

# Compare of long-read and short read sequencing

HO PHU QUY

[Order Article Reprints](#)[Open Access](#)[Article](#)

## Improving Bacterial Metagenomic Research through Long-Read Sequencing

by Noah Greenman<sup>1</sup>  , Sayf Al-Deen Hassounah<sup>1</sup> , Latifa S. Abdelli<sup>2</sup> , Catherine Johnston<sup>1</sup>  and Taj Azarian<sup>1,\*</sup>  

<sup>1</sup> College of Medicine, University of Central Florida, Orlando, FL 32827, USA


<sup>2</sup> Department of Health Science, College of Health Professions and Sciences, University of Central Florida, Orlando, FL 32816, USA

\* Author to whom correspondence should be addressed.

*Microorganisms* **2024**, *12*(5), 935; <https://doi.org/10.3390/microorganisms12050935>

Submission received: 9 April 2024 / Revised: 22 April 2024 / Accepted: 25 April 2024 / Published: 4 May 2024

(This article belongs to the Special Issue **Gut Microbiota: Metagenomics to Study Ecology**)

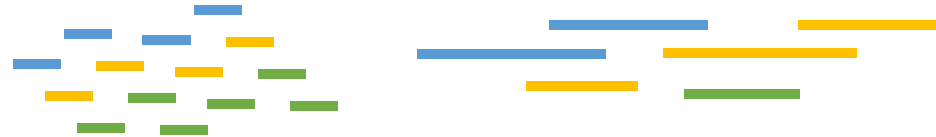
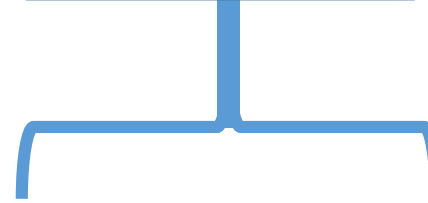
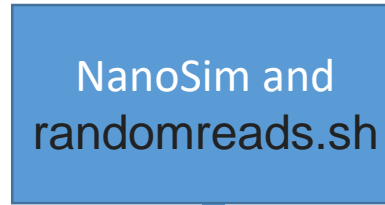
[Download](#) [Browse Figures](#)[Versions Notes](#)

### Abstract

Metagenomic sequencing analysis is central to investigating microbial communities in clinical and environmental studies. Short-read sequencing remains the primary approach for metagenomic research; however, long-read sequencing may offer advantages of improved metagenomic assembly and resolved taxonomic identification. To compare the relative performance for metagenomic studies, we simulated short- and long-read datasets using increasingly complex metagenomes comprising 10, 20, and 50 microbial taxa. Additionally, we used an empirical

<https://doi.org/10.3390/microorganisms12050935>

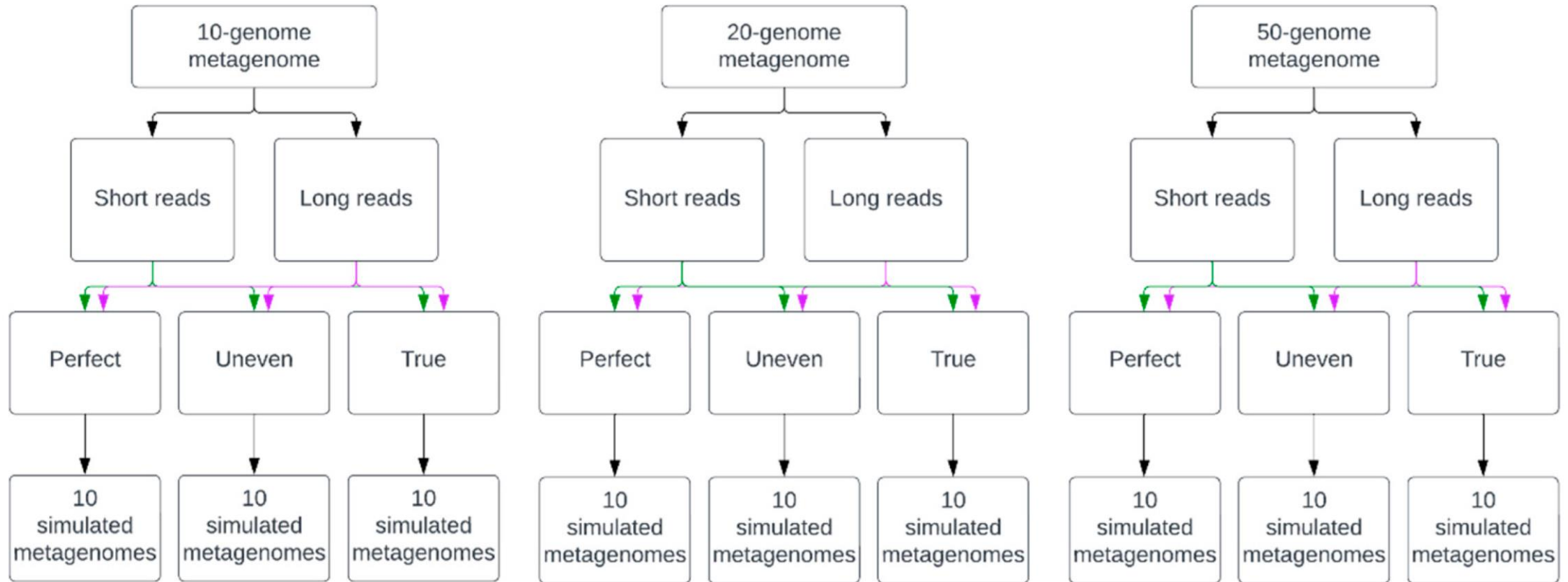
# Materials and Methods



- (1) genomes must be part of the latest RefSeq database,
- (2) they must have a complete genome assembly.
- (3) all anomalous assemblies are excluded.

- “Perfect” denotes that simulated data have no errors and abundances are evenly distributed for each organism in the metagenome.
- “Uneven” denotes that simulated data have no errors but variable abundances for each organism.
- “True” denotes that simulated data have simulated sequencing errors based on the read type, as well as variable abundances for each organism.

