Genome annotation

ILLINOIS INSTITUTE OF TECHNOLOGY

Jean-François Pombert, Ph.D.
Office PS 296 (Lab PS 340)

3 main steps in annotation

1. Gene prediction
   ## Where are the genes and other important elements?
2. Function prediction
   ## What are their respective functions?
   ## We can use homology searches to infer functions!
3. Database submission
   ## Usually involves a lot of reformatting
   ## and error checking (manual curation?!?)

Genomes: prokaryotes ne eukaryotes ne viruses

Different paradigms
Database rules are different
Rules for eukaryotic organelles derived
from endosymbiosis are a mix of
prokaryotes and eukaryotes

Prokaryotes

- **Haploid**
- Usually small
- Circular chromosome
- Plasmids?
- Few introns
- Density of ~ 1 gene/kb

Viruses

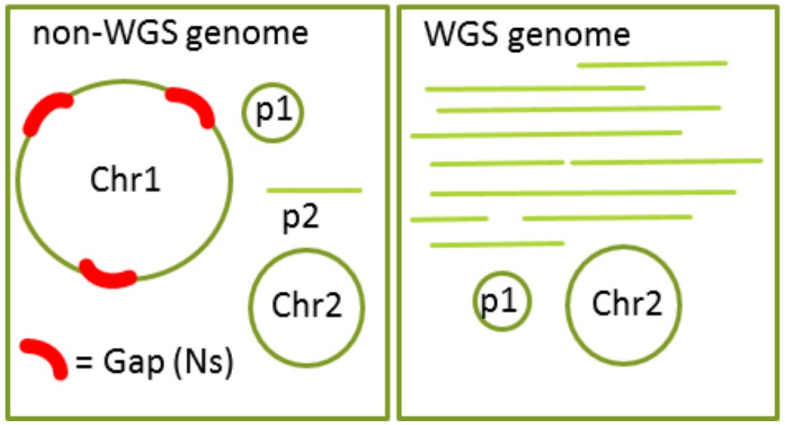
- Haploid, RNA or DNA, polypeptides

Eukaryotes

- **Haploid**, **diploid**, **polyploid**?
- Can be extremely large
- Linear chromosomes with telomeres
- introns + isoforms ## We need ## models to predict genes
- **RNA-seq** can be used to build models

Prokaryotic and Eukaryotic Genomes Submission Guide

Both WGS and non-WGS genomes, including gapless complete bacterial chromosomes, can be submitted via the Submission Portal. You will be asked to choose whether the genome being submitted is considered WGS or not. The differences for GenBank purposes are:



non-WGS

- Each chromosome is in a single sequence and there are no extra sequences
- Each sequence in the genome must be assigned to a chromosome or plasmid or organelle
- Plasmids and organelles can still be in multiple pieces.

WGS

- One or more chromosomes are in multiple pieces and/or some sequences are not assembled into chromosomes

In both cases

- There can still be gaps within the sequences; you will supply that information in the submission
- Plasmids and organelles can still be in multiple pieces.
- Internal sequences must be arranged in the correct order and orientation.
- Sequences concatenated in unknown order are not allowed.

NCBI's GenBank – Annotations

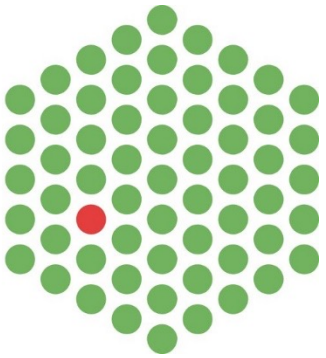
http://www.ncbi.nlm.nih.gov/genbank/genomesubmit/ ## General

https://www.ncbi.nlm.nih.gov/genbank/genomesubmit_annotation/ ## Prokaryotes

https://www.ncbi.nlm.nih.gov/genbank/eukaryotic_genome_submission/ ## Eukaryotes

Annotating? Read these first. Naming conventions differ.

EMBL



DDBJ
DNA Data Bank of Japan

Reminder – DB data is mirrored

NCBI **N**ational **C**enter for **B**iotechnology **I**nformation ## Americas

EMBL **E**uropean **M**olecular **B**iology **L**abs ## Europe

DDBJ **D**NA **D**ata**b**ase of **J**apan ## Asia

An heterotrimeric protein

3A1J

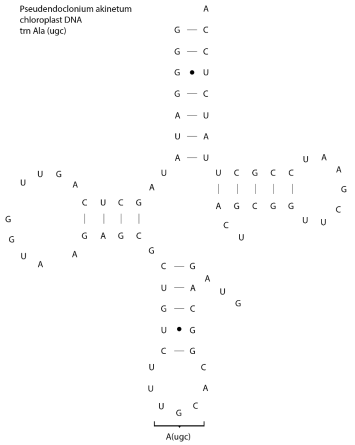
Crystal structure of the human Rad9-Hus1-Rad1 complex

PDB DOI: 10.2210/pdb3A1J/pdb

Classification: HYDROLASE/CELL CYCLE
Organism(s): Homo sapiens
Expression System: Escherichia coli
Mutation(s): No

Deposited: 2009-04-08 **Released:** 2009-06-02
Deposition Author(s): Sohn, S.Y., Cho, Y.

A transfer RNA



4TRA

RESTRAINED REFINEMENT OF TWO CRYSTALLINE FORMS OF YEAST ASPARTIC ACID AND PHENYLALANINE TRANSFER RNA CRYSTALS

PDB DOI: 10.2210/pdb4TRA/pdb **NDB:** TRNA09

Classification: T-RNA
Organism(s): Saccharomyces cerevisiae
Mutation(s): No

Deposited: 1987-11-06 **Released:** 1987-11-06
Deposition Author(s): Westhof, E., Dumas, P., Moras, D.

https://www.rcsb.org/structure/4TRA

What to look for?

Genes (proteins, rRNAs, tRNAs and ncRNAs) ## We need gene predictors

Repeats and invasive elements ## Repeat masking in eukaryotic annot.

Promoters and regulatory elements ## Not well defined in many cases

Nature Vol. 271 9 February 1978

news and views

Why genes in pieces?
Gilbert W
Nature. 1978. 271(9): 501
DOI: 10.1038/271501a0

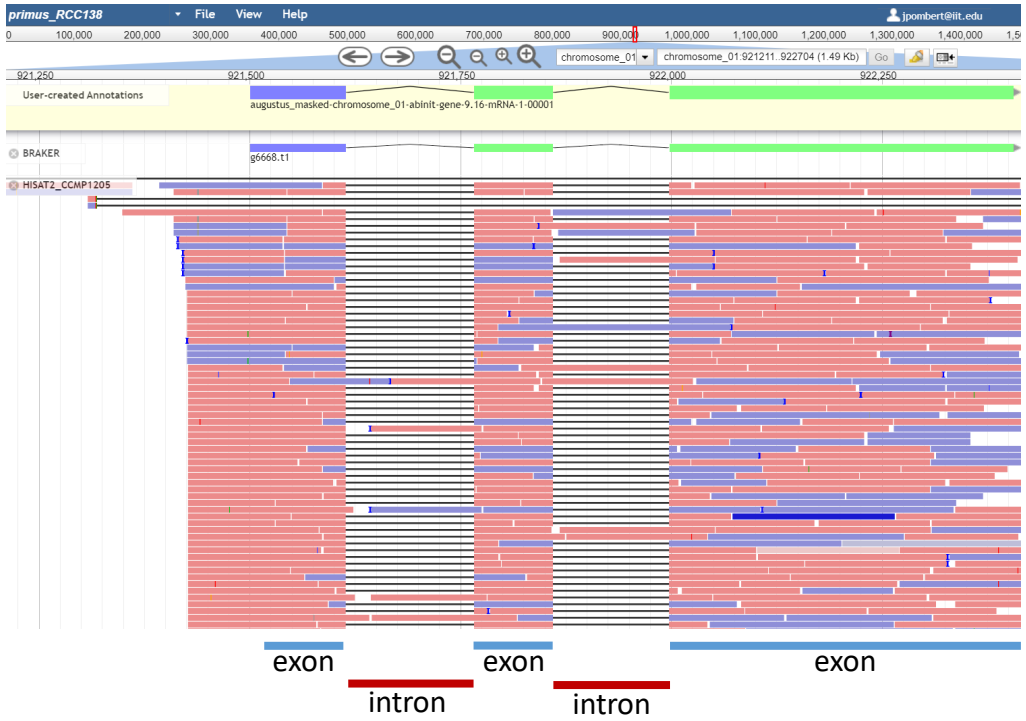
Why genes in pieces?

from Walter Gilbert

OUR picture of the organisation of genes in higher organisms has recently undergone a revolution. Analyses of eukaryotic genes in many laboratories [1-10], studies of globin, ovalbumin, immunoglobulin, SV40 and polyoma, suggest that in general the coding sequences on DNA, the regions that will ultimately be translated into amino acid sequence, are not continuous but are interrupted by

A gene, a contiguous region of DNA, now corresponds to one transcription unit, but that transcription unit can correspond to many polypeptide chains, of related or differing functions.

Recombination now becomes more rapid. Since the gene is spread out over a larger region of DNA, recombination, which should be hampered in higher cells by the inability of DNA molecules



