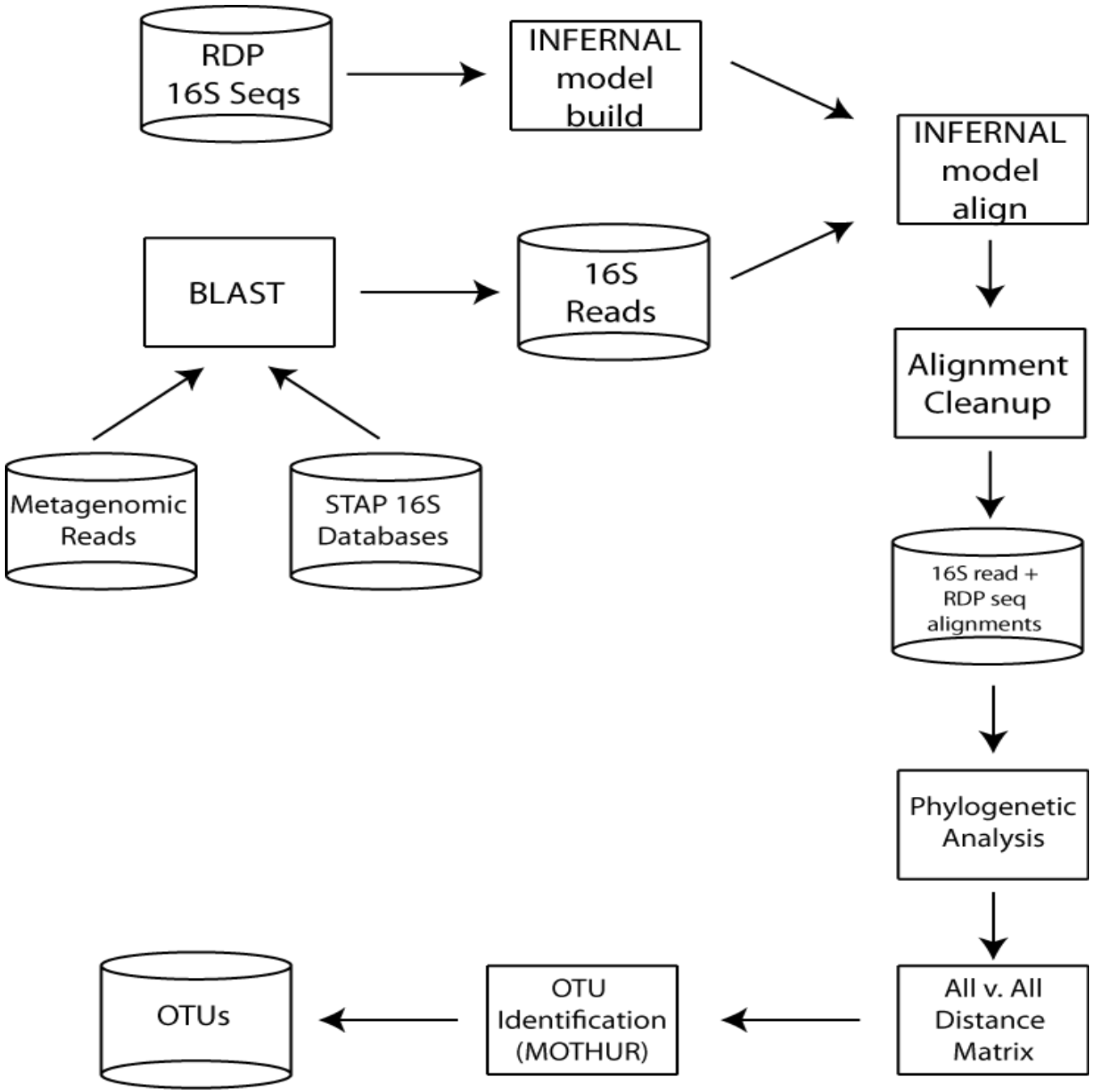
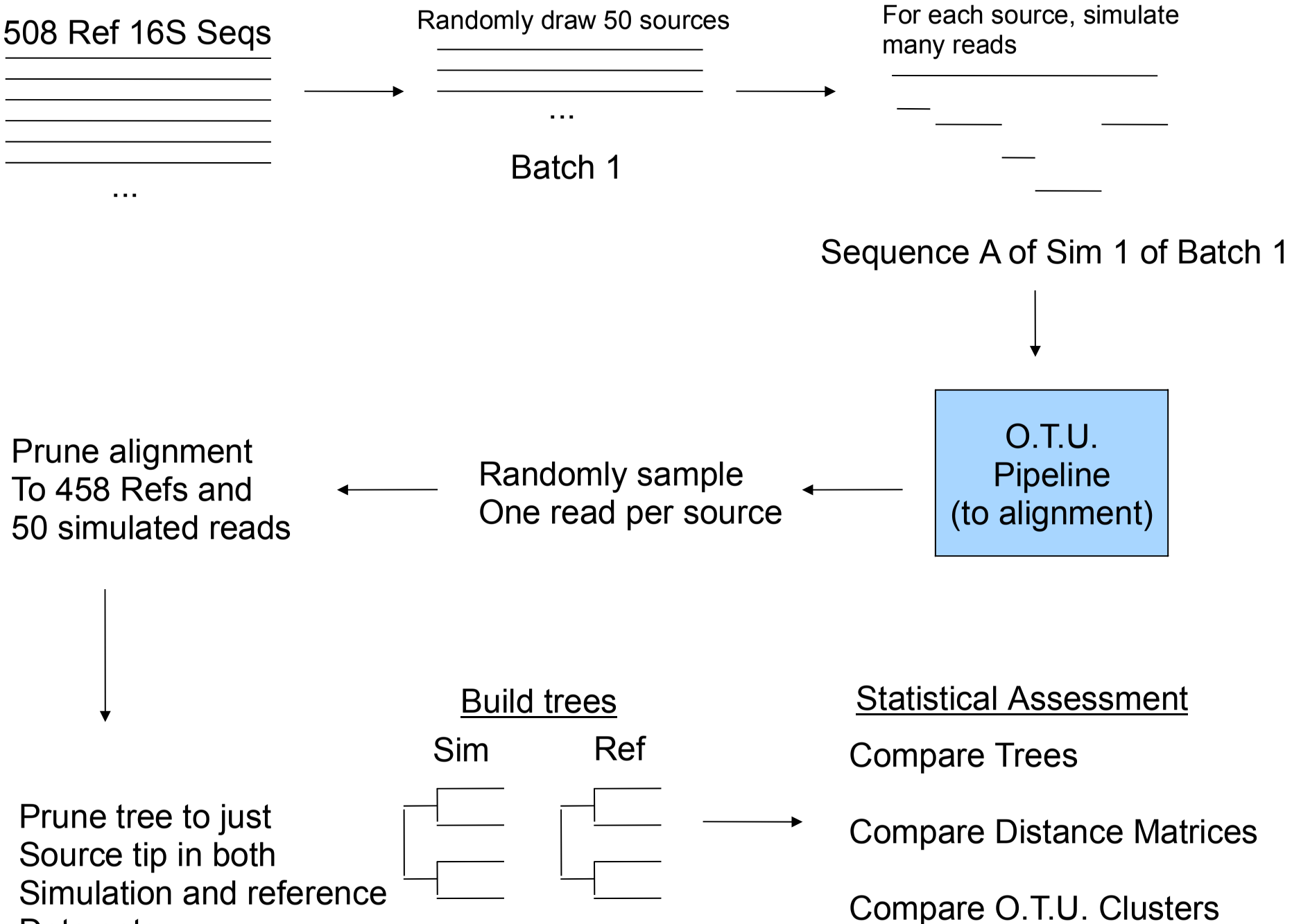
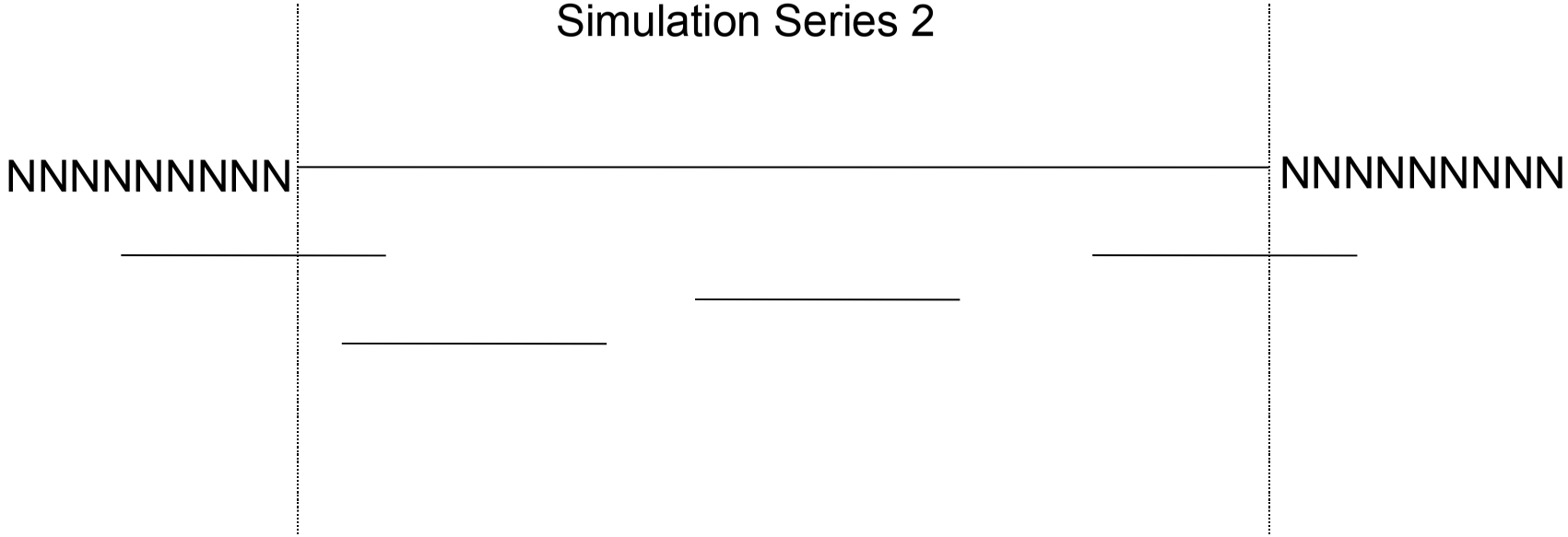

