

In order to understand the relationship between the environmental parameters and the temporal MG/MT trends we fitted a multivariate linear model and assessed the relevance of the explanatory variables. The results (**Fig. 1d**) show how the most relevant environmental factors are temperature and NO_3, linked with five EGs each, followed by dry matter with four EGs. Ammonium and pH are more specific with two EGs and conductivity and oxygen with one EG each. One of the main processes happening in WWTPs is the conversion of Ammonium into Nitrate ($NH_4 \rightarrow NH_3 \rightarrow NO_2 \rightarrow NO_3$), therefore it is hard to establish the causal direction of the link between these two compounds and gene copy number/expression. On the contrary the direction of the relation between temperature and the EG pattern is more intuitive, with the temperature acting as driver for the seasonality of Schif-LAO. Moreover, we fitted the MG-EG2 and temperature with a sine curve with a period (T) of one year and we obtained a perfect fit with F-statistics of 181 and 269, p-values $< 10^{15}$ same phase and inverted amplitude sign. The sine function is cyclical, assuming the same values at a distance of $2\pi/T$. The present fit, with a T of 365 days, points out the seasonal composition and behavior of the microbial community. This means that there is a set of genes whose presence in the Schif-LAO consortium depends on the temperature and is supposed to reach the same values every year.

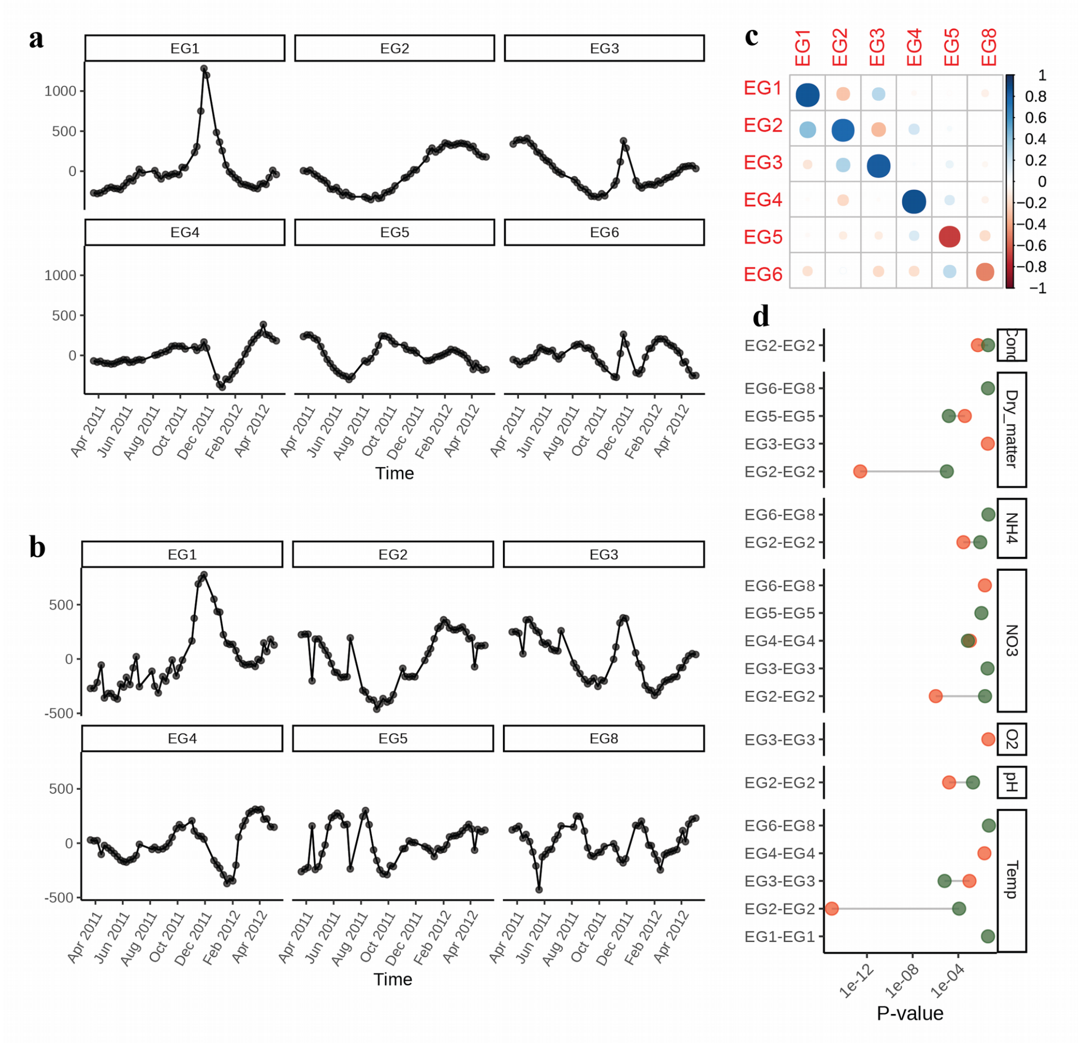


Fig. 1. Functional composition and time patterns in Schif-LAO. a. Non-stationary eigengenes (EG1-6) computed over time from the MG ORF reduced data set. The y axis is in arbitrary scale. **b.** Non-stationary eigengenes (EG1-5,8) computed over time from the MT ORF reduced data set. The y axis is in arbitrary scale. **c.** Pairwise correlations between the eigengenes from the MG (rows) and the MT (columns). Blue indicates a high level of positive correlation, red a high level of negative correlation. Size/opacity represent the absolute value of the correlation, with larger/opaque dots indicate values close to 1. **d.** p-value plot of the variable significance from the linear fit of the eigengene. Only the significant ones are shown ($p<0.05$), where orange marks if the eigengene comes from the MG eigengenes and green from the MT ones.

