PUG 2021

Pegasus Users Group

MEETING

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

Natasha Pavlovikj

PhD student
Department of Computer Science and Engineering
University of Nebraska-Lincoln

February 25th , 2021

Bacterial Population Genomics

- Group of individuals of the same bacterial species
- Populations evolve
- Differences between individuals are tiny
- Conclusions about populations

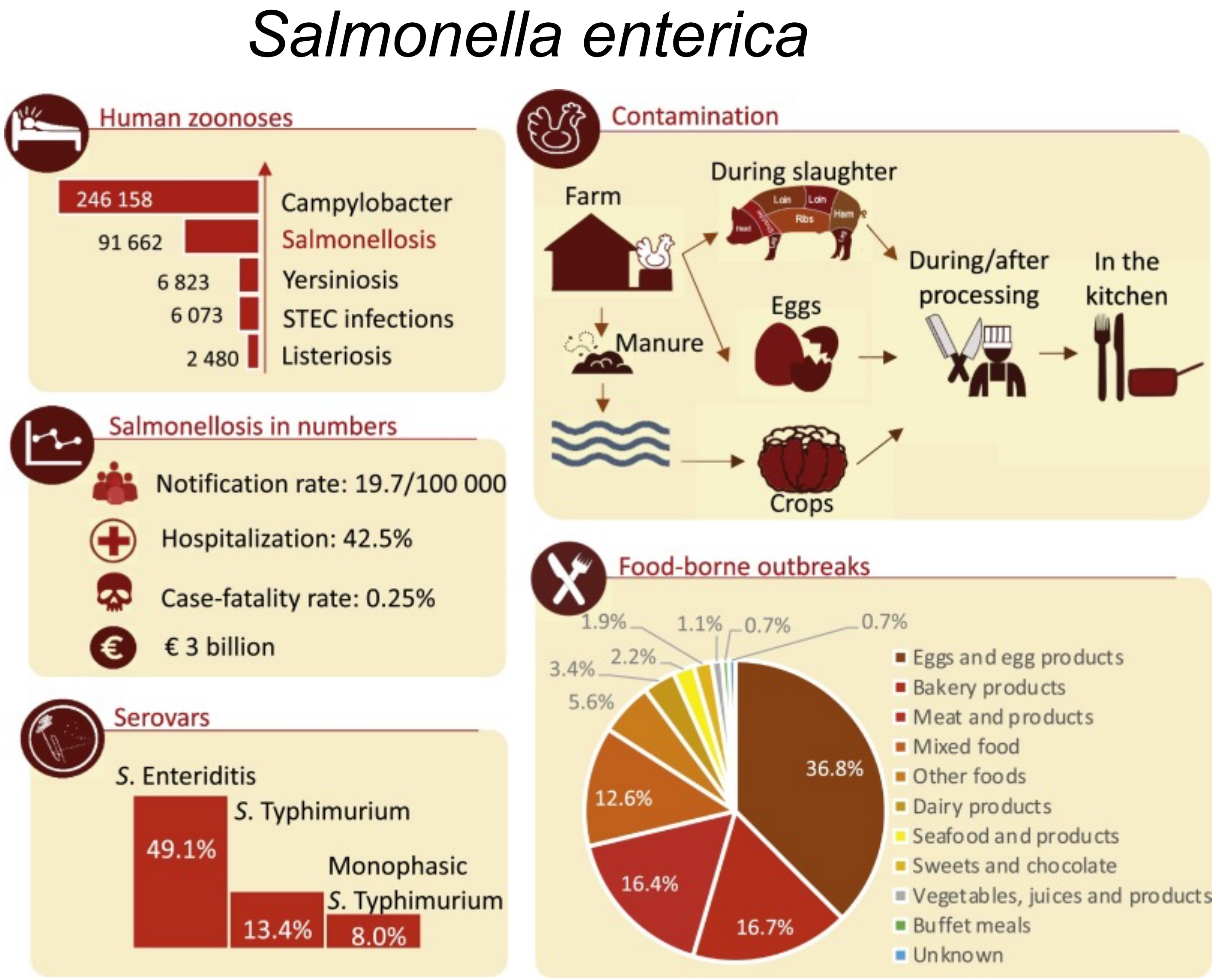
Bacterial Population Genomics

PUG 2021

- Group of individuals of the same bacterial species
- Populations evolve
- Differences between individuals are tiny
- Conclusions about populations

We need a comprehensive pipeline that allows us to do large-scale population genomics, because only then we can understand something about the population structure, distribution of genotypes, evolutionary history, accessory genome distribution that COUPLED with ecological and epidemiological information can bring better understanding of how populations diversify and evolve over time.