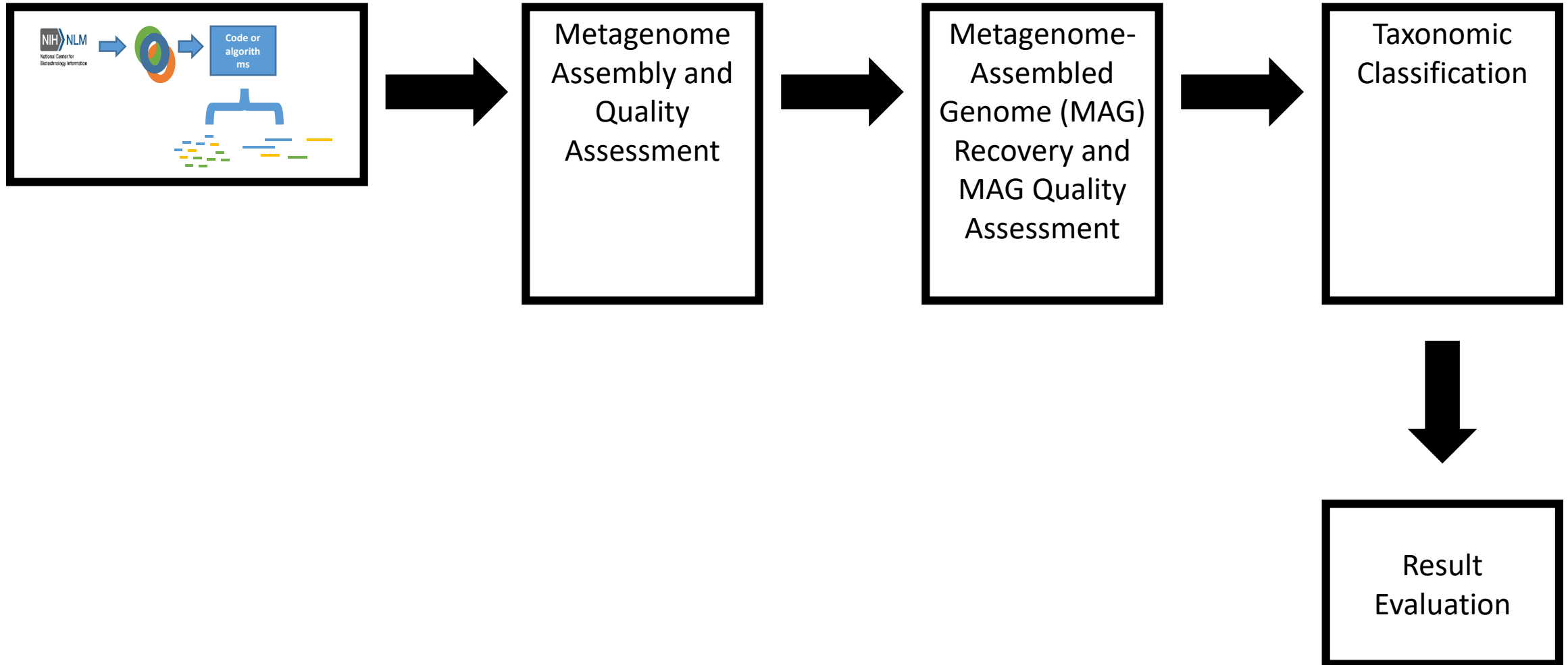
# Materials and Methods



ONT's MinION™: <https://github.com/bcgsc/NanoSim>

Illumina's™ HiSeq™ sequencer: <https://github.com/BioInfoTools/BBMap/blob/master/sh/randomreads.sh>

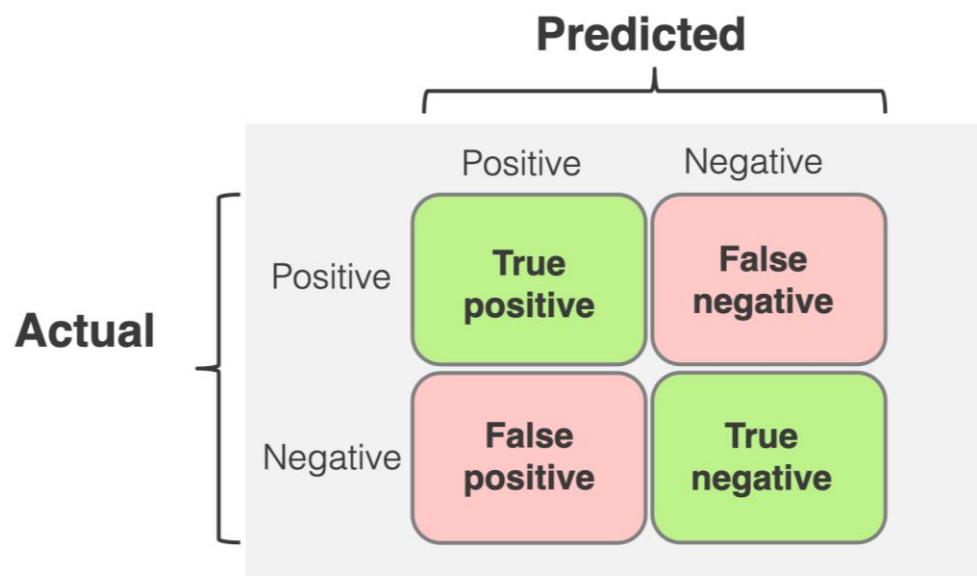
# Materials and Methods



# Materials and Methods

Tool	Version	Function in the Research
Flye	2.9-b1768	Long-read assembly using long reads for metagenomic samples.
SPAdes	3.15.5	Short-read metagenomic assembly with custom kmer sizes and PHRED quality adjustments.
metaQUAST	5.0.2	Quality assessment of metagenomic assemblies (genome fraction, NGA50, misassemblies).
minimap2	2.24-r1122	Long-read alignment using nanopore sequence-specific parameters for read mapping to assemblies.
seqtk	1.4-r122	Short-read header shortening for compatibility with downstream tools.
bwa-mem2	2.2.1	Short-read mapping to assemblies for generating BAM files.
Samtools	1.16.1	Conversion of SAM to BAM files and sorting of BAM files.
CoverM	0.6.1	Estimation of coverage for both long and short reads.
MetaBAT 2	2.15	Binning of contigs based on coverage information to generate metagenome-assembled genomes (MAGs).
CheckM2	1.0.1	Assessment of MAG completeness and contamination based on universal bacterial gene markers.
Kraken2	2.1.2	Taxonomic classification of metagenomic assemblies based on short and long reads.
Bracken	2.8	Estimation of relative abundance of taxa at genus and species levels from Kraken2 output.
Bandage	0.8.1	Visualization of assembly graphs for visual assessment of assembly quality.
SciPy	1.10.1	Used for linear regression analysis in comparing predicted versus actual abundance of taxa.
statannotations	N/A	Statistical comparison tool used for analyzing significance in precision, recall, F-scores, and MAG recovery rates.
Mann–Whitney U Test	N/A	Statistical test applied to compare MAG recovery and classification results between short- and long-read assemblies.

