WGS for Genotyping and Barcoding

PRASIT PALITTAPONGARNPIM, M.D.

Genotyping Microbes

- What are genotypes?
- Why genotyping?
- How to genotypes?
- WGS and genotypes?

What are genotypes?

- Classification of organisms based on genetic variations.
- The purposes of genotyping include
  - Taxonomy including subspecies classfication.
    - For correlation with virulence, epidemiological parameters, etc.
  - Non-taxonomic-related correlation with important properties such as
    - Drug resistance
    - Outbreak vs not outbreak
    - Virulence
    - Etc.

How to genotype haploid organisms? 4

- Ribotyping
- Restriction fragment length polymorphism
  - PFGE (pulsed field gel electrophoresis)
  - Southern hybridization
  - PCR-REA (-Restriction enzyme analysis)
- VNTR (variable number of tandem repeat) typing
- MLST (multilocus sequence typing)
- Others
- WGS

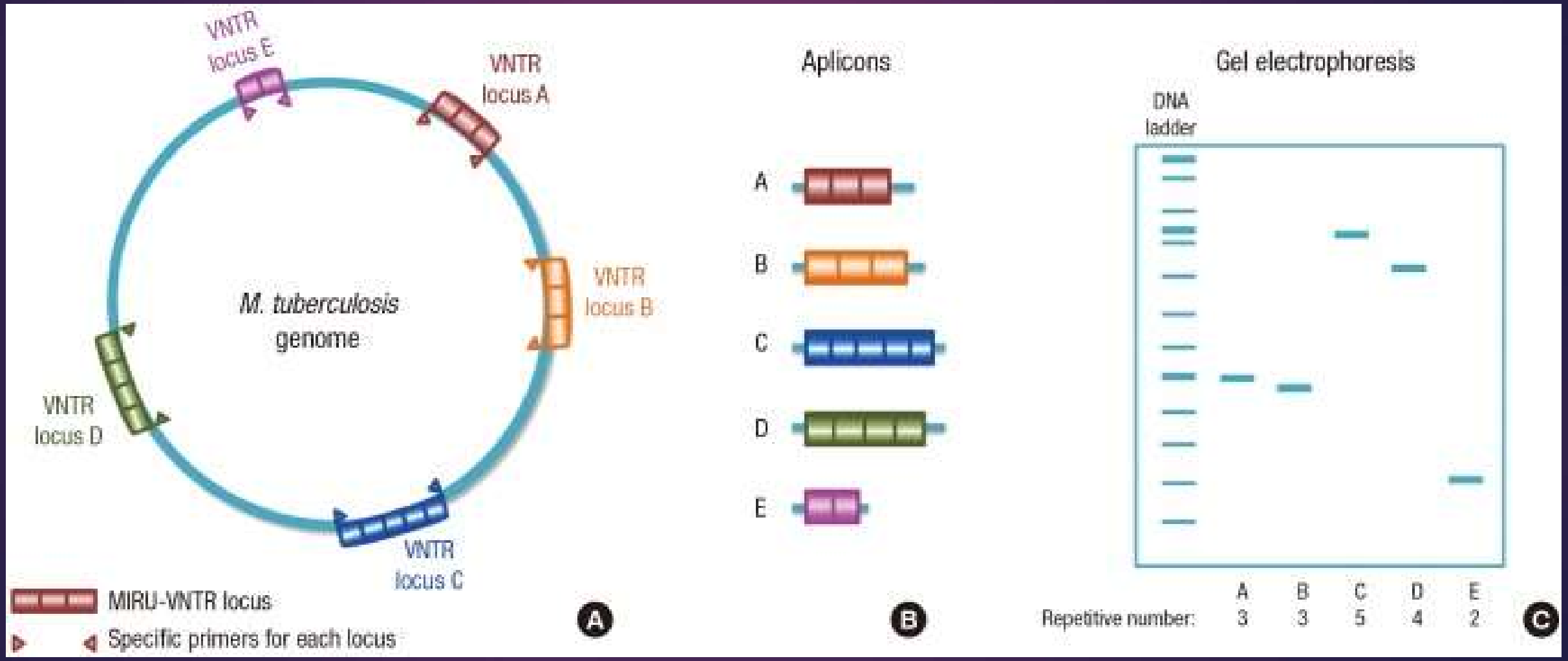
Ideal Properties of Genotyping

- High reproducibility
- High discriminatory power
- Digital results
- Cheap
- High throughput
- Ability to infer a common ancestor.

Genomic Variations Identified by WGS

- **Single nucleotide variants (SNV, point mutations)**
  - The term SNP is typically reserved for variants with >1% of population.
- **CNV** (copy number variations)(e.g. VNTR typing). Not commonly done by 2nd generation sequencing
- Gene profile variations
  - Large sequence polymorphism
  - Accessory gene variations
- Inversion/duplication of large chromosomal segments

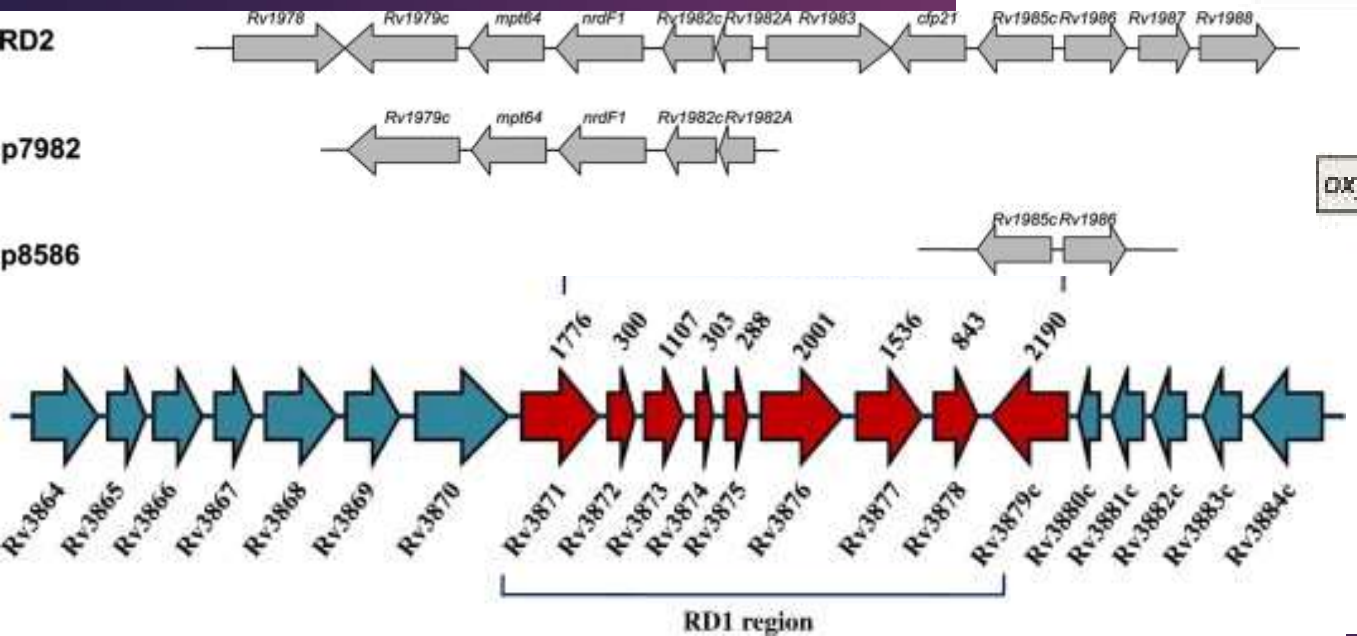
VNTR Typing of *Mycobacterium tuberculosis*

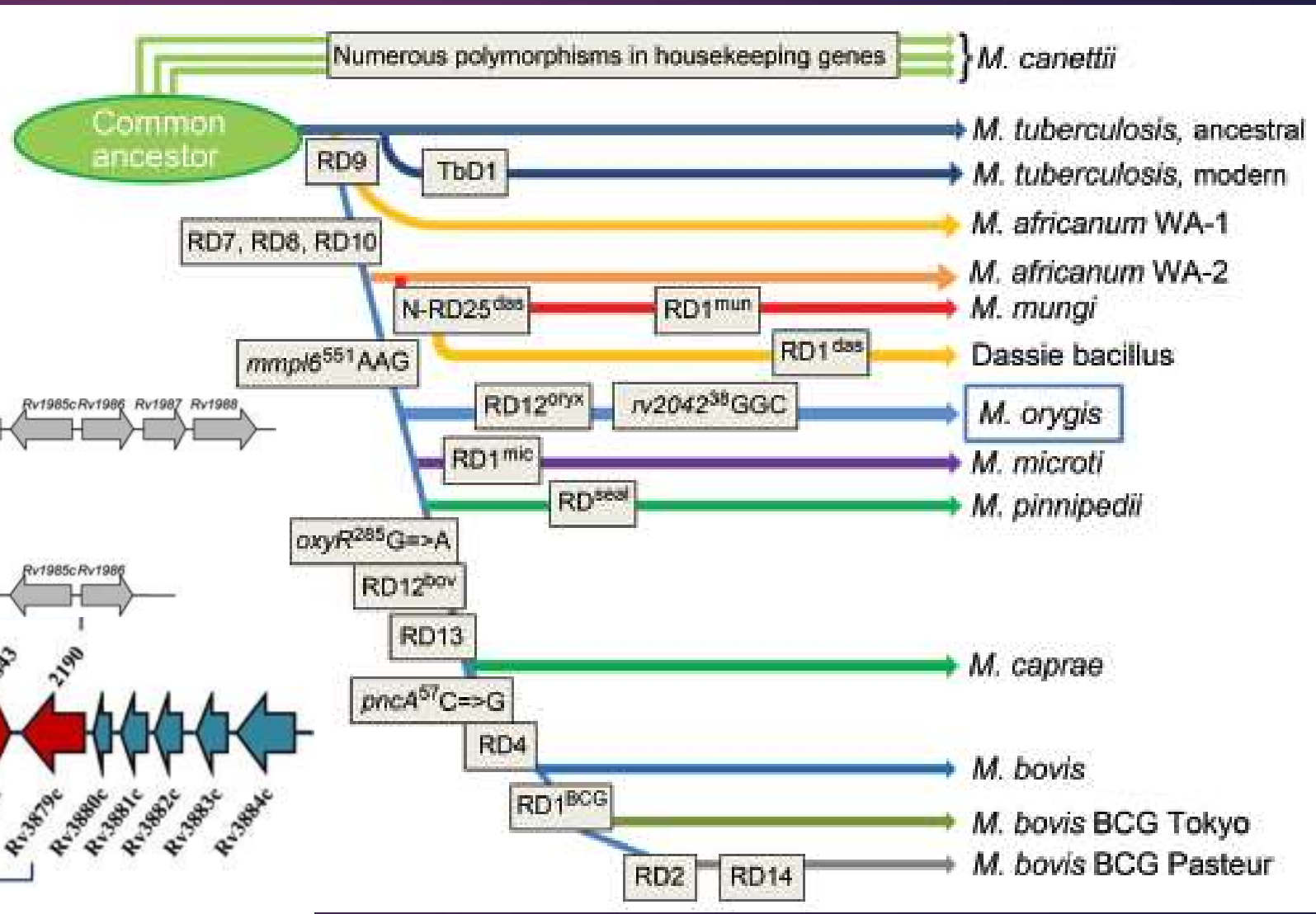


MTB LSP is used for species and lineage classifications.

