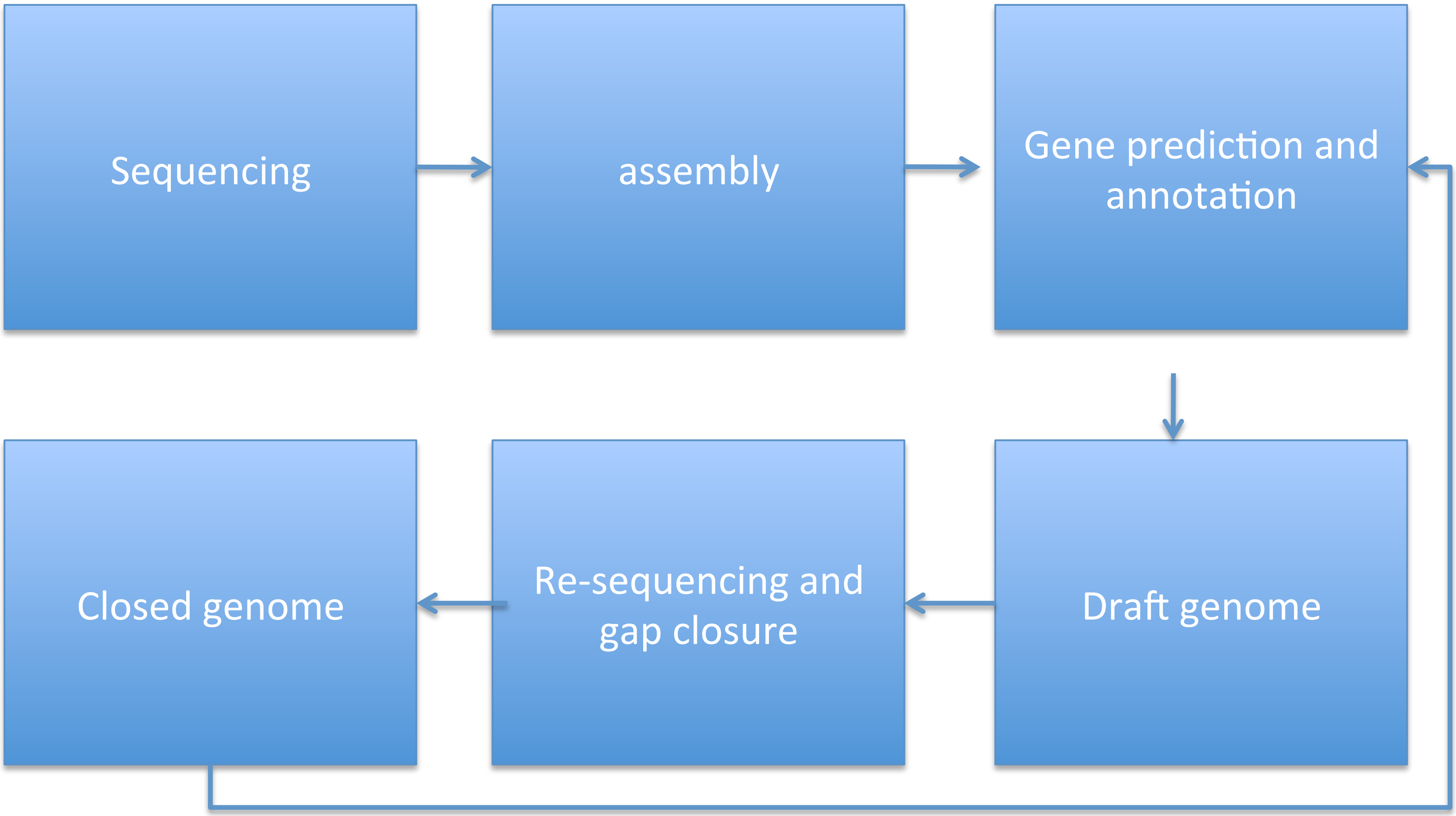
Lecture 7

Annotation

The whole process



Sequencing

assembly

Gene prediction and annotation

Draft genome

Re-sequencing and gap closure

Closed genome

Annotation

Finding genomic features using computation and experimental methods

1. Genes
2. Repeat sequences
3. Transcriptional regions
4. Other

levels of annotation

- Gene prediction
- Function: Predictions of domains
- General Role and interaction: What pathways involved in?

