



Biodiversity: Genome sequencing, bacterial population biology and genomic epidemiology

Emergence and evolution (of bacterial strains)

- Multidrug resistance
- Vaccine-escape
- Epidemiological surveillance
- One Health
- Links genotype-phenotype

Klebsiella pneumoniae

Multidrug resistance

Corynebacterium diphtheriae

Multidrug resistance

Bordetella pertussis

Vaccine-driven evolution



National Reference Center

Other pathogens

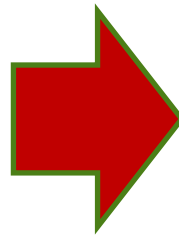
Public health importance

Genomic taxonomies of strains



Pedagogical objectives of the course

Biological concepts to understand strain diversity within bacterial species, and its epidemiological interpretation



- What are bacterial species?
- What are strains?
- How to interpret genetic diversity among strains within an epidemiological context
- Main strain definition approaches

Microbial diversity



Figures

- ~ 20 000 species of procaryotes
- $>10^7$ species estimated (?)
- Hundreds of pathogenic bacterial species
- Importance in medicine, agriculture, biotechnology, ...

Lopez-Garcia & Moreira 2007



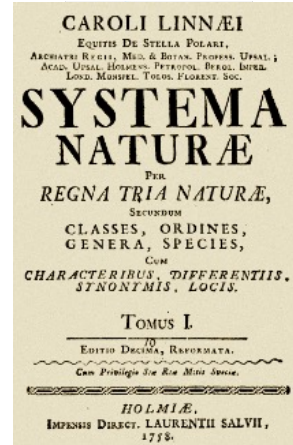
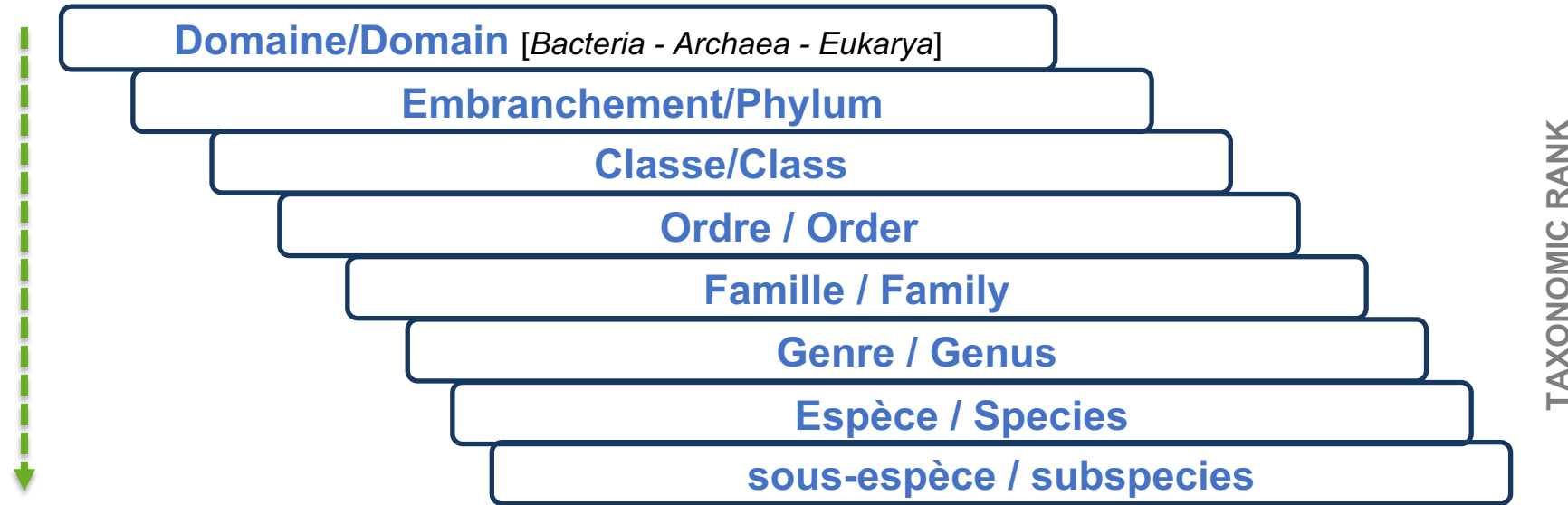
<https://lpsn.dsmz.de/>

LPSN - List of Prokaryotic names with Standing in Nomenclature

(Jean Euzéby *et al.*)

Taxonomy and nomenclature

Taxonomy : classification of life forms in a hierarchical system



Nomenclature : assignment of names to taxonomic objects

- Bi-nominal nomenclature system (Linnaeus, 1752)

Bordetella pertussis (whooping cough agent)

↑
genus

↑
species

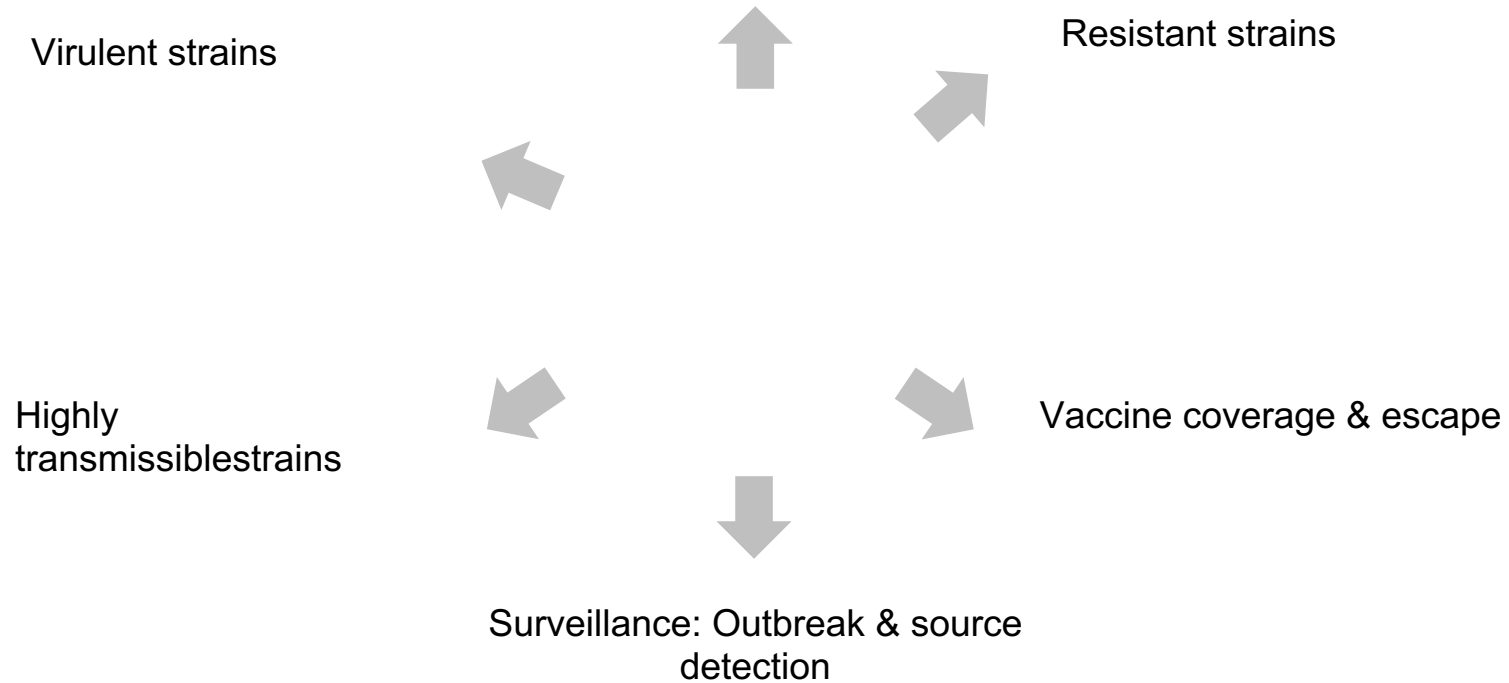
- Common language (lingua franca)
Diagnostics, Epidemiology, pathophysiology, ...

‘Biological esperanto’



Genotyping of microbial strains: why?

International dissemination



Why molecular typing?



To detect clusters or outbreaks

(Genomic surveillance,
Reverse epidemiology)

Figures

To find the infectious source

(Outbreak investigation)

Index case

Sources?

Question:

Figures

Strain typing

Figures

Same

Different

Linked (**rule-in**)

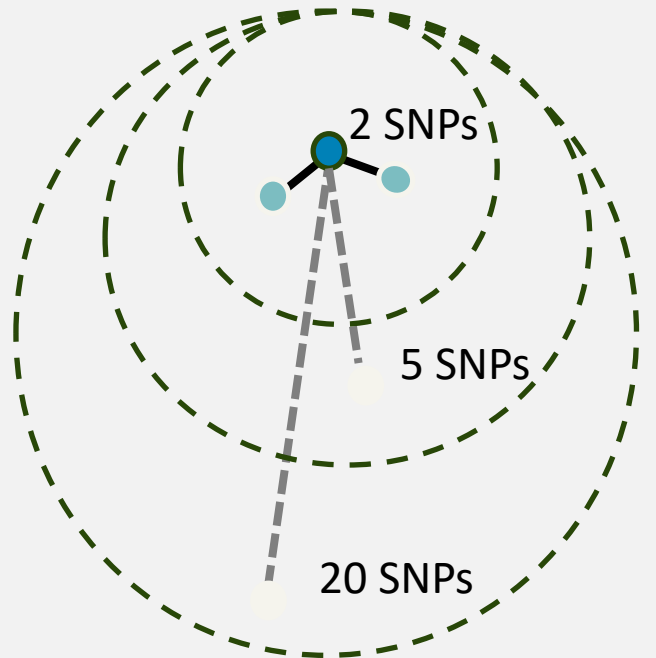
(**rule-out**)

... or maybe not:

Discrimination?