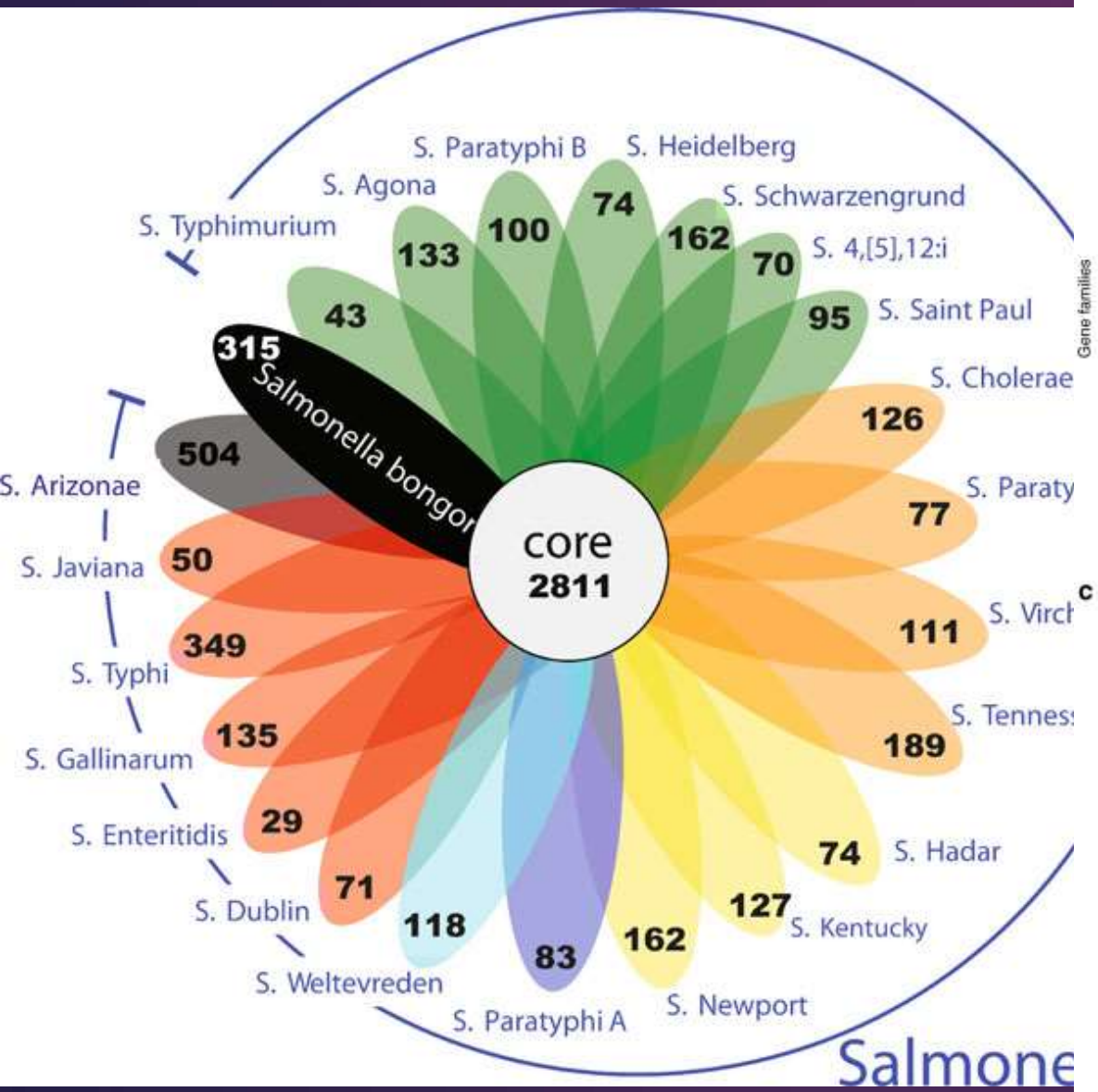
Ancestral MTBΔTbD1→ Modern MTB

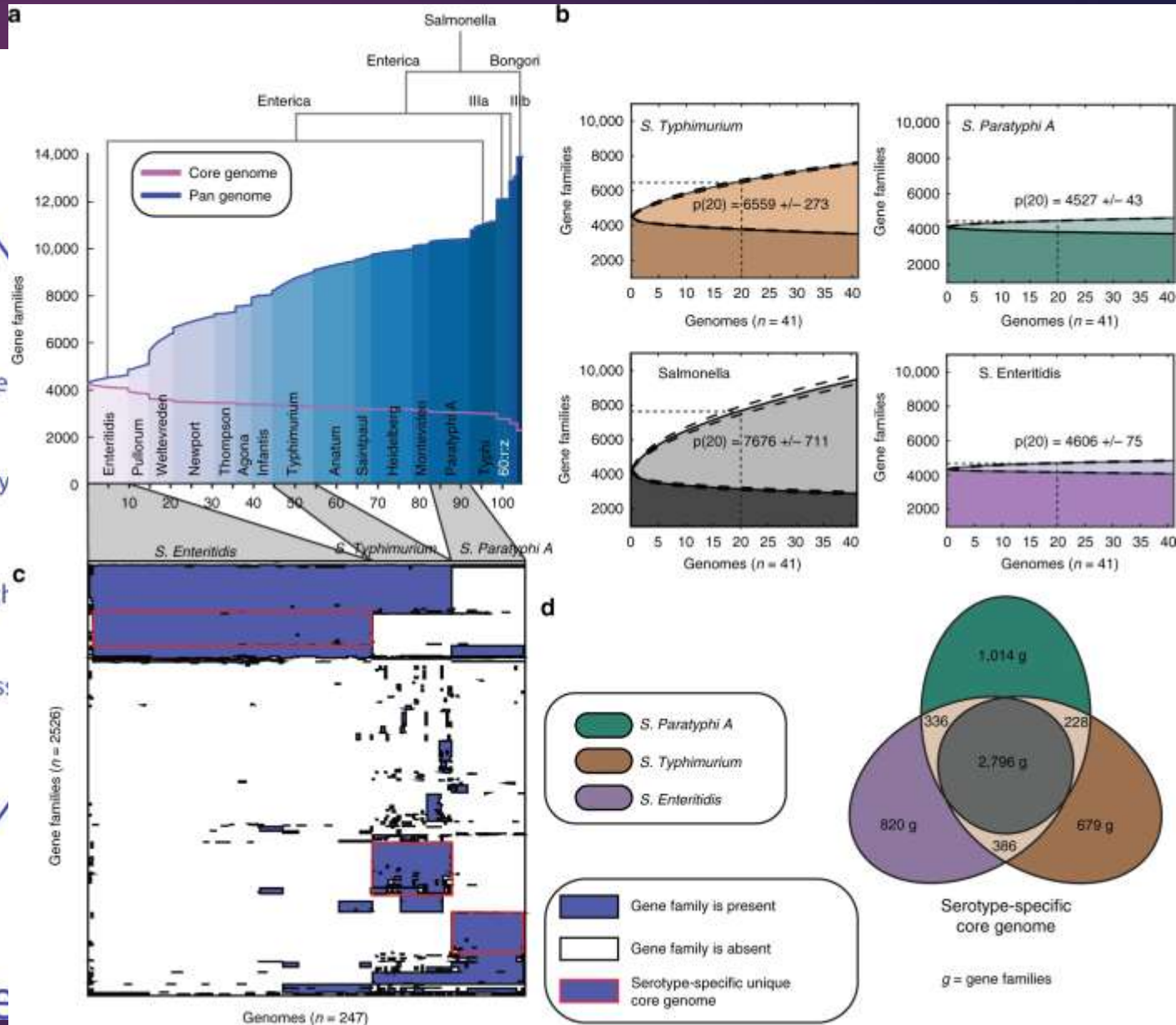
*M. bovis*ΔRD1→*M. bovis* BCG
Original BCGΔRD2→ BCG varia



