

# Introduction to genome annotation

---



22/09/2025

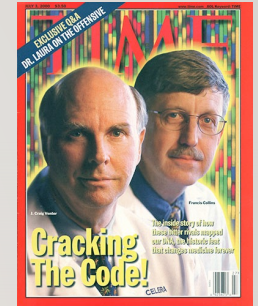


# ... prices go down

## Human genome sequencing:

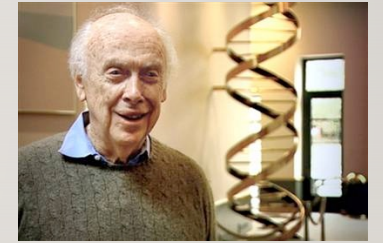
2004: Genome of Craig Wenter costs 70 mln \$

- Sanger's sequencing



2007: Genome of James Watson costs 2 mln \$

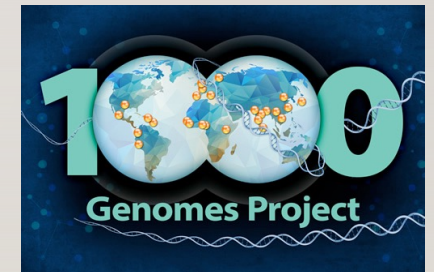
- 454 pyrosequencing



2014: Ultimate goal: 1000 \$ / individual

2016: Illumina Xten: Almost there! (1200 \$)

2017: NovaSeq: "Hold my beer..." (100 \$)



... scientific value diminishes



Science 5 September 1997:  
Vol. 277 no. 5331 pp. 1453-1462  
DOI: 10.1126/science.277.5331.1453

IF 31.6

[< Prev](#) | [Table of Contents](#) | [Next >](#)

#### ARTICLES

### The Complete Genome Sequence of *Escherichia coli* K-12

Frederick R. Blattner<sup>a</sup>, Guy Plunkett III<sup>a</sup>, Craig A. Bloch, Nicole T. Perna, Valerie Burland, Monica Riley, Julio Collado-Vides, Jeremy D. Glasner, Christopher K. Rode, George F. Mayhew, Jason Gregor, Nelson Wayne Davis, Heather A. Kirkpatrick, Michael A. Goeden, Debra J. Rose, Bob Mau and Ying Shao

Journal of Biotechnology  
Article in Press, Corrected Proof - Note to users

IF 2.9

doi:10.1016/j.jbiotec.2010.12.018 | [How to Cite or Link Using DOI](#)  
[Permissions & Reprints](#)

### The complete genome sequence of the dominant *Sinorhizobium meliloti* field isolate SM11 extends the *S. meliloti* pan-genome

Susanne Schneiker-Bekel<sup>a</sup>, Daniel Wibberg<sup>a</sup>, Thomas Bekel<sup>b</sup>, Jochen Blom<sup>b</sup>, Burkhard Linke<sup>b</sup>, Helko Neuweiger<sup>b</sup>, Michael Stiens<sup>a, c</sup>, Frank-Jörg Vorhölter<sup>a</sup>, Stefan Weidner<sup>a</sup>, Alexander Goesmann<sup>b</sup>, Alfred Pühler<sup>a</sup> and Andreas Schlüter<sup>a</sup>  



It is the **annotation** that bridges the gap from the sequence to the biology of the organism

# What is annotation ?

### Structural annotation:

Find out where the regions of interest (usually genes) are in the sequence data and what they look like.