Strain widely distributed?

Unlinked to index case



Why is there strain diversity?

some microbial population biology questions



- Mechanisms that generate diversity?
- How large is the diversity within species?
- What are strains, sublineages, clones, isolates?
- How do strains differ in their characteristics? (gene content, clinical phenotype?)
- How do they evolve in face of antimicrobials, vaccination?

Bacterial evolution: drivers

Mutation

Recombination

Gene transfer

Figures



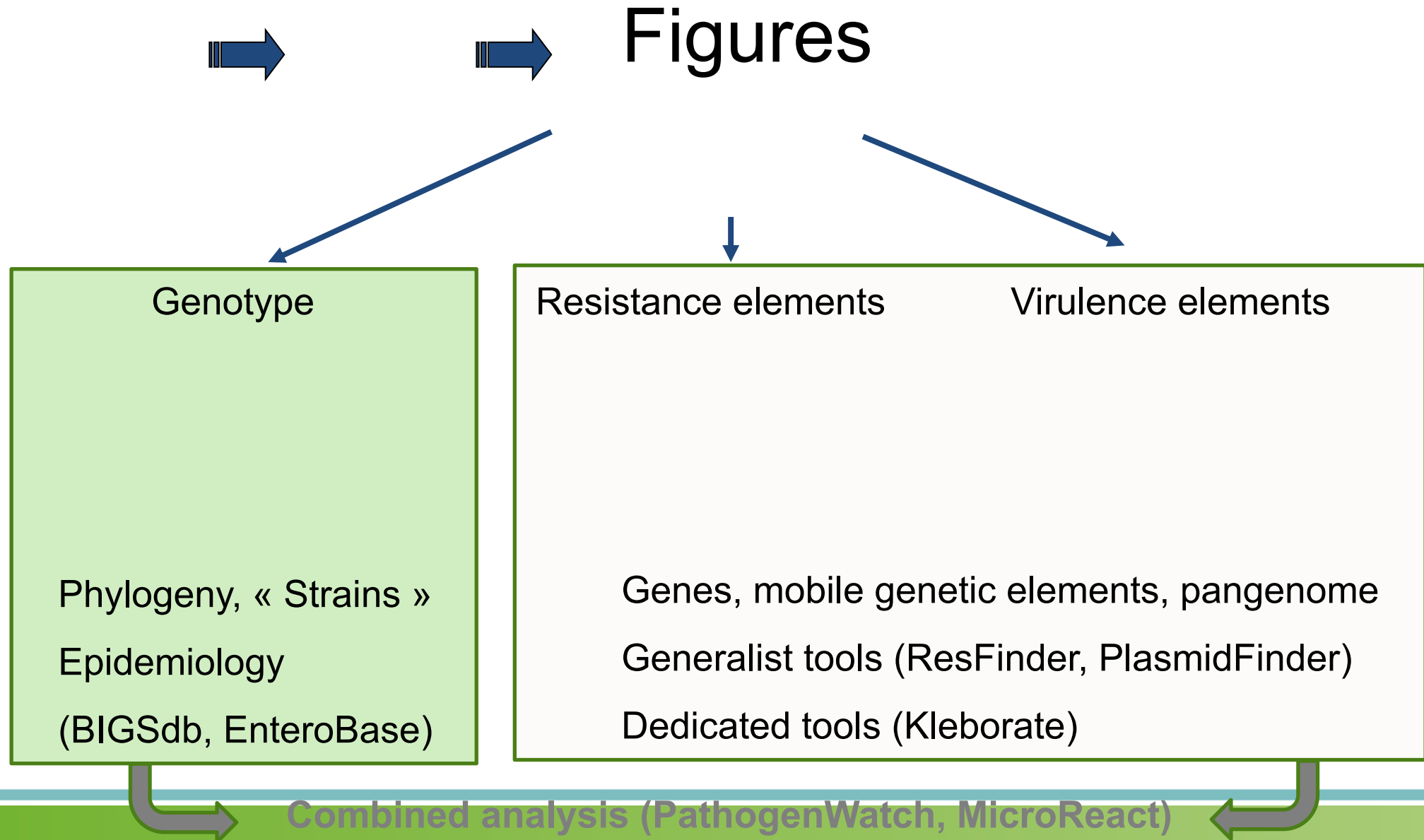
Large gene pool

Selection

Demographics & spread

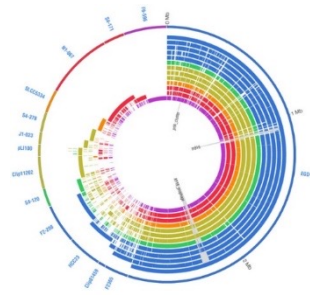
Figures

Genome sequencing: 'all in one technology'



Genomic data analysis strategies

Figures



De-novo assembly

Contig 1

Contig 2

Contig 3

whole genome sequence



Mapping



K-mers

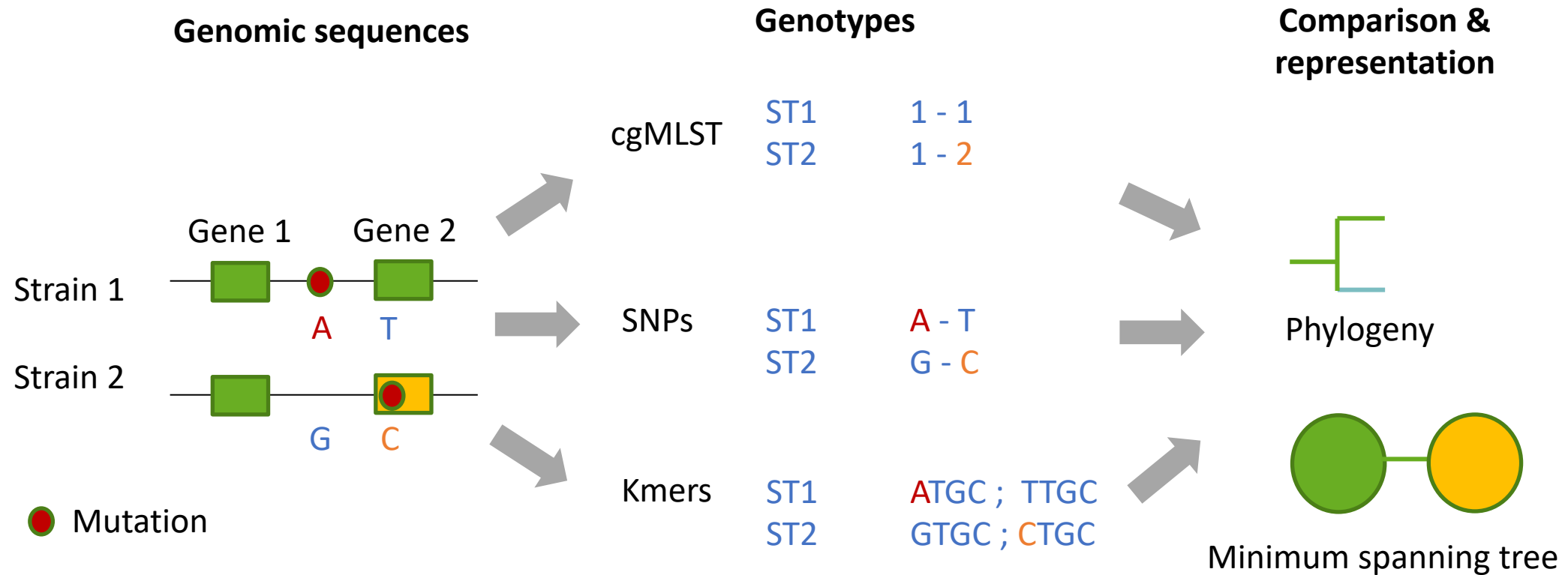


- Count & compare
- Derive genetic distance (e.g., MASH distance)



Single Nucleotide polymorphisms (SNPs)

