Approximating the Bacterial Ancestral Recombination Graph

Xavier Didelot (University of Warwick)

Daniel Lawson (University of Bristol)

Aaron Darling (University of California-Davis)

Peter Green (University of Bristol)

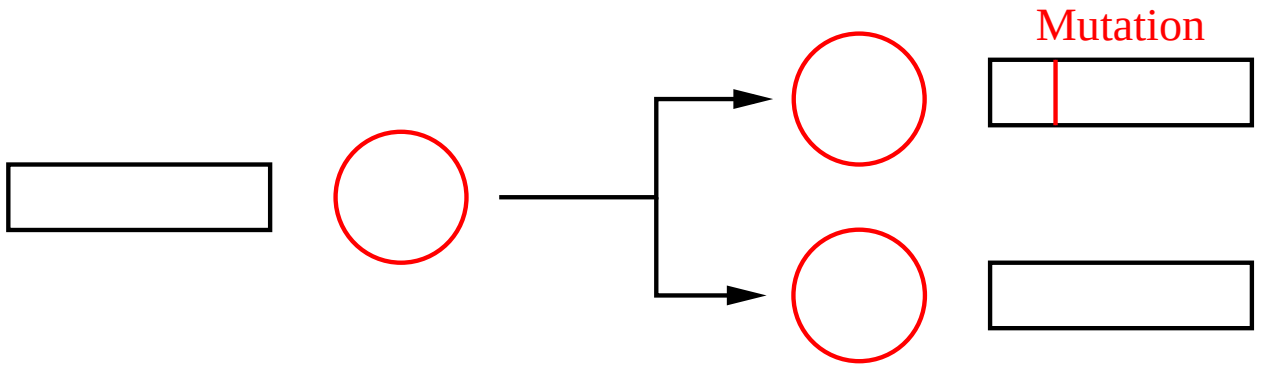
Daniel Falush (University College Cork)

Graphical Models and Genetic Applications

University of Warwick

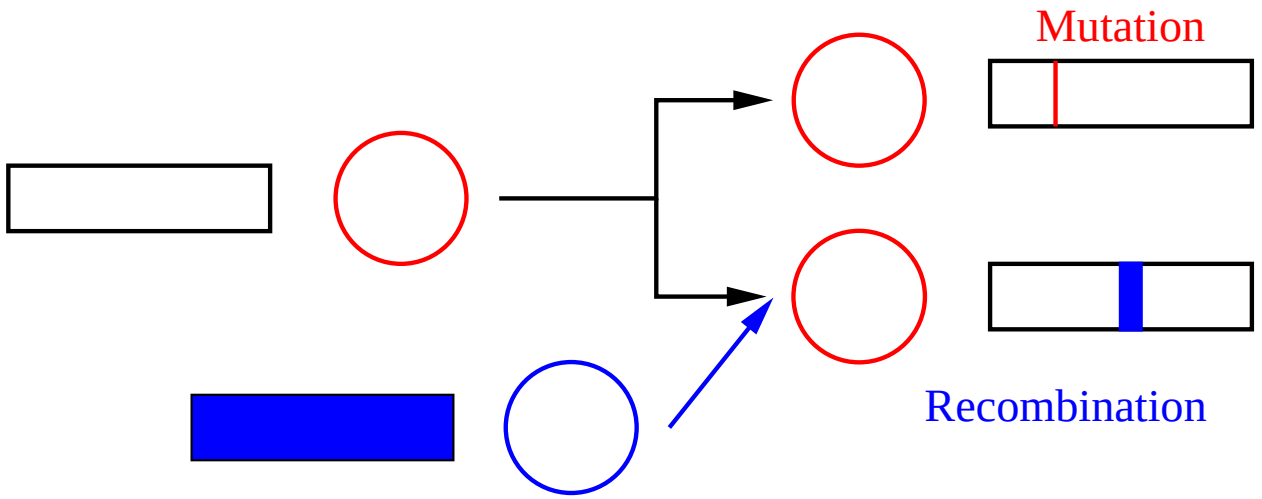
15-17 April 2009

Bacterial Reproduction



- Bacteria reproduce clonally:
  - Single bacteria splits into two identical copies

Bacterial Reproduction



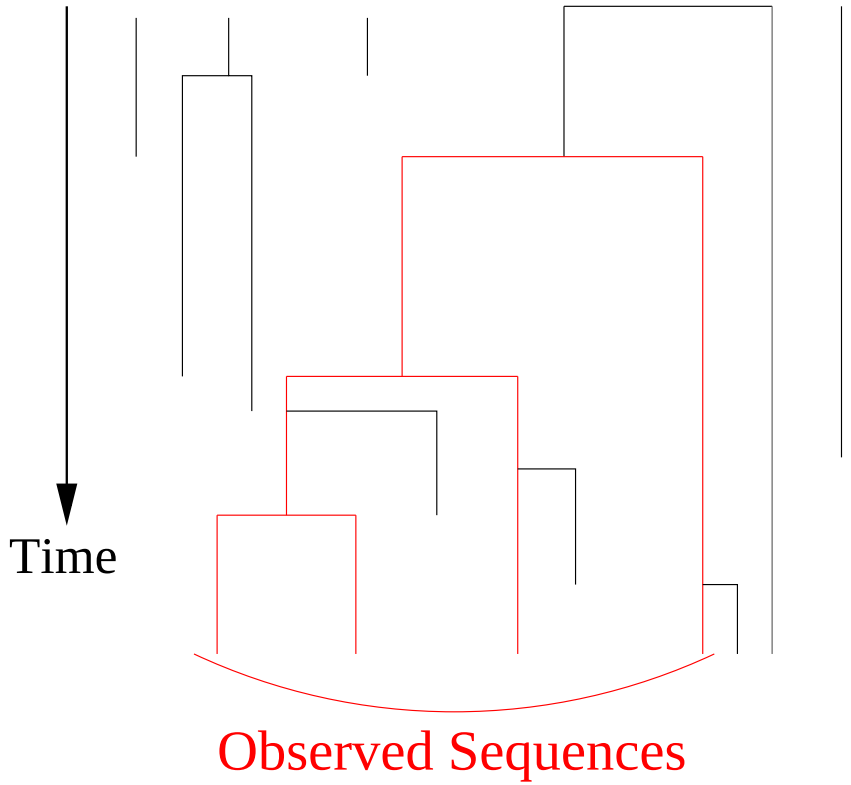
- Bacteria reproduce clonally:
  - Single bacteria splits into two identical copies
- But exchange DNA:
  - Transformation - from the environment
  - Transduction - transmission via bacteriophage
  - Conjugation - direct from another bacterium

Forward in time models

Moran model with constant population size $N = 6$.