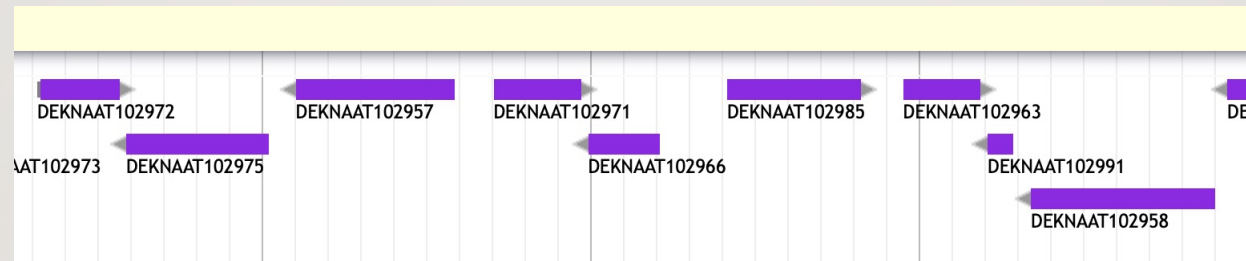
vs

### functional annotation:

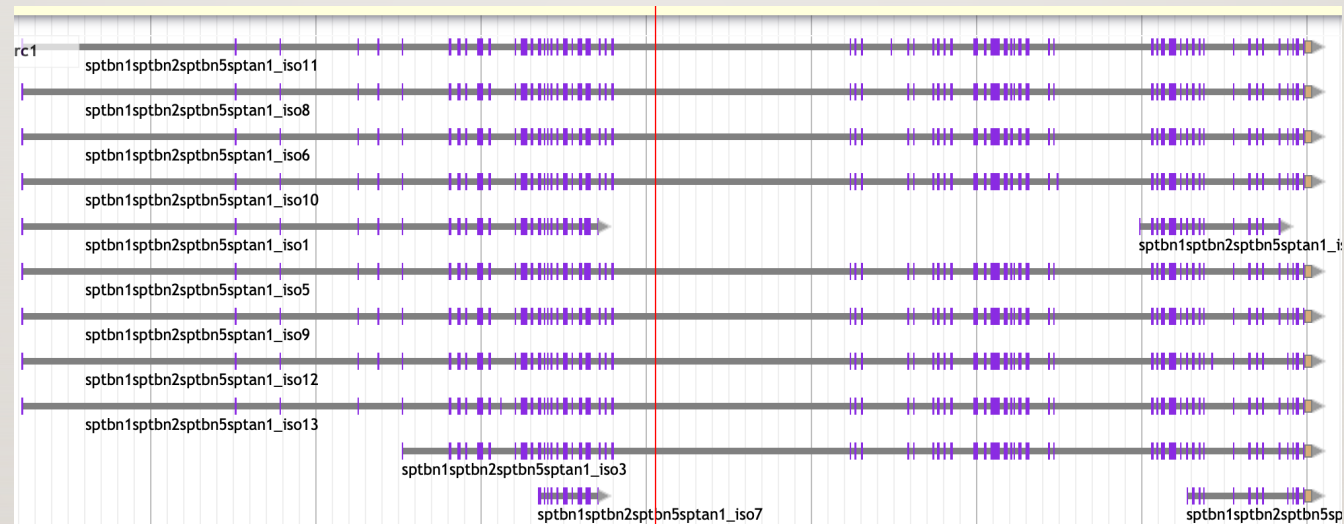
Find out what the regions do.  
What do they code for?