Comparative genomics: 3 approaches

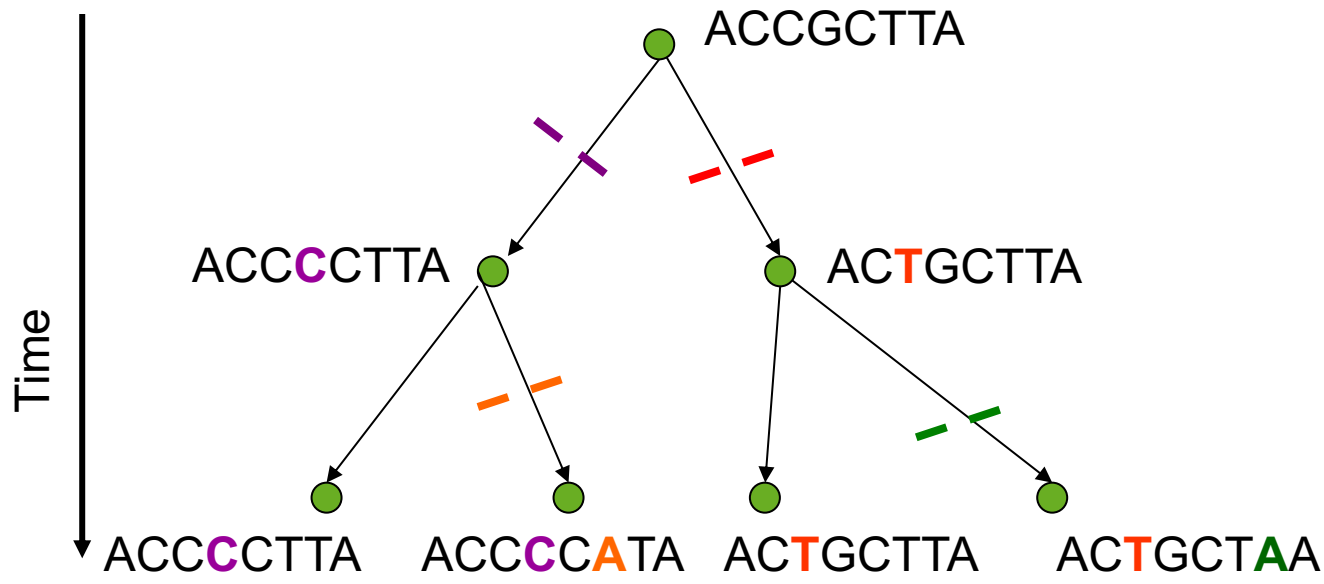
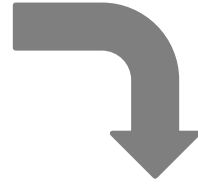


How many genetic differences define a 'strain' ?

Use of thresholds of genetic distance beyond which infectious agents can be defined as being unrelated to the transmission chain under investigation ('ruling-out')

Sequence-based phylogenetics

ACCCCTTA
ACCCATA
ACTGCTTA
ACTGCTAA



1. Sequence alignment
2. Evolutionary model

- Parsimony
- Distance
- Maximum likelihood
- Bayesian

- Groups interpreted as common ancestry (acquired changes)
- Branch lengths reflect number of changes

Phylogenies: dating evolution, inferring transmission



Haiti 2011

Cholera 7th pandemic

Mutreja *et al.*, 2011

— South East Asia
— Africa

Figures

Evolutionary rate:
~3 SNP per genome per year

Most bacterial species are very heterogeneous

Bacterial species threshold: ~5% different, i.e., 95% average nucleotide identity (ANI),
based on sequence divergence between common genes



Escherichia coli

Human- chimpanzee ~1.2%

Figures

Figures

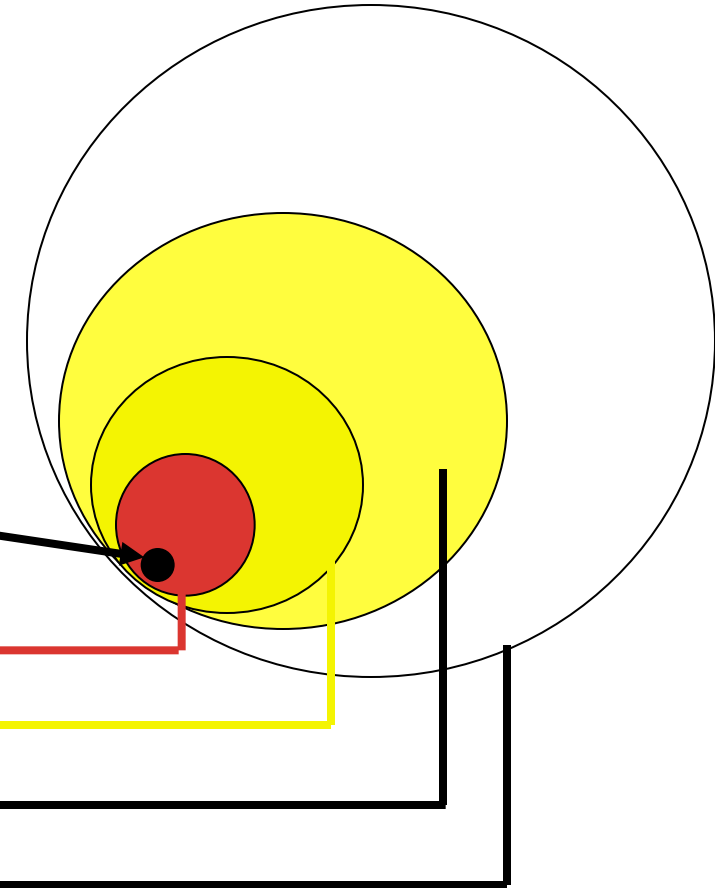
Many medically important bacterial species are 'monomorphic'

(π : Average number of nucleotide differences between two strains)

ATCGATGCCAGGCGTACAGGCGTAGGGTTTACGGGTTTAC
...T.....T.....C.....

$\pi \sim 1\text{-ANI}$

Species	π (%)	Age (y)
<i>M. tuberculosis</i>	< 0.01	35,000
<i>Salmonella</i> ser. Typhi	...	50,000
<i>Y. pestis</i>	...	10,000
<i>K. pneumoniae</i>	0.37	
<i>S. pyogenes</i>	1.02	
<i>E. coli</i>	1 - 2	
<i>L. monocytogenes</i>	3 clusters each 1-2%	
<i>N. meningitidis</i>	6.18	



[π in human species: 0.1 \Rightarrow 1 diff. (synonymous) every 1,000 nt]

Some taxonomic species are 'monomorphic'

Highly pathogenic taxonomic species derived from larger genomic species

Species	Infection	Ancestor	Reference
" <i>Salmonella typhi</i> "	Typhoid fever	<i>Salmonella enterica</i>	Selander et al. 1990
<i>Yersinia pestis</i>	Plague	<i>Yersinia pseudotuberculosis</i>	Achtman et al. 1999
<i>Shigella flexneri</i> , <i>S. boydii</i> , <i>S. dysenteriae</i> , <i>S. sonnei</i>	Shigellosis	<i>Escherichia coli</i>	Pupo et al. 2000
<i>Bacillus anthracis</i>	Anthrax	<i>Bacillus cereus</i>	Priest et al., 2004
<i>Burkholderia mallei</i>	Glanders	<i>Burkholderia pseudomallei</i>	Godoy et al. 2003
<i>Bordetella pertussis</i>	Whooping cough	<i>Bordetella bronchiseptica</i>	Diavatopoulos et al. 2005
<i>Bordetella parapertussis</i>	pertussis-like	<i>Bordetella bronchiseptica</i>	Diavatopoulos et al. 2005
<i>Mycobacterium ulcerans</i>	Buruli ulcer	<i>Mycobacterium marinum</i>	Stinear et al., 2007
<i>Mycobacterium tuberculosis</i>	Tuberculosis	<i>Mycobacterium prototuberculosis</i>	Gutierrez et al. 2005

➔ '*Nomen periculosum*' taxonomic code rule

Monomorphic species can be associated with discrepancies between phylogeny and nomenclature

- *Shigella* and *Escherichia coli* belong to same genomic species
- *Shigella* species are polyphyletic biovars

Figures

How to calculate ANI online: jSpeciesWS




(slow; but allows batches) <https://jspecies.ribohost.com/jspeciesws/>



Upload own genome (min 0.02MB
- max. 15MB)

Genome as (multi)-FASTA.

 Upload ZIP archive (New!)

 Select file

File WHOZ uploaded successfully at
2,105.63 kilobit per second

Choose genome from
GenomesDB 

Neisseria gonorrhoeae e03.04

ANId (Goris et al. IJSEM 2007)

Pieces of 1020 nt of query genome

BLASTN against reference

Take all hits > 30% identity over >70% length

Compute mean identity

ANIm:

based on MUMmer (Kurtz et al., Genome Biol. 2004)

Much faster

Comparison of own genomes and/or reference genomes

Running...



Genomes included (2/30)

Pairwise comparisons ?

Tetra correlation search ?

⚙️ Calculation in progress, click for status!

📊 Start ANIb

📊 Start ANIm

📊 Start Tetra

📋 Start TCS

Genome	Size	Contigs	GC [%]	N	Source	
WHOZ	2,229,351	1	52.4	0	Upload	✖
Neisseria gonorrhoeae e03.04	2,151,002	120	52.46	0	Public	✖

🛒 Genome Cart

📊 ANIb Result

📊 ANIm Result

📊 Tetra Result

📋 TCS Result

ANIb Matrix

[ANIb Result by Genome](#)

Show ANIb and [aligned nucleotides] [%] ▾

📄 Download as .csv

Legend:

Above cutoff (> 95%)

