INSTITUTE
OF TROPICAL
MEDICINE
ANTWERP

Phylodynamics of pathogenic mycobacteria

CONOR MEEHAN
UNIT OF MYCOBACTERIOLOGY

Acknowledgements

Mycobacteriology unit, ITM, Antwerp, Belgium

- Pauline Lempens
- Florian Gehre
- Suryia Akter
- **Bouke de Jong**

ADReM, UA, Antwerp, Belgium

- **Pieter Moris**

Computational Evolution, ETH Zurich, Switzerland

- Jūlija Pečerska
- Denise Kühnert
- Tanja Stadler

Mycobacteriology group, Research Center Borstel, Germany

- Thomas Kohl
- Matthias Merker
- **Stefan Niemann**

National TB program, Democratic Republic of Congo

- Michel Kaswa

European Research Council

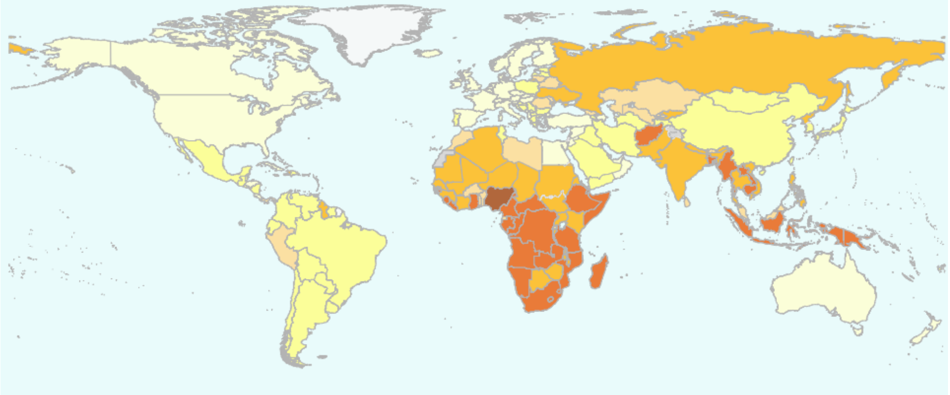
Established by the European Commission

Terminology

- Phylodynamics
  - The interface between evolutionary biology and epidemiology
  - Estimating pathogen evo/epi parameters from phylogenies
    - Mutation rates
    - Transmission rates and chains (R_0)
    - Population dynamics
- SNP
  - Single nucleotide polymorphism
  - Nucleotide that differs from the consensus/reference genome
  - SNP alignment contains only sites that have 2 or more nucleotides

Mycobacterium tuberculosis

- Causative agent of TB
  - ~1/3 of the world (supposedly) infected
  - 10.4M new cases per year



- Transmitted by aerosols
  - Close contact between people
  - Long latency before active disease
- Transmission important in drug resistance
  - Transmission > point mutations
  - Some multidrug resistant (MDR) strains circulating for decades

Mycobacterium tuberculosis clustering

- Several methods have been developed to look for transmission clusters of *M. tuberculosis*
- Classical genotyping methods:
  - Spoligotype (low resolution)
  - MIRU-VNTR (medium resolution)
  - SpoNC (high resolution)

Spoligotyping

- Direct repeat region of the genome
    - Repeats seperated by spacers
- PCR the spacers
    - Repeats used for primers
- Hybridize to a membrane with spacers
- Presence/absence of spacers gives genotype

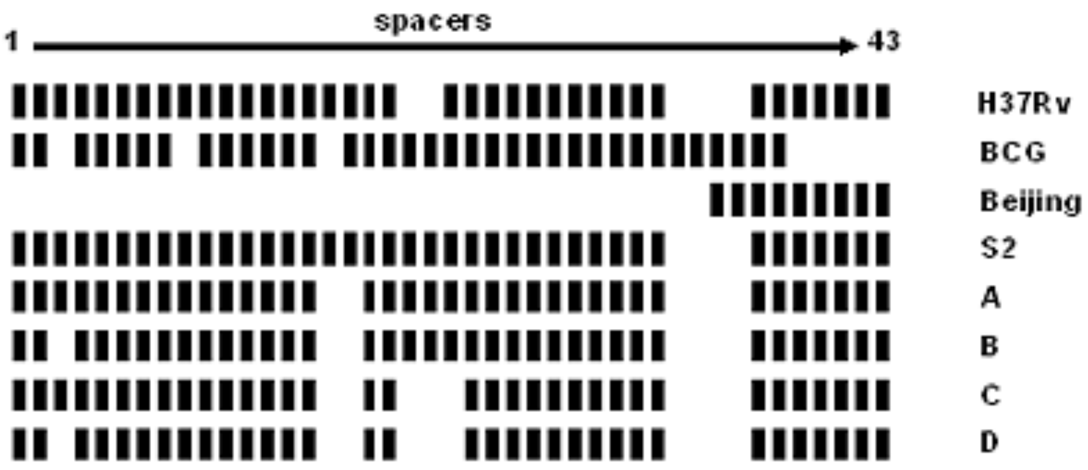
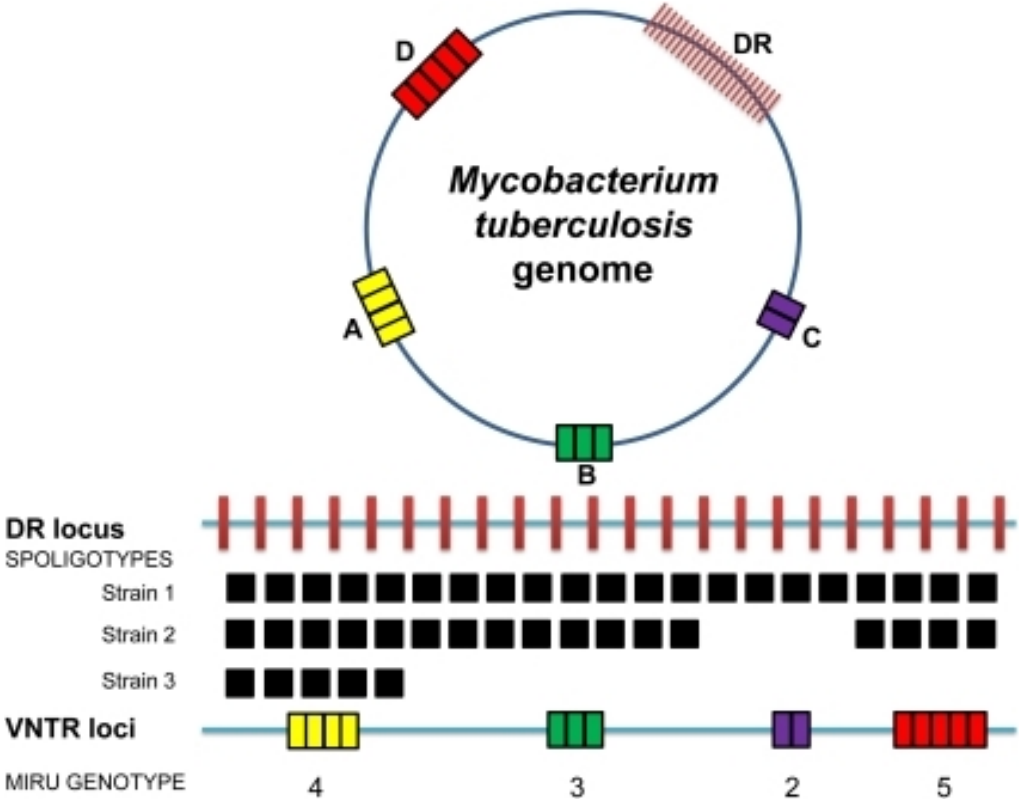


