ESTIMATING COMMUNITY COMPOSITION OF UNDERSTUDIED MICROBIOMES THROUGH NOVEL SPECIES GENOME ASSEMBLIES

Jonathan Rondeau-Leclaire*, Pierre-Étienne Jacques and Isabelle Laforest-Lapointe

Département de Biologie, Faculté des Sciences, Université de Sherbrooke.

*Corresponding author : jonathan.rondeau-leclaire@usherbrooke.ca

1. UNDERSTUDIED MICROBIOMES ABOUND

Most microbiome studies focus on humans, especially the gut. Microbes living elsewhere are underrepresented in genomic databases. Hence, community composition estimation is confined to **known species**, undermining the breadth of microbiome research. To better study microbiomes, unknown species should be accounted for.

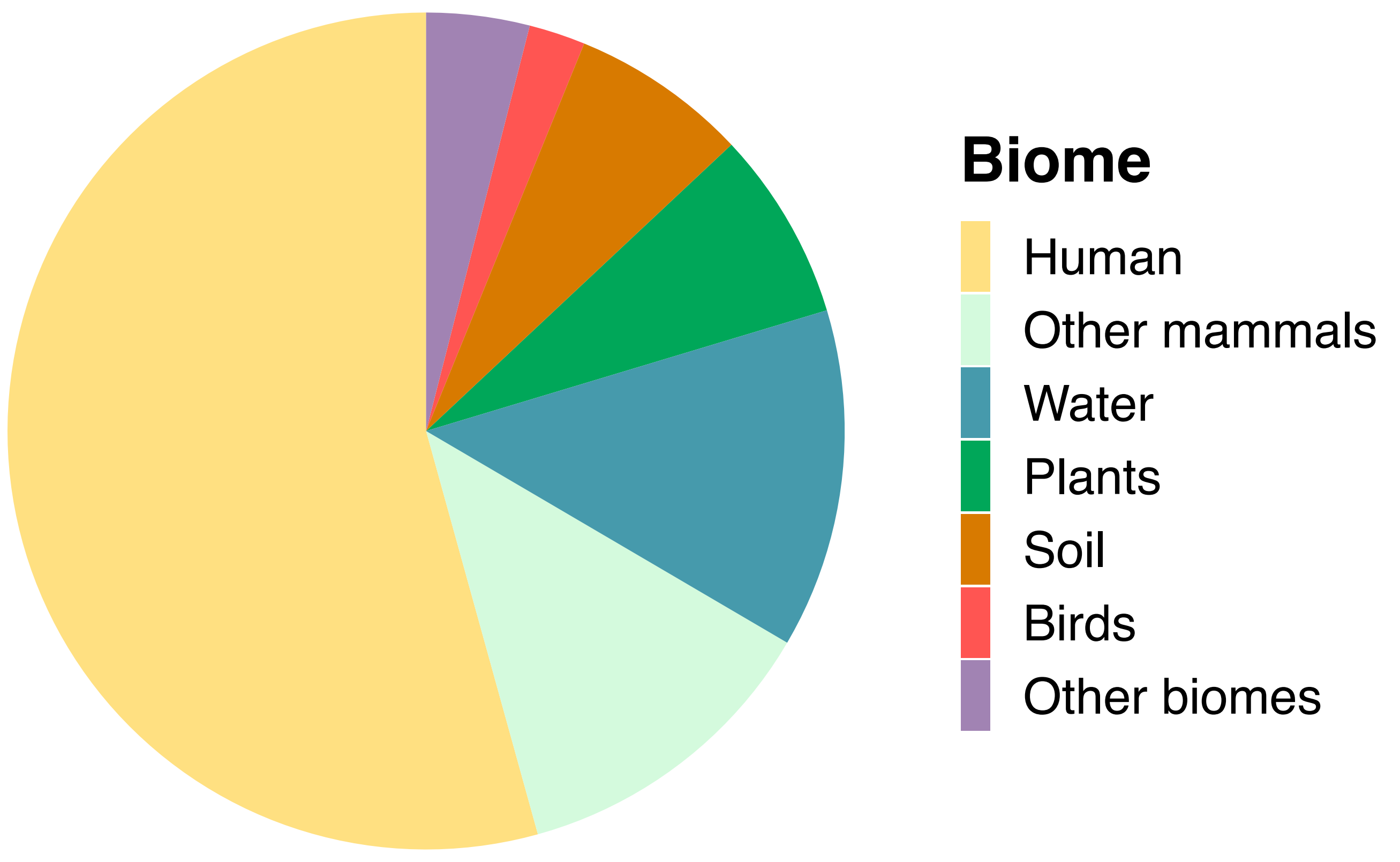


Fig. 1. Environmental and host-associated samples in MGnify database by biome (n = 393,469)

2. SPECIES LACKING A REFERENCE GENOME

Identified through *de novo* assembly of samples into **metagenome-assembled genomes** (MAGs), they can be added to reference databases to better estimate sample composition.