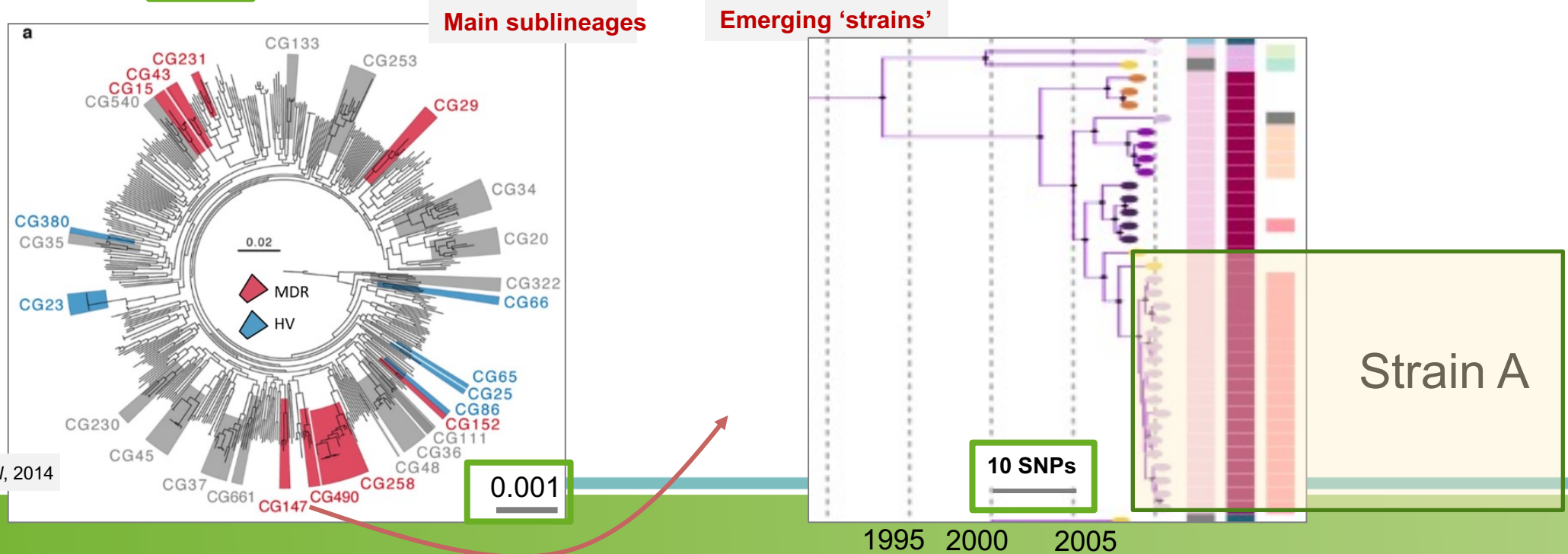
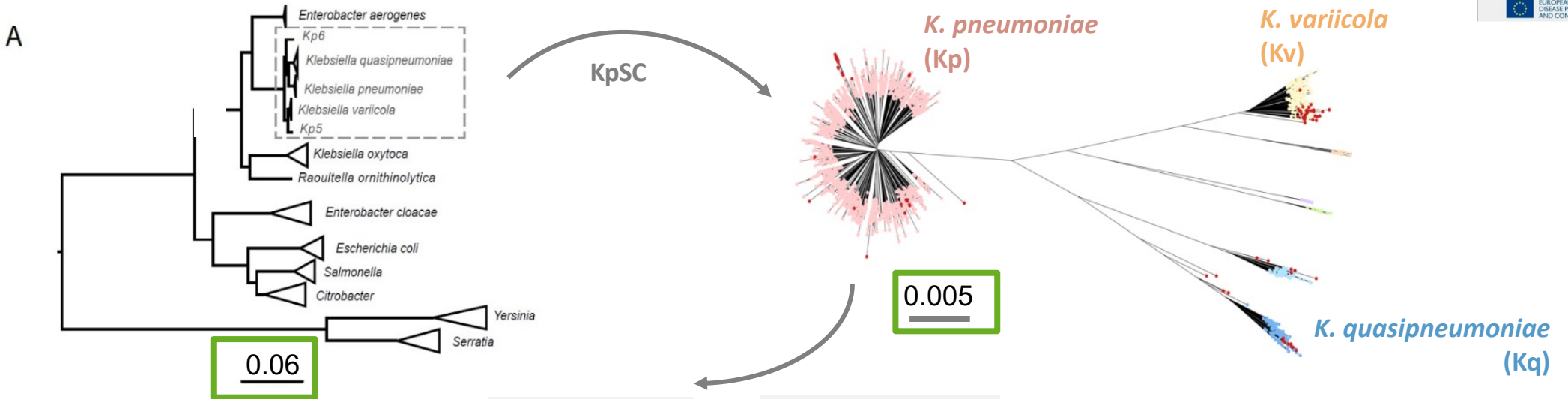
Below cutoff (< 95%)

Suspicious alignment!

	WHOZ	WHOA
WHOZ	*	99.26 (95.49)
WHOA	99.20 (95.75)	*

How distinct is *Escherichia coli* from *Salmonella enterica*?

The multiple levels of phylogenetic structure



Holt et al, 2014

Mind the scale!

Relationships between SNPs and genetic distance

SNPs: number of mutations

e.g., 10 SNPs

Genetic distance?

Relationships between SNPs and genetic distance

SNPs: number of mutations

e.g., 10 SNPs

Genetic distance?

Genome length: 5 MB (5 000 000 bases, or nucleotides)

$$10 / 5\,000\,000 = 0.000002$$



For small scales/distances, conversion into number of SNPs

The discrimination scale of typing methods

Diagnostic PCR,
MALDI-TOF MLST
RAPD rMLST
ERIC-PCR PFGE cgMLST Whole genome SNPs

Figures

→ All SNPs seen

One in 2

...

...

One in 1000

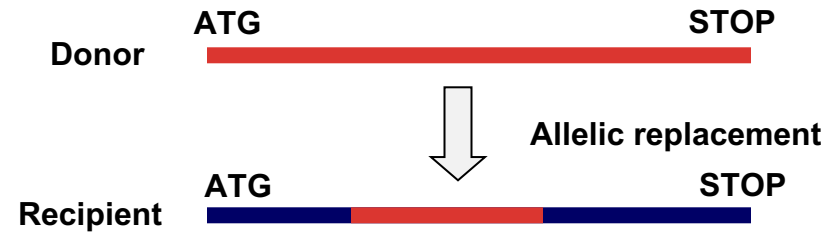


**Low
discrimination**

Image: Medini et al. 2008

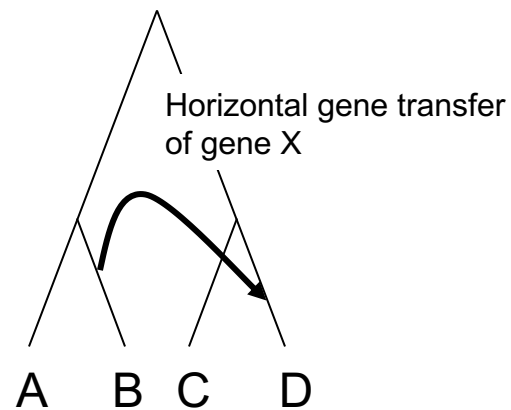
**High
discrimination**

Homologous recombination and phylogeny

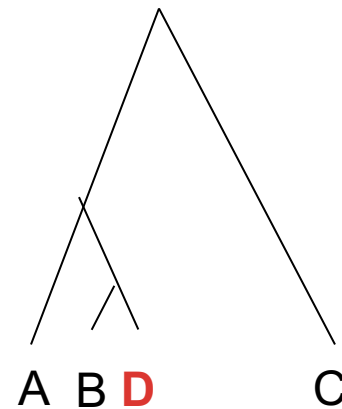


- Single-gene phylogeny: homologous recombination results in phylogenetic misplacement of recipients

True phylogeny



Gene X-based phylogeny

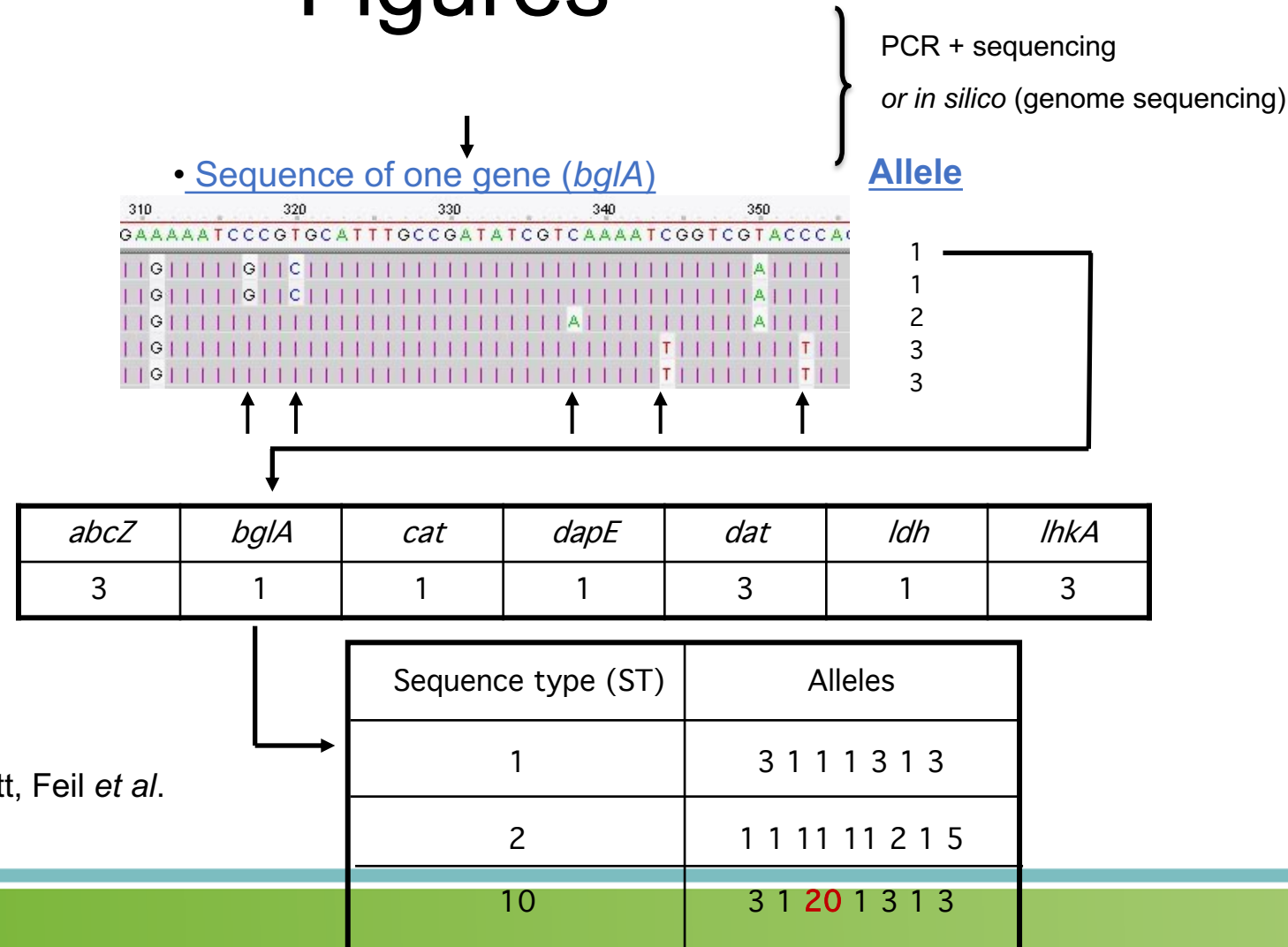


- Can we reconstruct the genealogy (phylogeny) of bacteria?

