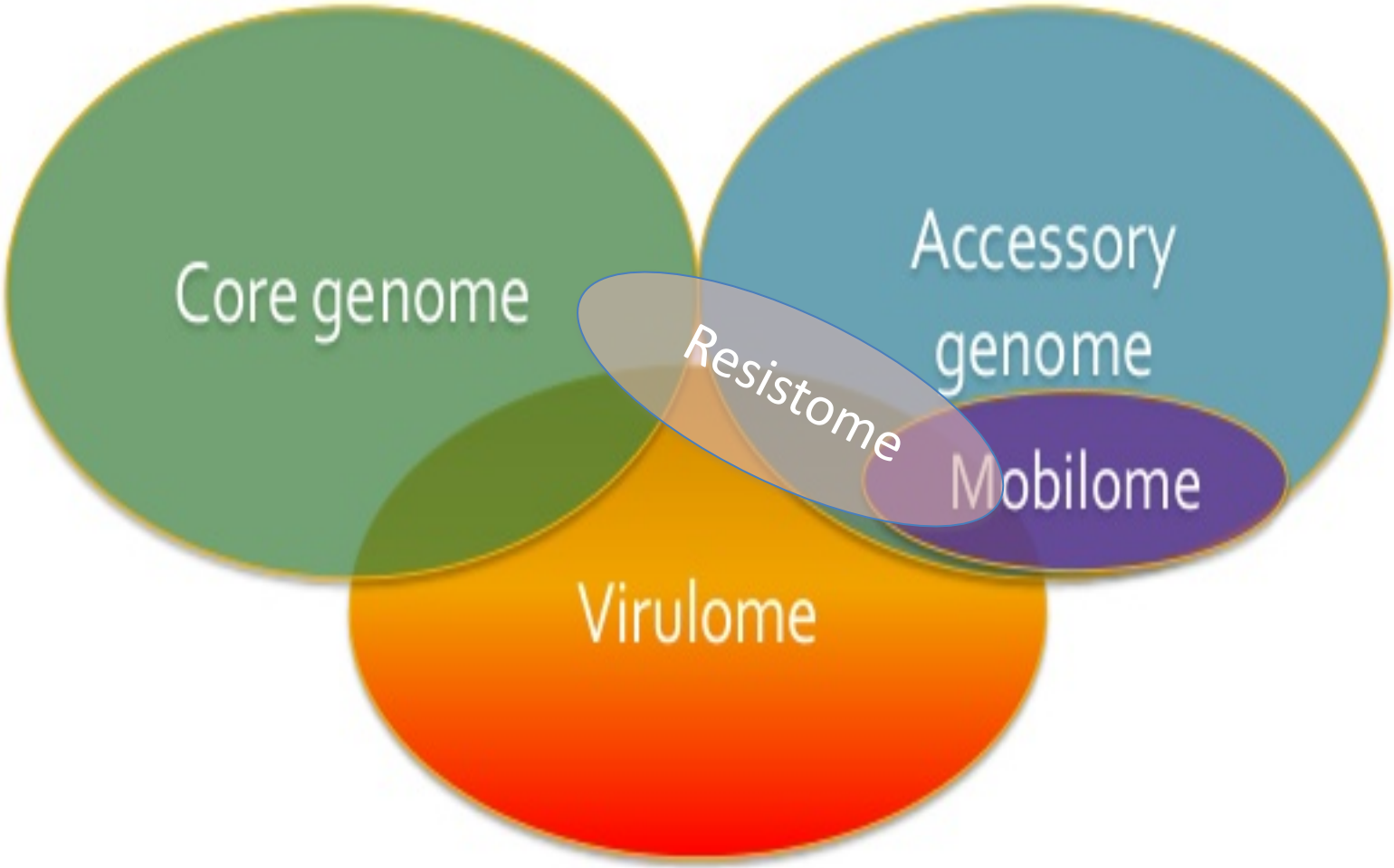
Notions de bases sur l'analyse des génomes II

(annotation, MLST, gènes de résistance et de virulence)

Théorie #4

Dr A. Jamet
Médecin Microbiologiste @Necker

Omics! Omics!