One application of ProkEvo is in the field of Public Health and Food Safety:
Salmonella enterica



Goals:

1. Determine the most Important genotypes
2. Infer the genomic events important for survival across the food chain in order to mitigate the risk of infection

Example:

sugE – resistance to quaternary ammonium
If that is a driver for a serovar to dominate, then it indicates a change in hygiene practice is needed, perhaps by switching disinfectants

Salmonella is one example only

3. Population heterogeneity tactics as driving force in Salmonella virulence and survival, Ines Staes, Ioannis Passaris, Alexander Cambré, Abram Aertsen

Applications

- Outbreak detection
- Source tracking
- Understanding epidemics
- Public Health Surveillance
- Pathogen transmission
- Discover ecological properties

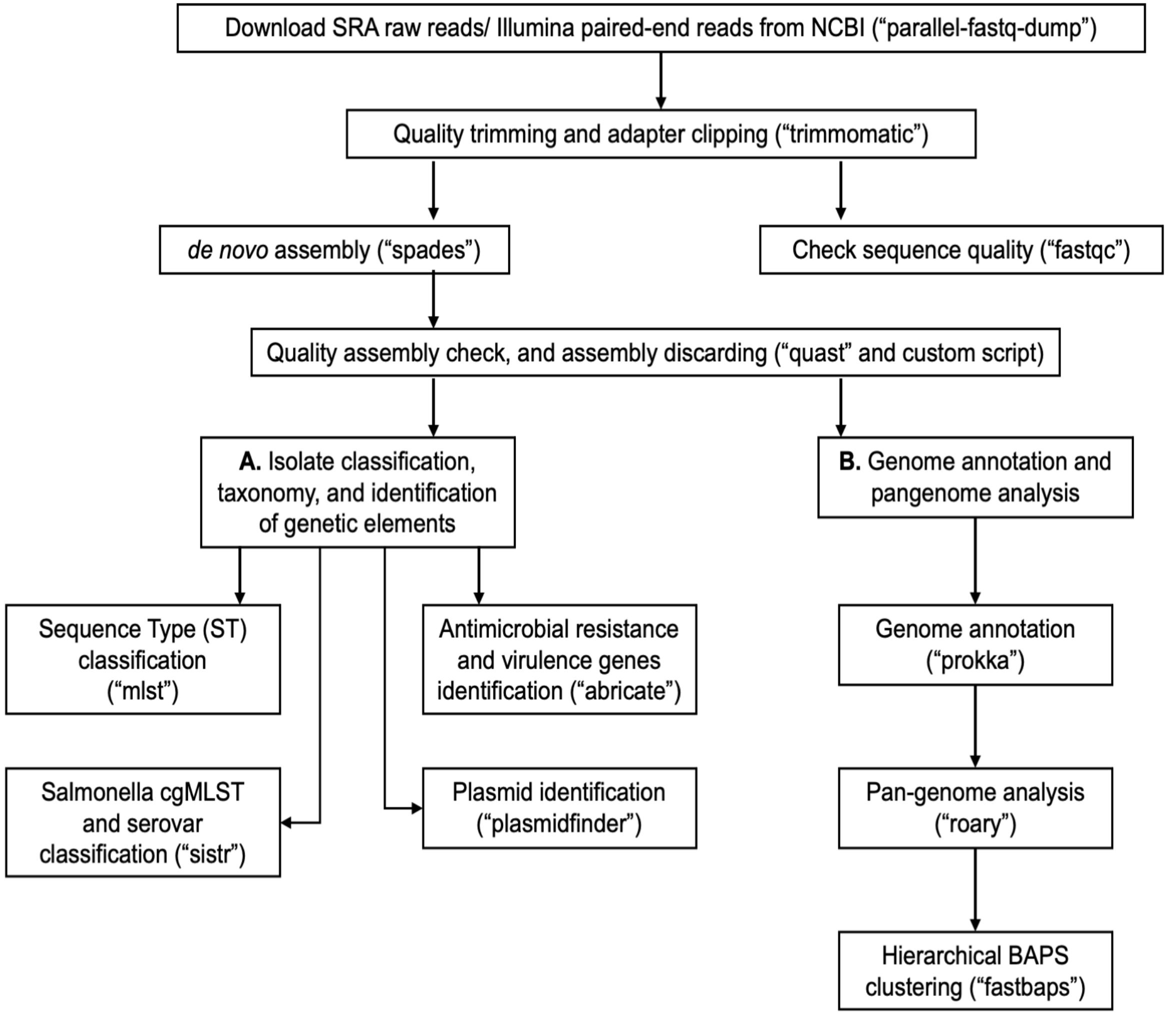
Challenges

- Scaling and automating WGS analyses can be a challenging task that comes with its own costs and benefits
- Several existing automated pipelines for analyses of bacterial genomes
- No usage of workflow management system
- Small number of used genomes

ProkEvo

- Automated, open-source pipeline for population genomics analyses
- Uses Pegasus Workflow Management System
- Analyses of few genomes, as well as tens of thousands using high-performance and high-throughput computational resources
- Easily modifiable and expandable pipeline to include additional steps, custom scripts, user databases, and species
- Modular pipeline that can run thousands of analyses concurrently if the resources are available
- Distributed with conda environment and Docker image for all bioinformatics tools and databases needed to perform population genomics analyses
- The output of ProkEvo is used to provide some guidance on how to perform population-based analyses with reproducible Jupyter Notebooks
- https://github.com/npavlovikj/ProkEvo

