**Katherine S. Pollard, Ph.d.**

**Director, Gladstone Institute of Data Science & Biotechnology**

**Professor, University of California, San Francisco**

**Investigator, Chan Zuckerberg Biohub**

***kpollard@gladstone.ucsf.edu***

May 3rd, 2024

Hillary Sussman

Executive Editor

*Genome Research*

Dear Editors,

We are excited to submit our manuscript entitled “Accurate estimation of intraspecific microbial gene content variation in metagenomic data with MIDAS v3 and StrainPGC” for consideration by *Genome Research*.

While microbial genome reference databases have expanded rapidly in recent years, they are still dwarfed by the strain diversity found in real microbiomes. Methods for understanding intraspecific variation in functional potential often rely on shotgun metagenomic data, yet challenges remain for accurate estimation of which genes are present in each strain. Solving this problem will enable well-powered studies interrogating the functional importance of this strain variation, with applications to biomedicine, ecology, and evolution.

In this manuscript, we describe a major upgrade to MIDAS for pangenome profiling, as well as a new tool, StrainPGC, which is able to accurately estimate gene content of individual strains found in shotgun metagenomes using the outputs of MIDAS. Our approach is specifically designed to overcome the limitations of traditional pangenome profiling without the need for isolation or assembly. The key innovation is the incorporation of strain tracking over multiple samples, which allows us to more accurately estimate the presence of genes in individual samples compared to only using sequencing depth. To demonstrate the potential of this approach, we apply MIDAS v3 and StrainPGC to a large collection of shotgun metagenomes (HMP2) and characterize strain diversity and pangenome dynamics across hundreds of microbial species.

Our manuscript reports the following novel results:

1. We harness a complex, synthetic microbial community of ~100 strains with complete genome sequences to benchmark the performance of StrainPGC on a highly realistic validation dataset. We show that StrainPGC is more accurate than both PanPhlAn and StrainPanDA for the vast majority of strains.
2. We comprehensively characterize the strain diversity in the HMP2 cohort, estimating the gene content of 3511 strains in 443 species across 12 phyla. In this one study alone, we find expansive strain diversity not captured by reference databases.
3. Finally, we compare the strains of *Escherichia coli* found in two FMT donors, finding distinct patterns of engraftment and functional potential.

We expect these results to be of broad interest to those working in microbial bioinformatics, statistical modeling for metagenomics, and ecology or evolution of the human microbiome.

All authors have read and approved the manuscript. All data used in these analyses are publicly available, and code to replicate our work is freely available on GitHub. All gene/protein names and symbols adhere to nomenclature guidelines, and no material has been published in a peer-reviewed journal or is under consideration elsewhere.

We suggest that reviewers cover expertise in (1) metagenomic analyses, (2) microbial population genetics, and (3) bioinformatic and statistical methods for microbiome analysis.

Sincerely,



Katherine S. Pollard