Slide credit: J. Carriço

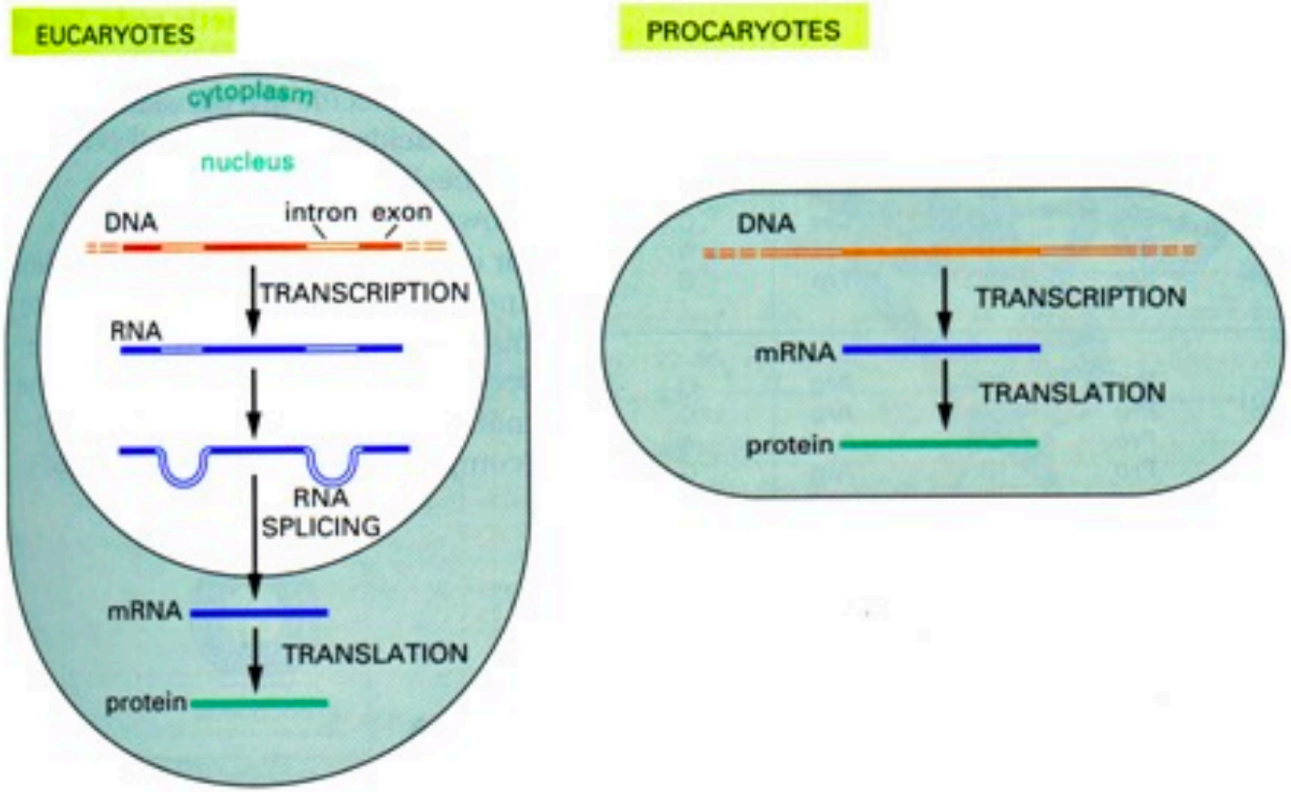
Size of Genomes

- Viral
  - 5 to 50 kilobases - KB - (1.000 bp)
- Prokaryotes
  - 0.5 to 12 megabases - MB - (1.000.000 bp)
- Eukaryotes
  - 8 megabases to 670 gigabases - GB- (1.000.000.000 bp)
  - High amount of repetitive DNA

