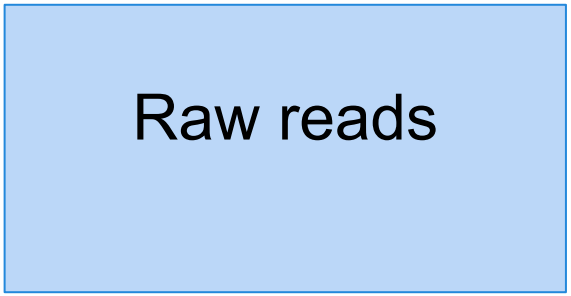
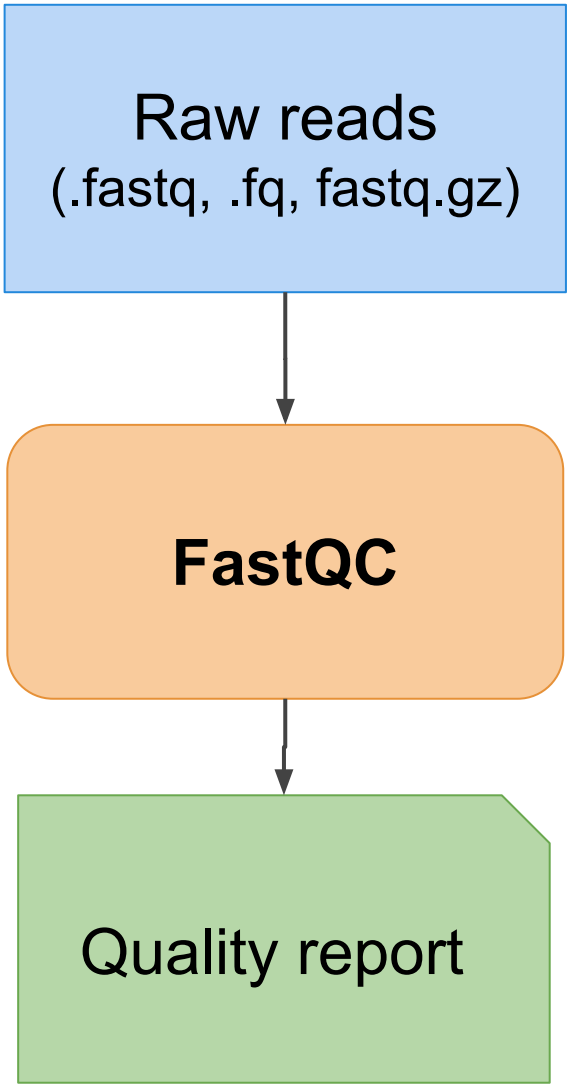
Genome annotation

