Georgia Tech

Team 1 Final Presentation

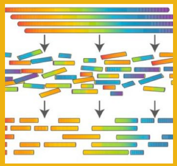
Assembly, Gene Prediction, Functional Annotation, Comparative Genomics, Predictive Webserver

CREATING THE NEXT®

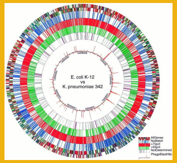
Outline

1 Overall Project Goals

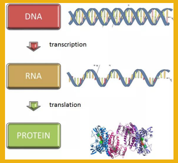
2 Genome Assembly



3 Gene Prediction



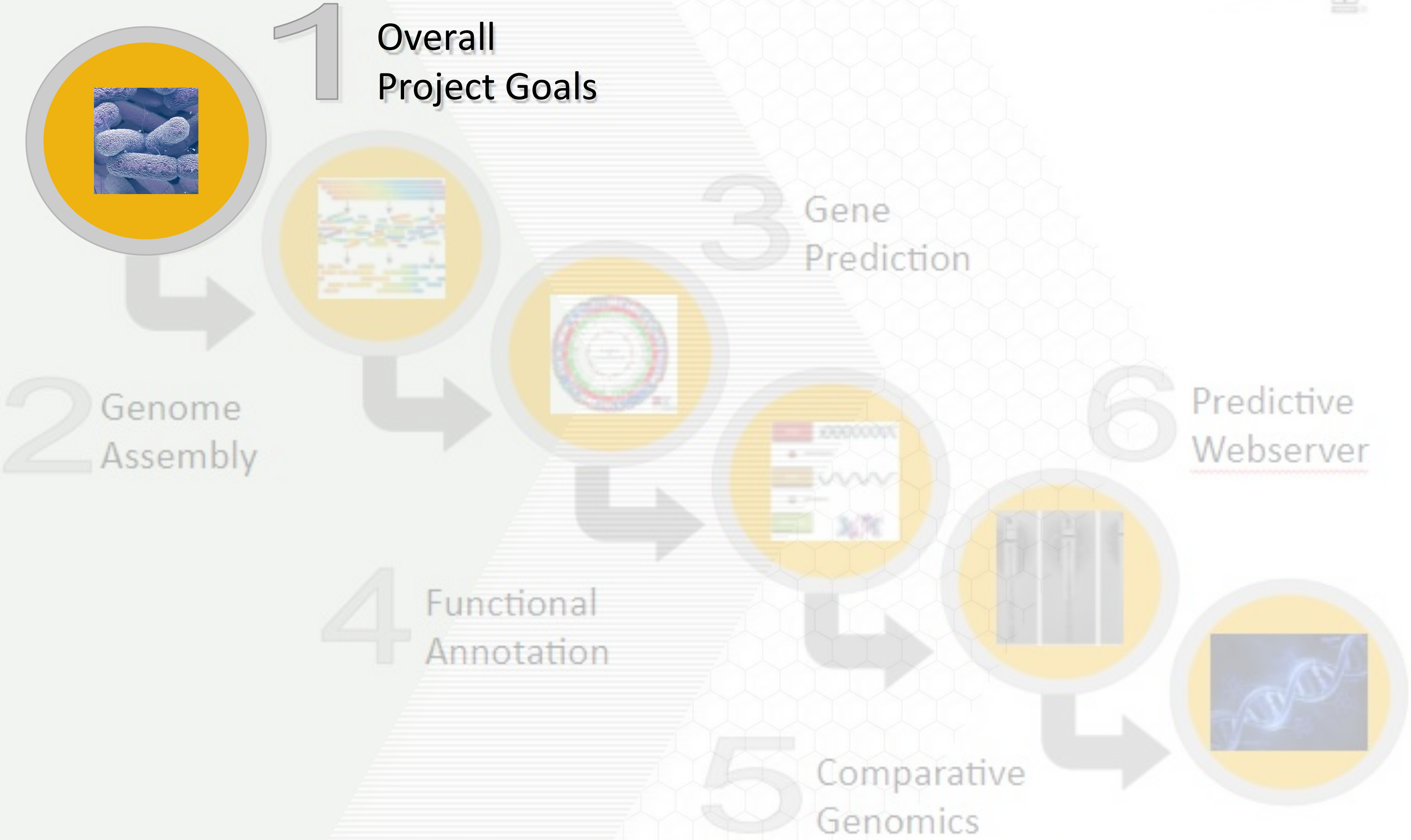
4 Functional Annotation



5 Comparative Genomics

6 Predictive Webserver

Outline



1 Overall Project Goals

2 Genome Assembly

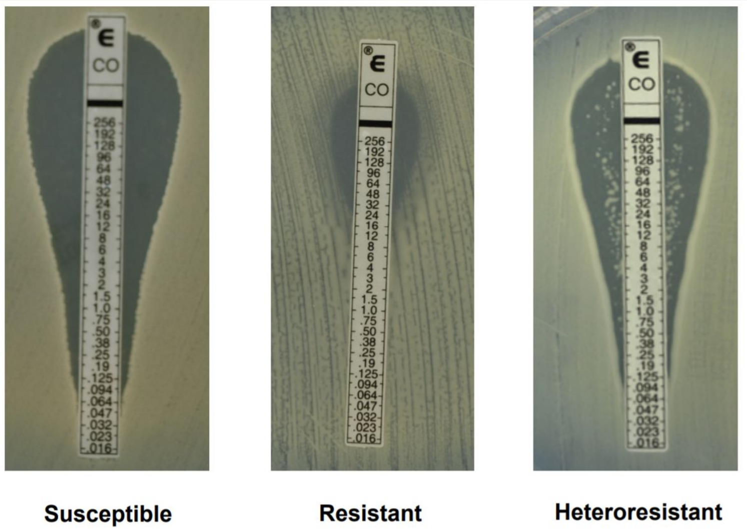
3 Gene Prediction

4 Functional Annotation

5 Comparative Genomics

6 Predictive Webserver

Overall Project Goals



Susceptible Resistant Heteroresistant

1 Vial
Colistimethate Sodium for Injection IP 1 Million IU
Colistin
कोलिस्टिन
1 MIU
Single Use Only
For Intravenous / Inhalation use only
cbc

Outline

1 Overall Project Goals

2 Genome Assembly

3 Gene Prediction

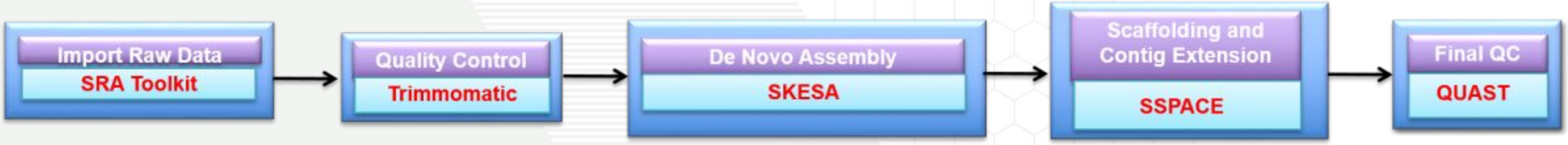
4 Functional Annotation

5 Comparative Genomics

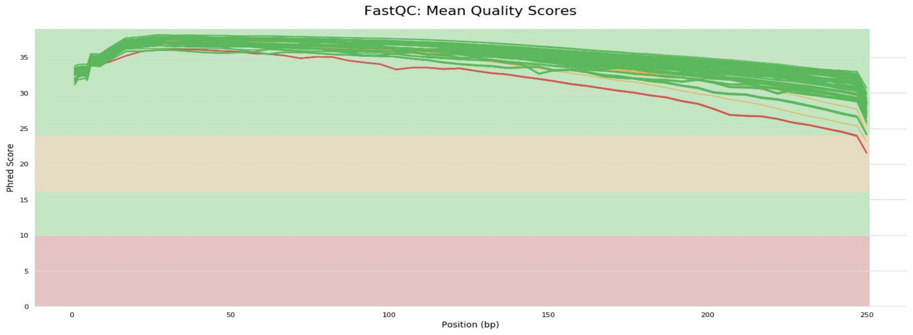
6 Predictive Webserver

Genome Assembly

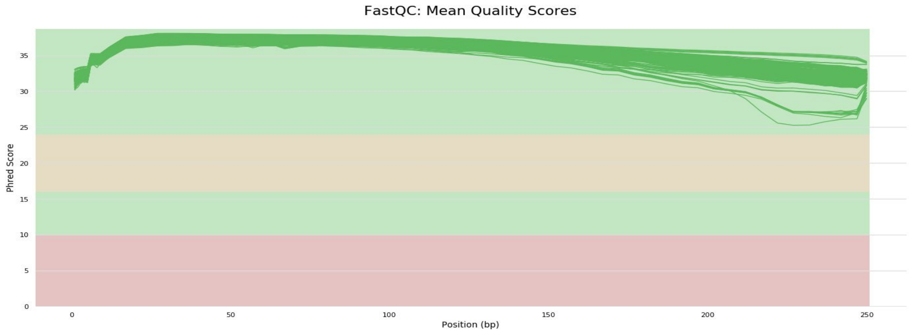
Assembly Pipeline



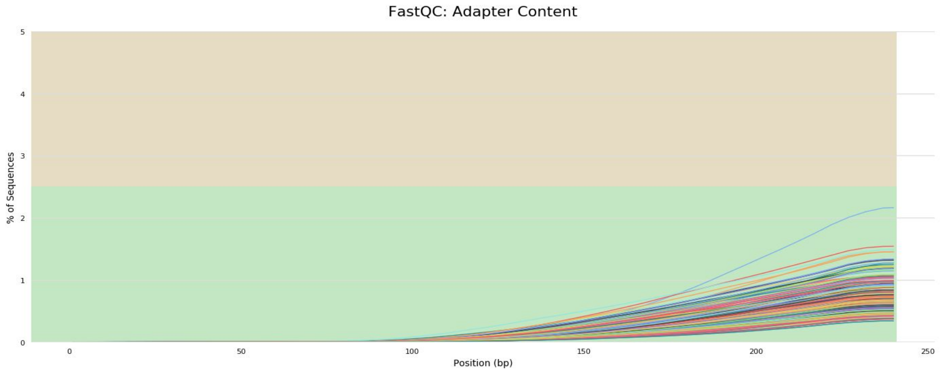
Trimming

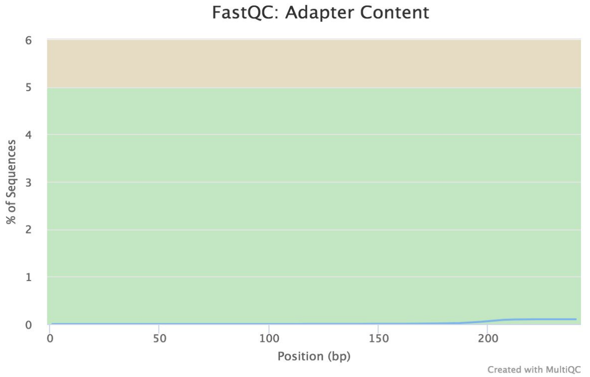


Before Trimming



After Trimming





De Novo Assembly

SKESA -

- The binary for Skesa was provided by CDC
- It is an assembler that works based on DeBruijn graphs
- Creates breaks at repeat regions in genomes
- It works for haploid genome
- Multi-threaded application - so good for scaling

