Genome assembly and annotation

Day 5: Genome annotation

Antti Karkman

Department of Microbiology – UH

antti.karkman@helsinki.fi

Slides modified from Igor Pessi, with permission

Aims for this part of MMB-114

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

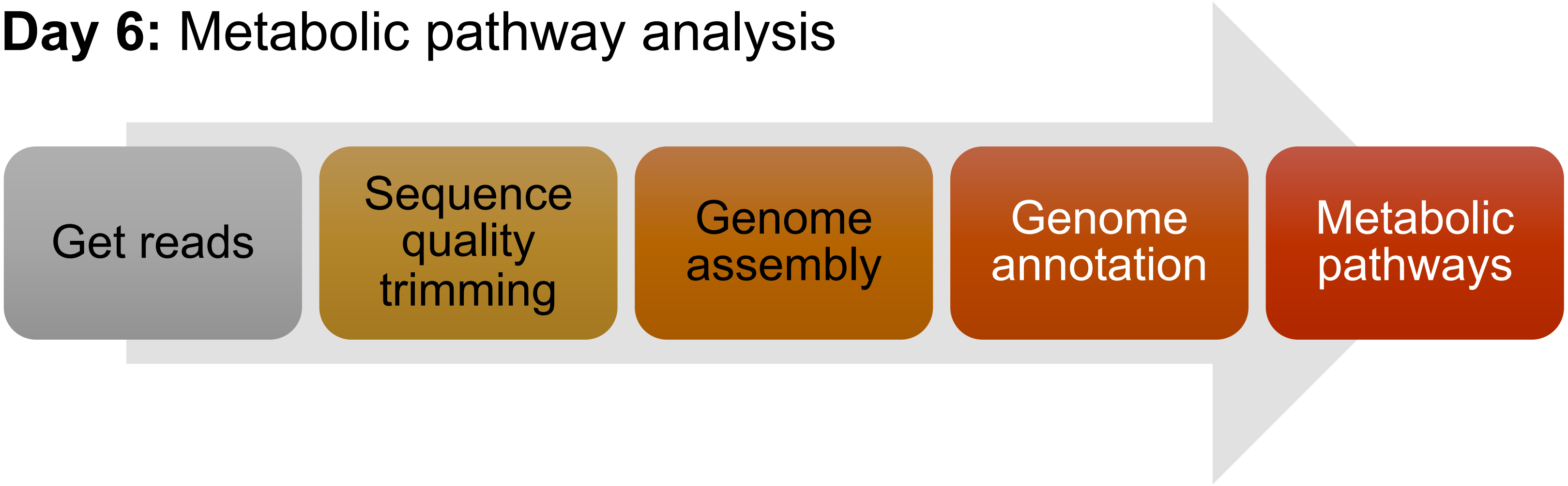
Day 2: Handling of Illumina data

Day 3: Genome assembly

Day 4: Check-up and report

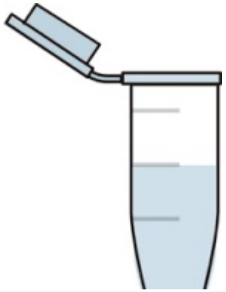
Day 5: Genome annotation

Day 6: Metabolic pathway analysis



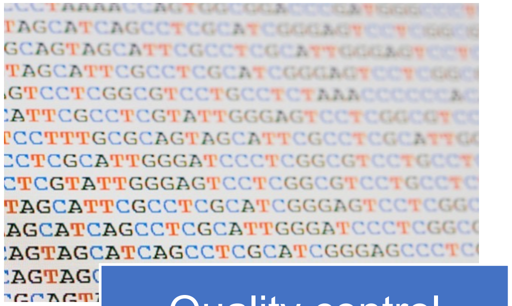
Recap from last week:

Isolation

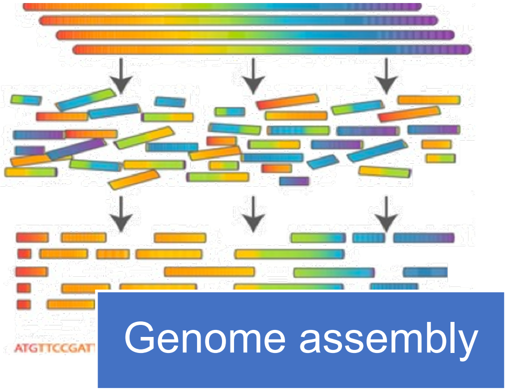


DNA extraction

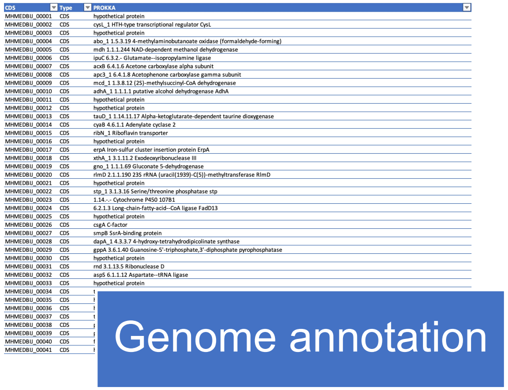
Sequencing



Quality control



Genome assembly



CDS	Type	PROKKA
MHMEDBIJ_00001	CDS	hypothetical protein
MHMEDBIJ_00002	CDS	cysL_1 HTH-type transcriptional regulator CysL
MHMEDBIJ_00003	CDS	hypothetical protein
MHMEDBIJ_00004	CDS	abo_1 1.5.3.19 4-methylaminobutanoate oxidase (formaldehyde-forming)
MHMEDBIJ_00005	CDS	mdh 1.1.1.244 NAD-dependent methanol dehydrogenase
MHMEDBIJ_00006	CDS	ipuC 6.3.2.- Glutamate--isopropylamine ligase
MHMEDBIJ_00007	CDS	acxB 6.4.1.6 Acetone carboxylase alpha subunit
MHMEDBIJ_00008	CDS	apc3_1 6.4.1.8 Acetophenone carboxylase gamma subunit
MHMEDBIJ_00009	CDS	mcd_1 1.3.8.12 (2S)-methylsuccinyl-CoA dehydrogenase
MHMEDBIJ_00010	CDS	adhA_1 1.1.1.1 putative alcohol dehydrogenase AdhA
MHMEDBIJ_00011	CDS	hypothetical protein
MHMEDBIJ_00012	CDS	hypothetical protein
MHMEDBIJ_00013	CDS	tauD_1 1.14.11.17 Alpha-ketoglutarate-dependent taurine dioxygenase
MHMEDBIJ_00014	CDS	cyaB 4.6.1.1 Adenylate cyclase 2
MHMEDBIJ_00015	CDS	ribN_1 Riboflavin transporter
MHMEDBIJ_00016	CDS	hypothetical protein
MHMEDBIJ_00017	CDS	erpA Iron-sulfur cluster insertion protein ErpA
MHMEDBIJ_00018	CDS	xthA_1 3.1.11.2 Exodeoxyribonuclease III
MHMEDBIJ_00019	CDS	gno_1 1.1.1.69 Gluconate 5-dehydrogenase
MHMEDBIJ_00020	CDS	rlmD 2.1.1.190 23S rRNA (uracil(1939)-C(5))-methyltransferase RlmD
MHMEDBIJ_00021	CDS	hypothetical protein
MHMEDBIJ_00022	CDS	stp_1 3.1.3.16 Serine/threonine phosphatase stp
MHMEDBIJ_00023	CDS	1.14.-.- Cytochrome P450 107B1
MHMEDBIJ_00024	CDS	6.2.1.3 Long-chain-fatty-acid--CoA ligase FadD13
MHMEDBIJ_00025	CDS	hypothetical protein
MHMEDBIJ_00026	CDS	csgA C-factor
MHMEDBIJ_00027	CDS	smpB SsrA-binding protein
MHMEDBIJ_00028	CDS	dapA_1 4.3.3.7 4-hydroxy-tetrahydrodipicolinate synthase
MHMEDBIJ_00029	CDS	gppA 3.6.1.40 Guanosine-5'-triphosphate,3'-diphosphate pyrophosphatase
MHMEDBIJ_00030	CDS	hypothetical protein
MHMEDBIJ_00031	CDS	rnd 3.1.13.5 Ribonuclease D
MHMEDBIJ_00032	CDS	aspS 6.1.1.12 Aspartate--tRNA ligase
MHMEDBIJ_00033	CDS	hypothetical protein
MHMEDBIJ_00034	CDS	t
MHMEDBIJ_00035	CDS	h
MHMEDBIJ_00036	CDS	h
MHMEDBIJ_00037	CDS	t
MHMEDBIJ_00038	CDS	p
MHMEDBIJ_00039	CDS	p
MHMEDBIJ_00040	CDS	f
MHMEDBIJ_00041	CDS	h

