Introduction to plant genome annotation

By

Pichahpuk Uthaipaisanwong, Ph. D

8 August 2018

Genome assembly and annotation workshop
at Kasetsart University, Kamphaeng Saen Campus

Content

- Overview of genome annotation
- Annotation of coding regions
  - Gene prediction
    - *ab initio* gene prediction
    - Homology-based
  - Functional gene annotation
- Annotation of non-coding regions
  - tRNA
  - rRNA
- Genome component
  - RepeatMasker
  - misa

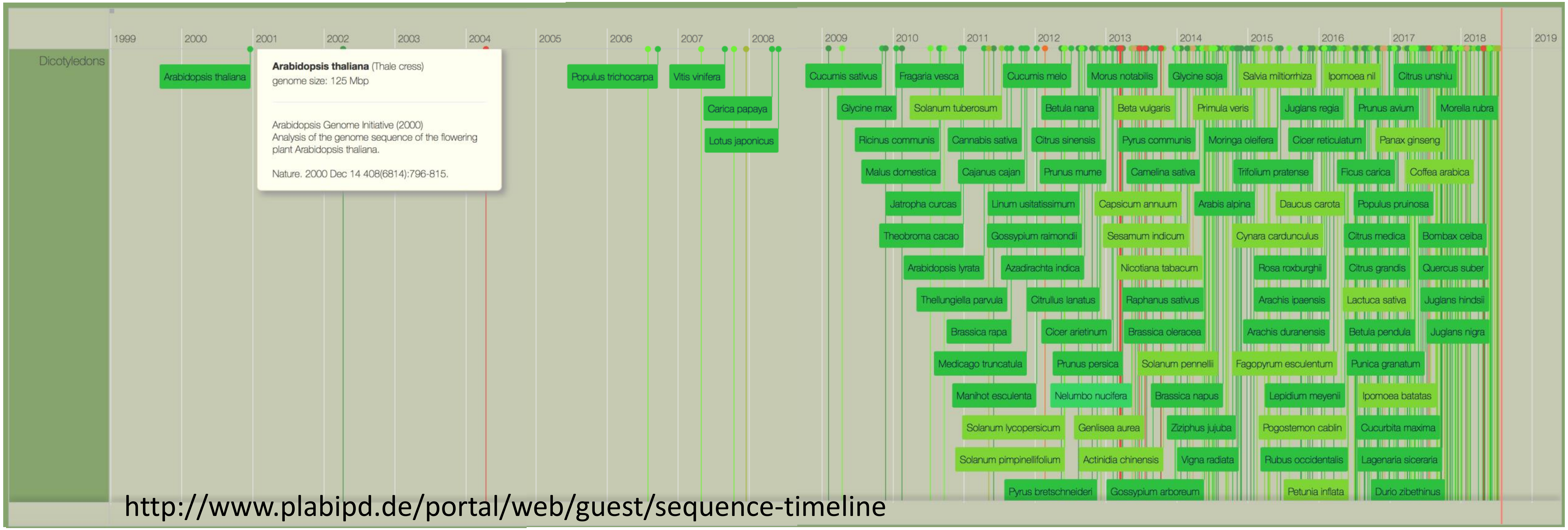
- CAB-Inhouse annotation pipeline (CABAnnot)
- Genome visualization by JBrowse

Plant genome sequencing

~ 180 plant genome sequences in NCBI

~ 450 plant transcriptome assemblies in NCBI

~1300 plant transcriptomes from the plant 1 KP project (onekp.com)



Arabidopsis thaliana (Thale cress)
genome size: 125 Mbp

Arabidopsis Genome Initiative (2000)
Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.

Nature. 2000 Dec 14 408(6814):796-815.

http://www.plabipd.de/portal/web/guest/sequence-timeline

Marie E Bolger et al. (2018) Plant genome and transcriptome annotations: from misconceptions to simple solutions. *Briefings in Bioinformatics*, Volume 19, Issue 3, 1 May 2018, Pages 437–449.

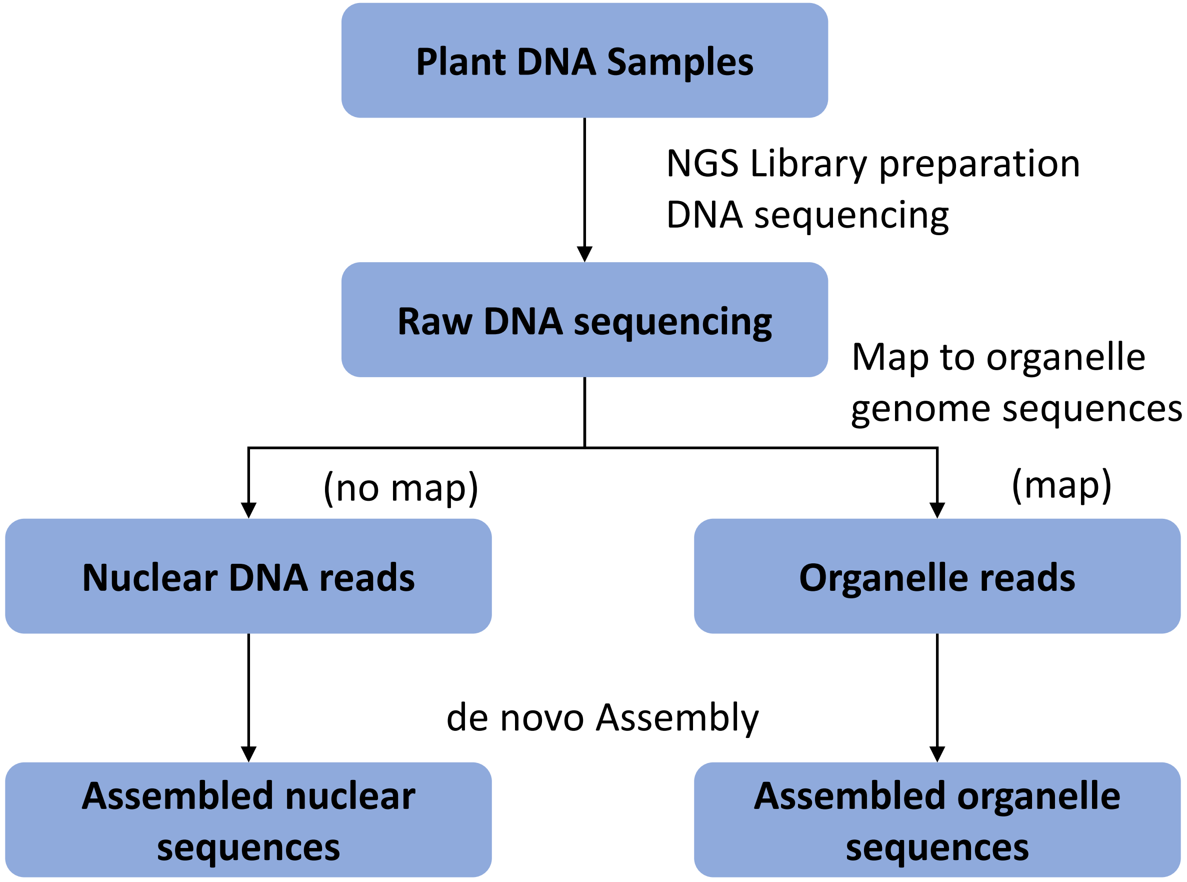
Naim Matasci *et.al.* (2014) Data access for the 1,000 plants (1KP) project. GigaScience, 2014, Volume 3, Page 1

What is genome annotation?

