# Genome assembly and annotation

## Day 5: Genome annotation

**Igor Pessi**

Department of Microbiology – UH

igor.pessi@helsinki.fi

# Aims for this part of MMB-114

**Day 1:** Basics of UNIX and working with the command line

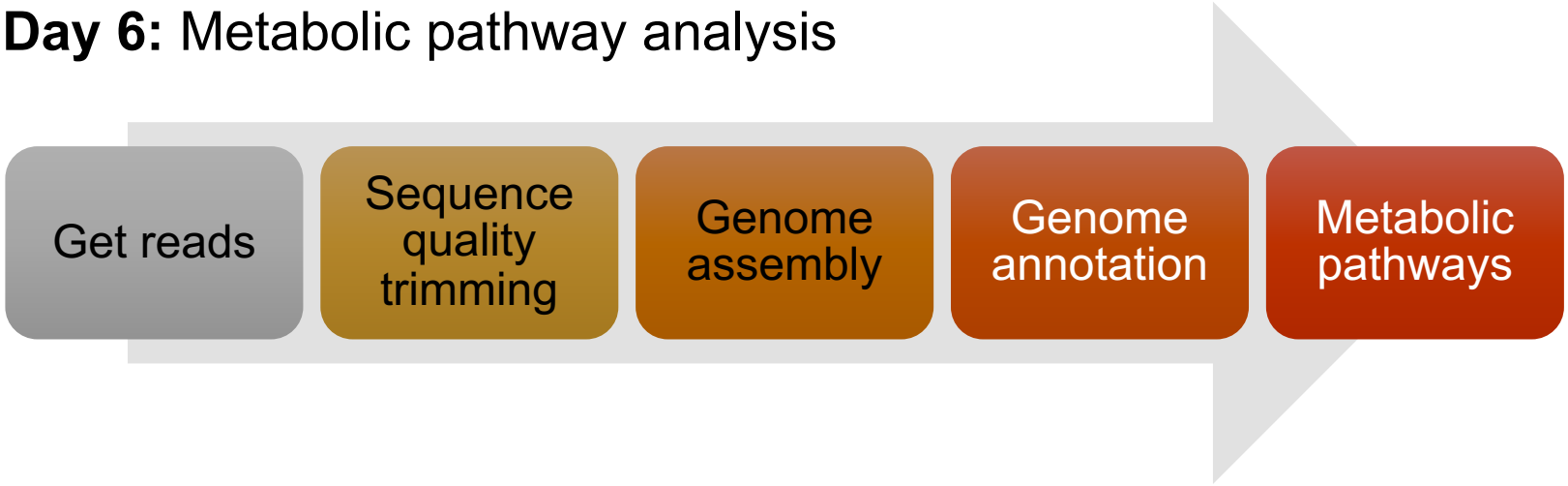
**Day 2:** Handling of Illumina data

**Day 3:** Genome assembly

**Day 4:** Check-up and report

**Day 5:** Genome annotation

**Day 6:** Metabolic pathway analysis



# Before we start...

**How does the assembly look like?**

How many contigs?

What is the longest contig?

Total size of the assembly? Is this more or less in the ballpark of what you expected for this genome?

# Three genomes

| Statistics without reference | SPADES_ALVAR_contigs | SPADES_ANTTON_contigs | SPADES_SUVI_contigs |
|---|---|---|---|
| # contigs | 44 | 27 | 1443 |
| # contigs (>= 0 bp) | 212 | 31 | 1777 |
| # contigs (>= 1000 bp) | 35 | 25 | 146 |
| # contigs (>= 5000 bp) | 30 | 20 | 1 |
| # contigs (>= 10000 bp) | 29 | 19 | 0 |
| # contigs (>= 25000 bp) | 27 | 18 | 0 |
| # contigs (>= 50000 bp) | 25 | 16 | 0 |
| Largest contig | 437 575 | 1 424 057 | 5980 |
| Total length | 4 663 117 | 5 299 459 | 1 038 776 |
| Total length (>= 0 bp) | 4 694 984 | 5 300 564 | 1 197 053 |
| Total length (>= 1000 bp) | 4 656 217 | 5 298 339 | 185 195 |
| Total length (>= 5000 bp) | 4 642 389 | 5 288 949 | 5980 |
| Total length (>= 10000 bp) | 4 634 843 | 5 283 472 | 0 |
| Total length (>= 25000 bp) | 4 593 937 | 5 262 819 | 0 |
| Total length (>= 50000 bp) | 4 522 542 | 5 206 047 | 0 |
| N50 | 227 477 | 429 888 | 702 |
| N75 | 136 681 | 242 271 | 584 |
| L50 | 8 | 4 | 554 |
| L75 | 14 | 8 | 962 |
| GC (%) | 48.06 | 56.02 | 59.41 |

