

Identification of novel biosynthetic gene clusters and bacterial taxa from a soil microbiome using metagenomic approaches

by

Alinne L. R. Santana-Pereira

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama

May 1, 2021

Keywords: Metagenomics, Functional Metagenomics, Shotgun Metagenomics, Next Generation Sequencing, Biosynthetic Gene clusters, Soil Bacterial Diversity, Bioinformatics

Copyright 2021® by Alinne L. R. Santana-Pereira

Approved by

Mark R. Liles, Chair, Professor of Biological Sciences
Paul A. Cobine, Co-chair, Professor of Biological Sciences
Laurie S. Stevison, Assistant Professor of Biological Sciences
Virginia Davis, Professor of Chemical Engineering

Abstract

The microbial world is home to incredible taxonomic and metabolic diversity. Microorganisms are prolific producers of secondary metabolites, which are an important source of bioactive compounds, namely antibiotics, which are encoded by biosynthetic gene clusters (BGCs). The majority of the microbial diversity cannot be cultured and thus their taxonomic and metabolic diversity need to be addressed via culture independent methods. Metagenomics provides a platform to explore microorganisms yet-to be-cultured. The study of metagenomic BGCs rely on cloning environmental DNA into metagenomic libraries and screening target genes or pathways of interest using DNA amplification. However, this approach is hypothesized to overlook the diversity of BGCs. In order to expand the current ability to recover BGCs from metagenomic libraries, a new platform for targeted screening was developed using next generation sequencing (NGS), in conjunction with bioinformatic algorithms, allowing the sequencing of an entire metagenomic library and assignment of assembled sequences to their clones of origin. Then the NGS approach was validated and contrasted to the traditional amplification approach, showing that the NGS approach increased the number and diversity of BGCs and taxonomic markers detected. Furthermore, the NGS strategy was applied to fully annotate pathways prior to selection offering an innovative platform to guide prospecting of metagenomic libraries. Taxonomic markers from the metagenomic library indicated the presence of bacteria affiliated with the Candidate Phyla Radiation (CPR), a newly described group believed to not colonize soils. In order to study these ultra-small members of the soil microbiome, soil samples for a crop rotation experiment were collected from plots under controlled nutrient deficiency. A protocol was developed to enrich for ultra-small cells from soil samples, which were shotgun sequenced. Genomes from members of

the ultra-small fraction were reconstructed to draft quality, showing a predominance of CPR. The MAGs were used to infer their complex carbohydrate degradation potential and reconstruct the community profiles across all soil samples to evaluate their responses to nutrient levels and pH in soils. Soil dwelling CPR were differentiated taxonomically and showed enrichment of genes associated with plant biomass degradation. No correlation between nutrient deficiency and community composition was found, but instead the community responded to changes in pH and metal levels in soil. Hence, they might play relevant roles in the soil microbiome while stressing the potential for metabolic and taxonomic diversity yet-to-be discovered in soil environments. Another relevant source of antibiotics are synthetic agents such as carbon. However, their ability to induce bacterial death is unclear, mainly due to the lack of appropriate experimental conditions to assess bacterial viability, namely the use of biologically incompatible dispersing agents and extraneous incubation conditions. To elucidate their toxicity, TSB growth medium, was employed as a dispersing agent for the first time and used to evaluate carbon nanotube toxicity against gram-positive and gram-negative bacteria, alongside other commonly used dispersing agents, such as surfactants. Bacterial cells are able to resist nanotube stress in biologically compatible media, however death is observed in the presence of sensitizing agents such as surfactants. Thus, carbon nanotubes, although unable to drive antimicrobial activity on their own are synergistic agents that can be used in conjunction with other compounds to increase antibiotic activity.

Acknowledgments

Despite all the words here in contained, I feel there are not enough words to truly express the depth of my gratitude, but an attempt is at the very least necessary.

I would like to start by thanking my advisor Dr. Mark Liles for his invaluable and compassionate guidance throughout my PhD. Thank you for being the example of a passionate researcher I strive to become. Likewise, the members of my doctoral committee, their input and expertise were also invaluable through the years. Additionally, I want to acknowledge all collaborators I have worked with throughout these years, and all the members of Liles lab, present and past, for providing a constructive, supportive, and often times simply delightful workplace. Go Power Grads! Thank you for making this possible.

A special thank you to all the siblings I found in Auburn. They have become, each on their time, part of my new foreign family and I part of theirs. Although our paths always seem to depart way too soon, their friendship has empowered me will continue to do so. Also, to the friends left behind, who are ever present despite the thousands of miles in our way. You all were, and are, incredibly important to me. Thank you for making this possible.

I want to express my deepest gratitude to my family whose unwavering support has always been an endless source of courage and inspiration. To my brother, who is a tonic that carried me through to many a time. Finally, I want to not only acknowledge, but also dedicate this work to the two giants on whose shoulders I stand: my mother and my father. They moved mountains so I could be here today. Thank you for making this possible.

This achievement is not my own, but one I gladly share with all the special people in my life. I hope you can be as proud of it as I am of you.

“Two roads diverged in a yellow wood,
And sorry I could not travel both
And be one traveler, long I stood
And looked down one as far as I could
To where it bent in the undergrowth;

Then took the other, as just as fair,
And having perhaps the better claim,
Because it was grassy and wanted wear;
Though as for that the passing there
Had worn them really about the same,

And both that morning equally lay
In leaves no step had trodden black.
Oh, I kept the first for another day!
Yet knowing how way leads on to way,
I doubted if I should ever come back.

I shall be telling this with a sigh
Somewhere ages and ages hence:
Two roads diverged in a wood, and I—
I took the one less traveled by,
And that has made all the difference.”

— **Robert Frost**

Table of Contents

Abstract.....	ii
Acknowledgments.....	iv
Table of Contents	vi
List of Tables.....	x
List of Figures	xi
List of Abbreviations.....	xviii
Chapter I: Introduction and Literature Review	1
1.1. Metagenomics	3
1.1.2. Metagenomic Libraries and Functional Metagenomics.....	3
1.1.3. Direct shotgun sequencing of eDNA	5
1.1.4. Genome resolved metagenomics	7
1.2. CPR: the candidate phyla radiation.....	9
1.3. Biotechnological potential of microorganisms: reservoirs of Natural Products ...	12
1.3.1. Secondary Metabolites.....	14
1.3.1.1. Non-Ribosomal Peptide (NRPs)	14
1.3.1.2. Polyketides	15
1.3.2. Assessing environmental BGCs through culture independent methods	17
1.3.2.1. Building metagenomic Libraries.....	18
1.3.2.2. Functional Screening of Metagenomic Libraries.....	21
1.3.2.3. Sequence-based screening of metagenomics libraries	22
1.4. Single walled carbon nanotubes as antimicrobial agents.....	23

Chapter 2: Discovery of novel biosynthetic gene cluster diversity from a soil metagenomic library

.....	27
2.1. Abstract	27
2.2. Introduction.....	28
2.3. Materials and Methods.....	30
2.3.1. Bacterial strains.....	30
2.3.2. Soil DNA isolation and metagenomic library construction	31
2.3.3. Metagenomic library pooling and sequencing.....	31
2.3.4. Assembly <i>de novo</i> of metagenomic contigs.....	32
2.3.5. Library deconvolution from contig to clone	33
2.3.6. Statistical analysis of library sequencing data	33
2.3.7. AntiSMASH prediction of BGCs from library contigs.....	33
2.3.8. Comparison of conserved domains from PKS and NRPS clusters with domains from databases	35
2.3.9. Taxonomic origin of PKS/NRPS containing library contigs	35
2.3.10. Re-sequencing identified clones, annotation, and taxonomic origin	36
2.3.11. Screening libraries for PKS clusters via PCR.....	36
2.3.12. Identification of 16S ribotypes.....	37
2.3.13. PCR amplification and sequencing of 16S rRNA genes from the Cullars metagenomic library	38
2.3.14. PCR amplification and sequencing of 16S rRNA genes from Cullars soil.....	38
2.4. Results.....	39
2.4.1. PCR amplification and sequencing of KS domains from the library.....	39

2.4.2. NGS of the library using a 3D pooling strategy	40
2.4.3. Mining the library for BGCs.....	43
2.4.4. Diversity of NRPS and PKS clusters identified from the library	44
2.4.5. Recovery of PKS and NRPS clusters from the metagenomic library.....	49
2.4.6. Phylogenetic Diversity of Cullars soil and the Cullars soil metagenomic library.....	52
2.5. Discussion	55

Chapter 3: Enrichment of ultra-small cells from topsoil sheds light into the unique characteristics

of soil dwelling CPR.....	60
3.1. Abstract	60
3.2. Introduction.....	61
3.3. Methods.....	63
3.3.1. Sampling site.....	63
3.3.2. Sample collection and soil analysis	63
3.3.3. Ultra-small cell enrichment.....	64
3.3.4. Shotgun Sequencing, Binning, and Genome Annotation.....	65
3.3.5. MAG taxonomic classification	66
3.3.6. CPR robust taxonomic classification	66
3.3.7. CPR MAGs carbohydrate degradation profiles	67
3.3.8. MAG count estimation.....	68
3.3.9. MAG community abundance and distribution across Cullars Plots	68
3.4. Results.....	69
3.4.1. Recovering ultra-small bacteria genomes from Cullars soil.....	69
3.4.2. Taxonomical Classification of Cullars MAGs	70

3.4.3. Carbohydrate Degradation potential of soil dwelling CPR	72
3.4.4. Distribution of the ultra-small bacteria in Cullars soil.....	73
3.5. Discussion.....	77
Chapter 4: Single-walled carbon nanotube antibacterial activity requires cell-wall sensitizing factors.....	83
4.1. Abstract	83
4.2. Introduction.....	84
4.3. Materials and Methods.....	86
4.3.1. SWNT dispersion preparations	86
4.3.2. Evaluation of SWNT dispersions.....	86
4.3.3. Bacterial Strains and cultivation	87
4.3.4. Bacterial Culture challenges	88
4.3.5. Bacterial viability measurements	88
4.3.6. Statistical analysis.....	89
4.4. Results.....	90
4.4.1. SWNT dispersions	90
4.4.2. SWNT effects on Bacterial Viability	93
4.5 Discussion.....	97
Chapter 5: Conclusions.....	101
References.....	107
Appendices.....	128

List of Tables

Table 1: List of oligonucleotide sequences used in this study.	37
Table 2: Biosynthetic gene clusters identified from the soil metagenomic library by PCR and NGS approach. “NGS Screening Deduplicated” refers to the corrected hit count after hits from contigs deconvoluted to the same clone of origin got deduplicated into a single hit. *Hybrid Pathways included all clusters of mixed types except T1PKS-NRPS.	44
Table 3: Phylogenetic marker genes used in this study.	67
Table 4: Initial concentrations and sonication protocols from each dispersion based on literature.	87
Table 5: SWNT concentrations, individual to bundled SWNT ratio, and Average length of individual SWNT for each dispersion.	92
Supplemental Table 1: Validated clones containing PKS and/ or NRPS pathways.	128

List of Figures

Figure 1: Pooling strategies used in sequencing Cullars metagenomic library. a. The Cullars metagenomic library composed of 50 384-well plates got divided into 5 sets of 10 plates that are pooled for Plate x Row x Column individually, for a total of 210 pools. b. Initial pooling strategy used in Set 5. All clones in each plate were combined to generate a plate pool, all clones from each row across the 10 plate set were combined to generate a row pool, and finally all clones from each column across the 10 plate set were combined to generate a column pool. Set 5 generated 50 pools, that were individually barcoded and sequenced. c. Second Pooling strategy applied to Set 1, 2, 3 and 4 individually. For each Set, each 384-well plate was arrayed into quadrants composed of 96 clones that got replicated into 96-well plates. Thus the initial 10 384-well plates got divided into 4 subsets of 10 96-well plates. All subsets are combined to generate the pools. All clones in each 96-well plate were combined to generate a plate pool, all clones from each row across the 96-well plate subsets were combined to generate a row pool, and finally all clones from each column across the 96-well plate subsets were combined to generate a column pool. Each Set from 1 through 4 generated 60 pools that were individually barcoded and sequenced. 34

Figure 2: Library construction and 3D pooling. a. Library construction. Superscript letters correspond to citations: a: (Liles et al., 2008); b: (Kakirde et al., 2011); c: (Nasrin et al., 2018). b. 3D Library Pooling strategy: Plate, Column, and Row are barcoded before sequencing. c. Contig to clone deconvolution and in silico mining. Reads from each pool are assembled separately into pool contigs. Highly homologous contigs from each dimension (Plate, Column, and Row) form triplets. Since their pool of origin is known, they form a coordinate system that points to their clone

of origin. Contigs can be analyzed bioinformatically for rational identification of candidates for functional screening.....41

Figure 3: Plots of contig Length x Coverage, both indicated as a log transform. No increase in contig length was observed past 55x coverage ($T_{coverage}=4$), with a very strong influence in final contig length up to 25X. a. Length x Coverage separated by SET (groups of 10 Library plates). b. Length x Coverage separated by deconvolution status. Higher amount of long contigs amongst the completely deconvoluted set. c. Basic statistics of contig lengths according to deconvolution status. 42

Figure 4: Maximum likelihood tree of KS domains recovered from the soil metagenomic library. Clades in magenta are formed uniquely by domains recovered in silico. a. KS domains recovered either by PCR or in silico (NGS). b. KS domains recovered from the soil metagenomic library and from the MIBiG database showing branch lengths. c. KS domains recovered from the soil metagenomic library and from the MIBiG database ignoring branch lengths to facilitate topography visualization. Circles in the tree represent branch bootstrap support > 70%. 45

Figure 5: Maximum likelihood tree of A-domains recovered from NRPS clusters identified from the soil metagenomic library and the MIBiG database. a. ML Tree showing branch lengths. b. ML Tree ignoring branch lengths to facilitate topography visualization. Circles in the tree represent branch bootstrap support > 70%. 46

Figure 6: Clustering of PKS (a) and NRPS (b) pathways against the MIBiG database using BiG-SCAPE. Library pathways are shown in orange and database pathways in blue. Circle size is proportional to the number of connections, and strike width is proportional to the similarity between pathways. 47

Figure 7: Taxonomic classification of the Contigs carrying PKS and/or NRPS pathways by k-mer frequencies. 48

Figure 8: Analysis of the clones containing viable PKS and NRPS pathways reconstructed and validated from the Cullars metagenomic library. a. Viable PKS and/or NRPS clusters recovered from the library compared by size (number of domains) and divergence from known BGCs (mean branch length and domain mean divergence). Each plot point represents a clone containing a viable BGC. Mean branch length for each clone is calculated as the average branch length of all the clone's KS and/or A domains, as extracted from the ML trees against the MIBiG database. Mean domain divergence for each clone is calculated as the average domain divergence of all clone's KS and/or A domains. Domain divergence refers is to the complementary percentage to the % Identity of the domain to the NCBI nr database (if a domain has % identity of 60, then the domain divergence is 40%). b. Density distribution of branch length, domain divergence and number of domains across the viable PKS and NRPS BGCs. c. Three interesting representatives, being the longest BGC from each type (PKS, NRPS, PKS-NRPS) to be recovered successfully in silico from the metagenomic library NGS contigs and validated upon clone resequencing. Annotation from the library contig was compared to the annotation from their corresponding validated insert sequence. ORFs are showed by arrows and the domains within each ORF are represented by colored stripes. 51

Figure 9: Taxonomic diversity of the Cullars metagenomic library and of the original soil sample used for library construction. Soil PCR: 16S rRNA gene PCR amplification of DNA isolated from the original Cullars soil sample used to construct the library; Library PCR: 16S rRNA gene PCR amplification of the DNA template from all metagenomic library clones pooled together; Library NGS: In silico mining for 16S rRNA genes from the metagenomic library NGS contigs. The phyla

contained in the section “Other phyla” are: RCP2-54, SAR324, Spirochaetes, Sumerlaeota, WPS-2; all with relative frequencies below 0.05%. 54

Figure 10: Representation of Cullars Rotation plots and Fertilization treatments. a. Schematic representation of Cullars Rotation Plot layout and Soil treatments administered. pH corr: pH correction with lime; mi Nut: Addition of micronutrient mix; S: Addition of Sulfur; P: Addition of Phosphorous; K: Addition of Potassium; L: Rotation with winter legumes; N: Addition of Nitrogen. Fractions represent the amount added to the plot with Standard Fertilization as reference (3/3). b. Table with the Standard Fertilization treatments applied to each plot.(Mitchell et al., 2005). c. Average soil nutrient parameters and pH from sampled bulk soil. 64

Figure 11: MAGs from the ultra-small bacteria fraction. a. MAGs completeness, strain heterogeneity and contamination (%). Red dashed lines represent the thresholds for keeping MAGs for subsequent analysis: 10%: maximum strain heterogeneity or contamination threshold; 65%: minimal completeness threshold. b. Genome size of Cullars MAGs. 70

Figure 12: Classification of all Cullars MAGs and the proportion of each phyla. *Total MAGs does not include MAGs filtered out by the completeness, strain heterogeneity and contamination thresholds. Concatenated maximum likelihood tree of 13 phylogenetic marker genes from Cullars CPR and 920 reference genomes allowed further classification of CPR MAGs. MAGs were classified as Saccharibacteria, ABY1 and Parcubacteria. The Parcubacteria phylum is expanded to show Classes since it had the highest number of Cullars MAGs and due to its intrinsic diversity. Zoomed in regions offers context to MAGs placement. Green dots denote Cullars MAGS. Boxes indicate clades formed only by Cullars MAGs and unclassified references. 71

Figure 13: Carbohydrate degradation potential of soil dwelling CPR in Cullars plots (MAG) and

of the associated reference genomes (REF), based on glycoside hydrolase (GH) families annotated from the CAZy database. Left panel is a presence/absence map of the annotated GHs across Cullars CPR (MAG) and the reference genomes separated by Phyla, with Parcubacteria (red) further separated into classes. Gemmatimonadetes were also included due to their correlation with Saccharibacteria abundances. Adler: Adlerbacteria (Parcubacteria); Kaiser: Kaiserbacteria (Parcubacteria); Nomura: Nomurabacteria (Parcubacteria); Taylor: Taylorbacteria (Parcubacteria); Uhr: Uhrbacteria (ABY1), Sacchari: Saccharibacteria; Gemmati: Gemmatimonadetes. Right panel shows the abundance of GHs according to their substrate specificity. Mixed: GH families with undefined specificity. OPP : GH families predicted to break other carbohydrates derived from plant biomass. Abundance describes the combined counts of identified GHs divided by the number of MAGs in each group. 73

Figure 14: MAG community reconstruction from Cullars soil at different plots. a. Soil macro-nutrient levels, pH. b. corresponding MAG average relative abundance, ranked by Phyla. Plots are organized by Varying nutrient/treatment: Plot 03: Positive control, Plot C: Negative control, Plot 8: No pH correction, Plot 10: Micronutrients, Plot 09: S varies; Plots 02/05: P varies; Plots B/A/01: N varies; Plots 06/11/07/04: K varies. 74

Figure 15: MAG absolute abundance across Cullars Rotation plots. a. Heatmap of MAG absolute abundance across Cullars plots for each triplicate, highlighting MAG specific fluctuations and sampling sparsity. C: Chlamydiae; D: Dependientiae; G: Gemmatimonadetes; Sacchari: Saccharibacteria; U: Uhrbacteria. b. Rarefaction curves for each one of the sampled plots from Cullars. 76

Figure 16: Ordination plots of all sample communities with Bray-Curtis distances. Colors

emphasize the varying nutrients and pH on the plots. Ctrl +: Plot 3 (Standard Fertilization); Ctrl -: Plot C (No fertilization or amendment). Black circles show Plot 8 samples. a. NMDS ordination of relative abundances. b. Canonical Correspondence Analysis and c. NMDS ordination with Al, Fe gradients fitted to the distribution. Purple lines: Aluminum levels; Green lines: Iron levels. Vectors are soil parameters significantly related to community changes ($p < 0.05$). 77

Figure 17: Experimental design of the SWNT antimicrobial challenges. a. Challenging solutions used in the experiment. Cultures were challenged against the dispersant and the dispersant with SWNTs. b. Challenge incubation. Overnight growth cultures of *S. enterica* and MRSA were diluted in fresh TSB to OD600 = 0.5 and incubated with the challenging solution for one hour. c. The challenge incubations were diluted in TSB and plated for CFU counting and grown in broth for growth curve construction. 90

Figure 18: AFM scans quantifying the number of individual versus bundled SWNTs. a) TSB-SWNT b) DNA-SWNT c) LSZ-SWNT, d) Pluronic-SWNT, and e) SDS-SWNT. 91

Figure 19: SWNT concentration analysis using UV-vis spectroscopy. (a) UV-vis spectra of DNA-SWNT dispersions with dispersion:water dilutions and (b) Beer-Lambert plot for DNA-SWNT with extinction coefficient. 92

Figure 20: Antimicrobial activity of different dispersants without SWNTs, normalized SWNT concentration and full-strength SWNT concentration. SWNT concentrations were normalized to the lowest concentration (lysozyme) for comparison. Antimicrobial activity measured as cell viability loss in relation to a water challenge. 94

Figure 21: Optical density growth curves of inoculums a) *S.aureus* (MRSA) and b) *S.*

typhymurium challenged with the 17 treatments into fresh TSB medium. Delayed exponential growths points towards a larger number of sick or unviable cells in the inoculum. SWNT-corrected refer.....96

Supplemental figure 1: Annotation of PKS and NRPS pathways predicted bioinformatically and from resequenced clones. 133

List of Abbreviations

A	Adenylation
ACP	Acyl Carrier Protein
AFM	Atomic Force Microscopy
ANOVA	Analysis Of Variance
AT	Acyl-Transferase
BAC	Bacterial Artificial Chromosome
BGC	Biosynthetic Gene Clusters
BHI	Brain Heart Infusion
BLAST	Basic Local Alignment Search Tool
CCA	Correlation Analysis
CFU	Colony Forming Units
CPR	Candidate Phyla Radiation
DH	Dehydrogenase
DNA	Deoxyribonucleic Acid
E	Epimerization
eDNA	Environmental DNA
GC	Guanine-Cysteine
gDNA	Genomic DNA
GH	Glycoside Hydrolase
HMW	High Molecular Weight
Kb	Kilobase
KR	Ketoreductase
KS	B-Ketoacylsynthase
LB	Luria Broth
MAG	Metagenomically Assembled Genome
Mb	Megabase
ML	Maximum Likelihood

mRNA	Messenger RNA
MRSA	Methicillin-Resistant <i>Staphylococcus Aureus</i>
MT	Methyltransferase
MWNT	Multi Walled Carbon Nanotube
NGS	Next Generation Sequencing
NMDS	Non-Metric Multidimensional Scaling
NRP	Non-Ribosomal Peptide
NRPS	Non-Ribosomal Peptide Synthetase
OD	Optical Density
ORF	Open Reading Frame
OUT	Operational Taxonomic Unit
PBS	Phosphate-Buffered Saline
PCP	Thiolation/Peptidyl-Carrier Protein
PCR	Polymerase Chain Reaction
PERMANOVA	Permutational Multivariate Analysis of Variance
PK	Polyketide
PKS	Polyketide Synthase
PPTase	Phosphopantetheinyl Transferase
Q-Q plots	Quantile-Quantile Plot
RuBisCOs	Ribulose-1,5-Bisphosphate Carboxylase-Oxygenase
SDS	Sodium Dodecyl Sulfate
SSU rRNA	Small Subunit Ribosomal Ribonucleic Acid
SWNT	Single Walled Carbon Nanotube
TE	Thioesterase
TSA	Tryptic Soy Agar
TSB	Tryptic Soy Broth

Chapter I: Introduction and Literature Review

Historically, microbiology had been limited to the study of microorganisms that could be cultured in the laboratory. Despite, this limitation, culture-dependent approaches generated a wealth of knowledge and shed light into microbial physiology, its roles in the environment from host-microbe interactions to biogeochemical cycles (Handelsman, 2004) as well as mining their metabolic potential in the form of natural products (Berdy, 2005).

However, the great plate count anomaly – the discrepancy between observed diversity through microscopy techniques and the organisms recovered through cultivation – already indicated that current culturing techniques were insufficient to grow every microorganism from an environment (Staley & Konopka, 1985). It wasn't until the introduction of technologies such as PCR and DNA sequencing, that microbiologists were able to circumvent cultivation to study the yet-unculturable microbiota.

The pioneering application of these technologies was seminal in establishing the use of the 16S rRNA (SSU rRNA) gene as an informative phylogenetic marker (Woese & Fox, 1977), leading to the revolutionary proposition of the 3 domains of life model (Woese *et al.*, 1990). Since then, 16S rRNA has been foundational to bacterial and archaeal taxonomy of both cultivated and uncultivated organisms (Lane *et al.*, 1985; Yarza *et al.*, 2014) with its own dedicated databases (Cole *et al.*, 2014; McDonald *et al.*, 2012; Quast *et al.*, 2012) and classification pipelines (Bolyen *et al.*, 2019; Pruesse *et al.*, 2012; Schloss, 2020; Wang *et al.*, 2007). This robust molecular classification tool allowed the survey of microbial diversity by mining for 16S rRNA genes directly from environmental genetic material without the need of cultivation, revealing that an

unknown 99% of the microorganisms were not recovered by available culturing techniques (Amann *et al.*, 1995).

The use of the 16S rRNA gene to classify the microbiota directly from an environment via targeted PCR, commonly referred as amplicon based surveys, led to an explosion of the catalogued microbial diversity including at phylum level (Dojka *et al.*, 1998; Hugenholtz, P. *et al.*, 1998; Kuske *et al.*, 1997; Ohkuma & Kudo, 1996; Rheims *et al.*, 1996). Indeed, since the first 16 rRNA based classification in the late seventies, the number of recognized bacterial phyla increased from 12 to 30 by the turn of the century (Pace, 1997). In the 21st century, the advent of next generation sequencing (NGS) platforms such as Roche 454 (now obsolete) and Illumina, massively expanded the depth of 16S rRNA sequences recovered from amplicon based surveys from a few dozens to thousands (Tringe & Hugenholtz, 2008), leading to an exponential increase of available sequences (Pace, 2009) and recognized operational taxonomic units (OTUs) (Schloss & Handelsman, 2004). More than half of the now known bacterial phyla diversity has no cultured representative, yet these candidate phyla are diverse and shrouded in mystery (Solden *et al.*, 2016), being referred to as the “microbial dark matter” (Rinke *et al.*, 2013). For some of these, representatives were later cultivated such as Acidobacteria (Quaiser *et al.*, 2003), Elusimicrobia (Herlemann *et al.*, 2009) and Saccharibacteria (He *et al.*, 2015). The discovery of this uncultured majority established culture independent approaches as a cornerstone of microbiology moving forward.

Despite the wealth of contribution towards understanding microbial diversity and distribution, amplicon-based surveys are not without limitations. Amplicon based surveys cannot provide functional insights into the taxonomic entities it lists, limiting its power in ecology and biotechnology endeavors (Handelsman, 2004). Moreover, it relies on “universal primers” capable of amplifying the majority of bacterial 16 rRNA genes. However, the available primer pairs often

show bias towards specific bacterial groups (Klindworth *et al.*, 2013), severely underestimate the often diverse rare biosphere (taxa with relative abundance below 1%) (Lynch & Neufeld, 2015), or worse yet, these primers are estimated to not amplify at least 5% the now known bacterial diversity altogether creating diversity “blind spots” (Eloe-Fadrosh, Ivanova, *et al.*, 2016).

1.1. Metagenomics

Meanwhile, another field strived to apply genomics to the study of uncultured organisms going beyond surveying the 16 rRNA gene in isolation, but rather trying to capture the genomic information of all microorganisms present in a sample and thus overcoming the inherent limitations of amplification-based and culture-dependent approaches (Handelsman, 2004). Thus the field of Metagenomics – the simultaneous analysis of a collection of genomes – was born (Handelsman *et al.*, 1998). Metagenomics offers the potential to not only answer the question of “Who is there” but also “What are they doing”, by reconstructing the metabolic capabilities present in a community (Rondon *et al.*, 2000; Schloss & Handelsman, 2003) and later even assign them to particular taxa (Sharon & Banfield, 2013), greatly expanding our understanding of the microbial dark matter (Solden *et al.*, 2016).

Metagenomics has evolved over the decades since its introduction at the turn of the century (Handelsman *et al.*, 1998), to encompass different technologies and approaches each tailored to different research objectives and with their own strengths and weaknesses. This section outlines the main metagenomic technologies roughly in the order of their introduction, since many groups have used them concomitantly.

1.1.2. Metagenomic Libraries and Functional Metagenomics

The very first iteration of metagenomics relied heavily on the construction of metagenomic libraries (Rondon *et al.*, 2000; Streit & Schmitz, 2004). A metagenomic library consists of random

pieces of environmental DNA (eDNA) which are attached to shuttle vectors, and then inserted into a host cell such as *Escherichia coli* (Rondon *et al.*, 2000). Because each cell carries only one environmental insert, the different clonal communities (clones) become a container for a particular eDNA fragment which can be accessed later. Metagenomic libraries can range from hundreds to thousands of clones produced in this manner, varying in regard to sampled representativeness compared to the original eDNA pool (Handelsman, 2004).

The fate of cloned eDNA depends on the research questions, and libraries can be revisited indefinitely in further studies. Studies focusing on characterizing functional aspects of a community often rely on heterologous expression of the eDNA insert and subsequent screening for activities of interest, thus being called Functional Metagenomics (Schloss & Handelsman, 2003). This approach is discussed in finer detail focusing on natural product discovery in the section “Functional Screening of Metagenomic Libraries” below. The biotechnological potential of this approach is at the heart of the establishment of the field of metagenomics (Schloss & Handelsman, 2003) and over the years has led to the discovery of novel antibiotics (Hover *et al.*, 2018; Kallifidas *et al.*, 2012; Kang & Brady, 2013; Kang & Brady, 2014a; Nasrin *et al.*, 2018), bioactive compounds (Chang & Brady, 2013; Owen *et al.*, 2015), thermostable cellulases (Lewin *et al.*, 2017), without the need for cultivation.

Clones displaying some activity of interest can then be sequenced and annotated, which has led to the discovery of new pathways and genes (Santana-Pereira *et al.*, 2020). Another early advantage of metagenomic libraries is the possibility to taxonomically classify a clone of interest based on the presence of phylogenetic markers in the insert (Liles *et al.*, 2003). A great example of the power of this approach was the discovery of the first proteorhodopsin, revealing a new and widespread type of phototrophy in oceanic bacteria previously only described in archaea (Beja *et*

al., 2000; Beja *et al.*, 2001). Alternatively, clones can be randomly sequenced offering insight into the metabolic potential of the community as a whole and allowing the inference of enriched functions and their ecological significance. This approach led to the reconstruction of the community and their relationships on an acid mine drainage biofilm (Tyson *et al.*, 2004) and 148 previously undescribed ribotypes, 1.2 million previously unknown genes, including 728 rhodopsin-like genes from a sargasso associated metagenome (Venter *et al.*, 2004).

Naturally, each one of the outlined steps can introduce their own set of limitations. For example, DNA extraction technique may favor the sampling of some bacterial groups in relation to others, overrepresenting them in the final library (Liles *et al.*, 2008). Shuttle vector of choice affects the size of the insert captured, which can truncate pathways and inflate estimates of abundance or hinder heterologous expression (Kakirde *et al.*, 2010). Additionally, heterologous expression is highly dependent on host machinery, and host metabolism can interfere with screening (Gomez-Escribano & Bibb, 2014). All these issues are even more pressing in the application of functional metagenomics for natural product discovery and will be discussed at length below.

1.1.3. Direct shotgun sequencing of eDNA

The field of metagenomics would be revolutionized with the advent of NGS. These new technologies allow the parallel sequencing of enormous amounts of DNA and consequently, generates enormous amounts of data. Therefore, NGS opened the doors for the direct sequencing of metagenomic DNA without the need of capturing it in metagenomic libraries first, increasing the sampling power of metagenomics while forfeiting the ability to chemically characterize and prospect products via heterologous expression. In order to study the metagenome, first, sequenced fragments (reads) need to be assembled in larger contiguous sequences (contigs) – and even

combined into scaffolds – which are then mined for features of interest using bioinformatic tools.

Theoretically, if all reads from a sequencing run could be fully assembled, each contig would correspond to one population composite genome from the metagenome. However, NGS shotgun sequencing – the untargeted sequencing of a DNA sample (Sharpton, 2014) – imposes a conundrum: it produces a vast amount of reads but they are very small. These reads are challenging to piece together and the task is further complicated by the presence of repetitive regions, mutations and heterogeneity which are escalated when assembling metagenomes (Ayling *et al.*, 2020). Therefore, a massive amount of biological information might be lost as the majority of reads remain unassembled (Howe *et al.*, 2014).

The rise and popularization of NGS prompted the development of increasingly more sophisticated bioinformatic tools to handle the complex and large datasets generated by shotgun metagenomics. Robust advancements made with de Bruijn graph-based genome assembly and metagenomic specific assemblers greatly improved the reconstruction of contigs to cover substantial portions of individual genomes (Ayling *et al.*, 2020). Alongside, other bioinformatic tools were developed/adopted to predict and mine genes of interest (Blin *et al.*, 2019; Eddy, 2011; Hyatt *et al.*, 2010; Zhang *et al.*, 2018). Community composition can be surveyed using 16S rRNA genes mined directly from contigs (Guo *et al.*, 2016; Nawrocki & Eddy, 2013), or using 16S specific assemblers (Yuan *et al.*, 2015). Tetranucleotide profiling allows the taxonomic classification of sequences in the absence of any phylogenetic marker genes (Dick *et al.*, 2009), expanding the reconstruction of ecological functions within a community (Gregor *et al.*, 2016; Wood & Salzberg, 2014).

These developments shed light on the wealth of microbiological metabolic diversity and microbiota interactions in varying environments such as soils (Edwards *et al.*, 2006; Eyice *et al.*,

2015; Howe *et al.*, 2014; Leff *et al.*, 2015), oceans (Biddle *et al.*, 2008; Sunagawa *et al.*, 2015) and host organisms (Brulc *et al.*, 2009; Palenik *et al.*, 2009; Wegley *et al.*, 2007), leading to the discovery of several novel genes, proteins and operational taxonomic units (OTUs). Most notably, this approach revealed the diversity and complexity of human associated microbiota (Almeida *et al.*, 2019; Human Microbiome Project, 2012; Qin *et al.*, 2010), revealing their association with human physiology (Kim *et al.*, 2015) and clinical conditions (Morgan *et al.*, 2012; Qin *et al.*, 2012; Turnbaugh *et al.*, 2006).

1.1.4. Genome resolved metagenomics

The advancement of bioinformatic tools and sequencing power has brought metagenomics full circle, with the increasing ability to reconstruct and analyze individual genomes from metagenomes, by reconstructing metagenomically assembled genomes (MAGs) (Chen *et al.*, 2020). This approach is key to link uncultured diversity to microbiome metabolic potential, marking the roles of individual populations in community dynamics and delineating the distribution of metabolic capabilities across microbial life (Sharon & Banfield, 2013). Furthermore, it allows better taxonomic placement by enabling the construction of phylogenetic trees based on concatenated marker genes (Hug *et al.*, 2016)

Traditionally, genomes are constructed from sequencing clonal lineages of cultured microorganisms or, more recently, through sequencing the genome of individual cells through single-cell genomics, which is a valuable culture independent tool in its own right that has led to the discovery of new bacterial and archaeal diversity (Becraft *et al.*, 2017; Elie-Fadrosh, Paez-Espino, *et al.*, 2016; Rinke *et al.*, 2013). Different from these approaches, MAGs are a composite from the genomes of closely related organisms, thus representing a population genome (Sangwan *et al.*, 2016). Thus, MAGs usually flatten out intra-population diversity (Luo *et al.*, 2012), unless

a combination of low complexity community and extensive sequencing power is applied (Brooks *et al.*, 2017). However, the very nature of metagenomics allows the resolution of strain specific polymorphisms by comparing contigs or mapping reads to MAGs, highlighting regions of variation (Simmons *et al.*, 2008).

Reconstruction of genomes from metagenomes relies on identifying discriminating signals within sequences that are indicative of shared origin and clustering these sequences in a genomic “bin” (Sangwan *et al.*, 2016). The first individual genomes from metagenomes were reconstructed from a low complexity acid mine drainage biofilm, where contigs were binned based on coverage and GC content to form two near complete genomic bins (Tyson *et al.*, 2004). Indeed, in very low complexity communities, one signal alone might be enough to correctly bin sequences and resolve genomes (Chivian *et al.*, 2008), but in most cases successful binning draws from a combination of signals (Chen *et al.*, 2020).

The introduction of tetranucleotide frequencies (Dick *et al.*, 2009) and differential coverage across multiple metagenomes (Albertsen *et al.*, 2013) as signals have greatly improved the ability to create genomic bins. Additionally, taxonomic affiliation inferred from predicted genes can be used to further separate sequences and assess contamination. Automated binning algorithms such as CONCOCT (Alneberg *et al.*, 2014), MaxBin (Wu *et al.*, 2016) and MetaBat (Kang *et al.*, 2019), analyze a collection of signals but perform differently depending on the dataset, and thus, binning is more robust when final bins are curated from combined outputs (Sieber *et al.*, 2018).

Genomic bins are commonly assessed for completion and contamination using single copy marker genes (Parks *et al.*, 2015) mainly due to its scalability, despite its shortcomings in addressing inconsistencies in GC content and coverage which might be indicative of contamination. Overall, MAGs tend to accrue to draft quality genomes, although they can be

greatly improved by manual curation and even result in full circularized genomes (Banfield *et al.*, 2017). Efforts to standardize best practices for MAG curation (Chen *et al.*, 2020) and submission (Bowers *et al.*, 2017) aim at the generation of consistent MAGs while minimizing the risk of wrong biological conclusions due to contamination or incompleteness.

Genome-resolved metagenomics flourished into a staple of metagenomics revealing novel metabolic pathways such as complete nitrification (comammox) (van Kessel *et al.*, 2015), methane metabolism outside Euryarchaeota (Evans *et al.*, 2015) and the secondary metabolite potential of rare soil taxa (Crits-Christoph *et al.*, 2018). Furthermore, it has been providing an integrated view of microbiomes from several environments such as bioreactors (Kantor *et al.*, 2017), subsurface sediments (Hernsdorf *et al.*, 2017), aquifers (Anantharaman *et al.*, 2016) and even complex environments such as soil (Kroeger *et al.*, 2018; Woodcroft *et al.*, 2018); culminating in the catalog of earth's microbiomes using more than 52,000 MAGs (Nayfach *et al.*, 2020). The possibility of generating concatenated phylogenetic trees including uncultured representatives has comprehensively changed the tree of life (Hug *et al.*, 2016; Parks *et al.*, 2017), in no small part due to the characterization of the Candidate Phyla Radiation (CPR) (Attar, 2016).

1.2. CPR: the candidate phyla radiation

Some members of the later to be the CPR group, such as candidate phyla TM7, OD11, OP1 and SR1, have been identified by 16S rRNA gene surveys (Harris *et al.*, 2004; Hugenholtz *et al.*, 2001), however their ecological roles and phylogenetic placement in the tree of life remained elusive (Pace, 2009). Comprehensive analysis of 789 draft MAGs from enriched aquifer samples showed that those four phyla were in fact members of a hyper diverse monophyletic group, comprised of more than 25 candidate phyla and representing 15% of the domain bacteria, thus granting them the name candidate phyla radiation (Brown *et al.*, 2015). Since then, the CPR has

reshaped the tree of life with not only new phyla but also relevant classes (Hug *et al.*, 2016). Their classification as a super-phyla or phyla is not completely resolved, varying depending on marker set chosen for tree building (Parks *et al.*, 2017). A comprehensive analysis of bacterial taxonomy, using 120 ubiquitous single-copy marker genes, places CPR as a monophyletic phyla and proposes the adoption of the name Patescibacteria to designate it (Parks *et al.*, 2018). Nevertheless, a more recent study using 381 markers still corroborates the placement of CPR as a super-phylum (Zhu *et al.*, 2019), and therefore will be adopted in this dissertation.

The first glimpse at CPRs unusual metabolic potential, was given by genome-resolved metagenomics applied to groundwater samples enriched for small cells by filtration. Genomes from four candidate phyla – Microgenomates (OD11), Parcubacteria (OD1), Gracilibacteria (BD1-5) and a new candidate called Peregrinibacteria (PER), were reconstructed and thoroughly annotated (Wrighton *et al.*, 2012). These organisms were found to have very small genomes (smaller than 1Mbp) with severed biosynthetic potential, lacking complete pathways for nucleotide, fatty-acid and amino-acid synthesis, and predicted to be strictly anaerobic fermenters.

Amongst the particular characteristics of the phyla were: the use of alternate enzymes in glycolysis, alternate means of production of proton-motor force for ATP synthesis, ability to produce H₂ and H₂S and presence of archaeal like ribulose-1,5-biphosphate carboxylase-oxygenase (RubisCO); later shown to be functional and fix CO₂ through a nucleoside salvaging pathway previously only described in Archaea (Wrighton *et al.*, 2016). Later studies with increasing number of MAGs, further corroborated the wide-spread limitation in CPR metabolisms and revealed the presence of intronic 16S rRNA genes (which hinders their detection by amplicon surveys), lack of some ribosomal proteins characteristic of symbiont genomes (Brown *et al.*, 2015), diversified carbon-processing potential and sparse ability to reduce nitrite (Danczak *et al.*, 2017).

These findings suggest that CPR can be key players in bio-geochemical processes (Anantharaman *et al.*, 2016) and carbon turn-over, despite their relatively low abundance (Lynch & Neufeld, 2015).

Due to these characteristics, CPR are predicted to rely heavily on the associated microbiota or specific hosts for survival (Brown *et al.*, 2015). However, CPR metabolism and life-styles are not monolithic. For instance, they have been observed as intra-cellular symbionts of *Paramecium bursaria* (Gong *et al.*, 2014), attached to other bacteria (Luef *et al.*, 2015), and as the extra-cellular symbionts of oral Actinobacteria, leading to the coculture of the Saccharibacterium TM7x alongside *Arachnia propionica* – the only successfully cultured representative to date (Murugkar *et al.*, 2020). Additionally, evidence suggests that some CPR (Saccharibacteria) form symbiont associations with *Hydrotalea sp* in an acid-mine drainage environment (Lemos *et al.*, 2019). However, studies tracking CPR abundances from soil effluent to aquifer failed to identify any specific correlations between members of the CPR and particular hosts, suggesting that these CPR rely on the community and can be supported by different cohorts (Herrmann *et al.*, 2019). Furthermore, CPR metabolism is marked by a patchy distribution of key enzymes across taxa, revealing a varied genetic inventory, indicating that gene loss and acquisition are in part adaptive (Jaffe *et al.*, 2020).

Although the majority of what is known about the CPR comes from aquifers and associated sediments, growing attention has been paid to their association with soil environments. These bacteria have been shown to be mobilized from soil into aquifers (Herrmann *et al.*, 2019) and possess plant-polysaccharide degradation capabilities that overlaps those found in overlying soils, such as cellulose and starch degradation associated with the groups Parcubacteria and Saccharibacteria, respectively (Taubert *et al.*, 2019), although no taxonomic overlap between the organisms performing these roles in soil and associated groundwater was found. However, another

study growing wild oats in the presence of labeled carbon $^{13}\text{CO}_2$ found that a rhizosphere associated *Saccharibacterium* MAG contained ^{13}C , indicating that this organism either consumed plant biomass and/or exudes (Starr *et al.*, 2018). Additionally, its genome was larger than what is observed in groundwater members and was enriched in genes related to complex carbon utilization (Starr *et al.*, 2018). Despite the presence of the genetic repertoire to process plant biomass, it is not yet clear whether or not it is their preferred carbon source or if they act as carbon sinks (Figueroa-Gonzalez *et al.*, 2020). Thus, recovery of more CPR directly from soil environments is paramount to understanding their roles and diversity in this high complexity ecosystem.

1.3. Biotechnological potential of microorganisms: reservoirs of Natural Products

Natural products have been an important source of bioactive compounds thorough the history of humanity. Mankind has always taken advantage of molecules synthesized by other forms of life, remarkably plants, in order to treat diseases, obtain pigments and other useful activities. In the modern days, natural products still play an ever so important role, especially bioactive compounds of microbiological origin.

Since the discovery of Penicillin in 1929, a myriad of drugs have been developed based solely, or inspired by, metabolites synthesized by microorganisms. All the main classes of antibiotics – tetracyclines, cephalosporins, aminoglycosides and macrolides –were isolated from microbes, mainly from actinomycetes (Berdy, 2005). Furthermore, microbial metabolites have been shown to have a myriad of useful medical applications such as anticancer drugs, antivirals, antiparasitic, immunosuppressive, lipid control agents and antidiabetics (Li & Vederas, 2009; Newman & Cragg, 2012), and in agriculture as an important source of pesticides and herbicides (Berdy, 2005).

However, the rate of discovery of novel bioactive compounds has decreased significantly

in the past decade, mainly due to two reasons: the pharmaceutical environment's lack of investment in natural product discovery, and technical limitations in identifying new compounds. The high re-isolation rates have led the pharmaceutical industry to turn to synthetic chemistry and drop its investments in screening endeavors (Li & Vederas, 2009).

Despite the long reliance on culture dependent methods for isolation of bioactive compounds, these approaches have been becoming less efficient as means of drug discovery (Tulp & Bohlin, 2005), and success depends on comprehensively sampling the microbiome diversity. The percentage of cultured bacteria recovered varies from environment to environment, with higher culture counts achieved from eutrophic environments such as gut microbiomes. Novel culture methods that utilize lower nutrient levels and extracts from natural environments can be used to improve the diversity cultured, particularly when culturing on solid media (Schoenborn *et al.*, 2004). Additionally, *in situ* incubation culturing methods are promising (Nichols *et al.*, 2010), and have led to the identification of teixobactin, the first new class of antibiotic in decades (Ling *et al.*, 2015). Despite the advances in culturing techniques, 99% of microorganisms remain recalcitrant to cultivation, thus natural product discovery can be greatly expanded by culture-independent methods.

In this scenario, metagenomic libraries combined with NGS and heterologous expression techniques are powerful tools to access the microbial 'dark matter' and exploit this reservoir of bioactive compounds (Cragg & Newman, 2013), while having a unique set of biases and limitations. In this section, we will discuss the use of metagenomic approaches in mining and expressing secondary metabolites from environmental samples, and how these inherent biases and limitations can be overcome to provide access to novel natural products.

1.3.1. Secondary Metabolites

The majority of bioactive compounds derived from microorganisms are classified as secondary metabolites (Helfrich *et al.*, 2014). Those are produced at the stationary phase and share characteristics such as (O'Brien & Wright, 2011): they are not essential for growth, nor reproduction; their formation is dependent on growth conditions and they are produced as a group of closely related compounds.

The characteristics of secondary metabolites and their overwhelming diversity and complexity, points towards a system that has long adapted to produce molecules that are able to stop or slow down competitors in complex environments (Challis & Hopwood, 2003) or as mediators of cross-species symbiosis with other microorganisms, plants and animals (O'Brien & Wright, 2011).

Secondary metabolites can be classified in many different ways, according to their biological activity, chemical structure or genetic organization (Berdy, 2005). From a metagenomic point of view, a classification based on biosynthetic gene clusters (BGCs) is more appropriate, since secondary metabolites will be dealt with on a genetic level, drawing upon information on genetic organization, active sites, and even prediction of final product structures to guide screening and heterologous expression projects (Cimermanic *et al.*, 2014; Weber *et al.*, 2015).

There are several families of BGCs that can often share genes or even biosynthetic pathways leading to a remarkably complex network (Cimermanic *et al.*, 2014). Two of the most promising types of BGCs are non-ribosomal peptide synthetases (NRPSs) and polyketide synthases (PKSs), which produce non-ribosomal peptides (NRPs) and polyketides (PKs), respectively.

1.3.1.1. Non-Ribosomal Peptide (NRPs)

Non-ribosomal peptides – synthesized by NRPSs – are usually 2 – 48 amino acids in length

and derive their diversity from the incorporation of more than 300 different amino-acids including unusual ones such as D-Amino acids and N-methylated amino acids (Du & Lou, 2010). They work as mega-synthases, composed of at least 3 modules: 1. Adenylation domain: recognizes and activates the entering amino-acid, 2. Peptidyl Carrier Protein (PCP) domain: covalently bonds to the amino acid and 3. Condensation (C) domain: catalyzes the elongation of the peptide. In addition, different domains may be present at the modules acting as tailoring enzymes (Rausch *et al.*, 2007). The termination is performed by a Thioesterase (TE) domain usually located at the end of the modular “assembly line”, however complete NRP synthesis is not always linear and may involve intra- and intermolecular domain interactions (Sundlov *et al.*, 2012).

1.3.1.2. Polyketides

Many secondary metabolites produced by microorganisms are classified as polyketides – synthesized by PKSs – and show a remarkable diversity in structure and bioactivities. Currently polyketide compounds have been developed for multiple applications such as chemotherapy (e.g., Rapamycin), anti-cholesterol (e.g., Lovastatin) and a number of antibiotics including Erythromycin, Rifamycin and Amphotericin B (Staunton & Weissman, 2001). Recently, metagenomics studies have led to the characterization of polyketides with antibiotic activity against Methicillin-resistant *Staphylococcus aureus* (MRSA) and Vancomycin-resistant *Enterococcus faecalis* (Feng *et al.*, 2011).

Polyketides are produced via successive decarboxylative Claisen thioester condensations of an activated acyl starter unit and extension by the addition of malonyl-CoA units or homologs (Hertweck, 2009; Khosla *et al.*, 2014). The elongation is performed by modules that are commonly composed of: an acyl-transferase (AT) domain that activates the acyl starter and loads it on the Acyl Carrier Protein (ACP) domain, and a β -ketoacylsynthase (KS) domain that catalyzes the

carbon-carbon condensation between molecules. Modules can contain other optional sites like ketoreductases and/or dehydrases, while the newly synthesized compounds can be later processed by methyltransferases/ estereomerases for example, to further contribute to the diversity of the compounds produced (Hertweck, 2009). PKSs can be divided into 3 main categories that differ significantly in the arrangement of synthetic modules. Type I PKSs use numerous modules in an assembly-line fashion to synthesize the final compound and each module performs only one elongation step. Type II PKSs use the same ACP domain throughout the entire elongation process and the modules function in an iterative way to produce highly oxidized aromatic compounds (Hertweck, 2009). Type III PKSs, are homodimeric enzymes, that acts directly on the Acyl-CoA substrates in an iterative manner to catalyze the condensations without the need of an ACP domain (Shen, 2003).

PKS pathways can also be found associated with other biosynthetic pathways such as NRPS and Fatty-Acid Synthases (Ansari *et al.*, 2004; Masschelein *et al.*, 2013; Milano *et al.*, 2013). Interestingly, all of these biosynthetic pathways share striking homology in their basic chemistry allowing their modules to interact to produce hybrid BGCs (Du *et al.*, 2001).

Given the organization and nature of PKS and NRPS pathways, they can be genetically engineered in the laboratory and serve as a tool for drug production and novel bioactive compound development (Siezen & Khayatt, 2008). Nevertheless, the lack of knowledge of the extant diversity of PKS pathways in environmental metagenomes has limited the polyketide diversity that can be obtained from engineering approaches. There are many unknowns even for the existing PKSs, such as lack of understanding of spatial interactions between enzymes, stereochemistry control or the correct flux of reagents and products between the active sites (Khosla *et al.*, 2014). Given these current constraints in our knowledge, the cloning and expression of environmental PKS and NRPS

pathways could provide insights into PKS/NRPS biochemistry and genetics, as well as provide new bioactive compounds of medicinal value.

1.3.2. Assessing environmental BGCs through culture independent methods

The culture independent methods used to study microbiomes can be readily borrowed for studying environmental BGCs. Akin to 16S rRNA surveys, microbiomes can be surveyed for the presence of genes related to specific BGCs by targeted amplification of conserved domains. This strategy can be applied well to the detection of PKSs and NRPSs by targeting the conserved β -ketosynthase (KS) domains and adenylation (A) domains, respectively, and using the amplicons to draw BGC relatedness and distribution in an environmental sample (Cacho *et al.*, 2014; Charlop-Powers *et al.*, 2016; Foerstner *et al.*, 2008; Wawrik *et al.*, 2005). However, the same biases and limitations from 16S rRNA surveys apply to this approach, including the overlook of divergent sequences and lack of genomic context.

Shotgun metagenomics overcome these constraints by direct sequencing of environmental DNA, however its application has been limited due to the difficulties in assembling large, contiguous genomic regions (Howe *et al.*, 2014), especially for low relative abundance taxa in complex microbiomes (Ladoukakis *et al.*, 2014). Advances in NGS technology and bioinformatics have greatly improved the recovery of large BGCs, such as NRPS and PKS from metagenomes, with the availability of powerful tools for the reconstruction (Meleshko *et al.*, 2019) and bioinformatic prediction of BGCs, such as AntiSMASH (Blin *et al.*, 2019). Similarly, genome-resolved metagenomics brings to the fore the ability to reconstruct environmental BGCs in their genomic context and has provided invaluable insights into BGC taxonomical origin, diversity and distribution across soils (Sharrar, Allison M *et al.*, 2020) including rare, understudied and uncultured representatives (Crits-Christoph *et al.*, 2018). Thus, direct sequencing and genome-

resolved metagenomics are important tools to assess the reservoir of BGCs from an environment, however these approaches do not integrate well with heterologous expression. For this reason, functional metagenomics, with the construction and screening of metagenomic libraries, is still the most common culture-independent strategy to prospect natural products.

1.3.2.1. Building metagenomic Libraries

Prior to constructing a metagenomic library it is important to first select the source DNA, isolation method, vector, insert size, screening method, and expression host so that the output of the screening match the desired project goals (Kakirde *et al.*, 2010). Since hosts differ in the relative degree of transformability and metabolic capacity, the development of broad host-range shuttle vectors is desirable since it allows libraries to be conjugally transferred into alternative hosts as plasmids (Kakirde *et al.*, 2010) or by integration into the recipient host chromosome (Aakvik *et al.*, 2009; Courtois *et al.*, 2003; Craig *et al.*, 2010; Heil *et al.*, 2012; Kakirde *et al.*, 2010)

Generally, the larger the insert a vector can accommodate, the lower its transformation efficiency, complicating library construction. Nevertheless, when capturing PKS and NRPS pathways from the environment, insert size is a key variable (Kakirde *et al.*, 2010), since these pathways can be quite large due to their modular architecture, exceeding 100Kb (Fu *et al.*, 2012). Common cloning strategies for the construction of metagenomic libraries include the use of cosmid, fosmid or bacterial artificial chromosome (BAC) vectors.

Cosmid vectors are plasmids containing the *cos* site of λ phage and an origin of replication allowing ligated sequences to be packed in λ phages and transduced into the host where the insert is maintained as a plasmid (Haley, 1988). Because of the small size of the plasmid and *in vitro* packaging restraints this technique can only successfully transform the host with inserts of usually

40kb. The lack of controlling mechanisms on plasmid copy number and recombination between plasmids often decrease the stability of cosmid vectors.

Furthermore, they perform well in a narrow-range of hosts, but its range can be widened via genetic manipulation. Broad-host-range cosmid vectors pJC8 and pJC24 were designed by cleverly adding the Gateway® homology site *attL1* and *attL2* to the cosmid along with the RK-2 *oriT* site (Cheng *et al.*, 2014). These cosmid vectors were able to transfer the insert to a Gateway® plasmid via recombination, and the RK-2 *oriT* site allows the transfer of the plasmid via conjugation to a number of other hosts, including gram-positive bacteria and even yeasts like *Saccharomyces cerevisiae*, enhancing the capabilities and usefulness of cosmid vectors in heterologous expression screening efforts.

Fosmid vectors are based on the same principles of cosmid vectors, but use a F-plasmid origin of replication. F-plasmids have advantages such as existing as single copy plasmids and having a tight fertility control which decreases host intolerance to repetitive sequences, while also limiting homologous recombination between copies. Therefore they are more stable and desirable when constructing metagenomic libraries (Quail *et al.*, 2011) and have been used for many studies of microbial ecology and natural product discovery (Felczykowska *et al.*, 2014; Parsley *et al.*, 2011). Nevertheless, fosmid vectors have the same restrictions on host-range faced by cosmid vectors, but similar approaches can be used to mitigate this issue, such as the introduction of RK-2 *oriV* and an *oriT* to the commercial fosmid vector pCC1FOS (Aakvik *et al.*, 2009). By doing so, the fosmid vector can be transferred to other hosts from *E. coli* via conjugation and the additional *oriV* allows for plasmid copy number regulation in hosts other than *E. coli* that does not recognize the original *ori2* site.

Even though cosmid or fosmid libraries provide a more efficient method to generate

metagenomic libraries, because of the necessity to use phage packaging, these vector systems are inherently limited in size and are ineffective at recovering larger BGCs, such as Type I PKS and NRPS pathways. In contrast, BAC vectors can be used to stably clone insert DNA fragments as large as 300kbp (Shizuya *et al.*, 1992), which is sufficient to contain the largest of known BGCs. Recently, metagenomic libraries containing inserts exceeding 100kbp on average have been constructed (Nasrin *et al.*, 2018). Although BAC vectors can accommodate complete BGCs, the transformation efficiency is significantly lower, making library construction more labor intensive and therefore are not a suitable and/or preferred vector for all applications (Quail *et al.*, 2011).

The first BAC vectors, such as pBELOBAC11 were kept as single copy plasmids in the host cell, granting high stability but decreasing the amount of product obtained by heterologous expression, requiring larger scale *E. coli* cultures for product extraction (Wild *et al.*, 2002). Inducible-copy systems were introduced into the vector, so that it carries an additional origin of replication *oriV* from RK2 that is dependent upon the expression of the TrfA protein. By expressing the *trfA* gene under the control of an arabinose-dependent promoter, the BAC vector is then conditionally copy-induced only in the presence of arabinose. This strategy allows the maintenance of vectors in single copy with triggered increase in copy-number for heterologous expression or DNA isolation purposes (Kakirde *et al.*, 2010; Wild *et al.*, 2002). The broad host-range vector pGNS-BAC has been shown to replicate in diverse Gram-negative bacteria and can be conjugally transferred from *E. coli* (Kakirde *et al.*, 2010). Therefore this vector is well suited for metagenomic library construction and screening, especially if the DNA isolated from an environmental sample is biased towards gram-negative bacteria (Liles *et al.*, 2003). Other derivatives of this vector have been generated that have the capacity to be introduced and stably integrated into Gram-positive hosts as well (Nasrin *et al.*, 2018).

1.3.2.2. Functional Screening of Metagenomic Libraries

The seminal work on soil metagenomics demonstrated that clone libraries could be screened for desired bioactivities by expressing metagenomic inserts within an *E. coli* host and resulted in discovery of various enzymatic activities (Rondon *et al.*, 2000). Since this method had been proposed, different groups have turned to functional screening to identify clones exhibiting a desired activity. High throughput function-based screenings often rely on indicators to rapidly identify the activity of interest, such as inhibition halos, color change in the media or colonies, degradation of media for example (Coughlan *et al.*, 2015).

A high throughput screening method for antibiotic activity was developed based on a sequential two antibiotic selection scheme (Brady & Clardy, 2000). First clones are selected for kanamycin resistance in media containing the antibiotic. Then antibiotic activity is verified by an overlay of kanamycin-resistant *Bacillus subtilis*. Positive hits are recovered directly from the assay plate and then streaked in media containing ampicillin killing *B. subtilis*. This screening method have led to the characterization of a number of compounds including several Long-chain N-acyl amino acid antibiotics (Brady *et al.*, 2004; Brady & Clardy, 2000), Isocyanide antibiotic (Brady & Clardy, 2005), antibiotic polyketides (Craig *et al.*, 2009), novel metalloproteases, serine proteases and lipolytic compounds (Iqbal *et al.*, 2014).

An interesting way to enrich for clones containing PKS and/or NRPS pathways, is to explore their inherent biochemical characteristics (Charlop-Powers *et al.*, 2013). PPTases, are a constitutive part of NRPS and PKS pathways and are also required for iron uptake in *E. coli*, namely by the PPTase encoded by *entD*. Therefore, metagenomic libraries can be enriched for PKS and NRPS pathways by the use of *entD*- mutants, since only clones expressing environmental PPTases are viable in iron limiting conditions (Charlop-Powers *et al.*, 2013). This approach

provides a fast way to select promising clones within a metagenomic library in an easy, fast and low-cost manner. However, there are still limitations intrinsic to heterologous expression, especially when using *E. coli* as a host.

The intricacies of transcription and translation of metagenome-derived genes often precludes heterologous expression in a particular host. For example, functional screening of a cosmid metagenomic library using 6 different *proteobacteria* hosts, showed that phenotypes were rarely observed twice in different host (Craig *et al.*, 2010), prompting the efforts to expand the range of bacteria for metagenomics library expression (Courtois *et al.*, 2003; Craig *et al.*, 2010). Additionally, incorporation of heterologous sigma factors, for example from *Acidobacteria* or *Lactobacillus* taxa, can enhance transcription from metagenomic clones (Gaida *et al.*, 2015; Sabree *et al.*, 2006).