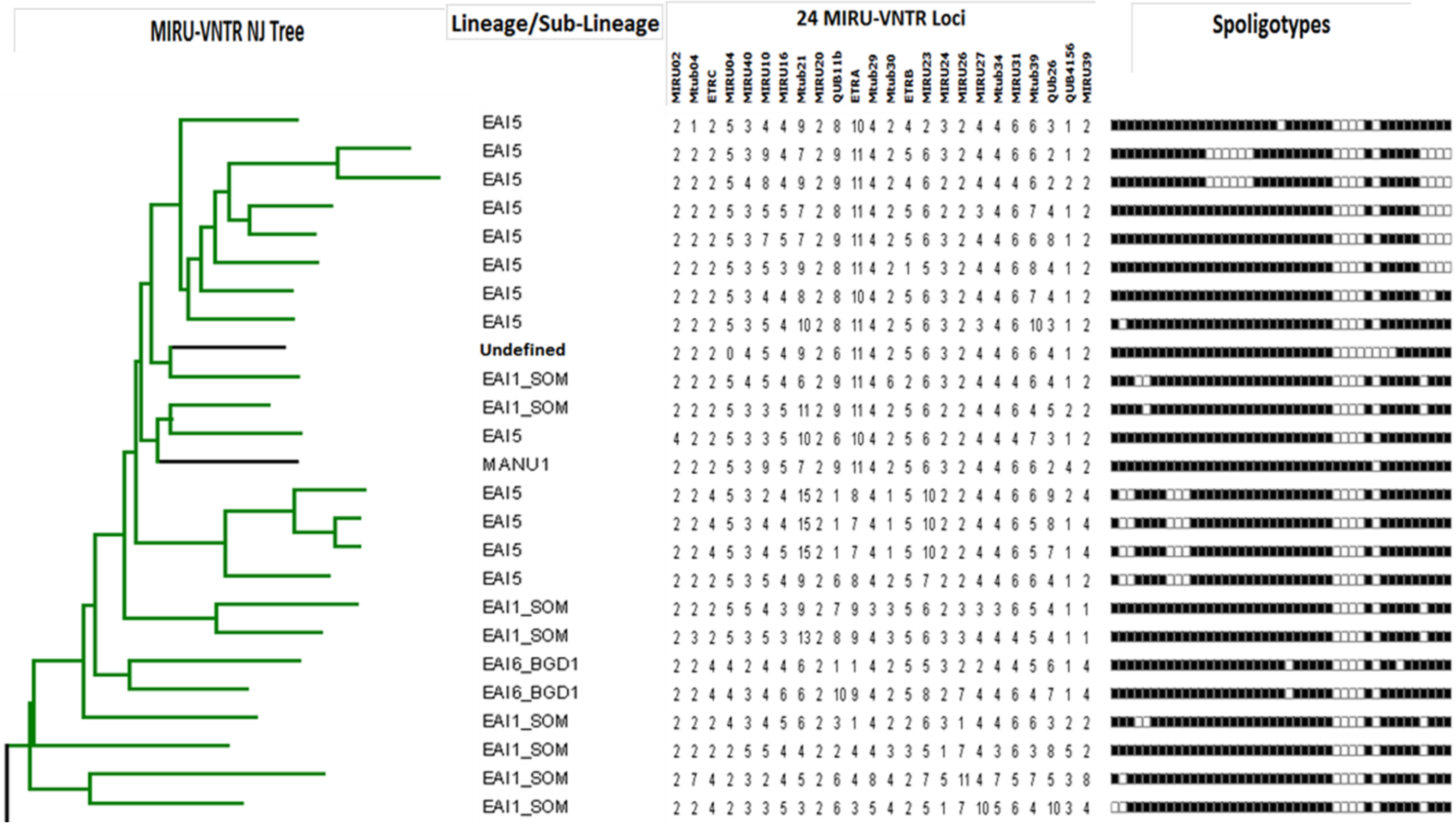
Image from CDC

MIRU-VNTR

- Mycobacterium Interspersed Repetitive Units- Variable Number Tandem Repeats
- Tandem repeat spread throughout the genome
  - 24 loci chosen for genotyping
- PCR the repeat
  - Size of amplicon indicates number of copies at location
- Pattern gives a fingerprint