ProkEvo



Download SRA raw reads/ Illumina paired-end reads from NCBI (“parallel-fastq-dump”)

Quality trimming and adapter clipping (“trimmomatic”)

de novo assembly (“spades”)

Check sequence quality (“fastqc”)

Quality assembly check, and assembly discarding (“quast” and custom script)

A. Isolate classification, taxonomy, and identification of genetic elements

B. Genome annotation and pangenome analysis

Sequence Type (ST) classification (“mlst”)

Antimicrobial resistance and virulence genes identification (“abricate”)

Salmonella cgMLST and serovar classification (“sistr”)

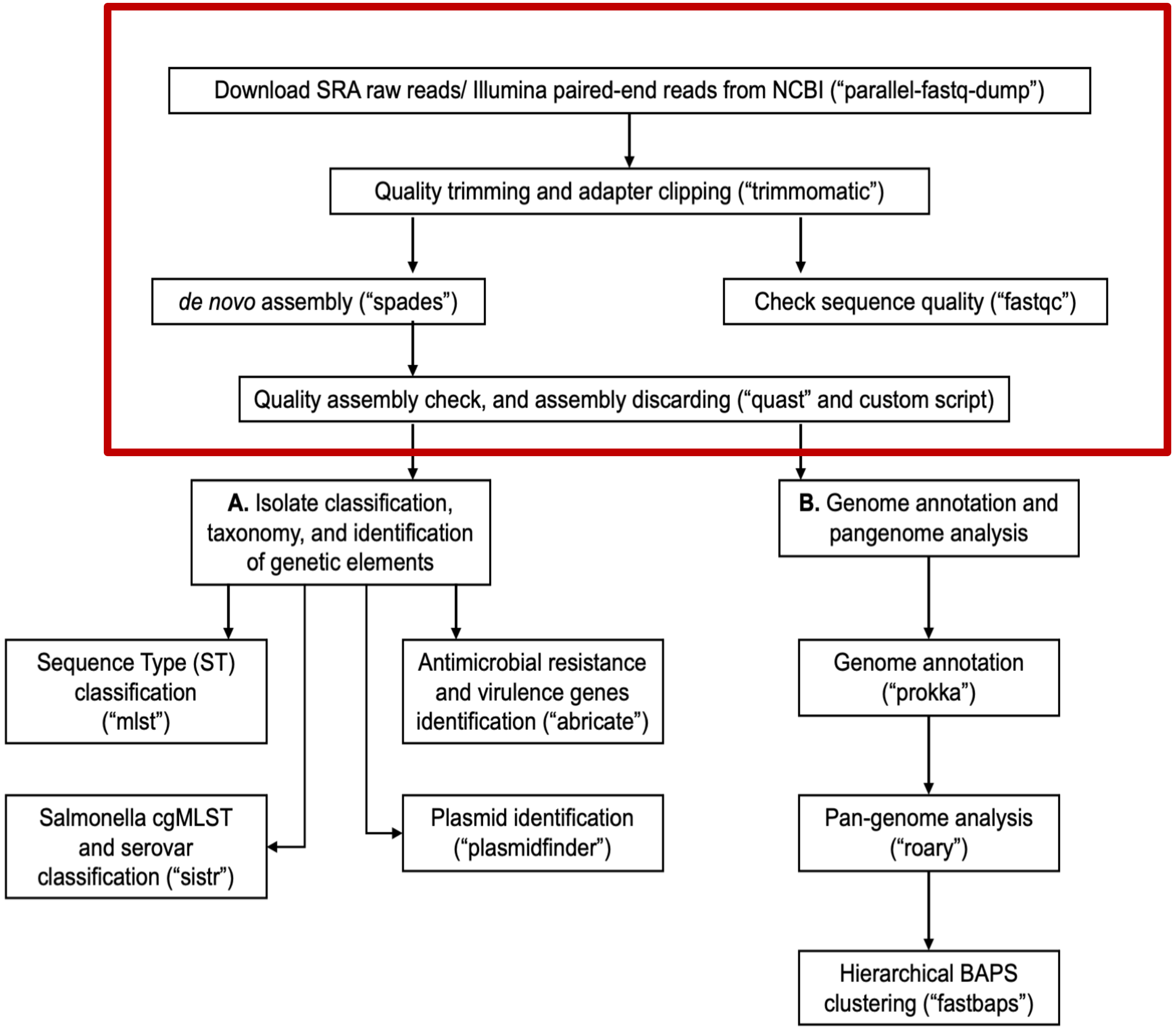
Plasmid identification (“plasmidfinder”)

Genome annotation (“prokka”)

Pan-genome analysis (“roary”)

Hierarchical BAPS clustering (“fastbaps”)

ProkEvo



Download SRA raw reads/ Illumina paired-end reads from NCBI (“parallel-fastq-dump”)

Quality trimming and adapter clipping (“trimmomatic”)

de novo assembly (“spades”)

Check sequence quality (“fastqc”)

Quality assembly check, and assembly discarding (“quast” and custom script)

