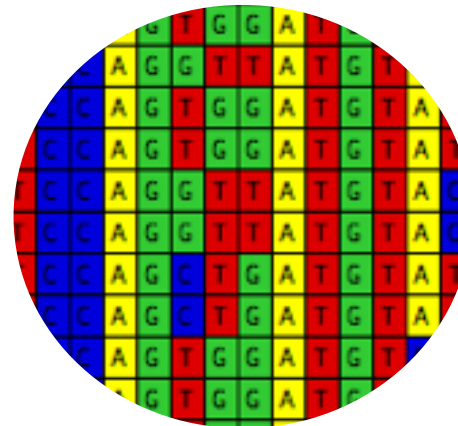
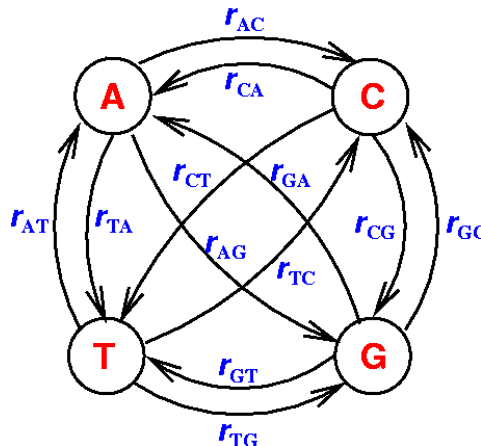


PMB2025

PATHOGEN MULTIOMICS AND BIOINFORMATICS

Recife PE 2025

Module 4: Introduction to Phylogenetics and Public Health



João Perdigão

*Phylogenetics pertains to the **study of the evolutionary relationships***

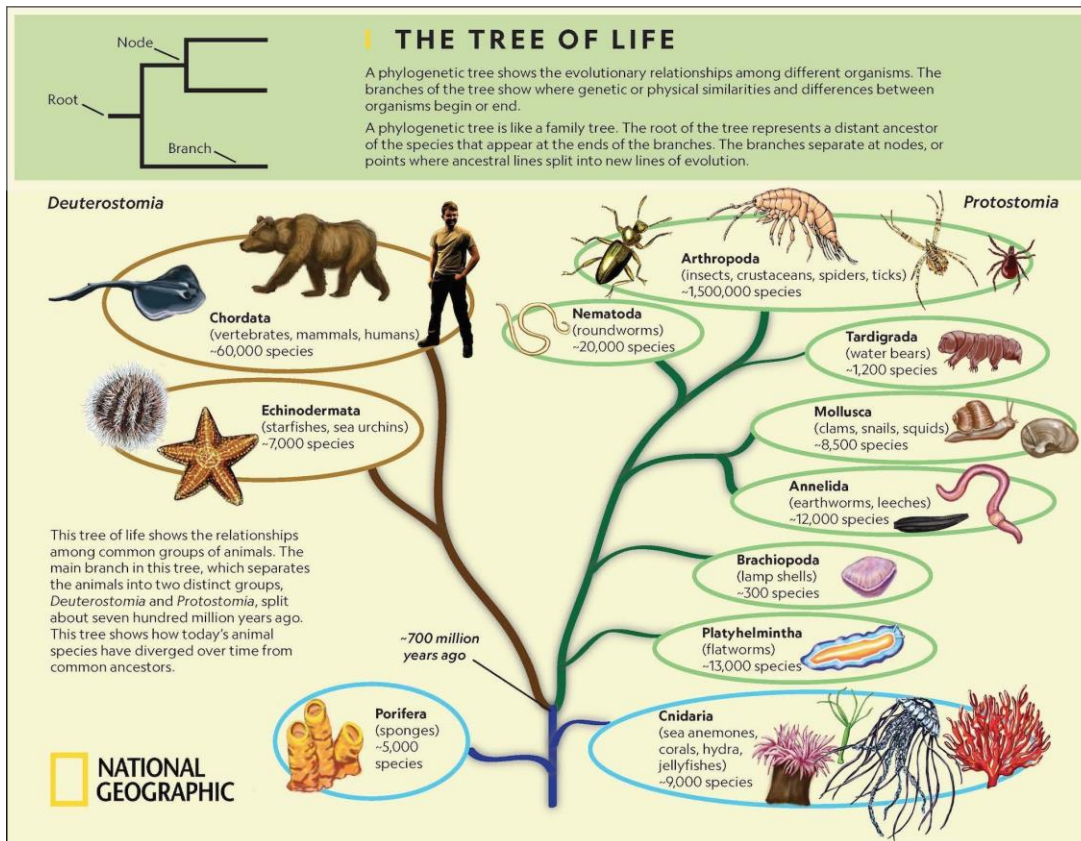
Between what?

- *organisms, e.g., species or strains*
- *genes*
- *genomes*
- *Etc.*

Phylogenetics should refer to how closely the taxa are

and...

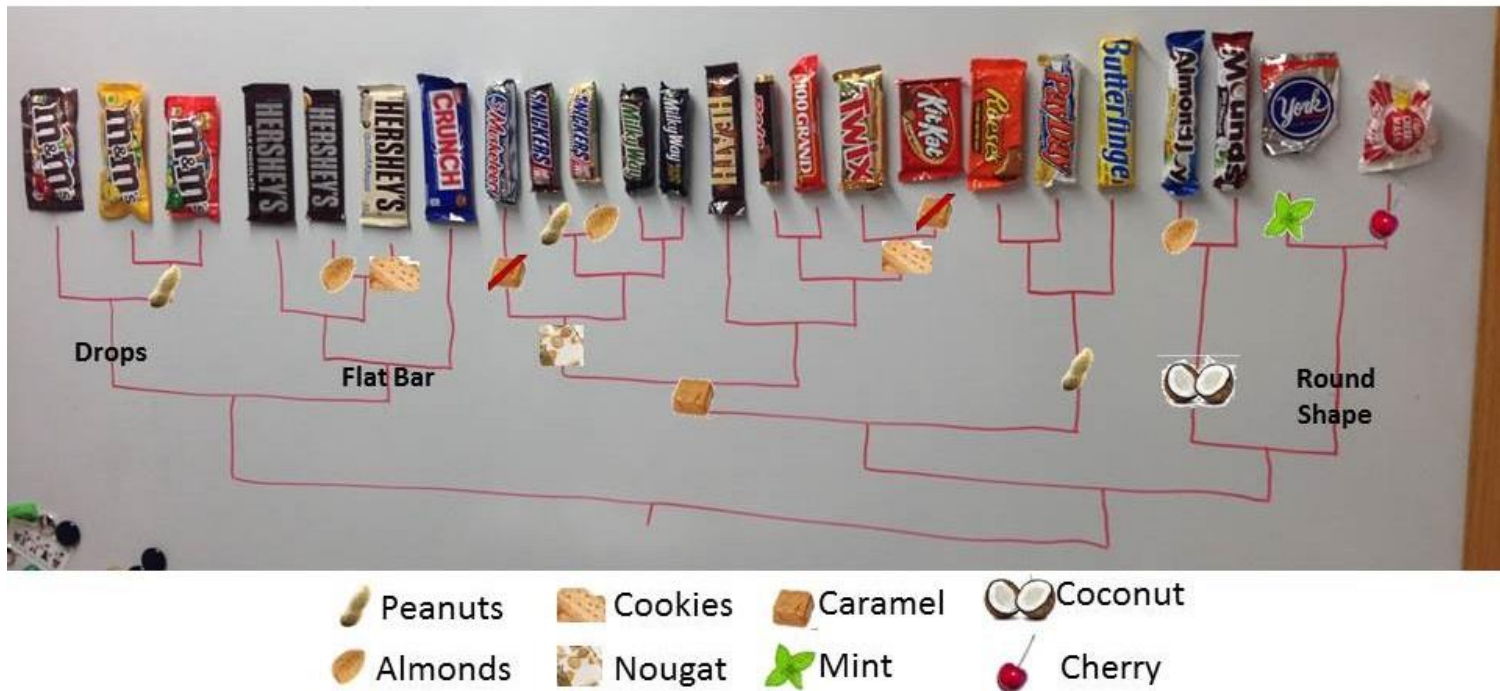
its evolutionary history



Phylogenetics

What characters to use to construct a tree?

Early characters for phylogenetic reconstruction relied upon morphological and physiological characteristics



Example: ability to grow at different temperatures, drug resistance, sugar fermentation

Problems?

Molecular Phylogenetics

Morphological/Physiological characters have two main problems associated:

Proneness to convergent evolution

Limited number of characters – poorly informative

However ... molecular data (DNA or Protein sequence),

are less prone to convergent evolution

can provide an increasing number of characters

```
      A T G C T T T G C
A T G T T T T G C
      A G G C T T T G C
A G G C T T T A C
```

But... which characters are these?

```
A T G C T T T G C
A T G T T T T G C
A G G C T T T G C
A G G C T T T A C
```

Each homologous position between sequences comprise a character?