RDP
16S Seqs
INFERNAL
model
build
INFERNAL
model
align
BLAST
16S
Reads
Alignment
Cleanup
Metagenomic
Reads
STAP 16S
Databases
16S read +
RDP seq
alignments
Phylogenetic
Analysis
All v. All
Distance
Matrix
OTU
Identification
(MOTHUR)
OTUs

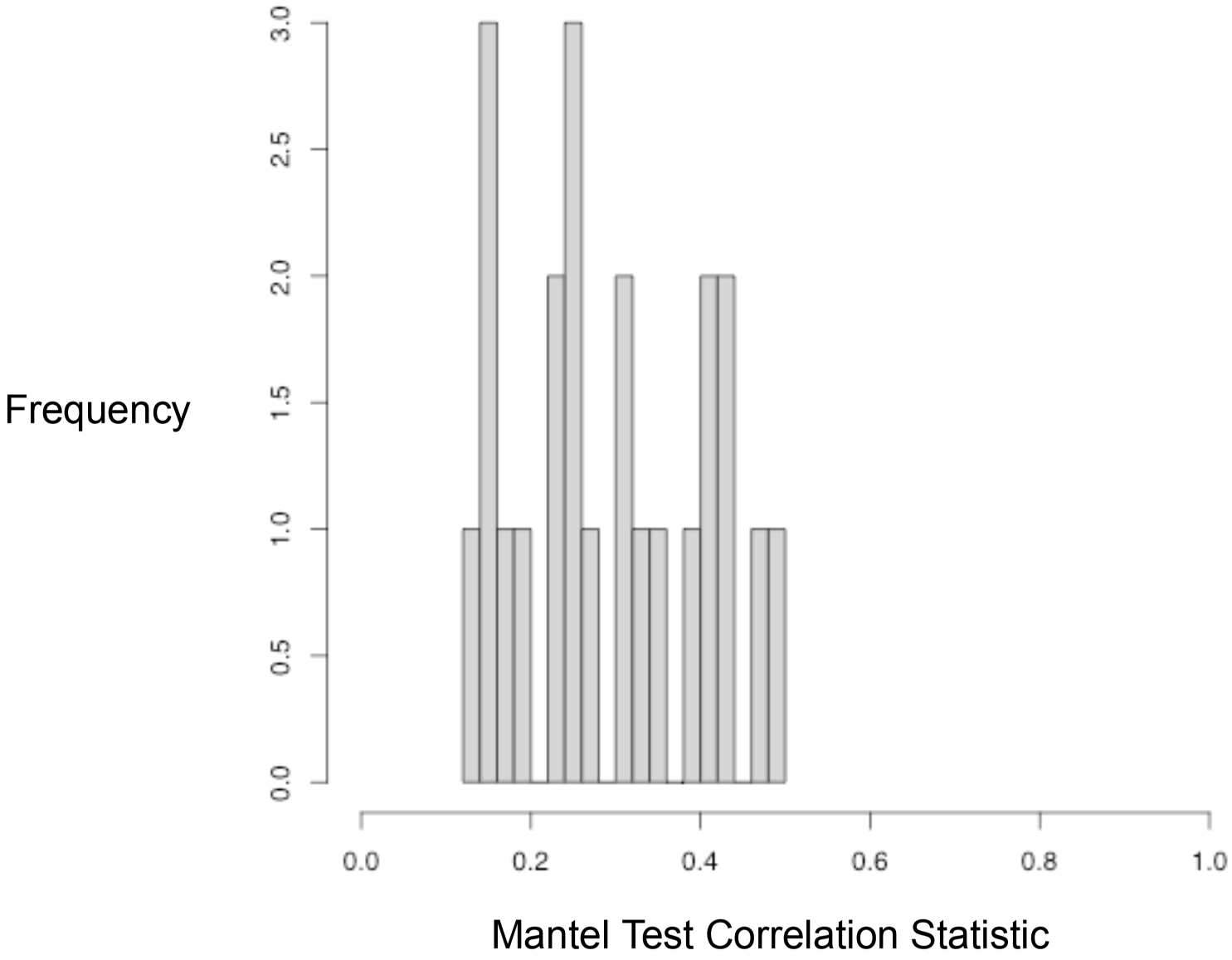

508 Ref 16S Seqs
...
Randomly draw 50 sources
...
Batch 1
For each source, simulate many reads
Sequence A of Sim 1 of Batch 1
O.T.U. Pipeline (to alignment)
Randomly sample One read per source
Prune alignment To 458 Refs and 50 simulated reads
Prune tree to just Source tip in both Simulation and reference Data sets
Build trees
Sim
Ref
Statistical Assessment
Compare Trees
Compare Distance Matrices
Compare O.T.U. Clusters

Simulation Series 1

Simulation Series 2

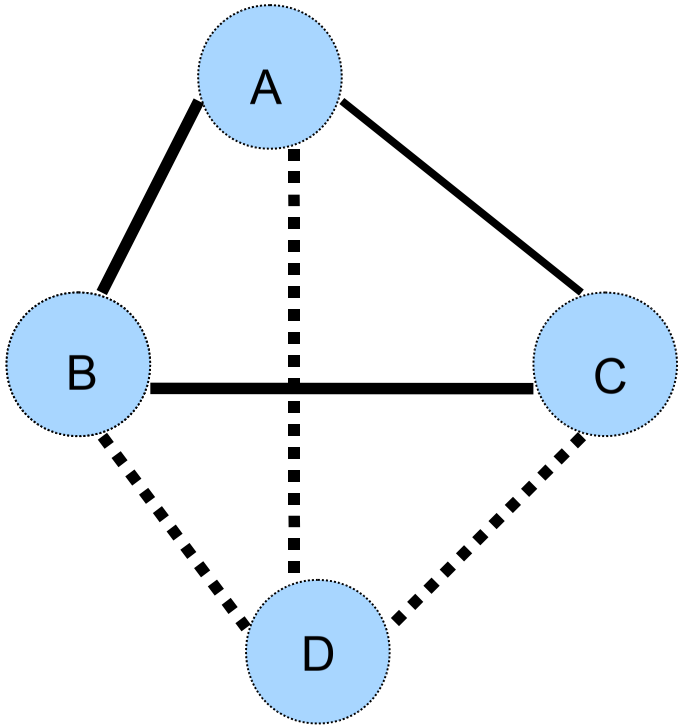


Matrix mantel tests show a significant positive correlation between Series 1 simulated distances and source distances

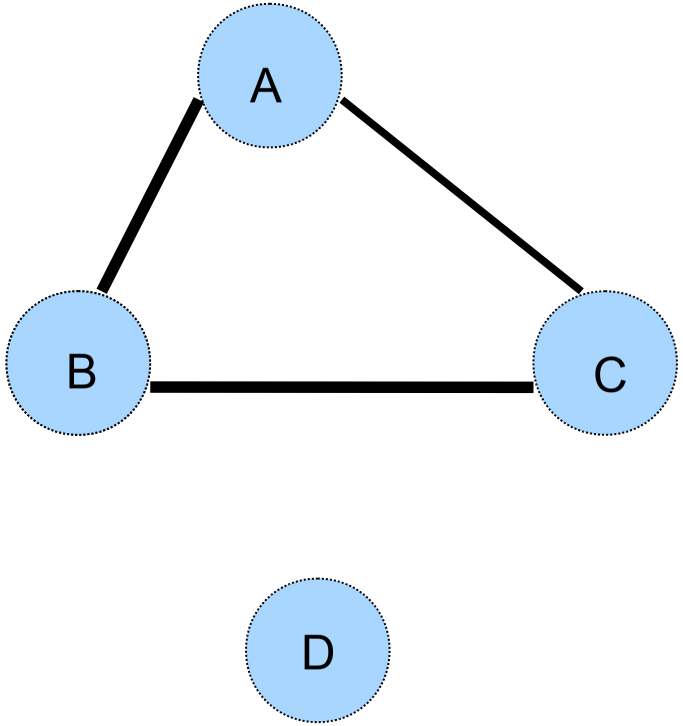


Error Rate Calculations

Sample (Simulation) Clusters



Reference (Full Length) Clusters



True Positive: Number of correctly called Sample Edges
True Negative: Number of correctly called Sample Cuts
False Positive: Number of incorrectly called Sample Edges
False Negative: Number of incorrectly called Sample Cuts

