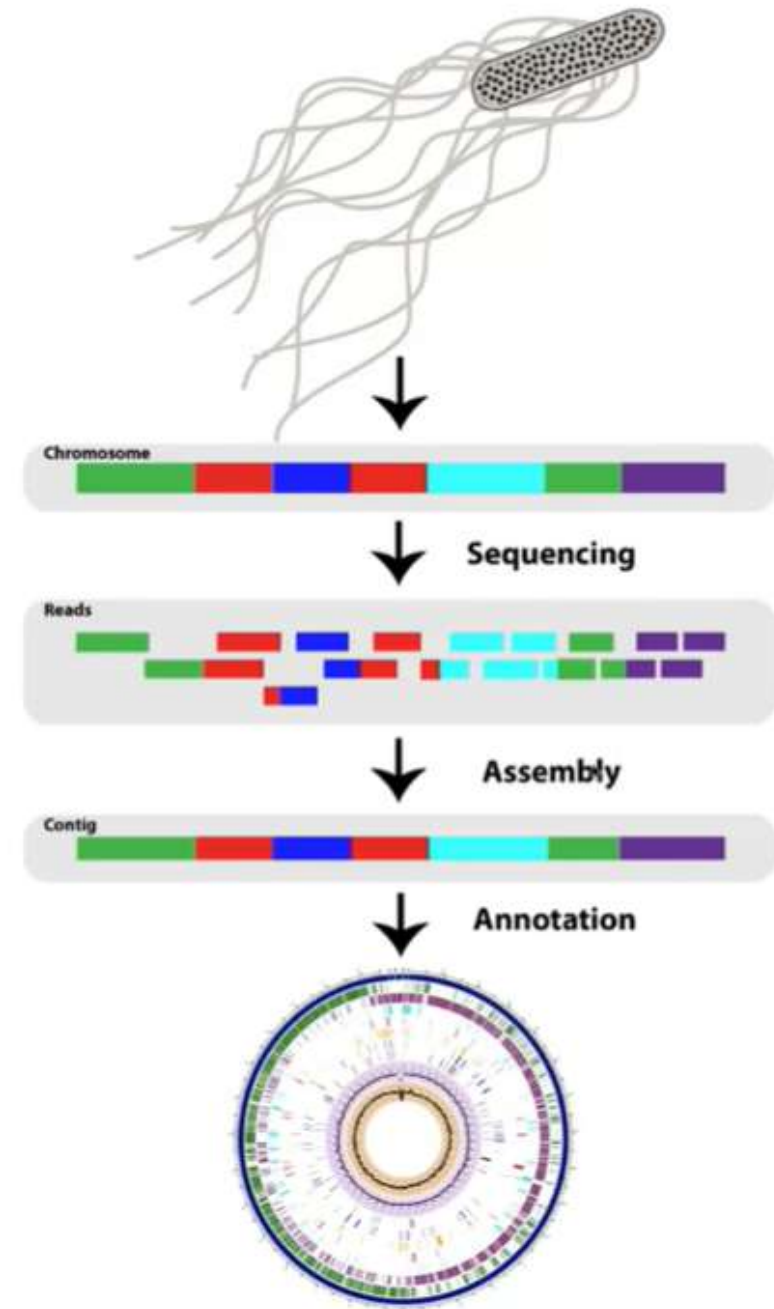


Genome Annotation

What happens when you annotate a genome?

