University of Zurich[UZH]

URPP Evolution in Action

BIO634 – Next-Generation Sequencing 2

Phylogenomics

Dr. Heidi E.L. Tschanz-Lischer
University of Zurich
Switzerland

21 April, 2017

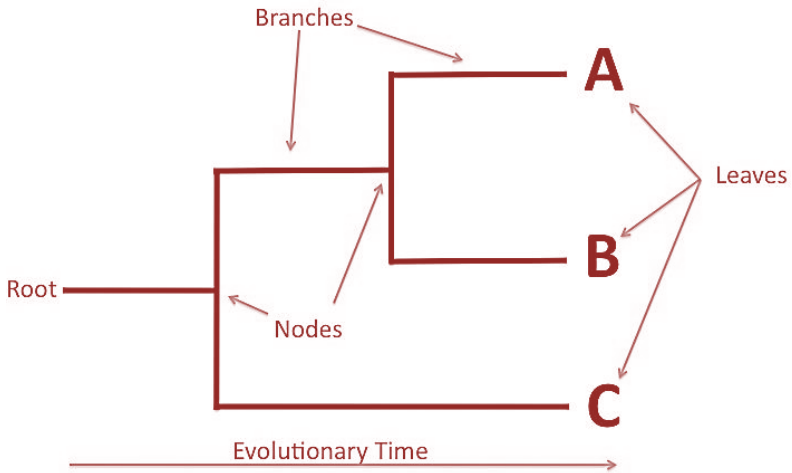
Phylogeny

Phylogenetics:

- Compares and analyzes sequences of single/small number of genes
- Study evolutionary history and relationships among individuals, populations or species

Phylogenetic tree:

- branching diagram or "tree"
- showing the inferred evolutionary relationships



Phylogenomics:

- Intersection between evolution and genomics
- Analysis of genome data to reconstruct the evolutionary history
  → Based on entire genomes or large parts

Multiple sequence alignments

Phylogenetic trees are based on multiple sequence alignments.

Two common used tools:

- ClustalW:
  1. Pairwise alignment → estimate pairwise distances
  2. Create a guide tree (neighbour-joining)
  3. Use the guide tree for multiple alignment (progressive alignment)
     1. first align most closely related pair of sequences
     2. the next most similar one to that pair
     3. …
- Muscle:
  1. Estimate pairwise distances
     (k-mer: count the number of in common short sub-sequences)
  2. Create a guide tree (UPGMA or NJ)
  3. Use the guide tree for multiple alignment
     (progressive alignment)
  4. Estimate pairwise distance

Repeat until tree stays stable

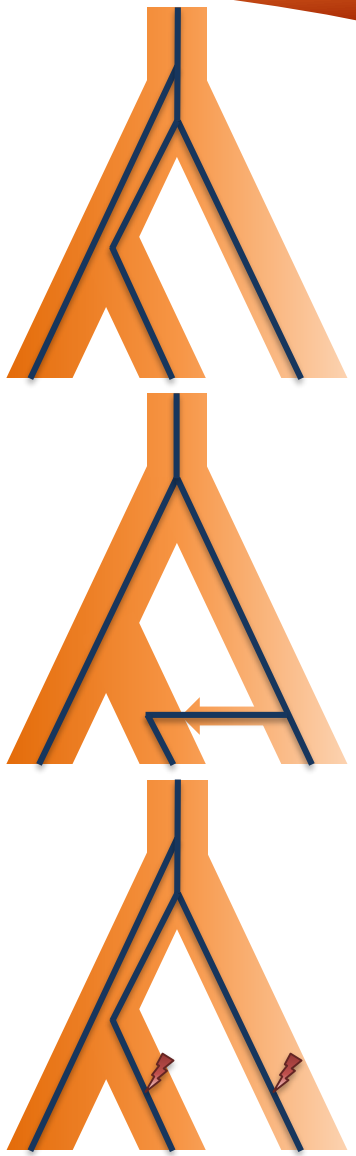
Phylogenomics

- Until few years ago most phylogenetic reconstructions were based on one locus

  → these trees can be distinct from species trees
  - Incomplete lineage sorting (deep coalescent)
  - Gene flow
  - Selection

- Phylogenetic reconstruction based on multiple loci can solve the problem
  - Resolved phylogenies in mammals, annelids or crustaceans (Meredith et al. 2011, Struck et al. 2011, von Reumont et al. 2012)

- NGS enables to obtain genomes-wide data sets



Mapping / de novo assembly

NGS-data:

- De novo assembly
  - Assemble each sample separately
    - Very small genomes: multiple sequence alignment
    - Multiple genome alignment (e.g. MAUVE, Mugsy, MUMmer, LAST)
  - Assemble all samples together
    - Closely related samples
    - Create alternative consensus sequence for each sample (e.g. Bcftools)
    - Call SNPs for each sample (e.g. GATK, freeBayes) → SNP filtering
- Reference mapping
  - Create alternative consensus sequence for each sample (e.g. Bcftools)
  - Call SNPs for each sample (e.g. GATK, freeBayes) → SNP filtering

Problems with reference mapping

- Mapping to a single reference genome — Bertels *et al.* 2014
  - Reads with more SNPs are less likely to be mapped → introduces a bias
  - Taxa with divergence >5-10%: distance to reference is systematically underestimated
    → wrong topology and branch length
  - Solution: map to multiple references (program REALPHY)

- Exclusion of non-polymorphic positions — Lewis 2001; Bertels *et al.* 2014; Leaché *et al.* 2015
  - Monomorphic positions provide information about the amount of evolution that has occurred along terminal lineages
  - Terminal branch length are overestimation
    → likelihood for particular tree topology is biased
  - Reduces accuracy: topology
  - Relative branch length are underestimated
  - → Using full sequence rather than SNPs is preferable
  - → But computational time increases
  - → RAxML: new acquisition bias correction method (conditional likelihood)