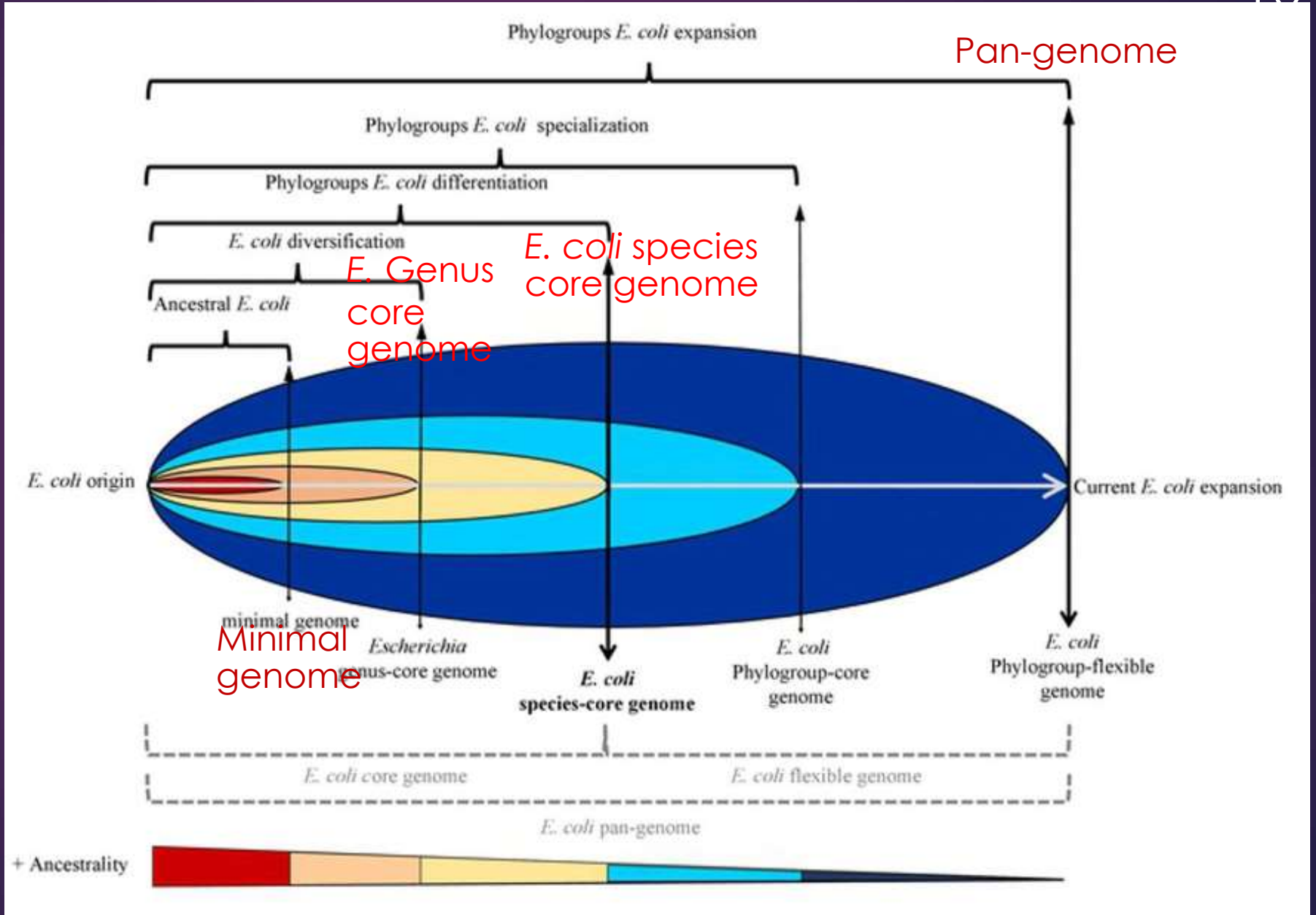


Salmonella Accessory genes/Pan-genome

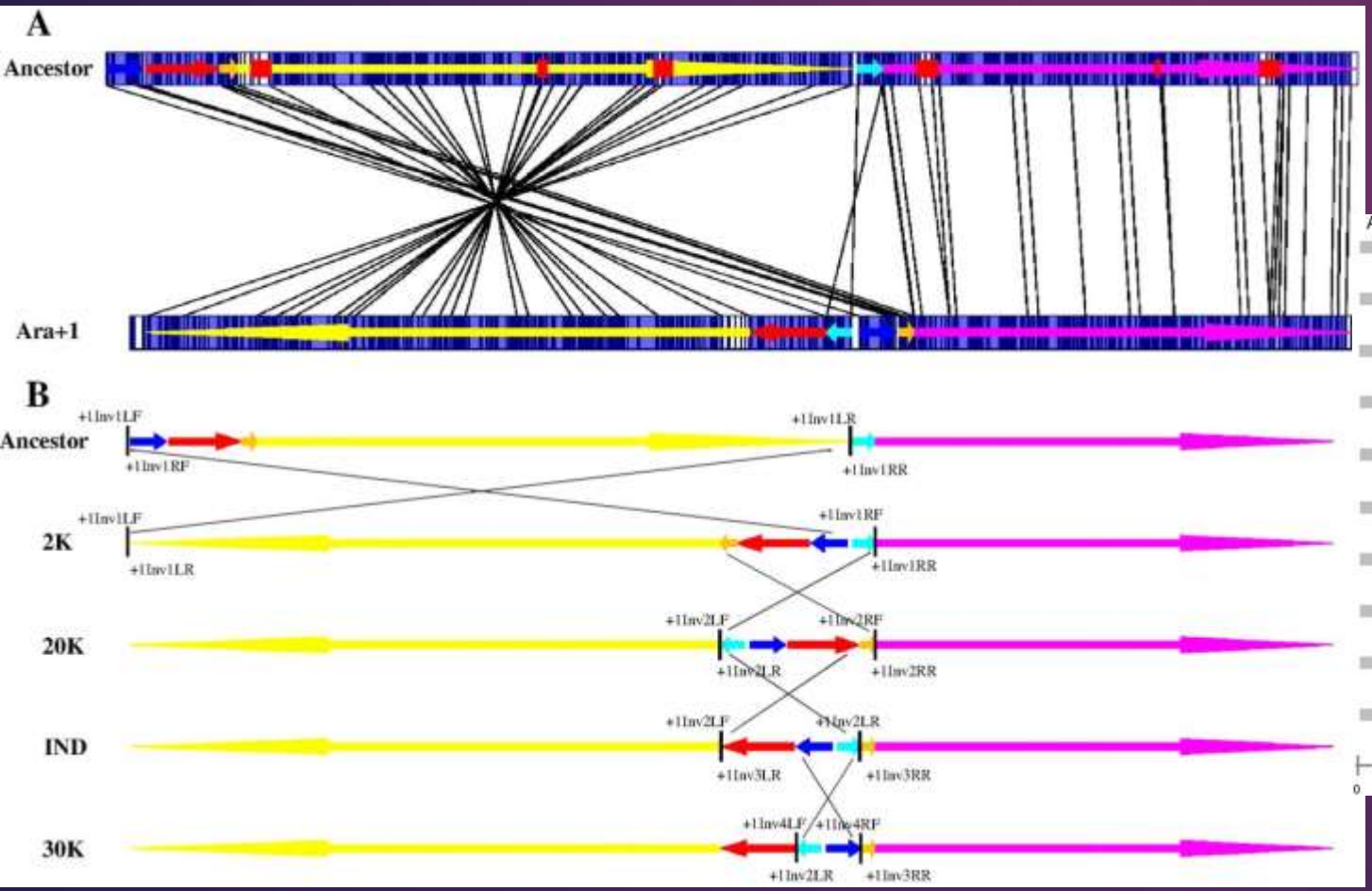


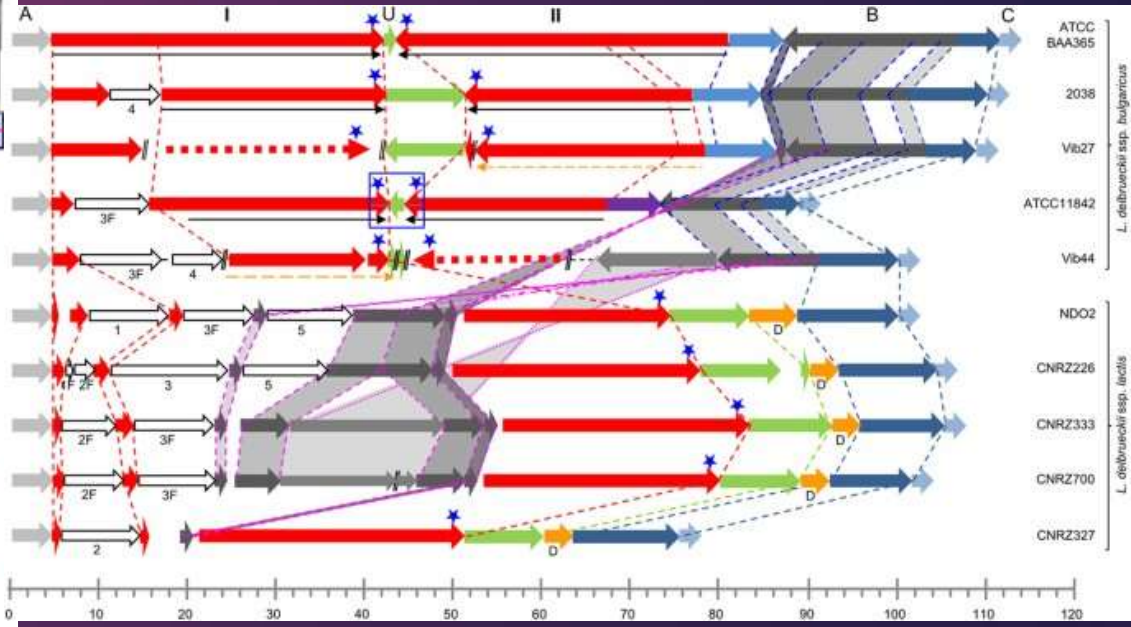


Gene sets in Pan-genome



Chromosomal rearrangement





Calling SNVs for Haploid Organisms

- SNV is always relative, depending on the reference genome.
- Reliability of calling depends on the depth of sequencing.
  - Regions with not enough depth may result in "NO VARIANT" calls.
- If the depth is enough, will all SNVs be identified?
  - Depending on % genome coverage. Uncovered genomic regions may be treated as "NO VARIANT".
- Should all SNVs be used?
  - Depending on purposes. For making phylogenetic trees, it is generally safer to confine only to the core genome and exclude
    - mobile genetic elements, prophages, etc.
    - families of similar genes (probable false SNV assignment)
    - drug resistance genes (probable convergence mutations).
  - Making phylogenetic trees based on WG SNVs may be fine for closely related samples.

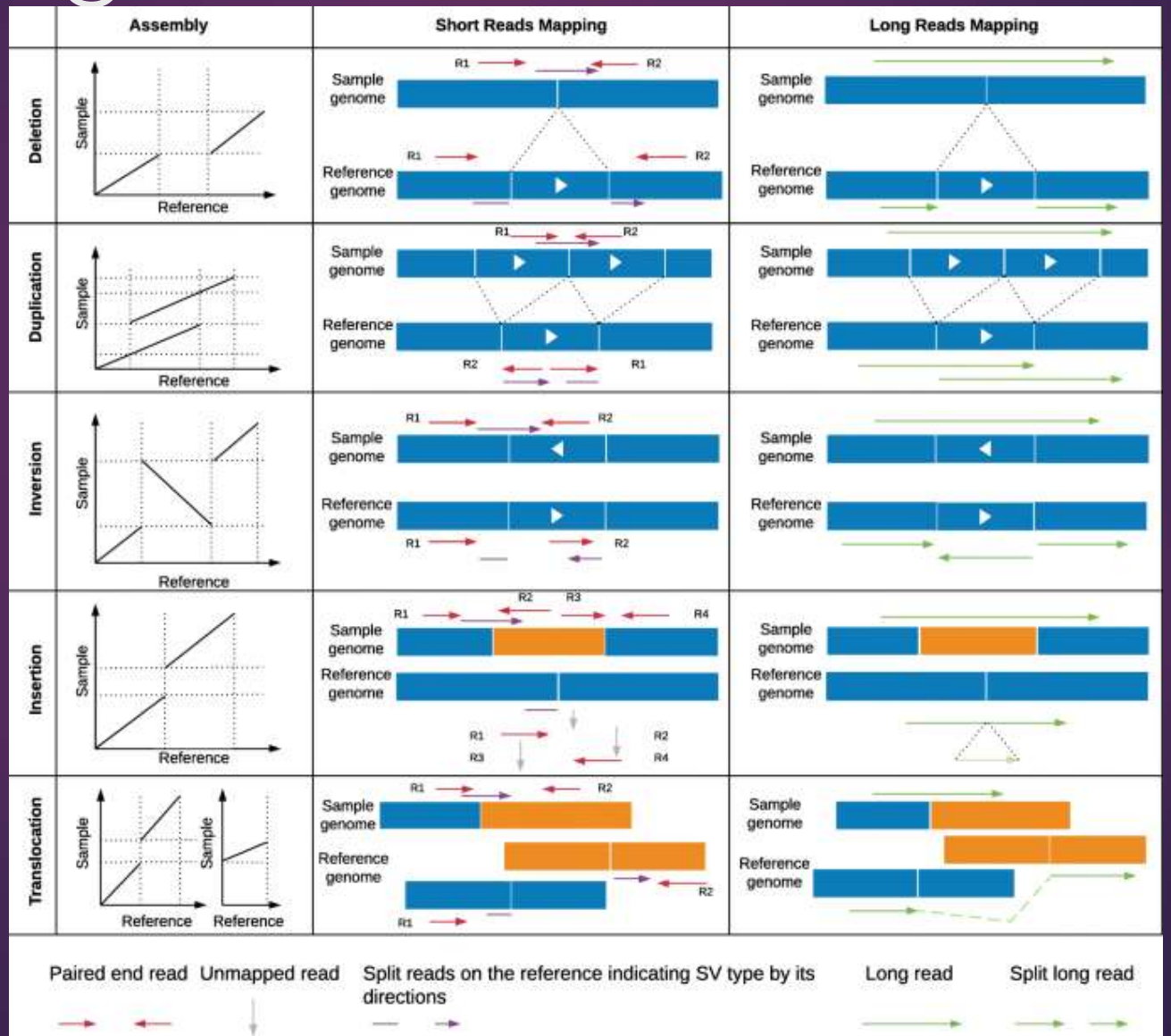
Selecting The Reference Genomes

- Use the **commonly used ones**. So that results may be easier to compare to other works.
- Use the ones **closely related to your samples**. Genomic structural differences between the reference and the samples affect the results of SNV calling.
- Use the ones **containing genes of interest**. The SNVs in DNA regions not present in the reference genome cannot be called.
- DNA regions present in the reference but not the samples may result in "NO VARIANT" calls.

The Meaning of SNVs

- Normally SNVs are reported if identified.
- NO VARIANTS may mean
  - Definitely no mutations
  - Not covered by sequencing
  - Numbers of reads too low
  - No genomic segments in samples
  - Variants not present in the reference genomes cannot be called.
- Typically in haploid organisms, only the major variant will be reported. There are situations that variants present in minor populations are also interedted, such as drug resistance mutations.
  - In MTB, If >1% of population are phenotypic resistant, the strains are considered resistant. Mutations of 1% cannot be detected by normal WGS analysis.

Problems of sequence read mapping and SNV calling due to structural variants.



Conceptual definition of species and subspecies: How do we define a genotype?

