Reference bias in *Phytophthora infestans* genomes

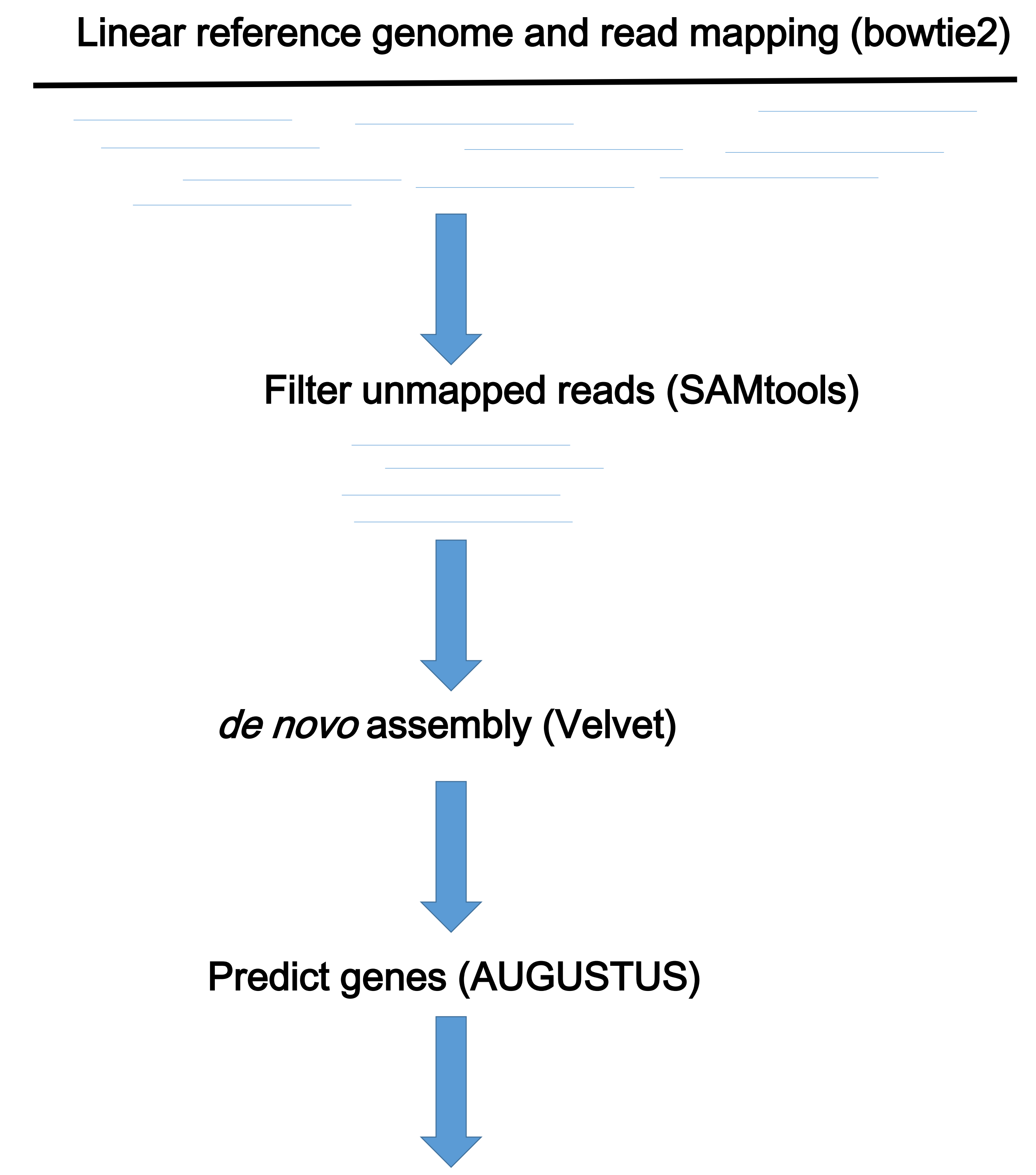
Shankar K. Shakya[1], Brian J. Knaus[1] and Niklaus J. Grünwald[2]

[1] Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR, 97331, USA
[2] Horticultural Crop Research Unit, United States Department of Agriculture, Agricultural Research Service, Corvallis, OR 97330, USA

Rationale

The majority of resequencing projects involve mapping reads to the linear reference genome. A reference genome hardly represents the genomic diversity of a population. Because of this, a **reference bias** occurs when mapping the reads. These unmapped reads are often neglected in the studies and no further analysis is done. There is growing evidence that the reads which fail to map to a reference genome might provide insight into the biological functions. Thus, we hypothesized that unmapped reads represented the uniqueness of the individual and could be assembled to predict isolate specific genes.

