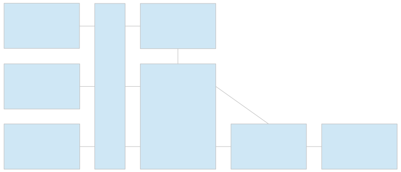
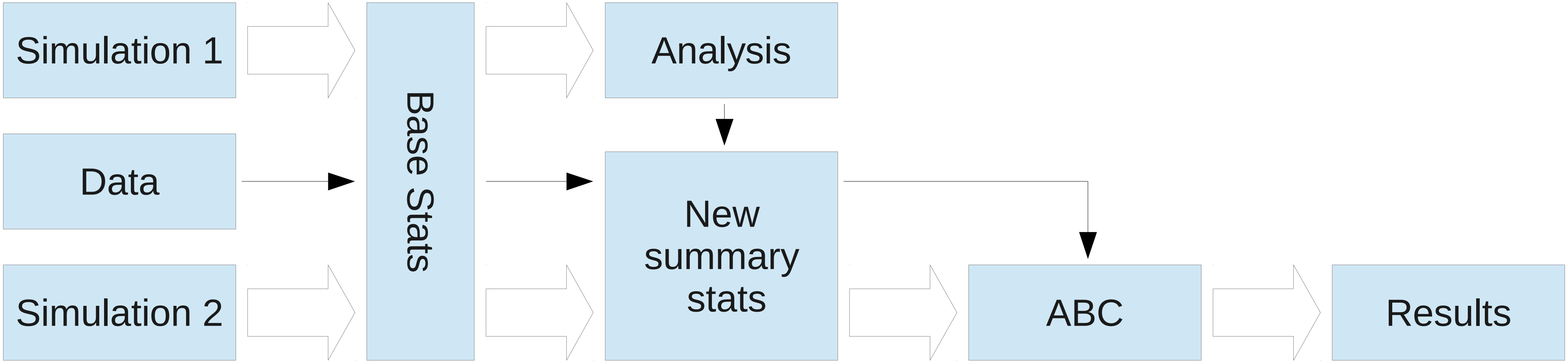
Seeking Signatures of Hybridization by Approximate Bayesian Computation