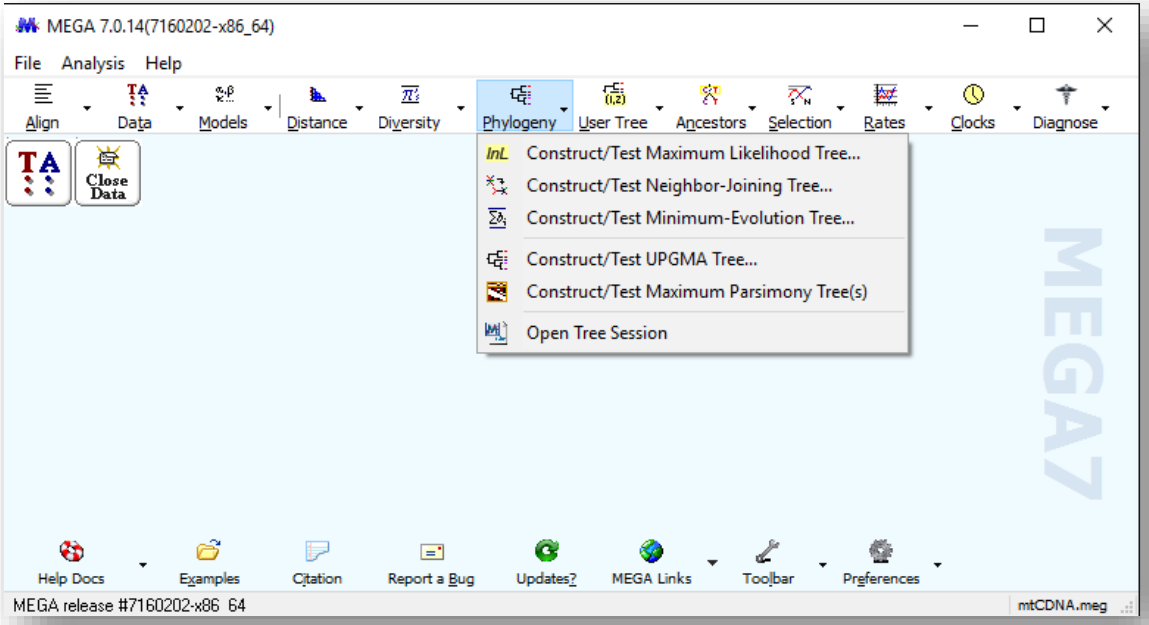
Phylogenetic tree reconstruction

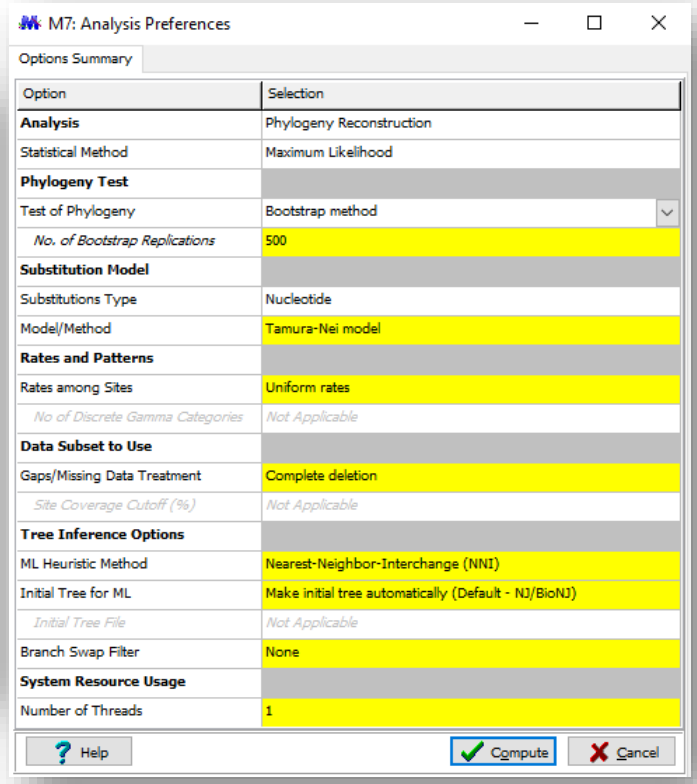
There exist several methods/algorithms for phylogenetic tree reconstruction:

- Distance matrix based methods
  - UPGMA → MEGA
  - Neighbor-joining → MEGA, PHYLIP
  - Minimum-evolution → MEGA
- Maximum Parsimony → MEGA
- Maximum likelihood → MEGA, PhyML, RAxML
- Bayesian inference → MrBayes, BEST, BEAST2, SNAPP

MEGA

- Runs on Windows, Mac OS X, Debian, RedHat, other Linux
- Graphical (GUI) (not for Linux systems) or command line (CC)
- http://www.megasoftware.net





UPGMA

Unweighted Pair Group Method with Arithmetic Mean

- simple hierarchical clustering method
- Based on a distance matrix (genetic distance between sequences)
- Assumes ultrametric tree: distances from root to every branch tip are equal

- Steps:
  1. Construct distance matrix
  2. Cluster the two shortest distance nodes into an internal nodes
  3. Recalculate the distance matrix
     → distance between clusters is the average distance between elements of each cluster
  4. Repeat the process until all nodes are grouped in a single cluster

