

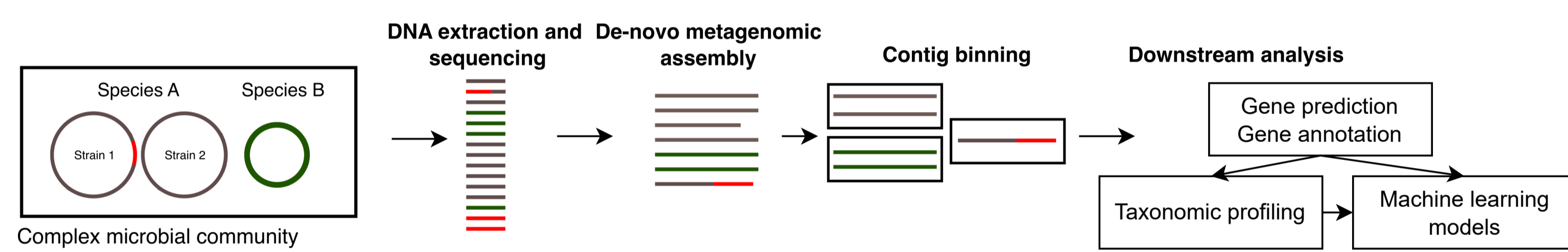
Comparative evaluation of short-read, long-read, and hybrid assemblies for MAG recovery in human fecal metagenomes

Baptiste HENNECART^{1,2}, Eugeni BELDA^{1,3}, Florian PLAZA-OÑATE², Raynald DE LAHONDES², Jean-Daniel ZUCKER^{1,3} and Edi PRIFTI^{1,3}

¹IRD, Sorbonne Université, UMMISCO, Bondy, France - ²GMT Science, Paris, France - ³INSERM, NutriOmique, AP-HP, Sorbonne Université, Paris, France | email: {baptiste.hennecart, edi.prifti}@ird.fr

1 CONTEXT

Shotgun metagenomic sequencing, assembly and contig binning enable the reconstruction of metagenome-assembled genomes (MAGs) from environmental DNA. Yet, strain-level genome assembly remains challenging, particularly for complex microbial communities. The literature highlights the advantages of long-read sequencing, particularly in improving contiguity and contig length. Despite this, short-read sequencing is typically favored in metagenomics due to its superior accuracy and throughput, accessibility and mature bioinformatics pipelines. Hybrid strategies have also emerged as a promising solution, which combines the strengths of different technologies to optimize genome recovery [2,3]. Here, we evaluate the performance of short-read, long-read, and hybrid assemblies, followed by contig binning, for recovering medium- and high-quality MAGs from human fecal microbiomes. To do so, we simulated metagenomic samples based on real fecal microbiome profiles ($n = 50$) and applied a standard de novo workflow, including assembly, binning, dereplication, and filtering.



3 ASSEMBLY STATISTICS (C-E)

Contiguity and fragmentation

- LR assemblies produce the longest contigs, followed by hybrid assemblies.
- SR assemblies are more fragmented but result in larger total assembly sizes.
- Hybrid and long-read assemblies show less fragmentation, but smaller total sizes.

Impact of metagenome complexity

- In richer and more complex metagenomes, all assemblies become more fragmented and larger overall.
- This pattern holds regardless of the sequencing strategy.

Genome fraction

- Overall, the genome fraction is low, indicating limited recovery of original reference genomes.
- SR and hybrid assemblies achieved the highest genome fractions, with a few exceptions.

4 MAG RECOVERY (F-K)

Bins matched to original genomes using Skani (>97% ANI).

SR assemblies gave the highest F1-scores* per sample

- Indicates best species/strain-level recovery.

No method recovered >1 strain of the same species within a single sample → justified bin pooling across samples for dereplication.

Out of 732 reconstructed strains:

- 244 had highest N50 from hybrid assemblies.
- 176 had lowest contamination.
- 117 were the most complete.
- Yet, majority of high-quality MAGs** came from short-read assemblies.

Some strains lost after dereplication (e.g., all *E. coli* strains → only one retained). May reflect ANI threshold settings.

Suggest:

- Increase dereplication ANI.
- Align ANI estimation between dRep and Skani.