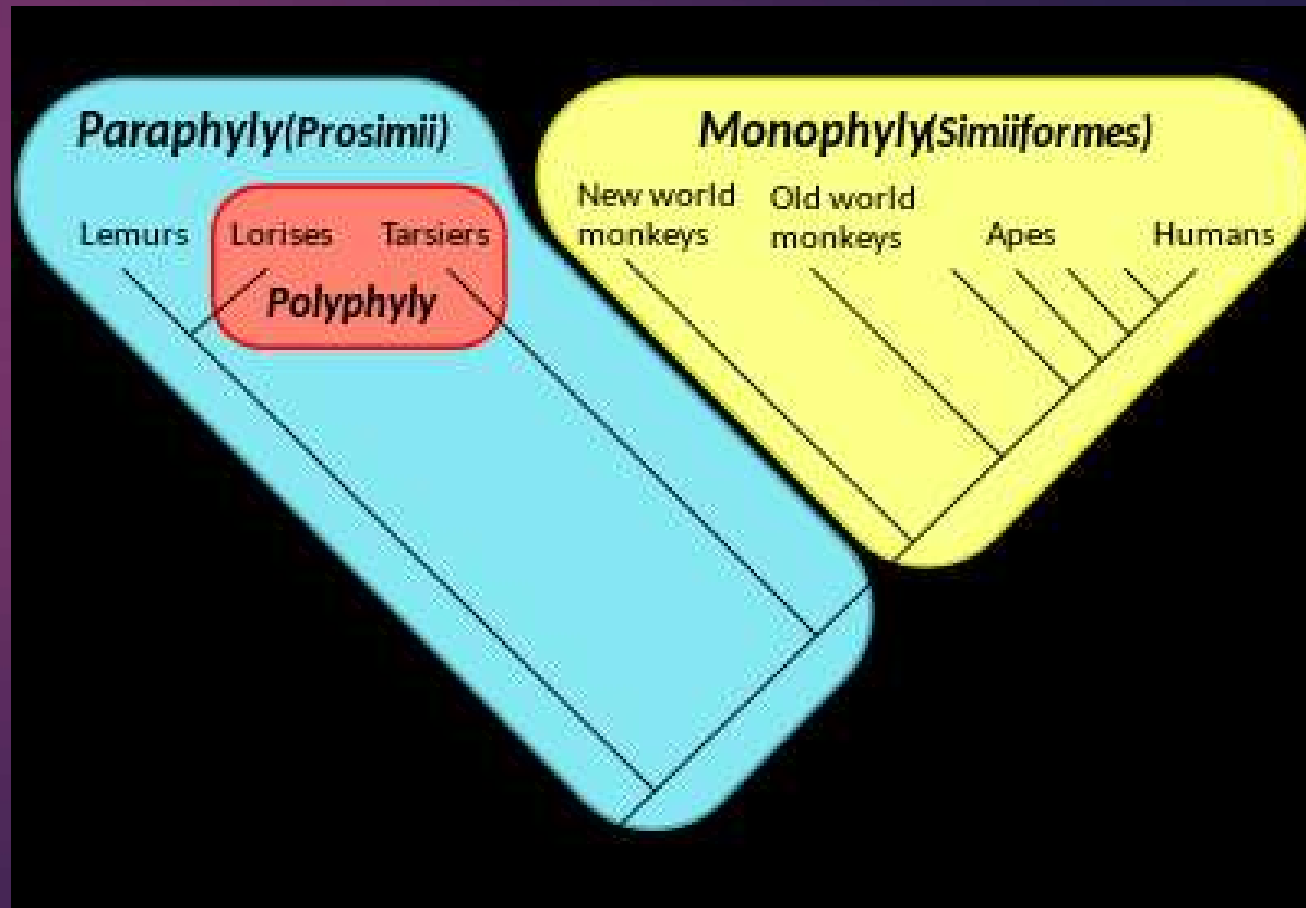
- Clustered characteristics (biochemical, immunological, anatomical, genetic (SNVs)) with clear separation from other clusters.
- Cohesive properties (e.g. sexual reproduction)
- Ecological homogeneity
  - Ecotypes: species- homogeneous ecological requirement.
  - Ephemeral ecotypes: subspecies: each with slightly different ecological requirement but still sharing and competing in the same ecological environment. One can completely replace another. Ephemeral ecotypes need not to correlate with genotypes.

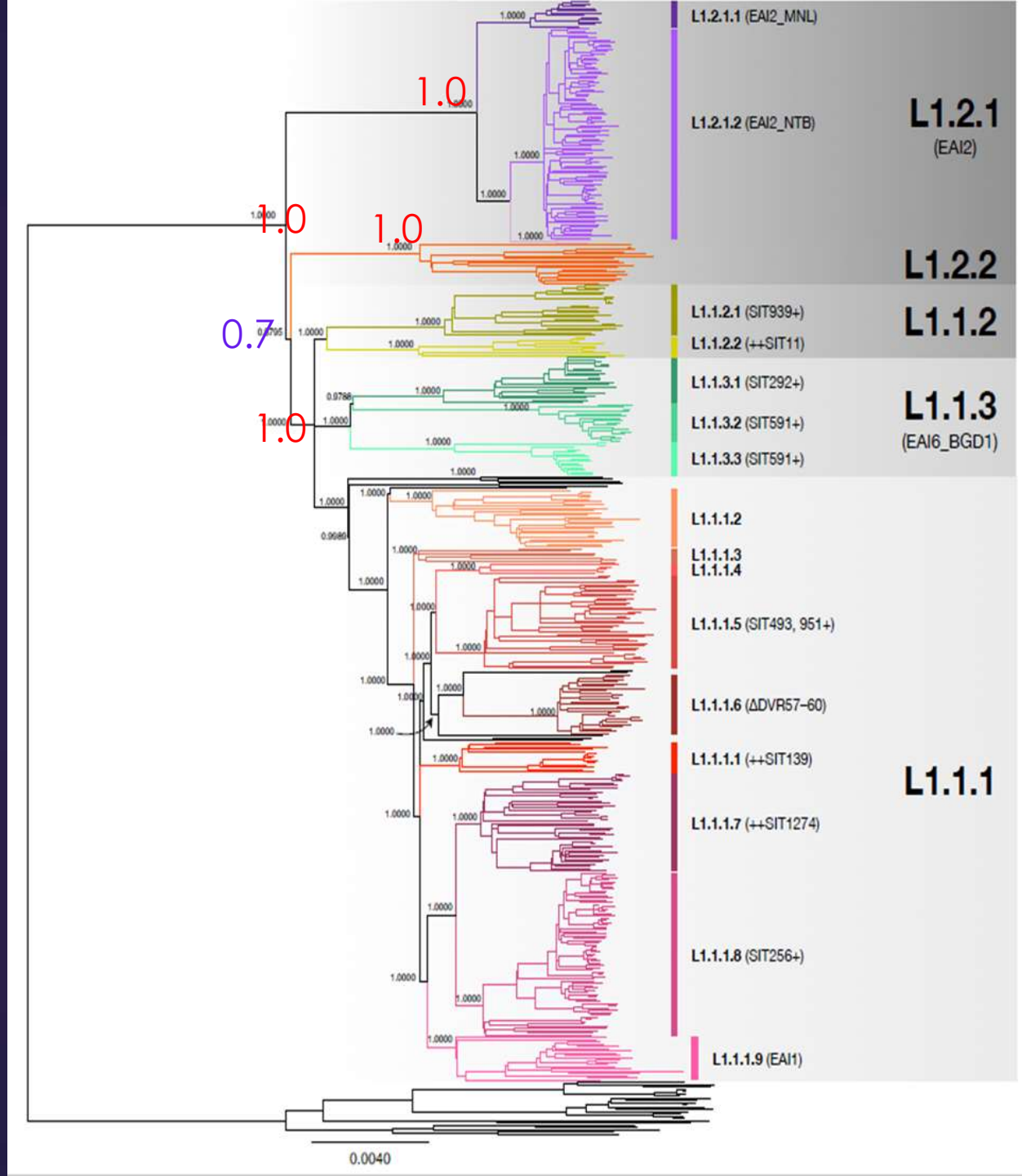
Taxonomic classification by SNV

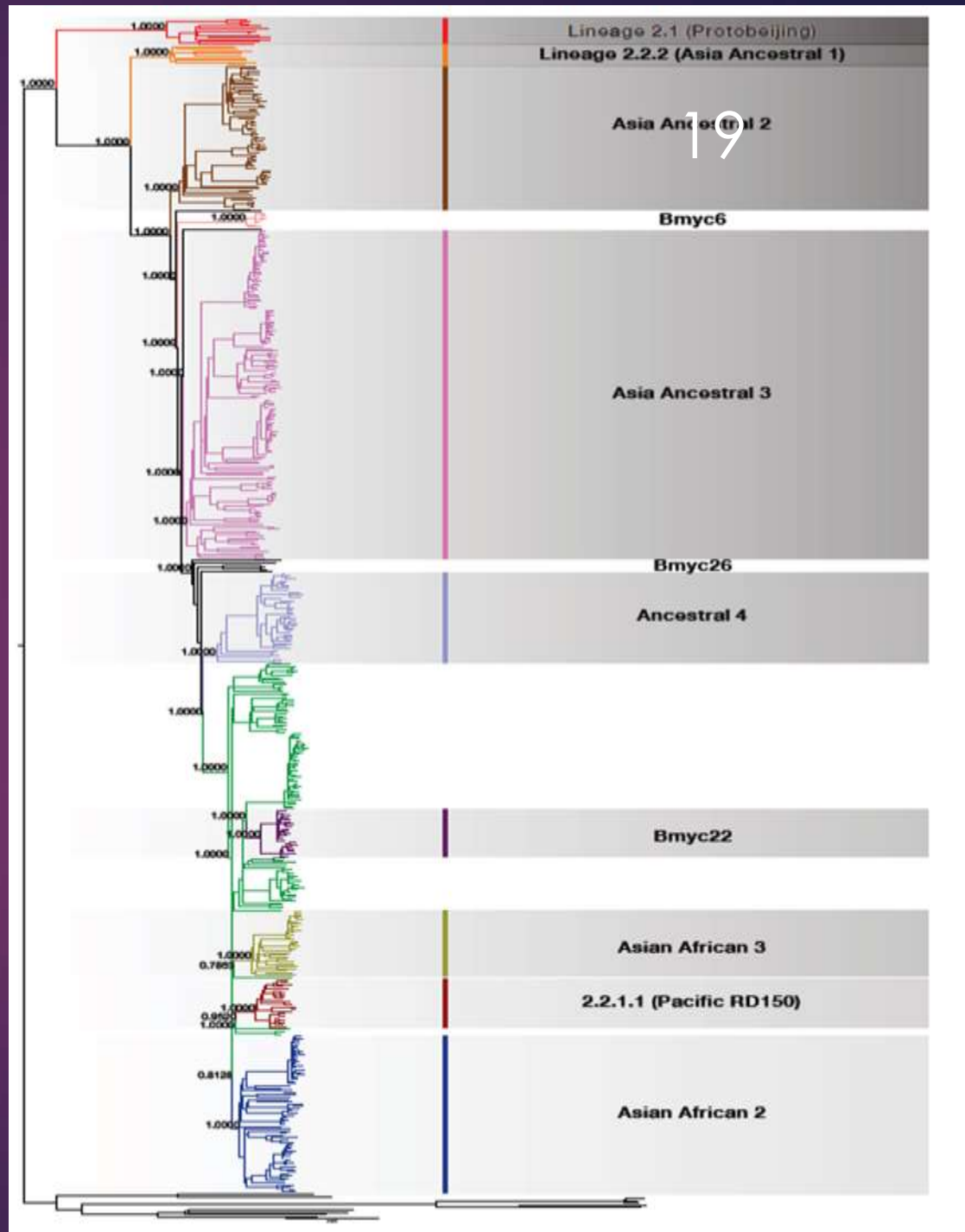
- The main idea is to use SNVs to "cluster" bacterial isolates that shared a common ancestor (best done by using core genome SNVs)
- Because of sharing a common ancestor, they may share (associate with) some phenotypes (host preferences, pathogenesis process, virulence, etc.).
  - As some virulence genes are in pathogenicity islands (accessory genes), some phenotypes may not correlate to the genotypes.
- Each classification method (e.g., serotyping, MLST, VNTR typing, deletion-based typing, etc.) needs **not** to be congruent
- There are usually a small number of samples that are not similar to any of the others (emerging and extinction genotypes).