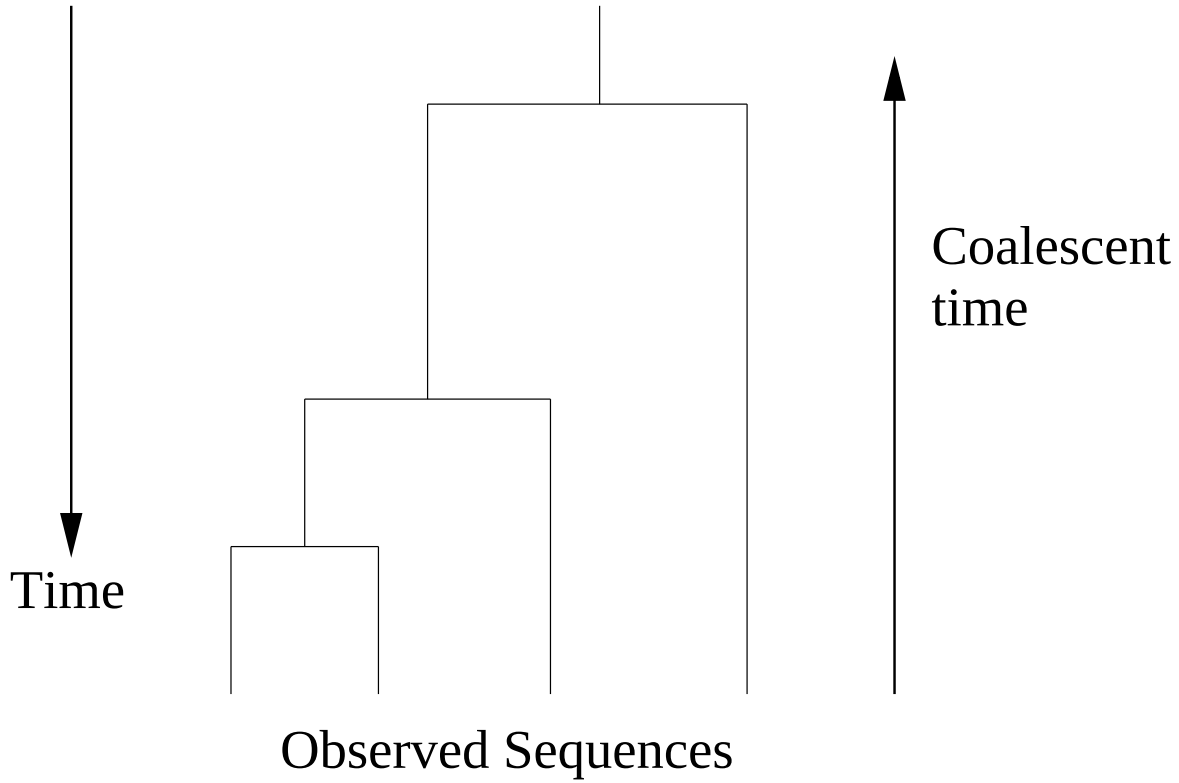
Timestep:

1. Each individual gives birth at rate b.
2. A randomly chosen individual is killed for each birth.



The coalescent

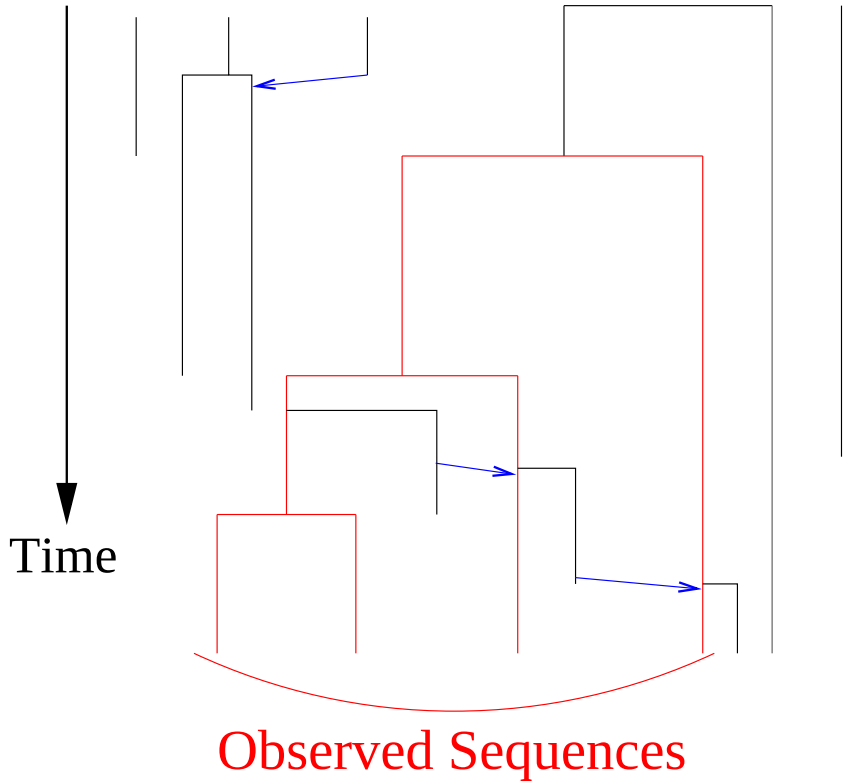
Pure death process with rate $k(k-1)/2$ where k is the number of lineages alive at a given time.



Allows us to calculate probability of a given tree.