## Confusion matrix



$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

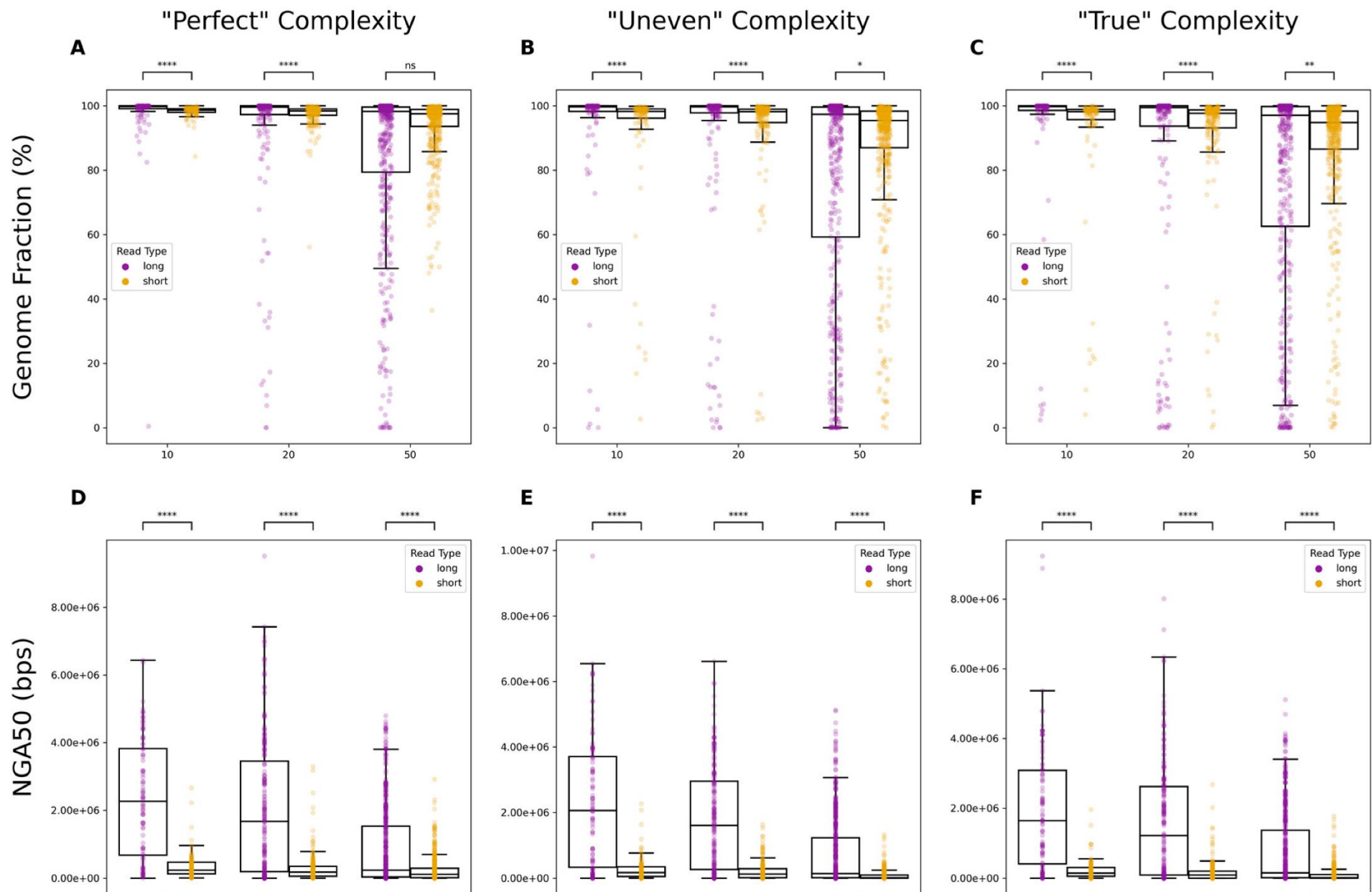
$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

ns: non-significant  $p$ -value ( $p > 0.05$ ), \*:  $p \leq 0.05$ , \*\*:  $p \leq 0.01$ , \*\*\*\*:  $p \leq 0.0001$ .



# Result: Comparison of Metagenomic Assembly Completeness

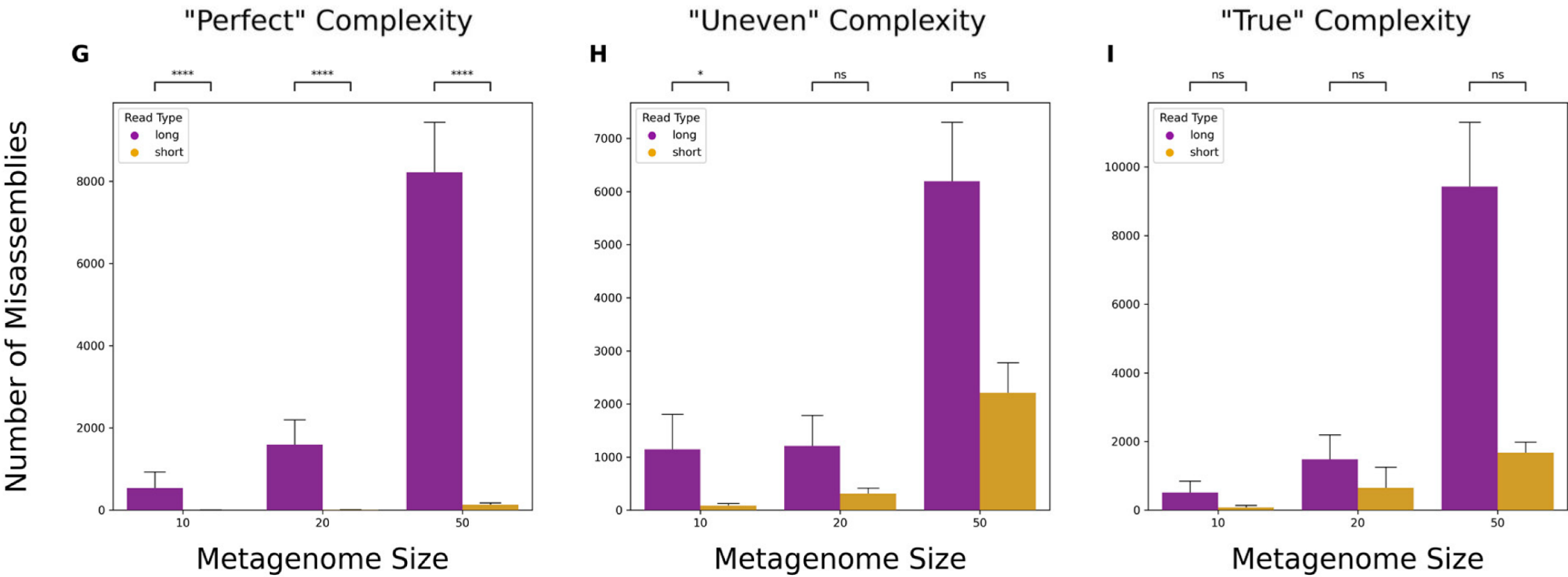


Assemblies from long reads demonstrated significantly higher genome fraction coverage