Community reaction network

We proceeded to explore the collective enzymatic capability of Schif-LAO via the study of its predicted biochemical reaction networks. Every chemical reaction takes one or more reactants and releases one of more products. The same compound may be a reactant for multiple reactions and a product for another, becoming the bridge between the two independent reactions. Following this idea, we took all the known reactions (in the form of KO numbers) and connected them to one another if they shared at least one associated compound, building a comprehensive reaction network (details in methods). We refer to the nodes of this network as "Collapsed KOs" (CKOs), which contain all the KO entries with the same associated compounds (see methods for complete explanation). We customized the general network for Schif-LAO using the list of annotated ORFs. We obtained a reaction network of 1,984 nodes and 13,350 edges. The number of ORFs per CKO varied greatly, with a maximum of 77,474 and a median of 284. Most functions are present in biological systems with a certain degree of redundancy, which give resilience to the system[14]. Nevertheless, it is known that sometimes certain functions are performed by irreplaceable populations/genes[15], which elect them to the status of keystone functions.

We speculated that the taxa contributing to a given function in the Schif-LAO community at any given point in the sampling time may change, therefore we sought to taxonomically identify those ORFs that are crucial in the carrying of their function. To do this, we computed the information entropy of the MT for every CKO at the Family level. In information theory, entropy is used to quantify the amount of uncertainty in a message, such that a high entropy score would indicate high uncertainty. Indeed, the maximum entropy is the one associated with all the possible outcomes having the same probability and defined as *log(n)*, where *n* is the number of outcomes. Therefore, a CKO in which a Family is contributing with the vast majority of the transcripts will have a very low entropy; in contrast, a widely shared CKO will have a high score. Moreover, we normalized the entropy scores by the maximum entropy in order to make all the CKOs comparable.