Not all the DNA codes genes

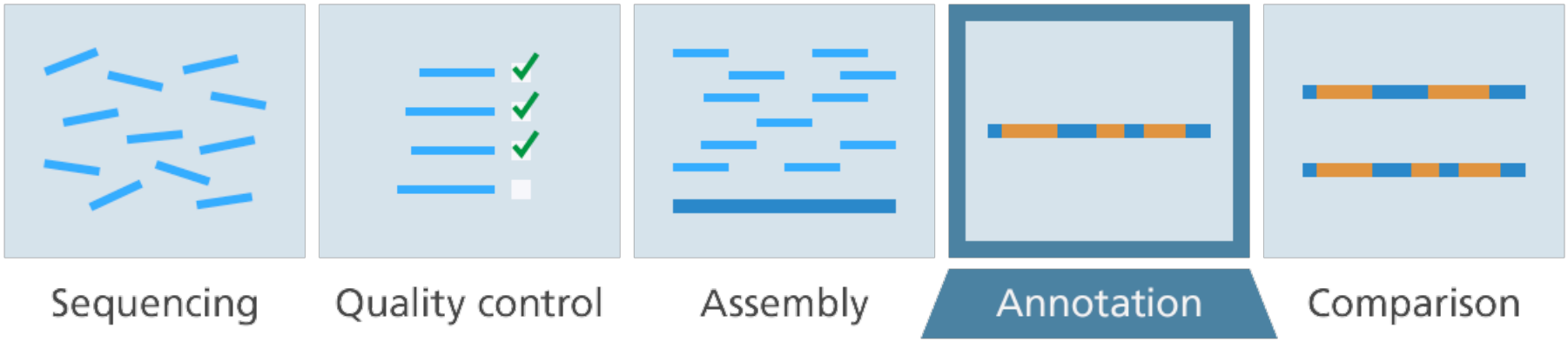
Organismo	Num. de pb	Genes	Descrição
ΦX-174	5 386	10	*E.coli* virus
Human mitochondrion	16 569	37	subcelular organell
Mycoplasma pneumoniae	816 394	680	Pneumonia
Hemophilus influenzae	1 830 138	1 738	Ear infection
E. coli	4.6 x 10^6	4 406	
Saccharomyces cerevisiae	12.1 x 10^6	5 885	yeast
C. elegans	95.5 x 10^6	19 099	worm
Drosophila melanogaster	180 x 10^6	13 601	House fly
Human	3 200 x 10^6	22 000 ?	Humans

It is simpler in prokaryotes !



Genome: 10Mbp-670Gbp Human: 3Gbp	Genome: 0.5-10Mbp
1% protein coding	>90% protein coding
Many repetitive sequences	Few repetitive sequences
Gene: exon structure	Gene: single contiguous stretch

Genome annotation



- The contigs we assembled contain different genomic features including
  - protein-coding genes, as well as other functional genome units such as structural
  - rRNAs, tRNAs, small RNAs, pseudogenes, insertion sequences, transposons

- The process of identifying and labelling those features is called genome annotation.
  - **It starts by identifying open reading frames** (ORFs).
  - **Functional annotation** of predicted sequences by searching for similarity to known elements (BLAST,...)


>Genomics DNA……..
atgcatgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaat
gcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgct
aagctgggatccgatgacaatgcatgcggctatgctaatgaatggtcttgggattt
accttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaat
ggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctgggat
ccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatccg
atgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctgggatc
cgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatcctg
cggctatgctaatgaatggtcttgggatttaccttggaatgctaagctgggatccg
atgacaatgcatgcggctatgctaatgaatggtcttgggatttaccttggaatatg
ctaatgcatgcggctatgctaagctgggaatgcatgcggctatgctaagctgggat
ccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatccg
atgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctcatgcg
gctatgctaagctgggaatgcatgcggctatgctaagctgggatccgatgacaatg
catgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgcta
agctgcggctatgctaatgcatgcggctatgctaagctcggctatgctaatgaatg
gtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggct
atgctaatgaatggtcttgggatttaccttggaatatgctaatgcatgcggctatg
ctaagctgggaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcg
gctatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgc
ggctatgctaatgcatgcggctatgctaagctcatgcgg
Gene!

ORF

Open Reading Frame (ORF) is a sequence of codons which starts with start codon, ends with an end codon and has no end codons in-between.

