AN EVALUATION OF BINNING METHODS TO RECOVER HUMAN GUT MICROBIAL PAN-GENOMES FROM NON-REDUNDANT REFERENCE GENE CATALOG

BORDERES M.[1,2,3], BOUCINHA L.[1], GASC C.[1], FERRARINI M.[2,4], VINGA S.[5], SAGOT M-F.[2,3]

[1] MaaT Pharma, Lyon; [2] Université de Lyon, Laboratoire de Biométrie et Biologie Évolutive UMR 5558; [3] Erable team, INRIA Grenoble Rhône-Alpes; [4] INSA-Lyon; [5] INESC-ID, Instituto Superior Técnico, Universidade de Lisboa

marie-france.sagot@inria.fr – mborderes@maat-pharma.com

BACKGROUND

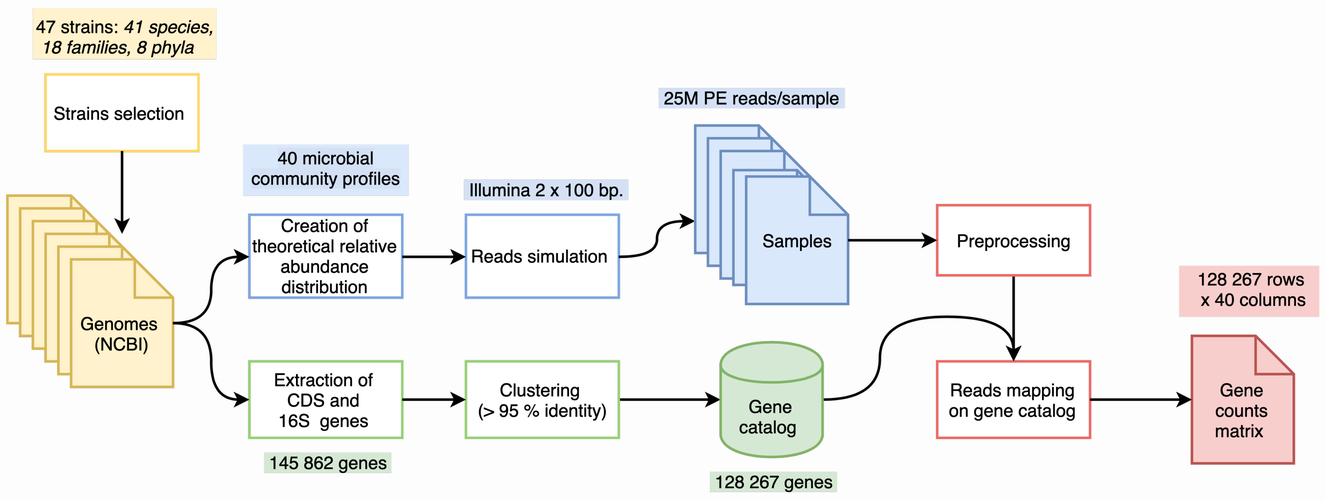
Human gut microbiota exerts functions essential for the maintenance of host physiology. However, a characterization of host-microbiota interactions remains challenging in reference-based quantitative metagenomics analyses. Taxonomic and functional analyses are realized independently, there is no link between genes and species. Although a first set of species-level bins (metagenomic species – MGS, *Nielsen 2014*) was built by clustering co-abundant genes, no reference MGS set is established based on the most comprehensive available human gut microbiota gene catalog – the Integrated Gene Catalog (IGC, *Li 2014*). The published benchmarking results focusing on the reconstruction of individual genomes have highlighted best-performing solutions but do not include methods based on binning co-abundant genes.

In order to identify the most suitable and accurate approach to cluster IGC genes, we benchmarked 13 recently developed or reviewed taxonomy-independent binners implementing abundance-based, hybrid (abundance and composition-based) or integrative approaches.

MATERIALS AND METHODS

Creation of a simulated gene catalog based on IGC construction workflow

To evaluate each binner, we designed a simulated gene catalog by selecting human gut microbial species and based on several characteristics: GC%, codon usage, genome & gene size, closely related genomes etc.