Lipid biosynthesis hints to resource competition in Schif-LAO

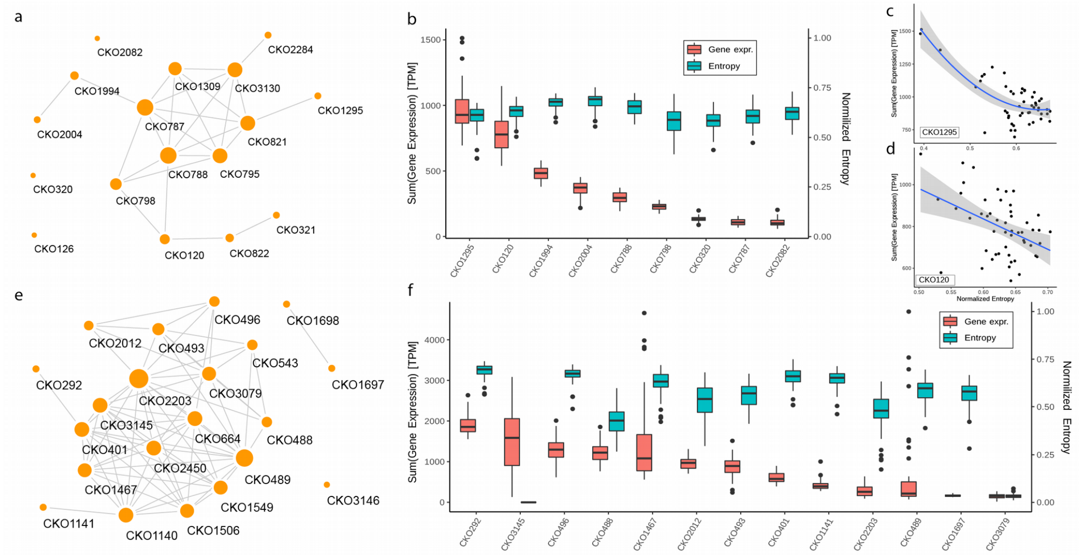


Fig 2. Community reaction networks in Schif-LAO. a. Metabolic network subset for the "fatty acid biosynthesis"-related CKOs. The size of the node is proportional to their degree. **b.** Boxplot of sum of gene expression per "fatty acid biosynthesis"-related CKO over time (scale on the left y axis) and the family-level taxonomic richness (normalized entropy) of the gene expression (scale on the right y axis). High values of gene expression indicate a large investment of the community to the given reaction (CKO). A high entropy score indicates that the reaction is performed in comparable amount among several taxonomic families, on the contrary, the lower the score, the more unequal the gene expression is, with one ore few taxa producing more transcripts than the others. **c.** Scatterplot of gene expression and entropy for each time point in CKO1295 with the quadratic fit and 0.95 confidence interval. **d.** Scatterplot of gene expression and entropy for each time point in CKO120 with the linear fit and 0.95 confidence interval. **e.** Metabolic network subset for the "nitrogen metabolism"-related CKOs. The size of the node is proportional to their degree. **f.** Boxplot of sum of gene expression per "nitrogen metabolism"-related CKO over time (scale on the left y axis) and the family-level taxonomic richness (normalized entropy) of the gene expression (scale on the right y axis).

Lipid biosynthesis hints to resource competition in Schif-LAO

Fatty acid synthesis of type I (FAS I) in mammals is a straightforward process, performed by a single structure containing all the catalytic centers required and originating from a single peptide (two in case of other Eukaryotes, such as yeast)[16]. On the contrary, the second type of FAS (FAS II) occurs in plants and bacteria and it is a complex task involving several soluble enzymes encoded in different ORFs[17]. The complexity and centrality of FAS II in those organisms is due to its ability to produce a wide array of lipids varying in length, unsaturation(s), branching, alongside the intermediates for other cellular components[17]. The core of FAS II is the Acyl-Carrier Protein (ACP) which shuttles the new fatty acid between several enzymes involved in the pathway[18]. In Schif-LAO there are 17 CKOs associated with the KO term "Fatty Acid Biosynthesis", connected by 26 metabolites (edges) (**Fig. 2a**), accounting for fatty acid initiation, elongation and termination. The most expressed reaction node is CKO1295, ranging between 695.3-1513.5 TPM and a median of 927.8 TPM per time point (**Fig. 2b**). CKO1295 embeds the two opposite reactions that attach and detach the cofactor A (CoA) to the fatty acid chain, accounting for the following KOs: K01068, K01074, K01076, K01897, K15013, K17360; which correspond to the genes ACOT1/2/4, PPT, ABHD17, ACSL, ACSBG, ACOT7 respectively. Interestingly the richness in taxa contributing to the node is inversely proportional to the gene expression (Spearman's ρ of -0.35, p=0.01) with a quadratic trend (**Fig. 2c**). The second largest expressed node is CKO120 with a range of 539.6-1147.4 TPM and a median equal to 778.3 TPM per time point. CKO120 encodes the fatty acid synthase reaction with the KOs K00059, K00665, K00667 and K11533; representing respectively the genes fabG, FASN, FAS2 and fas. Similar to the previous case, in CKO120 the gene expression is inversely proportional to the taxa richness (ρ=-0.38, p<0.01) but with a linear trend (**Fig. 2d**). CKO1295 and CKO120 cover two fundamental aspects of FAS: activation/deactivation of the fatty acid and its extension; however, our data would suggest that different taxa enact a competitive takeover of these functions. The most active families for CKO1295 are Leptospiraceae (32%), Comamonadaceae (13.2%) and Chitinophaga (7.3%) whilst the gene counts are different, with Leptospiraceae (7%) followed by Comamonadaceae (24.8%). The gene counts for CKO120 see the dominance of family Microthrixaceae (24.5%), followed by Acidomicrobiaceae (10%) and Comamonadaceae (8.4). Yet again in the gene expression Leptospiraceae (28.3%) contributes the most, followed by Moraxellaceae (11.8%) and Microthrixaceae (6.7%).