Center for Algorithmic Biotechnology
SPbU

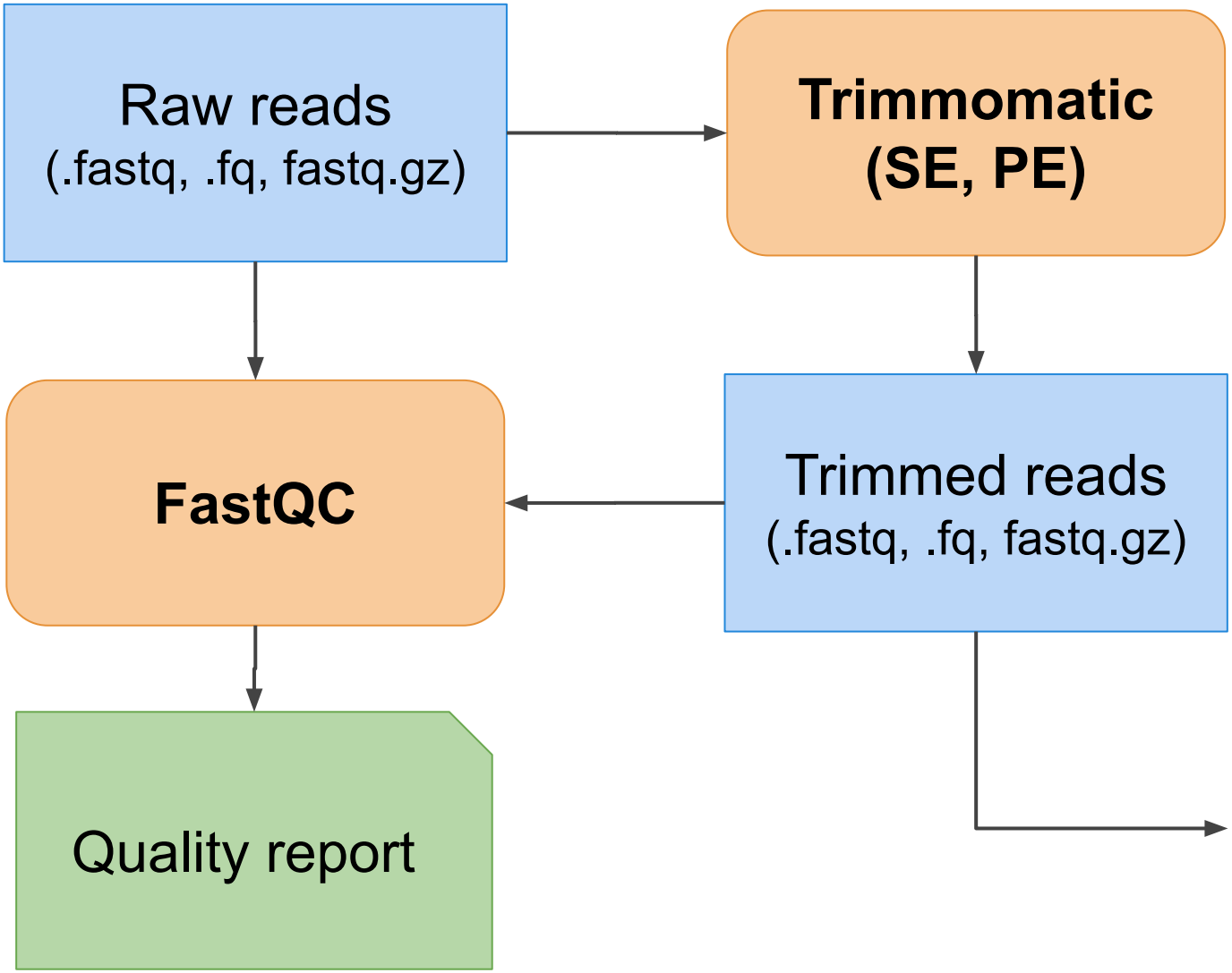
General pipeline



General pipeline



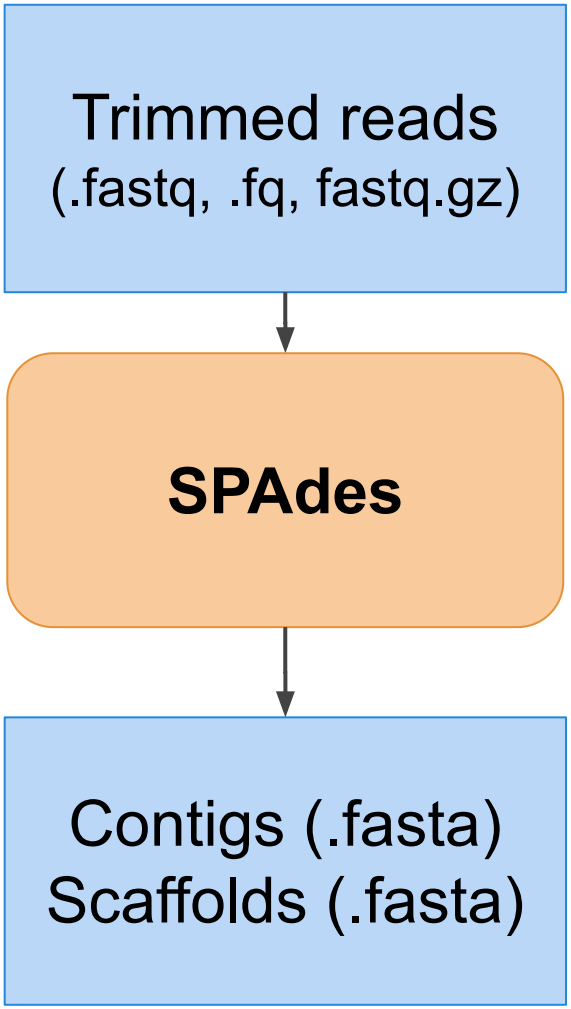
General pipeline



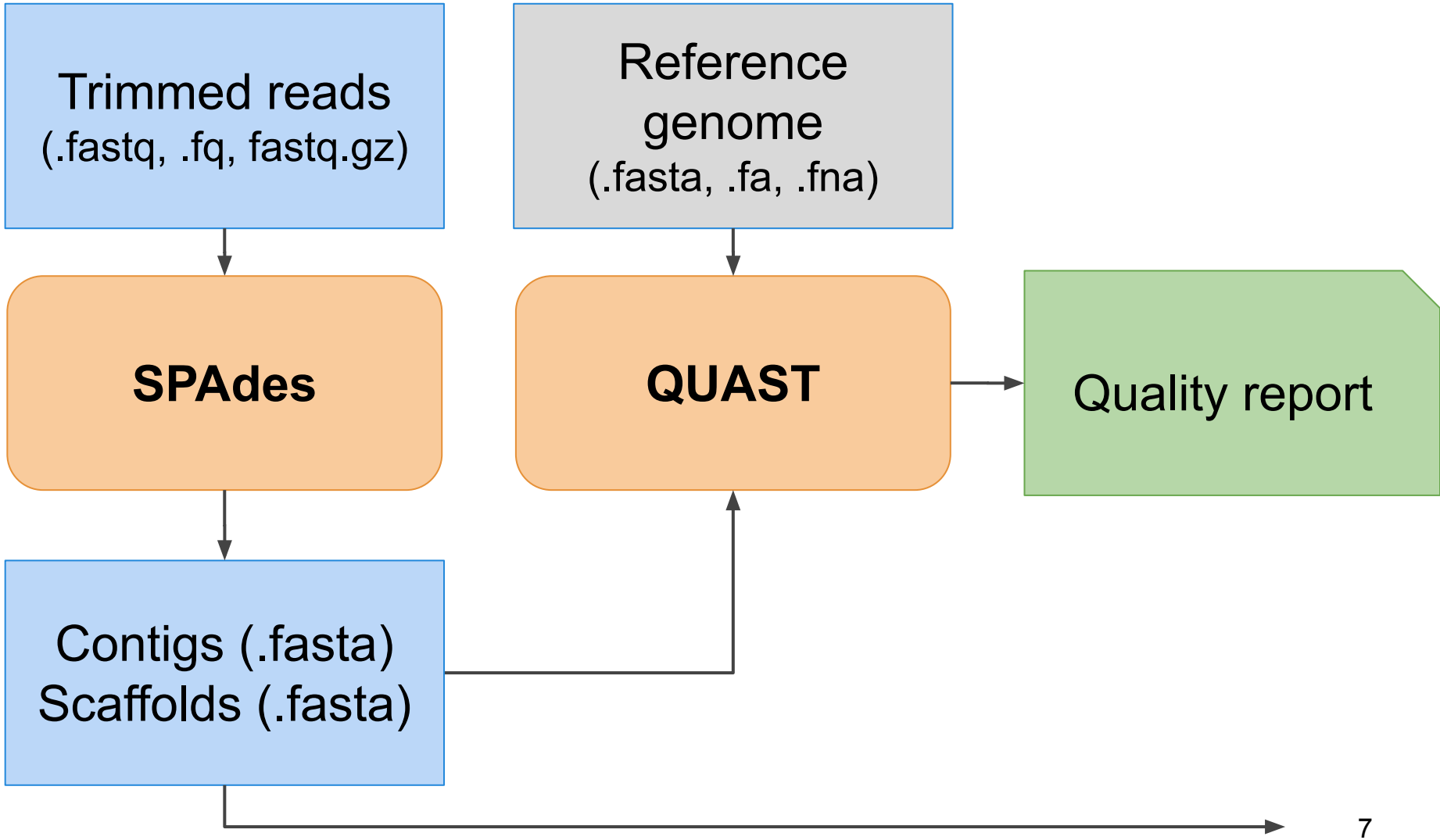
General pipeline

Trimmed reads
(.fastq, .fq, fastq.gz)