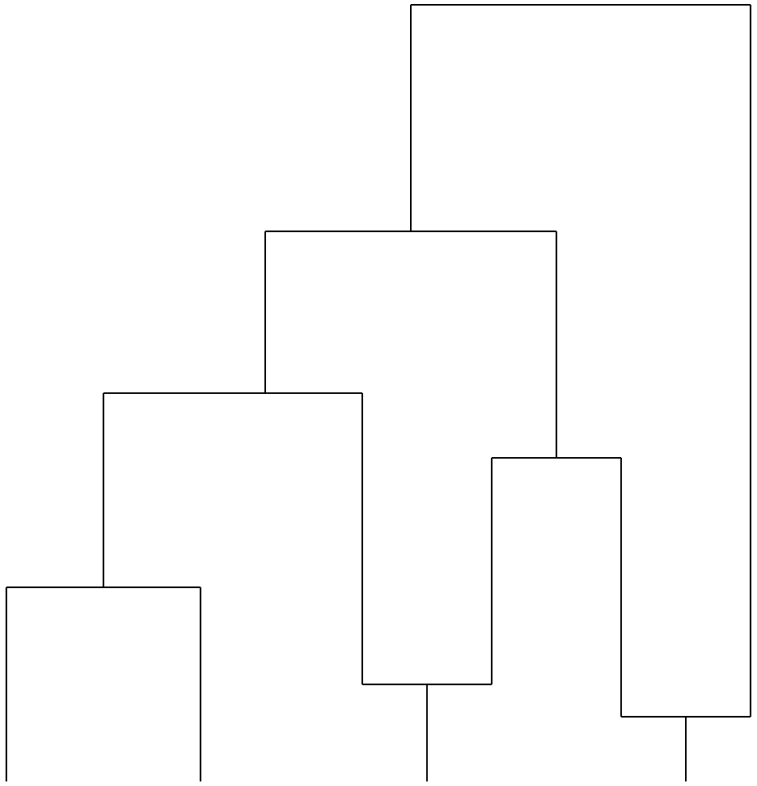
Recombination model

Allow recombination to occur with probability r for each birth; a second parent is chosen at random.



Ancestral recombination graph

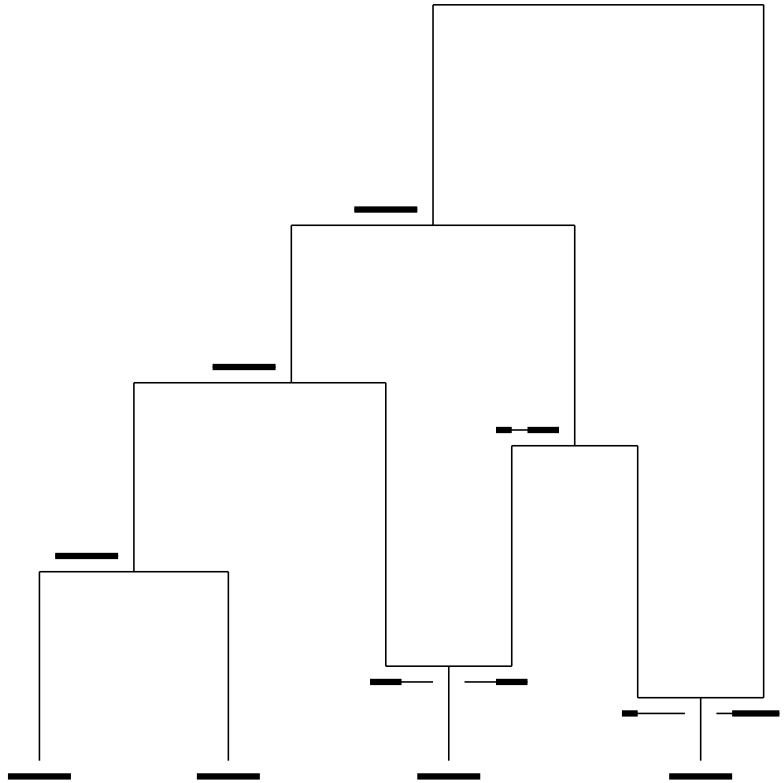
Birth-death process with recombination (birth) at rate $\rho k/2$ and coalescence (death) at rate $k(k-1)/2$.



Allows us to calculate probability of a given graph.

ARG with crossover

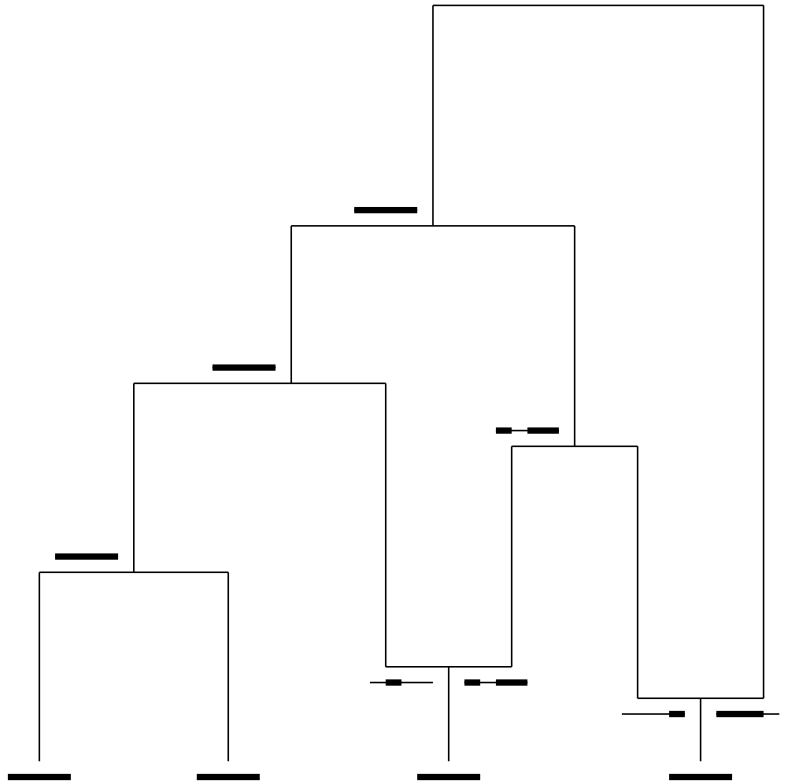
When a recombination occurs, a point S is chosen uniformly at random on the gene. The material on the left of S follows one line of descent and the material on the right the other.