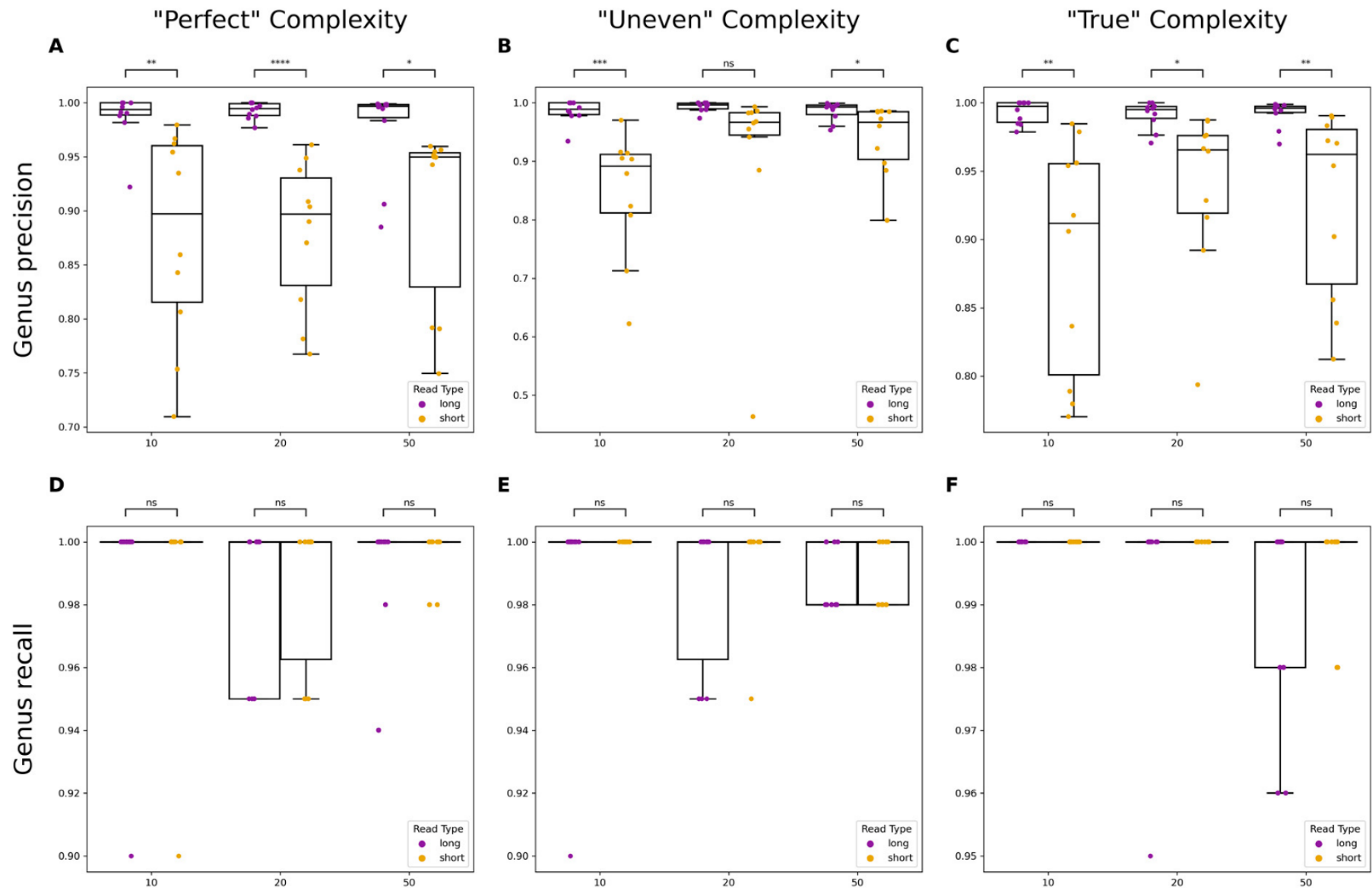
Similar results were observed when comparing assemblies by NGA50 values



# Result: Comparison of Metagenomic Assembly Completeness



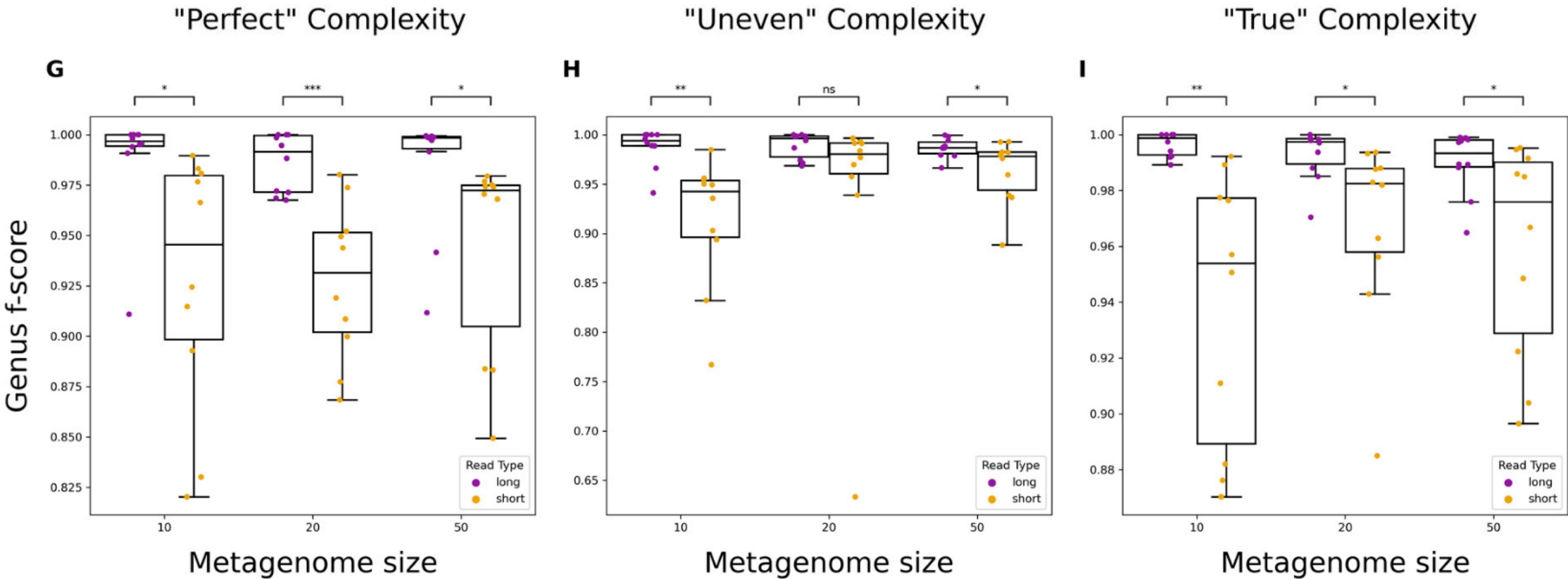
# Result: *Evaluating Taxonomic Classification*