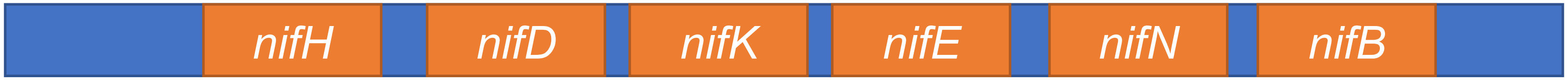
Genome annotation

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

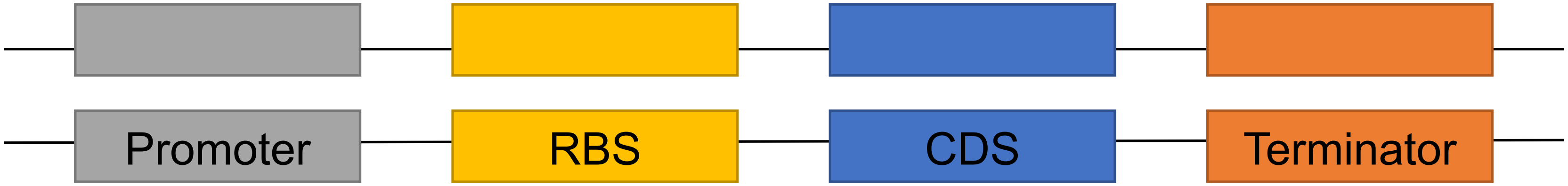
Ribosome binding site (RBS)

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

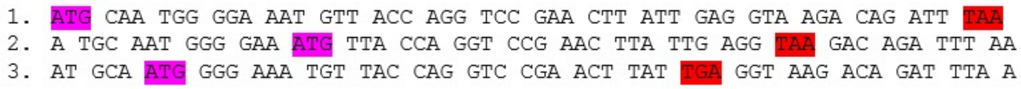
Start codon

- ATG

Open reading frame (ORF)

Stop codon

- TAA, TAG, TGA



```
1. ATG CAA TGG GGA AAT GTT ACC AGG TCC GAA CTT ATT GAG GTA AGA CAG ATT TAA
2. A TGC AAT GGG GAA ATG TTA CCA GGT CCG AAC TTA TTG AGG TAA GAC AGA TTT AA
3. AT GCA ATG GGG AAA TGT TAC CAG GTC CGA ACT TAT TGA GGT AAG ACA GAT TTA A
```