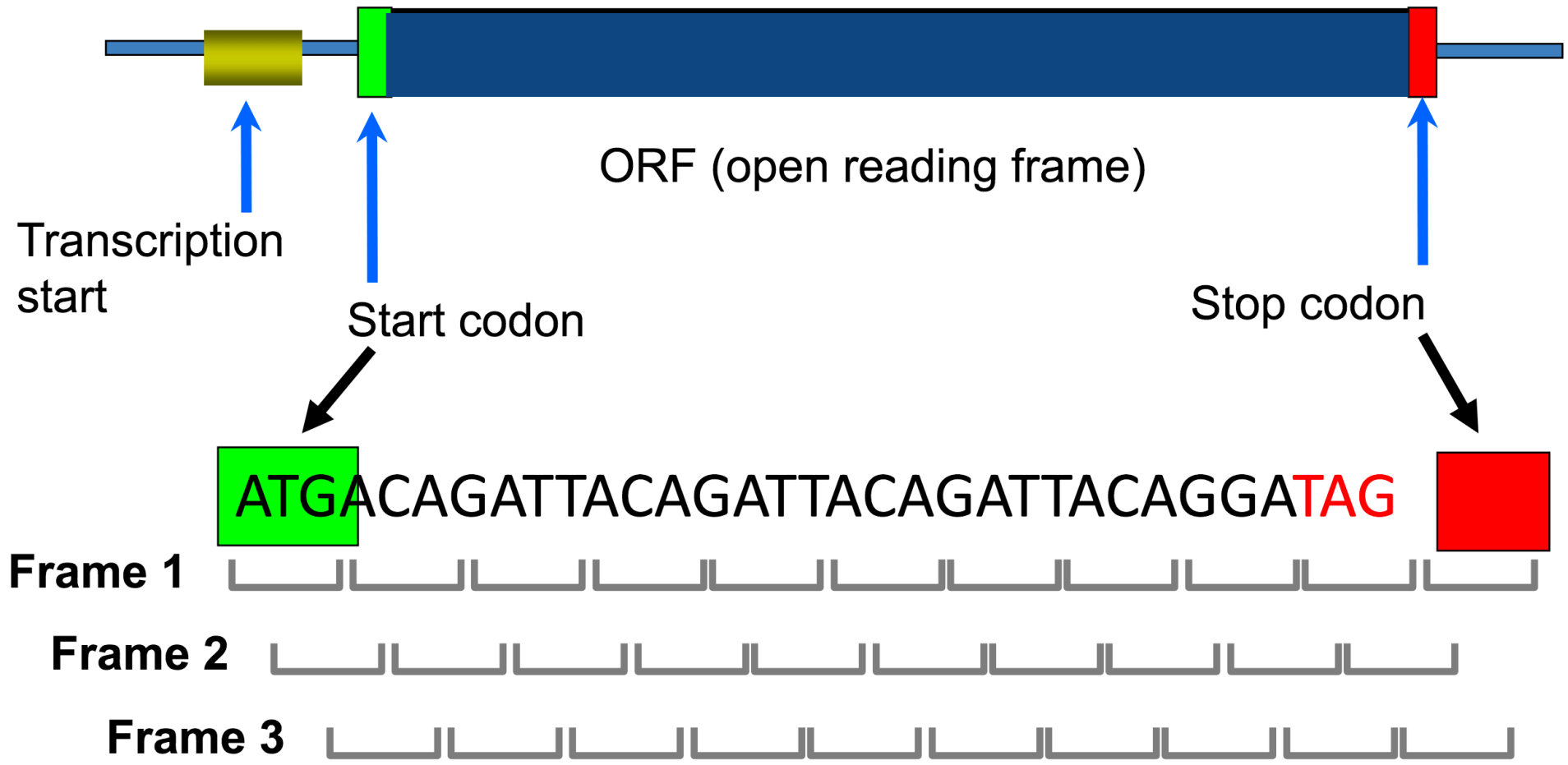
Searching for ORFs – consider all 6 possible reading frames: 3 forward and 3 reverse

Is the ORF a coding sequence?

1. Must be long enough (roughly 300 bp or more)
2. Should have average amino-acid composition specific for a give organism.
3. Should have codon use specific for the given organism.

In *E. coli* there are 6 500 ORFs but only 4 300 genes !!!

ORF


ORF (open reading frame)
Transcription start
Start codon
Stop codon
ATGACAGATTACAGATTACAGATTACAGGATAG
Frame 1
Frame 2
Frame 3