With specific regard to secondary metabolite producers, hosts such as *Streptomyces* spp. have been of particular interest, due to their prolific production of secondary metabolites. Several genetic engineering strategies have been developed to improve heterologous expression of library clones. Deletions of endogenous antibiotic gene clusters attenuate competing use of carbon, nitrogen and precursors, while facilitating the identification of compounds encoded by the eDNA by simplifying the host metabolic profile (Gomez-Escribano *et al.*, 2012). Introduction of point mutations in *rpoB* and *rpsL* genes in *S. coelicolor* resulted in increased heterologous antibiotic production (Gomez-Escribano & Bibb, 2014; Hu *et al.*, 2002).

1.3.2.3. Sequence-based screening of metagenomics libraries

Function-based screening has been an important way to mine metagenomic libraries for molecules and/or processes of interest, however one of its main limitations is redundant isolation (Tulp & Bohlin, 2005), which can be overcome by prior bioinformatical analyses of inserts

selected by sequence-based screening, making the process more efficient (Chang & Brady, 2013; Culligan *et al.*, 2014). BGCs can be targeted by PCR or hybridization-based methods in order to identify pathway-containing clones, utilizing conserved domains and/or regions.

Clones containing PKS and/or NRPS pathways can be identified by homology-based approaches targeting KS and A domains (Banik & Brady, 2010; Chang & Brady, 2013). Nevertheless, there is great sequence diversity among KS and A domain sequences, leading this approach to echo the bias associated with amplification surveys: biases towards sequences most similar to those from previously characterized pathways.

A powerful approach on the horizon is the utilization of NGS and bioinformatics approaches to screen for pathways of interest. In this approach, all of the clones in the library are sequenced and bioinformatic tools are used to identify and fully annotate clones containing the target pathways, which can be rationally selected for heterologous expression. This strategy has the potential to circumvent the biases associated with amplification-based screening, as well as the constraints associated with the lack of pathway information prior to expression, leading to the identification of more promising pathways within metagenomic libraries.

1.4. Single walled carbon nanotubes as antimicrobial agents

The increase of resistant pathogens has become a global health issue, and therefore, the discovery of new antibiotic agents has been deemed one of the milestones for human health. In conjunction with exploring the vast metabolic potential of microorganisms, other venues have also been pursued to improve our antibiotic arsenal.

Carbon nanotubes have gained lots of attention since its development due to its unique physicochemical, mechanical and electrical properties (Iijima, 1991; Jorio *et al.*, 2007). They have been widely explored as novel nanomaterials with diverse bio-applications such as biosensors

(Chen, R. J. *et al.*, 2003), cell delivery systems (Cai *et al.*, 2005b) and tissue engineering (Kam & Dai, 2005b). Another way of using carbon nanotubes in the biomedical field is exploiting their cytotoxicity. However, carbon-nanotube cytotoxicity studies have focused on mammalian cells with disputed conclusions, since nanotubes exhibit toxicity only at certain conditions (Kang *et al.*, 2008).

Nanotube toxicity and the need for novel antimicrobial treatments, led to the first report of antimicrobial activity from single walled carbon nanotubes (SWNTs) against *E. coli* K12 washed from media and challenged with pristine SWNTs in solution (5 µg/mL) and as a coated filter. Fluorescence microscopy showed that most of the bacterial cells that were adhered to SWNT aggregates in solution were dead and had damaged cell walls, however, solubilized nanotubes had no effect on the free-swimming bacteria (Kang *et al.*, 2007).

Investigation of differences between Single and Multi-walled carbon nanotubes (SWNTs and MWNTs) antimicrobial activity, suggested that aggregates are significantly more efficient at disrupting cells than MWNTs, except for free-swimming bacteria in the saline solution, which were not affected by treatment (Kang *et al.*, 2008).

Other factors also influence SWNT antimicrobial activity. Size was identified as an important factor as smaller nanotubes, with smaller diameters were more efficient at killing bacteria (Yang *et al.*, 2010), with length proving to be decisive in the production of antimicrobial biomaterials (Aslan *et al.*, 2010). Functionalization and contamination can also affect the observed activity. Raw Cu-SWNT show more activity than pristine nanotubes (Li *et al.*, 2014) without measurement of metal toxicity itself, functionalization with -OH and -COOH also seem to boost antimicrobial activity, while their results on the effects of NH₂ are conflicting (Arias & Yang, 2009; Li *et al.*, 2014).

Other study reported that individually dispersed nanotubes have higher toxicity than aggregates and probably affect the bacterial cells by puncturing them. Activity was increased in gram positive bacteria and was also positively related to concentration and shaking speed during incubation, also affecting free-swimming cells (Liu *et al.*, 2009).

The dispersing agents in which SWNTs are solubilized and antimicrobial assays medium also have a fundamental effect on observed activity, with no activity detected for functionalized nanotubes when assays were conducted in Phosphate buffer saline (PBS) and in Brain-heart Infusion (BHI) broth on either gram-positive nor gram-negative bacteria (Arias & Yang, 2009). Additionally, surfactants are routinely used as an assay medium either to increase SWNT activity (Arias & Yang, 2009; Bai *et al.*, 2011; Liu *et al.*, 2009; Yang *et al.*, 2010) or simply due to their efficiency at dispersing (Dong *et al.*, 2012a; Horn *et al.*, 2013; Martinez *et al.*, 2014). Indeed, Atomic force microscopy revealed that “softer” cells are more vulnerable to SWNT antimicrobial activity (Liu *et al.*, 2009).

An important step in understanding the interaction between bacterial cells and carbon nanotubes was the evaluation of gene expression under SWNT induced stress, showing that 91% of the upregulated genes after exposure were also upregulated in studies involving high-pressure stress (Kang *et al.*, 2008). There was an upregulation of general stress-response mechanisms, oxidative-stress mechanisms, fatty acid metabolism, glycolysis, and most importantly upregulation in genes related to cell-wall integrity; showing strong evidence that carbon nanotubes disrupt the cell-wall functionality and that cells actively resist the stress.

Rather counter intuitively, cell viability gene markers *dnaK*, *groEL*, *groES*, *grpE* (Kort *et al.*, 2008) were upregulated even with reported inhibition rates of 80%. Although there is a possibility that mRNA originated from dead cells (Trevors, 2012), the long incubation times make

it unlikely, suggesting that cells were still viable during treatment with carbon nanotubes. Interestingly, the assays for RNA expression were conducted in diluted LB medium as opposed to Live/dead assays where activity was observed, which were conducted in distilled water.

Nutrient availability is pivotal in the ability of bacterial cells to resist cell-wall stress. Surfactant resistance is intimately related to carbon source availability and ability to undergo anaerobic respiration since it is not dependent on cell wall integrity (Ishikawa, S. *et al.*, 2002). In addition, surfactant-resistant *E. coli* can grow in high concentrations of SDS but require more energy sources because of the constant cell-wall and membrane repair (Nakata *et al.*, 2010). Therefore, nutrient availability can be an important factor in bacterial ability to withstand nanotube stress. The development and use of biologically compatible dispersing agents such as DNA (Ao *et al.*, 2011) and more preferably TSB (Sloan *et al.*, 2017) offer a platform to assess SWNT toxicity without confounding variables such as inadequate osmotic pressure, nutrient starvation or sensitizing agents such as surfactants. Conciliating and robust information of SWNT activity is fundamental to advance the effective application of SWNTs either as dispersions or materials as antimicrobial agents.

Chapter 2: Discovery of novel biosynthetic gene cluster diversity from a soil metagenomic library

This work was published in the journal *Frontiers of Microbiology* (2020) under the title:

Discovery of novel biosynthetic gene cluster diversity from a soil metagenomic library

Alinne Santana-Pereira¹, Megan Sandoval-Powers¹, Scott Monsma², Jinglie Zhou¹, Scott R. Santos¹, David Mead^{3*} and Mark R. Liles^{1,3*}

¹ Department of Biological Sciences, Auburn University – Auburn AL, USA

² Lucigen Corporation – Middleton, WI, USA

³ Varigen Biosciences Corporation – Madison, WI, USA

2.1. Abstract

Soil microorganisms historically have been a rich resource for natural product discovery, yet the majority of these microbes remain uncultivated and their biosynthetic capacity is left underexplored. To identify the biosynthetic potential of soil microorganisms using a culture-independent approach, we constructed a large-insert metagenomic library in *Escherichia coli* from a topsoil sampled from the Cullars Rotation (Auburn, AL, United States), a long-term crop rotation experiment. Library clones were screened for biosynthetic gene clusters (BGCs) using either PCR or an NGS (next generation sequencing) multiplexed pooling strategy, coupled with bioinformatic analysis to identify contigs associated with each metagenomic clone. A total of 1,015 BGCs were detected from 19,200 clones, identifying 223 clones (1.2%) that carry a polyketide synthase (PKS) and/or a non-ribosomal peptide synthetase (NRPS) cluster, a dramatically improved hit rate compared to PCR screening targeting type I polyketide ketosynthase (KS) domains. The NRPS and PKS clusters identified by NGS were distinct from known BGCs in the MIBiG databases or those PKS clusters identified by PCR. Likewise, 16S rRNA gene sequences obtained by NGS of the library included many representatives that were not recovered by PCR, in concordance with

the same bias observed in KS amplicon screening. This study provides novel resources for natural product discovery and circumvents amplification bias to allow annotation of a soil metagenomic library for a more complete picture of its functional and phylogenetic diversity.

2.2. Introduction

There is a tremendous degree of microbial diversity in soils (Torsvik & Ovreas, 2002), and soilborne microorganisms are key to biogeochemical and ecological processes. Interactions between microorganisms are, to a large extent, mediated by bioactive natural products that are encoded by biosynthetic gene clusters (BGCs) (Hill *et al.*, 2014; Reddy *et al.*, 2012; Wang *et al.*, 2014). Natural products from soil microorganisms have historically been a rich resource for antibiotics (Hover *et al.*, 2018; Ling *et al.*, 2015; Schatz *et al.*, 1944); however, the vast majority of encoded BGCs have yet to be characterized since most soil microbes are recalcitrant to cultivation (Howe *et al.*, 2014; Janssen, 2006; Orellana *et al.*, 2018; Soliman *et al.*, 2017). Culture-independent, amplification-based surveys have resulted in significant insights into microbial ecology and population dynamics (Schöler *et al.*, 2017), while having their own inherent biases that obscure and distort our view of microbial diversity due to PCR amplification bias (Eloe-Fadrosh *et al.*, 2016). Therefore, adoption of metagenomic approaches that avoid amplification biases, such as cloning of nucleic acids from natural environments (Rondon *et al.*, 2000) or shotgun sequencing (Howe *et al.*, 2014), have provided access to hitherto underexploited reservoirs of microbial diversity and their encoded natural products (Daniel, 2004; Owen *et al.*, 2013; Sharrar *et al.*, 2020).

Shotgun metagenomic sequencing has allowed new insights into the functional and phylogenetic composition of complex microbial assemblages. However, assembly of short-read metagenomic sequences is especially challenging from phylogenetically diverse microbial

assemblages such as soils, and this had precluded the assembly of complete genomes and their encoded BGCs (Howe *et al.*, 2014; Kang & Brady, 2014b). The advent of genome-resolved metagenomics through binning approaches has now enabled recovery of intact BGCs from shotgun sequenced metagenomic DNA (Crits-Christoph *et al.*, 2018; Sharrar *et al.*, 2020). Despite these developments in uncovering BGC diversity, direct sequencing does not lend itself to heterologous expression of BGCs in order to identify the encoded secondary metabolites. In order to access and express complete BGCs, bacterial artificial chromosome (BAC) vectors can be used to clone contiguous genomic fragments that can exceed 300 Kb, thereby increasing the probability of obtaining intact BGCs from metagenomic sourced DNA, while also enabling inducible copy number in *E. coli* and conjugal transfer to heterologous expression hosts (Aakvik *et al.*, 2009; Kakirde *et al.*, 2011; Liu *et al.*, 2010; Liu *et al.*, 2016; Nasrin *et al.*, 2018; Wild *et al.*, 2002). Functional screening of metagenomic clones typically results in very low hit rates for enzymatic activities (Healy *et al.*, 1995; Lewin *et al.*, 2017; Thies *et al.*, 2016) and even lower rates for discovering bioactive metabolites (Tulp & Bohlin, 2005); thus, it is desirable to first identify clones containing the target gene or BGC of interest. Screening metagenomic libraries using degenerate oligonucleotides targeting conserved BGC domains by PCR or other molecular methods can be an effective strategy to identify BGCs of interest (Charlop-Powers *et al.*, 2016; Owen *et al.*, 2015; Owen *et al.*, 2013); however, because these methods rely on previously described BGCs for primer design, we hypothesized that probe or PCR-based approaches for BGC discovery would be biased and identify only a subset of the BGCs present within a metagenomic library. This hypothesized bias in BGC discovery using a PCR-based approach is analogous to the incomplete phylogenetic diversity observed using “universal” 16S rRNA gene-specific primers as compared to the vast phylogenetic diversity observed for the candidate phyla radiation (CPR) that has been recovered

only through non-amplification strategies (Hong *et al.*, 2009).

Among the classes of BGCs of greatest interest for producing diverse secondary metabolites are the PKS and NRPS clusters. These BGCs encode modular enzyme complexes that produce molecules of incredible chemical diversity that include metabolites with antibacterial, antifungal, anti-cancer, cholesterol-lowering and insecticidal bioactivities (Cragg & Newman, 2013). The modular arrangement of these mega-synthases and the large number of domains can result in long, contiguous clusters that can be challenging to reconstruct by direct sequencing or are too long to be captured intact by fosmid or cosmid metagenomic libraries (Kallifidas & Brady, 2012).

In this study we screened a large-insert soil metagenomic BAC library hosted in *E. coli* (Nasrin *et al.*, 2018) using multiple sequence-based approaches, including PCR amplification of 16S rRNA genes and KS domains associated with PKS clusters, as well as using a novel next-generation sequencing (NGS) strategy to mine the entire library for BGCs. A multiplexed pooling strategy was used to allow bioinformatic identification of sequences associated with each of the 19,200 clones formatted in 384-well plates. The collection of metagenomic clones of interest, identified by PCR or by NGS, were then compared both from phylogenetic and biosynthetic perspectives, demonstrating the power of the NGS approach in providing a source of previously undescribed BGCs that may be heterologously expressed to produce secondary metabolites with potential therapeutic applications.

2.3. Materials and Methods

2.3.1. Bacterial strains

The metagenomic library was cloned using the *E. coli* DH10B BAC-Optimized Cells (Lucigen Corporation, Middleton, WI), which are competent cells developed specifically for BAC

library construction. These cells contain an arabinose-inducible *trfA* gene for BAC vector copy-induction (Wild & Szybalski, 2004).

2.3.2. Soil DNA isolation and metagenomic library construction

Bulk soils were sampled from 10-30 cm below the soil surface, in a plot that had not received fertilizer for at least 100 years in the Cullars Rotation (Auburn, AL) and was at the time planted with a soybean crop. The soil samples were transported to the laboratory where subsamples were immediately frozen at -80°C, and the majority of the soil samples were maintained at 4°C until processed within one week for DNA isolation. A 10 g soil sample was processed for high molecular weight (HMW) DNA isolation and purification as previously described (Liles *et al.*, 2008), and the metagenomic DNA was randomly sheared. The fragmented DNA was end-repaired, and BstXI adaptors ligated on to the DNA, followed by gel-fractionation and ligation into the pSmartBAC-S vector as previously described (Nasrin *et al.*, 2018). The vector:insert ligation was transformed by electroporation into BAC-optimized *E. coli* DH10B (Lucigen Corp, Middleton, WI), and transformants were selected on LB agar (10 mg/mL of Tryptone, 10 mg/mL of NaCl, 5 mg/mL of Yeast Extract, 15 mg/mL of Agar in 1L of water) containing 12.5 µg/ml chloramphenicol. The soil metagenomic library consisted of 19,200 clones arrayed into 384 format (50 plates) with an average size of 113 Kb (Nasrin *et al.*, 2018).

2.3.3. Metagenomic library pooling and sequencing

The Library was divided into 5 sets of 10 plates: Set 1 (Plates 1-10); Set 2 (Plates 11-20); Set 3 (Plates 21-30); Set 4 (Plates 31-40) and Set 5 (41-50). For Set 5, the initial pooling strategy merged all 384 clones from each original library plate into a single plate pool (10 plate pools); row clones from the 10 Set 5 library plates into single row pools (16 row pools A-P); and column clones from the 10 Set 5 library plates into single column pools (24 column pools) (**Figure 1**). For the

remainder of the library (Sets 1 through 4), the 384-well plates from each Set were divided into 4 quadrants and replicated into 96-well plates, for a total of 40 x 96-well plates from each original Set. The resulting 96-well plates were pooled according to their Set of origin, resulting in 40 plate pools; 8 row pools A-H and 12 column pools (**Figure 1**).

For all pooling approaches, individual clones were grown in triplicate in 96-well plates using 1 ml LB containing 0.01% arabinose to amplify BAC copy number²¹. Pools were made by combining the liquid cultures as appropriate, pelleting the cells, and purifying BAC DNA, as previously described (Wu *et al.*, 1999). Fragment libraries for sequencing on Illumina were constructed with 100 ng purified BAC DNA from each pool using the multichannel protocol of the NxSeq® UltraLow DNA Library Preparation Kit (Lucigen, Middleton WI). Unique indexes were used for each library pool within each batch of 10 library plates (Sets). Libraries were multiplexed and sequenced on Illumina HiSeq 2500 with v3 chemistry at 2x 150bp.

2.3.4. Assembly *de novo* of metagenomic contigs

The raw HiSeq reads per each column, plate, or row pool were transferred to the Alabama Supercomputer (ASC) for processing. Reads were filtered for high quality reads (Q score > 30), trimmed, and clipped, and reads smaller than 30 bp discarded using the software Trimmomatic (Bolger *et al.*, 2014). To remove host and vector DNA sequences, all processed reads were mapped against *E. coli* DH10B and the vector pSmartBAC-S sequences using BWA (Li & Durbin, 2009), and those un-mapped to the reference were then assembled using metaSPAdes implementation of SPAdes 3.9.0 software (Nurk *et al.*, 2017). Reads corresponding to each respective sequencing pool were assembled individually, resulting in a set of contigs from each one of the 210 metagenomic library pools.

2.3.5. Library deconvolution from contig to clone

All contigs generated from the metaSPAdes assembly were tentatively deconvoluted to a clone location using a custom bash script. Briefly, the deconvolution process consisted of renaming each individual contig to include their pool of origin and a unique number identifier. Contigs from the plate pools were compared to those in the column or row pools via BLASTn with 95% identity and a 10^{-6} e-value cut-off. The BLAST hits were extracted and annotated into 3 categories: 1) completely deconvoluted – plate contigs with hits in both column and row pools; 2) partially deconvoluted – plate contigs with hits in only one other dimension; or 3) singletons – contigs with no significant hits. After each contig was annotated, the location information in the contig name was used to generate coordinates (plate, column, and row) for the respective clone of origin.

2.3.6. Statistical analysis of library sequencing data

A table containing contig name, length, coverage, pool dimension, deconvolution status, and set number for all contigs from the library were analysed in R studio version 1.2.1335-1. Normality checks were performed using Q-Q plots while correlation between contig length and contig coverage was evaluated using Spearman's test. Differences in contig length and coverage across sets and deconvolution status were evaluated using Kruskal-Wallis and Dunn tests. The analysis was conducted with and without outliers removed using the non-parametric Tukey's method with a factor of 1.5xIQR, in order to determine the impact of outliers on some correlations.

2.3.7. AntiSMASH prediction of BGCs from library contigs

A local version of antiSMASH 4.0 (Blin *et al.*, 2017) with prodigal (meta) for gene prediction was used to predict BGCs from plate pools, which had the greatest coverage per pool. The program was run in the Bioconda environment on the ASC to afford high-throughput detection. Annotations for the library contigs containing PKS and/or NRPS clusters were

conducted by importing the BiosynML antiSMASH 4.0 output of assembled library contigs and individually re-sequenced inserts into the Geneious software suite.

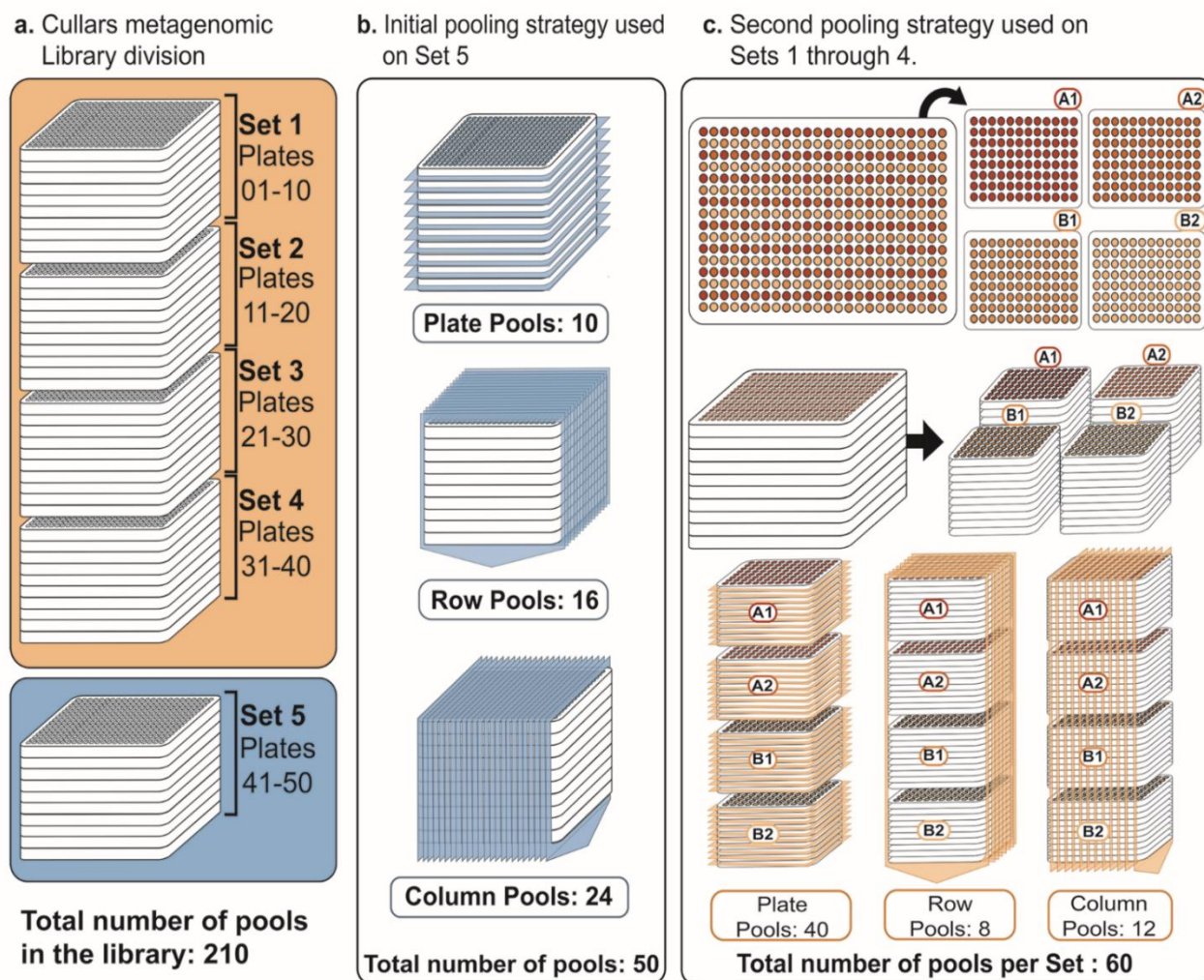


Figure 1. Pooling strategies used in sequencing Cullars metagenomic library. a. The Cullars metagenomic library composed of 50 384-well plates got divided into 5 sets of 10 plates that are pooled for Plate x Row x Column individually, for a total of 210 pools. b. Initial pooling strategy used in Set 5. All clones in each plate were combined to generate a plate pool, all clones from each row across the 10 plate set were combined to generate a row pool, and finally all clones from each column across the 10 plate set were combined to generate a column pool. Set 5 generated 50 pools, that were individually barcoded and sequenced. c. Second Pooling strategy applied to Set 1, 2, 3 and 4 individually. For each Set, each 384-well plate was arrayed into quadrants composed of 96 clones that got replicated into 96-well plates. Thus the initial 10 384-well plates got divided into 4 subsets of 10 96-well plates. All subsets are combined to generate the pools. All clones in each 96-well plate were combined to generate a plate pool, all clones from each row across the 96-well plate subsets were combined to generate a row pool, and finally all clones from each column across the 96-well plate subsets were combined to generate a column pool. Each Set from 1 through 4 generated 60 pools that were individually barcoded and sequenced.

2.3.8. Comparison of conserved domains from PKS and NRPS clusters with domains from databases

KS domains and A domains from predicted PKS and NRPS clusters, respectively, were extracted and translated using the bacterial translation table. For comparison to known NRPS or PKS clusters, the MIBiG database (Epstein *et al.*, 2018) (accessed 06/14/2019, version 1.4) was used to obtain a well-curated database of KS and A-domains. Prior to comparison, the KS and A domains database was clustered using mmseqs2 (Steinegger & Soding, 2017) to clusters of 75% amino acid similarity to decrease computational requirements for multiple alignment. Library sequences and MIBiG database sequences were aligned using MUSCLE (Edgar, 2004) and trimmed using TrimAl (Capella-Gutierrez *et al.*, 2009) automated algorithm to optimize signal x noise ratio. Following automated trimming additional sequences were dropped to ensure all alignments had the same set of library sequences. A maximum likelihood tree with 1000 Bootstraps was generated using RAxML 8.0 (Stamatakis, 2014). The annotations generated by antiSMASH 4.0 were combined with MIBiG GenBank files and analyzed using BiG-SCAPE (Navarro-Munoz *et al.*, 2020) with clustering set to 0.3. The obtained networks were imported into Cytoscape (Smoot *et al.*, 2011) for network visualization.

2.3.9. Taxonomic origin of PKS/NRPS containing library contigs

For taxonomic origin inference, k-mer profiles and taxonomic binning of all contigs containing BGCs and all the validated insert sequences was performed using Kraken (Wood & Salzberg, 2014). The predicted taxonomic origin of validated insert sequences was further assessed by matching gene annotation information and k-mer profiles to achieve a consensus taxonomic classification.

2.3.10. Re-sequencing identified clones, annotation, and taxonomic origin

Selected clones identified as containing PKS and/or NRPS clusters were individually grown from the *E. coli* cryostock and the presence of the targeted BGC confirmed by insert-specific PCR (data not shown). The isolated BAC DNA was re-sequenced by standard single-end fragment sequencing using a MiSeq sequencer (Illumina, San Diego, CA). Originally, trimming and assembly was conducted with CLC Genomics Workbench 8.5 followed by manual inspection and reassembly, and antiSMASH 4.0 was used for prediction of BGCs from fully-assembled clone insert sequences. Inserts with antiSMASH annotation matching that of their associated contig were considered validated.

Additionally, processed reads were reassembled using biosynthetic-SPAdes available in the package SPAdes 3.14.1 and BGCs were predicted by antiSMASH 5.0 (Blin *et al.*, 2019), in order to compare BGC assembly and detection from resequenced inserts with recently updated analysis tools. Annotations for the resequenced inserts were extracted from antiSMASH 5.0 output using the package Palantir (Meunier *et al.*, 2020) available from metaCPAN. Specifically, PKS and NRPS domains were extracted from the clusters to facilitate BGC annotations. The use of biosyntheticSPAdes for assembly performed better for BGC annotations and was therefore used for downstream annotations. For in-depth annotation of three BGCs, ORFs and domains encoded within each ORF as predicted by antiSMASH 5.0 were visualized using the R package gggenes v0.4.0 (Wilkins, 2019).

2.3.11. Screening libraries for PKS clusters via PCR

All 19,200 BAC clones were screened for the presence of KS domains using degenerate PCR primers 5LL and 4UU (Parsley *et al.*, 2011) (**Table 1**) Each 25- μ l PCR reaction contained 10 pmol of 5LL and 4UU, 12.5 μ l CloneID 1X colony PCR master mix with Taq DNA polymerase

(Lucigen Corp.), and 1 μ l overnight growth of supernatant containing *E. coli* BAC DNA. Amplification was performed by conducting an initial denaturation at 94°C for 1 min, followed by 30 rounds of thermal cycling at denaturation at 94°C for 30 s, annealing temperature of 60°C for 30 s and extension at 72°C for 1 min, followed by 5 min of extension at 72°C. Reactions were considered positive if a ~750 bp amplicon was visible upon agarose gel electrophoresis.

2.3.12. Identification of 16S ribotypes

16S rRNA gene candidates were mined from the Cullars metagenomic library by BLASTn search against the SILVA SSU nr database (Quast *et al.*, 2012) and top hits with E-value lower than 1×10^{-5} and minimal initial alignment of 45bp were computed. All hits smaller than 300bp were discarded, and the remaining were annotated to the phyla level according to SILVA taxonomy at 75% identity.

Table 1. List of oligonucleotide sequences used in this study.

Oligonucleotide name	Oligonucleotide sequence
PKSc1OligoA	CGAGCATGCCGTGTCGATCGCCA
PKSc1OligoB	TCAAAGGTCCGTGCATGGCGATC
PKSc2OligoA	GGGCCTCTATATATGTCACGTCT
PKSc2OligoB	CGAAGTCCATGCTCAAGACGTGA
PKSc3OligoA	CTCCAGTTGTGCATTTCAGCAGCG
PKSc3OligoB	GGTTAATTCGGGAAGCGCTGCTG
PKSc4OligoA	TCGCGCAACGCGTCGGAAAGCCT
PKSc4OligoB	ATCCTCGTGCTCAAGAGGCTTTC
PKSc9OligoA	CGTATAGCCGACTTTCGCCGACC
PKSc9OligoB	CCATCAACAACGACGGGTCTGGCG
PKSc11OligoA	ACATAGCCGATTTTCATTGGGGTT
PKSc11OligoB	GCTCGCAACGGCATCAACCCCAA
PKSc20OligoA	TGGCCATCCGGAGAAAGGATCAT
PKSc20OligoB	CTTTTCCAGGAAGGGATGATCCT
5LL	GGRTCNCCLARYTGIGTICCICTICCRTGIGC
4UU	MGIGARGCIYTICARATGGAYCCICARCARMG

2.3.13. PCR amplification and sequencing of 16S rRNA genes from the Cullars metagenomic library

All *E. coli* metagenomic library clones were grown separately in 96-well format overnight at 37°C and then the *E. coli* cultures were pooled into a single flask prior to DNA isolation using a Qiagen Large construct kit (Germantown, MD, United States). 16S rRNA genes were PCR amplified from the pooled Cullars metagenomic library DNA template with the “universal Bacteria-specific” primer set 27F and 1492R (Weisburg *et al.*, 1991), using the following conditions: 94°C for 2 min, followed by 30 cycles of 94°C for 30 seconds, 55°C for 15 seconds and 72°C for 1 minute, after which a final elongation step at 72°C for 5 minutes was performed. PCR products were visualized through gel electrophoresis, and the 16S rRNA gene amplicons were purified using the EZNA Cycle Pure kit (Omega Biotek, Norcross, GA) and cloned in *E. coli* using the TOPO-TA cloning kit (Invitrogen, Carlsbad, CA). Transformants were picked into a total number of eight 96-well plates and Sanger sequencing reactions were conducted with primer 27F (Lucigen Corp., Middleton, WI). Sequences were trimmed using the CLC Genomics Workbench (CLC bio, Cambridge, MA) and a BLAST search conducted against the SILVA SSU nr database. All 16S rRNA gene sequences with a top affiliation to *E. coli* were presumed to have been derived from host genomic DNA that contaminated the BAC DNA isolation and were eliminated from the analysis.

2.3.14. PCR amplification and sequencing of 16S rRNA genes from Cullars soil

The original Cullars soil sample used for library construction (stored at -80°C) was used for the 16S rRNA gene survey. Metagenomic DNA was isolated from 0.25 g of the soil sample using an EZNA Soil DNA kit (Omega Biotek, Norcross, GA, United States) and the gDNA was used as a template for bar-coded 16S rRNA gene sequencing targeting the V4 variable region

with PCR primers 515F and 806R (Liu *et al.*, 2007) and were used in a single-step 30 cycle PCR using the HotStarTaq Plus Master Mix Kit (Qiagen, United States) under the following conditions: 94°C for 3 minutes, followed by 28 cycles (5 cycle used on PCR products) of 94°C for 30 seconds, 53°C for 40 seconds and 72°C for 1 minute, and final elongation step at 72°C for 5 minutes. Sequencing was performed at Molecular Research (Shallowater, TX, United States) on an Ion Torrent PGM following the manufacturer's guidelines. Sequence data were processed by removing barcode and primer sequences, then ambiguous sequences, sequences with less than 150 bp or that had homopolymer runs exceeding 6 bp, were removed. The remaining sequences were denoised, OTUs generated and chimeras removed. Operational taxonomic units (OTUs) were defined by clustering at 97% similarity. Final OTUs were taxonomically classified using BLASTn against the SILVA SSU nr database.

2.4. Results

A large-insert metagenomic library was constructed in the shuttle vector pSMART-BAC-S using high molecular weight DNA isolated from the Auburn University's Cullars Rotation agricultural soil (Nasrin *et al.*, 2018). A total of 19,200 independent clones were picked as isolated *E. coli* DH10B colonies into 50 plates in 384-well format. The insert size of 215 randomly selected BAC clones was 12 Kb to >200 Kb, with an average insert size of 113 Kb (Nasrin *et al.*, 2018).

2.4.1. PCR amplification and sequencing of KS domains from the library

Degenerate primer sets targeting the conserved beta-ketoacyl synthase (KS) domain were used to screen for PKS clusters. Each KS domain primer set (**Table 1** Table 1. List of oligonucleotide sequences used in this study.) was tested for its ability to generate a PCR product using a pooled metagenomic library BAC DNA template, in which all 19,200 clones were pooled, and only the primer set 5LL/4UU was observed to give a PCR product (data not shown). The

5LL/4UU primer set was therefore used to amplify each of the 19,200 clones, in two rounds of screening to eliminate false positives, resulting in 925 clones selected, pooled, and sequenced. Contigs from the assembled pooled amplicons were evaluated by antiSMASH2.0, resulting in identification of 50 clones carrying PKS or PKS/NRPS hybrid gene clusters, for a final PCR-based screening hit rate of 0.26%.

2.4.2. NGS of the library using a 3D pooling strategy

Two three-dimensional pooling strategies were devised in order to generate bar-coded libraries for NGS of the 19,200 clone library. In each strategy, every clone in the library was sequenced three times in bar-coded pools for each plate, column, and row, with each pool assembled individually for a total of 210 pools of contigs (**Figure 2**). By comparing the occurrence of each assembled contig in each dimension, a three-coordinate system was generated to associate each clone with a specific well. The two different pooling approaches, one using 384-well formatted plates and the other using 96-well formatted plates, allowed the evaluation of the effects of different numbers of clones per pool for BGC detection. For each pooling strategy, the library was screened by pooling of a) *each plate separately*; b) all wells of *each column across plates* to create columns pools; and c) all wells of *each row across plates* to create a row pool (**Figure 2**).

A bar-coded library was constructed for each respective pool, and all libraries were sequenced together in a single Illumina HiSeq lane. The metagenomic library was divided into 5 sets that each contained 10 plates: Set 1 (Plates 1-10); Set 2 (Plates 11-20); Set 3 (Plates 21-30); Set 4 (Plates 31-40) and Set 5 (41-50). In the first pooling strategy we used a 384-well format for Set 5, while for Sets 1-4 we used a 96-well format in order to reduce the number of clones per each plate (**Figure 1**). This latter method significantly increased the coverage per plate pool, since each pool was composed of a smaller number of clones.

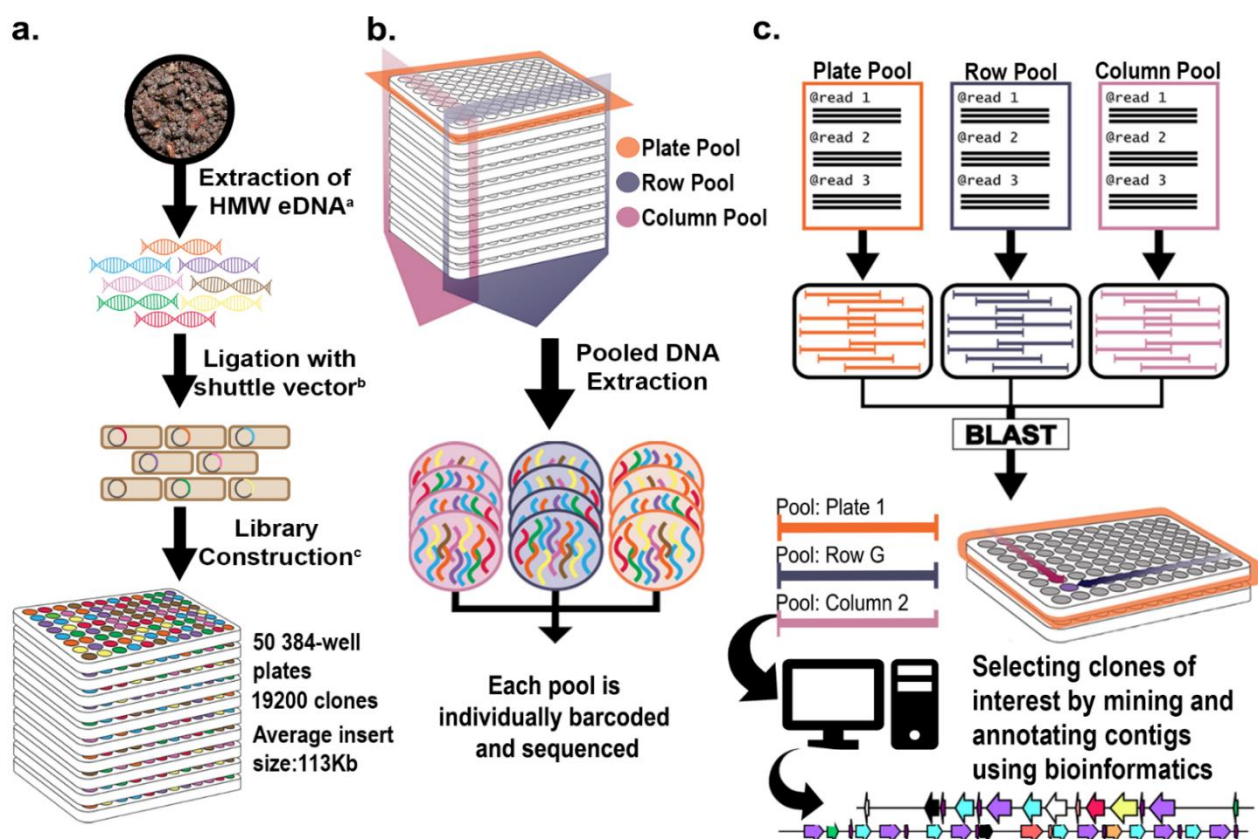


Figure 2. Library construction and 3D pooling. a. Library construction. Superscript letters correspond to citations: a: (Liles et al., 2008); b: (Kakirde *et al.*, 2011); c: (Nasrin *et al.*, 2018). b. 3D Library Pooling strategy: Plate, Column, and Row are barcoded before sequencing. c. Contig to clone deconvolution and *in silico* mining. Reads from each pool are assembled separately into pool contigs. Highly homologous contigs from each dimension (Plate, Column, and Row) form triplets. Since their pool of origin is known, they form a coordinate system that points to their clone of origin. Contigs can be analyzed bioinformatically for rational identification of candidates for functional screening.

A significant number of reads (average of 33%) were removed that corresponded to host or vector sequences. Since the library represents a collection of different genomes, *de novo* assembly was primarily conducted using metaSPAdes. From the 3,157,354 contigs generated, 38% were successfully deconvoluted to a specific BAC clone location, corresponding to 14,976 clones (78% of the library).

The effect of sequencing coverage on assembly statistics was determined. As expected,

higher sequencing coverage obtained with the second pooling approach provided larger contigs. This strong positive correlation ($\rho = 0.398$, $P < 2.2\text{e-}16$) was even stronger when outliers were removed ($\rho = 0.477$, $P < 2.2\text{e-}16$). Scatter plots indicated that increasing clone sequencing coverage resulted in increased contig length until approximately 55X coverage was achieved (**Figure 3**), while increasing sequencing coverage above 55X did not result in increased contig length. In fact, the majority of contigs with the greatest observed coverage had very small lengths, suggesting they correspond to repetitive or highly conserved sequence elements occurring in many microbial genomes.

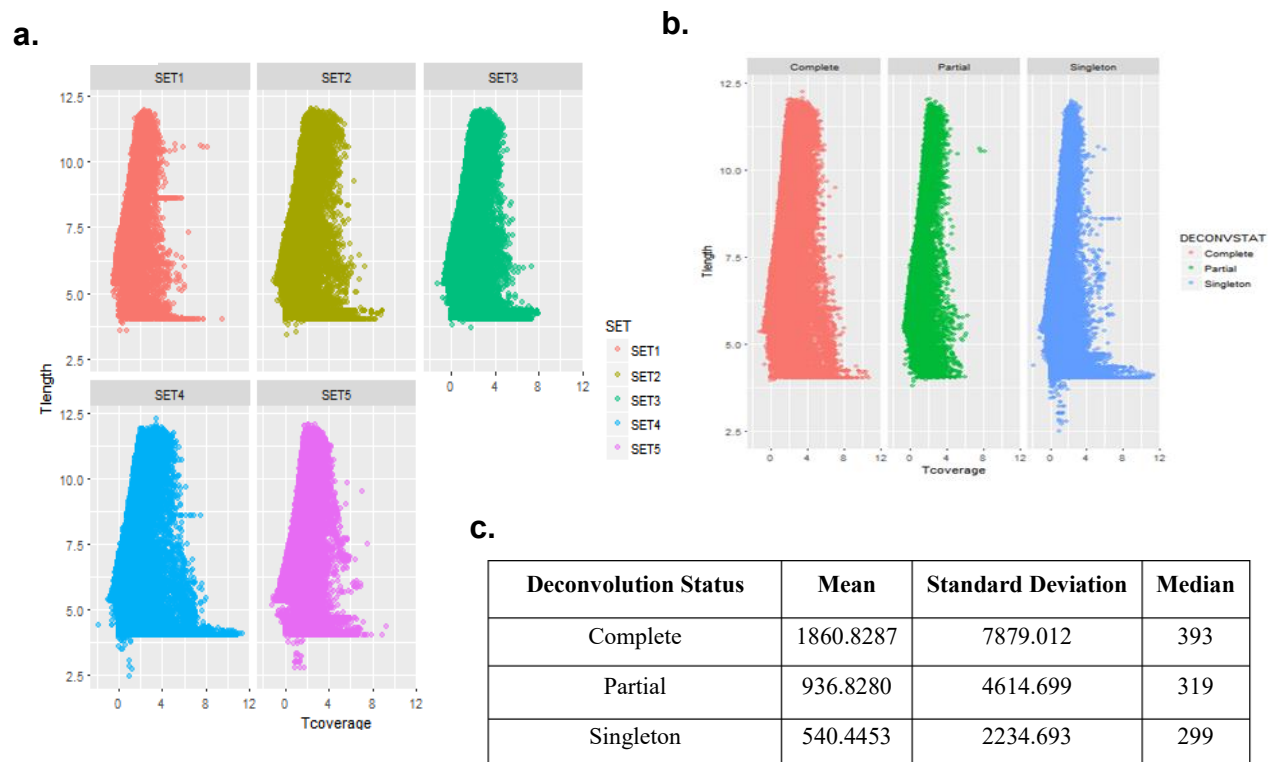


Figure 3. Plots of contig Length x Coverage, both indicated as a log transform. No increase in contig length was observed past 55x coverage (Tcoverage=4), with a very strong influence in final contig length up to 25X. a. Length x Coverage separated by SET (groups of 10 Library plates). b. Length x Coverage separated by deconvolution status. Higher amount of long contigs amongst the completely deconvoluted set. c. Basic statistics of contig lengths according to deconvolution status.

We hypothesized that longer contigs would be easier to deconvolute to specific BAC clones. Both the Kruskal-Wallis and the Dunn test showed that contig length was a significant ($P < 2.2\text{e-}16$) predictor of deconvolution success, with completely deconvoluted contigs having a much higher median and average length compared to singletons (**Figure 3**). Thus, the use of the second pooling strategy, which increased clone coverage, also generated larger contigs and in turn improved final BGC reconstruction from the metagenomic library.

2.4.3. Mining the library for BGCs

Contigs longer than 1 Kb were mined for BGCs using antiSMASH4.0, predicting the presence of 2,082 BGCs within the soil metagenomic library contigs. After deduplicating hits, referring to correcting the number of contigs belonging to the same clone when possible, there were a total of 1,015 BGCs predicted to be present in library clones, corresponding to 5.3% of all library clones (**Table 2**). Of these, 160 of the predicted BGCs corresponded to NRPS clusters, 134 were predicted as PKS clusters, among which were 39 Type I PKS, 12 Type II PKS, 71 Type III PKS, 3 Trans-PKS, and 9 Other KS; and 33 PKS-NRPS hybrid BGCs. From the Type I PKS and/or NRPS hits, 489 (71.5%) were successfully deconvoluted to 223 BAC clones, encompassing all but 9 PCR hits for PKS (the coordinates for which were already known, and these PCR-identified clones were also included within the NGS-discovered contigs). Screening the library for PKS-containing clones using NGS provided a greater number of clones ($n=158$) compared to PCR ($n=50$), which is to be expected given that any single PCR primer set would not be expected to amplify the broad diversity of PKS clusters identified by a homology-based search. Furthermore, NGS screening of the library allowed for parallel mining for different BGC types, leading to a much larger number of BGCs ($n=1,015$) and a diversity of BGC classes ($n=27$) that were recovered from the metagenomic library.

Table 2. Biosynthetic gene clusters identified from the soil metagenomic library by PCR and NGS approach. “NGS Screening Deduplicated” refers to the corrected hit count after hits from contigs deconvoluted to the same clone of origin got deduplicated into a single hit. *Hybrid Pathways included all clusters of mixed types except T1PKS-NRPS.

Biosynthetic Gene Cluster Type	PCR Screening	NGS Screening	NGS Screening Deduplicated	Deconvoluted (%)
Type I PKS	17	75	39	73.3
Type I PKS-NRPS	27	100	33	86.0
Type II PKS	0	22	12	90.9
Type III PKS	1	114	71	84.2
Transatpks	1	6	3	100.0
Other KS	4	14	9	78.6
NRPS	0	509	160	68.4
Other	0	213	110	70.0
Terpene	0	371	201	80.9
Bacteriocin	0	224	127	77.2
Arylpolyene	0	70	36	67.1
Lasso peptide	0	61	32	83.6
Lantipeptide	0	33	23	84.8
Hserlactone	0	27	16	74.1
Resorcinol	0	17	14	88.2
Phosphonate	0	15	11	93.3
Indole	0	13	5	46.2
Ladderane	0	17	16	82.4
Acyl Aminoacids	0	10	8	90.0
Butyrolactone	0	4	3	75.0
Microviridin	0	8	4	100.0
Siderophore	0	4	4	100.0
Cyanobactin	0	2	1	100.0
Thiopeptide	0	3	2	100.0
Linaridin	0	1	1	100.0
Phenazine	0	1	1	100.0
Ectoine	0	1	1	100.0
Hybrid Pathways*	0	147	72	100.0
Total	50	2082	1015	77.5

2.4.4. Diversity of NRPS and PKS clusters identified from the library

The conserved KS domains from PKS clusters were used to build maximum likelihood (ML) dendrograms for direct comparisons between PKS clusters identified either by PCR, by NGS

or present in the MIBiG database, a curated and comprehensive repository of fully annotated and characterized BGCs. NGS screening increased the number of BGCs identified and allowed the identification of more diverse PKS clusters (**Figure 4**). Importantly, some KS domains identified by NGS formed clades with strong bootstrap support that cluster only with library KS sequences, and were not discovered by PCR. These clades also typically had longer branch lengths, indicating greater divergence in their KS domains among each other and compared to representatives of clades identified by PCR (**Figure 4**).

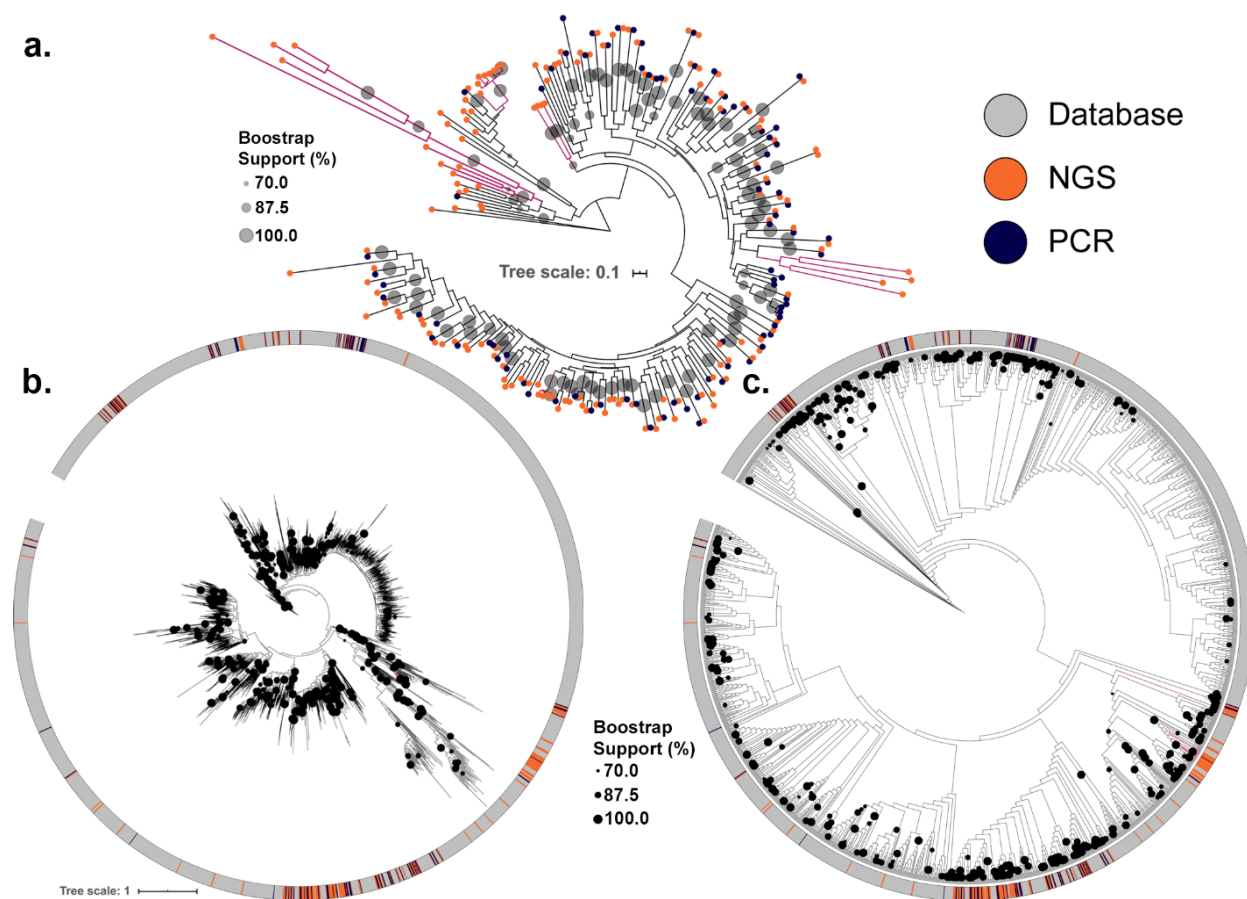


Figure 4. Maximum likelihood tree of KS domains recovered from the soil metagenomic library. Clades in magenta are formed uniquely by domains recovered *in silico*. **a.** KS domains recovered either by PCR or *in silico* (NGS). **b.** KS domains recovered from the soil metagenomic library and from the MIBiG database showing branch lengths. **c.** KS domains recovered from the soil metagenomic library and from the MIBiG database ignoring branch lengths to facilitate topography visualization. Circles in the tree represent branch bootstrap support > 70%.

Comparing the library KS domains to a curated database of known complete PKS clusters confirmed the trend of novelty among the BGCs identified by NGS. Four of the KS domains identified by NGS are maintained as monophyletic clusters that are divergent from sequences present in the MIBiG database (**Figure 5**). Interestingly, the deepest branching KS domains from the soil metagenomic library were solely identified by NGS screening. The library KSs also greatly expanded the diversity of some clades in the database, which correspond to BGCs that may produce novel secondary metabolites.

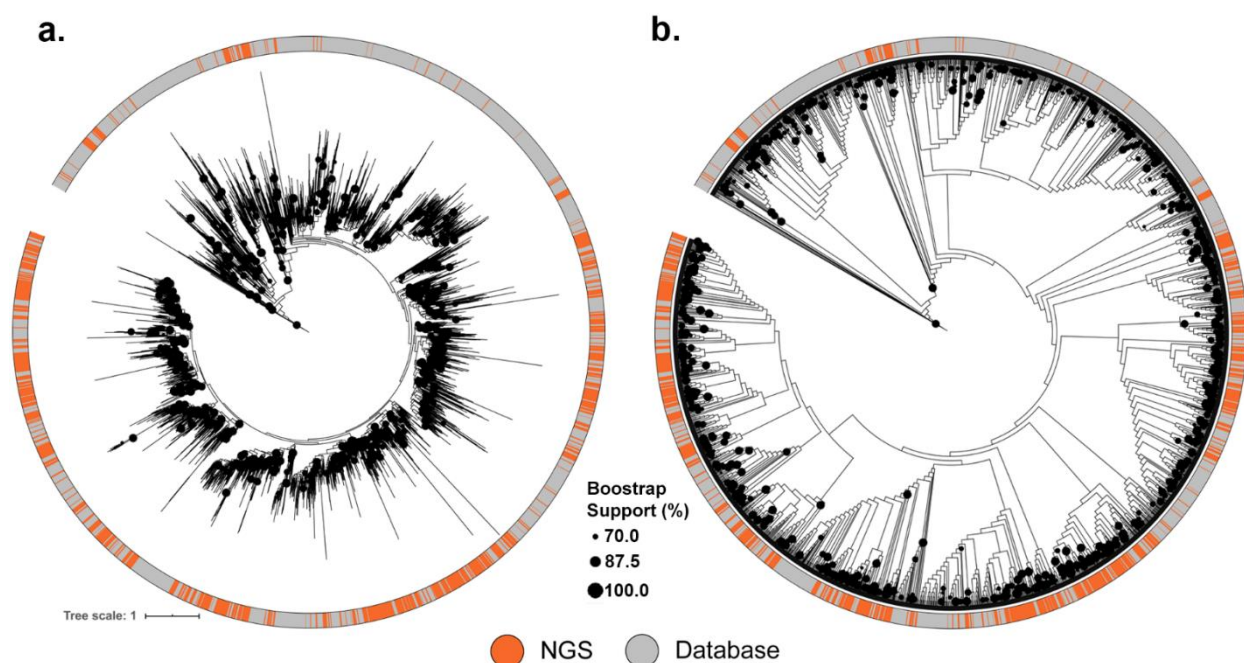


Figure 5. Maximum likelihood tree of A-domains recovered from NRPS clusters identified from the soil metagenomic library and the MIBiG database. a. ML Tree showing branch lengths. b. ML Tree ignoring branch lengths to facilitate topography visualization. Circles in the tree represent branch bootstrap support > 70%.

A similar approach was used to compare the NRPS clusters to the MIBiG database, using conserved adenylation (A) domain sequences. The diversity of A domain sequences identified from the metagenomic library were affiliated with almost every database clade of the tree (**Figure 5**).

Despite the comprehensiveness of the MIBiG database, monophyletic clades formed only by library-derived A domains further expand the diversity of described NRPS clusters.

To better visualize the relationships between library PKS and NRPS clusters versus those contained in the MIBiG database, the library-derived BGCs were clustered using the program BiG-SCAPE and visualized in Cytoscape. The library BGCs were shown to affiliate with diverse Pfam families and did not form links with MIBiG database entries, further highlighting their uniqueness relative to characterized BGCs (**Figure 6**).

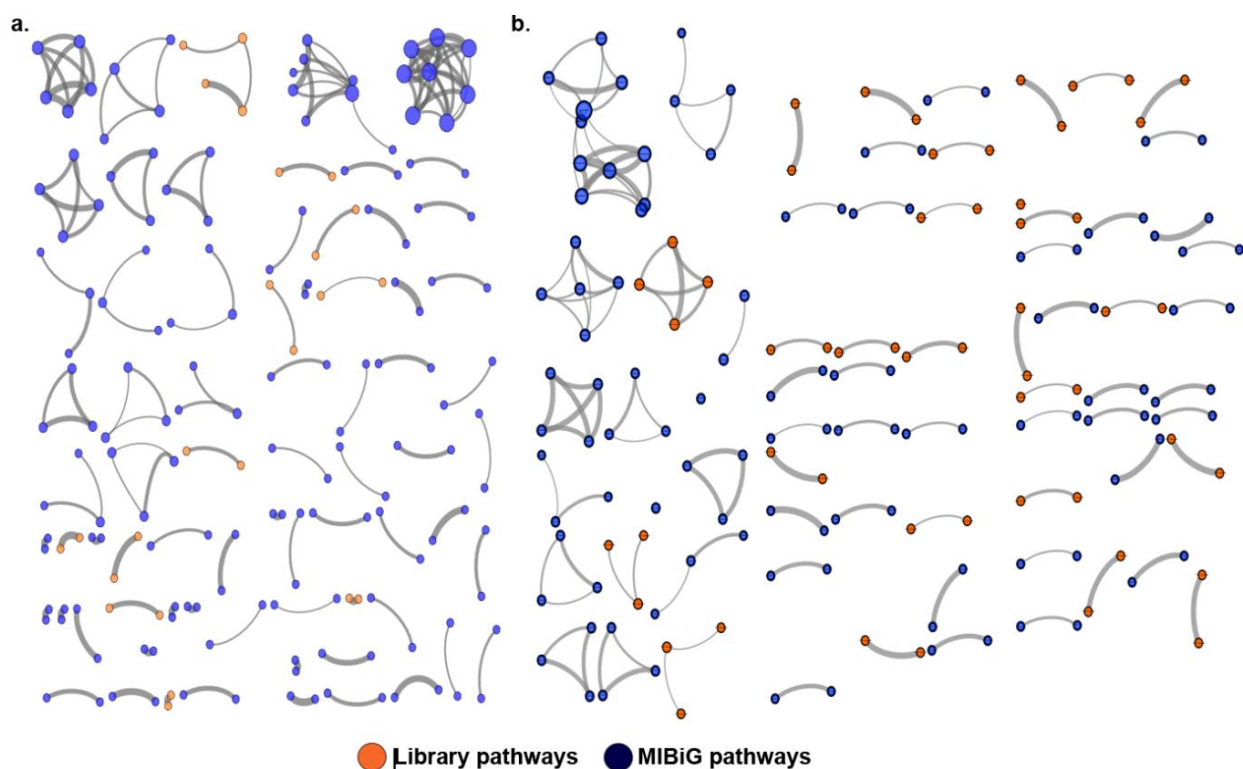


Figure 6. Clustering of PKS (a) and NRPS (b) pathways against the MIBiG database using BiG-SCAPE. Library pathways are shown in orange and database pathways in blue. Circle size is proportional to the number of connections, and strike width is proportional to the similarity between pathways.

The taxonomic origin of contigs containing PKS and/or NRPS gene clusters were inferred by comparing K-mer profiles using Kraken (**Figure 7**), since only one clone (P35L08) also carried

a 16S rRNA gene, impeding the general use of this gene for clone taxonomic classification. In the case of clone P35L08, it was predicted to be derived from bacteria in the genus *Pseudolabrys* within the *Rhizobiales* order of α -Proteobacteria, which aligns with the prediction of the 16S rRNA gene annotation. This result supports the use of k-mer profiling of BGC-containing clones to infer their taxonomic origin. Further taxonomic resolution may be possible for some metagenomic clones depending upon the availability of phylogenetically informative gene sequences.

