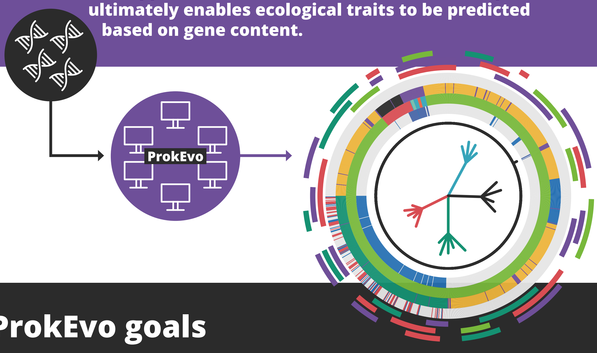
ProkEvo:
an automated, reproducible, and scalable framework for high-throughput bacterial population genomics analyses

Background

As we have seen with COVID-19, viruses and pathogenic bacteria evolve in real time, and genetic variants can quickly emerge and spread globally. However, **managing and scaling the analysis of microbial populations demands complex bioinformatics tools** in order to identify unique epidemiological and ecological patterns.

ProkEvo achieves that goal, and results in thousands of bacterial whole genomes being genotypically mapped in varying levels of resolution. This approach reveals their familial relationships, and ultimately enables ecological traits to be predicted based on gene content.



ProkEvo goals

ProkEvo was specifically developed to achieve the following goals:

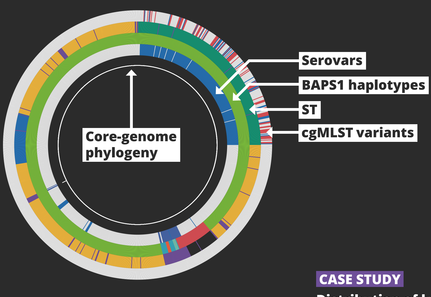
1. Automation and scaling of complex combinations of computational analyses for many thousands of bacterial genomes
2. Use of workflow management system (WMS) to ensure reproducibility, scalability, modularity, fault-tolerance, and robust file management
3. Use of high-performance and high-throughput computational platforms
4. Generation of hierarchical-based population structure analysis based on combinations of statistical approaches for classification of ecological and epidemiological inquiries
5. Association of antimicrobial resistance (AMR) genes, putative virulence factors, and plasmids from curated databases
6. Production of pan-genome annotations and data compilation that can be utilized for downstream analysis

Scalability of ProkEvo

The scalability of ProkEvo was measured with two datasets ranging from ~2,400 to ~23,000 genomes. Running time varied from ~3 to ~26 days. **ProkEvo can be used with virtually any bacterial species,** and the Pegasus WMS uniquely facilitates addition or removal of programs from the workflow or modification of options within them.

OVERVIEW OF POPULATION STRUCTURE

Hierarchical population structure mapping of *Salmonella enterica*

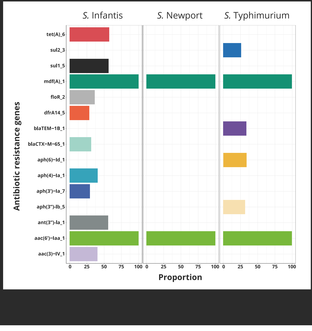


Case studies

To demonstrate the versatility of ProkEvo, we performed a hierarchical-based population structure analyses of available genomes from three distinct pathogenic bacterial species for individual case studies. The case studies illustrate how analyses of population structures, genotype frequencies, and distribution of specific gene functions can be integrated into an analysis.

CASE STUDY

Distribution of known AMR loci across the population structures of *S.* Infantis, *S.* Newport, and *S.* Typhimurium



Conclusion

Collectively, **our study shows that ProkEvo presents a practical and viable option for scalable, automated analyses of bacterial populations** with direct applications for basic microbiology research, clinical microbiological diagnostics, and epidemiological surveillance.

PeerJ

ProkEvo: an automated, reproducible, and scalable framework for high-throughput bacterial population genomics analyses
PeerJ 9:e11376 DOI: 10.7717/peerj.11376
http://peerj.com/articles/11376/