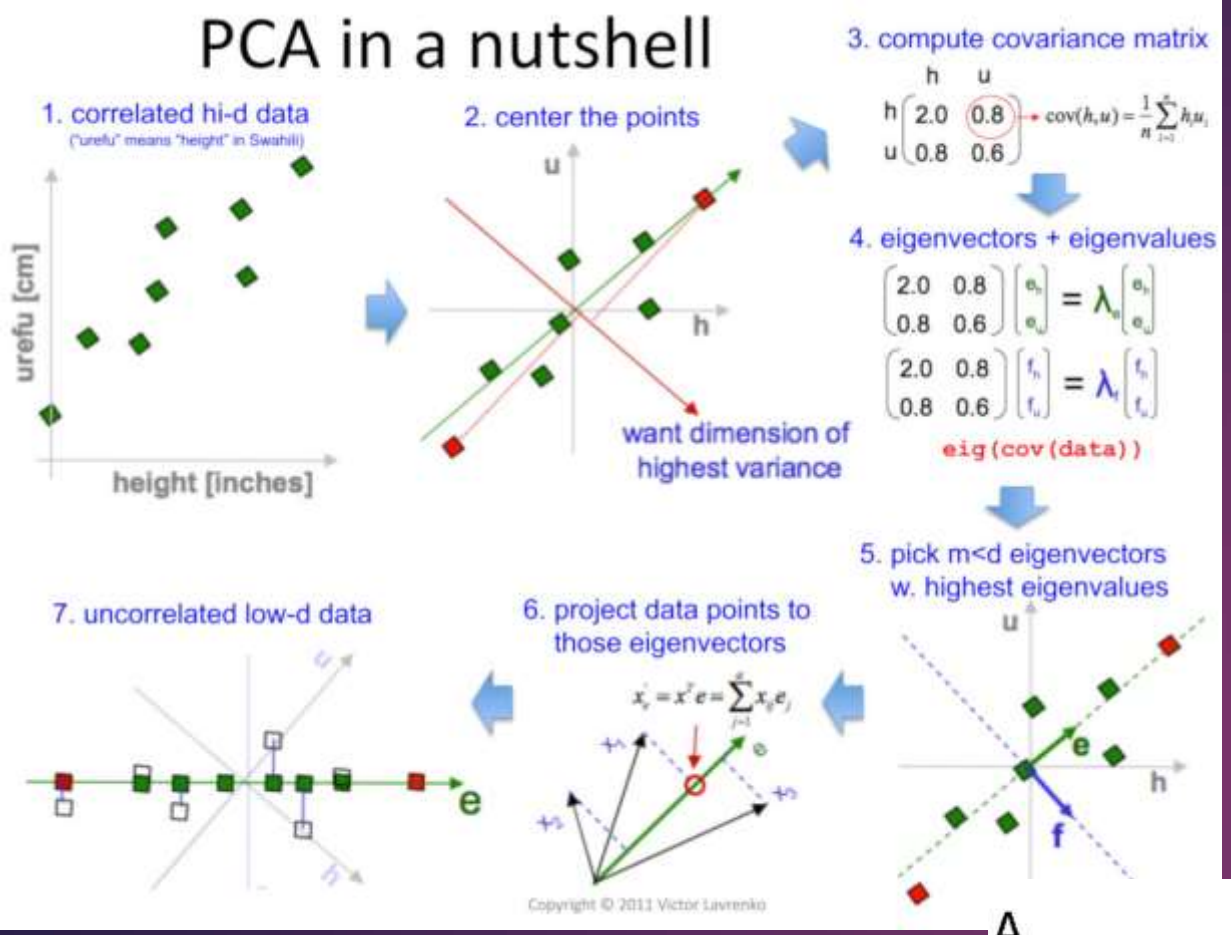
Formulating Genotyping Scheme based on SNVs Requires a few Criteria

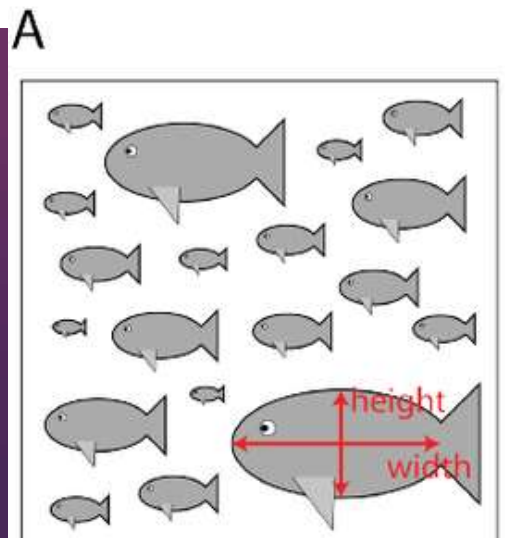
- A monophyletic group (clade) with high confidence (boot strap score)
- Clear separation from other groups.
  - Principal component analysis
  - Fixation index
  - Intragroup as intergroup average pairwise SNV distances.

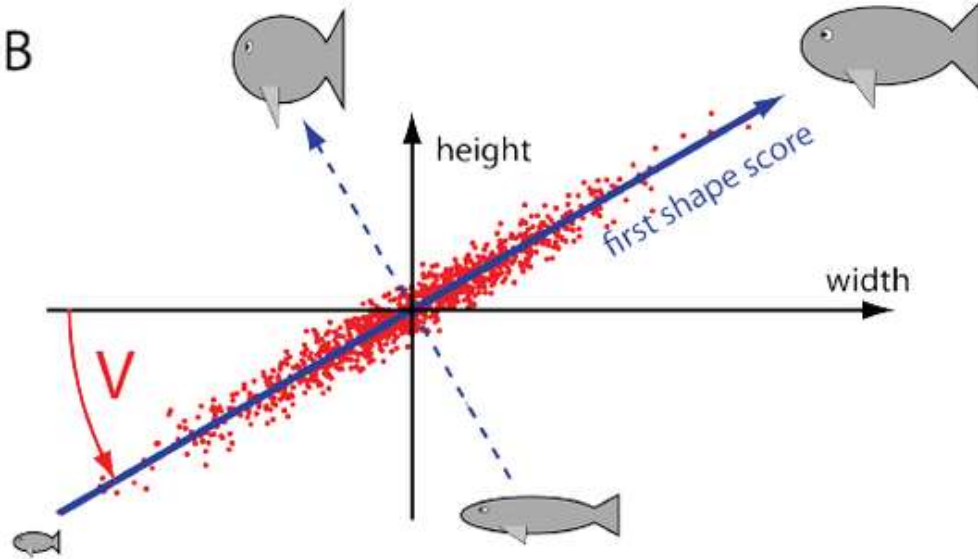



L1.2.1.1 (EAI2_MNL)
L1.2.1.2 (EAI2_NTB)
L1.2.1
(EAI2)
L1.2.2
L1.1.2.1 (SIT939+)
L1.1.2.2 (++SIT11)
L1.1.2
L1.1.3.1 (SIT292+)
L1.1.3.2 (SIT591+)
L1.1.3.3 (SIT591+)
L1.1.3
(EAI6_BGD1)
L1.1.1.2
L1.1.1.3
L1.1.1.4
L1.1.1.5 (SIT493, 951+)
L1.1.1.6 (ΔDVR57–60)
L1.1.1.1 (++SIT139)
L1.1.1.7 (++SIT1274)
L1.1.1.8 (SIT256+)
L1.1.1.9 (EAI1)
L1.1.1
1.0
1.0
1.0
0.7
1.0
0.0040

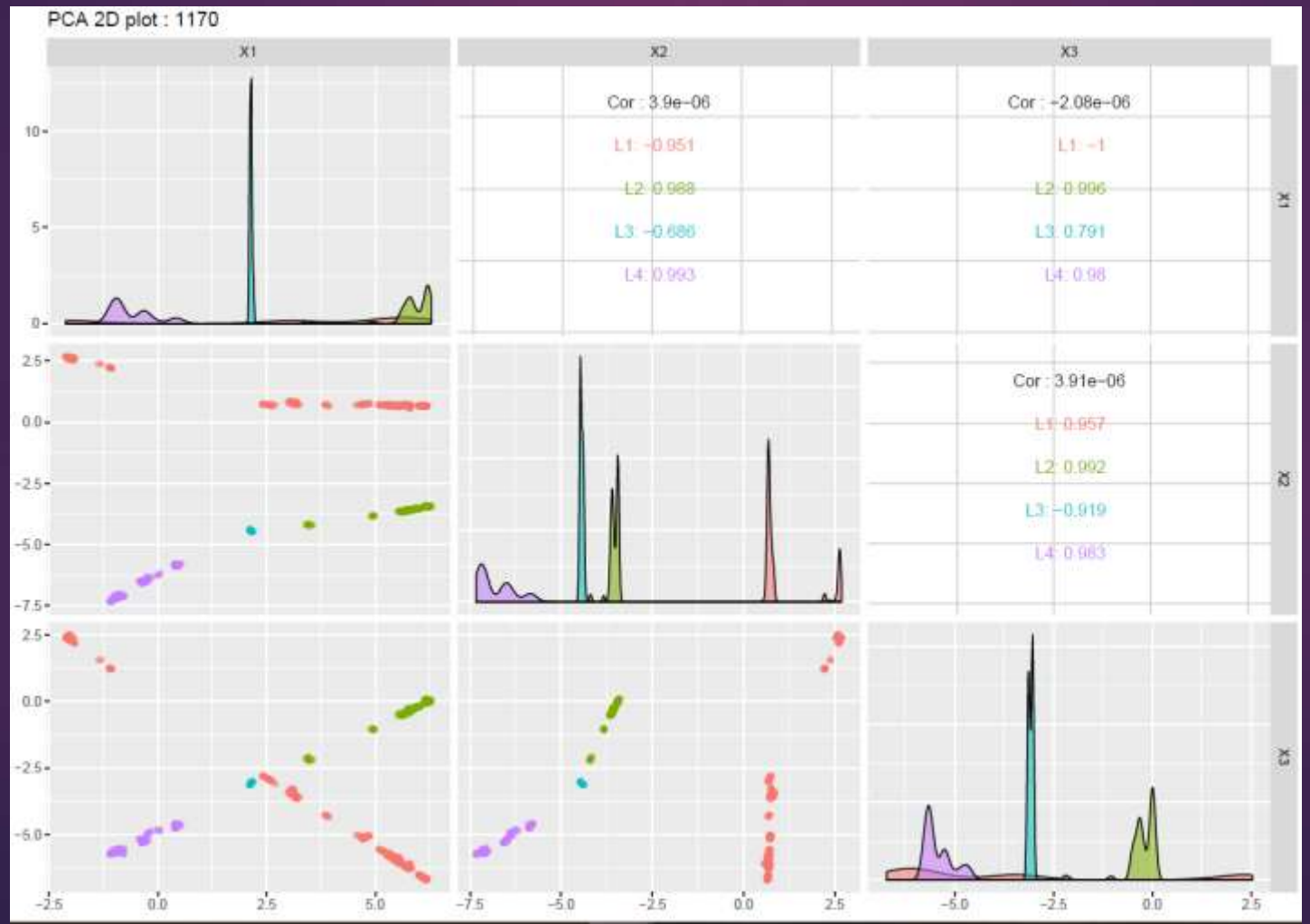

Lineage 2.1 (Protobeijing)
Lineage 2.2.2 (Asia Ancestral 1)
Asia Ancestral 2
Bmyc6
Asia Ancestral 3
Bmyc26
Ancestral 4
Bmyc22
Asian African 3
2.2.1.1 (Pacific RD150)
Asian African 2