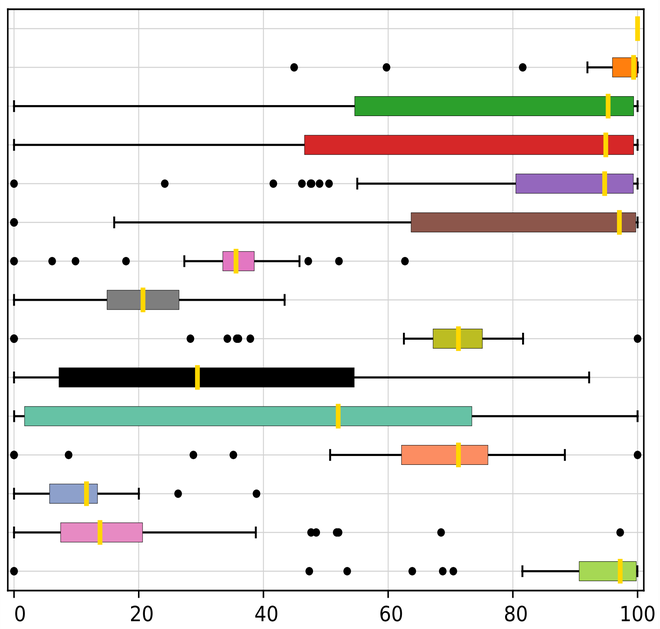
We created a Gold Standard (GS) composed of genes assigned to one or multiple expected bins and then compared this set of expected bins with the bins produced by each binner. Based on our adapted version of the quality assessment tool AMBER (*Meyer 2018*), we computed the following metrics to evaluate binning methods on a non-redundant gene set.

Purity	*Proportion of genes correctly assigned to a predicted bin*
Contamination	*Proportion of genes incorrectly assigned to a predicted bin (1-purity)*
Completeness	*Proportion of shared genes between a predicted and an expected bin*
Average completeness	*Average completeness per bin including unassigned expected bins associated to a completeness of zero*
Generalized NMI (GNMI)	*Extension of Normalized Mutual Information (NMI) to overlapping clusters compatible with conventional NMI values (Lutov 2019)*
Binned gene	*A gene assigned to at least one predicted bin*
High-Quality bins (HQ)	*Predicted bins with > 90% of completeness and < 5 % of contamination (Bowers 2017)*

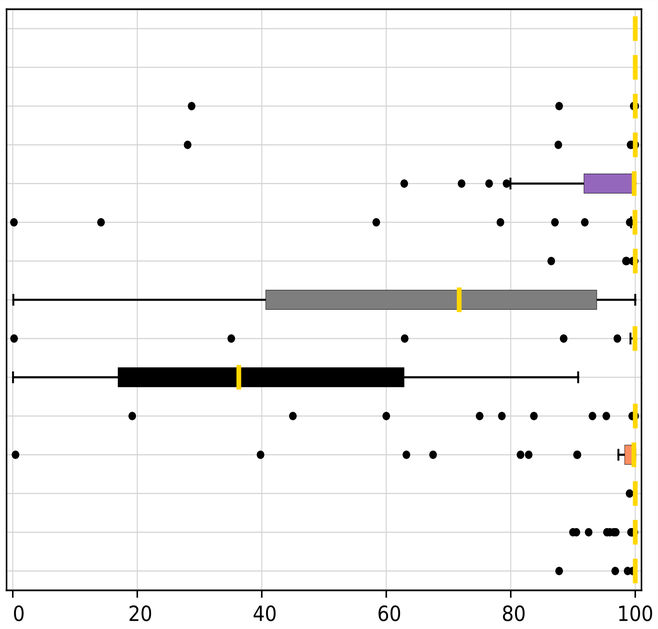
RESULTS

Binner	Publication	Type	Binned genes (assigned >1)	# Bins (# HQ)	Average Compl.	Average Purity
GS	-	◆	100% (3.9%)	41 (41)	100%	100%
GSS (Single assignations)	-	◆	100%	41 (38)	95.6%	100%
MGS-CANOPY_700	*Nielsen 2014*	●	96% (0.2%)	**42** (27)	74.2%	98%
MGS-CANOPY_FP_700	*Ad. from Nielsen 2014*	●	96.9% (0.5%)	44 (27)	71.9%	98%
MSP	*Plaza Oñate 2019*	●	99.3% (20.8%)	54 (25)	84.7%	94.9%
BINSANITY-LC_KMEANS[a,b]	*Graham 2017*	●	**100%**	**41** (26)	79%	93.3%
BMC3C3_1000[b]	*Yu 2018*	■	39.1%	**41**	34%	99.5%
SOLIDBIN_1000	*Wang 2019*	■	39.1%	38	19.9%	66.3%
MAXBIN2_500	*Wu 2016*	■	74.5%	39	62.5%	94.4%
COCACOLA_500	*Lu 2017*	■	**100%**	**40**	34.1%	39.9%
CONCOCT_500	*Alneberg 2014*	■	74.5%	68	42.9%	96.3%
MYCC_500_5p6mer	*Lin 2016*	■	74.5%	**40**	63.6%	92.6%
METABAT2_1500	*Kang 2019*	■	14%	43	10.9%	**100%**
METAWATT3[a]	*Strous 2012*	■	87%	210 (1)	16%	99.8%
DASTOOL_MGS-MSP	*Sieber 2018*	▲	91.6%	39 (**31**)	**87.4 %**	99.5 %

