PHASE GENOMICS

Application Note

Discover more metabolic features with proximity-guided metagenome deconvolution

The ProxiMeta™ Platform unlocks the functional potential of metagenomic samples by enabling significantly improved annotation. This is accomplished through improvements in metagenome deconvolution, enabled by proximity ligation technology.

Introduction

The analysis of microbial communities is becoming integral to our understanding of complex organisms and environments. Increasingly sophisticated sample preparation, sequencing and computational tools are enabling us to answer questions beyond the identity and relative abundance of the species and strains present in a metagenomic sample. Gene-level information for individual organisms sheds light on microbial evolution and horizontal gene transfer, global and local microbial migration patterns, and enables the identification and characterization of critical metabolic pathways.

Functional analysis of metagenomic samples is hampered by the loss of intracellular contiguity information during sample preparation, caused by the bulk extraction of input DNA. As a result, computational pipelines have to rely on *a priori* knowledge, statistical assumptions, and binning algorithms to reconstruct genomes. During annotation, genes and metabolic modules are either missed or cannot be assigned to the cells from which they originated. This leads to an incomplete and/or inaccurate analysis of metabolic pathways associated with individual or multiple members of a complex microbial community.

To overcome these challenges, Phase Genomics' ProxiMeta Platform[1] employs proximity ligation (Hi-C) data to guide metagenome deconvolution[2] (Figure 1). This technology is able to deconvolve DNA sequences from mixed communities using physical interactions captured prior to cell lysis. This results in significantly improved genome recovery and quality, as well as superior metabolic pathway analysis from metagenomic samples.

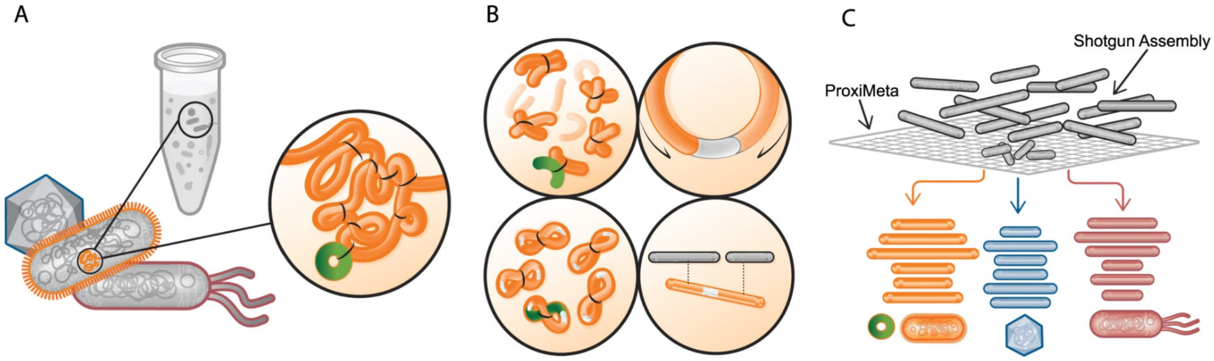

Figure 1. Overview of proximity-guided metagenome deconvolution. A. Proximity ligation (Hi-C) libraries are generated from a mixed microbial sample. Interactions between DNA fragments present in the same cell are captured by crosslinking. **B.** Digestion and ligation creates chimeric junctions that are sequenced and analyzed with short- or long-read shotgun assemblies. **C.** The proximity ligation data provides an additional layer of information used to deconvolve chromosomes and plasmids into complete genomes, with higher accuracy than traditional binning approaches.

Reproduced without change from Stadler T, et al. *ISME J* 2019; 13: 2437–2446. doi: 10.1038/s41396-019-0446-4, under the provisions outlined at http://creativecommons.org/licenses/by/4.0/.T, et al. ISME J 2019; 13: 2437–2446.

Library Preparation and Sequencing

To demonstrate the advantages of the ProxiMeta™ Platform for metabolic discovery, a fecal sample was obtained from a healthy human donor. A single proximity ligation library was prepared using the ProxiMeta Hi-C Kit. Sequencing (2 x 150 bp) was performed on an Illumina® NovaSeq™ 6000 instrument (S4 flow cell). A total of 493,334,300 reads yielded 43 Gb of data after trimming.