Performance evaluation of  
genus-level classification

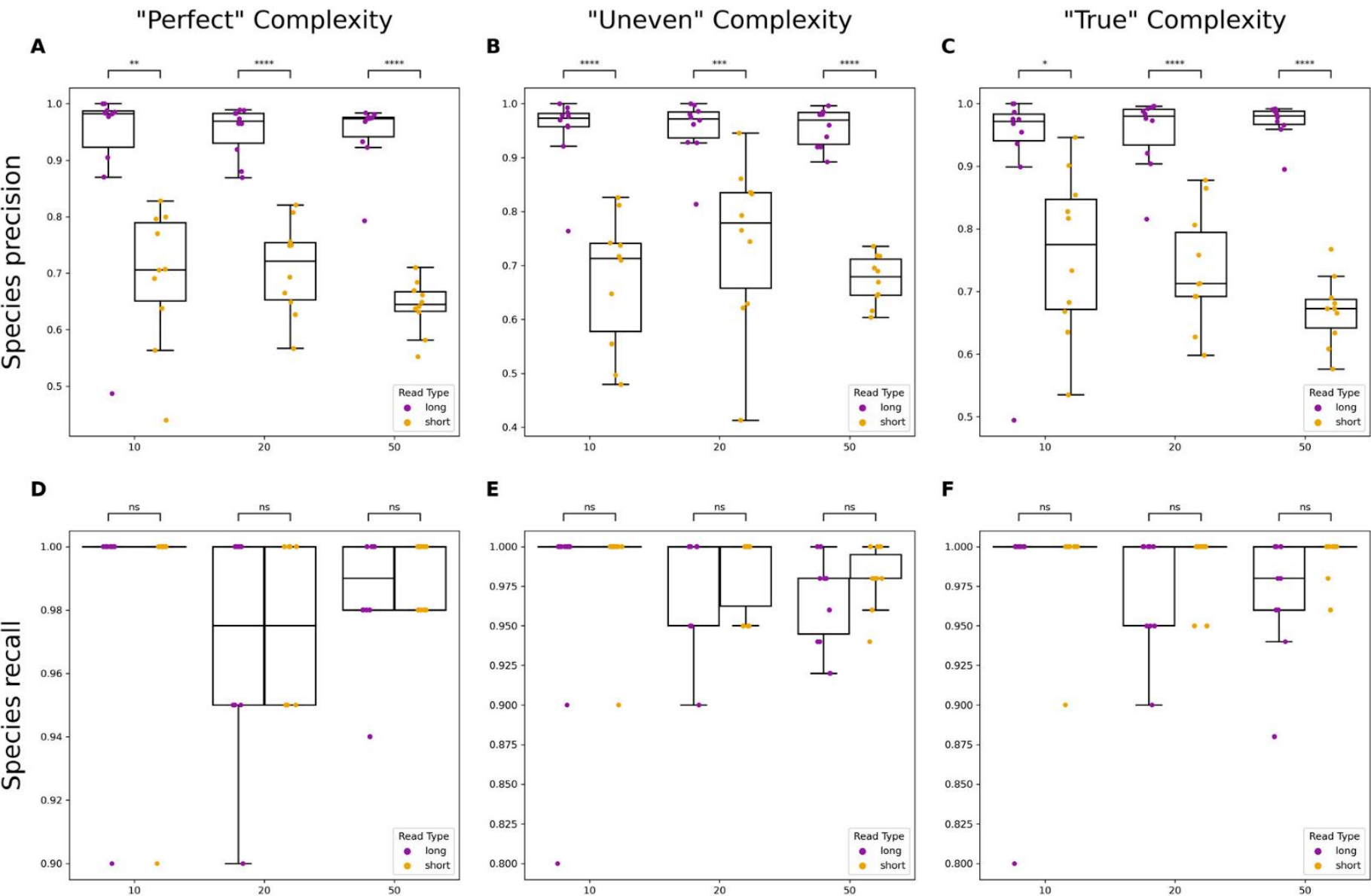
# Result: *Evaluating Taxonomic Classification*

Performance evaluation of  
genus-level classification



➡ In taxonomy classification, long reads exhibit significantly higher precision and F-score compared to short reads.

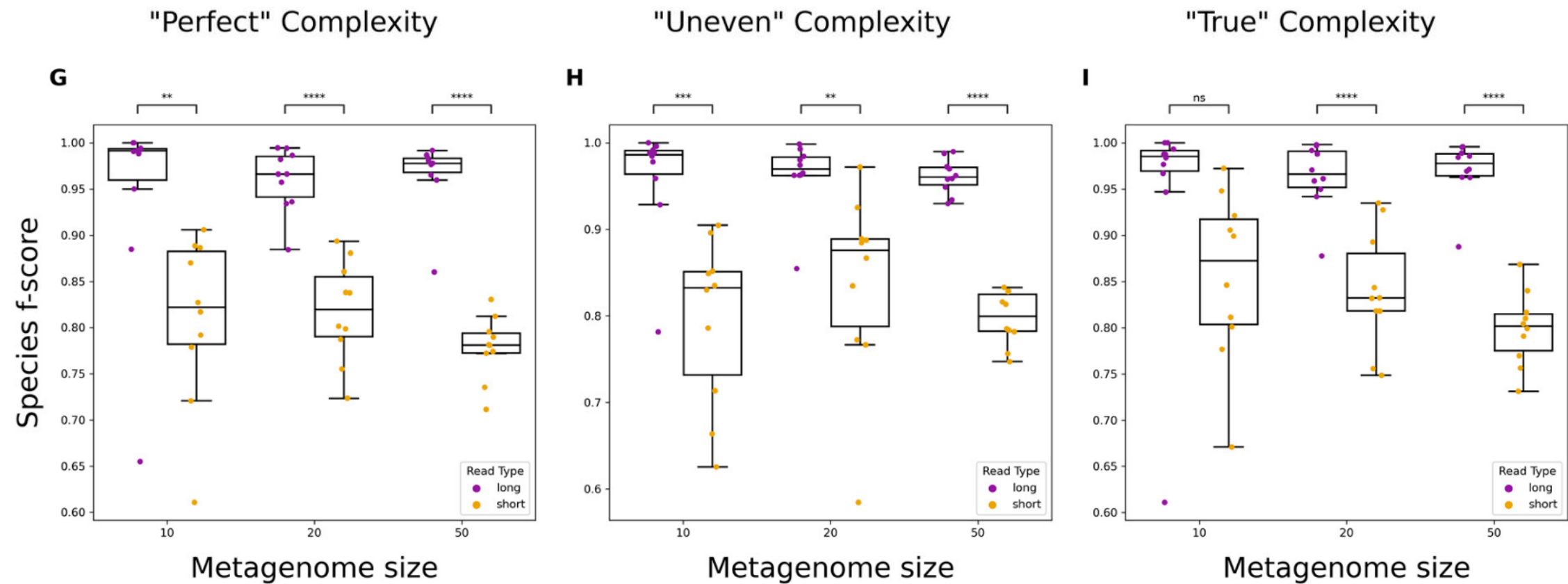
# Result: *Evaluating Taxonomic Classification*