What are genes?

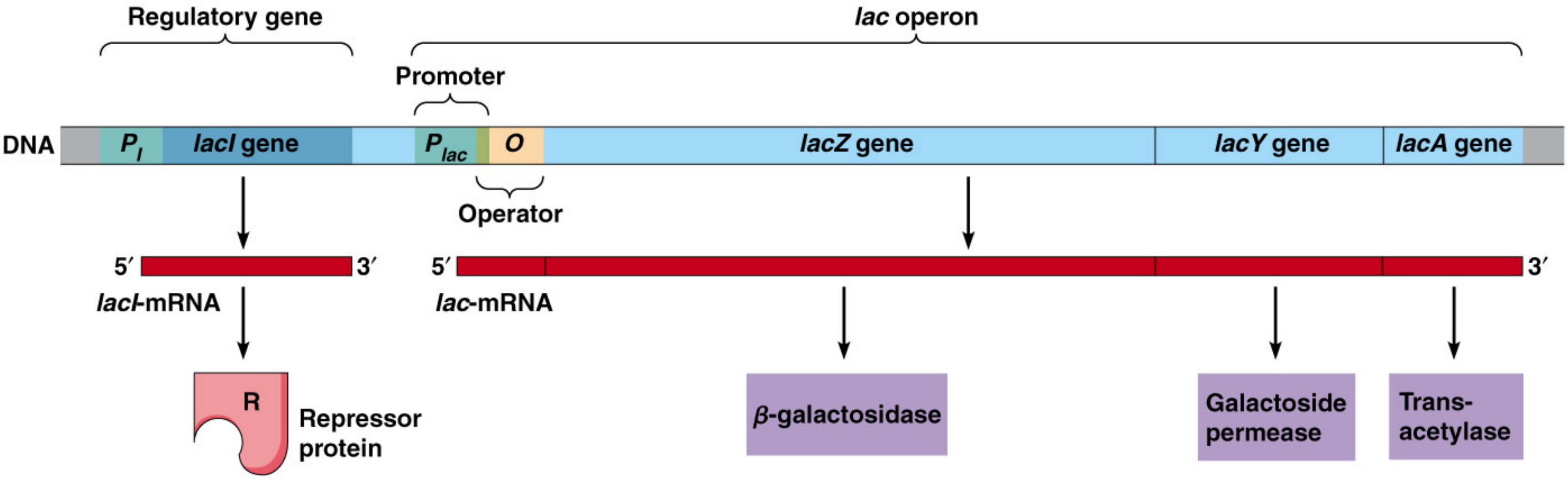
- DNA segments that produce functional products

- Manly proteins!
- RNA molecules!
- RNAi (interfering RNA)
- rRNA (ribosomal RNA)
- tRNA (trasnfer RNA)