Nitrogen metabolism is monopolized by the family Nitrosomonadaceae

Nitrogen removal is a crucial feature in treatment of wastewater, carried out by ammonia-oxidizing bacteria (AOB). Most of the nitrogen in the water is in the form of ammonia (NH_4) which is converted to nitrite (NO_2) during the two-steps *nitritation* process. The enzyme responsible of the first conversion is the ammonia oxidizing monooxygenase (*amo*), which is a close homolog to the particulate monooxygenase (*pmo*), which instead oxidizes methane. Both enzymes use copper ions to perform the oxidization of their substrates and they share the same KO number: K10940[19]. Interestingly *pmo* is a great resource in capturing the greenhouse gas methane, whilst on the other hand *amo*-encoding AOB, grown in sub-optimal conditions, tend to form the greenhouse nitric and nitrous gas[20]. The hydroxylamine produced by *amo* is then converted into nitrate (NO_2) by the enzyme Hao. Subsequently nitrite is used to produce nitrate (NO_3) via *nitratation* with the enzyme NarG-H. Both nitrate and nitrite can be transformed into nitrogen and oxygen gas (N_2+O_2) with the *denitritation* and *denitrification* processes respectively. The Nitrogen-related metabolism of Schif-LAO includes 21 reaction nodes and 71 metabolic edges (**Fig. 2e**). The entropy analysis of the contribution of different taxonomic families to the reactions points immediately to CKO3145 and CKO3079 as potential keystones in the system (**Fig. 2f**). The first reaction node is overwhelmingly dominated by the ORFs from the family Nitrosomonadaceae (MG 97.8%, MT 99.1%) and contains the *amo* gene subunits A-B (K10944-6, EC:1.14.16.3 and EC:1.14.99.39). The second node is dominated again by the transcripts from Nitrosomonadaceae (MG 63.9%, MT 91%) and encodes the hydroxylamine dehydrogenases (K10535, EC:1.7.2.6). Given the crucial importance of the presence of the gene *amo* in the environment to start the assimilation of ammonia, the main family producing transcripts from it, Nitrosomonadaceae, must be held carefully tuned to the optimal size to optimize the performance of Schif-LAO.

Conclusions

We present the temporal reconstruction of the surface microbial community of a BWWT plant over 1.5 years of weekly sampling. The gene count and gene expression show six distinct and linearly independent patterns (eigengenes) across time (**Fig. 1a-b**), many of which were linked to the physiochemical parameters (**Fig. 1d**). In particular, the MG/MT second eigengene show a cyclical behavior highly associated with the water temperature (**Fig. 1d**) and both of them can be fitted with sine functions of same phase (365 days) and opposite sign. Therefore, we can model the dynamics of the Schif-LAO community as a yearly cycle dictated by the temperature variation.

We reconstructed the enzymatic network of the community and inquired two specific functions important for the community: lipid accumulation and nitrogen metabolism. For lipid-associated functions, we show how there are no nodes that are both highly expressed and dominated by a single taxon, and therefore apparently there are no keystone nodes that dictate lipid accumulation. However, we show that the two higher expressed reaction nodes

(CKO1295 and CKO120) encode fundamental steps of the FAS and their expression is negatively correlated with taxa abundance, suggesting a competition for the substrate. Regarding nitrogen metabolism we found two more interesting reactions (CKO3145 and CKO3079) which are performed de facto only by the bacterial family Nitrosomonadaceae. In particular CKO3145 contains the *amo* gene, used in the first of the two-step reaction to convert NH_4 in NO_2, but can lead to the formation of nitrogen-based greenhouse gases depending on the oxygen level in the water and therefore must be taken into consideration when planning to alter the community or the physicochemical parameters of the system.

Data availability

The code used to build the static reaction network is available at: https://github.com/fdelogu/kegg_net .