Gene-by-gene analyses of bacterial genomes: Multilocus Sequence Typing (MLST)

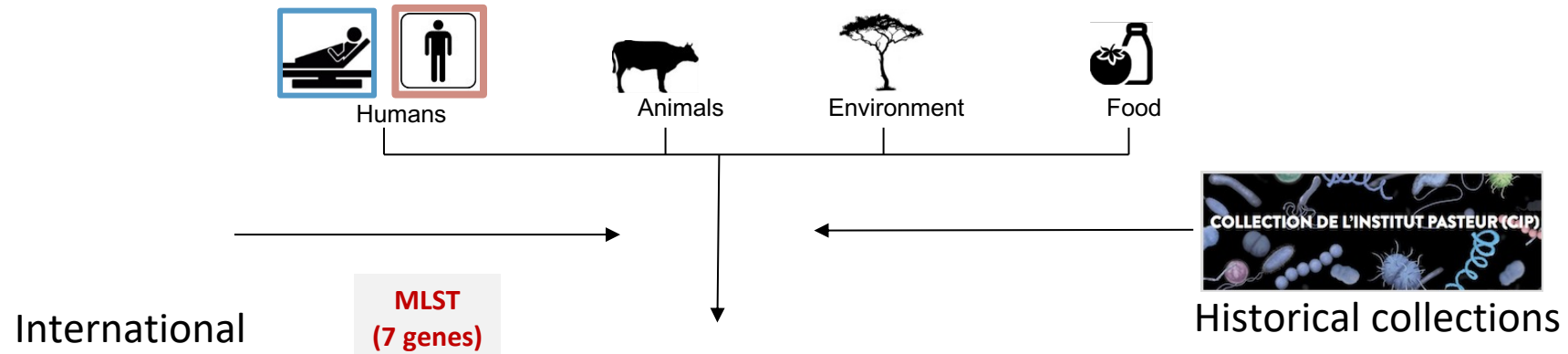
- Bacteria
- Fungi
- Parasites

Figures



Maiden, Achtman, Spratt, Feil *et al.*

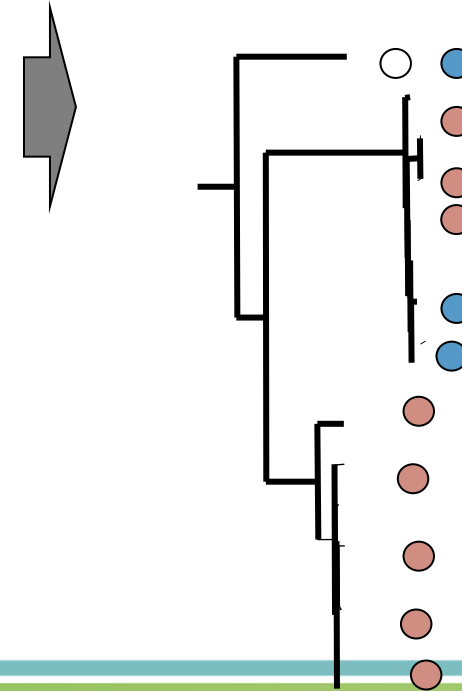
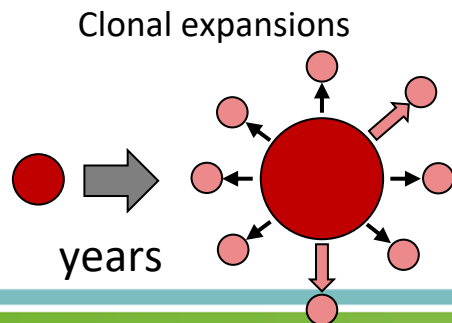
Defining the population structure of bacterial species: MLST *versus* phylogeny



Network analyses
(epidemiology)

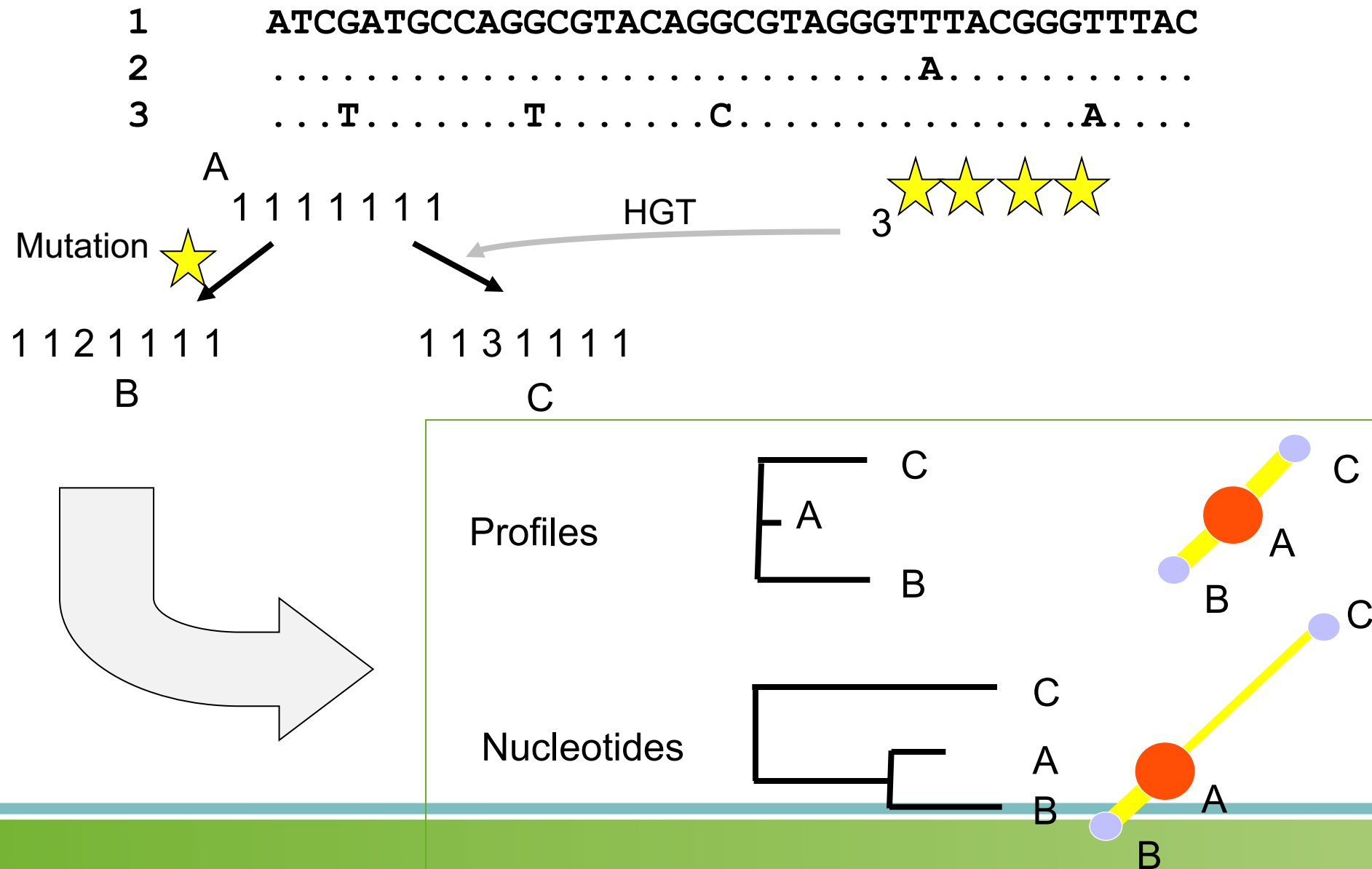
Phylogenetic analyses
(evolution)