✅ Simple and fast → suitable for large data sets

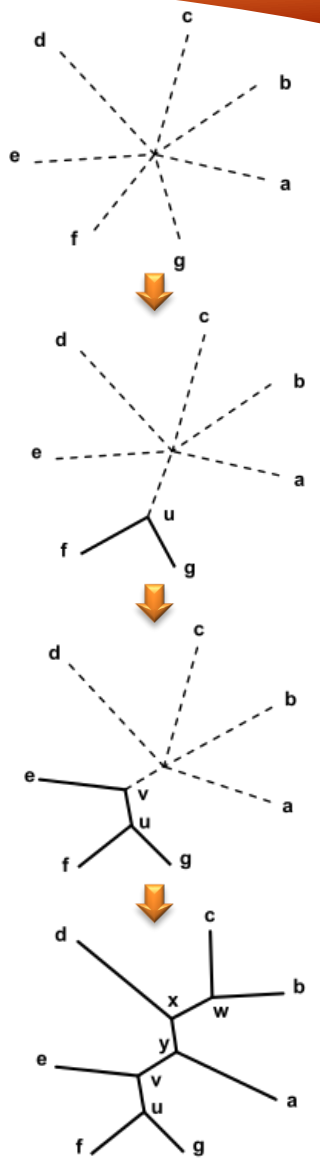
❌ Compress sequence information → distance data

❌ Very sensitive to unequal evolutionary rates

Neighbour-joining (NJ)

- Phylogenetic tree is constructed from a star-like tree by grouping nodes with shortest distances of branch length together

- Steps:
  1. Distance matrix
  2. Find two nodes with lowest value → create new node
  3. Calculate distance for the two nodes to the new node
  4. Calculate distance for each external node to the new node
  5. Repeat the process until only one terminal is present



✅ Fast → suitable for large data sets

✅ Allows for unequal rates of evolution
→ branch length are proportional to amount of changes

❌ Compress sequence information → distance data

❌ biased tree under some condition
→ no guarantee that the recovered tree best fits the data

Minimum-Evolution (ME)

- Find the tree that minimize the sum of branch lengths

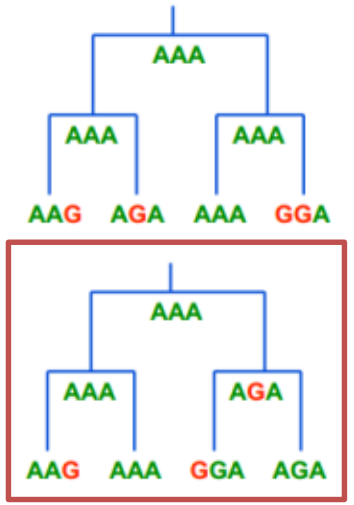
- Steps
  1. Distance matrix
  2. Construct all possible topologies and estimate total branch lengths
  3. Choose a tree with the smallest total branch length

✅ Better results for long DNA sequences than NJ

❌ For short sequences NJ method get more often the correct tree

❌ Slower than NJ

Maximum Parsimony (MP)



- Find the tree that **requires minimum number of changes** to explain the data
- Assumptions:
  - Any nucleotide can convert to any other nucleotide
  - Positions are independent

- Steps:
  1. Multiple alignment
  2. Construct all possible topologies and base on evolutionary changes to score each of these topologies
  3. Choose a tree with the fewest evolutionary changes as the final tree

✅ Reflect ancestral relationship
✅ Use all known evolutionary information

❌ Does not provide information on branch length
❌ Does not correct for multiple mutations
❌ Long computation time

Maximum likelihood (ML)

- Find the tree that maximizes the likelihood of the data
- Basic idea:
  - building a tree based on a mathematical model for nucleotide substitutions
  - find a tree based on probability calculations that best accounts for the data set

- Steps
  1. Multiple alignment
  2. List all possible topologies of each data partition (e.g.: column)
  3. Calculate probability of all possible topologies for each data partition
  4. Combine data partitions
  5. Identify tree with the highest overall probability as most likely phylogeny

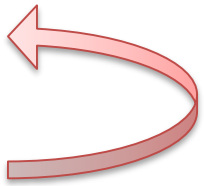
✅ More accurate than other methods
✅ All sequence information is used
✅ Allows varying rates of evolution across sites

❌ Very slow
❌ Computationally intensive

Bayesian inference

- Closely related to the maximum likelihood methods
- uses a likelihood function to create a posterior probability of trees using a model of evolution based on some prior probabilities
  → posterior probability indicate the probability of the tree to be correct

- Steps: use Markov chain Monte Carlo (MCMC) sampling algorithms
  1. Start at an arbitrary point (random tree)
  2. Estimate probability the tree is correct
  3. Make a small random move
     (new tree with changed parameters)
  4. Estimate probability the new tree is correct
     → probability larger: accept new tree
     → probability smaller: accept new tree with small probability,
     else stay at old tree