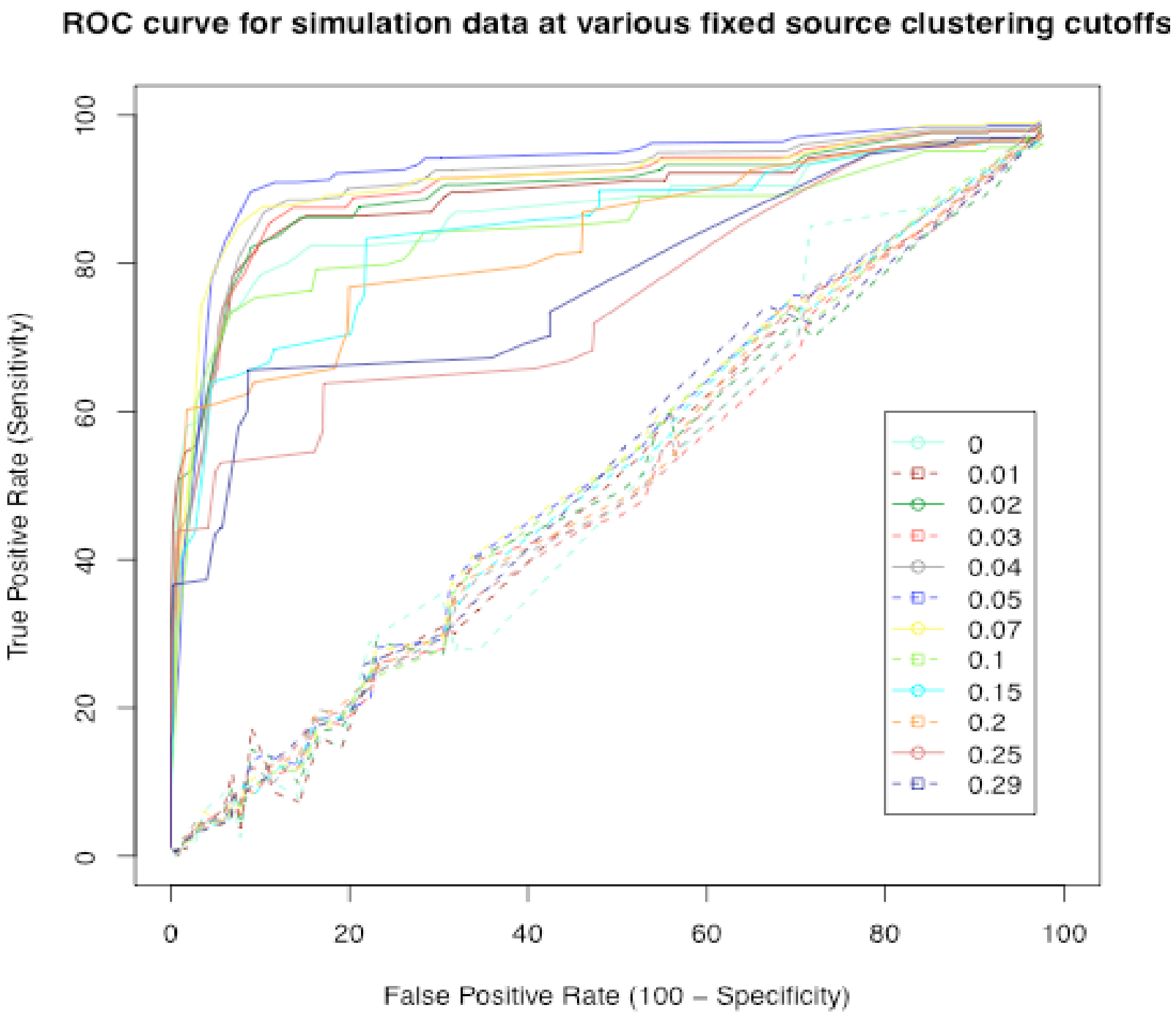
True Positive Rate (Sensitivity): TP / (TP + FN)

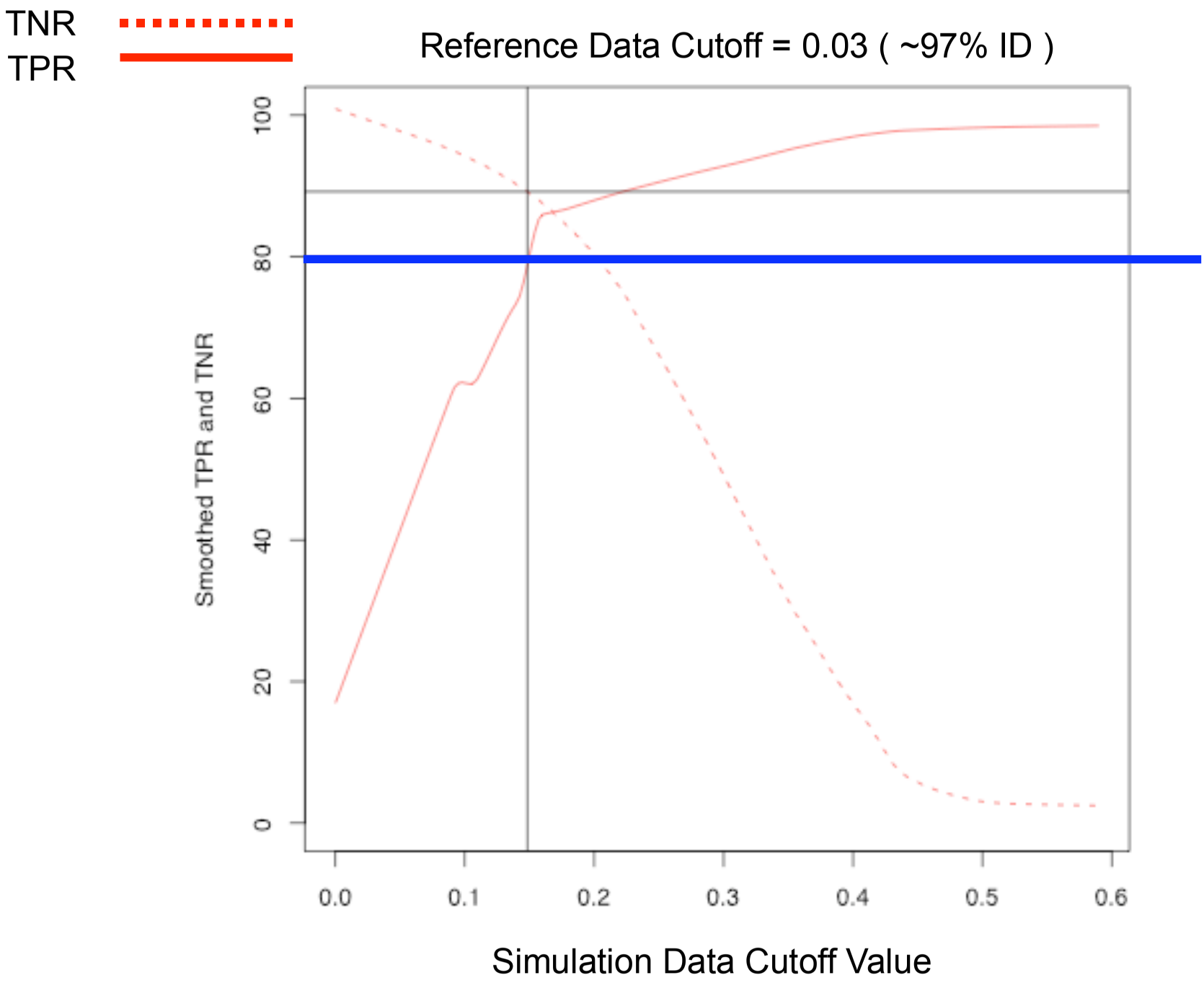
True Negative Rate (Specificity): TN / (TN + FP)