- **Genome annotation** is the process of finding and designating locations of individual genes and other features on raw DNA sequences. (NCBI Knowledge base)

https://support.nlm.nih.gov/knowledgebase/article/KA-03574/en-us

Separate nuclear and organellar sequences



Plant DNA Samples

NGS Library preparation
DNA sequencing

Raw DNA sequencing

Map to organelle genome sequences

(no map)

(map)

Nuclear DNA reads

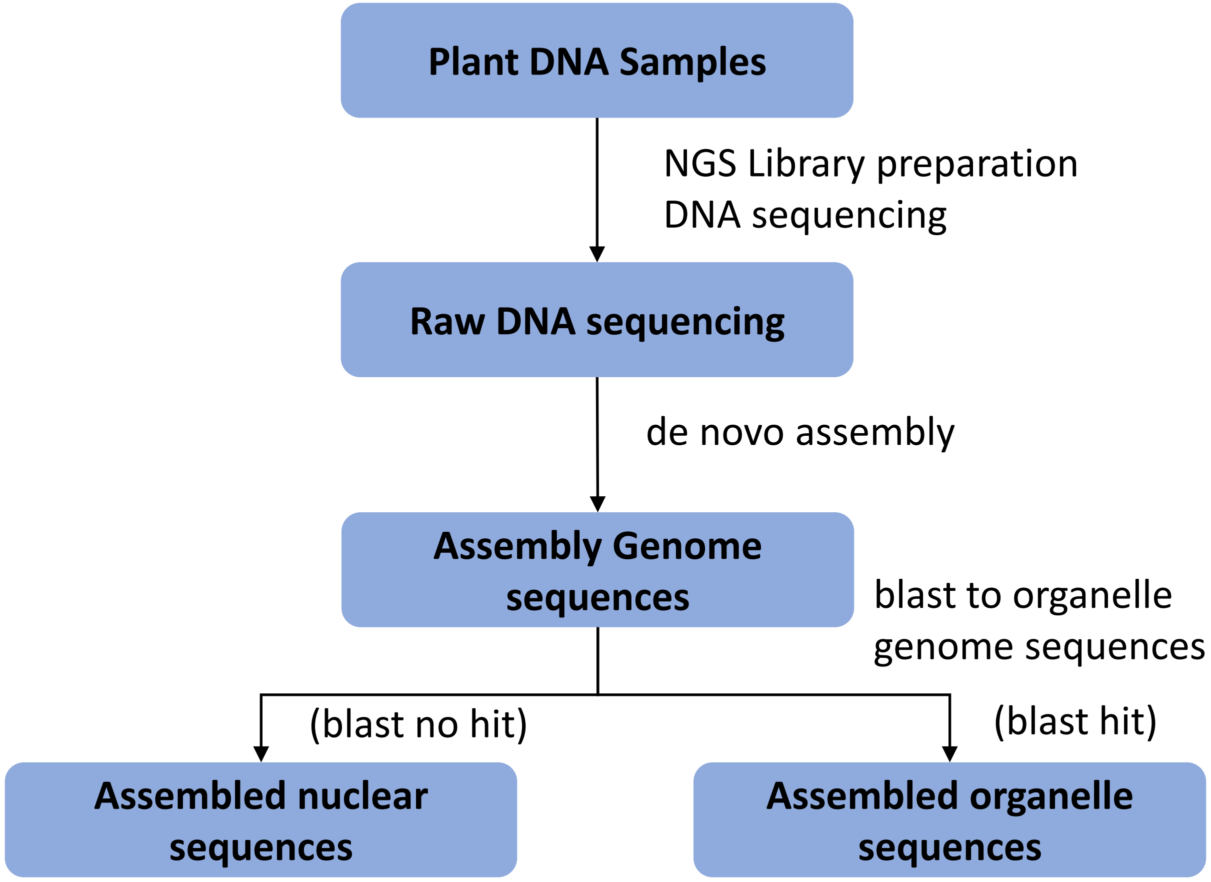
Organelle reads

de novo Assembly

Assembled nuclear sequences

Assembled organelle sequences

Methodology I



Plant DNA Samples

NGS Library preparation
DNA sequencing

Raw DNA sequencing

de novo assembly

Assembly Genome sequences

blast to organelle genome sequences

(blast no hit)

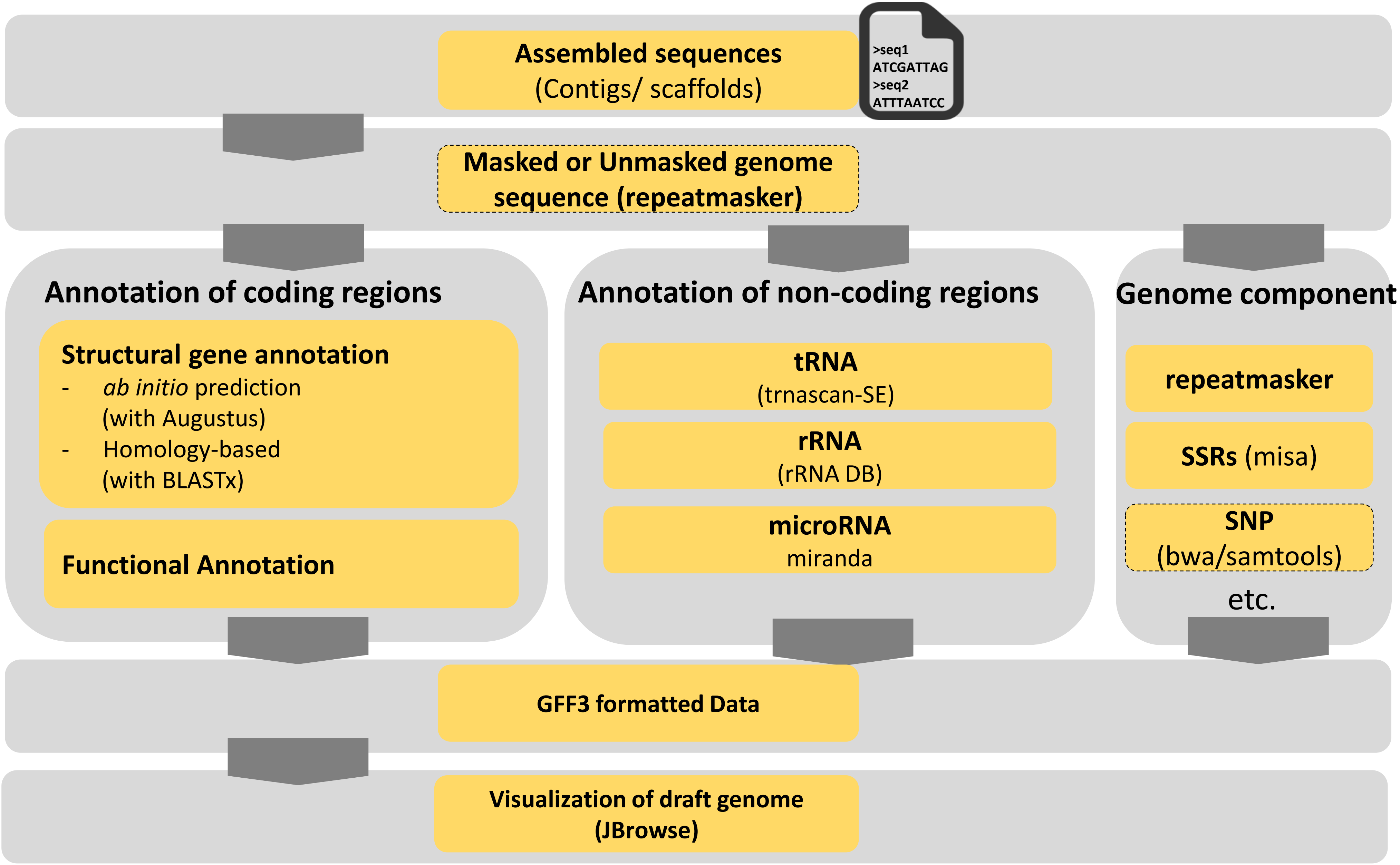
(blast hit)

Assembled nuclear sequences

Assembled organelle sequences

Methodology II

Overview of genome annotation and visualization



Assembled sequences
(Contigs/ scaffolds)

```
>seq1
ATCGATTAG
>seq2
ATTTAATCC
```

Masked or Unmasked genome sequence (repeatmasker)

Annotation of coding regions

Structural gene annotation

- *ab initio* prediction (with Augustus)
- Homology-based (with BLASTx)

Functional Annotation

Annotation of non-coding regions

tRNA
(trnascan-SE)

rRNA
(rRNA DB)

microRNA
miranda

Genome component

repeatmasker

SSRs (misa)

SNP
(bwa/samtools)

etc.