Michael Woodhams
with Barbara Holland

UNIVERSITY *of* TASMANIA





Simulation 1

Data

Simulation 2

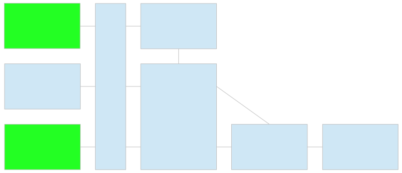
Base Stats

Analysis

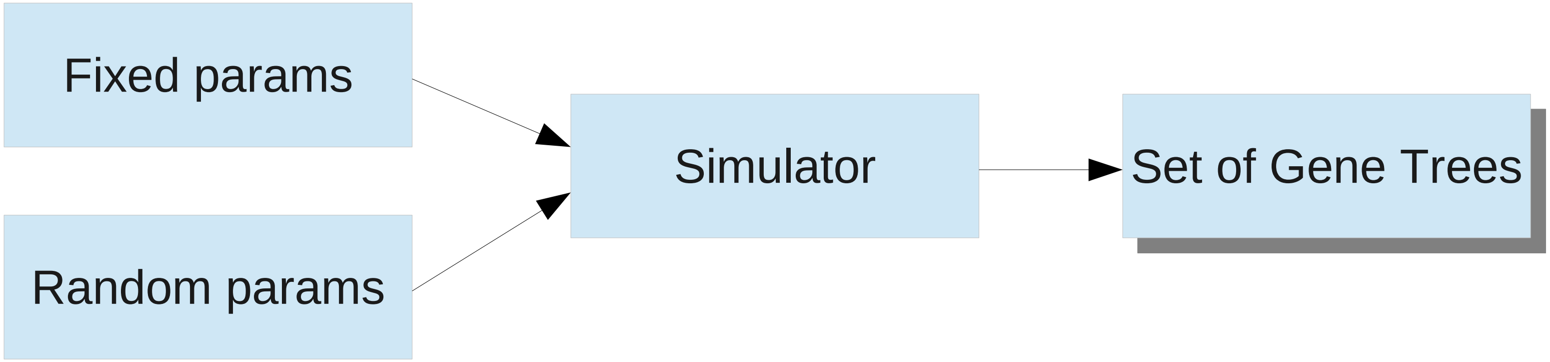
New summary stats

ABC

Results



Simulator

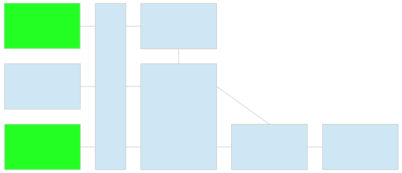


Fixed params

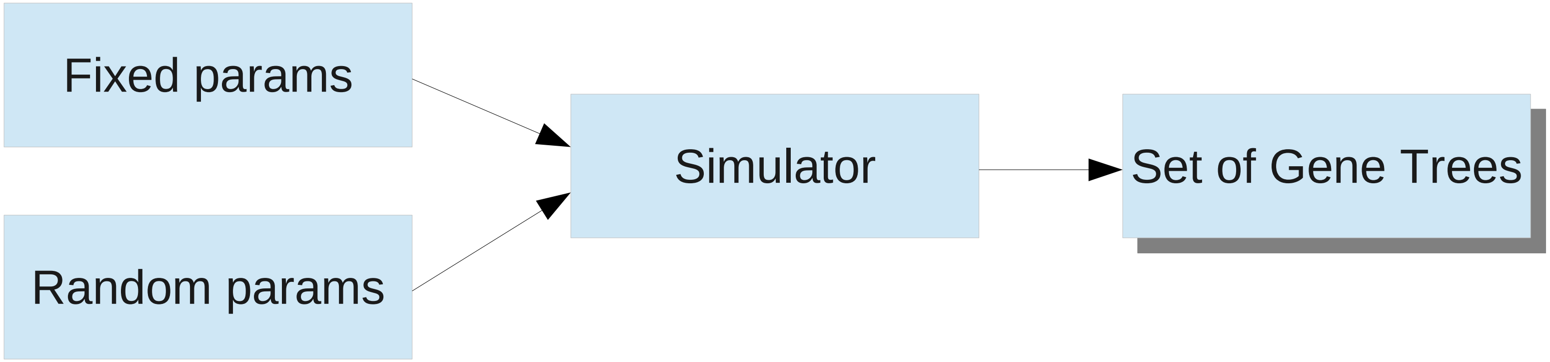
Random params

Simulator

Set of Gene Trees



Simulator

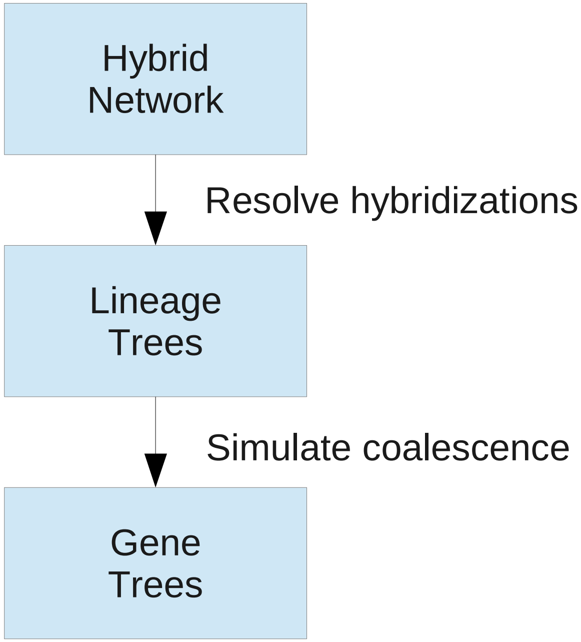


Fixed params

Random params

Simulator

Set of Gene Trees



Hybrid Network

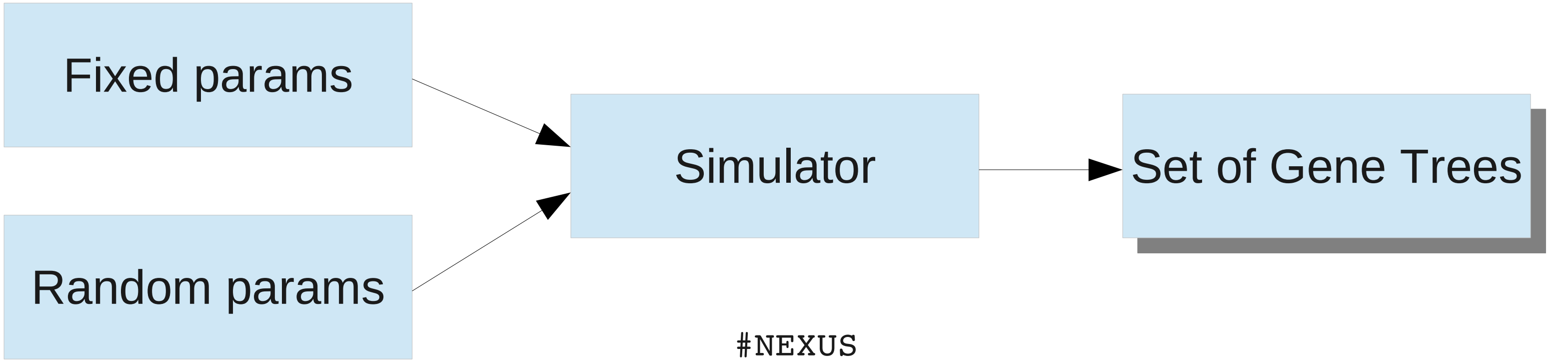
Resolve hybridizations

Lineage Trees

Simulate coalescence

Gene Trees

Simulator

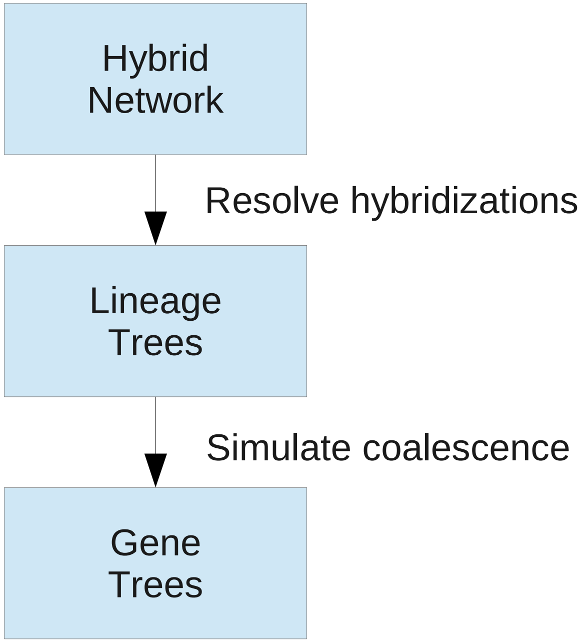


Fixed params

Random params

Simulator

Set of Gene Trees



Hybrid Network

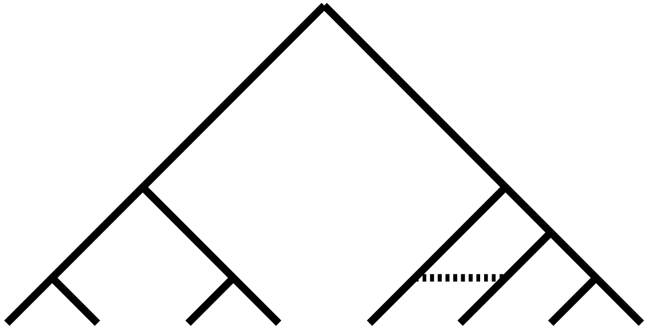
Resolve hybridizations

Lineage Trees

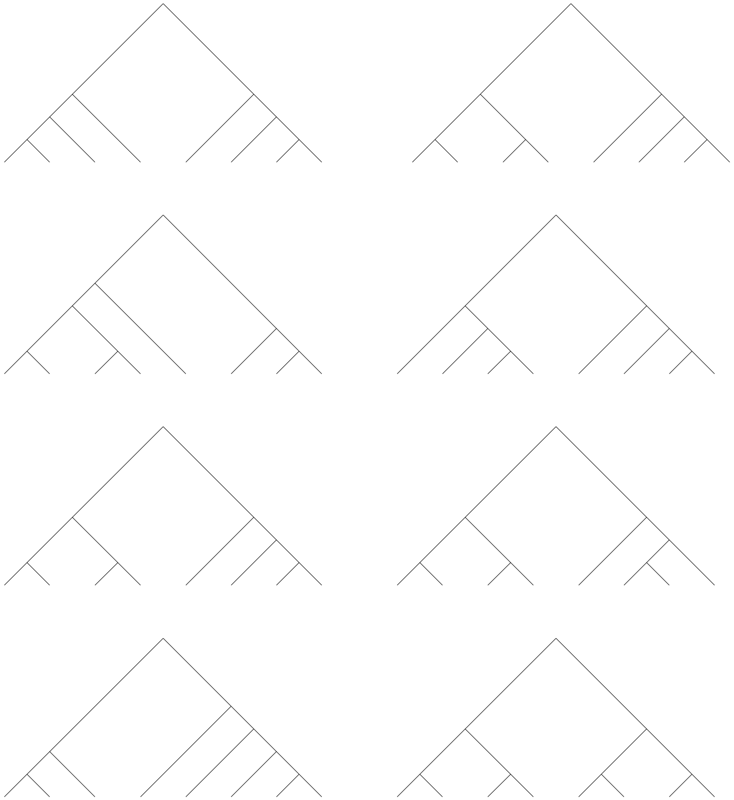
Simulate coalescence

Gene Trees

```
#NEXUS