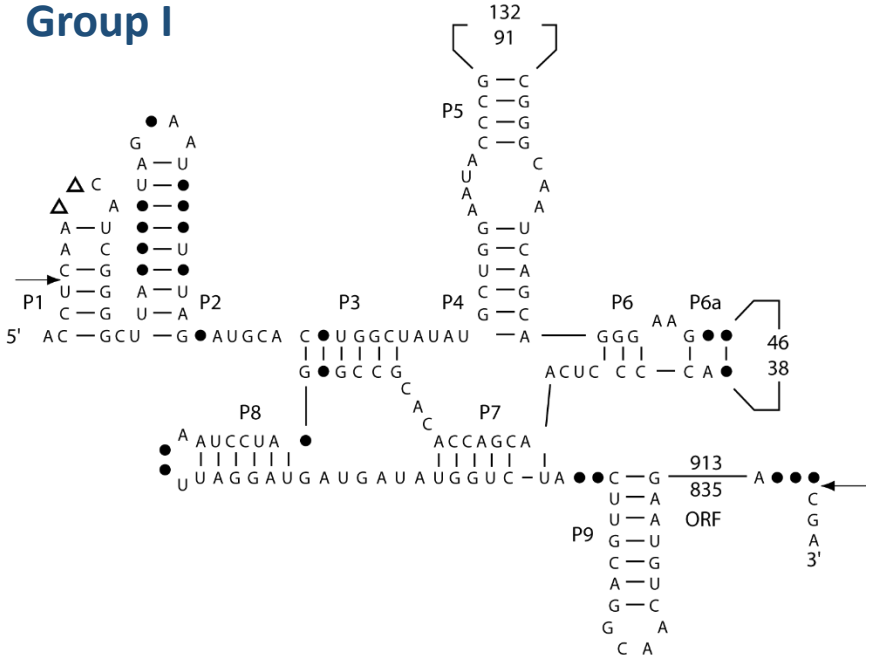
Genes – Exons and introns

Exons – **Ex**pressed regi**ons**

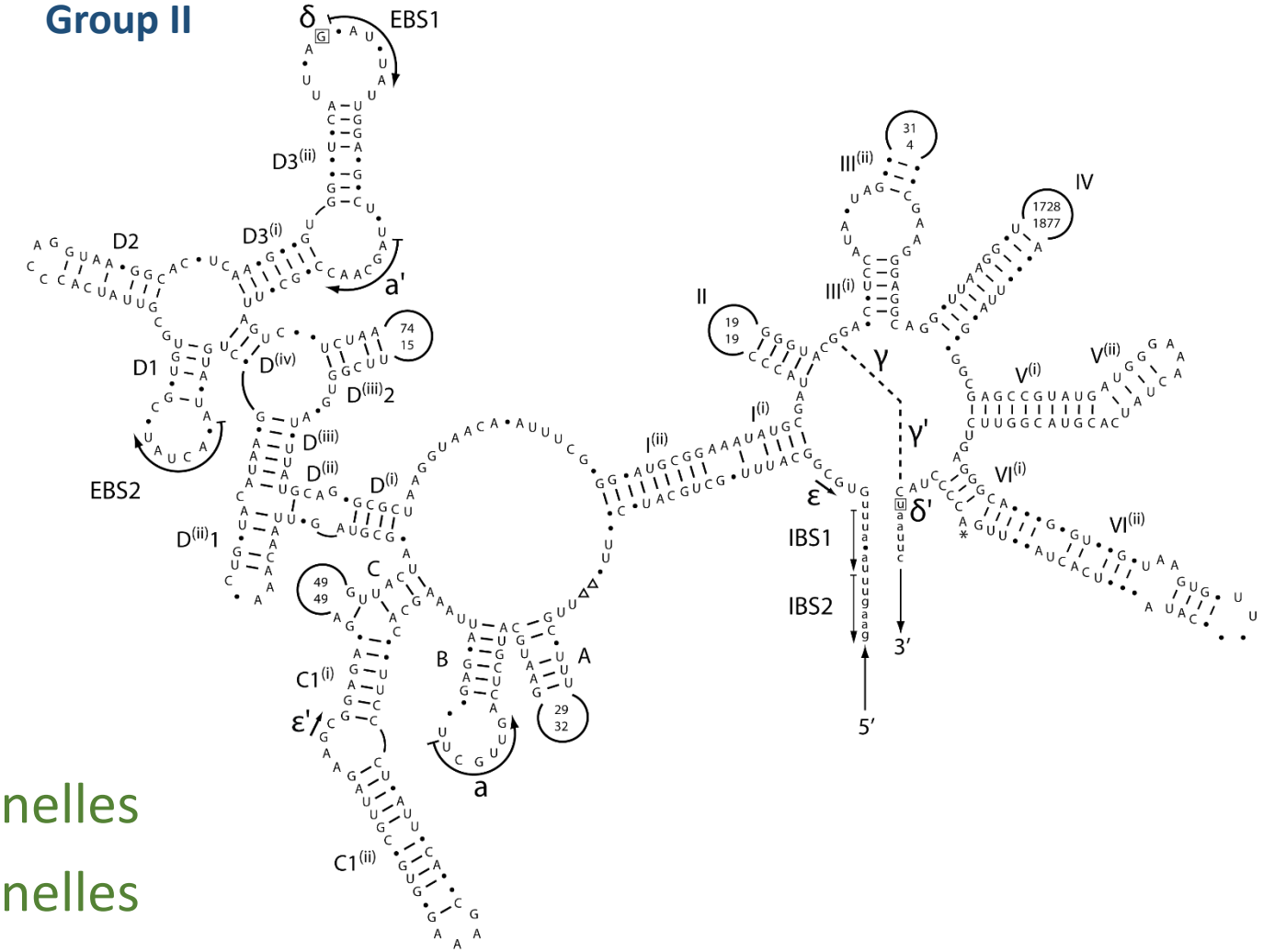
Introns – **Intr**agenic regi**ons**

Not true anymore, some introns are
located in UTRs (untranslated regions)

Group I



Group II



Introns

Group I	## Prokaryotes + organelles
Group II	## Prokaryotes + organelles
Spliceosomal	## Eukaryotes
	## Nuclear, derived from GII

The complete mitochondrial DNA sequence of the green alga *Oltmannsiellopsis viridis*: Evolutionary trends of the mitochondrial genome in the Ulvophyceae

Pombert JF, Beauchamp P, Otis C, Lemieux C, Turmel M

Current Genetics. 2006. 50:137-147. PMID: 16721603

Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes
Lomsadze A, Gemayel K, Tang S, Borodovsky M
Genome Res. 2018. 28(7):1079-1089. doi: 10.1101/gr.230615.117

Do we have introns?
Do we have RNAseq data?

Protein prediction

Machine learning is highly useful in genome annotation. A few tools:

Prodigal	https://github.com/hyattpd/Prodigal	## Prokaryotes
Augustus	https://github.com/Gaius-Augustus/Augustus	## Eukaryotes
GeneMark	http://exon.gatech.edu/GeneMark/	## Pro/Eukaryotes

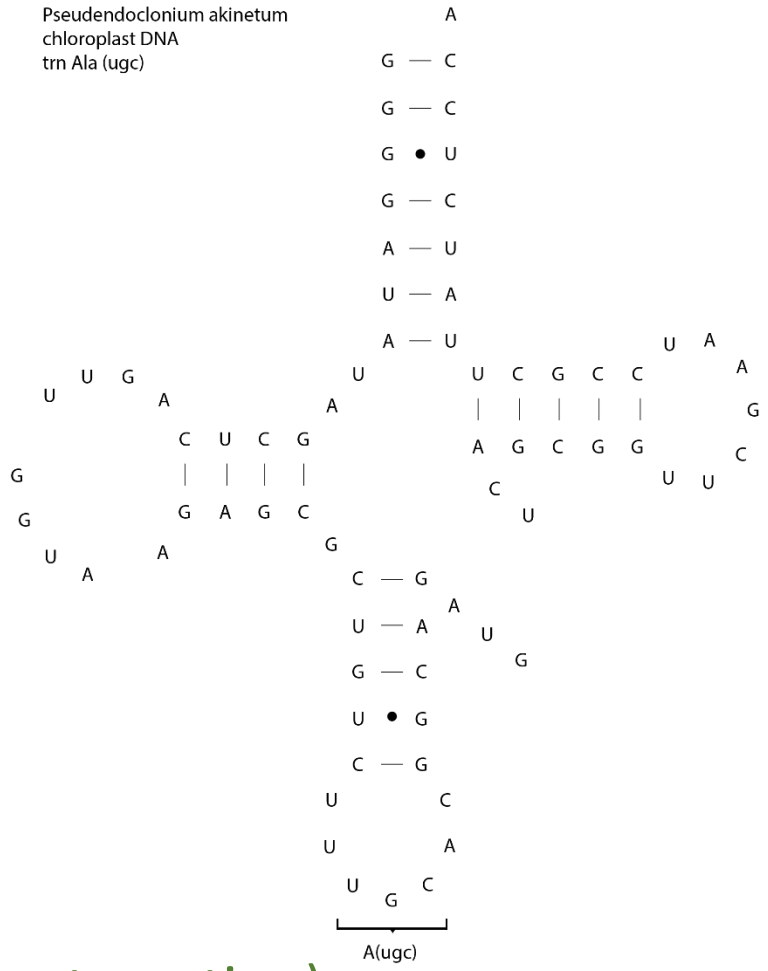
Prodigal: prokaryotic gene recognition and translation initiation site identification
Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ
BMC Bioinformatics. 2010. 11:119. doi: 10.1186/1471-2105-11-119

A novel hybrid gene prediction method employing protein multiple sequence alignments
Keller O, Kollmar M, Stanke M, Waack S
Bioinformatics. 2011. 27(6):757-63. doi: 10.1093/bioinformatics/btr010

The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs
Schattner P, Brooks AN, Lowe TM
Nucleic Acids Res. 2005. 33(Web Server issue):W686-W689

tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence
Lowe TM, Eddy SR
Nucleic Acids Res. 1997. 25(5):955-964

ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences
Laslett D, Canback B
Nucleic Acids Res. 2004 Jan 2;32(1):11-6



tRNA/tmRNA prediction

tRNAscan-SE http://lowelab.ucsc.edu/tRNAscan-SE/

Aragorn http://www.ansikte.se/ARAGORN/

Find tRNAs based on their canonical cloverleaf structure

Both can also be used from the command line (great for automation)

README.md

build passing License GPL v3

Barrnap

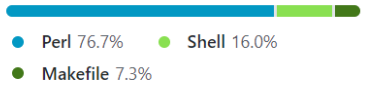
BAsic Rapid Ribosomal RNA Predictor

Description

Barrnap predicts the location of ribosomal RNA genes in genomes. It supports bacteria (5S,23S,16S), archaea (5S,5.8S,23S,16S), metazoan mitochondria (12S,16S) and eukaryotes (5S,5.8S,28S,18S).

It takes FASTA DNA sequence as input, and write GFF3 as output. It uses the new `nhmmer` tool that comes with HMMER 3.1 for HMM searching in RNA:DNA style. Multithreading is supported and one can expect roughly linear speed-ups with more CPUs.

Languages



https://github.com/tseemann/barrnap

Usage

```
% barrnap --quiet examples/small.fna
##gff-version 3
P.marinus       barrnap:0.8     rRNA    353314  354793  0       +       .       Name=16S_rRNA;pr
P.marinus       barrnap:0.8     rRNA    355464  358334  0       +       .       Name=23S_rRNA;pr
P.marinus       barrnap:0.8     rRNA    358433  358536  7.5e-07 +       .       Name=5S_rRNA;prc

% barrnap -q -k mito examples/mitochondria.fna
##gff-version 3
AF346967.1      barrnap:0.8     rRNA    643     1610    .       +       .       Name=12S_rRNA;pr
AF346967.1      barrnap:0.8     rRNA    1672    3228    .       +       .       Name=16S_rRNA;pr

% barrnap -o rrna.fa < contigs.fa > rrna.gff
% head -n 3 rrna.fa
>16S_rRNA::gi|329138943|tpg|BK006945.2|:455935-456864(-)
ACGGTCGGGGGCATCAGTATTCAATTGTCAGAGGTGAAATTCTTGGATT
TATTGAAGACTAACTACTGCGAAAGCATTTGCCAAGGACGTTTTCATTA
```

rRNA prediction

Barrnap https://github.com/tseemann/barrnap

BAsic Rapid Ribosomal RNA Predictor; meant as a replacement for RNAmmer

RNAmmer https://services.healthtech.dtu.dk/service.php?RNAmmer-1.2

Available as a web page or stand-alone program

Requires a license (free for academic purposes)

RNAmmer: consistent and rapid annotation of ribosomal RNA genes
Lagesen K, Hallin P, Rødland EA, Staerfeldt HH, Rognes T, Ussery DW
Nucleic Acids Res. 2007. 35(9):3100-3108

Automatic annotation of non-coding genes

http://useast.ensembl.org/info/genome/genebuild/ncrna.html

Non-coding RNAs (ncRNAs) are involved in many biological processes and are increasingly seen as important. As is the case with proteins, it is the overall structure of the molecule which imparts function. However, while similar protein structures are often reflected in a conserved amino acid sequence, sequences underlying RNA secondary structure are very variable; this makes ncRNAs difficult to detect using sequence alone.

Types of ncRNA

Abbreviation	Definition
tRNA	transfer RNA
Mt-tRNA	transfer RNA located in the mitochondrial genome
rRNA	ribosomal RNA
scRNA	small cytoplasmic RNA
snRNA	small nuclear RNA
snoRNA	small nucleolar RNA
miRNA	microRNA precursors
misc_RNA	miscellaneous other RNA
lincRNA	Long intergenic non-coding RNAs

Other RNAs? – It's complicated

There is no good consensus

In the absence of sequence homology, we must rely on folding structures

http://useast.ensembl.org/info/genome/genebuild/ncrna.html ## A good overview

Annotation pipelines

Because of spliceosomal introns, eukaryotic annotators work best with RNA-seq data

Prokaryotes

- Prokka
  https://github.com/tseemann/prokka
- DFAST
  https://dfast.nig.ac.jp/
- NCBI Prokaryotic Genome Annotation Pipeline (PGAP)
  https://www.ncbi.nlm.nih.gov/genome/annotation_prok

Eukaryotes

- BRAKER
  https://github.com/Gaius-Augustus/BRAKER
- MAKER
  https://www.yandell-lab.org/software/maker.html
- NCBI Eukaryotic Genome Annotation Pipeline
  https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/

Prokka: rapid prokaryotic genome annotation
Seemann T
Bioinformatics. 2014 Jul 15;30(14):2068-9. DOI: 10.1093/bioinformatics/btu153

DFAST: a flexible prokaryotic genome annotation pipeline for faster genome publication
Tanizawa Y, Fujisawa T, Nakamura Y
Bioinformatics. 2018 Mar 15;34(6):1037-1039. DOI: 10.1093/bioinformatics/btx713

BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS
Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M
Bioinformatics. 2016 Mar 1;32(5):767-9. DOI: 10.1093/bioinformatics/btv661

MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects
Holt C, Yandell M
BMC Bioinformatics. 2011 Dec 22;12:491. DOI: 10.1186/1471-2105-12-491

Proper Use of /locus_tag in Genome Submissions

At the International Nucleotide Sequence Database Collaborators meeting, it was agreed that we would require genome projects to be registered with the database. Each genome project would be assigned an ID in order to allow us to associate multiple sequences of a single genome project with each other. This Genome Project ID will appear in a new line type below ACCESSION and VERSION in the flat file. Registration of Genome Projects can be done at DDBJ, EBI or NCBI. A submitter can also register for a locus_tag prefix at the same time that they register their genome project.