## Organisms differ in genomic complexity

A yeast



A crustacean

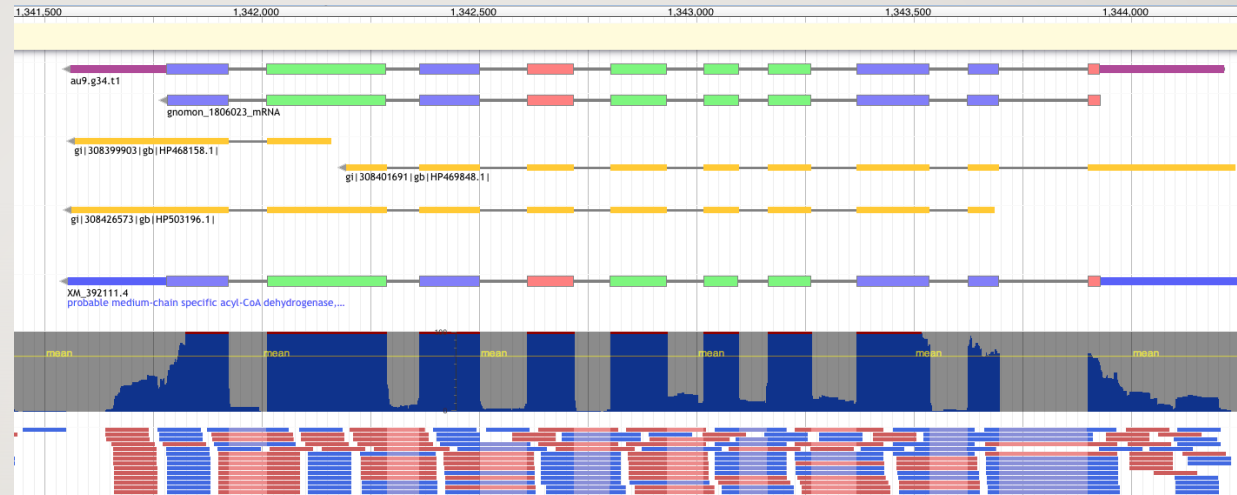


# Introduction to annotation

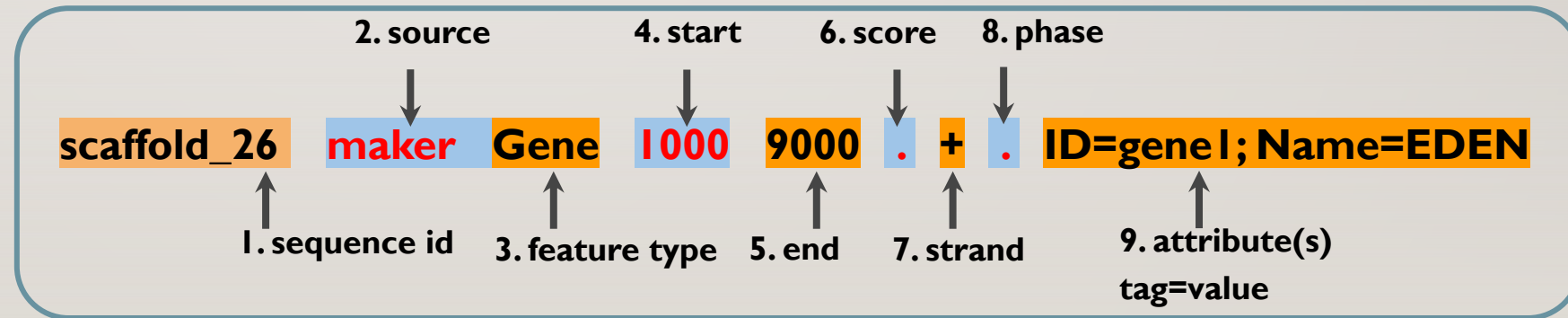
From a genome...  
**FASTA**

```
>scaffold_26
AGTCACACACCTTCAGCTTACACCTGACTGCAGCCCTTACTCAAAACA
TTCCAGCCAGGAAGATGCTCCGACACAGCTTCTGGATGCCGCTCTCGAC
GTCGAACGGCCCGCGCCGGGAAATCGGCAGCGTCGGTGACCGCGGAGAT
CCGAAGCCGCTCGGGACCTGCGAGACAACGGAGGCGGTCAACGAGAC
GCCGAGGGCTGGGAGTTATCCACACCGGGCCGTAAGTTTTCTACCCA
AAAACCCATAGAAAAGAGATGAACCACTAAGTTTGATAACTCTTCTACTT
AACCGTGACCTACGTGCCGGGGCAGGCGAGCTCTGACCTAAGCGGCAC
ACGAACAAGGTGGTGCCCAATATAACAAAGATGATGCAAGGCGTTGA
AATAAATCTCCGGAAGATTAATTCTCGAGCCGACACGCTTTGAGGCAGC
GGAACCTACAGAACCACCGCAGTCACGTGAGAAGAGTCTAATACTCTCCA
AAGAGAAGTCCAAGGGAATGGAACGTGAAAAGAAGTGCTTATCAAAAGC
GAGAAGGAAGATGGATGAGAATCTTTGTGACTTCTTCTGGTCTCAAAA
AGCAAAATGTAAGATGCCAGACTAAGCCCGATCTGAGAAAGTACGCGA
GCAGAGACCCCGCTGCCGATGTGGCCGAGAAGCATGCCGATAAAGCACC
GAGACATAACAAAGCCCTGTGACACACAAGACGATGGACACAACTACAT
