# Applications of TransPhylo

In this session we're going to explore several extensions of TransPhylo, and in particular consider them in the context of an upcoming paper.

First, some background...

BEAST

Bayesian Evolutionary Analysis Sampling Trees

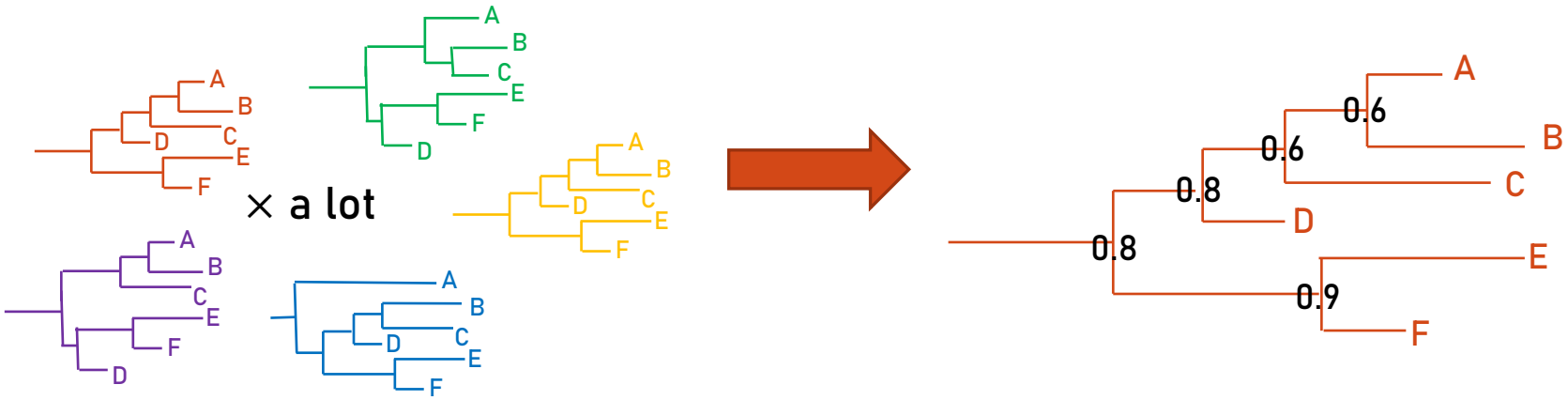
Beast2

Bayesian evolutionary analysis by sampling trees

Cross-platform programs for Bayesian analysis of molecular sequences using MCMC

Use MCMC to average over tree space, so that each tree is weighted proportional to its posterior probability i.e. rather than inferring a single phylogenetic tree, you infer a huge collection of phylogenetic trees

From that collection, you can then infer how often a given clade of pair of tips appears, or find the tree with the highest posterior probability, for example.



# There's a lot you can do with BEAST(2)…

Divergence Dating Tutorial with BF

Alexei Drummond, Andrew Rambaut, Remco Bouckaert

Model Averaging for Clocks Tutor

**Summary:** Bayesian model a

Using BETS to evaluate temporal sig

**Summary:** This tutorial describes the use of f Temporal Signal) to examine the temporal signal of a data set.

Ancestral Reconstruction/Discrete Phylogeography with BEAST 2.3.x

Remco Bouckaert remco@cs.a .nz and Walter Xie

Spherical Phylogeography with BEAST 2.5

Remco Bouckaert r.bouckaert@auckland.ac.nz

Epoch Model Tutorial

**Summary:** Setting up time-heterogeneous epoch substitution models in BEAST.

Estimating Species Trees from Multilocus Data

Joseph Heled, Remco Bouckaert, Alexei nmond and Walter Xie

*BEAST in BEAST 2.0

rarchical models for sets of parameters

hical phylogenetic models for different scenarios.

A Rough Guide to CladeAge

estimation of clade ages based on probabilities of fossil sampling

Analysing Continuous Traits

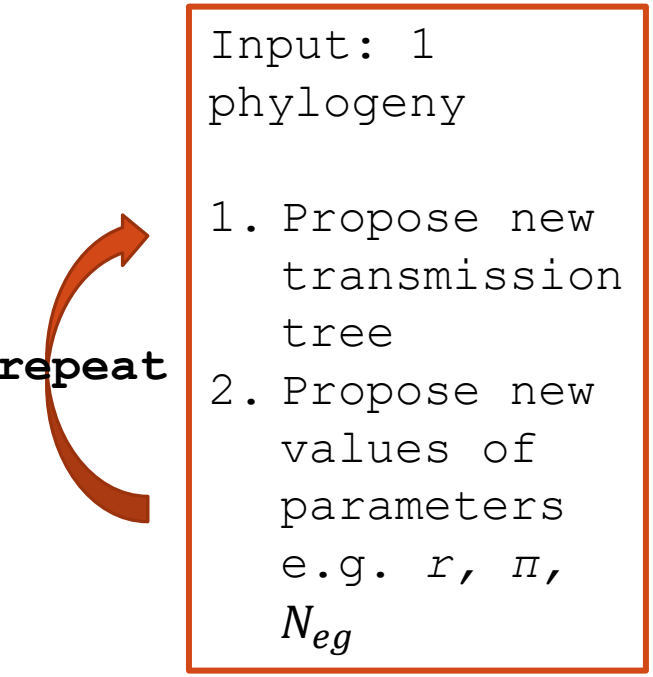
**Summary:** How to jointly estimate a molecular phylogenetic tree and the coevolutionar traits.

# Multi-tree capability

In the base TransPhylo methodology, users input a single phylogeny to the transmission reconstruction algorithm – this is assumed to be a 'ground truth'.