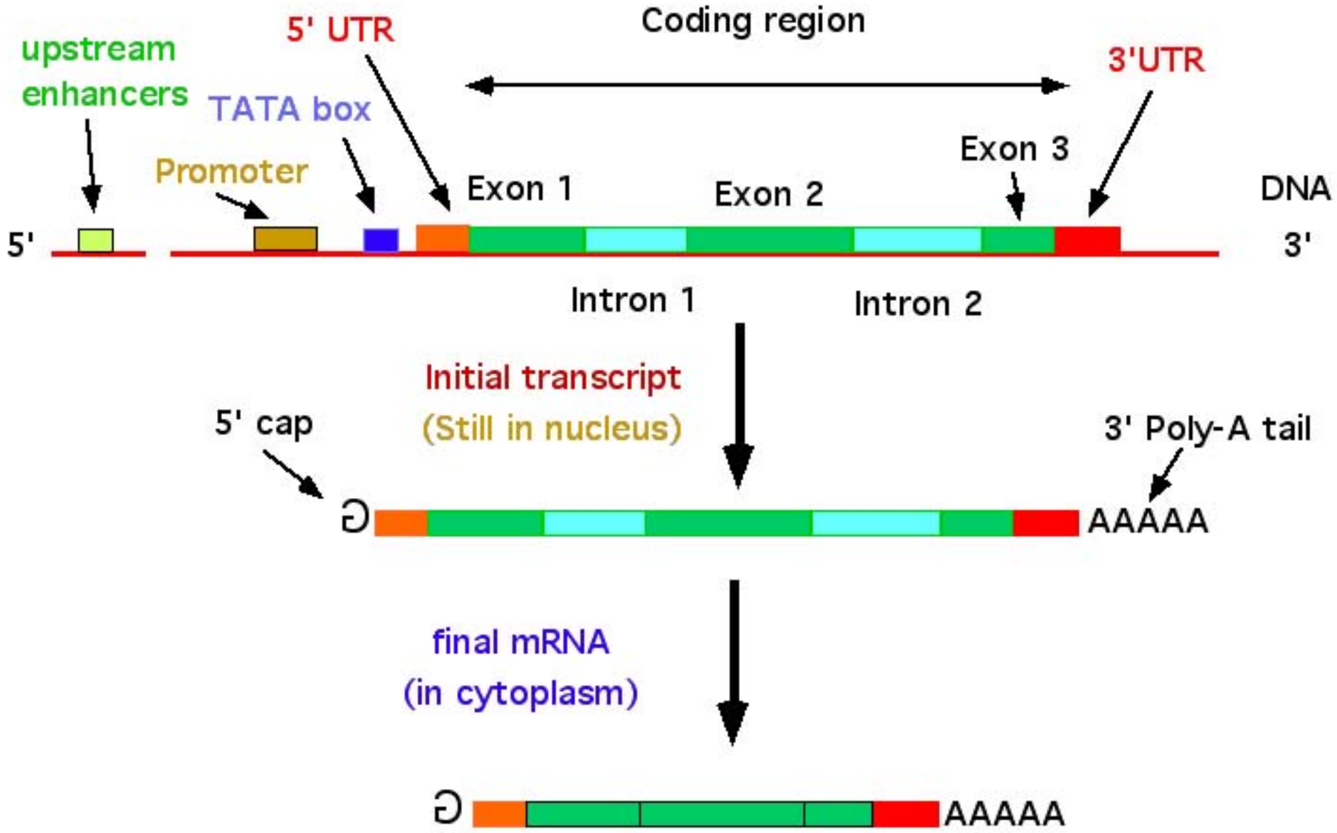
Prokaryotic vs Eukaryotic gene model

Prokaryotic	Eukaryotic
Small genomes, high gene density	Large genomes, many regions without coding
Operons, many genes in one transcript	Operons, monocistronic. Some exceptions
Open reading frames. One OFR per genes. ORFs begin with start (ATG, GTG, TTG), end with stop codon (TAG, TAA, TGA)	Posttranscriptional modification

Prokaryotic gene model



Eukaryotic gene model



Gene identification

1- Homology-based gene prediction

-Similarity search (BLAST!)

-Genome browsers

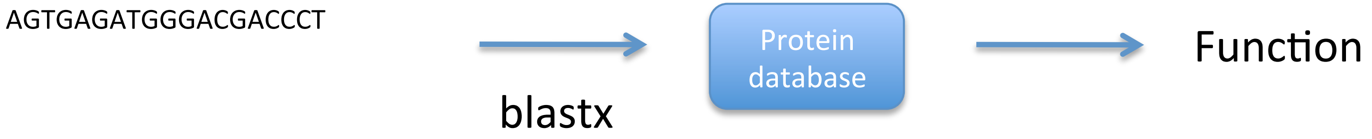
-RNA evidence (ETSs)

2- *Ab initio gene prediction*

Gene predictions programs

Homology-based gene prediction

- Similarity search:



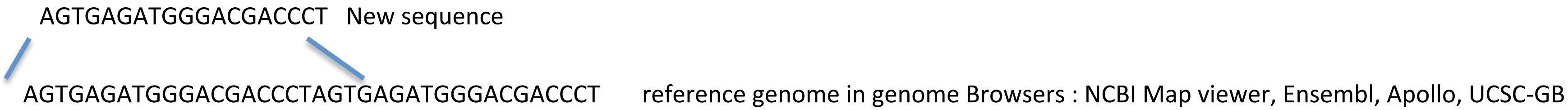
AGTGAGATGGGACGACCCT

blastx

Protein database

Function

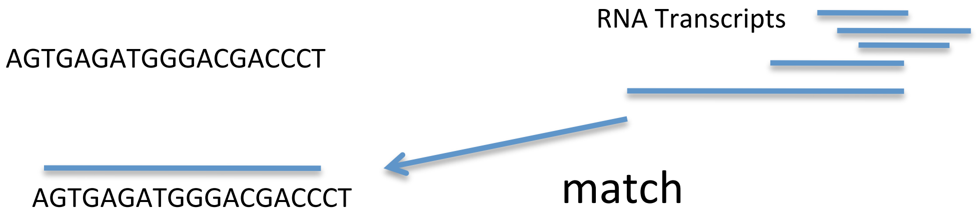
- Genome browsers:



AGTGAGATGGGACGACCCT New sequence

AGTGAGATGGGACGACCCTAGTGAGATGGGACGACCCT reference genome in genome Browsers : NCBI Map viewer, Ensembl, Apollo, UCSC-GB

- RNA evidence: ETS



RNA Transcripts

AGTGAGATGGGACGACCCT

AGTGAGATGGGACGACCCT

match

Ab initio gene prediction

- Prokaryotes

ORF detectors

ATGgtgtgg………………………………..ttggggTGA

Start   stop

- Eukaryotes

Position, extent and direction though promoter and polyA-signal

Promoter seq ………………..Splice-seq…………PolyAAA

Tools

- ORF detectors

Gorf : http://www.ncbi.nih.gov/gorf/gorf.html

- Promoter predictors

ICG TATA-Box predictor

BDGP fruitfly.org/seq_tools/promoter.html

Virtual footprint (for bacteria) http://www.prodoric.de/vfp

Softberry (for bacteria)
http://linux1.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb

- PolyA signal predictors

CSHL: argon.cshl.org/tabaska/polyadq_form.html

- Splice site predictors

BDGP : http://www.fruitfly.org/seq_tools/splice.html

Gene prediction programs

- Rule-based programs

Use specific set of rules to predict the genes(GeneFinder)

- Hidden Markov Model-based programs

Use probabilities of states and transitions to predict (genscan, genomeScan, glimmer)

Annotation tools

- Blast searches (Blast)
- HHM models of specific genes or gene families (PFAM)
- Cellular location (PBSORT, SignalP)
- Biochemical pathwayd/subsystem informtaion (KEGG)

Gene Ontology (GO)
http://www.geneontology.org/

The GO consortium describes gene productys with a standard vocabulary.

Biological process

Cellular component

Molecular function

For example: cytochrome C is described with

Biological process: phosphorylation GO:0016310

Cellular component: mitochondrial inner membrane GO:0005743

Molecular function: oxido-reductase activity GO:0016491

Each of these has a GO term

COG and KOG
Cluster of orthologs group

- http://www.ncbi.nlm.nih.gov/books/NBK21090/

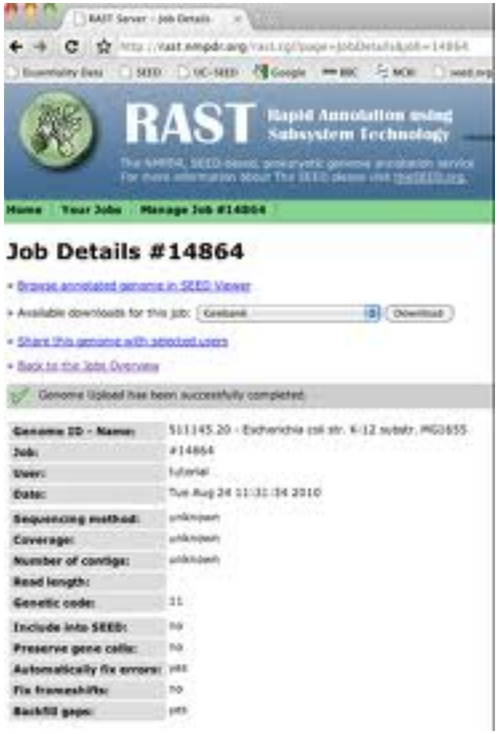
- **Phylogenetic Classification of Proteins from Complete Genomes**

Tools for annotation

Prokka

blast2go

MAKER
Annotate this!