Locus_tags are identifiers that are systematically applied to every gene in a genome. These tags have become surrogate gene names by the biological community. If two submitters of two different genomes use the same systematic names to describe two very different genes in two very different genomes, it can be very confusing. In order to prevent this from happening INSD has created a registry of locus_tag prefixes. Submitters of eukaryotic and prokaryotic genomes should register their prefix prior to submitting their genome. All components of a project (such as multiple chromosomes or plasmids, etc) should use the same locus_tag prefix.

A white paper about locus tags and their intended usage:
https://www.ncbi.nlm.nih.gov/genomes/locustag/Proposal.pdf

Genes *vs.* locus tags

Genes can be repeated in databases ## Many organism share the same genes!

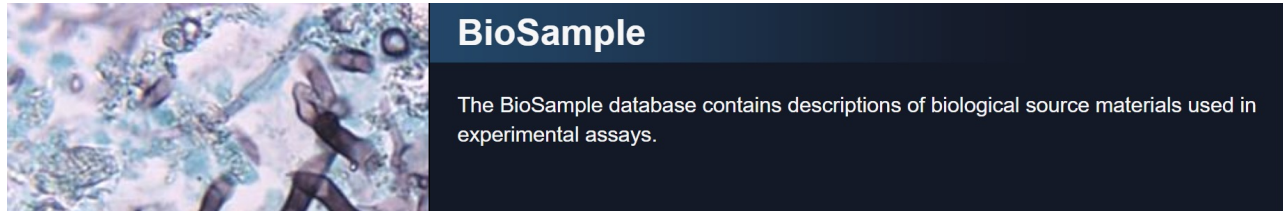
Gene names imply functions ## *e.g. recA* is involved in DNA recombination

Locus tags are unique database IDs ## Mandatory for NCBI/EMBL/DDBJ annotations

Locus tags names do not imply function ## b2699 is a simple database key

https://www.ncbi.nlm.nih.gov/gene/?term=reca
https://www.ncbi.nlm.nih.gov/gene/?term=b2699

https://www.ncbi.nlm.nih.gov/bioproject/



https://www.ncbi.nlm.nih.gov/biosample

Requesting locus_tag prefixes

https://submit.ncbi.nlm.nih.gov/subs/bioproject/

Each $prefix is unique, requested from databases

Should be requested early in the annotation stage; requires an NCBI account (free)

Do not request locus_tags in this course! We are not submitting anything to NCBI!

Submitting to NCBI? – A workflow

1) Request a Bioproject https://submit.ncbi.nlm.nih.gov/subs/bioproject/
2) Request one or more Biosample(s) https://submit.ncbi.nlm.nih.gov/subs/biosample/
3) Annotate contigs with NCBI or other pipelines ## Using the unique locus tag prefix

If non-NCBI pipeline:

4) Manual curation of gene predictions ## Optional
5) Manual curation of inferred functions ## Optional
6) Convert annotations to TBL format
7) Generate annotations in ASN (SQN) format with table2asn
8) Submit to proper NCBI portal (https://submit.ncbi.nlm.nih.gov/)

Exercise 01 – Prokka (A Perl pipeline)

We will use the genome from the well-known *Escherichia coli* K12 strain for this
exercise

1) Launch the command in a screen ## That way you can detach if needed
2) Annotate the *E. coli* genome (Ecoli_K12.fasta) with Prokka:
   prokka --addgenes --compliant -genus Escherichia --species coli --strain K12 \
   --gcode 11 --gram neg --cpus 4 ./Ecoli_K12.fasta
3) Look inside the output folder ## Prokka gives you multiple formats to facilitate editing
4) Compare the Prokka annotation with the original one using diff:
   diff Ecoli_K12.gb PROKKA_*/PROKKA_*.gbk | less ## Hard to see right?
   grep -c '/locus_tag=' Ecoli_K12.gb PROKKA_*/PROKKA_*.gbk ## Same number?

The GB/GBK/GBFF structure

```
LOCUS       NC_008256              56761 bp    DNA     circular PLN 28-JUL-2006
DEFINITION  Oltmannsiellopsis viridis mitochondrion, complete genome.
ACCESSION   NC_00856
VERSION     NC_008256.1  GI:110816043
DBLINK      Project: 16997
            BioProject: PRJNA16997
KEYWORDS    RefSeq.
SOURCE      mitochondrion Oltmannsiellopsis viridis
  ORGANISM  Oltmannsiellopsis viridis
            Eukaryota; Viridiplantae; Chlorophyta; Oltmannsiellopsis.
REFERENCE   1  (bases 1 to 56761)
  AUTHORS   Pombert,J.F., Beauchamp,P., Otis,C., Lemieux,C. and Turmel,M.
  TITLE     The complete mitochondrial DNA sequence of the green alga
            Oltmannsiellopsis viridis: evolutionary trends of the mitochondrial
            genome in the Ulvophyceae
  JOURNAL   Curr. Genet. 50 (2), 137-147 (2006)
   PUBMED   16721603
REFERENCE   2  (bases 1 to 56761)
  CONSRTM   NCBI Genome Project
  TITLE     Direct Submission
  JOURNAL   Submitted (27-JUL-2006) National Center for Biotechnology
            Information, NIH, Bethesda, MD 20894, USA
REFERENCE   3  (bases 1 to 56761)
  AUTHORS   Pombert,J.-F., Beauchamp,P., Otis,C., Lemieux,C. and Turmel,M.
  TITLE     Direct Submission
  JOURNAL   Submitted (17-JAN-2006) Biochimie et Microbiologie, Universite
            Laval, Pavillon Charles-Eugene Marchand, Quebec, QC G1K 7P4, Canada
COMMENT     PROVISIONAL REFSEQ: This record has not yet been subject to final
            NCBI review. The reference sequence was derived from DQ365900.
            COMPLETENESS: full length.
FEATURES             Location/Qualifiers
     source          1..56761
                     /organism="Oltmannsiellopsis viridis"
                     /organelle="mitochondrion"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:51324"
                     /cell_line="NIES 360"
     gene            complement(12603..12986)
                     /gene="rps12"
                     /locus_tag="OlviMp06"
                     /db_xref="GeneID:4200914"
     CDS             complement(12603..12986)
                     /gene="rps12"
                     /locus_tag="OlviMp06"
                     /codon_start=1
                     /product="ribosomal protein S12"
                     /protein_id="YP_684380.1"
                     /db_xref="GI:110816049"
ORIGIN
        1 taaaataata atgaaaagtg ctattaaata agtagtcttg ctttcaaaaa atatctacaa
       61 aatgatacgg caacgctact taatgttcca ttaagtagtt ttgtagcttc gctcgtagat
      121 atttttttaga cgttttagat ttatacaaat cacttttgt tttagattta tataaatcac
```