A study-specific effort is needed for understudied microbiomes !

3. OBJECTIVES

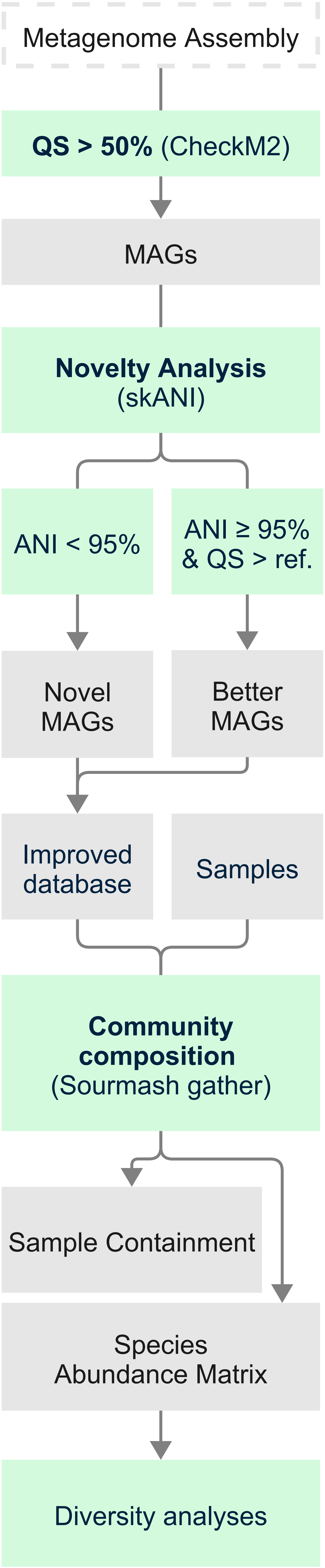
Better estimate community composition by…

1. Assembling metagenomes to find **novel species-level MAGs** and MAGs of better quality than references genomes;
2. Evaluating the effect of adding these to a reference database on increasing the **containment** therein of the samples from which the MAGs were assembled.

METHODS

1. MAG quality score (**QS**; Completeness – 5 × Contamination) > 0.50
2. Species-level MAGs is **novel** if average nucleotide identity (**ANI**) to every known representative genome is **< 95 %** (using skANI on GTDB r214)
3. MAGs with **>= 95 % ANI** are substituted if better QS than reference
4. Estimate sample containment* and community composition with default and improved reference database (Sourmash gather)
5. Explore effects of improvement on sample diversity (Shannon, rarefaction)

***Jaccard containment :** Abundance-weighted intersection of sample sequences *k*-mers and reference database *k*-mers.


Metagenome Assembly

QS > 50% (CheckM2)

MAGs

Novelty Analysis (skANI)

ANI < 95%

ANI ≥ 95% & QS > ref.

Novel MAGs

Better MAGs

Improved database

Samples

Community composition (Sourmash gather)

Sample Containment

Species Abundance Matrix

Diversity analyses

4. NOVEL AND BETTER MAGs IDENTIFIED FROM TWO BIOMES

	Human saliva 26 samples (13 persons)	**Boreal mosses** 131 samples (4 species)
Genome bins	488	260
Species-representative MAGs	130	153
Good quality MAGs (QS > 0.50)	110	107
Novel MAGs (ANI < 95%)	**12**	**102**
Non-novel species (ANI ≥ 95%)	98	5
Better MAGs (QS > ref. genome QS)	**20**	**1**

5. IMPROVEMENT IN SAMPLE CONTAINMENT Fig. 2A

Novel MAGs **heavily increase** sample containment in Moss dataset (mean: **4.84-fold** ± 2.46)

Better MAGs **marginally increase** sample containment in Saliva dataset (mean: **1.04-fold** ± 0.04)

Are samples worth assembling for the marginal gain of including better MAGs ?