AACACAGACACAACTAAATGACACAGAGAAGTTGAACTTCTGGGGA
AGTAAACATTTCTGAAACATCTACCAACATCCGTCATATATATTTCCA
TTCCACGGGACTCTTGGTTTGATATATGCGTGTAAACAGTAATCCCGCT
GTAGCAATCACCCTATGCATAATTCATTAATTTCTTGGAGTTGCTGAGT
ATCATCTTATCAGTCTTATTTTTTCTTGGCTCTGGTTTCGGGCTTTTT
TTTTTCTTCTGATAAGATTTTCCAGGAATGTGAAGACCCCTGCATCCT
TCCCAACTGACCACCAACTACAGACATTCATAGCATTACATTACAC
AACCTAGGCAAGTTTTTCTAACATTAAGGAACATGAAAAAGCCAACAT
CACAATATATTCATAACAATTATGGAACATGCGAAAAGCCAATACACAG
TACATTTATAACAATACCTCCCTTTCTTTCTTTAGAGATCATATGGCT
TGACCGCCGCTCTCGCCGCCACCGCTGAGTACTGCCGTGCCGGAGTC
ACGGAGCCAGTCCCCCGGGCCACCGCTCTCGCCGCCGCCACGGA
GATCGGCTGCGCACTCCGAGCTCGGCGTGCCATCGCCGCCCGCGCG
GGTGCCCCGCGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

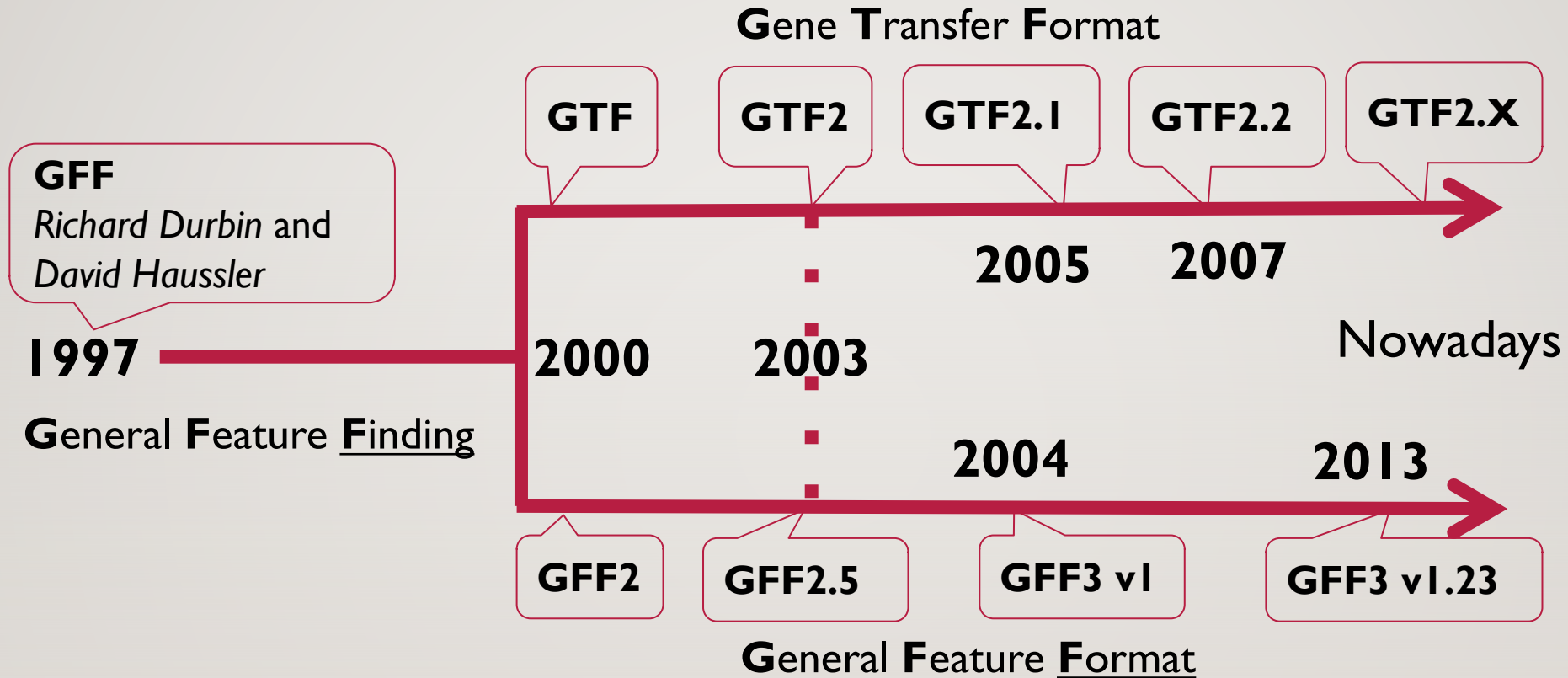
...to an annotated gene  
**GTF/GFF**



- 9 columns
- 1 feature = 1 line



## □ GFF / GTF formats



<https://agat.readthedocs.io/en/latest/gxf.html>



# Introduction to annotation: GFF3 / GTF