For metagenomic shotgun sequencing, DNA was extracted with the ZymoBIOMICS® DNA Miniprep Kit (Zymo Research). Three libraries were prepared:

- For shotgun sequencing on the Illumina platform, a library was prepared using the Nextera® XT DNA Library Preparation Kit (Illumina). A total of 1,113,374,660 reads (2 x 150 bp) was generated on an NovaSeq 6000 instrument (S4 flow cell). This yielded 94 Gb of data after trimming. A 618-Mb assembly was generated with MEGAHIT[3] using default parameters.
- Two libraries were prepared and sequenced on long-read platforms at the University of Idaho Core Facility (www.ibest.uidaho.edu/grc.php). For sequencing on the PacBio® platform, a HiFi SMRTbell® library was prepared and sequenced on the Sequel® II System. A single SMRT® cell yielded 8,439,956 reads, 68.1 Gb of data, and a 294-Mb assembly. For sequencing on the Oxford Nanopore (ONT) platform, a library prepared and sequenced on a single MinION® flow cell using standard protocols. A total of 1,913,344 reads yielded 15 Gb of data and a 339-Mb assembly.

Data Analysis and Results

For each of the three shotgun sequencing approaches, metagenome-assembled genomes (MAGs) were generated and annotated with the ProxiMeta pipeline (proximeta.phasegenomics.com), as outlined in Figure 2.

For benchmarking purposes, data were also analyzed using four conventional binning algorithms (CONCOCT[4], MetaBAT1[5], MetaBAT2[6] and MaxBin 2.0[7]). The outputs from these algorithms were combined and used as the input into DAS Tool[8] to obtain a merged, non-redundant set of bins for metabolic annotation. The number and quality of MAGs generated from the short-read assembly are shown Figure 3, which

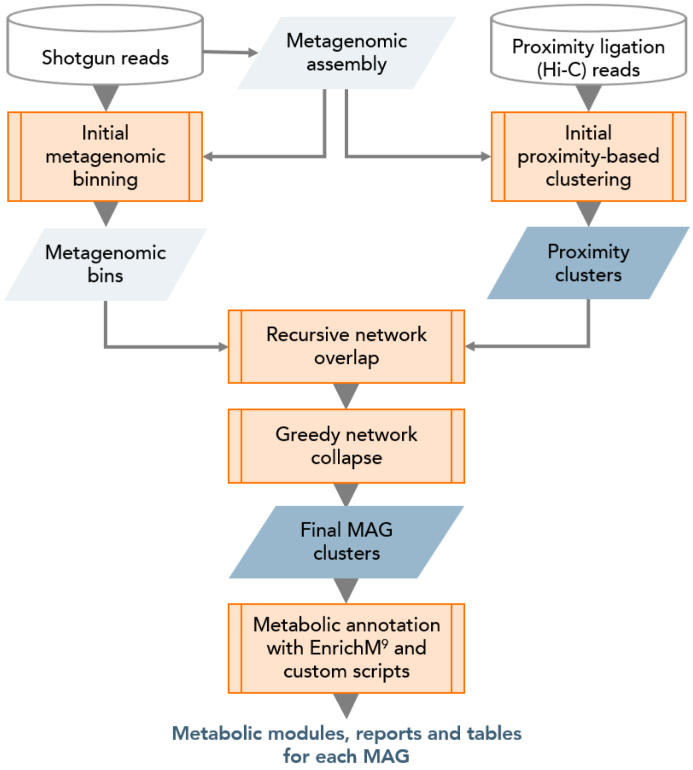
Shotgun reads
Metagenomic assembly
Proximity ligation (Hi-C) reads
Initial metagenomic binning
Initial proximity-based clustering
Metagenomic bins
Proximity clusters
Recursive network overlap
Greedy network collapse
Final MAG clusters
Metabolic annotation with EnrichM[9] and custom scripts
Metabolic modules, reports and tables for each MAG

Figure 2. Overview of the ProxiMeta analysis pipeline. Shotgun reads are assembled into a metagenomic assembly, which is putatively clustered using conventional metagenomic binning approaches in combination with inter-contig Hi-C linkages. Conflicts between the resulting groupings are subsequently resolved, and final MAG clusters are annotated with respect to major metabolic modules.