An alignment provides a way to:

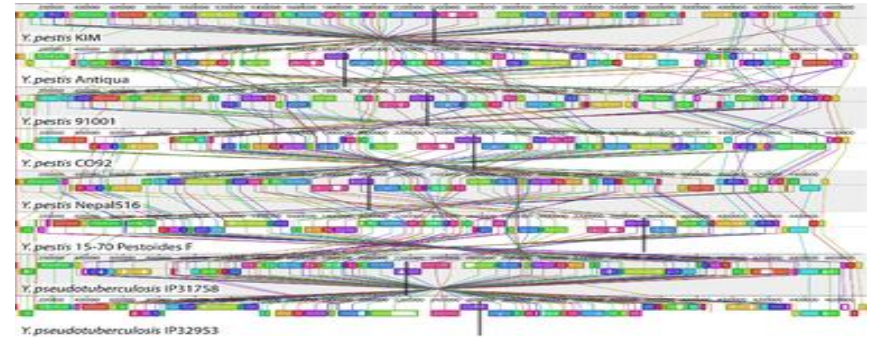
- Identify homology – regions of common ancestry
- Contrast regions

In summary, molecular phylogenetics require a Multiple Sequence Alignment

High amount of information – each column is an evolutionary marker

Possibility of going genome-wide – compare entire genomes

Problem: alignment of entire genomes consumes large amounts of memory, even more upon tree bulding



Research Questions & Trade-offs:

- Sequence length (full length vs gene fragments)
- Genetic variation (conserved vs variable regions)

Molecular Phylogenetics

A possible approach: construct a DNA pseudo-molecule based on SNPs

In theory (perfect world)

```

A T G C T T T G C
A T G T T T T G C
A G G C T T T G C
A G G C T T T A C
  
```



```

T C G
T T G
G C G
G C A
  
```

Real-life:

```

A T G C T T T G C
A T G T T T T G -
A G G - - T T G C
A G G C T T T A C
  
```



```

T C G
T T G
G - G
G C A
  
```

*wgSNP alignment
gapped alignment*

```

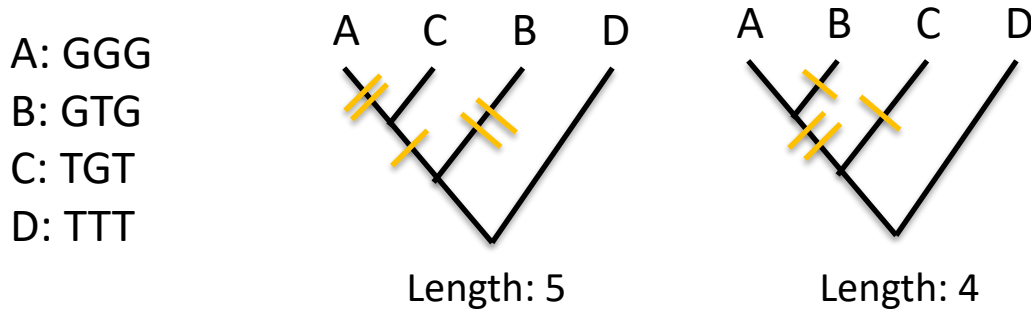
T G
T G
G G
G A
  
```

coreSNP alignment

Methods for Phylogenetic Reconstruction

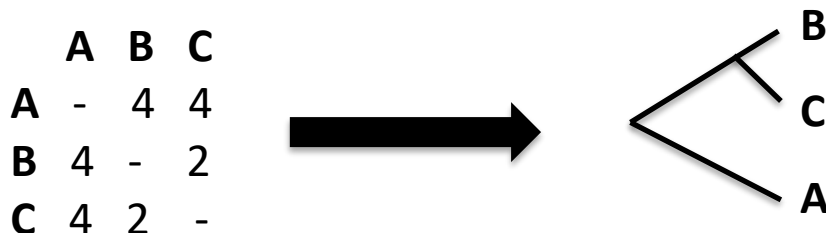
Maximum Parsimony

- *Simplest possible evolution scenario – the best tree is the shortest tree*
- *The rationale is to have the tree with the least homoplasy*



Distance Matrix Methods

- *Start by calculating pairwise distances from sequence data*
- *Tree is constructed from pairwise distances through clustering algorithms (e.g. Neighbour-Joining)*



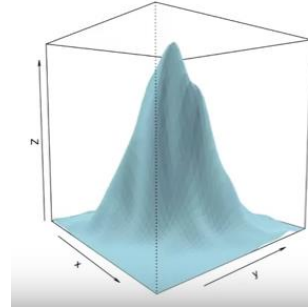
Methods for Phylogenetic Reconstruction

Maximum Likelihood

More robust approach with parameter estimation using a probabilistic model:

- *Tree topology and branch lengths*
- *Nucleotide frequencies*
- *Nucleotide substitution rates*
- *Measure how well the model fits the data*

Calculates the likelihood for every column first and for the entire alignment using different permutations, random starting values.



Likelihood (Model) = Probability (Data | Model)

Maximum Likelihood – Best set of parameter values yielding the highest possible likelihood

Software: PhyML [MEGA,Seaview, etc]

Bayesian Inference

- Based on the calculation of posterior probabilities given a set of prior parameter values
- Requires starting prior value
- Involves millions of iterations and measures the convergence to parameter values over the iterations

Software: MrBayes, BEAST

Nucleotide Substitution Models

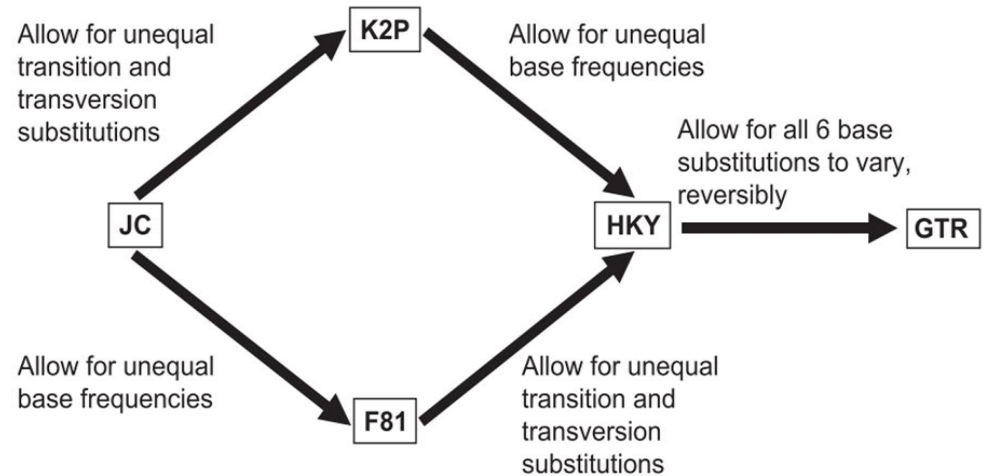
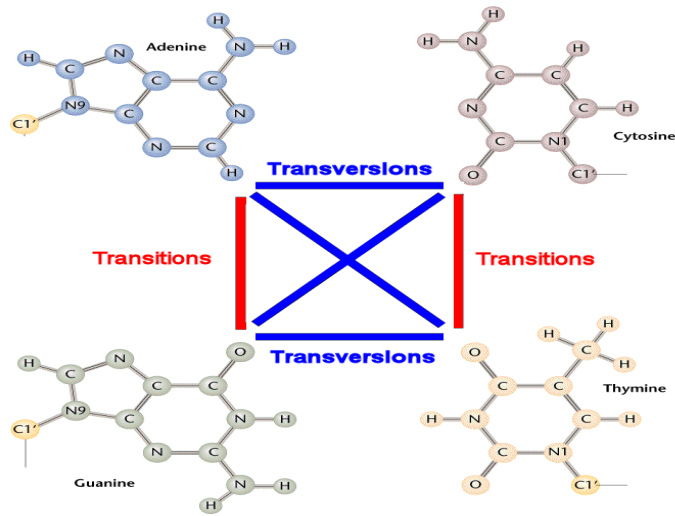
Models of Evolution

Aim to infer the number of real evolutionary events based on observed events!

Need to know how sequences are evolving – requires a model!!

Possibility of the existence of different substitution rates across site (gamma distribution).

The six possible substitution patterns for nucleotide data

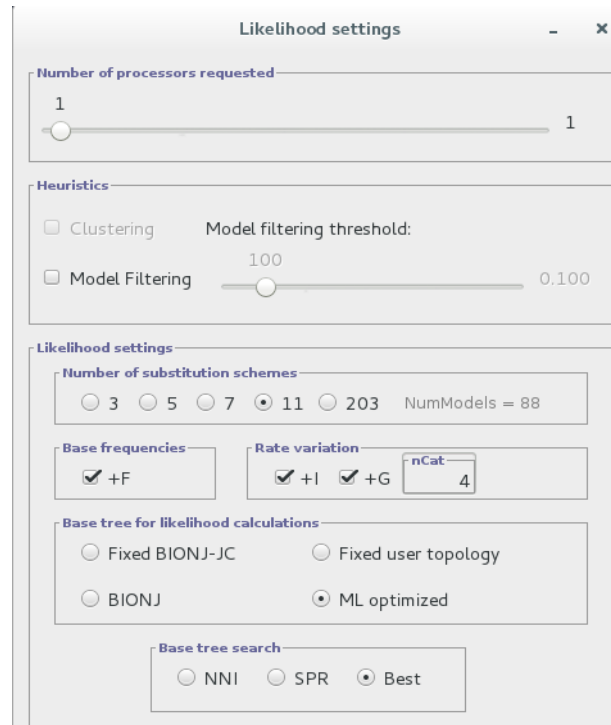
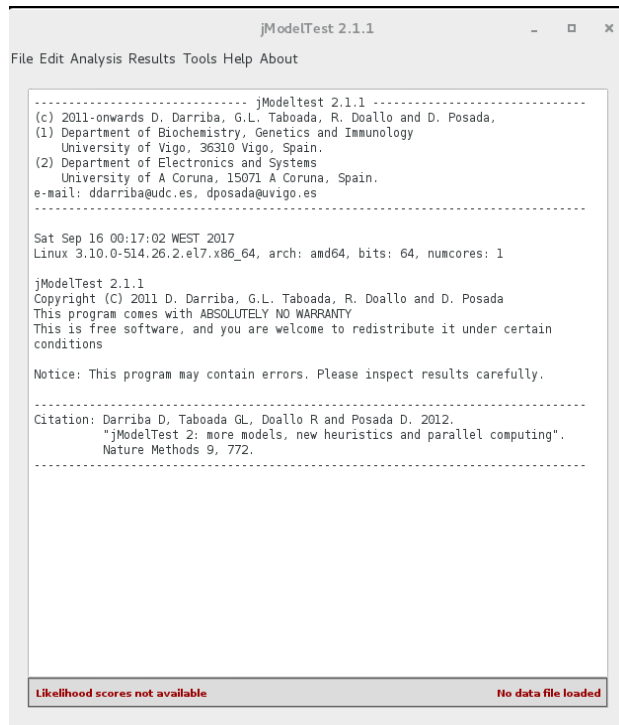


Transitions occur at higher frequency than Transversions!

How to choose the right parameters?