Materials and Methods

Sampling

Individual floating sludge islets within the anoxic tank of the Schifflange BWWT plant (Esch-sur-Alzette, Luxembourg; 49°30'48.29"N; 6°1'4.53"E) were sampled according to previously described protocols[11]. Samples are indicated as dates (YYYY-MM-DD). More frequent sampling of 51 time points was performed from 2011-03-21 to 2012-05-03, of which data from three samples (2011-10-05, 2011-10-05 and 2012-01-11) have been previously published[11].

Concomitant biomolecular extraction and high-throughput meta-omics

Concomitant biomolecular extraction of DNA, RNA and proteins as well as high-throughput measurements to obtain MG, MT, and MP data were carried out according to previously established protocols[11,12,21]. The raw MG and MT FASTQ files as well as the assembled contigs are available as NCBI BioProject PRJNA230567[11,12,21]. MP data has been deposited in the PRIDE database under the accession number PXD013655.

Co-assembly of metagenomic and metatranscriptomic data

Sample-wise integrated MG and MT data analyses were performed using IMP version 1.3[22] with customized parameters, i.e. i) Illumina Truseq2 adapters were trimmed, ii) the step involving the filtering of reads of human origin step was omitted for the preprocessing, and iii) the MEGAHIT de novo assembler[23] was used.

Gene prediction, annotation and clustering

Open reading frames were predicted using Prodigal[24] v2.6 with "meta" and "incomplete gene" settings. Predicted genes were annotated using hmmsearch[25] against an in-house licensed version of the KEGG KO database[26]. In order to build a coherent and non-redundant dataset of genes all ORF sequences (both complete and incomplete genes) from the 51 samples were pooled and clustered using CD-HIT-EST[27], with the parameters *-c 0.95 -d 0 -s 0.9 -aS 0.9 -aL 0.3 -uS 0.1*. A curated set of metagenome assembled genome (MAGs) from the same samples was used to infer taxonomy as reference database for a blastn (blast 2.2.28+[28]) search of the ORFs. The MAGs were in turn annotated using AMPHORA2[29], as previously described[30].

MG and MT quantification and filtering

The filtered MG and MT reads were pseudoaligned to the clustered set of genes using kallisto pseudo -pseudobam[31]. The whole dataset was then filtered to create the "Core dataset" containing only genes with at least 1 Transcript Per Million (TPM) in at least one sample.

Eigengenes and their analysis

The EG analysis was conducted in in R 3.5.3. Firstly we computed the EGs from the MG and MT core sets as the principal components obtained with the function prcomp. Subsequently the EGs were tested using the Ljung-Box test (Box.test), the augmented Dickey-Fuller test (adf.test) and the Kwiatkowski–Phillips–Schmidt–Shin (kpss.tests) tests with null hypotheses "trend" and "level". If at least two of the four tests were passed the EG was considered non-stationary. The seven physico-chemical parameters were used as explanatory variables in linear models to fit the non-stationary EGs using the lm function. Finally, we assessed the significance of the explanatory variables using ANOVA (anova).

Metabolic network construction

The construction of the metabolic network followed the steps from Roume et al. 2015[12], using the most updated version of the KEGG[26] rest repositories. Moreover, we released the code to enable other scientists to generate their networks locally, or as alternative to download the premade one and subset it to generate an experiment-specific network. In brief, we filtered the reaction file "rn" to remove those reactions which do not have a reaction class in "rc", then we removed the most common cofactors using the "cpd" database (10-Formil-THF, Acetyl-CoA, ADP, AMP, ATP, Co-A, GDP, L-Glutamine, L-Glutamate, GTP, NADH, NADPH, NAD, NADP, Phospho-enol-pyruvate, Propyonyl-CoA, Pyruvate, Suc-CoA, THF, Acceptor, Cytochromes-C-Reduced, Cytochromes-C, Donor-H2, Oxidized-Flavodoxins, Reduced-flavodoxins). Then we mapped the KEGG Orthology (KO) entries to the reactions and collapsed all the KOs with the same reactants into Collapsed KOs (CKOs) to act as nodes of the metabolic network. Finally, we connected the CKOs that share one or more compounds in their reactions. The final network is binary, with only entries accepted to be 0 and 1, where 1 indicates the existence of an edge between the nodes.

Bibliography

1. Martiny, J. B. H. *et al.* Microbial biogeography: putting microorganisms on the map. *Nat. Rev. Microbiol.* **4**, 102–112 (2006).