- Symmetry of the two parents;
- This is the version of the ARG found in most of the literature.

ARG with gene conversion

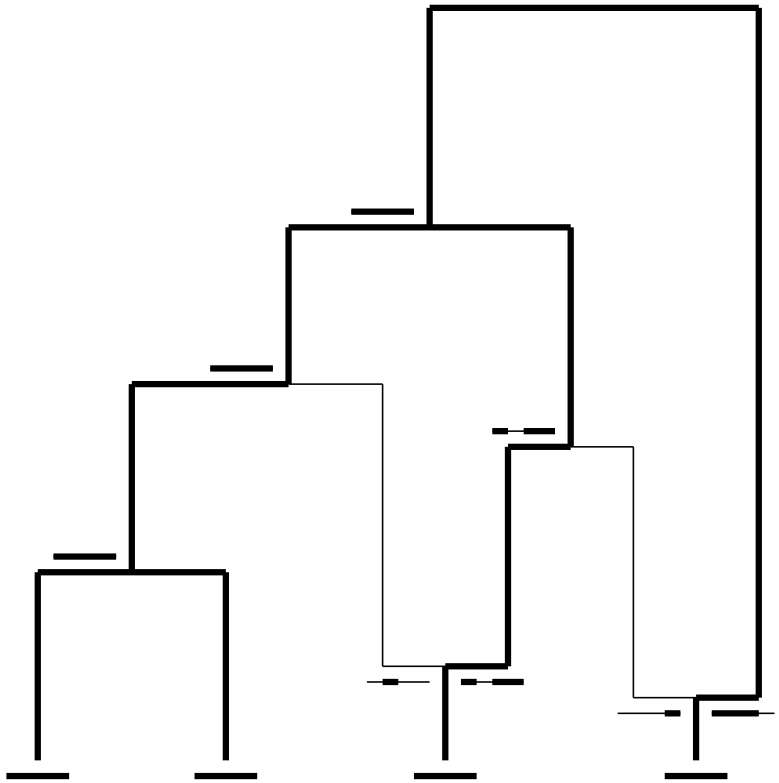
When a recombination occurs, a small fragment is contributed by the donor cell and the rest by the recipient.



- Asymmetry of the two parents (recipient/donor);
- This is the version of the ARG which is relevant to bacteria;
- Studied by Wiuf and Hein (2000).

Clonal genealogy

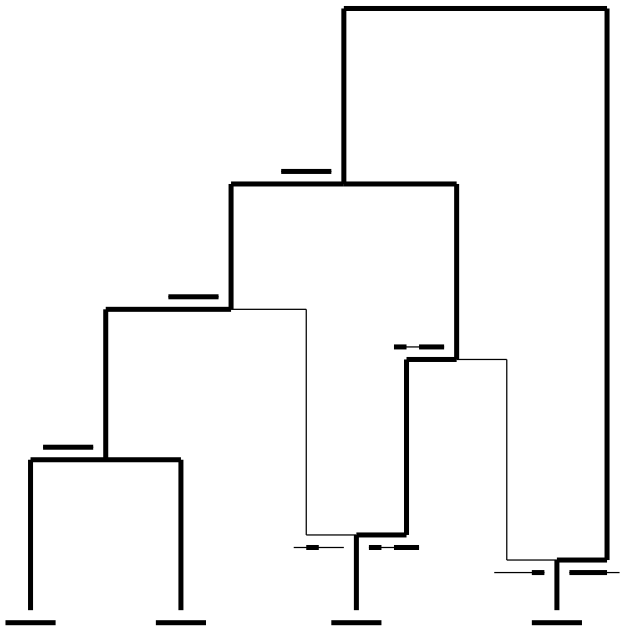
If we always follow the line of the recipient, we get a tree called “clonal genealogy”.



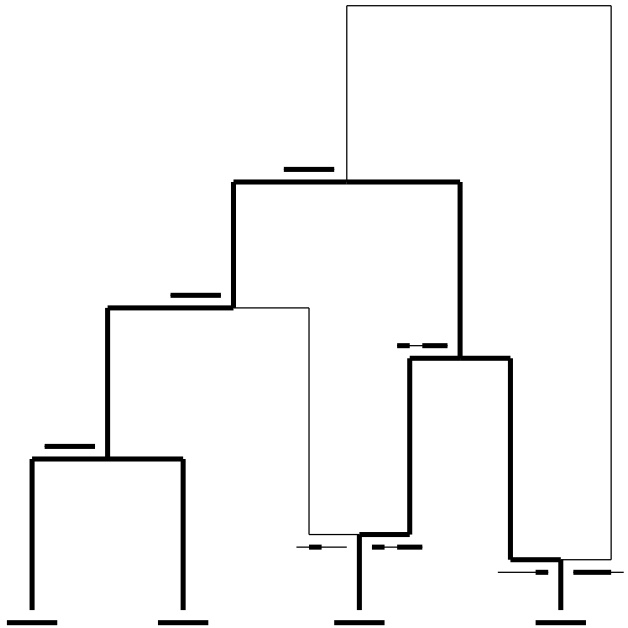
- Concept unique to ARG with gene conversion;
- The clonal genealogy is distributed as a coalescent tree.

Ancestral material and local trees

For each branch, the material that ends up in the leaves is called “ancestral material” and shown in bold. If we follow the ancestral material for a given site, we get the “local tree” of that site.



Local tree for the first DNA base



Local tree for the last DNA base

Full ARG inference

- Probability of the model:

$$P(Graph|Data) \propto P(Data|Graph)P(Graph)$$

- Inference under the full ARG model is difficult:
  - The state-space of ARGs is huge;
  - The data are uninformative about the actual ARG;
  - An ARG is more than the sum of its local trees.
- Importance sampling algorithm of Fearnhead and Donnelly (2001): a month for 31 sequences of 500bp;
- Can we find a good approximation?