The GFF3 structure

```
##gff-version 3
#!gff-spec-version 1.21
#!processor NCBI annotwriter
#!genome-build Helico_v1.0
#!genome-build-accession NCBI_Assembly:GCA_000690575.1
##sequence-region AYPS01000001.1 1 24678
##species https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1291522
AYPS01000001.1  Genbank region  1   24678   .   +   .   ID=id0;Dbxref=taxon:1291522;collection-date=1998;country=USA: Florida%2C Hatchet Creek%2C A
AYPS01000001.1  Genbank gene    1046    2301    .   -   .   ID=gene0;Name=H632_c1p0;end_range=2301,.;gbkey=Gene;gene_biotype=protein_coding;locus_t
AYPS01000001.1  Genbank mRNA    1046    2301    .   -   .   ID=rna0;Parent=gene0;end_range=2301,.;gbkey=mRNA;orig_protein_id=gnl|ubcbot|H632_c1p0;c
AYPS01000001.1  Genbank exon    2100    2301    .   -   .   ID=id1;Parent=rna0;end_range=2301,.;gbkey=mRNA;orig_protein_id=gnl|ubcbot|H632_c1p0;ori
AYPS01000001.1  Genbank exon    1919    2022    .   -   .   ID=id2;Parent=rna0;gbkey=mRNA;orig_protein_id=gnl|ubcbot|H632_c1p0;orig_transcript_id=g
AYPS01000001.1  Genbank exon    1729    1809    .   -   .   ID=id3;Parent=rna0;gbkey=mRNA;orig_protein_id=gnl|ubcbot|H632_c1p0;orig_transcript_id=g
AYPS01000001.1  Genbank exon    1465    1614    .   -   .   ID=id4;Parent=rna0;gbkey=mRNA;orig_protein_id=gnl|ubcbot|H632_c1p0;orig_transcript_id=g
AYPS01000001.1  Genbank exon    1046    1381    .   -   .   ID=id5;Parent=rna0;gbkey=mRNA;orig_protein_id=gnl|ubcbot|H632_c1p0;orig_transcript_id=g
AYPS01000001.1  Genbank CDS 2100    2301    .   -   0   ID=cds0;Parent=rna0;Dbxref=NCBI_GP:KDD77189.1;Name=KDD77189.1;gbkey=CDS;orig_transcript_id=
AYPS01000001.1  Genbank CDS 1919    2022    .   -   2   ID=cds0;Parent=rna0;Dbxref=NCBI_GP:KDD77189.1;Name=KDD77189.1;gbkey=CDS;orig_transcript_id=
AYPS01000001.1  Genbank CDS 1729    1809    .   -   0   ID=cds0;Parent=rna0;Dbxref=NCBI_GP:KDD77189.1;Name=KDD77189.1;gbkey=CDS;orig_transcript_id=
AYPS01000001.1  Genbank CDS 1465    1614    .   -   0   ID=cds0;Parent=rna0;Dbxref=NCBI_GP:KDD77189.1;Name=KDD77189.1;gbkey=CDS;orig_transcript_id=
AYPS01000001.1  Genbank CDS 1046    1381    .   -   0   ID=cds0;Parent=rna0;Dbxref=NCBI_GP:KDD77189.1;Name=KDD77189.1;gbkey=CDS;orig_transcript_id=
AYPS01000001.1  Genbank gene    3782    4823    .   -   .   ID=gene1;Name=H632_c1p1;end_range=4823,.;gbkey=Gene;gene_biotype=protein_coding;locus_t
AYPS01000001.1  Genbank mRNA    3782    4823    .   -   .   ID=rna1;Parent=gene1;end_range=4823,.;gbkey=mRNA;orig_protein_id=gnl|ubcbot|H632_c1p1;c
AYPS01000001.1  Genbank exon    4769    4823    .   -   .   ID=id6;Parent=rna1;end_range=4823,.;gbkey=mRNA;orig_protein_id=gnl|ubcbot|H632_c1p1;ori
AYPS01000001.1  Genbank exon    4035    4609    .   -   .   ID=id7;Parent=rna1;gbkey=mRNA;orig_protein_id=gnl|ubcbot|H632_c1p1;orig_transcript_id=g
AYPS01000001.1  Genbank exon    3782    3946    .   -   .   ID=id8;Parent=rna1;gbkey=mRNA;orig_protein_id=gnl|ubcbot|H632_c1p1;orig_transcript_id=g
AYPS01000001.1  Genbank CDS 4769    4823    .   -   0   ID=cds1;Parent=rna1;Dbxref=NCBI_GP:KDD77190.1;Name=KDD77190.1;gbkey=CDS;orig_transcript_id=
AYPS01000001.1  Genbank CDS 4035    4609    .   -   2   ID=cds1;Parent=rna1;Dbxref=NCBI_GP:KDD77190.1;Name=KDD77190.1;gbkey=CDS;orig_transcript_id=
AYPS01000001.1  Genbank CDS 3782    3946    .   -   0   ID=cds1;Parent=rna1;Dbxref=NCBI_GP:KDD77190.1;Name=KDD77190.1;gbkey=CDS;orig_transcript_id=
AYPS01000001.1  Genbank gene    7169    9142    .   -   .   ID=gene2;Name=H632_c1p2;Note=similar to PFAM: PF01490.13%2C 6.9E-049;end_range=9142,.;g
AYPS01000001.1  Genbank mRNA    7169    9142    .   -   .   ID=rna2;Parent=gene2;Note=similar to PFAM: PF01490.13%2C 6.9E-049;end_range=9142,.;gbke
AYPS01000001.1  Genbank exon    9078    9142    .   -   .   ID=id9;Parent=rna2;Note=similar to PFAM: PF01490.13%2C 6.9E-049;end_range=9142,.;gbkey=
AYPS01000001.1  Genbank exon    8856    8986    .   -   .   ID=id10;Parent=rna2;Note=similar to PFAM: PF01490.13%2C 6.9E-049;gbkey=mRNA;orig_protei
AYPS01000001.1  Genbank exon    8650    8758    .   -   .   ID=id11;Parent=rna2;Note=similar to PFAM: PF01490.13%2C 6.9E-049;gbkey=mRNA;orig_protei
AYPS01000001.1  Genbank exon    8295    8529    .   -   .   ID=id12;Parent=rna2;Note=similar to PFAM: PF01490.13%2C 6.9E-049;gbkey=mRNA;orig_protei
AYPS01000001.1  Genbank exon    7683    8210    .   -   .   ID=id13;Parent=rna2;Note=similar to PFAM: PF01490.13%2C 6.9E-049;gbkey=mRNA;orig_protei
AYPS01000001.1  Genbank exon    7442    7595    .   -   .   ID=id14;Parent=rna2;Note=similar to PFAM: PF01490.13%2C 6.9E-049;gbkey=mRNA;orig_protei
AYPS01000001.1  Genbank exon    7169    7359    .   -   .   ID=id15;Parent=rna2;Note=similar to PFAM: PF01490.13%2C 6.9E-049;gbkey=mRNA;orig_protei
AYPS01000001.1  Genbank CDS 9078    9142    .   -   0   ID=cds2;Parent=rna2;Dbxref=NCBI_GP:KDD77191.1;Name=KDD77191.1;Note=similar to PFAM: PF01490
AYPS01000001.1  Genbank CDS 8856    8986    .   -   1   ID=cds2;Parent=rna2;Dbxref=NCBI_GP:KDD77191.1;Name=KDD77191.1;Note=similar to PFAM: PF01490
AYPS01000001.1  Genbank CDS 8650    8758    .   -   2   ID=cds2;Parent=rna2;Dbxref=NCBI_GP:KDD77191.1;Name=KDD77191.1;Note=similar to PFAM: PF01490
AYPS01000001.1  Genbank CDS 8295    8529    .   -   1   ID=cds2;Parent=rna2;Dbxref=NCBI_GP:KDD77191.1;Name=KDD77191.1;Note=similar to PFAM: PF01490
AYPS01000001.1  Genbank CDS 7683    8210    .   -   0   ID=cds2;Parent=rna2;Dbxref=NCBI_GP:KDD77191.1;Name=KDD77191.1;Note=similar to PFAM: PF01490
AYPS01000001.1  Genbank CDS 7442    7595    .   -   0   ID=cds2;Parent=rna2;Dbxref=NCBI_GP:KDD77191.1;Name=KDD77191.1;Note=similar to PFAM: PF01490
AYPS01000001.1  Genbank CDS 7169    7359    .   -   2   ID=cds2;Parent=rna2;Dbxref=NCBI_GP:KDD77191.1;Name=KDD77191.1;Note=similar to PFAM: PF01490
AYPS01000001.1  Genbank gene    9601    11866   .   -   .   ID=gene3;Name=H632_c1p3;end_range=11866,.;gbkey=Gene;gene_biotype=protein_coding;locus_
AYPS01000001.1  Genbank mRNA    9601    11866   .   -   .   ID=rna3;Parent=gene3;end_range=11866,.;gbkey=mRNA;orig_protein_id=gnl|ubcbot|H632_c1p3;
AYPS01000001.1  Genbank exon    11609   11866   .   -   .   ID=id16;Parent=rna3;end_range=11866,.;gbkey=mRNA;orig_protein_id=gnl|ubcbot|H632_c1p3;c
AYPS01000001.1  Genbank exon    11351   11509   .   -   .   ID=id17;Parent=rna3;gbkey=mRNA;orig_protein_id=gnl|ubcbot|H632_c1p3;orig_transcript_id=
AYPS01000001.1  Genbank exon    11126   11261   .   -   .   ID=id18;Parent=rna3;gbkey=mRNA;orig_protein_id=gnl|ubcbot|H632_c1p3;orig_transcript_id=
AYPS01000001.1  Genbank exon    10948   11046   .   -   .   ID=id19;Parent=rna3;gbkey=mRNA;orig_protein_id=gnl|ubcbot|H632_c1p3;orig_transcript_id=
AYPS01000001.1  Genbank exon    10340   10800   .   -   .   ID=id20;Parent=rna3;gbkey=mRNA;orig_protein_id=gnl|ubcbot|H632_c1p3;orig_transcript_id=
AYPS01000001.1  Genbank exon    10117   10254   .   -   .   ID=id21;Parent=rna3;gbkey=mRNA;orig_protein_id=gnl|ubcbot|H632_c1p3;orig_transcript_id=
AYPS01000001.1  Genbank exon    9844    10031   .   -   .   ID=id22;Parent=rna3;gbkey=mRNA;orig_protein_id=gnl|ubcbot|H632_c1p3;orig_transcript_id=
AYPS01000001.1  Genbank exon    9601    9748    .   -   .   ID=id23;Parent=rna3;gbkey=mRNA;orig_protein_id=gnl|ubcbot|H632_c1p3;orig_transcript_id=
AYPS01000001.1  Genbank CDS 11609   11866   .   -   0   ID=cds3;Parent=rna3;Dbxref=NCBI_GP:KDD77192.1;Name=KDD77192.1;gbkey=CDS;orig_transcript_id=
AYPS01000001.1  Genbank CDS 11351   11509   .   -   0   ID=cds3;Parent=rna3;Dbxref=NCBI_GP:KDD77192.1;Name=KDD77192.1;gbkey=CDS;orig_transcript_id=
AYPS01000001.1  Genbank CDS 11126   11261   .   -   0   ID=cds3;Parent=rna3;Dbxref=NCBI_GP:KDD77192.1;Name=KDD77192.1;gbkey=CDS;orig_transcript_id=
AYPS01000001.1  Genbank CDS 10948   11046   .   -   2   ID=cds3;Parent=rna3;Dbxref=NCBI_GP:KDD77192.1;Name=KDD77192.1;gbkey=CDS;orig_transcript_id=
AYPS01000001.1  Genbank CDS 10340   10800   .   -   2   ID=cds3;Parent=rna3;Dbxref=NCBI_GP:KDD77192.1;Name=KDD77192.1;gbkey=CDS;orig_transcript_id=
AYPS01000001.1  Genbank CDS 10117   10254   .   -   0   ID=cds3;Parent=rna3;Dbxref=NCBI_GP:KDD77192.1;Name=KDD77192.1;gbkey=CDS;orig_transcript_id=
AYPS01000001.1  Genbank CDS 9844    10031   .   -   0   ID=cds3;Parent=rna3;Dbxref=NCBI_GP:KDD77192.1;Name=KDD77192.1;gbkey=CDS;orig_transcript_id=
AYPS01000001.1  Genbank CDS 9601    9748    .   -   1   ID=cds3;Parent=rna3;Dbxref=NCBI_GP:KDD77192.1;Name=KDD77192.1;gbkey=CDS;orig_transcript_id=
AYPS01000001.1  Genbank gene    12235   12901   .   -   .   ID=gene4;Name=H632_c1p4;Note=similar to PFAM: PF01105.19%2C 5.6E-037;end_range=12901,.;
AYPS01000001.1  Genbank mRNA    12235   12901   .   -   .   ID=rna4;Parent=gene4;Note=similar to PFAM: PF01105.19%2C 5.6E-037;end range=12901,.;gbk
```

The TBL structure

Start with >Feature tag

Each CDS, rRNA and tRNA feature also require a gene feature.

< &> signs signify that the feature is partial: i.e. it is either incomplete or its boundaries are not certain.

Values are always $start $end. Strandedness is decided comparing the two values.

Does not contain sequences. Sequences are in separate Fasta files.

```
>Feature KC121006.fsa
<413    >2322   gene
                gene    rnl
<413    >2322   rRNA
                gene    rnl
                product large subunit ribosomal RNA
                note    highly divergent, junctions unclear
3629    4531    gene
                gene    cox3
3629    4531    CDS
                gene    cox3
                product cytochrome oxidase subunit 3

10964   11034   gene
                gene    trnR(ccu)
10964   11034   tRNA
                gene    trnR(ccu)
                note    tRNA Type: Arg, Anti Codon: CCT, Cove Score: 63.69
                product tRNA-Arg

11119   11048   gene
                gene    trnA(ugc)
11119   11048   tRNA
                gene    trnA(ugc)
                note    tRNA Type: Ala, Anti Codon: TGC, Cove Score: 56.64
                product tRNA-Ala
```

\t (Tab delimited)

\t (3times)

Redundancy (not mandatory)

Reverse complement

The ASN (SQN) structure

Not intended to be human readable

Programming-like; curly-braced

```
Seq-submit ::= {
  sub {
    contact {
      contact {
        name
          name {
            last "Lemieux" ,
            first "Claude" ,
            initials "C." } ,
        affil
          std {
            affil "Universite Laval" ,
            div "Biochimie et Microbiologie" ,
            city "Quebec" ,
            sub "QC" ,
            country "Canada" ,
            street "Pavillon Charles-Eugene Marchand" ,
            email "claude.lemieux@rsvs.ulaval.ca" ,
            fax "(418) 656-7176" ,
            phone "(418) 656-2131 x5171" ,
            postal-code "G1K 7P4" } } } ,
    cit {
      authors {
        names
          std {
            {
              name
                name {
                  last "Pombert" ,
                  first "Jean-Francois" ,
                  initials "J.-F." } } ,
            {
              name
                name {
                  last "Lemieux" ,
                  first "Claude" ,
                  initials "C." } } ,
            {
              name
                name {
                  last "Turmel" ,
                  first "Monique" ,
                  initials "M." } } } ,
        affil
          std {
            affil "Universite Laval" ,
            div "Biochimie et Microbiologie" ,
            city "Quebec" ,
            sub "QC" ,
            country "Canada" ,
            street "Pavillon Charles-Eugene Marchand" ,
            postal-code "G1K 7P4" } } ,
      date
        std {
          year 2005 ,
          month 11 ,
          day 15 } } ,
    hup TRUE ,
```