A. Isolate classification, taxonomy, and identification of genetic elements

Sequence Type (ST) classification (“mlst”)

Salmonella cgMLST and serovar classification (“sistr”)

Antimicrobial resistance and virulence genes identification (“abricate”)

Plasmid identification (“plasmidfinder”)

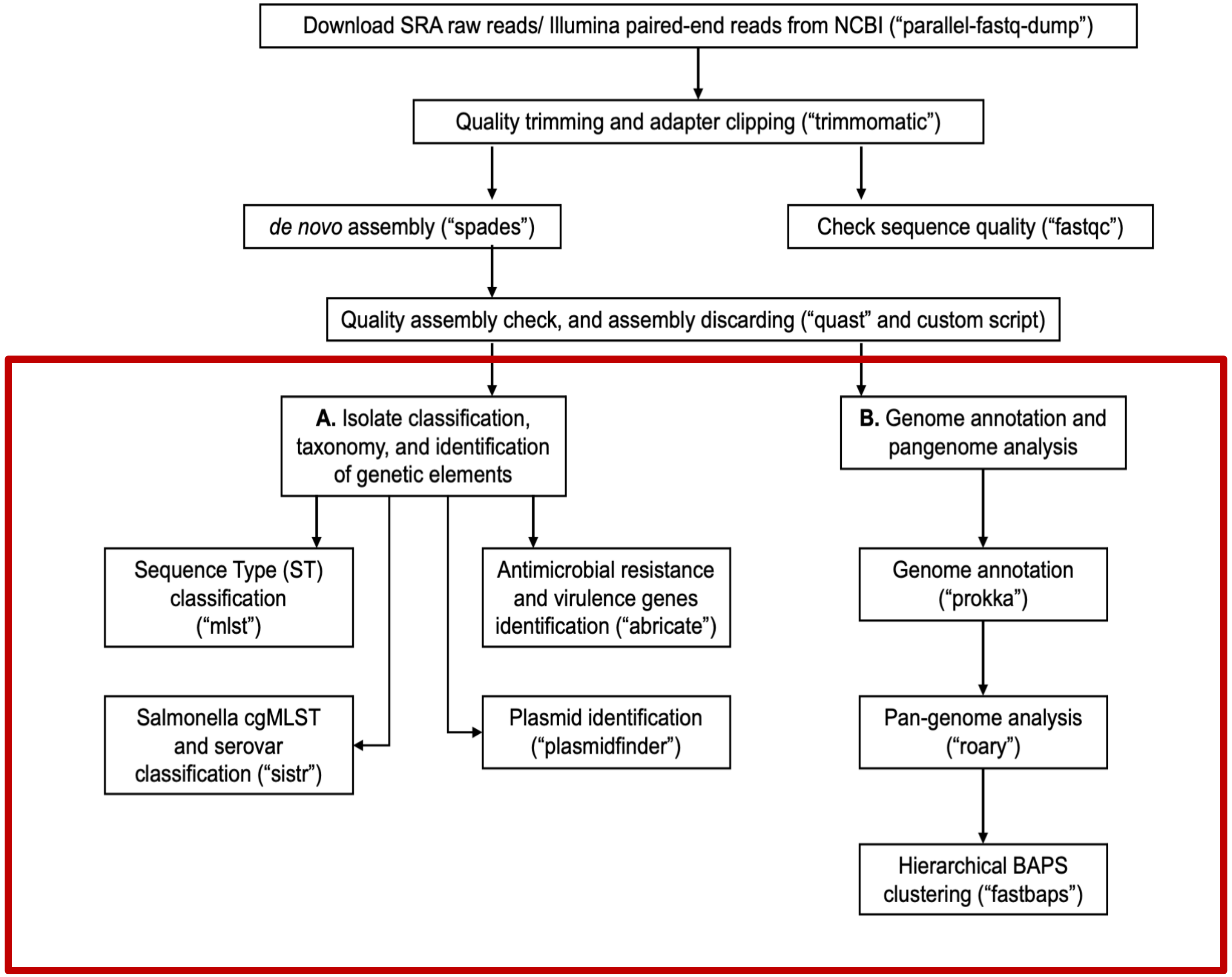
B. Genome annotation and pangenome analysis

Genome annotation (“prokka”)

Pan-genome analysis (“roary”)

Hierarchical BAPS clustering (“fastbaps”)

ProkEvo



Download SRA raw reads/ Illumina paired-end reads from NCBI (“parallel-fastq-dump”)

Quality trimming and adapter clipping (“trimmomatic”)

de novo assembly (“spades”)

Check sequence quality (“fastqc”)

Quality assembly check, and assembly discarding (“quast” and custom script)

A. Isolate classification, taxonomy, and identification of genetic elements

Sequence Type (ST) classification (“mlst”)

Salmonella cgMLST and serovar classification (“sistr”)

Antimicrobial resistance and virulence genes identification (“abricate”)

Plasmid identification (“plasmidfinder”)