Approximating the Bacterial Ancestral Recombination Graph

Xavier Didelot (University of Warwick)

Daniel Lawson (University of Bristol)

Aaron Darling (University of California-Davis)

Peter Green (University of Bristol)

Daniel Falush (University College Cork)

Graphical Models and Genetic Applications

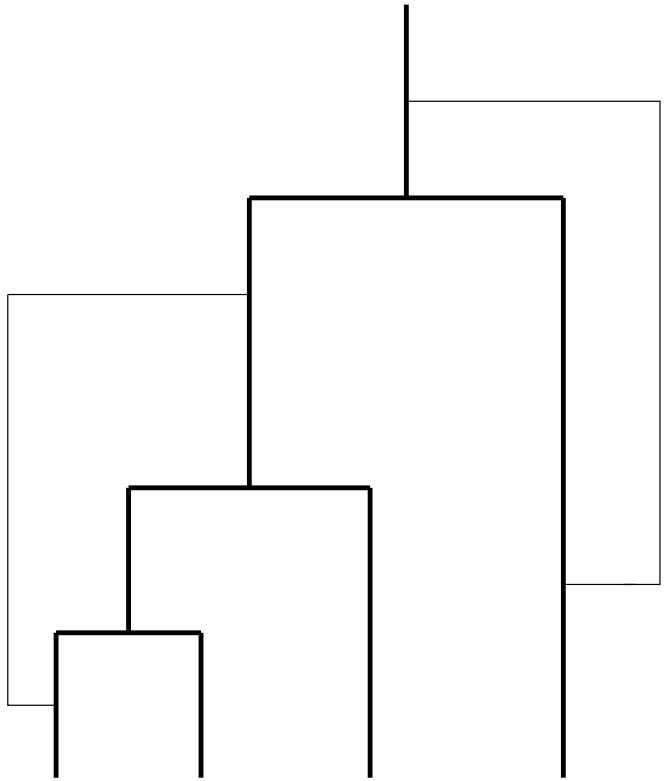
University of Warwick

15-17 April 2009

Approximating the ARG

- A good approximation to the ARG should:
  - Contain the parts of the ARG that are well informed and biologically important
  - Simplify the rest of the graph to make inference possible
- The clonal genealogy is of central interest in most studies. It is also likely to be well informed when the data is large and the recombination rate not too big.
- Thus we use the clonal genealogy as the centre of our approximation, and simplify the recombination process.

The weak ARG model



- Two differences with full ARG:
  - Recombinant edges not allowed to coalesce with each other;
  - Recombinant edges not allowed to recombine.
- The weak ARG is a “tree with added edges” rather than a graph

Probability of a wARG

- A weak ARG is made of a clonal genealogy $\mathcal{T}$ and a set $\mathcal{R}$ of recombinant edges.
- Each recombinant edge in $\mathcal{R}$ is characterized by an origin a_i on $\mathcal{T}$, a destination b_i on $\mathcal{T}$, a starting site x_i and an ending site y_i.
- Probability of a wARG:

$$P(Graph) = P(\mathcal{T})P(\mathcal{R}|\mathcal{T})$$

with:

$$P(\mathcal{T}) = \prod_{i=2}^{N} \exp\left(-\binom{i}{2} t_i\right)$$

$$P(\mathcal{R}|\mathcal{T}) = \exp(-\rho T/2)\left(\frac{\rho T}{2}\right)^{R} \prod_{i=1}^{R} P(x_i, y_i) T^{-1} \exp(-L(a_i, b_i))$$