```
##gff-version 3.2.1
##sequence-region ctg123 1 1497228
```

## GFF3

|        |   |      |      |      |   |   |   |                                 |
|--------|---|------|------|------|---|---|---|---------------------------------|
| ctg123 | . | Gene | 1300 | 1500 | . | + | . | ID=gene1;Name=eden              |
| ctg123 | . | mRNA | 1300 | 1500 | . | + | . | ID=mRNA1;Parent=gene1;Name=eden |
| ctg123 | . | exon | 1300 | 1500 | . | + | . | ID=exon1;Parent=mRNA1;Name=eden |
| ctg123 | . | CDS  | 1300 | 1500 | . | + | 0 | ID=cds1 ;Parent=mRNA1;Name=eden |

1 - sequence id      2 - source      3 - feature type      4 - start      5 - end      6 - score      7 - strand      8 - phase      9 - attributes  
tag=value

- Header
- 9 columns
- 1 feature = 1 line
- 1 record = features with relationship

```
#!genome-build GRCz11
#!genome-date 2017-05
```

## GTF

|        |   |            |      |      |   |   |   |                                             |
|--------|---|------------|------|------|---|---|---|---------------------------------------------|
| ctg123 | . | Gene       | 1500 | 1500 | . | + | . | gene_id gene1; name eden;                   |
| ctg123 | . | Transcript | 1500 | 1500 | . | + | . | gene_id gene1; transcript_id=t1; name eden; |
| ctg123 | . | exon       | 1300 | 1500 | . | + | . | gene_id gene1; transcript_id=t1; name eden; |
| ctg123 | . | CDS        | 1300 | 1500 | . | + | 0 | gene_id gene1; transcript_id=t1; name eden; |

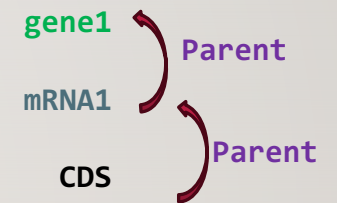
GTF and GFF formats are similar but very different...

## Different ways

- to establish the relationships between the features
- to deal with features types
- to deal with concepts (stop codon vs CDS)
- ...

GFF

|        |   |      |      |      |   |   |   |                       |        |
|--------|---|------|------|------|---|---|---|-----------------------|--------|
| ctg123 | . | gene | 1000 | 2000 | . | + | . | ID=gene1              | Level1 |
| ctg123 | . | mRNA | 1000 | 2000 | . | + | . | ID=mRNA1;Parent=gene1 | Level2 |
| ctg123 | . | exon | 1000 | 2000 | . | + | . | ID=exon1;Parent=mRNA1 | Level3 |
| ctg123 | . | CDS  | 1000 | 2000 | . | + | 0 | ID=cds01;Parent=mRNA1 | Level3 |

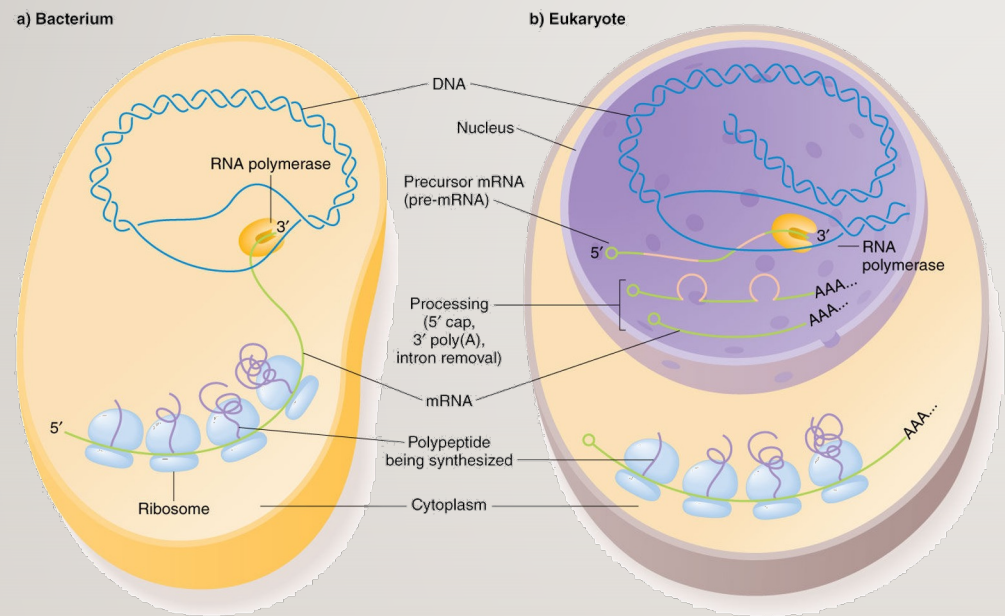


**2278 terms** accepted: Constrained to be either a term from the Sequence Ontology or an SO accession number

**9 terms** accepted: gene, transcript, exon, CDS, Selenocysteine, start\_codon, stop\_codon, three\_prime\_utr and five\_prime\_utr

GTF