B. Genome annotation and pangenome analysis

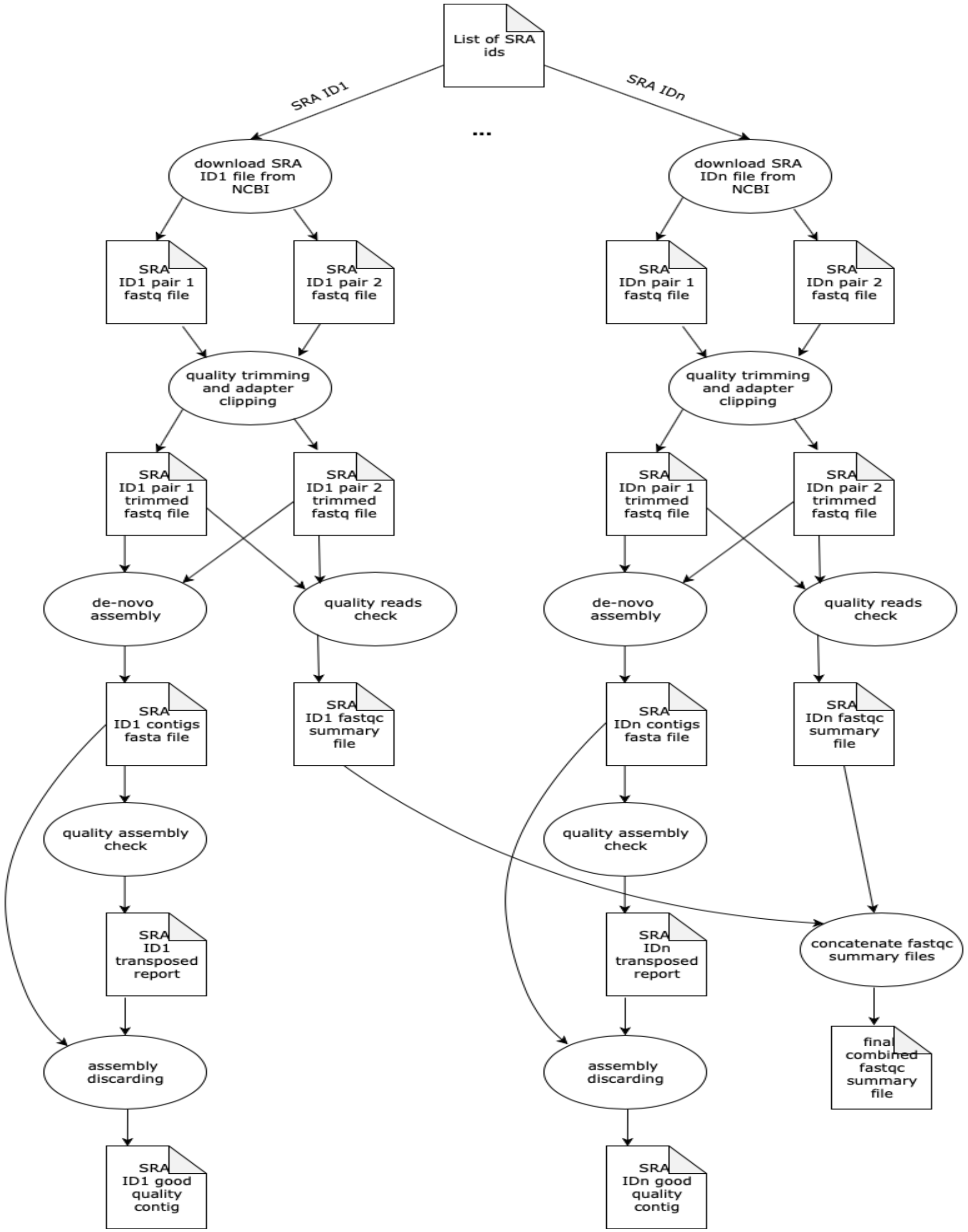
Genome annotation (“prokka”)

Pan-genome analysis (“roary”)

Hierarchical BAPS clustering (“fastbaps”)