GFF3 formatted Data

Visualization of draft genome (JBrowse)

Annotation of coding regions: ab initio prediction

ab initio approach: prediction of gene structure using only the genome sequence

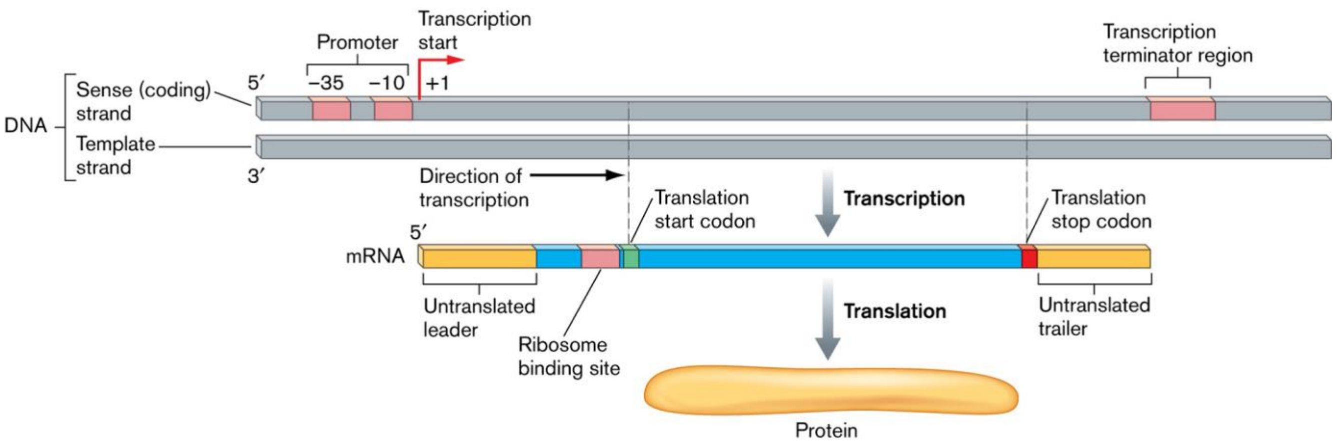


Figure1. Gene structure in Prokaryotes

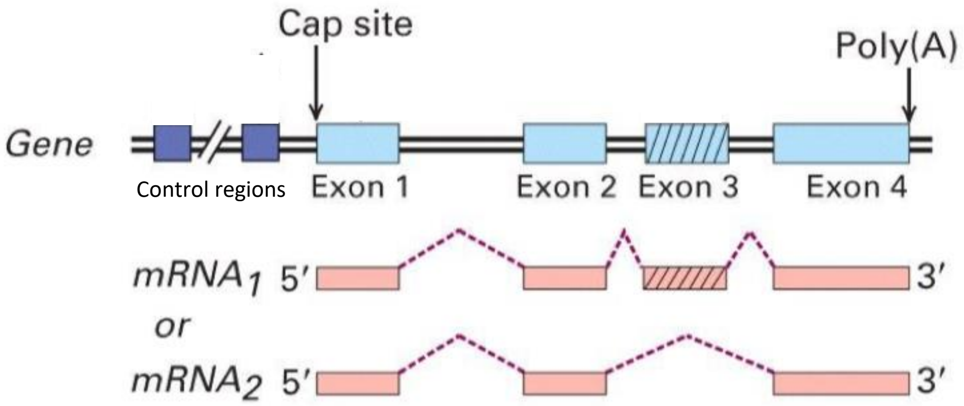


Figure2. Gene structure in Eukaryotes

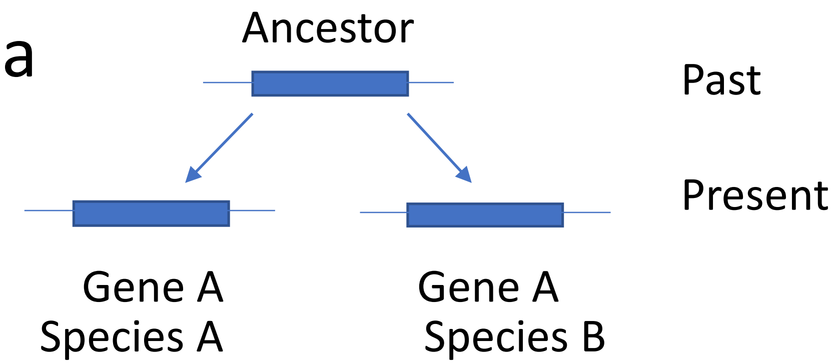
Errors includes: incorrect exon boundaries/ missed exons/ failure to detect entire genes

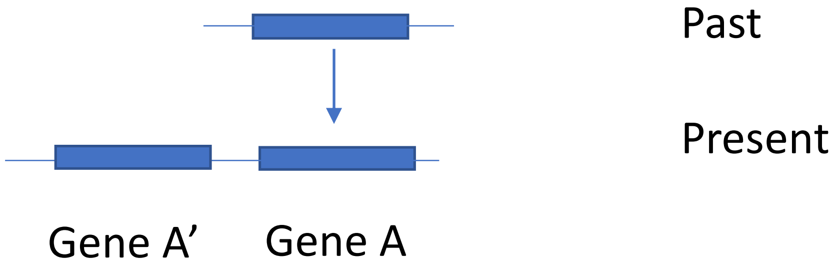
Wikibook
Figure was modified form Jorg Bungert

Annotation of coding regions: homology based

Finding genes in long sequences by looking for matches with sequences that are known to be transcribed such as cDNA, EST or protein.

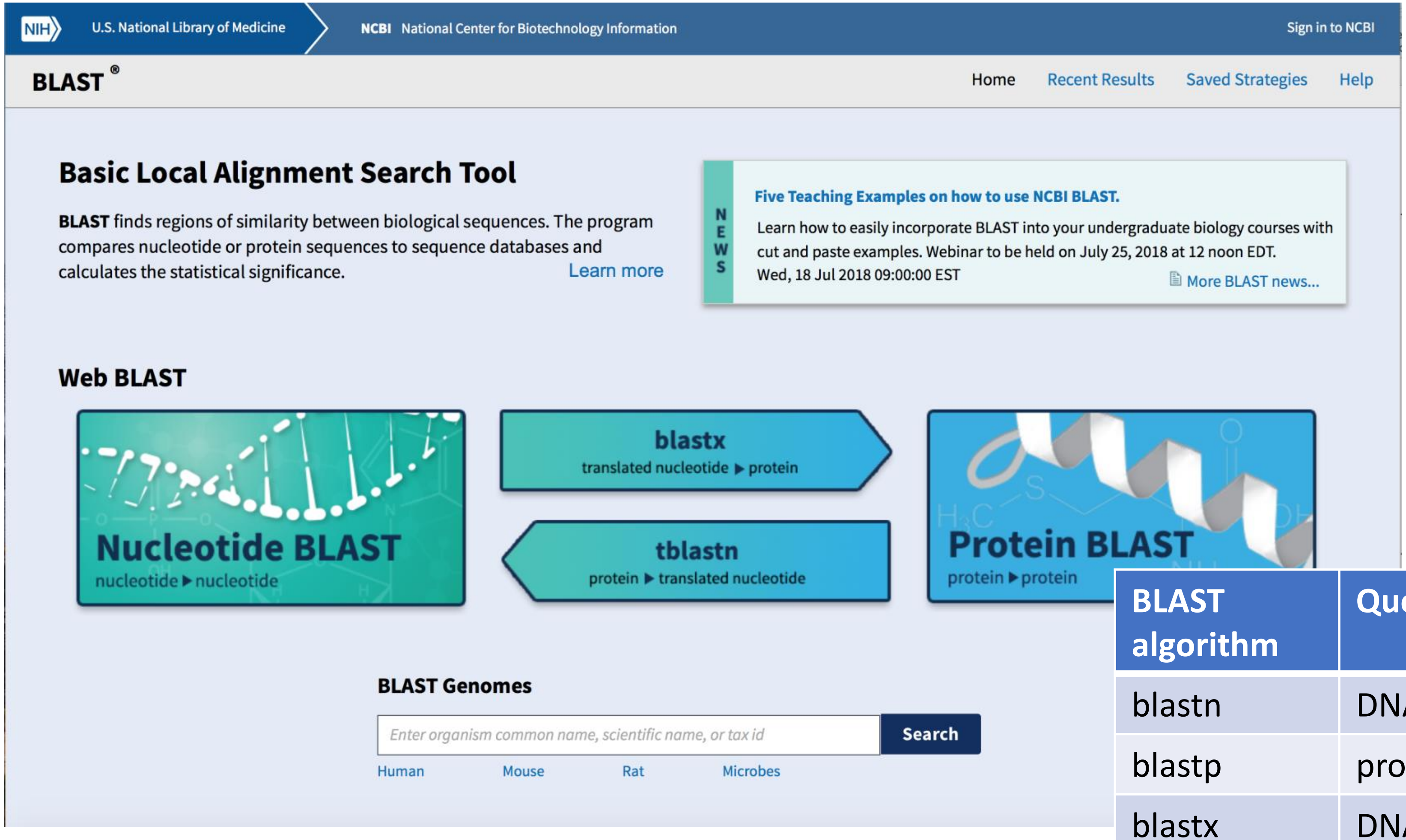
- Homologs: two genes related by descent from a **common ancestral** DNA sequence
- Orthologs: two genes in **different species**; evolved from a single ancestral gene by speciation
- Paralogs: two genes related by duplication **within a genome**
  - Mouse alpha globin and beta globin genes





Jean-François Dufayard. (2005). Tree Pattern Matching in Phylogenetic Trees Automatic Search for Orthologs or Paralogs in Homologous Gene Sequence Databases. Bioinformatics. 1; 21(11):2596-603

Basic Local Alignment Search Tools (BLAST)



NIH
U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
Sign in to NCBI
BLAST ®
Home
Recent Results
Saved Strategies
Help

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS

Five Teaching Examples on how to use NCBI BLAST.