begin hybridseq;
    epochs = ();
    speciation rate = (1);
    hybridization rate = (0.1);
    introgression rate = (0);
    hybridization function = step;
    hybridization threshold = 100;
    hybridization distribution =(0.5,1);
    minimum hybridizations = 3;
    coalesce = true;
    halt time = 100;
[   halt taxa = 23;]
    halt hybrid = 100;
[   number random trees = 1070;]
end;
begin ABC;
    iterations = 50000;
    reduce hybridizations to = HYBR(0,3);
    coalescence rate = COAL(1,20);
    ...
end;
begin trees;
...
```



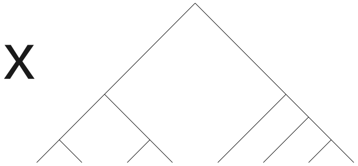
Coalescence



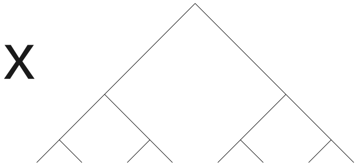
(we hope that other sources of phylogenetic
error will behave like coalescence)

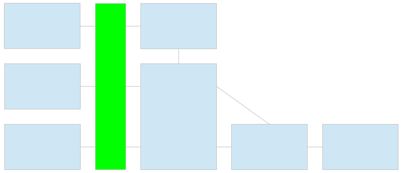
Hybridization

5 x



3 x





Base Stats

TE: Tree Entropy. Entropy of gene tree topologies

QE: Quartet Entropy: sum over quadruples of taxa, entropy of how that quadruple resolves into quartets.