RAST Server - Job Details

http://rast.nmpdr.org/rast.cgi?page=JobDetails&job=14864

RAST Rapid Annotation using Subsystem Technology

The NMPDR, SEED-based, prokaryotic genome annotation service
For more information about The SEED please visit theSEED.org.

Job Details #14864

- Browse annotated genome in SEED Viewer
- Available downloads for this job: Genbank Download
- Share this genome with selected users
- Back to the Jobs Overview

Genome Upload has been successfully completed.

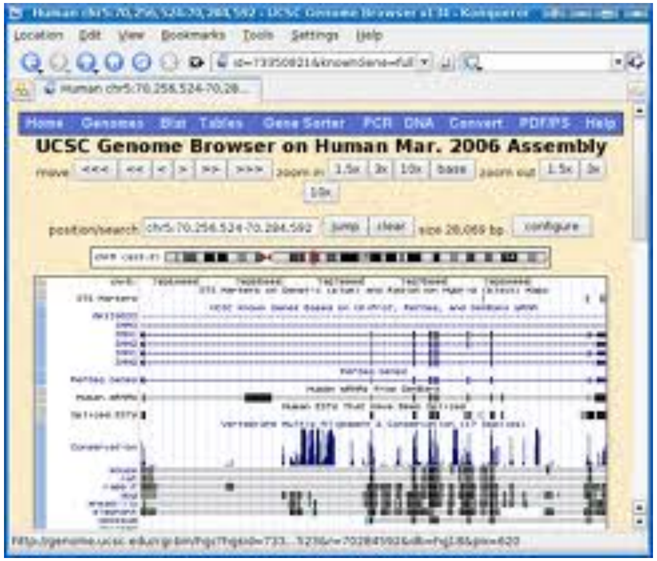
Genome ID - Name:	511145.20 - Escherichia coli str. K-12 substr. MG1655
Job:	#14864
User:	tutorial
Date:	Tue Aug 24 11:31:34 2010
Sequencing method:	unknown
Coverage:	unknown
Number of contigs:	unknown
Read length:	
Genetic code:	11
Include into SEED:	no
Preserve gene calls:	no
Automatically fix errors:	yes
Fix frameshifts:	no
Backfill gaps:	yes

xbase

RAST server

- http://rast.nmpdr.org/
- Rapid annotation using subsystem techology
- For prokaryotes.

Genome browsers, visualization and curation tools


Location Edit View Bookmarks Tools Settings Help
Home Genomes Blat Tables Gene Sorter PCR DNA Convert PDF/PS Help
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
jump clear configure