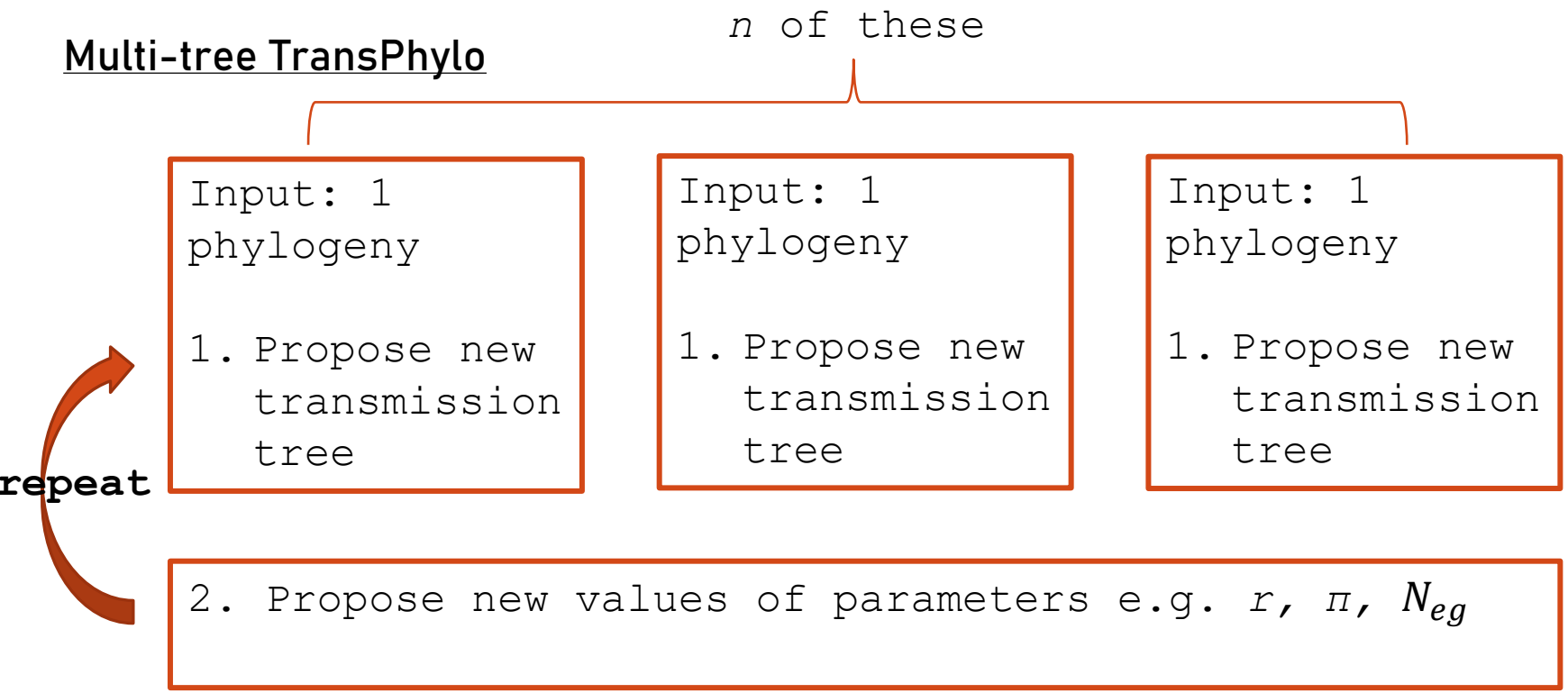
Recently, TransPhylo was extended to instead take a set of phylogenies as input. The transmission reconstruction is performed **on each of the input phylogenies, simultaneously, and with shared parameters across phylogenies.**

## ‘Base’ TransPhylo




Output: Collection of $x$ transmission trees

## Multi-tree TransPhylo




Output: Collection of *n*x* transmission trees

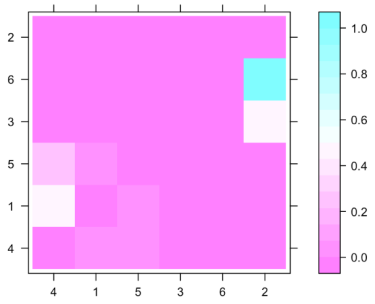
# Multi-tree capability

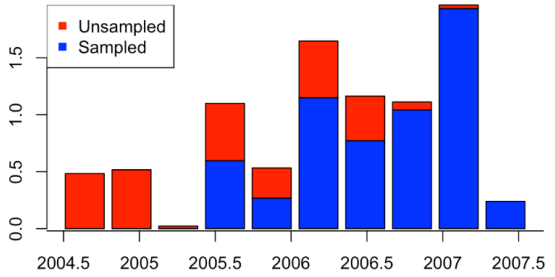
This allows us to **incorporate phylogenetic uncertainty**

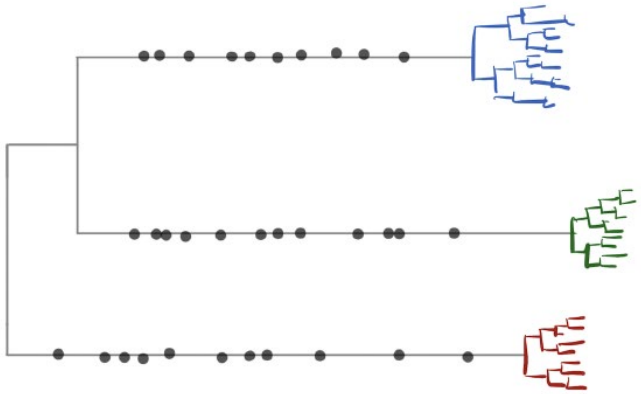
Multiple phylogenetic trees can be run from, for example, a posterior collection of trees generated in BEAST(2). Or, for multiple trees generated by different tree reconstruction methods or clock rate models, for example

The resulting transmission trees will have been generated from different phylogenetic trees, but will still include the same number of samples and so can be analysed and compared using the same methods we explored in exercise 4.

Alternately, we can use the multi-tree option to estimate transmission trees for multiple smaller clusters of a larger phylogeny – which we would expect to share parameters.



- We can use machine learning tools to compare inferred transmission links from TransPhylo analysis to additional covariate data, in order to predict possible indicators of infection/infectivity.

- We are also currently working on ways to use epidemiological data to constrain the space of possible transmission trees. For example, in nosocomial (hospital) or care settings, we might want to assume that an individual could only have been transmitting during their time there. Or, we could bias our transmission tree towards infections in line with contact tracing data.