Figures



MLST: Multilocus Sequence Typing

Allelic profiles can reflect more reliably the genetic relationships



Figures

Diancourt *et al.* 2005 JCM

Figures

➔ Lack of clear-cut groups
due to lack of resolution of
7-gene MLST

Core genome MLST

Figures

Step 1. Define core genome

Step 2. Define variation at
core genes

Step 3. Define allelic profiles
of genomes



1000-2000 gene loci

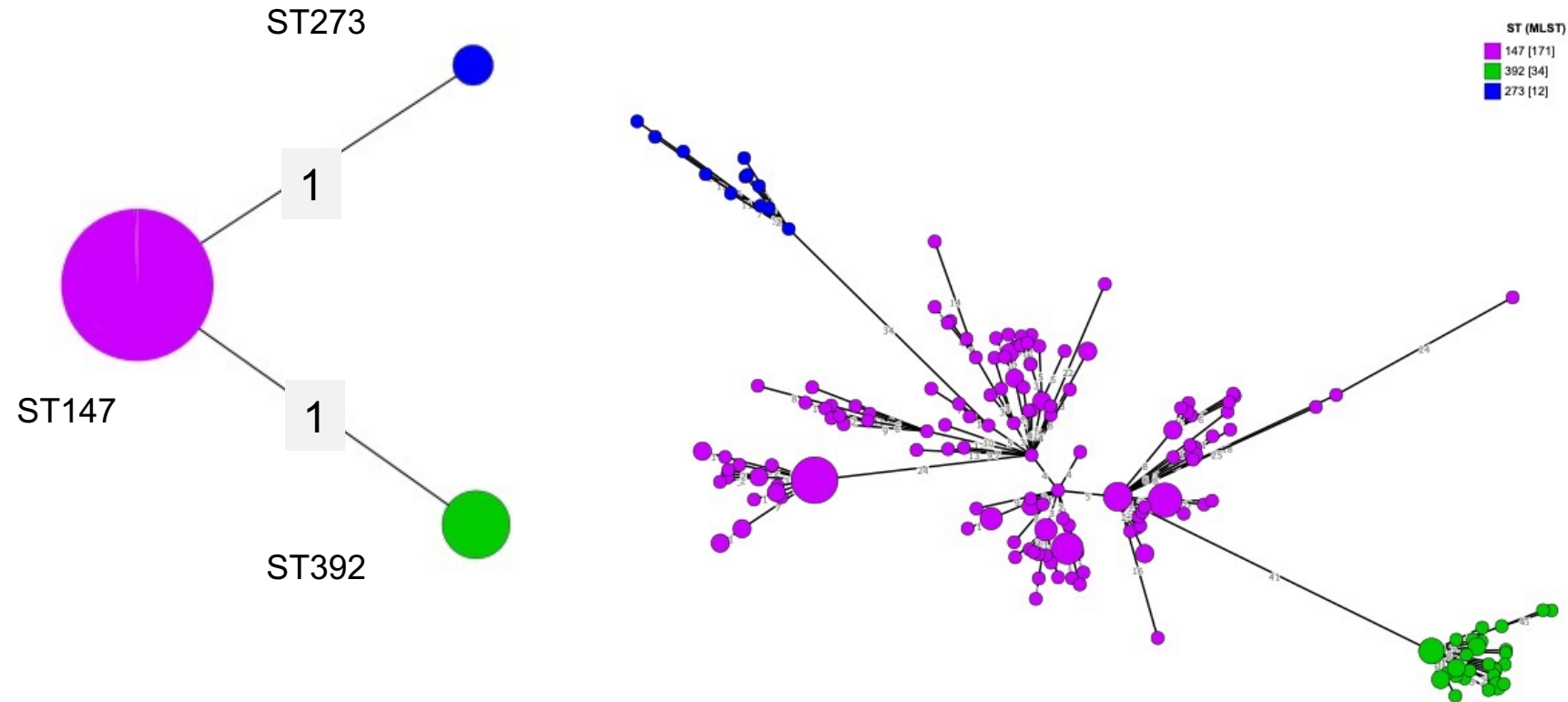
Discrimination is much improved compared to classical (7-gene) MLST

Klebsiella pneumoniae clonal group 147

MStree based on MLST (7 genes)

MStree based on cgMLST (629 genes)

(Hennart *et al.* bioRxiv, 2021)



MStree: minimum spanning tree

The discrimination scale of typing methods

Diagnostic PCR,
MALDI-TOF MLST
RAPD rMLST
ERIC-PCR PFGE cgMLST Whole genome SNPs

Figures

→ All SNPs seen

One in 2

...

...

One in 1000



Low
discrimination

Image: Medini et al. 2008

High
discrimination

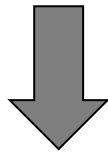
Bacterial species are heterogeneous, in two ways

Bacterial species
threshold ~5%

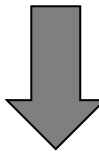
Large
pan-genomes

Figures

Jaureguy *et al.*, 2008



Epidemiology,
classification



Phenotypic variation

Clinical
phenotypes



Ecological
distribution



The inventory of sequence entities in a group of organisms

Figures

- Most generally, the pangenome is based on protein-coding genes (**Prokka/Bakta, BLAST**)

But it is also possible to define a pangenome based on:

- Intergenic regions (**Piggy**)
- All sequences (k-mers: **Mash, Panseq**) (mostly for eukaryotes)

Genome content variation inside bacterial species

Genomic island
Phages
Transposons, IS
ICE
Plasmids
...

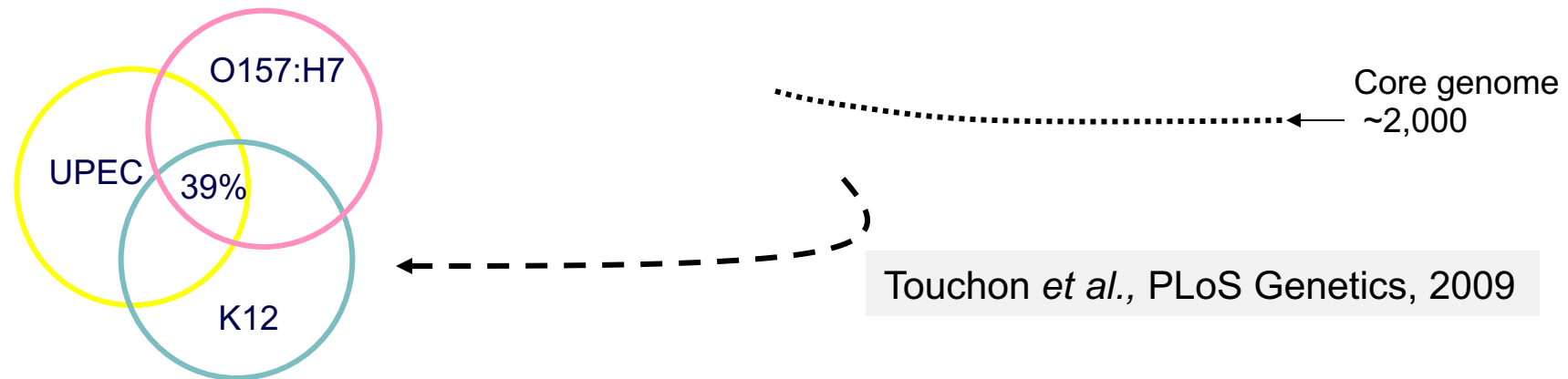
Escherichia coli

All genes
(pangenome)
~20,000

families
~11,500

IS excluded
~10,000

Figures



➔ Pangenome >> strain genome > core genome

The whole picture: core versus accessory gene variation

- Accessory genome evolves fast
- Its evolution is not predictable
- Some consistency with short-scale phylogeny

Figures

The duality of bacterial genomes

Core genome

Essential genes

Neutral evolution

Stability

Vertical inheritance

Sequence variation

MLST, SNPs



Phylogeny, classification
Definition of species, genotypes
Transmission, epidemiology

Accessory genome

Dispendable genes

Adaptation, virulence

Rapid changes

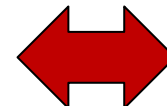
Horizontal transfer

Gene presence/absence

Annotation



Biology, phenotype
Virulence, resistance
Pathogenesis



Bacterial evolution: drivers

Mutation

Recombination

Gene transfer

Figures



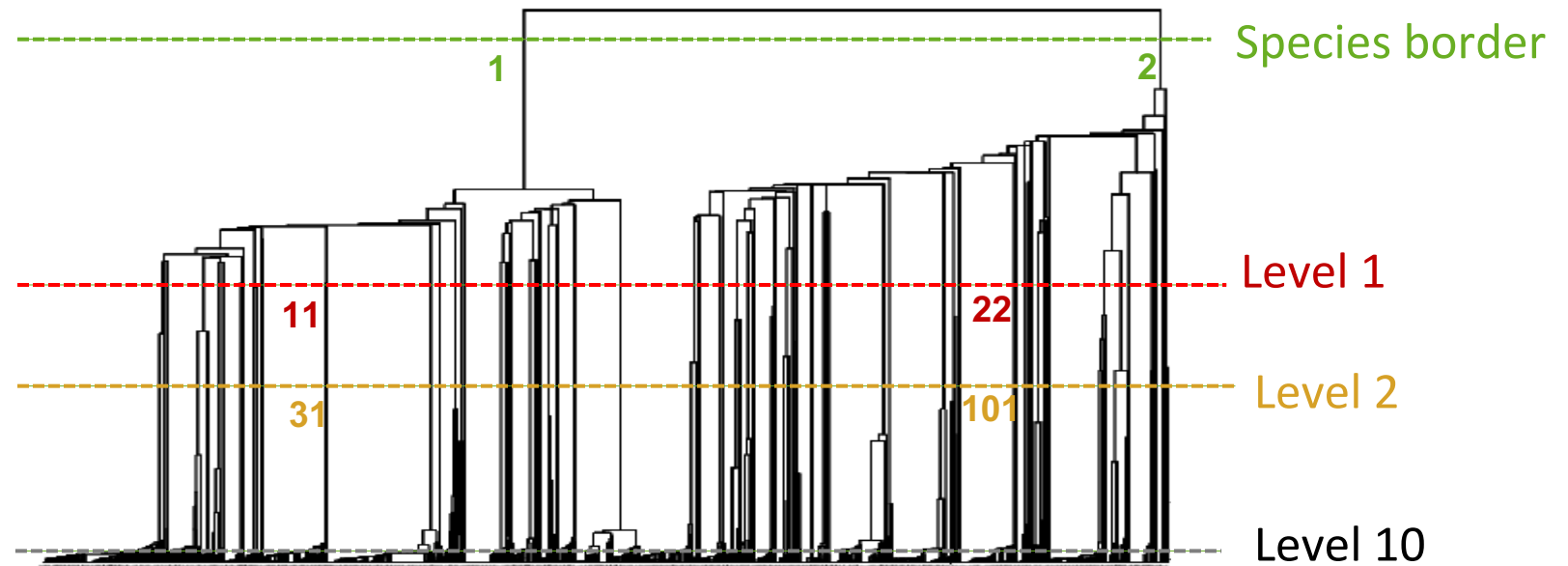
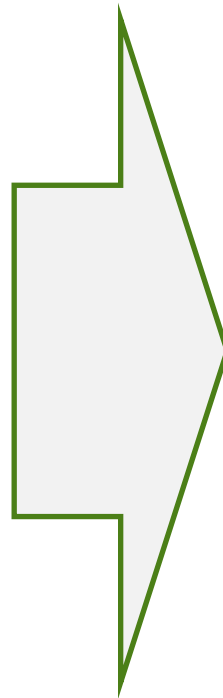
Is it the same strain?