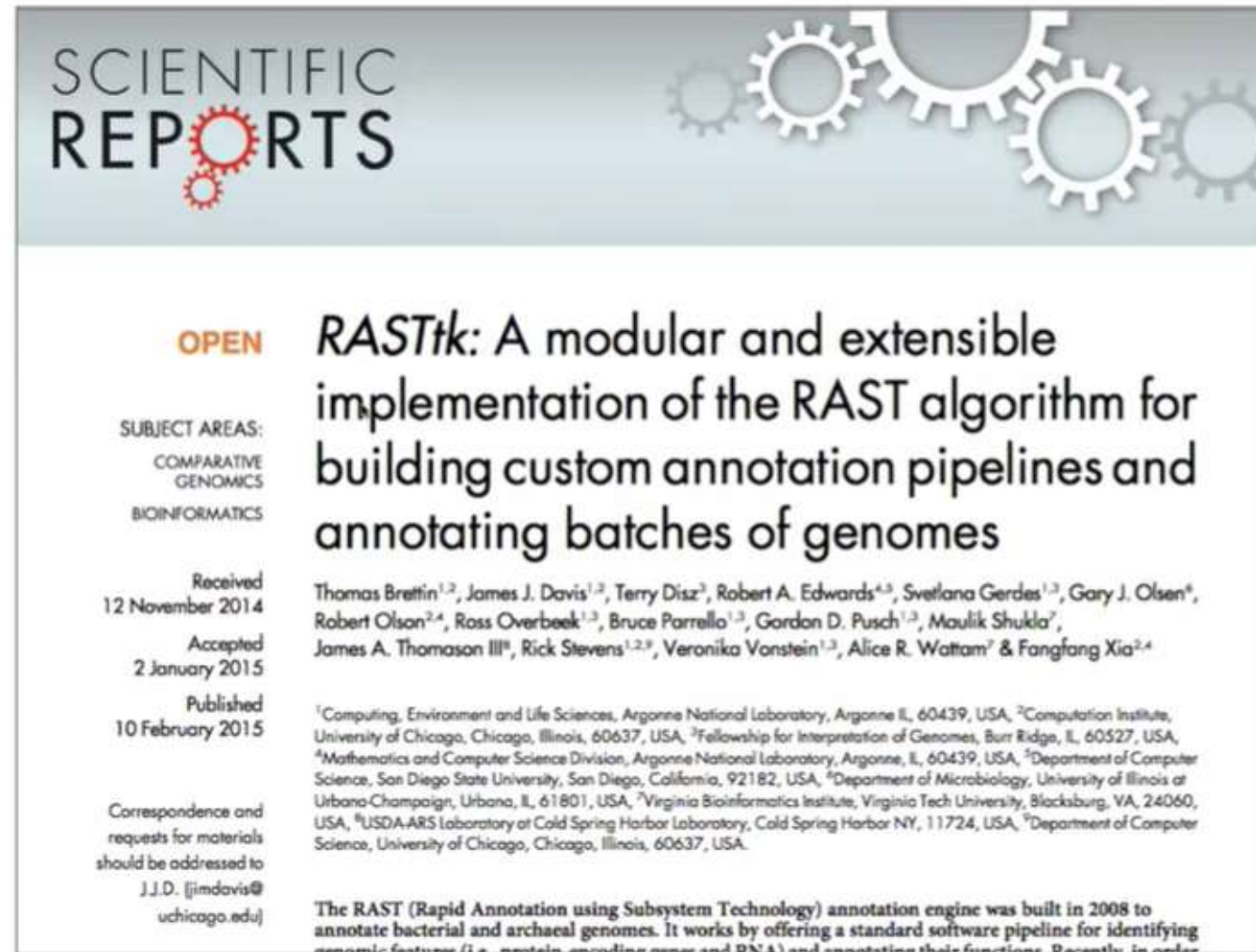
PATRIC Annotation Process

- 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



First steps in the annotation process

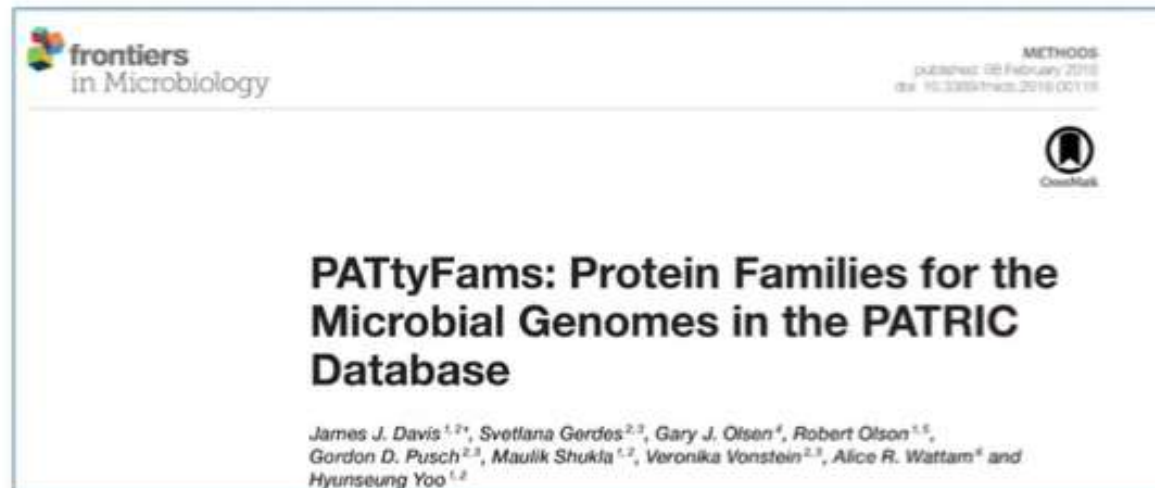
- Calling rRNAs (16S, 23S, 5S)
- Calling tRNAs with tRNAscanSE
 - (Lowe & Eddy 1997)
- Searching for repeat regions
- Finding special proteins
 - Selenoproteins
 - Pyrrolysylproteins
- Calling CRISPRs
 - clustered regularly interspaced short palindromic repeats

