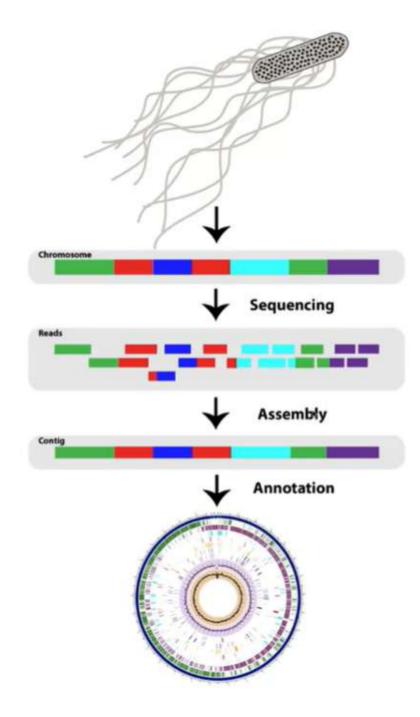
# Genome Annotation

What happens when you annotate a genome?

#### **PATRIC Annotation Process**

- Follows annotation
- Calls genes
- Assigns protein families
- Map to pathways





#### What's happening during the annotation job?

#### The RAST tool kit is customized for PATRIC





#### OPEN

SUBJECT AREAS: COMPARATIVE GENOMICS BIOINFORMATICS

Received 12 November 2014

Accepted 2 January 2015 Published 10 February 2015

Correspondence and requests for materials should be addressed to J.J.D. (jimdavis@uchicago.edu)

# RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes

Thomas Brettin<sup>1,2</sup>, James J. Davis<sup>1,3</sup>, Terry Disz<sup>3</sup>, Robert A. Edwards<sup>4,3</sup>, Svetlana Gerdes<sup>1,3</sup>, Gary J. Olsen<sup>6</sup>, Robert Olson<sup>2,4</sup>, Ross Overbeek<sup>1,3</sup>, Bruce Parrello<sup>1,3</sup>, Gordon D. Pusch<sup>1,3</sup>, Maulik Shukla<sup>7</sup>, James A. Thomason III<sup>8</sup>, Rick Stevens<sup>1,2,8</sup>, Veronika Vonstein<sup>1,3</sup>, Alice R. Wattam<sup>7</sup> & Fangfang Xia<sup>2,4</sup>

<sup>1</sup>Computing, Environment and Life Sciences, Argonne National Loboratory, Argonne II., 60439, USA, <sup>2</sup>Computation Institute, University of Chicago, Chicago, Illinois, 60637, USA, <sup>3</sup>Fellowship for Interpretation of Genomes, Burr Ridge, II., 60527, USA, <sup>4</sup>Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, II., 60439, USA, <sup>5</sup>Department of Computer Science, San Diego State University, San Diego, California, 92182, USA, <sup>6</sup>Department of Microbiology, University of Illinois at Urbano-Champaigri, Urbana, II., 61801, USA, <sup>7</sup>Virginia Bioinformatics Institute, Virginia Tech University, Blacksburg, VA, 24060, USA, <sup>6</sup>USDA-ARS Laboratory of Cold Spring Harbor Laboratory, Cold Spring Harbor NY, 11724, USA, <sup>9</sup>Department of Computer Science, University of Chicago, Chicago, Illinois, 60637, USA.

The RAST (Rapid Annotation using Subsystem Technology) annotation engine was built in 2008 to annotate bacterial and archaeal genomes. It works by offering a standard software pipeline for identifying genomic features (i.e. protein encoding comes and PNA) and annotating their functions. Recently, in order