Answer: test for a set of parameter permutations and choose on the combination that yields the combination that best fits your data

jModelTest



Also possible: R, phangorn package (model.test function); ModelTest online server
 Increasingly integrated in popular Tree building programs: RAxML, IQ-Tree

Tree Searching

Tree space – the number of all possible trees for a given dataset

**How many branches are present in a tree with 3 tips? But, how many possible trees:
And with 4 tips?**

$$\prod_{i=2}^{n-1} (2i - 3)$$

Number of branches in a tree with x tips = $2x-3$

3 tips ->	3 branches
4 tips ->	5 branches
5 tips ->	7 branches
10 tips ->	17 branches
20 tips ->	37 branches
40 tips ->	77 branches
100 tips ->	197 branches

*For every tree with x tips it is possible to
construct $2x-3$ derive trees by adding na extra tip*

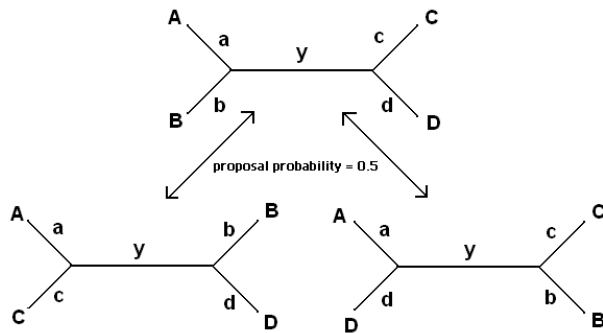
3 tips ->	1 tree
4 tips ->	3 trees
5 tips ->	15 trees
6 tips ->	105 trees
7 tips ->	945 trees
8 tips ->	10395 trees
9 tips ->	135135 trees
10 tips ->	2027025 trees
100 tips ->	1.7×10^{182} trees

It is not possible to exhaustively screen and search all possible trees in present day datasets.

Answer: start with random tree(s) and progress by making alterations and removing less likely pathways

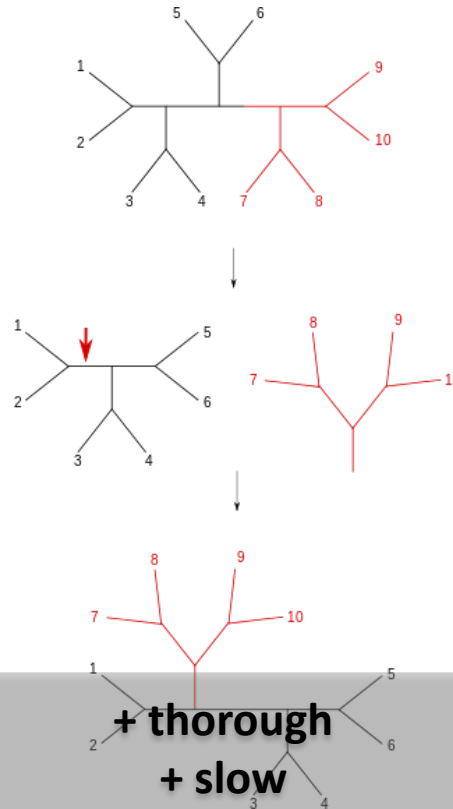
Tree Searching

Nearest Neighbour Interchange (NNI)



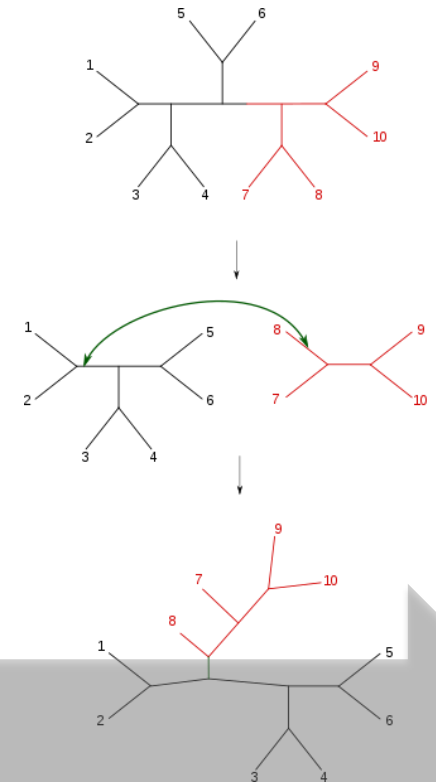
Swap neighbors at every internal branch

Subtree Pruning and Regrafting (SPR)



Cut a subtree at every possible point and regrafts at multiple points

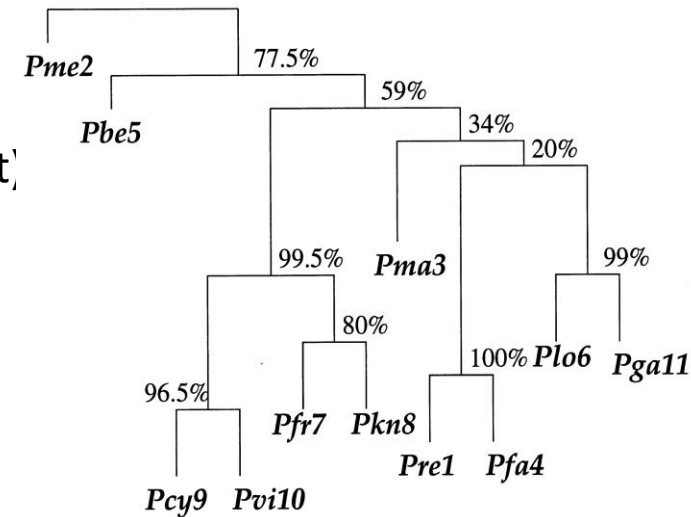
Tree Bisection and Reconnection (TBR)



Divides tree in two parts, regrafts using every possible branch of the detached tree at all possible branches of the other tree

Bootstrap

- Widely used;
- Random sampling from the alignment (with replacement) until achieving the original length;
- Tree reconstruction n times – for each new alignment;
- Calculation for each branch the occurrence of that same clade in each tree;
- Expressed as a percentage/fraction (0-100%).
- Can be extremely time consuming



Alternatives:

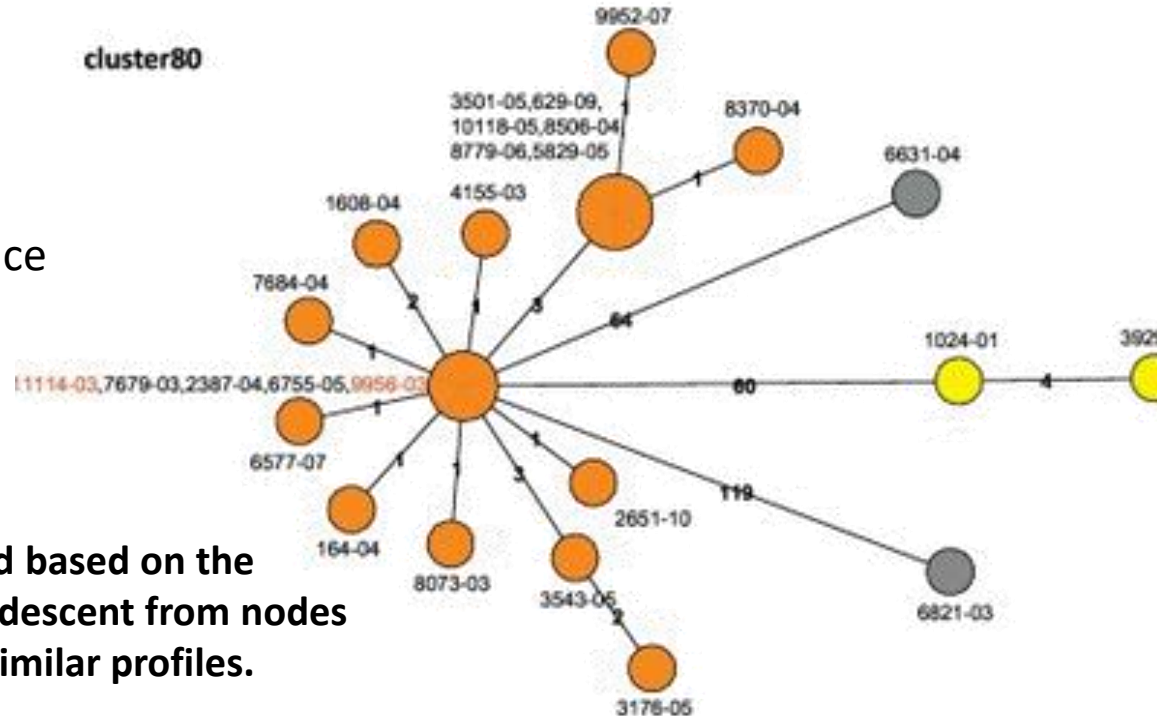
- Jackknife – removes a position from the alignment (leave one out)
- aLRT – approximate Likelihood Ratio Test – provides a p -value – likelihood gains between having a branch and not having a branch

Alternative Approaches

Distance based methods using the eBURST/goeBURST

- core genome Multi Locus Sequence Typing – cgMLST
- multiple sequence alignment

Trees/relationships can be inferred based on the goeBURST/eBURST most parsimonious descent from nodes aggregating isolates/strains with similar profiles.