Part 1

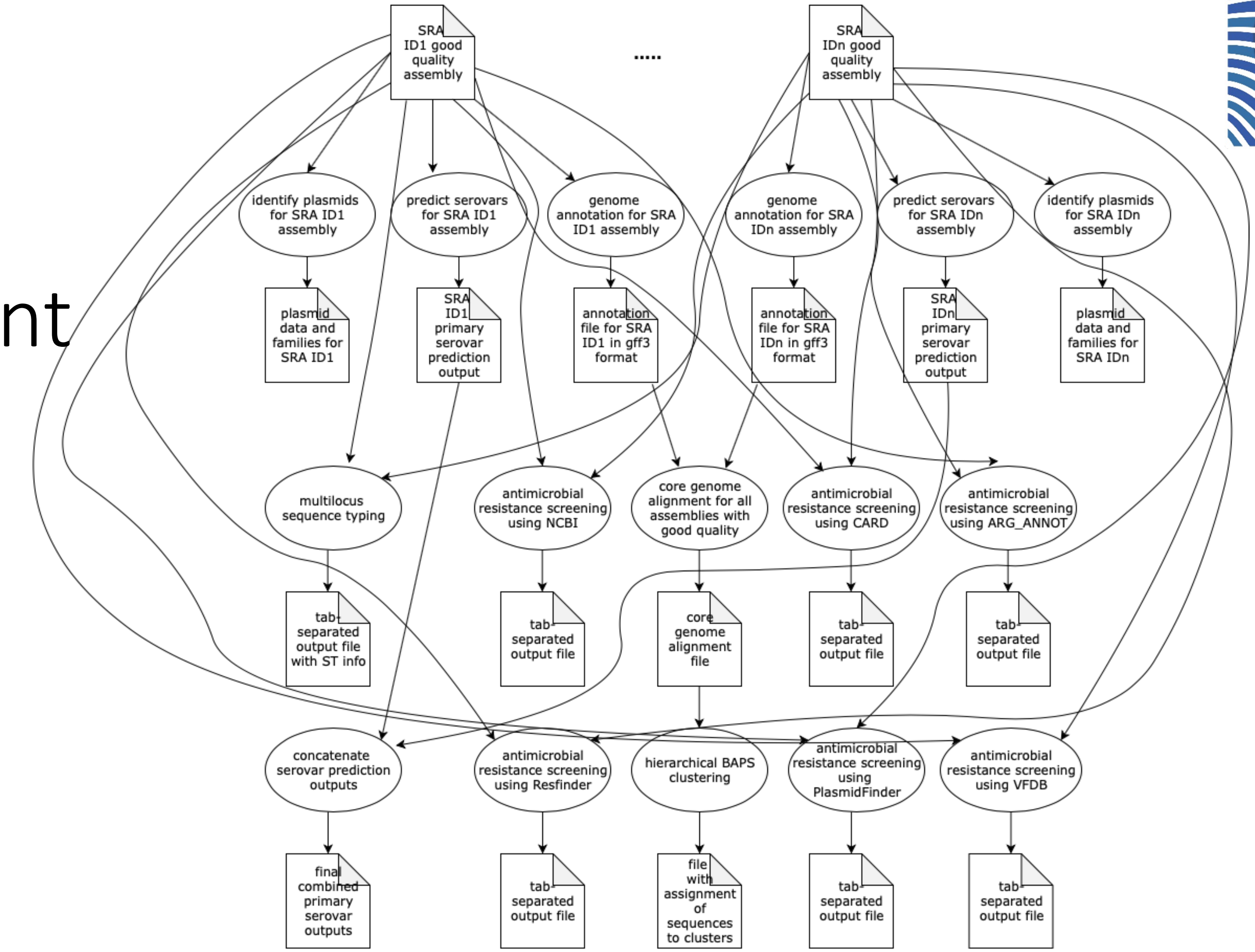
ProkEvo: Pegasus Workflow Management System


List of SRA ids
SRA ID1
SRA IDn
...
download SRA ID1 file from NCBI
download SRA IDn file from NCBI
SRA ID1 pair 1 fastq file
SRA ID1 pair 2 fastq file
SRA IDn pair 1 fastq file
SRA IDn pair 2 fastq file
quality trimming and adapter clipping
quality trimming and adapter clipping
SRA ID1 pair 1 trimmed fastq file
SRA ID1 pair 2 trimmed fastq file
SRA IDn pair 1 trimmed fastq file
SRA IDn pair 2 trimmed fastq file
de-novo assembly
quality reads check
de-novo assembly
quality reads check
SRA ID1 contigs fasta file
SRA ID1 fastqc summary file
SRA IDn contigs fasta file
SRA IDn fastqc summary file
quality assembly check
quality assembly check
concatenate fastqc summary files
SRA ID1 transposed report
SRA IDn transposed report
final combined fastqc summary file
assembly discarding
assembly discarding
SRA ID1 good quality contig
SRA IDn good quality contig