# Linking additional epidemiological data to transmission

# Machine Learning with TransPhylo

*"We can use machine learning tools to compare inferred transmission links from TransPhylo analysis to additional covariate data, in order to predict possible indicators of infection/infectivity. "*

We explored this in the following paper:

preprint at:
https://www.biorxiv.org/content/10.1101/761411v1

Transmission analysis of a large TB outbreak in London: mathematical modelling study using genomic data

Yuanwei Xu[*1], Jessica E Stockdale[†2], Vijay Naidu[‡2], Hollie Hatherell[§3], James Stimson[¶1,4], Helen R. Stagg[‖5], Ibrahim Abubakar[**6], and Caroline Colijn[††1,2]

[1]Centre for Mathematics of Precision Healthcare, Department of Mathematics, Imperial College London
[2]Department of Mathematics, Simon Fraser University
[3]University College London
[4]National Infection Service, Public Health England
[5]Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh
[6]Institute for Global Health, University College London

## Tuberculosis outbreak in London, UK

We used genomes collected from a large TB outbreak in London 1995-2006. By 2013, 501 UK cases belonged to his outbreak.

TB transmission chains are notoriously difficult to uncover (due to e.g. long time frames, short contact required for transmission and transmission often be centered on hard-to-reach populations). Transmission was also attempted to be inferred directly from the whole genome sequence data, bur due to small amounts of variation in the genomes, this was unsuccessful.

WGS data + sampling time data + modelling offers us opportunity to uncover transmission links

# Sequencing described in:

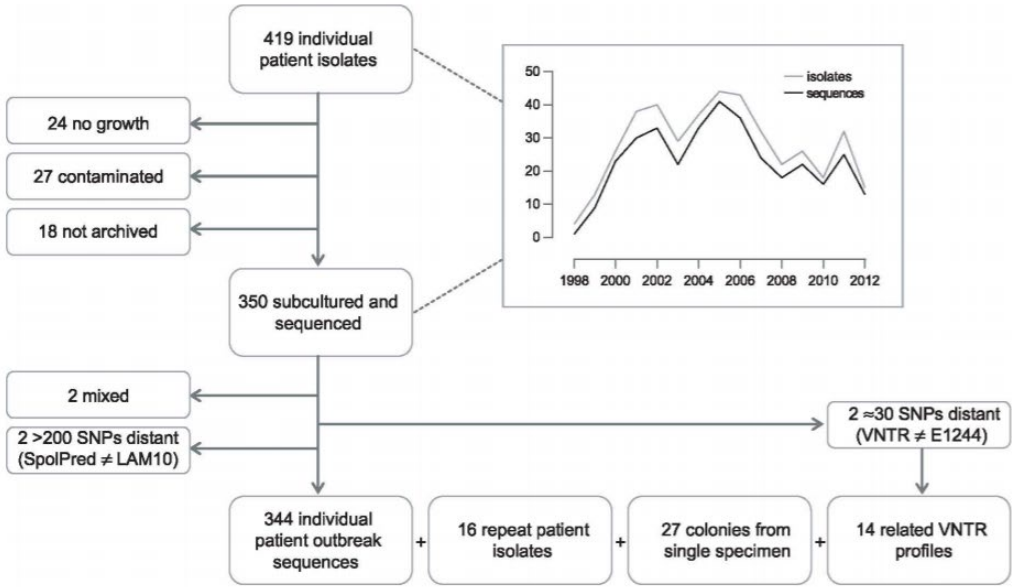
RESEARCH ARTICLE

## Whole Genome Sequence Analysis of a Large Isoniazid-Resistant Tuberculosis Outbreak in London: A Retrospective Observational Study

**Nicola Casali[1,2], Agnieszka Broda[1], Simon R. Harris[3], Julian Parkhill[3], Timothy Brown[4], Francis Drobniewski[1,4,5]***

**1** Department of Infectious Diseases and Immunity, Imperial College London, London, United Kingdom, **2** Centre for Immunology and Infectious Disease, Blizard Institute, Queen Mary University of London, London, United Kingdom, **3** Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom, **4** Public Health England National Mycobacterium Reference Laboratory, London, United Kingdom, **5** Departments of Microbiology and Respiratory Medicine, Barts Health NHS Trust, London, United Kingdom

* f.drobniewski@imperial.ac.uk



## Tuberculosis outbreak in London, UK

As well as sequencing the TB genomes, additional data was also collected via questionnaire and interview, including:

- age
- sex
- region of residence
- ethnicity
- country of birth
- occupation
- drug and alcohol use
- history of homelessness
- mental health concerns
- link to prison
- previous TB diagnosis

We use this data by training machine learning tools to predict which individuals were likely transmitters, using the covariate data alone