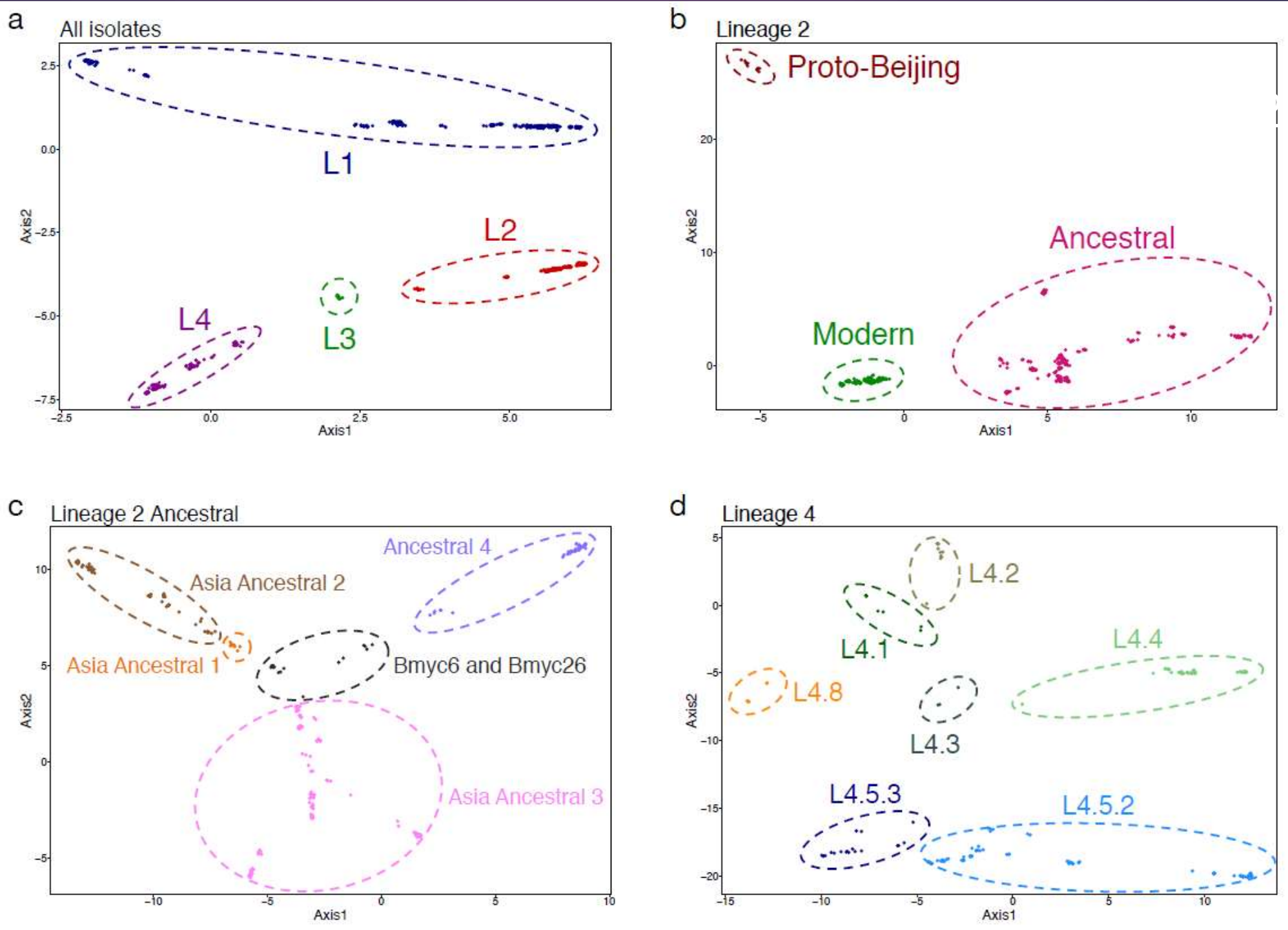

PCA in a nutshell
1. correlated hi-d data
("urefu" means "height" in Swahili)
urefu [cm]
height [inches]
2. center the points
u
h
want dimension of highest variance
3. compute covariance matrix
h u
h 2.0 0.8
u 0.8 0.6
4. eigenvectors + eigenvalues
eig(cov(data))
5. pick m<d eigenvectors w. highest eigenvalues
e
f
6. project data points to those eigenvectors
7. uncorrelated low-d data
Copyright © 2011 Victor Lavrenko

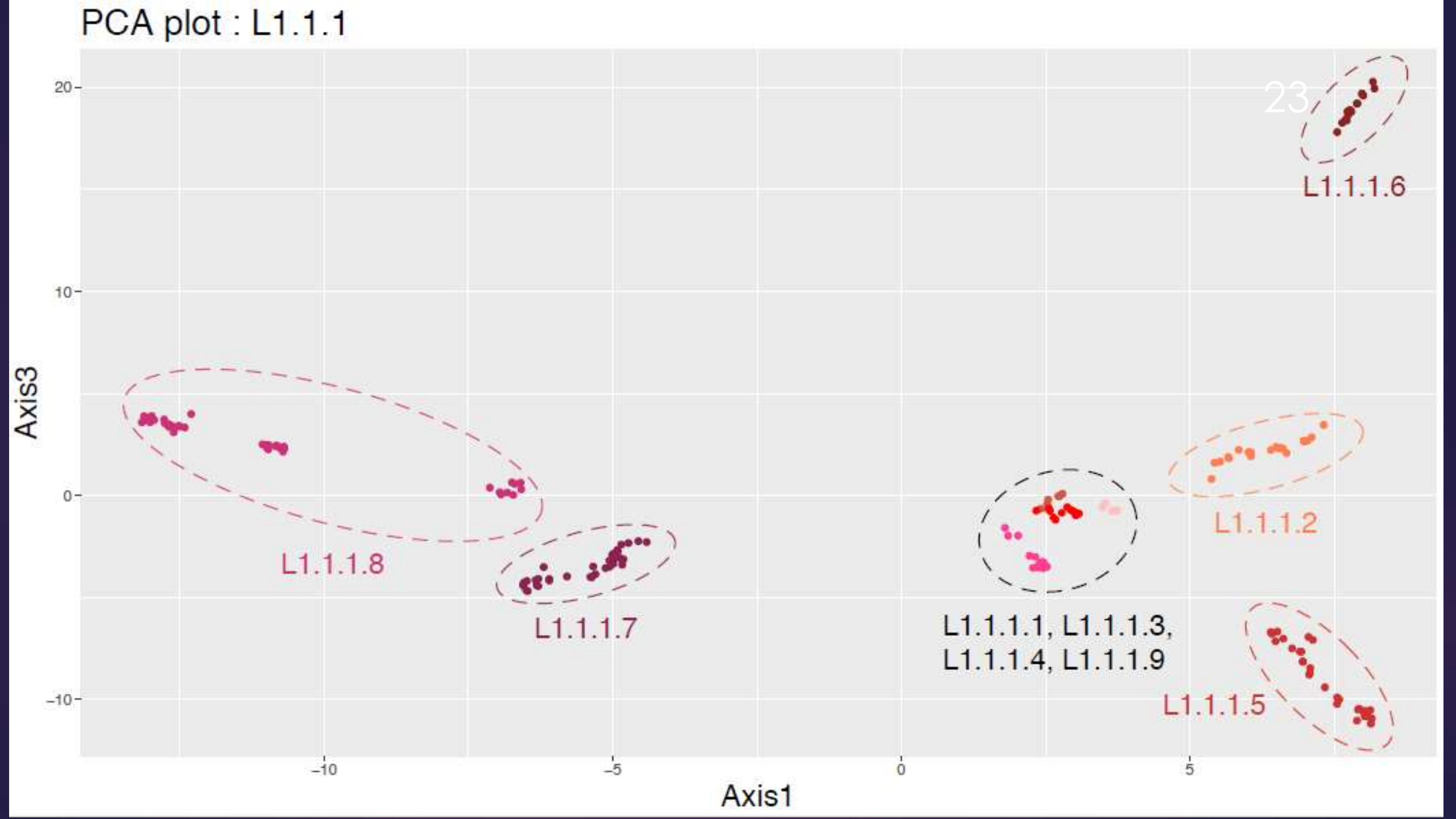

A
height
width


B
height
first shape score
width
V

In our set of data, each sample had about 1,000 SNVs (compared to 7-50 SNPs for MLST). The sample relationships are summarized in a 3D PCA plot.




a
All isolates
L1
L2
L3
L4
Axis1
Axis2
b
Lineage 2
Proto-Beijing
Ancestral
Modern
Axis1
Axis2
c
Lineage 2 Ancestral
Ancestral 4
Asia Ancestral 2
Asia Ancestral 1
Bmyc6 and Bmyc26
Asia Ancestral 3
Axis1
Axis2
d
Lineage 4
L4.2
L4.1
L4.4
L4.8
L4.3
L4.5.3
L4.5.2
Axis1
Axis2


PCA plot : L1.1.1
20
10
0
−10
Axis3
−10
−5
0
5
Axis1
L1.1.1.6
L1.1.1.8
L1.1.1.7
L1.1.1.1, L1.1.1.3,
L1.1.1.4, L1.1.1.9
L1.1.1.2
L1.1.1.5

Average Intra/inter-sublineage pairwise SNV distances

	1.1	1.1.1	1.1.1.1	1.1.1.2	1.1.1.3	1.1.1.4	1.1.1.5	1.1.1.6	1.1.1.7	1.1.1.8	1.1.1.9	1.1.2	1.1.2.1	1.1.2.2	1.1.3	1.1.3.1	1.1.3.2	1.1.3.3	1.2.1	1.2.1.1	1.2.2.2	1.2.1.3	1.2.2
1.1	570.0																						
1.1.1		451.1																					
1.1.1.1			252.9																				
1.1.1.2			530.0	340.5																			
1.1.1.3			473.0	549.5	431.8																		
1.1.1.4			450.5	547.7	491.0	224.2																	
1.1.1.5			478.9	575.6	517.8	464.8	323.9																
1.1.1.6			464.5	560.3	502.3	456.9	478.5	135.3															
1.1.1.7			444.5	542.4	485.0	463.7	490.6	475.5	269.8														
1.1.1.8			446.4	544.0	487.2	465.5	493.5	477.0	372.2	174.4													
1.1.1.9			459.5	557.6	500.1	479.2	505.1	492.3	457.1	459.2	284.6												
1.1.2		749.9										487.5											
1.1.2.1													322.0										
1.1.2.2													720.5	310.0									
1.1.3		749.6										755.0			502.1								
1.1.3.1																228.6							
1.1.3.2																671.3	131.4						
1.1.3.3																635.7	672.9	138.4					
1.2.1	800.8																		145.3				
1.2.1.1																				120.3			
1.2.1.2																				286.5	106.4		
1.2.1.3																				296.8	189.7	76.00	
1.2.2	853.6																		846.1				430.2

Fixation Index

- Definition:

$$F_{ST} = \frac{\sigma_S^2}{\sigma_T^2} = \frac{\sigma_S^2}{\bar{p}(1-\bar{p})}$$

Alternatively,[2]

$$F_{ST} = \frac{f_0 - \bar{f}}{1 - \bar{f}}$$

- σ_S^2 = Variance of samples
- σ_T^2 = Variance of total populations
- Fst can have values between 0 and 1. The higher means the more separation of the subgroup.
- **Estimation from WGS data**

$$F_{ST} = \frac{\pi_{\text{Between}} - \pi_{\text{Within}}}{\pi_{\text{Between}}}$$

- $\pi_{\text{within group}}$ is the average pairwise SNV distance within a group
- $\pi_{\text{between group}}$ is the average pairwise SNV distance between all members in the group and all members not in the same group but in the same level of grouping.
- The statistical tests for the differences between $\pi_{\text{within group}}$ and $\pi_{\text{between group}}$ were done by Wilcoxon rank-sum test at $a=0.05$ with Bonferroni correction. *Significance usually require an adequate number of the samples.*

Formulating Genotyping Scheme based on SNVs Usually Requires a few Criteria

- A monophyletic group (clade) with high confidence (boot strap score)
- Clear separation from other groups.
  - Principal component analysis
  - Fixation index
  - Intragroup as intergroup average pairwise SNV distances.
- Supported by independent criteria (e.g., specific deletions or other genetic markers)

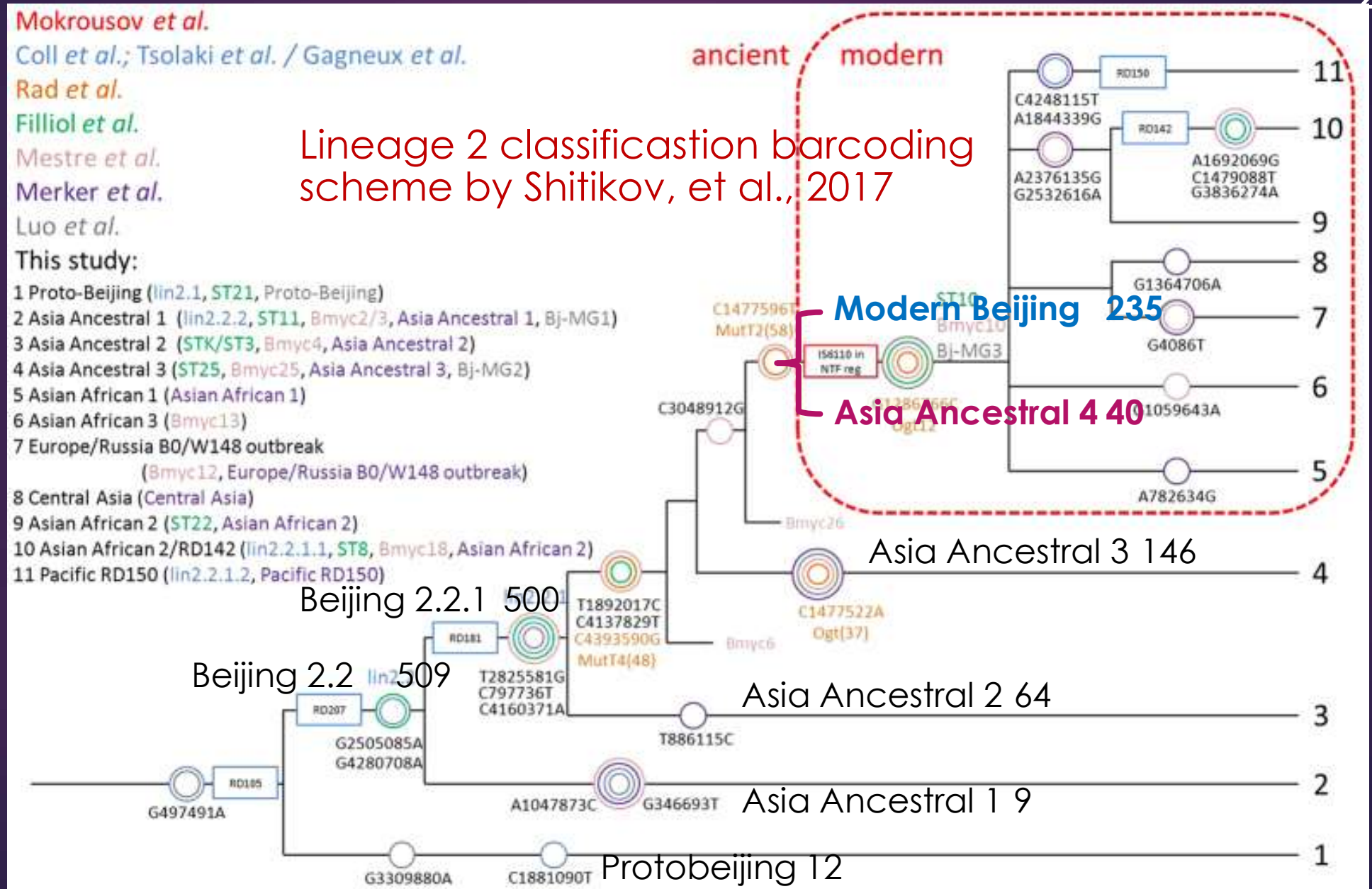
Do we need to do all these exercise every WGS project?