ProkEvo: Pegasus Workflow Management System

Part 2

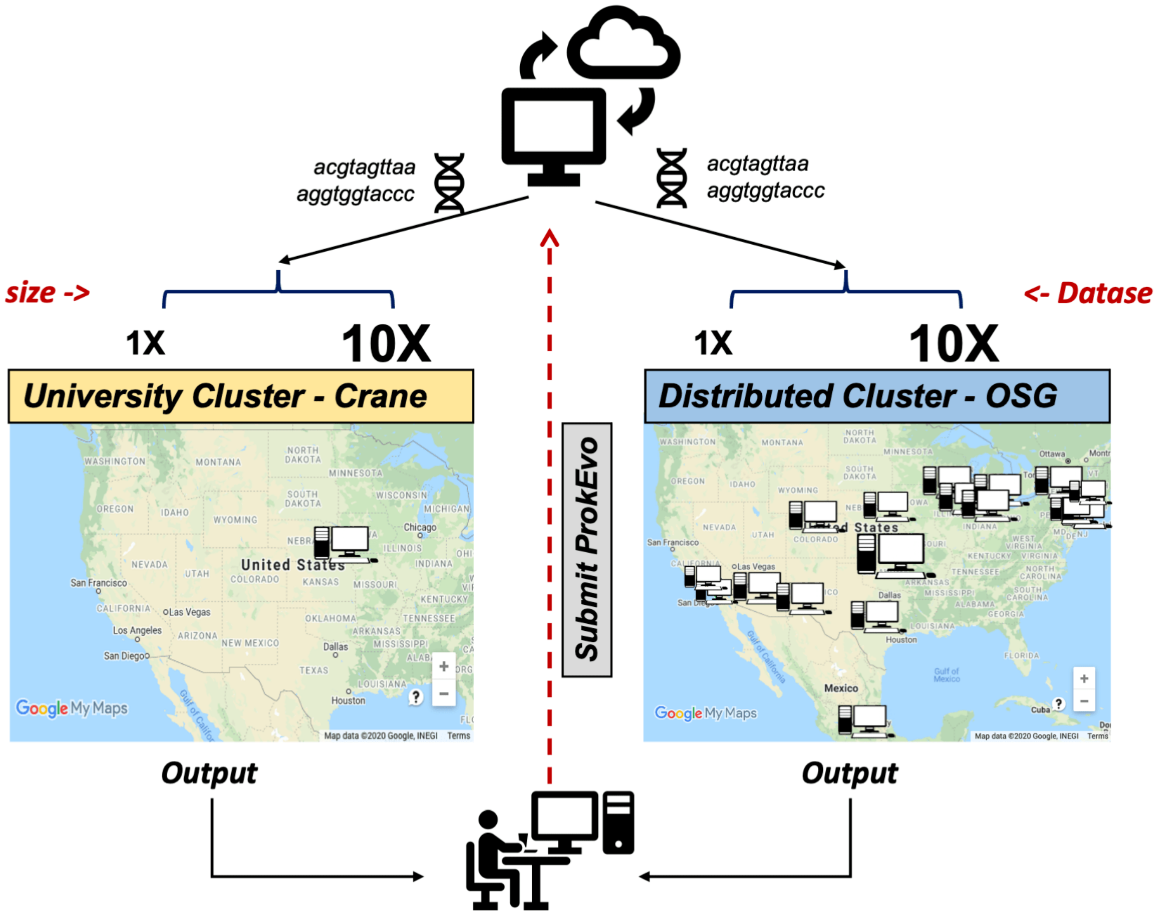
PUG 2021


SRA ID1 good quality assembly
.....
SRA IDn good quality assembly
identify plasmids for SRA ID1 assembly
predict serovars for SRA ID1 assembly
genome annotation for SRA ID1 assembly
genome annotation for SRA IDn assembly
predict serovars for SRA IDn assembly
identify plasmids for SRA IDn assembly
plasmid data and families for SRA ID1
SRA ID1 primary serovar prediction output
annotation file for SRA ID1 in gff3 format
annotation file for SRA IDn in gff3 format
SRA IDn primary serovar prediction output
plasmid data and families for SRA IDn
multilocus sequence typing
antimicrobial resistance screening using NCBI
core genome alignment for all assemblies with good quality
antimicrobial resistance screening using CARD
antimicrobial resistance screening using ARG_ANNOT
tab-separated output file with ST info
tab-separated output file
core genome alignment file
tab-separated output file
tab-separated output file
concatenate serovar prediction outputs
antimicrobial resistance screening using Resfinder
hierarchical BAPS clustering
antimicrobial resistance screening using PlasmidFinder
antimicrobial resistance screening using VFDB
final combined primary serovar outputs
tab-separated output file
file with assignment of sequences to clusters
tab-separated output file
tab-separated output file

Computational execution platforms

Characteristics:

- one resource with many nodes and cores
- jobs with long running time
- jobs with low and high memory usage
- jobs with low and high disk usage
- fair-share priority
- limited storage