Learn how to easily incorporate BLAST into your undergraduate biology courses with cut and paste examples. Webinar to be held on July 25, 2018 at 12 noon EDT.

Wed, 18 Jul 2018 09:00:00 EST

More BLAST news...

Web BLAST

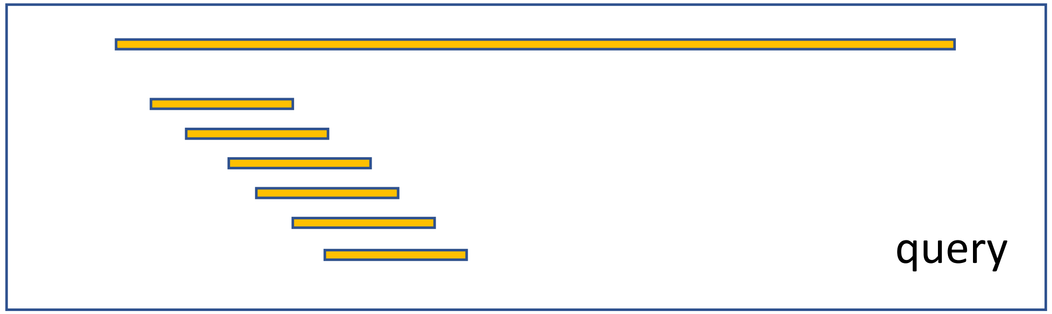
BLAST Genomes

Enter organism common name, scientific name, or tax id

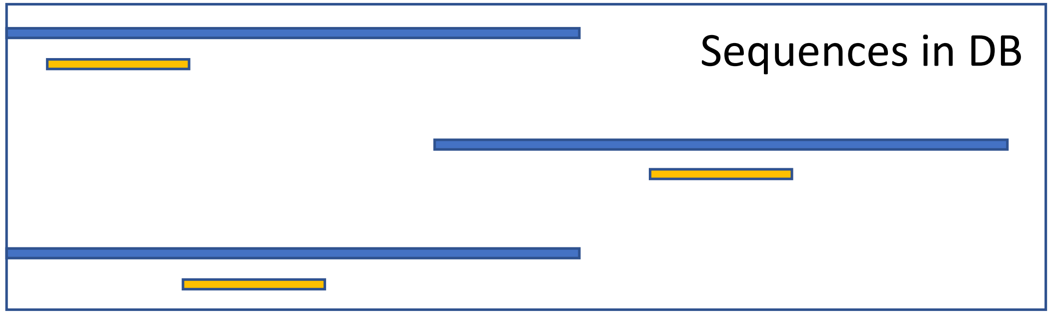
BLAST algorithm	Query sequence	DATABASE
blastn	DNA	DNA
blastp	protein	protein
blastx	DNA (translated)	protein
tblastn	Protein	DNA (translated)
tblastx	DNA (translated)	DNA (translated)

https://blast.ncbi.nlm.nih.gov/Blast.cgi

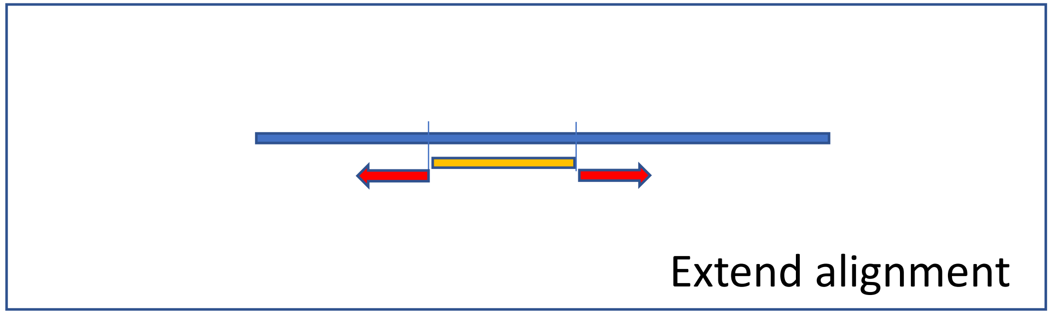
How BLAST works?



1 “word” (subsequence of query sequence)



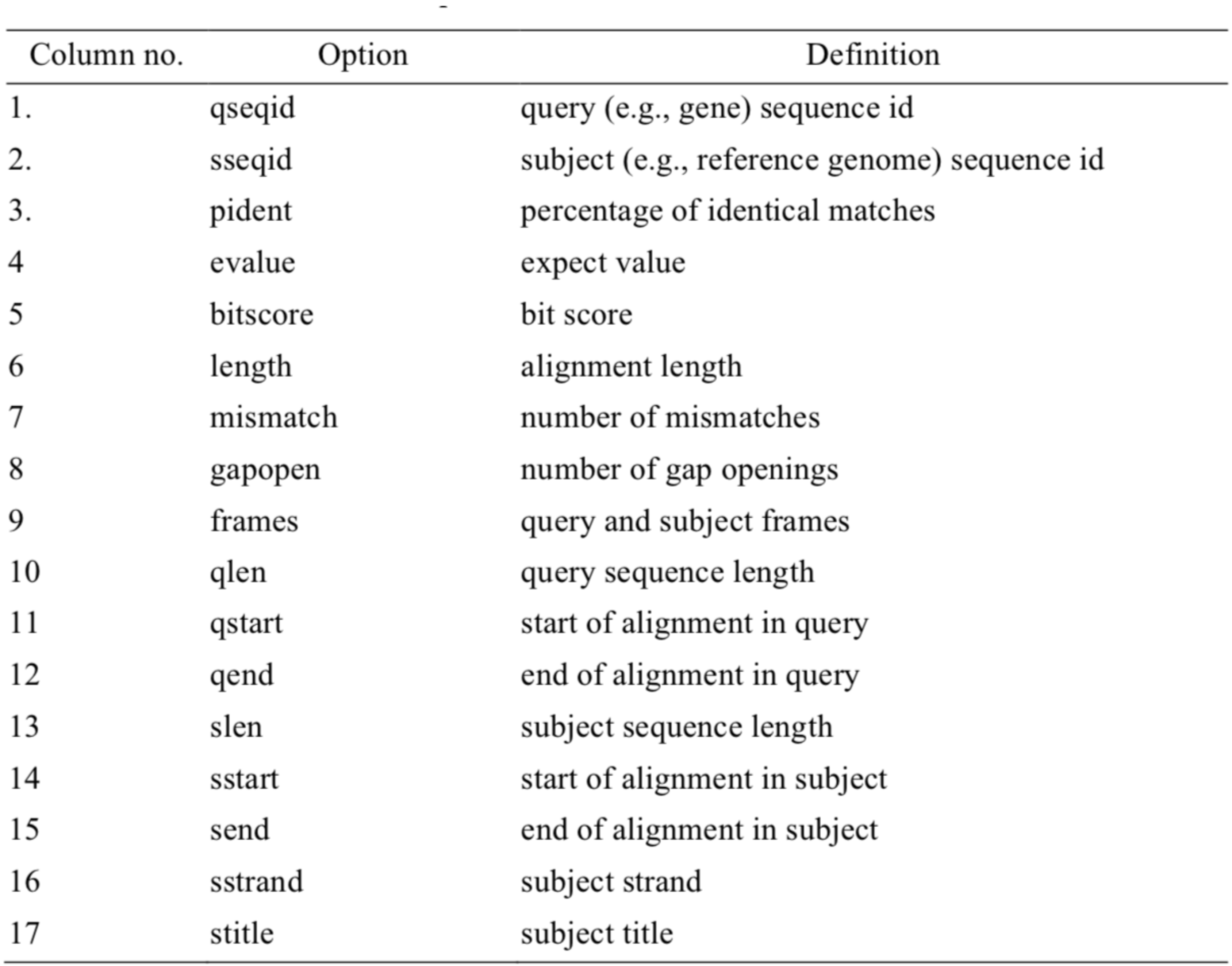
2. Query “words” are compared to the database (target sequences) and exact matches identified