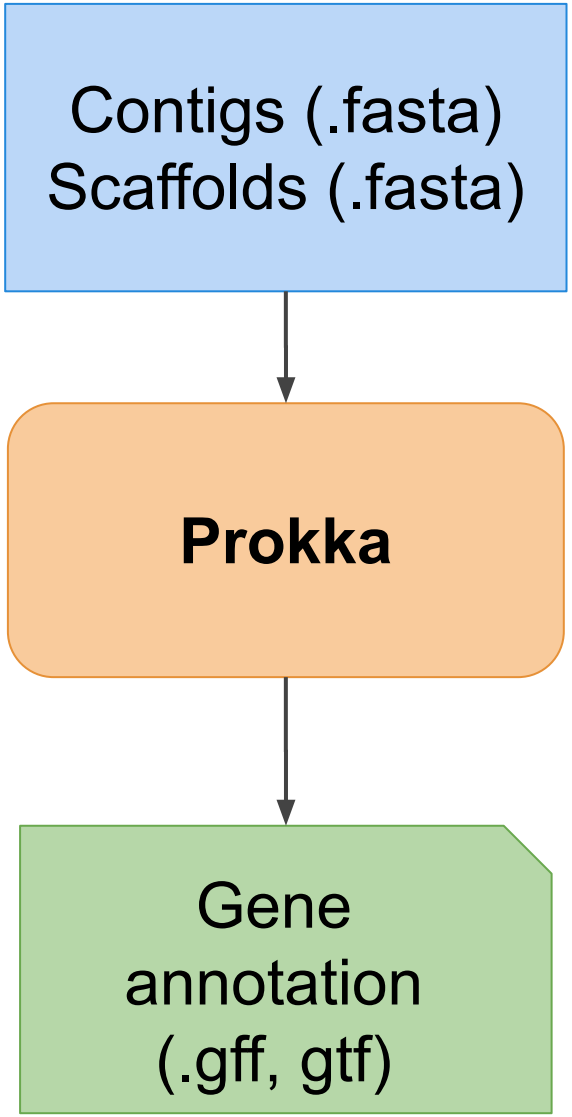
General pipeline



General pipeline

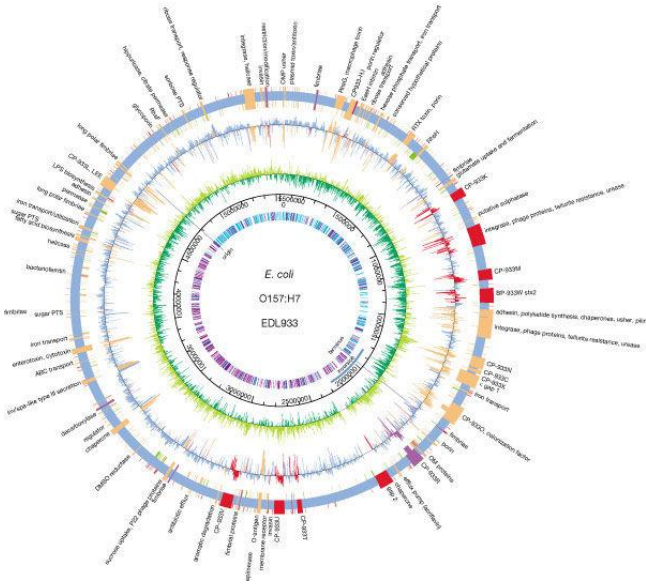


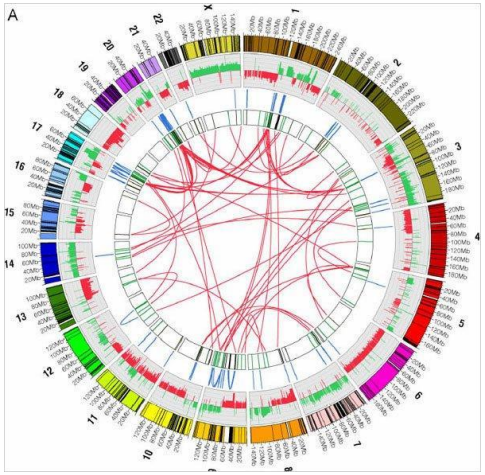
General pipeline



Genome Annotation Questions

- Which genes are present?
- How did they get there (evolution)?
- Are the genes present in more than one copy?
- Which genes are not there that we would expect to be present?



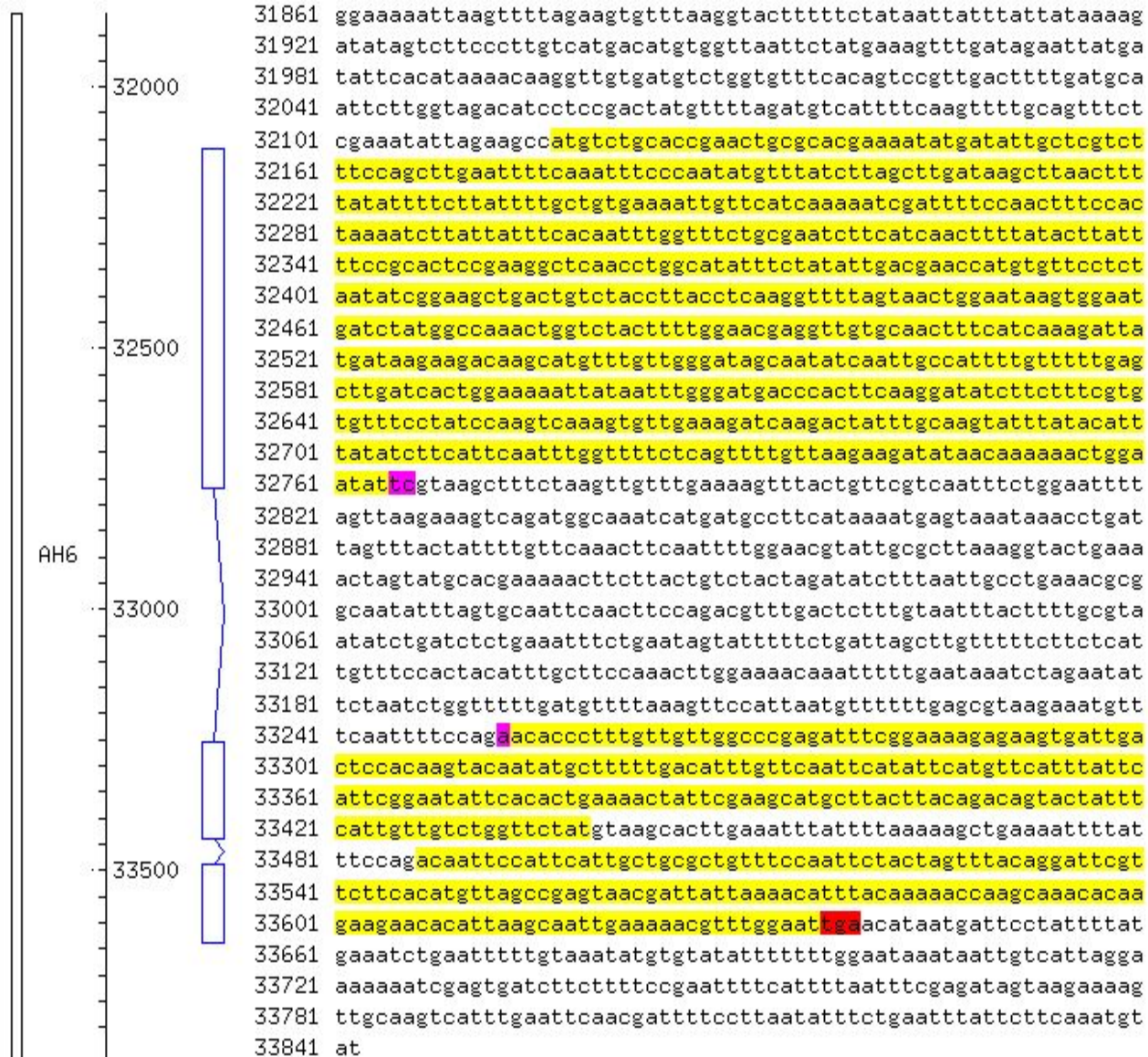


- What is the order are the genes and does this have any significance?
- How similar is the genome of one organism to that of another?

After completing the human genome we faced 3 Gigabytes of this:

Genome sequence does not give you list of all genes

Not immediately apparent where the genes are…



Genomic Features

- **Protein coding genes.**

In long open reading frames
ORFs interrupted by introns in eukaryotes

- **RNA-only genes**

Transfer RNA, ribosomal RNA, ncRNA, other small RNAs

- **Gene control sequences**

Promoters
Regulatory elements

- **Transposable elements, both active and defective**

DNA transposons and retrotransposons

- **Repeated sequences**

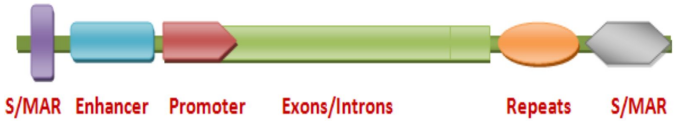
Centromeres and telomeres
Many with unknown (or no) function