SI: Split incompatibility. Sum over pairs of gene trees of their Robinson-Foulds distance. Equivalently, number of incompatible pairs of splits from the gene trees

SI-*k*: Threshold split incompatibility: like SI but subtract *k* from number of times each split occurs

RS*k*: Rare splits. The number of splits occurring *k* or fewer times

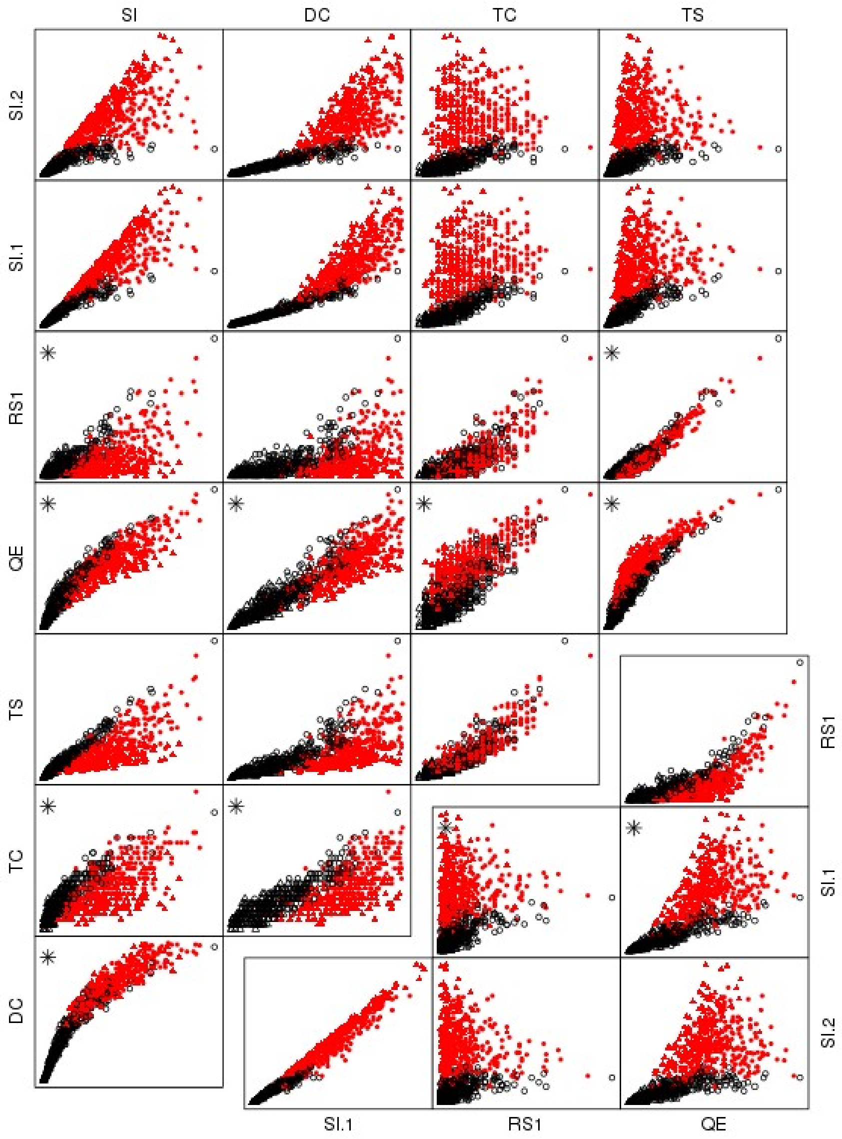
DC: Distance to Consensus. The sum over gene trees of Robinson-Foulds distance to majority-rule consensus tree.

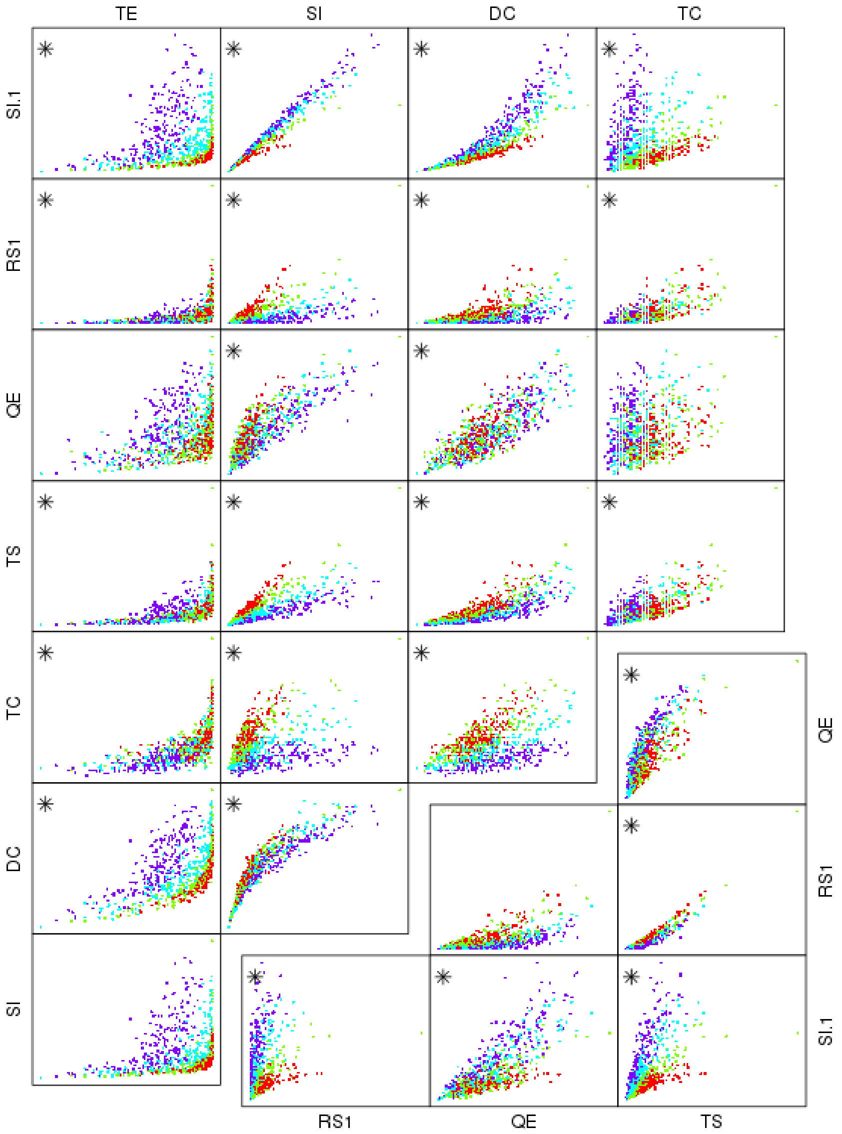
TS: Total Splits. The number of distinct splits in the gene trees