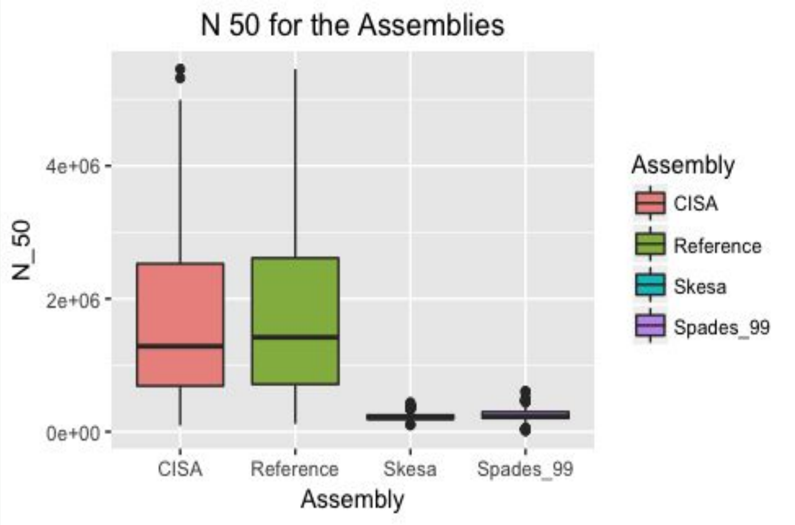
Scaffolding -

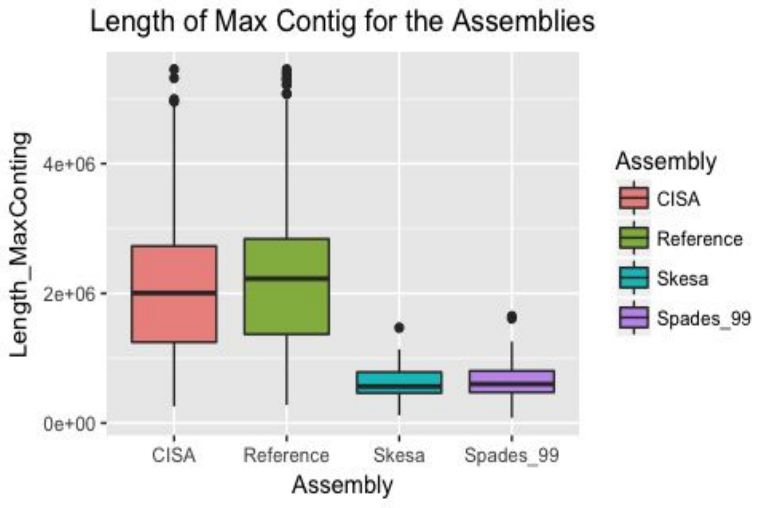
- Scaffolding was performed using SSPACE
- Extends and scaffolds pre-assembled contigs

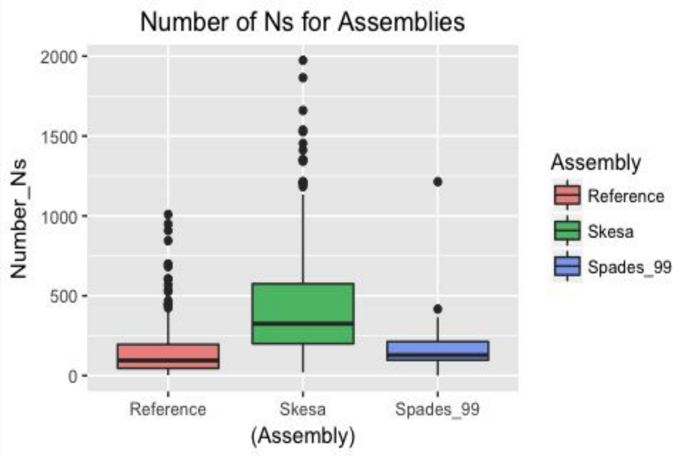
The output contains final scaffolds in fasta format, scaffolds with initial numbered contigs, a log file and a summary file

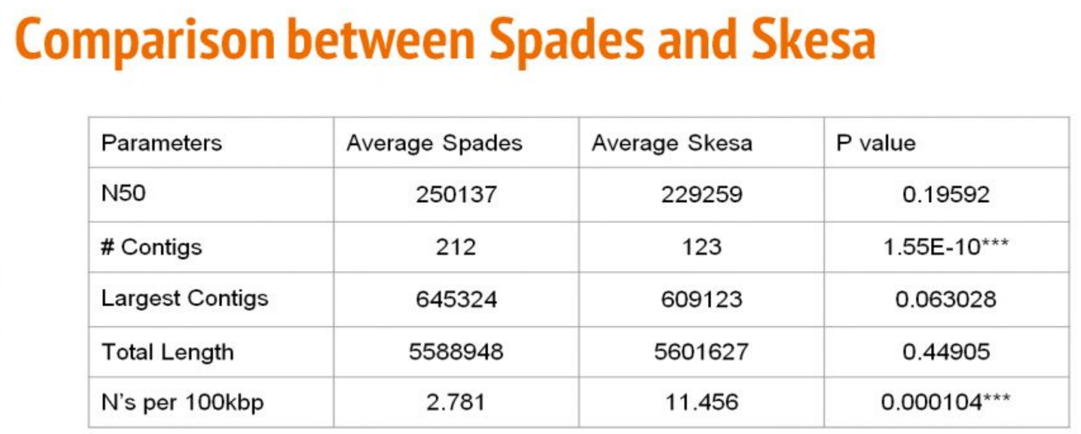
De Novo Assembly

Quality of assemblies





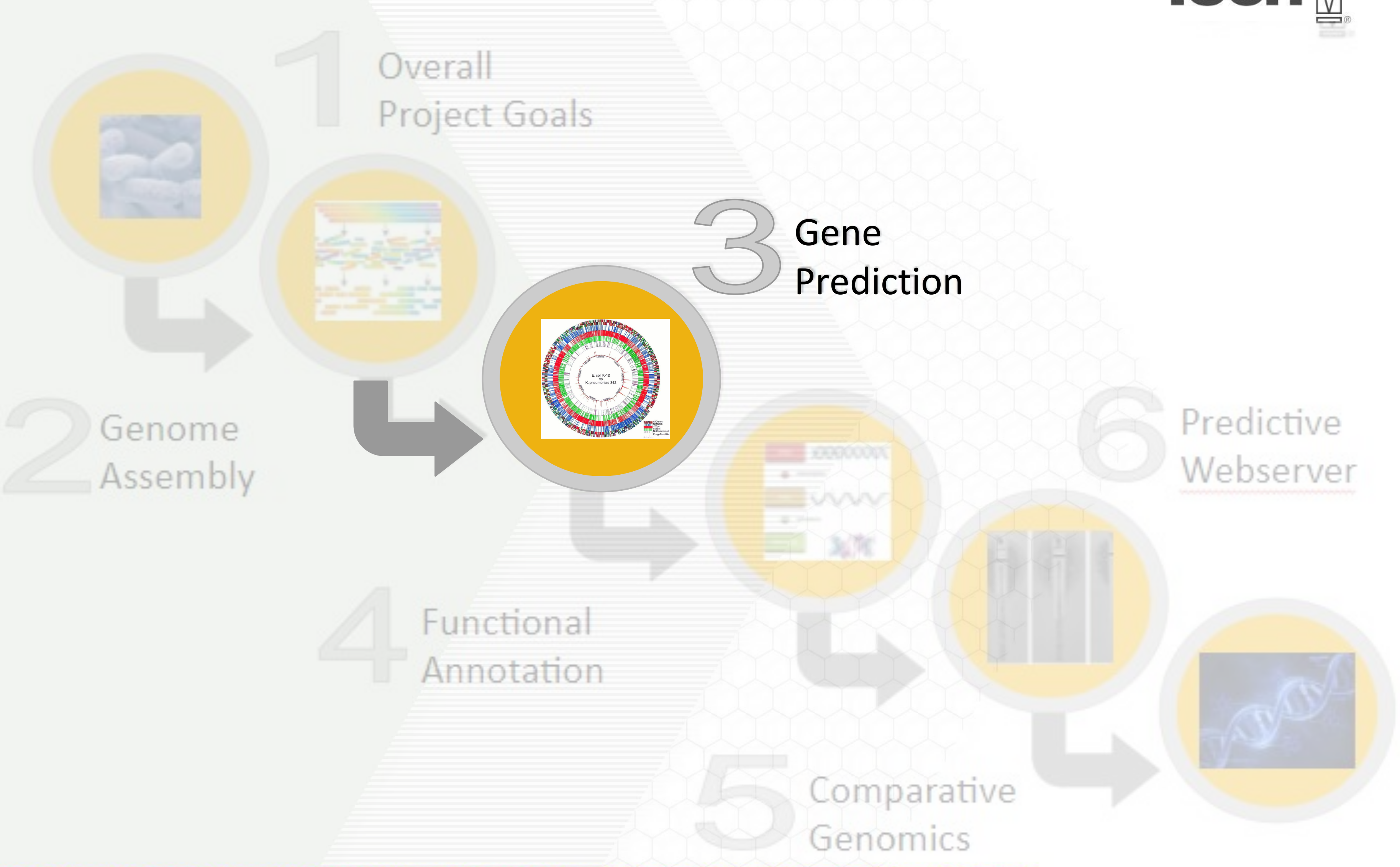




Comparison between Spades and Skesa

Parameters	Average Spades	Average Skesa	P value
N50	250137	229259	0.19592
# Contigs	212	123	1.55E-10***
Largest Contigs	645324	609123	0.063028
Total Length	5588948	5601627	0.44905
N's per 100kbp	2.781	11.456	0.000104***