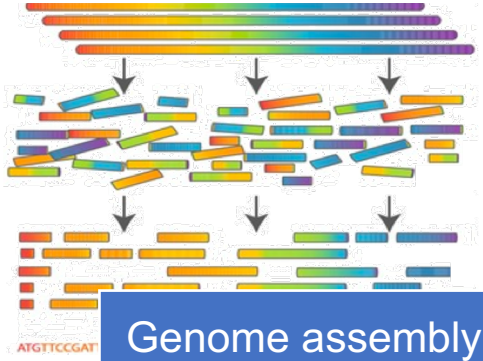
# Recap from last week:

Isolation

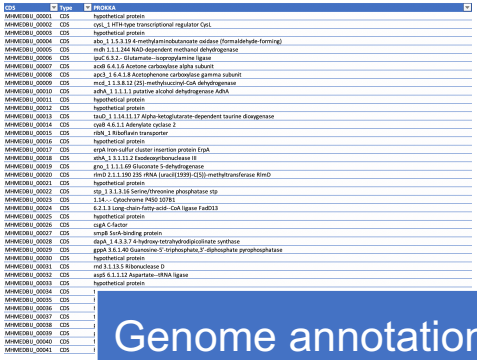
DNA extraction

Sequencing

Quality control




Genome assembly



Genome annotation

# Annotation

Adding biological information to sequences (contigs)

Information that there is a gene x in contig y at location z

- Size of the gene
- Name of the gene
- Protein product


Contig y = 2,035 bp

# Bacterial genes

Promoter

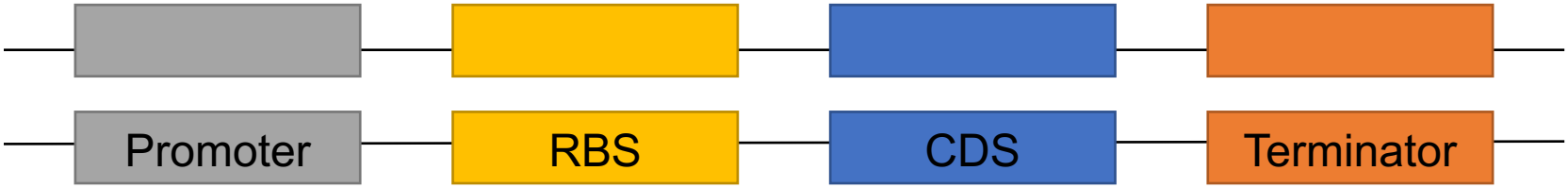
Ribosome binding site (RBS)

Coding sequence (CDS)

Terminator

Also non-coding genes

- tRNA
- rRNA



# Two ways to identify protein-coding genes

**Sequence alignment (e.g. BLAST)**

Search contigs against a database

Computationally-intensive

**Gene finding**

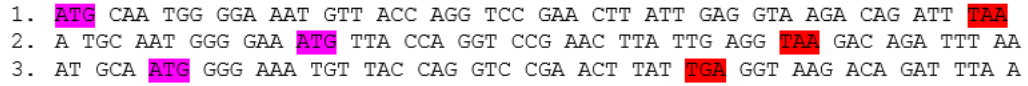
Start codon

- ATG

Open reading frame (ORF)