MIRU-VNTR and Spoligotyping



Lineage/Sub-Lineage	MIRU02	Mtub04	ETRC	MIRU04	MIRU40	MIRU10	MIRU16	Mtub21	MIRU20	QUB11b	ETRA	Mtub29	Mtub30	ETRB	MIRU23	MIRU24	MIRU26	MIRU27	Mtub34	MIRU31	Mtub39	QUb26	QUB4156	MIRU39
EAI5	2	1	2	5	3	4	4	9	2	8	10	4	2	4	2	3	2	4	4	6	6	3	1	2
EAI5	2	2	2	5	3	9	4	7	2	9	11	4	2	5	6	3	2	4	4	6	6	2	1	2
EAI5	2	2	2	5	4	8	4	9	2	9	11	4	2	4	6	2	2	4	4	4	6	2	2	2
EAI5	2	2	2	5	3	5	5	7	2	8	11	4	2	5	6	2	2	3	4	6	7	4	1	2
EAI5	2	2	2	5	3	7	5	7	2	9	11	4	2	5	6	3	2	4	4	6	6	8	1	2
EAI5	2	2	2	5	3	5	3	9	2	8	11	4	2	1	5	3	2	4	4	6	8	4	1	2
EAI5	2	2	2	5	3	4	4	8	2	8	10	4	2	5	6	3	2	4	4	6	7	4	1	2
EAI5	2	2	2	5	3	5	4	10	2	8	11	4	2	5	6	3	2	3	4	6	10	3	1	2
Undefined	2	2	2	0	4	5	4	9	2	6	11	4	2	5	6	3	2	4	4	6	6	4	1	2
EAI1_SOM	2	2	2	5	4	5	4	6	2	9	11	4	6	2	6	3	2	4	4	4	6	4	1	2
EAI1_SOM	2	2	2	5	3	3	5	11	2	9	11	4	2	5	6	2	2	4	4	6	4	5	2	2
EAI5	4	2	2	5	3	3	5	10	2	6	10	4	2	5	6	2	2	4	4	4	7	3	1	2
MANU1	2	2	2	5	3	9	5	7	2	9	11	4	2	5	6	3	2	4	4	6	6	2	4	2
EAI5	2	2	4	5	3	2	4	15	2	1	8	4	1	5	10	2	2	4	4	6	6	9	2	4
EAI5	2	2	4	5	3	4	4	15	2	1	7	4	1	5	10	2	2	4	4	6	5	8	1	4
EAI5	2	2	4	5	3	4	5	15	2	1	7	4	1	5	10	2	2	4	4	6	5	7	1	4
EAI5	2	2	2	5	3	5	4	9	2	6	8	4	2	5	7	2	2	4	4	6	6	4	1	2
EAI1_SOM	2	2	2	5	5	4	3	9	2	7	9	3	3	5	6	2	3	3	3	6	5	4	1	1
EAI1_SOM	2	3	2	5	3	5	3	13	2	8	9	4	3	5	6	3	3	4	4	4	5	4	1	1
EAI6_BGD1	2	2	4	4	2	4	4	6	2	1	1	4	2	5	5	3	2	2	4	4	5	6	1	4
EAI6_BGD1	2	2	4	4	3	4	6	6	2	10	9	4	2	5	8	2	7	4	4	6	4	7	1	4
EAI1_SOM	2	2	2	4	3	4	5	6	2	3	1	4	2	2	6	3	1	4	4	6	6	3	2	2
EAI1_SOM	2	2	2	2	5	5	4	4	2	2	4	4	3	3	5	1	7	4	3	6	3	8	5	2
EAI1_SOM	2	7	4	2	3	2	4	5	2	6	4	8	4	2	7	5	11	4	7	5	7	5	3	8
EAI1_SOM	2	2	4	2	3	3	5	3	2	6	3	5	4	2	5	1	7	10	5	6	4	10	3	4

Devi et al PLOS ONE 2015

SpoNC

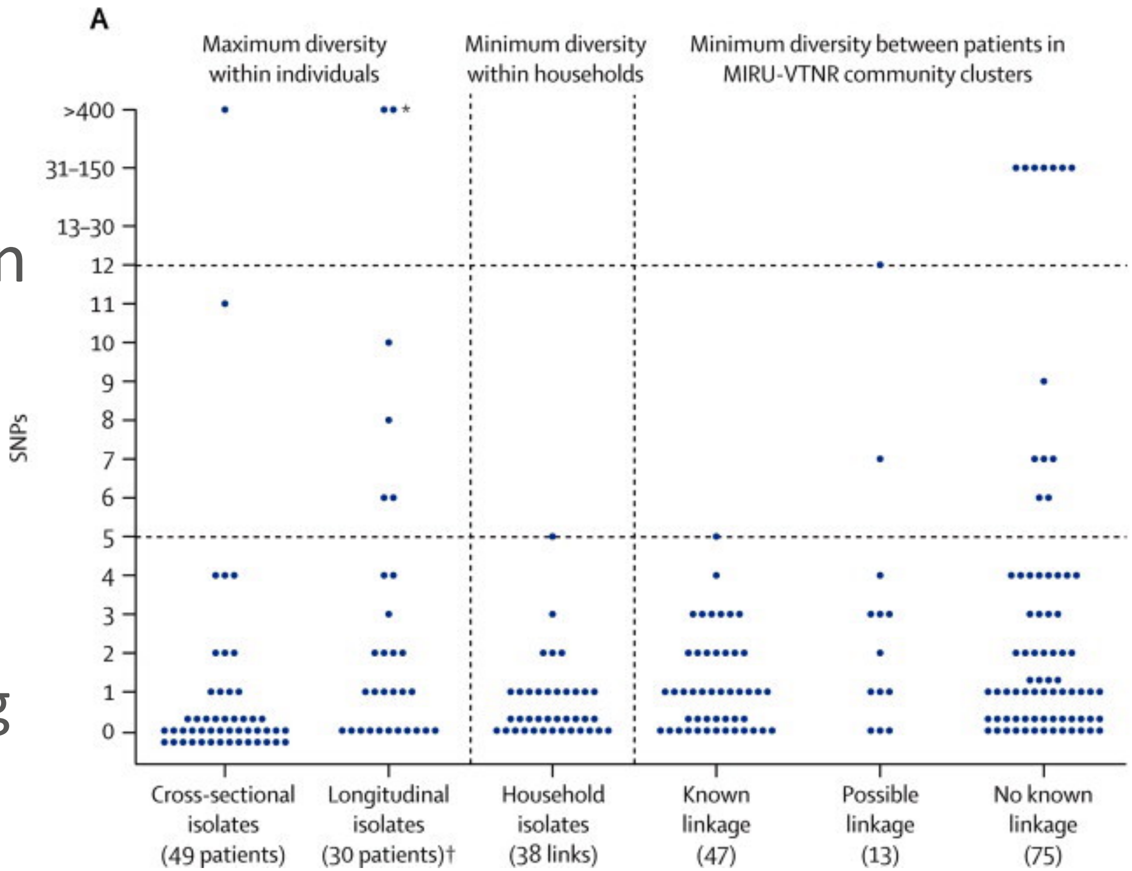
- Mutations tend to occur often within the *pncA* gene
  - Many associated with resistance to pyrazinimide
- Thought that convergence of these mutations is low
- Combination of *pncA* mutations with Spoligotyping increases resolution
  - Referred to as the SpoNC method
  - Only works for those with *pncA* mutations
  - Quite sparse

Mycobacterium tuberculosis clustering

- Several methods have been developed to look for transmission clusters of *M. tuberculosis*
- Classical genotyping methods:
  - Spoligotype (low resolution)
  - MIRU-VNTR (medium resolution)
  - SpoNC (high resolution)
- Whole genome sequencing allows for high resolution clustering
  - 4.4Mb circular genome
  - Chains of transmission
  - SNP thresholds