“eBURST approach subdivides large MLST data sets into nonoverlapping groups of related STs or clonal complexes and then discerns the most parsimonious patterns of descent of isolates within each clonal complex from the predicted founder”

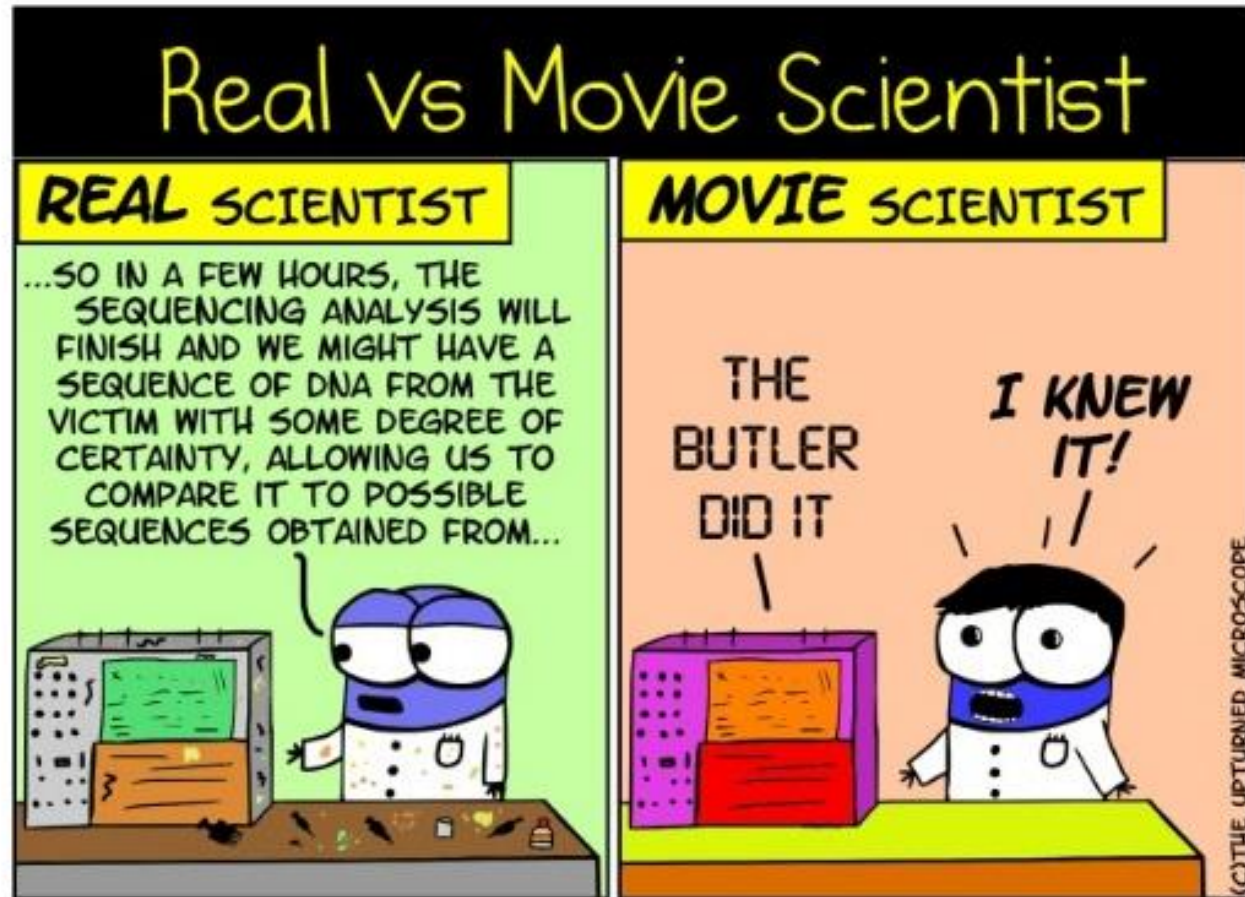
Feil et al 2004

“goeBURST is a globally optimized implementation of the eBURST algorithm, that identifies alternative patterns of descent for several bacterial species. Furthermore, the algorithm can be applied to any multilocus typing data based on the number of differences between numeric profiles.”

Francisco et al 2009

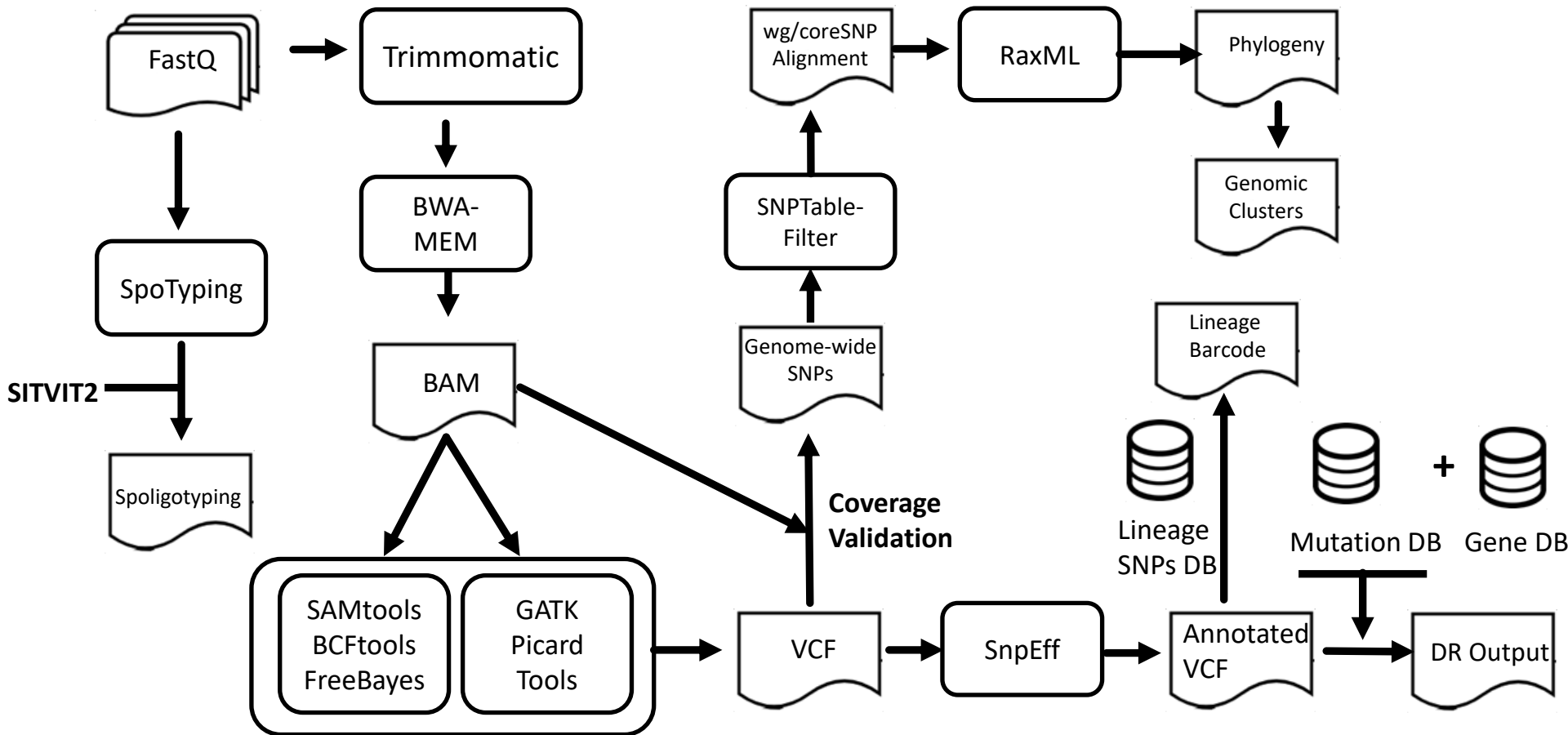
DNA sequence analysis

- |
- |
- |
- |
- |

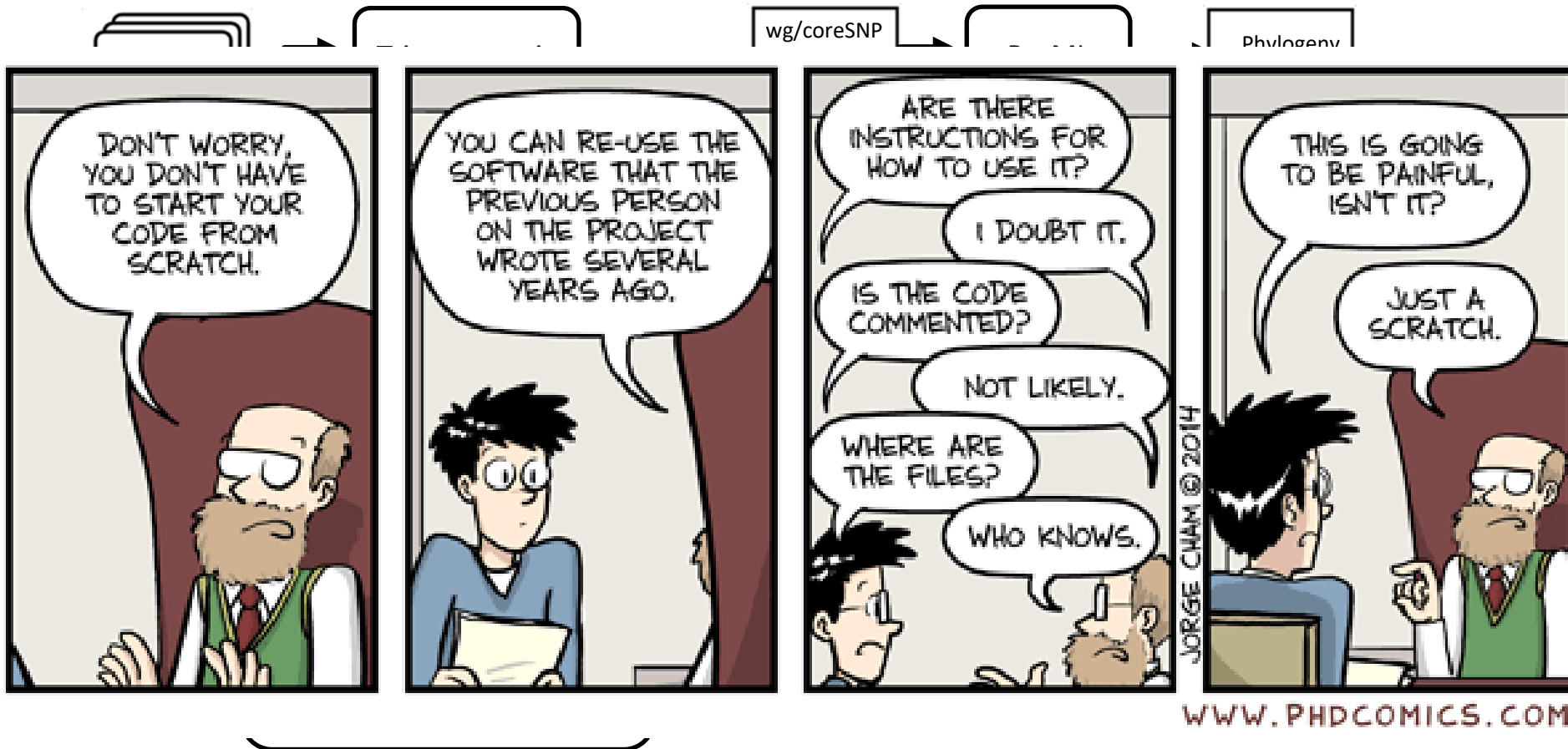


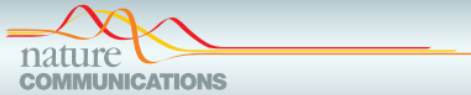
es;