TC: Total Cherries. The number of distinct cherries in the gene trees

SPR, NNI distances would be ideal, but too computationally expensive.
Suggestions welcome.

Base Stats

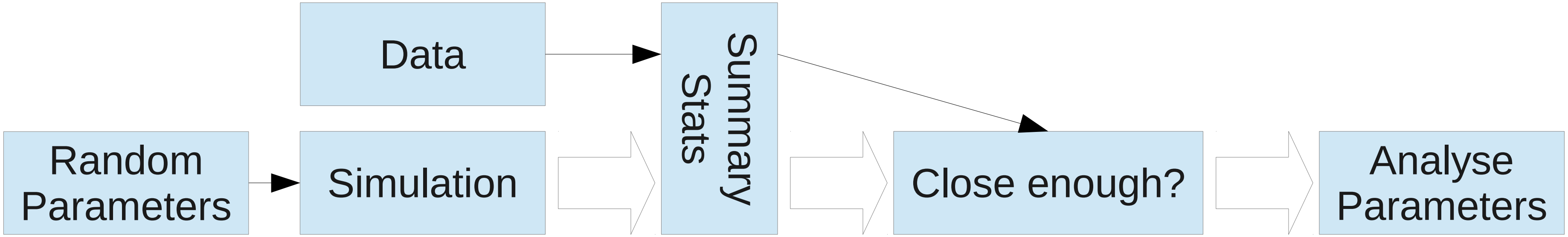




No hybrid, two hybrid
▲fast coal, ● slow coal

Coal. rate and hybr rate: high, med, low, tiny

ABC Overview



Data

Random Parameters

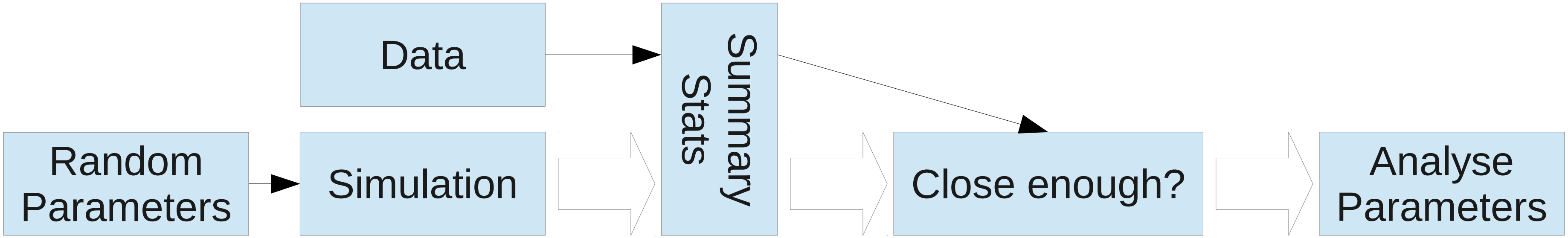
Simulation

Summary Stats

Close enough?

Analyse Parameters

ABC Overview



Data

Random Parameters

Simulation

Summary Stats

Close enough?

Analyse Parameters

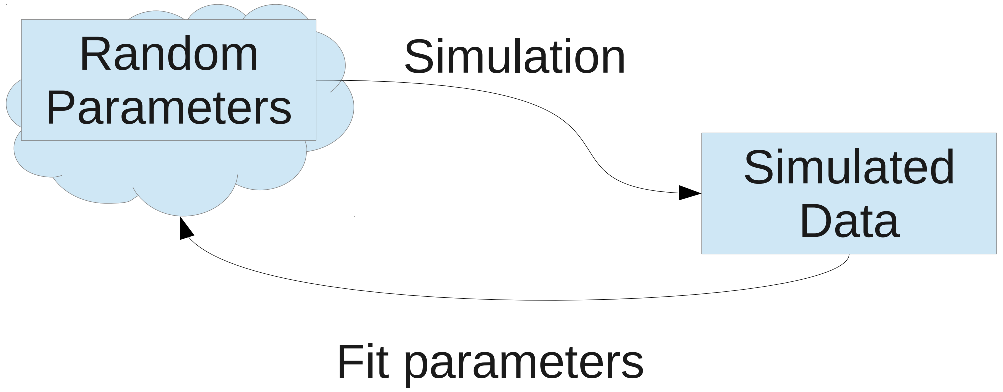
Randomized over what range?

Which summary stats?

How close?