Human chr5:70,256,524-70,284,592 - UCSC Genome Browser v131 - Konqueror

UCSC Genome Browser on Human Mar. 2006 Assembly

position/search chr5:70,256,524-70,284,592 size 28,069 bp

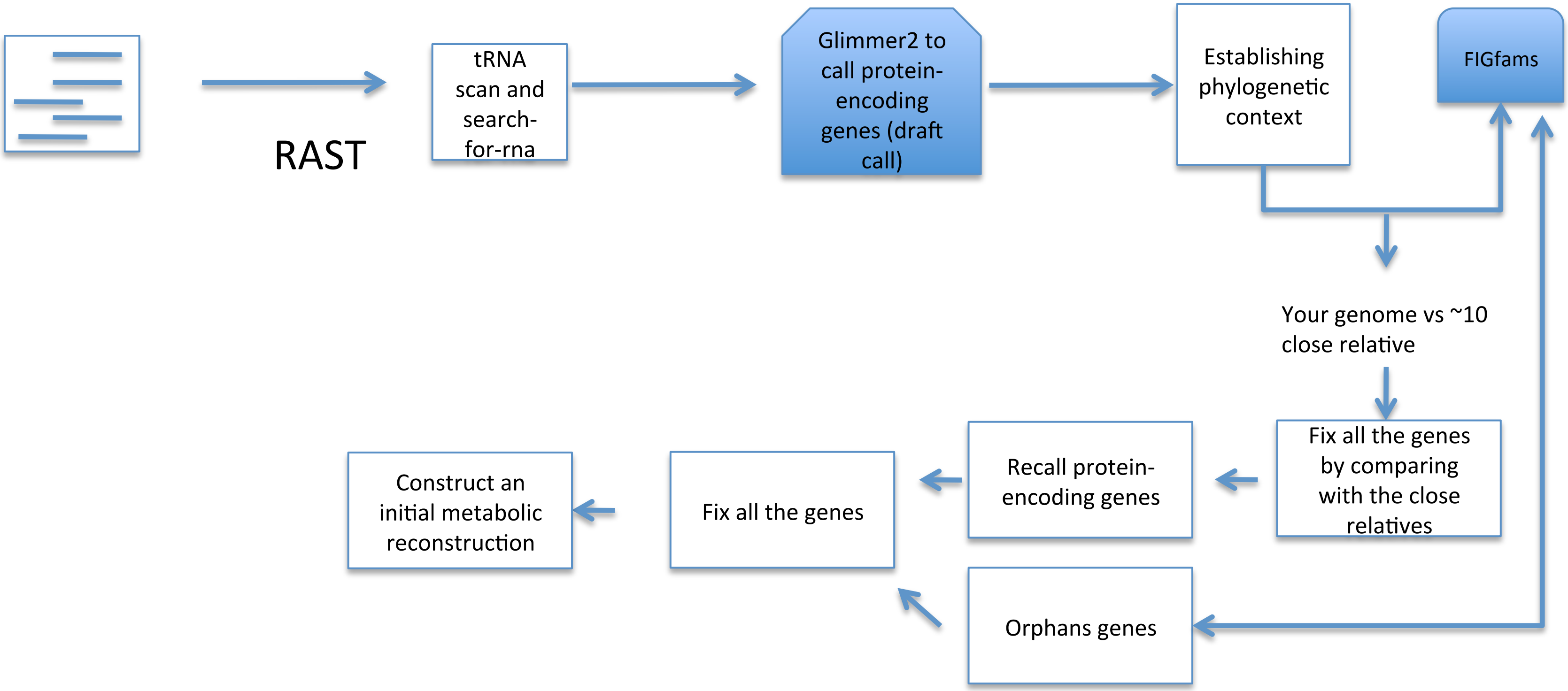
Artemis Release 8

Artemis
Release 8
1. Standard

Copyright 1998 - 2006
Genome Research Limited

The Sanger Institute

The basic steps in annotating a genome using RAST



RAST

tRNA scan and search-for-rna

Glimmer2 to call protein-encoding genes (draft call)

Establishing phylogenetic context

FIGfams

Your genome vs ~10 close relative

Fix all the genes by comparing with the close relatives

Recall protein-encoding genes

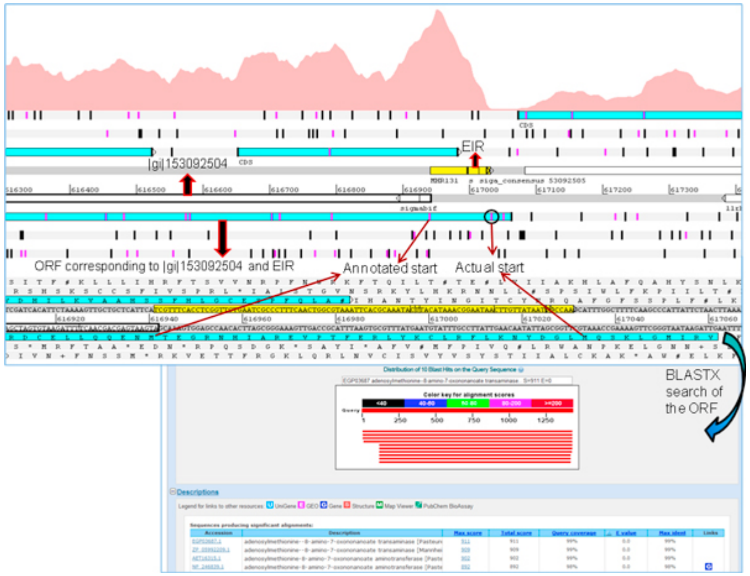
Orphans genes

Fix all the genes

Construct an initial metabolic reconstruction

Artemis genome browser

- https://www.sanger.ac.uk/resources/software/artemis/



Artemis