WGS analytic workflow: an overview



WGS analytic workflow: an overview





ARTICLE

Received 11 Apr 2014 | Accepted 25 Jul 2014 | Published 1 Sep 2014

DOI: 10.1038/ncomms5812

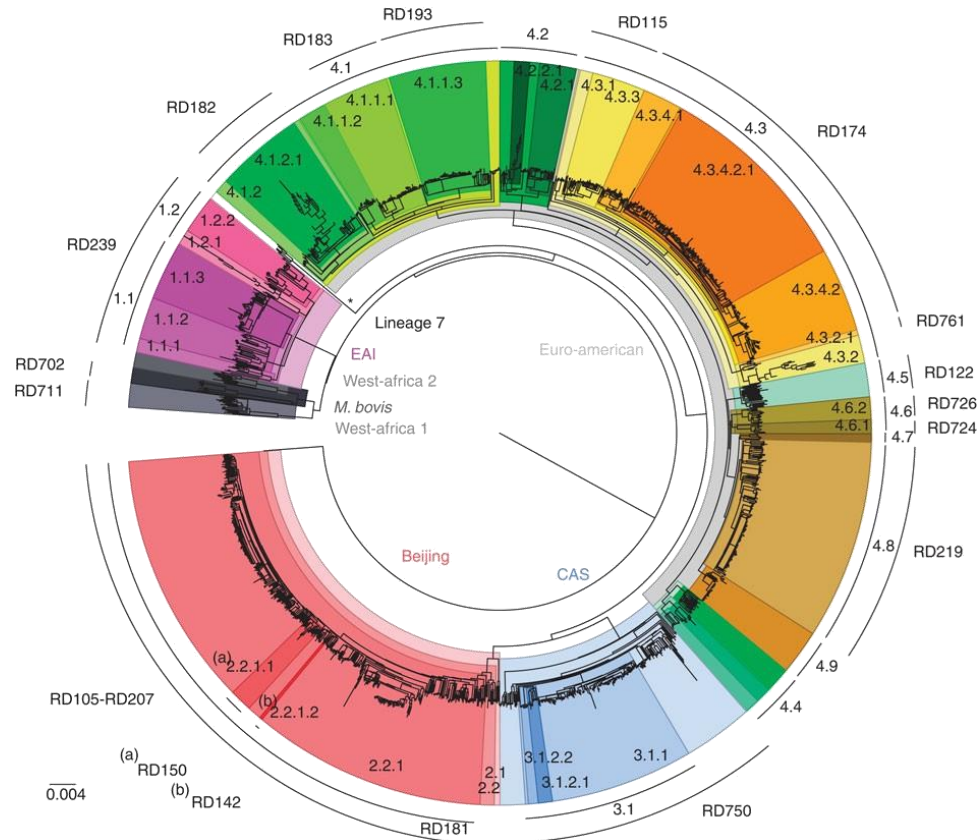
OPEN

A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains

Francesc Coll¹, Ruth McNerney¹, José Afonso Guerra-Assunção², Judith R. Glynn², João Perdigão³, Miguel Viveiros⁴, Isabel Portugal³, Arnab Pain⁵, Nigel Martin⁶ & Taane G. Clark^{1,2}

:: 1601 *M. tuberculosis* genomes

:: 62 SNP markers for strain barcoding



Coll et al, 2014

A Framework for Genome-Wide Association Studies (GWAS):



Genome-wide analysis of multi- and extensively drug-resistant *Mycobacterium tuberculosis*

Francesc Coll¹, Jody Phelan¹, Grant A. Hill-Cawthorne^{2,3}, Mridul B. Nair², Kim Mallard¹, Shahjahan Ali², Abdallah M. Abdallah², Saad Alghamdi⁴, Mona Alsomali², Abdallah O. Ahmed⁵, Stephanie Portelli^{1,6}, Yaa Oppong¹, Adriana Alves⁷, Theolis Barbosa Bessa⁸, Susana Campino⁹, Maxine Caws¹⁰, Anirvan Chatterjee¹¹, Amelia C. Crampin^{12,13}, Keertan Dheda¹⁴, Nicholas Furnham¹⁵, Judith R. Glynn^{12,13}, Louis Grandjean¹⁶, Dang Minh Ha¹⁰, Rumina Hasan¹⁶, Zahra Hasan¹⁶, Martin L. Hibberd¹⁷, Moses Joboba¹⁸, Edward C. Jones-López¹⁹, Tomoshige Matsumoto²⁰, Anabela Miranda²¹, David J. Moore²², Nora Mocillo²³, Stefan Panaiotov²⁴, Julian Parkhill²⁵, Carlos Penha²⁶, João Perdigão²⁶, Isabel Portugal²⁶, Zineb Rchiad²⁷, Jaime Robledo²⁸, Patricia Sheen²⁹, Nashwa Talaat Shesha³⁰, Frik A. Sirgel³¹, Christophe Solà³², Erivelton Oliveira Sousa³³, Elizabeth M. Streicher³⁴, Paul Van Helden³⁵, Miguel Viveiros³⁶, Robert M. Warren³⁷, Ruth McNerney³⁸, Arnab Pain^{39,40} and Taane G. Clark⁴¹

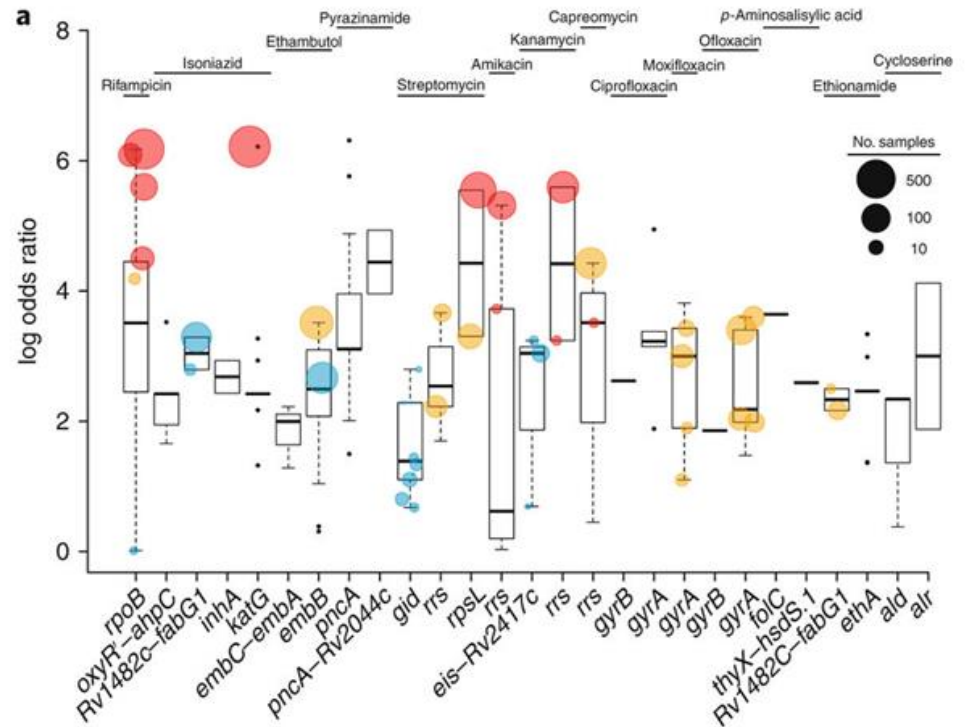
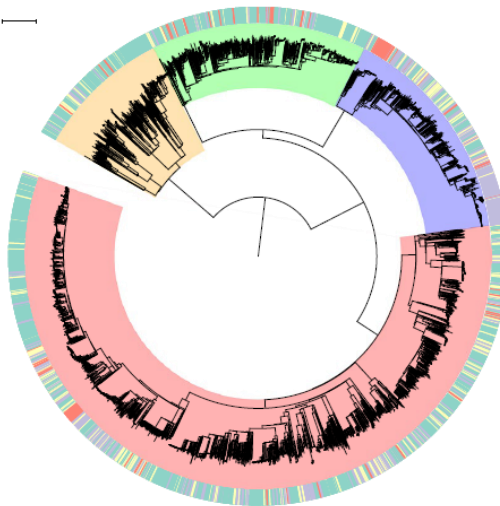
Tree scale: 0.001

Lineages

- Lineage 1
- Lineage 3
- Lineage 2
- Lineage 4

Phenotype

- Susceptible
- Drug resistant
- MDR-TB
- XDR-TB



:: 6 465 clinical isolates

:: > 35 countries

:: 102 106 SNPs

:: 11 122 Indels

:: 284 large deletions

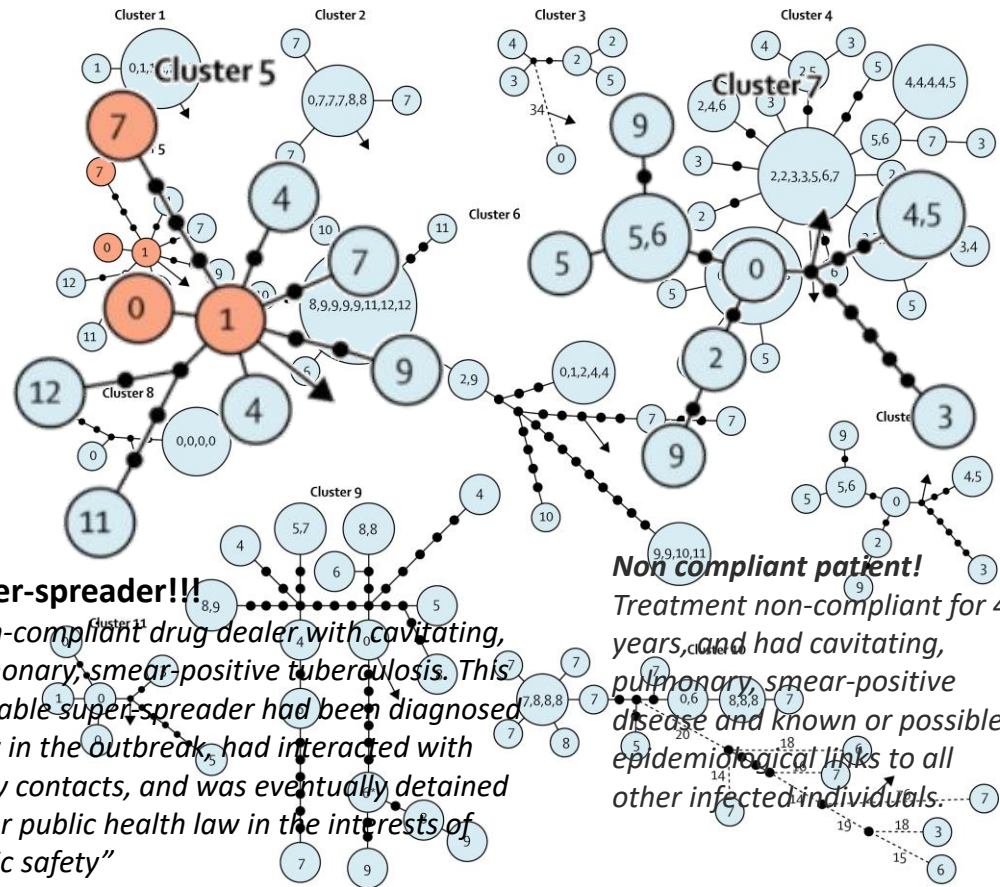
WGS as a Public Health Tool

Whole-genome sequencing can delineate outbreaks of tuberculosis and allows inference about direction of transmission between cases.