False Positive Rate: 1 - TPR

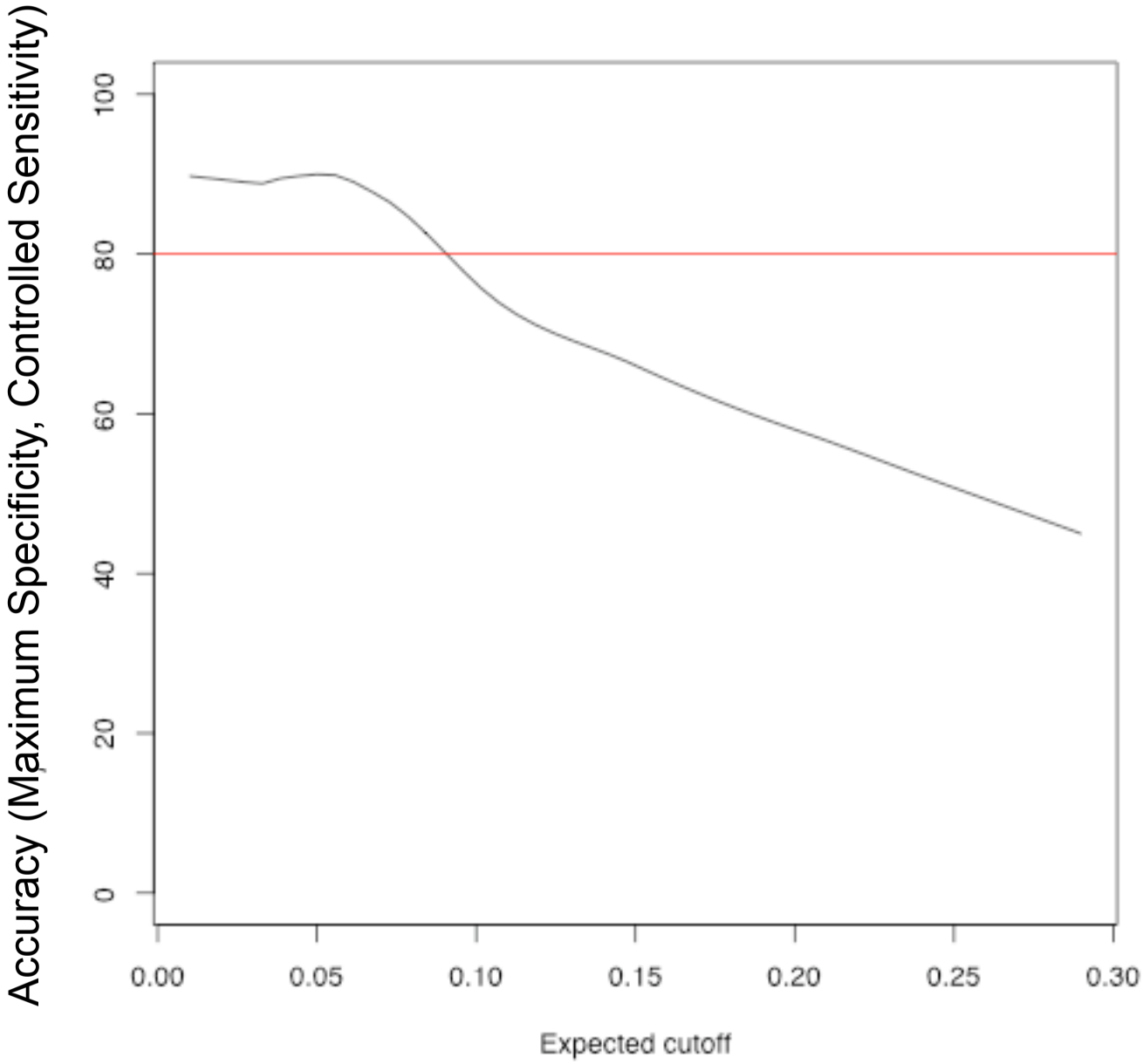
False Negative Rate: 1 - TNR

Considering replacing these terms with “True Clustering Rate” and “True Cutting Rate”

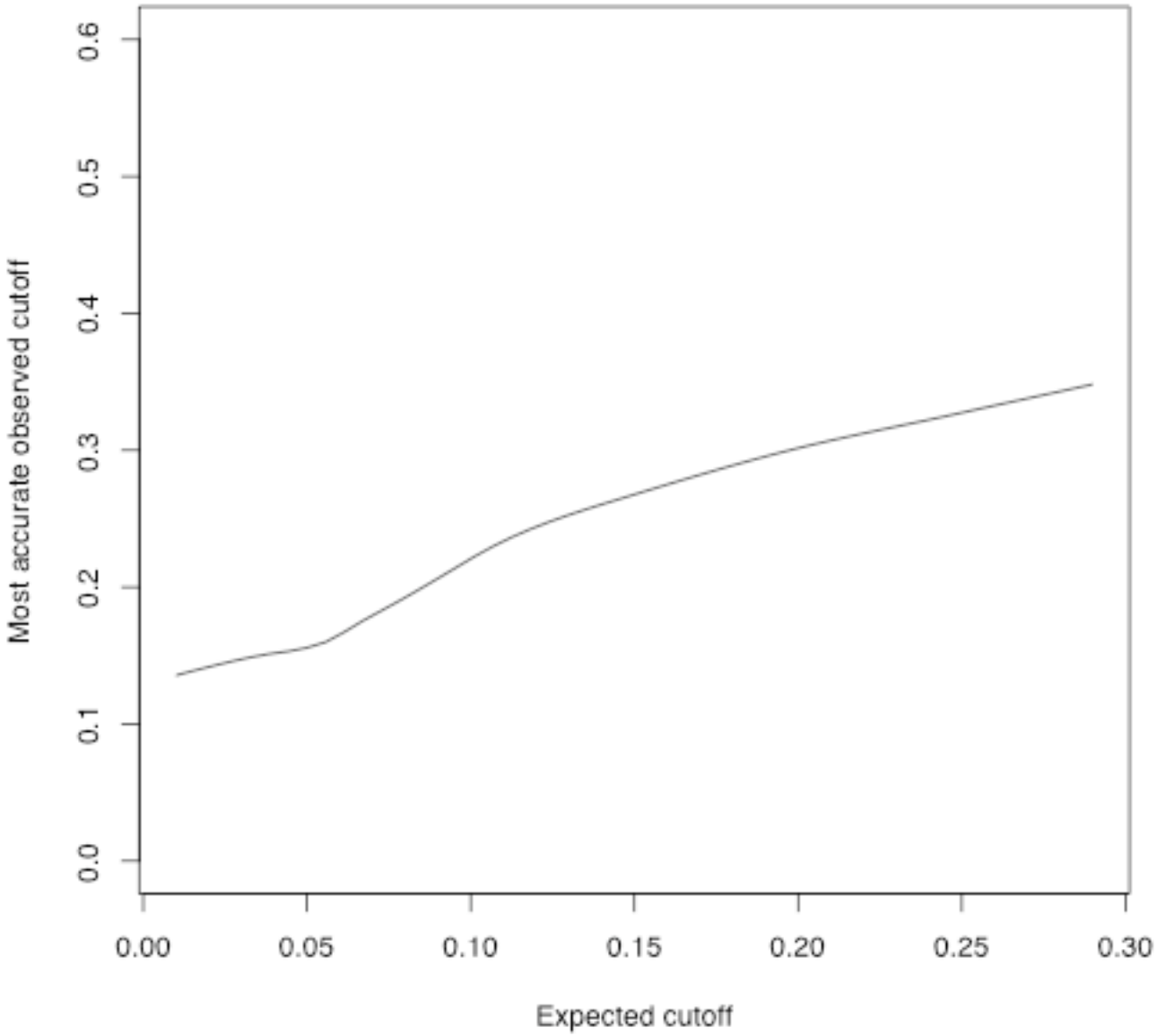

ROC curve for simulation data at various fixed source clustering cutoffs
True Positive Rate (Sensitivity)
False Positive Rate (100 – Specificity)
0
20
40
60
80
100
0
0.01
0.02
0.03
0.04
0.05
0.07
0.1
0.15
0.2
0.25
0.29