Completeness per predicted bin and unassigned expected bin (%)



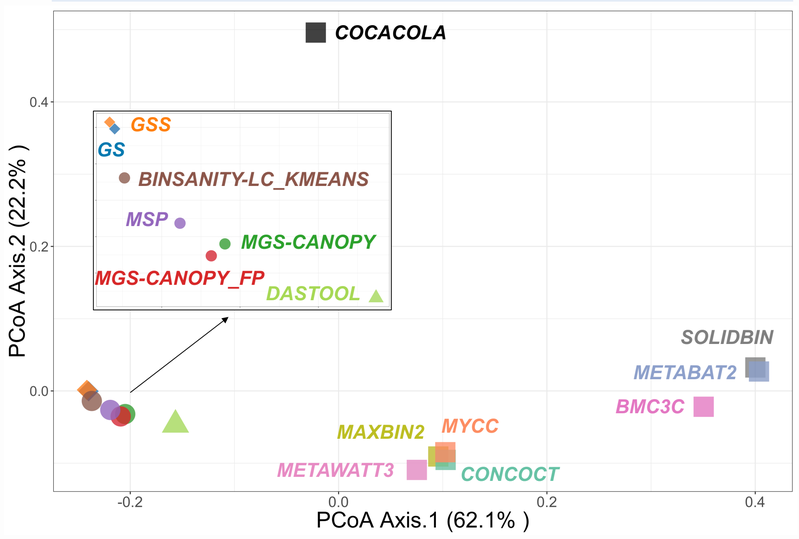
Purity per bin (%)



We launched each binner with different parameters (such as the gene length filter or the kmer size) and selected one of the best benchmarking runs based on the completeness and purity results. The main parameters used are detailed in the suffix of each binner's name, corresponding to min. gene length for all hybrid tools except for MYCC which specifies also kmer size; and to min. bin size for MGS_CANOPY and MGS_CANOPY_FP. Zero values are not represented in the table for HQ bins and genes assigned more than once. Binning type: ◆ - gold standard; ● - abundance-based, ■ - hybrid, ▲ - integrative approach.

[a] Intermediate set of bins generated by the binner : BINSANITY_LC - the set of bins generated by the first clustering step (kmeans) showed better results than the final set of bins; METAWATT - optimization of bins failed.
[b] The number of bins was fixed to 41 for BINSANITY_LC_KMEANS and BMC3C.

PCoA based on GNMI metric computed between each pair of binners



While most binners have an average purity above 90%, abundance-based and integrative binners show a higher average completeness and number of HQ bins (best values reached by DASTOOL). Hybrid binners discard short genes and therefore tend to bin less genes. Moreover, they assign a sequence to a single bin only whereas abundance-based methods assign up to 20.8% (MSP) of the genes more than once. Except for CONCOCT, METAWATT3 and MSP, the number of bins was well estimated by all methods. Similar trends can be observed on the PCoA based on GNMI scores, on which abundance-based and integrative binners are closer to our GS than hybrid methods. It should be noted that the closest point to GS corresponds to a method (BINSANITY-LC_KMEANS) requiring the user to provide the number of bins. As for hybrids binners, a group of three overlapping points representing three hybrids methods sharing high GNMI scores can be distinguished (MAXBIN2, MYCC, CONCOCT).

CONCLUSIONS AND PERSPECTIVES

Quality assessment results show that no hybrid or abundance-based binner performs best on all metrics with our simulated catalog. Overall, the best combination of average purity and completeness per bin was achieved by an integrative method. Ultimately, this approach seems promising but still requires some additional work to find optimal parameters.

References:

Nielsen, HB et al. (2014) *Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes.* Nat Biotechnol. Aug;32(8):822-8. doi: 10.1038/nbt.2939.

Li, J et al. (2014) *An integrated catalog of reference genes in the human gut microbiome*. Nat Biotechnol. Aug;32(8):834-41. doi: 10.1038/nbt.2942.

Meyer, F et al. (2018) *AMBER: Assessment of Metagenome BinnERs*. GigaScience Jun; 7(6): giy069. doi:10.1093/gigascience/giy069.

Lutov, A et al. (2019) *Accuracy Evaluation of Overlapping and Multi-Resolution Clustering Algorithms on Large Datasets.* IEEE International Conference on Big Data and Smart Computing. 1-8. doi: 10.1109/BIGCOMP.2019.8679398.

Bowers, RM et al. (2017) *Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea*. Nat Biotechnol. Aug 8;35(8):725-731. doi: 10.1038/nbt.3893.

This work was performed using the computing facilities of the CC LBBE/PRABI.