Questions biologiques

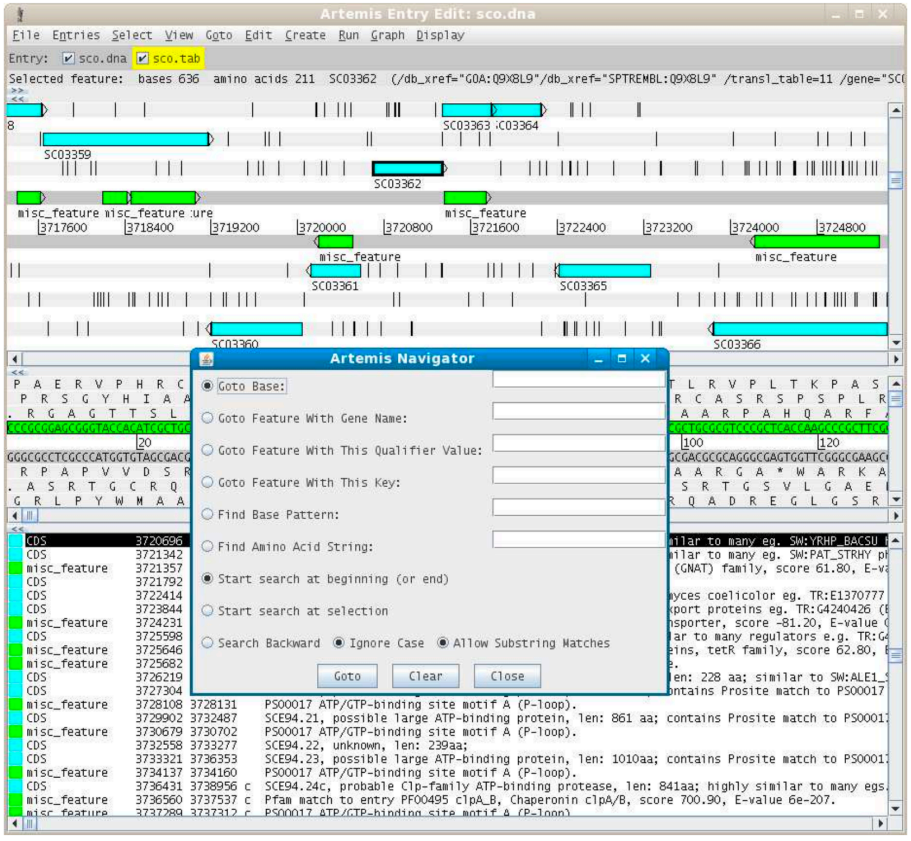
- Cette séquence contient-elle des gènes ?
- Ces gènes codent-ils pour des protéines, pour des ARNs ?
- Les protéines putatives codées par ces gènes sont-elles déjà connues ?
- Quelles sont leurs fonctions ?

Automated Annotation

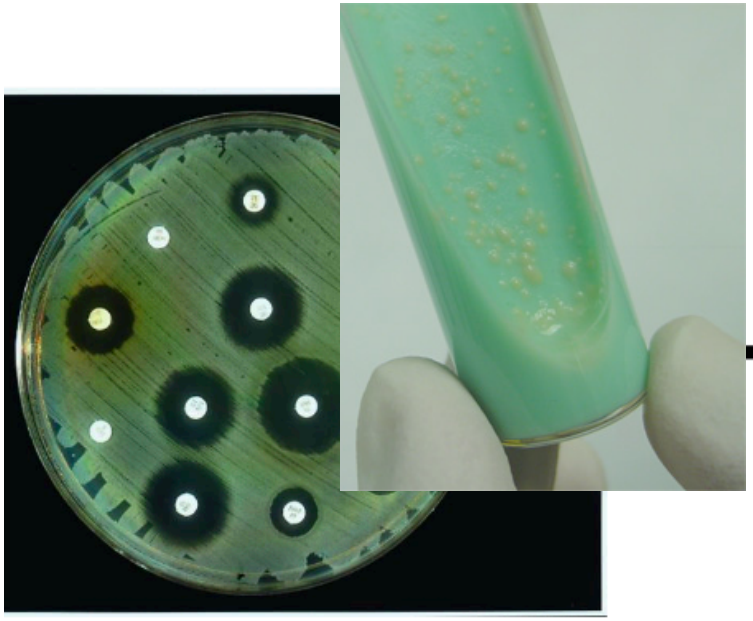
- We will annotate all genomes with an automated approach
- **Prokka or RAST** are pipeline scripts which coordinate a series of tools
  - locating ORFs and RNA regions on contigs,
  - translating ORFs to protein sequences,
  - searching for protein homologs
    - via BLAST and HMMER on the translated protein sequences as queries against a set of databases (CDD, PFAM, TIGRFAM...)
  - producing standard output files.
- The names of the contigs produced by SPades are quite long. PROKKA needs name which are shorter than 20 characters…

Annotation visualization

- With Artemis
- Artemis is a free genome viewer and annotation tool that allows visualization of sequence features and the results of analyses within the context of the sequence, and its six-frame translation. Artemis is written in Java, and is available for UNIX, GNU/Linux, BSD, Macintosh and MS Windows systems. It can read complete EMBL and GENBANK database entries or sequence in FASTA or raw format. Extra sequence features can be in EMBL, GENBANK or GFF format.



ATB



?