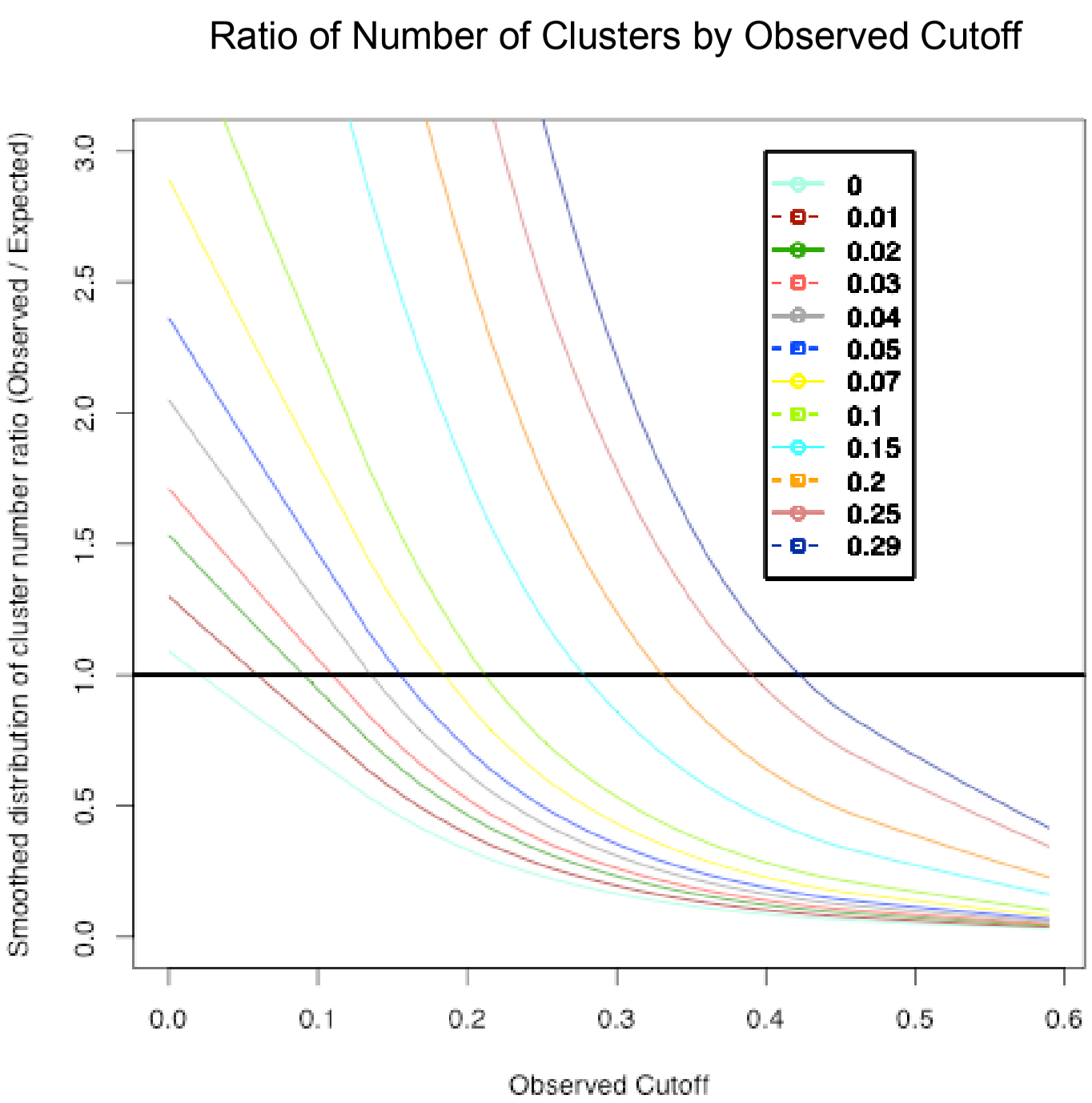

TNR
TPR
Reference Data Cutoff = 0.03 (~97% ID)
Smoothed TPR and TNR
0
20
40
60
80
100
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Simulation Data Cutoff Value

Accuracy Falls off as Expected Cutoff Increases



Clustering Cutoff Scaling Adjustment Plot (Max Accuracy)




Ratio of Number of Clusters by Observed Cutoff
Smoothed distribution of cluster number ratio (Observed / Expected)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Observed Cutoff
0
0.01
0.02
0.03
0.04
0.05
0.07
0.1
0.15
0.2
0.25
0.29