5 CONCLUSIONS

Study Focus

- Comparison of de-novo **short-read (SR)**, **long-read (LR)**, and **hybrid** metagenomic assembly.
- Target: **MAG recovery** from fecal metagenomes, with **strain-level resolution**.
- Based on **simulated datasets** reflecting real microbial abundance.

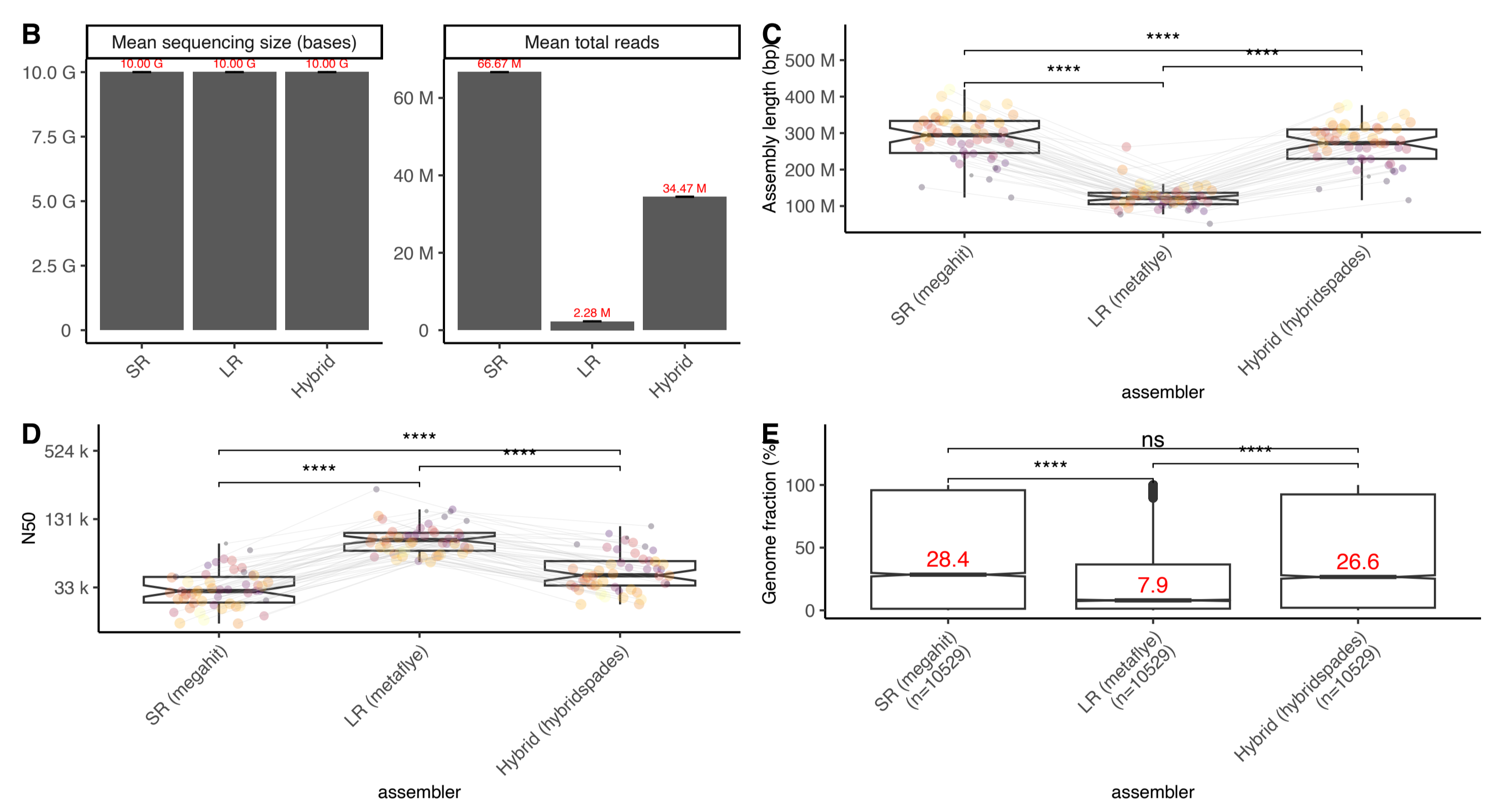
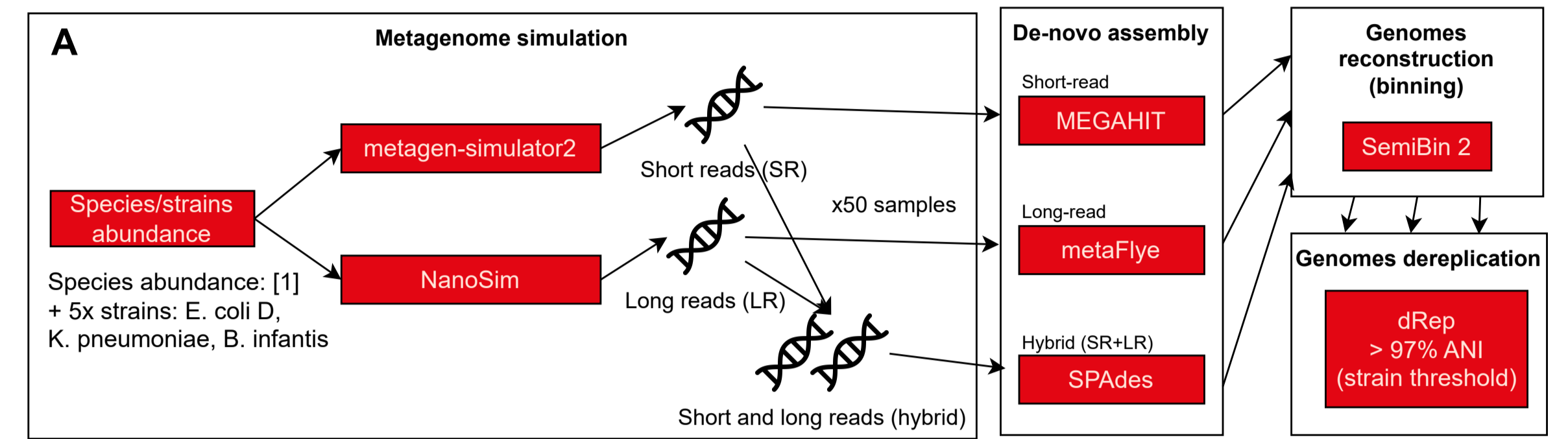
Findings