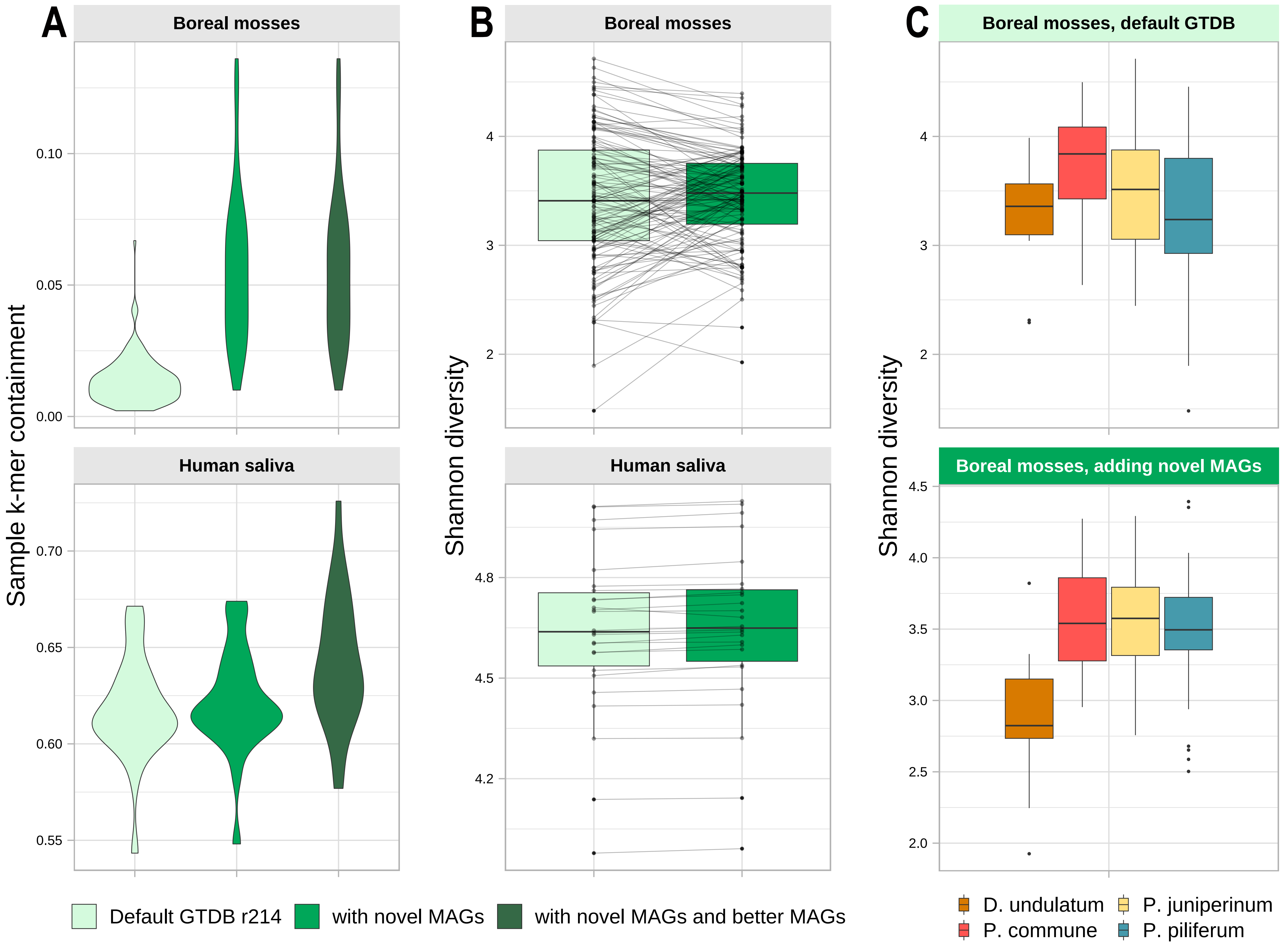


Fig. 2. Effect of reference genome database improvement on sample containment (A) and taxonomic diversity analysis (B) across two biomes, and across sample metadata for boreal moss samples (C).

6. EFFECTS ON DIVERSITY ESTIMATION Fig. 2B,C

For boreal mosses, adding novel MAGs **randomly affects diversity** (mean change 2.7 % ± 16.7 %), **nearly halves** diversity variance (from 0.37 to 0.20), and **quadruples** diversity variance explained by Host species through linear regression (R^2: 0.04, $p < 0.05$ to 0.18, $p < 10^{-5}$).

Including novel MAGs reduces noise and could provide a better estimate of microbial diversity in understudied microbiomes.

We acknowledge the support of :

CRSNG NSERC

Fonds de recherche Québec

Fonds d'appui à l'engagement étudiant (FAEE) de l'Université de Sherbrooke

REFERENCES

1. Gurbich, T. A. et al. MGnify Genomes: A Resource for Biome-specific Microbial Genome Catalogues. Journal of Molecular Biology 435, 168016 (2023).
2. Titus Brown, C. & Irber, L. sourmash: a library for MinHash sketching of DNA. JOSS 1, 27 (2016).
3. Shaw, J. & Yu, Y. W. Fast and robust metagenomic sequence comparison through sparse chaining with skani. Nat Methods 20, 1661–1665 (2023).

We thank T. A. Gurbich and R. D. Finn from the EMBL-EBI Microbiome Bioinformatics team for the ideas and guidance that inspired this project.