- **Unique sequences that have no obvious function**

Genome annotation

STRUCTURAL ANNOTATION

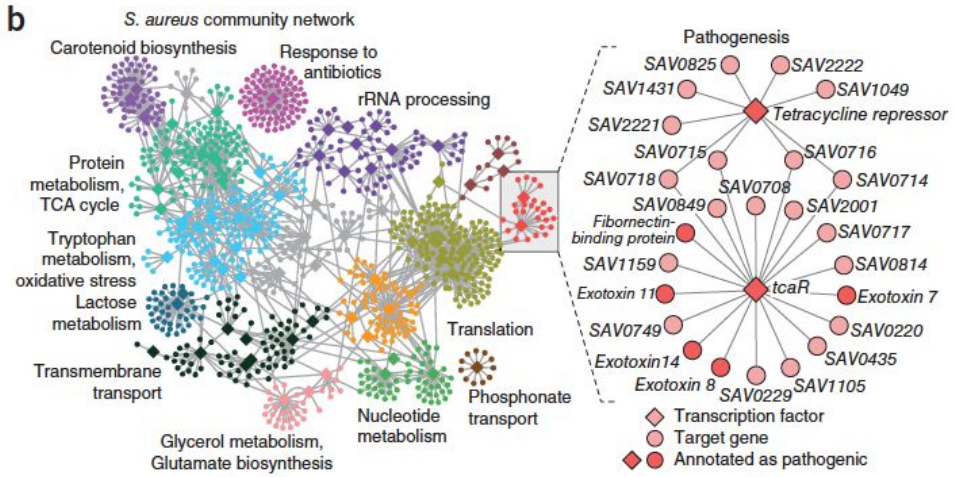
- Open reading frame and their localization
- Exons, introns, UTRs
- Start/Stop
- Location of regulatory motifs
- Splice Sites
- Non coding Regions
- Transposable elements
- tRNA, miRNA, rRNA, ncRNA



FUNCTIONAL ANNOTATION

Gene function prediction: attaching biological information to these elements

- Biochemical function
- Biological function
- Involved regulation and interactions



http://geneontology.org

Structural annotation

- **Open reading frame and their localization**
  ORFfinder, personal scripts

- **Exons, introns, UTRs, Start/Stop, Splice Sites, Non coding Regions**
  from GFF annotation file (gene prediction programs) using personal scripts

- **Location of regulatory motifs**
  PEAKS, MEME, and other …

- **Transposable elements**
  RepeatModeler, RepeatMasker

- **tRNA, miRNA, rRNA, ncRNA**
  tRNA-ScanSE, Arwen, sRNAbench, and other …

Automatic annotation approaches

Similarity based

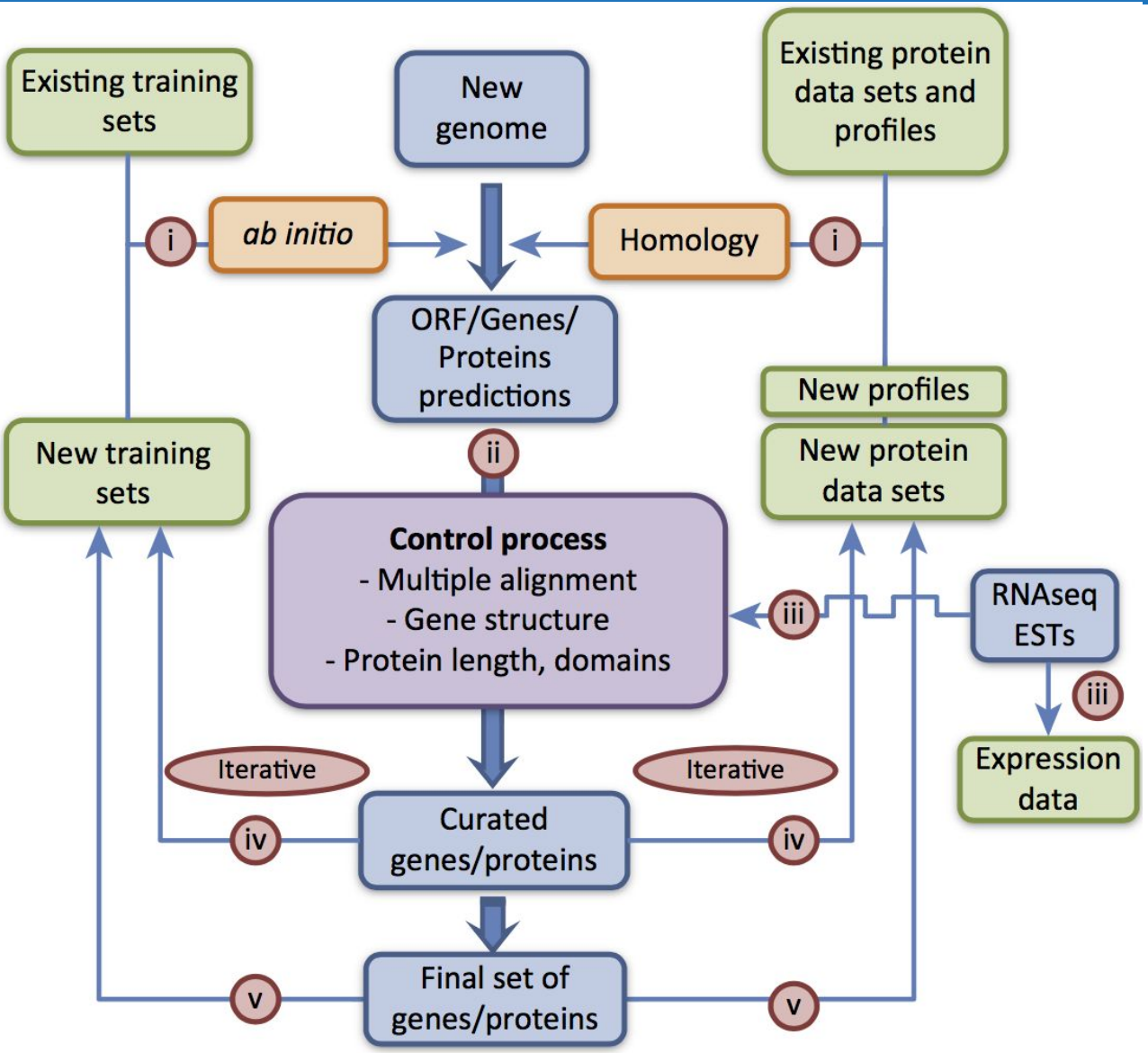
- Alignment of the known protein coding genes to contigs
- Will miss proteins not in your database (unique)
- May miss partial proteins

Ab initio

- Predict coding regions using mathematical models
- Training sets are required
- overprediction of small genes
- untypical coding sequences

Examples: Genefinder, Augustus, Glimmer, SNAP, fgenesh

Pipeline for ideal annotation



Useful databases and web-browsers

EnsEMBL -http://www.ensembl.org/index.html

Vega (Vertebrate and Genome Annotation) - http://vega.sanger.ac.uk/index.html

UCSC Genome Browser - http://genome.ucsc.edu/

MGC (Mammalian Gene Collection) - http://genecollectio...ci.nih.gov/MGC/

NCBI Map Viewer - http://www.ncbi.nlm.nih.gov/mapview/

GOLD (Genomes OnLine Database) - http://www.genomesonline.org/

Useful online annotation pipelines

NCBI Prokaryotic Genomes Automatic Annotation Pipeline.
- http://www.ncbi.nlm....nnotation_prok/

IGS Prokaryotic Annotation Pipeline - http://www.igs.umary...hole_genome.php

MAKER Web Annotation Service (MWAS) - http://www.yandell-l...tware/mwas.html

AMIGene - http://www.genoscope...e/Form/form.php

xBASE bacterial genome annotation service - http://xbase.bham.ac.uk/

MITOS - http://mitos.bioinf....zig.de/index.py

.
GenSAS (Genome Sequence Annotation Server) - http://gensas.bioinfo.wsu.edu/

BEACON (automated tool for Bacterial gEnome Annotation ComparisON) - http://www.cbrc.kaust.edu.sa/BEACON/