# A multi-stage process for analysis:

1. SNP calling and phylogenetic reconstruction
2. Transmission inference
3. Patient-level prediction from metadata

# 1. SNP calling and phylogenetic reconstruction

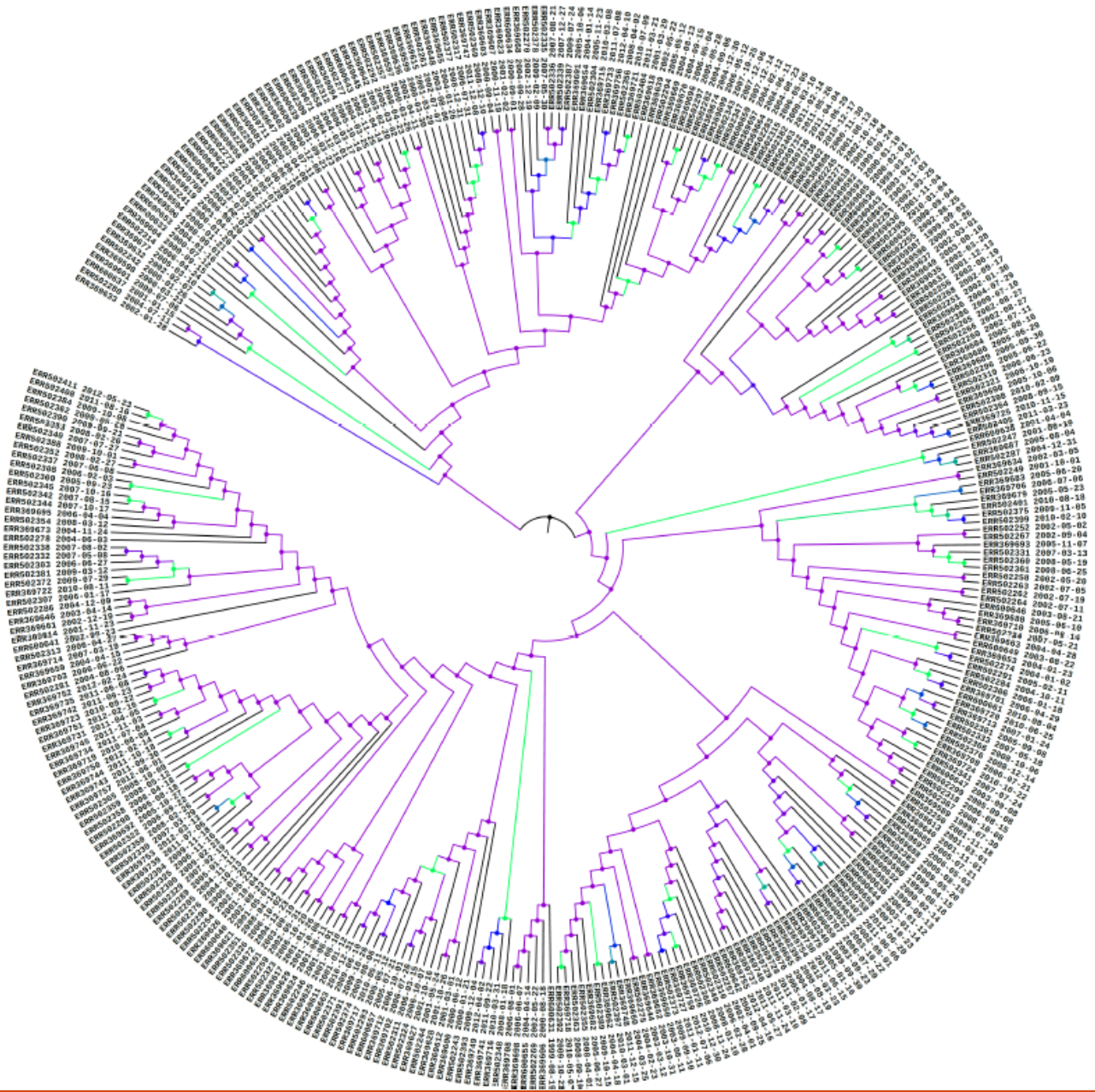
Created a variable-site alignment of 269 sites, for 329 genomes (after excluding low-quality samples and multiple samples from the same individual)

Used BEAST2 to build a collection of timed phylogenetic trees

Various checks and tools for making sure we produce 'good' trees:

- **TempEst** verifies temporal signal i.e. a positive correlation between between genetic distance and sampling time
- **Model testing** calculates BIC scores under different tree-generating models
- **Verify chain convergence** by confirming multiple independent MCMC runs starting from different locations end up at the same place
- **Verify good mixing** of the MCMC, by visual inspection of trace plots and calculation of the ESS
- **Treespace** analysis to compare posterior trees, revealing no multi-modality

**Result:** 9,000 phylogenetic trees – of which we randomly sampled 50



## 2. Transmission inference

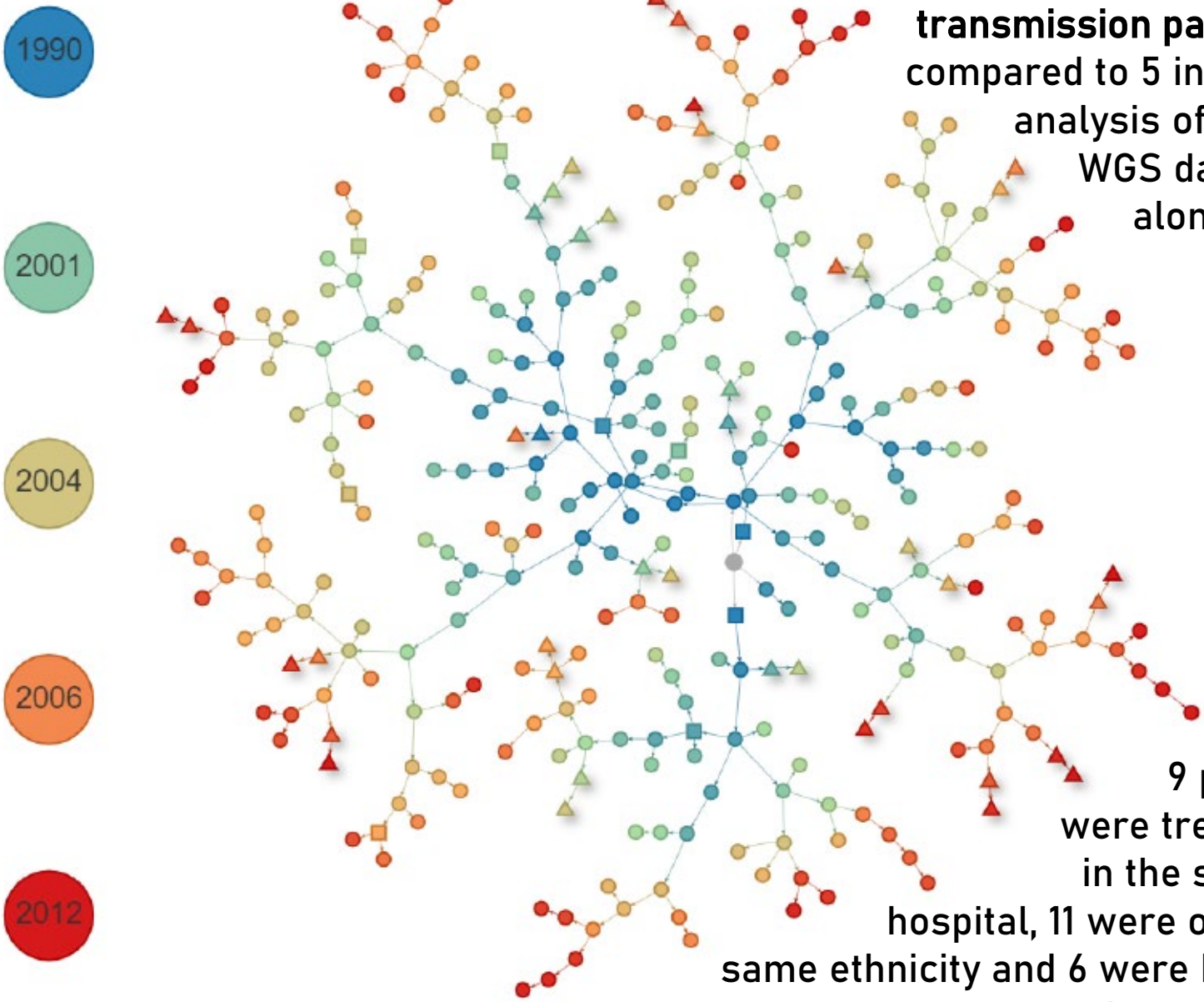
Use multi-tree TransPhylo to simultaneously infer transmission across all 50 BEAST2 trees – to capture phylogenetic uncertainty, whilst sharing parameters for better MCMC mixing.