- LR assemblies
 - Highest **contiguity** and **contig length**.
 - But **smaller total assembly sizes**.
- Hybrid assemblies
 - Highest **number of bins per sample**.
 - But **SR assemblies** recovered more **non-redundant medium- and high-quality MAGs** after dereplication.
- Dereplication trade-off
 - Some strains assembled across samples/methods were lost
 - Suggests **current thresholds may be too strict**.

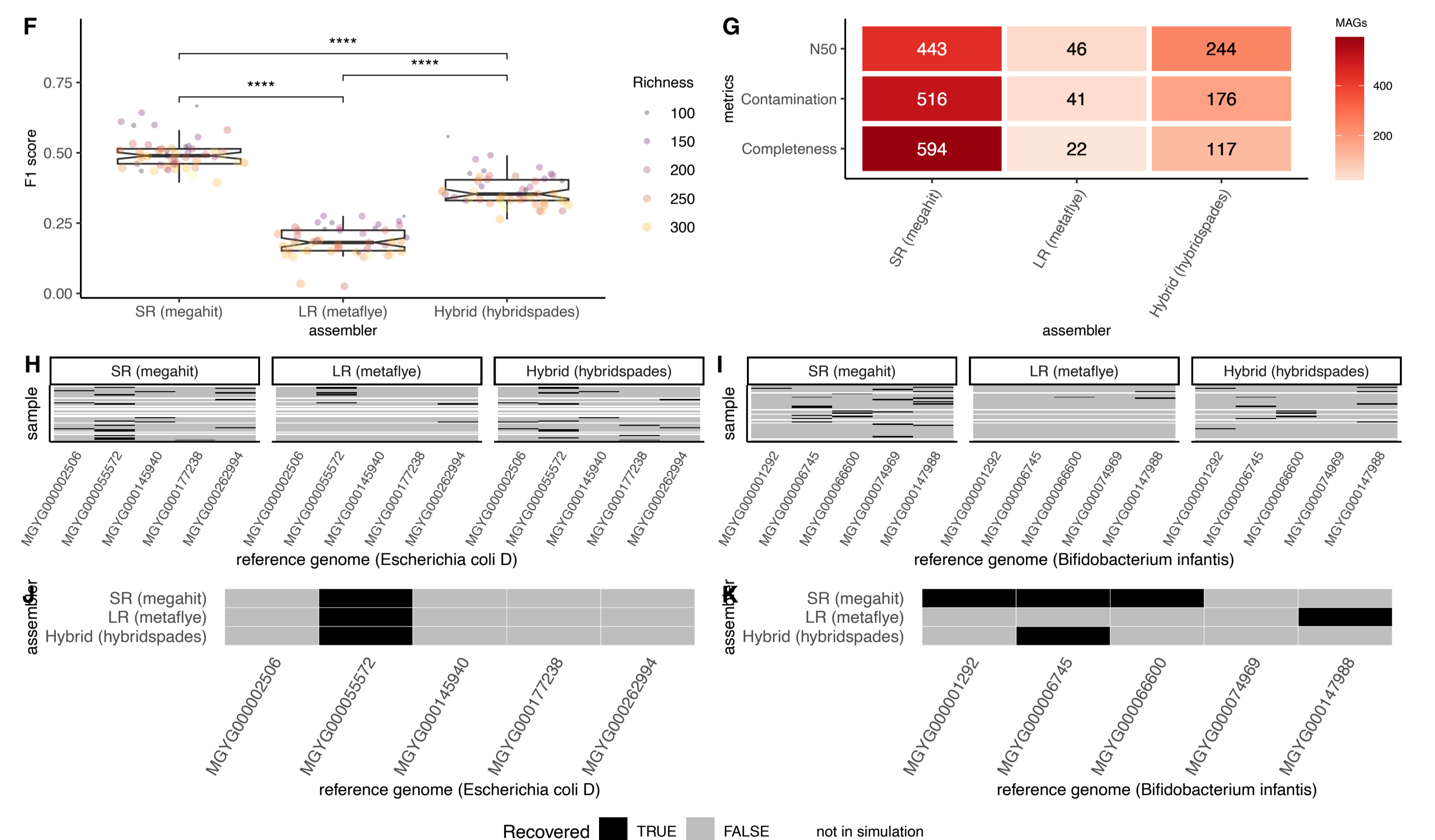
Take-home message

- **SR** sequencing and assembly remains **effective** for MAG recovery.
- **Hybrid assemblies** still face **binning and dereplication challenges**.
- There is a need for:
 - Improvement of metagenomic **hybrid/LR assembly**.
 - Better **binning tools** adapted to hybrid/LR data.
 - **Optimized dereplication thresholds**.

2 METHOD (A-B)



A. Methodology. B. Sequencing depth of simulated samples, here in mean bases pair and in mean total reads by sample (LR: long-read, SR short-read, Hybrid: short and long-read downsized SR and LR samples). C-E. Boxplots of total size of the assembly (sum of contigs length), N50 (contiguity metrics), and genome fraction (% of bases of the assembly mapping back to references). The size and color of point depict sample's richness in terms of species. ns: $p > 0.05$, *: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$, ****: $p \leq 0.0001$



F. Recovery performance of reference genomes per sample and assembly strategy before dereplication. Bins were classified as true positives (TP), true negatives (TN), false positives (FP), or false negatives (FN) based on ANI comparisons with references. F1-scores* were then computed per sample and assembly approach. G. Number of best bins by assembly approach after dereplication of MAGs at 97% ANI level. H-I. Recovery of reference genomes by sample and assembly strategy, before dereplication, focusing on species for which multiple strains were simulated (here *E. coli* D, *B. infantis*). J-K. Same as H-I, but after pooling bins of all samples and dereplicating and filtering. ns: $p > 0.05$, *: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$, ****: $p \leq 0.0001$