PEDANT - http://pedant.gsf.de/

Bacterial genome annotation

Eukaryote vs Prokaryote Genomes

	Eukaryote	Prokaryote
Size	∀ Large (10 Mb – 100,000 Mb) ∀ There is not generally a relationship between organism complexity and its genome size (many plants have larger genomes than human!)	∀ Generally small (<10 Mb; most < 5Mb) ∀ Complexity (as measured by # of genes and metabolism) generally proportional to genome size
Content	∀ Most DNA is non-coding	∀ DNA is "coding gene dense"
Telomeres/ Centromeres	∀ Present (Linear DNA)	∀ Circular DNA, doesn't need telomeres ∀ Don't have mitosis, hence, no centromeres.
Number of chromosomes	∀ More than one, (often) including those discriminating sexual identity	∀ Often one, sometimes more, -but plasmids, not true chromosome.
Chromatin	∀ Histone bound (which serves as a genome regulation point)	∀ No histones ∀ Uses supercoiling to pack genome

Eukaryote vs Prokaryote Genomes

	Eukaryote	Prokaryote
Genes	• Often have introns • Intraspecific gene order and number generally relatively stable • many non-coding (RNA) genes • There is NOT generally a relationship between organism complexity and gene number	• No introns • Gene order and number may vary between strains of a species
Gene regulation	• Promoters, often with distal long range enhancers/silencers, MARS, transcriptional domains • Generally mono-cistronic	• Promoters • Enhancers/silencers rare • Genes often regulated as polycistronic operons
Repetitive sequences	• Generally highly repetitive with genome wide families from transposable element propagation	• Generally few repeated sequences • Relatively few transposons
Organelle (subgenomes)	• Mitochondrial (all) • chloroplasts (in plants)	• Absent

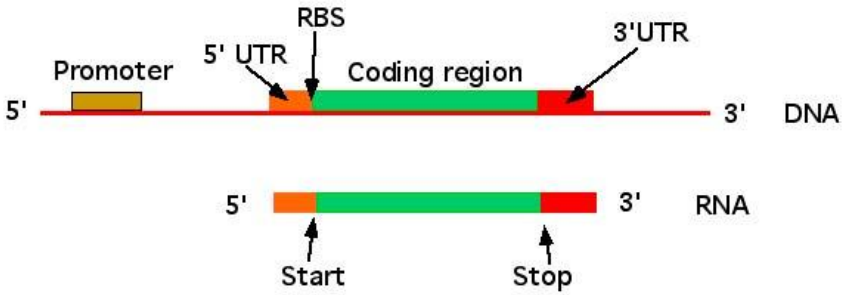
Prokaryotic Genes

- **ATG** is main start codon, but GTG and TTG are also common
- start codons are also used internally: the actual start codon may not be the first one in the ORF.
- The stop codons are the same as in eukaryotes: TGA, TAA, TAG
- stop codons are absolute (the stop codon at the end of an ORF is the end of protein translation): except for a few cases of programmed frameshifts and the use of TGA for selenocysteine.
- Genes can overlap by a small amount. Not much, but a few codons of overlap is common enough so that you can't just eliminate overlaps as impossible.

Cross-species homology works well for many genes. It is very unlikely that non-coding sequence will be conserved.

But, a significant minority of genes (say 20%) are unique to a given species.

Translation start signals (ribosome binding sites) are often found just upstream from the start codon



Bacterial feature types

- **protein coding genes**
  promoter (-10, -35)
  ribosome binding site (RBS)
  coding sequence (CDS)
    - signal peptide, protein domains, structure
  terminator
- **non coding genes**
  transfer RNA (tRNA)
  ribosomal RNA (rRNA)
  non-coding RNA (ncRNA)
- **Other**
  repeat patterns, operons, origin of replication, ...

Gene-finding in Prokaryotes: Easy? ….or not?

ORF Finder

- Open reading frame (ORF) from methionine codon to first Stop codon
- ORFs linked to BLAST
  http://www.ncbi.nlm.nih.gov/gorf/gorf.html

Problem: not All ORFs are genes.
How can this be improved?

Gene-finding in Prokaryotes: Improving predictions…

Common way to search by content

- build Markov models of coding & noncoding regions apply to ORFs or fixed-sized sequence windows

Markov Model approaches: prokaryotic gene prediction

- Glimmer

http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer_3.cgi
http://cbcb.umd.edu/software/glimmer/
open source

- GeneMark

http://opal.biology.gatech.edu/GeneMark/
not open source

Another existing tools for genome annotation:

Software	Ab initio	alignment	Availability	Speed
RAST	Yes	Yes	Web only	12-24 hours
xBASE	Yes	No	Web only	>4 hours
BG7	No	Yes	standalone	>10 hours
PGAAP (NCBI)	Yes	Yes	Email/we	>1 month

BASys Genome Submission

For assistance on running BASys you may wish to check out the BASys HOWTO.

Email Address (Required)

An email address is required to notify you of progress and results.

*Email Address:

Taxonomy (Fields marked with * are required)

*Genome / Contig Identifier: (for identifying output files)

*Gram Stain: Positive Negative

Genus:

Species:

Strain:

Description:

Contig (Required)

Upload your FASTA-formatted bacterial genomic sequence file (Example) : Выбрать файл файл не выбран