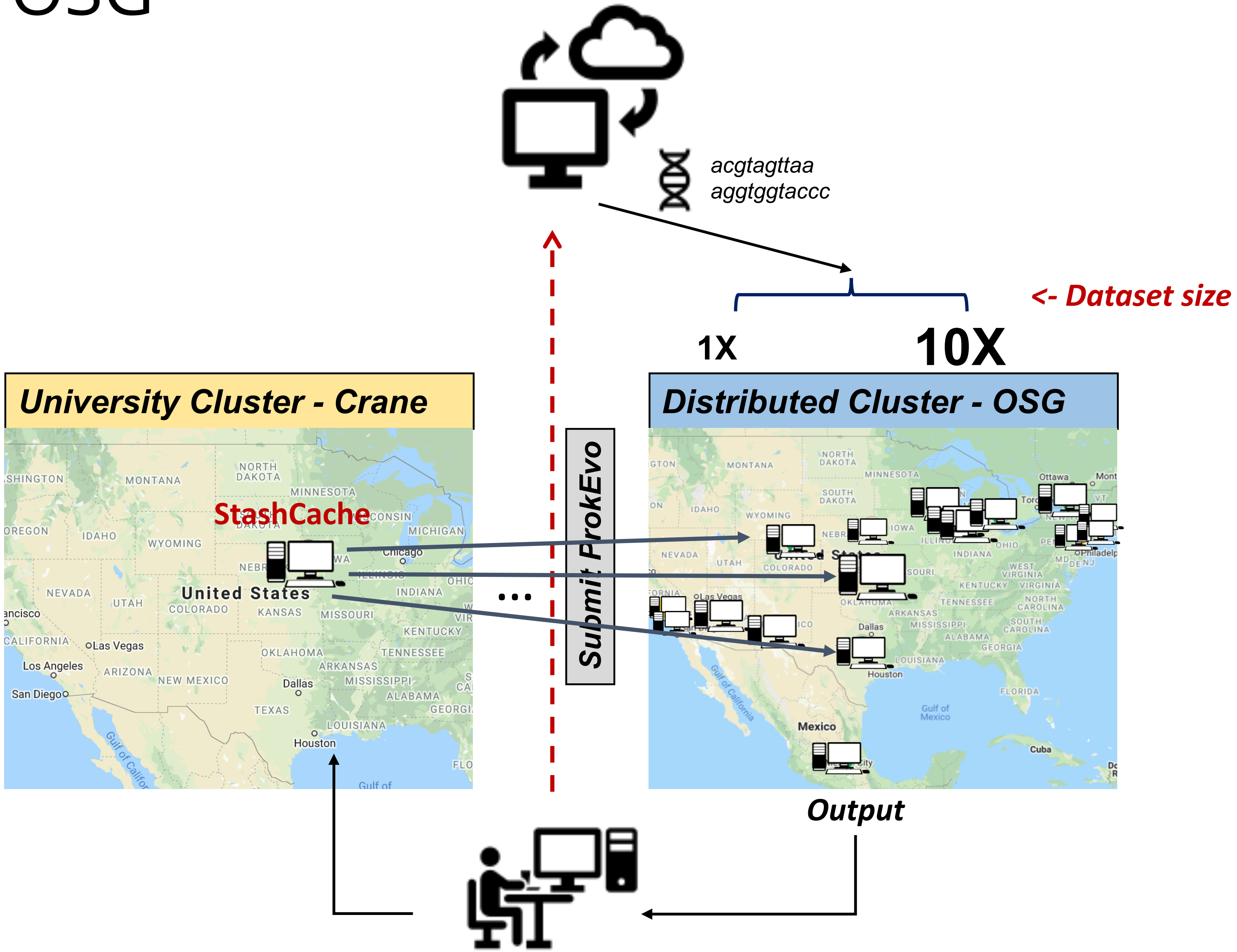
Characteristics:

- many distributed resources
- jobs with short running time
- jobs with low memory usage
- jobs with low disk usage
- jobs with single core
- job preemption
- frequent job restarts

ProkEvo on OSG

- OSG Connect
- Copy data to Crane
- Non-shared file-system
- Singularity
- Downloading data from NCBI
  - Intermittent network errors
  - Download limitations

ProkEvo on OSG



ProkEvo on HCC

- Store data to Crane
- Shared file-system
- Anaconda package manager
- Downloading data from NCBI
  - Intermittent network errors
  - Download limitations

Results

- Performance evaluation

Performance evaluation

	Crane	**OSG**	**Crane**	**OSG**
Number of genomes	2,392		23,045	
Total distributed running time*	3 days 15 hours	7 days 4 hours	15 days 22 hours	26 days 6 hours
Total estimated sequential running time**	115 days 18 hours	1 year 69 days	2 years 268 days	13 years 5 days
Maximum jobs ran in a day***	2,377	8,608	12,382	25,540
Total number of jobs ran	9,281	16,624	217,942	232,422
Output data size	131 GB		1.2 TB	

*Total distributed running time is calculated when many independent tasks are executed simultaneously utilizing single cores. This is the default behavior of ProkEvo.

**Total estimated sequential running time is calculated when all steps from the pipeline are assumed to be run sequentially, on one single core.

***The number of maximum jobs ran in a day depends on the type and length of the job, and is not linear.

Conclusion

- ProkEvo
  - Automated, open-source pipeline for population genomics analyses that uses Pegasus Workflow Management System
  - Utilizing high-performance and high-throughput computational resources
- Applications

Acknowledgement

- IANR Agricultural Research Division and the National Institute for Antimicrobial Resistance Research and Education
- Holland Computing Center of the University of Nebraska
- Open Science Grid

- Mats Rynge
- Dr. Derek Weitzel
- Karan Vahi

Future plans

- Submit ProkEvo from GUI (Science Gateway or OpenOnDemand) and run it on different cyberinfrastructures

Our Pegasus Feedback

- Pegasus is awesome ☺

Notes:

- Modify executables after failure and re-run
- Use only Slurm instead of HTCondor?
- Singularity support for “sharedfs” data configurations
- If a job fails, don’t mark the pipeline as failure, but only skip the tasks that follow after the failed job
- Command line options/wrappers (the tasks defined in the config files can be overridden from the command line)

Thank You!

PUG 2021