Outline



1 Overall Project Goals

2 Genome Assembly

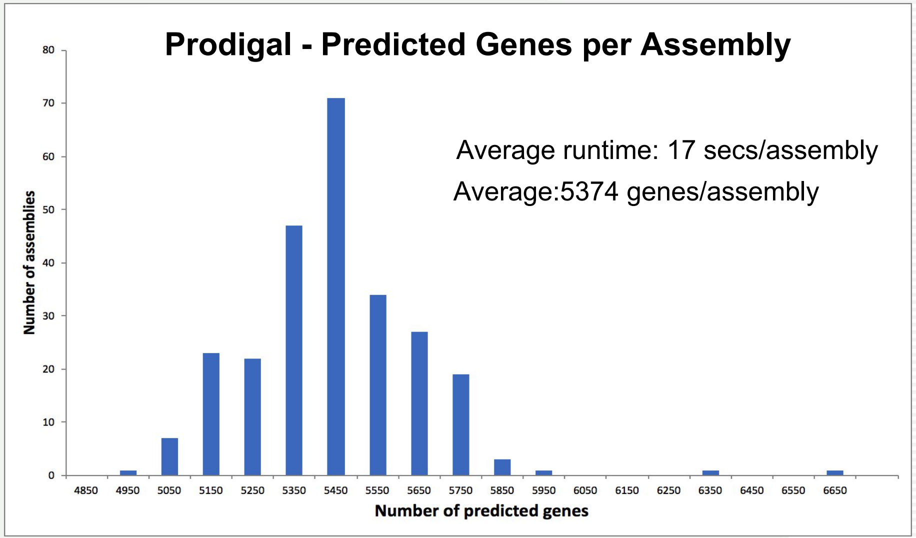
3 Gene Prediction

4 Functional Annotation

5 Comparative Genomics

6 Predictive Webserver

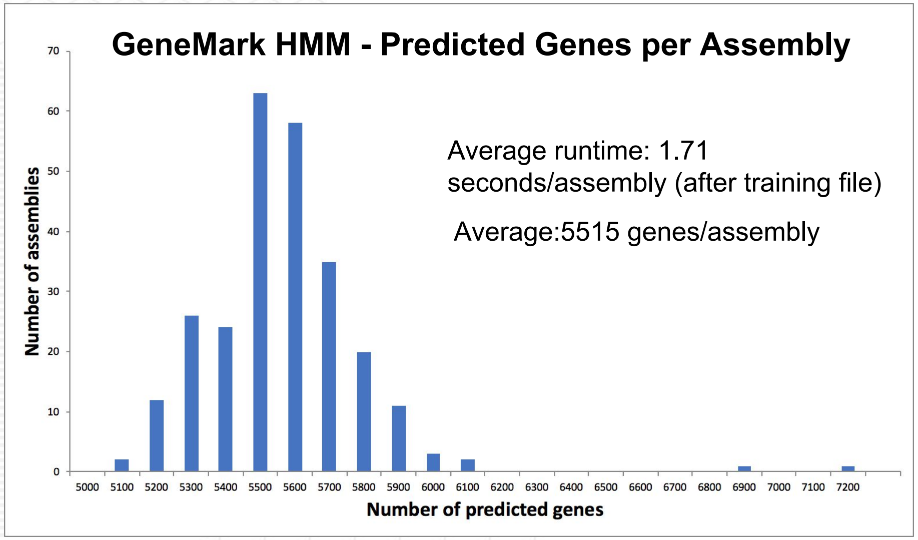
Gene Prediction Methods



Prodigal - Predicted Genes per Assembly

Average runtime: 17 secs/assembly

Average:5374 genes/assembly

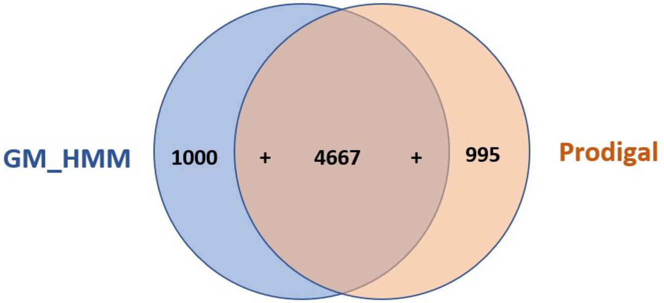


GeneMark HMM - Predicted Genes per Assembly

Average runtime: 1.71 seconds/assembly (after training file)

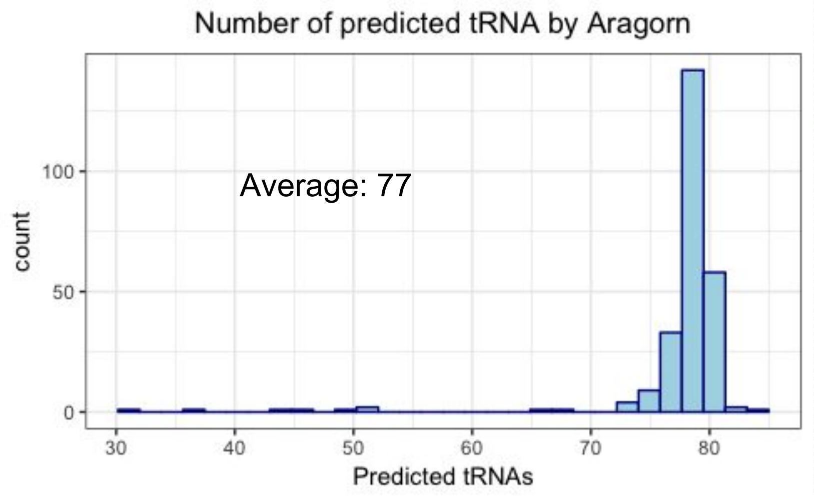
Average:5515 genes/assembly

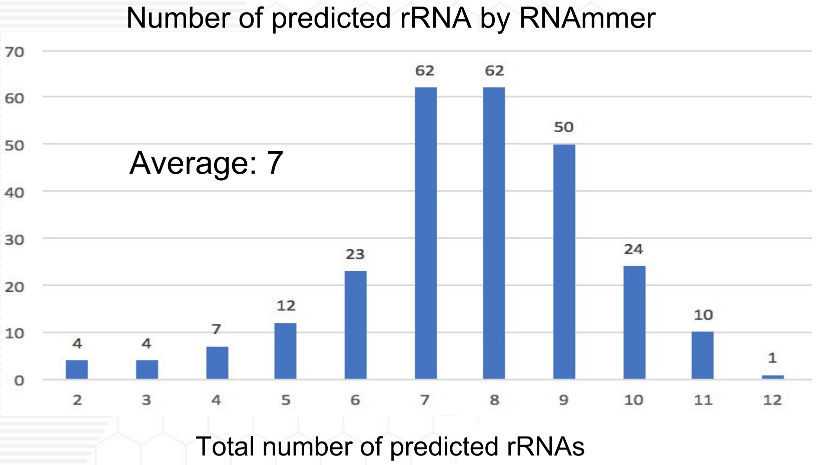
Method	True Positives	False Positives	False negatives	Sensitivity	PPV
Prodigal	5015.8	437.7	480.6	91.2	92.0
GeneMark HMM	5061.5	507.1	456.4	91.7	91.1
Intersect	4383.4	323.6	1096.4	80.0	93.1
Union	5693.9	618.5	423.1	93.1	90.25



Gene Prediction Methods

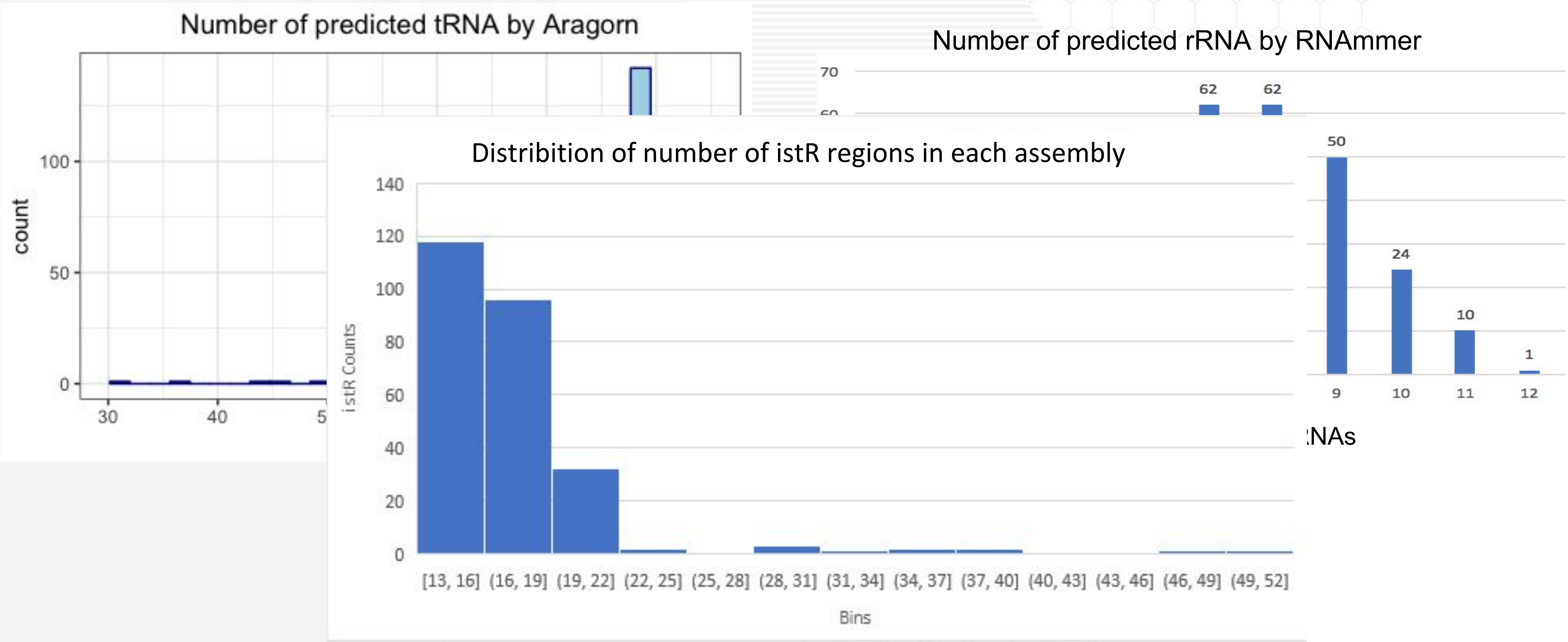
	Aragon	RNAmmer	Infernal
Method	Homology Base	ab initio	ab initio
Run Time	<10 sec/assembly	~3min/assembly	<10 min/assembly
Type of RNA	tRNA	rRna	istR





Gene Prediction Methods

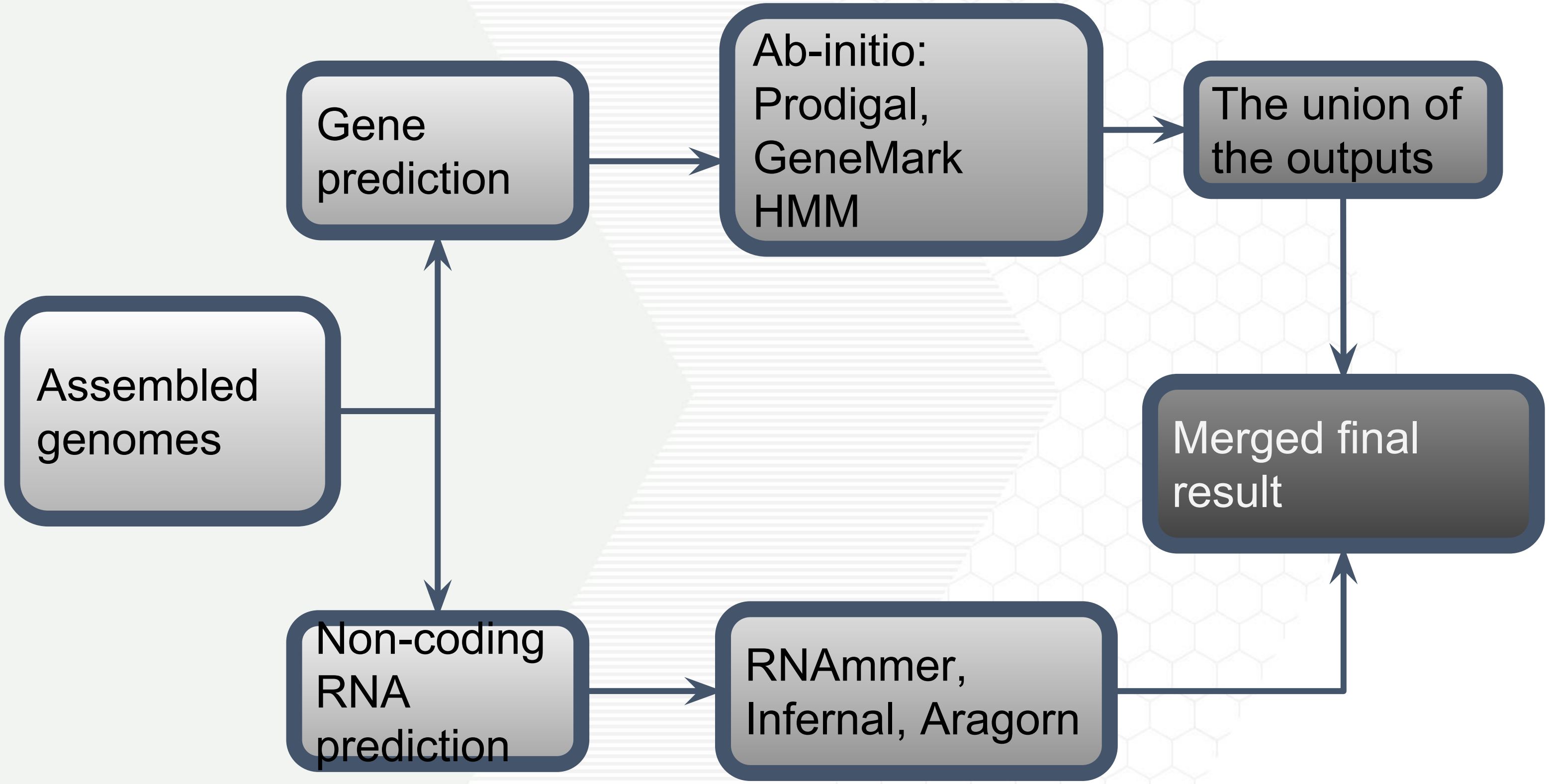
	Aragon	RNAmmer	Infernal
Method	Homology Base	ab-initio	ab-initio
Run Time	<10 sec/assembly	~3min/assembly	<10 min/assembly
Type of RNA	tRNA	rRna	istR