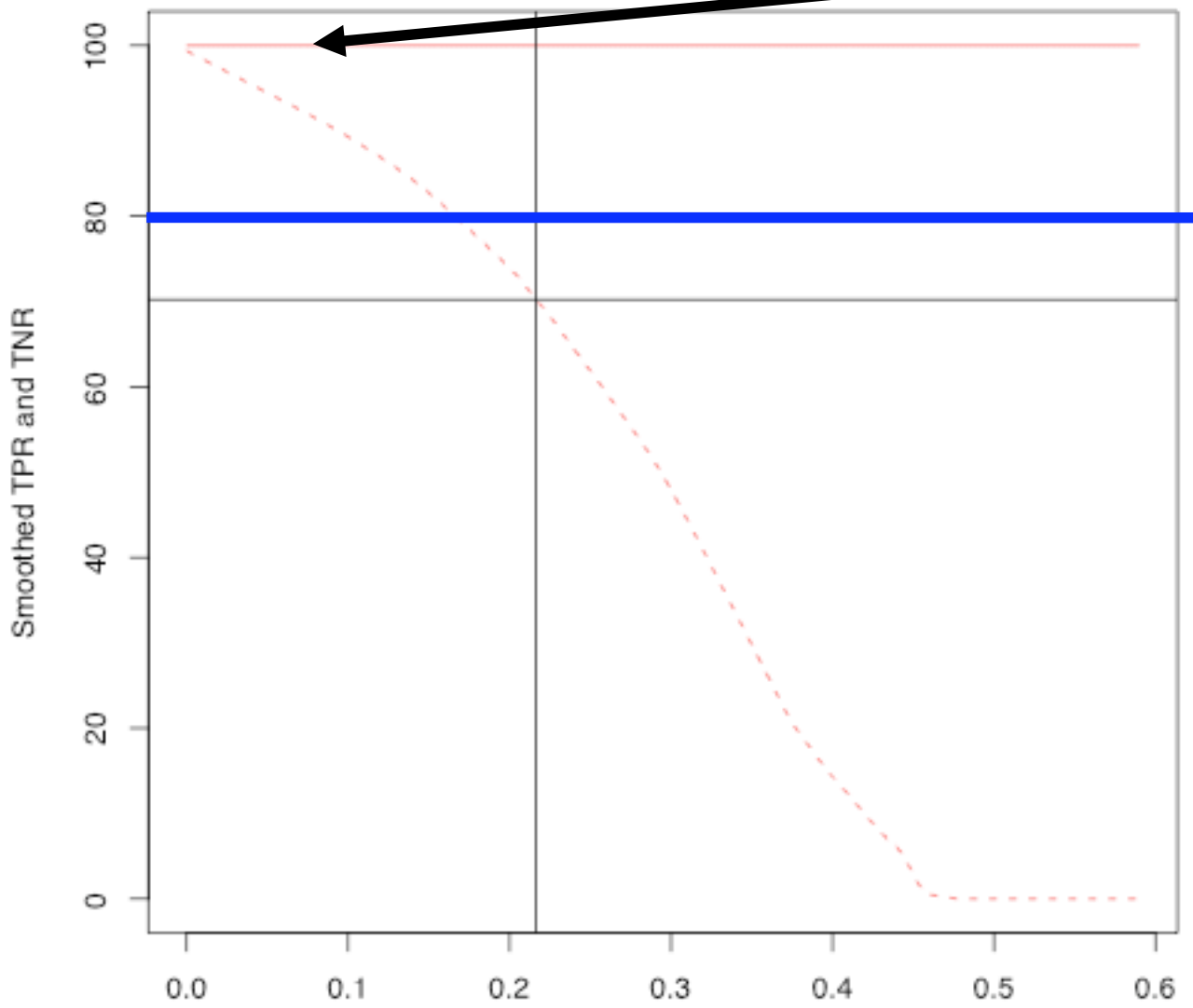
TNR

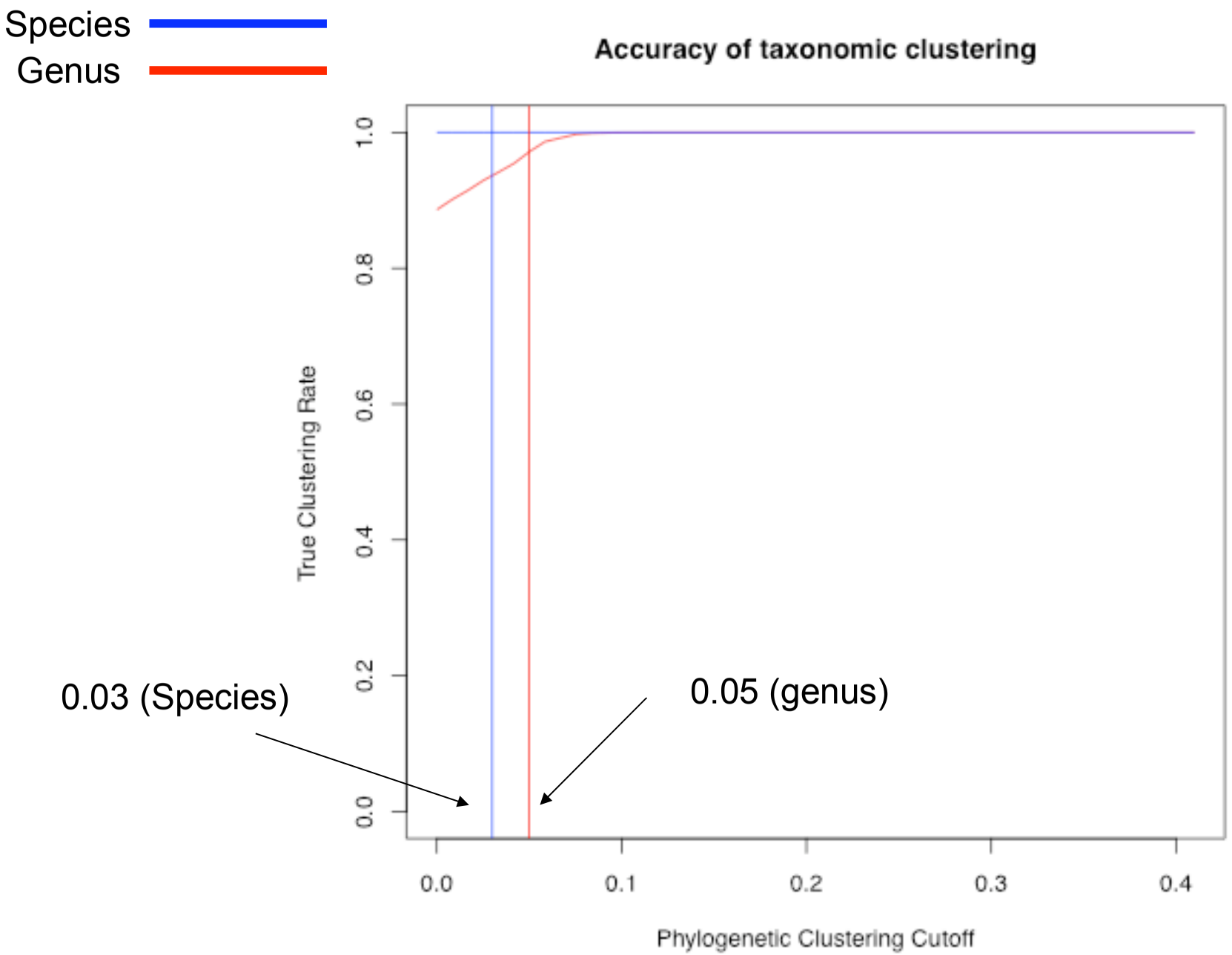
TPR

Phylogenetic distance v. Percent ID

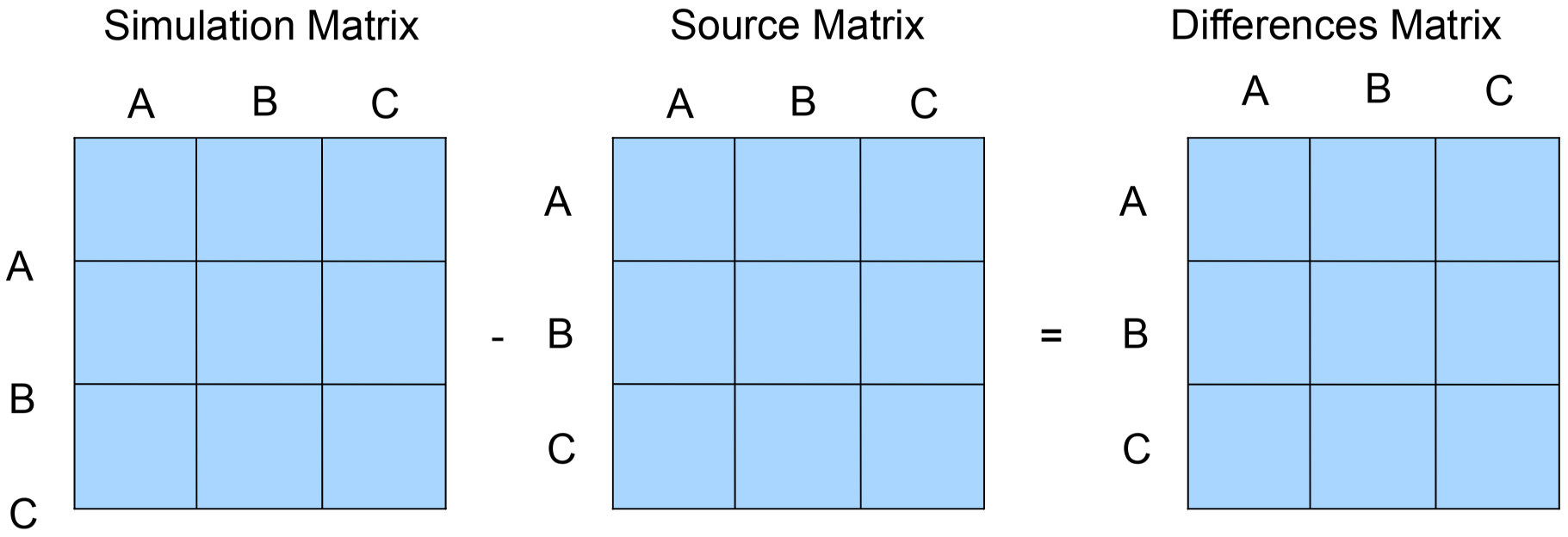
Reference Data Cutoff = 0.03 (~97% ID)

This happens because we underestimate biodiversity



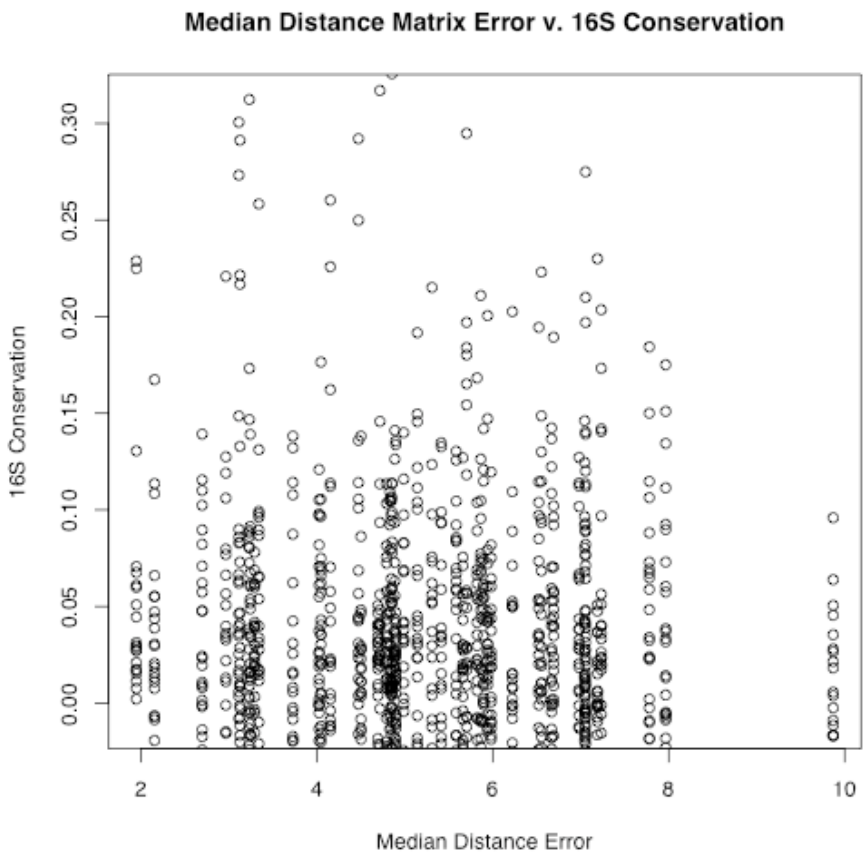

Species
Genus
Accuracy of taxonomic clustering
True Clustering Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.1
0.2
0.3
0.4
Phylogenetic Clustering Cutoff
0.03 (Species)
0.05 (genus)