The EMBL structure

```
ID   DQ365900; SV 1; circular; genomic DNA; STD; PLN; 56761 BP.
AC   DQ365900;
DT   26-JUL-2006 (Rel. 88, Created)
DT   11-OCT-2012 (Rel. 114, Last updated, Version 4)
DE   Oltmannsiellopsis viridis mitochondrion, complete genome.
KW   .
OS   Oltmannsiellopsis viridis
OC   Eukaryota; Viridiplantae; Chlorophyta; Oltmannsiellopsis.
OG   Mitochondrion
RN   [1]
RP   1-56761
RX   DOI; 10.1007/s00294-006-0076-z.
RX   PUBMED; 16721603.
RA   Pombert J.F., Beauchamp P., Otis C., Lemieux C., Turmel M.;
RT   "The complete mitochondrial DNA sequence of the green alga
RT   Oltmannsiellopsis viridis: evolutionary trends of the mitochondrial genome
RT   in the Ulvophyceae";
RL   Curr. Genet. 50(2):137-147(2006).
RN   [2]
RP   1-56761
RA   Pombert J.-F., Beauchamp P., Otis C., Lemieux C., Turmel M.;
RT   ;
RL   Submitted (17-JAN-2006) to the INSDC.
RL   Biochimie et Microbiologie, Universite Laval, Pavillon Charles-Eugene
RL   Marchand, Quebec, QC G1K 7P4, Canada
DR   RFAM; RF00005; tRNA.
DR   RFAM; RF00029; Intron_gpII.
DR   RFAM; RF00177; SSU_rRNA_bacteria.
DR   RFAM; RF01118; PK-G12rRNA.
DR   SILVA-LSU; DQ365900.
FH   Key             Location/Qualifiers
FT   source          1..56761
FT                   /organism="Oltmannsiellopsis viridis"
FT                   /organelle="mitochondrion"
FT                   /mol_type="genomic DNA"
FT                   /cell_line="NIES 360"
FT                   /db_xref="taxon:51324"
FT   gene            complement(12603..12986)
FT                   /gene="rps12"
FT   CDS             complement(12603..12986)
FT                   /codon_start=1
FT                   /gene="rps12"
FT                   /product="ribosomal protein S12"
FT                   /db_xref="GOA:Q0QIR5"
FT                   /db_xref="InterPro:IPR005679"
FT                   /db_xref="UniProtKB/TrEMBL:Q0QIR5"
FT                   /protein_id="ABC96339.1"
FT                   /translation="MVTKNQLLRKSTKRVKKVKKNKVPALANHGPFRRGTCVRVFTRTP
FT                   KKPNSALRKVAKVRLSNKTTVIVAIPGEGHNLKEHAIILFRGGRVRDLPGVKYKAVRGV
FT                   LDLAGVQKRKTARSKYGRRRDDV"SQ   Sequence 56761 BP; 18397 A; 9537 C;
9418 G; 19409 T; 0 other;
     taaaataata atgaaaagtg ctattaaata agtagtcttg ctttcaaaaa atatctacaa        60
     aatgatacgg caacgctact taatgttcca ttaagtagtt ttgtagcttc gctcgtagat       120
```

Exercise 02 – DFAST (A Python pipeline)

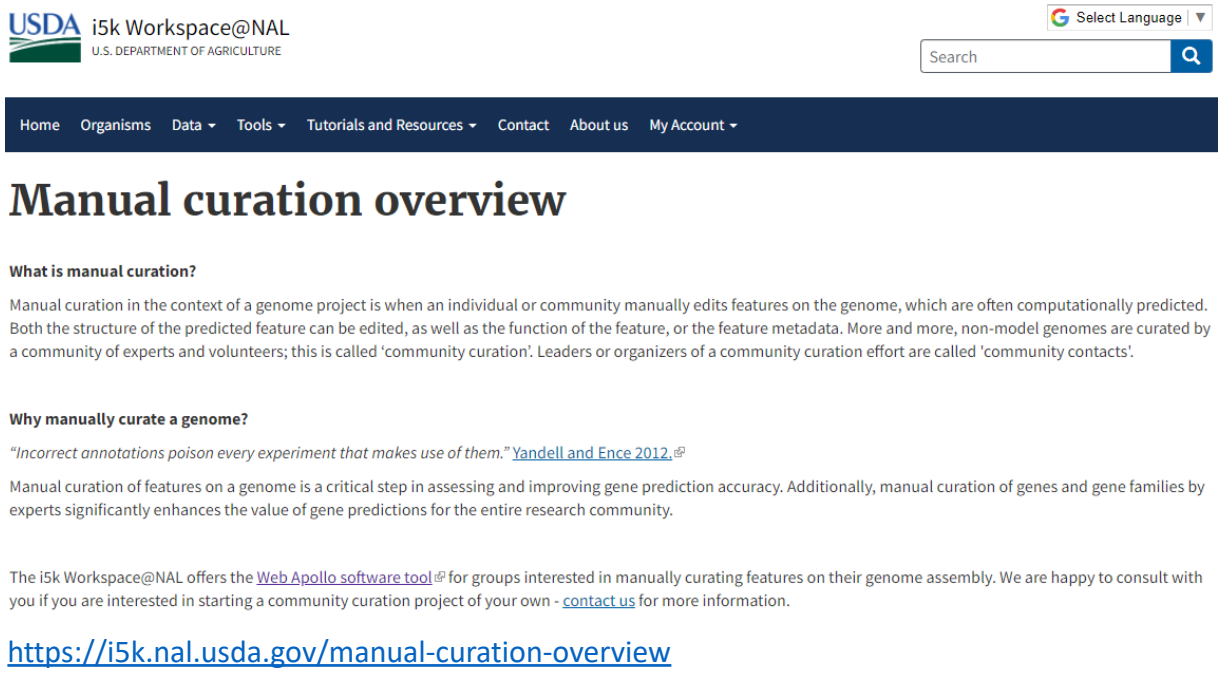
Same genome, different pipeline

1) Launch the command in a screen ## That way you can detach if needed
2) Annotate the *E. coli* genome (Ecoli_K12.fasta) with DFAST:
   dfast -g ./Ecoli_K12.fasta -o DFAST --strain K12 --cpu 4
3) Look inside the output folder ## DFAST gives you multiple formats to facilitate editing
4) Compare the DFAST annotation with the original/Prokka ones:
   grep -c '/locus_tag=' Ecoli_K12.gb ../Exercise_01/PROKKA_*/PROKKA_*.gbk \
   DFAST/genome.gbk

DFAST uses MetaGeneAnnotator by default instead of Prodigal as protein predictor
but that is not the cause of this huge difference, can you explain it? Hint: look
inside the files

Optional

Additional slides for the curious



https://i5k.nal.usda.gov/manual-curation-overview

A2GB – Annotations to GenBank

A2GB is a pipeline that will transform the genome annotation files exported from Apollo (fomerly known as WebApollo) in preparation for sequence submission to NCBI's GenBank. While its primary goal is to submit annotations to GenBank for the generation of accession numbers, the conversion of file formats will permit the use of different tools in downstream analyses.

In a stepwise approach, the A2GB pipeline converts annotation files from GFF3 -> EMBL -> TBL -> ASN. Each format will be useful for diagnostic quality checks of the annotations or become the springboard for other analyses, such as protein function prediction.

Furthermore, A2GB acts as a guide to prepare sequence submissions according to NCBI's guidelines, including project registration with BioSample and BioProject for the generation of locus_tag prefixes. Upon acceptance from NCBI, these essential steps will ensure that sequence data will be made publicly available through GenBank and other member databases within the International Nucleotide Sequence Database Collaboration.

https://github.com/PombertLab/A2GB

Expert curation

Requires expert knowledge and complementary information (*e.g.* RNAseq, comparative analyses)

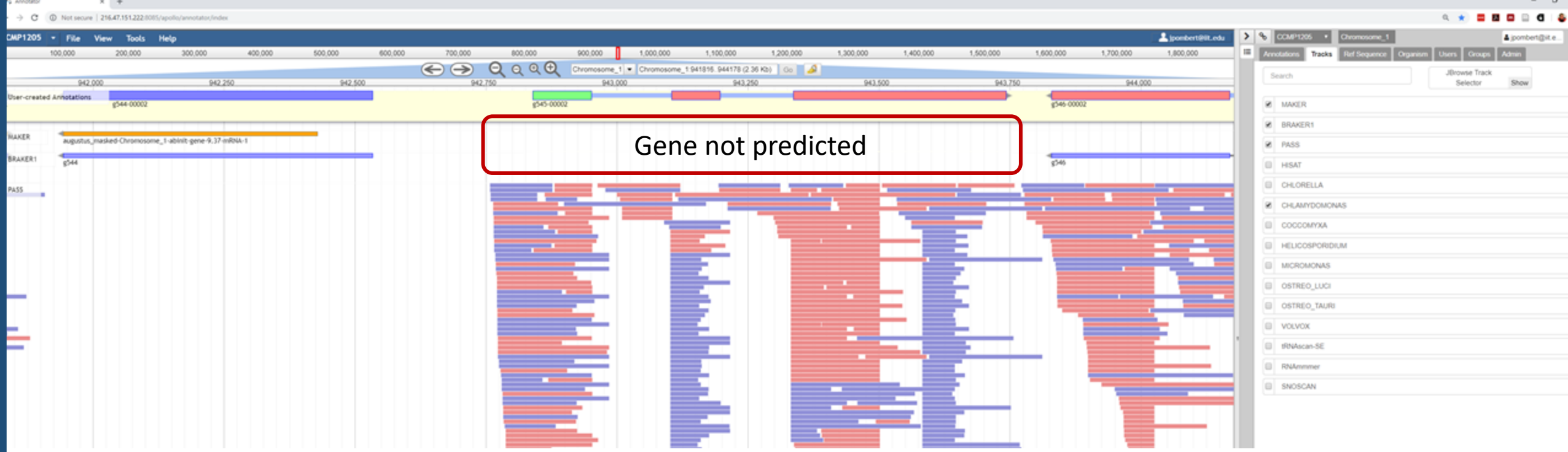
Especially important with complex genomes

e.g. human genome

Time consuming

Often a manual process

For an example of the process, see A2GB https://github.com/PombertLab/A2GB



Problems to look for:

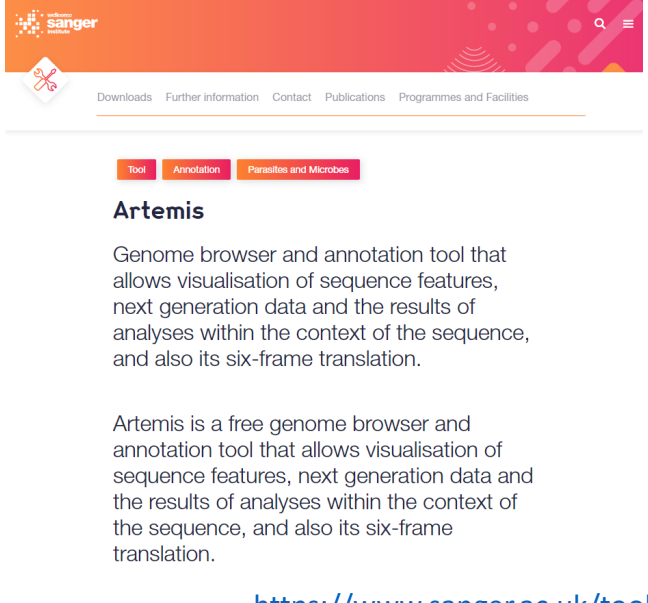
Missing genes
Missing start/stop codons
Exon/intron junction issues
Pseudogenes/frameshifts

gene prediction

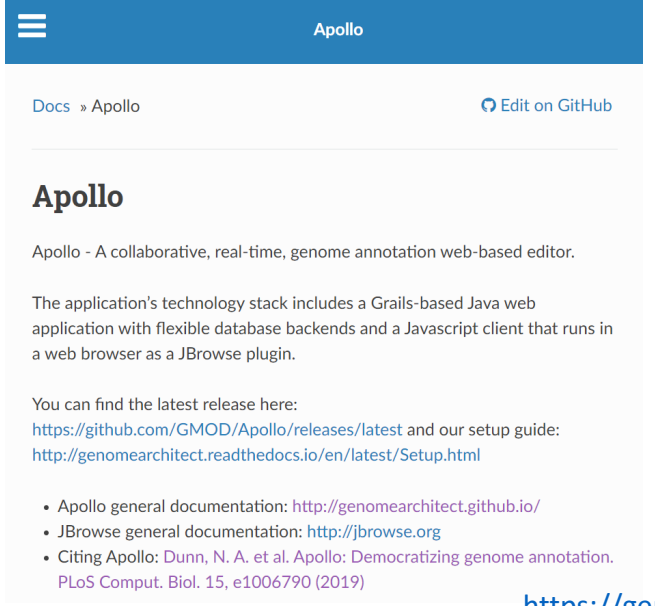
Wrong functions

function prediction

Machine learning can miss genes even with good models
Incomplete predictions?
GT/AG rule in eukaryotes
Are those real or due to sequencing/assembly errors?
Were the functions predicted accurate?