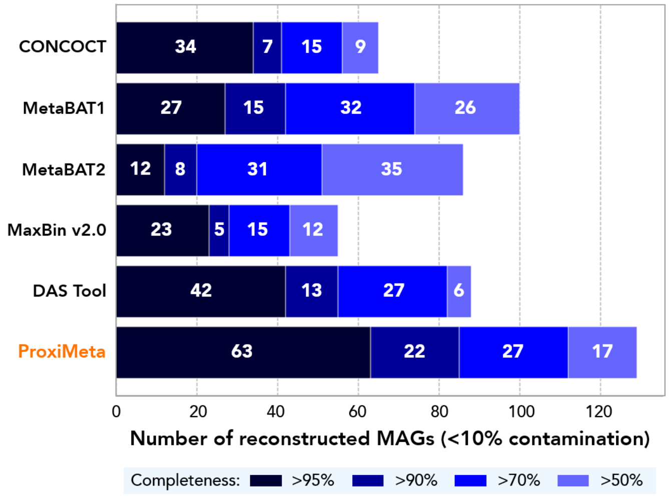

Figure 3. Binning result comparison, showing the number of low contamination MAGs reconstructed from the Illumina® assembly using different binning approaches. Completion thresholds are indicated in different shades of blue. Results from the four binning tools were consolidated and optimized with DAS Tool. Completion and contamination of MAGs was estimated with CheckM.[10]

clearly shows that proximity-guided metagenome deconvolution produces significantly more high-quality MAGs than conventional binning.

To illustrate the benefits of more accurate and complete MAG reconstruction in the context of metabolic pathway analysis, heatmaps were constructed to compare the metabolic modules discovered in ProxiMeta™ MAGs with those identified using shotgun sequencing and conventional binning. The improvements achieved with the ProxiMeta Platform are shown in Figure 4 (heat map for short read assembly), in which each metabolic module (y-axis) was color-coded based on whether it was identified (i) with both strategies (green); (ii) with **ProxiMeta only** (yellow, potential gains); (ii) with **DAS Tool only** (black, potential false positives from conventional binning); or (iv) not at all (blue). A summary of the annotation comparisons for all three shotgun assemblies is given in Figure 5 on the next page, and confirms that the ProxiMeta Platform consistently identified more modules in more MAGs, irrespective of whether the shotgun assembly was derived from short- or long-read sequencing.

To validate these annotations, MAGs were mapped to reference genomes (where available from NCBI), to confirm the presence of genes associated with each module. As shown in Table 1, a high percentage