3. For each word match, alignment is extended in both directions to find alignments that score greater than some threshold (maximal segment pairs, or MSPs)

BLAST output

outfmt=6



Column no.	Option	Definition
1.	qseqid	query (e.g., gene) sequence id
2.	sseqid	subject (e.g., reference genome) sequence id
3.	pident	percentage of identical matches
4	evalue	expect value
5	bitscore	bit score
6	length	alignment length
7	mismatch	number of mismatches
8	gapopen	number of gap openings
9	frames	query and subject frames
10	qlen	query sequence length
11	qstart	start of alignment in query
12	qend	end of alignment in query
13	slen	subject sequence length
14	sstart	start of alignment in subject
15	send	end of alignment in subject
16	sstrand	subject strand
17	stitle	subject title

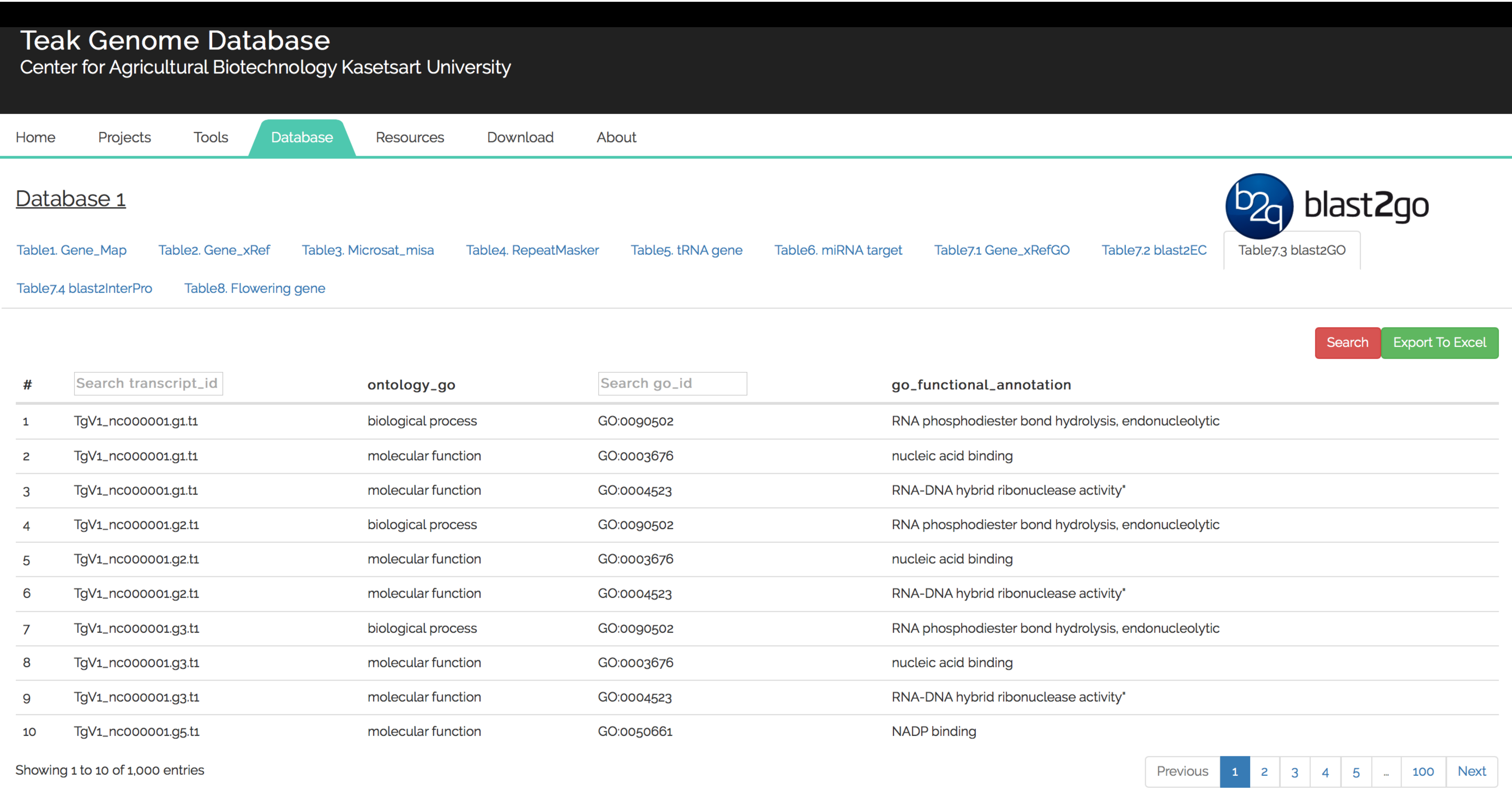
Modified from: https://www.ncbi.nlm.nih.gov/books/NBK279684/

Functional annotation (gene function prediction)

- ‘**Structural**’ genome annotation is the process of identifying genes and their intron–exon structures.

- ‘**Functional**’ genome annotation is the process of attaching meta-data such as gene ontology terms to structural annotations.

Mark Yandell and Daniel Ence (2012) A beginner’s guide to eukaryotic genome annotation. Nature Reviews Genetics 13(5):329-42



Database 1

#	Search transcript_id	ontology_go	Search go_id	go_functional_annotation
1	TgV1_nc000001.g1.t1	biological process	GO:0090502	RNA phosphodiester bond hydrolysis, endonucleolytic
2	TgV1_nc000001.g1.t1	molecular function	GO:0003676	nucleic acid binding
3	TgV1_nc000001.g1.t1	molecular function	GO:0004523	RNA-DNA hybrid ribonuclease activity"
4	TgV1_nc000001.g2.t1	biological process	GO:0090502	RNA phosphodiester bond hydrolysis, endonucleolytic
5	TgV1_nc000001.g2.t1	molecular function	GO:0003676	nucleic acid binding
6	TgV1_nc000001.g2.t1	molecular function	GO:0004523	RNA-DNA hybrid ribonuclease activity"
7	TgV1_nc000001.g3.t1	biological process	GO:0090502	RNA phosphodiester bond hydrolysis, endonucleolytic
8	TgV1_nc000001.g3.t1	molecular function	GO:0003676	nucleic acid binding
9	TgV1_nc000001.g3.t1	molecular function	GO:0004523	RNA-DNA hybrid ribonuclease activity"
10	TgV1_nc000001.g5.t1	molecular function	GO:0050661	NADP binding

Showing 1 to 10 of 1,000 entries

Ribosomal ribonucleic acid (rRNA)

ribosomal RNA in Prokaryotic genome

- Ribosome large subunit

  23S rRNA
  5S rRNA

- Ribosome small subunit

  16S rRNA (bp)

ribosomal RNA in Eukaryotic genome

- Ribosome arge subunit

  28S rRNA in mammals, 25S rRNA in plant
  5S rRNA
  5.8S rRNA (154 bp)

- Ribosome small subunit

  18S rRNA (1753 bp)