SNP thresholds

- Walker *et al* (2013) compared 390 isolates to look for SNP thresholds of likely tranmission
  - < 5 very likely; > 12 very unlikely
  - ~0.3-0.5 SNPs per year
- Unknown how these relate to phylogenetic distances/clustering



cgMLST

- Comparison of SNPs is heavily reliant on accurate SNP calling
  - Different pipelines/versions can give small/large differences in SNP calls

- Core Group Multi-Locus Sequence Typing (cgMLST)
  - Pre-selected set of genes to call alleles in
  - Create matrix based on allele distances
  - Applied to *M. tuberculosis* by Kohl et al (2015)
    - Stated that 5/12 cgMLST grouping is similar to 5/12 SNP groupings

Partitional SNP clustering

- Tight clusters
  - Every isolate is within the SNP cut-off distance of all others
  - Isolates may belong to 2 clusters
- Loose clustering
  - Every isolate is within the SNP cut-off distance of at least one other
  - Isolates belong to one cluster only
  - May create long chains of sparse connections

	T1	T2	T3	T4
T1	-	3	6	7
T2		-	3	4
T3			-	1
T4				-

Applying 5 SNP cut-off

Tight clusters:

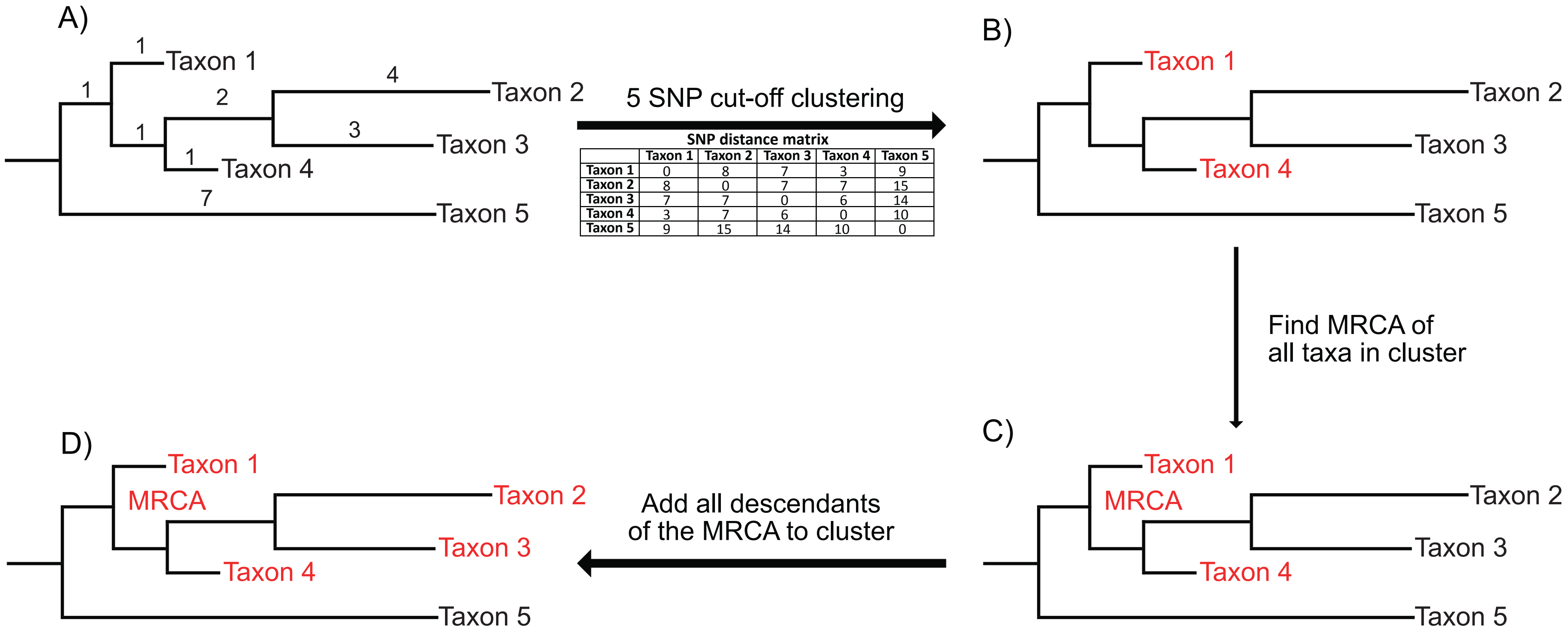
C1: T1 T2

C2: T2 T3 T4

Loose clusters:

C1: T1 T2 T3 T4

Phylogenetic inclusion method (extension of loose clusters)



5 SNP cut-off clustering

SNP distance matrix

	Taxon 1	Taxon 2	Taxon 3	Taxon 4	Taxon 5
Taxon 1	0	8	7	3	9
Taxon 2	8	0	7	7	15
Taxon 3	7	7	0	6	14
Taxon 4	3	7	6	0	10
Taxon 5	9	15	14	10	0

Find MRCA of
all taxa in cluster

Add all descendants
of the MRCA to cluster

Aim

- Aid public health initiatives in selecting the best approach for tracking transmissions
  - Good resolution in the time scale of their study
  - Cost/resolution trade-off

- What is the best method for finding chains of recent transmission?

- What time period is covered by each clustering method?

Transmission clusters estimated by 20 methods

- Spoligotype
  - Membrane-based (Mem-Spo)
  - Genome-Based (Gen-Spo)
- MIRU-VNTR
- Combination of both
  - MemSpo-MIRU
  - GenSpo-MIRU
- Above methods with pncA mutation clustering
  - MemSpo-NC
  - GenSpo-NC
  - MIRU-NC
- 4 different SNP cut-offs
  - 0, 1, 5, 12
- Extension of these cut-offs (Phylogenetic inclusion method; clades)
- 4 different cgMLST cut-offs
  - 0, 1, 5, 12