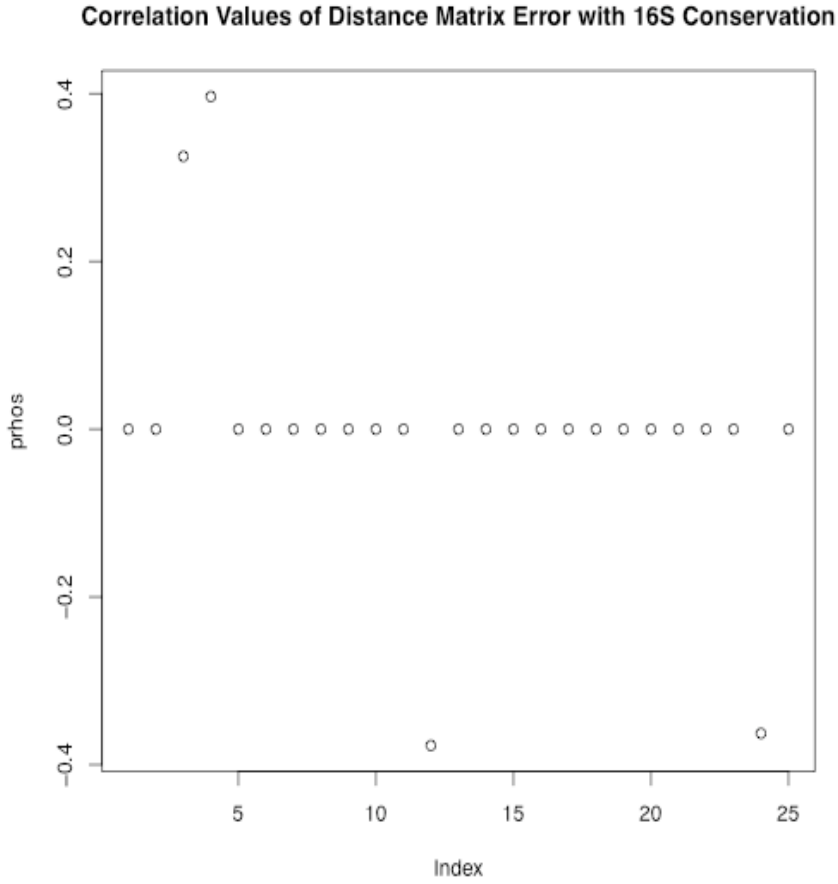
How we quantified per read error



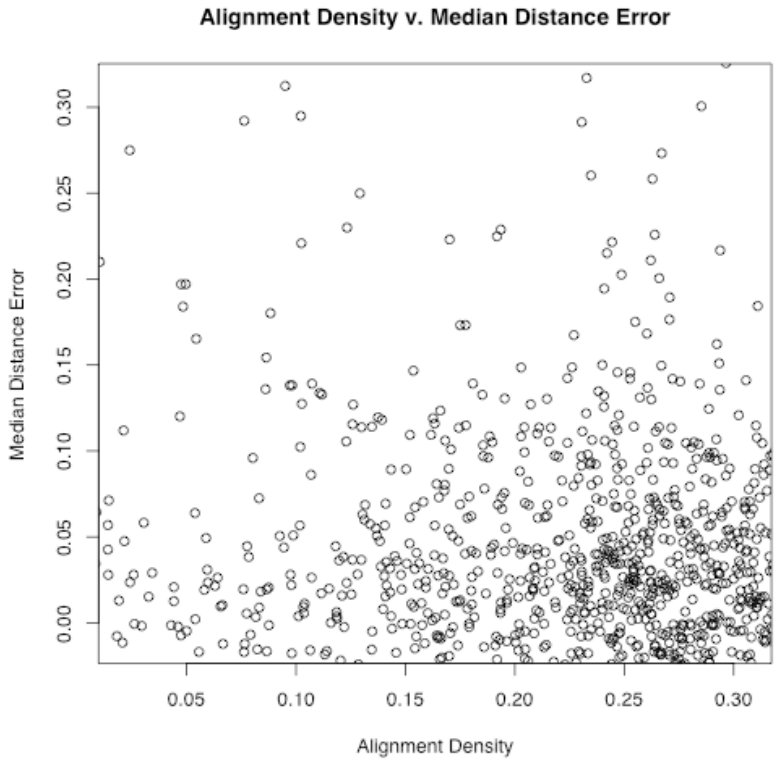
L1: The median value of the difference in distances
L2: A read's contribution to weighted path difference

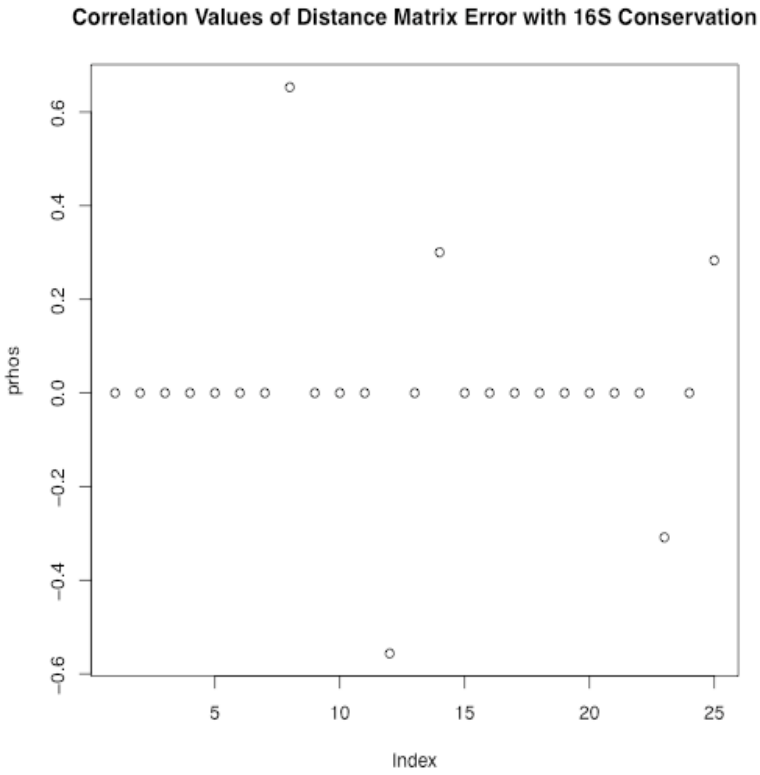
Median Distance Error V. 16S Evolutionary Rate





Alignment Density V. Median Distance Error

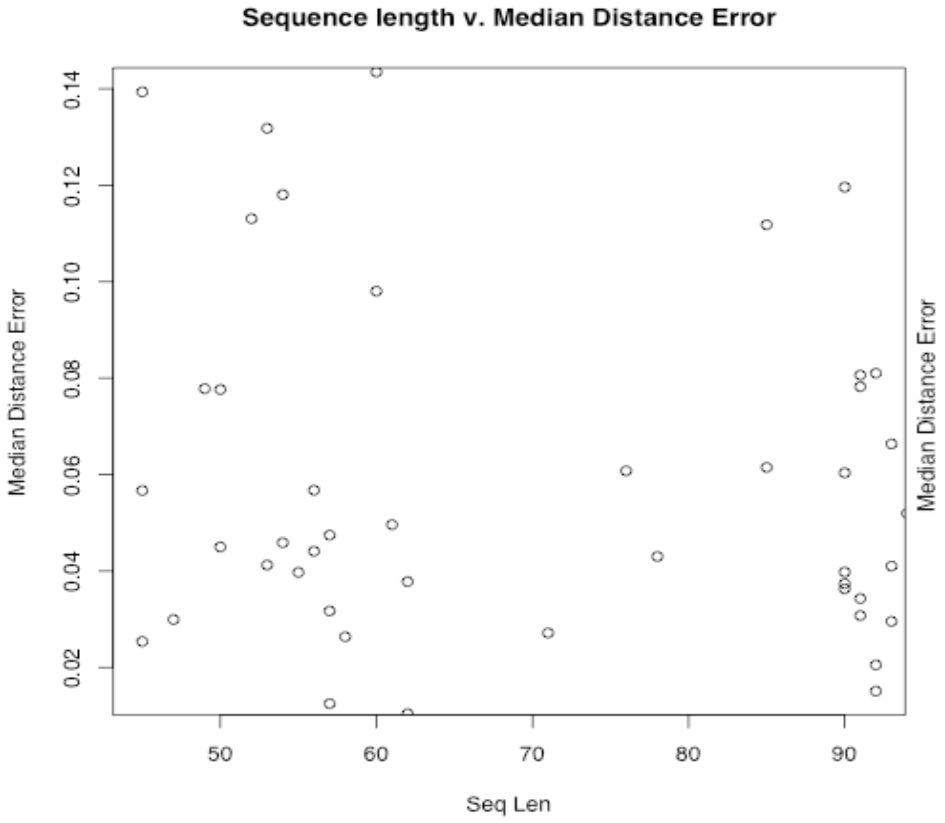




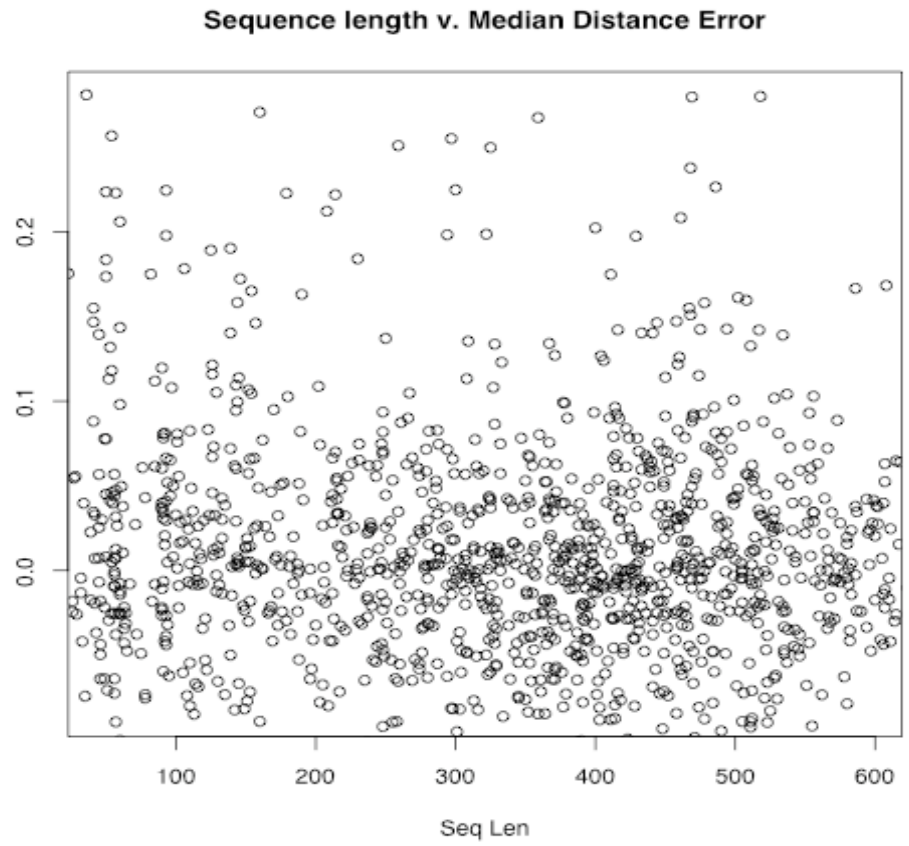
Read length bias?