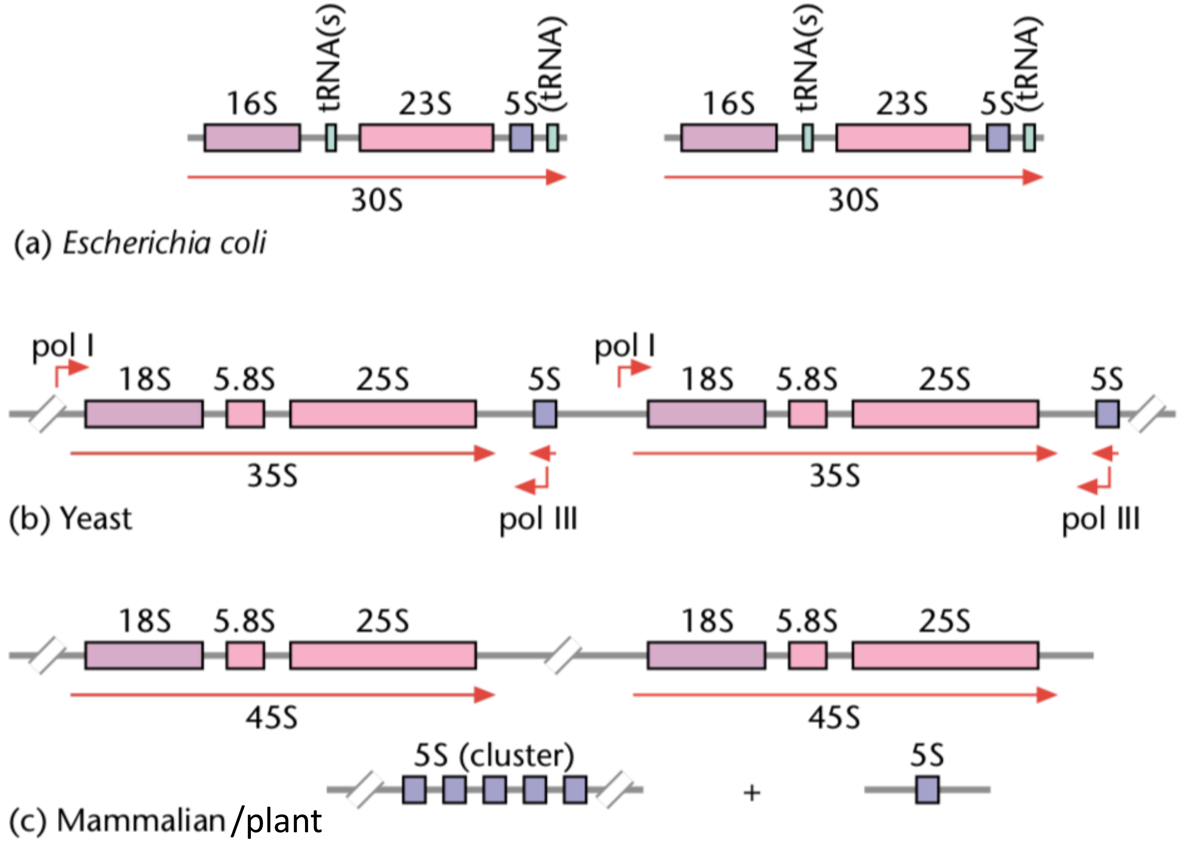
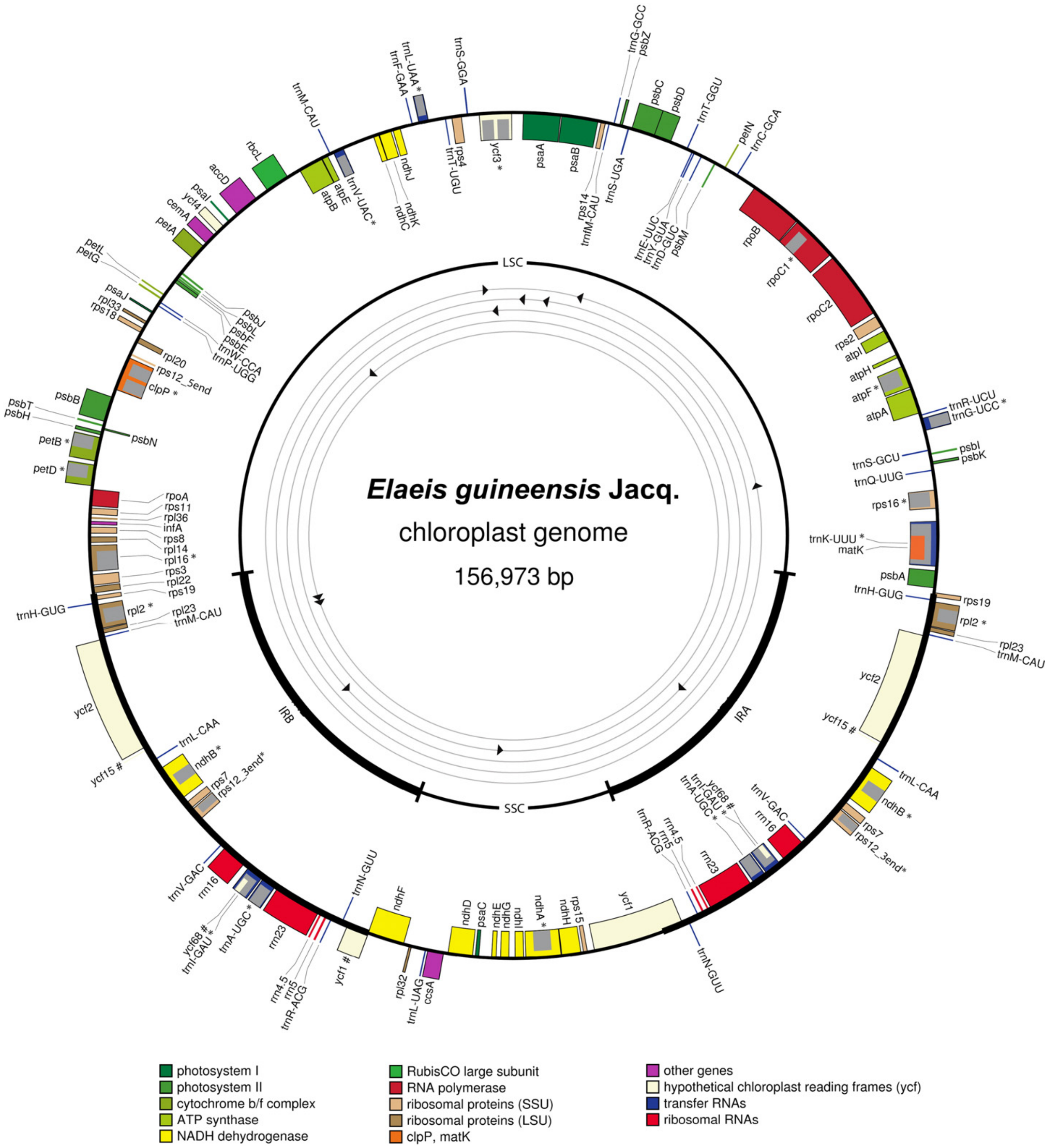


Figure 1 Ribosomal DNA (rDNA) organization in different species.

Lafontaine, D. and Tollervey D. (2001) ENCYCLOPEDIA OF LIFE SCIENCES Nature Publishing Group



Oil palm chloroplast genome

112 unique genes
79 protein-coding genes
29 tRNA genes

4 ribosomal RNA genes
Large subunit
23S rRNA (2,805 bp)
5S rRNA (121 bp)
4.5S rRNA (103 bp)
Small subunit
16S rRNA (1,491 bp)

Uthaipaisanwong P. etal. (2012) Characterization of the chloroplast genome sequence of oil palm (*Elaeis guineensis* Jacq.). Gene. Volume 500, Issue 2, Pages 172-180

Annotation of non-coding regions: Repeats

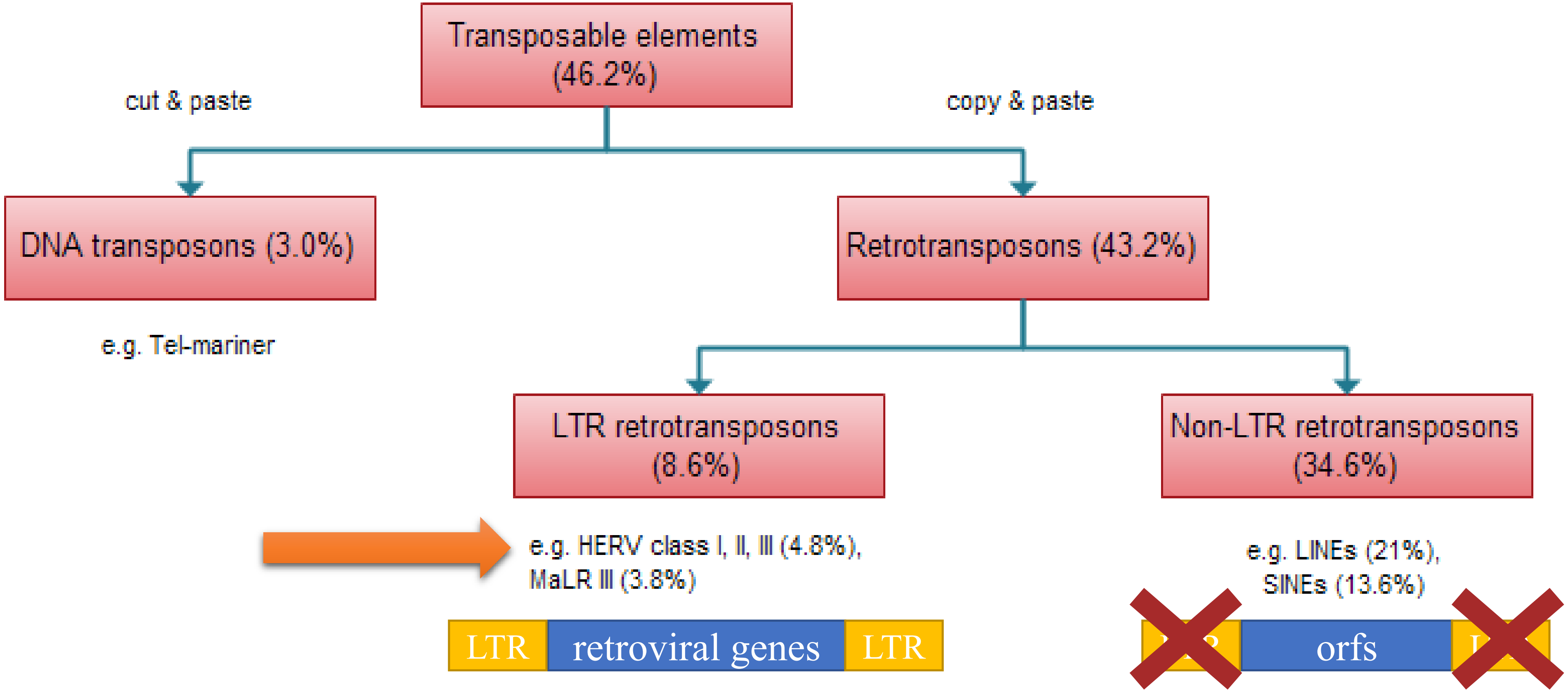
- Transposable elements
  - DNA transposons: cut and paste mechanism
  - Retrotransposons: copy itself and paste mechanism
  - Transposable element fractions in plant genomes are variable, as low as ~3% in small genomes and as high as ~85% in large genomes.