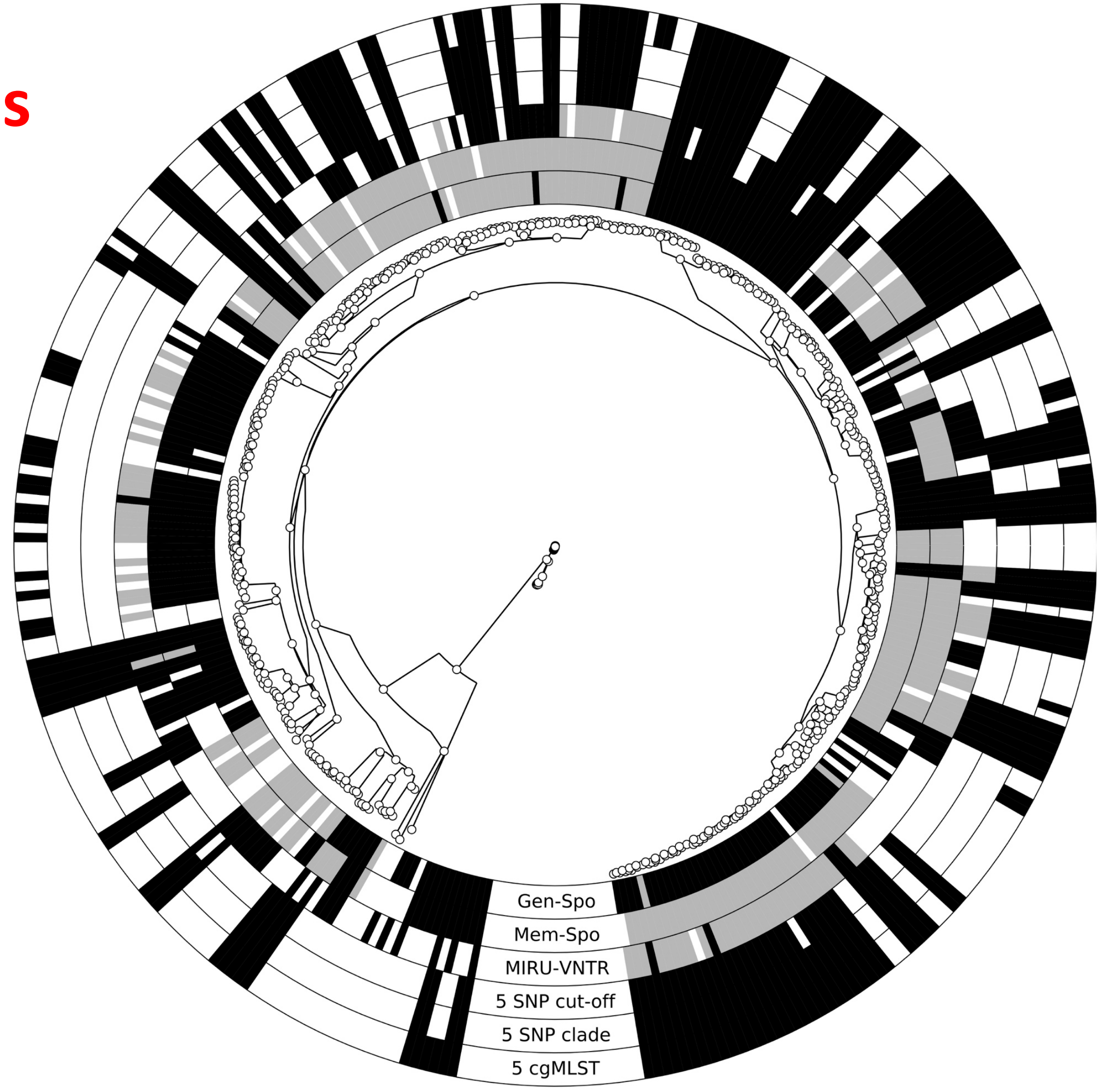
Kinshasa dataset

- Democratic Republic of Congo
  - One of 22 high prevelance countries
- MDR-TB dataset
  - 324 samples
  - 2005-2010
  - Retreatment cases
  - Rifampicin resistant (+ isoniazid for most)
- WGS reads -> SNPs with MTBseq

Clustering method comparisons

- Maximum likelihood tree with RAxML-NG
  - GTR+γ model with Stamatakis ascertainment bias correction
  - 10 starting trees, 100 bootstraps
- Map clusters from different methods on to tree
- Visualise with GraPhlan

- Black: in a cluster
- Grey: in a cluster affected by convergence

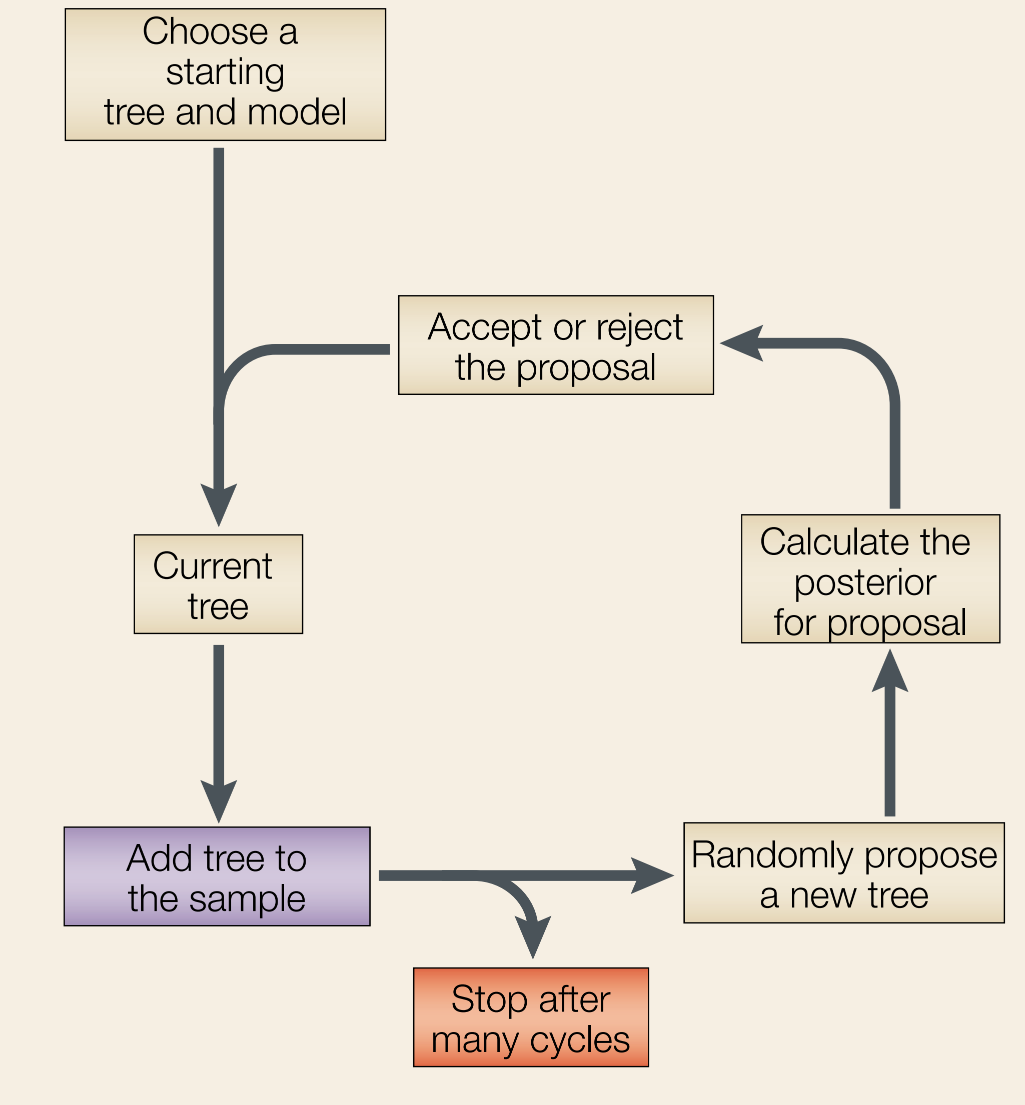


Assigning timespans

- Phylogenetic dating method
- Tree built using the Bayesian method in BEAST-2

Bayesian analysis (MCMC)