≤ 5 SNPs – strong likelihood for epidemiological link

5-12 SNPs – uncertain!

≥ 12 SNPs - Epidemiological link unlikely



Walker et al 2013

Investigate cross-contamination, exogenous reinfection, relapse and outbreaks.

- Negative acid-fast smears;
- Only one specimen culture positive;
- Clinical findings not supportive of TB;

Genotype concurrently handled
strains in the laboratory

Contamination?

If yes,

Discontinuation of therapy!

- Same patient;
- Different patterns of drug susceptibility (after or during treatment);
- Resistance development or reinfection or cross-contamination

- Second TB episode;
- Relapse or reinfection?

Resistance Development

Reinfection

**Highly important
in drug trials!**

Relapse

Non-adherence?
Low drug concentration?
(malabsorption, drug interactions)

Identify Source!!

Evaluate treatment failure
Susceptibility of original
isolate, etc.

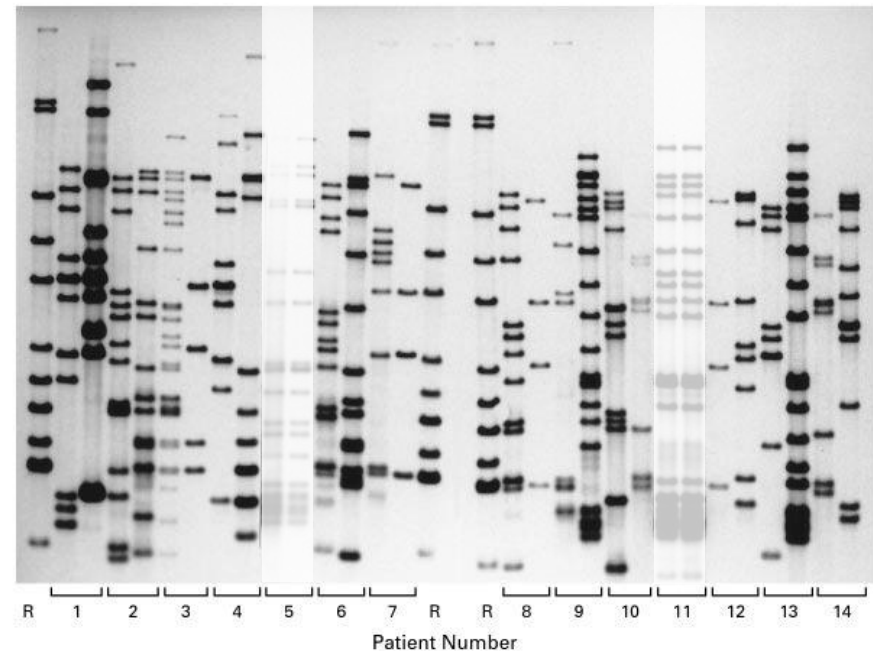
ORIGINAL ARTICLE

Exogenous Reinfection as a Cause of Recurrent Tuberculosis after Curative Treatment

Annelies van Rie, M.D., Robin Warren, Ph.D., Madeleine Richardson, M.Sc., Thomas C. Victor, Ph.D., Robert P. Gie, M.D., Donald A. Enarson, M.D., Nulda Beyers, Ph.D., and Paul D. van Helden, Ph.D.

*We performed DNA fingerprinting with restriction-fragment-length polymorphism analysis on pairs of isolates of *Mycobacterium tuberculosis* from 16 compliant patients who had a relapse of pulmonary tuberculosis after curative treatment of postprimary tuberculosis.*

*For **12 of the 16 patients**, the restriction-fragment-length polymorphism banding patterns for the isolates obtained after the relapse were different from those for the isolates from the initial tuberculous disease. This finding indicates that reinfection was the cause of the recurrence of tuberculosis after curative treatment.*



THE LANCET

ARTICLES | VOLUME 358, ISSUE 9294, P1687-1693, NOVEMBER 17, 2001

HIV-1 and recurrence, relapse, and reinfection of tuberculosis after cure: a cohort study in South African mineworkers

Dr Pamela Sonnenberg, MBBCh   Jill Murray, FFPATH • Judith R Glynn, PhD • Stuart Shearer, FFOM •
Bupe Kambashi, DipMLT • Peter Godfrey-Faussett, FRCP

Published: November 17, 2001 • DOI: [https://doi.org/10.1016/S0140-6736\(01\)06712-5](https://doi.org/10.1016/S0140-6736(01)06712-5)

Patients were examined 3 and 6 months after cure, and then were monitored by the routine tuberculosis surveillance system until December, 1998. IS6110 DNA fingerprints from initial and subsequent episodes of tuberculosis were compared to determine whether recurrence was due to relapse or reinfection

Paired DNA fingerprints were available in 39 of 65 recurrences: 25 pairs were identical (relapse) and 14 were different (reinfection). 93% (13/14) of recurrences within the first 6 months were attributable to relapse compared with 48% (12/25) of later recurrences.

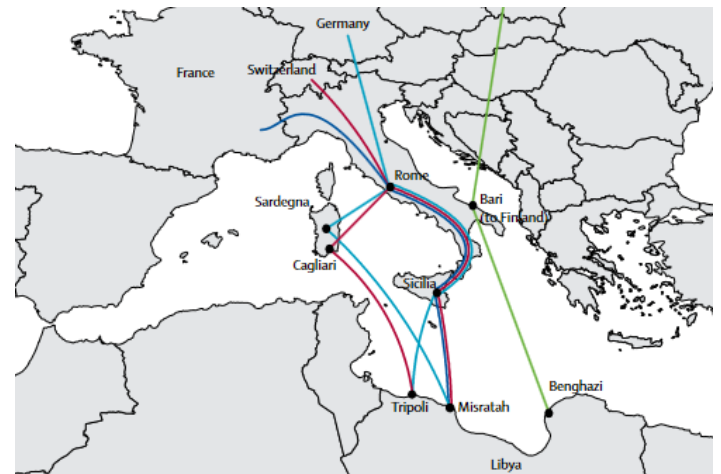
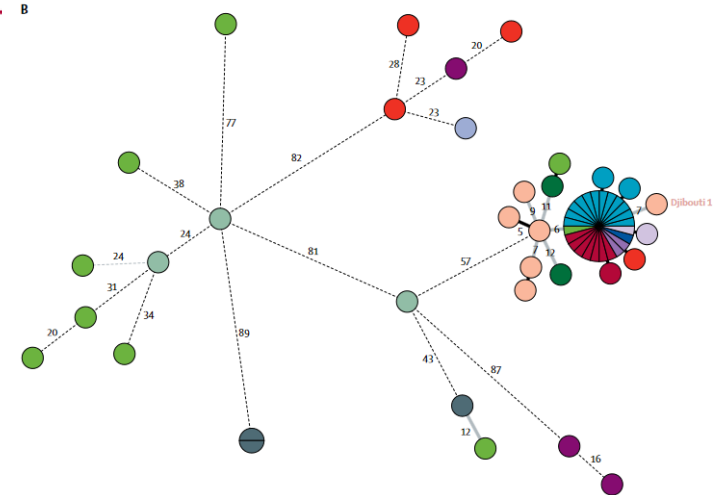
Case Study III: International, Migration-Driven Clone Dissemination

A cluster of multidrug-resistant *Mycobacterium tuberculosis* among patients arriving in Europe from the Horn of Africa: a molecular epidemiological study

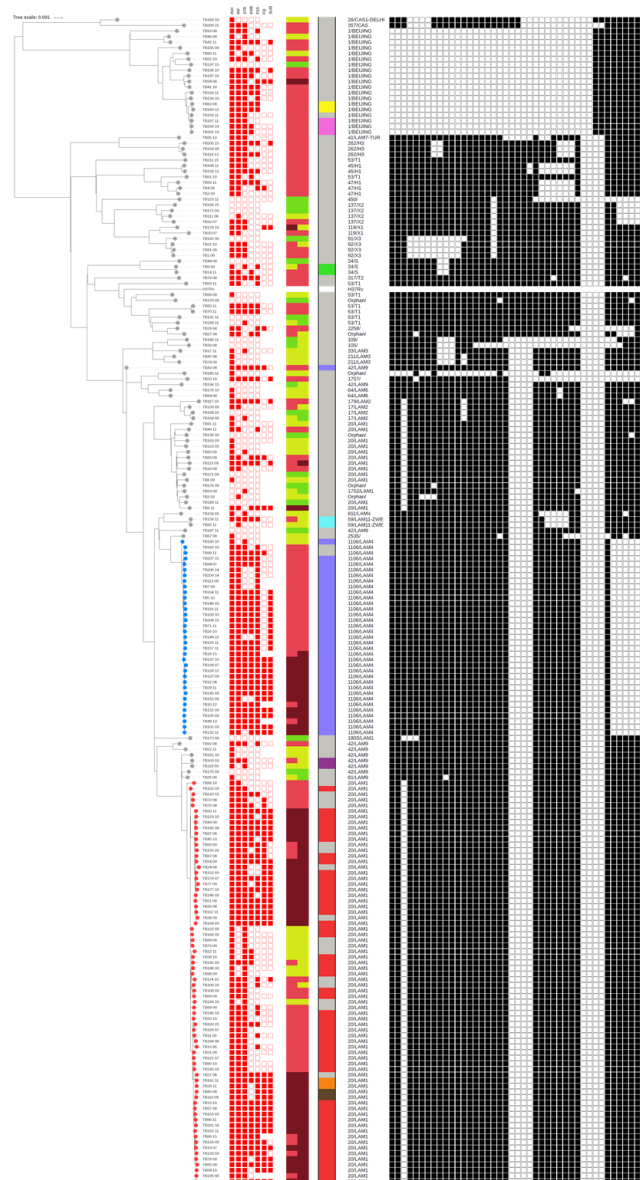
Timothy M Walker*, Matthias Merker*, Astrid M Knoblauch*, Peter Helbling, Otto D Schoch, Marieke J van der Werf, Katharina Kranzer, Lena Fiebig, Stefan Kröger, Walter Haas, Harald Hoffmann, Alexander Indra, Adrian Egli, Daniela M Cirillo, Jérôme Robert, Thomas R Rogers, Ramona Groenheit, Anne T Mengschoel, Vanessa Mathys, Marjo Haanperä, Dick van Soolingen, Stefan Niemann†, Erik C Böttger†, Peter M Keller†, and the MDR-TB Cluster Consortium‡