Contig is: Circular Linear

Genetic Code: Bacterial

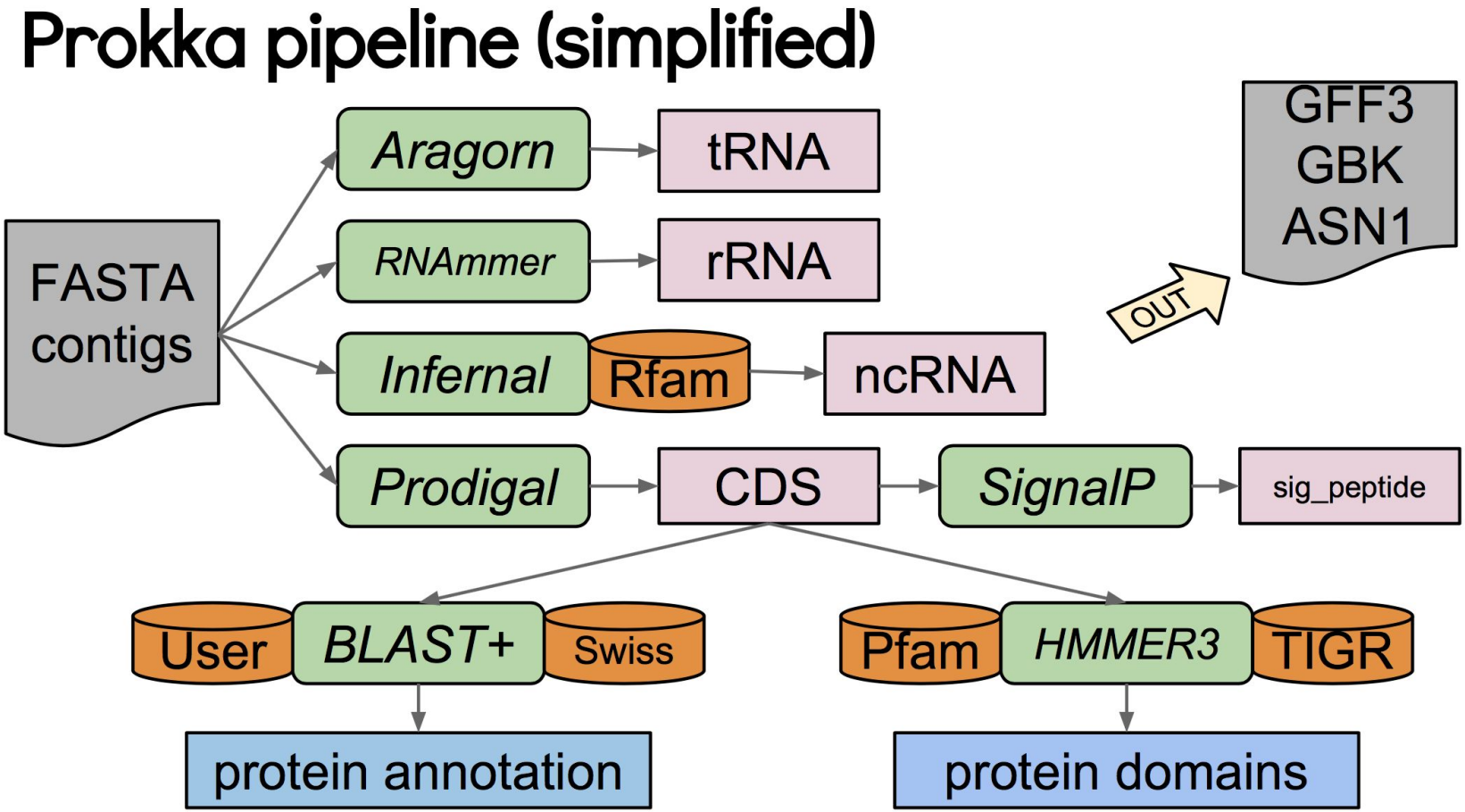
Submit

https://www.basys.ca/

Prokka: rapid prokaryotic genome annotation

- designed for Bacteria, Archaea and Viruses. It can't handle multi-exon gene models
- your own custom "trusted" set (optional)
- core bacterial proteome (default)
- genus-specific proteome (optional)
- whole protein HMMs: PRK clusters, TIGRfams
- protein domain HMMs: Pfam

Prokka: rapid prokaryotic genome annotation



Prokka output

`.fna` FASTA file of original input contigs (nucleotide)
`.faa` FASTA file of translated coding genes (protein)
`.ffn` FASTA file of all genomic features (nucleotide)
`.fsa` Contig sequences for submission (nucleotide)
`.tbl` Feature table for submission
`.sqn` Sequin editable file for submission
`.gbk` Genbank file containing sequences and annotations
`.gff` GFF v3 file containing sequences and annotations
`.log` Log file of Prokka processing output
`.txt` Annotation summary statistics

Prokka

prokka --help
prokka --docs Show full manual/documentation
prokka --setupdb
prokka --listdb List all configured databases

- **prokka --outdir mydir --prefix mygenome contigs.fasta**

Another options:

--addgenes	Add 'gene' features for each 'CDS' feature
--setupdb	Index all installed databases
--kingdom	Annotation mode: Archaea|Bacteria|Mitochondria|Viruses (default 'Bacteria')
--gram	Gram: -/neg +/pos
--fast	Fast mode - skip CDS /product searching (default OFF)
--cpus	Number of CPUs to use [0=all] (default '8')

etc…

http://www.vicbioinformatics.com/software.prokka.shtml
https://github.com/tseemann/prokka/blob/master/README.md

GFF: a standard annotation format

- GFF - General Feature Format (V2, V2.5, **V3**)

- Designed as a single line record for describing features on DNA sequence - originally used for gene prediction output

- The GFF files are text files and every line represents a region on the annotated sequence and these regions are called features

- Features can be functional elements (e.g., genes), genetic polymorphisms (e.g. SNPs, INDELs, or structural variants), or any other annotations

- 9 tab-delimited fields common to all versions

  seq source feature begin end score strand phase group

GFF-version 3

GROUP tag different for ALL versions

- **GFF2:** group is a unique description, usually the gene name. `NCOA1`

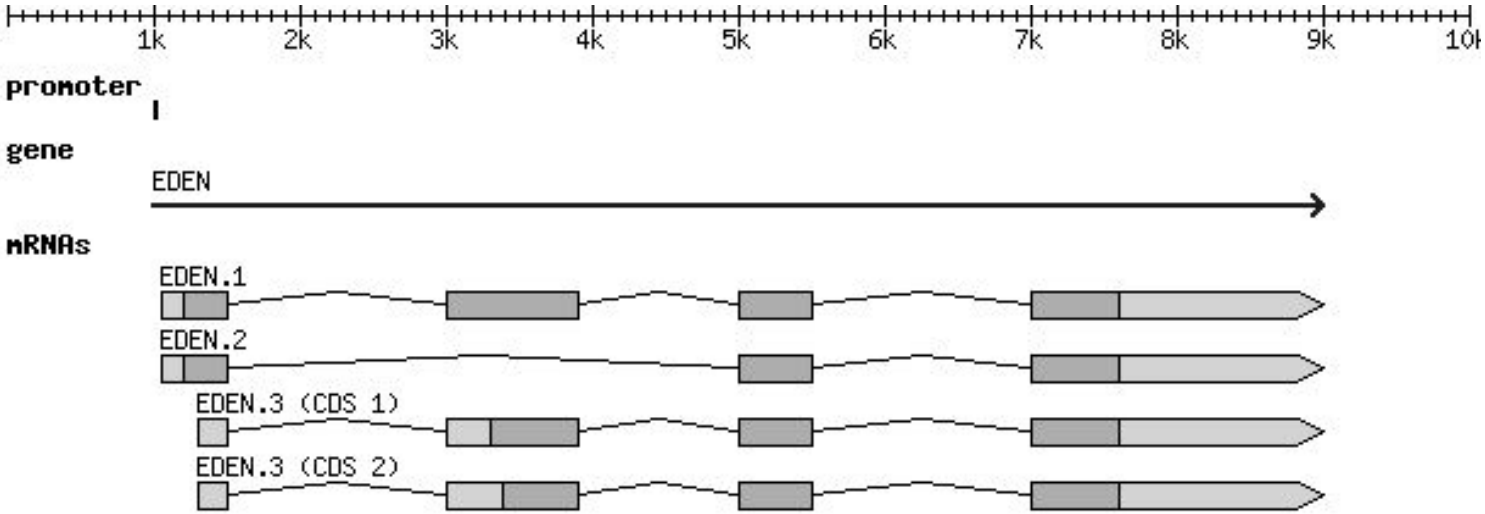
- **GFF2.5 / GTF** (Gene Transfer Format):
  - tag-value pairs introduced,
  - start_codon and stop_codon are required features for CDS

```
transcript_id "NM_056789"; gene_id "NCOA1"
```

- **GFF3:**
  - FASTA seqs can be embedded
  - New tag "Parent" – nested multilevel structure