Summary Stats

Semi-automatic ABC: Fearnhead & Prangle, JRStatS B, **74** 419-474 (2012)



Random Parameters

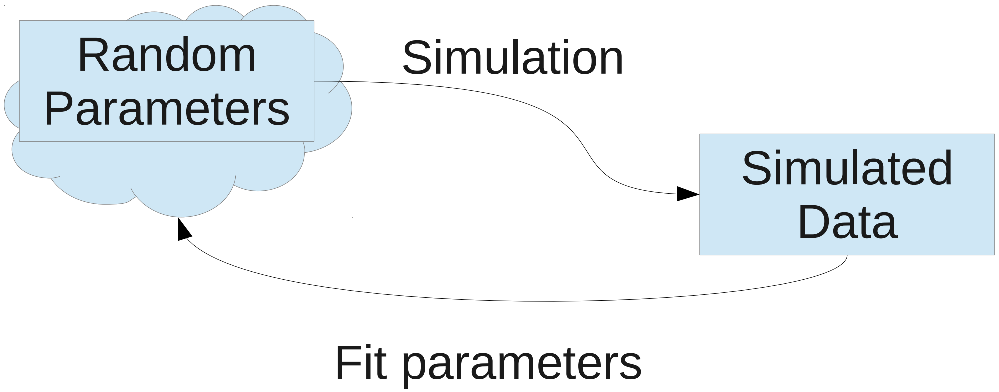
Simulation

Simulated Data

Fit parameters

Summary Stats

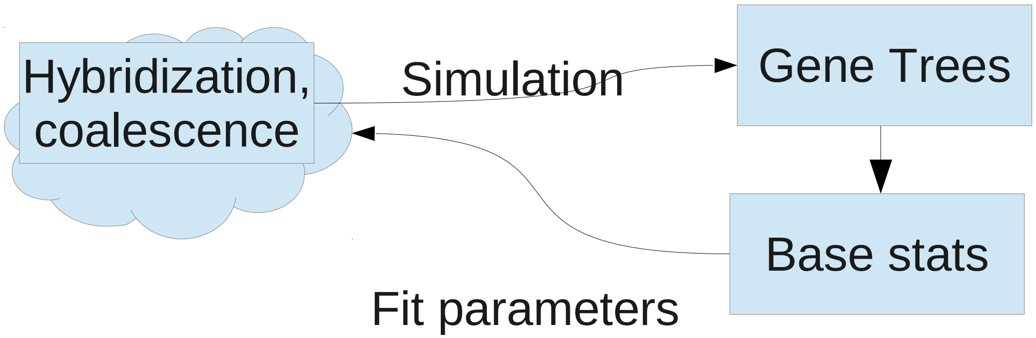
Semi-automatic ABC: Fearnhead & Prangle, JRStatS B, **74** 419-474 (2012)