On April 29 and May 30, 2016, the Swiss and German National Mycobacterial Reference Laboratories independently triggered an outbreak investigation after four patients were diagnosed with multidrug-resistant tuberculosis. In this molecular epidemiological study, we prospectively defined outbreak cases with 24-locus mycobacterial interspersed repetitive unit-variable number tandem repeat (MIRU-VNTR) profiles

Between Feb 12, 2016, and April 19, 2017, 29 patients were diagnosed with multidrug-resistant tuberculosis in seven European countries. All originated from the Horn of Africa or Sudan, with all isolates two SNPs or fewer apart. 22 (76%) patients reported their travel routes, with clear spatiotemporal overlap between routes. We identified a further 29 MIRU-VNTR-linked cases from the Horn of Africa that predated the outbreak, but all were more than five SNPs from the outbreak.



Genomic Population Structure in Lisbon, Portugal: an Overview



Genome-wide Phylogenetic Scenario based on 28 051 SNPs

:: N= 207 Clinical Isolates

:: Analysis Period: 1995-2016

:: Two main clades: **Lisboa3** (n=72) and **Q1** (n=35)

:: 16 Genomic Clusters (≤ 5 SNPs): 92 (44.4%) isolates

Drug Resistance:

101 MDR-TB

49 XDR-TB

38 Other Resistance

19 Susceptible

GC67/Q1

N=26

SIT1106/LAM4

GC8 Lisboa3

N=17

SIT20/LAM1

GC5/ Lisboa3

N=13

SIT20/LAM1

GC3/Lisboa3

N=6

SIT20/LAM1

GC160/Q1

N=4

SIT1106/LAM4

GC7/Lisboa3

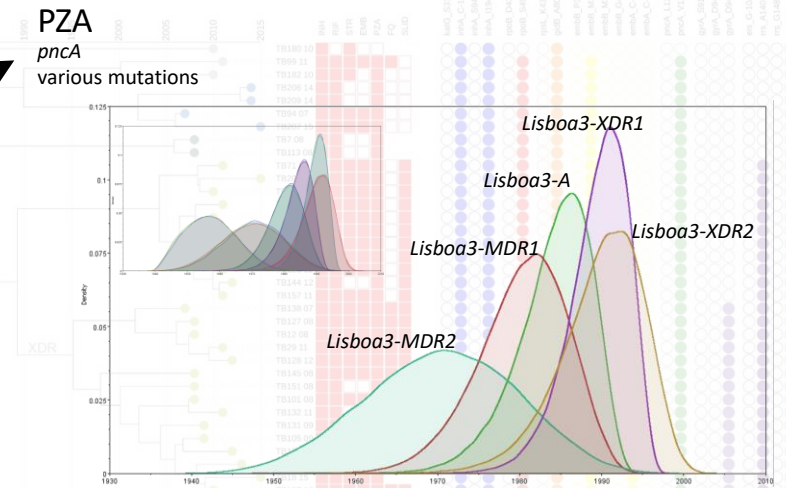
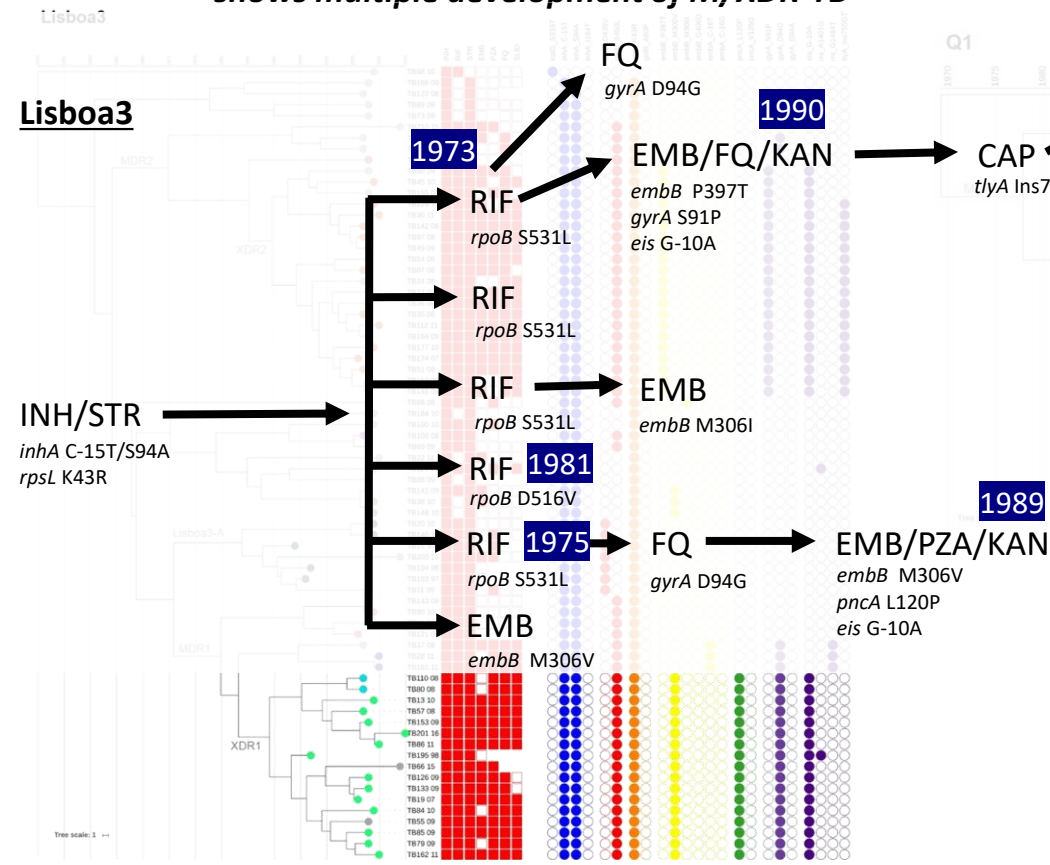
N=4

SIT20/LAM1

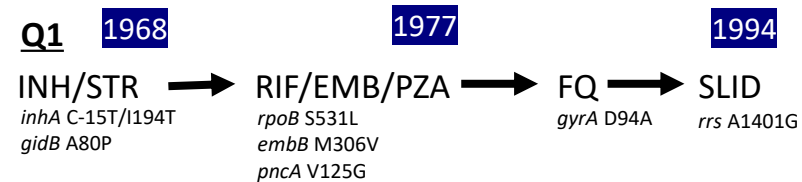
WGS and Microevolution towards Drug Resistance

Microevolutionary trajectory of the Lisboa3 clade shows multiple development of M/XDR-TB

Microevolutionary trajectory of the Q1 clade shows sequential development of M/XDR-TB



Bayesian Dating Estimates (BEAST)
Uncorrelated lognormal clock
Four 50 million MCMC Runs



MDR-TB Trends in Portugal: comparing two time periods...

Two time-periods: 2008-2011 and 2013-2017

Whole-genome sequence-based genotyping for 193 MDR-TB isolates nationwide (115.6% [+26] of notified isolates)

Four main clades of interest: Lisboa3, Q1, SIT92/X3 and Beijing strains (Total: 141/193 [73.1%])