Generation time ~ Gamma(shape 1.3, scale 2.5)
Sampling time ~ Gamma(shape 1.1, scale 6)

Estimated $R_0$ and sampling fraction π – with a strong Beta(mean = 0.8) prior to reflect our belief that case finding was effective in the UK at the time

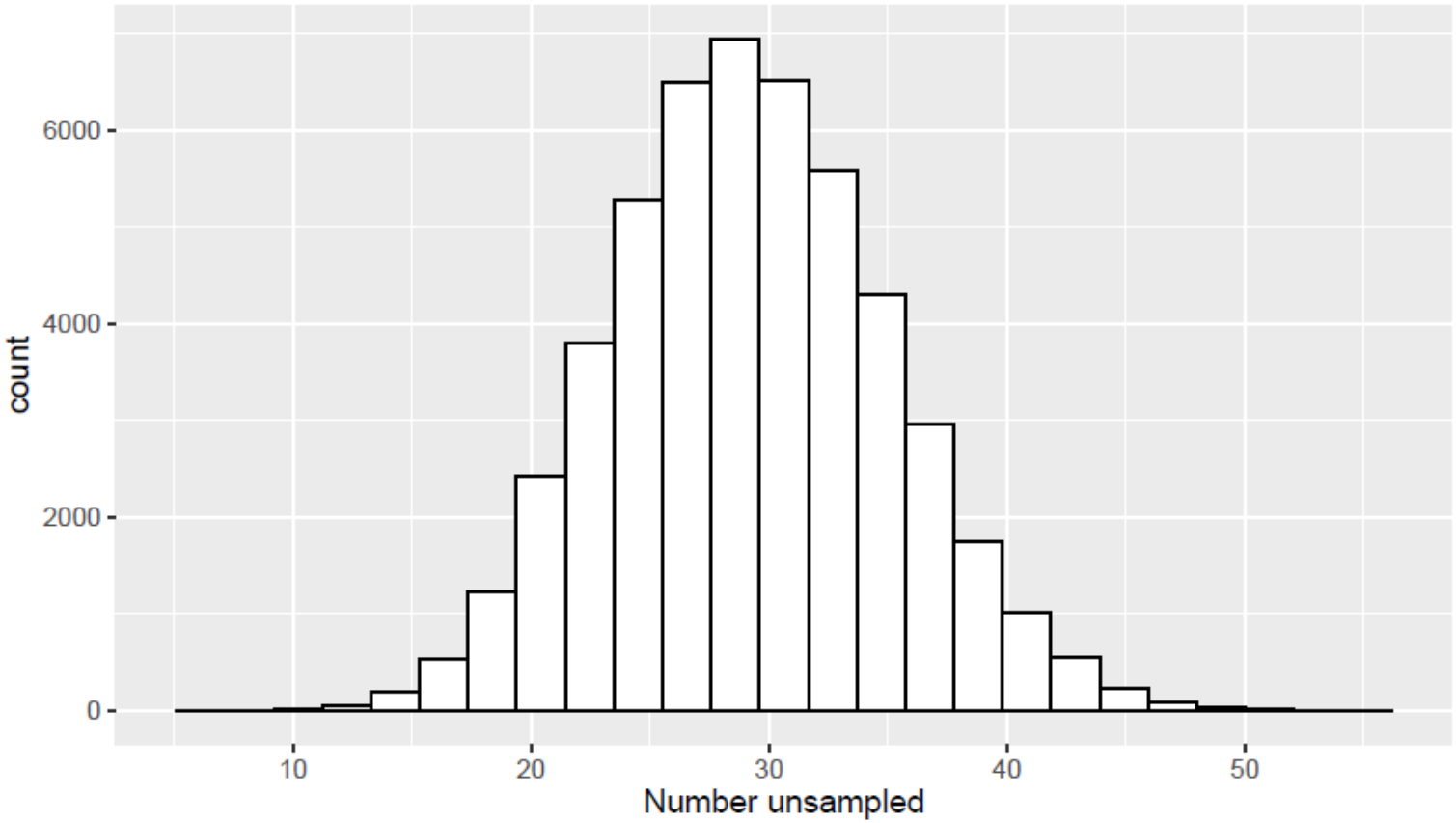
**Result:** After 10^5 iterations, we collect 1,000 transmission trees per phylogenetic tree, for a total of **50,000 transmission trees**