Fast growers (Staphylococcus, Enterobacteria...) ?
Slow growers (Mycobacteria) ?

Willem van Schaik

Variations conferring ATB resistance

- acquired genes encoding functions conferring resistance
- variation in genes/regulatory sequences through nucleotide substitution, insertion and deletion
- The « gold standard » issue !
  - Many of the discrepant results were found to be phenotypic errors in the routine laboratory.

What is an antibiotic resistance gene?

"a gene that confers resistance to an antibiotic in an otherwise susceptible microbial host"

Horizontally acquired resistance genes: e.g. *bla*KPC, *ermB*, *vanA*
associated with mobile genetic elements

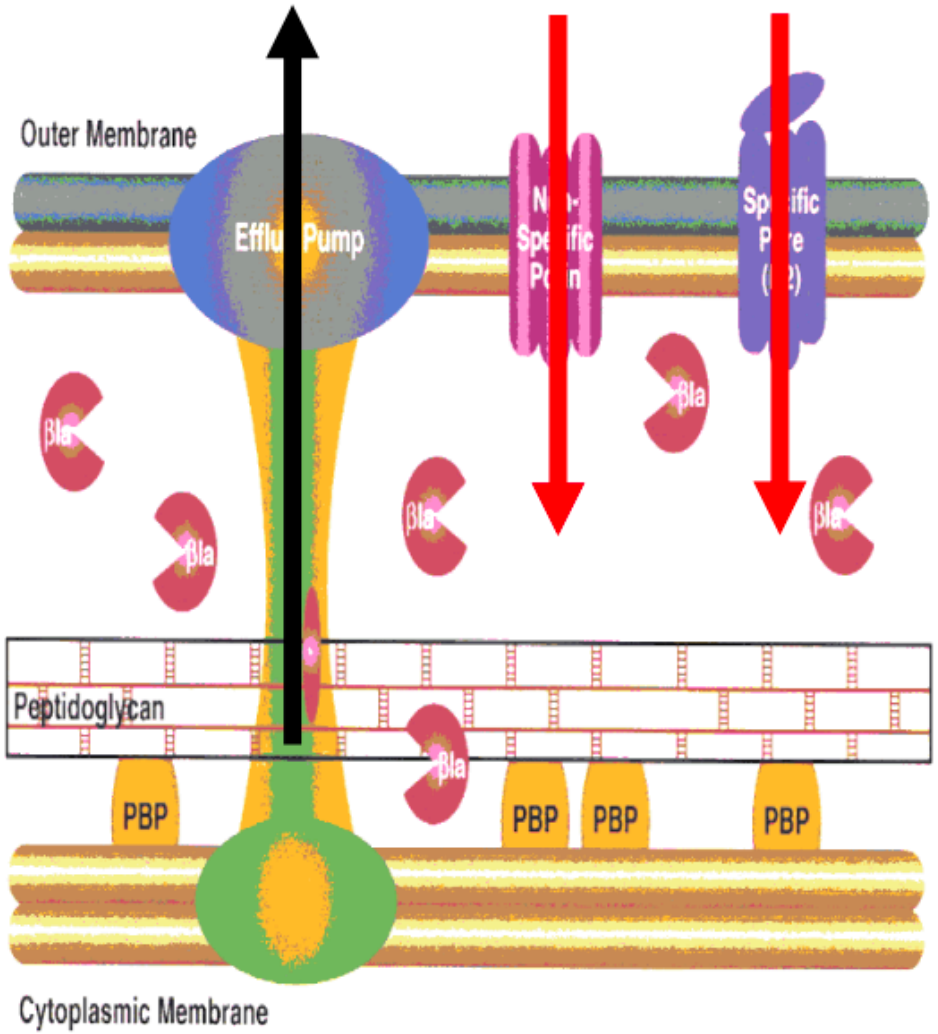
Mutations in housekeeping genes that confer resistance
e.g. target modification, porin mutation

Conserved genes that make a species intrinsically resistant
e.g. efflux pumps

A. Jamet

Willem van Schaik

Complex interplays determine an MIC



$V_{Entry} + V_{Efflux}$

External [drug]

Periplasmic [drug]

$V_{Hydrolysis}$

$V_{Binding}$

It's a lot more complex than gene presence / absence

Combinatorial resistance: WGS vs. AST

Table 1

Comparison ofWGS and Reference Laboratory Testing of Carbapenem-Resistant Gram-Negative Bacteria