Number of predicted tRNA by Aragorn

Number of predicted rRNA by RNAmmer

Distribition of number of istR regions in each assembly

Gene Prediction Final Pipeline


Assembled genomes

Gene prediction

Ab-initio: Prodigal, GeneMark HMM

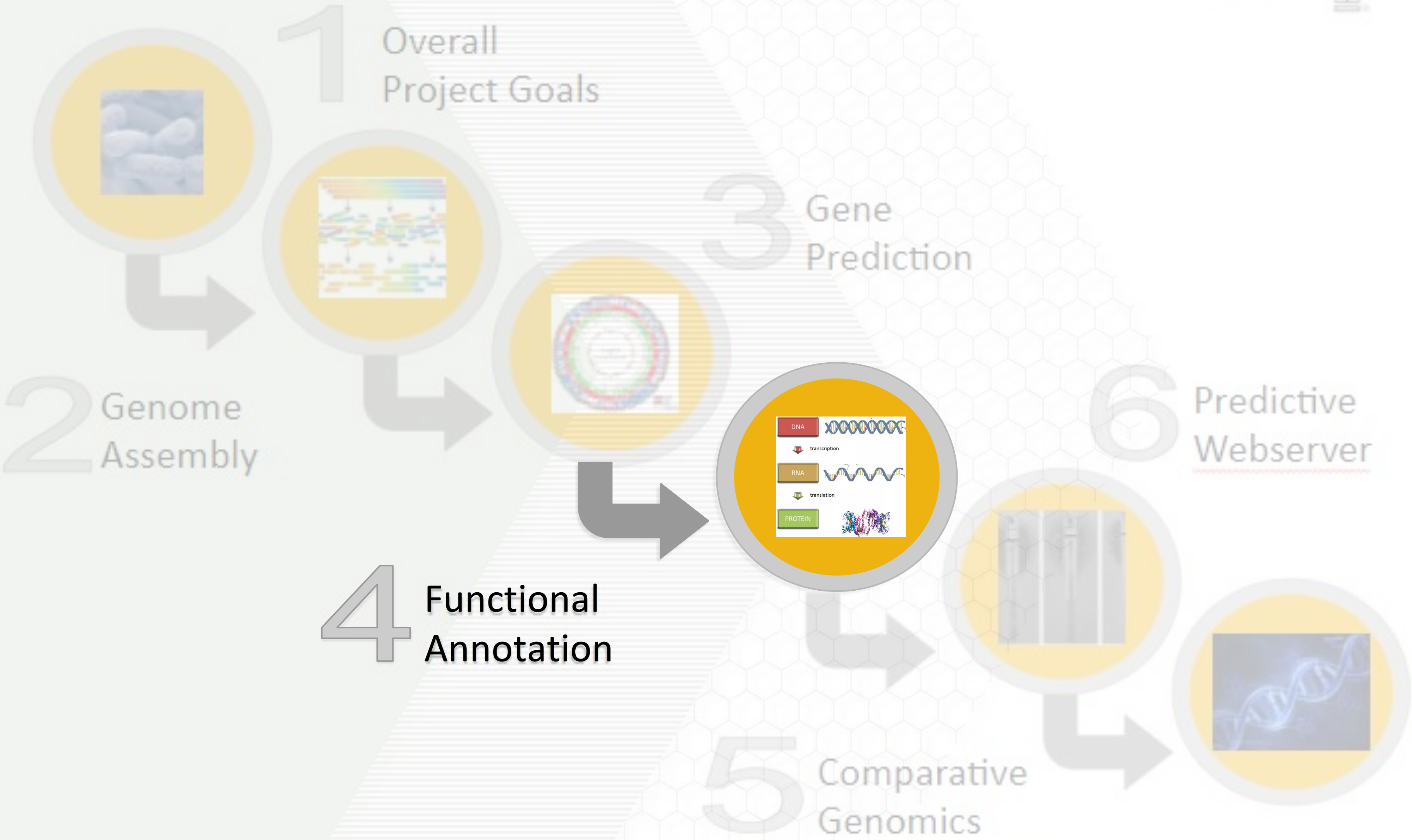
The union of the outputs

Non-coding RNA prediction

RNAmmer, Infernal, Aragorn

Merged final result

Outline



1 Overall Project Goals

2 Genome Assembly