How to define strains?

Escherichia coli

Figures

Jaureguy *et al.*, 2008



How to define strains? Conclusions

- There is no standard definition of a strain, clone, sublineage,...
- 2 broad approaches: gene-by-gene (MLST) or sequence alignment-based
- Best based on phylogenetic approaches
- Need to account for recombination within species
- Multiple levels are needed, depending on the question

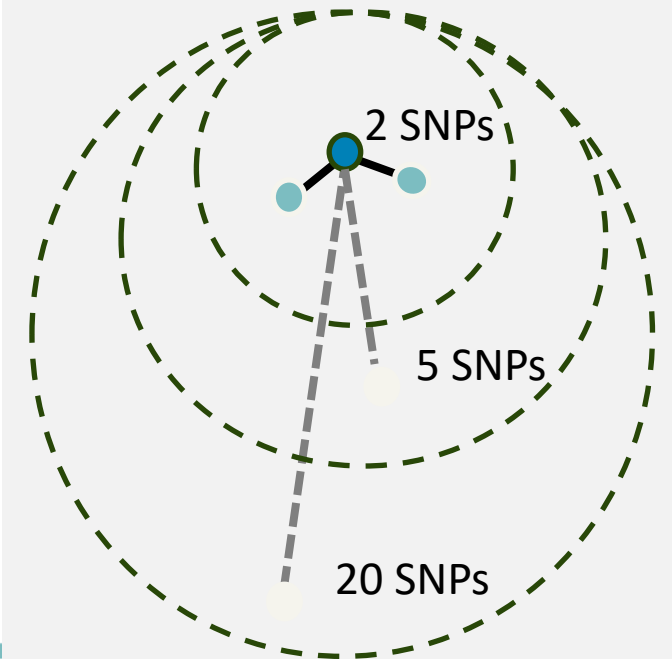
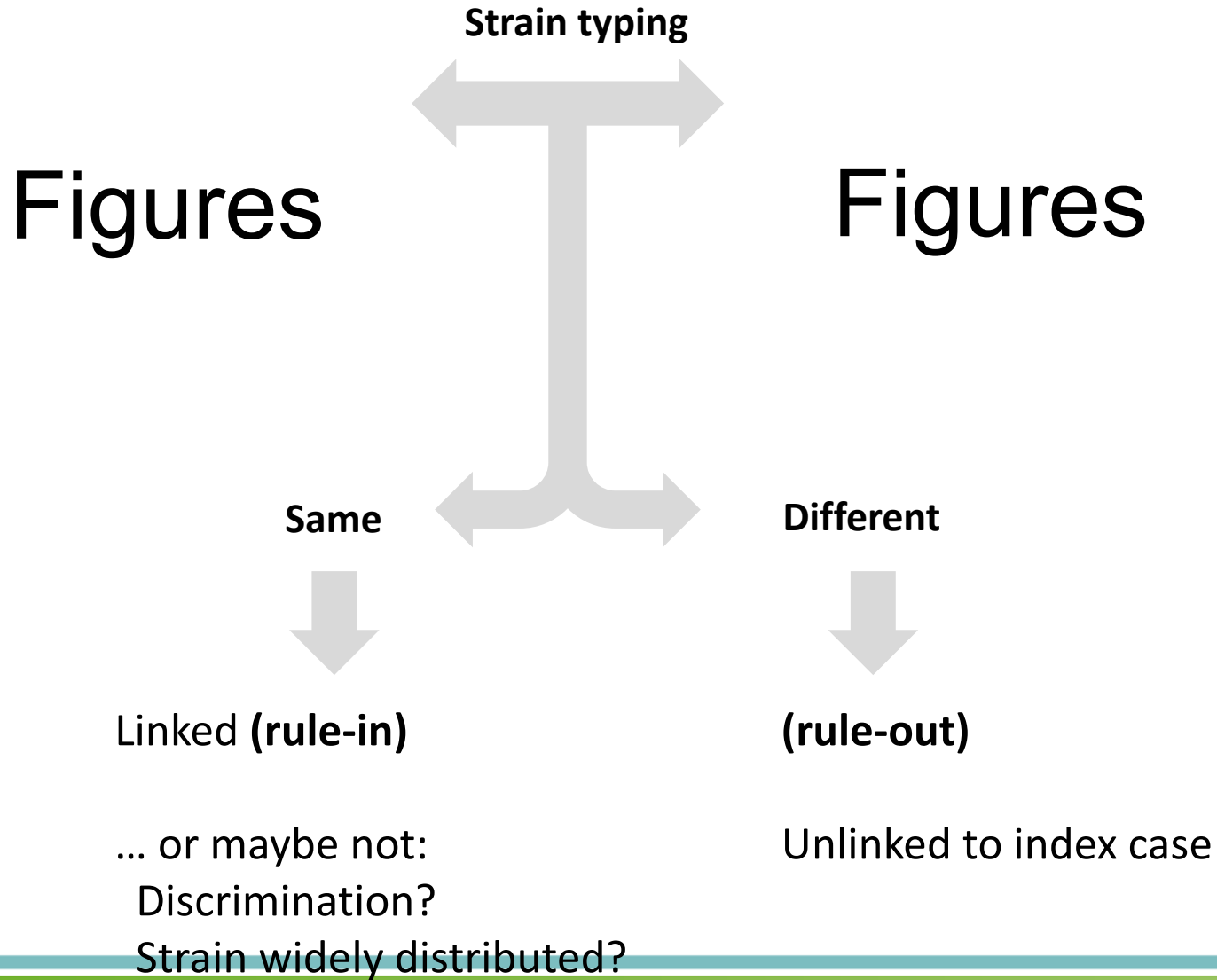
Is. It. The. Same. Strain ?



Index case

Sources?

Question:



'Magic threshold approach': Tenover criteria (PFGE)



Figures

[PDF] Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing

FC Tenover, RD Arbeit, RV Goering... - Journal of clinical ..., 1995 - Am Soc Microbiol

Clinical microbiologists are often asked to determine the relatedness of a group of bacterial isolates, that is, to type them. During the last decade, traditional methods of strain typing, ...

☆ Enregistrer Citer Cité 10549 fois Autres articles Les 14 versions Web of Science: 7052

Widely applied:

- All species
- All cluster types
- Any duration

Strain definition: magic SNP threshold?