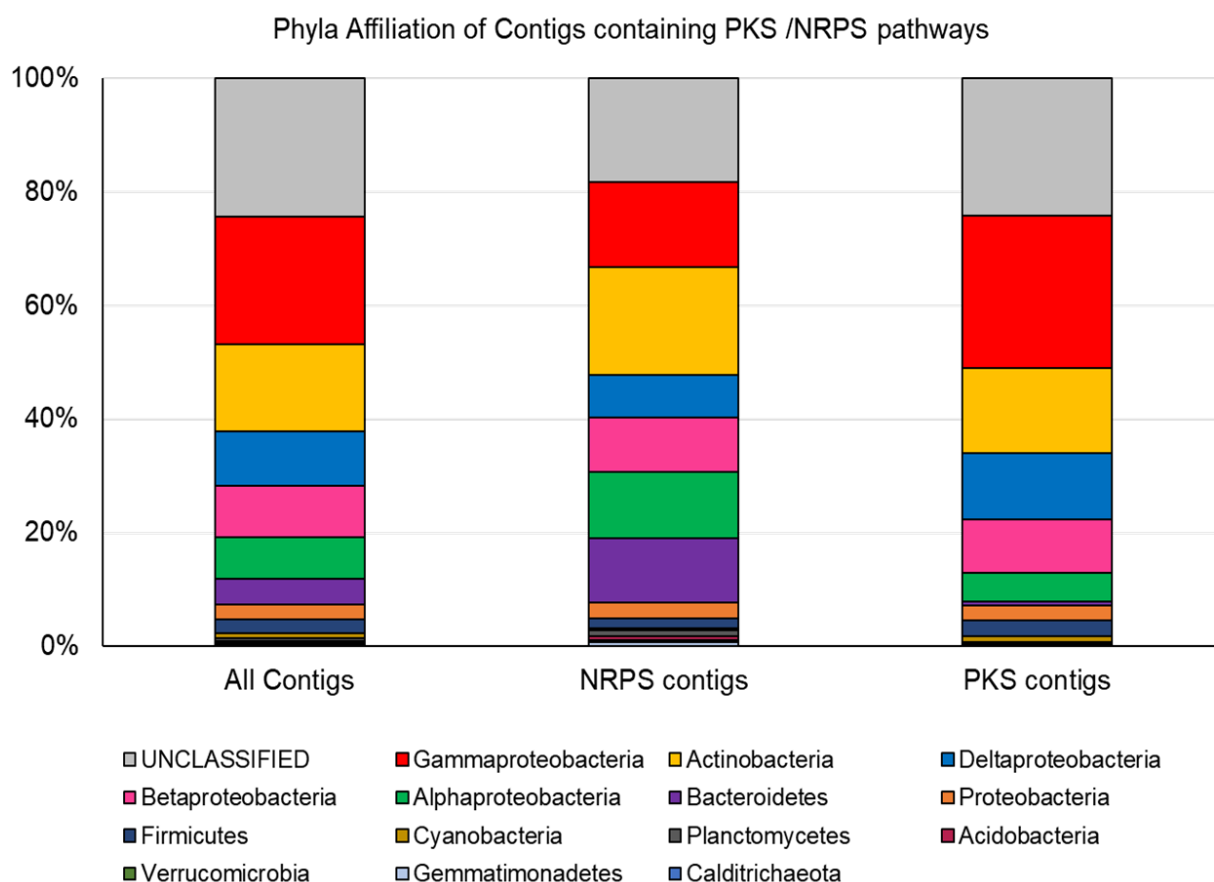


Figure 7. Taxonomic classification of the Contigs carrying PKS and/or NRPS pathways by kmer frequencies.

The largest taxonomic category for BGC origin was indicated as “unclassified”, with the phyla Proteobacteria and Actinobacteria being the most frequent among the 11 phyla identified,

followed by the phyla Bacteroidetes and Firmicutes (**Figure 7**). Interestingly, members of the phylum Actinobacteria were the most commonly predicted source of NRPS clusters, whereas Gamma-Proteobacteria were the most common source of PKS clusters. Bacteroidetes and Firmicutes were also frequently predicted to be sources of NRPSs.

2.4.5. Recovery of PKS and NRPS clusters from the metagenomic library

From the deconvoluted clones predicted to contain PKS or NRPS clusters, BAC inserts of 221 clones were resequenced and annotated to validate the predictions from the metagenomic library contigs. Insert sequences were assembled using biosyntheticSPAdes and these BGCs were then annotated using antiSMASH 5.0. Validation was done by comparing antiSMASH 5.0 predictions for each clone and its associated contigs, and by comparing sequence similarity using BLAST. The BGCs from validated clones were then manually inspected and deemed viable if they contained at least 3 complete adjacent modules (for Type I PKS and NRPS). A total of 195 clones from the 221 clones resequenced were validated, with 117 of the inserts carrying viable PKS or NRPS BGCs while the rest were predicted by antiSMASH5.0 to contain truncated modules (**Appendices, Supplemental Table 1; Appendices, Supplemental Figure 1**).

Visualization of library BGCs in terms of their divergence from BGCs in the MIBiG database and their size can aid the process of selecting clones of interest that encode viable PKS/NRPS clusters (**Figure 8**). The majority of the viable BGCs recovered had 10-20 domains with larger clusters being less frequent. Hence, the importance of increased insert size, which allowed for the recovery of very large clusters (i.e. the largest having 77 predicted PKS-NRPS domains) was demonstrated despite their relative sparsity in the metagenomic library. The A domains and KS domains from the library BGCs were compared to the GenBank nr/nt database by BLASTp and the majority showed moderate similarity (mean of 54.7% identity, ranging from

23.8% to 95.0% identity) to other A and KS domains in the nr/nt database. This moderate similarity can also be observed in terms of the branch lengths of the conserved KS and A domains carried by each viable BGC when compared to the MIBiG database. Most of the KS domains had a branch length of around 2, whereas with the A domains a longer branch length around 3 was more frequently observed (**Figure 8**).

The A domain and KS domain dendograms allowed for selection of three interesting representatives for more detailed annotation, including PKS-containing clone P33I07, NRPS-containing clone P39I17, and PKS-NRPS-containing clone P29O18. These three clones were the largest recovered BGCs of their type to be identified. The PKS-NRPS hybrid representative, clone P29O18, harbored the longest BGC recovered (>100 Kb) and the synteny between the contig and the insert clusters were highly similar (**Figure 8**). The BGC consisted of 22 complete modules and 4 incomplete modules. Most of the modules were assigned as NRPS machinery; however, one complete PKS module was included. Several methyltransferase (MT) domains were predicted throughout the BGC which could have an important role in modification of the peptide intermediate. Additionally, the presence of thioesterase (TE) domains in the BGC may drive the release of the putative product.

The largest PKS recovered was a 30 Kb Type I PKS BGC from clone P33I07 which, if active, is predicted to synthesize a linear polyketide. Annotation of this gene cluster found that it contained 4 complete modules consisting of the minimum three catalytic domains: an acyltransferase (AT), a KS domain, and an acyl-carrier protein (ACP). Each module carried the characteristic reductive ketoreductase (KR) domains which often double as epimerization domains and can alter the linear structure predicted by antiSMASH. In addition, one module carries a dehydrogenase (DH) domain to further tailor the nascent polyketide.

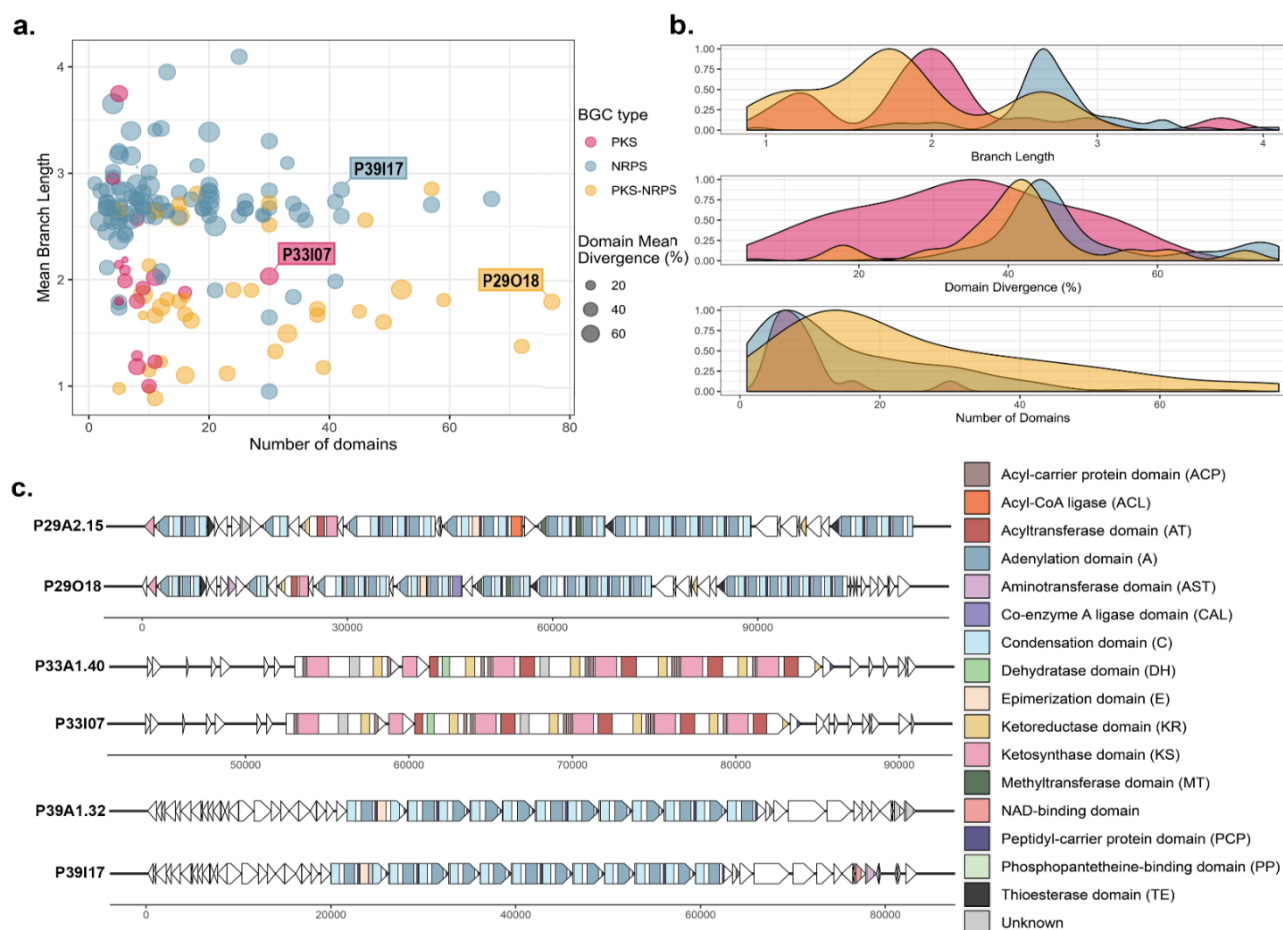


Figure 8. Analysis of the clones containing viable PKS and NRPS pathways reconstructed and validated from the Cullars metagenomic library. **a.** Viable PKS and/or NRPS clusters recovered from the library compared by size (number of domains) and divergence from known BGCs (mean branch length and domain mean divergence). Each plot point represents a clone containing a viable BGC. Mean branch length for each clone is calculated as the average branch length of all the clone's KS and/or A domains, as extracted from the ML trees against the MIBiG database. Mean domain divergence for each clone is calculated as the average domain divergence of all clone's KS and/or A domains. Domain divergence refers to the complementary percentage to the % Identity of the domain to the NCBI nr database (if a domain has % identity of 60, then the domain divergence is 40%). **b.** Density distribution of branch length, domain divergence and number of domains across the viable PKS and NRPS BGCs. **c.** Three interesting representatives, being the longest BGC from each type (PKS, NRPS, PKS-NRPS) to be recovered successfully *in silico* from the metagenomic library NGS contigs and validated upon clone resequencing. Annotation from the library contig was compared to the annotation from their corresponding validated insert sequence. ORFs are showed by arrows and the domains within each ORF are represented by colored stripes.

Clone P39I17 was selected as an NRPS representative for annotation based on the completeness of the BGC and >40% domain divergence. This BGC spanned 40 Kb with 12

complete modules consisting of the three core domains-a condensation (C) domain, an A domain, and a thiolation/peptidyl-carrier protein (PCP) domain. The starter module also carries an additional enzymatic, epimerization (E) domain capable of converting amino acids between the L- and D- isomers (Felnagle *et al.*, 2008). One incomplete module lacking an A domain was predicted by antiSMASH 5.0 and no tailoring domains were evident on the BGC. Core structures predicted from antiSMASH 5.0 suggests that a linear NRP could be produced with low similarity to the antitumor peptide hemiasterlin B (Anderson *et al.*, 1997).

2.4.6. Phylogenetic Diversity of Cullars soil and the Cullars soil metagenomic library

To determine the phylogenetic diversity captured in the metagenomic library, 16S rRNA genes were obtained from three sources: (1) PCR amplification of DNA isolated from the original Cullars soil sample used to construct the library (Original Cullars soil survey); (2) PCR amplification of the pooled DNA template from all metagenomic library clones (Cullars metagenomic library survey); and (3) Screening for 16S rRNA genes from the metagenomic library NGS contigs (Cullars metagenomic library NGS). The Original Cullars soil survey revealed a microbiota composition dominated by Proteobacteria (32.4%), Actinobacteria (15.6%), Acidobacteria (14.3%) and Bacteroidetes (10.0%), with a significant presence of Planctomycetes, Verrucomicrobia, Mixococcus, Chloroflexi and, at lower frequencies, Gemmatimonadetes (2.5%), Firmicutes (1.0%) and Cyanobacteria (0.5%). In addition, 23 low abundance phyla were identified in the original soil sample including most notably Nitrospirae, Armatimonadetes, Methylomirabilota and CPR, in addition to several other understudied candidate taxa (Lynch & Neufeld, 2015) (**Figure 9**).

For the Cullars metagenomic library survey, 295 16S rRNA gene amplicons were annotated after excluding *E. coli* host amplicons. The library 16S rRNA gene amplicon sequences were

affiliated with 10 bacterial phyla, with Acidobacteria, Actinobacteria, Bacteroidetes, Gemmatimonadetes, and Proteobacteria accounting for 97.0% of the total observed ribotypes. Chloroflexi, Firmicutes, Planctomycetes, Armatimonadetes, and Verrucomicrobia constituted the remaining ribotypes (**Figure 9**). It should be noted that, while the phyla representation was generally similar for the PCR amplified 16S rRNA gene sequences obtained from the original soil, or the metagenomic library, that two different primer sets were used (due to the need for short amplicons for NGS in the case of the original soil), and that these two 16S amplicon data sets are not directly comparable.

16S rRNA sequences recovered from the Cullars metagenomic library NGS of the metagenomic library identified 457 ribotypes affiliated with 20 different phyla. There was a predominance of Proteobacteria (32.9%), Bacteroidetes (22.7%), Acidobacteria and Actinobacteria (9.5%). NGS of the library was also more sensitive in detecting members of the PVC superphylum (Planctomycetes, Verrucomicrobia, and Chlamydiae), Cyanobacteria and Chloroflexi, as well as revealing that genomic inserts from the rare soil biosphere were captured, albeit in low abundance (**Figure 9**). Due to the differences in numbers of 16S rRNA genes sequenced from the library PCR amplicons vs. the library NGS, it is not surprising that a greater number of phyla were observed; however, these ribotypes identified solely from library NGS are affiliated with six classes of CPR, in addition to Armatimonadetes, Abditibacteria and Candidate division WS2, all of which were also not identified from the original soil 16S survey and are typically missed by PCR amplicon surveys and only observed via direct metagenomic sequencing (Brown, C. T. *et al.*, 2015). Hence, just as we had observed based on BGC diversity, using NGS provided a more exhaustive picture of the phylogenetic diversity contained within the Cullars metagenomic library.

When compared to the original soil sample, the metagenomic library successfully captured representatives of the most common phyla in roughly similar proportions, with a noticeable reduction in the relative abundance of Actinobacteria when compared to the original soil. Recovery of rare taxa is a challenge when building metagenomic libraries, however the Cullars metagenomic library was able to capture some representatives of lower abundance phyla present in the original soil sample (**Figure 9**).

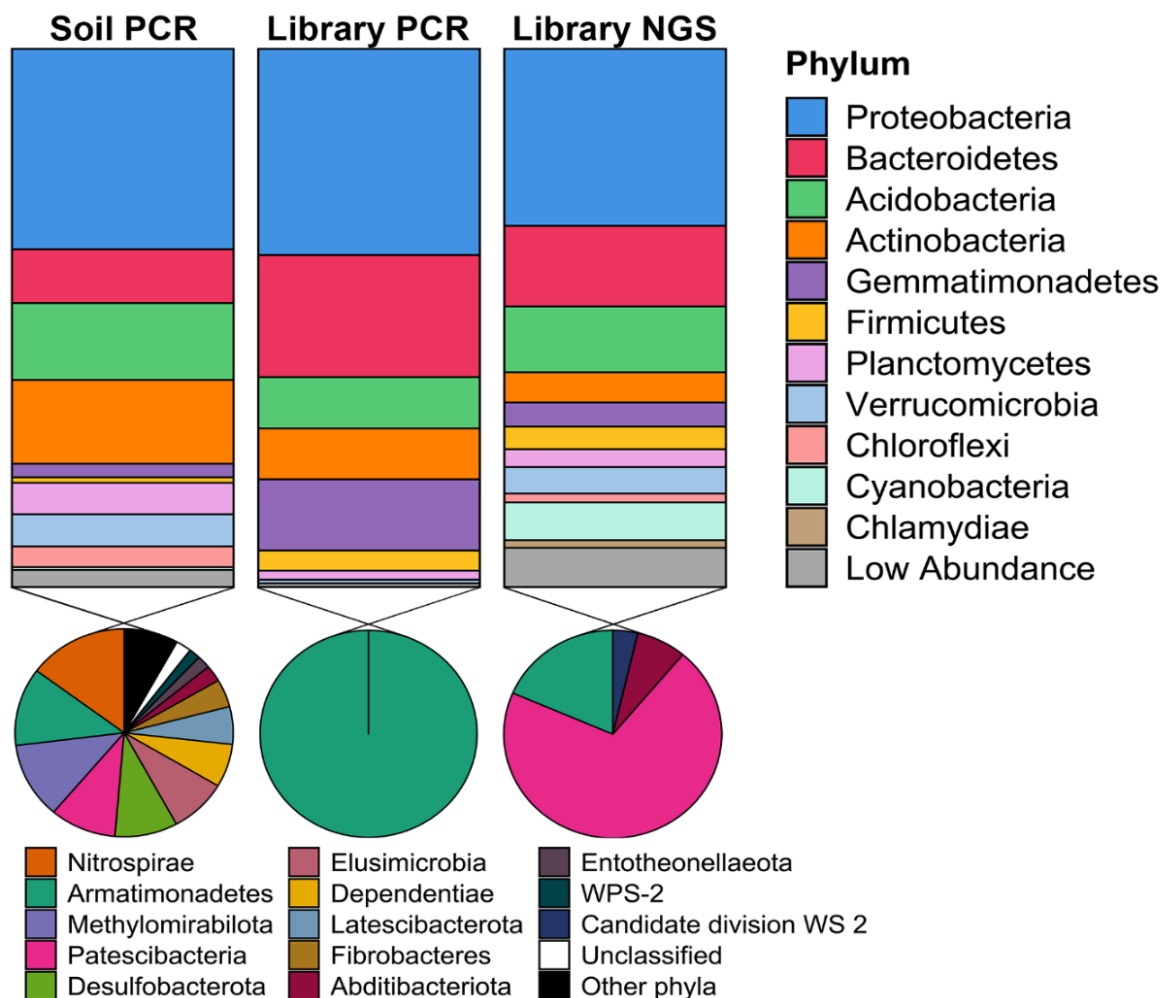


Figure 9. Taxonomic diversity of the Cullars metagenomic library and of the original soil sample used for library construction. Soil PCR: 16S rRNA gene PCR amplification of DNA isolated from the original Cullars soil sample used to construct the library; Library PCR: 16S rRNA gene PCR amplification of the DNA template from all metagenomic library clones pooled together; Library NGS: *In silico* mining for 16S rRNA genes from the metagenomic library NGS contigs. The phyla contained in the section “Other phyla” are: RCP2-54, SAR324, Spirochaetes, Sumerlaeota, WPS-2; all with relative frequencies below 0.05%.

2.5. Discussion

There is no better crucible for the generation of novel natural products than microbial competition in natural environments, especially among soil microbiota, from which many and diverse natural products have been derived (Wohlleben *et al.*, 2016). *In situ* culturing techniques have been successful in cultivating previously uncultured microorganisms that can produce bioactive metabolites (Ling *et al.*, 2015). Nevertheless, the vast majority of environmental microbiota can only be exploited for their natural products via culture-independent approaches (Milshteyn *et al.*, 2014). Indeed, functional screening of metagenomic clone libraries has led to the discovery of structurally and functionally diverse antimicrobial compounds (Hover *et al.*, 2018; Kallifidas *et al.*, 2012; Kang & Brady, 2013; Kang & Brady, 2014a; Nasrin *et al.*, 2018), bioactive compounds (Chang & Brady, 2013; Fisch *et al.*, 2009; Owen *et al.*, 2015; Zimmermann *et al.*, 2009), novel enzymes (Cheng *et al.*, 2017; Kusnezowa & Leichert, 2017; Lewin *et al.*, 2017; Nasser *et al.*, 2018; Wang *et al.*, 2010), and novel promoters (Westmann *et al.*, 2018).

Unfortunately, many technical hurdles such as inadequate metagenomic clone insert sizes, biased detection of BGCs, and poor heterologous expression, render functional screening suboptimal, laborious and inefficient (Palazzotto & Weber, 2018). Therefore, rational selection of clones that are known to encode BGCs of interest for subsequent targeted heterologous expression could greatly improve the efficiency of identifying secondary metabolites expressed from metagenomic BGCs. Targeted screening of metagenomic libraries based on arrayed PCR or other amplification methods has been developed to enrich clones containing genes of interest, which can then be heterologously expressed and screened for bioactivity (Charlop-Powers *et al.*, 2016; Owen *et al.*, 2013). Other approaches relying on functional complementation can further isolate clones of interest (Bitok *et al.*, 2017; Charlop-Powers *et al.*, 2013). However, these ingenious approaches

do not provide insight into the BGCs or other functional genes carried by the metagenomic inserts in advance of their expression, so that selection of an ideal heterologous host for a specific BGC of interest is not possible.

This study demonstrates a bioinformatic pipeline for complete metagenomic library sequencing and mining for BGC-containing clones, providing full annotation of inserts prior to targeted heterologous expression. Attempts to leverage NGS capabilities to sequence collections of metagenomic clones have so far relied on a hybrid approach, for example using “tags” for each clone that are created via amplification and Sanger sequencing of individual clones, then pooling the clones and performing sequencing using NGS (Calderon *et al.*, 2019; Dzunkova *et al.*, 2012). In contrast, this study presents a strategy capable of directly sequencing large collections, while maintaining insert positional information within the library, a critical step in targeting specific cloned BGCs for conjugal transfer to a compatible heterologous host and functional screening for bioactive secondary metabolites. Thus, this pipeline expands the capabilities of previous targeted screening strategies by circumventing PCR bias and individual tagging, while allowing the use of bioinformatics to mine for diverse genomic features of interest. Given the known biases associated with PCR, it is plausible to assume that this approach would hasten the discovery of other classes of BGCs as well. Not only does NGS-based screening generate more hits, but this strategy results in discovery of highly diverse BGCs, which was evident for the PKS and NRPS clusters that are very different from those currently characterized. Hence, it is evident that our current exploitation of bioactive natural products from metagenomic libraries have been limited by earlier methods, and environmental BGCs remain a largely understudied and untapped resource for natural product discovery.

Furthermore, BGC annotation can be extended to prediction of metabolite structure, prior

to heterologous expression. Indeed, we were able to annotate and deconvolute complete PKS and NRPS clusters, the longest of which spanned over 77 domains, and predict the linear structure of their encoded metabolites. While the metabolites expressed from these metagenome-derived BGCs in a heterologous host may be significantly different from the metabolites expressed in the original host, and the heterologous host may significantly affect BGC expression and metabolite chemistry (Gillespie *et al.*, 2002), this approach does allow targeted, rational selection of BGC and expression host combinations. This approach can also improve the ability to detect and express other relevant classes of gene products, such as biotechnologically relevant enzymes. While this study focused on characterizing NRPS and/or PKS clusters, future research will explore the diversity of enzymes and other natural products encoded within the soil metagenomic library.

Most of the library PKS and NRPS clusters were not affiliated with any particular phylum, stressing the biotechnological potential of as-yet-uncultured bacteria such as CPR or Acidobacteria taxa, as potentially novel sources of BGCs to be further explored (Crits-Christoph *et al.*, 2018; Parsley *et al.*, 2011; Sharrar, Allison M *et al.*, 2020). Among those clusters that were classified, we observed a predominance of well-studied phyla, such as Proteobacteria, Actinobacteria, Bacteroidetes, and Firmicutes, whereas abundant yet less studied phyla, such as Acidobacteria, were underrepresented. The paucity of BGCs previously obtained from Acidobacteria and other abundant but poorly characterized soil microbiota is expected to have biased the prediction of BGC origin. In support of that hypothesis, manual inspection of the complete insert sequence for BGC-containing clones that have been shown to express antibacterial metabolites in a *Streptomyces coelicolor* host have suggested an Acidobacteria origin (data not shown). Nonetheless, there are notable shifts in relative abundance between phyla within the metagenomic library when compared to the original soil, which suggests there were biases associated with library construction (Lam &

Charles, 2015). Improvements in DNA extraction and cloning techniques may increase the representativeness of microbial genomes accessed from a metagenomic library, widening the scope of BGCs available for functional screening. In spite of such biases, the Cullars soil metagenomic library was found to contain cloned genome fragments from diverse bacteria, including underrepresented taxa such as the CPR.

The Cullars metagenomic library survey of 16S rRNA genes using PCR amplification failed to recover many phyla that were identified by library NGS. The 16S survey of the metagenomic library therefore had the same biases encountered with KS amplicon screening, which were likewise overcome using a NGS approach. While efforts to revise universal primers have been proven useful (Caporaso *et al.*, 2011; Frank *et al.*, 2008; Klindworth *et al.*, 2013), the inherent biases associated with PCR amplification will continue to plague efforts to catalogue BGC and phylogenetic diversity. Akin to the unveiling of microbial diversity such as CPR and the Asgard superphylum of Archaea by shotgun metagenomic sequencing (Zaremba-Niedzwiedzka *et al.*, 2017), the application of NGS to a soil metagenomic library has exposed an unprecedented diversity of BGCs from soil microorganisms that have been previously overlooked because of culture- and PCR-biases. The application of this NGS strategy to other metagenomic libraries would therefore be predicted to reveal heretofore undiscovered functional and phylogenetic diversity that have been captured in these genomic resources.

This pipeline provides a powerful way to annotate metagenomic libraries for rational exploration and screening of genomic features of interest. In the context of BGCs, we demonstrated that library sequencing can significantly improve hit rate, BGC diversity, and elucidation of pathways that may be then targeted for heterologous expression and production of bioactive compounds. The ability to fully leverage NGS and bioinformatic tools will enable a more

complete assessment of the functional and phylogenetic diversity among metagenomic libraries and permit their full exploitation for biotechnological research.

Chapter 3: Enrichment of ultra-small cells from topsoil sheds light into the unique characteristics of soil dwelling CPR

3.1. Abstract

Candidate Phyla Radiation (CPR) were first observed in amplicon-based surveys of diverse and often inhospitable aquatic environments. The advances in genome resolved metagenomics revealed that CPR is in fact a hyper-diverse group of ultra-small bacteria, with versatile and yet sparse metabolisms, playing roles in biogeochemical cycles and complex biomass degradation. However, the insights into this cryptic group comes from a surprisingly small number of environments, namely aquifers, sediments and groundwater. Recovery of CPR from soils has been hindered due to their extremely low abundance in the sea of a complex community.

Here we developed a protocol for enrichment of the ultra-small bacterial fraction from soil into a mini-metagenome prior to shotgun sequencing. This strategy was applied to study ultra-small bacteria from soil samples from the Cullars Rotation, by surveying plots under different controlled nutrient deficiency experiments. Fourteen plots were surveyed and 42 mini-metagenomes sequenced enabling the reconstruction of 26 quality CPR metagenomically assembled genomes (MAGs), in addition to representative genomes from Gemmatimonadetes, Dependentiae and Chlamydiae phyla, with average completeness of 86.05% and contamination below 5%. Further taxonomic characterization against 920 reference genomes, showed that the CPR MAGs belonged to the phyla Parcubacteria, Saccharibacteria and ABY1, while forming monophyletic clades within these groups, and a deep branching clade related to Saccharibacteria which may correspond to a novel CPR group.

Cullars Saccharibacteria was differentiated in their ability to process biomass, being

enriched in enzymes associated with cellulose, starch, chitin and xylan degradation when compared to the reference genomes, indicating a role in soil biomass degradation and carbon turn over. Although CPR were recovered in all plots, community fluctuations were often not taxa dependent suggesting diversity of adaptational strategies within the same groups. Moreover, nutrient was not correlated with community shifts, but instead pH, Aluminum and Iron concentrations. These findings corroborate the hypothesis that soil-dwelling CPR are differentiated from their aquatic counterparts, with phylogenetic and metabolic fingerprints unique to soils, expanding the extant diversity within this group.

3.2. Introduction

Since the discovery of Microgenomatia (OD11) from the Obsidian pools in Yellowstone National Park (Hugenholtz, Philip *et al.*, 1998), they have been reported in myriad environments leading to the discovery of the Parcubacteria (OD1) and Asconditabacteria (SR1) taxa (Harris *et al.*, 2004). Since then, the systematic study of these uncultured microorganisms led to the proposition of the hyper-diverse monophyletic Candidate Phyla Radiation (CPR) (Brown, C. T. *et al.*, 2015), comprising more than 20% of the known bacterial diversity (Hug *et al.*, 2016; Parks *et al.*, 2017)

Members of this group have small cells ($0.009 \pm 0.002 \mu\text{m}^3$) (Luef *et al.*, 2015) and unusual metabolisms. They commonly lack or have severe loss of major pathways such as nucleotide, amino acid and lipid metabolism, lack TCA cycle or complete electron transport chains relying on alternate mechanisms to generate and exploit the proton motor force (Kantor *et al.*, 2013; Wrighton *et al.*, 2012). Despite these losses, CPR have innovative features such as alternate fermentation enzymes (Kantor *et al.*, 2013), hybrid RuBisCOs that function in an archaeal-like fashion (Wrighton *et al.*, 2016) and deployment of other archaeal like enzymes (Ghuneim *et al.*, 2018).

Additionally, they have been predicted to participate in Nitrogen, Sulfur and Carbon cycling (Danczak *et al.*, 2017; Wrighton *et al.*, 2014). The foundation of our current understanding of this taxa focused on aquifers, groundwater, and associated sediment (Brown, *et al.*, 2015; Castelle *et al.*, 2013; Danczak *et al.*, 2017; Luef *et al.*, 2015; Wrighton *et al.*, 2012). Therefore, the study of these ultra-small bacteria in other environmental contexts is imperative to elucidate the extent of their diversity, metabolic capabilities, and ecological roles.

Soil environments harbor enormous bacterial diversity (Torsvik & Ovreas, 2002), however members of the CPR group have only sparsely been reported in soil microbiome surveys (Santana-Pereira *et al.*, 2020; Wagner *et al.*, 2009; Yeates *et al.*, 2000). Recovery of soil dwelling CPR is challenging due to their extremely low abundances (Herrmann *et al.*, 2019) and intronic 16 rRNAs genes that evade common amplicon-based surveys (Brown, C. T. *et al.*, 2015).

Genome-resolved metagenomics strives to reconstruct complete genomes from metagenomes using a range of techniques to assort assembled sequences with the same genomic origin into “bins” allowing the study of the metabolic potential of uncultured microorganisms (Sharon & Banfield, 2013). However, the challenges in assembling sequences from complex soil metagenomes (Howe *et al.*, 2014) has hindered the reconstruction of these low abundant CPR genomes from soil. To date only two studies achieved this goal (Kroeger *et al.*, 2018; Woodcroft *et al.*, 2018)

In this study, an ultra-small cell enrichment protocol was developed to allow the preferential DNA extraction of low abundant CPR and other members of the rare biosphere from soil samples of the Cullars Rotation experimental station (Auburn, AL, USA). Direct sequencing of the enrichments allowed the reconstruction of 29 Metagenomically Assembled Genomes (MAGs) predominantly from CPR. CPR were accessed for their biomass degrading potential and

abundance across environments with controlled nutrient deficiency shedding light into their possible ecological roles and distribution patterns in soil.

3.3. Methods

3.3.1. Sampling site

Soil samples were obtained from the Cullars Rotation experimental station in Auburn, AL on March 2019. Cullars Rotation is a century old crop rotation experiment alternating between five crops during the course of a year (cotton, crimson clover, corn, wheat and soybean) (Mitchell *et al.*, 2011). The rotation is composed of 14 plots for the study of the effects of controlled deficiency of phosphorus (P), potassium (K), nitrogen (N) and sulfur (S) – as well as other soil amendments such as lime (for pH control) and addition of a micronutrient mix containing boron (B), zinc (Zn), manganese (Mn), copper (Cu) and iron (Fe) – in crop yield (Mitchell *et al.*, 2005) (**Figure 10**).

3.3.2. Sample collection and soil analysis

Three topsoil samples from each one of the 14 plots in Cullars Rotation were collected for soil DNA extraction, alongside ~400g of bulk soil for analysis (**Figure 10**). The samples for DNA extraction were immediately transported to the lab and kept refrigerated at 4 °C until processing. Processing of the 42 resulting samples were completed in the same day as sampling. The bulk soil samples were submitted for analysis at Auburn University's Soil Testing Laboratory the same day. Soil nutrients were obtained using Mehlich I Extraction and analyzed by ICP. Measurements were obtained for: calcium (Ca), K, magnesium (Mg), P, aluminum (Al), B, Cu, Fe, Mn, sodium (Na), Zn, Nitrate-Nitrogen (NO₃-N), all in parts per million (ppm). Percent carbon (C) and nitrogen (N), and pH were also measured.

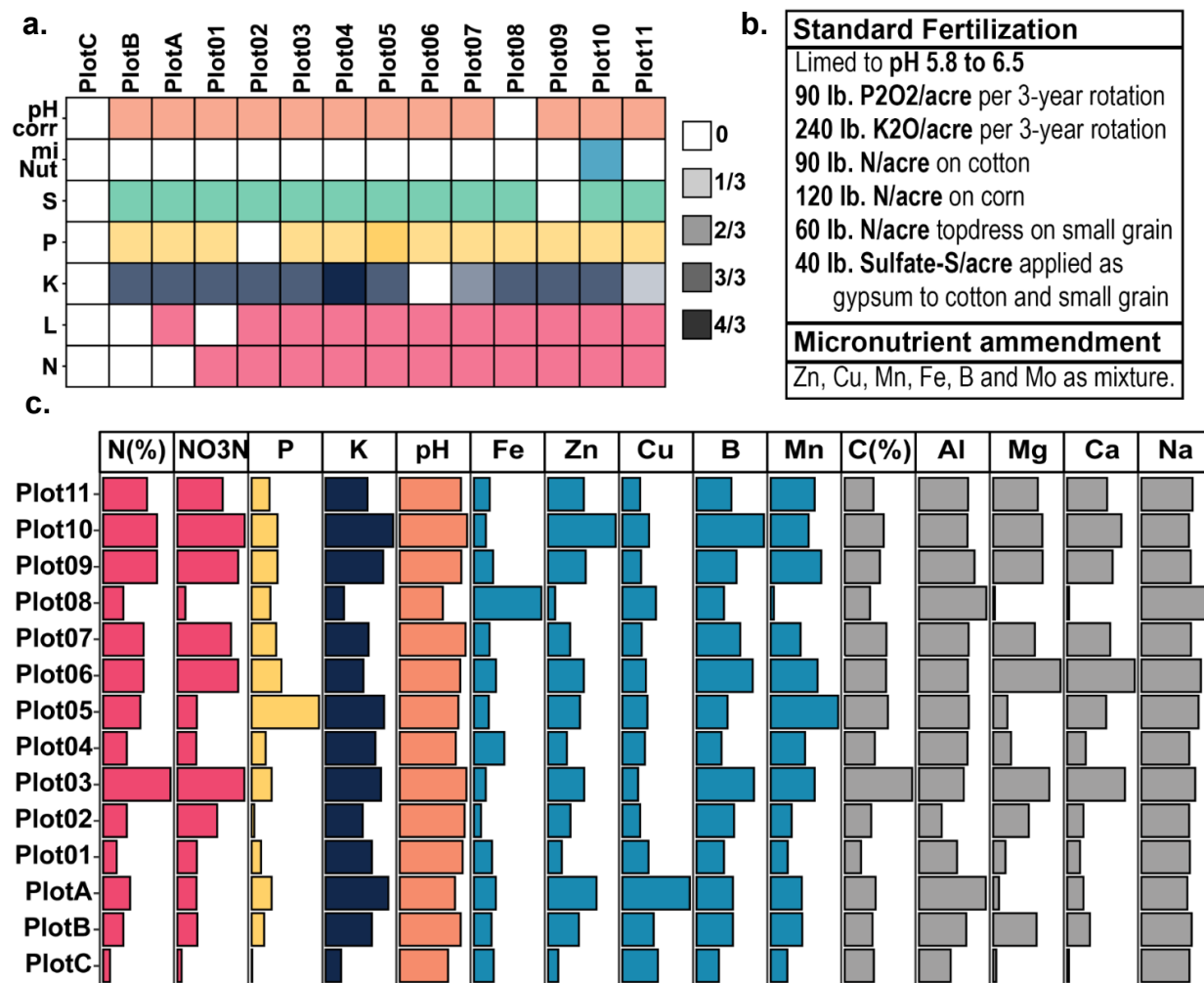


Figure 10. Representation of Cullars Rotation plots and Fertilization treatments. **a.** Schematic representation of Cullars Rotation Plot layout and Soil treatments administered. **pH corr:** pH correction with lime; **mi Nut:** Addition of micronutrient mix; **S:** Addition of Sulfur; **P:** Addition of Phosphorous; **K:** Addition of Potassium; **L:** Rotation with winter legumes; **N:** Addition of Nitrogen. Fractions represent the amount added to the plot with Standard Fertilization as reference (3/3). **b.** Table with the Standard Fertilization treatments applied to each plot.(Mitchell *et al.*, 2005). **c.** Average soil nutrient parameters and pH from sampled bulk soil.

3.3.3. Ultra-small cell enrichment

For each sample, 5g of soil was suspended in 25 ml of 0.2% tetrasodium pyrophosphate buffer (pH 8.0), which has been shown to dissociate microbial cells from soil particulates (e.g. clay, sand and silt) while retaining the microbial community structure associated with bulk soil (Portillo *et al.*, 2013b; Yuan *et al.*, 2018). Suspensions were homogenized for 10 min with gentle

inversions, followed by 10 cycles of intermittent vortexing for 5 seconds. The soil slurry was subjected to a low-speed centrifugation (5,000 x g for 3 min) to precipitate larger soil particulates and eukaryotic cells. The supernatant was carefully collected and serially filtered through 5.0µm cellulose membrane, then 1.2µm and 0.45µm SUPOR® membranes (Paul Corporation) in tandem. The < 0.45 µm fraction was then pelleted by centrifugation (20,000 x g for 20 min) and DNA was isolated from the cell pellet using the Power Soil DNA isolation kit (MoBIO Labs).

3.3.4. Shotgun Sequencing, Binning, and Genome Annotation

The DNAs extracted from the enrichments (“mini-metagenomes”) were pair-end sequenced in an Illumina HiSeq platform in one dedicated run. Raw reads were processed using Trimmomatic (Bolger *et al.*, 2014) to clip adapters, trim low quality ends, and discard short and low-quality reads. The remaining reads from all 42 samples were combined. To aid in the assembly and binning, an additional set of reads from a similar sequencing run of a pilot enrichment for Plot C was added to the read pool (Gupta *et al.*, 2016). The combined reads were assembled with MEGAHIT (Li *et al.*, 2015). The reads from each sample were mapped back against the assembled contigs using BWA (Li & Durbin, 2009). Parallel binning was conducted with MetaBAT2 (Kang *et al.*, 2019), MaxBin 2.0 (Wu *et al.*, 2016) and CONCOCT (Alneberg *et al.*, 2014). Resulting genomic bins were curated using DasTool (Sieber *et al.*, 2018). The completeness and contamination of each bin was accessed using CheckM (Parks *et al.*, 2015) using the standard bacterial marker set and a CPR specific marker set (Brown, *et al.*, 2015). Bins with completeness below 65%, contamination above 10% or strain heterogeneity above 10% were dropped from subsequent analysis. Bins with completeness above 85%, contamination below 5% and strain heterogeneity below 5% were considered high-quality MAGs, while bins with completeness above 75%, contamination below 5% and strain heterogeneity below 5% were considered good-quality

MAGs and the remainder were considered draft MAGs.

3.3.5. MAG taxonomic classification

MAGs were taxonomically classified using two strategies: 16S homology and Phylogenetic Marker Gene voting. First, 16 rRNA sequences were extracted using CheckM (Parks *et al.*, 2015) and a custom package for CPR 16S rRNA extraction and intron removal (Brown, *et al.*, 2015). The MAG's 16S rRNA genes were classified against the SILVA database (Quast *et al.*, 2012) (accessed February, 2020) using SINA (Pruesse *et al.*, 2012). A homology threshold of 75% was used for phylum classification and 80% for class classification. For the second approach, 13 phylogenetic informative markers (**Table 3**) corresponding to ribosomal proteins with similar phylogenetic signal to 16S rRNA (Gregor *et al.*, 2016), were extracted from each bin using CheckM (selected markers are a subset of CheckM's bacterial marker set). The markers were annotated against the nr/nt NCBI database (accessed February 2020). Then each marker “casts a vote” according to their taxonomic annotation. The taxonomy with the higher number of “votes” is assigned to the MAG. Finally, the consensus classification from the two strategies were applied to the MAG.

3.3.6. CPR robust taxonomic classification

Due to CPR hyper diversity and the presence of some MAGs with low 16rRNA thresholds, MAGS identified as CPR were further classified by constructing a concatenated Maximum Likelihood tree. A collection of 1200 CPR genomes was downloaded from NCBI (accessed October 2019). Genome quality was accessed with CheckM as previously described, and good-quality MAGs were used as reference genomes. The set of 13 phylogenetic markers was extracted from the reference genomes as described. Markers from the Dependitiae MAGs were used to root the tree. For each marker, the MAG's sequences and the references were aligned using

MAFFT (Kato & Standley, 2013), trimmed with TrimAl (Capella-Gutierrez *et al.*, 2009) and concatenated using Phylemon 2.0 (Sanchez *et al.*, 2011). The final concatenated alignment was manually inspected with Geneious 10.3 (<https://www.geneious.com>) and used to generate a maximum likelihood tree with 1000 bootstraps using RAxML (Stamatakis, 2014). The final tree was visualized and annotated using iTOL (Letunic & Bork, 2019).

Table 3. Phylogenetic marker genes used in this study.

Phylogenetic Markers	Gene	Protein
PF00181.18	rplB	Ribossomal protein L2
PF00189.15	rpsC	Ribossomal protein S3
PF00203.16	rpsS	Ribossomal protein S19
PF00238.14	rpl14	Ribossomal protein L14
PF00252.13	rplP	Ribossomal protein L16
PF00276.15	rplW	Ribossomal protein L23
PF00281.14	rplE	Ribossomal protein L5
PF00333.15	rpsE	Ribossomal protein S5
PF00366.15	rpdQ	Ribossomal protein S17
PF00573.17	rplD	Ribossomal protein L4/L1
PF00673.16	rplE	Ribossomal protein L5_C
PF03719.10	rpsE	Ribossomal protein S5_C
TIGR01079	rplX	Ribossomal protein L4

3.3.7. CPR MAGs carbohydrate degradation profiles

CPR MAGs and reference genomes from the same taxonomic group were annotated. ORFs were called using Prodigal (Hyatt *et al.*, 2010) and searched against the CazY database (Lombard *et al.*, 2014) using dbCan2 (Zhang *et al.*, 2018). For carbohydrate degradation profiles only glycoside hydrolases (GHs) were considered. Substrate specificity for the recovered GH families were defined by information available on the CAZy database website (<http://www.cazy.org/GlycosylTransferases.html>, accessed 03/21/21) and publications on specific GH families when available (Baldrian *et al.*, 2016; Nguyen *et al.*, 2018; Uversky *et al.*, 2010; Xia

et al., 2016). GH families with undefined substrate specificity were labeled as mixed (Baldrian *et al.*, 2016). Frequency of GHs associated to a substrate degradation was calculated for Cullars MAGs and associated reference genomes, grouped by phylum or class (when possible). GH frequency was calculated by summing the number of GHs recovered across the genomes in each group then dividing by the number of genomes in each group.

Statistical difference in GH frequency profiles between Cullars MAGs and reference genomes were calculated for each phylum using non-parametric, pair-wise Dunn tests with Bonferroni correction in Rstudio.

3.3.8. MAG count estimation

To estimate MAG counts, reads from each sample were individually mapped back against each MAG using Bowtie2 (Langmead & Salzberg, 2012) with high sensitivity. Well mapped reads were counted using SAMTools (Li *et al.*, 2009). The matrix with read counts per sample was imported into Rstudio. Library sizes were normalized using the cumulative scale sum strategy (CSS) implemented in the R package metagenomeSeq (Paulson *et al.*, 2013). Library-size normalized counts were then normalized by genome length by dividing the total read length (calculated as library-size normalized count x 150bp), by the MAG genome size (bp). Coverages below 10% were discarded as inconclusive and assigned to zero. Fractions were rounded up to generate a list of normalized natural counts.

3.3.9. MAG community abundance and distribution across Cullars Plots

Normalized counts were used to calculate MAG relative abundance for each sample, averaged by plot to generate a consensus MAG relative abundance per plot. Absolute abundances were used to create a heatmap of MAG per sample allowing diverse representation of the same information. Rarefaction curves for each sample were calculated using the package Ranacapa

(Kandlikar *et al.*, 2018) For ordination analysis, relative abundances per plot were used to avoid library size issues. Non-metric Multidimensional Scaling (NMDS) and Correlation Analysis (CA) were performed with Bray-Curtis calculated distances using the package PhyloSeq (McMurdie & Holmes, 2013). Differences in community composition across plots were accessed by performing PERMANOVA with 999 permutations and subsequent multivariate homogeneity of groups dispersions (Betadisper) using the package Vegan (Oksanen *et al.*, 2013). The same package was used to perform vector fitting and surface fitting of Cullars soil parameters with significant correlation to community changes ($p < 0.05$). Plots were rendered in ggplot2 (Wickham, 2011) and schematics and illustrations were rendered in Adobe Illustrator (Adobe Inc.).

3.4. Results

3.4.1. Recovering ultra-small bacteria genomes from Cullars soil

Previous studies of Cullars soil metagenome revealed that CPR were present in the soil in extremely low abundances (Santana-Pereira *et al.*, 2020). Therefore, for each of three samples from the 14 plots at Cullars rotation, ultra-small cells were carefully detached from soil particles and enriched through serial filtration to final 0.45 μm fraction. These “mini-metagenomes” were shotgun sequenced. Due to the challenges in recovering these rare soil taxa, an additional set of reads obtained from a dedicated sequencing run of Plot C was spiked alongside the mini-metagenomes to aid in the assembly and genome binning process (Gupta *et al.*, 2016).

From the 122 total bins, a total of 40 genomic bins were curated ranging from 400 Kb to 3.5 Mb (average 1.17Mb Kb), containing twenty-five high quality MAGs (completeness over 80% and contamination below 5%), (**Figure 11**). MAGs with completeness below 65%, contamination over 10% or strain heterogeneity over 10% were not considered acceptable, resulting in a final set of 29 MAGs. These MAGs had an average completeness of 85.60% with 4 bins, later identified as

CPR, with 100% completeness. Average genome size was 1.06Mbp in accordance with the reduced size reported specially for CPR. The largest MAG had 3.55Mbp and completeness of 78.61%, however this “giant” was found to belong to the phylum Gemmatimonadetes which are not a ultra-small bacterium per-se, but rather a filterable one (Portillo *et al.*, 2013a).

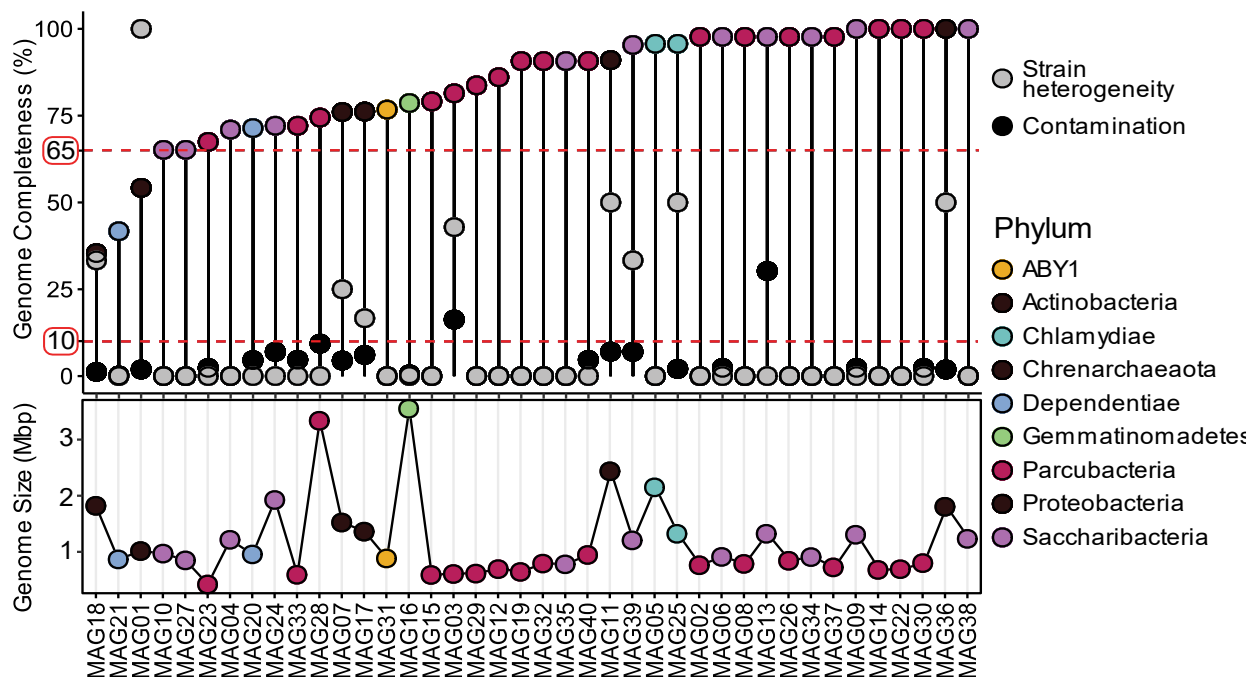


Figure 11. MAGs from the ultra-small bacteria fraction. **a.** MAGs completeness, strain heterogeneity and contamination (%). Red dashed lines represent the thresholds for keeping MAGs for subsequent analysis: **10%:** maximum strain heterogeneity or contamination threshold; **65%:** minimal completeness threshold. **b.** Genome size of Cullars MAGs.

3.4.2. Taxonomical Classification of Cullars MAGs

A collection of 13 phylogenetic marker genes (**Table 3**) and 16S rRNA genes were extracted from Cullars MAGs. The final MAG classification was done by consensus annotation from two approaches: **1.** 16S rRNA annotation against the SILVA database using the web-server ACT; **2.** Marker gene taxonomy voting: for this approach the 13 phylogenetic marker genes were classified separately against the nr/nt database in NCBI. Then each marker “casts a vote” according

to their taxonomic classification. The taxonomy with the highest number of votes is assigned as the MAG taxonomic classification. All curated MAGs were classified, comprising 6 bacterial phyla, from which 90% belong to the CPR superphylum with representatives of the Parcubacteria, Saccharibacteria and ABY1 phyla (**Figure 12**). In addition, one representative MAG was recovered for Chlamydiae, Dependientiae and Gemmatimonadetes phyla.

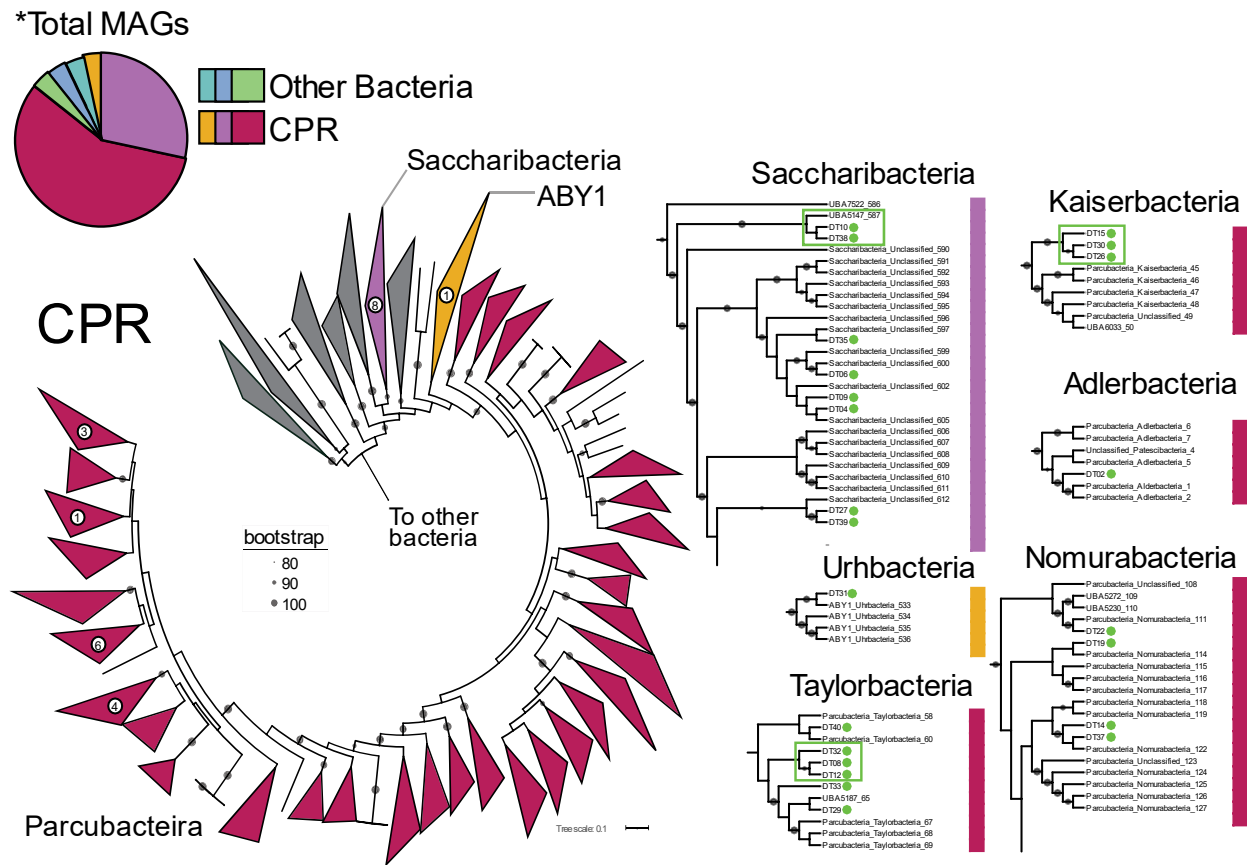


Figure 12. Classification of all Cullars MAGs and the proportion of each phyla. *Total MAGs does not include MAGs filtered by the completeness, strain heterogeneity and contamination thresholds. Concatenated maximum likelihood tree of 13 phylogenetic marker genes from Cullars CPR and 920 reference genomes allowed further classification of CPR MAGs. MAGs were classified as Saccharibacteria, ABY1 and Parcubacteria. The Parcubacteria phylum is expanded to show Classes since it had the highest number of Cullars MAGs and due to its intrinsic diversity. Zoomed in regions offers context to MAGs placement. Green dots denote Cullars MAGs. Boxes indicate clades formed only by Cullars MAGs and unclassified references.

Due to the high degree of diversity and novelty in CPR, the candidate CPR MAGs were

further analyzed by constructing a concatenated maximum likelihood tree using all 13 phylogenetic markers and 920 reference genomes, allowing further classification of CPR MAGs to the Class level when possible (**Figure 12**). Most Cullars CPR MAGs were affiliated with Parcubacteria being further classified as Nomurabacteria, Taylorbacteria, Adlerbacteria and Kaiserbacteria. Both in Taylorbacteria and Kaiserbacteria, Cullars MAGs formed unique clades. One MAG belonging to the ABY1 phyla was further classified into the Uhrbacteria group. Due to the lack of standardized nomenclature the Saccharibacteria MAGs could not be assigned further classification. Interestingly, two MAGs clustered with an unclassified CPR MAG, to form a deep branch related to Saccharibacteria which may represent a new CPR lineage.

3.4.3. Carbohydrate Degradation potential of soil dwelling CPR

Members of the CPR have been implicated in carbon turnover by breaking complex sugars. To access the possible carbon utilization capabilities of the soil dwelling CPR, all CPR MAGs and their corresponding reference genomes were mined for glycoside hydrolases (GHs) and against the CazY database (**Figure 13**).

Cullars MAGs showed similar profiles to that of the references (p values > 0.0610) Saccharibacteria showed indications of enrichment in GHs related to starch degradation when compared to the references, despite not reaching the statistically significant threshold ($p = 0.0610$).

Interestingly, some soil Saccharibacteria also showed an increase in GHs associated with Peptidoglycan, Chitin, Xylan degradation when compared to the reference genomes. In contrast, the Uhrbacteria (Phylum ABY1) MAG lacked GHs sometimes found in the taxa associated with the breakage of mannan, oligosaccharides and peptidoglycan. Members of the phylum Parcubacteria showed no distinct profile pattern when compared to the references, with the exception of the Nomurabacteria MAGs which carried five families GHs with uncharacterized or

undefined substrates not found in any of the reference genomes.

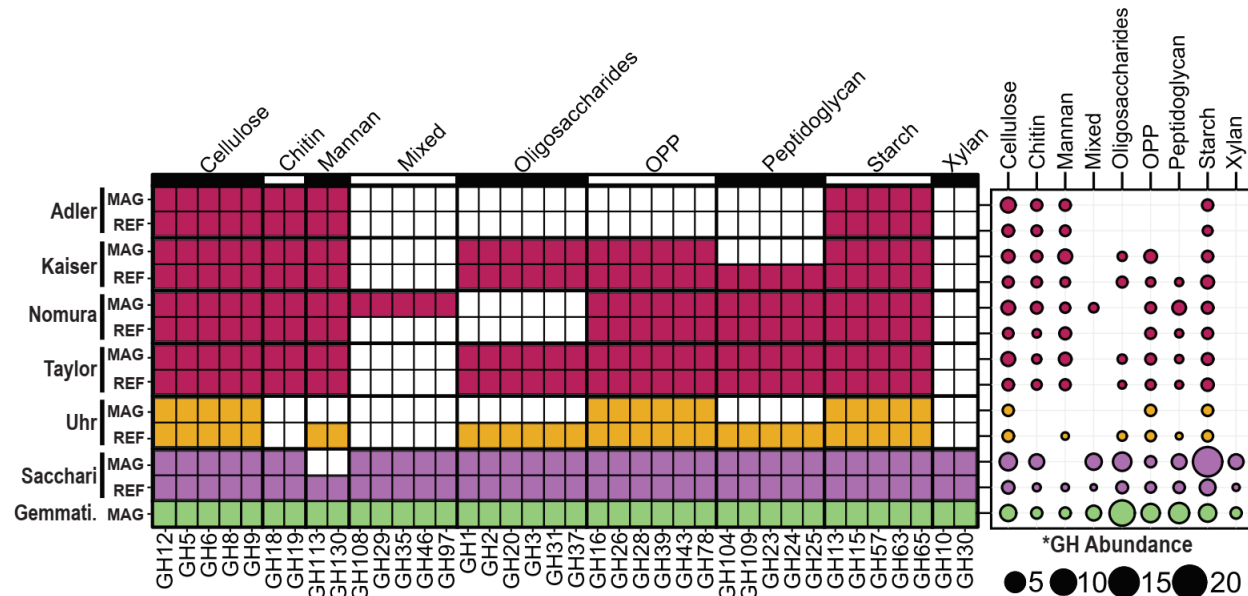


Figure 13. Carbohydrate degradation potential of soil dwelling CPR in Cullars plots (MAG) and of the associated reference genomes (REF), based on glycoside hydrolase (GH) families annotated from the CAZy database. Left panel is a presence/absence map of the annotated GHs across Cullars CPR (MAG) and the reference genomes separated by Phyla, with Parcubacteria (red) further separated into classes. Gemmatimonadetes were also included due to their correlation with Saccharibacteria abundances. **Adler:** Adlerbacteria (Parcubacteria); **Kaiser:** Kaiserbacteria (Parcubacteria); **Nomura:** Nomurabacteria (Parcubacteria); **Taylor:** Taylorbacteria (Parcubacteria); **Uhr:** Uhrbacteria (ABY1), **Sacchari:** Saccharibacteria; **Gemmati:** Gemmatimonadetes. Right panel shows the abundance of GHs according to their substrate specificity. **Mixed:** GH families with undefined specificity. **OPP :** GH families predicted to break other carbohydrates derived from plant biomass. Abundance describes the combined counts of identified GHs divided by the number of MAGs in each group.