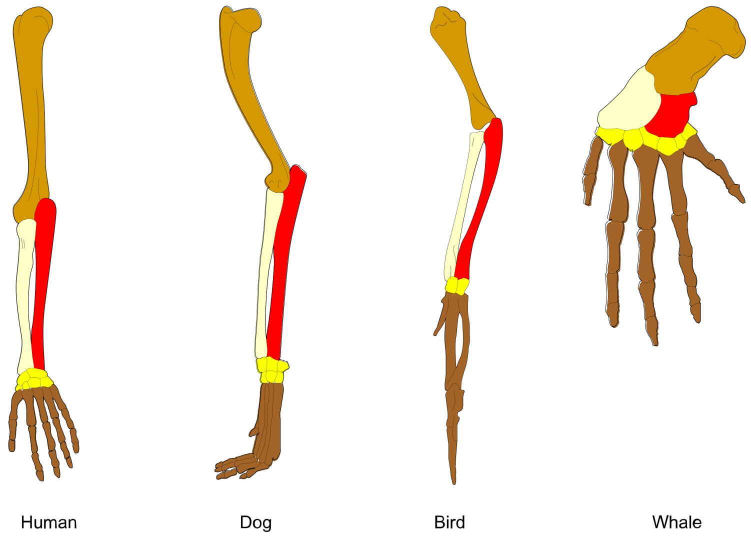
Annotating genes

Homology

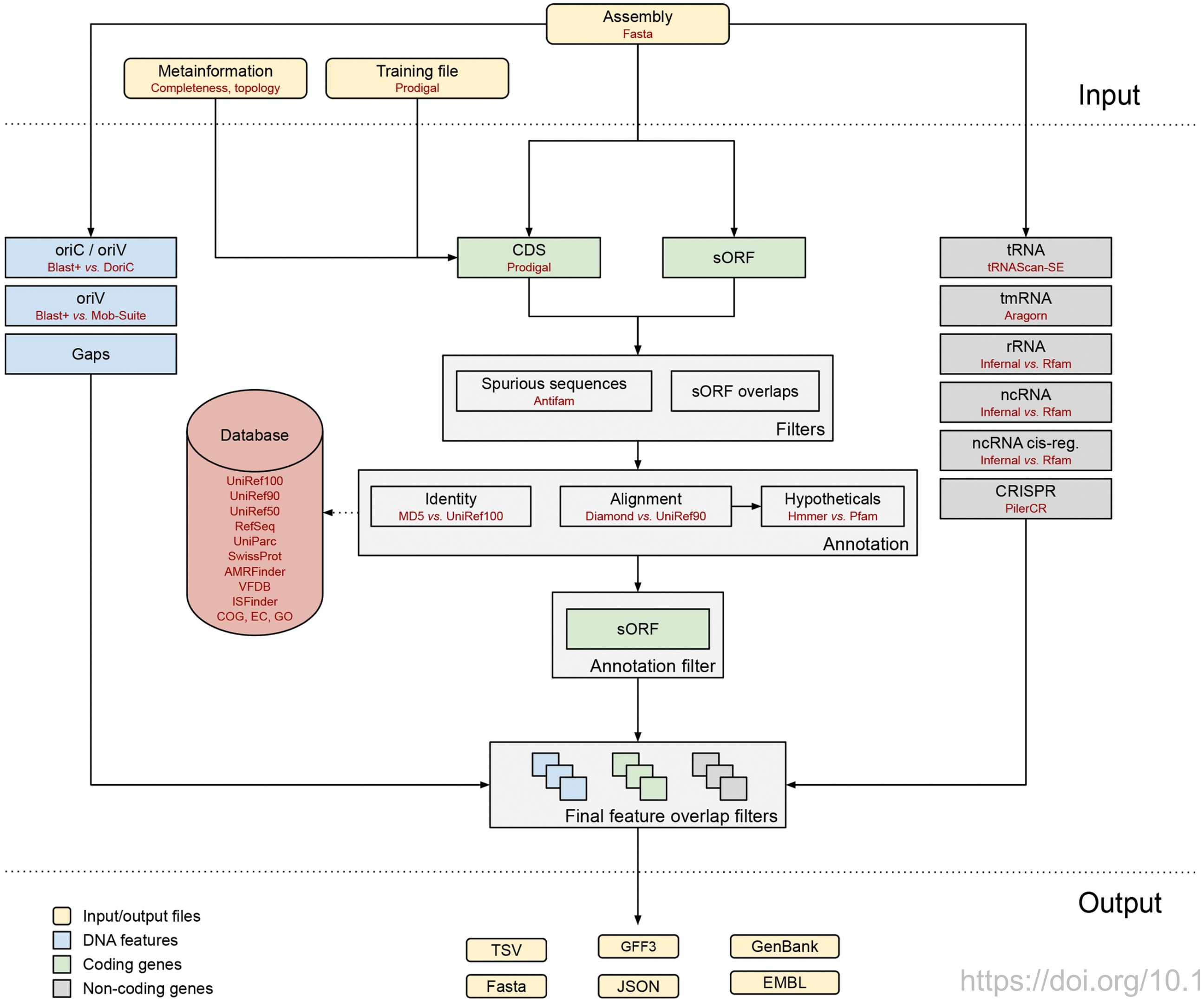
Statistical modelling of protein families/domains

Annotated databases

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



BAKTA: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification



https://doi.org/10.1099/mgen.0.000685

Let’s annotate our genome

https://github.com/karkman/MMB-114_Genomics

(**Day 5:** Genome annotation)