Bioinformatics pipeline



Linear reference genome and read mapping (bowtie2)

Filter unmapped reads (SAMtools)

de novo assembly (Velvet)

Predict genes (AUGUSTUS)

- Orthologs and paralogs clustering (OrthoFinder)
- Identification of RxLR and CRN genes (effectR)
- Amino acid identity analysis (AAI profiler)
- Identification of singletons

Take home message

- Reference bias is a commonly observed phenomenon but is often overlooked.
- Assembly of unmapped reads can help in the improvement of the existing reference genome and identify isolate specific genes.
- In the case of *P. infestans*, the reference bias led to huge variation in percentage of unmapped reads and gene content. Majority of those predicted genes had homologs in other *Phytophthora* species and bacteria *Paenibacillus*.
- Presence of *Paenibacillus* genes in *P. infestans* genomes could be due to bacterial infection or horizontal gene transfer.

Results

The percentage of reads that did not map to the reference strain ranged from **3-58**. Assembly size of unmapped region ranged from **1.43 - 16 Mbp** and the gene content ranged from **394-3938** genes (Fig 1). We were able to identify 62 orthologous gene clusters with genes present from all 22 *P. infestans* isolates and a total of **1554** singletons. These newly identified genes had homologs in other *Phytophthora* species which suggests the unmapped reads are not just contaminations and can be used to improve the reference genome (Fig 2). Average amino acid identity also identified homologs in the bacterial genus *Paenibacillus* (Fig 3).