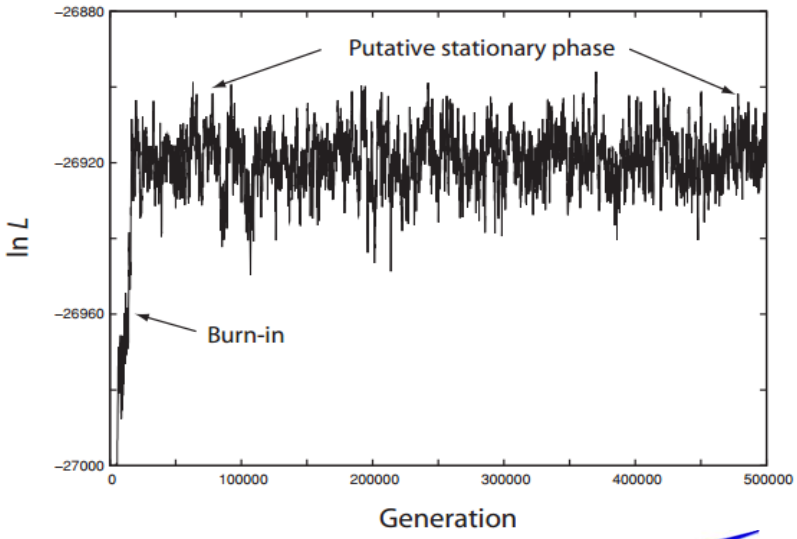


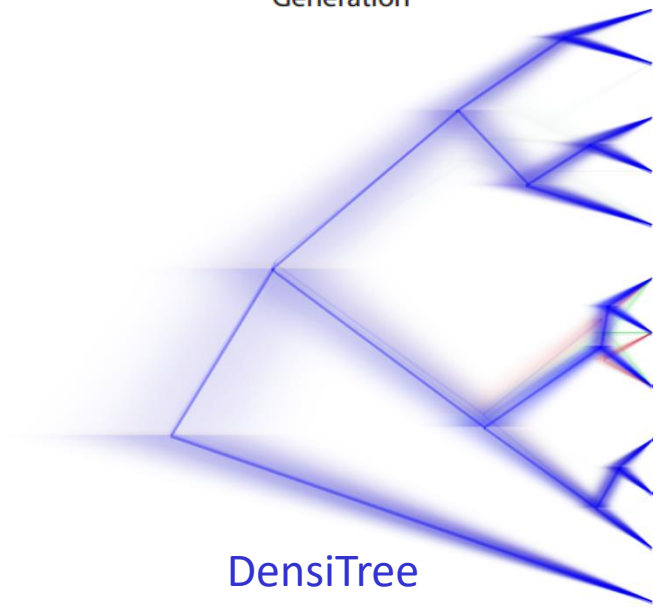
Repeat thousands or millions of times

Bayesian inference

- More accurate than other methods
- All sequence information is used
- Allows varying rates of evolution across sites
- Better able to accommodate missing data
- Possibility to account for the phylogenetic uncertainty
- Use of prior information

- Very slow
- Computationally intensive
- Use of prior information
  → not independent





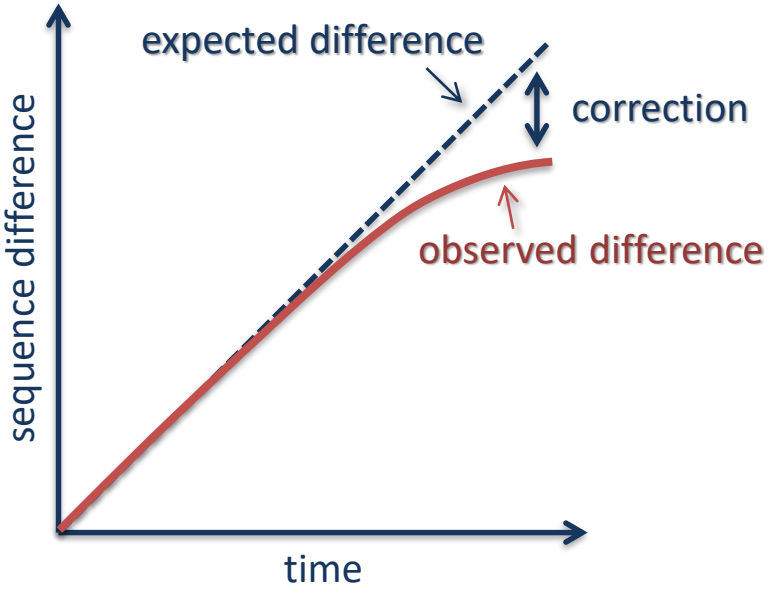
DensiTree

Substitution models

Substitution model: hypothesis about the relative rates of mutation at various sites

- assign a set of weights to each possible change
  → e.g.: correct for differences in the rates of transitions and transversions

- correct number of substitutions
  - The longer the time after divergence, the more likely two mutations occur at the same site
    → Simple genetic distance calculations will undercount the number of mutation
  - increases with increasing divergence time
    → can lead to long branch attraction (misassignment of distantly related but convergently evolving sequences as closely related)



Substitution models

Several methods have been proposed, all with different assumptions about the nature of the evolutionary processes:

Model	Unequal base frequencies	Different transition / transversion rates	Two different transition rates	Different substitution rates for each pair
Jukes-Cantor				
Tajima-Nei	x			
Kimura's 2 parameter		x		
Felsenstein / Hasegawa-Kishino-Yano (HKY)	x	x		
Tamura-Nei	x	x	x	
Generalized-time-reversible (GTR)	x			x

Substitution models

- Models may also allow for the variation of rates with positions in the input sequence
  → often the gamma distribution or log-normal distribution
- Methods of model selection:
  - likelihood ratio test (LRT): expresses how many times more likely the data are under one model than the other
  - Akaike information criterion (AIC): likelihood estimate with a correction factor to penalize over parameterized models
  - Bayesian information criterion (BIC): similar basic interpretation as AIC, but penalizes complex models more heavily

  → MEGA (Models → find best DNA/Protein Models)

  → jmodeltest2
- Some programs also allow partitioning of sequences:
  - estimating independent models of molecular evolution for different subsets of sites in a sequence alignment
  - has been shown to improve phylogenetic inference