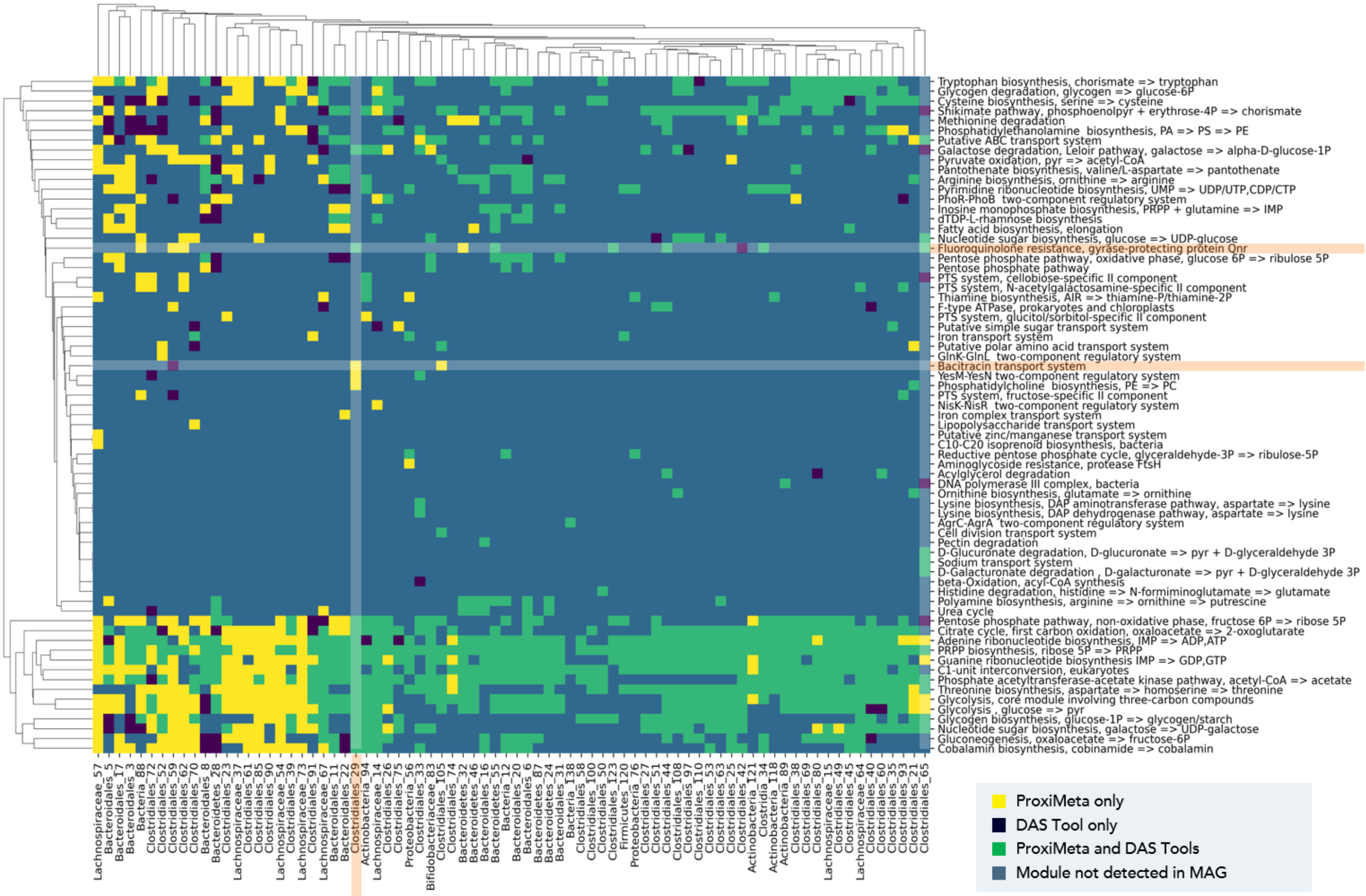

Figure 4. Comparison of metabolic modules identified in the short read assembly. Of the modules discovered with the ProxiMeta Platform only (yellow), 89.5% were confirmed to be accurate (see Table 1). Of the modules identified using binning/DAS Tool only (black), 84.8% were confirmed to be false positives introduced by conventional binning. Highlighted modules provide evidence of antibiotic resistance in several MAGs.

Only high-quality MAGs (completion >50%, contamination <10%) were included in the analysis. The MAGs listed on the x-axis exclude ProxiMeta MAGs for which a corresponding MAG with a sequence overlap ≥50% could not be found in the DAS Tool bin set.

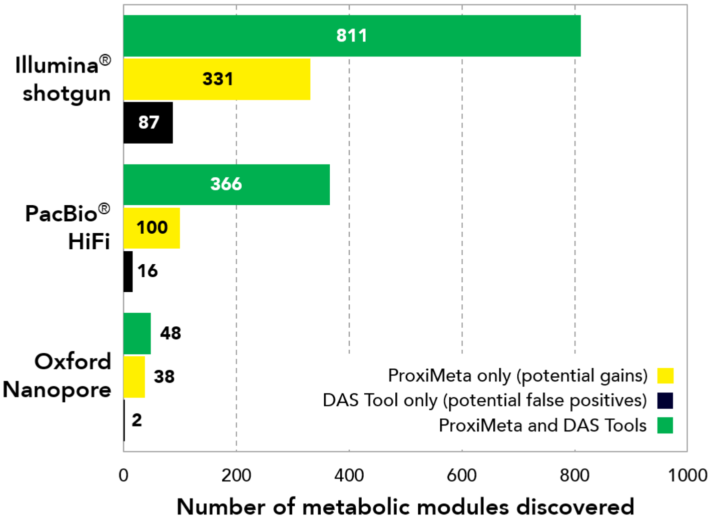

Figure 5. Number of metabolic modules found in MAGs re-constructed from Illumina, PacBio and Oxford Nanopore metagenomic assemblies. Numbers for the Illumina assembly were derived from the heat map in Figure 4. Numbers for the PacBio and Oxford Nanopore assemblies were generated using the same methodology.

Table 1. Summary of metabolic annotation validation.

Assembly	% of ProxiMeta-only modules confirmed to be accurate	% of binning/DAS Tool-only modules confirmed to be inaccurate
Illumina shotgun	89.5% (77 of 86)	84.8% (31 of 37)
PacBio HiFi	74.9% (17 of 23)	No binning/DAS Tool only modules
Oxford Nanopore	52.2% (12 of 23)	100% (1 of 1)

The second number in each set of parentheses correspond to the number of metabolic modules (from the summary in Figure 5) that could be validated (using ≥92% average nucleotide identity (ANI) over ≥50% of the sequence between a MAG and reference genome in the NBCI_nt v4 database). For example, of the 331 ProxiMeta-only (yellow) modules in the Illumina shotgun assembly, 86 modules could be validated and 77 of those (89.5%) were found to be accurately assigned with the ProxiMeta Platform. Similarly, 37 of the 87 binning/DAS Tool-only (black) modules could be investigated, and 31 of those (84.8%) were confirmed to be false (inaccurately identified in the MAG). Validation rates for other samples or sample types may be higher or lower, depending on the novelty and complexity of the sample.

of ProxiMeta™-only modules (potential gains) were confirmed to be accurate, for both the Illumina shotgun and PacBio HiFi assemblies. Support for gains from proximity-guided ligation was less strong for the Oxford Nanopore assembly, presumably due to the inherently lower sequencing fidelity. In addition, the validation process confirmed most of the binning/DAS Tool-only modules to be inaccurate (false positives), indicating that they should be removed from the analysis.