3.4.4. Distribution of the ultra-small bacteria in Cullars soil

Cullars Rotation offers a unique opportunity to access the influence of nutrient availability and soil parameters in the ultra-small bacterial community, and thus in the soil dwelling CPR community, due to its century old multiple treatments (**Figure 11**). In order to reconstruct MAG community in each plot, reads from each sample (3 samples/plot) were mapped back against the curated MAGs and normalized to account for variations in library size and genome length. For each sample, MAGs recruiting less than 10% of its coverage in reads were filtered out. The

normalized, filtered counts were used to reconstruct the communities in each of the plots in Cullars Rotation (**Figure 14**). Additionally, soil nutrient content for each plot were also measured (**Figure 10**).

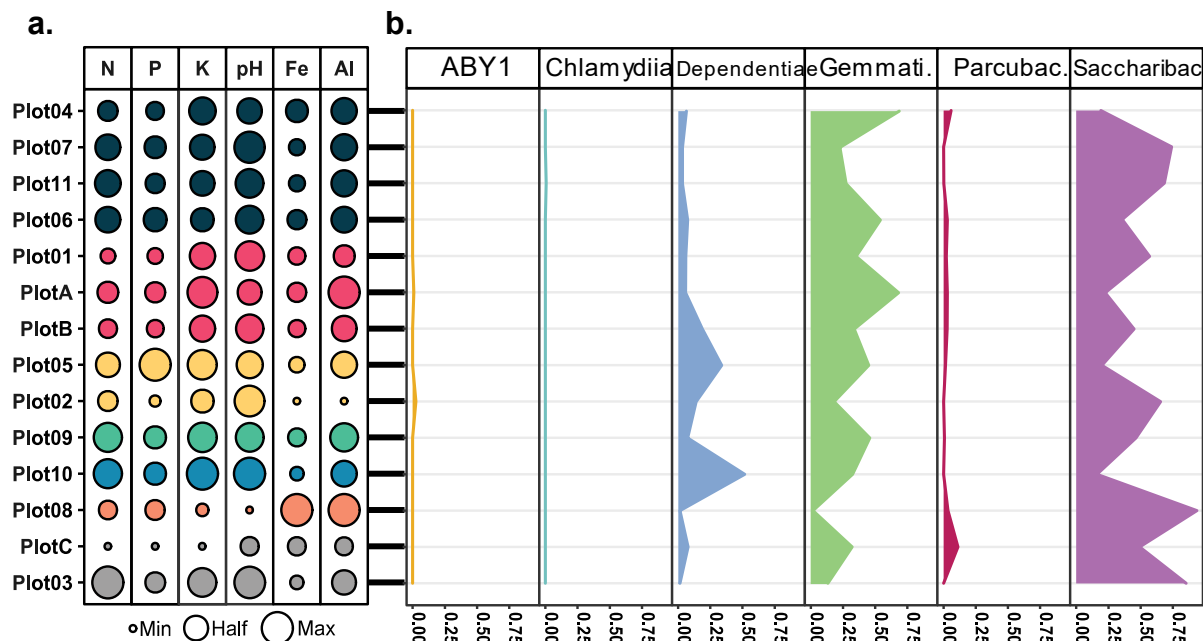


Figure 14. MAG community reconstruction from Cullars soil at different plots. **a.** Soil macro-nutrient levels, pH. **b.** corresponding MAG average relative abundance, ranked by Phyla. Plots are organized by Varying nutrient/treatment: **Plot 03:** Positive control, **Plot C:** Negative control, **Plot 8:** No pH correction, **Plot 10:** Micronutrients, **Plot 09:** S varies; **Plots 02/05:** P varies; **Plots B/A/01:** N varies; **Plots 06/11/07/04:** K varies.

MAG communities are composed predominantly of Saccharibacteria, Gemmatimonadetes and Dependenciae, even though the latter two have only one representative MAG each. (**Figure 14**). Despite having the higher and most diverse number of representatives, Parcubacteria accounts for a significantly smaller proportion of the community, indicating that taxa dominance is still observable in the mini-metagenome community. Indeed, in each plot, at least 70% of the relative abundance can be attributed to the Gemmatimonadetes and Saccharibacteria taxa alone.

The visualization of individual MAG absolute abundance allows the evaluation of intra-

taxa fluctuation and diversity (**Figure 15**). Interestingly, MAG abundance within a particular taxa varies independently from plot to plot, suggesting a reliance on different survival strategies, although more extensive sampling is necessary to resolve sparsity from true absence (**Figure 15**). While in Plot 03 (Standard fertilization) the Saccharibacteria comprises 83% of the community and as much as 93% in Plot 08 (No pH correction), these percentages are represented by a different set of MAGs in each plot. In particular, the plot with no pH correction shows the greatest shift in community structure with 90% of the relative abundance corresponding to MAG35 which is dominant exclusively in this plot, while also had the highest number of recovered Parcubacteria genomes.

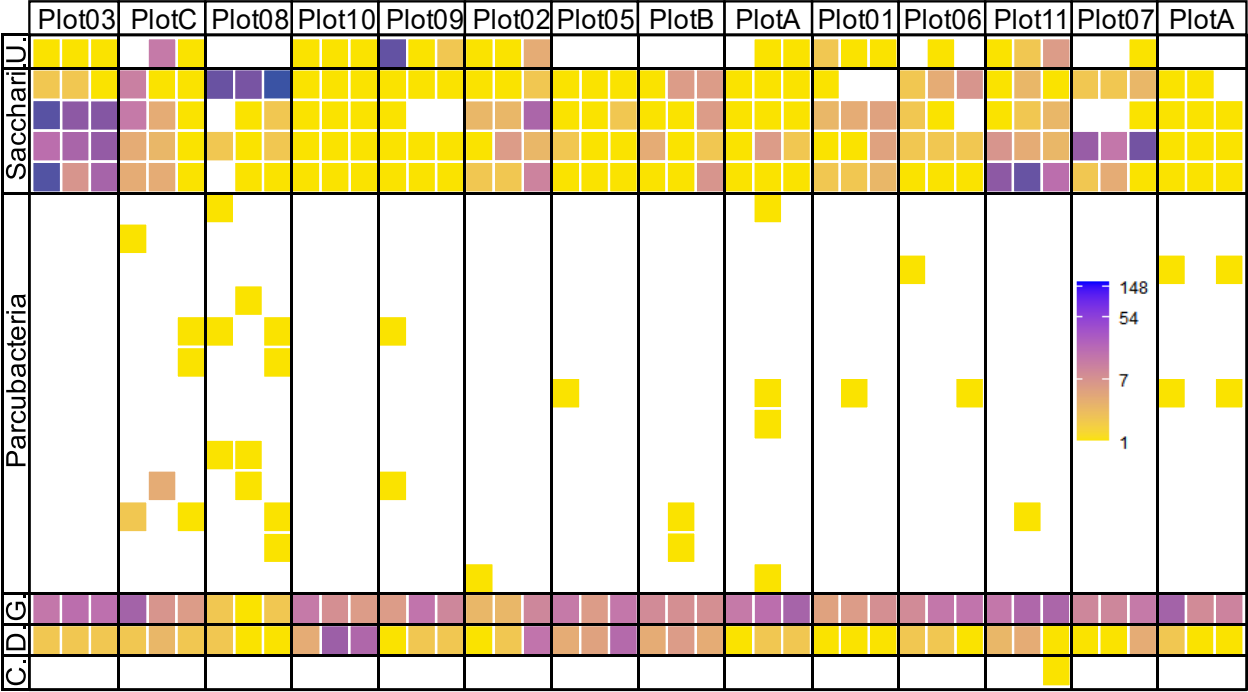
Dependentiae and Gemmatimonadetes were sampled across all plots, with the later peaking under standard fertilization treatments (Plot03) and dropping to marginal abundance when there was no soil pH correction (Plot08) indicating a susceptibility to very acidic soils (DeBruyn *et al.*, 2011). Otherwise, its abundance oscillated without fitting a Plot-by-Plot pattern.

NDMS ordination and CA analysis were used to visualize communities' dissimilarities as a function of Fertilization treatment (**Figure 16**). Community profiles did vary significantly across plots (PERMANOVA, $p = 0.001$ and Betadisper, $p = 0.8537$). While different treatments did affect the community compositions to some extent, Lack of pH correction (Plot 08) had the most dramatic effect in clustering, in both NMDS and CA analysis (**Figure 16**).

The lack of pH correction coupled with the addition of standard fertilization treatment, resulted in the most alkaline pH across all Plots ($\text{pH} = 4.08 \pm 0.08$), surpassing even Plot C which receives no amendment or fertilization ($\text{pH} = 4.58 \pm 0.04$). Although average pH levels seem to be linked to microbiome profile shifts (PERMANOVA, $p = 0.015$), it cannot alone account for the observed differences, since a similar trend was not observed for Plot C, which also has very acidic

soil.

a.



b.

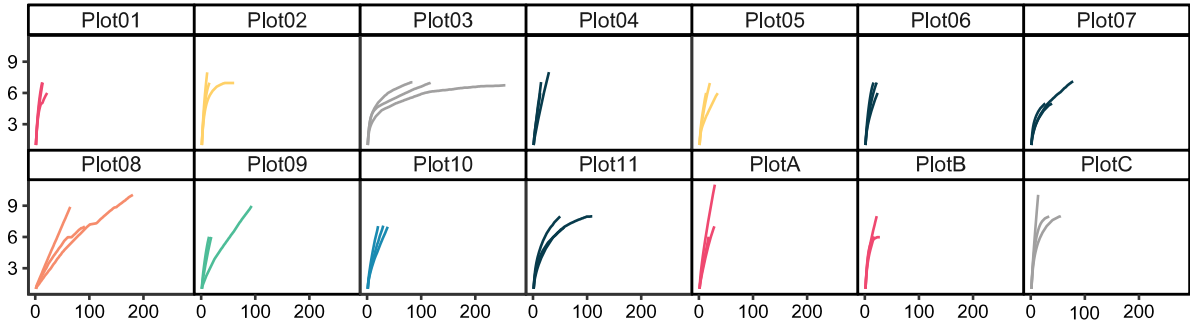
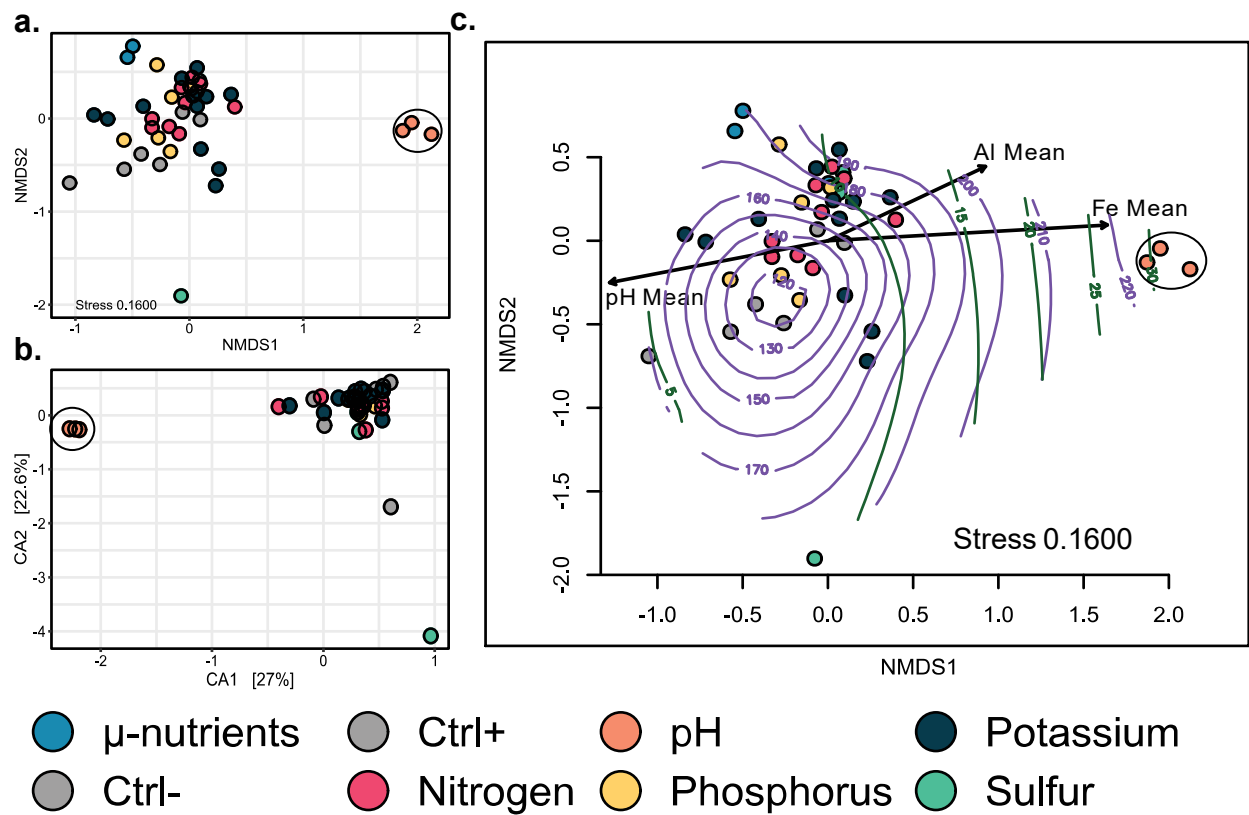


Figure 15. MAG absolute abundance across Cullars Rotation plots. **a.** Heatmap of MAG absolute abundance across Cullars plots for each triplicate, highlighting MAG specific fluctuations and sampling sparsity. **C:** Chlamydiae; **D:** Dependientiae; **G:** Gemmatimonadetes; **Sacchari:** Saccharibacteria; **U:** Uhrbacteria. **b.** Rarefaction curves for each one of the sampled plots from Cullars.

To further explore the role of soil parameters in community compositions, the measurements of soil macro and micro-nutrients were modeled to the community distance matrix. The environmental fit analysis reveals that although pH levels are significantly correlated with data ordination ($p = 0.003$), the environmental factor with the strongest correlation to the changes

in the community is Aluminum concentrations (PERMANOVA $p = 0.001$, Vector fit $p > 0.000$), followed by Iron concentrations (PERMANOVA $p = 0.001$, Vector fit $p = 0.0002$) (**Figure 16**).



3.5. Discussion

Most of the study of ultra-small bacteria has centered around the group CPR (Ghuneim *et al.*, 2018) due to their unusual metabolic potential (Brown, *et al.*, 2015; Castelle *et al.*, 2018) and astounding in phylogenetic diversity (Hug *et al.*, 2016; Parks *et al.*, 2018). However, the majority of what we know about them comes from studies on aquatic environments or associated sediment

where they tend to have higher abundances (Anantharaman *et al.*, 2016; Danczak *et al.*, 2017; Kantor *et al.*, 2013; Luef *et al.*, 2015; Wrighton *et al.*, 2012), limiting our understanding of their distribution and diversity. Recently it was shown that groundwater CPR are mobilized from soils (Herrmann *et al.*, 2019) and genome resolved metagenomics with deep-sequencing reconstructed CPR genomes from Amazon soil (Kroeger *et al.*, 2018) and Antarctic soil (Woodcroft *et al.*, 2018).

This study reveals that these organisms are widespread in Cullars soils and this environment harbor uncharacterized metabolic diversity. We reconstructed the genomes of 26 representatives from three different CPR containing further class diversity. In addition, the mini-metagenomes also contained other ultra-small bacterial groups such as environmental Chlamydia and the parasitic group Dependitiae (TM6) (Deeg *et al.*, 2019), indicating a prevalence of symbionts in ultra-small fraction diversity as naturally these lifestyles tend to lead to genome reduction (McCutcheon & Moran, 2011). Interestingly, the largest recovered genome (3.5Mbp) belongs to Gemmatimonadetes which is a known member of soil microbiota (DeBruyn *et al.*, 2011) and was also sampled in Cullars soil previously (Santana-Pereira *et al.*, 2020). Despite the large genome size, and by implication cell-size, Gemmatimonadetes are filterable, being observed in fractions of 0.45 μ m up to 0.2 μ m pore size (Portillo *et al.*, 2013b).

While some CPR have been empirically shown to live a endosymbiotic life-style (Bor *et al.*, 2018; Gong *et al.*, 2014), others still present a strong reliance on their communities whether or not they form specific symbiotic relationships due to their severe loss of housekeeping metabolic functions (Brown, *et al.*, 2015; Castelle *et al.*, 2018; Herrmann *et al.*, 2019). These assumptions can be extended to Cullars CPR which also have very small genome sizes (< 1.5Mbp) and were previously shown to share the typical losses or lack in nucleotide, amino-acid, lipid and respiratory pathways (Data not shown) (Brown, *et al.*, 2015) Thus, shifts in ultra-small bacteria

distribution is intimately linked to host availability, and shifts in microbial consortia which they depend upon (Anantharaman *et al.*, 2016).

CPR has been predicted to be able to ferment complex substrates (Kantor *et al.*, 2013) and participate in carbon turnover and other biogeochemical cycles (Danczak *et al.*, 2017; Wrighton *et al.*, 2014) which can be translated to soil environments as well, since Saccharibacteria and Parcubacteria were found amongst starch and (hemi)cellulose degraders, respectively (Starr *et al.*, 2018; Taubert *et al.*, 2019; Woodcroft *et al.*, 2018).

Although soil dwelling Parcubacteria and ABY1 were similar to the reference genomes in their GH profiles, soil dwelling Saccharibacteria were enriched in GHs associated to breakage of other complex plant derived carbohydrates, including cellulose, and even more pronounced in the enrichment in starch specific GHs, which suggests an adaptation to fermentation of broad soil biomass (Starr *et al.*, 2018), although it is unclear if these organisms preferentially derive their energy from direct degradation of biomass or from scavenging for processed derivatives (Figueroa-Gonzalez *et al.*, 2020; Taubert *et al.*, 2019). The higher frequency of peptidoglycan degradation GHs in some Saccharibacteria MAGs may be indicative of endosymbiotic lifestyles which have been previously reported within this group (Bor *et al.*, 2020; Lemos *et al.*, 2019). Furthermore, soil dwelling Nomurabacteria (Parcubacteria Phyla) were also enriched in GH families not found in the references which may play specific roles in the soil context. Gemmatimonadetes showed a versatile GH profile with the potential to utilize all listed substrates.

These findings offer insight into the possible roles played by soil dwelling CPR and points towards discrete niche occupation despite their similar metabolic limitations (Nascimento Lemos *et al.*, 2020), due to adaptative acquisition (Jaffe *et al.*, 2020) or maintenance of distinct accessory traits. Furthermore, Cullars MAGs formed unique clades within Kaiserbacteria class and

Taylorbacteria class, including a deep branching clade related to Saccharibacteria which can correspond to a novel, Patescibacteraia group. The unique metabolic profiles combined with phylogenetic analysis indicates soil-dwelling CPR specific fingerprints.

Cullars Rotation offers a unique setting to observe how the ultra-small bacterial community responds to controlled nutrient deficiency in soil. The distribution of MAGs across treatment plots does not follow marked taxa specific trends, suggesting individual diversity in survival strategies (Jaffe *et al.*, 2020). Saccharibacteria is strongly negatively related to Gemmatimonadetes, although their more enriched GH does not overlap, the Gemmatimonates MAG was enriched in GHs related to oligosaccharide utilization which may place these two groups in competition for scavenging metabolic byproducts. Nonetheless, these changes are a complex interplay between plant, soil microbiome patterns of nutrient utilization and symbiont abundances which in turn affects the ultra-small bacterial community. Thus, more extensive exploration of Cullars microbiome is necessary to fully elucidate CPR relations.

Soil without pH correction (Plot 08) has the most distinct microbial community composition and interestingly also has the highest concentration of Al and Fe. These high concentrations in combination with low pH causes metal to solubilize leading to potential toxicity. Indeed, this plot has a dramatic decrease in yield despite receiving fertilization (Mitchell *et al.*, 2005). CPR are no strangers to extraneous environments, including extremes in pH, carbon availability and heavy-metal (Gagen *et al.*, 2020; Hosokawa *et al.*, 2021; Nelson & Stegen, 2015; Rinke *et al.*, 2013). However, the most commonly found groups in these habitats are Microgenomates and Parcubacteria, which might explain the increased diversity of the latter in soil with higher metal concentrations. However, the high abundance of Saccharibacterium MAG35 was surprising and further genome annotation is necessary to evaluate the presence of specific

mechanisms to deal with these environmental conditions. Alternatively, the increase in MAG35 abundance might be related to symbiotic associations, since pH has been reported to shift overall bacterial composition in Cullars soil (Zhao *et al.*, 2015). Thus, further studies focusing on characterizing the associated microbial communities in Cullars soil will offer key insight into the specificities of soil dwelling CPR, greatly expanding our understanding of the most diverse group of ultra-small bacteria.

Although novel taxa were recovered, more extensive sampling could have the potential to improve diversity discovery. Rarefaction curves shows that the majority of replicates were severely under-sampled, which is corroborated by the sparsity with which Parcubacteria MAGs were recovered. Moreover, a large metagenomic library constructed from a soil sample from PlotC, was able to capture eDNA fragments associated with six CPR phyla, albeit in very low abundances (below 0.01%), pointing to the presence of other CPR phyla such as Microgenomatia, Gracilibacteria and Dojkabacteria in Cullars soil (Santana-Pereira *et al.*, 2020). Taxa dominance was also observed in the mini-metagenome where the community was dominated by Gemmatimonadetes and Saccharibacteria, hindering the observation of more rare CPR phyla such as Parcubacteria, ABY1 and possibly others.

Adequate recovery of the rare biosphere is a multi-faceted issue, that usually stems from insufficient DNA isolation coupled with insufficient sequencing depth. The enrichment protocol was efficient at recovering ultra-small taxa DNA, however achieving appropriate yields was challenging. Nevertheless, due to taxa dominance in the mini-metagenome, some CPR were still represented in low abundances and thus more stringent filtering (Luef *et al.*, 2015) or increased sequencing depth could be implemented. Indeed, most reads were not assembled or assembled into short contigs (< 2kb) – common issues in soil shotgun metagenomics (Howe *et al.*, 2014) – not

due to complexity but sparsity. Despite these issues, the enrichment allowed for recovery of the rare taxa using much modest resources than the otherwise terabyte efforts previously reported (Anantharaman *et al.*, 2016; Woodcroft *et al.*, 2018).

Despite the lowering costs of NGS and high-performance computing, genome-resolved metagenomics of complex communities still relies on massive sequencing power and thus, might remain inaccessible to many a researcher. Therefore, enrichments are valuable tools for the exploration of CPR and other rare microorganisms in soil. This study shows that enrichments allow the reconstruction of genomes from a rare bacterial group generally not recovered from soil environments. Study of their genomes indicates that they are differentiated phylogenetically and in their metabolic potential from counterparts sampled from aquatic environments. Furthermore, changes in CPR communities correlate to pH and metal levels, instead of nutrient fluctuations. These findings highlight the importance of expanding the environments of study of the seemingly cosmopolitan CPR.

Chapter 4: Single-walled carbon nanotube antibacterial activity requires cell-wall sensitizing factors

This work was developed in collaboration with Dr. Virginia Davis, Professor of Chemical Engineering, and her research group at Auburn University. All the chemistry work including handling and preparing SWNT dispersions, analyzing SWNT dispersions and associated images were performed by Matthew M. Noor, PhD. This work is to be submitted as the manuscript:

Single-walled carbon nanotube antibacterial activity requires cell-wall sensitizing factors

Matthew M. Noor^{1‡}, Alinne L. R. Santana-Pereira^{2‡}, Mark R. Liles², and Virginia A. Davis^{1*}

¹Department of Chemical Engineering, Auburn University – Auburn University, AL, USA

²Department of Biological Sciences, Auburn University – Auburn University, AL, USA

[‡]Both authors contributed equally to this paper

4.1. Abstract

The intrinsic antibacterial properties of single-walled carbon nanotubes (SWNTs) have long been disputed in the literature. Incongruent results from different sources are often due to differences in SWNT purity, type, or functionalization, or differences in the dispersing agent, cell type, or cell culture medium. In this research, aqueous SWNT dispersions were prepared using agents that have previously been used to study SWNT antibacterial activity as well as tryptic soy broth and DNA. The dispersions were used to assess growth inhibition of methicillin-resistant *Staphylococcus aureus* (MRSA) or *Salmonella enterica*. Our results showed that SWNTs have a synergistic effect with dispersants that cause significant cell stress. However, the SWNT treatments had no effect in *S. enterica*, and there was no loss of MRSA viability when SWNTs were dispersed in DNA or tryptic soy broth (TSB). Sodium dodecyl sulfate (SDS) proved fatal to MRSA regardless of the presence of SWNTs, whereas pluronic and lysozyme rendered MRSA sensitive to SWNTs. Therefore, we propose that the antibacterial activity of SWNTs requires bacterial

sensitization by dispersing agents that impair the integrity of bacterial cell walls. The results of this research provide new insights SWNTs potential use in antimicrobial applications and provides a protocol and considerations for assessing the antimicrobial activity attributed to other nanomaterials.

4.2. Introduction

SWNTs have been studied extensively due to their mechanical, electrical, thermal, and optical properties.(Dresselhaus *et al.*, 1995) Interactions of SWNTs with biomolecules and biological systems is a field of great interest, with researchers studying diverse biological applications such as biosensing(Chen *et al.*, 2003), cell delivery systems(Cai *et al.*, 2005a), tissue engineering,(Kam & Dai, 2005a) and cytotoxicity. Cytotoxicity research concerning SWNTs has been primarily conducted using mammalian cell culture models and the results have been disparate, with nanotubes reported to exhibit toxicity only under certain conditions.(Kang *et al.*, 2008) In addition, significant research has been devoted to understanding whether SWNTs have intrinsic antimicrobial activity.

The need for novel antibacterial materials is becoming more pressing with the rise of drug-resistant bacteria, especially in the healthcare environment. These types of infections cause an estimated 700,000 deaths worldwide per year according to the United Nations; this is predicted to rise to 10 million by 2050.(Willyard, 2017) Antimicrobial research involving SWNTs has yielded conflicting results due to the complex nature of this research, which involves the disparate fields of chemistry, nanomaterial dispersion, and microbiology. For example, SWNT purity, functionalization, chirality distribution, and dispersion state need to be considered(Chen *et al.*, 2006; Dumortier *et al.*, 2006; Wick *et al.*, 2007), alongside the microbiological effects of the cell culture media, (Zhu & Li, 2008) cell type, (Kostarelos *et al.*, 2007) and microbial susceptibility.

A number of authors have reported that SWNTs have antimicrobial activity and attributed their findings to a range of hypotheses including metabolic disruption or inhibition,(Nel *et al.*, 2006) oxidative stresses,(Kang *et al.*, 2008; Manna *et al.*, 2005; Nel *et al.*, 2006) and physical piercing damage to the cell membrane.(Kang *et al.*, 2008; Kang *et al.*, 2007; Liu *et al.*, 2009) Studies of different dispersants,(Dong *et al.*, 2012b) functionalization,(Arias & Yang, 2009) layering,(Nepal *et al.*, 2008) length,(Yang *et al.*, 2010) and diameter(Kang *et al.*, 2008) have been conducted to determine the factors that influencing SWNTs' apparent antibacterial activity. An inherent challenge in these studies has been that dispersing SWNTs in aqueous media requires either surface modification (typically oxidation) or the use of dispersing agents such as surfactants.

For this reason, previous reports of SWNT antibacterial activity have included sensitizing factors in the experimental design, usually the use of surfactants to disperse nanotubes, but also the use of hypotonic media such as DI water to conduct the experiments. The use of surfactants, other SWNT dispersing agents, and/or non-physiologic media that induce physiological stress on bacterial cells when conducting antibacterial assays complicate data interpretation. For example, surfactants have the same effect on bacteria as the one proposed for SWNT, i.e. they disrupt cell walls and cellular membranes, leading to cell lysis. Thus, in many studies it has been difficult to discern if SWNTs are inherently capable of inhibiting bacterial growth in the absence of environmental stressors or sensitizing agents. Interestingly, even studies that have shown antimicrobial activity under some conditions, reported an absence of activity in isotonic solutions such as saline water and buffer solutions.(Arias & Yang, 2009; Kang *et al.*, 2008) These findings along with the discovery that SWNTs can be dispersed in tryptic soy broth (TSB),(Sloan, *et al.*, 2017) a standard biological growth media inspired exploration of potential role of physiological stress on bacteria from the media and/or dispersing agents on the perceived antibacterial activity

of SWNTs.

4.3. Materials and Methods

4.3.1. SWNT dispersion preparations

SWNTs (CG300L35) for this research were purchased from CHASM (Norman, OK, United States) and used as received. Pristine SWNTs, as opposed to -OH or -COOH functionalized SWNT, were used in this research due to two major factors. First, the goal of this research was to probe the intrinsic antibacterial activity of pristine SWNT without the potentially complicating effects of sidewall or end functionalization as additional factors. Second, this enabled comparison to the hypothesis of Liu et al., that a higher proportion and population of individual pristine SWNT exhibits more antibacterial activity, without the use of known antibacterial agents, such as Tween 20 or sodium cholate.(Liu *et al.*, 2009)

Lysozyme from hen egg white, double-stranded DNA sodium salt from salmon testes, and pluronic F108 were purchased from Sigma-Aldrich (St. Louis, MO, United States). Sodium dodecyl sulfate (SDS) was purchased from Fisher Scientific (Hampton, NH, United States). Finally, BD Bacto (Franklin Lakes, NJ, United States) TSB mixture was prepared according to vendor instructions.

The dispersants and SWNT were dispersed according to initial concentrations in the literature (**Table 4**). After sonication, each sample was centrifuged at 17000 x g for 3 hours to remove large SWNT bundles and aggregates. The supernatant was collected and used for the experiments. For normalized concentrations, the SWNT dispersions were diluted with their respective dispersant controls.

4.3.2. Evaluation of SWNT dispersions

A Thermo Scientific (Waltham, MA, United States) NanoDrop 2000c UV-visible

spectrophotometer was used to measure absorbance and subsequently determine concentration by the Beer-Lambert equation, as described in the Results and Discussion section. Samples were diluted with varying ratios of DI water to maintain absorbance values lower than 1 in the range of the observed van Hove peaks for this purpose.

Atomic force microscopy (AFM) was performed with a Pacific Nanotechnology (Berkley, CA, United States) Nano-R SPM. AFM samples were prepared by dropping a single drop of 50 ppm (0.005 vol%) dispersion, diluted by DI water, onto freshly cleaved mica sheets. The sample was allowed to rest on the surface for 10 min, then rinsed with DI water and vacuum dried at 50 °C for at least 12 h. The AFM was operated in non-contact mode at a resolution of 1024 and a scan rate of 0.25 s⁻¹. The diameter of over 100 entities were measured for each sample with some samples requiring multiple scans. Entities with measured heights greater than 3nm were considered to be bundled SWNTs, lower than 3 nm was considered to be an individual SWNT.

Table 4. Initial concentrations and sonication protocols from each dispersion based on literature.

Dispersants	Dispersant concentration (WT %)	Swnt concentration (WT %)	Sonication procedure		
			Time (min)	Amplitude (%)	Pulse (5 s on, 2 s off)
SDS(Bai <i>et al.</i> , 2011; Dong <i>et al.</i> , 2012b)	1	0.2	60	60	no
Pluronic(Chowdhury <i>et al.</i> , 2012; Moore <i>et al.</i> , 2003)	2	0.2	60	60	no
TSB(Sloan, A. W. <i>et al.</i> , 2017)	N/A	0.1	30	60	yes
Lysozyme(Horn <i>et al.</i> , 2012)	0.5	0.1	30	60	yes
DNA(Ao <i>et al.</i> , 2011; Nepal <i>et al.</i> , 2008)	0.75	0.1	30	50	no

4.3.3. Bacterial Strains and cultivation

Bacterial strains of medical relevance were used in this study. For a representative Gram-

positive bacterium, a methicillin-resistant *Staphylococcus aureus* (MRSA) strain Xen 29 was selected, and for a Gram-negative bacterium *Salmonella enterica subsp. enterica serovar Enterica* (ATCC® 14028™) was selected. Each bacterial species was grown from a glycerol stock onto Tryptic Soy Agar (TSA) medium to obtain isolated colonies and incubated at 37 °C overnight. An isolated colony from each culture was used to inoculate 2 mL of Tryptic Soy Broth (TSB), and the culture tubes were incubated at 37 °C with shaking at 200 rpm overnight.

4.3.4. Bacterial Culture challenges

The overnight broth cultures were normalized to an optical density at 600 nm (OD₆₀₀) of 0.5 by dilution with TSB. Then 360 µL of the normalized bacterial suspensions were inoculated with 40 µL of a specific treatment group containing either sonicated dispersants without SWNT (pluronic, SDS, lysozyme, DNA or TSB), SWNT dispersions (in pluronic, SDS, lysozyme, DNA or TSB), or SWNT dispersions with their concentration normalized (**Figure 17**). The antibiotic vancomycin was used as a positive control to inhibit MRSA growth and chloramphenicol was used as a positive control to inhibit *S. enterica* growth, while sterile MilliQ water was used as a negative control, resulting in a total of 17 treatments per bacterial strain. Each of the bacterial treatment group suspensions were incubated for 1 h at 37 °C with shaking at 200 rpm.

4.3.5. Bacterial viability measurements

An aliquot of each treatment group sample was then inoculated at a 10⁻¹ dilution into fresh TSB and transferred to a clear 96-well plate in tetraplicate, with 200 µL per well. The plate containing all of the treatment groups was incubated for 24 h with shaking at 37°C in a microtitre plate reader and the OD₆₀₀ was measured every 30 min to determine growth curves. Mean and standard deviation of the replicates were used to build the reported growth curves.

Another aliquot of each sample was serially diluted and plated onto TSA plates to

determine the number of colony forming units (CFU) per mL, also in triplicate. The plates were incubated for 24 h, and the dilutions yielding roughly 30-300 colonies were picked for CFU determination. Reported inhibition was calculated as the difference between mean CFU count between challenges and negative control treatment (water).

A second run of the experiment was performed for the TSB and DNA treatments after the inoculums failed in the first 24 h incubation attempt due to excessive dehydration at the border of the 96-well plate.

4.3.6. Statistical analysis

For the OD measurements, 4 points were chosen for statistical difference calculations: 3h, 6h, 9h and 12h; given that all inoculums reached stationary phase after 12h. For each dispersant, first an ANOVA test were conducted to evaluate if there were differences between the means of the treatments: dispersant, dispersant + SWNTs with corrected concentration, dispersant + SWNTs full strength and water. Following a positive ANOVA, Tukey tests were performed to identify the treatments significantly different from each other. Inhibition is considered as a significant difference between a treatment and the water control.

In the case of CFU counts, non-parametric methods were preferred since the data was not normally distributed. For each dispersant, first the Kruskal Wallis test (Analogous to ANOVA) were conducted to evaluate if there were differences between the means of the treatments: dispersant, dispersant + SWNTs with corrected concentration, dispersant + SWNTs full strength and water. Following a positive Kruskal-Wallis, Dunn tests (analogous to Tukey's test) were performed, in the same manner as described above. Again, inhibition is considered as a significant difference between a treatment and the water control. All analysis and plots were performed and drawn in R-studio, respectively. Charts were drawn in Adobe Illustrator.

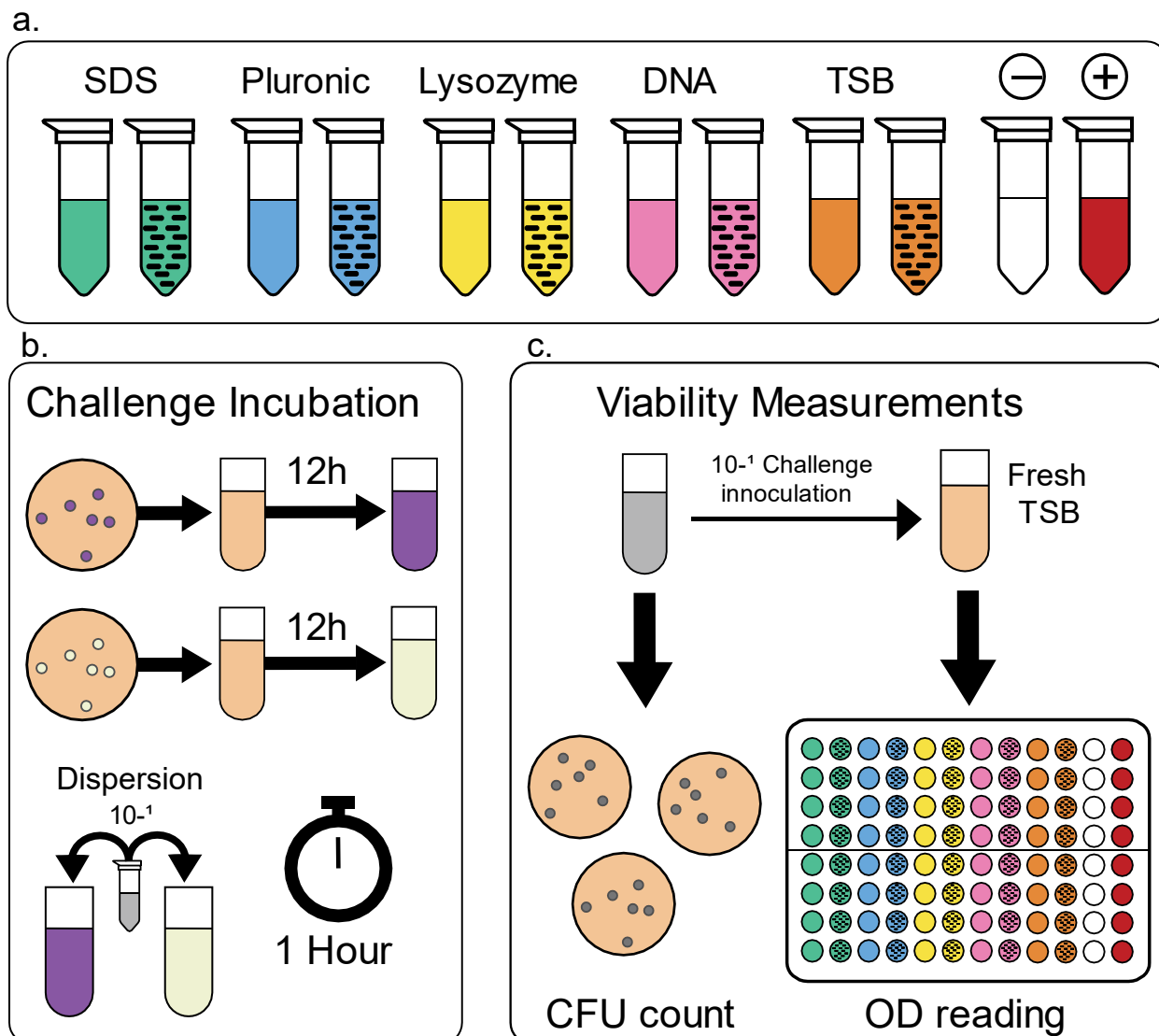


Figure 17. Experimental design of the SWNT antimicrobial challenges. a. Challenging solutions used in the experiment. Cultures were challenged against the dispersant and the dispersant with SWNTs. b. Challenge incubation. Overnight growth cultures of *S. enterica* and MRSA were diluted in fresh TSB to OD600 = 0.5 and incubated with the challenging solution for one hour. c. The challenge incubations were diluted in TSB and plated for CFU counting and grown in broth for growth curve construction.

4.4. Results

4.4.1. SWNT dispersions

In order to evaluate the degrees of SWNT dispersion achieved by each of the chosen dispersants: pluronic, SDS, DNA, lysozyme and TSB; atomic force microscopy (AFM) was used

to quantify the number of individual SWNT versus bundled SWNTs (**Figure 18**). UV-vis spectroscopy was performed to determine the SWNT concentration in each supernatant (**Figure 19**).

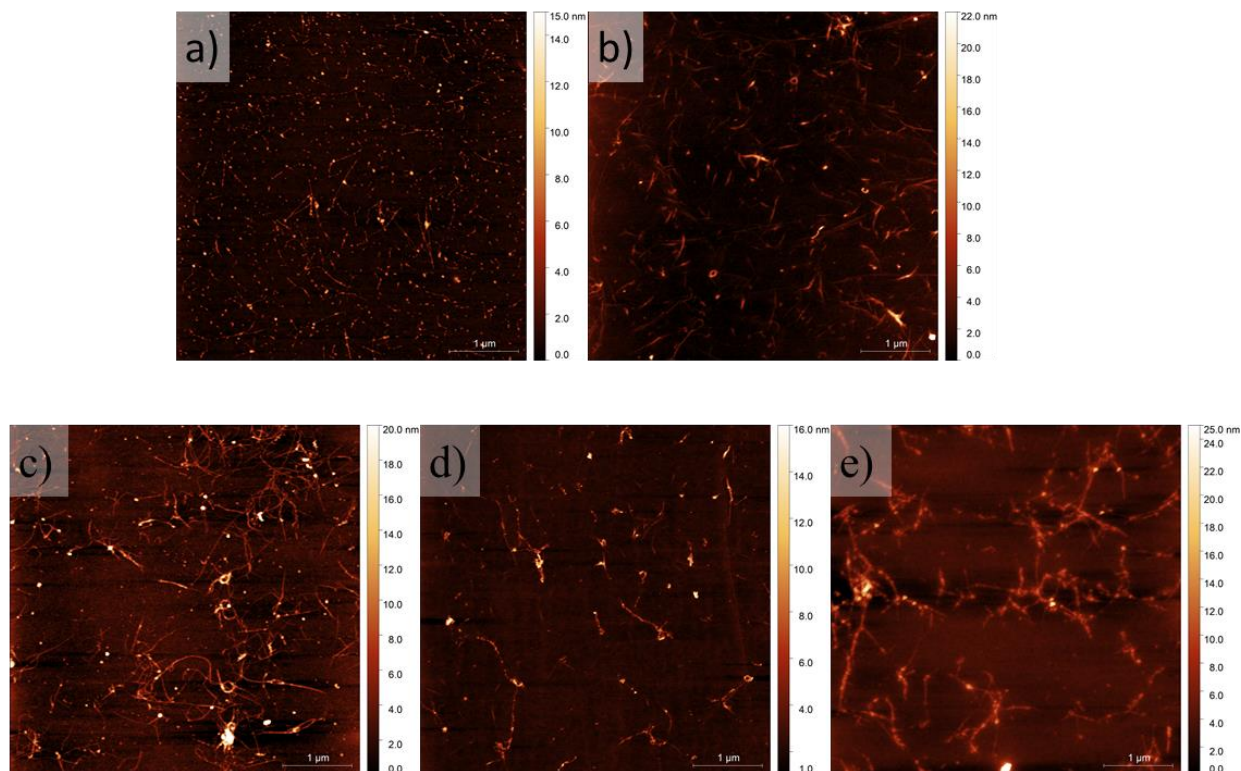


Figure 18. AFM scans quantifying the number of individual versus bundled SWNTs. a) TSB-SWNT b) DNA-SWNT c) LSZ-SWNT, d) Pluronic-SWNT, and e) SDS-SWNT.

*Images generated by Matthew Noor – PhD, in collaboration with Dr. Virginia Davis group.

The SWNT concentrations were determined using the Beer-Lambert law $A = \epsilon lc$ where ϵ is the extinction coefficient, A is the absorbance at a specified wavelength, l is the path length, and c is the concentration of the dispersion. The optical density at 660 nm (OD_{660}) was determined for each sample because this was the location of a strong van Hove singularity peak, an indicator of individual SWNT dispersion. Each dispersion was diluted with ultrapure water and the OD_{660} was recorded. The extinction coefficient was obtained by taking the OD_{660} slope of dispersions serially

diluted with DI water plotted against SWNT concentration.

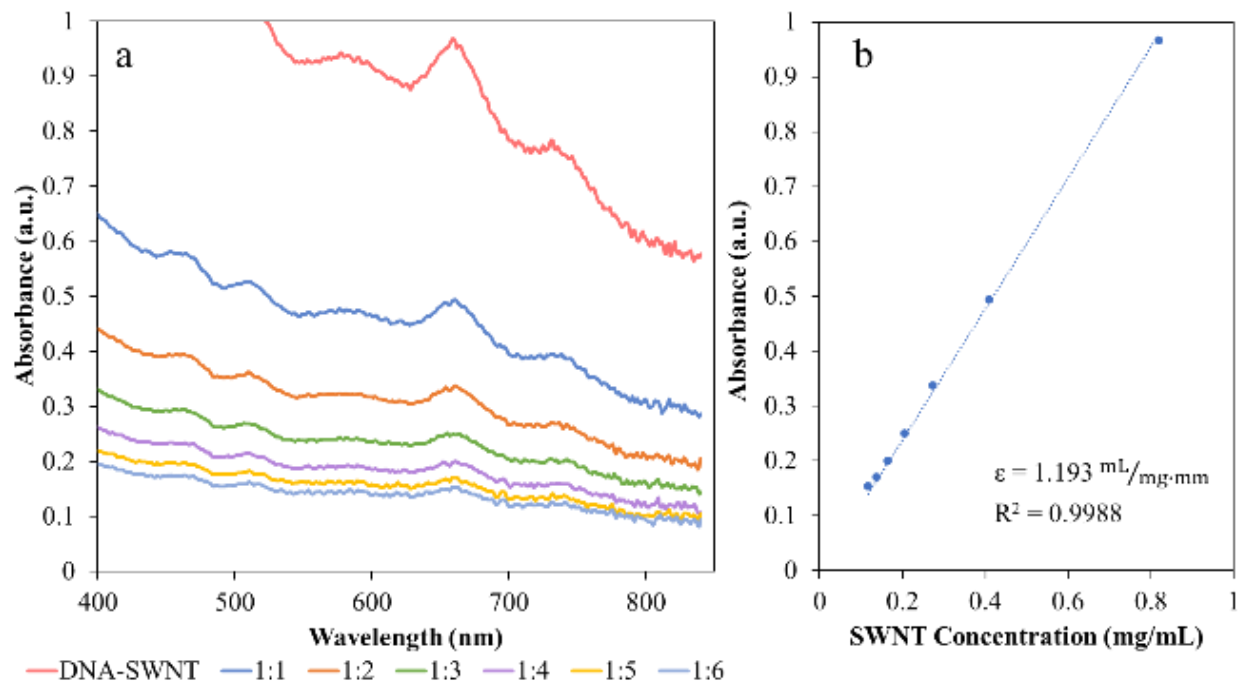


Figure 19. SWNT concentration analysis using UV-vis spectroscopy. (a) UV-vis spectra of DNA-SWNT dispersions with dispersion:water dilutions and (b) Beer-Lambert plot for DNA-SWNT with extinction coefficient.

*Images generated by Matthew Noor – PhD, in collaboration with Dr. Virginia Davis group.

In order to differentiate dispersant and SWNT activity, the SWNT dispersions were also diluted so that SWNT concentrations were the same in each treatment group (**Table 5**).

Table 5. SWNT concentrations, individual to bundled SWNT ratio, and Average length of individual SWNT for each dispersion.

Sample	Dispersion SWNT Concentration (mg/mL)	Individual to Bundle Ratio	Average Length of Individuals (nm)
SWNT-LSZ	0.227	3.2	190 ± 49
SWNT-DNA	0.723	5.7	222 ± 52
SWNT-Pluronic	0.577	6.7	172 ± 47
SWNT-SDS	1.670	0.4	171 ± 35
SWNT-TSB	0.445	9.0	234 ± 62

Because the lysozyme-SWNT dispersions had the lowest concentration, all dispersions were diluted with their respective controls to match that concentration (**Table 5**). The SDS-SWNT dispersion had a much higher concentration but this is compensated for by a much lower individual to bundle ratio.

4.4.2. SWNT effects on Bacterial Viability

The growth of *S. aureus* and *S. enterica*, after a one hour incubation with different SWNT-dispersant combinations (**Figure 17**), was correlated with bacterial cell wall morphology and dispersant type (**Figure 20**). Determining CFU counts allowed the assessment of viable bacterial cells while avoiding SWNT interference in fluorescence- or absorbance-based bacterial viability assays.

Gram-positive bacteria like *S. aureus* have a thick cell wall comprised of many layers of peptidoglycan which gives the cell structural support, is the target of lysozyme and is directly affected by surfactants. In contrast, Gram-negative bacteria like *S. enterica* (which is affiliated with the Gamma-Proteobacteria such as *Escherichia coli*) have an inner and outer membrane between which is the periplasm that includes a thinner peptidoglycan layer. Because of their double membrane structure, Gram-negative bacteria are less vulnerable to lysozyme or surfactants.

As expected, *S. enterica* was not affected by any of the dispersant and/or SWNT treatments, with no decrease in CFU count detected. In contrast, the Gram-positive bacterium MRSA was observed to be inhibited in its growth in a dispersant-dependent way (**Figure 21**). MRSA was very susceptible to the surfactant SDS regardless of the presence of SWNT ($p = 0.023$). Interestingly, the mild and non-cytotoxic surfactant pluronic had no effect on MRSA viability except when coupled with high SWNT concentrations ($p = 0.201$), and the same was also observed for lysozyme ($p = 0.083$), although not at statistically significant thresholds. (**Figure 20**).

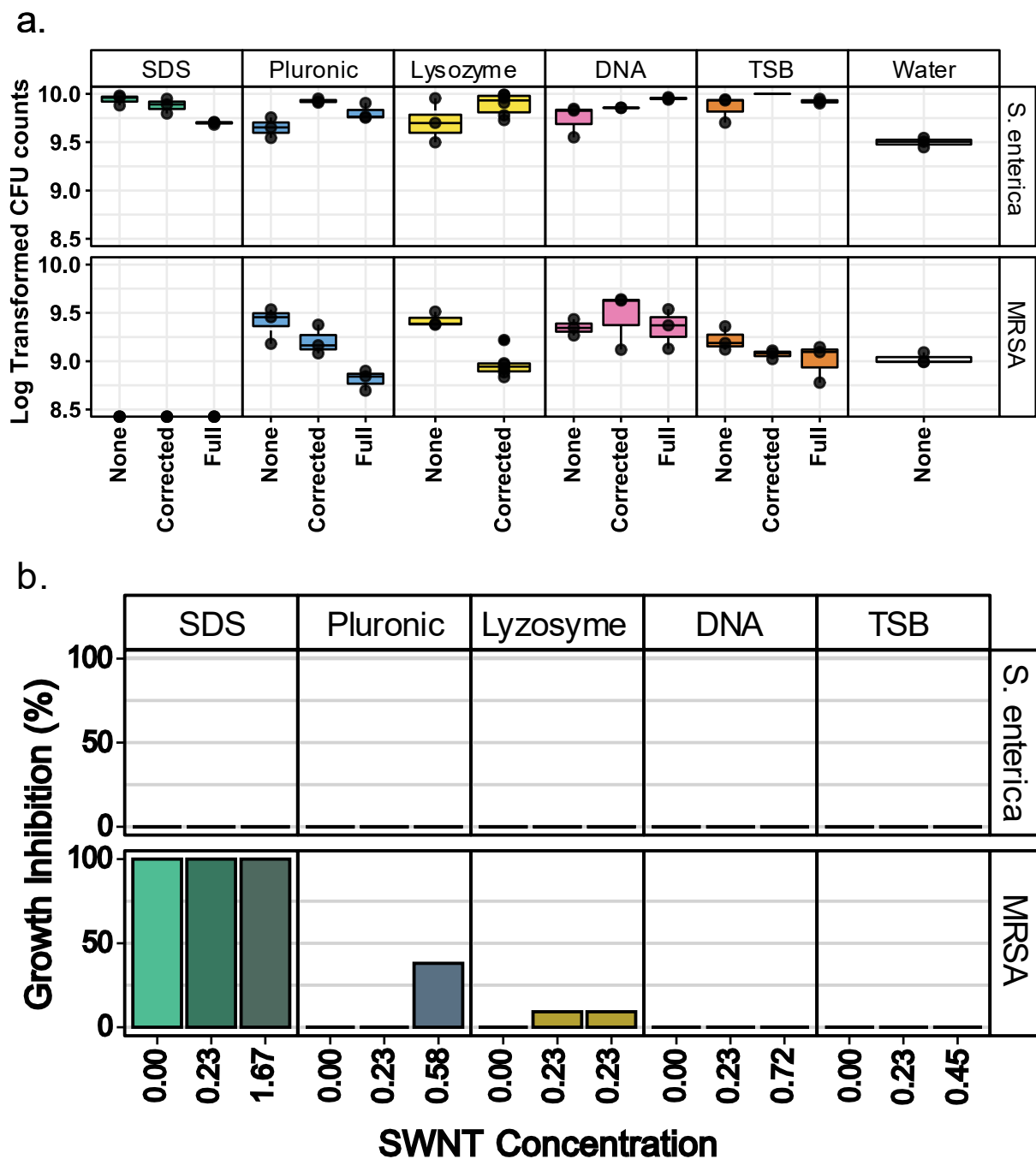


Figure 20. Cell viability measured by CFU counts. a. CFU counts in triplicate of the challenges for MRSA and *S. enterica*. Log 10 transform was used to coerce the data into a more normal distribution and facilitate visualization. b. Antimicrobial activity of different dispersants dependent on SWNT concentration. SWNT concentrations were normalized to the lowest concentration (lysozyme) for comparison. Antimicrobial activity measured as cell viability loss in relation to a water challenge.

The dispersants that singly or in combination with SWNTS resulted in an observed inhibition of MRSA growth: SDS, lysozyme, and pluronic; are all known to act directly on the bacterial cell wall, destabilizing the peptidoglycan structure. Importantly, MRSA was unaffected by SWNTs, regardless of their concentration, when they were dispersed in TSB or DNA, both of which are biologically compatible dispersants with no negative effect on bacterial viability. To appropriately compare the effect of SWNT concentration on bacterial viability, additional bacterial incubations were conducted with SWNT concentrations normalized to the lowest concentration amongst the dispersions. The normalized dispersions showed no interference in cell viability except for SDS and lysozyme, although the SDS antibacterial activity was independent of SWNT presence.

In addition to measuring bacterial growth inhibition after a 1 h exposure to SWNT-dispersant combinations, the effect of these treatments on bacterial growth in fresh TSB media for 24 h was also assessed by monitoring OD₆₀₀ (**Figure 21**). Optical density readings can be valuable indicators of antibiotic activity when growth is tracked in the presence of the bioactive compound. However, this approach was prohibitive given the high absorbance of light by SWNT especially at 600 nm which is the absorption range of bacterial cells. Consequently, bacterial suspensions were incubated with a 10-fold decrease in SWNT concentration after the challenge to control for absorbance interference.

Albeit indirect, determining bacterial growth using OD₆₀₀ indicates the growth curve relative to controls and any impact of treatments on the rate of growth. The *S. enterica* growth curve was highly similar to the negative control (water added to the TSB growth medium) control and no significant differences were observed in different treatment groups ($p > 0.05$).

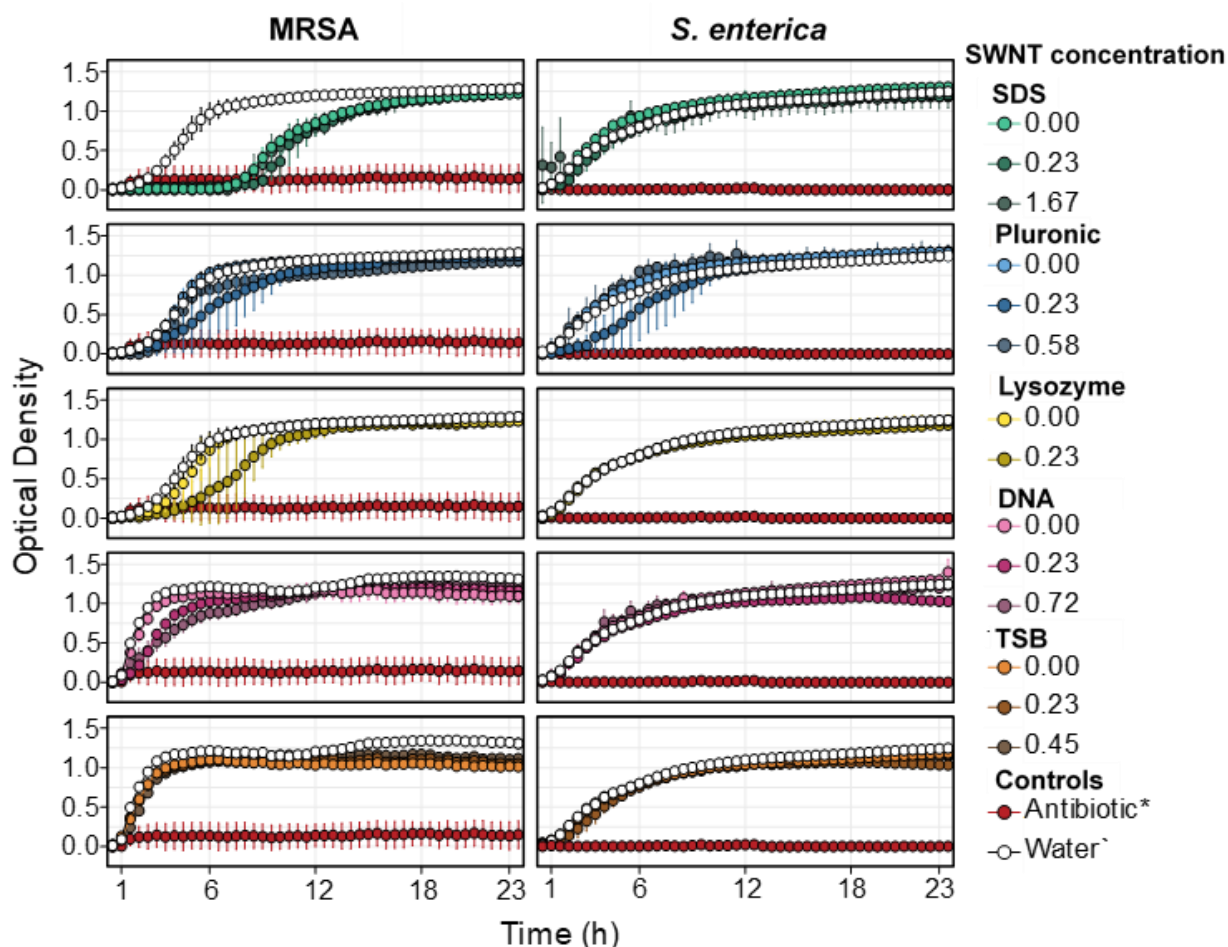


Figure 21. Optical density growth curves of inoculums of *S.aureus* (MRSA) and *S. enterica* challenged with the 17 treatments into fresh TSB medium. Delayed exponential growths points towards a larger number of sick or unviable cells in the inoculum. SWNT concentrations were normalized to the lowest concentration (lysozyme) for comparison.

Gram-positive bacteria like *S. aureus* seem to be more susceptible to SWNT interference and corroborate the observation that SDS is highly toxic to *S. aureus*. The growth of *S. aureus* was reduced by incubation with pluronic or lysozyme but was enhanced in the presence of SWNTS when TSB or DNA was used as a dispersant. MRSA growth was significantly reduced relative to the negative control after treatment with pluronic at 3h of growth ($p = 1.8 \times 10^{-6}$) and lysozyme ($p = 1.7 \times 10^{-6}$) alone, however there was no significant difference in growth after 6 h ($p > 0.05$) indicating that the cells were able to overcome the initial stress imposed by these treatments. As observed in the previous experiment, MRSA treated with SDS was observed to have a heavily

reduced growth even at 12 h incubation ($p < 0.007$), taking approximately 8 hours for logarithmic growth to be initiated post-SDS exposure.

4.5 Discussion

When evaluating antibacterial activity of any compound it is important to avoid sensitizing agents or stressful environmental conditions in order to provide a more accurate assessment of the antibacterial activity for the compound of interest. While much of the previous literature has used sensitizing agents in an effort to achieve dispersion of the SWNTs, this research used dispersions that were sensitizing, non-sensitizing, and known to support growth. SDS and lysozyme both have antibacterial activity against Gram-positive bacteria, and are used extensively in the literature as SWNT dispersing agents (Bai *et al.*, 2011; Dong *et al.*, 2012b; Horn *et al.*, 2012; Moore *et al.*, 2003; Nyankima *et al.*, 2013). Pluronic, a mild triblock copolymeric surfactant with low cytotoxicity, and double-stranded DNA, with neutral activity are excellent SWNT dispersing agents (Ao *et al.*, 2011; Chowdhury *et al.*, 2012; Nepal *et al.*, 2008). More recently, TSB containing casein is routinely used for bacterial cultivation was recently discovered to be an excellent dispersant for SWNTs (Sloan, *et al.*, 2017).