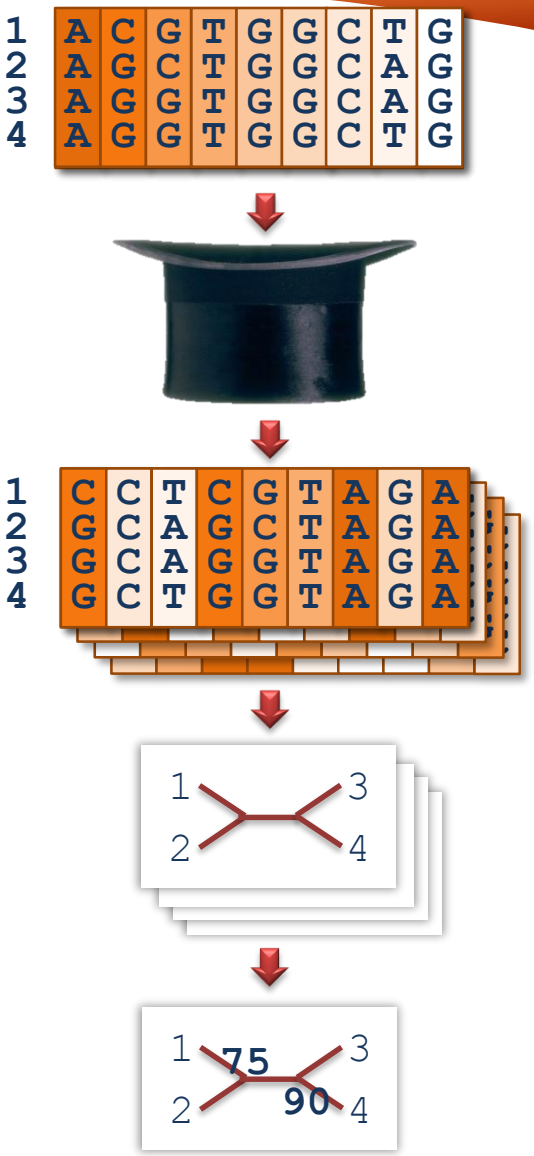
  → PartitionFinder:
  select best-fit partitioning schemes and models of molecular evolution

Evaluating tree support

- Bootstrapping
  - Method to test stability and consistency of a tree topology
    → does not test the accuracy of a tree

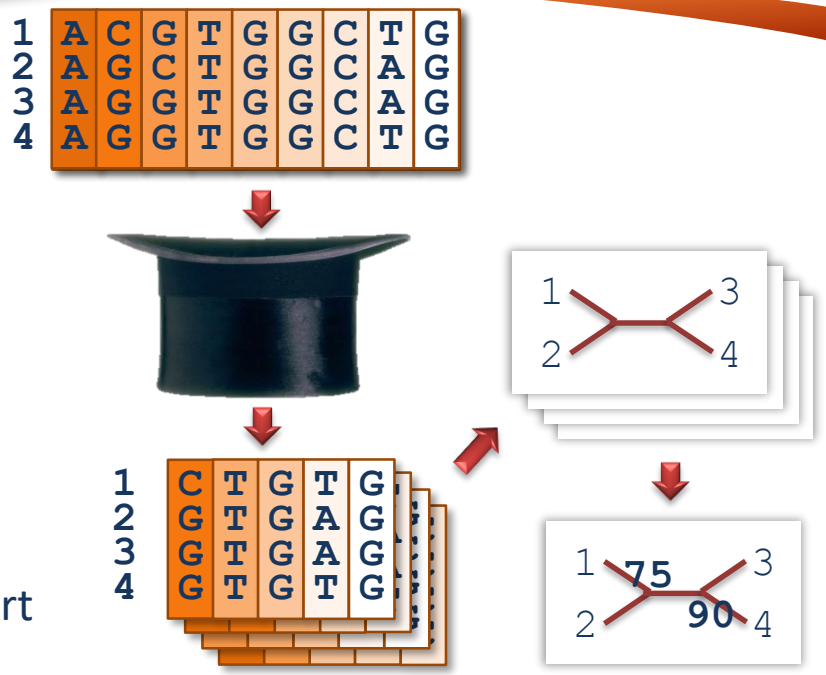
- Steps:
  1. Re-samples columns (with replacement) in a multiple sequence alignment
     → creates new alignment
  2. Get phylogenetic tree of the new alignment
  3. Repeat many times

- Results will show the percentage of times a particular branch point occurred out of all the trees
- Bootstrap values between 90-100 are considered statistically significant

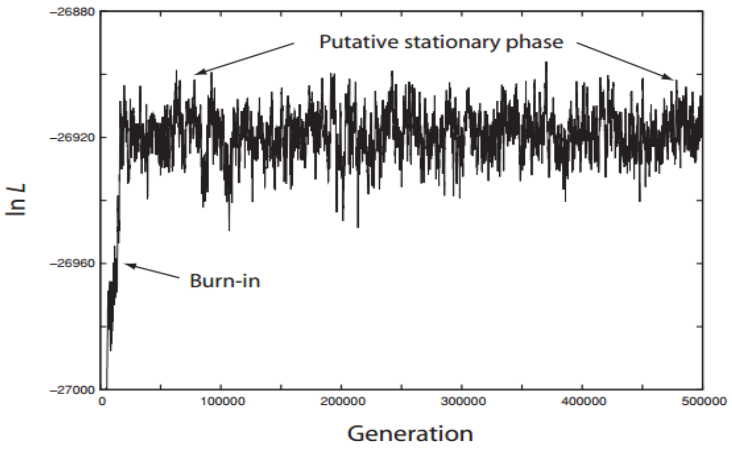


Evaluating tree support

- Jackknifing:
  - similar procedure to bootstrapping
  - columns of the matrix are sampled without replacement
  - Pseudo replicates are generated by randomly subsampling the data
  - → 50% jackknife: randomly sample 50% of the matrix many times to get nodal support