Median Distance Error

Ultrashort (series 2), < 100 bp coor



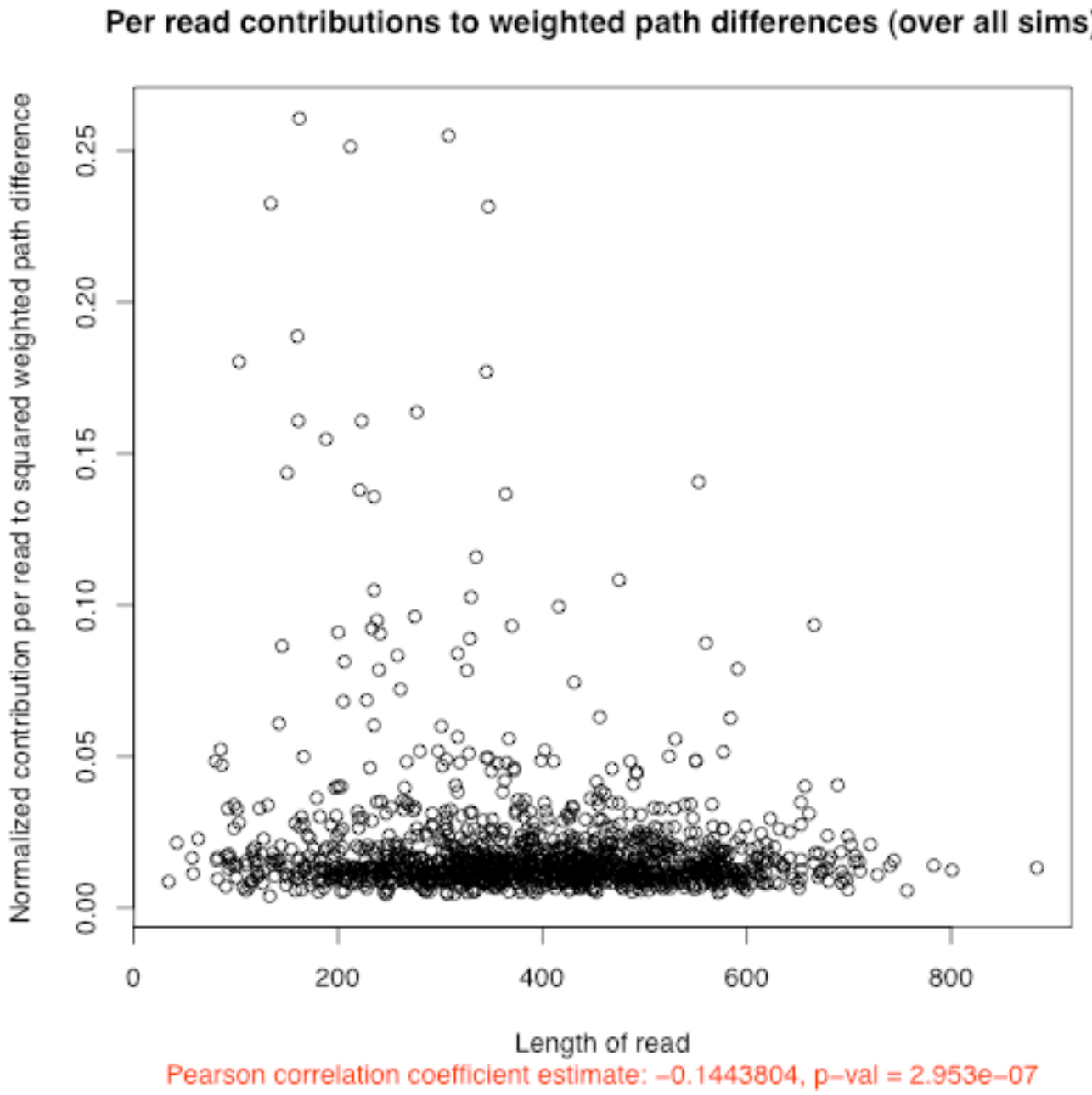
Series 2 all reads



No significant correlations (Spearman)

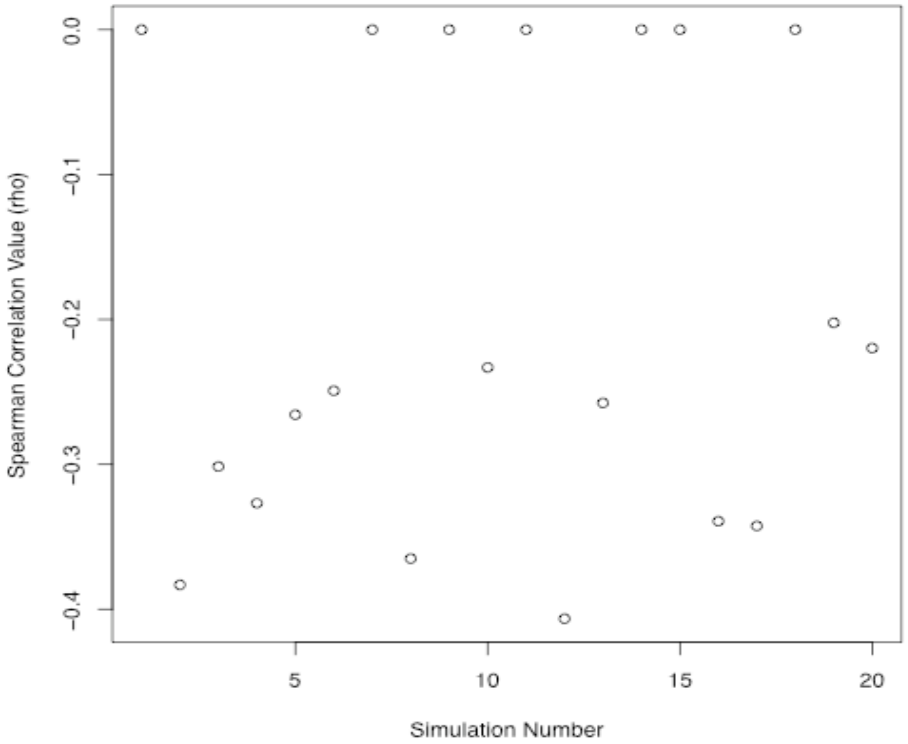
Read length bias?

Weighted path differences

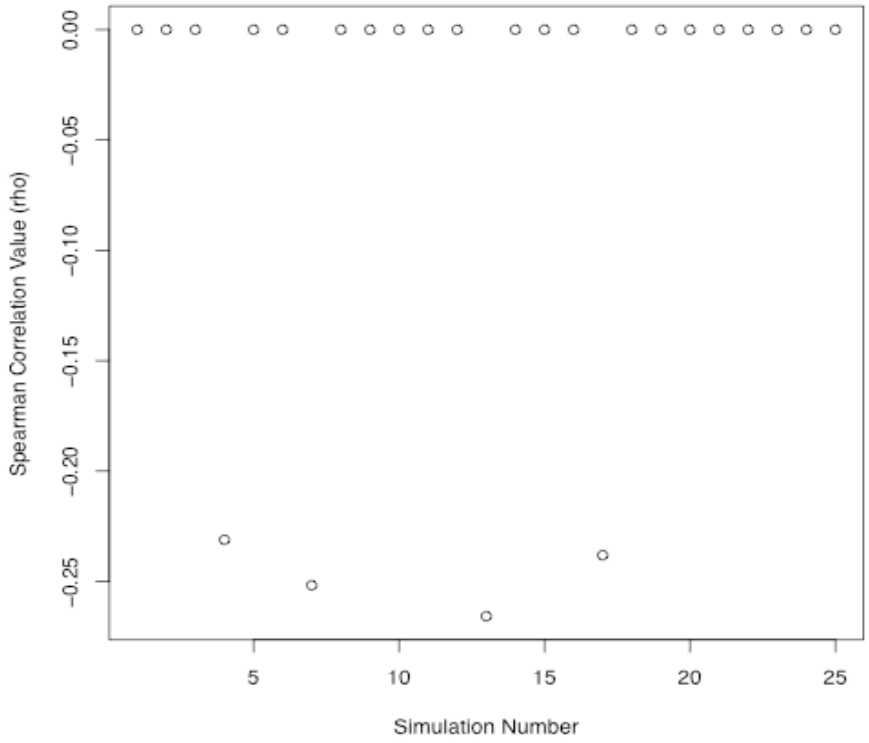


Reads tend to co-cluster in a tree

Length v. Tip-to-tip Distance Spearman Correlation Values



Length (>100bp) v. Tip-to-tip Distance Spearman Correlation Values



Summary

Error (at least for median difference) does not appear to be a function of:

- Position in the alignment
- Molecular rate variation across 16S locus
- Amount of information in an alignment column

There may be an association between read length and error:

- Reads <100bp are clearly a source of error
- Sam and Tom find, using different error metrics, different answers for reads > 100bp
  - Sam's test is more sensitive and shows a correlation between length and error

Multiple reads sampled from the same source tend to co-cluster on the tree

Reads tend to cluster in a taxonomically meaningful way

In general, we underestimate diversity from the standpoint of %ID and taxonomy terms

To Do List:

1. Amend pipeline to correctly scale read distances and reduce error
2. Reanalyze GOS and distal gut and see if improve biodiversity resolution remains true
3. Write the manuscript!
4. Others?

Call to authors

1. Who is on this manuscript?
2. Will require assistance in the next 1.5 months to put the manuscript together
3. Might request assistance on few remaining analyses
4. Might be an early question, but where should we go with this story?