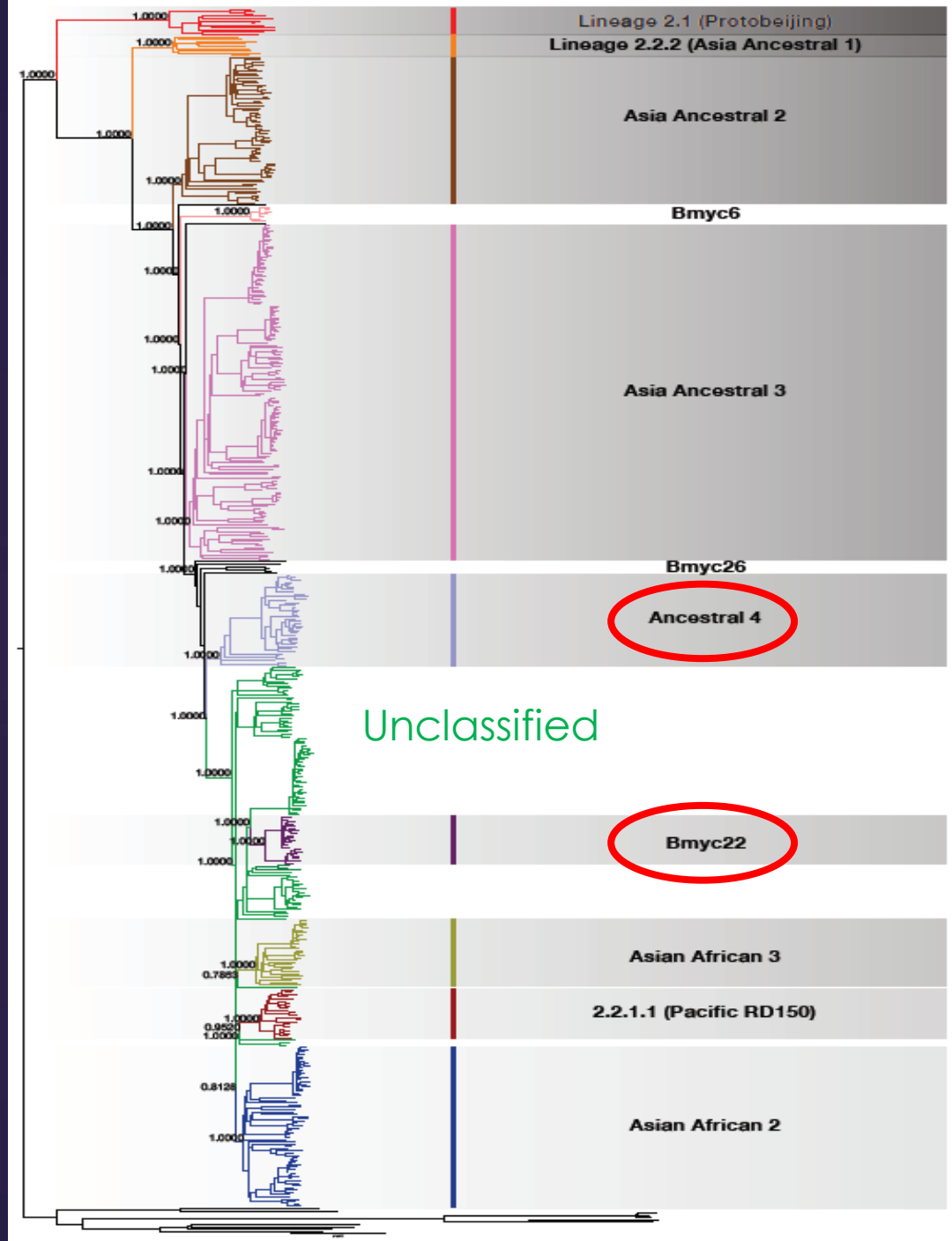
- If one add substantial numbers of new samples, especially from new locations, it may be a good idea to do so.
- If the number of samples were small or come from the same location (no new genotype is expected), it may be more convenient to match to lists of (sub)lineage-specific (bar-coding) SNPs.
- However, lists of LS-SNPs will be continuously improving, indicating the needs to re-genotype the samples periodically.

Summary: Identifying Genotypes of Samples with new WGS data.

- Constructing a phylogenetic tree. If the number of samples is too small, related samples may be added to make a tree.
- The identity of each clade is identified using lists of genotype-specific SNPs or bar-coding SNPs
- OR adding reference sequences of known genotypes into the tree. All the isolates that form the same clade as the reference should be the same genotype.
- **A clade that cannot be matched to any genotypes may be a new genotypes.**
- If only a few samples have to be identified- may use the lists of genotype specific SNPs or barcoding SNPs. This has the risks of missing new genotypic variants.

Examples of Barcoding SNPs and Pitfalls





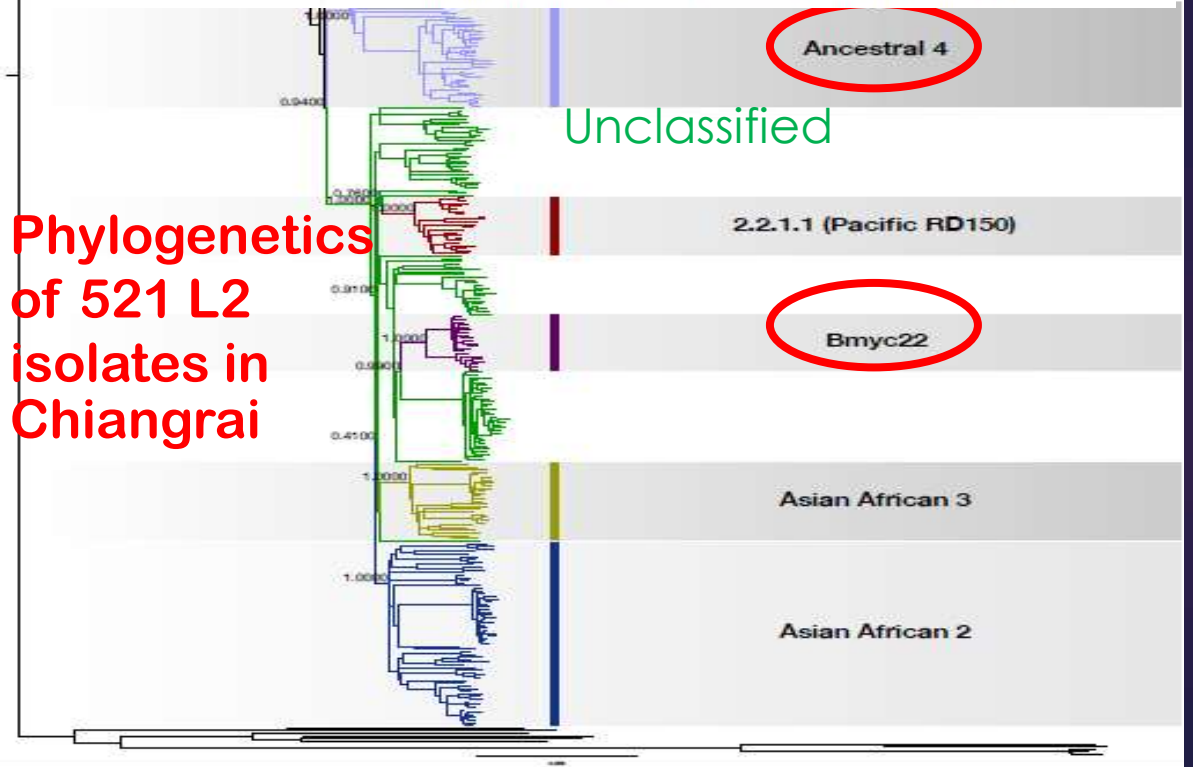
SCIENTIFIC REPORTS
nature research

OPEN

A novel Ancestral Beijing sublineage of *Mycobacterium tuberculosis* suggests the transition site to Modern Beijing sublineages

Received: 19 March 2019
Accepted: 27 August 2019
Published online: 23 September 2019

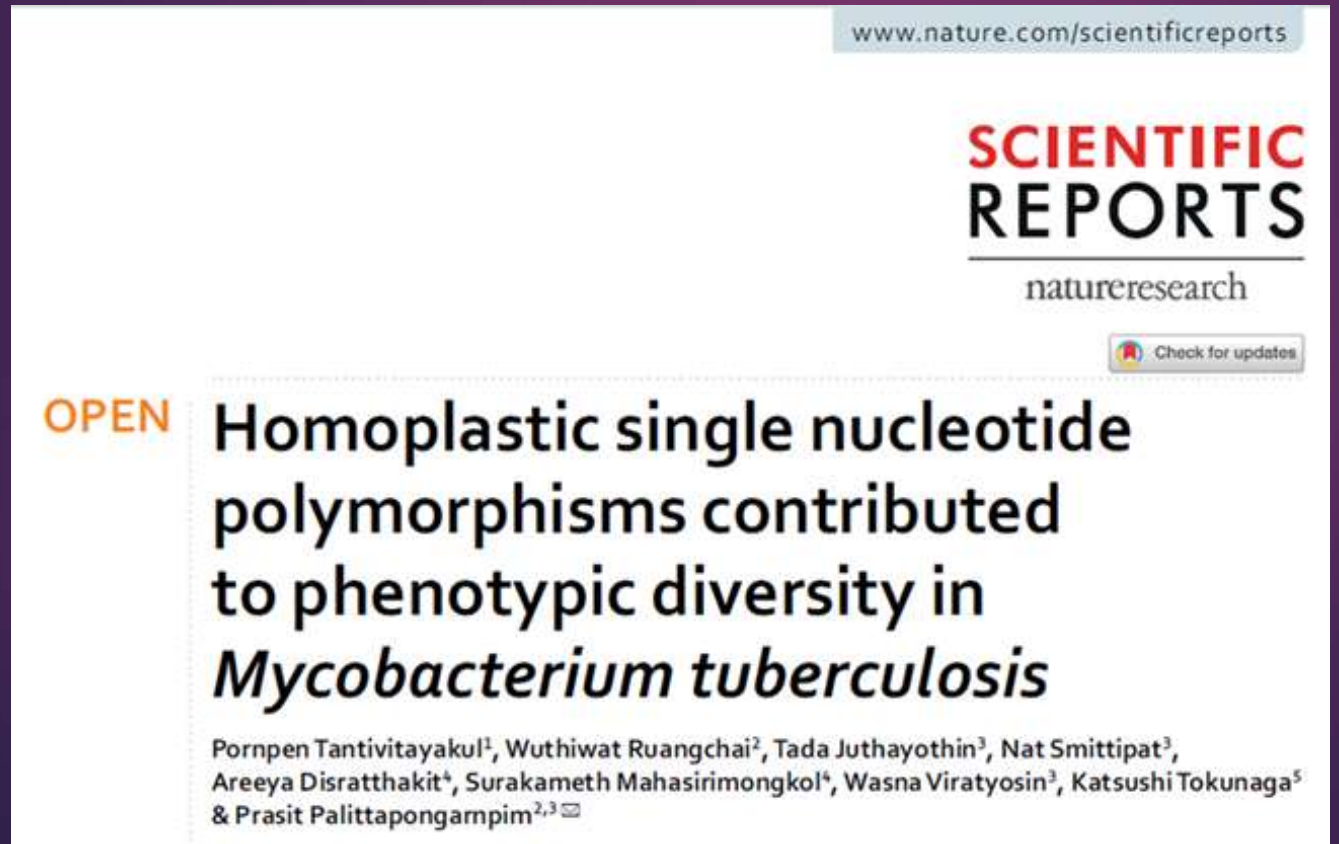
Pravech Ajawatanawong[1], Hideki Yanai[2], Nat Smittipat[3], Areeya Disratthakit[4], Norio Yamada[5], Reiko Miyahara[6], Supalert Nedsuwan[7], Worarat Imasanguan[7], Pacharee Kantipong[7], Boonchai Chaiyasirinroje[8], Jiraporn Wongyai[8], Supada Plitphonganphim[9], Pornpen Tantivitayakul[10], Jody Phelan[11], Julian Parkhill[12], Taane G. Clark[11], Martin L. Hibberd[11], Wuthiwat Ruangchai[1], Panawun Palittapongarnpim[1], Tada Juthayothin[3], Yuttapong Thawornwattana[1], Wasna Viratyosin[3], Sissades Tongsima[3], Surakameth Mahasirimongkol[4], Katsushi Tokunaga[6] & Prasit Palittapongarnpim[1,3]