Random Parameters

Simulation

Simulated Data

Fit parameters


Hybridization, coalescence

Simulation

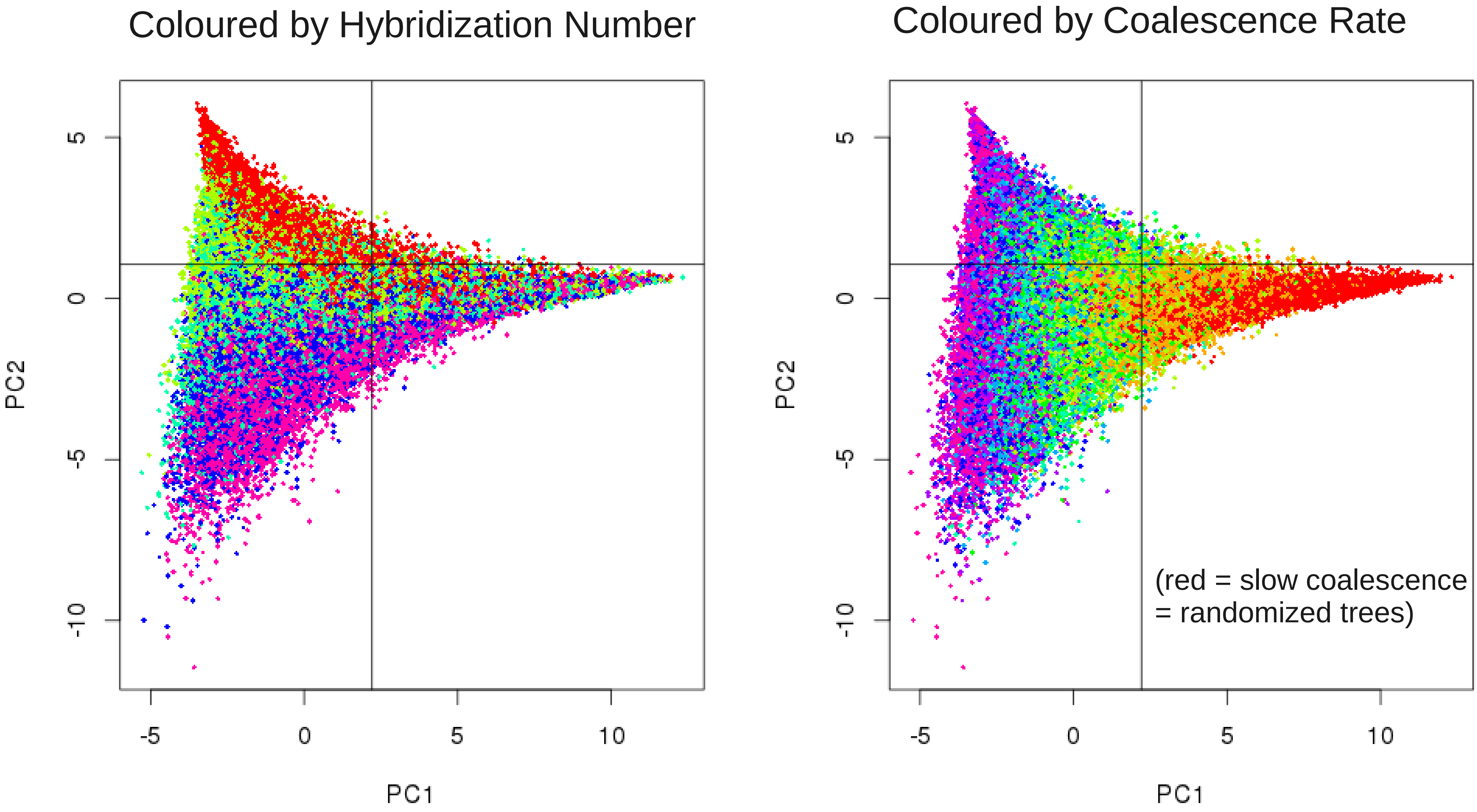
Gene Trees

Base stats

Fit parameters

Fitted hybridization, coalescence = summary statistics for ABC

Summary Stats



Coloured by Hybridization Number

Coloured by Coalescence Rate

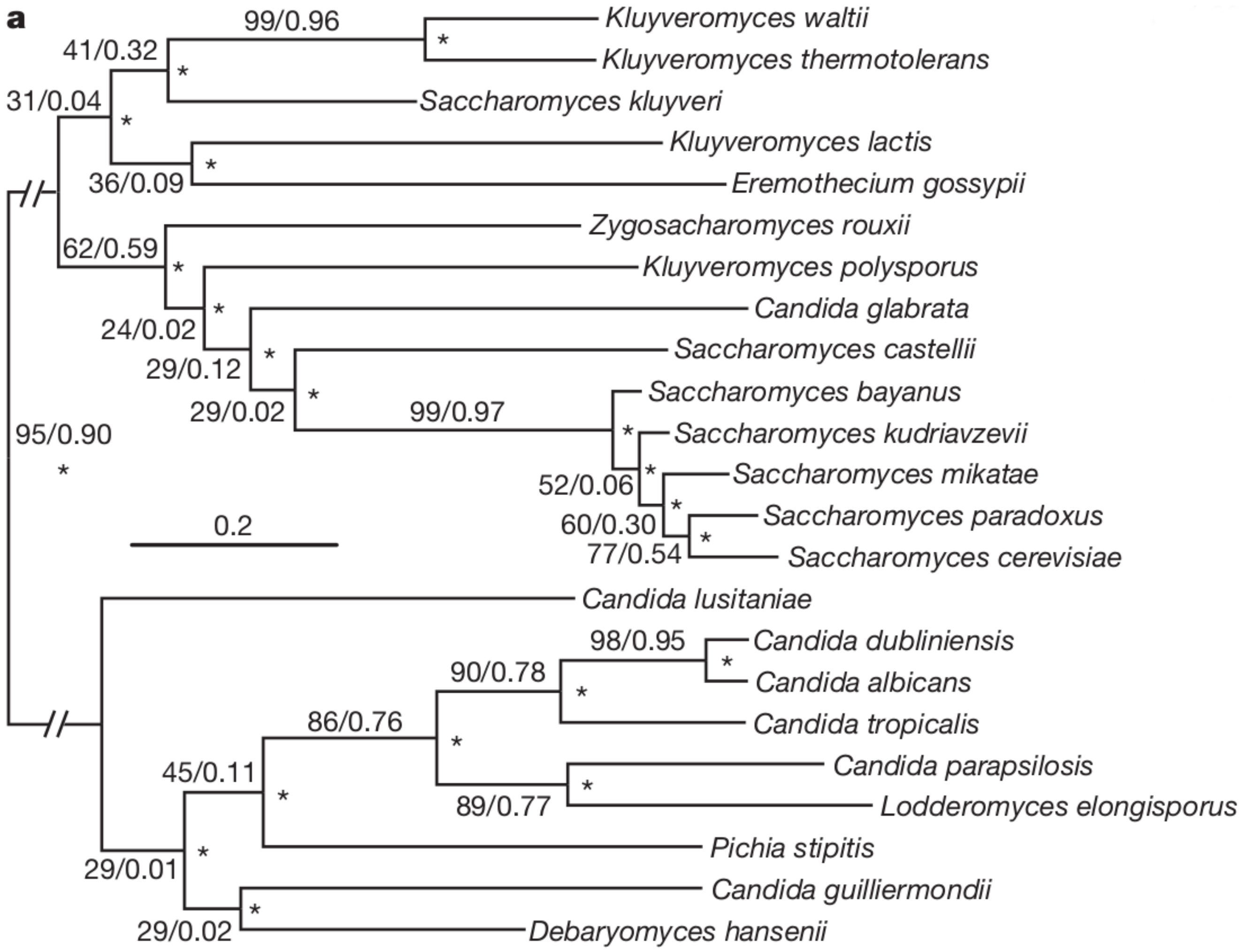
(red = slow coalescence
= randomized trees)