GFF-version 3

GFF3: New tag “Parent” – nested multilevel structure



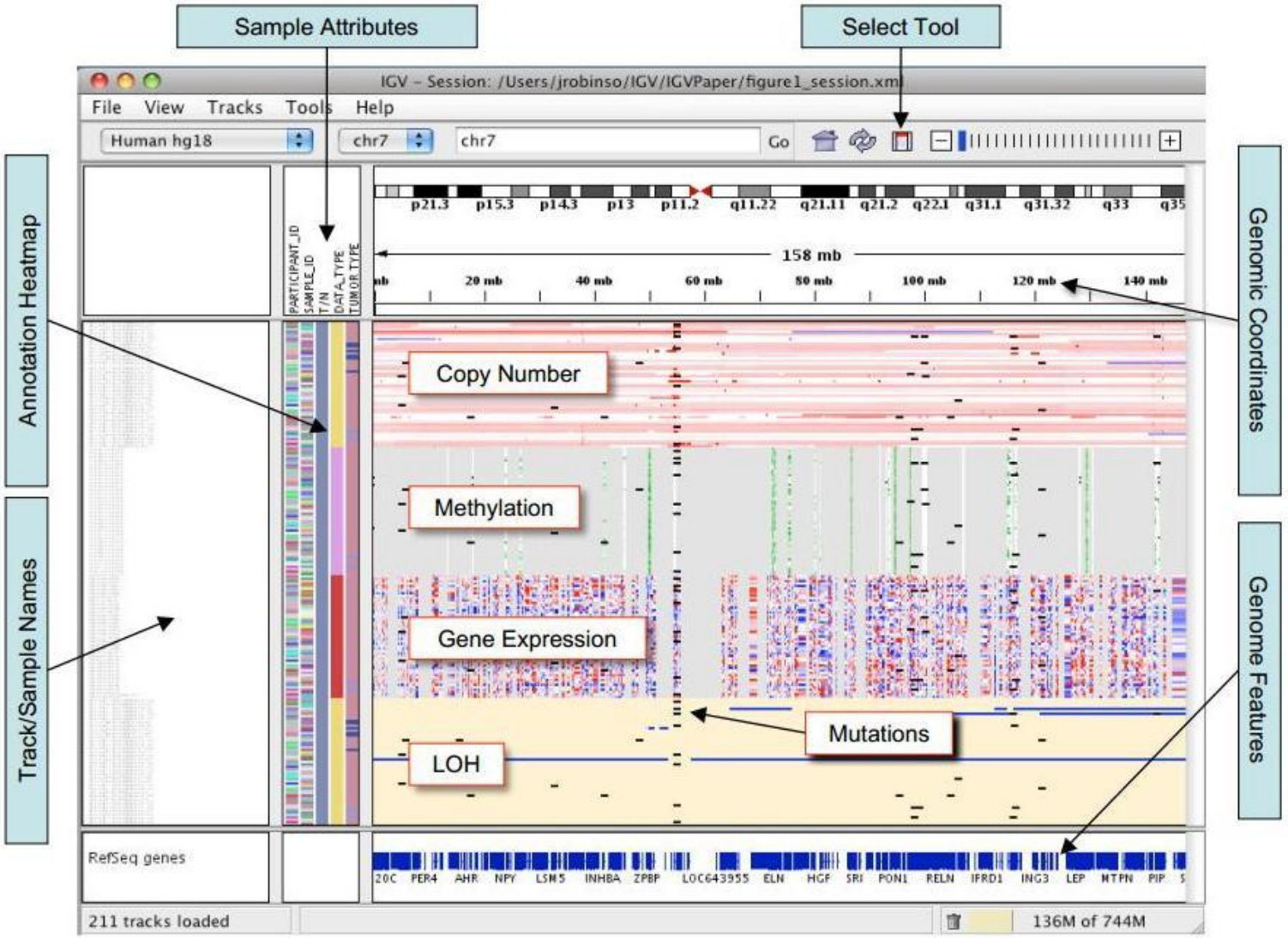

```
0  ##gff-version 3.2.1
1  ##sequence-region ctg123 1 1497228
2  ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDEN
3  ctg123 . TF_binding_site 1000  1012  .  +  .  ID=tfbs00001;Parent=gene00001
4  ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001;Name=EDEN.1
5  ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00002;Parent=gene00001;Name=EDEN.2
6  ctg123 . mRNA            1300  9000  .  +  .  ID=mRNA00003;Parent=gene00001;Name=EDEN.3
7  ctg123 . exon            1300  1500  .  +  .  ID=exon00001;Parent=mRNA00003
8  ctg123 . exon            1050  1500  .  +  .  ID=exon00002;Parent=mRNA00001,mRNA00002
9  ctg123 . exon            3000  3902  .  +  .  ID=exon00003;Parent=mRNA00001,mRNA00003
```

GFF-version 3

GFF3: FASTA seqs can be embedded

```
##gff-version 3.2.1
##sequence-region ctg123 1 1497228
ctg123 . gene                1000  9000  .  +  .  ID=gene00001;Name=EDEN
ctg123 . TF_binding_site     1000  1012  .  +  .  ID=tfbs00001;Parent=gene00001
ctg123 . mRNA                1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001;Name=EDEN.1
ctg123 . five_prime_UTR      1050  1200  .  +  .  Parent=mRNA00001
ctg123 . CDS                 1201  1500  .  +  0  ID=cds00001;Parent=mRNA00001
ctg123 . CDS                 3000  3902  .  +  0  ID=cds00001;Parent=mRNA00001
ctg123 . CDS                 5000  5500  .  +  0  ID=cds00001;Parent=mRNA00001
ctg123 . CDS                 7000  7600  .  +  0  ID=cds00001;Parent=mRNA00001
ctg123 . three_prime_UTR     7601  9000  .  +  .  Parent=mRNA00001
ctg123 . cDNA_match          1050  1500  5.8e-42  +  . ID=match00001;Target=cdna0123+12+462
ctg123 . cDNA_match          5000  5500  8.1e-43  +  . ID=match00001;Target=cdna0123+463+963
ctg123 . cDNA_match          7000  9000  1.4e-40  +  . ID=match00001;Target=cdna0123+964+2964
##FASTA
>ctg123
cttctgggcgtacccgattctcggagaacttgccgcaccattccgccttg
tgttcattgctgcctgcatgttcattgtctacctcggctacgtgtggcta
tctttcctcggtgccctcgtgcacggagtcgagaaaccaaagaacaaaaa
aagaaattaaaatatttattttgctgtggttttgatgtgtgtttttat
aatgatttttgatgtgaccaattgtacttttcctttaaatgaaatgtaat
```