Even though a fecal sample from a healthy donor was expected to be relatively "unremarkable" in terms of metabolic annotation, this study produced evidence of multi-drug resistance. A module associated with fluoroquinolone resistance was identified in six MAGs with both analytical approaches, in three additional MAGs with the ProxiMeta Platform only, and in one MAG as a potential false positive with binning/DAS Tool. In addition, a bacitracin transport system was identified with the ProxiMeta Platform only in two MAGs (and was incorrectly attributed to a third MAG using binning/DAS Tool). Only the ProxiMeta Platform identified **both** antibiotic resistance mechanisms in one of the MAGs (*Clostridiales_29*). These results suggest that the limitations of conventional shotgun sequencing and binning could have considerable consequences in applications such as microbiome analysis or pathogen surveillance.

Visit proximeta.phasegenomics.com to view the MAGs and metabolic analyses generated in this study (Example Reports > Human Fecal Microbiome), or to perform your own analysis

Summary

The ProxiMeta™ Platform is the only commercially available technology designed for the application of proximity ligation data to the deconvolution of complex metagenomic assemblies. By incorporating information about the physical relationship of sequence fragments in the biological sample, the platform allows for more accurate binning than traditional methods that rely purely on shotgun data and statistical approaches.

In this study we have illustrated the key advantages of the ProxiMeta Platform for metabolic pathway analysis. Specifically, improved MAG recovery and quality were confirmed to :

- offer significant gains in the identification of true (validated) metabolic modules, as compared to the best available conventional approach, and
- enable the removal (with high confidence) of inaccurate annotations resulting from lower quality MAGs reconstruction when using conventional binning methods.

It should be noted that higher than recommended Hi-C sequencing coverage was employed for this study. This was done to match the depth of the shotgun assembly (which was also unusually high), and extended the ProxiMeta Platform's binning and annotation improvements to the less abundant organisms in the sample. When focusing on the more abundant members of a complex microbial community, similar advantages can be achieved with significantly less data.

The discovery of metabolic modules associated with antibiotic resistance underlined the power of genome-resolved metagenomics in detecting the reservoirs of antibiotic resistance and identification of novel biosynthetic pathways.

References

1. Capture a complete picture of complex microbial communities, including the moving parts. Phase Genomics Application Note 2020. phasegenomics.com/wp-content/uploads/2020/09/ProxiMeta-Application-Note_Aug-2020.pdf.
2. Burton JN, et al. Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. *G3* (Bethesda, Md.) 2014; 4(7):1339–1346. doi: 10.1534/g3.114.011825.
3. Li D, et al. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 2015; 31(10):1674-1676. doi: 10.1093/bioinformatics/btv033.
4. Alneberg J, et al. Binning metagenomic contigs by coverage and composition. *Nat Methods* 2014; 11(11):1144-1146. doi: 10.1038/nmeth.3103.
5. Kang DD, et al. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *Peer J* 2015. 3:e1165. doi: 10.7717/peerj.1165.
6. Kang DD, et al. MetaBAT2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 2019; 7:e7359. doi: 10.7717/peerj.7359.
7. Wu, Y-W et al. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets, *Bioinformatics* 2016; 32(4):605–607. doi: 10.1093/bioinformatics/btv638.
8. Sieber, CMK, et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol* 2018; 3:836–843. doi: 10.1038/s41564-018-0171-1.
9. Boyd JA et al. Comparative genomics using EnrichM. 2019. In preparation.
10. Parks DH, et. al. Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2014; 25: 1043–1055. doi: 10.1101/gr.186072.114.

Learn more about the ProxiMeta Platform at
phasegenomics.com/products/proximeta/

info@phasegenomics.com
www.phasegenomics.com

Phone: 1-833-742-7436
Twitter: @PhaseGenomics

Unless otherwise stated, data on file.
For Research Use Only. Not for use in diagnostic procedures.
PROXIMETA is a trademark of Phase Genomics, Inc. All other product names and trademarks are the property of their respective owners.