|        |   |      |      |      |   |   |   |                               |        |
|--------|---|------|------|------|---|---|---|-------------------------------|--------|
| ctg123 | . | gene | 1000 | 2000 | . | + | . | gene_id g1;                   | Level1 |
| ctg123 | . | mRNA | 1000 | 2000 | . | + | . | gene_id g1; transcript_id t1; | Level2 |
| ctg123 | . | exon | 1000 | 2000 | . | + | . | gene_id g1; transcript_id t1; | Level3 |
| ctg123 | . | CDS  | 1000 | 2000 | . | + | 0 | gene_id g1; transcript_id t1; | Level3 |



© 2010 Pearson Education, Inc.

|        |   |      |      |      |   |   |   |                       |
|--------|---|------|------|------|---|---|---|-----------------------|
| ctg123 | . | gene | 1000 | 2000 | . | + | . | ID=gene1              |
| ctg123 | . | mRNA | 1000 | 2000 | . | + | . | ID=mRNA1;Parent=gene1 |
| ctg123 | . | exon | 1000 | 2000 | . | + | . | ID=exon1;Parent=mRNA1 |
| ctg123 | . | CDS  | 1000 | 2000 | . | + | 0 | ID=cds01;Parent=mRNA1 |

## AGAT (Another GFF/GTF Analysis Toolkit)

|        |   |      |      |      |   |   |   |                       |
|--------|---|------|------|------|---|---|---|-----------------------|
| ctg123 | . | gene | 1000 | 2000 | . | + | . | ID=gene1              |
| ctg123 | . | CDS  | 1000 | 2000 | . | + | 0 | ID=cds01;Parent=mRNA1 |

/!\ different type of gff: **alignment** / **annotation** / other

## Alignment

protein1 or  
transcript1 {  
protein2 or  
transcript2 {

match, match\_part

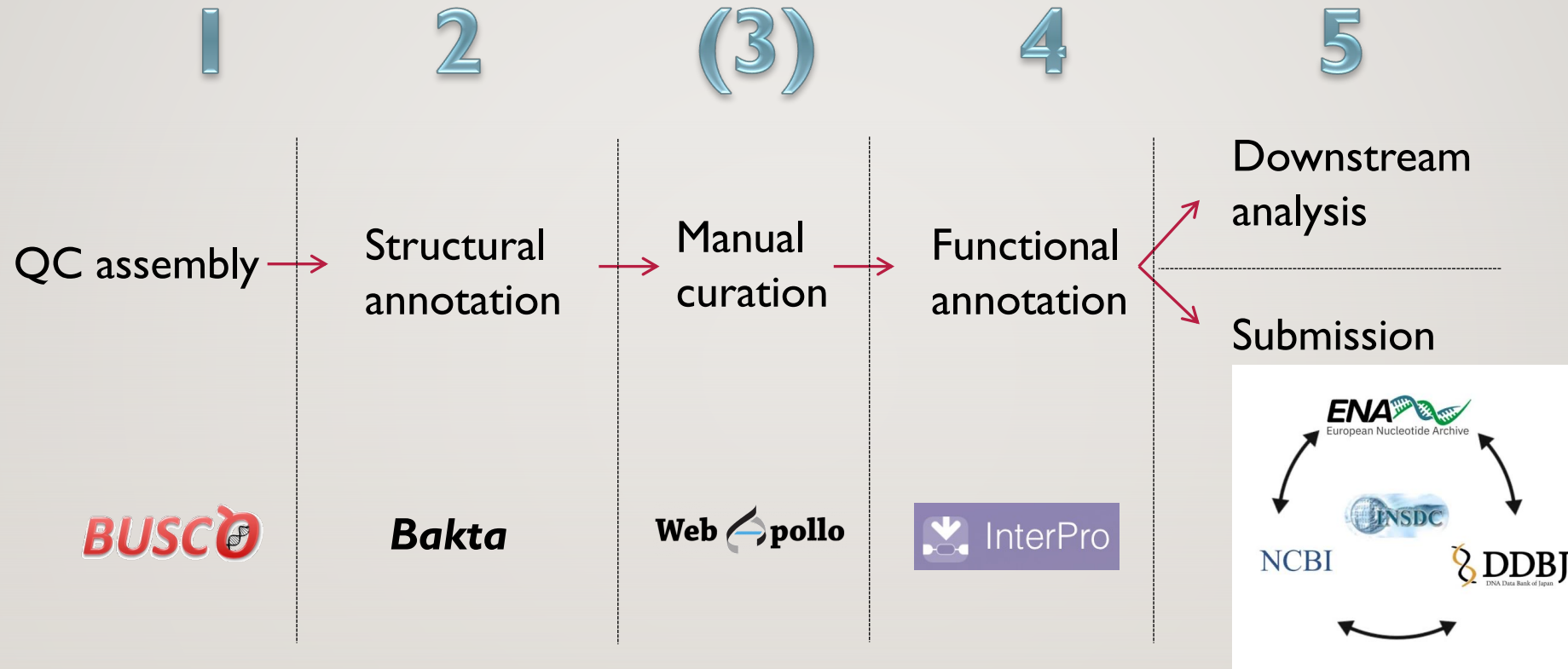
DNA

gene {

Gene, mRNA, exon, intron,  
CDS, splice site, UTR, ...  
isoforms

## Annotation

# The main steps in genome annotation



The quality of the assembly will heavily influence the quality of the annotation

- ☐ SNP-errors can change start/stop-codons/splice sites
- ☐ Indels can cause frame-shifts
- ☐ High fragmentation could break loci
- ☐ missing loci cannot be annotated

=> Annotation tools have difficulties to deal with those problems

## Before to start a project check:

- ☐ Fragmentation (N50, number of sequences, how many small contigs)
- ☐ Sanity of the fasta file (Ns, IUPAC, **lowercase nucleotides**)
- ☐ Completeness / duplication / fragmentation
- ☐ **Presence of Organelles, plasmids, etc.**
- ☐ Other (GC content, how distant from other species)