https://www.sanger.ac.uk/tool/artemis/



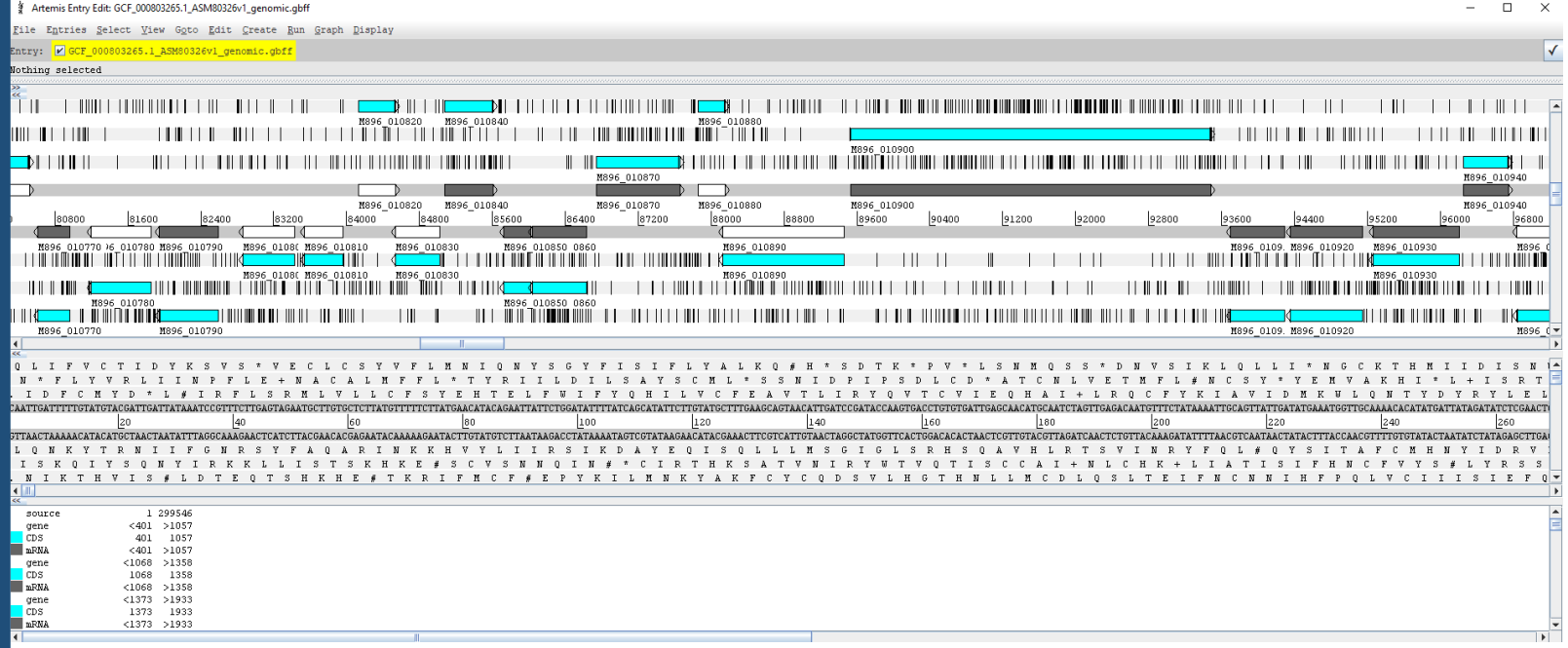
https://genomearchitect.readthedocs.io/en/latest/

Genome browsers/annotation editors

Expert curation requires proper tools

Best genome browsers/annotators are:

Artemis (2000)	Local	## Works best on single contigs/chromosomes, great at fixing granular issues
Apollo (2013)	Web-based	## Database-like, collaborative, great at working with multiple evidence sources
		## Formerly known as WebApollo



Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data
Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA
Bioinformatics. 2012. 28(4):464-469

Artemis: sequence visualization and annotation
Rutherford K, Parkhill J, Crook J,
Horsnell T, Rice P, Rajandream MA, Barrell B
Bioinformatics. 2000. 16(10):944-945

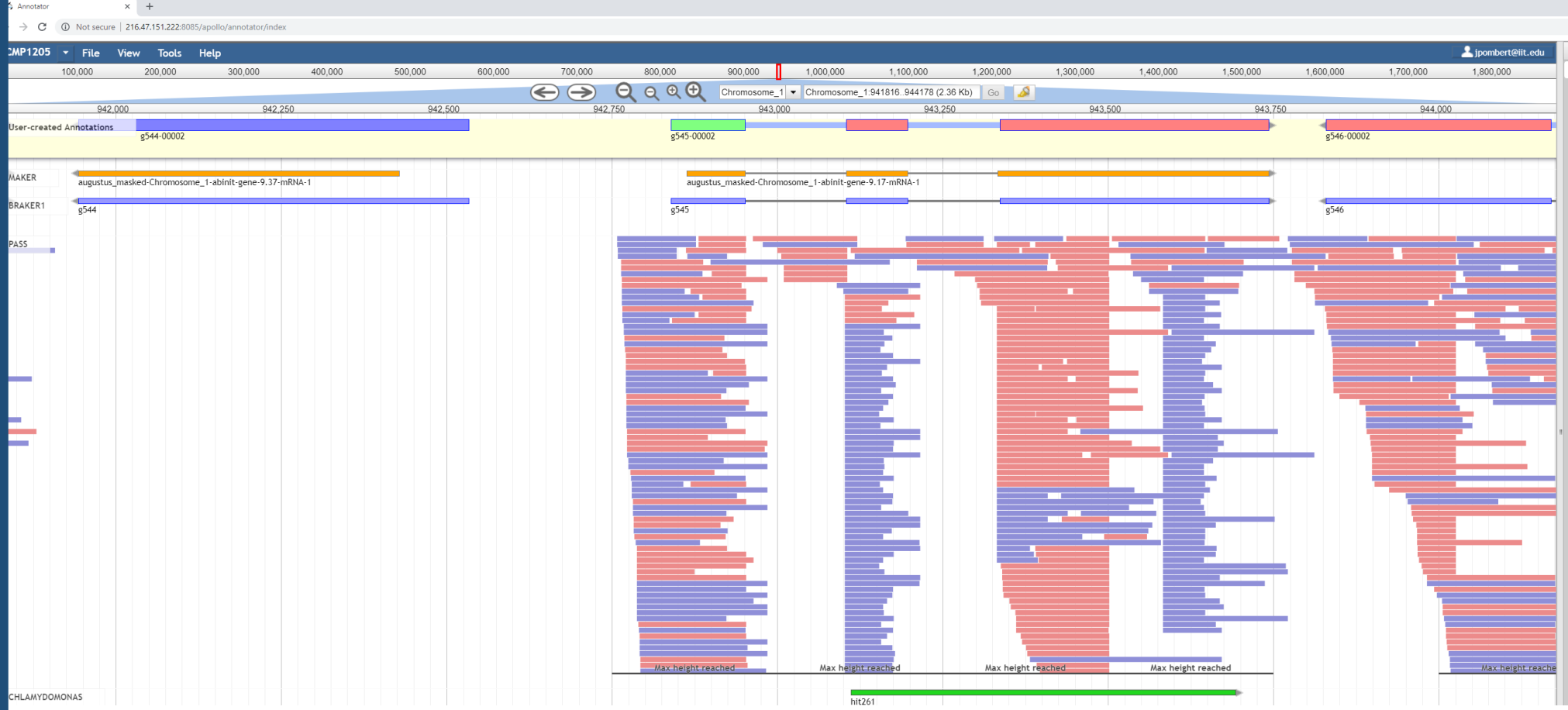
Artemis – Annotation swiss knife

https://www.sanger.ac.uk/tool/artemis/

Free, portable Java application (Linux, OSX & Windows)

Can be used to annotate, curate and/or browse genome

Developed by EMBL + Sanger Institute



Apollo – Collaborative annotation

Great for collaborative work

Built for manual curation

Can combine multiple datasets/analyses in a very visual format

Apollo: Democratizing genome annotation

Dunn NA, Unni DR, Diesh C, Munoz-Torres M, Harris NL, Yao E, Rasche H, Holmes IH, Elsik CG, Lewis SE

PLoS Comput Biol. 2019 Feb 6;15(2):e1006790

DOI: 10.1371/journal.pcbi.1006790

```
#!/usr/bin/bash
## On Spartacus
chown -R tomcat:jpombert /media/Data_1/apollo/
chown -R jpombert:jpombert /media/Data_1/apollo/data
mkdir /media/Data_1/apollo/data
mkdir /media/Data_1/apollo/data/E_hellem_50604

## Loading Fasta
/home/jpombert/Downloads/Apollo-2.4.1/web-app/jbrowse/bin/prepare-refseqs.pl \
--fasta /media/Data_4/jpombert/Microsporidia/E_hellem_50604/WebApollo/E_hellem_basecalled_v3_FINAL.fasta \
--out /media/Data_1/apollo/data/E_hellem_50604

# http://216.47.151.222:8085/apollo/annotator/index
```

```
/home/jpombert/Downloads/Apollo-2.4.1/docs/web_services/examples/groovy/add_organism.groovy \
-name E_hellem_50604 \
-url http://localhost:8085/apollo/ \
-directory /media/Data_1/apollo/data/E_hellem_50604 \
-username jpombert@iit.edu \
-password 'xxx'

## Prodigal
prodigal -c -f gff -i ../E_hellem_basecalled_v3_FINAL.fasta -o E_hellem_basecalled_v3_FINAL.gff
../../../scripts/splitGFF3.pl E_hellem_basecalled_v3_FINAL.gff

for k in {01..11}; do /home/jpombert/Downloads/Apollo-2.4.1/jbrowse/bin/flatfile-to-json.pl \
--gff /media/Data_4/jpombert/Microsporidia/E_hellem_50604/WebApollo/prodigal/chromosome_$k.gff3 \
--type CDS --subfeatureClasses '{"CDS": "orange-80pct"}' \
--trackLabel PRODIGAL \
--out /media/Data_1/apollo/data/E_hellem_50604; done

## TBLASTN
makeblastdb -in ../../E_hellem_basecalled_v3_FINAL.fasta -dbtype nucl -out Ehel_50604
tblastn -num_threads 10 -query Eint_50506_GCF_000146465.1_ASM14646v1_protein.faa -db DB/Ehel_50604 -outfmt 6 -out Eint_50506_GCF_000146465.1_ASM14646v1_protein.tblastn.6
tblastn -num_threads 10 -query Ehel_50504_GCF_000277815.2_ASM27781v3_protein.faa -db DB/Ehel_50604 -outfmt 6 -out Ehel_50504_GCF_000277815.2_ASM27781v3_protein.tblastn.6
tblastn -num_threads 10 -query Erom_SJ2008_GCF_000280035.1_ASM28003v2_protein.faa -db DB/Ehel_50604  -outfmt 6 -out Erom_SJ2008_GCF_000280035.1_ASM28003v2_protein.tblastn.6

## Loading TBLASTN
../../../scripts/getProducts.pl *.faa
../../../scripts/TBLASTN_to_GFF3.pl *.tblastn.6
/home/jpombert/Downloads/Apollo-2.4.1/jbrowse/bin/flatfile-to-json.pl --gff /media/Data_4/jpombert/Microsporidia/E_hellem_50604/WebApollo/BLAST/Eint_50506_GCF_000146465.1_ASM14646v1_protein.gff \
--type match,match_part --subfeatureClasses '{"match_part": "orange-80pct"}' --trackLabel Eint_50506_tblastn --out /media/Data_1/apollo/data/E_hellem_50604
/home/jpombert/Downloads/Apollo-2.4.1/jbrowse/bin/flatfile-to-json.pl --gff /media/Data_4/jpombert/Microsporidia/E_hellem_50604/WebApollo/BLAST/Ehel_50504_GCF_000277815.2_ASM27781v3_protein.gff \
--type match,match_part --subfeatureClasses '{"match_part": "green-80pct"}' --trackLabel Ehel_50504_tblastn --out /media/Data_1/apollo/data/E_hellem_50604
/home/jpombert/Downloads/Apollo-2.4.1/jbrowse/bin/flatfile-to-json.pl --gff /media/Data_4/jpombert/Microsporidia/E_hellem_50604/WebApollo/BLAST/Erom_SJ2008_GCF_000280035.1_ASM28003v2_protein.gff \
--type match,match_part --subfeatureClasses '{"match_part": "blue-80pct"}' --trackLabel Erom_SJ2008_tblastn --out /media/Data_1/apollo/data/E_hellem_50604

## tRNAscan-2.0
../../../scripts/run_tRNAscan.pl ../E_hellem_basecalled_v3_FINAL.fasta
../../../scripts/tRNAscan_to_GFF3.pl *.tRNAs
/home/jpombert/Downloads/Apollo-2.4.1/jbrowse/bin/flatfile-to-json.pl \
--gff /media/Data_4/jpombert/Microsporidia/E_hellem_50604/WebApollo/tRNAscan/E_hellem_basecalled_v3_FINAL.fasta.tRNAs.gff \
--type tRNA \
--trackLabel tRNAscan-SE \
```