Lisboa3

73 isolates
39 XDR-TB cases

Q1

37 isolates
16 XDR-TB cases

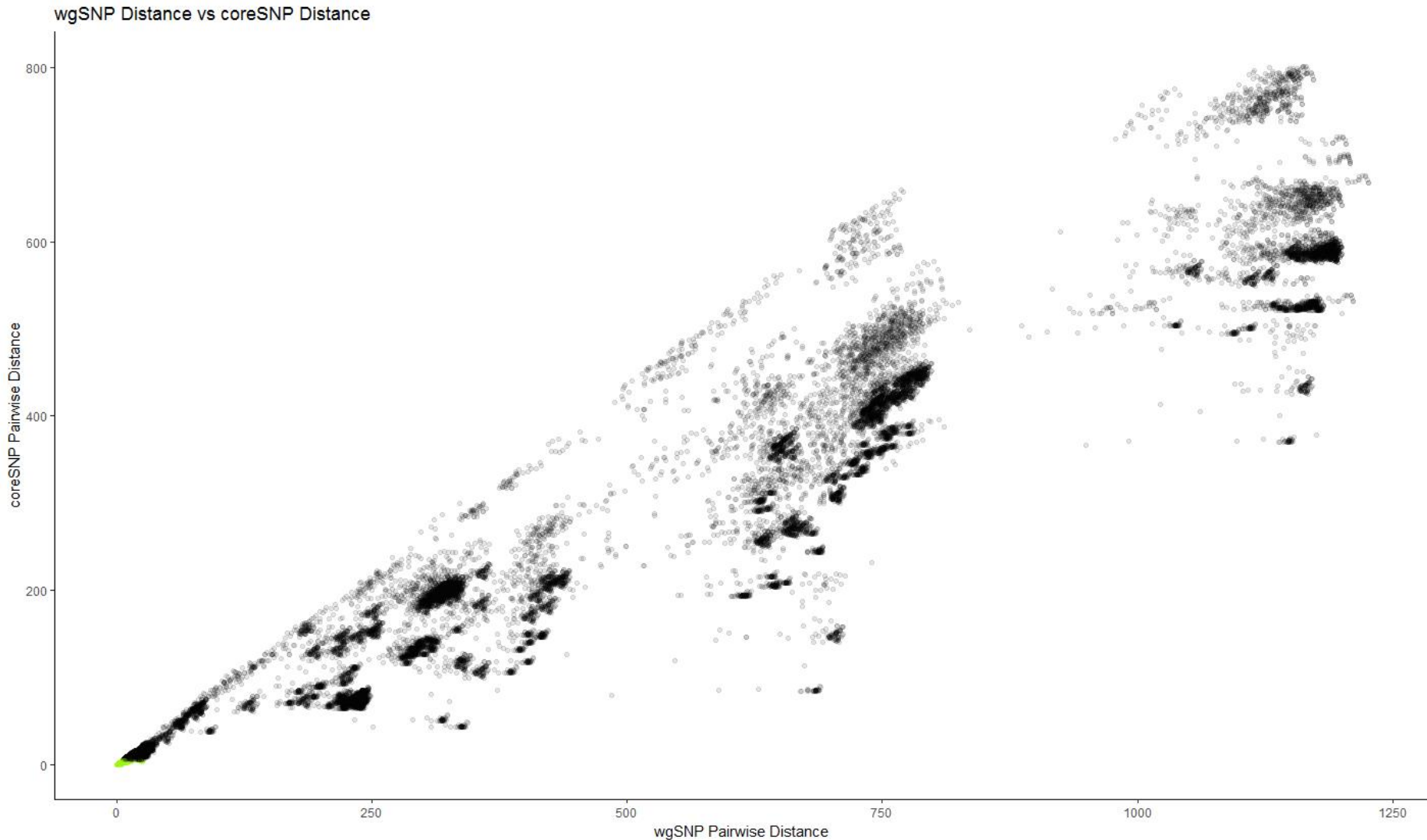
SIT92/X3

9 isolates
1 XDR-TB case

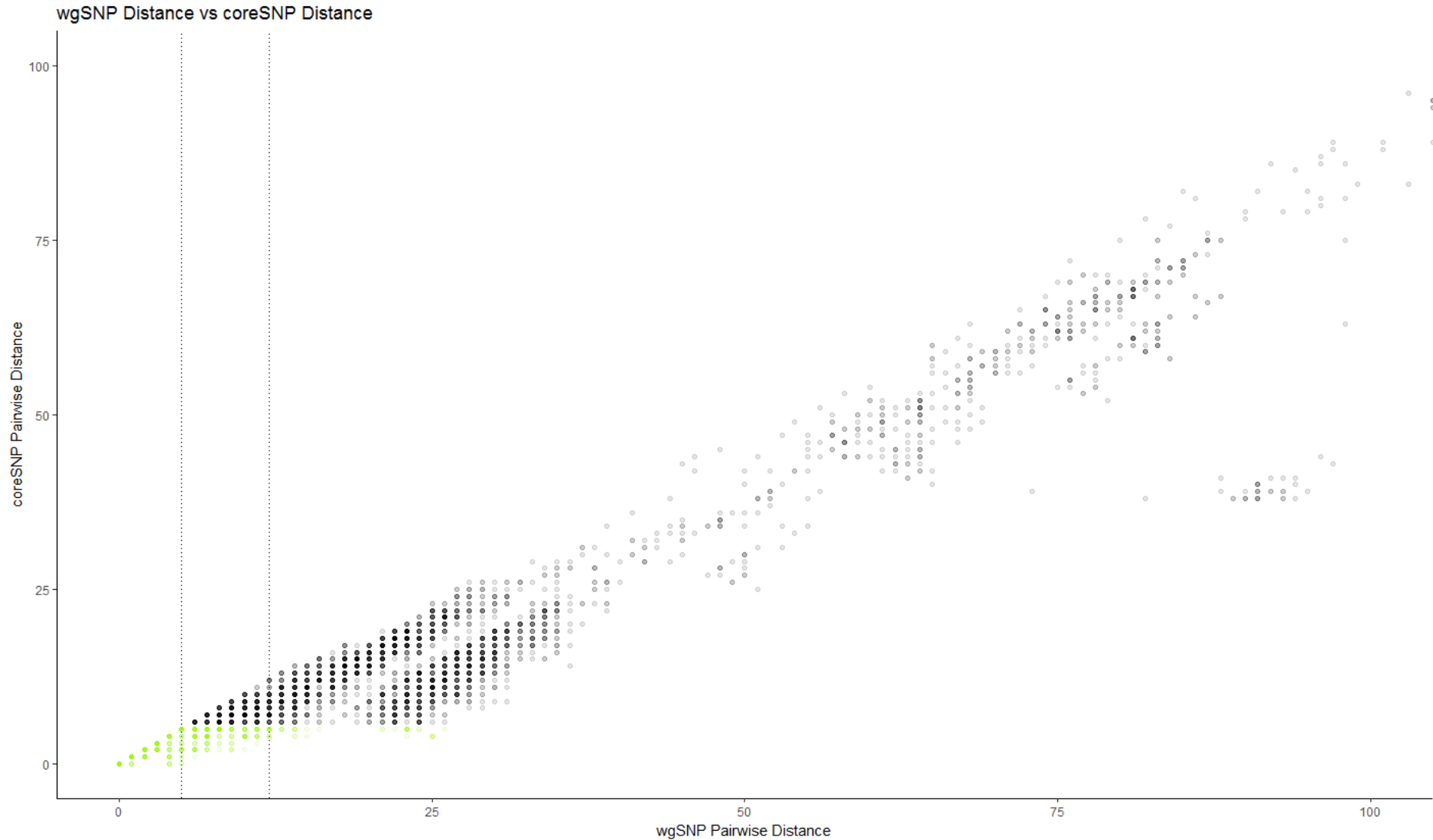
Beijing

22 isolates
1 XDR-TB case

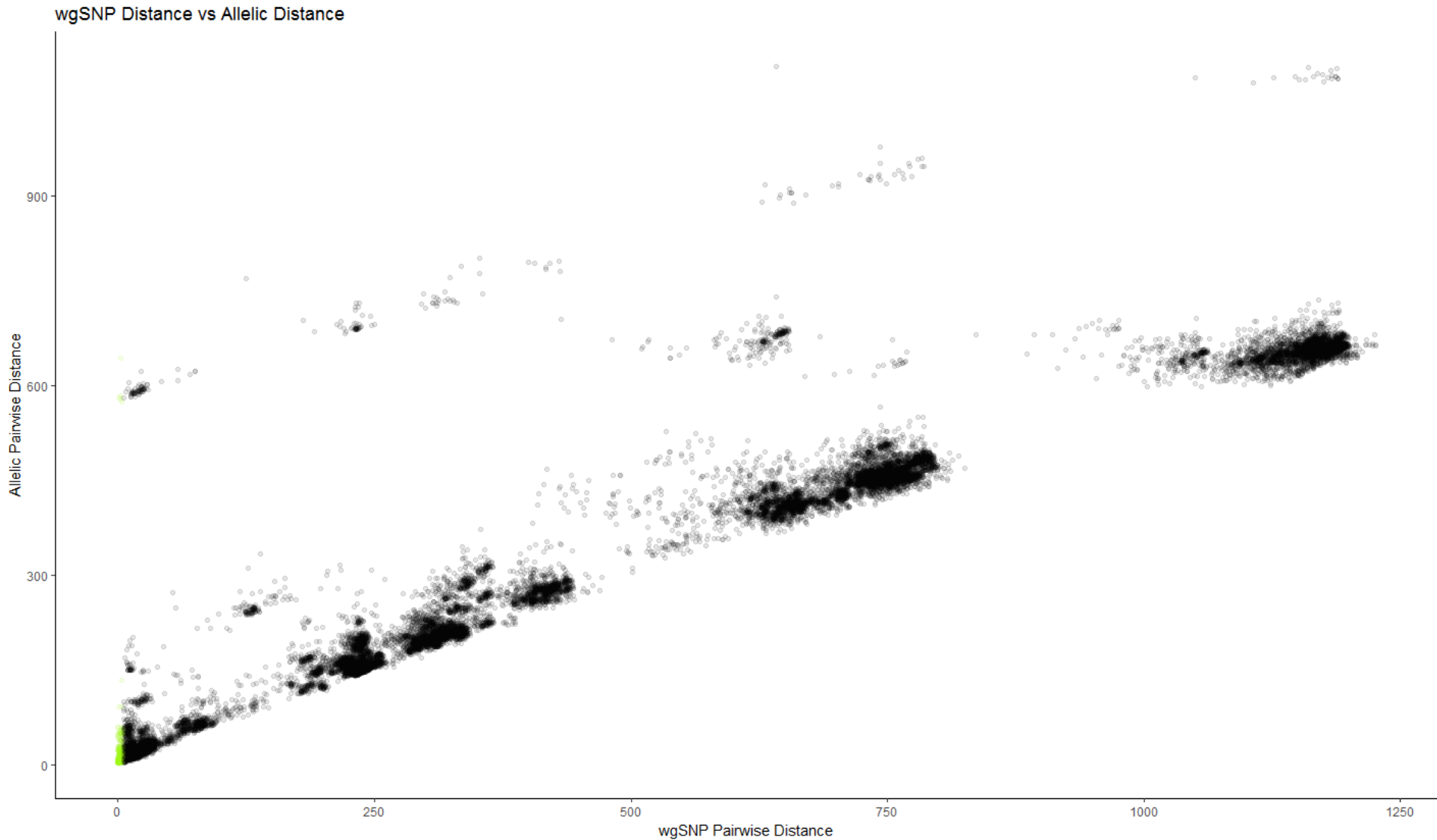
wgSNP Distance vs coreSNP distance: global view



wgSNP Distance vs coreSNP distance: closer look



wgSNP Distance vs Allelic distance: global view



wgSNP Distance vs Allelic distance: closer look