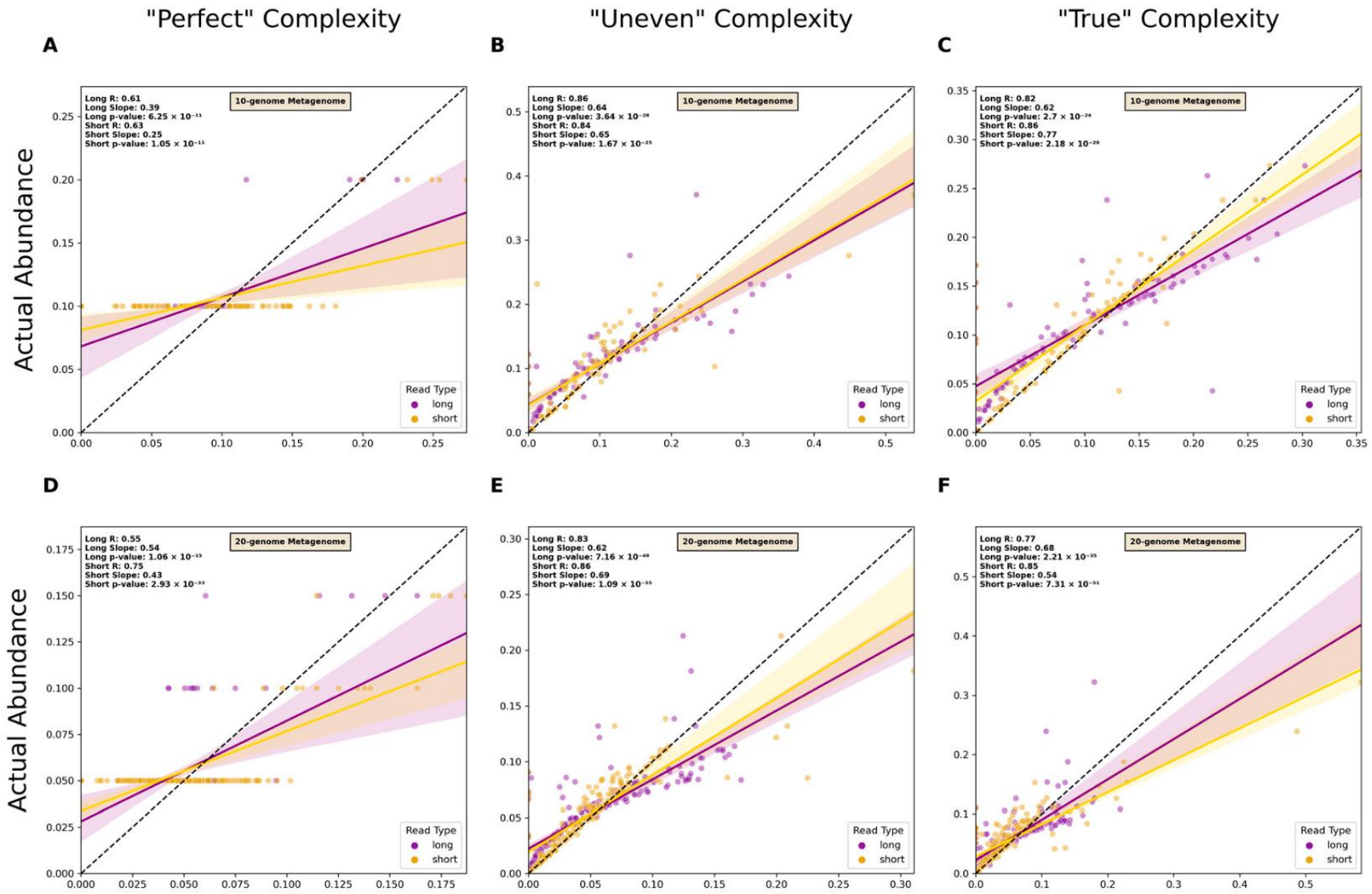
Performance evaluation of  
species-level classification

Result: *Evaluating Taxonomic Classification*

Performance evaluation of  
species-level classification



# Result: Estimating Relative Abundance Using Short- and Long-Read Data



Comparison of short- and long-read capacity for species-level relative abundance estimation.

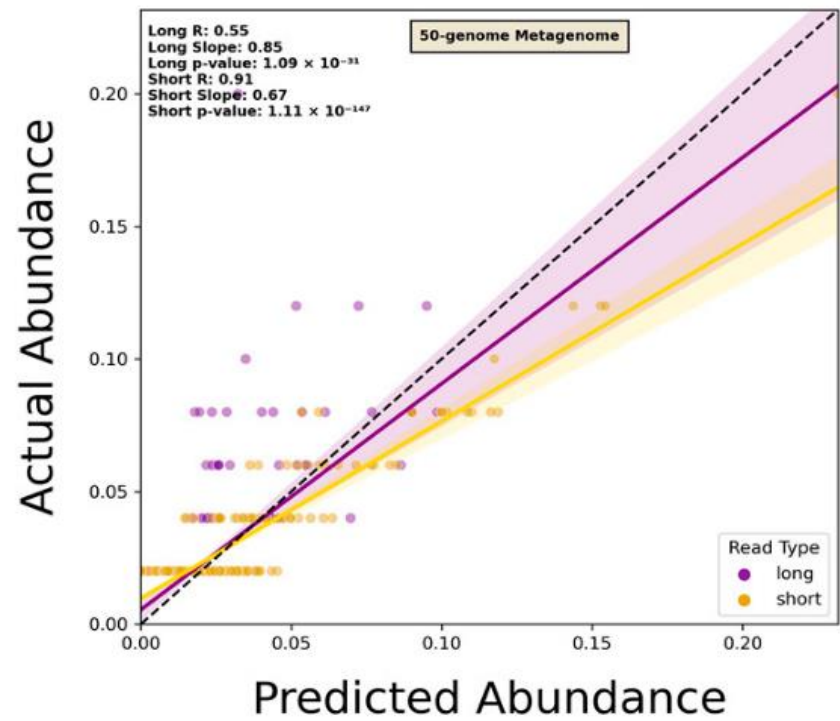
Long reads outperformed short reads in most cases



