www.bhsai.org

AGeS: A Software System for Annotation and Analysis of Genome Sequences

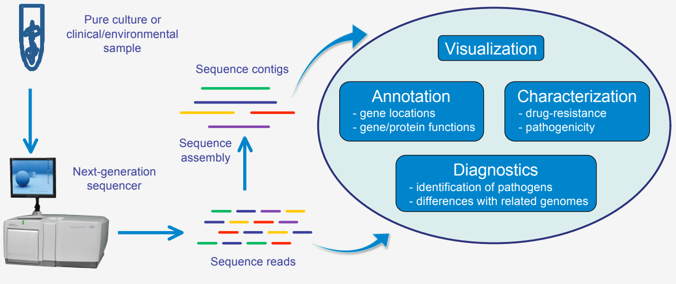
Kamal Kumar, Valmik Desai, Li Cheng, Maxim Khitrov, Deepak Grover, Ravi Vijaya Satya, Chenggang Yu, Nela Zavaljevski, and Jaques Reifman*

DoD Biotechnology HPC Software Applications Institute, Telemedicine and Advanced Technology Research Center,
U.S. Army Medical Research and Materiel Command, Ft. Detrick, MD, USA

*jaques.reifman@us.army.mil (301) 619-7915

Problem

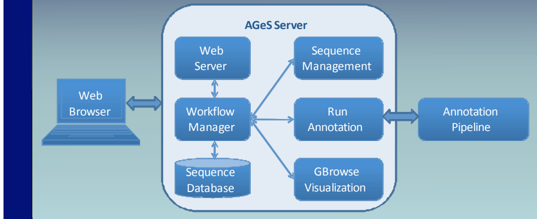
Analysis of next-generation sequencing data



Salient Features

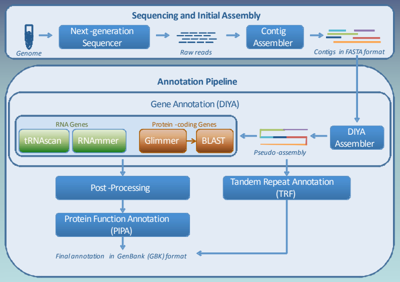
- Fully automated annotation of completed and draft bacterial genomes by combining the configurable annotation framework DIYA [1] with the protein function annotation pipeline PIPA [2]
- Compliant with Minimum Information about a Genome Sequence [3] standard for genomic sequence information and Gene Ontology [4] for protein function annotations
- Repeat identification based on Tandem Repeats Finder (TRF) [5]
- User-friendly visualization based on the familiar open source genome browser GBrowse [6] with option to download annotated genomes in the GenBank format
- High-throughput annotation accomplished through efficient utilization of high-performance computing

AGeS Architecture



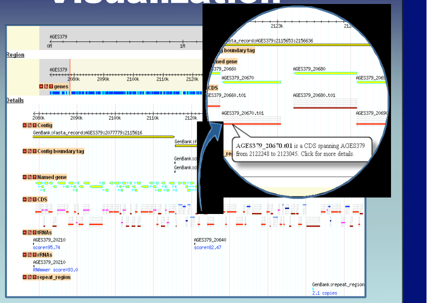
The **Web server** hosts the AGeS Web application. The **workflow manager** handles **sequence management**, **annotation pipeline** and **GBrowse visualization**. The **sequence database** stores all sequence and job-related data.

Annotation Pipeline



Given assembled contigs, the pipeline starts with **DIYA** gene annotation, followed by **post-processing**, **tandem repeat annotation**, and **protein function annotation** with PIPA.

Visualization



The annotation of an 86.5 Kbp region of *S. hominis* SK119 genome, showing the locations of various features. The inset shows the zoomed-in view of a 2.2 Kbp.

Validation

Genomes selected for comparison

Genome	Previous annotation source	Genome status	Size (No. of contigs)
Staphylococcus hominis SK119	J. Craig Venter Institute (JCVI)	Draft	2.2 Mbp (37 contigs)
Staphylococcus aureus subsp. *aureus* TCH60	Baylor College of Medicine (BCM)	Draft	2.8 Mbp (68 contigs)
Yersinia pestis CO92	Sanger Institute	Complete	4.6 Mbp

Summary of genomic features predicted by AGeS and the other three annotation methods

	S. hominis SK119		*S. aureus* subsp. *aureus* TCH60		*Y. pestis* CO92	
Annotation Feature	AGeS	JCVI	AGeS	BCM	AGeS	Sanger
Genes	2229	2244	2652	2805	4336	4103
CDS	2172	2182	2591	2738	4249	3885
rRNA	4	4	4	4	19	19
tRNA	53	52	57	57	68	70
Tandem Repeats	60	NA*	123	NA*	780	NA*

*Annotation was not available for this feature from the source

Detailed comparison of gene overlaps for the three genomes analyzed

	S. hominis SK119	*S. aureus* subsp. *aureus* TCH60	*Y. pestis* CO92
Category	No. of genes (percentage)	No. of genes (percentage)	No. of genes (percentage)
Identical start and end	1753 (78.7%)	2037 (76.8%)	2639 (60.9%)
Identical start only	252 (11.3%)	286 (10.8%)	634 (14.6%)
Identical end only	210 (9.4%)	283 (10.7%)	655 (15.1%)
Overlap	10 (0.4%)	20 (0.7%)	201 (4.6%)
Unique to AGeS	4 (0.2%)	26 (1.0%)	207 (4.8%)

AGeS did not annotate 1% of *S. hominis* SK119 genes annotated by JCVI, 5.8% of *S. aureus* subsp. *aureus* TCH60 genes annotated by BCM, and 5.6% of *Y. pestis* CO92 genes annotated by the Sanger Institute.

Conclusions and Future Work

- **AGeS is a fully integrated, user-friendly HPC system that:**
  - **Provides a Web-based interface to store and retrieve sequence data**
  - **Annotates genomic sequences**
  - **Assigns functions to predicted protein-coding regions**
  - **Provides a visualization of the annotation using GBrowse**
  - **Currently compatible with bacterial genomes**
- **Future work**
  - **Annotation of viral genomes, clinical samples and metagenomic samples**
  - **Addition of tools for diagnostics, characterization, and comparative genomics**

Acknowledgments

This work was partially sponsored by the U.S. Department of Defense High Performance Computing Modernization Program office, under the High Performance Computing Software Applications Institutes Initiative.

Disclaimer

The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the U.S. Army or the U.S. Department of Defense. This poster has been approved for public release with unlimited distribution.

References

1. Stewart AC *et al.*, *Bioinformatics*, 2009, 25:7.
2. Yu C *et al.*, *BMC Bioinformatics*, 2008, 9:52.
3. Field D *et al.*, *Nat Biotechnol*, 2008, 26:5
4. Ashburner M *et al.*, *Nat Genet*, 2000, 25:1.
5. Benson G, *Nucleic Acids Res*, 1999, 27:2.
6. Donlin MJ, *Curr Protoc Bioinformatics*, 2009 9:9.