Stop codon

- TAA, TAG, TGA

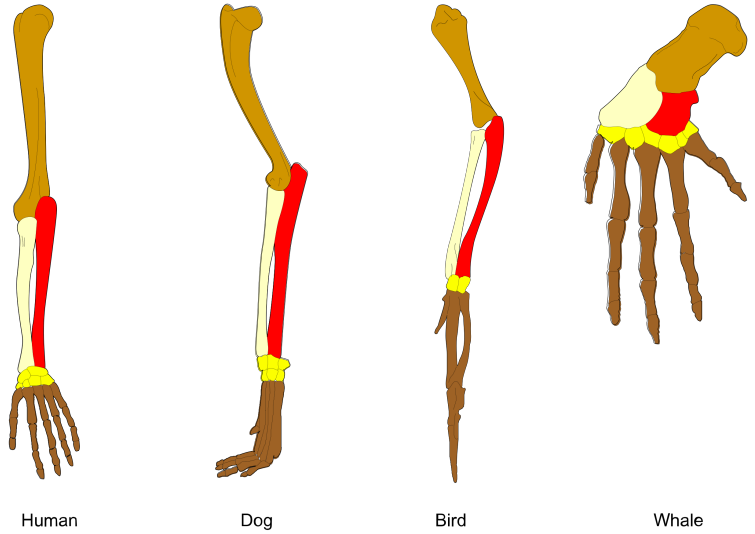


# Annotating genes

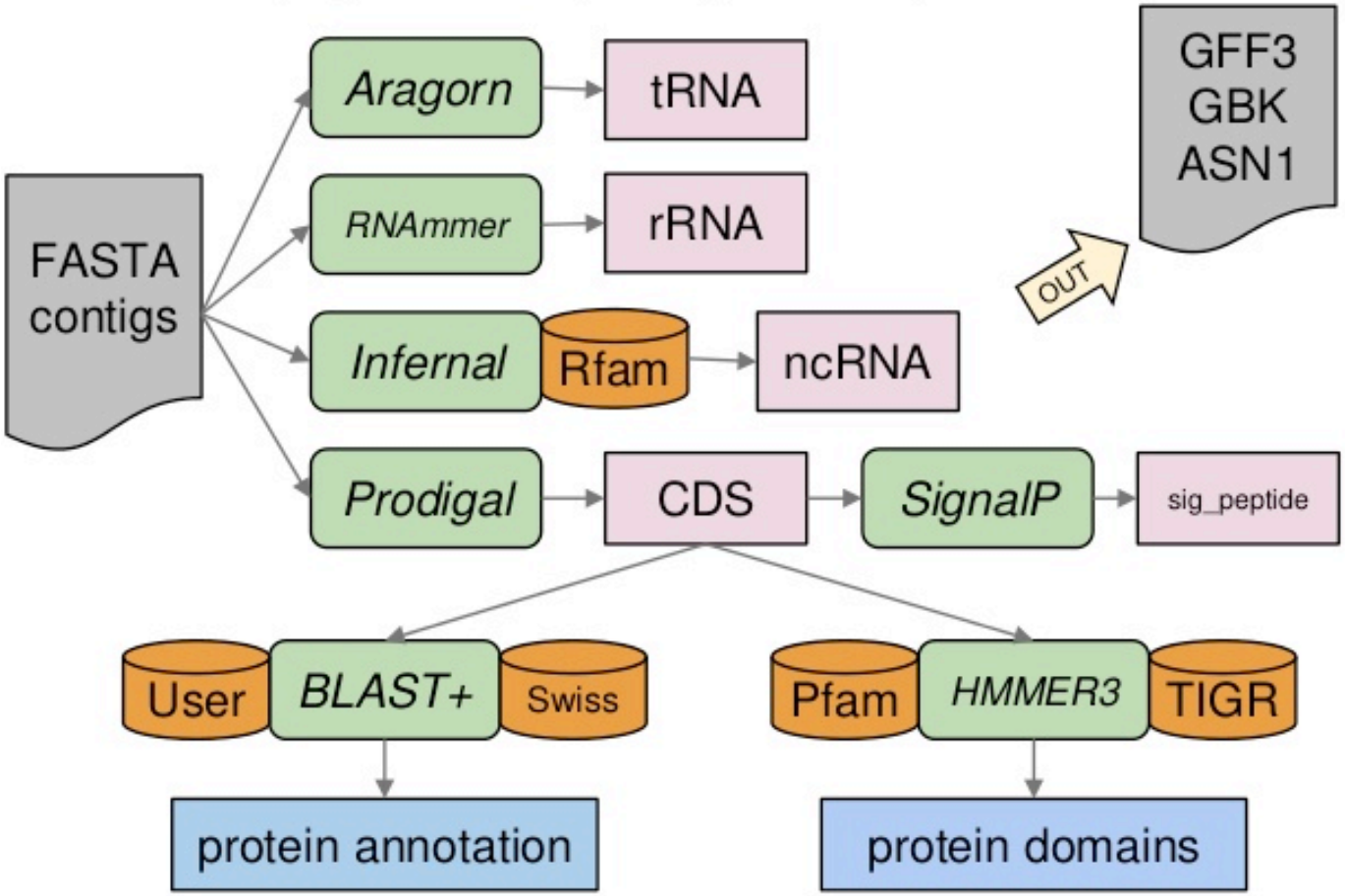
Homology

Similarity search against an annotated database

- NCBI
- KEGG
- COG
- SEED
- GO
- UNIPROT
- INTERPRO
- PFAM
- TIGR



# PROKKA: Rapid prokaryotic genome annotation




https://github.com/tseemann/prokka

# Let’s annotate our genome

https://github.com/igorspp/MMB-114

(**Day 5:** Genome annotation)