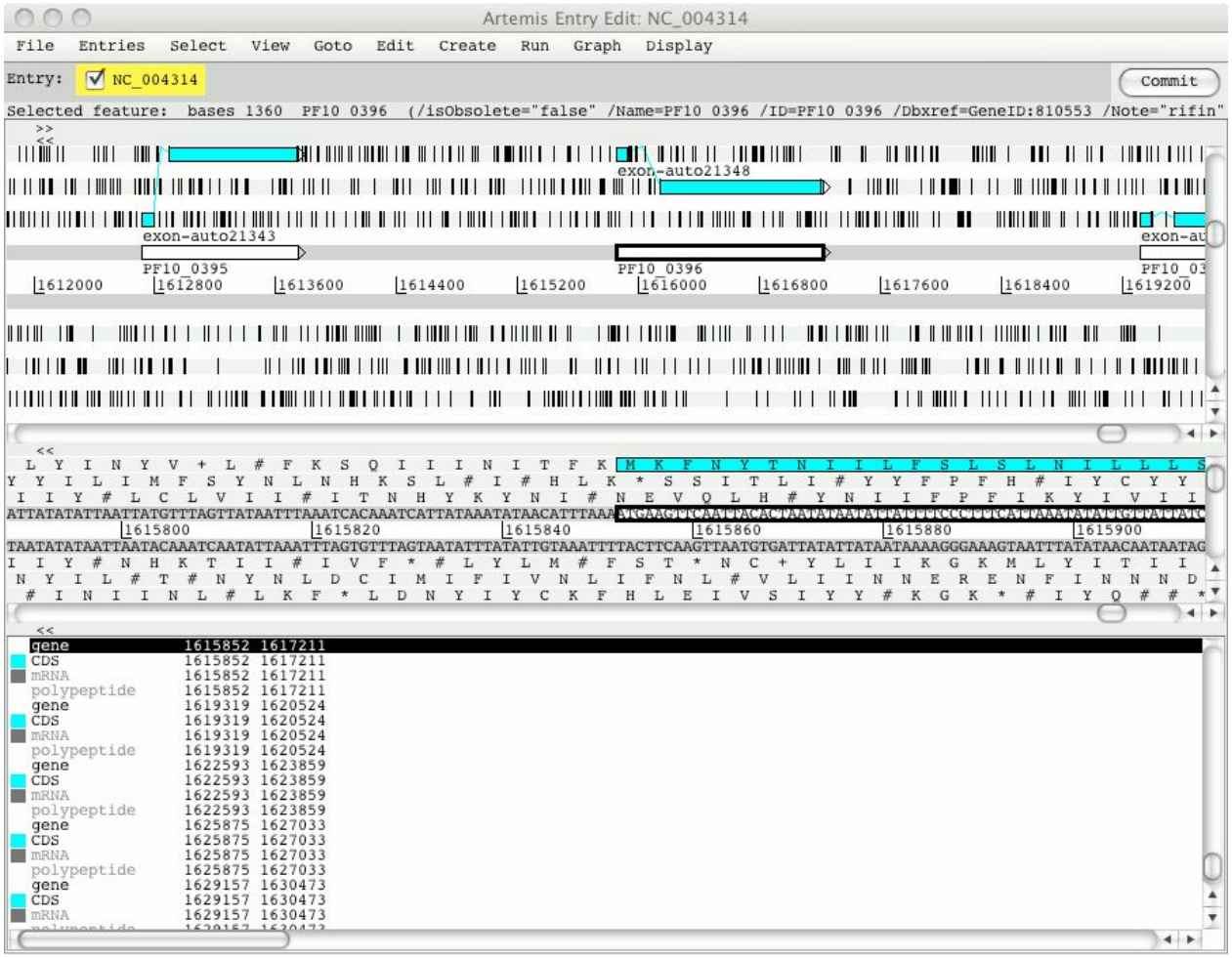
Integrative Genomics Viewer (IGV)



http://software.broadinstitute.org/software/igv/home

genome viewer Artemis

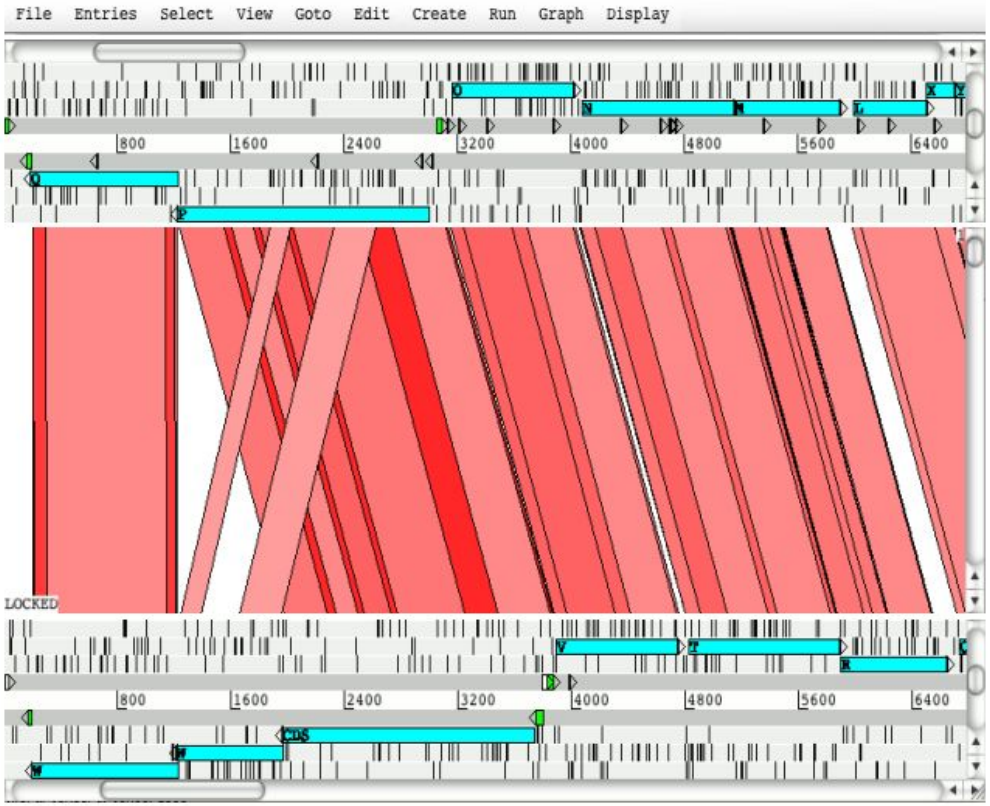
Free genome browser and annotation tool that allows visualization of sequence features, next generation data and the results of analyses within the context of the sequence, and also its six-frame translation

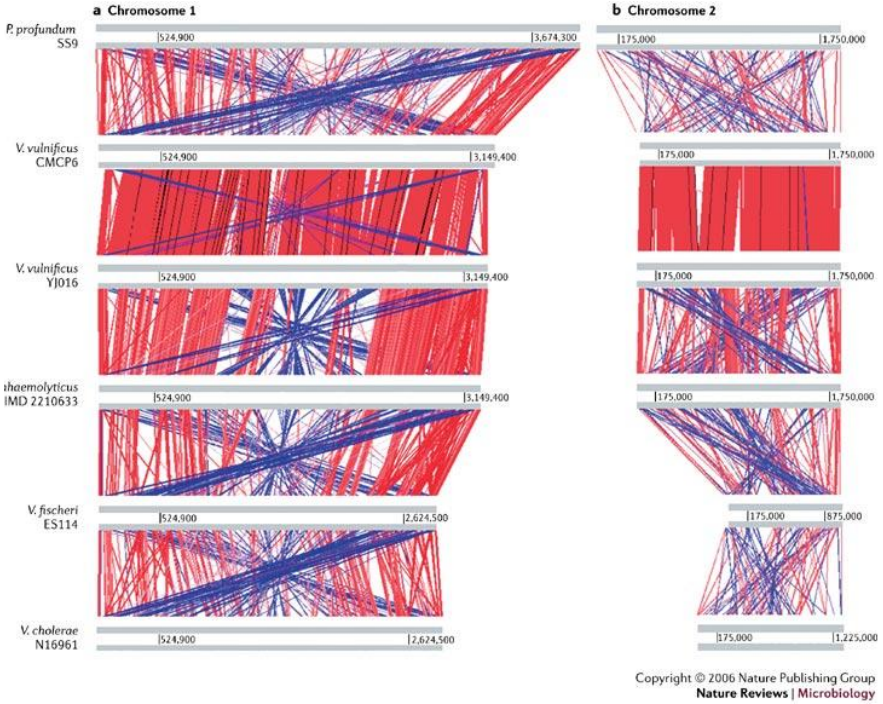


http://www.sanger.ac.uk/science/tools/artemis

Artemis comparison tool (ACT)

Display pairwise comparisons between two or more DNA sequences. Can read complete EMBL, GENBANK and GFF entries or sequences in FASTA or raw format





http://www.sanger.ac.uk/science/tools/artemis-comparison-tool-act