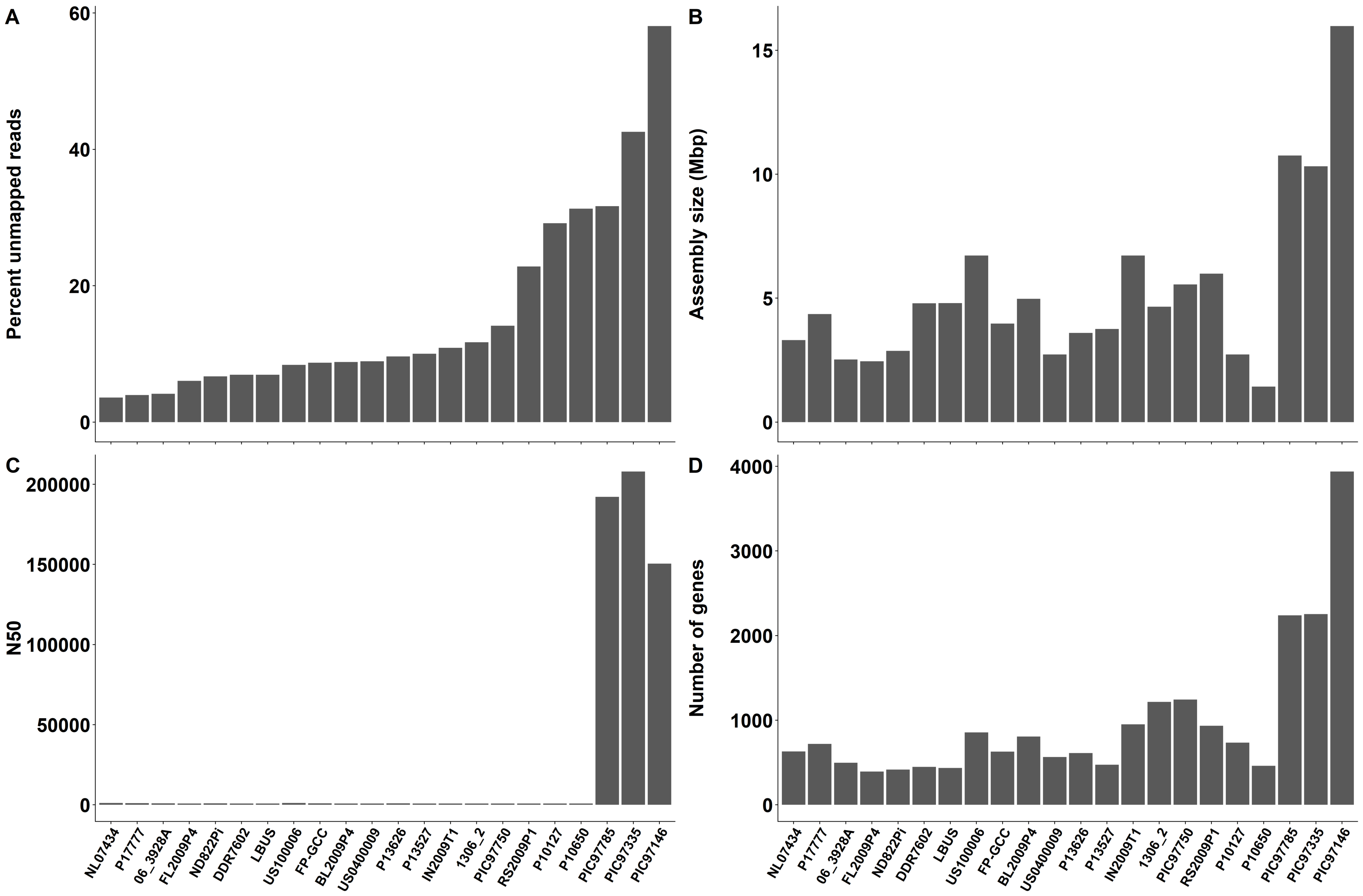


Fig 1. Barplots showing **A.** Percentage of reads that fail to map to reference T30-4 isolate; **B.** Assembly size of unmapped reads; **C.** Contig N50; **D.** Number of predicted genes from assembly of unmapped reads.

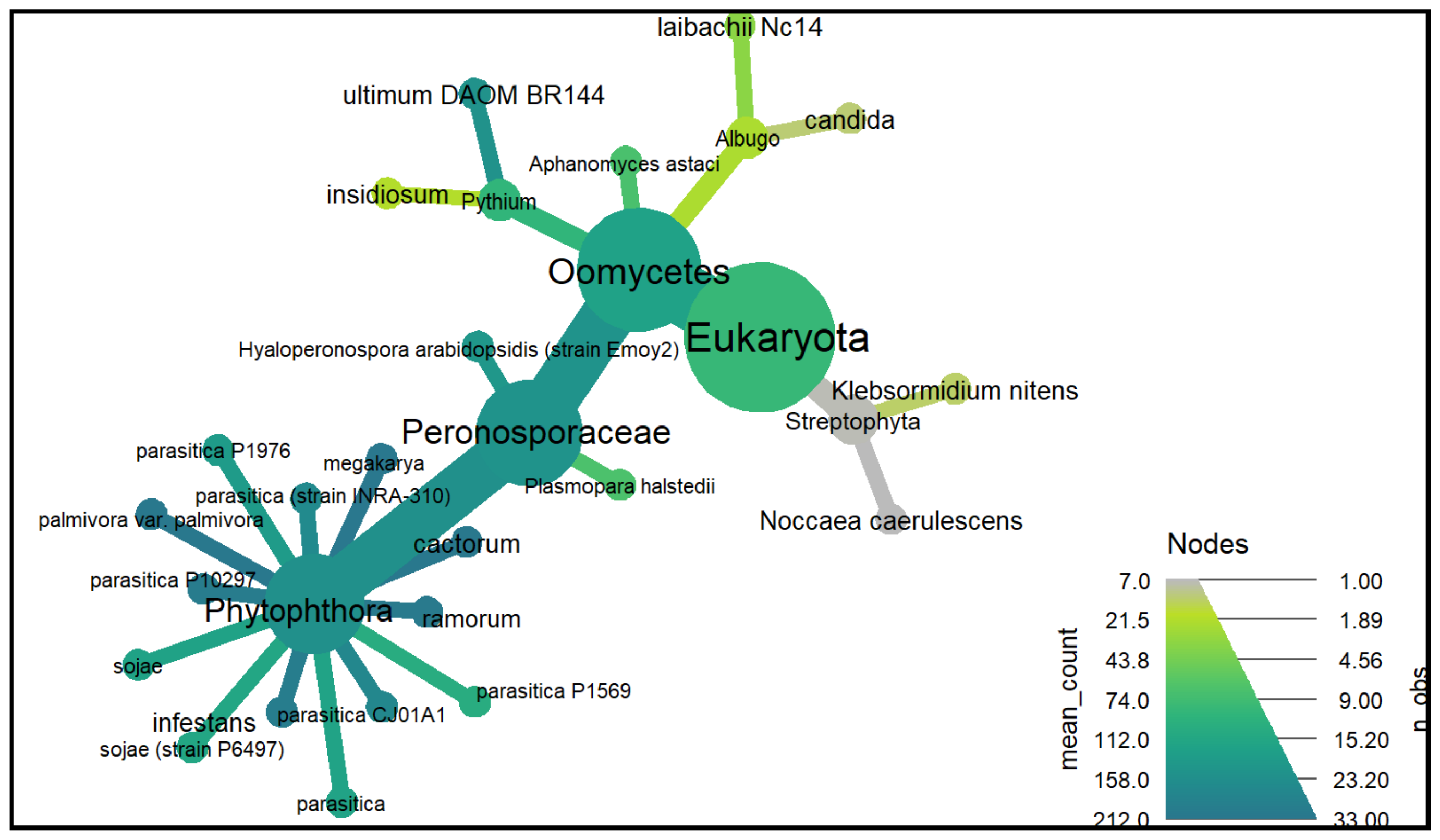


Fig 2. Newly predicted genes have homologs in *Phytophthora* species.