2. Bar-On, Y. M., Phillips, R. & Milo, R. The biomass distribution on Earth. *Proc. Natl. Acad. Sci.* **115**, 6506–6511 (2018).

3. Sheik, A. R., Muller, E. E. L. & Wilmes, P. A hundred years of activated sludge: time for a rethink. *Front. Microbiol.* **5**, (2014).

4. Murphy, D. J. The dynamic roles of intracellular lipid droplets: from archaea to mammals. *Protoplasma* **249**, 541–585 (2012).

5. Raunkjær, K., Hvitved-Jacobsen, T. & Nielsen, P. H. Measurement of pools of protein, carbohydrate and lipid in domestic wastewater. *Water Res.* **28**, 251–262 (1994).

6. Chen, J. *et al.* Economic assessment of biodiesel production from wastewater sludge. *Bioresour. Technol.* **253**, 41–48 (2018).

7. Winkler, M. K. & Straka, L. New directions in biological nitrogen removal and recovery from wastewater. *Curr. Opin. Biotechnol.* **57**, 50–55 (2019).

8. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).

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

10. Poretsky, R. S. *et al.* Analysis of Microbial Gene Transcripts in Environmental Samples. *Appl. Environ. Microbiol.* **71**, 4121–4126 (2005).

11. Muller, E. E. L. L. *et al.* Community-integrated omics links dominance of a microbial

generalist to fine-tuned resource usage. *Nat. Commun.* **5**, 5603 (2014).

12. Roume, H. *et al.* Comparative integrated omics: identification of key functionalities in microbial community-wide metabolic networks. *npj Biofilms Microbiomes* **1**, 15007 (2015).

13. Holter, N. S. *et al.* Fundamental patterns underlying gene expression profiles: Simplicity from complexity. *Proc. Natl. Acad. Sci.* **97**, 8409–8414 (2000).

14. Ricotta, C. *et al.* Measuring the functional redundancy of biological communities: a quantitative guide. *Methods Ecol. Evol.* **7**, 1386–1395 (2016).

15. Dee, L. E. *et al.* When Do Ecosystem Services Depend on Rare Species? *Trends Ecol. Evol.* **34**, 746–758 (2019).

16. Smith, S., Witkowski, A. & Joshi, A. K. Structural and functional organization of the animal fatty acid synthase. *Prog. Lipid Res.* **42**, 289–317 (2003).

17. White, S. W., Zheng, J., Zhang, Y.-M. & Rock, C. O. THE STRUCTURAL BIOLOGY OF TYPE II FATTY ACID BIOSYNTHESIS. *Annu. Rev. Biochem.* **74**, 791–831 (2005).

18. Byers, D. M. & Gong, H. Acyl carrier protein: structure–function relationships in a conserved multifunctional protein family. *Biochem. Cell Biol.* **85**, 649–662 (2007).

19. Fisher, O. S. *et al.* Characterization of a long overlooked copper protein from methane- and ammonia-oxidizing bacteria. *Nat. Commun.* **9**, 4276 (2018).

20. Chandran, K., Stein, L. Y., Klotz, M. G. & van Loosdrecht, M. C. M. Nitrous oxide production by lithotrophic ammonia-oxidizing bacteria and implications for engineered nitrogen-removal systems. *Biochem. Soc. Trans.* **39**, 1832–1837 (2011).

21. Muller, E. E. L. *et al.* First draft genome sequence of a strain belonging to the Zoogloea genus and its gene expression in situ. *Stand. Genomic Sci.* **12**, 64 (2017).

22. Narayanasamy, S. *et al.* IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome Biol.* **17**, 260 (2016).

23. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).

24. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).

25. Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11**, 431 (2010).

26. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–

D462 (2016).

27. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

28. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

29. Wu, M. & Scott, A. J. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics* **28**, 1033–1034 (2012).

30. Laczny, C. C. *et al.* Identification, Recovery, and Refinement of Hitherto Undescribed Population-Level Genomes from the Human Gastrointestinal Tract. *Front. Microbiol.* **7**, (2016).

31. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

ISBN: 978-82-575-1696-3
ISSN: 1894-6402

Norwegian University
of Life Sciences

Postboks 5003
NO-1432 Ås, Norway
+47 67 23 00 00
www.nmbu.no

106685 / ANDVORDGRAFISK.NO