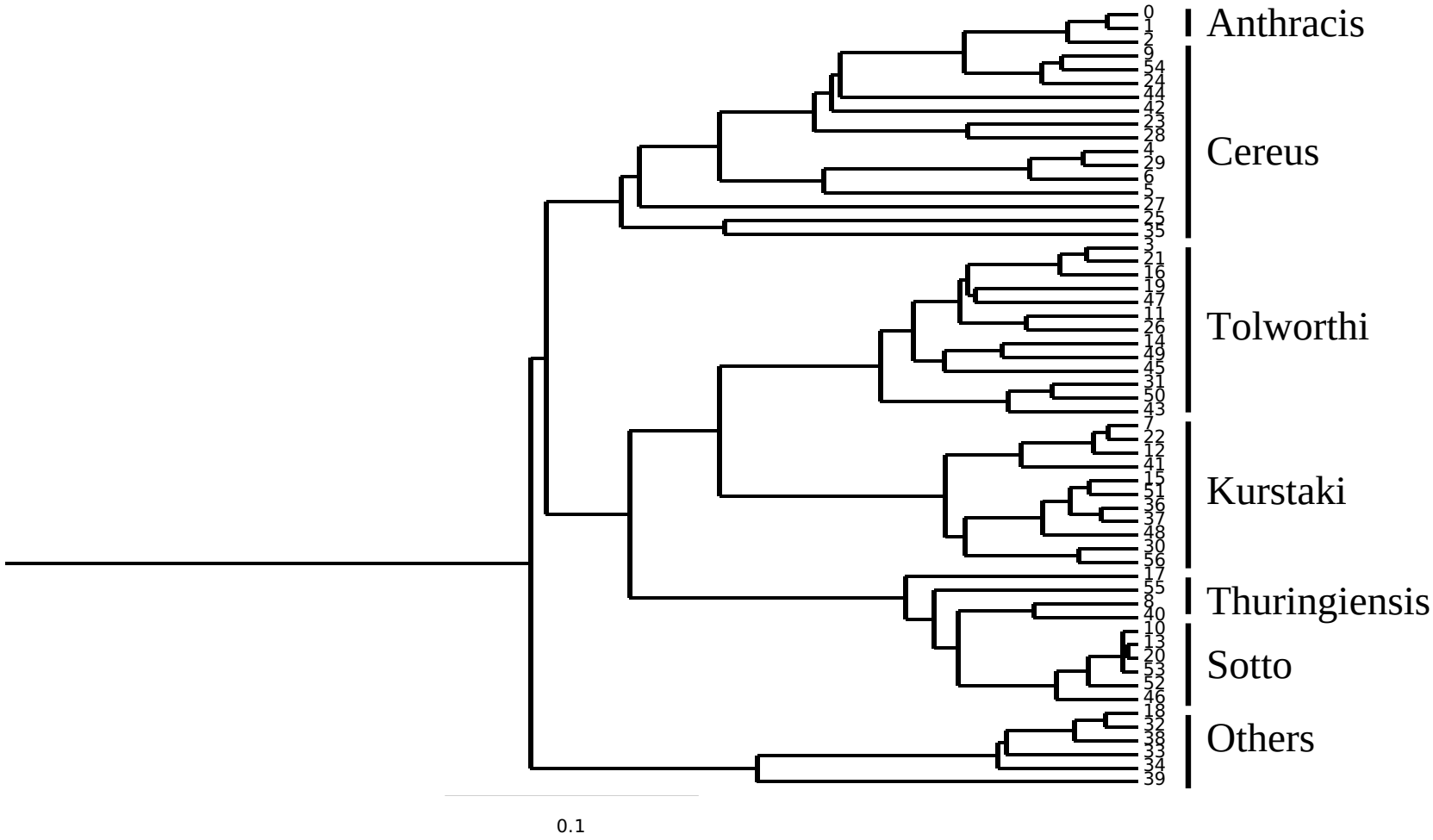
Bayesian Inference

- Inference problem: given a set of observed sequences, can we infer the ancestry graph?
- Likelihood calculation: as for the full ARG, the wARG defines a local tree for each site, so we can use Felsenstein's pruning algorithm.
- Reversible-jump Monte-Carlo Markov Chain
- Moves:
  - Subtree pruning and regrafting for $\mathcal{T}$;
  - Add and remove recombinant edges in $\mathcal{R}$ (only the local likelihood needs recalculating);
  - Updates for mutation rate θ, recombination rate ρ and mean import length δ.

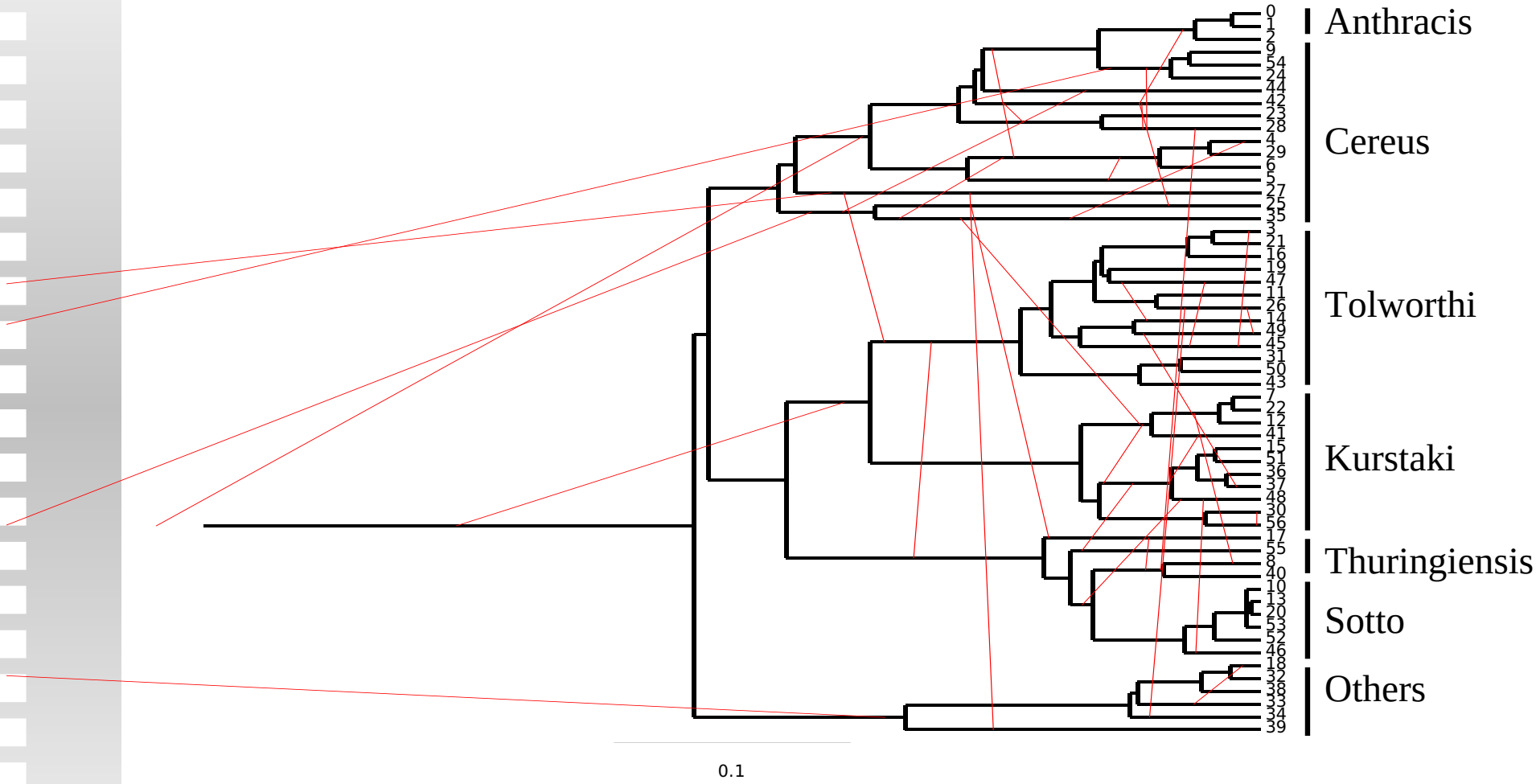
Implementation

- Implementation in progress, should be ready this year
- Current status: infer the recombination process given the clonal genealogy
- In progress: jointly infer the clonal genealogy
- The current implementation is interesting in itself, when the clonal genealogy is clear (eg. because ρ is low or lots of data) and we want to characterize patterns of recombination
- Example: Multi-Locus Sequence Typing (MLST) data for 57 isolates from the *Bacillus cereus* group (Priest *et al.*, 2004).
- These are very preliminary results!