Backend is Perl, Tomcat + SQL-like databases

Bash proficiency is required to load data

Exercise 03 – Artemis (prokaryote)

Will be slow on Mozart over SSH -X (art), better to use it on your laptop

1) Open the NCBI annotation:
   art Ecoli_K12.gb & ## The & allows you to reuse the same shell after launching graphical user interface (GUI) tools from the command line
2) Open the Prokka annotation:
   art PROKKA_10112018.gbk &
3) Open the DFAST annotation:
   art genome.gbk &
4) Briefly compare them. Notice the lack of white boxes in the DFAST annotations? You can see the content of a feature by clicking on it and using ctrl+e to edit the feature

Exercise 04 – Artemis (eukaryote)

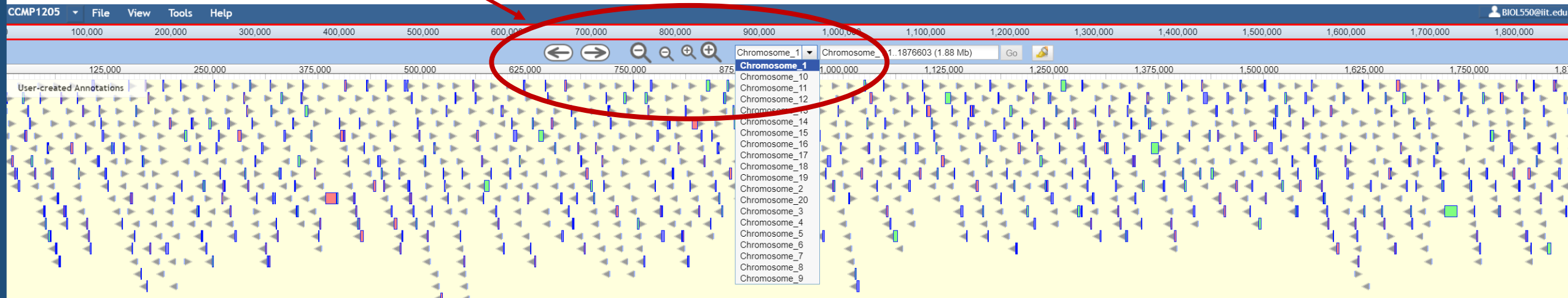
Eukaryote genomes are more complex, this one is relatively simple

1) Open the NCBI annotation: art Chromosome_01.gbk &
2) Look at the different feature types:
   ## CDS, mRNA and gene features are required for eukaryote genomes
   ## deposited in GenBank; exons and introns are optional
3) Double click on CDS feature A3770_01p00020
   Click on the first intron. Look at the ^GT.*AG$ structure ## This structure is typical of spliceosomal introns
4) This locus has 3 exons: join(2873..4104,4189..4329,4491..4731)
   Each annotation format uses a different way to write that information.
   ## A good way to interconvert formats with Perl is to use a hash of arrays
   ## where the hash keys are locus_tags and the arrays are the lists of exons

Exercise 05a – Apollo (manual curation)

The Ex 04 annotation was curated with Apollo before conversion to GenBank

1) Using a web browser, go to http://216.47.151.222:8085/apollo/annotator/index
2) Enter the following username and password:
   username: BIOL550@iit.edu
   password: Coll4bor8_G3nome
3) Use the scrolling menu to select chromosome 1; use the magnifying glass icons to zoom/in or out



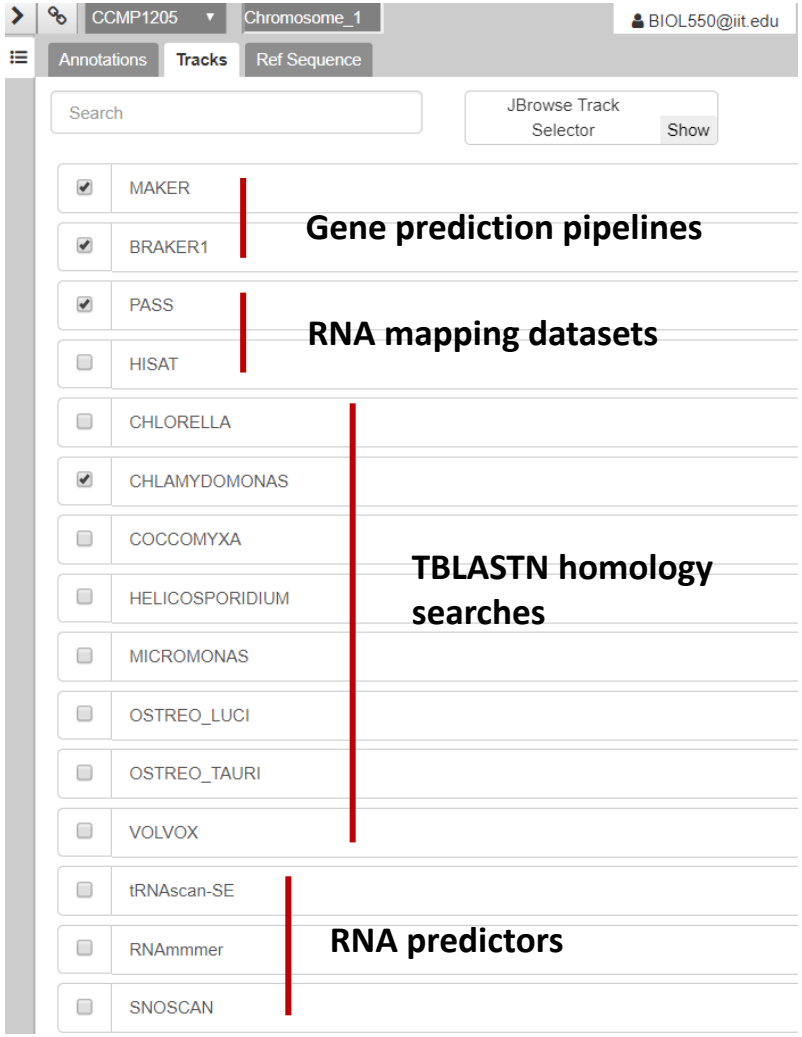
Exercise 05b – Apollo (manual curation)

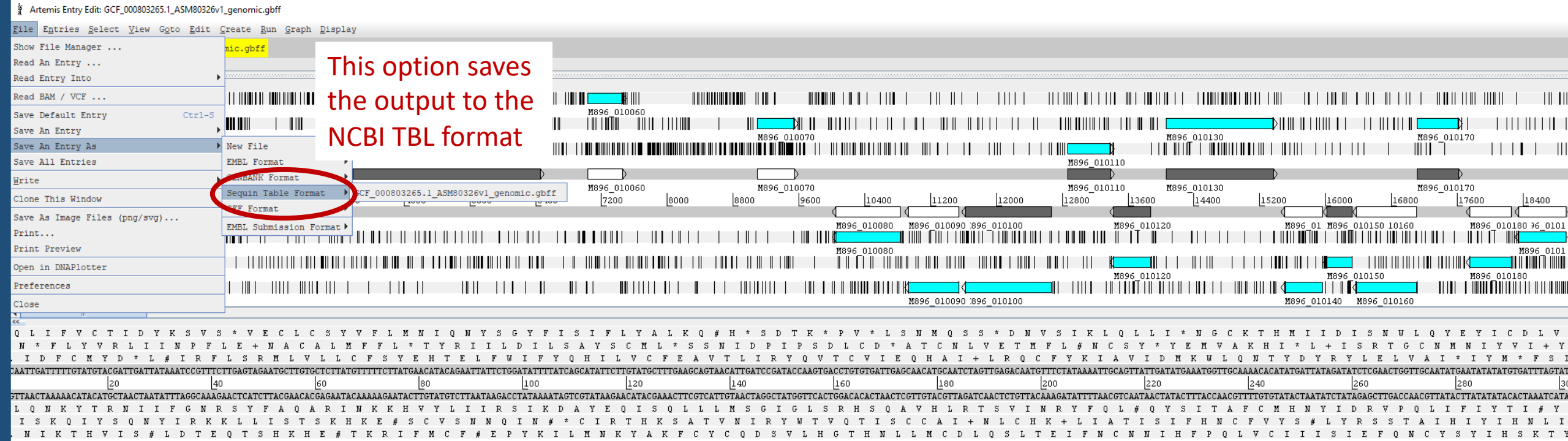
4) In menu to the right click the Tracks tab
   ## We can select/hide the evidence/analyses
   ## track to display

5) Select:
   MAKER
   BRAKER1
   PASS
   CHLAMYDOMONAS

6) Look at the different information displayed
   ## We'll talk about it in the class

7) We can export data using the Ref Sequence tab
   GFF3 -> Genome File Format 3
   VCF -> Variant Calling Format
   FASTA -> Genomic, cDNA, CDS or peptide





Artemis/Apollo outputs need reformatting

Artemis/Apollo default format: EMBL/GFF3

GenBank required format: ASN (SQN)

NCBI provides a tool called TBL2ASN that converts from TBL to ASN format

We can save to TBL format from Artemis (a bit tedious if many contigs/chromosomes)

We can also convert from EMBL/GFF3 to TBL format using scripts

Conversion to TBL

```
if ($format eq 'Artemis') {print "EMBL => TBL";}
if ($format eq 'Apollo') {print "GFF3 => TBL";}
Perl is GREAT for these conversions
# Prokka and DFAST also generate TBL files when
# performing automatic annotations
```

```
>Feature KC121006.fsa
<413    >2322   gene
                gene    rnl
<413    >2322   rRNA
                gene    rnl
                product large subunit ribosomal RNA
                note    highly divergent, junctions unclear

3629    4531    gene
                gene    cox3
3629    4531    CDS
                gene    cox3
                product cytochrome oxidase subunit 3

10964   11034   gene
                gene    trnR(ccu)
10964   11034   tRNA
                gene    trnR(ccu)
                note    tRNA Type: Arg, Anti Codon: CCT, Cove Score: 63.69
                product tRNA-Arg

11119   11048   gene
                gene    trnA(ugc)
11119   11048   tRNA
                gene    trnA(ugc)
                note    tRNA Type: Ala, Anti Codon: TGC, Cove Score: 56.64
                product tRNA-Ala
```

Exercise 06 – EMBL to TBL

Let's practice Perl a bit using a very simple EMBL file as input

1) Create a single Perl script named my_converter.pl that will take one or more EMBL files and convert them to TBL format:
   a. The input file is K12_short.embl
   b. The desired output file is K12_short_desired_output.tbl
2) To test, make sure to convert all files to Unix format
   dos2unix *
3) Run your script on the .embl file(s)
   ./my_converter.pl *.embl
4) Look for potential discrepancies between your output and the desired one
   diff K12_short.tbl K12_short_desired_output.tbl

Reality is a bit more complex

Introns?

i.e. code must allow multiple exons if introns

Stop codon/start methionine?

Predicted proteins can be incomplete

Phased open reading frame?

Reading frame starts at position 0, 1 or 2?

Strandedness?

In the previous exercise, for simplification, all
genes were on the same strand

External list of locus_tags/products?

Easier for maintenance and easy to implement;
hash of products

https://github.com/PombertLab/A2GB/blob/master/EMBLtoTBL.pl

C:\Users\jfpombert\Desktop\EMBLtoTBL.pl - Notepad++

File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?