Phylogenetics of 521 L2 isolates in Chiangrai

What do we do with the genotypes?

- Association with phenotypes
- Identification of homoplastic SNPs, indicating possible convergence evolution: drug resistance, T cell epitopes.



www.nature.com/scientificreports

SCIENTIFIC REPORTS

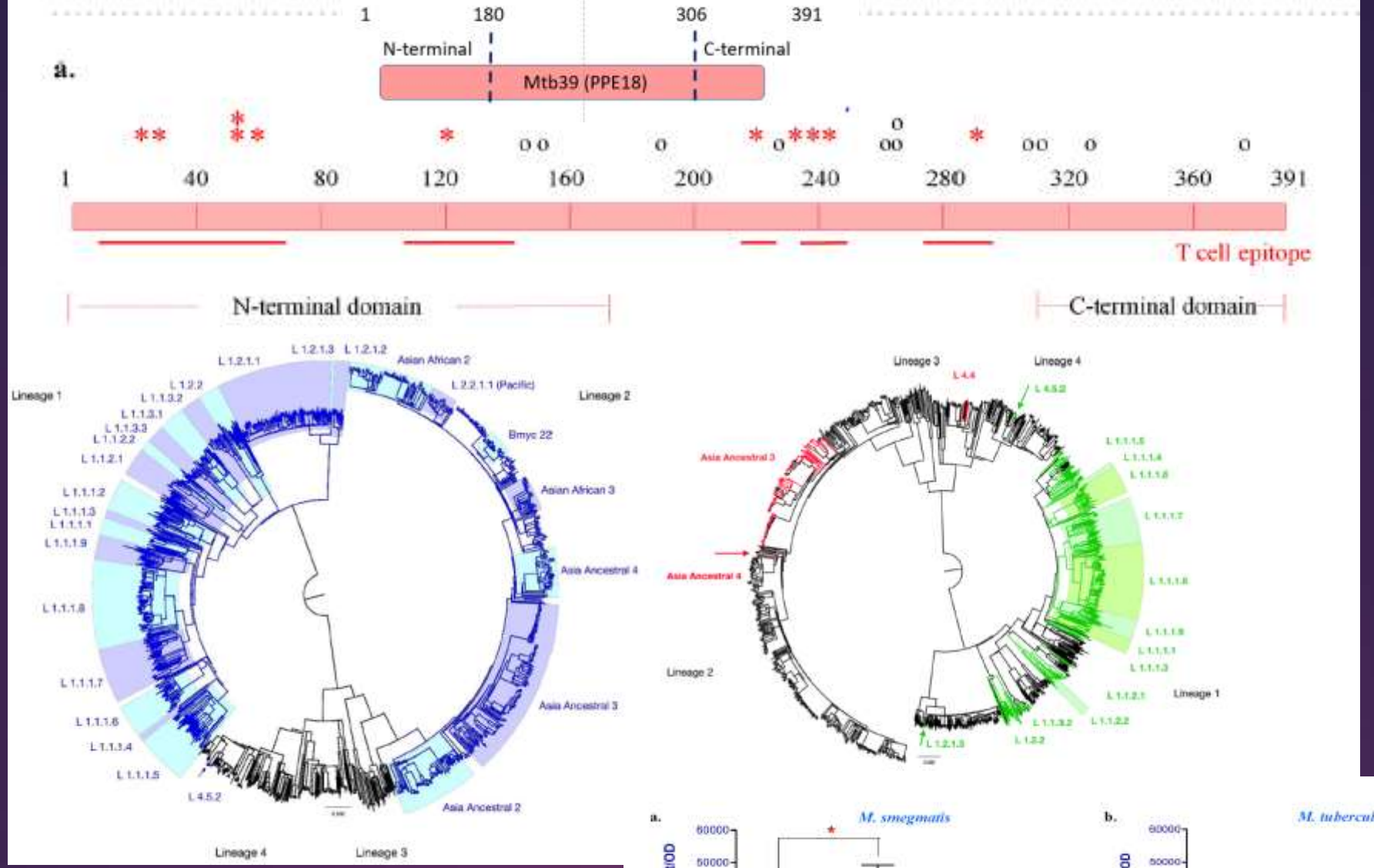
natureresearch

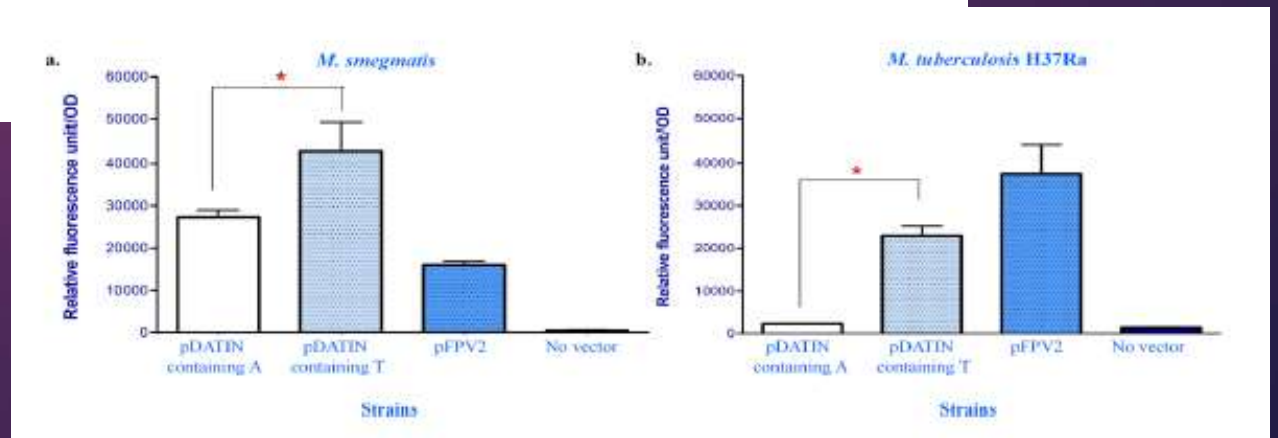
Check for updates

OPEN

Homoplastic single nucleotide polymorphisms contributed to phenotypic diversity in *Mycobacterium tuberculosis*

Pornpen Tantivitayakul[1], Wuthiwat Ruangchai[2], Tada Juthayothin[3], Nat Smittipat[3], Areeya Disratthakit[4], Surakameth Mahasirimongkol[4], Wasna Viratyosin[3], Katsushi Tokunaga[5] & Prasit Palittapongarnpim[2,3]

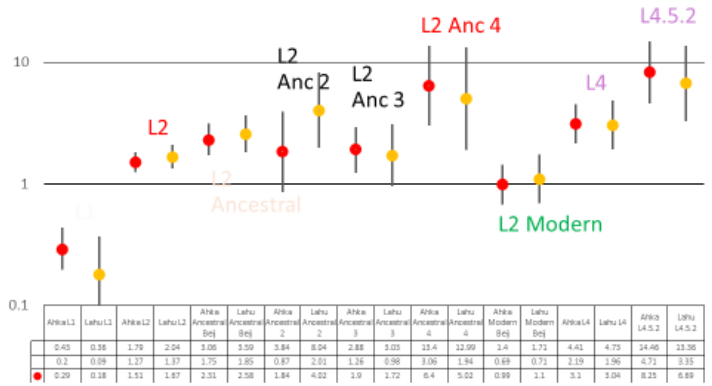

1
180
306
391
N-terminal
C-terminal
a.
Mtb39 (PPE18)
1
40
80
120
160
200
240
280
320
360
391
T cell epitope
N-terminal domain
C-terminal domain
Lineage 1
Lineage 2
Lineage 3
Lineage 4
Asian African 2
Asian African 3
Asia Ancestral 2
Asia Ancestral 3
Asia Ancestral 4
Bmyc 22
L 2.2.1.1 (Pacific)


a.
M. smegmatis
b.
M. tuberculosis H37Ra
Relative fluorescence unit/OD
pDATIN containing A
pDATIN containing T
pFPV2
No vector
Strains

Association with phenotypes

- Common associations: Age, Ethnicity, Geography
- Rationale: An established genotype should result from phenotypic fitness. The genes accountable for the fitness may be among the LS-SNPs

Risk Ratios of infections by selected lineages and sublineages of Akha and Lahu populations compared to Thais.

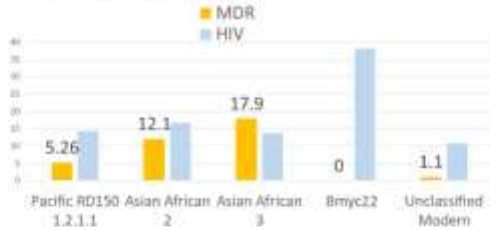


INT J TUBERC LUNG DIS 23(9):000–000
© 2019 The Union
http://dx.doi.org/10.5588/ijtld.18.0710

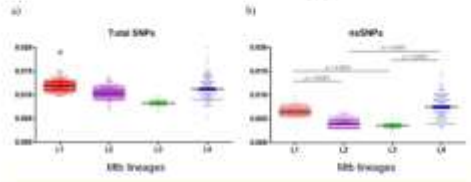
Indo-Oceanic *Mycobacterium tuberculosis* strains from Thailand associated with higher mortality

N. Smittipat,* R. Miyahara,† T. Juthayothin,* P. Billamas,* K. Dokladda,* W. Imsanguan,‡ D. Intralawan,‡ K. Rukseree,§ S. Jaitrong,* B. Chaiyasirinroje,¶ J. Wongjai,¶ A. Disratthakit,# A. Chaiprasert,** S. Nedsuwan,‡ S. Mahasirimongkol,# L. Toyo-oka,† K. Tokunaga,† N. Yamada,†† P. Palittapongarnpim,*‡‡ H. Yanai‡‡§§

% MDR and HIV infections among various sublineages of Modern Beijing strains. Therefore, association of Beijing strains to phenotypes varied by studies due to different composition of sublineages.



Lineage Specific SNPs in 50 DosR Regulon Genes AND 5 *rpf* GENES



Types of SNVs identified by WGS

- Singleton: 33527 SNVs (47%)
- Lineage/sublineage-specific SNPs, probably indicating adapting mutations.
- Non-lineage Clade specific SNVs- outbreak identification
- Homoplastic SNPs (>0.5%, 5 isolates)
  - Drug resistance mutations
  - There were 1229 SNVs in 1170 MTB isolates: (1121 in 589 genes 440 SNVs in PE-PPE genes)
- SNPs occurring in only one subset of sublineage.
- Others