Bacillus: clonal genealogy

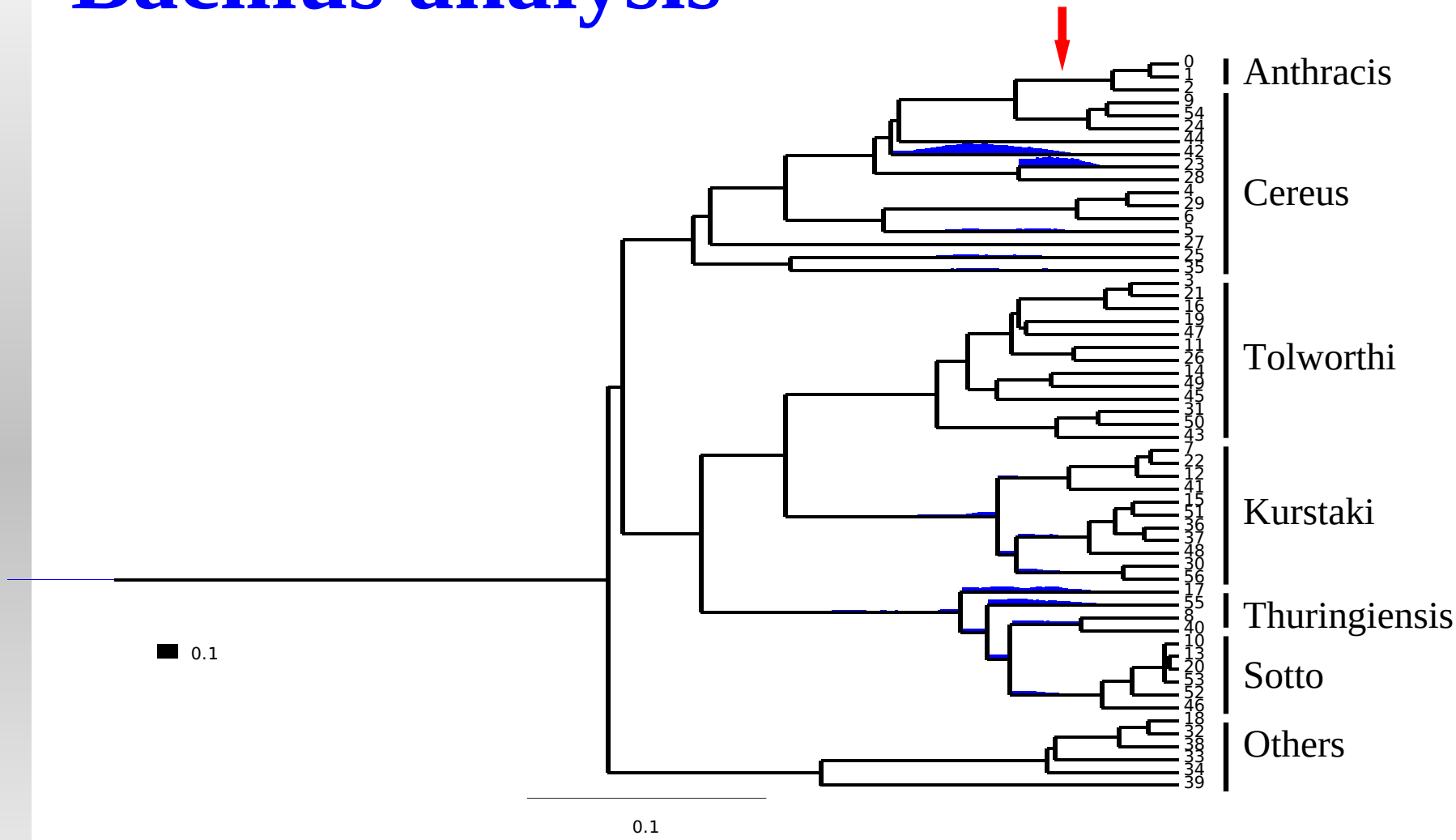


Bacillus: a single iteration



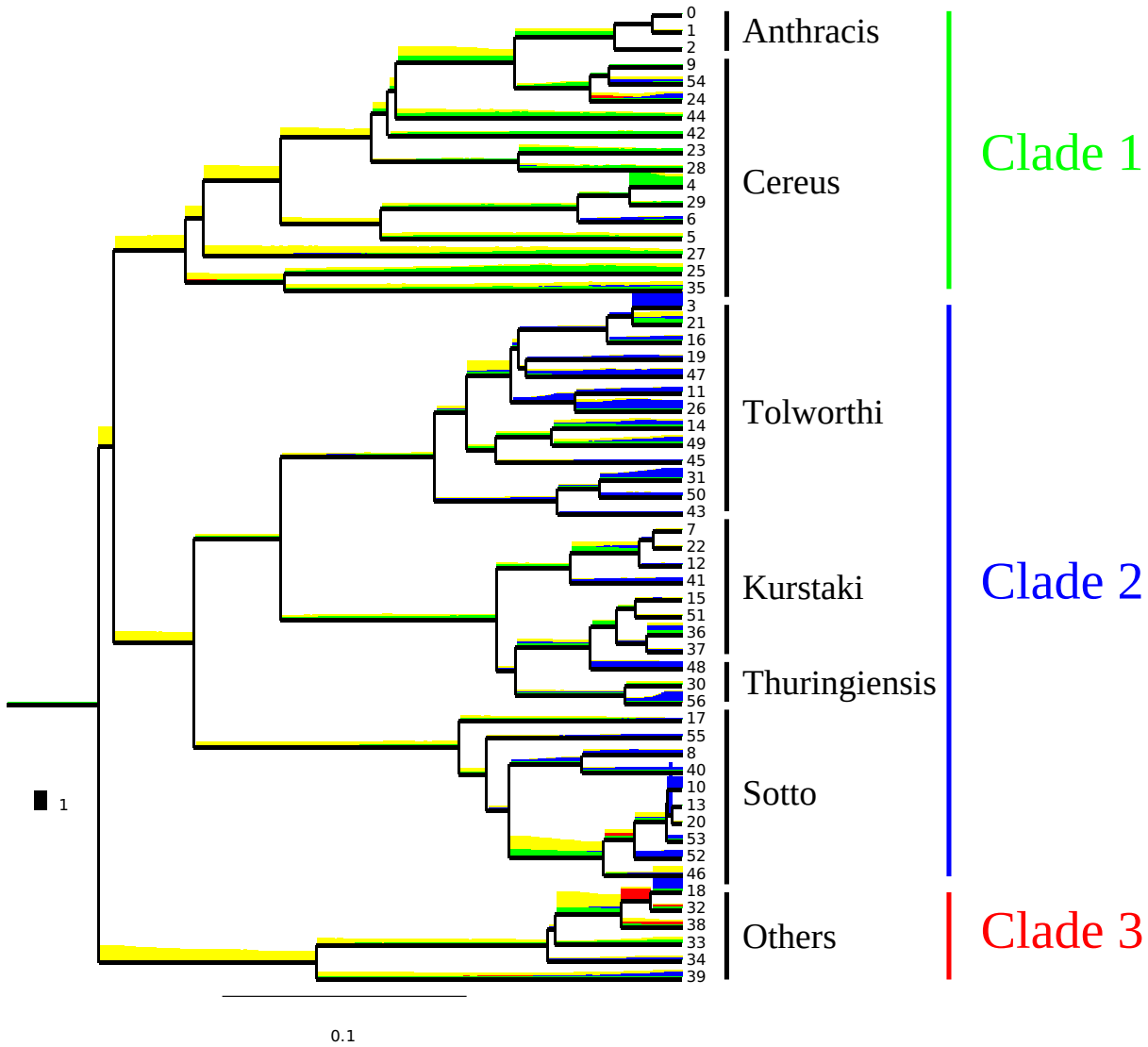
Problem: how can we summarize all iterations?