### First steps in the annotation process

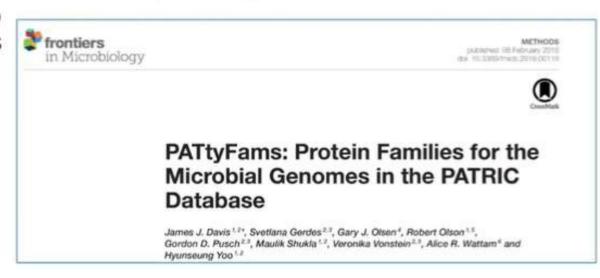
- Calling rRNAs (16S, 23S, 5S)
- Calling tRNAs with tRNAscanSE
  - (Lowe & Eddy 1997)
- Searching for repeat regions
- Finding special proteins
  - Selenoproteins
  - Pyrrolysylproteins
- Calling CRISPRs
  - clustered regularly interspaced short palindromic repeats

#### Additional steps

- Annotates matches to:
  - NCBI
  - Comprehensive Antibiotic Resistance Database CARD (McArthur et al. 2013)
  - Virulence Factor Database VFDB (Chen et al. 2012)
  - Victors (Xiang et al. 2007)
  - PATRIC virulence factors (Mao et al. 2014)
  - DrugBank (Law et al. 2014)
  - Therapeutic Targets Database TTD (Qin et al. 2014)
- Assigns proteins to families
- Finds closest neighbors

#### **Protein Families**

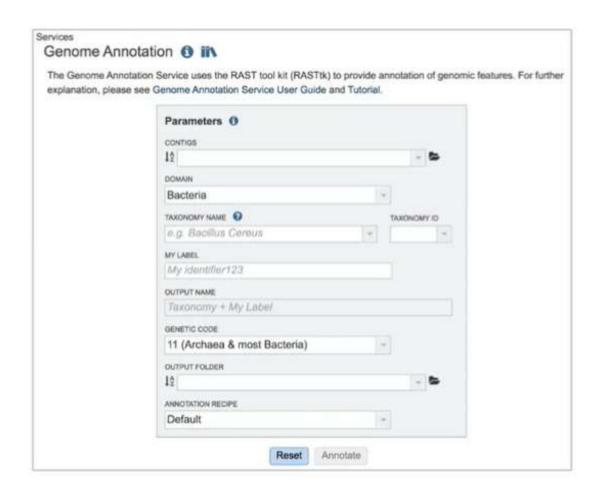
- PATtyFams two flavors of protein families come with each annotation
  - Genus-specific protein families (PLfams)
  - · "Global" protein families (PGfams) cross the genus boundary
    - Used in Codon Tree service
- Steps for assigning protein families include
  - · Computation of local protein families for each genus
  - Merger of protein families across genera in order to provide global families
  - Projection of global family membership to genera with very few sequenced genomes