The results of this research suggest that SWNTs have a complex effect on cell viability. First, they seem to not be able to trigger cell death alone. SWNTs have been shown to trigger stress responses in bacterial cells; (Kang *et al.*, 2008) however, they appear to not be enough to kill cells, hence the lack of antibacterial activity when dispersed in DNA or TSB. Other groups reported cell death when performing experiments dispersing oxidized SWNT in innocuous dispersants like water, (Arias & Yang, 2009; Kang *et al.*, 2008; Yang *et al.*, 2010) but their experiments were conducted with bacterial cells suspended in distilled water. The use of a hypotonic media sensitizes the cells to cell-wall disruption by significantly increasing osmotic stress. Indeed, highly hypotonic

solutions can themselves cause cells to lyse.

No cell death was previously observed when conducting antibacterial assays in isotonic media such as phosphate-buffered saline (PBS) or Brain Heart Infusion (BHI) Broth (Arias & Yang, 2009). Resistance to surfactants, which also act on the cell-wall, were observed among Gram-negative bacteria (Ishikawa, S *et al.*, 2002), but it was found to require nutrient uptake (Ishikawa, S *et al.*, 2002; Nakata *et al.*, 2010). Consequently, using DI water as a challenge medium might exacerbate loss of viability by denying living cells the nutrients to withstand stress. It should be noted that SWNTs have very low dispersibility in water, BHI or PBS without chemical modification so more suitable dispersants, like TSB, should be chosen for additional investigations of pristine SWNTs potential interactions with bacteria.

We propose that SWNT antibacterial activity requires a synergistic interaction between the nanotubes and sensitized cells. Published observations of SWNT antibacterial activity were observed using Tween-20 (Liu *et al.*, 2009), a surfactant routinely used in cell lysis protocols, SDS (Dong *et al.*, 2012b) or other known antimicrobial agents like sodium cholate (Liu *et al.*, 2009). Liu *et al.* showed using AFM revealed that “softer” cells are more vulnerable to SWNT piercing (Liu *et al.*, 2009). In our study, loss of cell viability was only observed when the SWNT were dispersed in agents that disrupt cell walls such as surfactants or lysozyme. This effect was especially evident with SWNTs dispersed in SDS, which was enough to completely kill MRSA on its own, but it was also observed with the use of pluronic. Despite being described as non-cytotoxic, pluronic promoted cell death in conjunction with a high concentration of SWNTs. Our findings are in accordance with the previous literature where functionalized SWNT dispersed in Brain-Heart Infusion (BHI) media and PBS had no observable antibacterial activity.

Notably, in this study, the effect of SWNT was limited to Gram-positive bacteria that have

less resilience to cell-wall disruption due to SDS, pluronic, or lysozyme. MRSA had loss of viability only when challenged by SWNTs dispersed in sensitizing dispersants. In contrast, *S. enterica*, a Gram-negative bacterium, was able to withstand the challenges with no loss in viability, in accordance with other studies that identified no loss of viability in *E. coli* (Kang *et al.*, 2008; Kang *et al.*, 2007) and increased viability in *Salmonella enterica* when compared to *Bacillus subtilis* (Arias & Yang, 2009). Thus, neither of the bacteria in this study was susceptible to SWNTs when the dispersions were conducted in biologically compatible media.

Our results indicate that antimicrobial activity is not an intrinsic property of SWNTs, but instead a synergistic interaction between the nanotubes and sensitizing factors which exert significant cell stress such as dispersants. Therefore, SWNTs alone are not able to exert enough stress to kill healthy microbial cells under non-stressful environmental conditions. These synergistic effects are more pronounced in Gram-positive bacteria which are more vulnerable to cell-wall stresses and disruption. To accurately assess the intricate effects of SWNTs on microbial cells, close care should be taken to avoid bias due to sensitizing factors in the experimental design as well as to include appropriate, biologically compatible dispersing agents such as TSB or DNA.

It is likely that our observations can be generalized to other Gram-positive and Gram-negative bacteria, but further research is needed to determine how widely these results may be extended to diverse microbial species. Such studies will similarly need to take into consideration the bacterial type and environment, the SWNT type and dispersion state, as well as synergistic interactions with cytotoxic and growth supporting compounds. It is noted that the question of antibacterial activity is applicable to many materials besides SWNTs and it is hoped that these results will highlight the need for considering physiological stress in the study of other nanomaterials including graphenes and MXenes.

Chapter 5: Conclusions

The microbial world is home to incredible taxonomic and metabolic diversity, although 99% of it cannot, yet, be cultured. Metagenomics has proved to be a robust culture-independent platform which is pervasive to improvements, innovations, and adaptable to different research goals. The research here in performed applied different metagenomic approaches in a complementary manner to answer distinct research questions. The appropriation of tools, pipelines and concepts across metagenomic disciplines allowed the combination of approaches in a manner that expanded their power and provided deeper insight into the research questions.

This combination of approaches was the forefront of the first research question which aimed at improving natural product discovery through functional metagenomics by introducing NGS and bioinformatic tools associated with shotgun metagenomics. Natural product discovery relies heavily on screening metagenomic libraries using PCR-based approaches, which are expected to carry the same biases and limitations encountered in 16S rRNA surveys: its tendency to favor highly homologous templates, overlooking divergent genes/pathways and thus, creating blind spots where the highest potential for novelty lies. On the other hand, direct sequencing with NGS can harness the power of bioinformatics to overcome PCR limitations, however its application for screening of metagenomic libraries has been precluded by the lack of a strategy that allows parallel sequencing of all clones.

Here, a pipeline was developed for the application of NGS to large scale sequencing of metagenomic libraries. A metagenomic library comprising tens of thousands of clones of large inserts to maximize modular BGC recovery was constructed from a soil sample from the Cullars Rotation field. The use of a multiplexed sequencing strategy allowed the entire library to be

sequenced by creating a three-dimensional pooling coordinate system. Then, an algorithm was developed to deconvolute the coordinates for each sequence allowing their assignment to their clone of origin. This innovation circumvents the need for individual clone barcoding or tagging, which constrains the application of NGS to only a few hundreds of clones at a time, precluding its use in screening.

Hence, metagenomic library sequencing open the doors for the implementation of bioinformatics to library screening. Instead of relying on PCR amplification to identify clones carrying interesting pathways, these same pathways can be identified and fully annotated *in silico*. Indeed, the NGS approach was able to recover a higher number of and more diverse BGCs and 16S rRNA genes from the library when compared to traditional PCR, due to the application of bioinformatic tools. Therefore, the hypothesis that the proposed approach would be able to overcome biases associated with PCR screening is strongly supported. Furthermore, the NGS approach allows the parallel screening of several different pathways, while PCR approaches are limited to one target.

Another constraint associated with functional metagenomics is the lack of substantial information on the genetic contents moved to heterologous expression, despite the employment of prior PCR screening. This issue is related to high re-isolation rates and difficulties in heterologous expression. Thus, clone annotation with the NGS approach prior to heterologous expression enables the filtering of already characterized pathways, while providing information to guide the selection/engineering of more appropriate hosts and vectors, increasing the chances of successful expression. In fact, the NGS approach allows the complete annotation of all deconvoluted clones and the screening of virtually any genetic feature of interest, consequently the concept of a metagenomic library annotation is an achievable feat and a significant contribution to the field of

functional metagenomics.

The reconstruction of the bacterial diversity captured in the soil metagenomic library using the NGS approach indicated the presence of bacteria affiliated with the CPR in extremely low abundances. At that point, they had not been described in soils. Moreover, the low homology between the library markers and the database suggested that the library captured potential taxonomic novelty. Due to being a newly described group, which remains severely understudied, we hypothesized that Cullars rotation soil carried uncharacterized members of CPR.

The controlled nutrient deficiency experiment being carried out in Cullars Rotation provides a unique opportunity to evaluate the impacts of nutrient and pH levels on the microbiome. Thus, empowered by the information contained in the metagenomic library, samples from all plots of Cullars Rotation were taken for the implementation of genome-resolved metagenomics. The expected low abundance of CPR would preclude CPR genome reconstruction from direct sequencing of soil DNA. An enrichment protocol was developed to enrich ultra-small bacteria from soil samples prior to sequencing, since this fraction is predicted to be dominated by CPR.

This strategy allowed the reconstruction of good quality CPR genomes which are the foundation of all the analysis performed downstream. Phylogenetic analysis indicated that soil CPR forms distinct clades, including a deep-branching clade which potentially represents a novel candidate phyla within the CPR. Cultivated soils are rich in plant derived biomass which is composed of complex carbohydrates which are often hard to metabolize. In order to glimpse at their possible involvement in carbon turn-over, their ability to degrade carbohydrates was assessed, showing a prevalence of cellulose and starch degradation, which suggests they might be biomass degraders. Pointedly, soil Saccharibacteria MAGs were enriched in glycoside hydrolases associated with starch, cellulose, chitin, and xylan degradation when compared to the reference

genomes. Although further studies are necessary to fully elucidate whether or not CPR are scavengers or primary degraders of complex soil biomass, the findings from this study point towards the presence of soil specific metabolic capability profiles in Saccharibacteria, thus corroborating the view that they participate in direct degradation of soil biomass.

Reconstruction of ultra-small bacteria communities revealed that even this fraction is dominated by few taxa: Gemmatimonadetes, Saccharibacteria and Dependenteiae, and even there, members of CPR appear in low abundances. However, this study suggests that Saccharibacteria is the most abundant member of the CPR in Cullars soil and is strongly negatively correlated to Gemmatimonadetes. The application of this observation to other soil environments will be verified as soil CPR starts to be more studied. Moreover, the ultra-small community changes were not correlated to nutrient levels across plots but to pH and metal concentrations (which are themselves affected by soil pH thresholds). This observation suggests that soil CPR are broadly equipped to deal with nutrient deficiency, a characteristic previously described in the group. Metabolic mechanisms related to low-nutrient, low-pH and high metal concentration tolerance can be further investigated by performing complete genome annotation and metabolic reconstructions. CPR are believed to rely on symbionts on the community for survival as they have severe losses in their biosynthetic potential. Therefore, conclusions about the forces shaping CPR community must take in consideration the complex interplay between CPR populations, environmental factors and whole microbiome shifts.

Sampling of CPR bacteria was sparse even with enrichment. Therefore, the distribution profiles of low abundance CPR such as Parcubacteria could not be reconstructed. Additionally, not all bacteria phyla found initially in the metagenomic library were recovered by the study, indicating that this approach would benefit from improvements in the enrichment and sequencing strategies

to fully characterize CPR distribution and their relationships to soil parameters. Nonetheless, this study provides initial insights that pave the way for more extensive investigations, striving to tackle these complex variables. It points towards the presence of soil specific taxonomic and metabolic profiles in CPR, while suggesting that soil low pH – and consequently high metal concentration – have a higher impact in their community either directly or indirectly, offering a starting point for further studies.

The use of SWNTs as antibiotic agents can lead to breakthrough developments in materials with antimicrobial properties. Therefore, it is paramount that their antibiotic potential be assessed appropriately. The contention in the literature and identification of confounding factors in the experimental design of reports of SWNT activity, prompted the study to re-evaluate SWNT antimicrobial activity. It was hypothesized that SWNT alone was insufficient to kill healthy cells in the absence of sensitizing factors. The study strived to eliminate the major confounding factors: sensitizing dispersing agents and strenuous incubation conditions, hence measuring activity in healthy cells with disposable carbon and energy to deploy stress resistance mechanisms.

Pristine SWNTs tend to bundle together due to strong van der Waals attraction and thus require dispersing agents to be solubilized. These dispersants are frequently surfactants which are known to act on bacterial cell walls, the site of interaction with SWNTs. Therefore, the first step was the innovative utilization of TSB as a SWNT dispersant. The use of TSB provided an ideal standpoint to contrast against other commonly used dispersing agents, surfactants, and also other biomolecules such as DNA and lysozyme. SWNT challenges against gram-positive and gram-negative bacteria, conducted fully in TSB detected no effect of SWNTs in the gram-negative model. SDS intrinsic cytotoxicity was enough to inhibit gram-positive model growth on its own. However, growth inhibition and cell death were observed only when SWNT was dispersed in

pluronic or lysozyme, which are known to destabilize the cell wall. Therefore, although SWNT induced stress was not enough to drive cell-death, it can establish synergistic interactions and display antibiotic activity in the presence of sensitizing factors that destabilize the bacterial cell-wall. Thus, these findings can help design efficient applications of SWNTs as antimicrobial agents or aseptic materials. Additionally, it invites further exploration of bacterial cell responses to SWNT induced stress in order to elucidate the mechanisms of SWNT antimicrobial activity.

The research here in contained dealt with applied and basic science questions and contributed with findings that pave the way to more research fronts and projects with relevant end goals. The search for novel antibacterial agents is a pressing health issue with the increase and spread of bacterial resistance. It is our hope that the advancements in metagenomic library screening here in developed, can be applied not only in the fight against bacterial infections, but also other diseases such cancer and malaria, and even used for industrial applications. Complementary, the appropriate use of SWNTs to enhance antimicrobial activity can lead to groundbreaking applications in the development of improved sterilizing solutions and antimicrobial surfaces, tackling the issue of contamination and biofilm formation in health care settings and apparel.

Our basic research highlighted the soil microbiome diversity. Members of the CPR phyla have unique metabolisms and phylogeny, defying the pre-conceptions of bacterial evolution. Moreover, these enigmatic bacteria can themselves be a source of novel chemistry with biotechnological impact. Ultimately, the elucidation of the forces behind CPR adaptations and niche occupation in soils can lead to a broader understanding of the processes of bacteria adaptation, metabolic capabilities and evolution which is inextricable from the natural history of life itself.

References

- Aakvik, T., Degnes, K. F., Dahlsrud, R., Schmidt, F., Dam, R., Yu, L. H., . . . Valla, S. (2009). A plasmid RK2-based broad-host-range cloning vector useful for transfer of metagenomic libraries to a variety of bacterial species. *FEMS Microbiol Lett*, 296(2), 149-158. doi: 10.1111/j.1574-6968.2009.01639.x
- Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K. L., Tyson, G. W., & Nielsen, P. H. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol*, 31(6), 533-538. doi: 10.1038/nbt.2579
- Almeida, A., Mitchell, A. L., Boland, M., Forster, S. C., Gloor, G. B., Tarkowska, A., . . . Finn, R. D. (2019). A new genomic blueprint of the human gut microbiota. *Nature*, 568(7753), 499-504. doi: 10.1038/s41586-019-0965-1
- Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., . . . Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nat Methods*, 11(11), 1144-1146. doi: 10.1038/nmeth.3103
- Amann, R. I., Ludwig, W., & Schleifer, K. H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev*, 59(1), 143-169.
- Anantharaman, K., Brown, C. T., Hug, L. A., Sharon, I., Castelle, C. J., Probst, A. J., . . . Banfield, J. F. (2016). Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat Commun*, 7, 13219. doi: 10.1038/ncomms13219
- Anderson, H. J., Coleman, J. E., Andersen, R. J., & Roberge, M. (1997). Cytotoxic peptides hemiasterlin, hemiasterlin A and hemiasterlin B induce mitotic arrest and abnormal spindle formation. *Cancer Chemother Pharmacol*, 39(3), 223-226. doi: 10.1007/s002800050564
- Ansari, M. Z., Yadav, G., Gokhale, R. S., & Mohanty, D. (2004). NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. *Nucleic Acids Res*, 32(Web Server issue), W405-413. doi: 10.1093/nar/gkh359
- Ao, G., Nepal, D., Aono, M., & Davis, V. A. (2011). Cholesteric and Nematic Liquid Crystalline Phase Behavior of Double-Stranded DNA Stabilized Single-Walled Carbon Nanotube Dispersions. *ACS Nano*, 5(2), 1450-1458. doi: 10.1021/nn103225r
- Arias, L. R., & Yang, L. (2009). Inactivation of Bacterial Pathogens by Carbon Nanotubes in Suspensions. *Langmuir*, 25(5), 3003-3012. doi: 10.1021/la802769m
- Aslan, S., Loebick, C. Z., Kang, S., Elimelech, M., Pfefferle, L. D., & Van Tassel, P. R. (2010). Antimicrobial biomaterials based on carbon nanotubes dispersed in poly(lactic-co-glycolic acid). *Nanoscale*, 2(9), 1789. doi: 10.1039/c0nr00329h
- Attar, N. (2016). Bacterial evolution: CPR breathes new air into the tree of life. *Nat Rev Microbiol*, 14(6), 332-333. doi: 10.1038/nrmicro.2016.63
- Ayling, M., Clark, M. D., & Leggett, R. M. (2020). New approaches for metagenome assembly with short reads. *Brief Bioinform*, 21(2), 584-594. doi: 10.1093/bib/bbz020
- Bai, Y., Park, I. S., Lee, S. J., Bae, T. S., Watari, F., Uo, M., & Lee, M. H. (2011). Aqueous dispersion of surfactant-modified multiwalled carbon nanotubes and their application as an antibacterial agent. *Carbon*, 49(11), 3663-3671. doi: 10.1016/j.carbon.2011.05.002
- Baldrian, P., Berlemont, R., & Martiny, A. C. (2016). Glycoside Hydrolases across Environmental Microbial Communities. *PLoS Comput Biol*, 12(12), e1005300. doi:

10.1371/journal.pcbi.1005300

- Banfield, J. F., Anantharaman, K., Williams, K. H., & Thomas, B. C. (2017). Complete 4.55-Megabase-Pair Genome of "Candidatus Fluvicola riflensis," Curated from Short-Read Metagenomic Sequences. *Genome Announc*, 5(47). doi: 10.1128/genomeA.01299-17
- Banik, J. J., & Brady, S. F. (2010). Recent application of metagenomic approaches toward the discovery of antimicrobials and other bioactive small molecules. *Curr Opin Microbiol*, 13(5), 603-609. doi: 10.1016/j.mib.2010.08.012
- Becraft, E. D., Woyke, T., Jarett, J., Ivanova, N., Godoy-Vitorino, F., Poulton, N., . . . Stepanauskas, R. (2017). Rokubacteria: Genomic Giants among the Uncultured Bacterial Phyla. *Front Microbiol*, 8, 2264. doi: 10.3389/fmicb.2017.02264
- Beja, O., Aravind, L., Koonin, E. V., Suzuki, M. T., Hadd, A., Nguyen, L. P., . . . DeLong, E. F. (2000). Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science*, 289(5486), 1902-1906. doi: 10.1126/science.289.5486.1902
- Beja, O., Spudich, E. N., Spudich, J. L., Leclerc, M., & DeLong, E. F. (2001). Proteorhodopsin phototrophy in the ocean. *Nature*, 411(6839), 786-789. doi: 10.1038/35081051
- Berdy, J. (2005). Bioactive microbial metabolites. *J Antibiot (Tokyo)*, 58(1), 1-26. doi: 10.1038/ja.2005.1
- Biddle, J. F., Fitz-Gibbon, S., Schuster, S. C., Brenchley, J. E., & House, C. H. (2008). Metagenomic signatures of the Peru Margin seafloor biosphere show a genetically distinct environment. *Proc Natl Acad Sci U S A*, 105(30), 10583-10588. doi: 10.1073/pnas.0709942105
- Bitok, J. K., Lemetre, C., Ternei, M. A., & Brady, S. F. (2017). Identification of biosynthetic gene clusters from metagenomic libraries using PPTase complementation in a Streptomyces host. *FEMS Microbiol Lett*, 364(16). doi: 10.1093/femsle/fnx155
- Blin, K., Shaw, S., Steinke, K., Villebro, R., Ziemert, N., Lee, S. Y., . . . Weber, T. (2019). antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res*, 47(W1), W81-W87. doi: 10.1093/nar/gkz310
- Blin, K., Wolf, T., Chevrette, M. G., Lu, X., Schwalen, C. J., Kautsar, S. A., . . . Medema, M. H. (2017). antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res*, 45(W1), W36-W41. doi: 10.1093/nar/gkx319
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114-2120. doi: 10.1093/bioinformatics/btu170
- Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., . . . Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol*, 37(8), 852-857. doi: 10.1038/s41587-019-0209-9
- Bor, B., Collins, A. J., Murugkar, P. P., Balasubramanian, S., To, T. T., Hendrickson, E. L., . . . Dewhirst, F. E. (2020). Insights Obtained by Culturing Saccharibacteria With Their Bacterial Hosts. *J Dent Res*, 99(6), 685-694. doi: 10.1177/0022034520905792
- Bor, B., McLean, J. S., Foster, K. R., Cen, L., To, T. T., Serrato-Guillen, A., . . . He, X. (2018). Rapid evolution of decreased host susceptibility drives a stable relationship between ultrasmall parasite TM7x and its bacterial host. *Proc Natl Acad Sci U S A*, 115(48), 12277-12282. doi: 10.1073/pnas.1810625115
- Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., . . . Woyke, T. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol*, 35(8), 725-731. doi: 10.1038/nbt.3893

- Brady, S. F., Chao, C. J., & Clardy, J. (2004). Long-chain N-acyltyrosine synthases from environmental DNA. *Applied and environmental microbiology*, 70(11), 6865-6870. doi: 10.1128/AEM.70.11.6865-6870.2004
- Brady, S. F., & Clardy, J. (2000). Long-Chain N-Acyl Amino Acid Antibiotics Isolated from Heterologously Expressed Environmental DNA. *Journal of the American Chemical Society*, 122(51), 12903-12904. doi: 10.1021/ja002990u
- Brady, S. F., & Clardy, J. (2005). Cloning and heterologous expression of isocyanide biosynthetic genes from environmental DNA. *Angewandte Chemie (International ed. in English)*, 44(43), 7063-7065. doi: 10.1002/anie.200501941
- Brooks, B., Olm, M. R., Firek, B. A., Baker, R., Thomas, B. C., Morowitz, M. J., & Banfield, J. F. (2017). Strain-resolved analysis of hospital rooms and infants reveals overlap between the human and room microbiome. *Nat Commun*, 8(1), 1814. doi: 10.1038/s41467-017-02018-w
- Brown, C. T., Hug, L. A., Thomas, B. C., Sharon, I., Castelle, C. J., Singh, A., . . . Banfield, J. F. (2015). Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature*, 523(7559), 208-211. doi: 10.1038/nature14486
- Brulc, J. M., Antonopoulos, D. A., Miller, M. E., Wilson, M. K., Yannarell, A. C., Dinsdale, E. A., . . . White, B. A. (2009). Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. *Proc Natl Acad Sci U S A*, 106(6), 1948-1953. doi: 10.1073/pnas.0806191105
- Cacho, R. A., Tang, Y., & Chooi, Y. H. (2014). Next-generation sequencing approach for connecting secondary metabolites to biosynthetic gene clusters in fungi. *Front Microbiol*, 5, 774. doi: 10.3389/fmicb.2014.00774
- Cai, D., Mataraza, J. M., Qin, Z.-H., Huang, Z., Huang, J., Chiles, T. C., . . . Ren, Z. (2005a). Highly efficient molecular delivery into mammalian cells using carbon nanotube spearing. *Nature Methods*, 2(6), 449.
- Calderon, D., Pena, L., Suarez, A., Villamil, C., Ramirez-Rojas, A., Anzola, J. M., . . . Mongui, A. (2019). Recovery and functional validation of hidden soil enzymes in metagenomic libraries. *Microbiologyopen*, 8(4), e00572. doi: 10.1002/mbo3.572
- Capella-Gutierrez, S., Silla-Martinez, J. M., & Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15), 1972-1973. doi: 10.1093/bioinformatics/btp348
- Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., . . . Knight, R. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci U S A*, 108 Suppl 1, 4516-4522. doi: 10.1073/pnas.1000080107
- Castelle, C. J., Brown, C. T., Anantharaman, K., Probst, A. J., Huang, R. H., & Banfield, J. F. (2018). Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nat Rev Microbiol*, 16(10), 629-645. doi: 10.1038/s41579-018-0076-2
- Castelle, C. J., Hug, L. A., Wrighton, K. C., Thomas, B. C., Williams, K. H., Wu, D., . . . Banfield, J. F. (2013). Extraordinary phylogenetic diversity and metabolic versatility in aquifer sediment. *Nat Commun*, 4, 2120. doi: 10.1038/ncomms3120
- Challis, G. L., & Hopwood, D. A. (2003). Synergy and contingency as driving forces for the evolution of multiple secondary metabolite production by *Streptomyces* species. *Proc Natl Acad Sci U S A*, 100 Suppl 2(suppl 2), 14555-14561. doi: 10.1073/pnas.1934677100

- Chang, F.-Y. Y., & Brady, S. F. (2013). Discovery of indolotryptoline antiproliferative agents by homology-guided metagenomic screening. *Proceedings of the National Academy of Sciences of the United States of America*, 110(7), 2478-2483. doi: 10.1073/pnas.1218073110
- Chang, F. Y., & Brady, S. F. (2013). Discovery of indolotryptoline antiproliferative agents by homology-guided metagenomic screening. *Proc Natl Acad Sci U S A*, 110(7), 2478-2483. doi: 10.1073/pnas.1218073110
- Charlop-Powers, Z., Banik, J. J., Owen, J. G., Craig, J. W., & Brady, S. F. (2013). Selective enrichment of environmental DNA libraries for genes encoding nonribosomal peptides and polyketides by phosphopantetheine transferase-dependent complementation of siderophore biosynthesis. *ACS Chem Biol*, 8(1), 138-143. doi: 10.1021/cb3004918
- Charlop-Powers, Z., Pregitzer, C. C., Lemetre, C., Ternei, M. A., Maniko, J., Hover, B. M., . . . Brady, S. F. (2016). Urban park soil microbiomes are a rich reservoir of natural product biosynthetic diversity. *Proc Natl Acad Sci U S A*, 113(51), 14811-14816. doi: 10.1073/pnas.1615581113
- Chen, L. X., Anantharaman, K., Shaiber, A., Eren, A. M., & Banfield, J. F. (2020). Accurate and complete genomes from metagenomes. *Genome Res*, 30(3), 315-333. doi: 10.1101/gr.258640.119
- Chen, R. J., Bangsaruntip, S., Drouvalakis, K. A., Wong Shi Kam, N., Shim, M., Li, Y., . . . Dai, H. (2003). Noncovalent functionalization of carbon nanotubes for highly specific electronic biosensors. *Proceedings of the National Academy of Sciences*, 100(9), 4984-4989. doi: 10.1073/pnas.0837064100
- Chen, X., Tam, U. C., Czapinski, J. L., Lee, G. S., Rabuka, D., Zettl, A., & Bertozzi, C. R. (2006). Interfacing carbon nanotubes with living cells. *Journal of the American Chemical Society*, 128(19), 6292-6293.
- Cheng, J., Pinnell, L., Engel, K., Neufeld, J. D., & Charles, T. C. (2014). Versatile broad-host-range cosmids for construction of high quality metagenomic libraries. *J Microbiol Methods*, 99, 27-34. doi: 10.1016/j.mimet.2014.01.015
- Cheng, J., Romantsov, T., Engel, K., Doxey, A. C., Rose, D. R., Neufeld, J. D., & Charles, T. C. (2017). Functional metagenomics reveals novel beta-galactosidases not predictable from gene sequences. *PLoS One*, 12(3), e0172545. doi: 10.1371/journal.pone.0172545
- Chivian, D., Brodie, E. L., Alm, E. J., Culley, D. E., Dehal, P. S., DeSantis, T. Z., . . . Onstott, T. C. (2008). Environmental genomics reveals a single-species ecosystem deep within Earth. *Science*, 322(5899), 275-278. doi: 10.1126/science.1155495
- Chowdhury, I., Duch, M. C., Gits, C. C., Hersam, M. C., & Walker, S. L. (2012). Impact of synthesis methods on the transport of single walled carbon nanotubes in the aquatic environment. *Environmental science & technology*, 46(21), 11752-11760.
- Cimermancic, P., Medema, M. H., Claesen, J., Kurita, K., Wieland Brown, L. C., Mavrommatis, K., . . . Fischbach, M. A. (2014). Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell*, 158(2), 412-421. doi: 10.1016/j.cell.2014.06.034
- Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., . . . Tiedje, J. M. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res*, 42(Database issue), D633-642. doi: 10.1093/nar/gkt1244
- Coughlan, L. M., Cotter, P. D., & Hill, C. (2015). Biotechnological applications of functional metagenomics in the food and pharmaceutical industries. *Frontiers in* doi:

10.3389/fmicb.2015.00672

- Courtois, S., Cappellano, C. M., Ball, M., Francou, F. X., Normand, P., Helynck, G., . . . Pernodet, J. L. (2003). Recombinant environmental libraries provide access to microbial diversity for drug discovery from natural products. *Appl Environ Microbiol*, 69(1), 49-55. doi: 10.1128/AEM.69.1.49-55.2003
- Cragg, G. M., & Newman, D. J. (2013). Natural products: a continuing source of novel drug leads. *Biochim Biophys Acta*, 1830(6), 3670-3695. doi: 10.1016/j.bbagen.2013.02.008
- Craig, J. W., Chang, F.-Y. Y., & Brady, S. F. (2009). Natural products from environmental DNA hosted in *Ralstonia metallidurans*. *ACS chemical biology*, 4(1), 23-28. doi: 10.1021/cb8002754
- Craig, J. W., Chang, F. Y., Kim, J. H., Obiajulu, S. C., & Brady, S. F. (2010). Expanding small-molecule functional metagenomics through parallel screening of broad-host-range cosmid environmental DNA libraries in diverse proteobacteria. *Appl Environ Microbiol*, 76(5), 1633-1641. doi: 10.1128/AEM.02169-09
- Crits-Christoph, A., Diamond, S., Butterfield, C. N., Thomas, B. C., & Banfield, J. F. (2018). Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis. *Nature*, 558(7710), 440-444. doi: 10.1038/s41586-018-0207-y
- Culligan, E. P., Sleator, R. D., Marchesi, J. R., & Hill, C. (2014). Metagenomics and novel gene discovery: promise and potential for novel therapeutics. *Virulence*, 5(3), 399-412. doi: 10.4161/viru.27208
- Danczak, R. E., Johnston, M. D., Kenah, C., Slattery, M., Wrighton, K. C., & Wilkins, M. J. (2017). Members of the Candidate Phyla Radiation are functionally differentiated by carbon- and nitrogen-cycling capabilities. *Microbiome*, 5(1), 112. doi: 10.1186/s40168-017-0331-1
- Daniel, R. (2004). The soil metagenome - a rich resource for the discovery of novel natural products. *Curr Opin Biotechnol*, 15(3), 199-204. doi: 10.1016/j.copbio.2004.04.005
- DeBruyn, J. M., Nixon, L. T., Fawaz, M. N., Johnson, A. M., & Radosevich, M. (2011). Global biogeography and quantitative seasonal dynamics of Gemmatimonadetes in soil. *Appl Environ Microbiol*, 77(17), 6295-6300. doi: 10.1128/AEM.05005-11
- Deeg, C. M., Zimmer, M. M., George, E. E., Husnik, F., Keeling, P. J., & Suttle, C. A. (2019). Chromulinavorax destructans, a pathogen of microzooplankton that provides a window into the enigmatic candidate phylum Dependotriaceae. *PLoS Pathog*, 15(5), e1007801. doi: 10.1371/journal.ppat.1007801
- Dick, G. J., Andersson, A. F., Baker, B. J., Simmons, S. L., Thomas, B. C., Yelton, A. P., & Banfield, J. F. (2009). Community-wide analysis of microbial genome sequence signatures. *Genome Biol*, 10(8), R85. doi: 10.1186/gb-2009-10-8-r85
- Dojka, M. A., Hugenholtz, P., Haack, S. K., & Pace, N. R. (1998). Microbial diversity in a hydrocarbon- and chlorinated-solvent-contaminated aquifer undergoing intrinsic bioremediation. *Appl Environ Microbiol*, 64(10), 3869-3877. doi: 10.1128/AEM.64.10.3869-3877.1998
- Dong, L., Henderson, A., & Field, C. (2012a). Antimicrobial Activity of Single-Walled Carbon Nanotubes Suspended in Different Surfactants. *Journal of Nanotechnology*, 2012, 1-7. doi: 10.1155/2012/928924
- Dresselhaus, M., Dresselhaus, G., & Saito, R. (1995). Physics of carbon nanotubes. *Carbon*, 33(7), 883-891.
- Du, L., & Lou, L. (2010). PKS and NRPS release mechanisms. *Nat Prod Rep*, 27(2), 255-278. doi: 10.1039/b912037h

- Du, L., Sanchez, C., & Shen, B. (2001). Hybrid peptide-polyketide natural products: biosynthesis and prospects toward engineering novel molecules. *Metab Eng*, 3(1), 78-95. doi: 10.1006/mben.2000.0171
- Dumortier, H., Lacotte, S., Pastorin, G., Marega, R., Wu, W., Bonifazi, D., . . . Bianco, A. (2006). Functionalized carbon nanotubes are non-cytotoxic and preserve the functionality of primary immune cells. *Nano Letters*, 6(7), 1522-1528.
- Dzunkova, M., D'Auria, G., Perez-Villarroya, D., & Moya, A. (2012). Hybrid sequencing approach applied to human fecal metagenomic clone libraries revealed clones with potential biotechnological applications. *PLoS One*, 7(10), e47654. doi: 10.1371/journal.pone.0047654
- Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Comput Biol*, 7(10), e1002195. doi: 10.1371/journal.pcbi.1002195
- Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32(5), 1792-1797. doi: 10.1093/nar/gkh340
- Edwards, R. A., Rodriguez-Brito, B., Wegley, L., Haynes, M., Breitbart, M., Peterson, D. M., . . . Rohwer, F. (2006). Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics*, 7, 57. doi: 10.1186/1471-2164-7-57
- Eloe-Fadrosh, E. A., Ivanova, N. N., Woyke, T., & Kyrpides, N. C. (2016). Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. *Nat Microbiol*, 1(4). doi: Artn 15032
10.1038/Nmicrobiol.2015.32
- Eloe-Fadrosh, E. A., Paez-Espino, D., Jarett, J., Dunfield, P. F., Hedlund, B. P., Dekas, A. E., . . . Ivanova, N. N. (2016). Global metagenomic survey reveals a new bacterial candidate phylum in geothermal springs. *Nat Commun*, 7, 10476. doi: 10.1038/ncomms10476
- Epstein, S. C., Charkoudian, L. K., & Medema, M. H. (2018). A standardized workflow for submitting data to the Minimum Information about a Biosynthetic Gene cluster (MIBiG) repository: prospects for research-based educational experiences. *Stand Genomic Sci*, 13. doi: ARTN 16
10.1186/s40793-018-0318-y
- Evans, P. N., Parks, D. H., Chadwick, G. L., Robbins, S. J., Orphan, V. J., Golding, S. D., & Tyson, G. W. (2015). Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics. *Science*, 350(6259), 434-438. doi: 10.1126/science.aac7745
- Eyice, O., Namura, M., Chen, Y., Mead, A., Samavedam, S., & Schafer, H. (2015). SIP metagenomics identifies uncultivated Methylophilaceae as dimethylsulphide degrading bacteria in soil and lake sediment. *ISME J*, 9(11), 2336-2348. doi: 10.1038/ismej.2015.37
- Felczykowska, A., Dydecka, A., Bohdanowicz, M., Gasior, T., Sobon, M., Kobos, J., . . . Wegrzyn, G. (2014). The use of fosmid metagenomic libraries in preliminary screening for various biological activities. *Microb Cell Fact*, 13(1), 105. doi: 10.1186/s12934-014-0105-4
- Felnagle, E. A., Jackson, E. E., Chan, Y. A., Podevels, A. M., Berti, A. D., McMahon, M. D., & Thomas, M. G. (2008). Nonribosomal peptide synthetases involved in the production of medically relevant natural products. *Mol Pharm*, 5(2), 191-211. doi: 10.1021/mp700137g
- Feng, Z., Kallifidas, D., & Brady, S. F. (2011). Functional analysis of environmental DNA-derived type II polyketide synthases reveals structurally diverse secondary metabolites. *Proc Natl Acad Sci U S A*, 108(31), 12629-12634. doi: 10.1073/pnas.1103921108
- Figuroa-Gonzalez, P. A., Bornemann, T. L. V., Adam, P. S., Plewka, J., Revesz, F., von Hagen, C.

- A., . . . Probst, A. J. (2020). Saccharibacteria as Organic Carbon Sinks in Hydrocarbon-Fueled Communities. *Front Microbiol*, 11, 587782. doi: 10.3389/fmicb.2020.587782
- Fisch, K. M., Gurgui, C., Heycke, N., van der Sar, S. A., Anderson, S. A., Webb, V. L., . . . Piel, J. (2009). Polyketide assembly lines of uncultivated sponge symbionts from structure-based gene targeting. *Nat Chem Biol*, 5(7), 494-501. doi: 10.1038/nchembio.176
- Foerstner, K. U., Doerks, T., Creevey, C. J., Doerks, A., & Bork, P. (2008). A computational screen for type I polyketide synthases in metagenomics shotgun data. *PLoS ONE*, 3(10), e3515. doi: 10.1371/journal.pone.0003515
- Frank, J. A., Reich, C. I., Sharma, S., Weisbaum, J. S., Wilson, B. A., & Olsen, G. J. (2008). Critical evaluation of two primers commonly used for amplification of bacterial 16S rRNA genes. *Appl Environ Microbiol*, 74(8), 2461-2470. doi: 10.1128/AEM.02272-07
- Fu, J., Bian, X., Hu, S., Wang, H., Huang, F., Seibert, P. M., . . . Zhang, Y. (2012). Full-length RecE enhances linear-linear homologous recombination and facilitates direct cloning for bioprospecting. *Nat Biotechnol*, 30(5), 440-446. doi: 10.1038/nbt.2183
- Gagen, E. J., Levett, A., Paz, A., Bostelmann, H., Valadares, R., Bitencourt, J. A. P., . . . Southam, G. (2020). Accelerating microbial iron cycling promotes re-cementation of surface crusts in iron ore regions. *Microb Biotechnol*, 13(6), 1960-1971. doi: 10.1111/1751-7915.13646
- Gaida, S. M., Sandoval, N. R., Nicolaou, S. A., Chen, Y., Venkataramanan, K. P., & Papoutsakis, E. T. (2015). Expression of heterologous sigma factors enables functional screening of metagenomic and heterologous genomic libraries. *Nat Commun*, 6, 7045. doi: 10.1038/ncomms8045
- Ghuneim, L. J., Jones, D. L., Golyshin, P. N., & Golyshina, O. V. (2018). Nano-Sized and Filterable Bacteria and Archaea: Biodiversity and Function. *Front Microbiol*, 9, 1971. doi: 10.3389/fmicb.2018.01971
- Gillespie, D. E., Brady, S. F., Bettermann, A. D., Cianciotto, N. P., Liles, M. R., Rondon, M. R., . . . Handelsman, J. (2002). Isolation of antibiotics turbomycin a and B from a metagenomic library of soil microbial DNA. *Appl Environ Microbiol*, 68(9), 4301-4306. doi: 10.1128/aem.68.9.4301-4306.2002
- Gomez-Escribano, J. P., & Bibb, M. J. (2014). Heterologous expression of natural product biosynthetic gene clusters in *Streptomyces coelicolor*: from genome mining to manipulation of biosynthetic pathways. *J Ind Microbiol Biotechnol*, 41(2), 425-431. doi: 10.1007/s10295-013-1348-5
- Gomez-Escribano, J. P., Song, L. J., Fox, D. J., Yeo, V., Bibb, M. J., & Challis, G. L. (2012). Structure and biosynthesis of the unusual polyketide alkaloid coelimycin P1, a metabolic product of the cpk gene cluster of *Streptomyces coelicolor* M145. *Chemical Science*, 3(9), 2716-2720. doi: DOI 10.1039/c2sc20410j
- Gong, J., Qing, Y., Guo, X., & Warren, A. (2014). "Candidatus *Sonnebornia yantaiensis*", a member of candidate division OD1, as intracellular bacteria of the ciliated protist *Paramecium bursaria* (Ciliophora, Oligohymenophorea). *Syst Appl Microbiol*, 37(1), 35-41. doi: 10.1016/j.syapm.2013.08.007
- Gregor, I., Droge, J., Schirmer, M., Quince, C., & McHardy, A. C. (2016). PhyloPythiaS+: a self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. *PeerJ*, 4, e1603. doi: 10.7717/peerj.1603
- Guo, J., Cole, J. R., Zhang, Q., Brown, C. T., & Tiedje, J. M. (2016). Microbial Community Analysis with Ribosomal Gene Fragments from Shotgun Metagenomes. *Appl Environ Microbiol*, 82(1), 157-166. doi: 10.1128/AEM.02772-15

- Gupta, A., Kumar, S., Prasoodanan, V. P. K., Harish, K., Sharma, A. K., & Sharma, V. K. (2016). Reconstruction of Bacterial and Viral Genomes from Multiple Metagenomes. *Front Microbiol*, 7. doi: 10.3389/fmicb.2016.00469
- Haley, J. D. (1988). Cosmid library construction. *Methods Mol Biol*, 4, 257-283. doi: 10.1385/0-89603-127-6:257
- Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev*, 68(4), 669-685. doi: 10.1128/MMBR.68.4.669-685.2004
- Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol*, 5(10), R245-249. doi: 10.1016/S1074-5521(98)90108-9
- Harris, J. K., Kelley, S. T., & Pace, N. R. (2004). New perspective on uncultured bacterial phylogenetic division OP11. *Appl Environ Microbiol*, 70(2), 845-849. doi: 10.1128/aem.70.2.845-849.2004
- He, X., McLean, J. S., Edlund, A., Yooseph, S., Hall, A. P., Liu, S. Y., . . . Shi, W. (2015). Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. *Proc Natl Acad Sci U S A*, 112(1), 244-249. doi: 10.1073/pnas.1419038112
- Healy, F. G., Ray, R. M., Aldrich, H. C., Wilkie, A. C., Ingram, L. O., & Shanmugam, K. T. (1995). Direct Isolation of Functional Genes Encoding Cellulases from the Microbial Consortia in a Thermophilic, Anaerobic Digester Maintained on Lignocellulose. *Appl Microbiol Biotechnol*, 43(4), 667-674.
- Heil, J. R., Cheng, J., & Charles, T. C. (2012). Site-specific bacterial chromosome engineering: PhiC31 integrase mediated cassette exchange (IMCE). *J Vis Exp*(61). doi: 10.3791/3698
- Helfrich, E. J., Reiter, S., & Piel, J. (2014). Recent advances in genome-based polyketide discovery. *Curr Opin Biotechnol*, 29, 107-115. doi: 10.1016/j.copbio.2014.03.004
- Herlemann, D. P., Geissinger, O., Ikeda-Ohtsubo, W., Kunin, V., Sun, H., Lapidus, A., . . . Brune, A. (2009). Genomic analysis of "Elusimicrobium minutum," the first cultivated representative of the phylum "Elusimicrobia" (formerly termite group 1). *Appl Environ Microbiol*, 75(9), 2841-2849. doi: 10.1128/AEM.02698-08
- Hernsdorf, A. W., Amano, Y., Miyakawa, K., Ise, K., Suzuki, Y., Anantharaman, K., . . . Banfield, J. F. (2017). Potential for microbial H₂ and metal transformations associated with novel bacteria and archaea in deep terrestrial subsurface sediments. *ISME J*, 11(8), 1915-1929. doi: 10.1038/ismej.2017.39
- Herrmann, M., Wegner, C. E., Taubert, M., Geesink, P., Lehmann, K., Yan, L., . . . Kusel, K. (2019). Predominance of Cand. Patescibacteria in Groundwater Is Caused by Their Preferential Mobilization From Soils and Flourishing Under Oligotrophic Conditions. *Front Microbiol*, 10, 1407. doi: 10.3389/fmicb.2019.01407
- Hertweck, C. (2009). The biosynthetic logic of polyketide diversity. *Angew Chem Int Ed Engl*, 48(26), 4688-4716. doi: 10.1002/anie.200806121
- Hill, P., Piel, J., Aris-Brosou, S., Kristufek, V., Boddy, C. N., & Dijkhuizen, L. (2014). Habitat-specific type I polyketide synthases in soils and street sediments. *J Ind Microbiol Biotechnol*, 41(1), 75-85. doi: 10.1007/s10295-013-1362-7
- Hong, S. H., Bunge, J., Leslin, C., Jeon, S., & Epstein, S. S. (2009). Polymerase chain reaction primers miss half of rRNA microbial diversity. *Isme Journal*, 3(12), 1365-1373. doi: 10.1038/ismej.2009.89
- Horn, D. W., Ao, G., Maugey, M., Zakri, C., Poulin, P., & Davis, V. A. (2013). Dispersion State

- and Fiber Toughness: Antibacterial Lysozyme-Single Walled Carbon Nanotubes. *Advanced Functional Materials*, 23(48), 6082-6090. doi: 10.1002/adfm.201300221
- Horn, D. W., Tracy, K., Easley, C. J., & Davis, V. A. (2012). Lysozyme dispersed single-walled carbon nanotubes: interaction and activity. *The Journal of Physical Chemistry C*, 116(18), 10341-10348.
- Hosokawa, S., Kuroda, K., Narihiro, T., Aoi, Y., Ozaki, N., Ohashi, A., & Kindaichi, T. (2021). Cometabolism of the Superphylum Patescibacteria with Anammox Bacteria in a Long-Term Freshwater Anammox Column Reactor. *Water*, 13(2), 208.
- Hover, B. M., Kim, S. H., Katz, M., Charlop-Powers, Z., Owen, J. G., Ternei, M. A., . . . Brady, S. F. (2018). Culture-independent discovery of the malacidins as calcium-dependent antibiotics with activity against multidrug-resistant Gram-positive pathogens. *Nat Microbiol*, 3(4), 415-422. doi: 10.1038/s41564-018-0110-1
- Howe, A. C., Jansson, J. K., Malfatti, S. A., Tringe, S. G., Tiedje, J. M., & Brown, C. T. (2014). Tackling soil diversity with the assembly of large, complex metagenomes. *Proc Natl Acad Sci U S A*, 111(13), 4904-4909. doi: 10.1073/pnas.1402564111
- Hu, H., Zhang, Q., & Ochi, K. (2002). Activation of antibiotic biosynthesis by specified mutations in the *rpoB* gene (encoding the RNA polymerase beta subunit) of *Streptomyces lividans*. *J Bacteriol*, 184(14), 3984-3991. doi: 10.1128/JB.184.14.3984-3991.2002
- Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., . . . Banfield, J. F. (2016). A new view of the tree of life. *Nat Microbiol*, 1, 16048. doi: 10.1038/nmicrobiol.2016.48
- Hugenholtz, P., Goebel, B. M., & Pace, N. R. (1998). Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *Journal of bacteriology*, 180(18), 4765-4774.
- Hugenholtz, P., Pitulle, C., Hershberger, K. L., & Pace, N. R. (1998). Novel division level bacterial diversity in a Yellowstone hot spring. *J Bacteriol*, 180(2), 366-376. doi: 10.1128/JB.180.2.366-376.1998
- Hugenholtz, P., Tyson, G. W., Webb, R. I., Wagner, A. M., & Blackall, L. L. (2001). Investigation of candidate division TM7, a recently recognized major lineage of the domain Bacteria with no known pure-culture representatives. *Appl Environ Microbiol*, 67(1), 411-419. doi: 10.1128/AEM.67.1.411-419.2001
- Human Microbiome Project, C. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402), 207-214. doi: 10.1038/nature11234
- Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11, 119. doi: 10.1186/1471-2105-11-119
- Iijima, S. (1991). Helical microtubules of graphitic carbon. *Nature*, 354(6348), 56-58. doi: 10.1038/354056a0
- Iqbal, H. A., Craig, J. W., & Brady, S. F. (2014). Antibacterial enzymes from the functional screening of metagenomic libraries hosted in *Ralstonia metallidurans*. *FEMS microbiology letters*, 354(1), 19-26. doi: 10.1111/1574-6968.12431
- Ishikawa, S., Matsumura, Y., Katoh-Kubo, K., & Tsuchido, T. (2002). Antibacterial activity of surfactants against *Escherichia coli* cells is influenced by carbon source and anaerobiosis. *Journal of Applied Microbiology*, 93(2), 302-309. doi: 10.1046/j.1365-2672.2002.01690.x
- Jaffe, A. L., Castelle, C. J., Matheus Carnevali, P. B., Gribaldo, S., & Banfield, J. F. (2020). The rise of diversity in metabolic platforms across the Candidate Phyla Radiation. *BMC Biol*,

- 18(1), 69. doi: 10.1186/s12915-020-00804-5
- Janssen, P. H. (2006). Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. *Appl Environ Microbiol*, 72(3), 1719-1728. doi: 10.1128/Aem.72.3.1719-1728.2006
- Jorio, A., Kauppinen, E., & Hassanien, A. (2007). Carbon-Nanotube Metrology. *III*, 63-100. doi: 10.1007/978-3-540-72865-8_3
- Kakirde, K. S., Wild, J., Godiska, R., Mead, D. A., Wiggins, A. G., Goodman, R. M., . . . Liles, M. R. (2011). Gram negative shuttle BAC vector for heterologous expression of metagenomic libraries. *Gene*, 475(2), 57-62. doi: 10.1016/j.gene.2010.11.004
- Kallifidas, D., & Brady, S. F. (2012). Reassembly of Functionally Intact Environmental DNA-Derived Biosynthetic Gene Clusters. *Natural Product Biosynthesis by Microorganisms and Plants, Pt C*, 517, 225-239. doi: 10.1016/B978-0-12-404634-4.00011-5
- Kallifidas, D., Kang, H. S., & Brady, S. F. (2012). Tetarimycin A, an MRSA-active antibiotic identified through induced expression of environmental DNA gene clusters. *J Am Chem Soc*, 134(48), 19552-19555. doi: 10.1021/ja3093828
- Kam, N. W. S., & Dai, H. (2005a). Carbon nanotubes as intracellular protein transporters: generality and biological functionality. *Journal of the American Chemical Society*, 127(16), 6021-6026.
- Kam, N. W. S., & Dai, H. (2005b). Carbon Nanotubes as Intracellular Protein Transporters: Generality and Biological Functionality. *J Am Chem Soc*, 127(16), 6021-6026. doi: 10.1021/ja050062v
- Kandlikar, G. S., Gold, Z. J., Cowen, M. C., Meyer, R. S., Freise, A. C., Kraft, N. J. B., . . . Curd, E. E. (2018). ranacapa: An R package and Shiny web app to explore environmental DNA data with exploratory statistics and interactive visualizations. *F1000Res*, 7, 1734. doi: 10.12688/f1000research.16680.1
- Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., & Wang, Z. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, 7, e7359. doi: 10.7717/peerj.7359
- Kang, H. S., & Brady, S. F. (2013). Arimetamycin A: improving clinically relevant families of natural products through sequence-guided screening of soil metagenomes. *Angew Chem Int Ed Engl*, 52(42), 11063-11067. doi: 10.1002/anie.201305109
- Kang, H. S., & Brady, S. F. (2014a). Arixanthomycins A-C: Phylogeny-guided discovery of biologically active eDNA-derived pentangular polyphenols. *ACS Chem Biol*, 9(6), 1267-1272. doi: 10.1021/cb500141b
- Kang, H. S., & Brady, S. F. (2014b). Mining soil metagenomes to better understand the evolution of natural product structural diversity: pentangular polyphenols as a case study. *J Am Chem Soc*, 136(52), 18111-18119. doi: 10.1021/ja510606j
- Kang, S., Herzberg, M., Rodrigues, D. F., & Elimelech, M. (2008). Antibacterial Effects of Carbon Nanotubes: Size Does Matter! *Langmuir*, 24(13), 6409-6413. doi: 10.1021/la800951v
- Kang, S., Pinault, M., Pfefferle, L. D., & Elimelech, M. (2007). Single-walled carbon nanotubes exhibit strong antimicrobial activity. *Langmuir*, 23(17), 8670-8673.
- Kantor, R. S., Huddy, R. J., Iyer, R., Thomas, B. C., Brown, C. T., Anantharaman, K., . . . Banfield, J. F. (2017). Genome-Resolved Meta-Omics Ties Microbial Dynamics to Process Performance in Biotechnology for Thiocyanate Degradation. *Environ Sci Technol*, 51(5), 2944-2953. doi: 10.1021/acs.est.6b04477
- Kantor, R. S., Wrighton, K. C., Handley, K. M., Sharon, I., Hug, L. A., Castelle, C. J., . . . Banfield,

- J. F. (2013). Small genomes and sparse metabolisms of sediment-associated bacteria from four candidate phyla. *mBio*, 4(5), e00708-00713. doi: 10.1128/mBio.00708-13
- Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*, 30(4), 772-780. doi: 10.1093/molbev/mst010
- Khosla, C., Herschlag, D., Cane, D. E., & Walsh, C. T. (2014). Assembly line polyketide synthases: mechanistic insights and unsolved problems. *Biochemistry*, 53(18), 2875-2883. doi: 10.1021/bi500290t
- Kim, Y., Koh, I., & Rho, M. (2015). Deciphering the human microbiome using next-generation sequencing data and bioinformatics approaches. *Methods*, 79-80, 52-59. doi: 10.1016/j.ymeth.2014.10.022
- Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., & Glockner, F. O. (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res*, 41(1), e1. doi: 10.1093/nar/gks808
- Kort, R., Keijser, B. J., Caspers, M. P. M., Schuren, F. H., & Montijn, R. (2008). Transcriptional activity around bacterial cell death reveals molecular biomarkers for cell viability. *BMC Genomics*, 9(1), 590. doi: 10.1186/1471-2164-9-590
- Kostarelos, K., Lacerda, L., Pastorin, G., Wu, W., Wieckowski, S., Luangsivilay, J., . . . Muller, S. (2007). Cellular uptake of functionalized carbon nanotubes is independent of functional group and cell type. *Nature Nanotechnology*, 2(2), 108.
- Kroeger, M. E., Delmont, T. O., Eren, A. M., Meyer, K. M., Guo, J., Khan, K., . . . Nusslein, K. (2018). New Biological Insights Into How Deforestation in Amazonia Affects Soil Microbial Communities Using Metagenomics and Metagenome-Assembled Genomes. *Front Microbiol*, 9, 1635. doi: 10.3389/fmicb.2018.01635
- Kuske, C. R., Barns, S. M., & Busch, J. D. (1997). Diverse uncultivated bacterial groups from soils of the arid southwestern United States that are present in many geographic regions. *Appl Environ Microbiol*, 63(9), 3614-3621. doi: 10.1128/AEM.63.9.3614-3621.1997
- Kusnezowa, A., & Leichert, L. I. (2017). In silico approach to designing rational metagenomic libraries for functional studies. *BMC Bioinformatics*, 18(1), 267. doi: 10.1186/s12859-017-1668-y
- Ladoukakis, E., Kolisis, F. N., & Chatziioannou, A. A. (2014). Integrative workflows for metagenomic analysis. *Front Cell Dev Biol*, 2, 70. doi: 10.3389/fcell.2014.00070
- Lam, K. N., & Charles, T. C. (2015). Strong spurious transcription likely contributes to DNA insert bias in typical metagenomic clone libraries. *Microbiome*, 3, 22. doi: 10.1186/s40168-015-0086-5
- Lane, D. J., Pace, B., Olsen, G. J., Stahl, D. A., Sogin, M. L., & Pace, N. R. (1985). Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci U S A*, 82(20), 6955-6959. doi: 10.1073/pnas.82.20.6955
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 9(4), 357-359. doi: 10.1038/nmeth.1923
- Leff, J. W., Jones, S. E., Prober, S. M., Barberan, A., Borer, E. T., Firn, J. L., . . . Fierer, N. (2015). Consistent responses of soil microbial communities to elevated nutrient inputs in grasslands across the globe. *Proc Natl Acad Sci U S A*, 112(35), 10967-10972. doi: 10.1073/pnas.1508382112
- Lemos, L. N., Medeiros, J. D., Dini-Andreote, F., Fernandes, G. R., Varani, A. M., Oliveira, G., &

- Pyro, V. S. (2019). Genomic signatures and co-occurrence patterns of the ultra-small Saccharimonadia (phylum CPR/Patescibacteria) suggest a symbiotic lifestyle. *Mol Ecol*, 28(18), 4259-4271. doi: 10.1111/mec.15208
- Letunic, I., & Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res*, 47(W1), W256-W259. doi: 10.1093/nar/gkz239
- Lewin, A., Zhou, J. L., Pham, V. T. T., Haugen, T., El Zeiny, M., Aarstad, O., . . . Liles, M. R. (2017). Novel archaeal thermostable cellulases from an oil reservoir metagenome. *AMB Express*, 7. doi: Art 183
- Li, A., Wang, L., Li, M., Xu, Z., Li, P., & Zhang, Z. (2014). The antibacterial activity of covalent-functionalized carbon nanotubes against Escherichia coli. *Digest Journal of Nanomaterials and Biostructures*, 9(2), 599-607.
- Li, D., Liu, C. M., Luo, R., Sadakane, K., & Lam, T. W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, 31(10), 1674-1676. doi: 10.1093/bioinformatics/btv033
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), 1754-1760. doi: 10.1093/bioinformatics/btp324
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), 2078-2079. doi: 10.1093/bioinformatics/btp352
- Li, J. W., & Vederas, J. C. (2009). Drug discovery and natural products: end of an era or an endless frontier? *Science*, 325(5937), 161-165. doi: 10.1126/science.1168243
- Liles, M. R., Manske, B. F., Bintrim, S. B., Handelsman, J., & Goodman, R. M. (2003). A census of rRNA genes and linked genomic sequences within a soil metagenomic library. *Applied and Environmental Microbiology*, 69(5), 2684-2691. doi: Doi 10.1128/Aem.69.5.2684-2691.2003
- Liles, M. R., Williamson, L. L., Rodbumrer, J., Torsvik, V., Goodman, R. M., & Handelsman, J. (2008). Recovery, purification, and cloning of high-molecular-weight DNA from soil microorganisms. *Appl Environ Microbiol*, 74(10), 3302-3305. doi: 10.1128/AEM.02630-07
- Ling, L. L., Schneider, T., Peoples, A. J., Spoering, A. L., Engels, I., Conlon, B. P., . . . Lewis, K. (2015). A new antibiotic kills pathogens without detectable resistance. *Nature*, 517(7535), 455-459. doi: 10.1038/nature14098
- Liu, J. A., Li, J. Q., Feng, L., Cao, H., & Cui, Z. L. (2010). An Improved Method for Extracting Bacteria from Soil for High Molecular Weight DNA Recovery and BAC Library Construction. *Journal of Microbiology*, 48(6), 728-733. doi: 10.1007/s12275-010-0139-1
- Liu, S., Wei, L., Hao, L., Fang, N., Chang, M. W., Xu, R., . . . Chen, Y. (2009). Sharper and Faster “Nano Darts” Kill More Bacteria: A Study of Antibacterial Activity of Individually Dispersed Pristine Single-Walled Carbon Nanotube. *ACS Nano*, 3(12), 3891-3902. doi: 10.1021/nn901252r
- Liu, Y., Yang, D., Zhang, N., Chen, L., Cui, Z., Shen, Q., & Zhang, R. (2016). Characterization of Uncultured Genome Fragment from Soil Metagenomic Library Exposed Rare Mismatch of Internal Tetranucleotide Frequency. *Front Microbiol*, 7, 2081. doi: 10.3389/fmicb.2016.02081
- Liu, Z., Lozupone, C., Hamady, M., Bushman, F. D., & Knight, R. (2007). Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Res*, 35(18), e120. doi: 10.1093/nar/gkm541

- Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M., & Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res*, 42(Database issue), D490-495. doi: 10.1093/nar/gkt1178
- Luef, B., Frischkorn, K. R., Wrighton, K. C., Holman, H. Y., Birarda, G., Thomas, B. C., . . . Banfield, J. F. (2015). Diverse uncultivated ultra-small bacterial cells in groundwater. *Nat Commun*, 6, 6372. doi: 10.1038/ncomms7372
- Luo, C., Tsementzi, D., Kyrpides, N. C., & Konstantinidis, K. T. (2012). Individual genome assembly from complex community short-read metagenomic datasets. *ISME J*, 6(4), 898-901. doi: 10.1038/ismej.2011.147
- Lynch, M. D., & Neufeld, J. D. (2015). Ecology and exploration of the rare biosphere. *Nature Reviews Microbiology*, 13(4), 217-229.
- Manna, S. K., Sarkar, S., Barr, J., Wise, K., Barrera, E. V., Jejelowo, O., . . . Ramesh, G. T. (2005). Single-Walled Carbon Nanotube Induces Oxidative Stress and Activates Nuclear Transcription Factor- κ B in Human Keratinocytes. *Nano Letters*, 5(9), 1676-1684. doi: 10.1021/nl0507966
- Martinez, D. S. T., Faria, A. F., Berni, E., Souza Filho, A. G., Almeida, G., Caloto-Oliveira, A., . . . Alves, O. L. (2014). Exploring the use of biosurfactants from *Bacillus subtilis* in bionanotechnology: A potential dispersing agent for carbon nanotube ecotoxicological studies. *Process Biochemistry*, 49(7), 1162-1168. doi: 10.1016/j.procbio.2014.04.006
- Masschelein, J., Mattheus, W., Gao, L. J., Moons, P., Van Houdt, R., Uytterhoeven, B., . . . Lavigne, R. (2013). A PKS/NRPS/FAS hybrid gene cluster from *Serratia plymuthica* RVH1 encoding the biosynthesis of three broad spectrum, zeamine-related antibiotics. *PLoS ONE*, 8(1), e54143. doi: 10.1371/journal.pone.0054143
- McCutcheon, J. P., & Moran, N. A. (2011). Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol*, 10(1), 13-26. doi: 10.1038/nrmicro2670
- McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., . . . Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J*, 6(3), 610-618. doi: 10.1038/ismej.2011.139
- McMurdie, P. J., & Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One*, 8(4), e61217. doi: 10.1371/journal.pone.0061217
- Meleshko, D., Mohimani, H., Tracanna, V., Hajirasouliha, I., Medema, M. H., Korobeynikov, A., & Pevzner, P. A. (2019). BiosyntheticSPAdes: reconstructing biosynthetic gene clusters from assembly graphs. *Genome Res*, 29(8), 1352-1362. doi: 10.1101/gr.243477.118
- Meunier, L., Tocquin, P., Cornet, L., Sirjacobs, D., Leclere, V., Pupin, M., . . . Baurain, D. (2020). Palantir: a springboard for the analysis of secondary metabolite gene clusters in large-scale genome mining projects. *Bioinformatics*, 36(15), 4345-4347. doi: 10.1093/bioinformatics/btaa517
- Milano, T., Paiardini, A., Grgurina, I., & Pascarella, S. (2013). Type I pyridoxal 5'-phosphate dependent enzymatic domains embedded within multimodular nonribosomal peptide synthetase and polyketide synthase assembly lines. *BMC Struct Biol*, 13(1), 26. doi: 10.1186/1472-6807-13-26
- Milstejn, A., Schneider, J. S., & Brady, S. F. (2014). Mining the Metabiome: Identifying Novel Natural Products from Microbial Communities. *Chem Biol*, 21(9), 1211-1223. doi: 10.1016/j.chembiol.2014.08.006

- Mitchell, C. C., Delaney, D., & Balkcom, K. S. (2005). Cullars Rotation: The South's oldest continuous soil fertility experiment. *Better Crops*, 89(4), 5-9.
- Mitchell, C. C., Delaney, D. P., & Balkcom, K. S. (2011). Centennial of Alabama's Cullars Rotation, the South's oldest, continuous soil fertility experiment.
- Moore, V. C., Strano, M. S., Haroz, E. H., Hauge, R. H., Smalley, R. E., Schmidt, J., & Talmon, Y. (2003). Individually suspended single-walled carbon nanotubes in various surfactants. *Nano letters*, 3(10), 1379-1382.
- Morgan, X. C., Tickle, T. L., Sokol, H., Gevers, D., Devaney, K. L., Ward, D. V., . . . Huttenhower, C. (2012). Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol*, 13(9), R79. doi: 10.1186/gb-2012-13-9-r79
- Murugkar, P. P., Collins, A. J., Chen, T., & Dewhirst, F. E. (2020). Isolation and cultivation of candidate phyla radiation Saccharibacteria (TM7) bacteria in coculture with bacterial hosts. *J Oral Microbiol*, 12(1), 1814666. doi: 10.1080/20002297.2020.1814666
- Nakata, K., Koh, M. M., Tsuchido, T., & Matsumura, Y. (2010). All genomic mutations in the antimicrobial surfactant-resistant mutant, Escherichia coli OW66, are involved in cell resistance to surfactant. *Appl Microbiol Biotechnol*, 87(5), 1895-1905. doi: 10.1007/s00253-010-2638-8
- Nascimento Lemos, L., Manoharan, L., William Mendes, L., Monteiro Venturini, A., Satler Pylro, V., & Tsai, S. M. (2020). Metagenome assembled-genomes reveal similar functional profiles of CPR/Patescibacteria phyla in soils. *Environ Microbiol Rep*, 12(6), 651-655. doi: 10.1111/1758-2229.12880
- Nasrin, S., Ganji, S., Kakirde, K. S., Jacob, M. R., Wang, M., Ravu, R. R., . . . Liles, M. R. (2018). Chloramphenicol Derivatives with Antibacterial Activity Identified by Functional Metagenomics. *J Nat Prod*, 81(6), 1321-1332. doi: 10.1021/acs.jnatprod.7b00903
- Nasseri, S. A., Betschart, L., Opaleva, D., Rahfeld, P., & Withers, S. G. (2018). A Mechanism-Based Approach to Screening Metagenomic Libraries for Discovery of Unconventional Glycosidases. *Angew Chem Int Ed Engl*, 57(35), 11359-11364. doi: 10.1002/anie.201806792
- Navarro-Munoz, J. C., Selem-Mojica, N., Mullooney, M. W., Kautsar, S. A., Tryon, J. H., Parkinson, E. I., . . . Medema, M. H. (2020). A computational framework to explore large-scale biosynthetic diversity. *Nat Chem Biol*, 16(1), 60-68. doi: 10.1038/s41589-019-0400-9
- Nawrocki, E. P., & Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22), 2933-2935. doi: 10.1093/bioinformatics/btt509
- Nayfach, S., Roux, S., Seshadri, R., Udwy, D., Varghese, N., Schulz, F., . . . Eloe-Fadrosh, E. A. (2020). A genomic catalog of Earth's microbiomes. *Nat Biotechnol*. doi: 10.1038/s41587-020-0718-6
- Nel, A., Xia, T., Madler, L., & Li, N. (2006). Toxic Potential of Materials at the Nanolevel. *Science*, 311(5761), 622-627. doi: 10.1126/science.1114397
- Nelson, W. C., & Stegen, J. C. (2015). The reduced genomes of Parcubacteria (OD1) contain signatures of a symbiotic lifestyle. *Front Microbiol*, 6, 713. doi: 10.3389/fmicb.2015.00713
- Nepal, D., Balasubramanian, S., Simonian, A. L., & Davis, V. A. (2008). Strong antimicrobial coatings: single-walled carbon nanotubes armored with biopolymers. *Nano Letters*, 8(7), 1896-1901.
- Newman, D. J., & Cragg, G. M. (2012). Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J Nat Prod*, 75(3), 311-335. doi: 10.1021/np200906s

- Nguyen, S. T. C., Freund, H. L., Kasanjian, J., & Berlemont, R. (2018). Function, distribution, and annotation of characterized cellulases, xylanases, and chitinases from CAZy. *Appl Microbiol Biotechnol*, 102(4), 1629-1637. doi: 10.1007/s00253-018-8778-y
- Nichols, D., Cahoon, N., Trakhtenberg, E. M., Pham, L., Mehta, A., Belanger, A., . . . Epstein, S. S. (2010). Use of ichip for high-throughput in situ cultivation of "uncultivable" microbial species. *Appl Environ Microbiol*, 76(8), 2445-2450. doi: 10.1128/AEM.01754-09
- Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome research*, 27(5), 824-834.
- Nyankima, A. G., Horn, D. W., & Davis, V. A. (2013). Free-Standing Films from Aqueous Dispersions of Lysozyme, Single-Walled Carbon Nanotubes, and Polyvinyl Alcohol. *ACS Macro Letters*, 3(1), 77-79.
- O'Brien, J., & Wright, G. D. (2011). An ecological perspective of microbial secondary metabolism. *Curr Opin Biotechnol*, 22(4), 552-558. doi: 10.1016/j.copbio.2011.03.010
- Ohkuma, M., & Kudo, T. (1996). Phylogenetic diversity of the intestinal bacterial community in the termite *Reticulitermes speratus*. *Appl Environ Microbiol*, 62(2), 461-468. doi: 10.1128/AEM.62.2.461-468.1996
- Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'hara, R., . . . Wagner, H. (2013). Package 'vegan'. *Community ecology package, version*, 2(9), 1-295.
- Orellana, L. H., Chee-Sanford, J. C., Sanford, R. A., Löffler, F. E., & Konstantinidis, K. T. (2018). Year-Round Shotgun Metagenomes Reveal Stable Microbial Communities in Agricultural Soils and Novel Ammonia Oxidizers Responding to Fertilization. *Appl Environ Microbiol*, 84(2). doi: UNSP e01646-17
- Owen, J. G., Charlop-Powers, Z., Smith, A. G., Ternei, M. A., Calle, P. Y., Reddy, B. V., . . . Brady, S. F. (2015). Multiplexed metagenome mining using short DNA sequence tags facilitates targeted discovery of epoxyketone proteasome inhibitors. *Proc Natl Acad Sci U S A*, 112(14), 4221-4226. doi: 10.1073/pnas.1501124112
- Owen, J. G., Reddy, B. V. B., Ternei, M. A., Charlop-Powers, Z., Calle, P. Y., Kim, J. H., & Brady, S. F. (2013). Mapping gene clusters within arrayed metagenomic libraries to expand the structural diversity of biomedically relevant natural products. *Proc Natl Acad Sci U S A*, 110(29), 11797-11802. doi: 10.1073/pnas.1222159110
- Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. *Science*, 276(5313), 734-740. doi: 10.1126/science.276.5313.734
- Pace, N. R. (2009). Mapping the tree of life: progress and prospects. *Microbiol Mol Biol Rev*, 73(4), 565-576. doi: 10.1128/MMBR.00033-09
- Palazzotto, E., & Weber, T. (2018). Omics and multi-omics approaches to study the biosynthesis of secondary metabolites in microorganisms. *Curr Opin Microbiol*, 45, 109-116. doi: 10.1016/j.mib.2018.03.004
- Palenik, B., Ren, Q., Tai, V., & Paulsen, I. T. (2009). Coastal *Synechococcus* metagenome reveals major roles for horizontal gene transfer and plasmids in population diversity. *Environ Microbiol*, 11(2), 349-359. doi: 10.1111/j.1462-2920.2008.01772.x
- Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarszewski, A., Chaumeil, P. A., & Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol*, 36(10), 996-1004. doi: 10.1038/nbt.4229
- Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and

- metagenomes. *Genome research*, 25(7), 1043-1055.
- Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P. A., Woodcroft, B. J., Evans, P. N., . . . Tyson, G. W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol*, 2(11), 1533-1542. doi: 10.1038/s41564-017-0012-7
- Parsley, L. C., Linneman, J., Goode, A. M., Becklund, K., George, I., Goodman, R. M., . . . Liles, M. R. (2011). Polyketide synthase pathways identified from a metagenomic library are derived from soil Acidobacteria. *FEMS Microbiol Ecol*, 78(1), 176-187. doi: 10.1111/j.1574-6941.2011.01122.x
- Paulson, J. N., Stine, O. C., Bravo, H. C., & Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nat Methods*, 10(12), 1200-1202. doi: 10.1038/nmeth.2658
- Portillo, M. C., Leff, J. W., Lauber, C. L., & Fierer, N. (2013). Cell size distributions of soil bacterial and archaeal taxa. *Appl Environ Microbiol*, 79(24), 7610-7617. doi: 10.1128/AEM.02710-13
- Pruesse, E., Peplies, J., & Glockner, F. O. (2012). SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics*, 28(14), 1823-1829. doi: 10.1093/bioinformatics/bts252
- Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., . . . Wang, J. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285), 59-65. doi: 10.1038/nature08821
- Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., . . . Wang, J. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418), 55-60. doi: 10.1038/nature11450
- Quail, M. A., Matthews, L., Sims, S., Lloyd, C., Beasley, H., & Baxter, S. W. (2011). Genomic libraries: I. Construction and screening of fosmid genomic libraries. *Methods Mol Biol*, 772, 37-58. doi: 10.1007/978-1-61779-228-1_3
- Quaiser, A., Ochsenreiter, T., Lanz, C., Schuster, S. C., Treusch, A. H., Eck, J., & Schleper, C. (2003). Acidobacteria form a coherent but highly diverse group within the bacterial domain: evidence from environmental genomics. *Mol Microbiol*, 50(2), 563-575. doi: 10.1046/j.1365-2958.2003.03707.x
- Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., . . . Glöckner, F. O. (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41(D1), D590-D596. doi: 10.1093/nar/gks1219
- Rausch, C., Hoof, I., Weber, T., Wohlleben, W., & Huson, D. H. (2007). Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. *BMC Evol Biol*, 7(1), 78. doi: 10.1186/1471-2148-7-78
- Reddy, B. V., Kallifidas, D., Kim, J. H., Charlop-Powers, Z., Feng, Z., & Brady, S. F. (2012). Natural product biosynthetic gene diversity in geographically distinct soil microbiomes. *Appl Environ Microbiol*, 78(10), 3744-3752. doi: 10.1128/AEM.00102-12
- Rheims, H., Rainey, F., & Stackebrandt, E. (1996). A molecular approach to search for diversity among bacteria in the environment. *Journal of industrial microbiology*, 17(3), 159-169.
- Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J. F., . . . Woyke, T. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 499(7459), 431-437. doi: 10.1038/nature12352
- Rondon, M. R., August, P. R., Bettermann, A. D., Brady, S. F., Grossman, T. H., Liles, M. R., . . .

- Goodman, R. M. (2000). Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl Environ Microbiol*, 66(6), 2541-2547. doi: 10.1128/aem.66.6.2541-2547.2000
- Sabree, Z. L., Bergendahl, V., Liles, M. R., Burgess, R. R., Goodman, R. M., & Handelsman, J. (2006). Identification and characterization of the gene encoding the *Acidobacterium capsulatum* major sigma factor. *Gene*, 376(1), 144-151. doi: 10.1016/j.gene.2006.02.033
- Sanchez, R., Serra, F., Tarraga, J., Medina, I., Carbonell, J., Pulido, L., . . . Dopazo, H. (2011). Phylemon 2.0: a suite of web-tools for molecular evolution, phylogenetics, phylogenomics and hypotheses testing. *Nucleic Acids Res*, 39(Web Server issue), W470-474. doi: 10.1093/nar/gkr408
- Sangwan, N., Xia, F., & Gilbert, J. A. (2016). Recovering complete and draft population genomes from metagenome datasets. *Microbiome*, 4, 8. doi: 10.1186/s40168-016-0154-5
- Santana-Pereira, A. L. R., Sandoval-Powers, M., Monsma, S., Zhou, J., Santos, S. R., Mead, D. A., & Liles, M. R. (2020). Discovery of Novel Biosynthetic Gene Cluster Diversity From a Soil Metagenomic Library. *Front Microbiol*, 11, 585398. doi: 10.3389/fmicb.2020.585398
- Schatz, A., Bugle, E., & Waksman, S. A. (1944). Streptomycin, a Substance Exhibiting Antibiotic Activity Against Gram-Positive and Gram-Negative Bacteria.*. *Proceedings of the Society for Experimental Biology and Medicine*, 55(1), 66-69.
- Schloss, P. D. (2020). Reintroducing mothur: 10 Years Later. *Appl Environ Microbiol*, 86(2). doi: 10.1128/AEM.02343-19
- Schloss, P. D., & Handelsman, J. (2003). Biotechnological prospects from metagenomics. *Current opinion in biotechnology*, 14(3), 303-310.
- Schloss, P. D., & Handelsman, J. (2004). Status of the microbial census. *Microbiol Mol Biol Rev*, 68(4), 686-691. doi: 10.1128/MMBR.68.4.686-691.2004
- Schoenborn, L., Yates, P. S., Grinton, B. E., Hugenholtz, P., & Janssen, P. H. (2004). Liquid serial dilution is inferior to solid media for isolation of cultures representative of the phylum-level diversity of soil bacteria. *Appl Environ Microbiol*, 70(7), 4363-4366. doi: 10.1128/AEM.70.7.4363-4366.2004
- Schöler, A., Jacquiod, S., Vestergaard, G., Schulz, S., & Schlöter, M. (2017). Analysis of soil microbial communities based on amplicon sequencing of marker genes: Springer.
- Sharon, I., & Banfield, J. F. (2013). Microbiology. Genomes from metagenomics. *Science*, 342(6162), 1057-1058. doi: 10.1126/science.1247023
- Sharpton, T. J. (2014). An introduction to the analysis of shotgun metagenomic data. *Front Plant Sci*, 5, 209. doi: 10.3389/fpls.2014.00209
- Sharrar, A. M., Crits-Christoph, A., Meheust, R., Diamond, S., Starr, E. P., & Banfield, J. F. (2020). Bacterial Secondary Metabolite Biosynthetic Potential in Soil Varies with Phylum, Depth, and Vegetation Type. *mBio*, 11(3). doi: 10.1128/mBio.00416-20
- Shen, B. (2003). Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. *Curr Opin Chem Biol*, 7(2), 285-295. doi: 10.1016/S1367-5931(03)00020-6
- Shizuya, H., Birren, B., Kim, U. J., Mancino, V., Slepak, T., Tachiiri, Y., & Simon, M. (1992). Cloning and Stable Maintenance of 300-Kilobase-Pair Fragments of Human DNA in *Escherichia-Coli* Using an F-Factor-Based Vector. *Proceedings of the National Academy of Sciences of the United States of America*, 89(18), 8794-8797. doi: DOI 10.1073/pnas.89.18.8794
- Sieber, C. M. K., Probst, A. J., Sharrar, A., Thomas, B. C., Hess, M., Tringe, S. G., & Banfield, J.

- F. (2018). Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol*, 3(7), 836-843. doi: 10.1038/s41564-018-0171-1
- Siezen, R. J., & Khayatt, B. I. (2008). Natural products genomics. *Microb Biotechnol*, 1(4), 275-282. doi: 10.1111/j.1751-7915.2008.00044.x
- Simmons, S. L., Dibartolo, G., Deneff, V. J., Goltsman, D. S., Thelen, M. P., & Banfield, J. F. (2008). Population genomic analysis of strain variation in *Leptospirillum* group II bacteria involved in acid mine drainage formation. *PLoS Biol*, 6(7), e177. doi: 10.1371/journal.pbio.0060177
- Sloan, A. W. N., Santana-Pereira, A. L. R., Goswami, J., Liles, M. R., & Davis, V. A. (2017). Single-Walled Carbon Nanotube Dispersion in Tryptic Soy Broth. *ACS Macro Letters*, 6(11), 1228-1231. doi: 10.1021/acsmacrolett.7b00656
- Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L., & Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27(3), 431-432. doi: 10.1093/bioinformatics/btq675
- Solden, L., Lloyd, K., & Wrighton, K. (2016). The bright side of microbial dark matter: lessons learned from the uncultivated majority. *Curr Opin Microbiol*, 31, 217-226. doi: 10.1016/j.mib.2016.04.020
- Soliman, T., Yang, S. Y., Yamazaki, T., & Jenke-Kodama, H. (2017). Profiling soil microbial communities with next-generation sequencing: the influence of DNA kit selection and technician technical expertise. *PeerJ*, 5. doi: Artn E417810.7717/Peerj.4178
- Staley, J. T., & Konopka, A. (1985). Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu Rev Microbiol*, 39, 321-346. doi: 10.1146/annurev.mi.39.100185.001541
- Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9), 1312-1313. doi: 10.1093/bioinformatics/btu033
- Starr, E. P., Shi, S., Blazewicz, S. J., Probst, A. J., Herman, D. J., Firestone, M. K., & Banfield, J. F. (2018). Stable isotope informed genome-resolved metagenomics reveals that *Saccharibacteria* utilize microbially-processed plant-derived carbon. *Microbiome*, 6(1), 122. doi: 10.1186/s40168-018-0499-z
- Staunton, J., & Weissman, K. J. (2001). Polyketide biosynthesis: a millennium review. *Nat Prod Rep*, 18(4), 380-416. doi: 10.1039/a909079g
- Steinegger, M., & Soding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol*, 35(11), 1026-1028. doi: 10.1038/nbt.3988
- Streit, W. R., & Schmitz, R. A. (2004). Metagenomics--the key to the uncultured microbes. *Curr Opin Microbiol*, 7(5), 492-498. doi: 10.1016/j.mib.2004.08.002
- Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., . . . Bork, P. (2015). Ocean plankton. Structure and function of the global ocean microbiome. *Science*, 348(6237), 1261359. doi: 10.1126/science.1261359
- Sundlov, J. A., Shi, C., Wilson, D. J., Aldrich, C. C., & Gulick, A. M. (2012). Structural and functional investigation of the intermolecular interaction between NRPS adenylation and carrier protein domains. *Chem Biol*, 19(2), 188-198. doi: 10.1016/j.chembiol.2011.11.013
- Taubert, M., Stahly, J., Kolb, S., & Kusel, K. (2019). Divergent microbial communities in groundwater and overlying soils exhibit functional redundancy for plant-polysaccharide degradation. *PLoS One*, 14(3), e0212937. doi: 10.1371/journal.pone.0212937
- Thies, S., Rausch, S. C., Kovacic, F., Schmidt-Thaler, A., Wilhelm, S., Rosenau, F., . . . Jaeger, K.

- E. (2016). Metagenomic discovery of novel enzymes and biosurfactants in a slaughterhouse biofilm microbial community. *Sci Rep*, 6. doi: Artn 2703510.1038/Srep27035
- Torsvik, V., & Ovreas, L. (2002). Microbial diversity and function in soil: from genes to ecosystems. *Curr Opin Microbiol*, 5(3), 240-245.
- Trevors, J. T. (2012). Can dead bacterial cells be defined and are genes expressed after cell death? *Journal of Microbiological Methods*, 90(1), 25-28. doi: 10.1016/j.mimet.2012.04.004
- Tringe, S. G., & Hugenholtz, P. (2008). A renaissance for the pioneering 16S rRNA gene. *Curr Opin Microbiol*, 11(5), 442-446. doi: 10.1016/j.mib.2008.09.011
- Tulp, M., & Bohlin, L. (2005). Rediscovery of known natural compounds: nuisance or goldmine? *Bioorg Med Chem*, 13(17), 5274-5282. doi: 10.1016/j.bmc.2005.05.067
- Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., & Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444(7122), 1027-1031. doi: 10.1038/nature05414
- Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., . . . Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978), 37-43. doi: 10.1038/nature02340
- Uversky, V. N., Wohlkönig, A., Huet, J., Looze, Y., & Wintjens, R. (2010). Structural Relationships in the Lysozyme Superfamily: Significant Evidence for Glycoside Hydrolase Signature Motifs. *PLoS One*, 5(11), e15388. doi: 10.1371/journal.pone.0015388
- van Kessel, M. A., Speth, D. R., Albertsen, M., Nielsen, P. H., Op den Camp, H. J., Kartal, B., . . . Lucker, S. (2015). Complete nitrification by a single microorganism. *Nature*, 528(7583), 555-559. doi: 10.1038/nature16459
- Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., . . . Smith, H. O. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304(5667), 66-74. doi: 10.1126/science.1093857
- Wagner, D., Kobabe, S., & Liebner, S. (2009). Bacterial community structure and carbon turnover in permafrost-affected soils of the Lena Delta, northeastern Siberia. *Can J Microbiol*, 55(1), 73-83. doi: 10.1139/W08-121
- Wang, H., Fewer, D. P., Holm, L., Rouhiainen, L., & Sivonen, K. (2014). Atlas of nonribosomal peptide and polyketide biosynthetic pathways reveals common occurrence of nonmodular enzymes. *Proc Natl Acad Sci U S A*, 111(25), 9259-9264. doi: 10.1073/pnas.1401734111
- Wang, K., Li, G., Yu, S. Q., Zhang, C. T., & Liu, Y. H. (2010). A novel metagenome-derived beta-galactosidase: gene cloning, overexpression, purification and characterization. *Appl Microbiol Biotechnol*, 88(1), 155-165. doi: 10.1007/s00253-010-2744-7
- Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol*, 73(16), 5261-5267. doi: 10.1128/AEM.00062-07
- Wawrik, B., Kerkhof, L., Zylstra, G. J., & Kukor, J. J. (2005). Identification of unique type II polyketide synthase genes in soil. *Appl Environ Microbiol*, 71(5), 2232-2238. doi: 10.1128/AEM.71.5.2232-2238.2005
- Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H. U., Brucoleri, R., . . . Medema, M. H. (2015). antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res*, 43(W1), W237-243. doi: 10.1093/nar/gkv437
- Wegley, L., Edwards, R., Rodriguez-Brito, B., Liu, H., & Rohwer, F. (2007). Metagenomic

- analysis of the microbial community associated with the coral *Porites astreoides*. *Environ Microbiol*, 9(11), 2707-2719. doi: 10.1111/j.1462-2920.2007.01383.x
- Weisburg, W. G., Barns, S. M., Pelletier, D. A., & Lane, D. J. (1991). 16S ribosomal DNA amplification for phylogenetic study. *J Bacteriol*, 173(2), 697-703.
- Westmann, C. A., Alves, L. F., Silva-Rocha, R., & Guazzaroni, M. E. (2018). Mining Novel Constitutive Promoter Elements in Soil Metagenomic Libraries in *Escherichia coli*. *Front Microbiol*, 9, 1344. doi: 10.3389/fmicb.2018.01344
- Wick, P., Manser, P., Limbach, L. K., Dettlaff-Weglikowska, U., Krumeich, F., Roth, S., . . . Bruinink, A. (2007). The degree and kind of agglomeration affect carbon nanotube cytotoxicity. *Toxicology Letters*, 168(2), 121-131.
- Wickham, H. (2011). ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(2), 180-185.
- Wild, J., Hradecna, Z., & Szybalski, W. (2002). Conditionally amplifiable BACs: Switching from single-copy to high-copy vectors and genomic clones. *Genome Res*, 12(9), 1434-1444. doi: 10.1101/gr.130502
- Wild, J., & Szybalski, W. (2004). Copy-control tightly regulated expression vectors based on pBAC/oriV *Recombinant Gene Expression* (pp. 155-167): Springer.
- Wilkins, D. (2019). gggenes: draw gene arrow maps in 'ggplot2'. R package version 0.4. 0.
- Willyard, C. (2017). The drug-resistant bacteria that pose the greatest health threats. *Nature News*, 543(7643), 15.
- Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A*, 74(11), 5088-5090. doi: 10.1073/pnas.74.11.5088
- Woese, C. R., Kandler, O., & Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci U S A*, 87(12), 4576-4579. doi: 10.1073/pnas.87.12.4576
- Wohlleben, W., Mast, Y., Stegmann, E., & Ziemert, N. (2016). Antibiotic drug discovery. *Microb Biotechnol*, 9(5), 541-548. doi: 10.1111/1751-7915.12388
- Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, 15(3), R46.
- Woodcroft, B. J., Singleton, C. M., Boyd, J. A., Evans, P. N., Emerson, J. B., Zayed, A. A. F., . . . Tyson, G. W. (2018). Genome-centric view of carbon processing in thawing permafrost. *Nature*, 560(7716), 49-54. doi: 10.1038/s41586-018-0338-1
- Wrighton, K. C., Castelle, C. J., Varaljay, V. A., Satagopan, S., Brown, C. T., Wilkins, M. J., . . . Banfield, J. F. (2016). RubisCO of a nucleoside pathway known from Archaea is found in diverse uncultivated phyla in bacteria. *ISME J*, 10(11), 2702-2714. doi: 10.1038/ismej.2016.53
- Wrighton, K. C., Castelle, C. J., Wilkins, M. J., Hug, L. A., Sharon, I., Thomas, B. C., . . . Banfield, J. F. (2014). Metabolic interdependencies between phylogenetically novel fermenters and respiratory organisms in an unconfined aquifer. *ISME J*, 8(7), 1452-1463. doi: 10.1038/ismej.2013.249
- Wrighton, K. C., Thomas, B. C., Sharon, I., Miller, C. S., Castelle, C. J., VerBerkmoes, N. C., . . . Banfield, J. F. (2012). Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science*, 337(6102), 1661-1665. doi: 10.1126/science.1224041
- Wu, C., Asakawa, S., Shimizu, N., Kawasaki, S., & Yasukochi, Y. (1999). Construction and characterization of bacterial artificial chromosome libraries from the silkworm, *Bombyx*

- mori. *Mol Gen Genet*, 261(4-5), 698-706. doi: 10.1007/s004380050013
- Wu, Y. W., Simmons, B. A., & Singer, S. W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, 32(4), 605-607. doi: 10.1093/bioinformatics/btv638
- Xia, W., Lu, H., Xia, M., Cui, Y., Bai, Y., Qian, L., . . . Kelly, R. M. (2016). A Novel Glycoside Hydrolase Family 113 Endo- β -1,4-Mannanase from Alicyclobacillus sp. Strain A4 and Insight into the Substrate Recognition and Catalytic Mechanism of This Family. *Appl Environ Microbiol*, 82(9), 2718-2727. doi: 10.1128/aem.04071-15
- Yang, C., Mamouni, J., Tang, Y., & Yang, L. (2010). Antimicrobial Activity of Single-Walled Carbon Nanotubes: Length Effect. *Langmuir*, 26(20), 16013-16019. doi: 10.1021/la103110g
- Yarza, P., Yilmaz, P., Pruesse, E., Glockner, F. O., Ludwig, W., Schleifer, K. H., . . . Rossello-Mora, R. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol*, 12(9), 635-645. doi: 10.1038/nrmicro3330
- Yeates, C., Holmes, A. J., & Gillings, M. R. (2000). Novel forms of ring-hydroxylating dioxygenases are widespread in pristine and contaminated soils. *Environmental microbiology*, 2(6), 644-653.
- Yuan, C., Lei, J., Cole, J., & Sun, Y. (2015). Reconstructing 16S rRNA genes in metagenomic data. *Bioinformatics*, 31(12), i35-43. doi: 10.1093/bioinformatics/btv231
- Yuan, Q., Xueyu, P., Wei, J., Lianqing, C., & Zhilin, Y. (2018). Comparison of Four Extraction Methods of Soil Microbiome in Poplar Plantation. *Scientia Silvae Sinicae*, 54(9), 169-176.
- Zaremba-Niedzwiedzka, K., Caceres, E. F., Saw, J. H., Backstrom, D., Juzokaite, L., Vancaester, E., . . . Ettema, T. J. (2017). Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature*, 541(7637), 353-358. doi: 10.1038/nature21031
- Zhang, H., Yohe, T., Huang, L., Entwistle, S., Wu, P., Yang, Z., . . . Yin, Y. (2018). dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res*, 46(W1), W95-W101. doi: 10.1093/nar/gky418
- Zhao, C., Fu, S., Mathew, R. P., Lawrence, K. S., & Feng, Y. (2015). Soil microbial community structure and activity in a 100-year-old fertilization and crop rotation experiment. *Journal of Plant Ecology*, 8(6), 623-632.
- Zhu, Q., Mai, U., Pfeiffer, W., Janssen, S., Asnicar, F., Sanders, J. G., . . . Knight, R. (2019). Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat Commun*, 10(1), 5477. doi: 10.1038/s41467-019-13443-4
- Zhu, Y., & Li, W. (2008). Cytotoxicity of carbon nanotubes. *Science in China Series B: Chemistry*, 51(11), 1021-1029.
- Zimmermann, K., Engeser, M., Blunt, J. W., Munro, M. H., & Piel, J. (2009). Pederin-type pathways of uncultivated bacterial symbionts: analysis of o-methyltransferases and generation of a biosynthetic hybrid. *J Am Chem Soc*, 131(8), 2780-2781. doi: 10.1021/ja808889k

Appendices

Supplemental Table 1. Validated clones containing PKS and/ or NRPS pathways.

Clone	Contig	BLmean	Identified	Domains	Insert Length	BGC Length	GC (%)	Set	Type	Truncated	Domain Ident	Dom Div	DivT
P01H15	P01A1.21	2.59749	NGS	15	64339	4,722	71.3	Set 1	PKS-NRPS	Complete	27.70	72.30	94.84
P01N03	P01B1.33 P01B1.73	2.5038	NGS	21	40763	41,349	58.1	Set 1	NRPS	Complete	24.41	75.59	99.15
P02O05	P02A1.12	2.715957	NGS	9	106305	13,762	56.3	Set 1	NRPS	Complete	27.94	72.06	94.53
P04H02	P04B1.15 P04B1.67	1.374	NGS	72	140742	97,614	54.8	Set 1	PKS-NRPS	Complete	60	40.00	52.47
P05D13	P05B1.29	2.631503	NGS	35	71074	51,314	54.7	Set 1	NRPS	Complete	26.99	73.01	95.77
P05G19	P05A1.676	3.38761	NGS	20	90869	39,049	54.2	Set 1	NRPS	Complete	24.84	75.16	98.59
P06G05	P06A1.1	2.8049	NGS	18	102475	68,150	56	Set 1	PKS-NRPS	Complete	52.69	47.31	62.06
P06G11	P06A1.285	1.9081	NGS	52	129397	52,348	55.8	Set 1	PKS-NRPS	Complete	28.95	71.05	93.20
P07D12	P07B2.21	2.805976	NGS	20	115226	24,317	57.2	Set 1	NRPS	Complete	31.72	68.28	89.57
P07H07	P07B1.72	2.750188	NGS	20	30585	37,239	54.2	Set 1	NRPS	Complete	28.01	71.99	94.44
P07H14	P07B2.159 P07B2.164	2.688972	NGS	20	130799	114,768	61.9	Set 1	NRPS	Complete	34.45	65.55	85.99
P07O09	P07A1.56	1.67399	NGS	16	108907	87,047	55.6	Set 1	PKS-NRPS	Complete	58.64	41.36	54.25
P08A19	P08A1.2	1.7239	NGS	38	96916	43,697	56	Set 1	PKS-NRPS	Complete	55.84	44.16	57.93
P08C17	P08A2.46 P08A1.540 P08A1.626 P08A1.683	3.096427	NGS	33	124248	6,693	55.2	Set 1	NRPS	Complete	67.09	32.91	43.18
P08J17	P08A2.29	2.714854	NGS	34	106688	42,699	54.1	Set 1	NRPS	Complete	57.13	42.87	56.24
P08M24	P08A2.142	2.85275	NGS	57	133951	43,032	58.2	Set 1	PKS-NRPS	Complete	61.46	38.54	50.56
P09L06	P09B2.14	1.899525	NGS	27	75828	26,416	54.8	Set 1	PKS-NRPS	Complete	61.22	38.78	50.87
P09M22	P09A2.103	2.558026	NGS	46	99134	1,624	54.7	Set 1	PKS-NRPS	Complete	58.84	41.16	53.99
P09P02	P09B2.417 P09B2.194 P09B2.157	2.667553	NGS	19	143164	1,040	70.4	Set 1	NRPS	Complete	58.17	41.83	54.87
P11H15	P11B1.15	4.09459	NGS	25	77085	66,479	54.8	Set 2	NRPS	Complete	54.90	45.10	59.16
P11I24	P11A2.58	1.700657	NGS	45	68208	57,478	54	Set 2	PKS-NRPS	Complete	64.22	35.78	46.94
P11J18	P11B2.361 P11B2.716 P11B2.149 P11B2.133 P11B2.172 P11B2.474	2.727884	NGS	11	42994	87,614	50.6	Set 2	NRPS	Complete	55.59	44.41	58.26
P11N21	P11B1.351 P11B1.159	3.406645	NGS	11	49818	31,688	50.8	Set 2	NRPS	Complete	65.23	34.77	45.61
P15C22	P15A1.77 P15A2.23	3.071734	NGS	18	83712	39,703	62.8	Set 2	NRPS	Complete	63.31	36.69	48.13
P17G08	P17A2.40	2.65185	NGS	11	65444	2,884	56.5	Set 2	NRPS	Complete	52.67	47.33	62.08
P19N16	P19B2.404	2.59971	NGS	20	60112	2,302	58.3	Set 2	NRPS	Complete	58.39	41.61	54.59
P20P17	P20B1.264 P20B1.518	2.589025	NGS	11	26094	49,110	63.6	Set 2	NRPS	Complete	54.24	45.76	60.03
P22K24	P22A2.182	1.8373	NGS	34	105147	149,469	54	Set 3	NRPS	Complete	56.29	43.71	57.34
P22P12	P22B2.239	2.652798	NGS	13	44041	3,014	53.2	Set 3	NRPS	Complete	59.41	40.59	53.24
P23K14	P20B1.242	2.51521	NGS	30	43775	48,479	71.7	Set 3	PKS-NRPS	Complete	59.63	40.37	52.95

Clone	Contig	BLmean	Identified	Domains	Insert Length	BGC Length	GC (%)	Set	Type	Truncated	Domain Ident	Dom Div	DivT
P24C17	P24A1.31	2.667338	NGS	26	79331	77,061	53.6	Set 3	NRPS	Complete	51.55	48.45	63.55
P25M03	P25A1.10	2.607858	NGS	29	87165	71,332	54.8	Set 3	NRPS	Complete	56.34	43.66	57.28
P26P07	P26B1.150 P26B1.124 P26B1.102 P26B1.83 P26B1.192	2.66871	NGS	26	37748	25,461	58.4	Set 3	NRPS	Complete	55.06	44.94	58.96
P27K06	P27A2.75	1.9834	NGS	41	139647	20,072	57	Set 3	NRPS	Complete	58.92	41.08	53.89
P27O06	P27A2.39	2.733196	NGS	41	139647	2,970	57.4	Set 3	NRPS	Complete	56.45	43.55	57.13
P27O09	P27A1.38	1.60135	NGS	49	132241	20,272	54.8	Set 3	PKS-NRPS	Complete	56.00	44.00	57.72
P27P01	P27B1.24	2.588137	NGS	15	113937	10,545	52.9	Set 3	NRPS	Complete	57.04	42.97	56.36
P28L15	P28B1.439 P28B1.133 P28B1.257 P28B1.63 P28B1.237 P28B1.390	2.837996	NGS	20	110926	23,644	55.7	Set 3	NRPS	Complete	53.37	46.63	61.17
P28O04	P28A2.202	1.64563	NGS	30	85938	77,087	61	Set 3	NRPS	Complete	56.14	43.86	57.53
P29F04	P29B2.3	2.672458	NGS	30	133026	48,625	53.1	Set 3	NRPS	Complete	58.91	41.09	53.90
P29O18	P29A2.15	1.80395	NGS	77	114397	117,998	56.6	Set 3	PKS-NRPS	Complete	54.79	45.21	59.30
P35G10	P35A2.492 P35A2.22	2.596979	NGS	42	69371	5,532	56.7	Set 4	NRPS	Complete	55.83	44.17	57.94
P39F17	P39B1.116	2.641278	NGS	13	25361	50,960	58.6	Set 4	NRPS	Complete	55.45	44.55	58.44
P39I02	P39A1.8	1.79775	NGS	15	135991	70,579	57.6	Set 4	PKS-NRPS	Complete	57.94	42.06	55.17
P39I17	P39A1.32 P39B1.24	2.844745	NGS	42	92944	73,906	64.7	Set 4	NRPS	Complete	53.72	46.28	60.72
P42D19	P42.1469 P42.1108 P42.2018 P42.1840	2.888025	NGS	20	95303	112,607	55	Set 5	NRPS	Complete	51.27	48.73	63.93
P42K11	P42.2025	3.30237	NGS	30	116916	30,675	56.5	Set 5	NRPS	Complete	54.19	45.81	60.09
P42P24	P42.2475	3.95117	NGS	13	19065	65,394	53.9	Set 5	NRPS	Complete	51.40	48.60	63.75
P43D10	P43.1515	2.558	NGS	36	129280	99,228	53.9	Set 5	NRPS	Complete	57.25	42.75	56.08
P44L04	P44.129	1.1188	NGS	23	53866	101,858	56	Set 5	PKS-NRPS	Complete	56.14	43.86	57.54
P45H03	P45.341	2.668237	NGS	19	57229	72,087	59.2	Set 5	NRPS	Complete	52.39	47.61	62.45
P48M16	P48.926 P48.1895 P48.1710 P48.988 P48.831	2.738415	NGS	26	123238	14,386	53.9	Set 5	NRPS	Complete	56.64	43.36	56.88
P49O09	P49.2215 P49.1560 P49.2119	2.70316	NGS	57	124799	2,251	71.9	Set 5	NRPS	Complete	53.84	46.16	60.56
P50O17	P50.1556 P50.1984	2.75925	NGS	67	99486	105,370	42.3	Set 5	NRPS	Complete	54.06	45.94	60.27
P11I06	P11A2.14	2.606004	NGS	29	106068	69,570	52.8	Set 2	NRPS	Complete	71	29.00	38.04
P15P18	P15B2.5	2.600735	NGS	25	81084	70,052	54.6	Set 2	NRPS	Complete	58	42.00	55.10
P16O06	P17B2.38	2.754474	NGS	17	67482	42,261	54.6	Set 2	NRPS	Complete	56	44.00	57.72
P20H02	P20A2.2	1.66405	NGS	9	134998	30,345	57	Set 2	PKS-NRPS	Complete	83	17.00	22.30
P21E03	P21A1.38	2.818178	NGS	15	91188	53,798	54.8	Set 3	NRPS	Complete	57	43.00	56.41
P28O18	P28A2.23	0.950076	NGS	30	85938		61	Set 3	NRPS	Complete	55	45.00	59.03
P32P04	P32B2.12	1.2293	NGS	12	45372	44,670	61	Set 4	PKS-NRPS	Complete	70	30.00	39.35
P33I07	P33A1.40	2.031962	NGS	30	91908	53,131	40	Set 4	PKS	Complete	44	56.00	73.46
P35K06	P35A2.331	1.81645	NGS	13	142359	50,486	66	Set 4	PKS-NRPS	Complete	62	38.00	49.85
P36D09	P36B1.54 P36B1.60	2.83628	NGS	30	76453	75,565	58.8	Set 4	NRPS	Complete	55	45.00	59.03

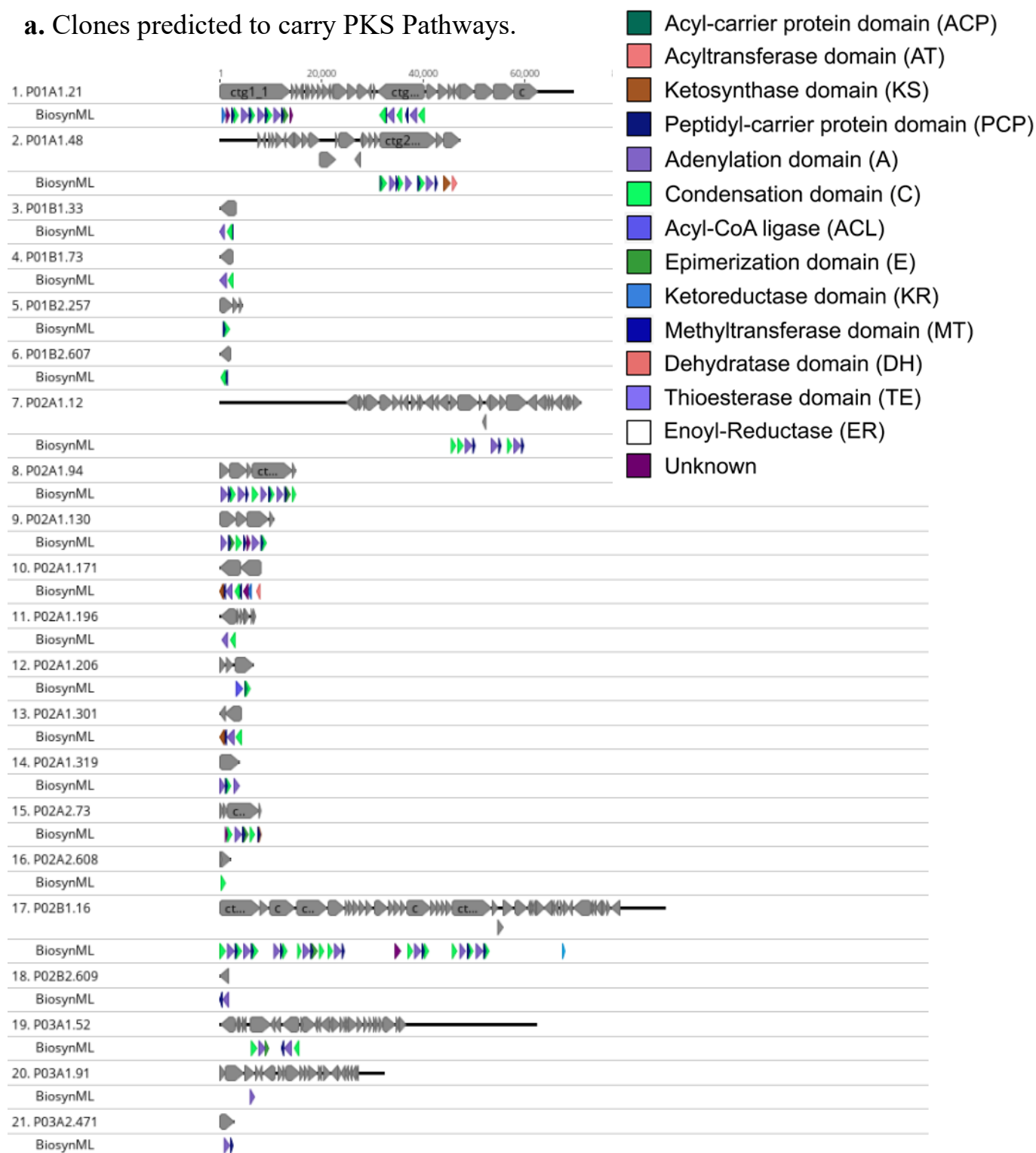
Clone	Contig	BLmean	Identified	Domains	Insert Length	BGC Length	GC (%)	Set	Type	Truncated	Domain Ident	Dom Div	DivT
P01A07	P01A1.48	1.4943	NGS-PCR	33	128761	96,149	69	Set 1	PKS-NRPS	Complete	38.51	61.49	80.67
P07H03	P07B1.70	1.103	NGS-PCR	16	31569	79,592	70.7	Set 1	PKS-NRPS	Complete	43.77	56.23	73.76
P09D05	P09B1.149	1.1741	NGS-PCR	39	63459	1,116	59	Set 1	PKS-NRPS	Complete	62.38	37.62	49.36
P09I20	P09B2.14	1.899525	NGS-PCR	21	30553	73,864	59.6	Set 1	NRPS	Complete	56.71	43.29	56.79
P10B06	P10B2.42	1.87925	NGS-PCR	16	70533	49,433	62	Set 1	PKS	Complete	69.23	30.77	40.37
P11O04	P11A2.44	1.32454	NGS-PCR	31	77067	74,462	67.7	Set 2	PKS-NRPS	Complete	58.19	41.81	54.85
P11P20	P11B2.532	1.808	NGS-PCR	59	150553	31,118	65	Set 2	PKS-NRPS	Complete	65.98	34.02	44.62
P17F15	P17B1.340	1.1501	NGS-PCR	10	71189	88,261	70	Set 2	PKS-NRPS	Complete	64.33	35.67	46.79
P37L16	P37L16	1.614667	NGS-PCR	17	57334	4,799	63.8	Set 4	PKS-NRPS	Complete	54.39	45.61	59.83
P38G02	P38G02	1.9029	NGS-PCR	24	91476	3,192	54.6	Set 4	PKS-NRPS	Complete	56.40	43.60	57.19
P45H08	P45.362	1.6673	NGS-PCR	38	146800	70,241	53	Set 5	PKS-NRPS	Complete	59.02	40.98	53.76
P02K21	P02A1.196	2.55024	NGS	2	107433	14,523	65.5	Set 1	NRPS	Truncated	23.78	76.22	99.99
P03C03	P03A1.52	2.82462	NGS	6	62165	118,628	54.6	Set 1	NRPS	Truncated	23.77	76.23	100.00
P03G12	P03A2.665	2.37516	NGS	5	90962	104,170	65.1	Set 1	NRPS	Truncated	28.06	71.94	94.37
P03G13	P03A1.91	2.90367	NGS	1	123646	1,510	51.6	Set 1	NRPS	Truncated	64.59	35.41	46.45
P04F05	P04B1.68	2.896505	NGS	8	129359	98,035	71.1	Set 1	NRPS	Truncated	28.59	71.41	93.68
P04O01	P04A1.29	1.1822	NGS	8	53001	2,609	57.7	Set 1	PKS	Truncated	51.06	48.94	64.20
P05D19	P05A1.115 P05A1.255	2.652045	NGS	5	103884	153,088	52.1	Set 1	NRPS	Truncated	26.35	73.65	96.62
P05K23	P05A1.752	3.65239	NGS	4	108335	1,278	54.6	Set 1	NRPS	Truncated	24.32	75.68	99.27
P05P18	P05B2.4	2.68926	NGS	5	121566	58,216	57	Set 1	NRPS	Truncated	47.34	52.66	69.08
P06A05	P06A1.18	1.85877	NGS	9	87898	83,133	57	Set 1	PKS-NRPS	Truncated	38.07	61.93	81.24
P07E15	P07A1.117	2.88421	NGS	10	133243	66,517	54.1	Set 1	NRPS	Truncated	26.01	73.99	97.07
P07L16	P07B2.36	2.562933	NGS	8	70028	93,770	65	Set 1	PKS	Truncated	67.76	32.24	42.29
P07O04	P07A2.48	0.972874	NGS	10	61286	106,177	66	Set 1	PKS-NRPS	Truncated	81.67	18.33	24.05
P07P02	P07B2.19	2.437705	NGS	6	83740	145,642	66.1	Set 1	NRPS	Truncated	60.60	39.40	51.69
P09G10	P09A2.159	2.64345	NGS	11	112164	15,283	67.7	Set 1	PKS-NRPS	Truncated	50.24	49.76	65.27
P09J07	P09A1.116 P09B1.70	2.86057	NGS	5	46842	80,512	70.5	Set 1	NRPS	Truncated	60.80	39.20	51.42
P11I13	P11A1.18	2.02242	NGS	11	110052	78,569	63	Set 2	PKS	Truncated	47.14	52.86	69.34
P19O08	P19A2.56	1.228433	NGS	11	93369	58,388	41.9	Set 2	PKS	Truncated	63.87	36.13	47.39
P20F15	P20B1.242 P20B1.293	2.721065	NGS	30	75278	53,104	68.5	Set 2	PKS-NRPS	Truncated	59.63	40.37	52.95
P24H02	P24B2.16	1.98759	NGS	6	87129	67,260	47.8	Set 3	PKS	Truncated	62.06	37.94	49.77
P24J10	P24B2.83	2.11339	NGS	3	40322	24,265	67.7	Set 3	NRPS	Truncated	58.61	41.39	54.29
P24L04	P24B2.4	3.008657	NGS	10	142547	1,664	41.7	Set 3	NRPS	Truncated	55.41	44.59	58.49
P24L06	P24B2.26	1.79834	NGS	8	97272	97,771	40	Set 3	PKS	Truncated	59.07	40.93	53.69
P25I09	P25A1.149	2.5978	NGS	8	70076	4,483	55.5	Set 3	NRPS	Truncated	55.03	44.97	58.99
P27L24	P27B2.506	2.71706	NGS	12	112758	5,913	55.1	Set 3	NRPS	Truncated	54.73	45.27	59.38
P28I16	P28A2.16	2.1338	NGS	10	74990	38,208	55	Set 3	PKS-NRPS	Truncated	64.65	35.35	46.37
P29G19	P29A1.75	2.5445	NGS	5	63402	2,653	63.9	Set 3	NRPS	Truncated	53.79	46.21	60.62
P32A16	P32A2.546	2.66837	NGS	9	59698	83,726	68.7	Set 4	NRPS	Truncated	55.84	44.16	57.93
P33A17	P33A1.19	2.755103	NGS	8	114632	34,816	54.2	Set 4	NRPS	Truncated	54.59	45.41	59.57

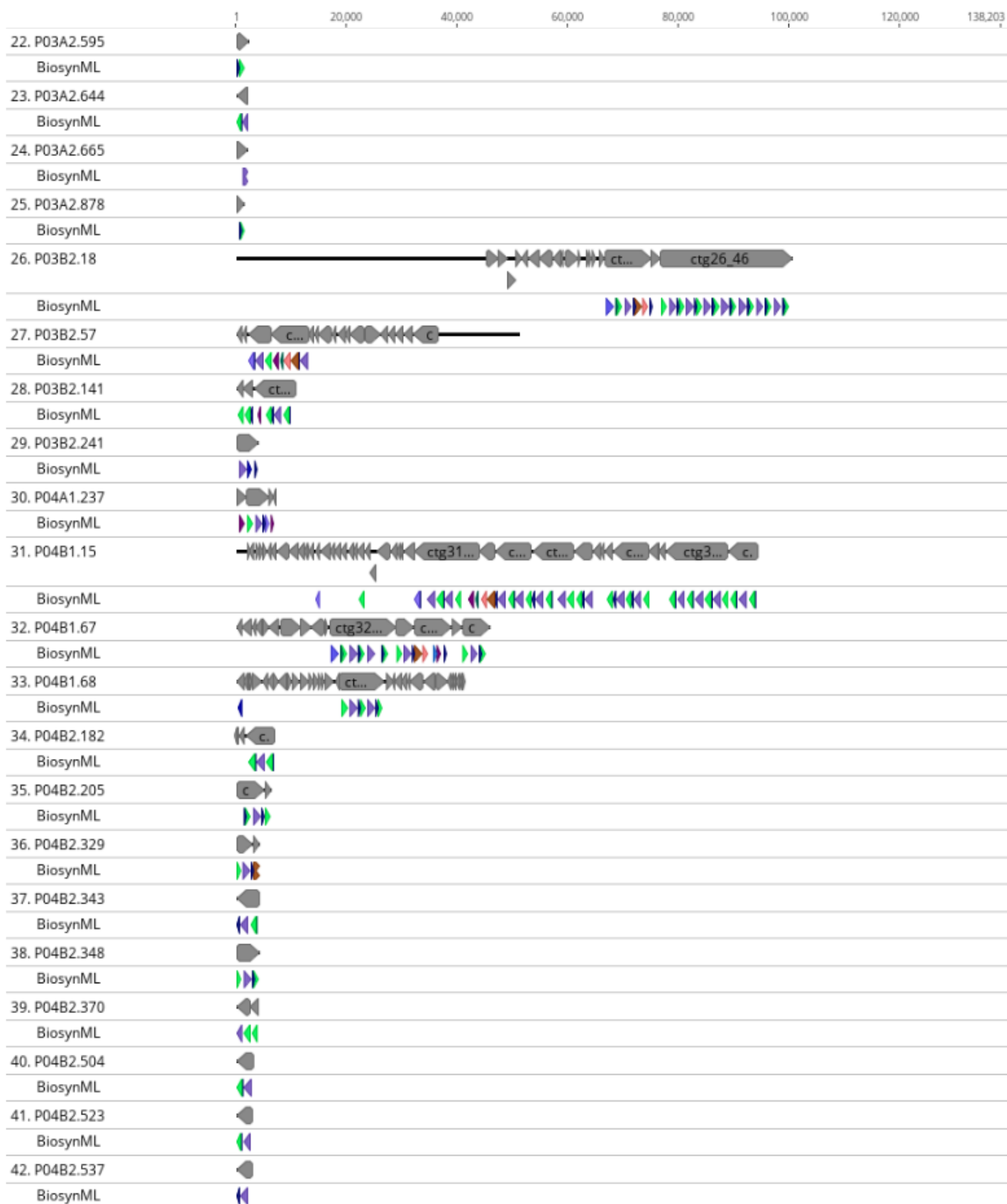
Clone	Contig	BLmean	Identified	Domains	Insert Length	BGC Length	GC (%)	Set	Type	Truncated	Domain Ident	Dom Div	DivT
P33E15	P33A1.78 P33A1.142	2.141635	NGS	5	125482	36,538	63.9	Set 4	PKS	Truncated	83.99	16.01	21.01
P38F23	P38B1.47	2.8209	NGS	9	35834	3,270	64.8	Set 4	NRPS	Truncated	55.15	44.85	58.84
P38H14	P38B2.118	2.41933	NGS	6	106831	32,332	53.8	Set 4	NRPS	Truncated	52.42	47.58	62.41
P38N20	P38B2.22	2.57864	NGS	5	96293	5,942	54	Set 4	NRPS	Truncated	47.90	52.10	68.34
P39E24	P39A2.5 P39A1.8	2.648825	NGS	5	57207	5,643	62	Set 4	PKS-NRPS	Truncated	57.73	42.27	55.45
P39I13	P39A1.70	3.20332	NGS	5	51616	26,123	58.1	Set 4	NRPS	Truncated	50.75	49.25	64.61
P40C07	P40A1.53	2.680755	NGS	9	142632	103,818	58.3	Set 4	NRPS	Truncated	57.64	42.36	55.56
P40C19	P40A1.32	1.7382	NGS	5	113503	9,815	65.9	Set 4	NRPS	Truncated	54.23	45.77	60.04
P40O18	P40A2.5	2.004605	NGS	12	136500	1,137	69	Set 4	NRPS	Truncated	76.74	23.26	30.51
P41O18	P41.7	1.92096	NGS	9	61544	52,280	62.7	Set 5	PKS	Truncated	62.82	37.18	48.77
P44H15	P44.1391	2.70924	NGS	15	100008	41,973	53.5	Set 5	PKS-NRPS	Truncated	58.85	41.15	53.98
P47I11	P47.1080	2.55545	NGS	4	26351	76,463	58.4	Set 5	NRPS	Truncated	37.71	62.30	81.72
P48G08	P48.1239 P48.682	3.42013	NGS	12	76978	46,342	71	Set 5	NRPS	Truncated	52.15	47.86	62.78
P48N16	P48.126	2.67936	NGS	3	37665	12,723	57.7	Set 5	NRPS	Truncated	49.78	50.22	65.88
P49N01	P49.44	0.885196	NGS	11	58826	10,790	62	Set 5	PKS-NRPS	Truncated	57.67	42.33	55.53
P50M18	P50.95	2.18771	NGS	6	84718	1,755	57.1	Set 5	PKS	Truncated	90.23	9.77	12.82
P50N06	P50.4	2.0896	NGS	6	104104	59,363	37.8	Set 5	PKS	Truncated	78.34	21.66	28.42
P03B10	P03B2.241	2.67331	NGS	4	80352	24,854	52.8	Set 1	NRPS	Truncated	55	45.00	59.03
P15P01	P15B1.63	2.68009	NGS	8	132498	45,875	56.1	Set 2	NRPS	Truncated	67	33.00	43.29
P17B06	P17B2.114	2.07415	NGS	12	100101	105,964	60	Set 2	NRPS	Truncated	47	53.00	69.53
P17I02	P17A2.30	2.586536	NGS	10	125265	43,874	66.2	Set 2	NRPS	Truncated	67	33.00	43.29
P18H23	P18A1.398	2.67571	NGS	3	51470	25,813	61.6	Set 2	NRPS	Truncated	45	55.00	72.15
P19H07	P19B1.221	2.83341	NGS	2	16484	16,484	52.7	Set 2	NRPS	Truncated	53	47.00	61.66
P19L22	P19A2.82 P19B2.34	3.16285	NGS	7	87188	51,653	56.3	Set 2	NRPS	Truncated	35	65.00	85.27
P21A05	P21A1.60	2.960735	NGS	4	54244	26,263	31.3	Set 3	NRPS	Truncated	57	43.00	56.41
P23I15	P21A1.65	2.9474	NGS	4	61549	42,000	55	Set 3	PKS	Truncated	70	30.00	39.35
P25F10	P25B2.32	3.17367	NGS	5	64797	45,350	62.7	Set 3	NRPS	Truncated	37	63.00	82.64
P25O21	P25A1.147	2.758465	NGS	8	37775	29,896	62.4	Set 3	NRPS	Truncated	63	37.00	48.54
P26D23	P26B1.113	2.728533	NGS	3	51817	24,124	57.3	Set 3	NRPS	Truncated	59	41.00	53.78
P27C17	P27A1.12	3.39678	NGS	7	102918	47,322	41.2	Set 3	NRPS	Truncated	33	67.00	87.89
P28C19	P28A1.563	2.72299	NGS	3	127612	43,223	64	Set 3	NRPS	Truncated	47	53.00	69.53
P28H21	P28B1.114	2.84304	NGS	3	28688	25,219	55.4	Set 3	NRPS	Truncated	56	44.00	57.72
P29M03	P29A1.109	3.7476	NGS	5	97291	47,644	63.9	Set 3	PKS	Truncated	49	51.00	66.90
P33F15	P33B1.42	1.78757	NGS	5	125482	35,404	63.9	Set 4	NRPS	Truncated	57	43.00	56.41
P34D08	P34B2.15	2.780065	NGS	7	29301	27,435	53.8	Set 4	NRPS	Truncated	61	39.00	51.16
P34F16	P34B2.447	3.01348	NGS	4	58820	25,347	56.1	Set 4	NRPS	Truncated	58	42.00	55.10
P35M06	P35A2.5	1.79834	NGS	5	96287	30,405	40	Set 4	PKS	Truncated	84	16.00	20.99
P38B06	P38B2.77	2.54513	NGS	7	91764	47,012	70.1	Set 4	NRPS	Truncated	56	44.00	57.72
P38J21	P38B1.620	3.05963	NGS	8	97705	34,263	62.5	Set 4	NRPS	Truncated	95	5.00	6.56
P48E04	P49.261	2.61969	NGS	6	88481	49,571	63.6	Set 5	NRPS	Truncated	44	56.00	73.46