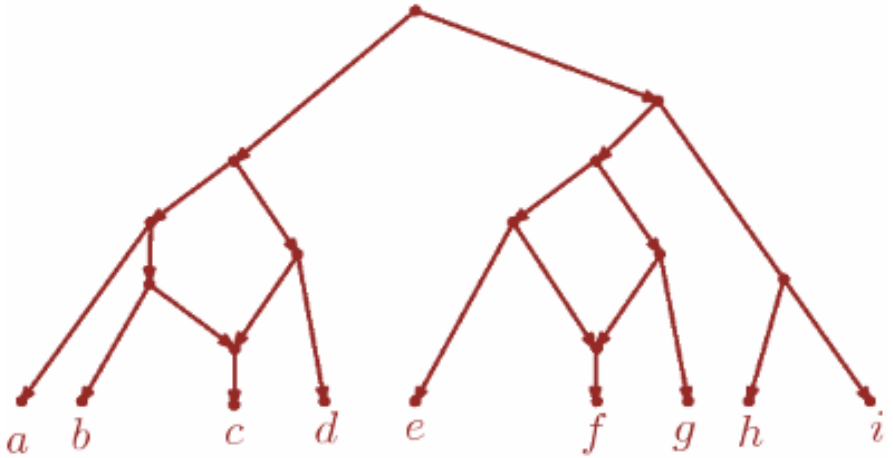


- Bayesian trees: Posterior Probability
  - posterior distribution of highly probable trees (not one single best tree)
  - nodal support:
    percentage of trees in the posterior distribution (post-burn-in)
    which contain the node



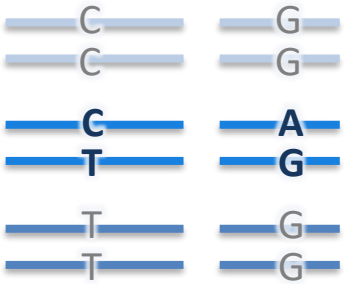
Phylogenetic network

- Used to visualize evolutionary relationships
- Used when reticulate events such as hybridization, horizontal gene transfer, recombination, or gene duplication and loss are involved
- They differ from phylogenetic trees by the addition of hybrid nodes (nodes with two parents) instead of only tree nodes (nodes with only one parent)



Heterozygous sites

- Diploid organisms harbour a difficulty:
  - Heterozygous sites



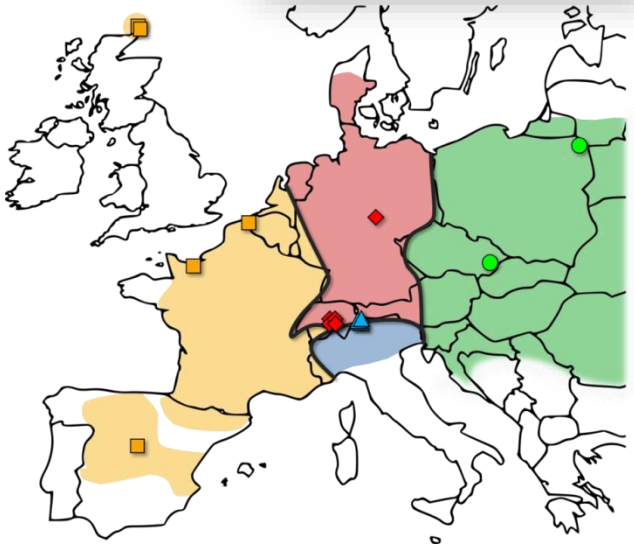
- Most phylogenomic methods require a concatenation, but how?
  - Encoding with ambiguity codes
    → partial integration
    → usually not handled well in phylogenomic methods
  - Exclude these positions

- Can be common in analyses within closely related taxa
  (Sota and Vogler 2003, Kelleher et al. 2007)