# MAP (maximum a posteriori) tree

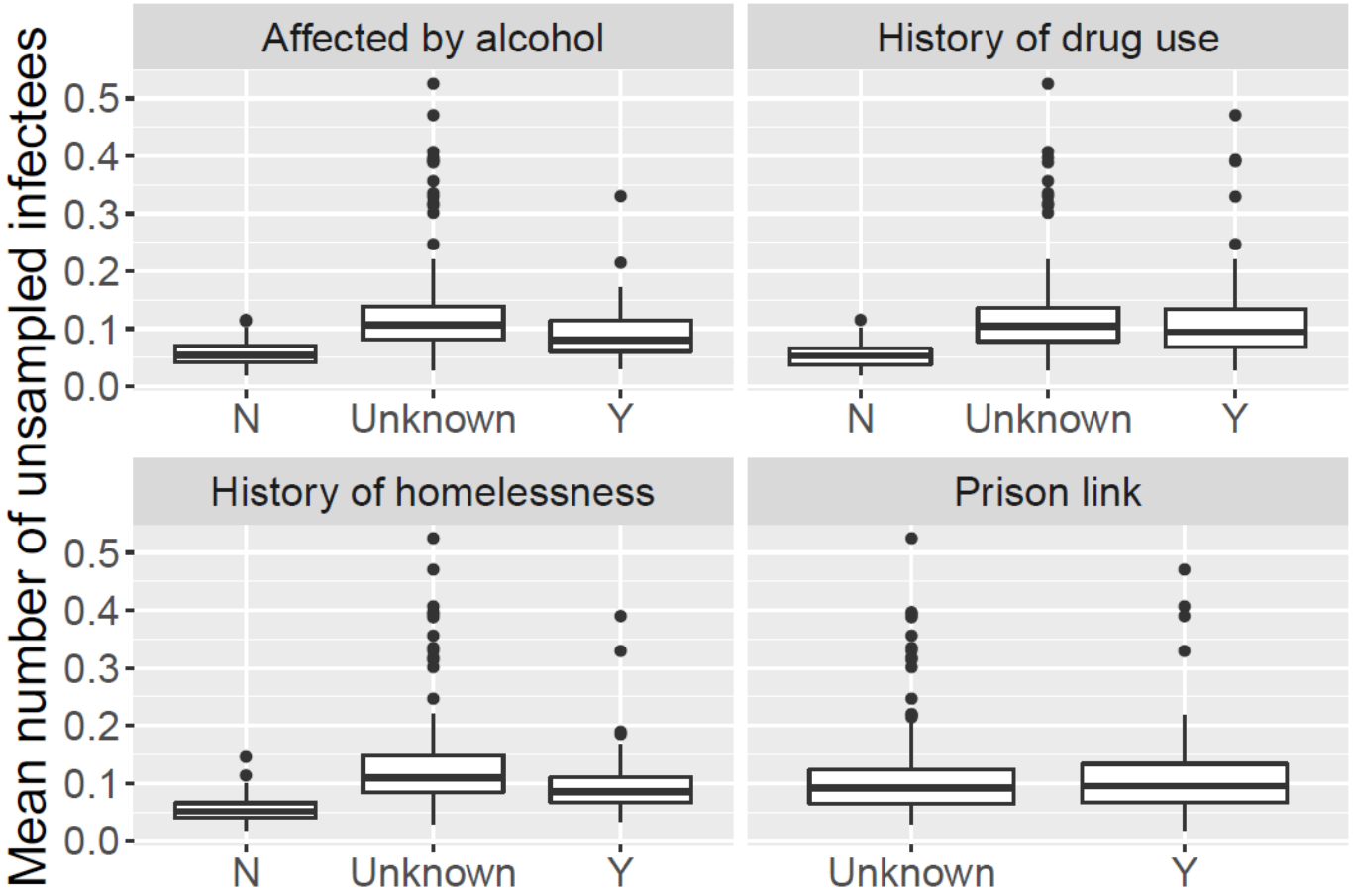



We found **21 credible transmission pairs** – compared to 5 in the analysis of the WGS data alone

9 pairs were treated in the same hospital, 11 were of the same ethnicity and 6 were both drug users




# How many unsampled cases were there?




# Who infects unsampled cases?

# 3. Patient-level prediction from metadata

Assume posterior TransPhylo trees = 'ground truth'

We assume that a credible transmitter is a sampled individual who was estimated to infect others in at least 50% of the posterior trees – we found this to be **62%** of cases.

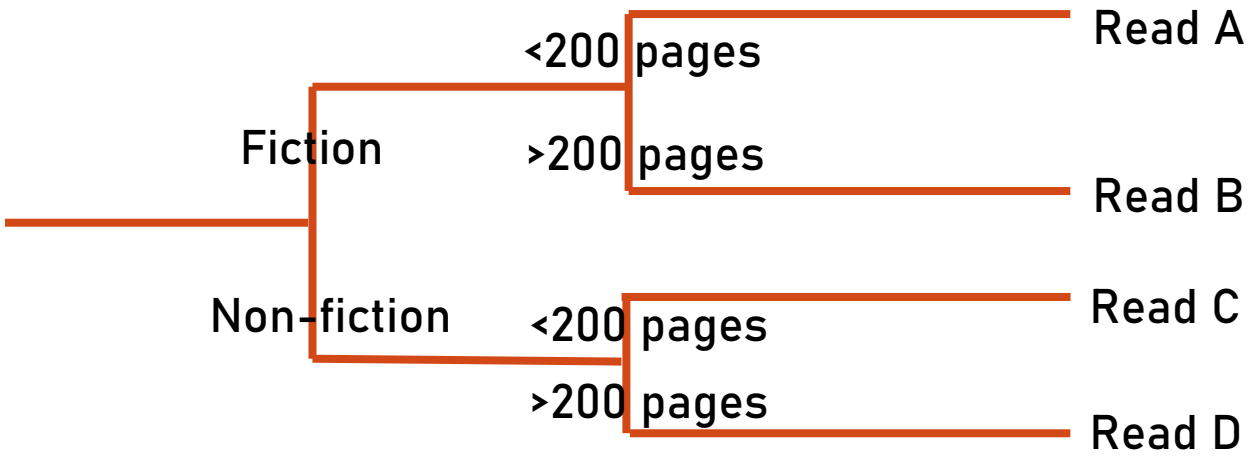
then

**Use a Random Forest classifier to use the metadata to predict whether a case is a credible TB transmitter**

- age
- sex
- region of residence
- ethnicity
- country of birth
- occupation
- drug and alcohol use
- history of homelessness
- mental health concerns
- link to prison
- previous TB diagnosis