Figures

Why general genetic thresholds are not applicable

Local environmental conditions

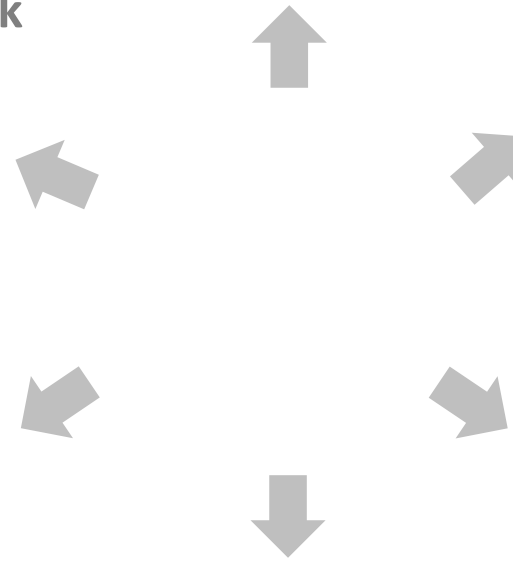
Duration of outbreak

Mutation rates & recombination

Selection on
genetic markers

Sampling dates

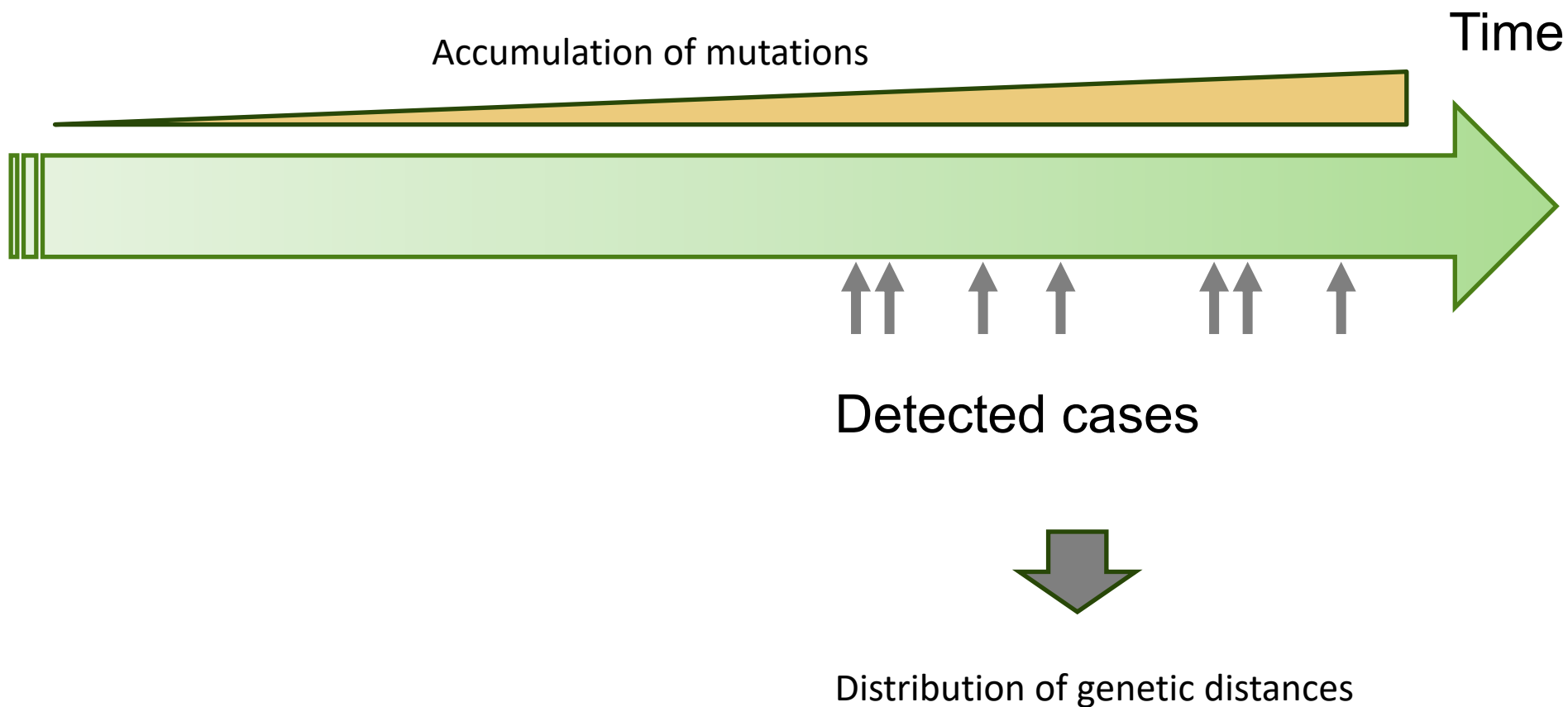
Genetic markers



General SNP thresholds do not exist

Figures

Evolution within the contaminated source



SAMESTRAIN: Outbreak- and risk-specific threshold



- Use of outbreak parameters to run a simulation and get the expected distribution of pairwise distances
- Define the threshold based on exclusion percentile
- Rule-out outbreak isolate using the inferred threshold (with risk taken)

Figures

The SAMESTRAIN framework: An evolutionary approach



- We proposed a hypothesis-driven genetic threshold approach, using a forward Wright-Fisher evolutionary model
- Outbreak-specific features such as *a priori* **pathogen mutation rate** and **duration of source contamination** are considered, and can also be refined through a MCMC-based estimation
- **SNP or cgMLST pairwise distances** and **sample collection dates** of the outbreak of interest are directly used as input parameters

SAMESTRAIN website



Inputs

SimulPopB

HomeSimulationEstimationF.A.QCredits

© - Institut Pasteur -

Getting started

Enter the **Type** of cut-off you want. If you want a genetic cut-off based on SNP data enter the **Genome size** or if you want a genetic cut-off based on cgMLST data enter the **Average size of genes** and the **Number of genes**. Enter the duration of outbreak in **Duration** and the **Number of mutations** per site per year.

You need a csv file with all swab dates to get the cut-off and/or a csv file with the pairwise distance between each isolates to discriminate outbreak to non outbreak cases according to the genetic cut-off found in the **CSV Files** box.

Epidemiological value

Type

☒ SNP
☐ cgMLST

Genome size

4857450

Time step

☒ day
☐ month

Duration (days)

120

Number of mutations

0,0000012

Genome length

Duration

Mutation rate

CSV Files

Choose CSV Date File

Browse...

octavia_dates.csv

Upload complete

Choose CSV SNP File

Browse...

octavia_snp.csv

Upload complete

SNPs

Sampling dates

Run Simulation

Find Cut-off

Reset Data

Download

Manual modification of the threshold:

1

8

50

Inputs:

Length: 4857450 bp

Duration: 120 days

Number of mutation: 1.2e-06 per site per years

Time step: day

Type: SNP

Threshold: 8

Isolates linked by less than 8 SNPs are considered epidemiologically related (see graph)

Distribution

Proposed threshold

Cluster

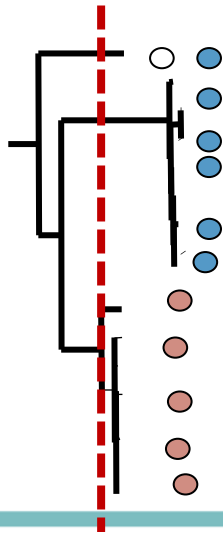
Cluster & outliers

Outputs

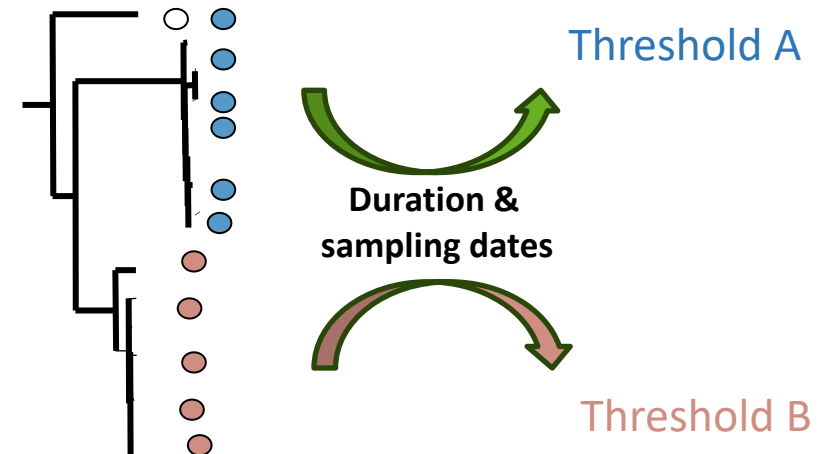
SAMESTRAIN framework: moving away from magic thresholds



Species-specific thresholds



Outbreak-specific thresholds



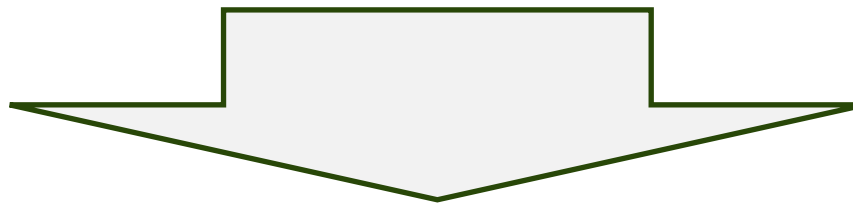
Bacterial evolution: drivers

Mutation

Recombination

Horizontal gene transfer (HGT)

Figures



Is it the same strain?

- Mutations can be modelled (mutation rate)
- Need to purge recombination (Gubbins, ClonalFrameML)
- Gene content is not a reliable strain comparison given HGT

Contributors

Audrey Duval (1,2,3), Lulla Opatowski (1,2), Sylvain Brisse (3)

Institut Pasteur, Paris

1 Epidemiology and modelling of bacterial escape to antimicrobials, Institut Pasteur, Paris, France

2 Anti-infective Evasion and Pharmacoepidemiology Team, CESP, Université Paris-Saclay, UVSQ, INSERM U1018, Montigny-le-Bretonneux, France

3 Institut Pasteur, Université Paris Cité, **Biodiversity and Epidemiology of Bacterial Pathogens, Paris, France**



**Lulla
Opatowski**

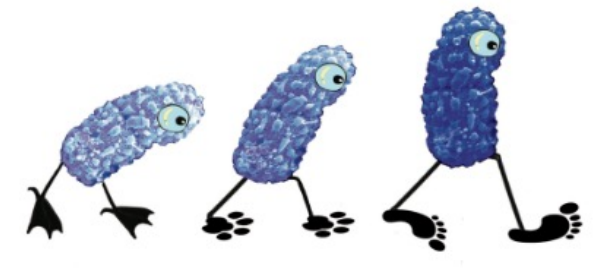


**Audrey
Duval**

THE LANCET
Microbe



Funding



MedVetKlebs project

Contact: sylvain.brisse@pasteur.fr



@sylvainbrisse



Epidemiological surveillance
Clinical data, Diagnostics

Global diversity
Population biology

Acknowledgements

The creation of this training material was commissioned by ECDC to Institut Pasteur with the direct involvement of Pr. Sylvain Brisse.