```
# BUSCO version is: 3.0.2
# The lineage dataset is: fungi_odb9 (Creation date: 2016-02-13, number of species: 85, number of BUSCOs: 290)
#
# Summarized benchmarking in BUSCO notation for file genome.fa
# BUSCO was run in mode: genome
```

**C:98.6%** [S:97.9%,D:0.7%],F:0.0%,M:1.4%,n:290

```
286 Complete BUSCOs (C)
284 Complete and single-copy BUSCOs (S)
2 Complete and duplicated BUSCOs (D)
0 Fragmented BUSCOs (F)
4 Missing BUSCOs (M)
290 Total BUSCO groups searched
```

## What to annotate?

| Feature type | DB associated               | Tool example                                                                                                          |
|--------------|-----------------------------|-----------------------------------------------------------------------------------------------------------------------|
| Coding genes | Ensembl, NCBI, EBI, Uniprot | <a href="https://juke34.github.io/awesome-genome-annotation/">https://juke34.github.io/awesome-genome-annotation/</a> |
| ncRNA        | Rfam                        | infernal                                                                                                              |
| tRNA         | Sprinzl database            | tRNAscan-SE                                                                                                           |
| snoRNA       |                             | snoscan                                                                                                               |
| miRNA        | miRBase                     | Splign<br>miR-PREFeR (for plant)                                                                                      |
| Repeats      | Repbase, Dfam               | RepeatModeler, repeatMasker, REPET, EDTA                                                                              |
| Pseudogenes  |                             | pseudopipe                                                                                                            |
| ...          |                             |                                                                                                                       |

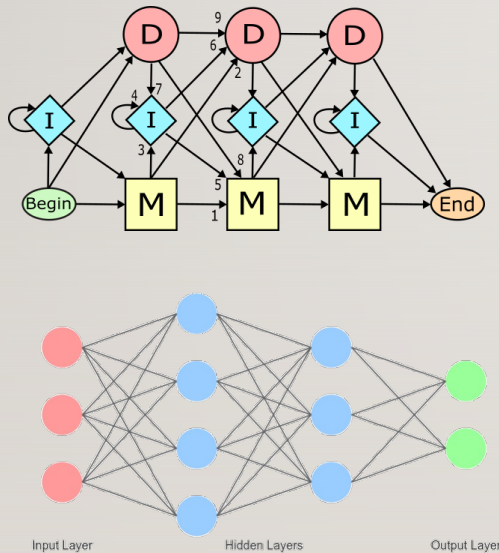


=> Focus on Repeats and coding genes  
Why repeats?

## Types of data needed

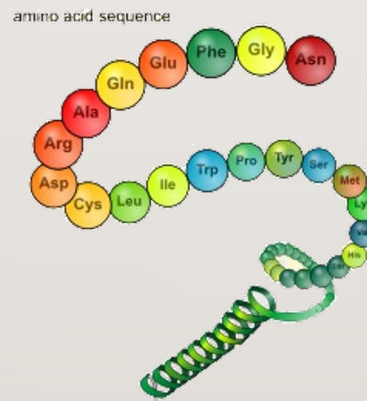
### Machine Learning

- HMM
- Deep Learning



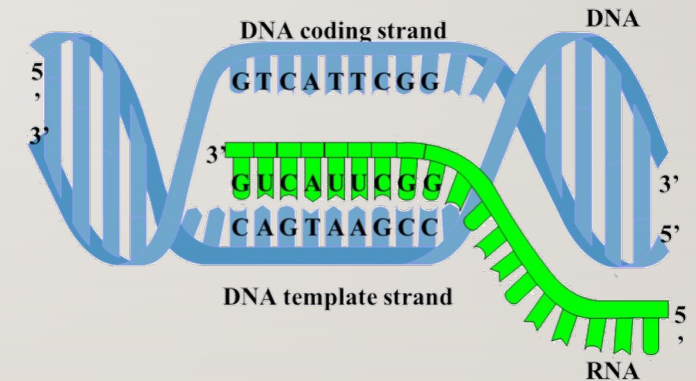
### Proteins

- Known amino acid sequences from other organisms



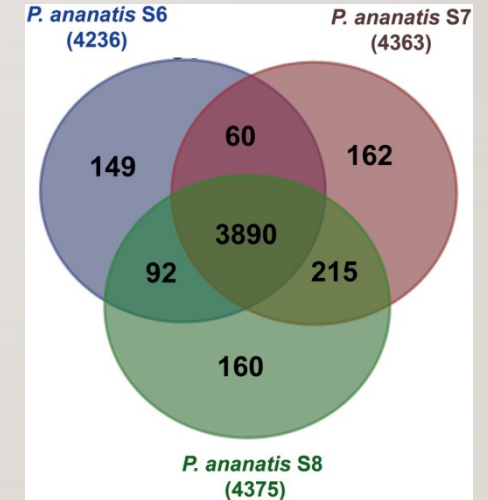
### Transcripts

- Assembled from RNA-seq or downloaded ESTs



No annotation method is perfect, they do mistakes !!

- ☐ Simple statistics (number genes / number exon per gene)
- ☐ **BUSCO** (and compare against assembly result )
- ☐ Protein/transcript evidence (AED score in MAKER)
- ☐ Comparative genomics