Organism	Isolate No.	Phenotypic Resistance to Carbapenems and Third-Generation Cephalosporins	Attributable Resistance Mechanism According to Reference Laboratory[a]	Dominant Resistance Mechanism Based on WGS[b]
Acinetobacter baumannii	AB223	MEM, IPM[c]	OXA-23 carbapenemase	OXA-23 carbapenemase
Enterobacter cloacae	EC1a[d]	ETP, MEM, IPM, CTX, CAZ	IMP-1 carbapenemase	IMP-1 carbapenemase
E cloacae	EC302	ETP, CTX, CAZ	No carbapenemase genes detected. AmpC activity present	No carbapenemase genes detected. OmpF porin loss
Klebsiella pneumoniae	KP652	ETP, CTX, CAZ	No carbapenemase genes detected. ESBL activity consistent with CTX-M. ETP resistance consistent with porin loss	No carbapenemase genes detected. CTX-M-15 ESBL with OmpK36 porin loss
Escherichia coli	Eco216	ETP, CTX, CAZ	No carbapenemase genes detected. ESBL activity present. ETP resistance consistent with porin loss	No carbapenemase genes detected. CTX-M-15 ESBL with OmpF porin loss

Reuter et al., 2013. JAMA Intern Med 12;173:1397-404

Antibiotic Resistance Gene Databases

ARDB: no updates, many intrinsic resistance genes: do not use

CARD: frequently updated, based on ARDB. Contains some intrinsic resistance genes; database for 'resistance SNPs' but can give false positives

ResFinder

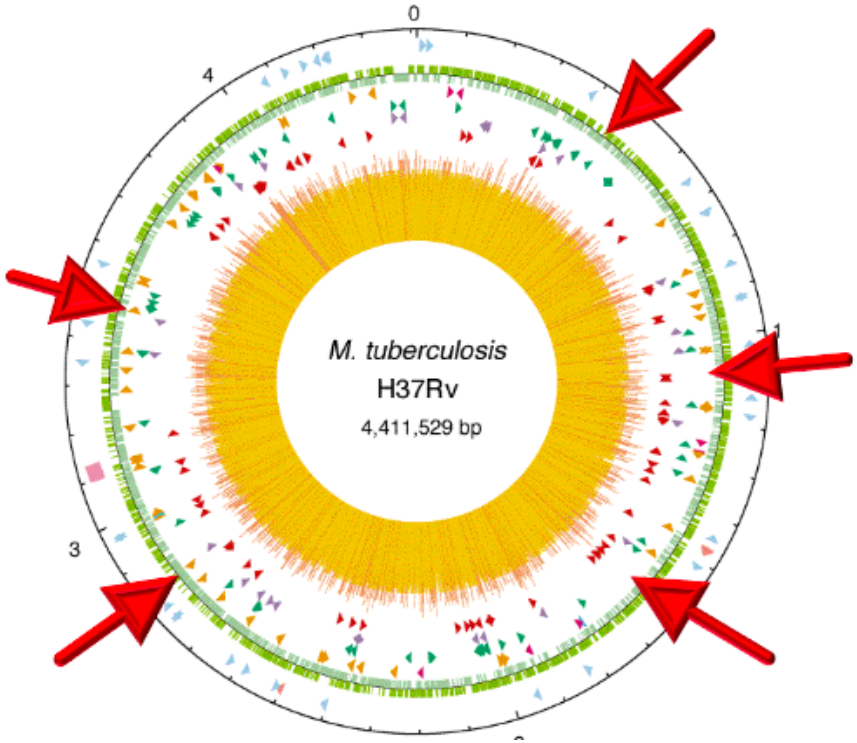
ARG-ANNOT

} Frequently updated. Acquired resistance genes only, small number of intrinsic resistance genes.

ResFams: prediction of resistance genes from metagenomes, many false positives or unvalidated genes

Tuberculosis

- No resistance plasmids
- No horizontal gene transfer



SNPs identification can allow resistance prediction

Gagneux

Resistance et lineage

Name: 5a99822338467

Sample: 5a99822338467

Drug[1]		Resistance	Supporting Mutations
Isoniazid	*katG inhA*		
Rifampicin	*rpoB*	R	rpoB (Gln432Pro)
Ethambutol	*EmbCAB*		
Pyrazinamide	*pncA*		