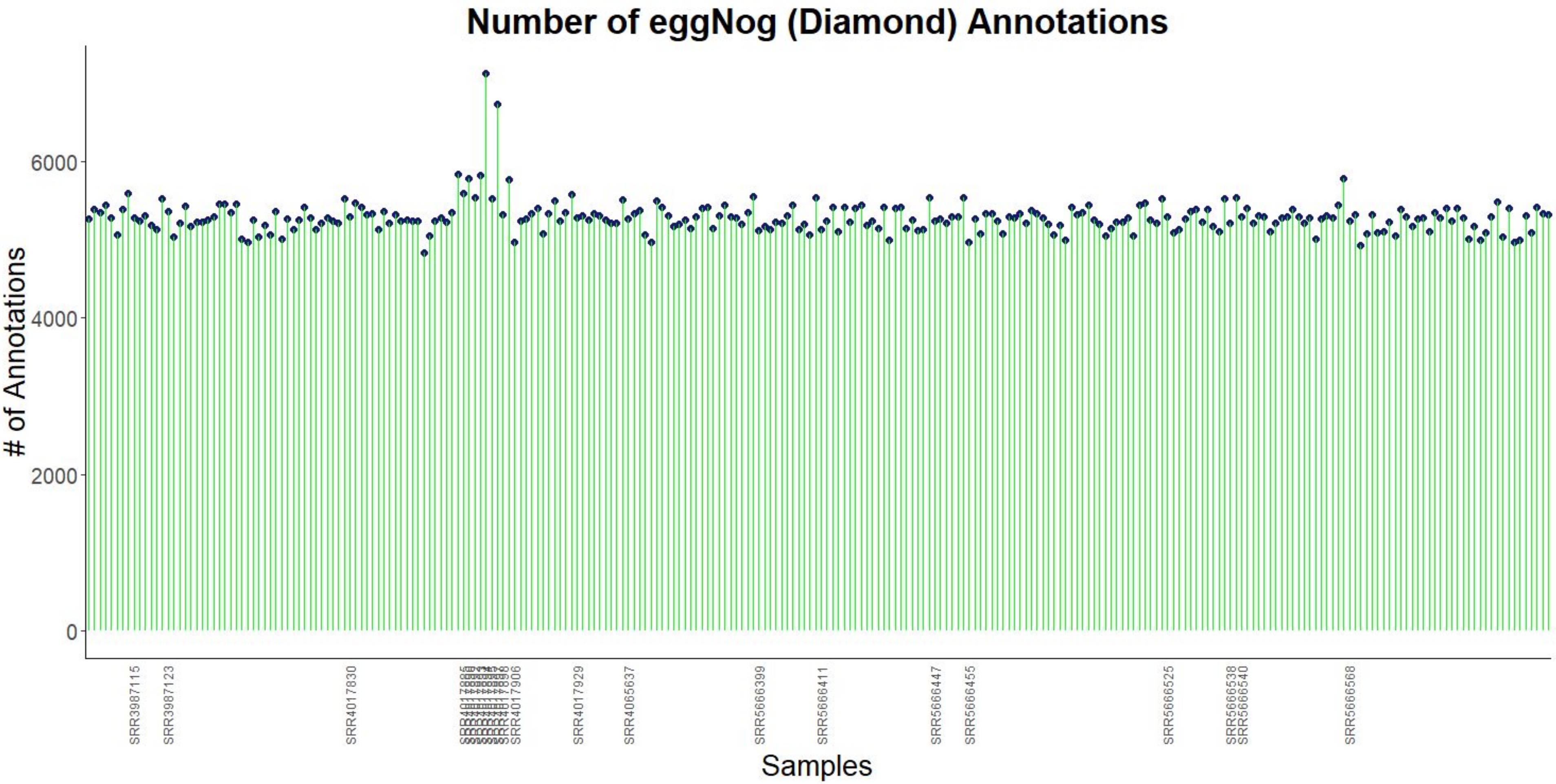
3 Gene Prediction

4 Functional Annotation

5 Comparative Genomics

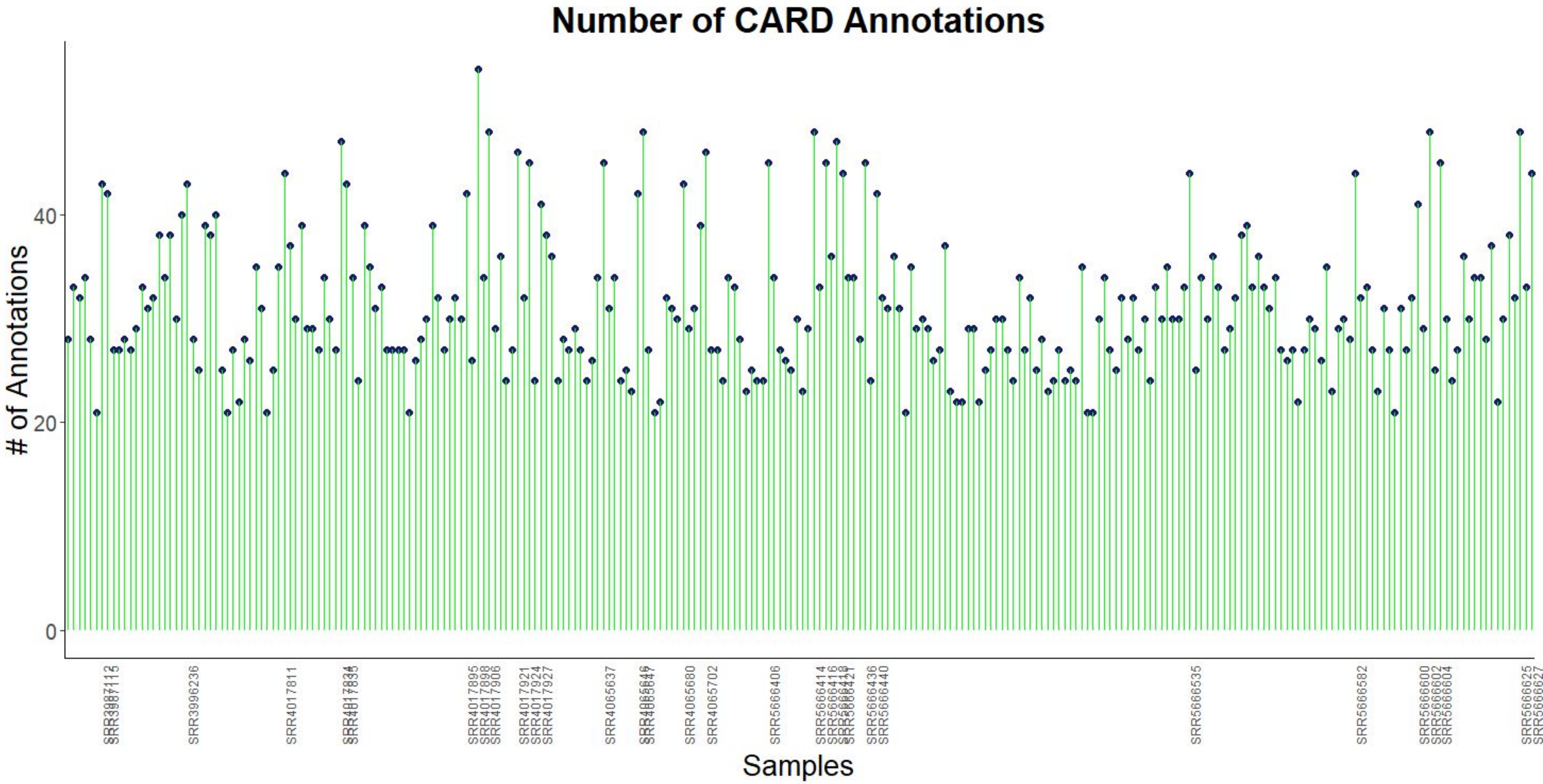
6 Predictive Webserver

Functional Annotation Methods



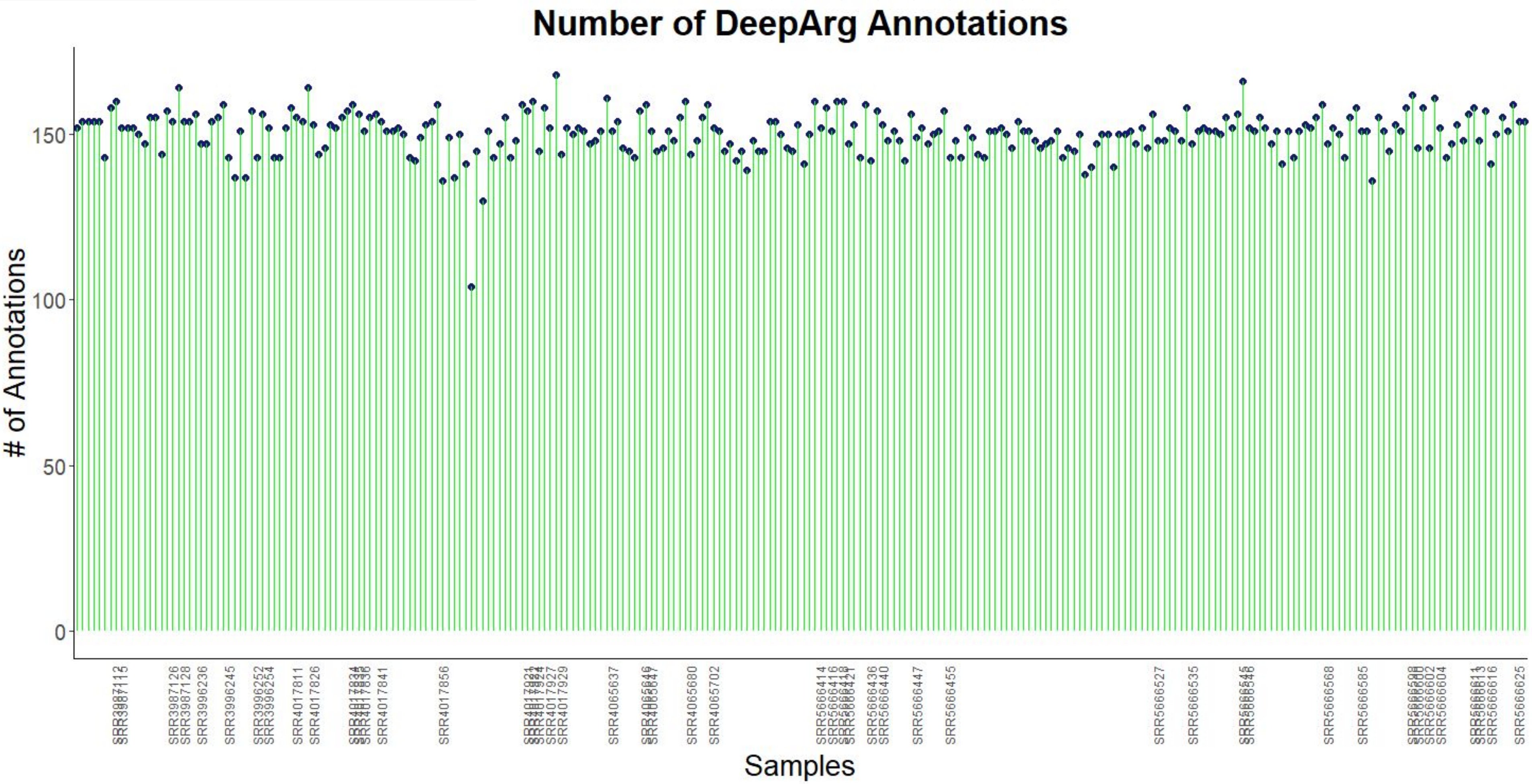
CREATING THE NEXT®

Functional Annotation Methods

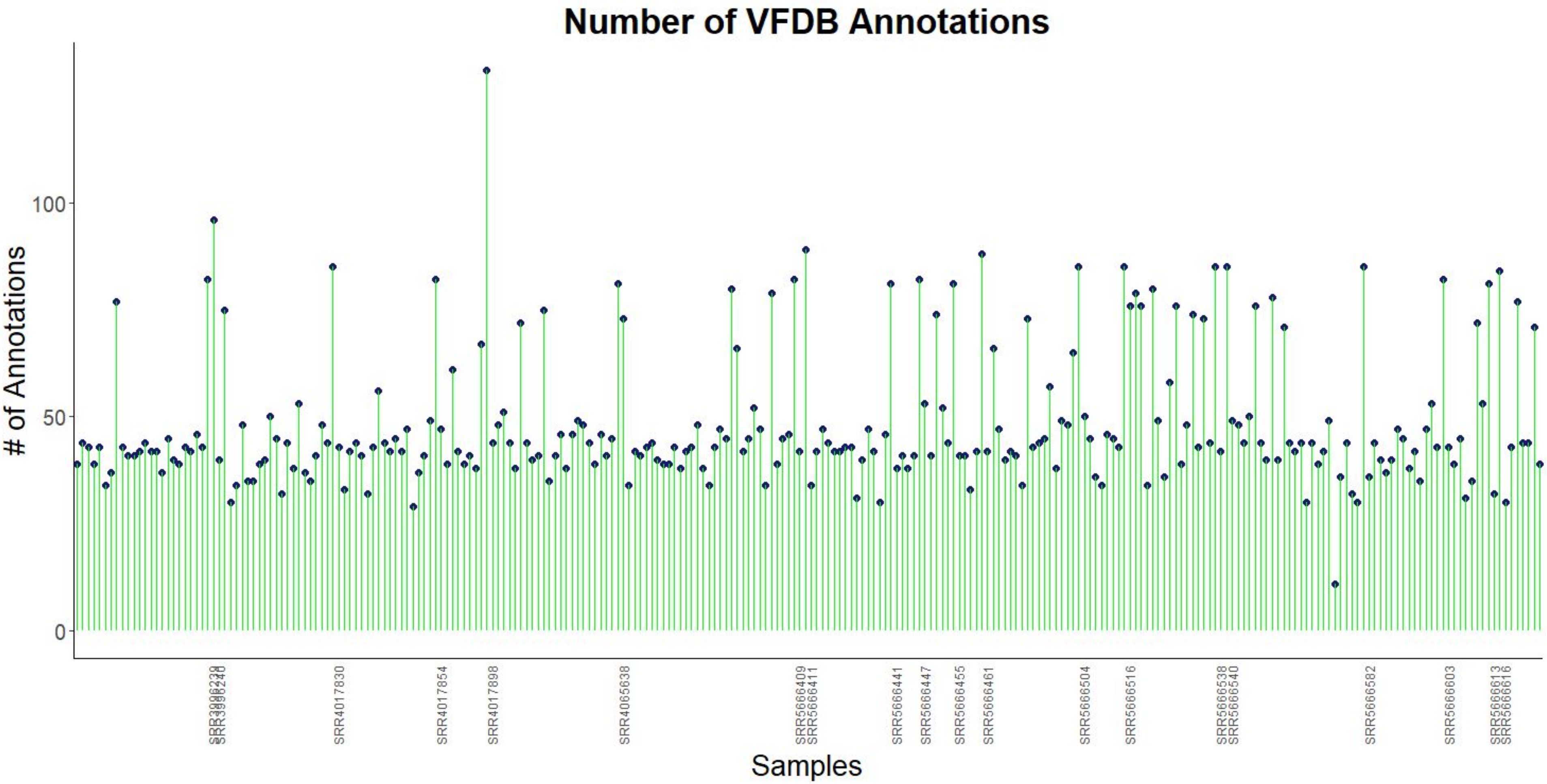


Number of CARD Annotations

Functional Annotation Methods



Functional Annotation Methods