- ☐ Domain / Function attached
- ☐ Visualization



## Visualization / Manual curation tools

| Name                    | Standalone | Web tool | Manual curation | year | comment                                                                               |
|-------------------------|------------|----------|-----------------|------|---------------------------------------------------------------------------------------|
| Artemis                 | X          |          | X               | 2000 | Can save annotation in EMBL format                                                    |
| IGV                     | X          |          |                 | 2011 | Popular                                                                               |
| Savant                  | X          |          |                 | 2010 | Sequence Annotation, Visualization and Analysis Tool. enable Plug-ins                 |
| Tablet                  | X          |          | X               | 2013 |                                                                                       |
| IGB                     | X          |          |                 | 2008 | enable Plug-ins. Can load local and remote data (dropbox, UCSC genome, etc)           |
| Jbrowse                 |            | X        |                 | 2010 | GMOD (successor of Gbrowse)                                                           |
| Web Apollo              |            | X        | X               | 2013 | Active community (gmod). Based on Jbrowse. Real-time collaboration                    |
| UCSC                    |            | X        |                 | 2000 | A large amount of locally stored data must be uploaded to servers across the internet |
| Ensembl genome browsers |            | X        |                 | 2002 | A large amount of locally stored data must be uploaded to servers across the internet |

Annotation can be improved by **manual curation**

Most methods focus primarily on gene prediction, without **functional annotation**. On the other hand, some pipelines incorporate additional steps to address this essential aspect of genome annotation.

Accurate

- Experimentally  
=> Mutants, knockout, etc.



Mice homozygous for the diabetes 3J spontaneous mutation

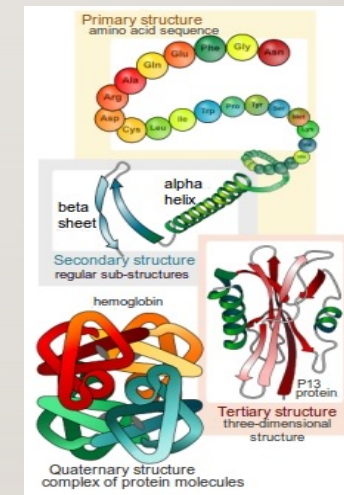
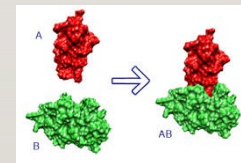
- Computationally
  - Sequence-based



limited accuracy

- Structure based

- Protein-protein interaction data



Submit your annotation in public archives e.g. EBI, NCBI, etc.

# GFF



|                 | EBI              | NCBI                      |
|-----------------|------------------|---------------------------|
| converter       | EMBLmyGFF3       | GAG<br>Manual + table2asn |
| Expected format | Flat file (EMBL) | ASN.I format              |



# *The End*

***Thank you for your attention***