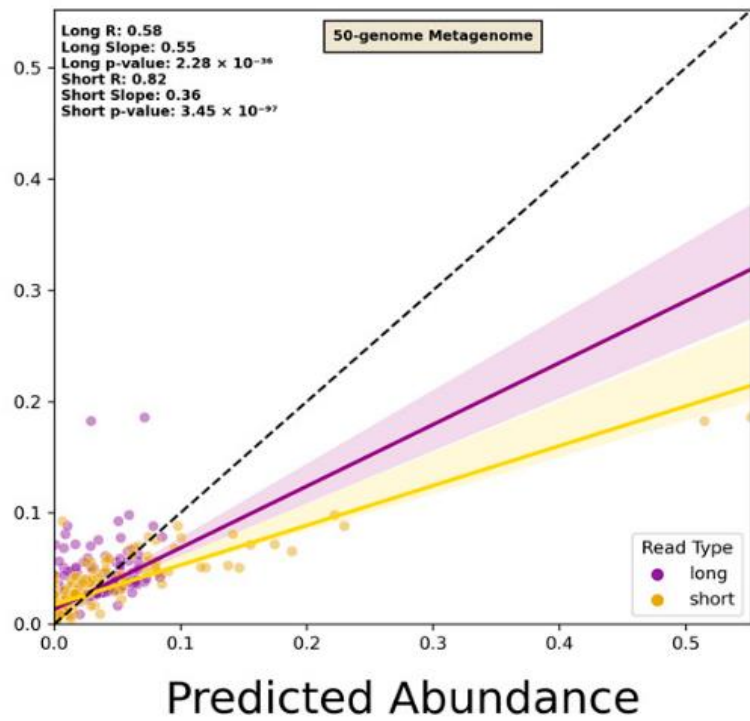
Result: *Estimating Relative Abundance Using Short- and Long-Read Data*

Comparison of short- and long-read capacity for species-level relative abundance estimation.