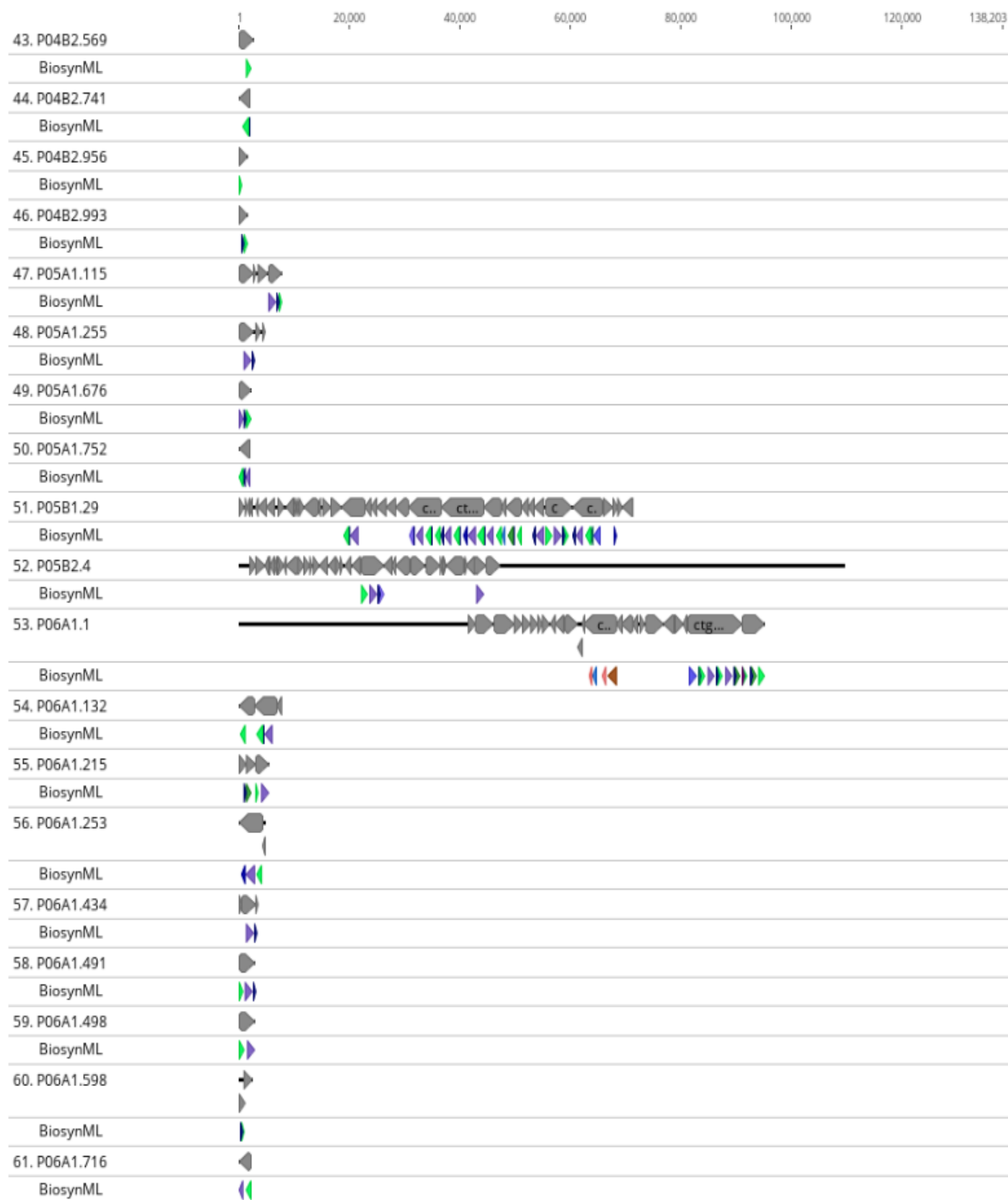
Clone	Contig	BLmean	Identified	Domains	Insert Length	BGC Length	GC (%)	Set	Type	Truncated	Domain Ident	Dom Div	DivT
P04E17	P04A1.356	1.74144	NGS-PCR	12	117009	37,401	47.9	Set 1	PKS-NRPS	Truncated	43.94	56.06	73.54
P07M04	P07A1.173	0.977724	NGS-PCR	5	11583	39,662	64.8	Set 1	PKS-NRPS	Truncated	71.17	28.83	37.82
P09P03	P09B1.20	1.66243	NGS-PCR	11	110970	73,677	63.8	Set 1	PKS-NRPS	Truncated	56.22	43.78	57.43
P22N20	P22B2.30	1.2852	NGS-PCR	8	103714	98,719	52	Set 3	PKS	Truncated	78.12	21.88	28.71
P24H13	P24B1.29	1.83181	NGS-PCR	16	99579	43,697	51.6	Set 3	PKS-NRPS	Truncated	81.82	18.18	23.85
P35L08	P35B2.21	1	NGS-PCR	10	81817	4,886	59	Set 4	PKS	Truncated	62.95	37.05	48.60
P34D04	P34B2.375	2.45349	NGS-PCR	3	68400	31,669	59.9	Set 4	NRPS	Truncated	43	57.00	74.77

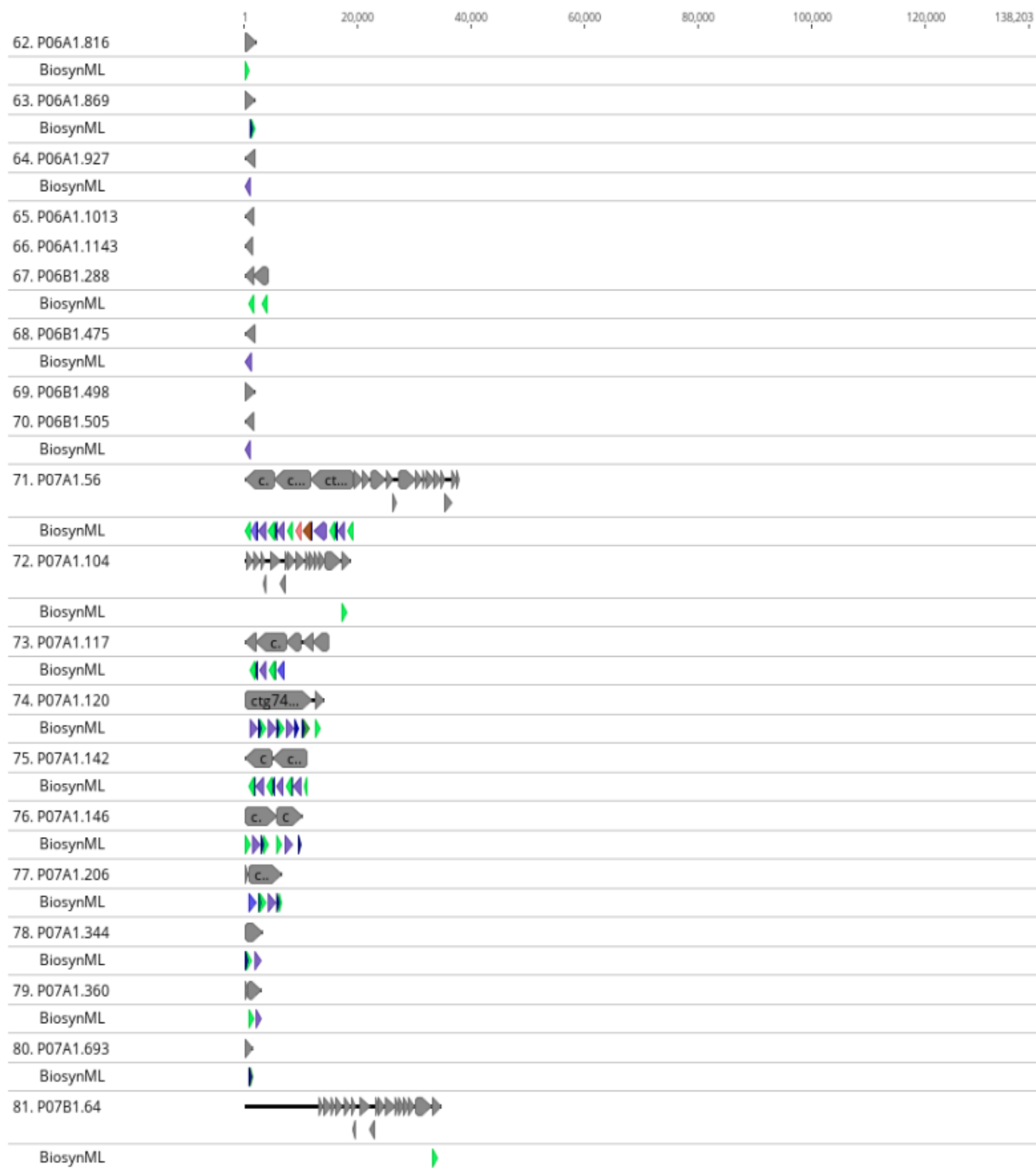
Supplemental figure 1. Annotation of PKS and NRPS pathways predicted bioinformatically and from resequenced clones

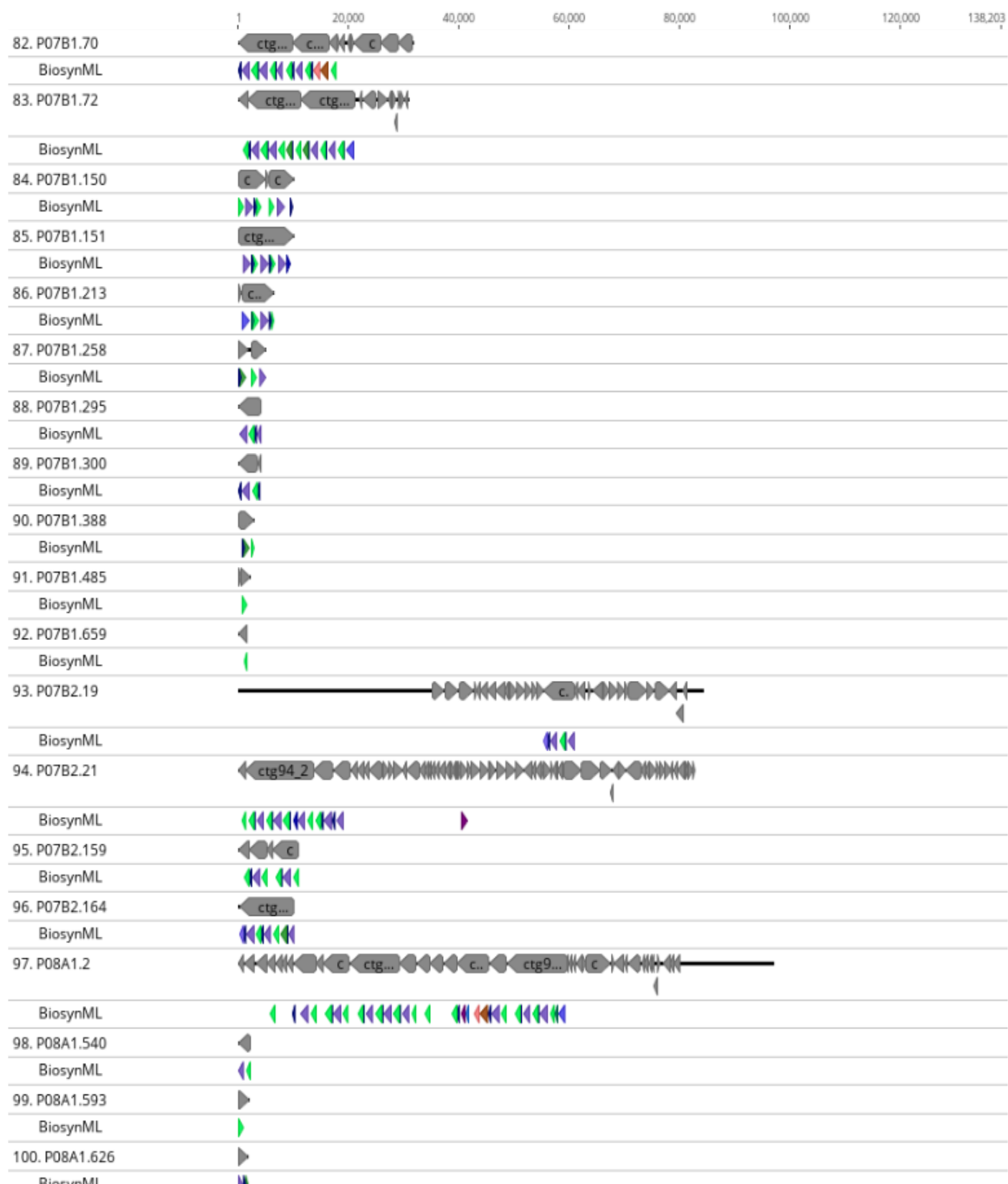
a. Clones predicted to carry PKS Pathways.