- Each cycle is a step in a Monte Carlo Markov Chain (MCMC)
- Can store the parameters (tree, rates etc) at each step or subset of steps
- Must assess for convergence to see if the cycles are finding the same answer repeatedly and thus can be stopped


Choose a starting tree and model
Accept or reject the proposal
Current tree
Calculate the posterior for proposal
Add tree to the sample
Randomly propose a new tree
Stop after many cycles

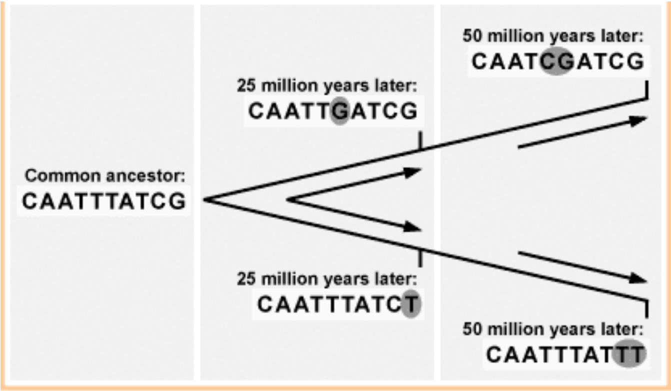
Assigning timespans

- Phylogenetic dating method
- Tree built using the Bayesian method in BEAST-2
- Constant coalescent population model
- Relaxed clock model
  - Diffuse prior between 1x10^{-7} and 1x10^{-8}
    - Started previously as the rate for *M. tuberculosis*

Molecular Clock

- Assumes that there is a stocastic relationship between time and the rate of mutation of a gene

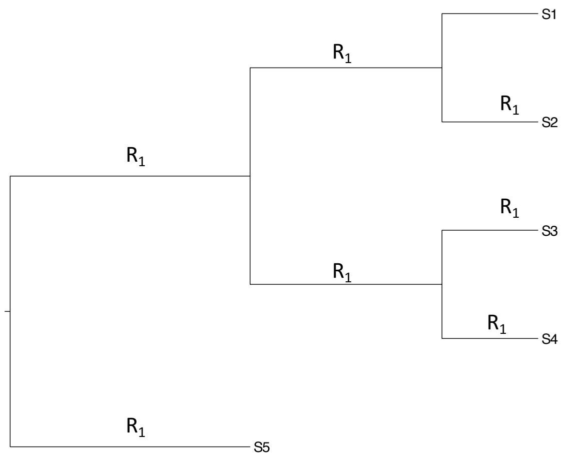
- If we know this rate, we can correlate it to changes through time and estimate a divergence time



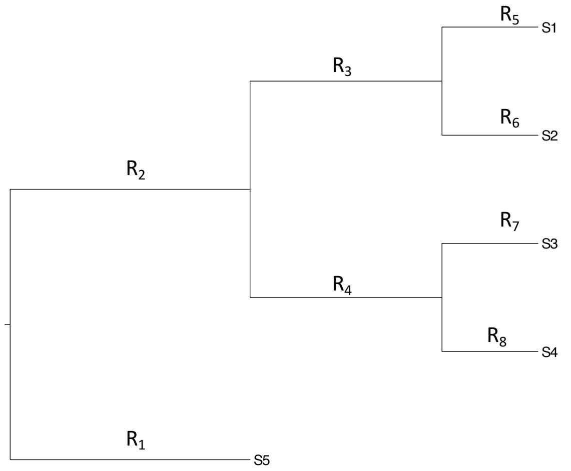
http://evolution.berkeley.edu/evosite/evo101/IIE1cMolecularclocks.shtml

Strict vs Relaxed

- Strict:
  - One rate of mutation for every branch in tree

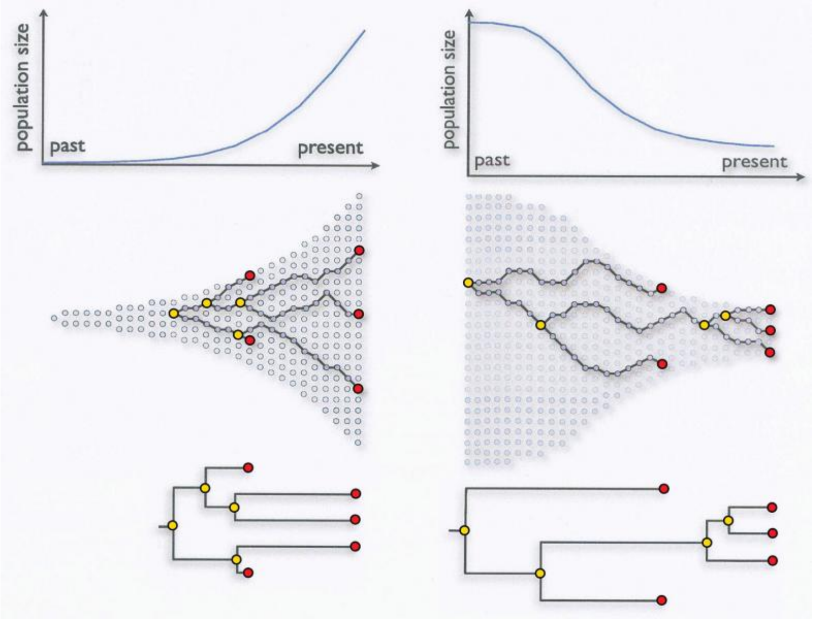


- Relaxed
  - Different rate for each branch
  - Correlated:
    - Rates influenced by the rate of the parent branch
  - Uncorrelated
    - Rates independent



Coalescent

- A model of population evolution
- When two alleles/individuals come together to form an ancestor (i.e. two branches meet) this is called a coalescent event



Sainani Biomedical computation review (2009)