- Repeated sequences
  - Microsatellite or single sequence repeats (SSRs):
    - tandem repeats of short 1-6 bp DNA sequence motifs
    - total size is less than 1000 bp
  - Minisatellite
    - tandem repeat of 10-60 bp DNA sequences motifs (5-50 times)
    - total size is 1k-20kbp

Sung-Il Lee and Nam-Soo Kim (2014). Transposable Elements and Genome Size Variations in Plants. Genomics and informatics. 12(3): 87-97

Percent of human transposable element

Transposable elements (TEs) are fragments of DNA that can insert into new chromosomal locations and often make duplicate copies of themselves in the process. (Nature Review Genetics, 2002, volume 3 329-341)



Transposable elements (46.2%)

cut & paste

copy & paste

DNA transposons (3.0%)

e.g. Tel-mariner

Retrotransposons (43.2%)

LTR retrotransposons (8.6%)

e.g. HERV class I, II, III (4.8%), MaLR III (3.8%)

Non-LTR retrotransposons (34.6%)

e.g. LINEs (21%), SINEs (13.6%)

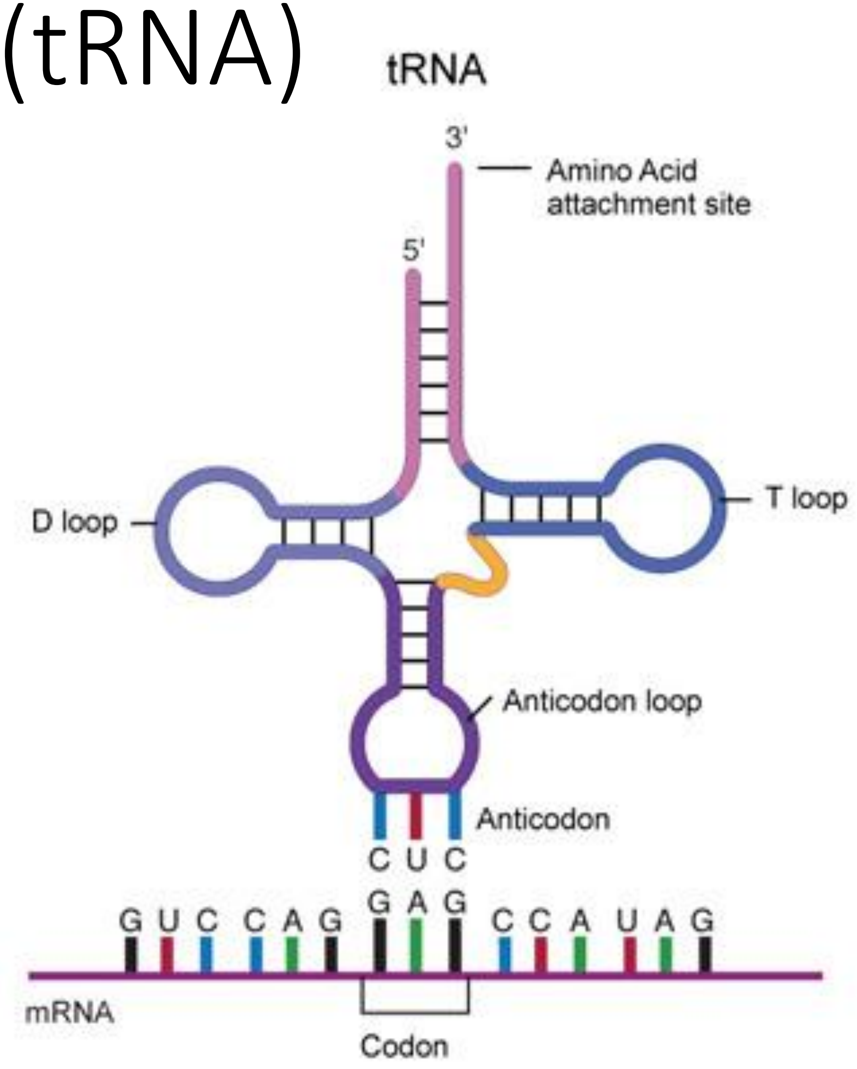
Ph.D. QE of Thongyoo P. 2013

Derived from Nelson, P. N., et al., 2004, **Clin Exp Immunol**, Vol.138, pp.1-9.

transfer ribonucleic acid (tRNA)

- tRNA is a type of RNA molecule that helps decode a messenger RNA (mRNA) sequence into a protein.

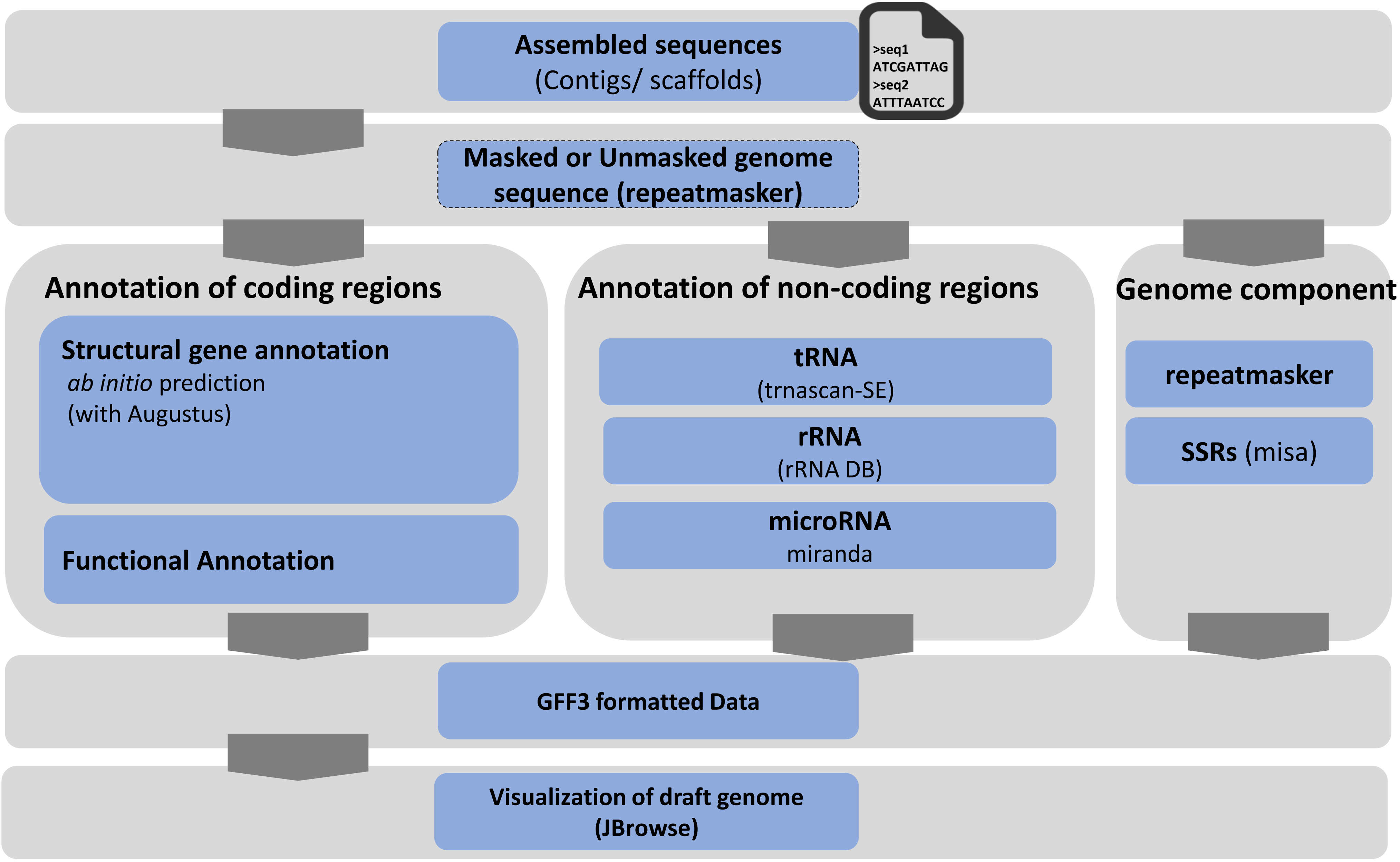
- Each codon represents a particular amino acid, and each codon is recognized by a specific tRNA.