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

Services


Genome Annotation

The Genome Annotation Service uses the RAST tool kit (RASTtk) to provide annotation of genomic features. For further explanation, please see [Genome Annotation Service User Guide and Tutorial](#).

Parameters

CONTIGS



DOMAIN

TAXONOMY NAME  TAXONOMY ID

MY LABEL

OUTPUT NAME

GENETIC CODE

OUTPUT FOLDER


ANNOTATION RECIPE

Files returned by the annotation job

ARWattam / home / ANL_March_2020 / Annotation / Escherichia SRR3584989 (17 items)

VIEW

LOG

SHARE

GenomeAnnotation Job Result

Genome

Feature count (5168), Organism (Escherichia SRR3584989), Domain (Bacteria), Annotation ID (5611832)

Job ID

475666

Start time

2/21/20, 4:40 PM

End time

2/21/20, 4:58 PM

Run time

18m1s

Parameters

Name	Size	Owner	Members	Created
Parent folder			-	
Escherichia SRR3584989.contigs.fasta	5.3 MB	me	Only me	2/21/20, 4:48 PM
Escherichia SRR3584989.embl	11.4 MB	me	Only me	2/21/20, 4:48 PM
Escherichia SRR3584989.feature_dna.fasta	5.2 MB	me	Only me	2/21/20, 4:48 PM
Escherichia SRR3584989.feature_protein.fasta	2.1 MB	me	Only me	2/21/20, 4:48 PM
Escherichia SRR3584989.features.txt	687.5 kB	me	Only me	2/21/20, 4:48 PM
Escherichia SRR3584989.gb	11.0 MB	me	Only me	2/21/20, 4:48 PM
Escherichia SRR3584989.genome	18.4 MB	me	Only me	2/21/20, 4:48 PM
Escherichia SRR3584989.gff	714.3 kB	me	Only me	2/21/20, 4:48 PM
Escherichia SRR3584989.merged.gb	11.0 MB	me	Only me	2/21/20, 4:48 PM
Escherichia SRR3584989.tar.gz	2.9 MB	me	Only me	2/21/20, 4:48 PM
Escherichia SRR3584989.txt	7.0 MB	me	Only me	2/21/20, 4:48 PM
Escherichia SRR3584989.xls	7.8 MB	me	Only me	2/21/20, 4:48 PM
GenomeReport.html	54.1 kB	me	Only me	2/21/20, 4:48 PM
alignments		me	Only me	2/21/20, 4:48 PM
genome_quality_details.txt	52.8 kB	me	Only me	2/21/20, 4:48 PM
load_files		me	Only me	2/21/20, 4:48 PM
quality.json	526.4 kB	me	Only me	2/21/20, 4:58 PM



Downloading data from private genomes

ARWattam / home / ANL_March_2020 / Annotation / Escherichia SRR3584989 (17 items)

GenomeAnnotation Job Result

Genome	Feature count (5168), Organism (Escherichia SRR3584989), Domain (Bacteria), Annotation ID (561.1832)
Job ID	475686
Start time	2/21/20, 4:40 PM
End time	2/21/20, 4:58 PM
Run time	18m1s

Parameters

Name	Size	Owner	Members	Created
Parent folder				
Escherichia SRR3584989.contigs.fasta	5.3 MB	me	Only me	2/21/20, 4:48 PM
Escherichia SRR3584989.embl	11.4 MB	me	Only me	2/21/20, 4:48 PM
Escherichia SRR3584989.feature_dna.fasta	5.2 MB	me	Only me	2/21/20, 4:48 PM
Escherichia SRR3584989.feature_protein.fasta	2.1 MB	me	Only me	2/21/20, 4:48 PM
Escherichia SRR3584989.features.txt	687.5 kB	me	Only me	2/21/20, 4:48 PM

VIEW CDS BROWSER

HIDE GUIDE DOWNLOAD DELETE RENAME COPY MOVE EDIT TYPE

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".