# A 1-minute intro to Random Forests

A supervised machine learning technique that fits decision tree classifiers on sub-samples of a dataset

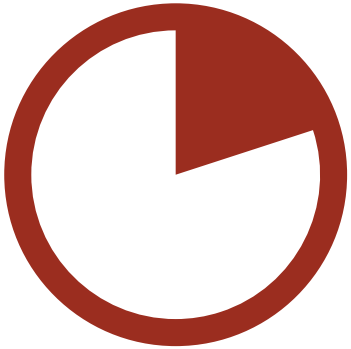



Builds many decision trees by randomly subsampling from the available questions, and then picking the questions that best divide the data. This gives us a collection of X decision trees, which we average over to find the best one.

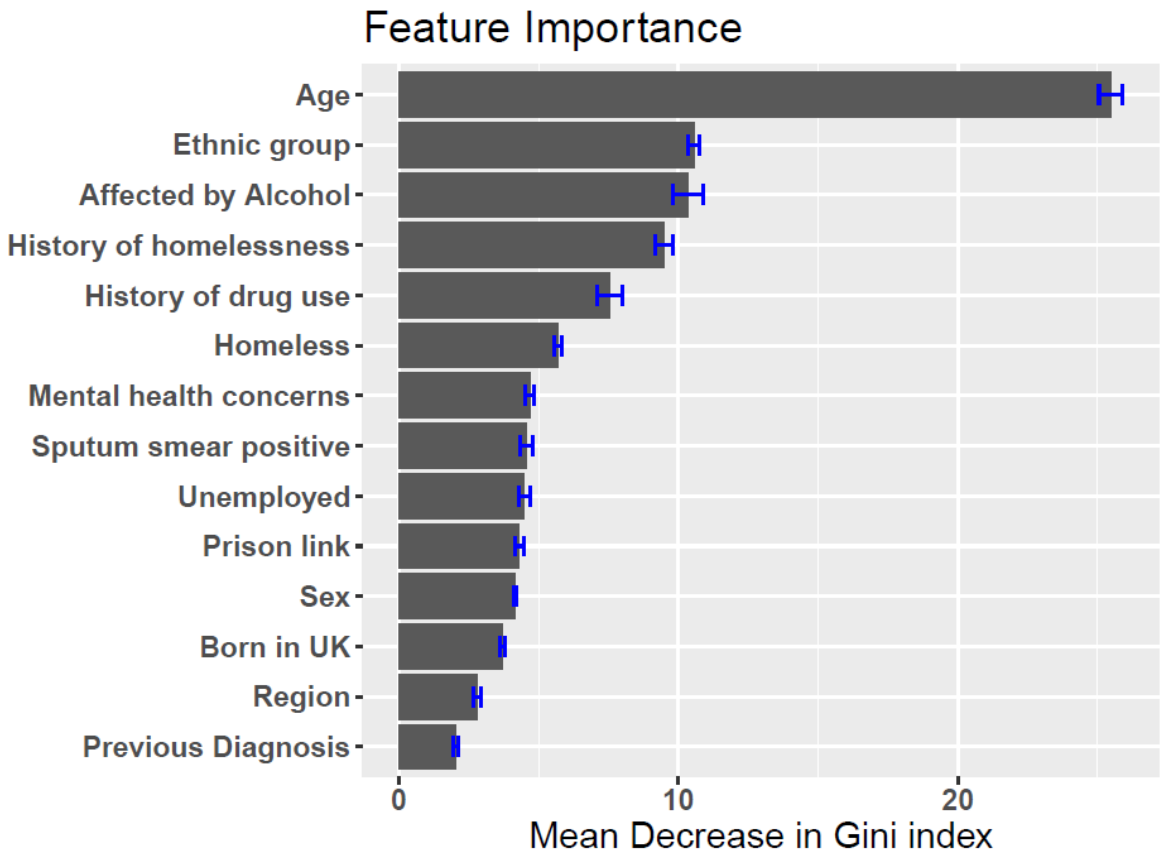
# 3. Patient-level prediction from metadata

**Use a Random Forest classifier to use the metadata to predict whether a case is a credible TB transmitter**

We used 5-fold cross validation to ensure the model is robust/ minimise the effect of outliers