Structure of tRNA

Scitable by nature education
https://rarediseases.info.nih.gov/GlossaryDescription/474/0
Lowe, T.M. and Chan, P.P. (2016) tRNAscan-SE On-line: Search and Contextual Analysis of Transfer RNA Genes. *Nucl. Acids Res.* **44**: W54-57.

CAB-Inhouse annotation pipeline (CABAnnot)



Assembled sequences
(Contigs/ scaffolds)

```
>seq1
ATCGATTAG
>seq2
ATTTAATCC
```

Masked or Unmasked genome sequence (repeatmasker)

Annotation of coding regions

Structural gene annotation
ab initio prediction
(with Augustus)

Functional Annotation

Annotation of non-coding regions

tRNA
(trnascan-SE)

rRNA
(rRNA DB)

microRNA
miranda

Genome component

repeatmasker

SSRs (misa)

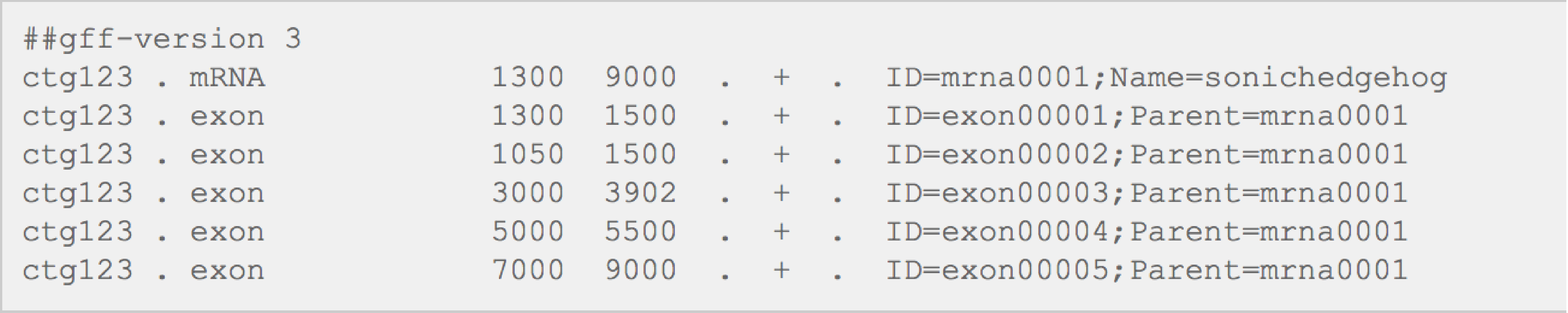
GFF3 formatted Data

Visualization of draft genome (JBrowse)

Generic Feature Format Version 3 (Gff3) file format

- Standard file format for storing genomic features in a text file.
- GFF3 format consists of one line per feature, each containing 9 columns of data.

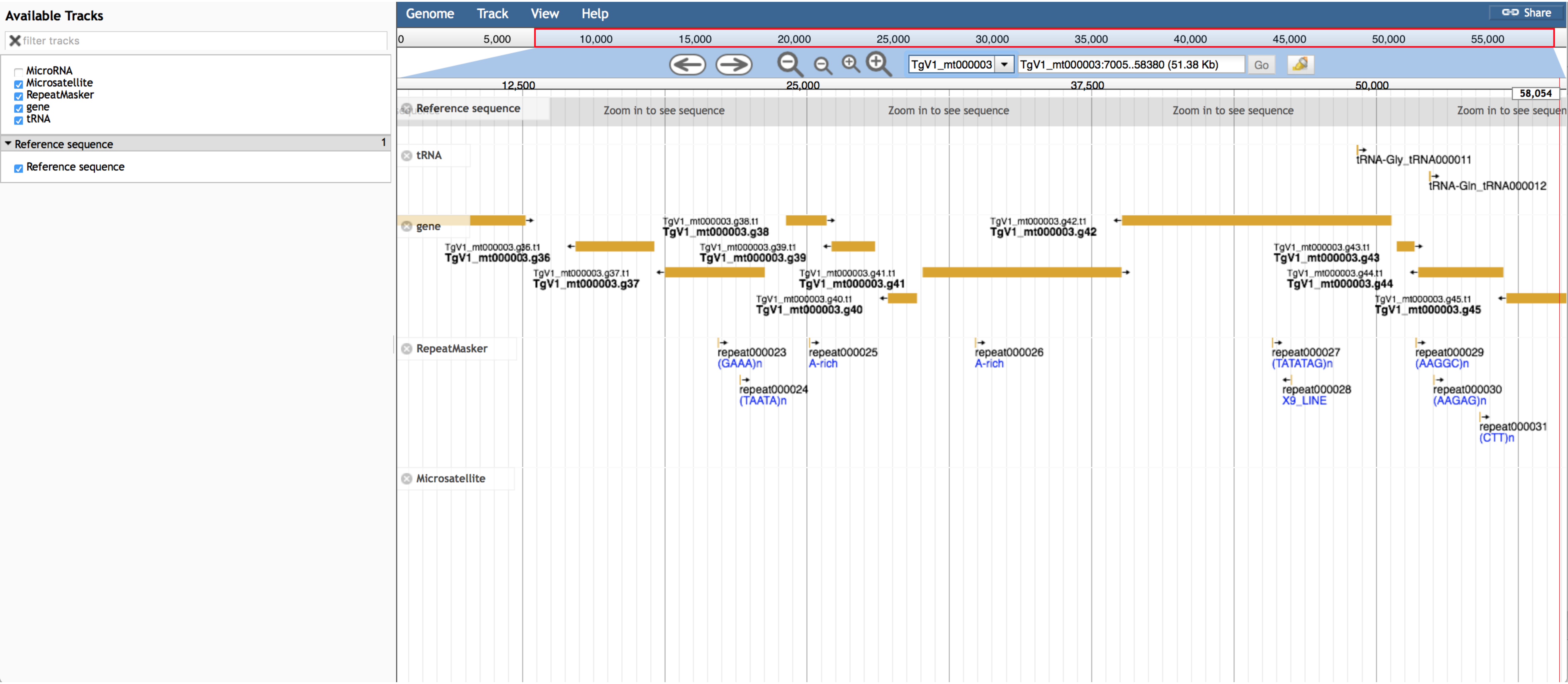
Col.1	Col.2	Col.3	Col.4	Col.5	Col.6	Col.7	Col.8	Col.9
seqid	source	type	start	end	score	stand	phase	attributes



```
##gff-version 3
ctg123 . mRNA            1300  9000  .  +  .  ID=mrna0001;Name=sonichedgehog
ctg123 . exon            1300  1500  .  +  .  ID=exon00001;Parent=mrna0001
ctg123 . exon            1050  1500  .  +  .  ID=exon00002;Parent=mrna0001
ctg123 . exon            3000  3902  .  +  .  ID=exon00003;Parent=mrna0001
ctg123 . exon            5000  5500  .  +  .  ID=exon00004;Parent=mrna0001
ctg123 . exon            7000  9000  .  +  .  ID=exon00005;Parent=mrna0001
```

https://asia.ensembl.org/info/website/upload/gff3.html

JBrowse Genome Browser



http://www.breedserve.cab.kps.ku.ac.th/teakgenomedb/index.php/home

Practical section: genome annotation