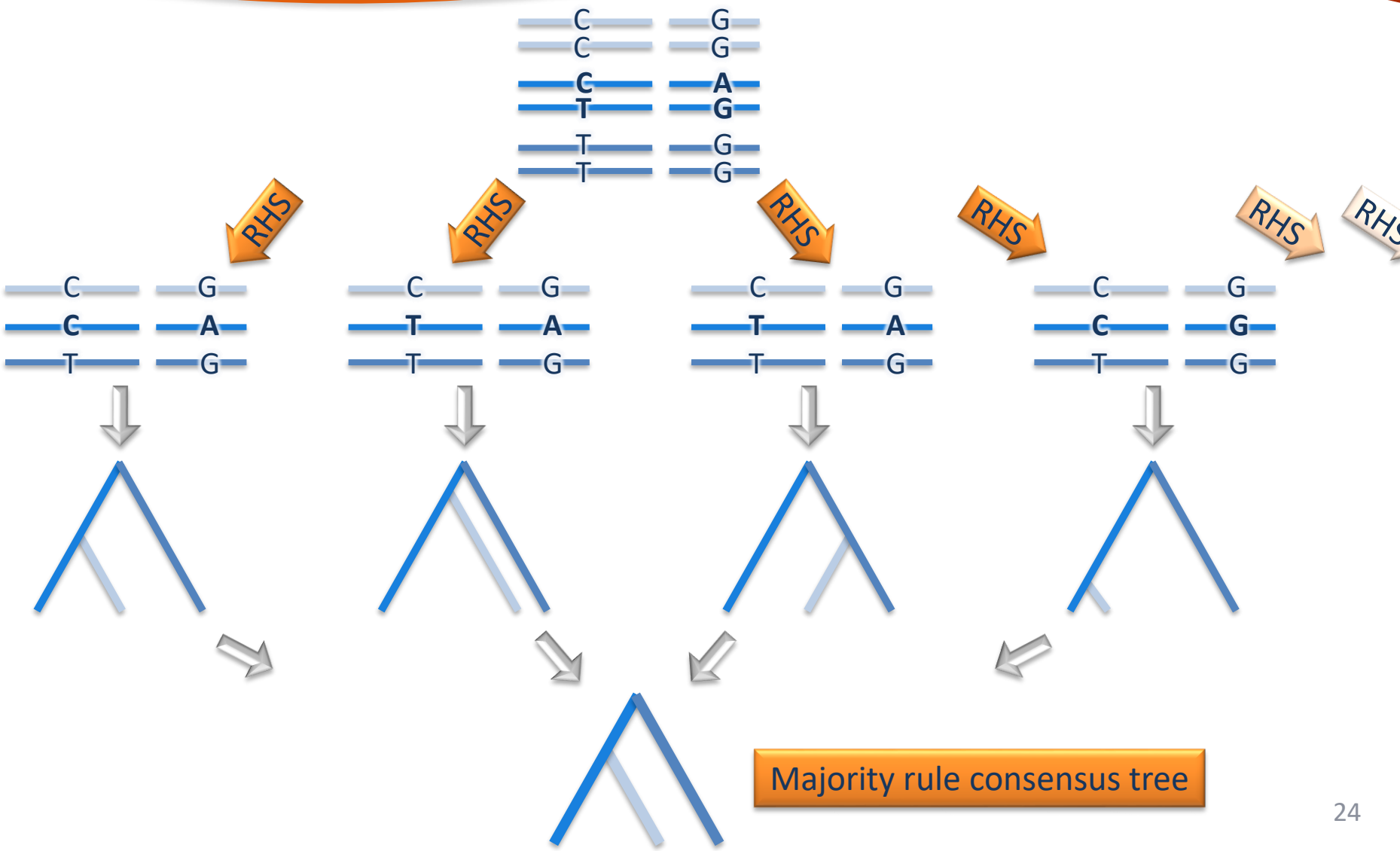
→ Lischer *et al*. 2014, MBE:
Influence of heterozygosity handling on phylogenomic estimations

Genome-wide data set

- Sequence AFLP fragments of 15 *M. arvalis* individuals with Roche 454 FLX technology
  - 1,552 polymorphic contigs
  - 6,807 SNPs (88.4% heterozygous)
- Distance-based methods (NJ) on concatenated data (Liu and Edwards 2009)
  → confirmed by simulation studies
- Handle heterozygosity:
  - Total removal
  - Ambiguity codes
  - Average number of base differences
  - New: Repeated Random Haplotypes Sampling



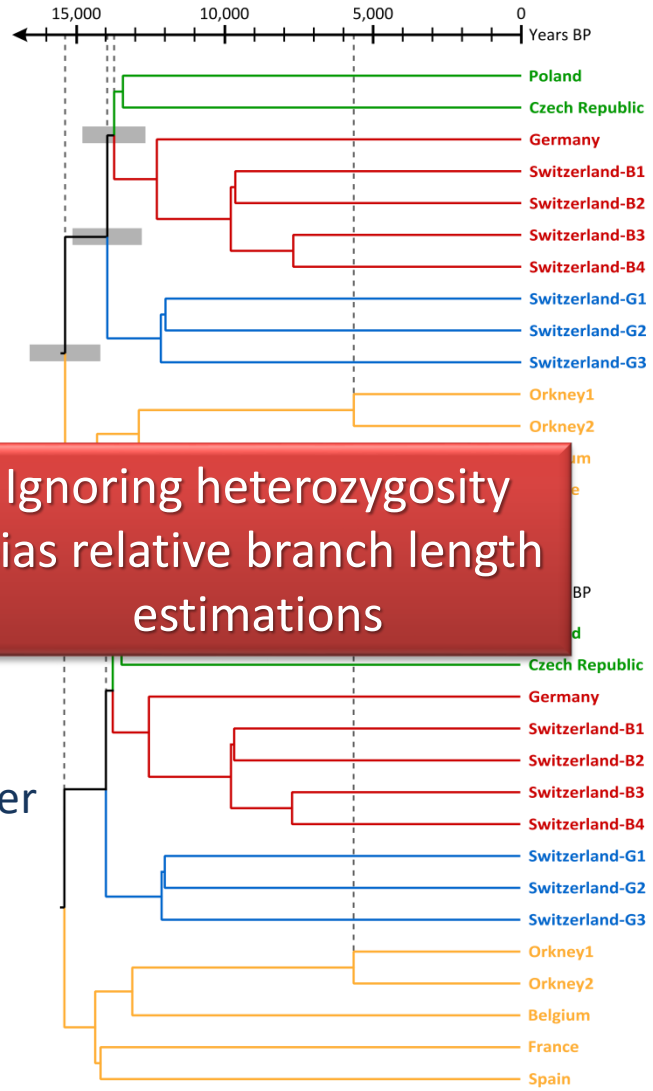
Repeated Random Haplotype Sampling (RRHS)