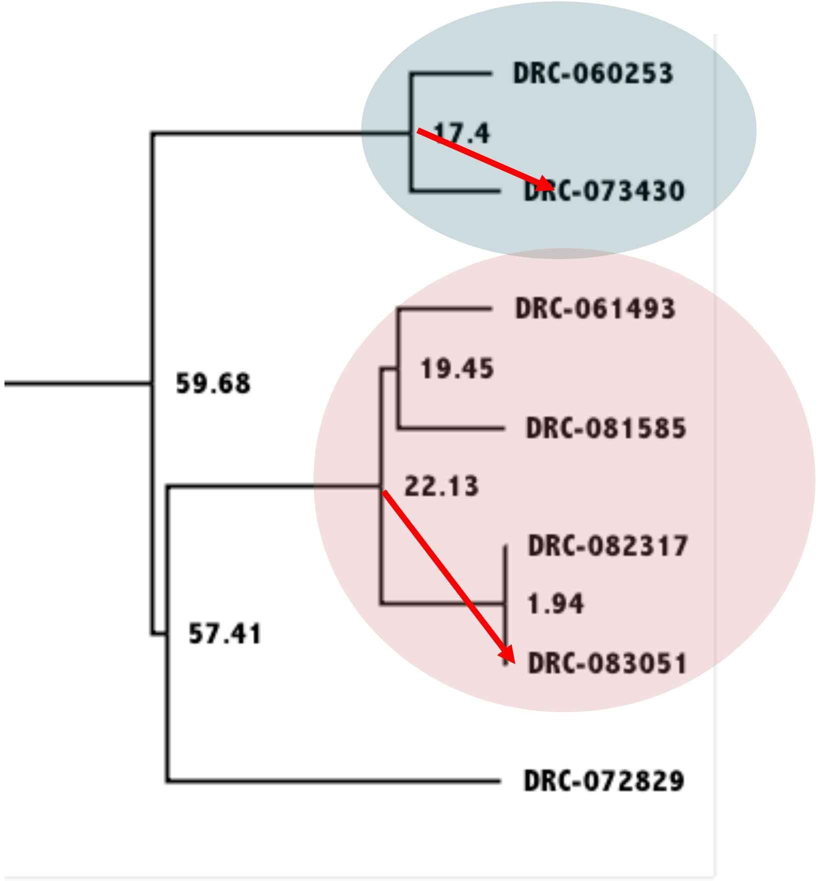
Assigning timespans

- Phylogenetic dating method
- Tree built using the Bayesian method in BEAST-2
- Constant coalescent population model
- Relaxed clock model
  - Diffuse prior between $1x10^{-7}$ and $1x10^{-8}$

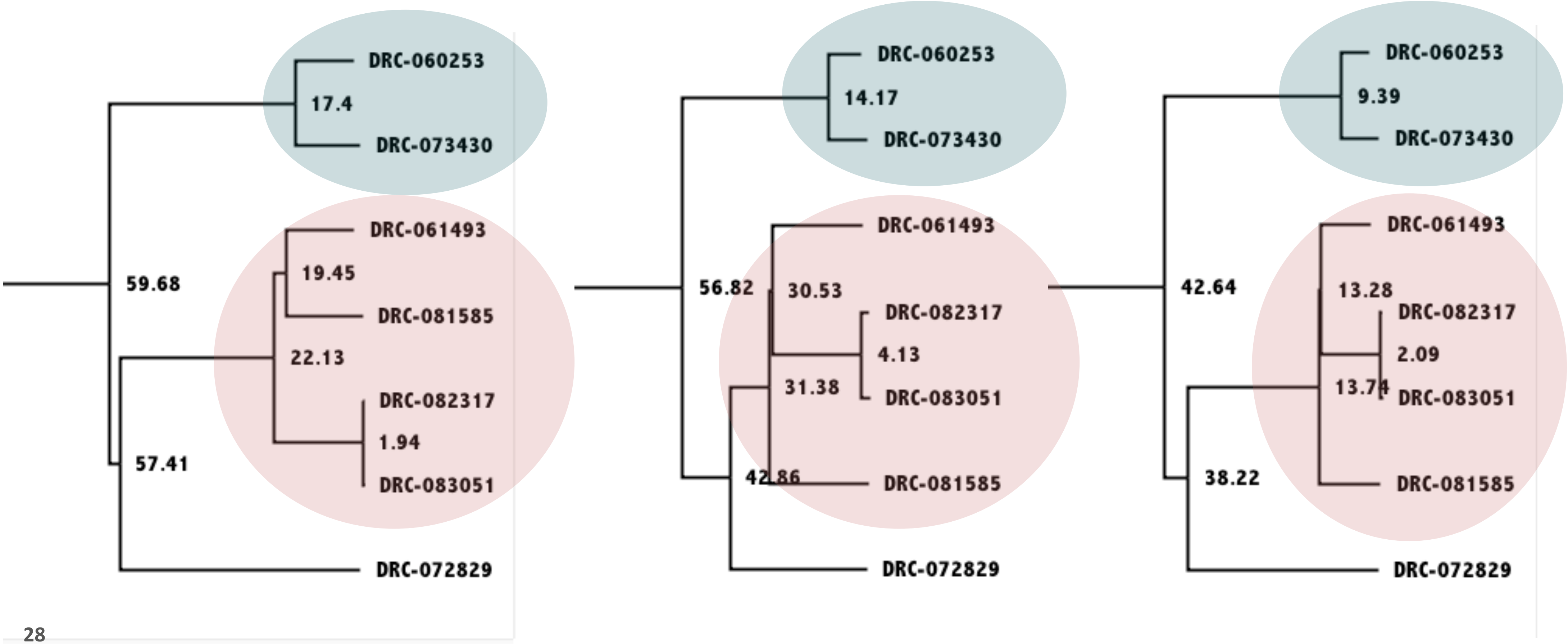
- Find the age of the MRCA of each cluster

Assessing ages from MCMC

- Place clusters on tree according to method
- Get the MRCA of each cluster
- Age is the difference in time between the MRCA and the furthest isolate
  - Internal node (MRCA) times are time before the last sample in the dataset (2010)
    - 17.4=~1993
    - 22.13=~1988
  - Age:
    - 2007 - 1993 = 14
    - 2008 - 1988 = 20
- Aggregate over all the trees in the MCMC run