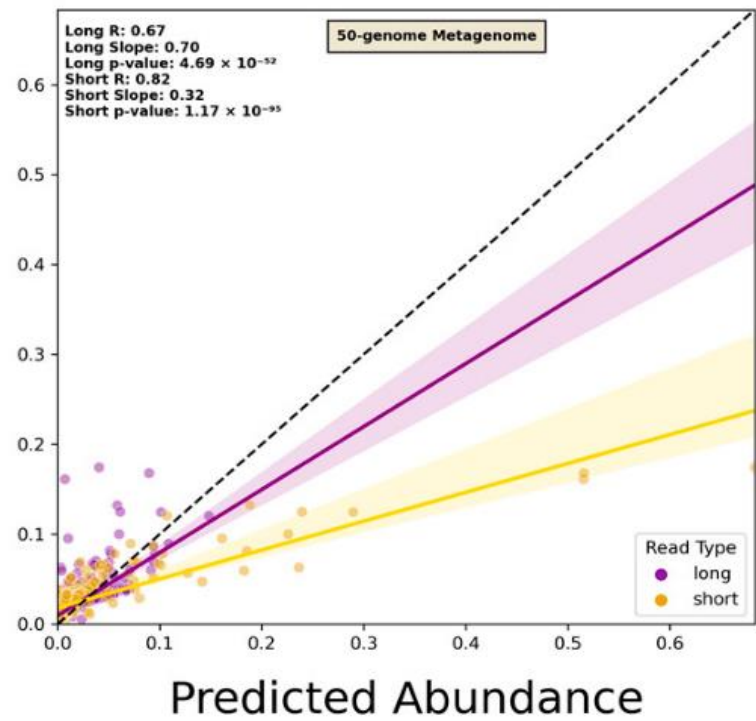
"Perfect" Complexity



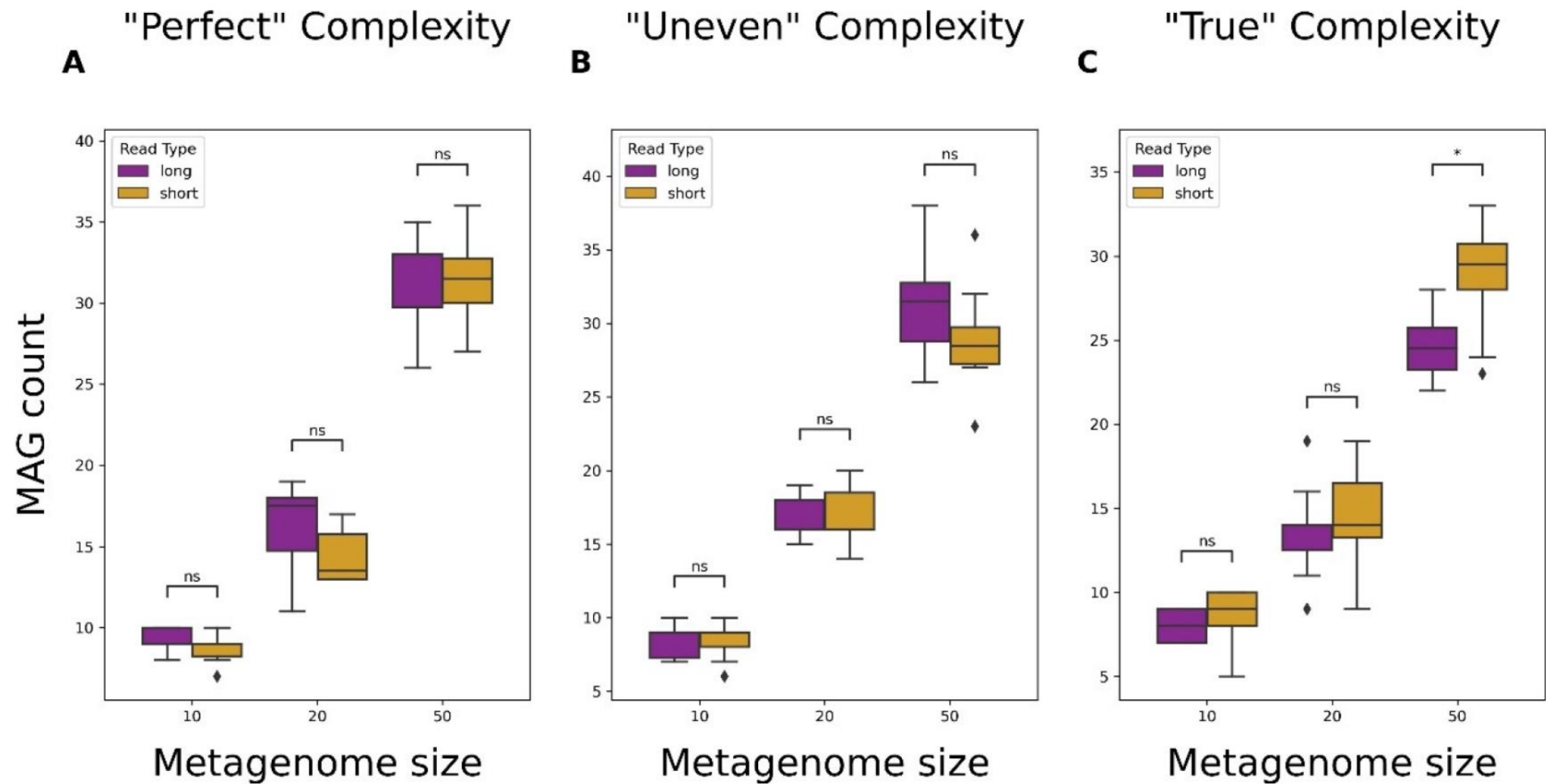
"Uneven" Complexity



"True" Complexity

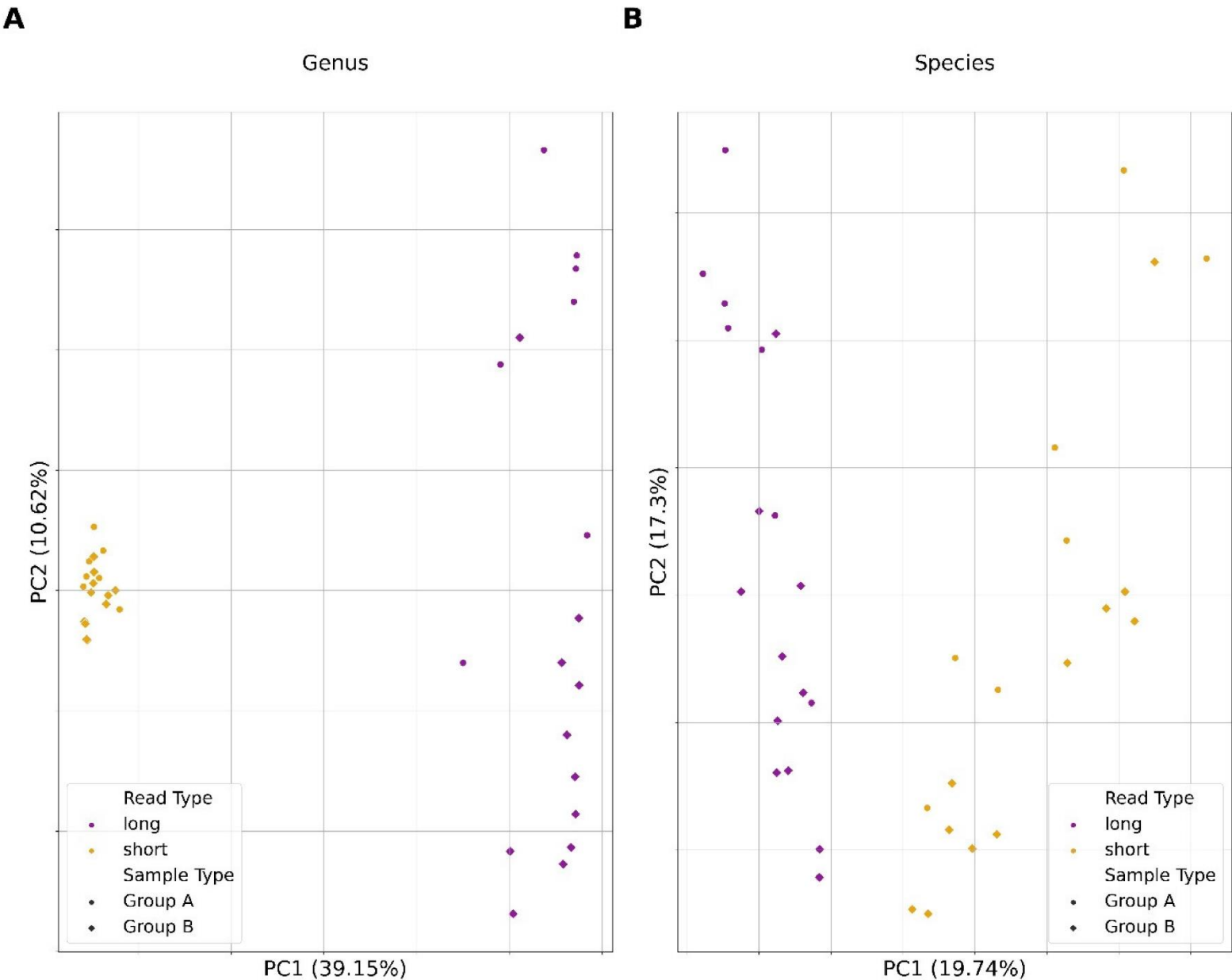


Result: *Metagenome-Assembled Genome (MAG) Recovery from Short- and Long-Read Data*



In terms of actual complexity, across 50 genomes, the number of Mags generated by short reads was higher than that of long reads.

Result: *Microbial Compositional Differences between Short and Long Reads*



Using experimental data  
obtained from mouse fecal  
pellets

samples clustered  
around similar read  
types instead of  
around similar  
sample types

- **Assembly Quality:** Long-read assemblies exhibited significantly higher genome coverage compared to short-read assemblies
- **Taxonomic Classification:** Long-read assemblies provided higher precision in taxonomic classifications at both genus and species levels due to their contiguous nature
- **Relative Abundance Estimation:** Predictions of species abundance from long-read data closely matched actual abundances, benefiting from more accurate classifications
- **MAG Recovery:** Both read types had similar metagenome-assembled genome (MAG) recovery rates, although short reads identified more MAGs in simpler datasets. Long reads, while more contiguous, suffered from misassembly rates that impacted MAG identification.
- **Beta Diversity Assessment:** Samples clustered by sequencing type rather than sample type, indicating that short-read assemblies had inflated misclassification rates.