EMBLtoTBL.pl

```
#!/usr/bin/perl
## Pombert Lab, IIT, 2020
my $name = 'EMBLtoTBL.pl';
my $version = '1.5a';

use strict; use warnings; use Bio::SeqIO; use File::Basename; use Getopt::Long qw(GetOptions);

my $usage = <<"OPTIONS";
die "$usage\n" unless @ARGV;

my $instID = 'IITBIO'; ##
my $products; ## protein_list.txt
my @embl;
GetOptions(
    'id=s' => \$instID,
    'p=s' => \$products,
    'embl=s@{1,}' => \@embl,
);


sub numSort {if ($a < $b){return -1;} elsif ($a == $b){return 0;} elsif ($a > $b) {return 1;}}

### Filling the products database
my %hash = ();
open HASH, "<$products";
while(my $dbkey = <HASH>){chomp $dbkey;if($dbkey =~ /^(\S+)\t(.*)$/){my $prot = $1;my $prod = $2;$hash{$prot}=$prod;}}

### Working on EMBL files
my $locus_tag;
while(my $file = shift@embl){
    open IN, "<", "$file" or die "Can't open EMBL file file: $file\n";
    $file =~ s/.embl$//;
    my ($head, $dir) = fileparse($file);
    open DNA, "<", "$file.fsa" or die "Can't open FASTA file file: $file.fsa\n";
    open TBL, ">", "$file.tbl" or die "Can't create TBL output file: $file.tbl\n";
    print TBL ">Feature $head\n"; ## Generate TBL header

    ### Creating a single DNA string for codon verification
    my $DNAseq = undef;
    while (my $dna = <DNA>){chomp $dna;if ($dna =~ /^>/){next;}else{$DNAseq.= $dna;}}
    my $DNAsequence = lc($DNAseq); ## Changing to lower case to fit with the codon check
    my $contig_length = length($DNAsequence); ## Calculating the contig size
    $locus_tag = undef;

    while(my $line = <IN>){
        chomp $line;
        my @start = ();
        my @stop = ();
        my $asize = undef;
        my $num = undef;
        my $dum = undef;

        ### Defining the locus tags
        if ($line =~ /FT\s+\/locus_tag="(\S+)"/){$locus_tag = $1;}

        ### Working on tRNAs/rRNAs
        elsif ($line =~ /FT\s+(tRNA|rRNA)\s+(\d+)..(\d+)/){ ## Forward, single exon
            my $type = $1;
            my $start = $2;
            my $stop = $3;
            print TBL "$start\t$stop\tgene\n";
            print TBL "\t\t\tlocus_tag\t$locus_tag\n";
            print TBL "$start\t$stop\t$type\n";
            RNA();
        }
        elsif ($line =~ /FT\s+(tRNA|rRNA)\s+join\((.*)\)/){ ## Forward, multiple exons
            my $feat = $1;
            my @array = split(',',$2);
```

What is table2asn?

table2asn is a command-line program that creates sequence records for submission to GenBank. It uses many of the same functions as Genome Workbench but is driven generally by data files, and the records it produces do not necessarily require additional manual editing before submission to GenBank.

6 types of input data files

1. Template file containing a text ASN.1 Submit-block object (suffix .sbt). [Required]
2. Nucleotide sequence data in FASTA format (suffix .fsa). [Required]
3. 5-column Feature Table (suffix .tbl). [Required only if including annotation in this format]
4. Protein sequence (suffix .pep). [Optional; these are rarely needed.]
5. Quality Scores (suffix .qvl.) [Optional]
6. Source Table (suffix .src.) [Optional]

tbl2asn is now obsolete and has
been replaced by table2asn

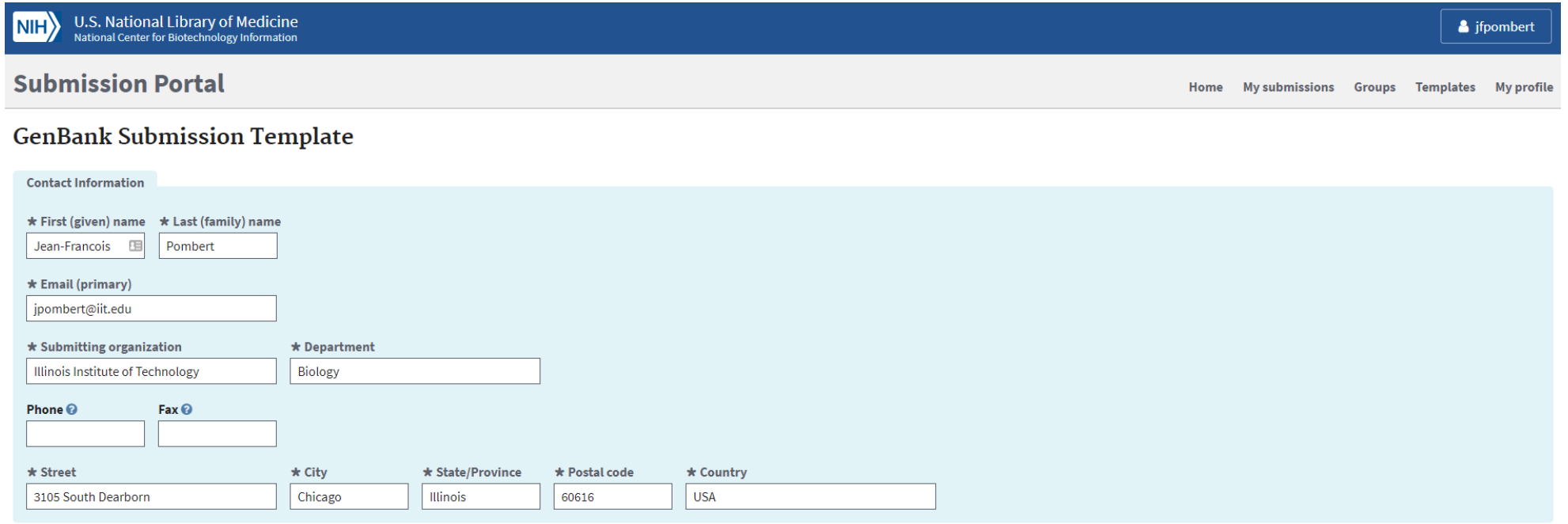
table2asn – Converts to ASN (SQN)

https://www.ncbi.nlm.nih.gov/genbank/table2asn/

The NCBI ASN (SQN) format required for depositing sequences ne GBK!

```
table2asn -t template.sbt -indir Files/ -outdir Outdir/      ## Files must contain FASTA and TBL files
```

The -euk switch is mandatory for eukaryotic genomes   ## table2asn -help

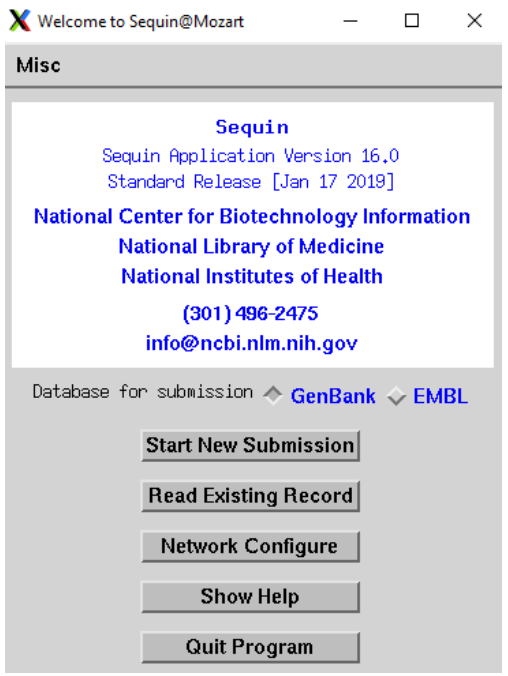


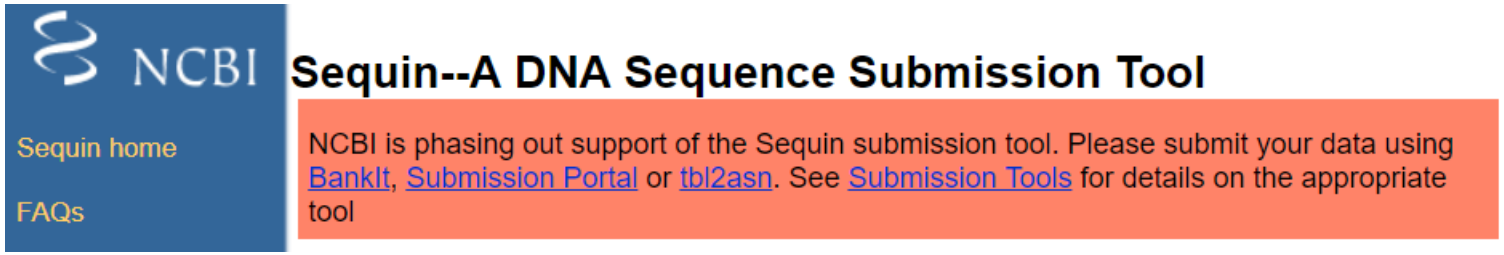
Template.sbt

http://www.ncbi.nlm.nih.gov/WebSub/template.cgi

A header describing the submission files(s). Includes authorship.

Can be generated from Sequin too ## Sequin is being phased out





Genome Workbench Submission Wizard to replace Sequin for prokaryotic and eukaryotic genome submissions in January 2021

https://www.ncbi.nlm.nih.gov/tools/gbench

Sequin – NCBI GUI tool ## Obsolete

https://ncbiinsights.ncbi.nlm.nih.gov/tag/sequin/

Used in the past to generate entries for GenBank submission, but:

- Buggy, laborious manual entries
- Wasn't accepted for large genomes.
- Discontinued. No longer supported by NCBI.

Exercise 07 – Using table2asn

From the CMD line: table2asn -help; www.ncbi.nlm.nih.gov/genbank/table2asn/

1) Let's generate ASN (SQN) files; they will be created inside Files/
   table2asn -t template.sbt -w genome.asm -indir Files -outdir SQN -euk
   ## -euk -> Eukaryote
   ## -w -> Structured comments include information about the assemblies
2) Let's generate and validate SQN files
   table2asn -t template.sbt -w genome.asm -indir Files/ -outdir SQN -euk -M n -V v -Z
   ## -V v -> validates
   ## -Z -> creates a log of potential issues
3) Let's check at the .val files and the log of discrepancies. The .val files in SQN/ should contain an error about lineage. The File.dr (discrepancy report) contain warnings and maybe errors. ## Lineage is an old FASTA identifier no longer used by NCBI

```
curate_annotations.pl -i $ANNOT/proteins.annotations

Putative annotation(s) found for protein #0002: HOP50_01g00020:
1.      SWISSPROT:      2.5e-50         Hybrid signal transduction histidine kinase J
2.      TREMBL:         0.0e+00         Signal transduction histidine kinase
3.      Pfam:           5.5E-30         Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase
4.      TIGRFAM:        NA              hypothetical protein
5.      HAMAP:          NA              hypothetical protein
6.      CDD:            1.8907E-11      HisKA

Please enter selection [1-6] to assign annotation, [0] to annotate as 'hypothetical protein', [m] for manual annotation, or [x] to exit
m
Enter desired annotation: signal transduction histidine kinase

Putative annotation(s) found for protein #0003: HOP50_01g00030:
1.      SWISSPROT:      NA              hypothetical protein
2.      TREMBL:         1.0e-07         Insulin-like growth factor binding, N-terminal
3.      Pfam:           1.0E-6          Putative ephrin-receptor like
4.      TIGRFAM:        NA              hypothetical protein
5.      HAMAP:          NA              hypothetical protein
6.      CDD:            6.31891E-8      TNFRSF

Please enter selection [1-6] to assign annotation, [0] to annotate as 'hypothetical protein', [m] for manual annotation, or [x] to exit
x
```

https://github.com/PombertLab/A2GB/blob/master/Function_prediction/parse_annotators.pl
https://github.com/PombertLab/A2GB/blob/master/Function_prediction/curate_annotations.pl

Are predicted functions congruent?

Function predictors may yield different results

Congruency elevates confidence in your predictions

Manually checking output files is tedious

Writing scripts to aggregate the results in a concise fashion is helpful

You can create databases (*e.g.* hashes) of curated products