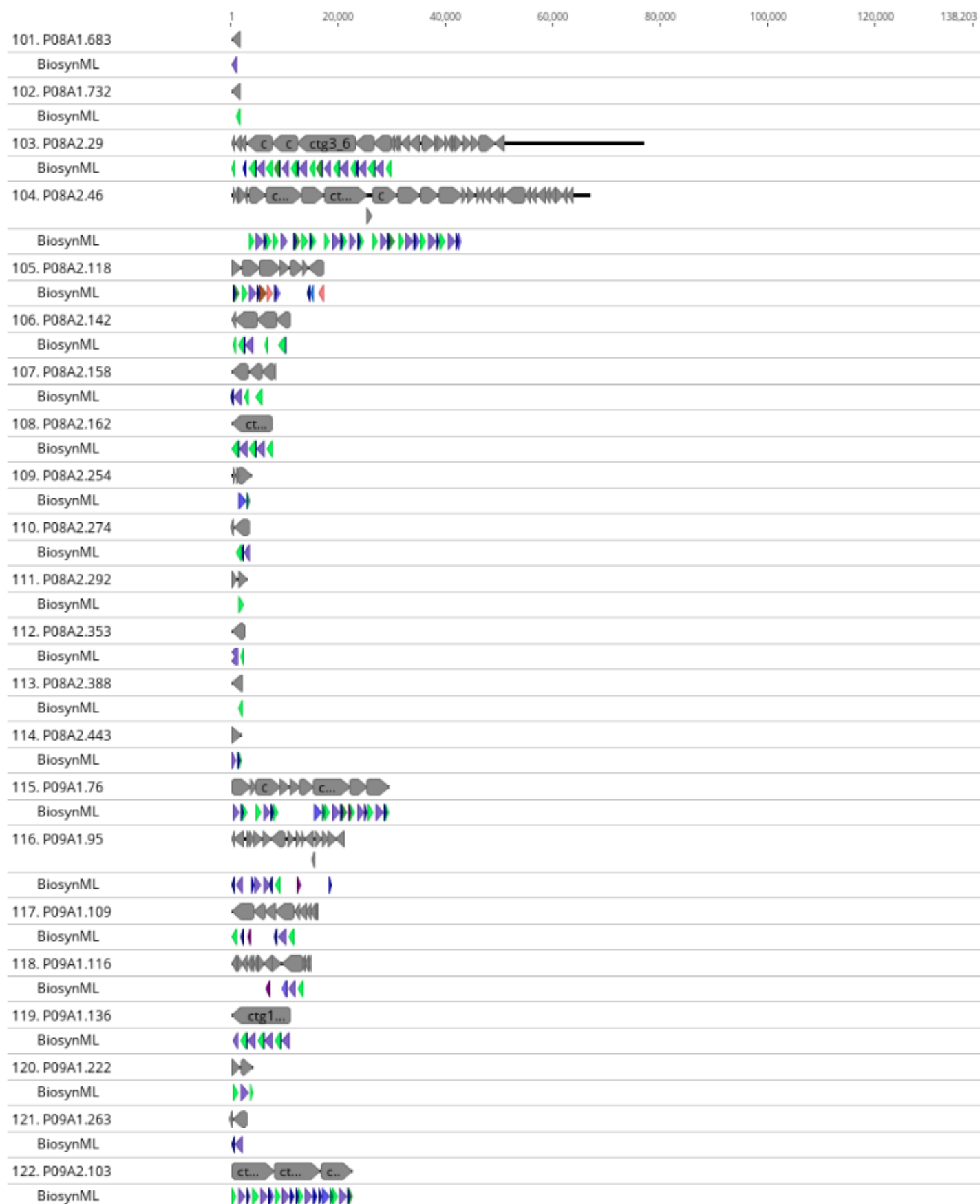




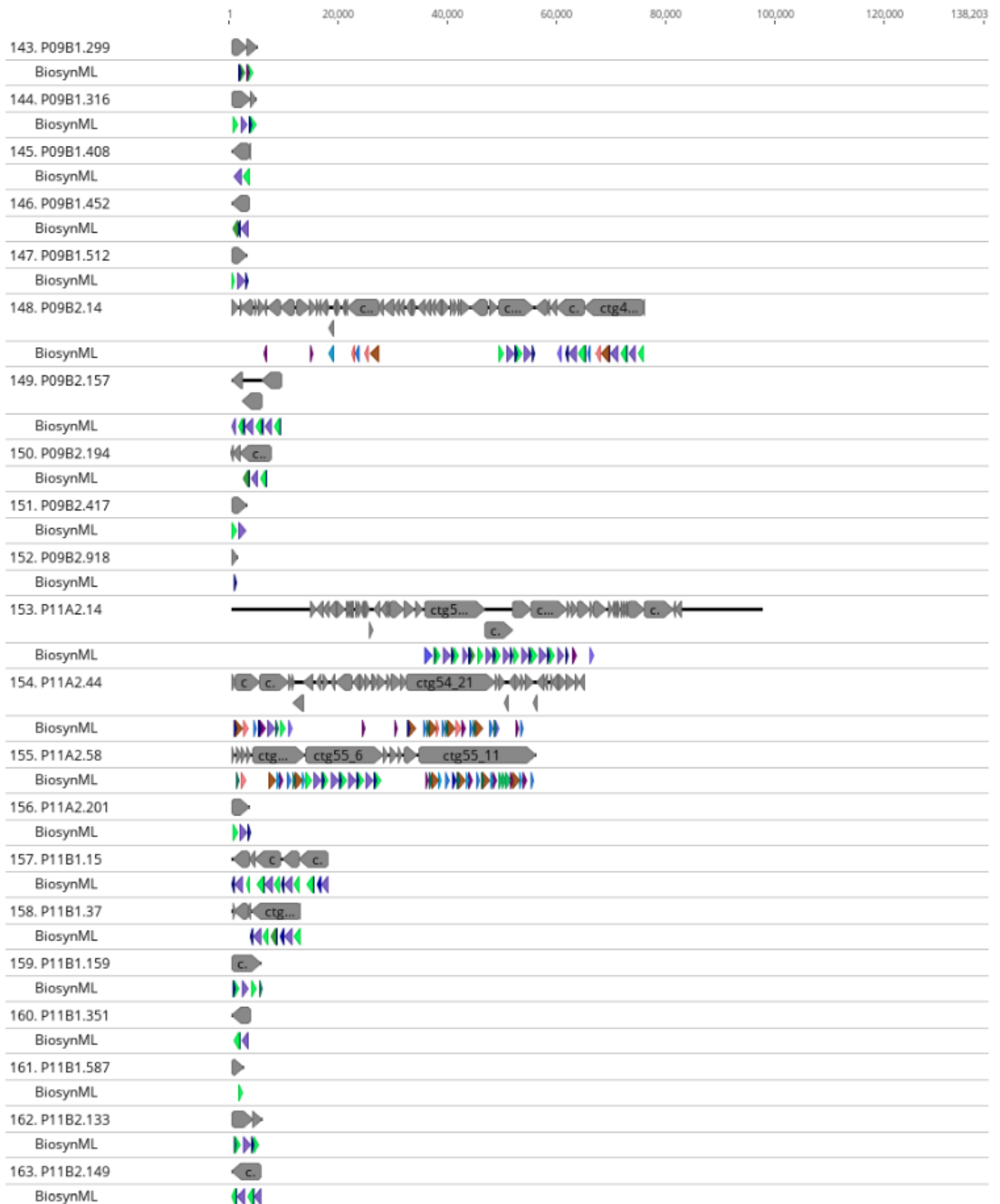


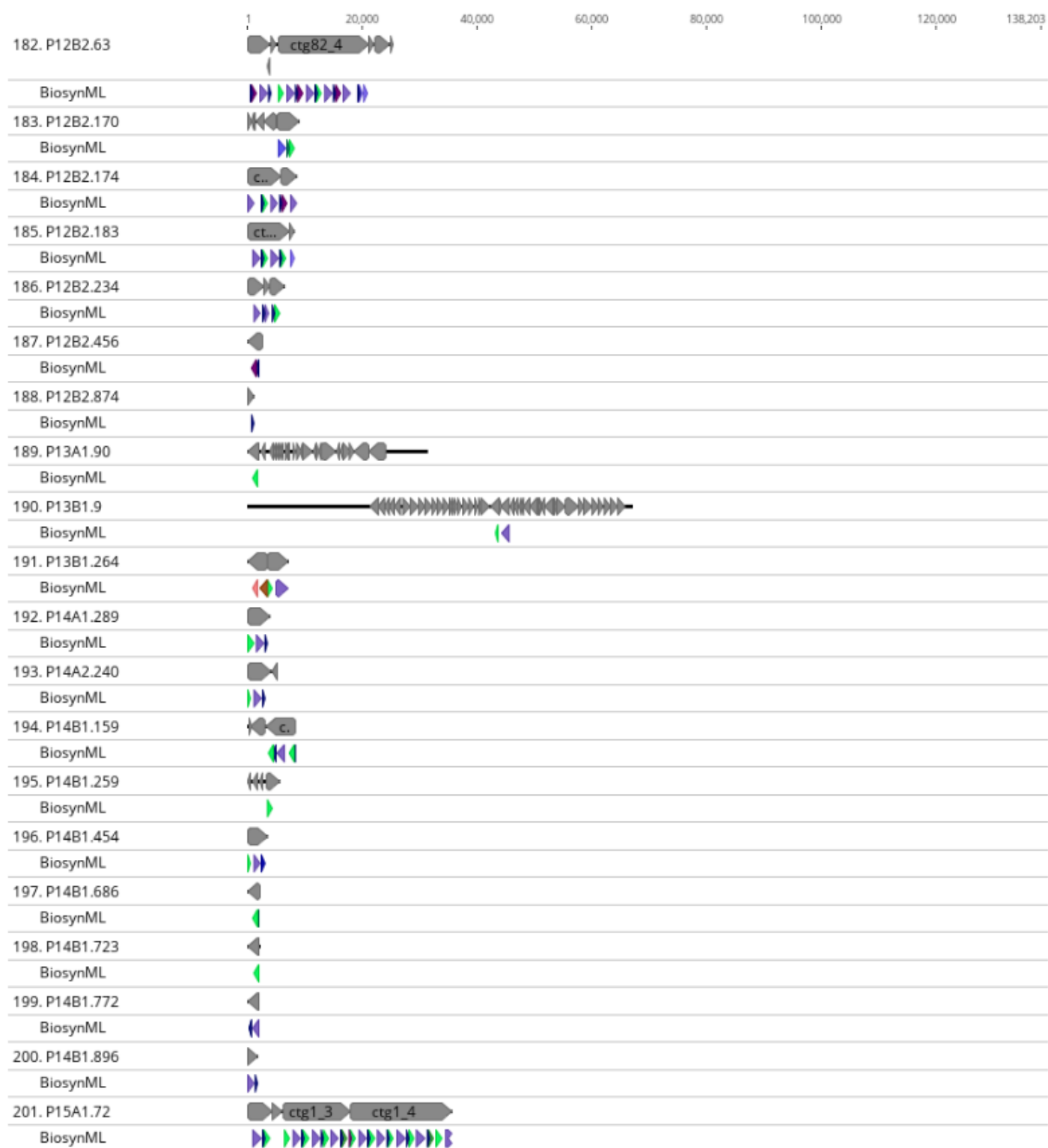


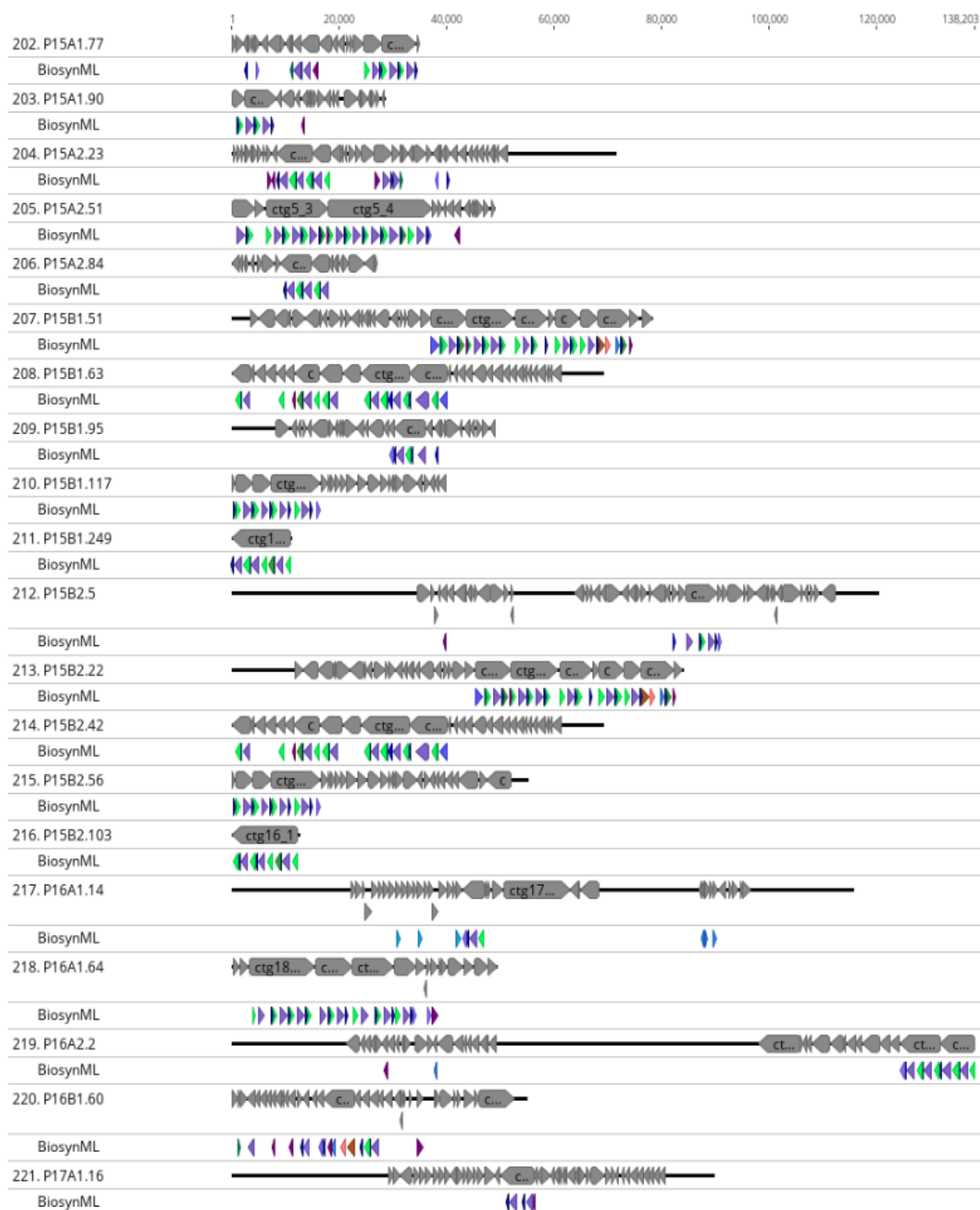


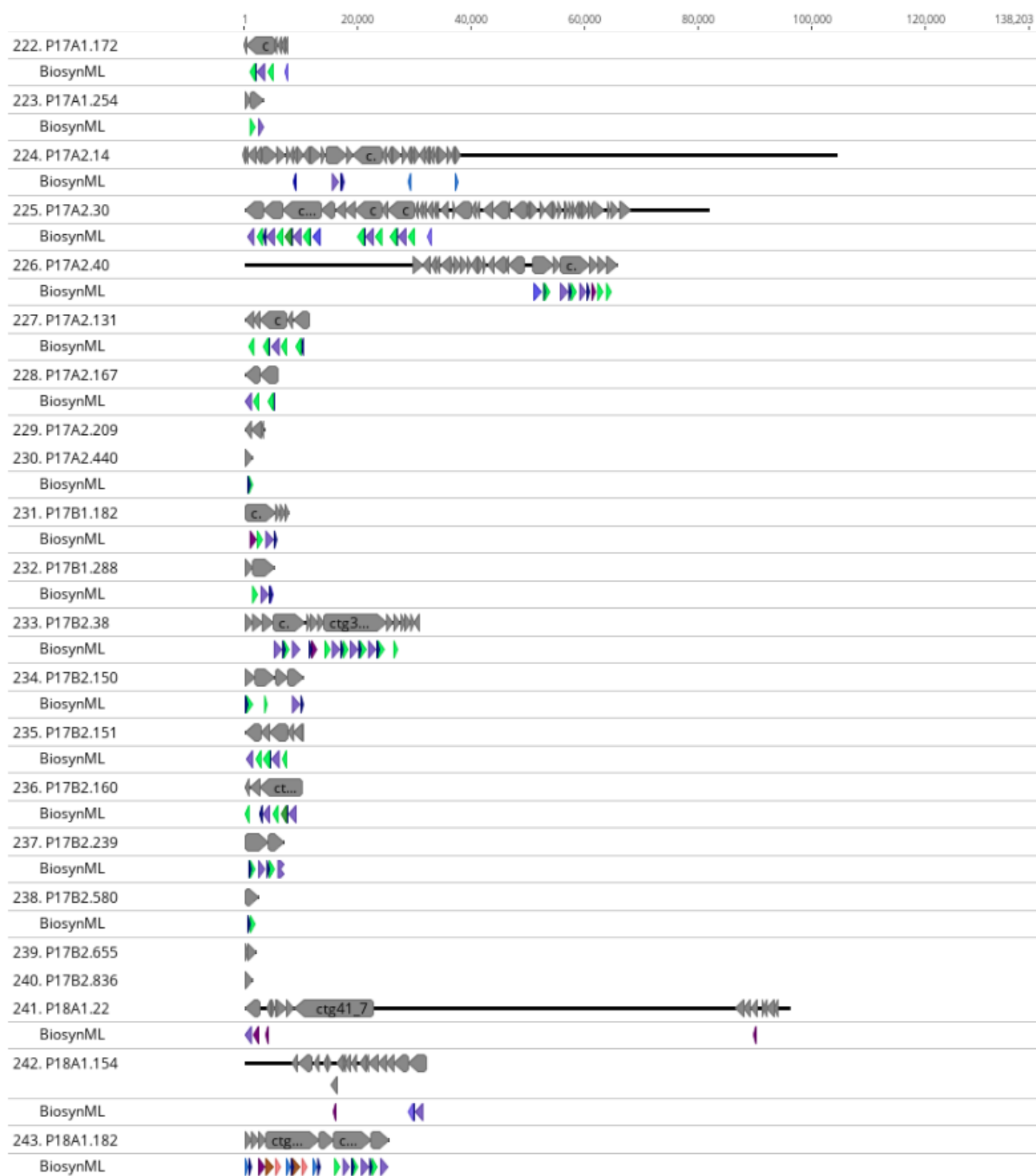




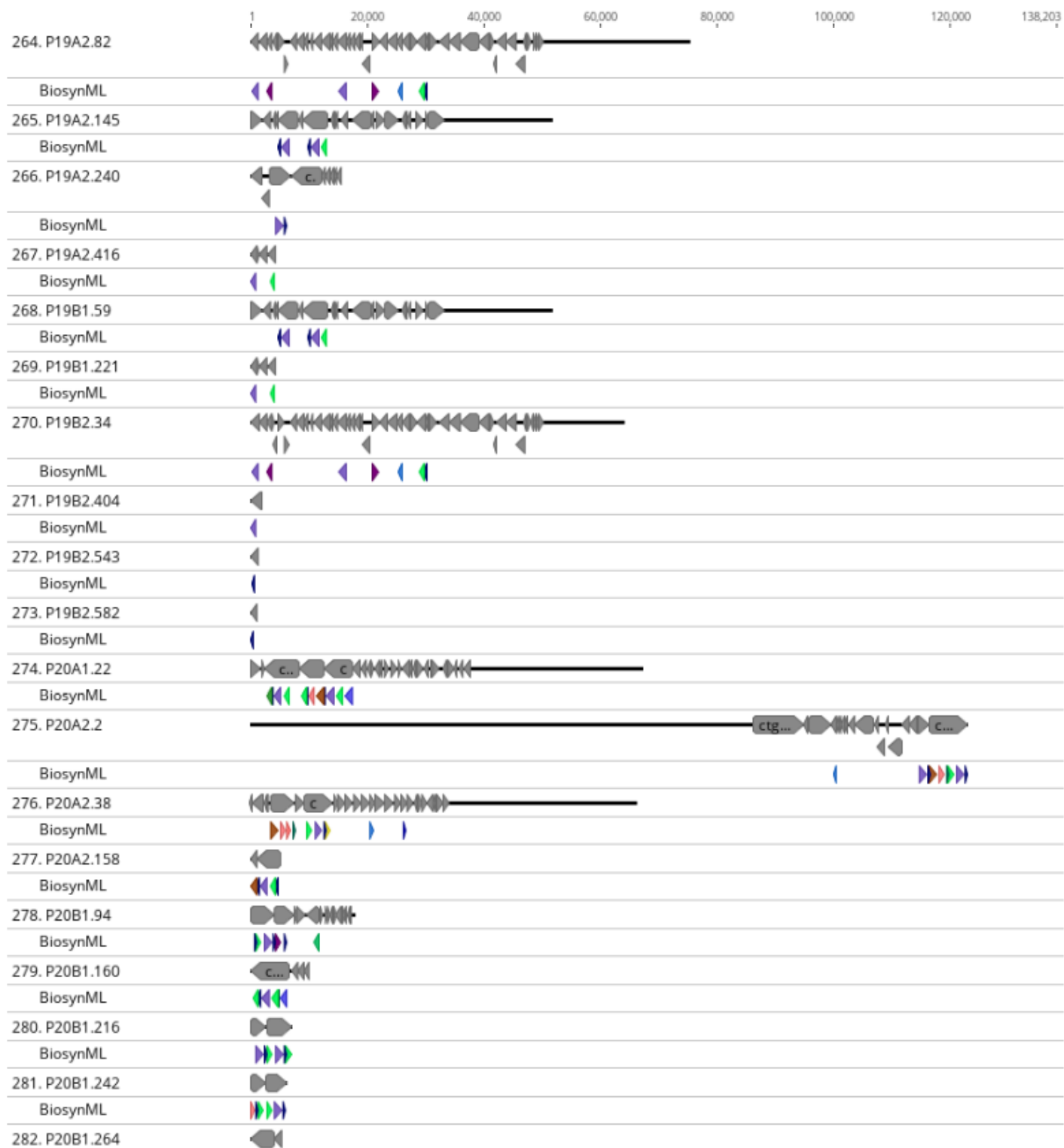


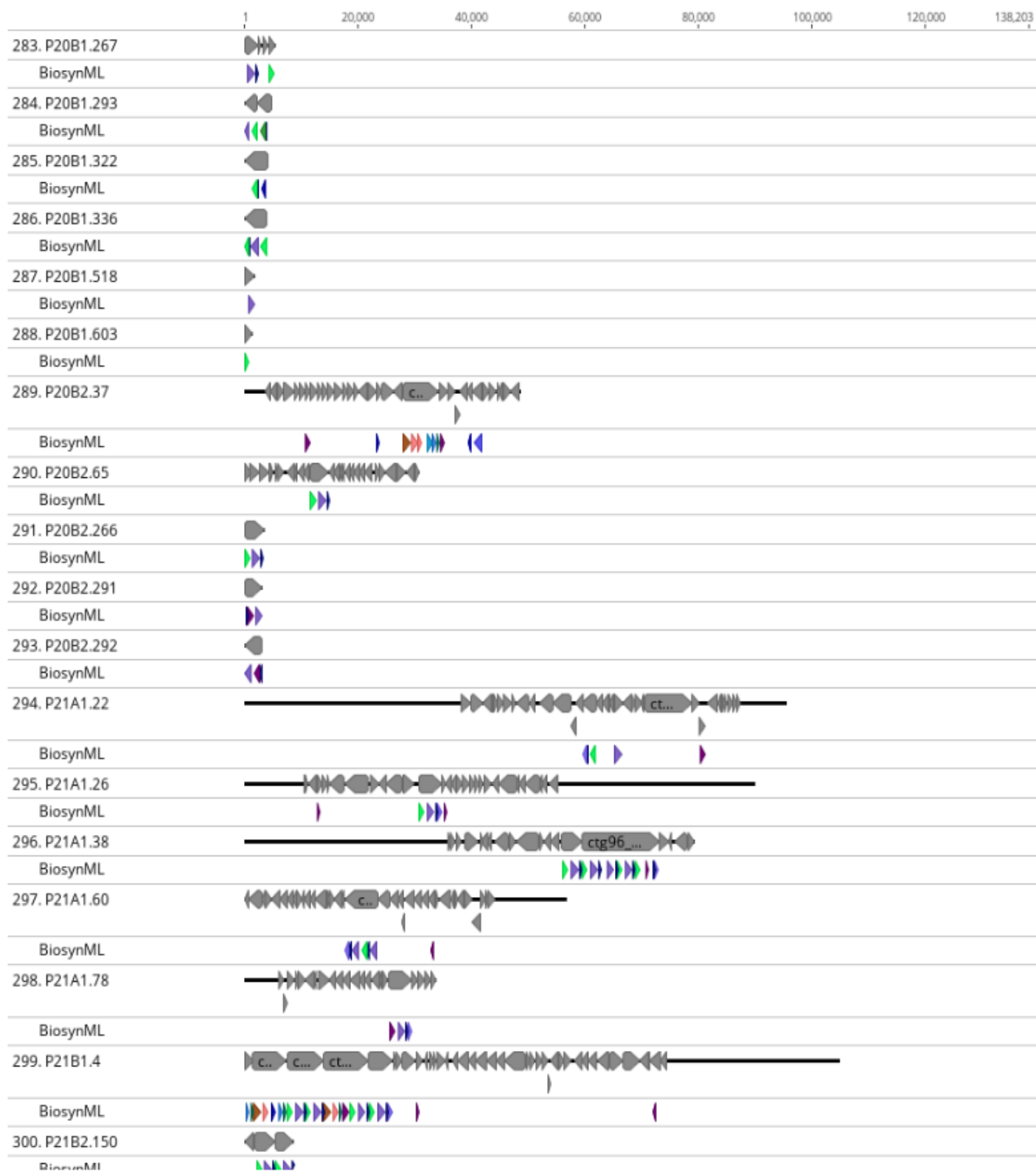


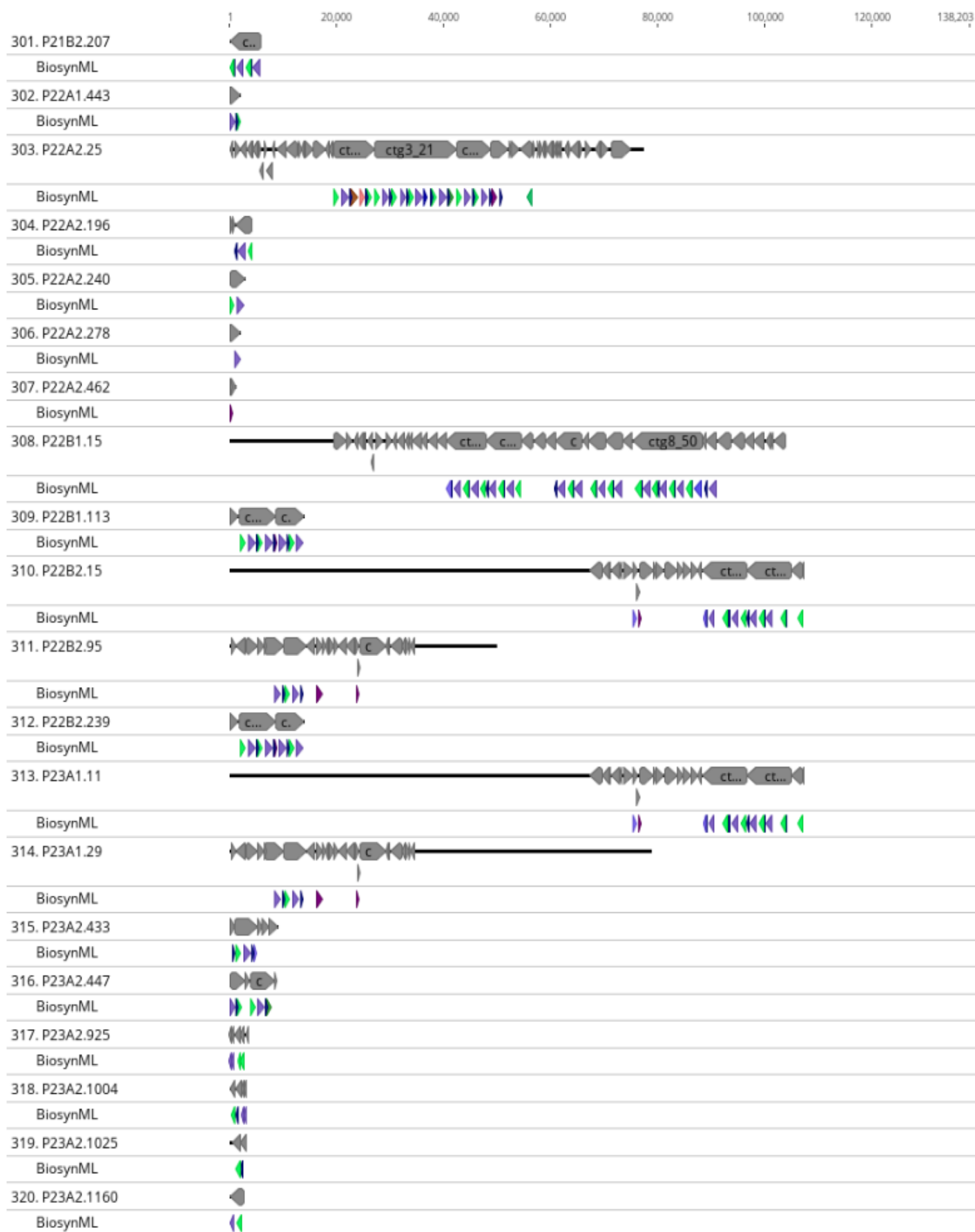


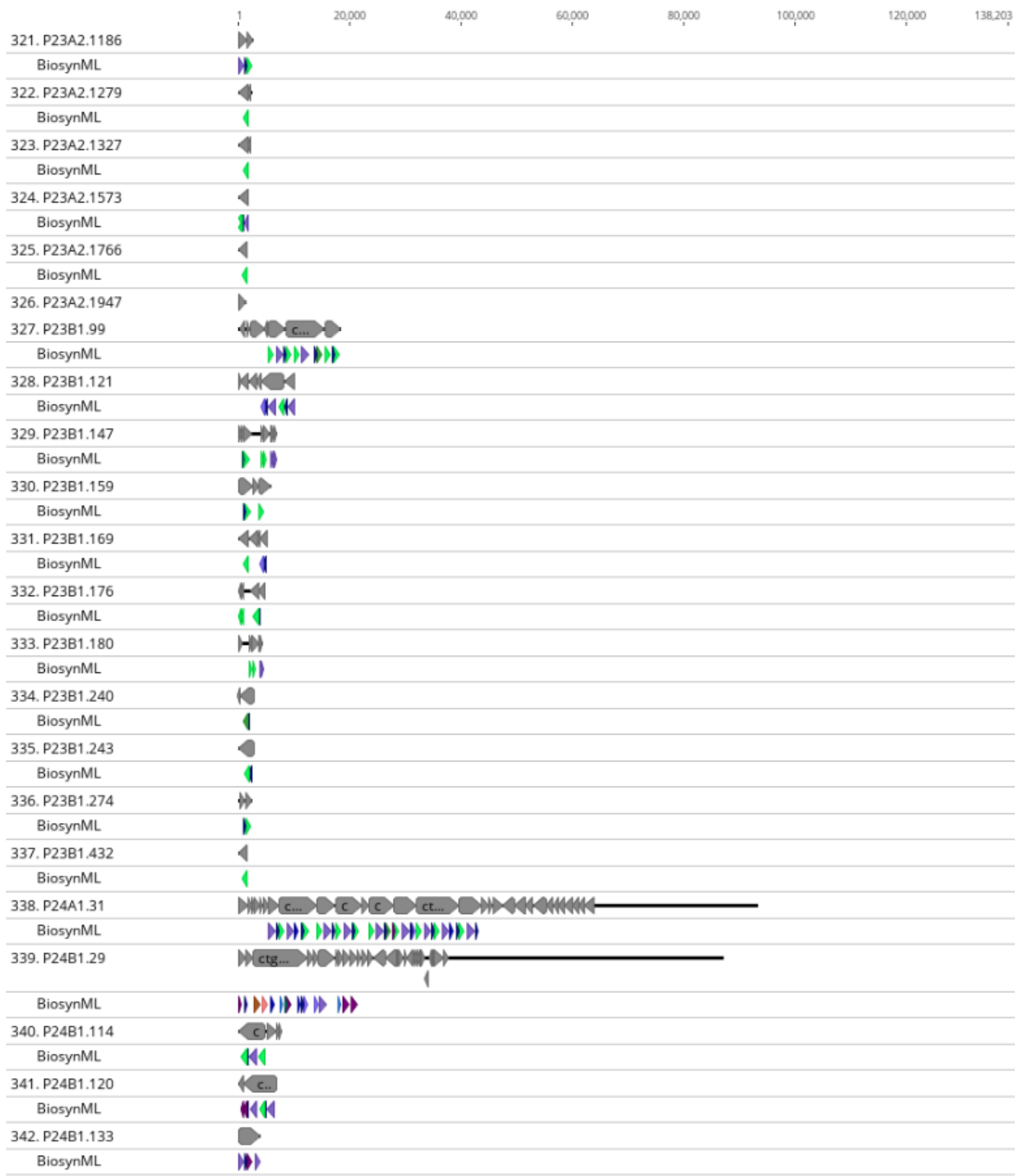


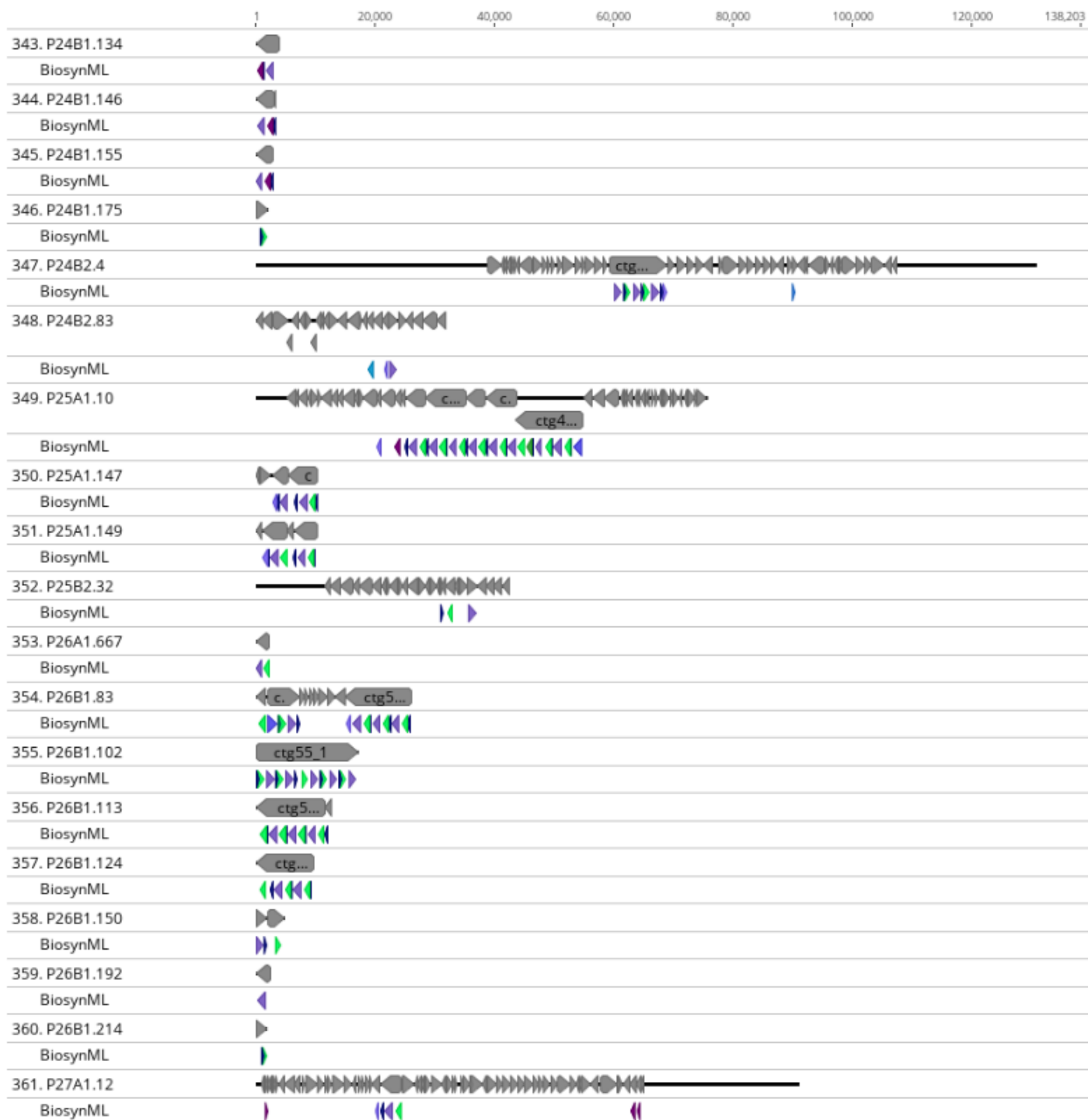


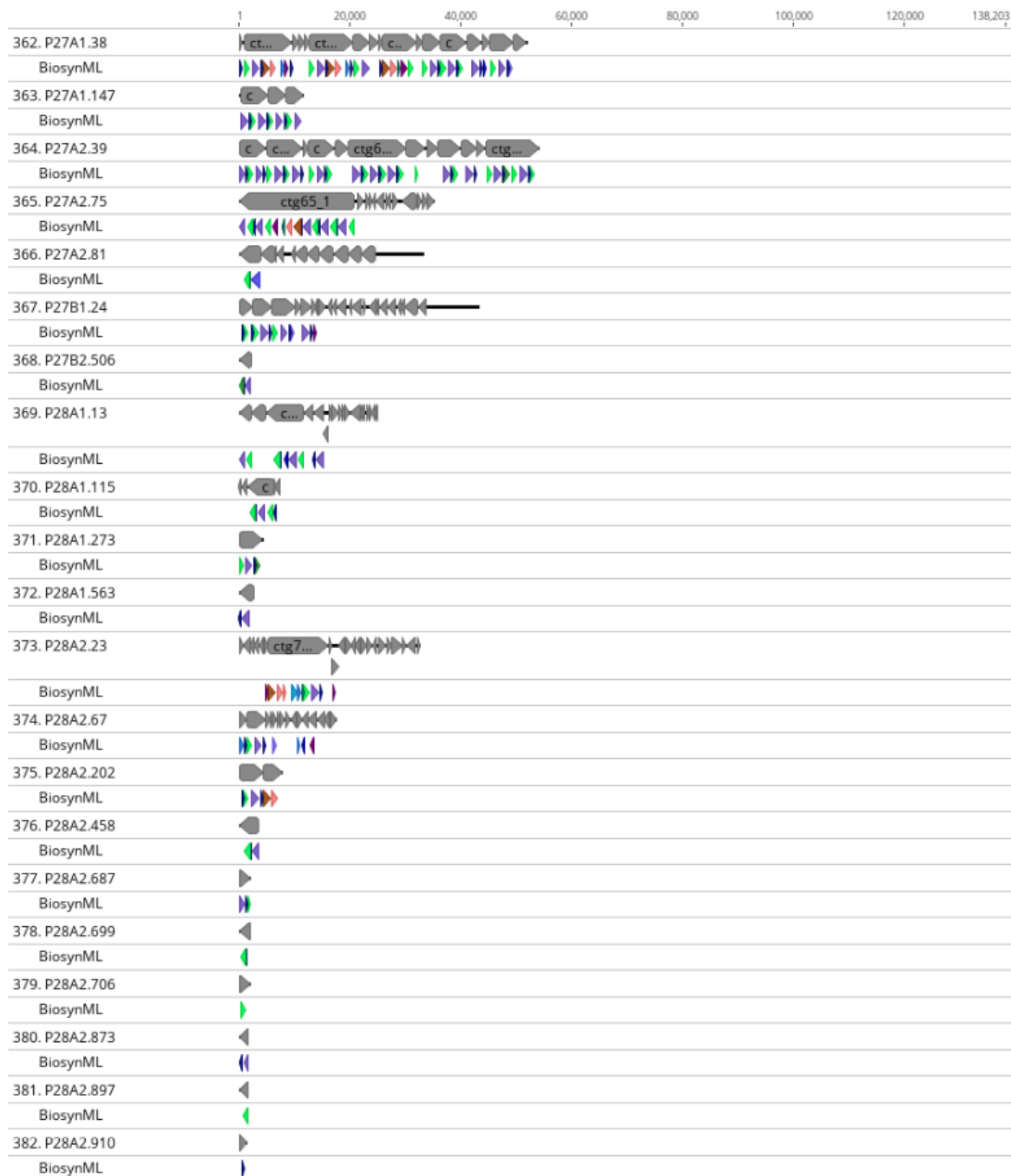


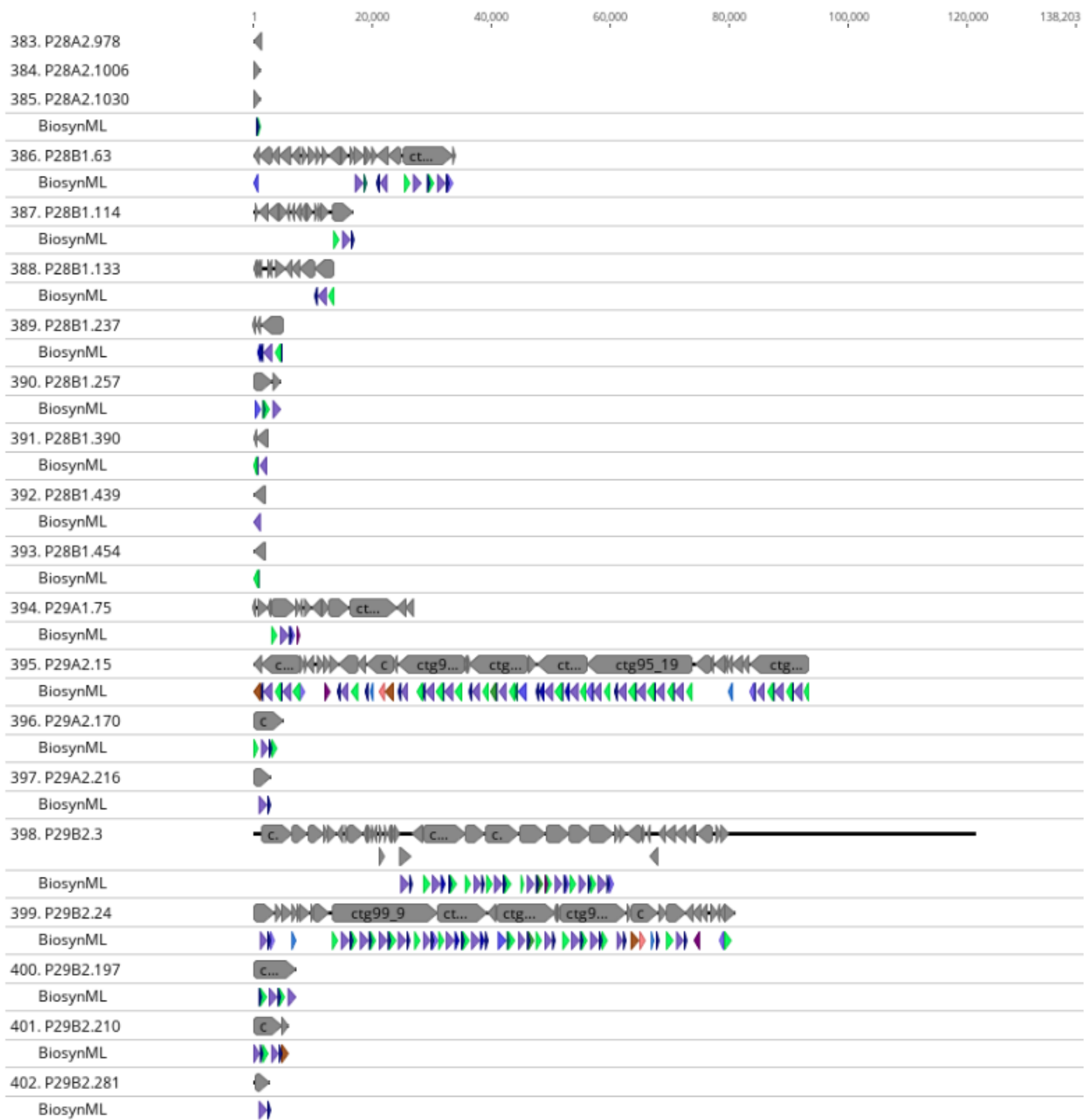


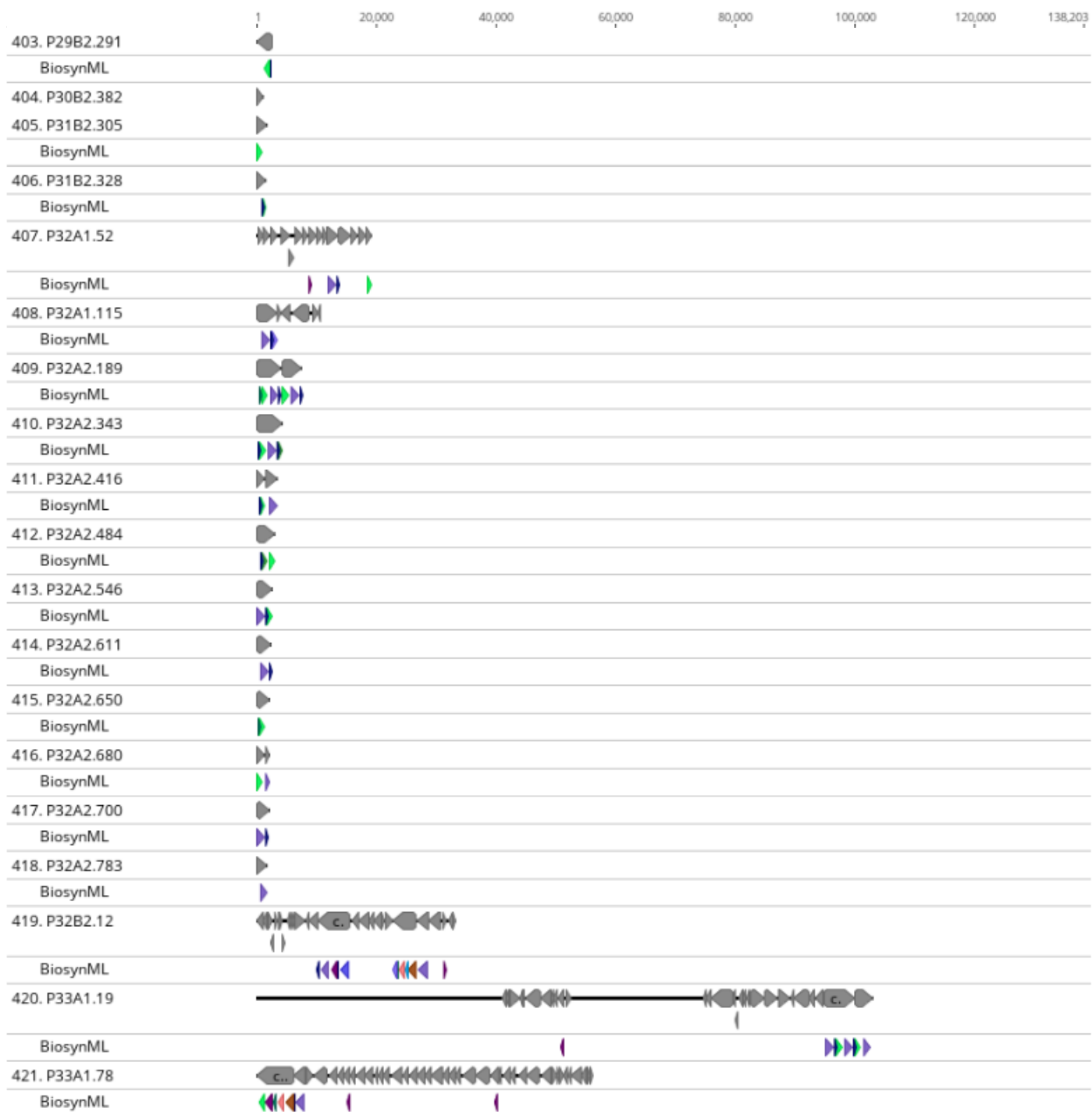


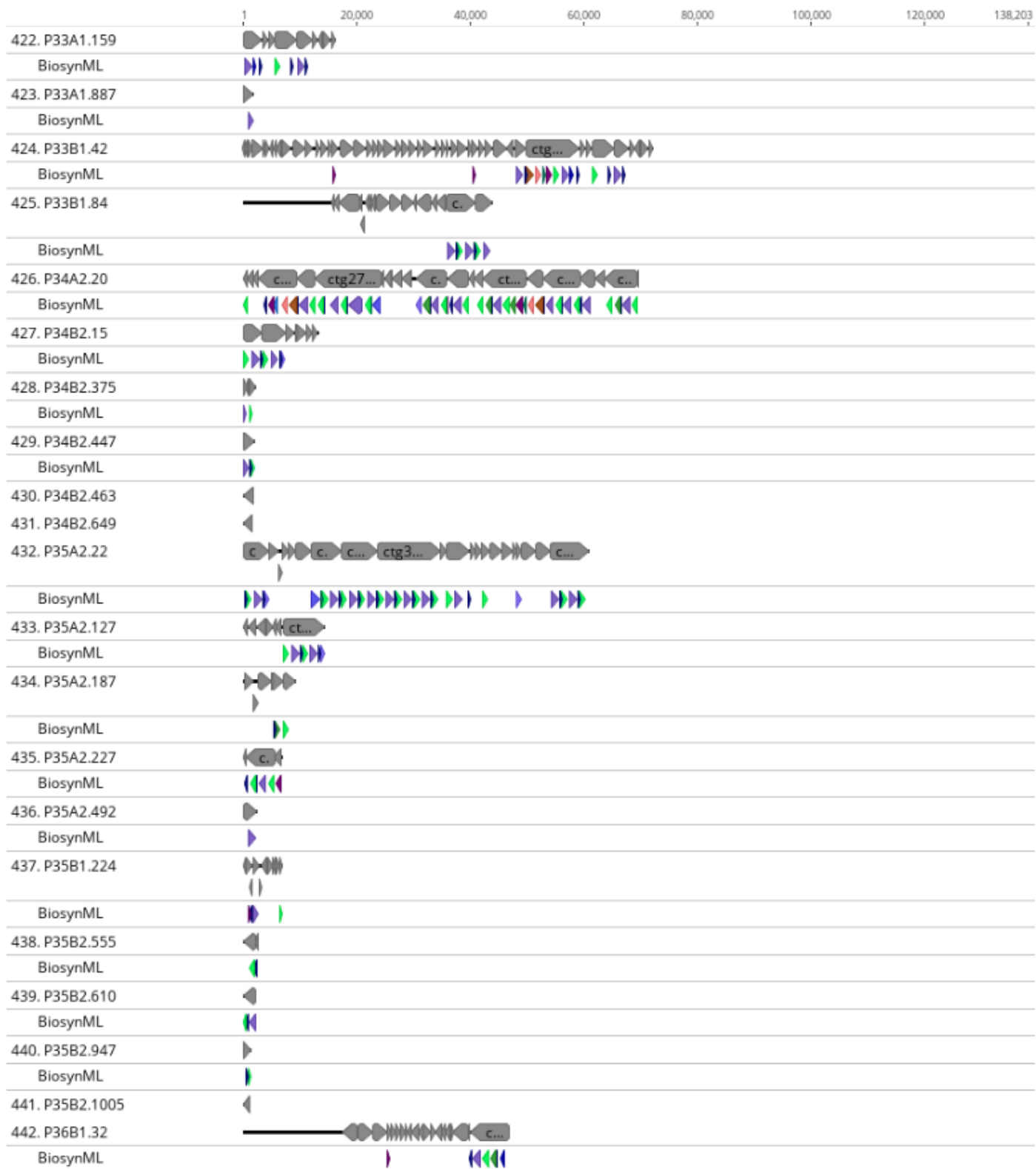


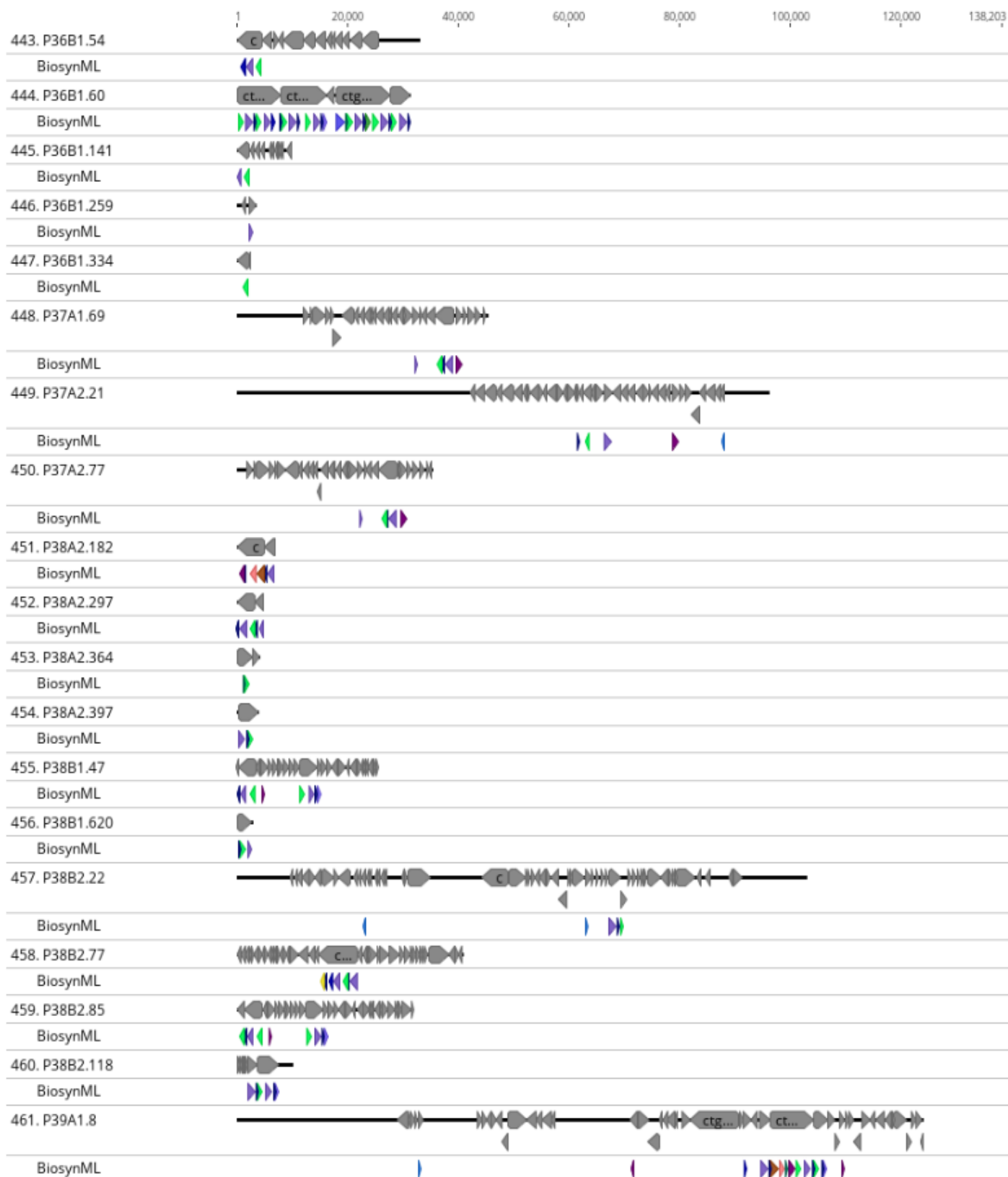


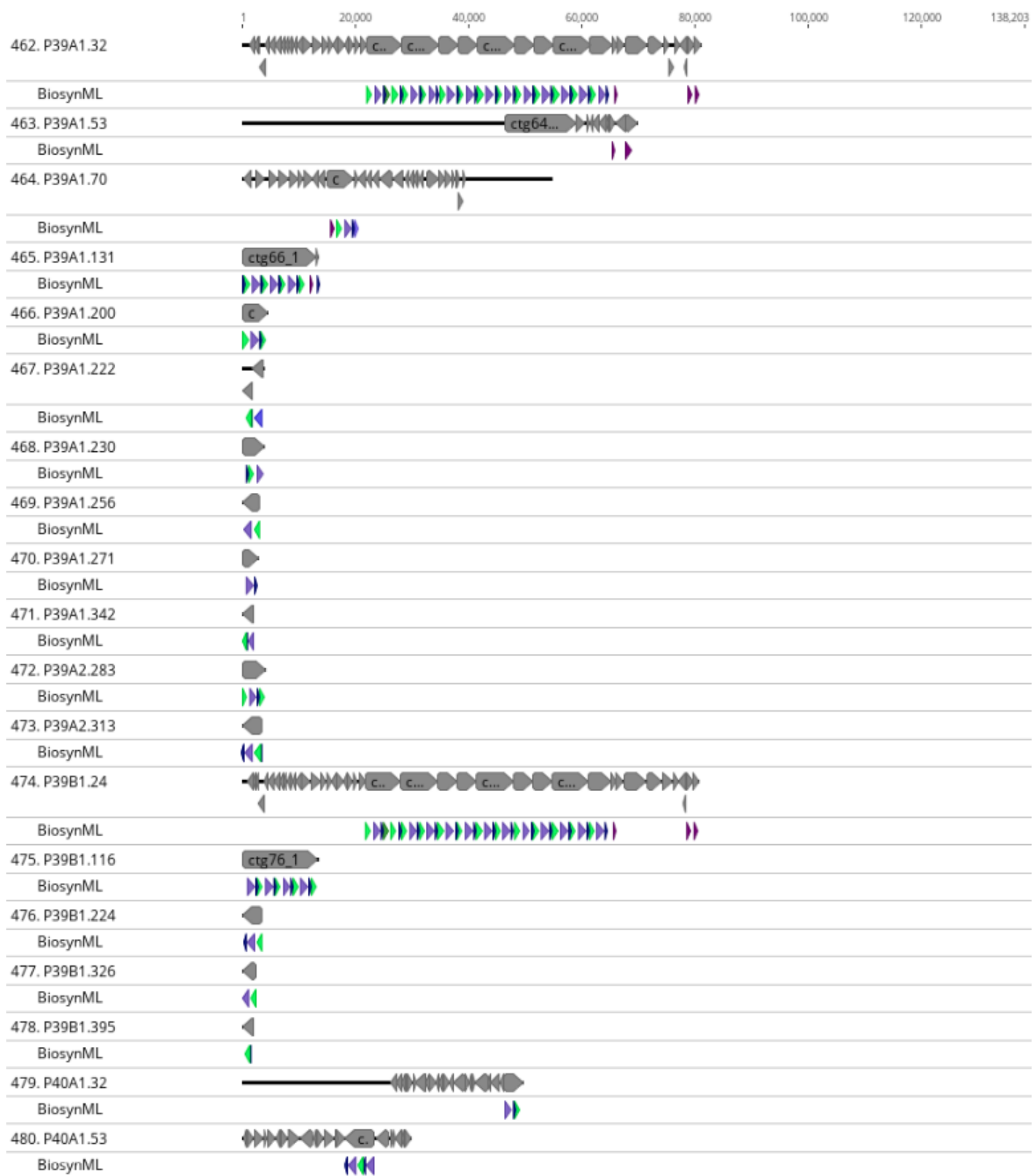


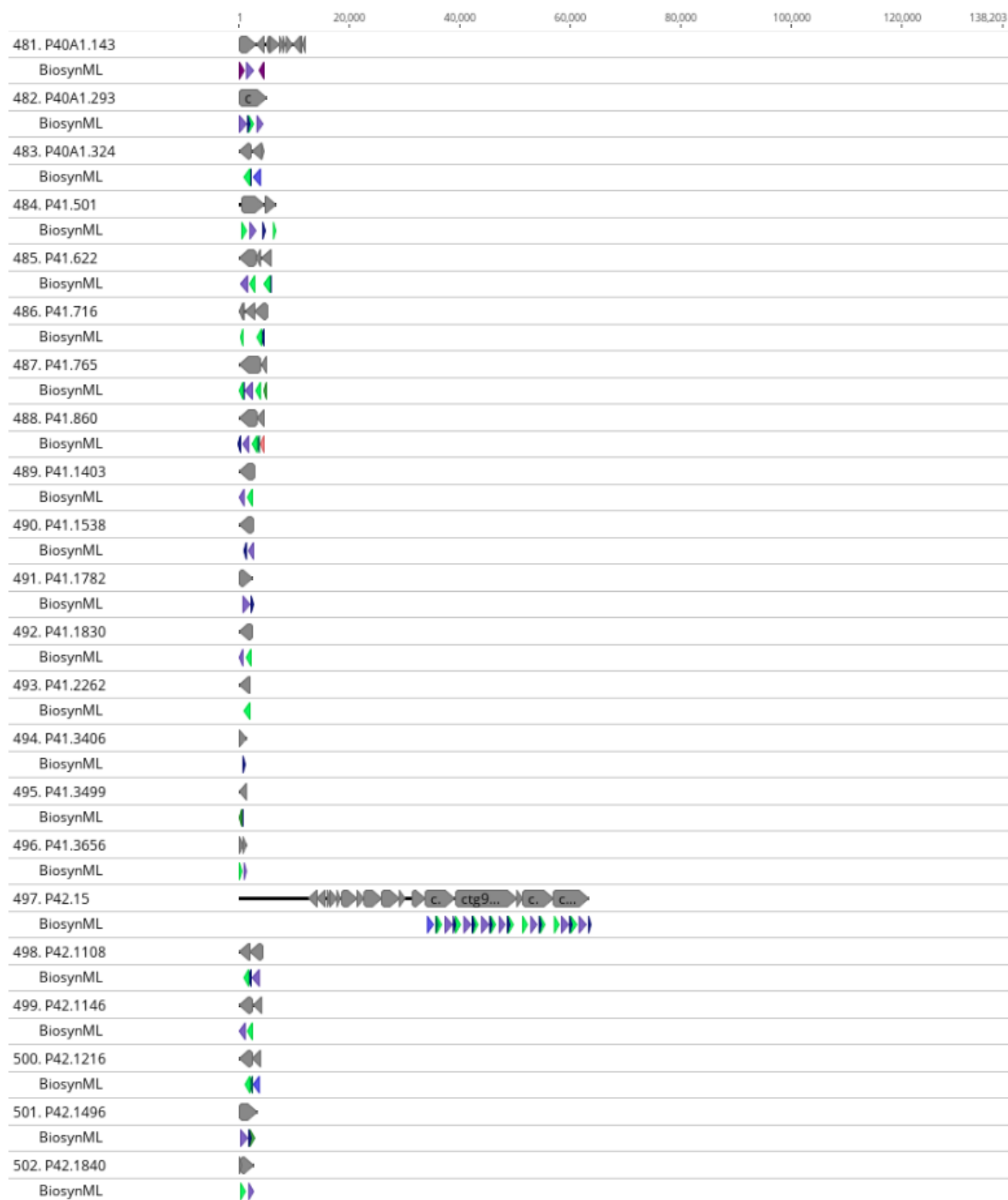


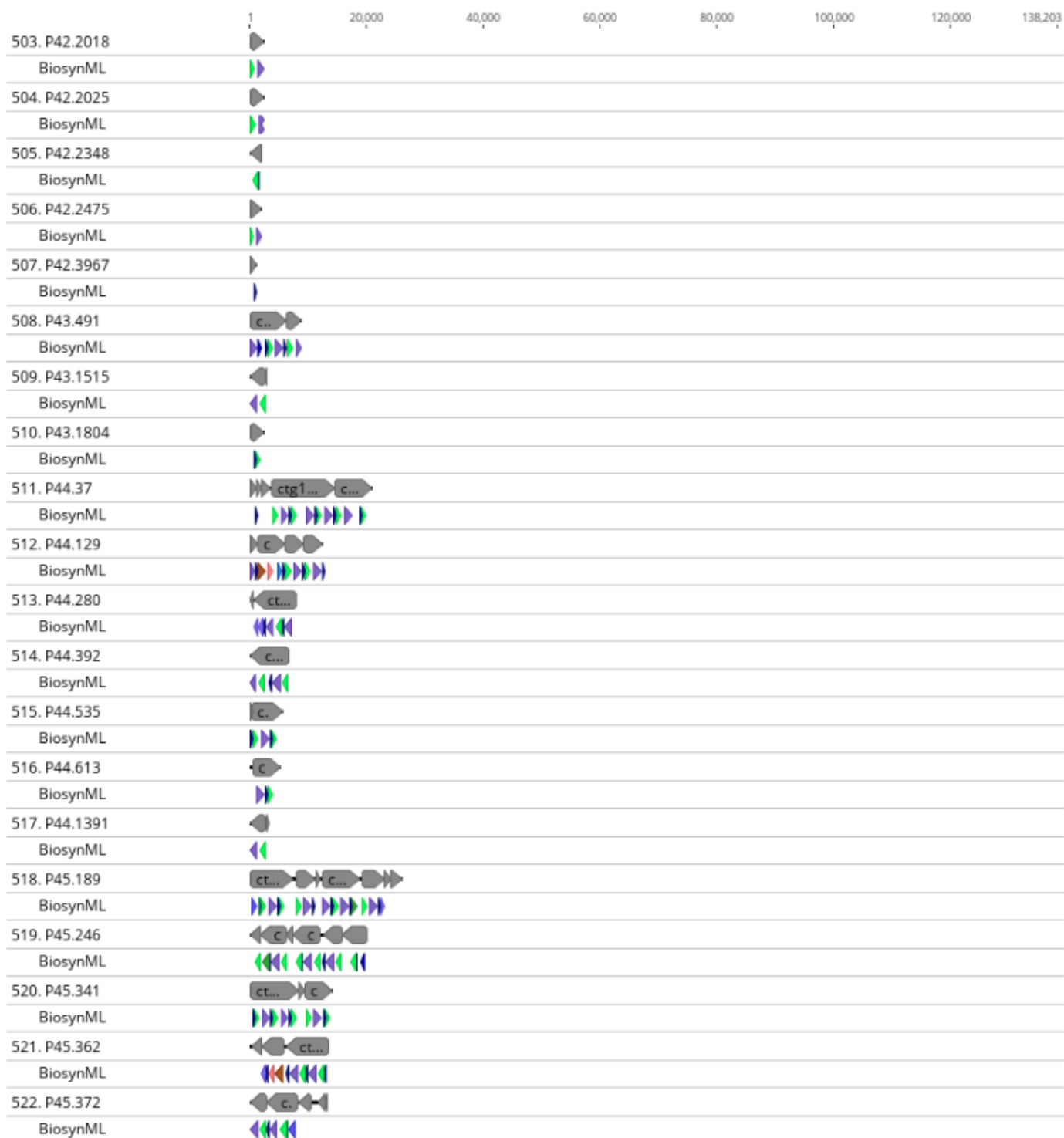


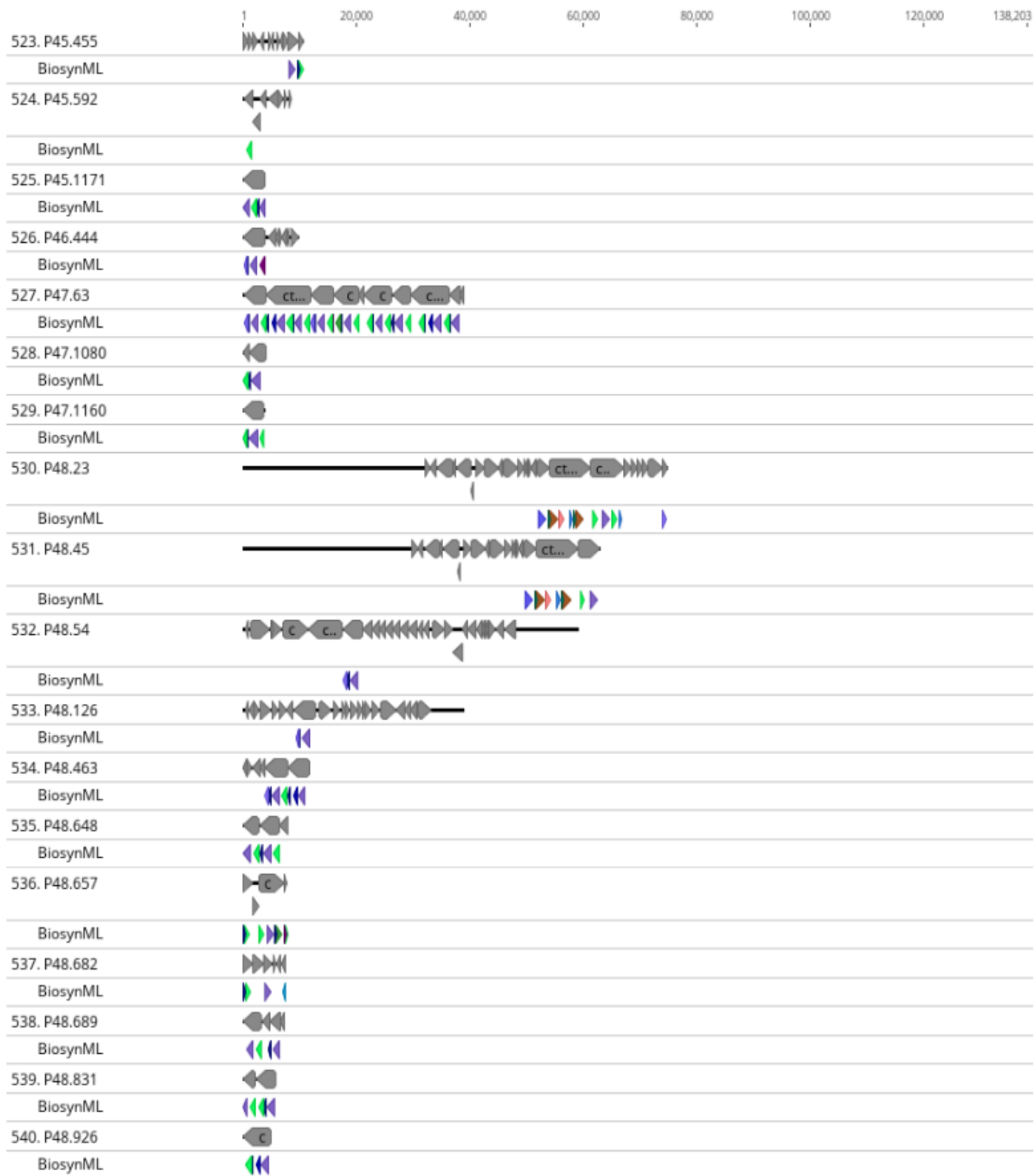






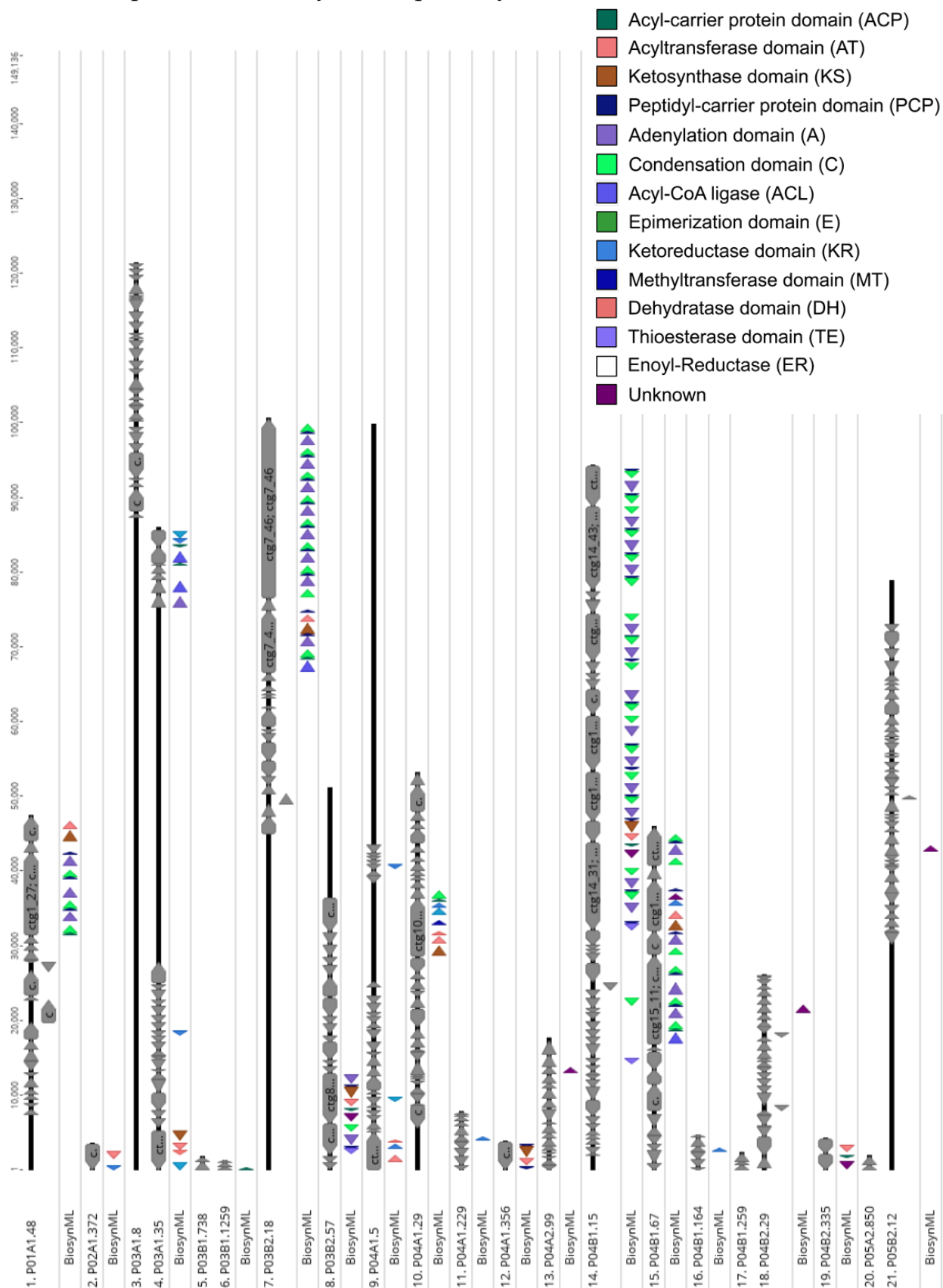


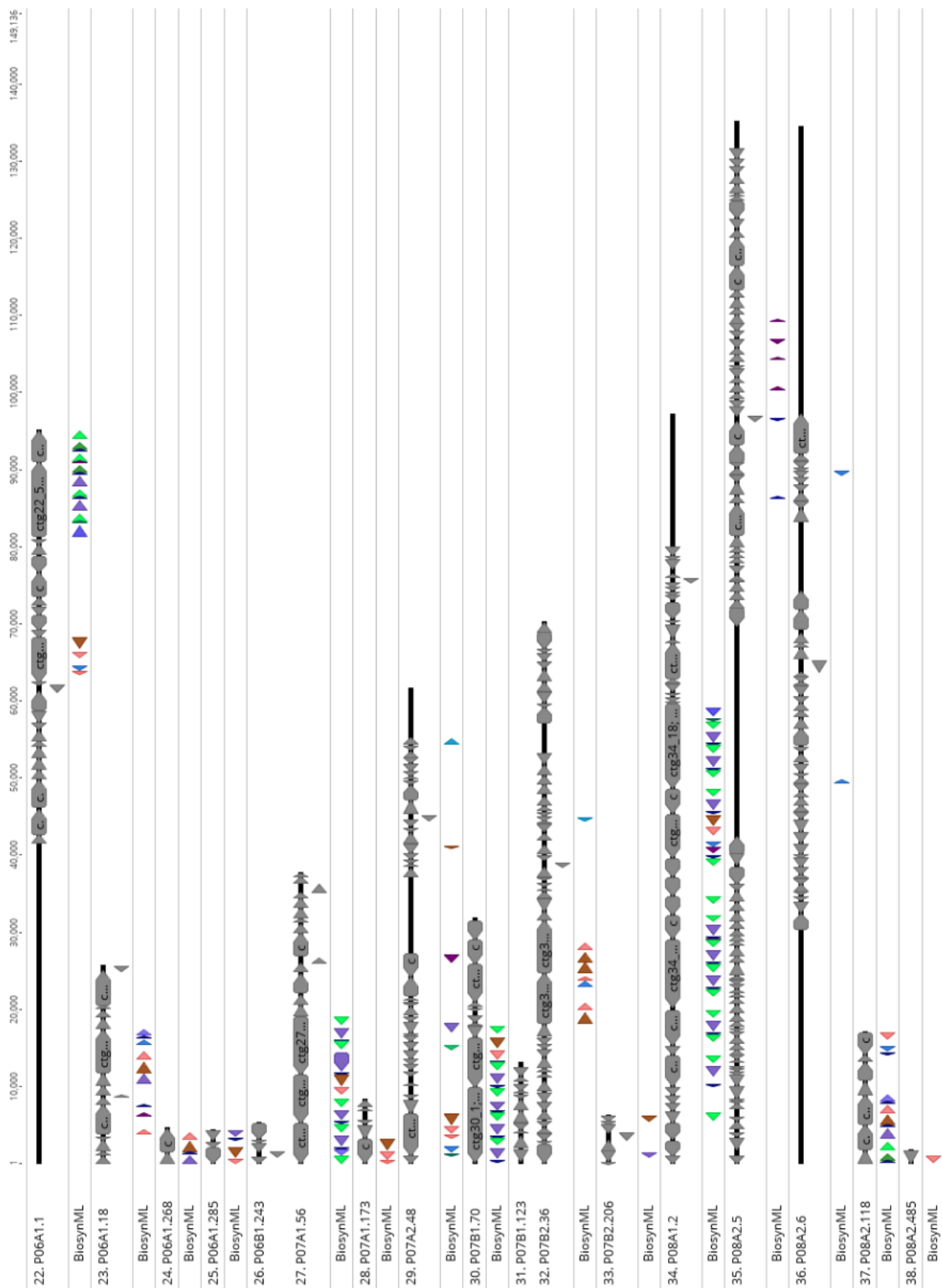


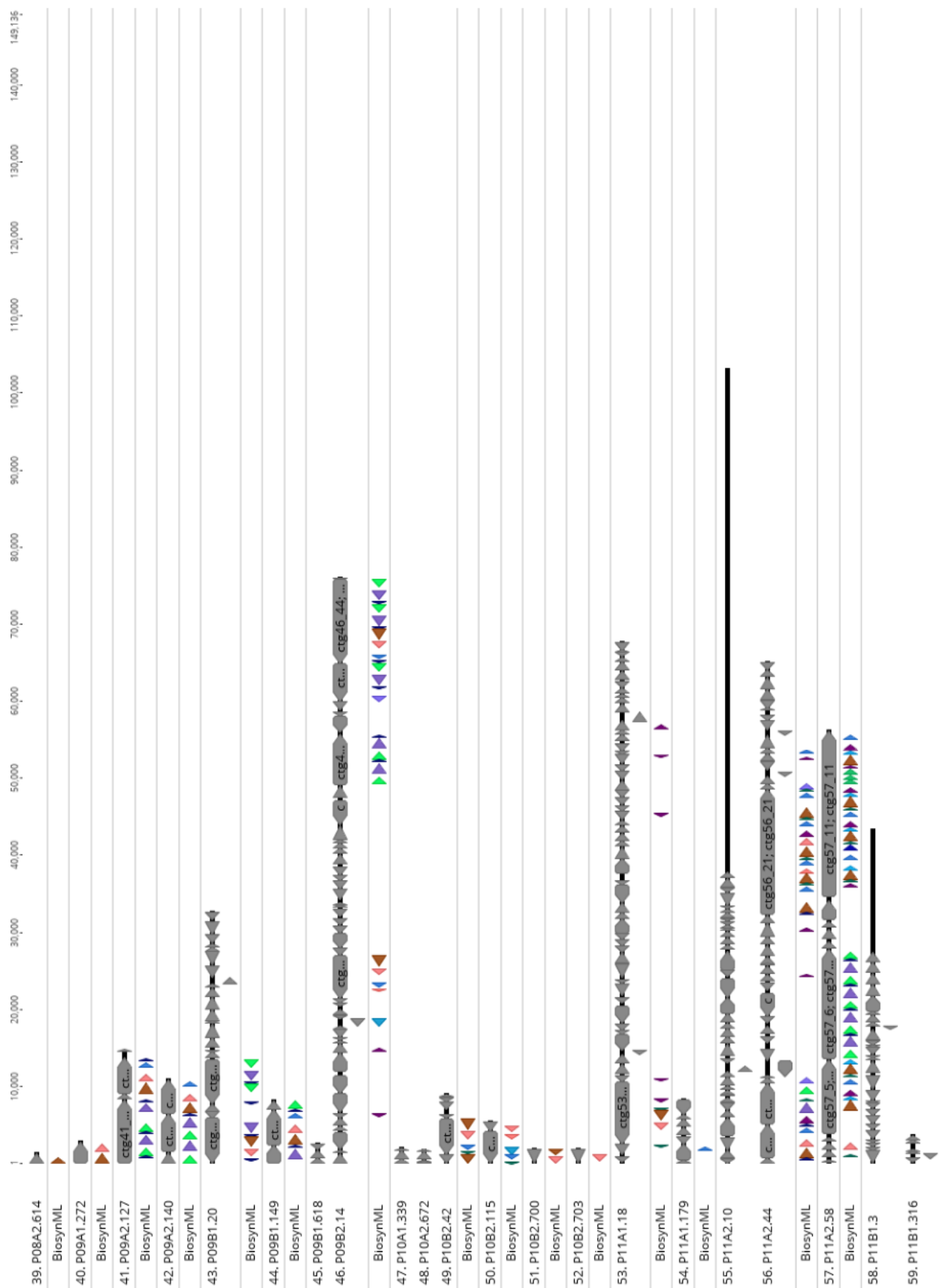


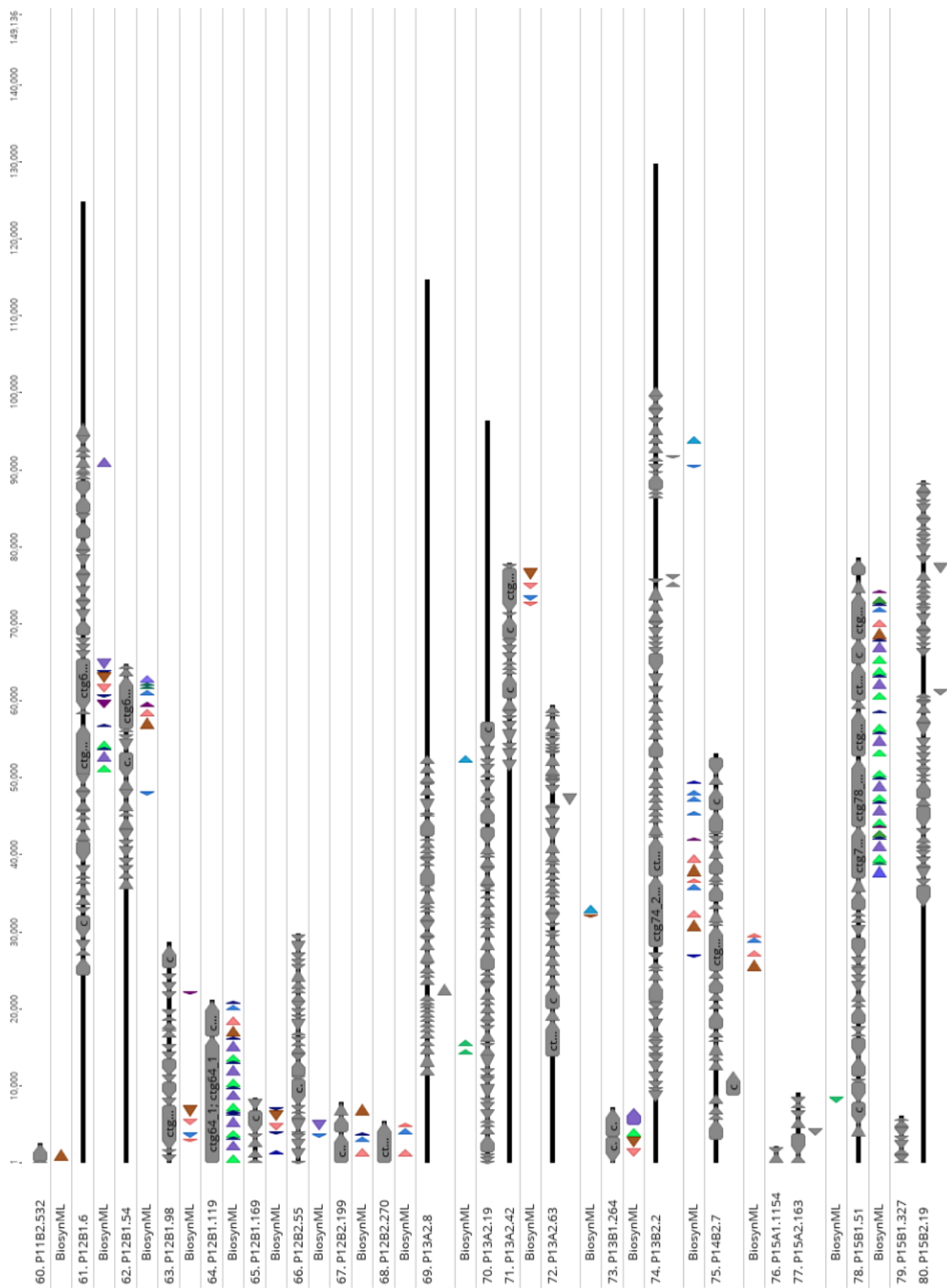


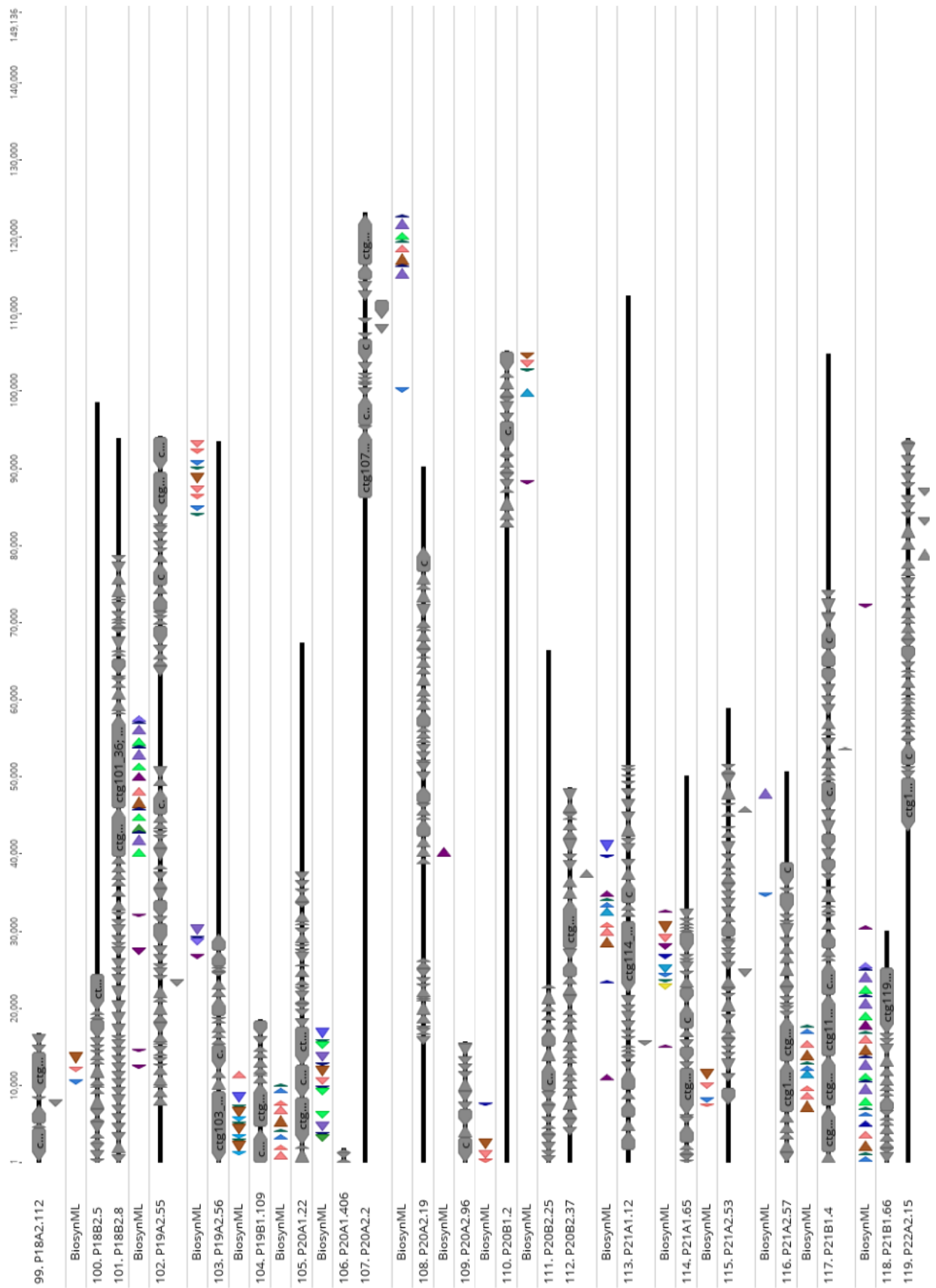
b. Clones predicted to carry NRPS pathways

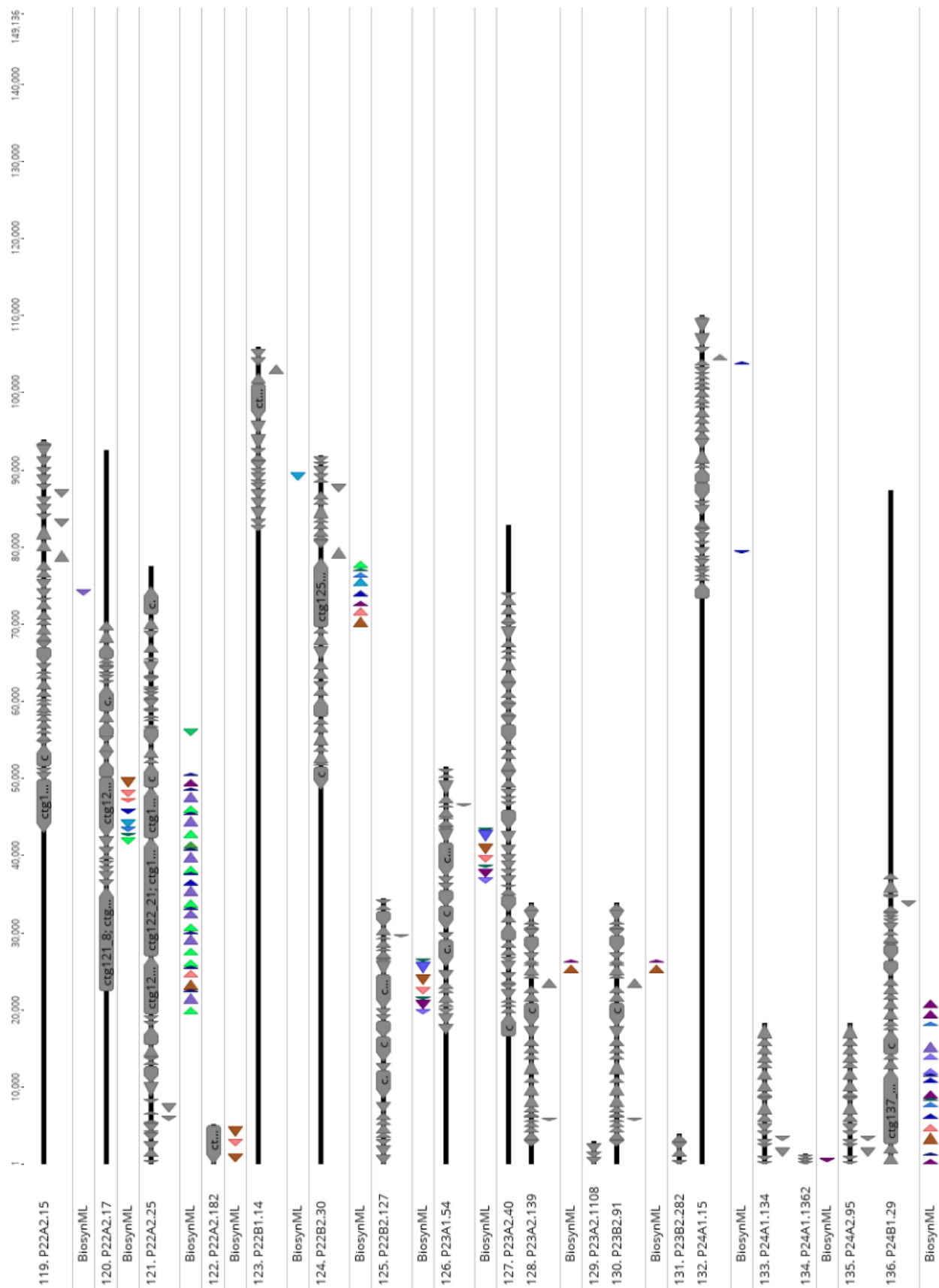


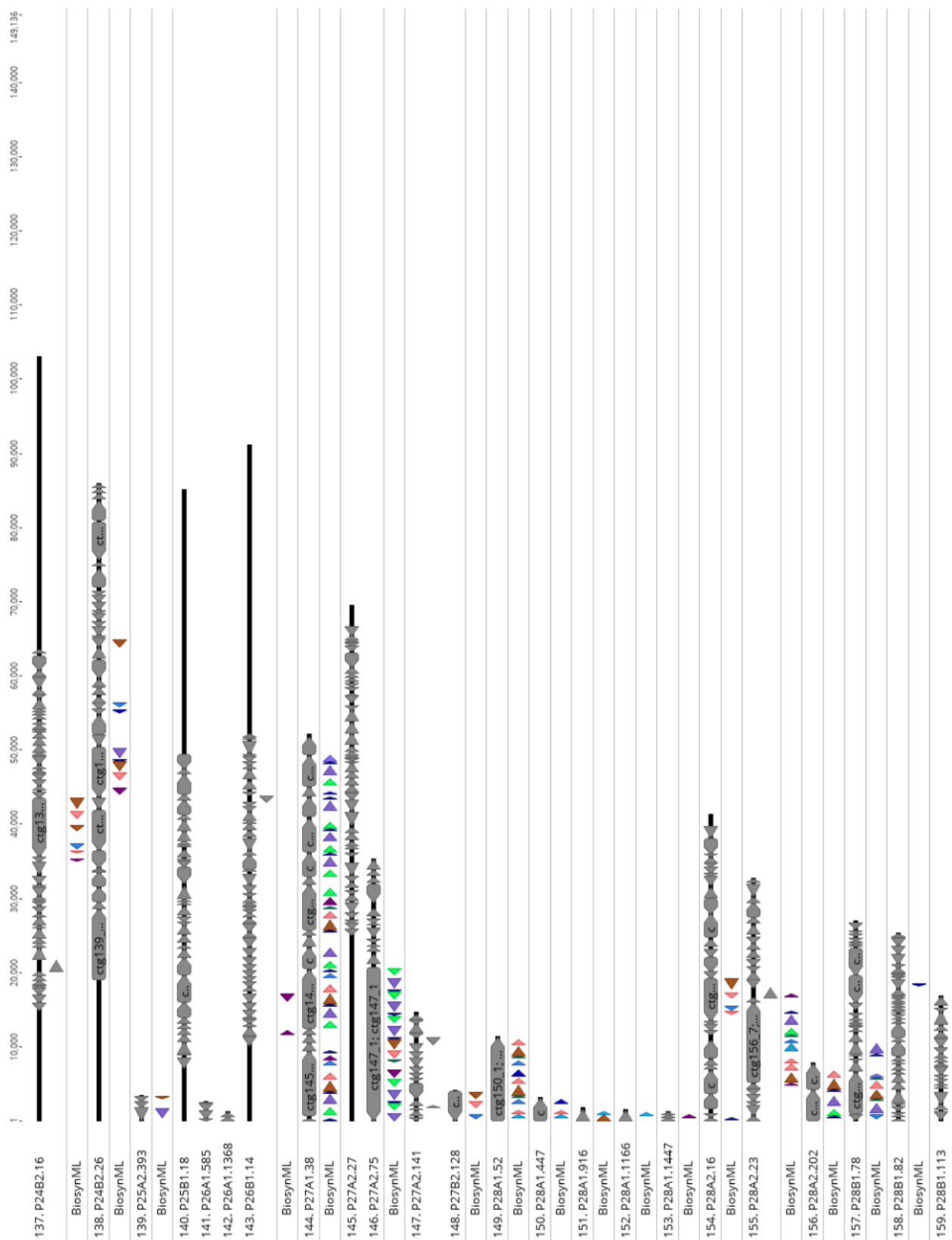


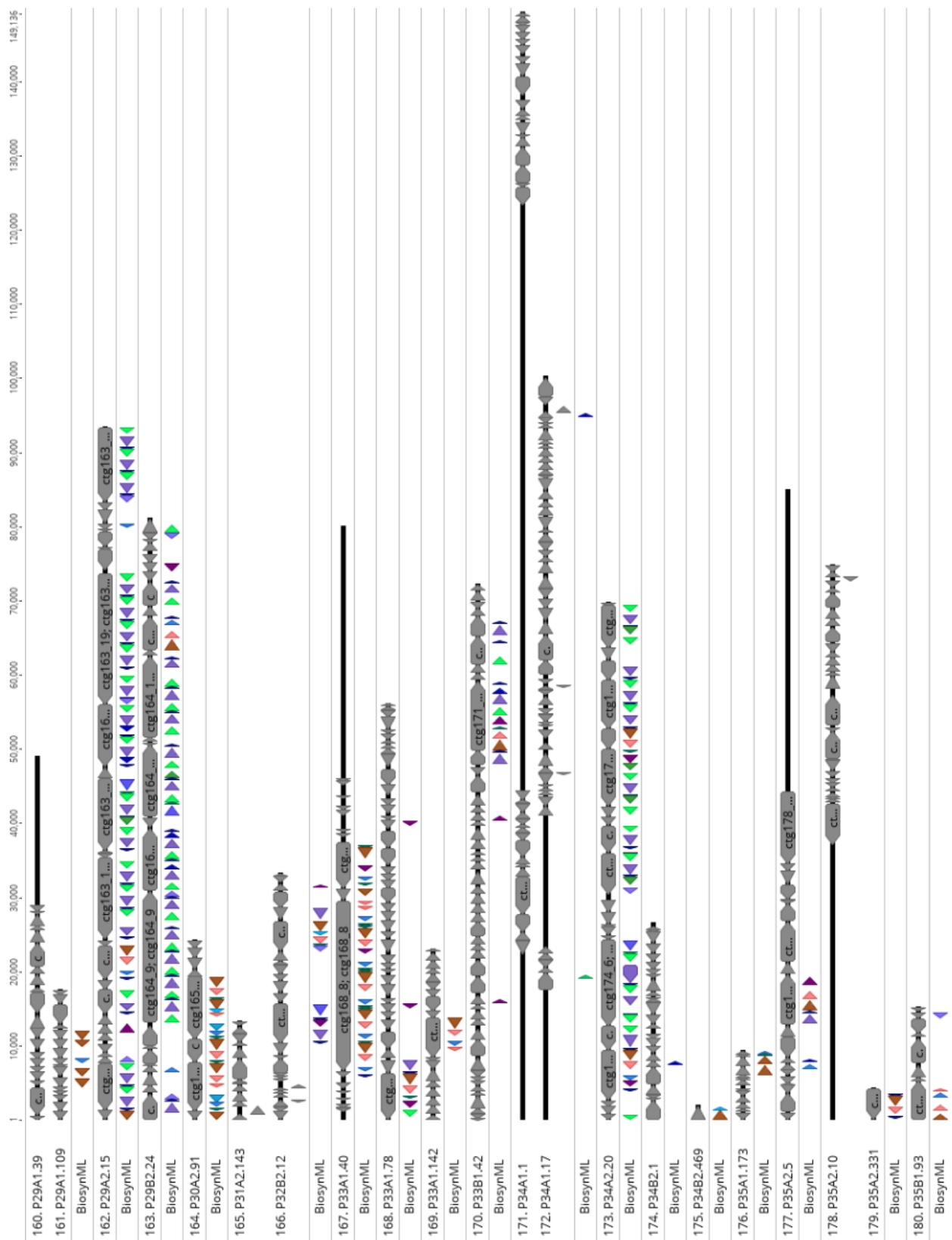


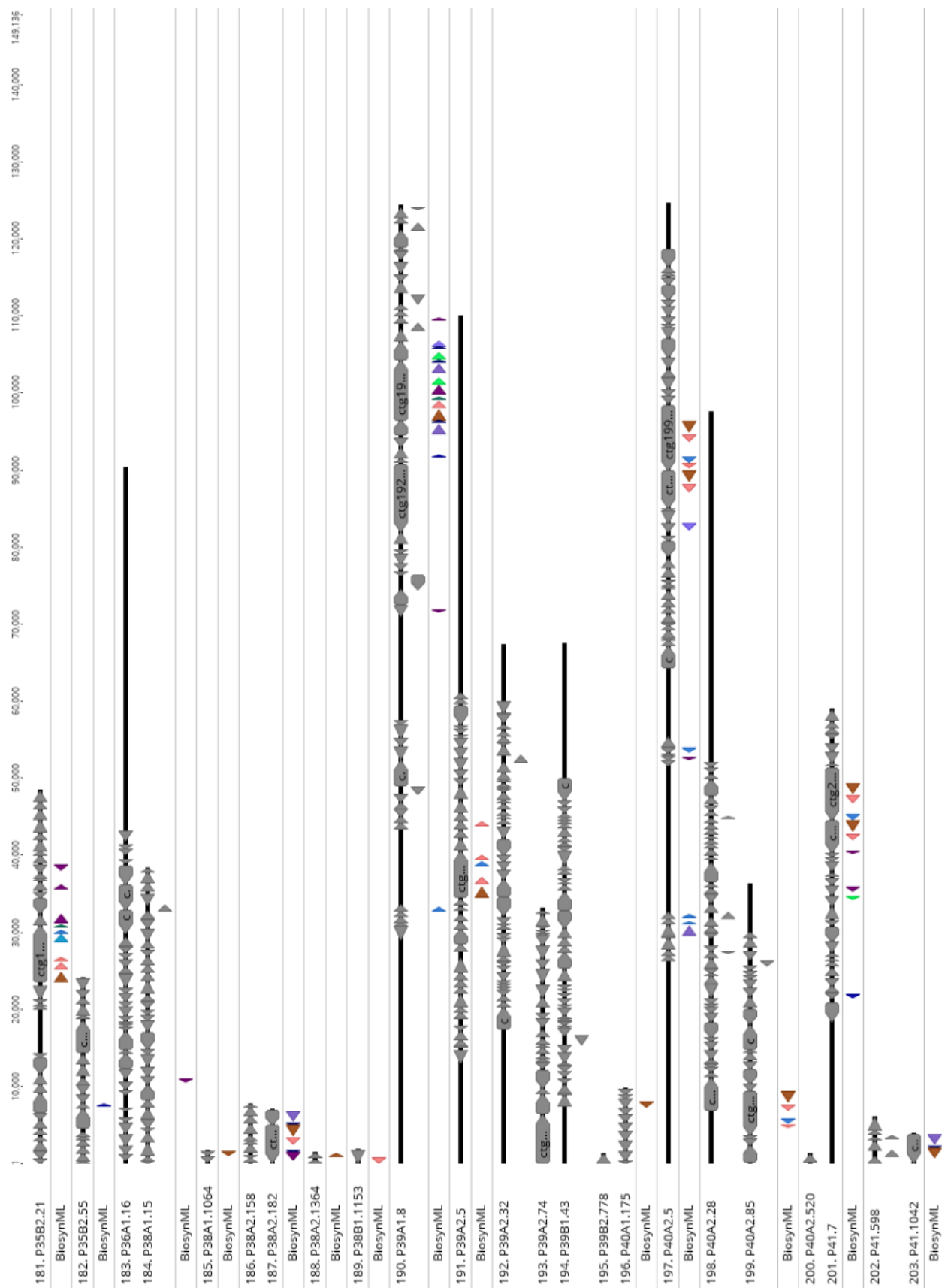


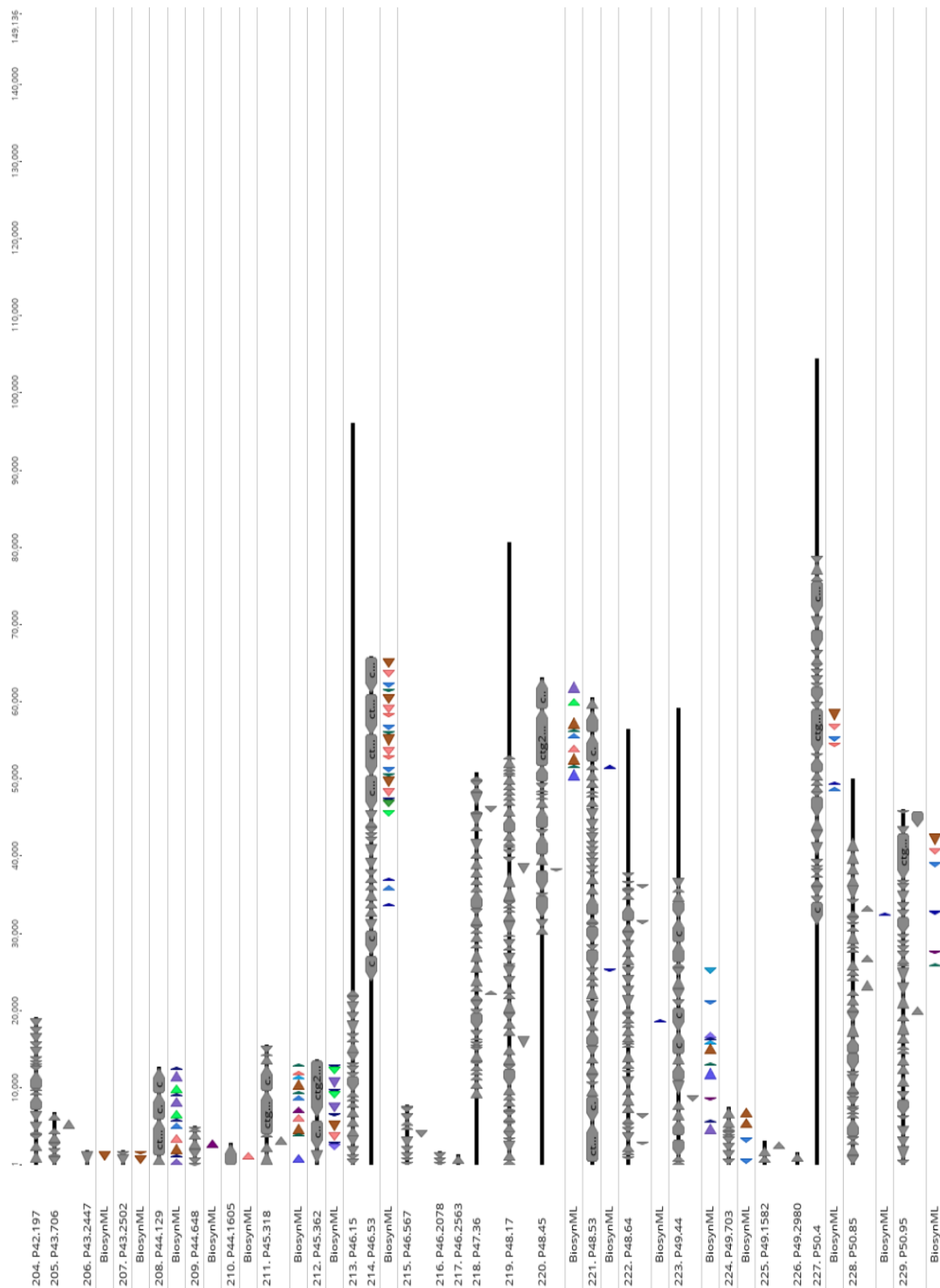












c. Resequenced clones