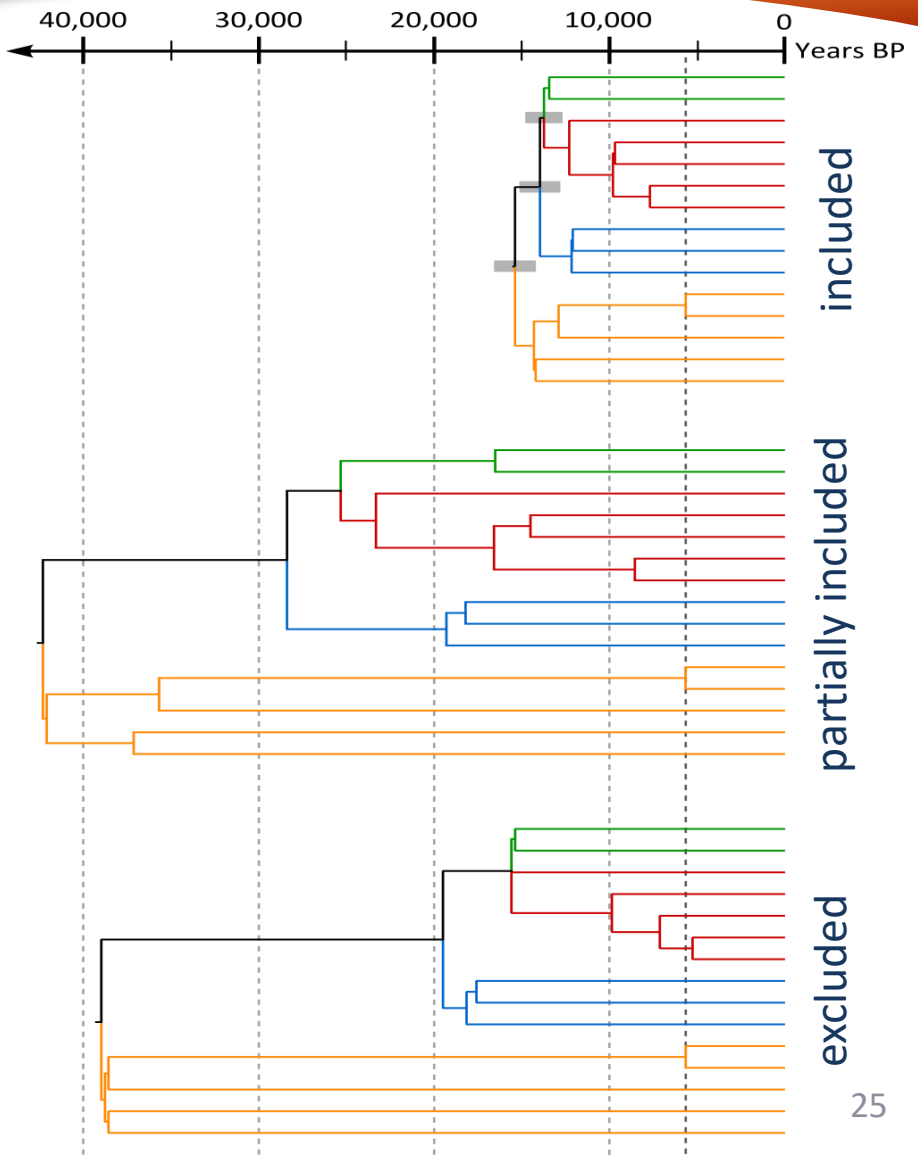
Influence of heterozygosity handling

RRHS

Average number of differences

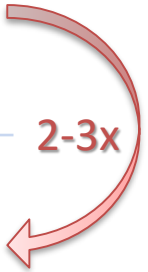


Ignoring heterozygosity bias relative branch length estimations



Influence of heterozygosity handling

Heterozygous position handling	Program	# runs	Divergence times between lineages (years BP ± standard deviation)		
			W – (I, C, E)	I – (C, E)	C - E
Random haplotypes	RAxML	5000	20,421 ± 1,653	18,739 ± 1,666	17,735 ± 1,426
	MrBayes	1000	19,730 ± 2,130	18,277 ± 2,087	17,009 ± 1,808
	NJ	5000	15,306 ± 1,188	13,947 ± 1,165	13,703 ± 1,060
Ambiguity codes	RAxML	100	65,610 ± 1,709	50,642 ± 1,951	43,625 ± 1,185
	MrBayes	100	61,619 ± 8,559	48,083 ± 7,259	40,567 ± 6,059
	NJ	1	42,164	28,352	25,278
Totally removed	RAxML	100	37,348	17,332	15,539
	MrBayes	100	38,258 ± 18,931	17,170 ± 8,739	15,771 ± 8,363
	NJ	1	36,299	19,467	15,524



Influence of heterozygosity handling

- Heterozygous SNP handling has a large impact on relative branch length estimations
  - → Ignoring heterozygosity may bias estimations of divergence times
  - → Correct integration of heterozygous information is needed
  - → RRHS approach integrates heterozygous SNP information into phylogenetic analyses (http://www.cmpg.iee.unibe.ch/services/computer_programs/rrhs/index_eng.html)

Lischer *et al*. 2014, MBE

Acknowledgment

- LAST: http://last.cbrc.jp/
- MAUVE: http://darlinglab.org/mauve/mauve.html
- Mugsy: http://mugsy.sourceforge.net/
- MUMmer: http://mummer.sourceforge.net/
- DensiTree: https://www.cs.auckland.ac.nz/~remco/DensiTree/
- MEGA: http://www.megasoftware.net/
- MrBayes: http://mrbayes.sourceforge.net/
- BEST: https://www.stat.osu.edu/~dkp/BEST/introduction/
- Beast2: http://beast2.org/
- SNAPP: http://beast2.org/snapp/
- PhyML: http://www.atgc-montpellier.fr/phyml/
- RAxML: http://sco.h-its.org/exelixis/web/software/raxml/
- PHYLIP: http://evolution.genetics.washington.edu/phylip.html
- jmodeltest2: https://github.com/ddarriba/jmodeltest2
- PartitionFinder: http://www.robertlanfear.com/partitionfinder/

- http://www.nature.com/protocolexchange/protocols/2740