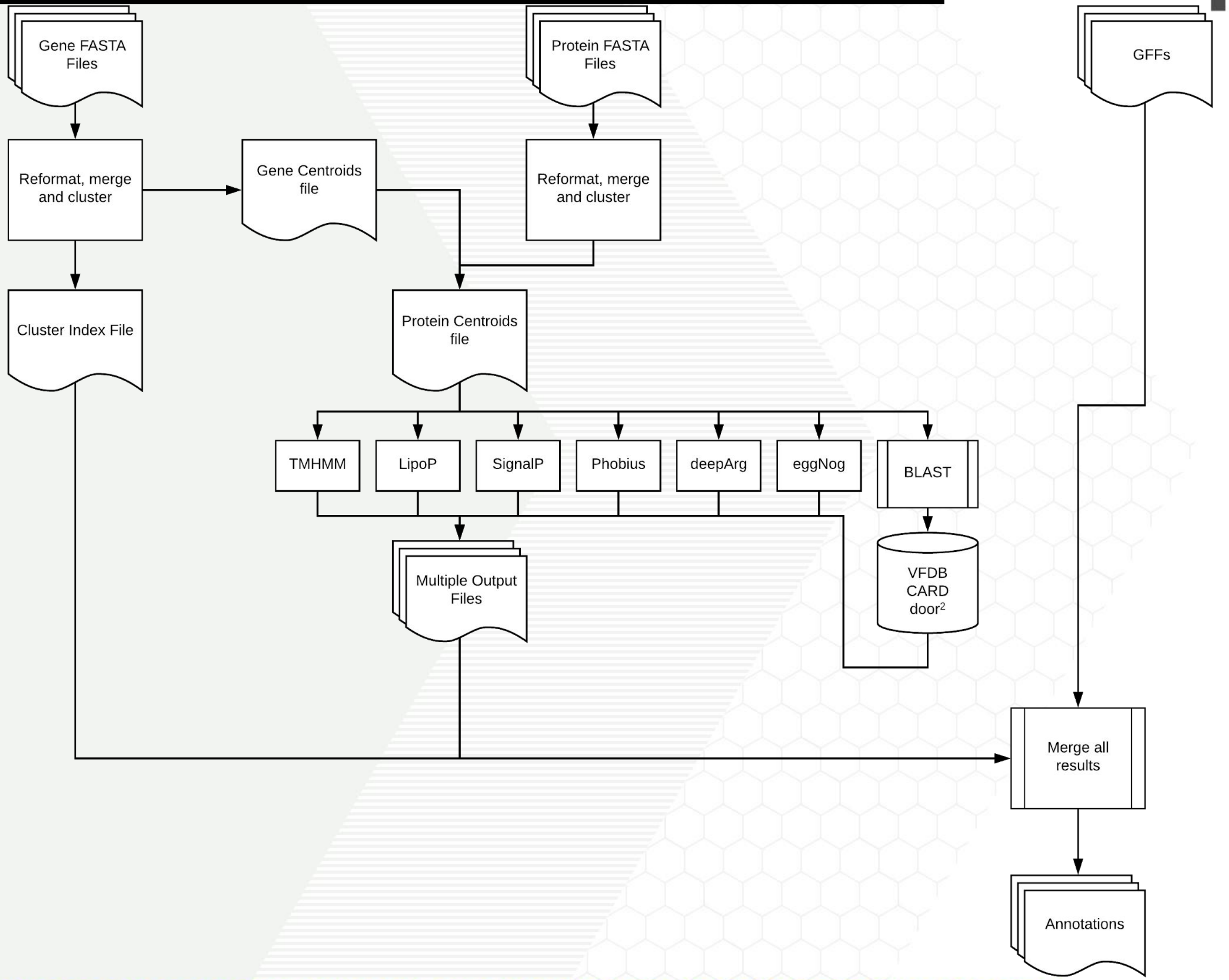


Functional Annotation Methods

Overall statistics

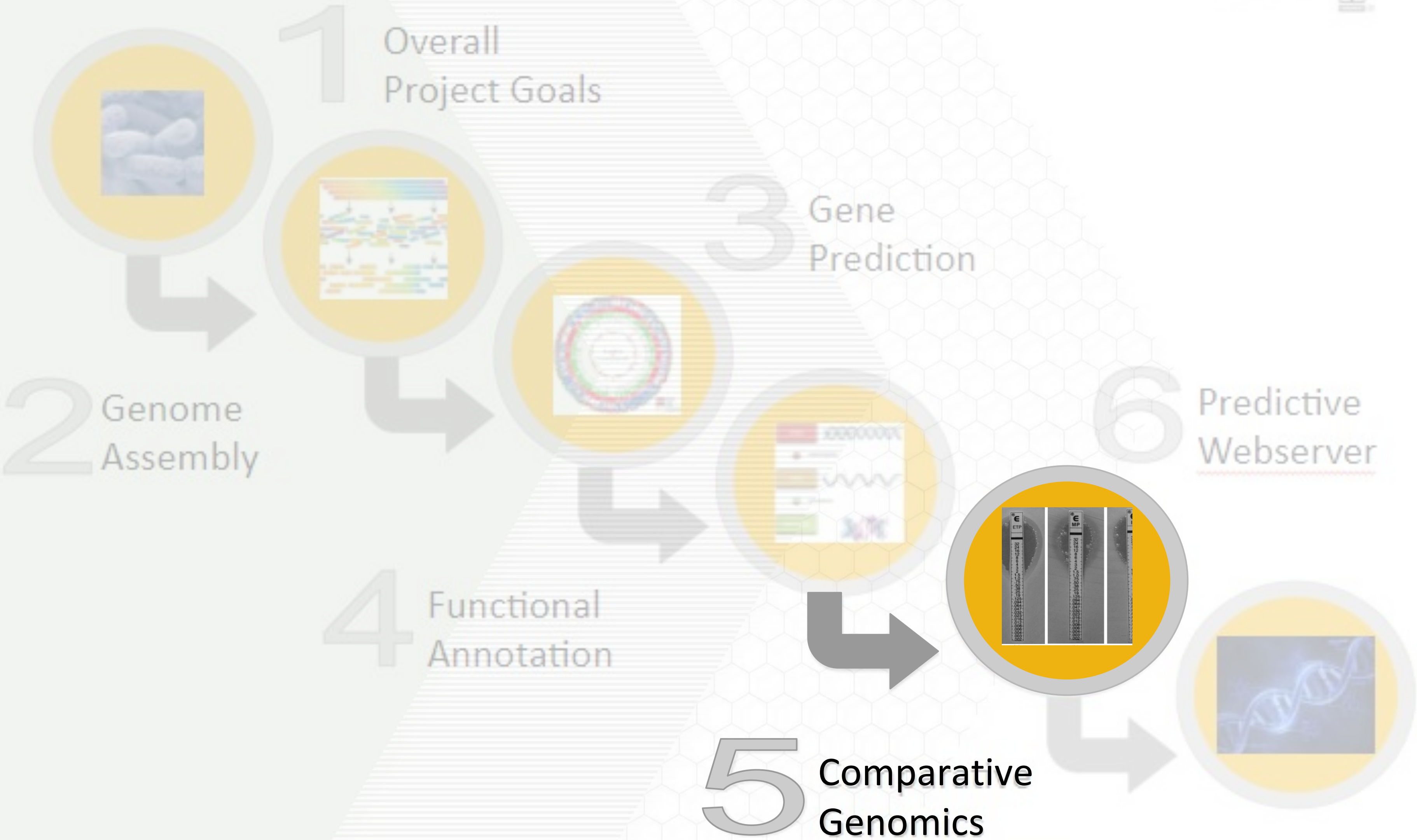
Tools / Database	Total # of annotations	Average # of annotations
CARD	8,051	31
DeepARG	38,799	150
Door2	1,330,879	5158
Eggnog (diamond)	1,364,546	5289
GeneMark.hmm	110,235	427
LipoP	1,502,024	5822
Phobius	1,501,560	5820
Prodigal	1,391,789	5395
SignalP	1,501,569	5820
TMHMM	1,501,577	5820
VFDB	12,517	49