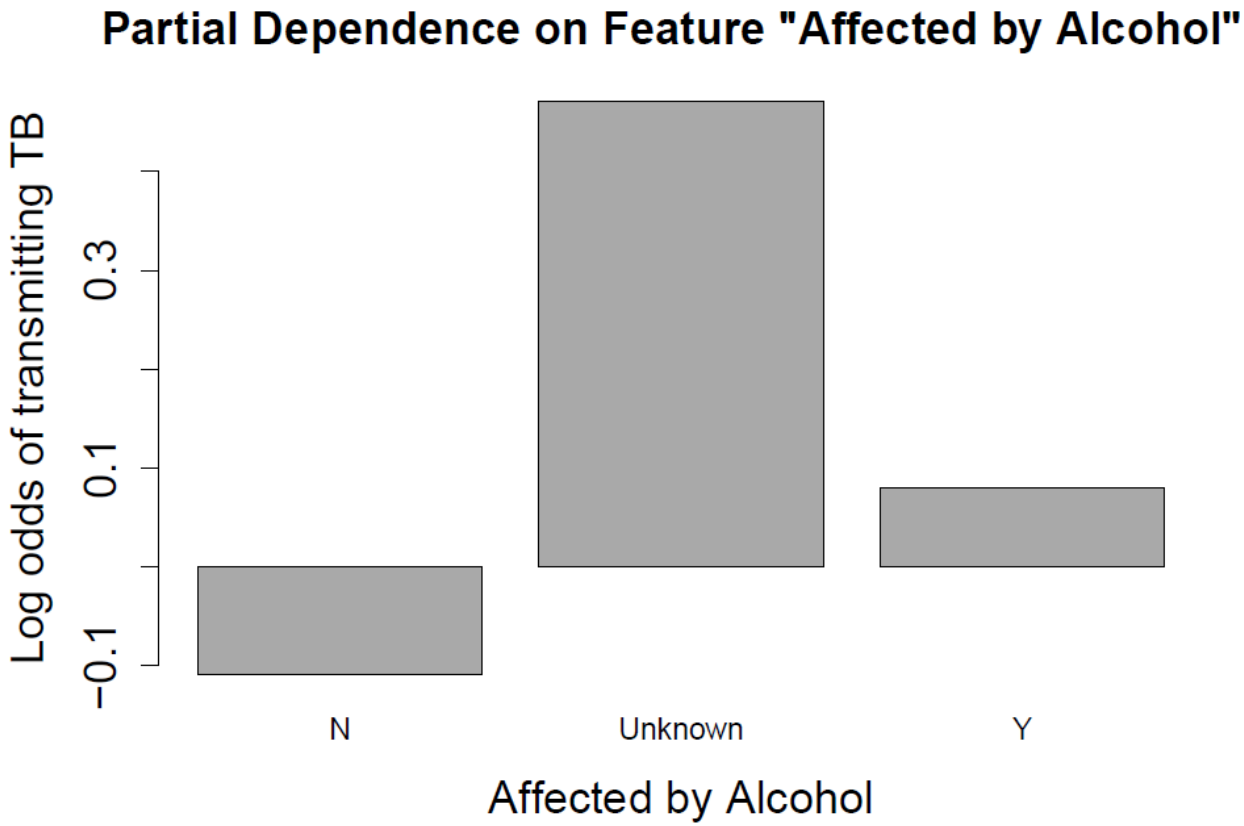


- Split the data into 5 chunks
- Remove 1 chunk, and train the RF on the remaining 4
- Test the RF on the chunk you removed
- REPEAT for the other 4 chunks

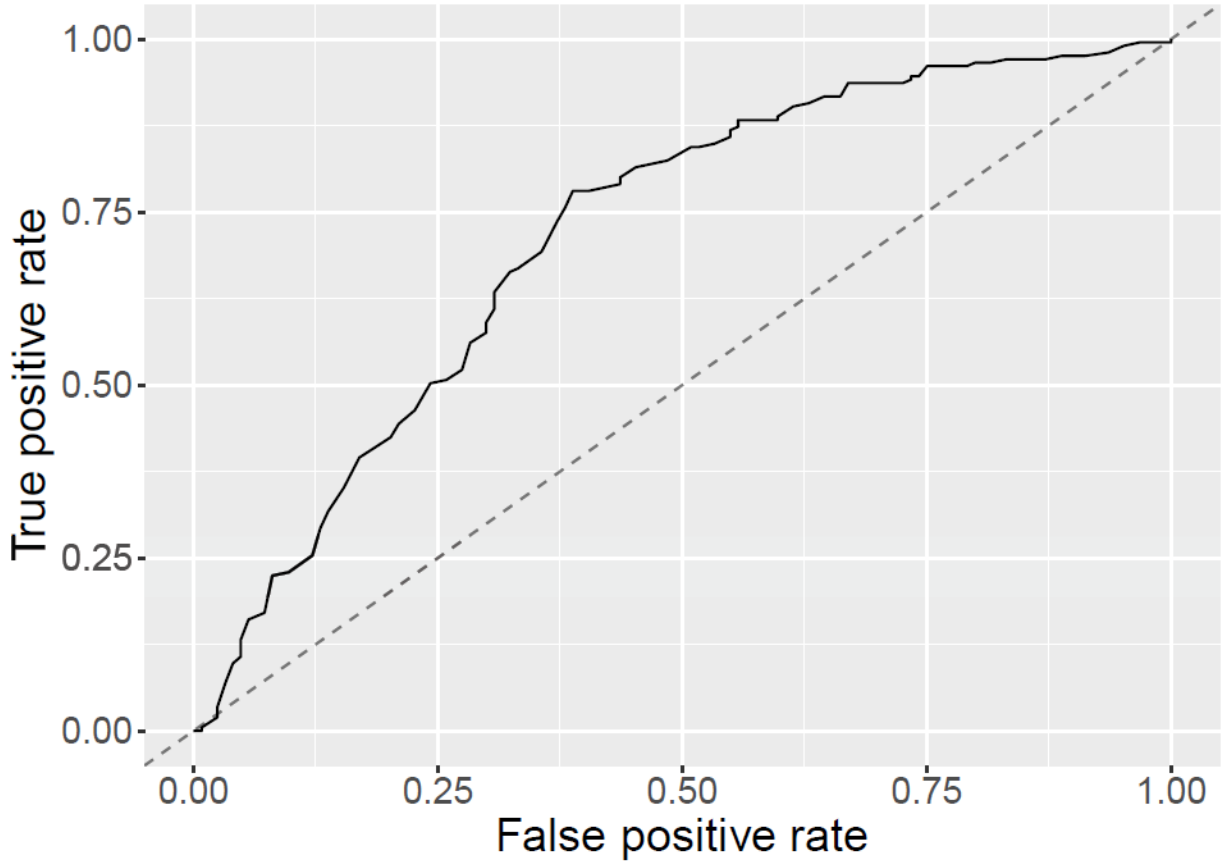



"The mean decrease in Gini coefficient is a measure of how each variable contributes to the homogeneity of the nodes"

# How important are the different metadata features?




# How important are the different metadata features?




AUC
0.72

# How well is the classifier performing? (ROC curve)

# Limitations

- All the machine learning based on inferred TP trees – these are not really ground truth
- We got a relatively high false positive rate from the Random Forest (0.28) – however, since TB prevalence is low this may be acceptable.
- Although we tried to take uncertainty into account at each stage, our 3-stage approach still loses some nuance at each step. Could we do this more directly?