#### Submitting the annotation job

- Annotation service accepts contigs only
- · Contigs come as a fasta file

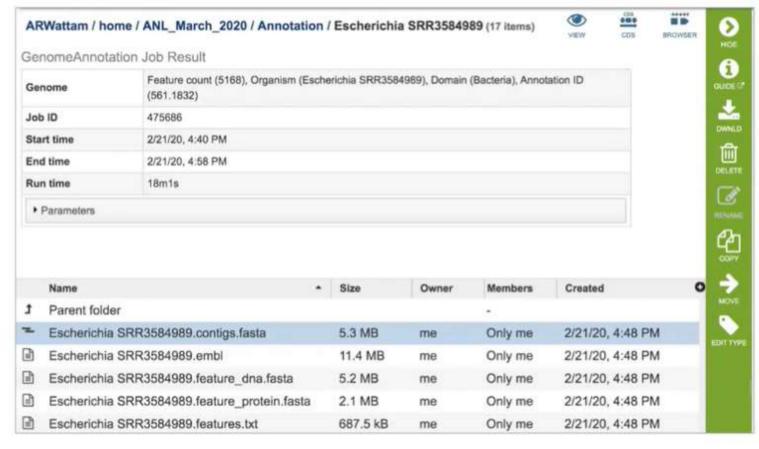


# Files returned by the annotation job

AF	RWattam / home / ANL_March_2020 / Annotation / Escherichia SF	RR3584989 (17 fiams)			-		ATTE	-
Ger	nomeAnnotation Job Result							
Ge	nome Feature count (5168), Organism (Escherichia SRR	0584989), Domain (Bacteria), Annoti	ation ID (561.16	132)				
Jol	b ID 475666							
Sta	ort time 2/21/20, 4:40 PM							
End time 2/21/20, 4:58 PM Run time 18/21/2								
	Parameters							
	Name		Size	Owner	Members	Cres	ated	
,	Parent folder		-	- Guitai		-		
-	Escherichia SRR3584989.contigs.fasta		5.3 MB	me	Only me	2/2	1/20. 4	48 PM
3	Escherichia SRR3584989.embl		11.4 MB	me	Only me	2/21/20, 4:48 PM		
(0)	Escherichia SRR3584989 feature, dna fasta		5.2 MB	me	Only me	2/21/20, 4:48 PM		
3	Escherichia SRR3584989.feature_protein.fasta		2.1 MB	me	Only me		2/21/20, 4:48 PM	
3	Escherichia SRR3584989 features bit		687.5 kB	me	Only me			48 PM
0	Escherichia SRR3584989.gb		11.0 MB	me	Only me			48 PM
3	Escherichia SRR3584989.genome		18.4 MB	me	Only me			48 PM
	Escherichia SRR3584989.gff		714.3 kB	me	Only me			48 PM
	Escherichia SRR3S84989.merged.gb		11.0 MB	me	Only me	2/2	1/20, 4	48 PM
	Escherichia SRR3584989.tar.gz		2.9 MB	me	Only me	2/21/20, 4:48 PM		
3	Escherichia SRR3584989.txt		7.0 MB	me	Only me	2/2	1/20, 4	48 PM
(2)	Escherichia SRR3584989.xls		7.8 MB	me	Only me	2/2	1/20, 4	48 PM
3	GenomeReport.html		54.1 kB	me	Only me	2/2	1/20, 4	48 PM
-	alignments			me	Only me	2/21	1/20, 4	48 PM
(a)	genome_quality_details.txt		52.8 kB	me	Only me			48 PM
-	load files			me	Only me			48 PM
(A)	quality.json		526.4 kB	me	Only me			58 PM



#### Downloading data from private genomes



Select row and then click download in vertical green bar

## Quality scores on annotated genomes

Genome ID	561.1832
Genome Name	Escherichia SRR3584989
Reference Genomes	511145.180
Coarse consistency (%)	97.0
Fine consistency (%)	96.1
Completeness (%)	99.5
Contamination (%)	2.1
Evaluation Group	Enterobacteriaceae
Contig count	100
DNA size (bp)	5183531
Contigs N50 (bp)	170302
Contigs L50	9
Overpresent Roles	49
Underpresent Roles	6
Predicted Roles	2039
Completeness Roles	1306
Total Distinct Roles	4448
Protein-Encoding Genes with Functional Assignment	3669
Protein-Encoding Genes without Functional Assignment	1425
% Protein-Encoding Feature Coverage	98.27
% Features that are Hypothetical	11.94
% Features that are in Local Protein Families	96.9



#### Quality scores on annotated genomes

Genome ID	561.1832
Genome Name	Escherichia SRR3584989
Reference Genomes	511145.180
Coarse consistency (%)	97.0
Fine consistency (%)	96.1
Completeness (%)	99.5
Contamination (%)	2.1
Evaluation Group	Enterobacteriaceae
Contig count	100
DNA size (bp)	5183531
Contigs N50 (bp)	170302
Contigs L50	9
Overpresent Roles	49
Underpresent Roles	6
Predicted Roles	2039
Completeness Roles	1306
Total Distinct Roles	4448
Protein-Encoding Genes with Functional Assignment	3669
Protein-Encoding Genes without Functional Assignment	1425
% Protein-Encoding Feature Coverage	98.27
% Features that are Hypothetical	11.94
% Features that are in Local Protein Families	96.9

A high coarse consistency indicates either a single genome or a cluster of very close strains.

A fine consistency of 87% is the minimum required for a genome to be considered "good".

A completeness of 80% is the minimum required for a genome to be considered "good".

A contamination of 10% is the maximum allowed for a genome to be considered "good".