En ligne

TGS-TB et TB profiler

https://gph.niid.go.jp/tgs-tb/

http://tbdr.lshtm.ac.uk/

Strain	Lineage
MT0011_S8_L001	lineage4 Euro-American LAM;T;S;X;H None lineage4.1 Euro-American T;X;H None lineage4.1.2 Euro-American (X-type) T;H None lineage4.1.2.1 Euro-American (X-type) T1;H1 RD182
MT0010_S7_L001	lineage4 Euro-American LAM;T;S;X;H None lineage4.3 Euro-American (LAM) mainly-LAM None lineage4.3.1 Euro-American (LAM) LAM9 None
MT0008-2_S4_L001	lineage1 Indo-Oceanic EAI RD239 lineage1.1 Indo-Oceanic EAI3;EAI4;EAI5;EAI6 RD239 lineage1.1.2 Indo-Oceanic EAI3;EAI5 RD239
MT0005-2_S2_L001	lineage4 Euro-American LAM;T;S;X;H None lineage4.1 Euro-American T;X;H None lineage4.1.2 Euro-American (X-type) T;H None lineage4.1.2.1 Euro-American (X-type) T1;H1 RD182
1AB_S12_L001	lineage4 Euro-American LAM;T;S;X;H None lineage4.6 Euro-American T;LAM None lineage4.6.2 Euro-American T;LAM RD726 lineage4.6.2.2 Euro-American (Cameroon) LAM10-CAM RD726

Resistance et lineage

- Prédiction de la résistance aux ATB
  - *M. tuberculosis* (Rif, INH, Emb)

MYKROBE PREDICTOR TB Se 82.6% Sp 98.5%

En local

R RIFAMPICIN

Resistance mutation found: Q432X in gene rpoB
Resistant allele seen 34 times
Susceptible allele seen 0 times

S SPECIES

M. tuberculosis (lineage: European/American)

SUSCEPTIBLE

Isoniazid
Ethambutol
Quinolones
Streptomycin
Amikacin
Capreomycin
Kanamycin

RESISTANT

Rifampicin

Drug	Associated MIC (mg/L)	Mutation frequency among resistant isolates (%)	Compensatory mechanisms
Isoniazid: inhibition of cell wall mycolic acid synthesis			
katG	0.02–0.2	70	*oxyR'* and *ahpC*
inhA		~10	
kasA		~10	
Rifampicin: inhibition of RNA synthesis			
rpoB	0.05–1	95	*rpoA* and *rpoC*
Ethambutol: inhibition of cell wall arabinogalactan biosynthesis			
embB	1–5	~70	unknown
ubiA		~45, occurs with *embB* mutations	
Pyrazinamide: reduction of membrane energy; inhibition of trans-translation; inhibition of pantothenate and coenzyme A synthesis			
pncA	16–100	~99	unknown
rpsA		no clinical evidence	
panD		no clinical evidence	
Streptomycin: inhibition of protein synthesis			
rpsL	2–8	~6	unknown
rrs		<10	
gidB		clinical relevance to be determined	
Fluoroquinolones: inhibition of DNA synthesis			
gyrA	0.5–2.5	~90	*gyrA* (T80A and A90G)
gyrB		<5	putative *gyrB*
Capreomycin, amikacin and kanamycin: inhibition of protein synthesis			
rrs	2–4	60–70	*rrs* (C1409A and G1491T)
eis		~80 (low-level kanamycin)	
tlyA		~3 (capreomycin)	
Ethionamide: inhibition of cell wall mycolic acid synthesis			
ethA	2.5–25	mutations occurring in various combinations in these genes account for ~96% of ethionamide resistance	unknown
mshA			
ndh			
inhA			
inhA promoter			
Para-aminosalicylic acid: inhibition of folic acid and thymine nucleotide metabolism			
thyA	1–8	~40	unknown
folC		to be determined	
ribD		~90	

Issues

Only datasets passing QC metrics should be used for AST predictions, since resistance genes or mutations might be missed in sequences of poor quality.

Before WGS can be routinely implemented into accredited clinical practice there is a need to establish necessary minimum QC-thresholds

An inadequate limit of detection of WGS

- when detection is direct from clinical specimens e.g. TB
- for most organisms WGS is likely to use cultured (high titre) bacteria.

Short reads don't allow the assembly of plasmids

Issues

Need better standardisation of annotation of AMR genes

- BLAST analysis retrieves hits that are inconsistently annotated even where the actual sequences are identical.

Need a single, regularly updated 'challenge database' containing all validated AMR genes and chromosomal point mutations linked with AMR

Need international consensus on the criteria used to define genes as "new" or as variants of known genes.