Functional Annotation Final Pipeline



CREATING THE NEXT®

Outline



1 Overall Project Goals

2 Genome Assembly

3 Gene Prediction

4 Functional Annotation

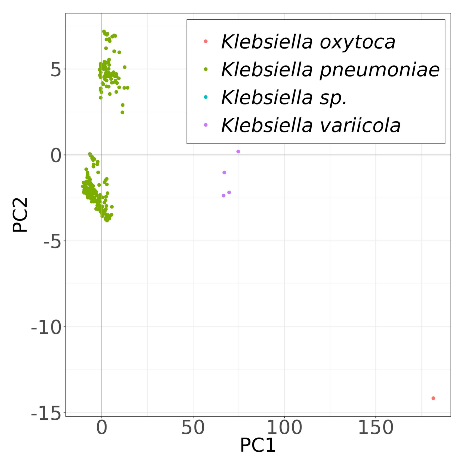
5 Comparative Genomics

6 Predictive Webserver

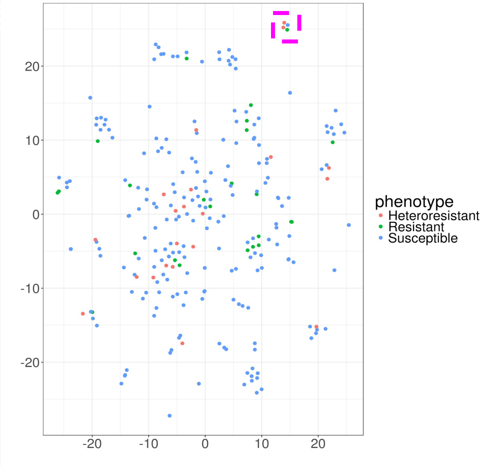
Comparative Genomics Final Pipeline

Georgia Tech

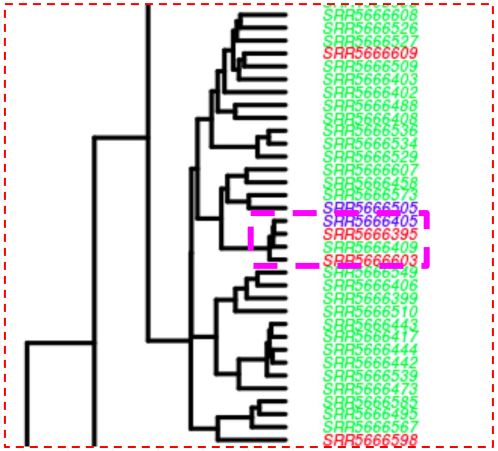
PCA analysis



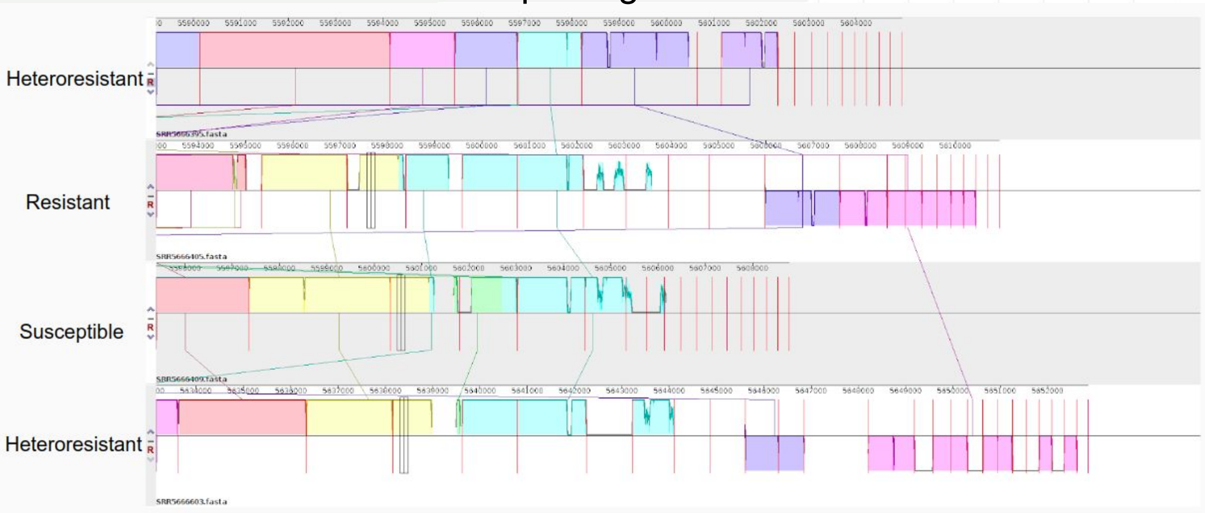
tSNE



Hierarchical clustering

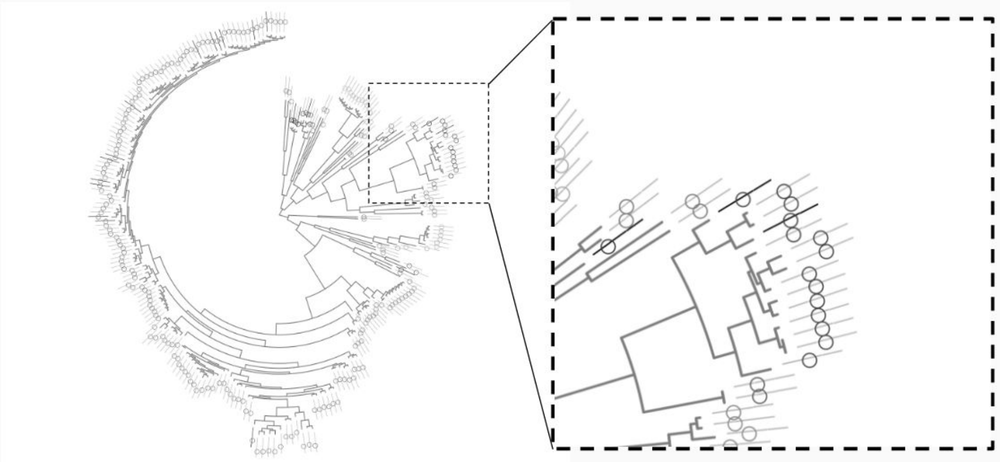


Multiple alignment

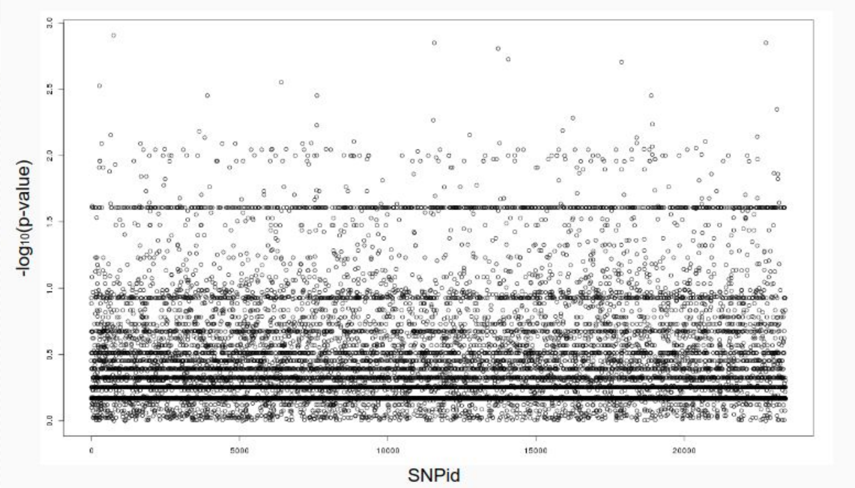


Comparative Genomics Methods

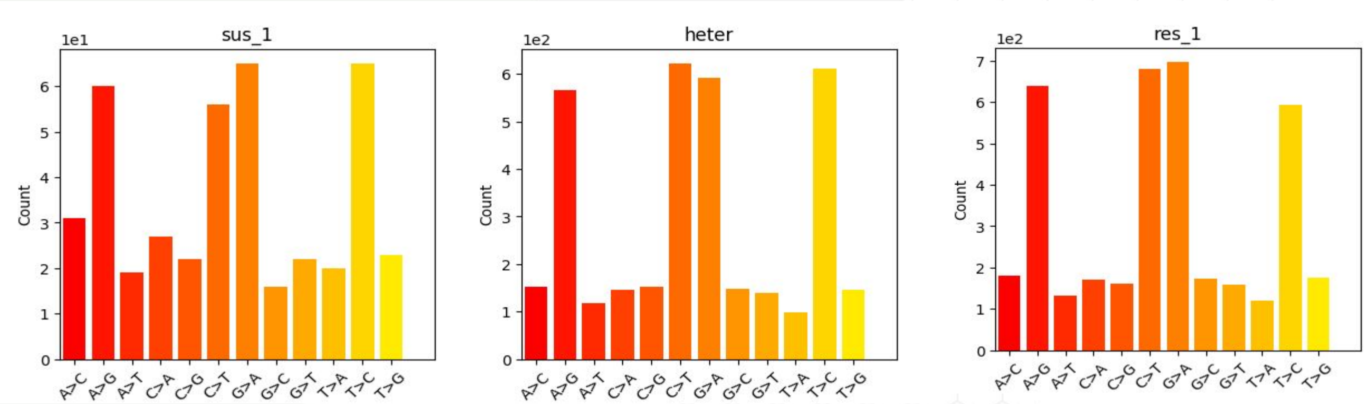
buGWAS



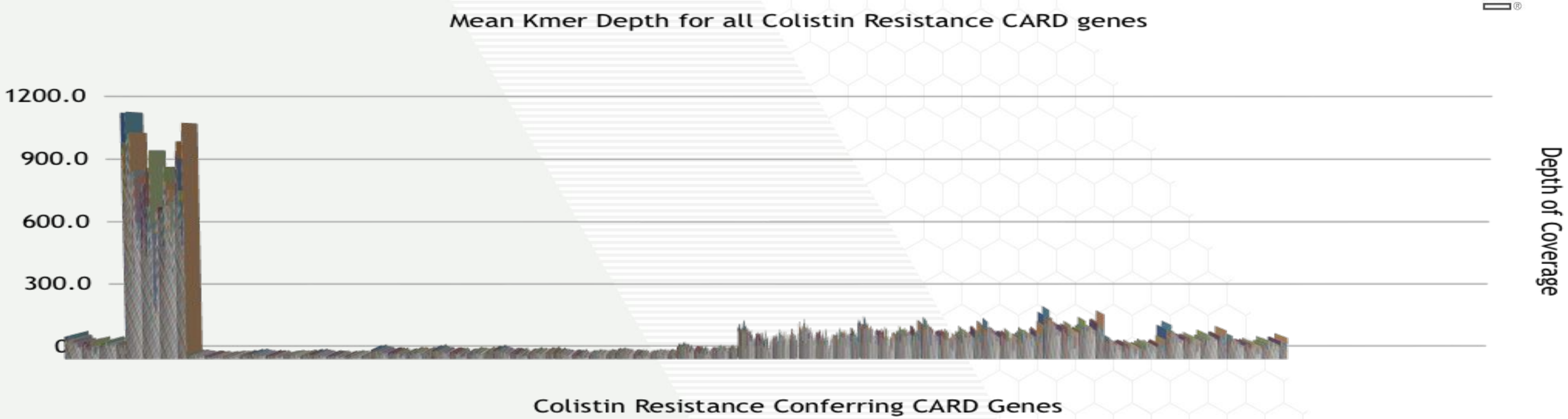
bacterialGWAS

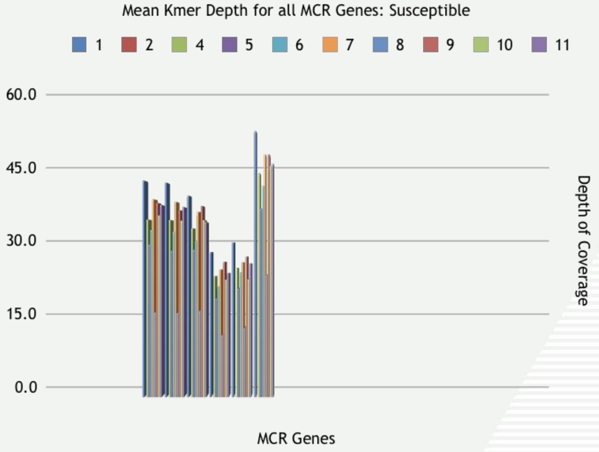


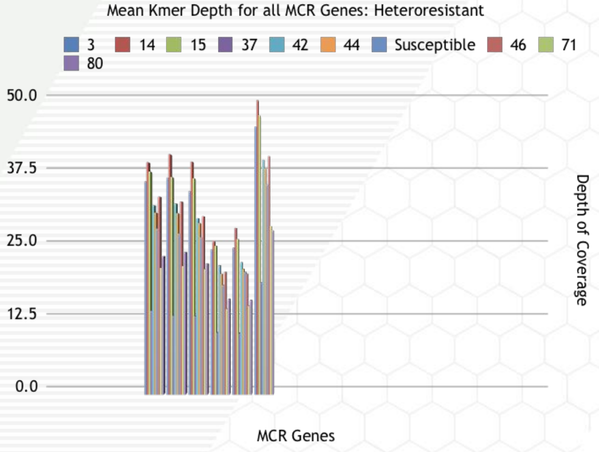
Substitution rates

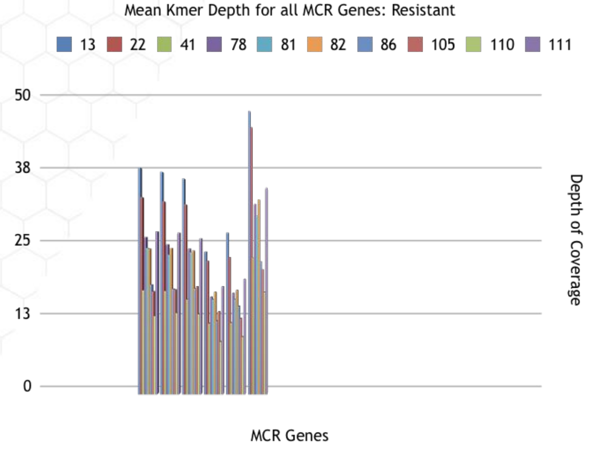


Comparative Genomics Methods


Mean Kmer Depth for all Colistin Resistance CARD genes

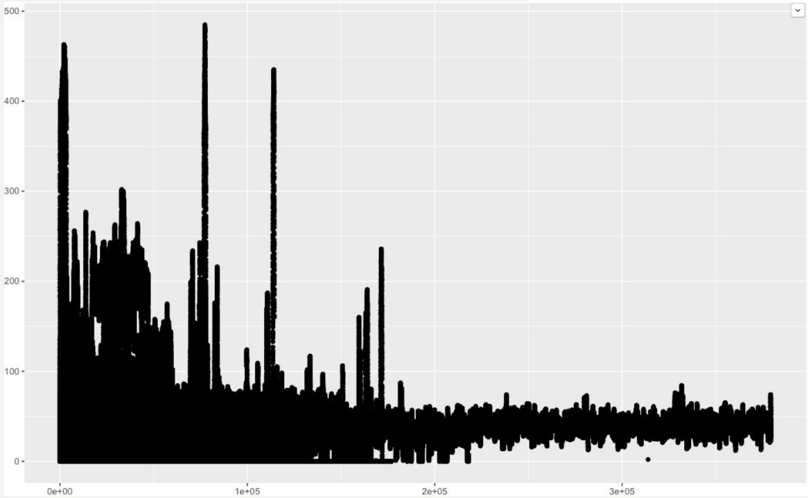

Mean Kmer Depth for all MCR Genes: Susceptible


Mean Kmer Depth for all MCR Genes: Heteroresistant

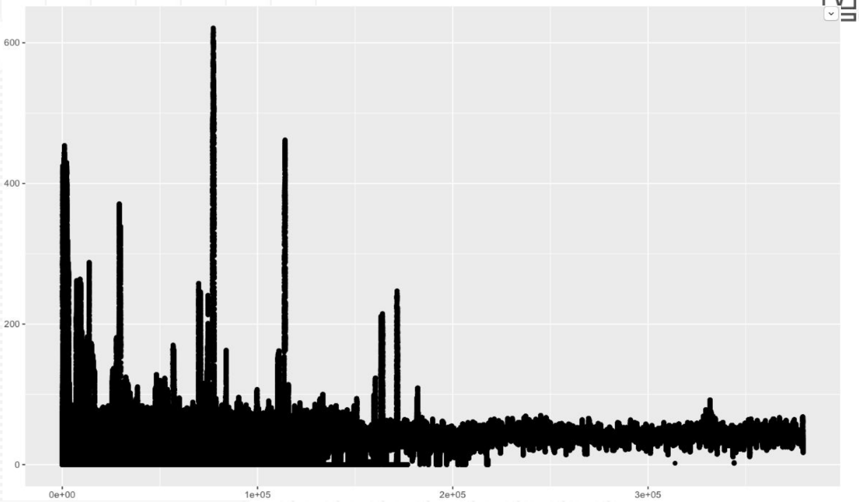

Mean Kmer Depth for all MCR Genes: Resistant

CREATING THE NEXT®

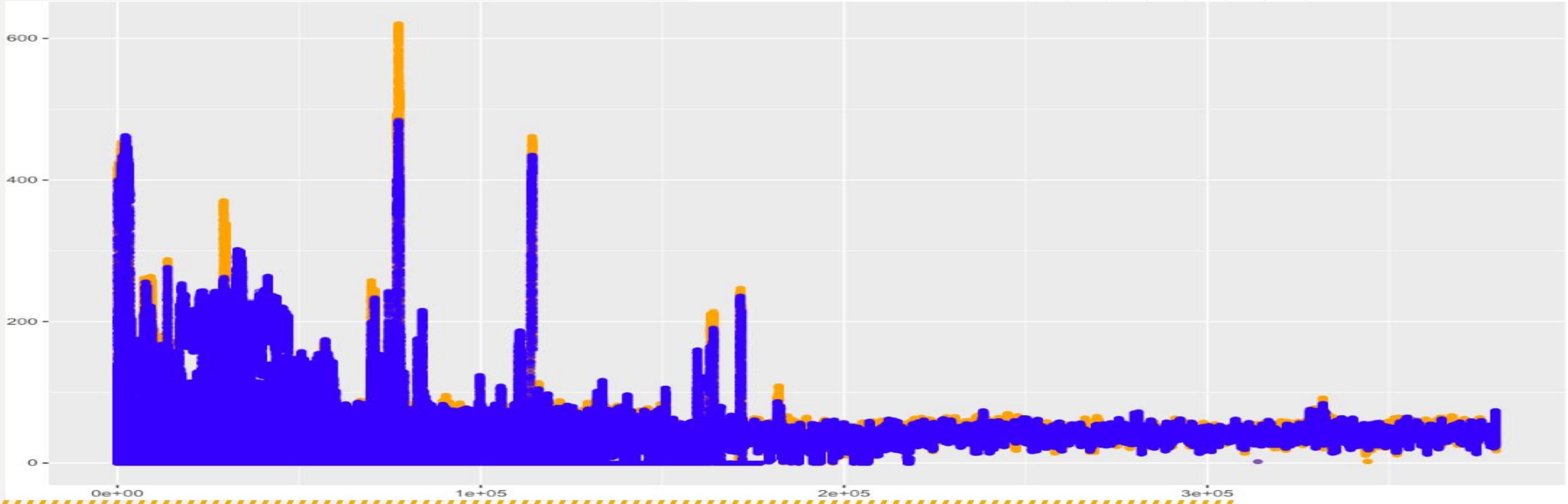
Comparative Genomics Methods



Susceptible



Heteroresistant



Outline

1 Overall Project Goals

2 Genome Assembly

3 Gene Prediction

4 Functional Annotation

5 Comparative Genomics

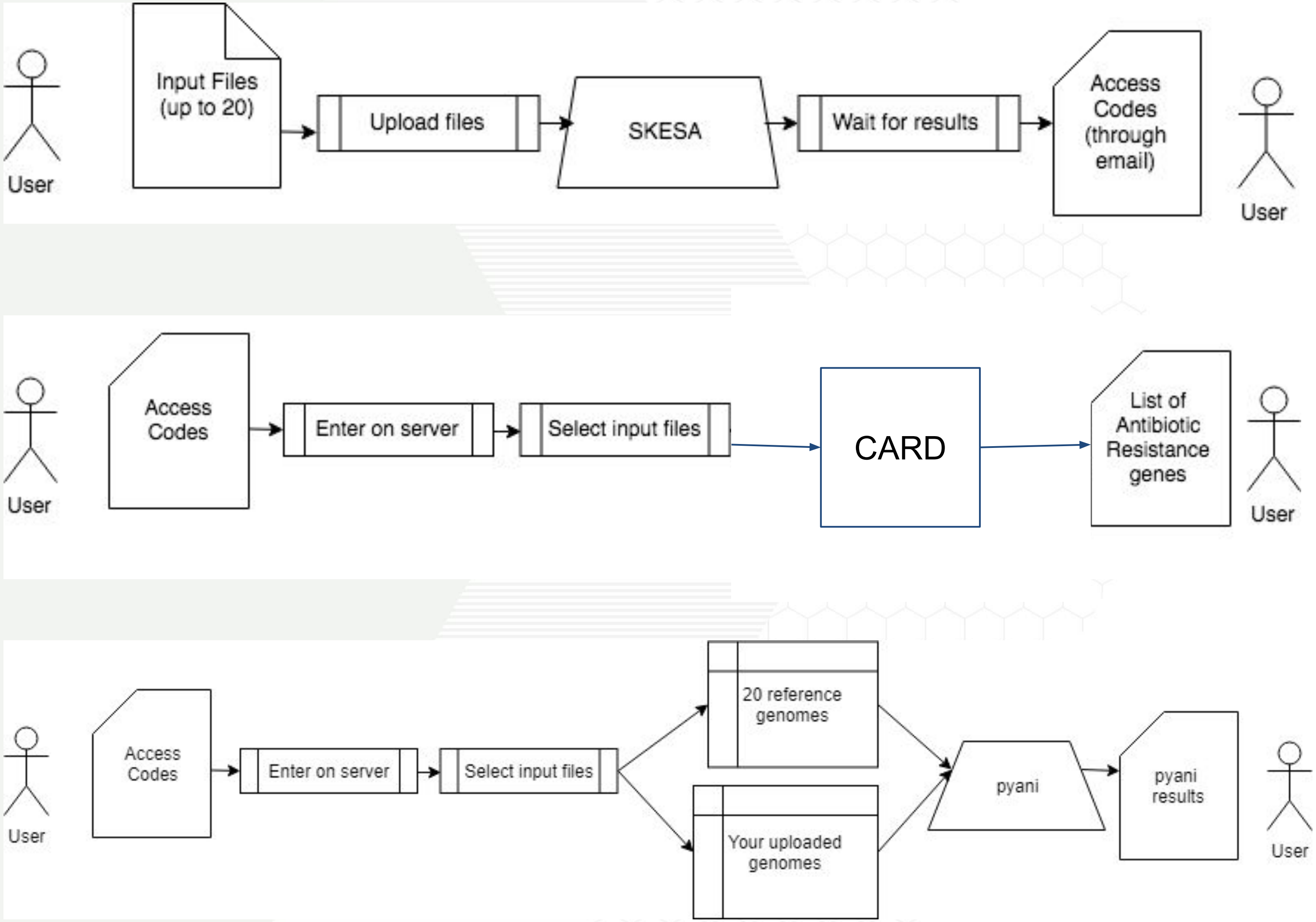
6 Predictive Webserver

Predictive Webserver Goals

Goal - to build a tool that can provide high quality assemblies and relevant information about genomes

- Computational phenotyping
- Assembly
- Average Nucleotide Identity
- Strain Seeker

Predictive Webserver Methods



Predictive Webserver Demo

CREATING THE NEXT®

Sources

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. "Trimmomatic: a flexible trimmer for Illumina sequence data." *Bioinformatics* 30.15 (2014): 2114-2120.

SRA Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2010-. Download Guide. 2009 Sep 9 [Updated 2016 Jan 14]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK242621/

Boetzer, Marten, et al. "Scaffolding pre-assembled contigs using SSPACE." *Bioinformatics* 27.4 (2010): 578-579.

Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi and Glenn Tesler, QUAST: quality assessment tool for genome assemblies, *Bioinformatics* (2013) 29 (8): 1072-1075. doi: 10.1093/bioinformatics/btt086. First published online: February 19, 2013

Hyatt, Doug, et al. "Prodigal: prokaryotic gene recognition and translation initiation site identification." *BMC bioinformatics* 11.1 (2010): 119.

Lukashin, Alexander V., and Mark Borodovsky. "GeneMark. hmm: new solutions for gene finding." *Nucleic acids research* 26.4 (1998): 1107-1115.

Quinlan, Aaron R., and Ira M. Hall. "BEDTools: a flexible suite of utilities for comparing genomic features." *Bioinformatics* 26.6 (2010): 841-842.

Laslett, Dean, and Bjorn Canback. "ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences." *Nucleic acids research* 32.1 (2004): 11-16.

Lagesen, Karin, et al. "RNAmmer: consistent and rapid annotation of ribosomal RNA genes." *Nucleic acids research* 35.9 (2007): 3100-3108.

Nawrocki, Eric P., Diana L. Kolbe, and Sean R. Eddy. "Infernal 1.0: inference of RNA alignments." *Bioinformatics* 25.10 (2009): 1335-1337.

Jensen, Lars Juhl, et al. "eggNOG: automated construction and annotation of orthologous groups of genes." *Nucleic acids research* 36.suppl_1 (2007): D250-D254.

Arango-Argoty, Gustavo, et al. "DeepARG: A deep learning approach for predicting antibiotic resistance genes from metagenomic data." *Microbiome* 6.1 (2018): 23.

Käll, Lukas, Anders Krogh, and Erik LL Sonnhammer. "Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server." *Nucleic acids research* 35.suppl_2 (2007): W429-W432.

Petersen, Thomas Nordahl, et al. "SignalP 4.0: discriminating signal peptides from transmembrane regions." *Nature methods* 8.10 (2011): 785.

Seemann, Torsten. "Prokka: rapid prokaryotic genome annotation." *Bioinformatics* 30.14 (2014): 2068-2069.

Roosaare, Märt, et al. "StrainSeeker: fast identification of bacterial strains from raw sequencing reads using user-provided guide trees." *PeerJ* 5 (2017): e3353.

Bushnell, Brian. BBMap: a fast, accurate, splice-aware aligner. No. LBNL-7065E. Ernest Orlando Lawrence Berkeley National Laboratory, Berkeley, CA (US), 2014.

Thank you

Questions?