Bacillus analysis



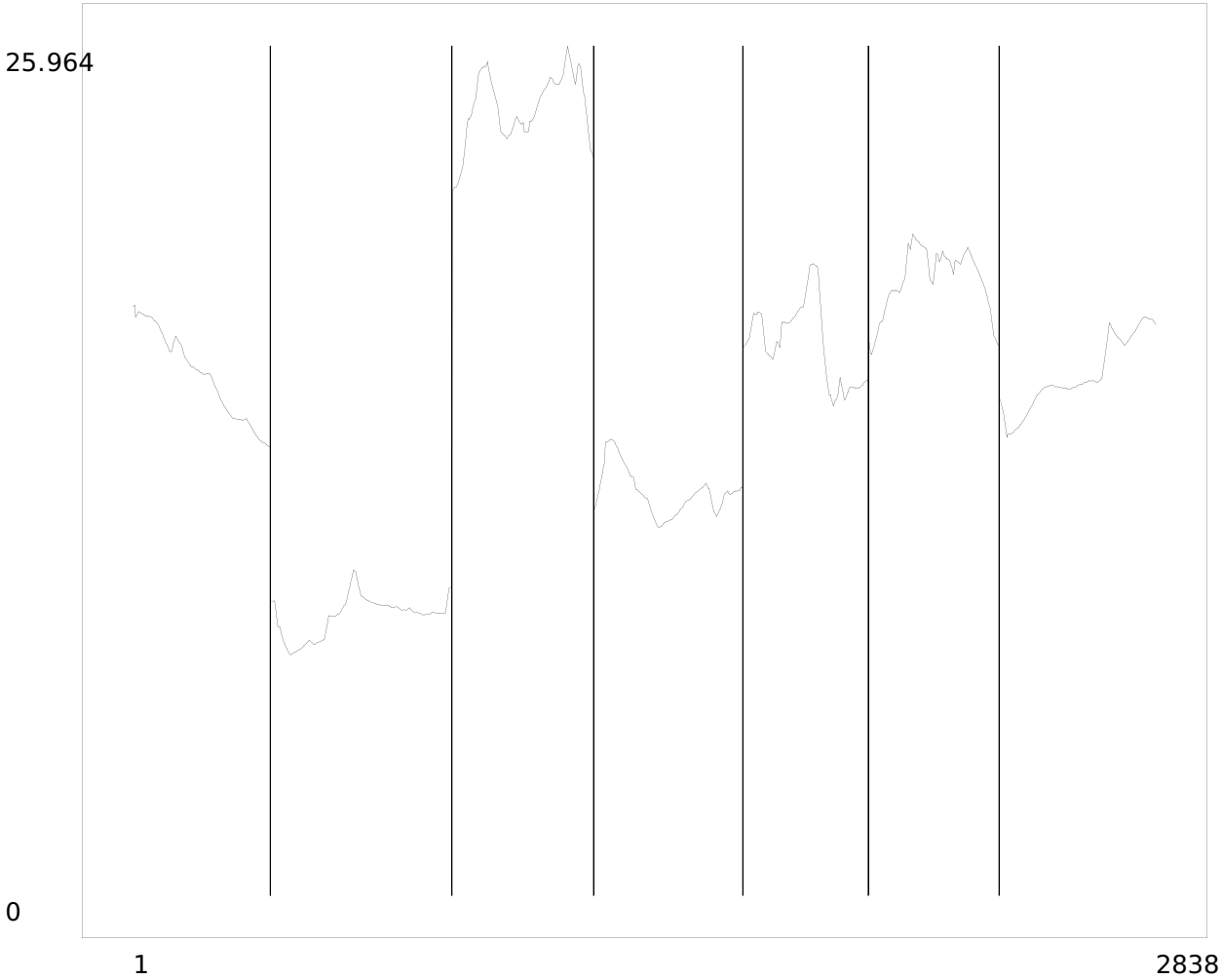
Distribution of orgins for the imports on a given branch

Bacillus analysis



Distribution of arrivals colour-coded by origin

Bacillus analysis



Distribution of recombination along the genes

Conclusions

- The ancestral recombination graph is a powerful model of ancestry, but is not usable in inference
- Need for efficient approximations
- Ignoring recombination completely (as most methods do) is dangerous
- The weak ARG is based on the existence in bacteria of a clonal genealogy, and approximates the recombination process by modelling the origin of each import but not its full ancestry
- Our approximation is simple enough to perform inference
- Our program can reveal patterns of recombination
- It can also be used to infer the clonal genealogy