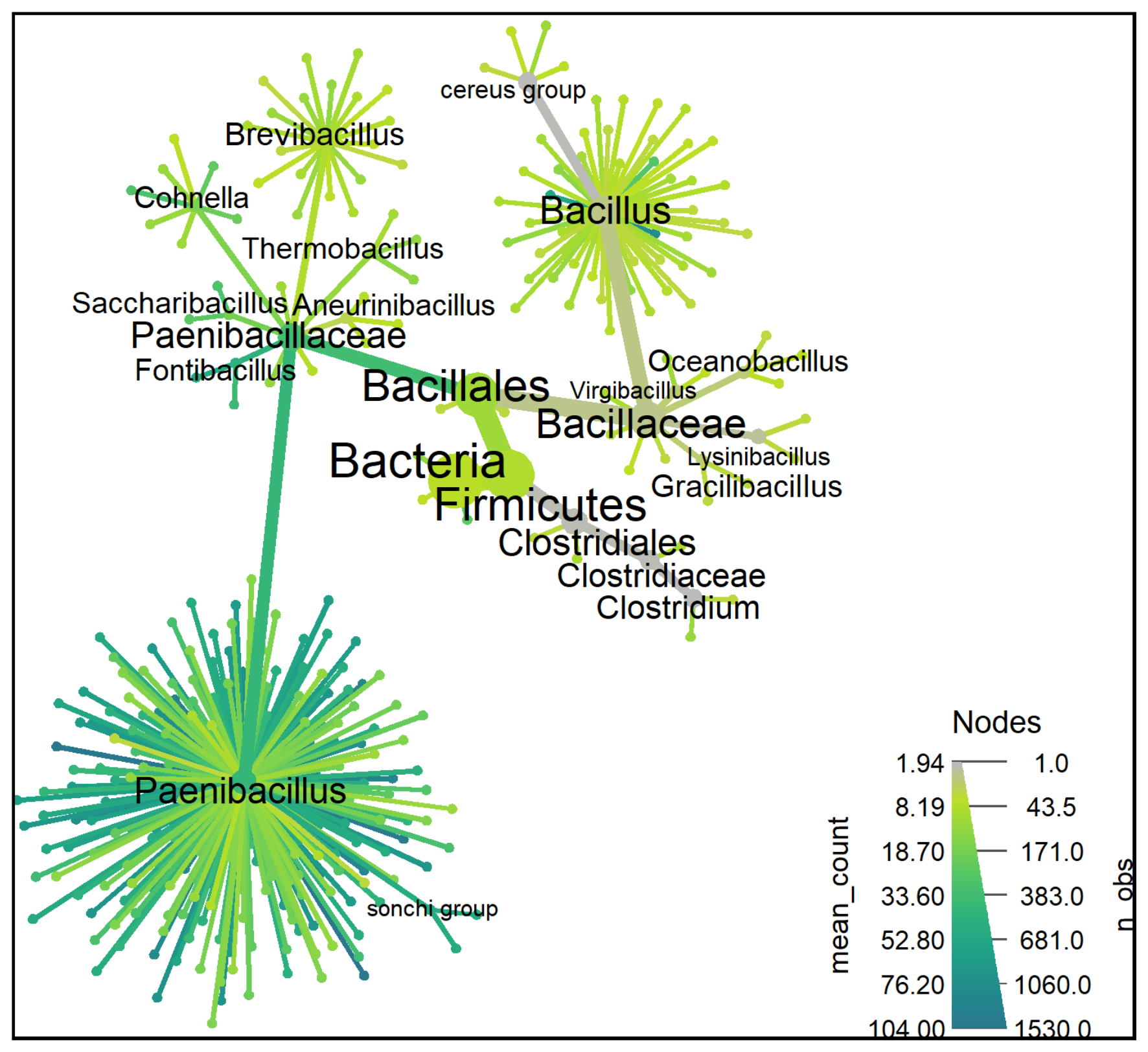


Fig 3. Newly predicted genes have homologs in *Paenibacillus* bacteria.

Contact information: Shankar K. Shakya (shakyash@oregonstate.edu) and Niklaus J. Grünwald (grunwaln@oregonstate.edu)
Website: https://shankarkshakya.github.io