Assessing ages from MCMC



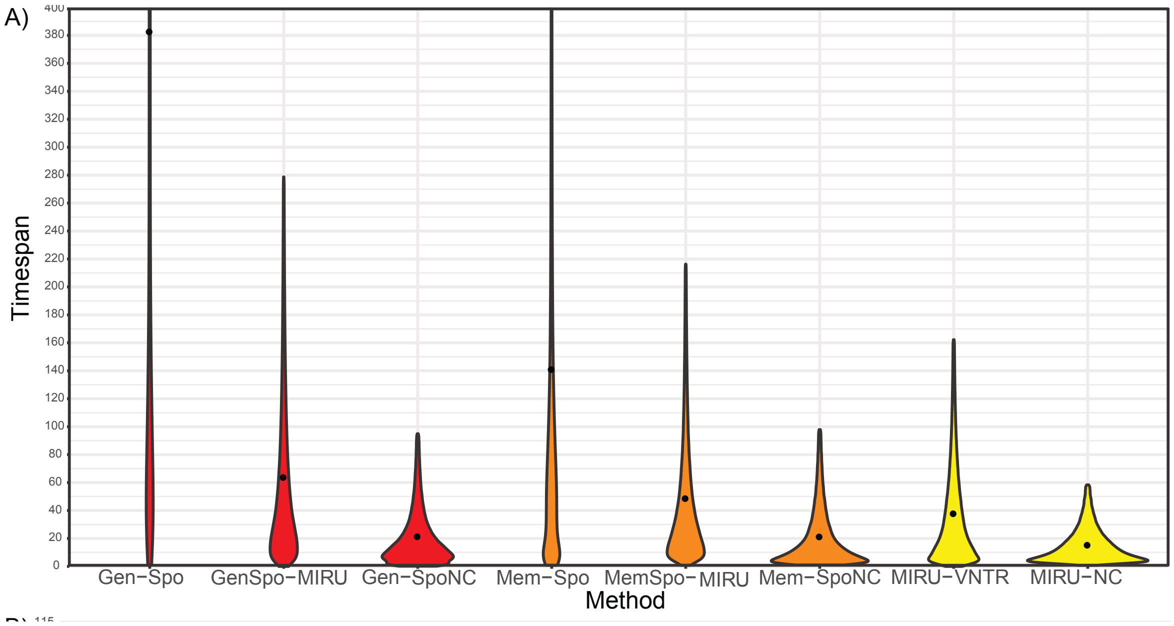
Assessing ages from MCMC

Cluster	Age 1	Age 2	Age 3	…
1	13	11	6	…
2	20	19	11	…

- Aggregate all ages of all clusters in all MCMC samples per method
  - ~20,000 samples per method
- Use R to get the mean and the 95% HPD
- Visualise with violin plots

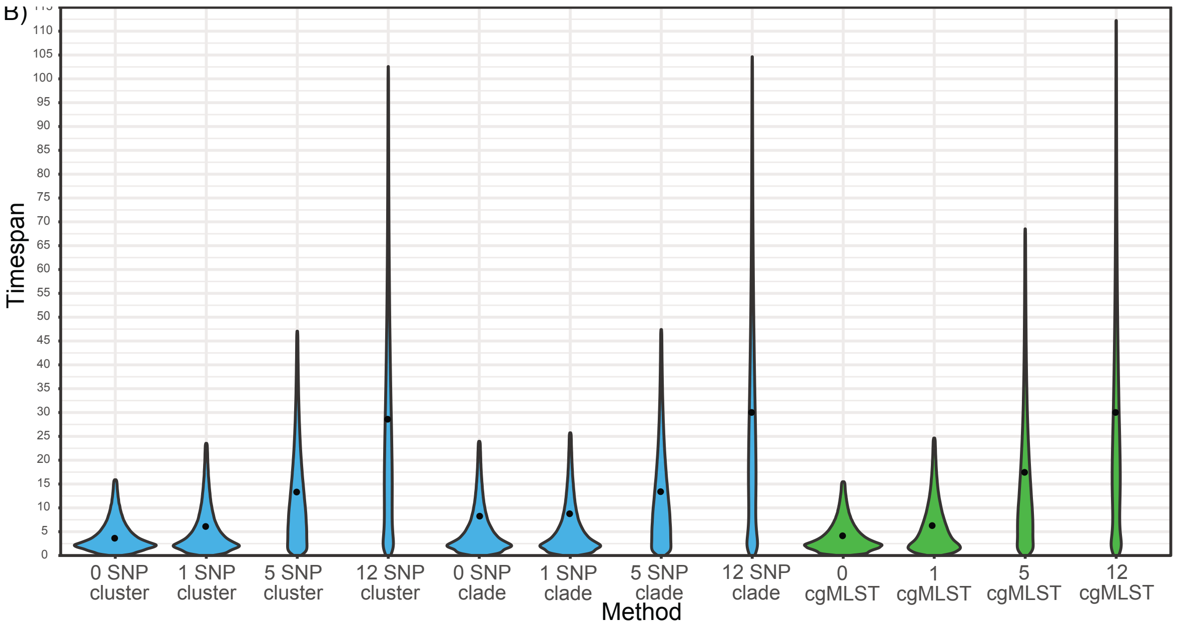
Spoligotype and MIRU-VNTR (classical methods)

Method	Mean Timespan	Max Timespan 95% HPD
Gen-Spo	383	1893
GenSpo-MIRU	64	278
Gen-SpoNC	22	94
Mem-Spo	141	823
MemSpo-MIRU	49	216
Mem-SpoNC	21	97
MIRU-VNTR	38	162
MIRU-NC	15	58



SNP based methods (WGS-based methods)

Method	Mean Timespan	Max Timespan 95% HPD
0 SNP cluster	4	16
1 SNP cluster	6	24
5 SNP cluster	13	47
12 SNP cluster	29	103
0 SNP clade	6	24
1 SNP clade	6	26
5 SNP clade	13	47
12 SNP clade	30	105
0 allele cgMLST	4	15
1 allele cgMLST	6	25
5 allele cgMLST	18	69
12 allele cgMLST	30	112



Drawbacks and improvements

- SNP –based methods have their own issues
  - No standard pipelines
    - Reference genomes and SNP calls
  - Exclude repetitive regions
    - Known to have many mutations
    - Cause structural rearrangements
  - Solution?
    - Long read sequencing (maybe)
- Drug resistance and mutation rates
- Confirmation of transmission links

Integrating different methods into public health

- Spoligotype
  - Cheap and quick
  - High convergence rates
  - Suitable for assigning 7 primary lineages
- MIRU-VNTR
  - Somewhat quick
  - Medium/low convergence rates
  - Suitable for country surveillance
- 12 SNP/cgMLST
  - Slower method (for now) and most expensive
  - Best for country surveillance
- 0, 1, 5 SNP/cgMLST
  - Slower method (for now) and most expensive
  - Best for recent transmission studies
    - Different cut-offs for different timespans

INSTITUTE OF TROPICAL MEDICINE ANTWERP

Conor Meehan

cmeehan@itg.be