Principal Component Analysis

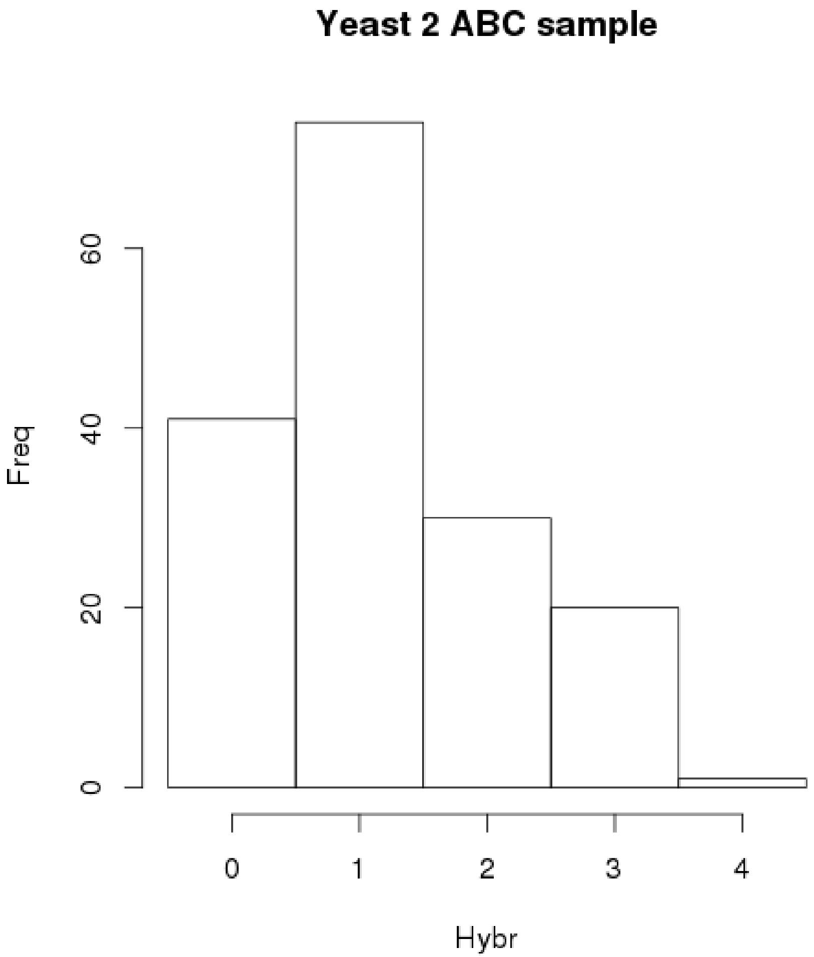
Data

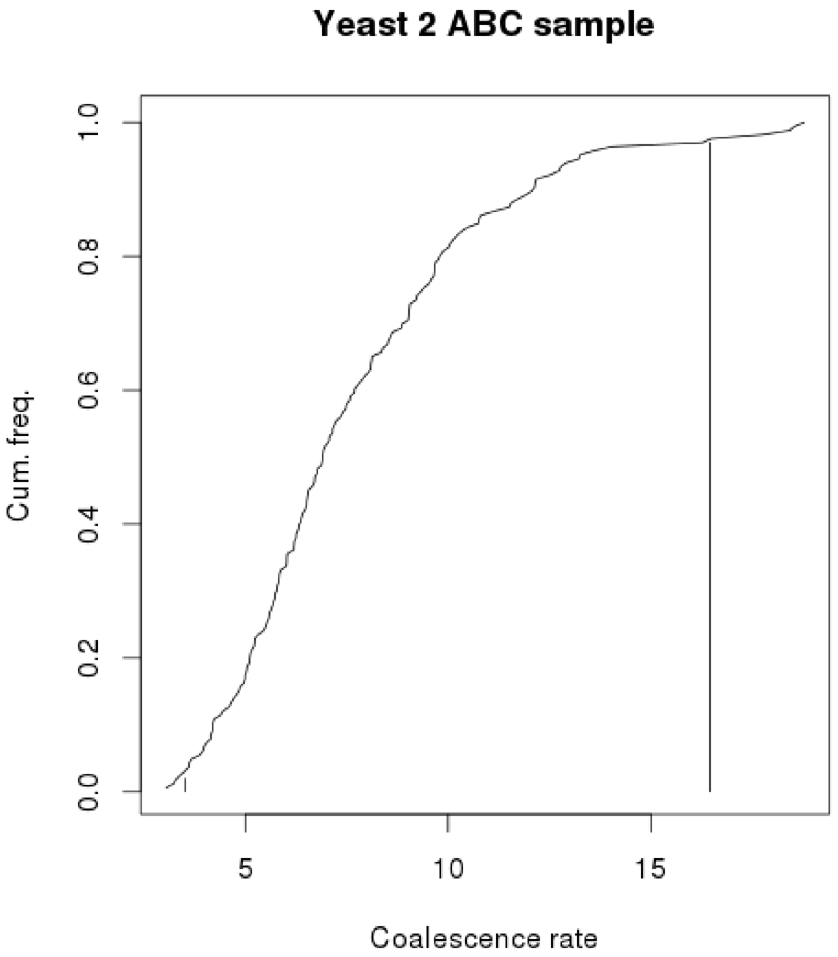
Inferring ancient divergences...: Salichos & Rokas, Nature, **497** 327-331 (2013)

Yeast
23 taxa
1070 genes



Results


Yeast 2 ABC sample

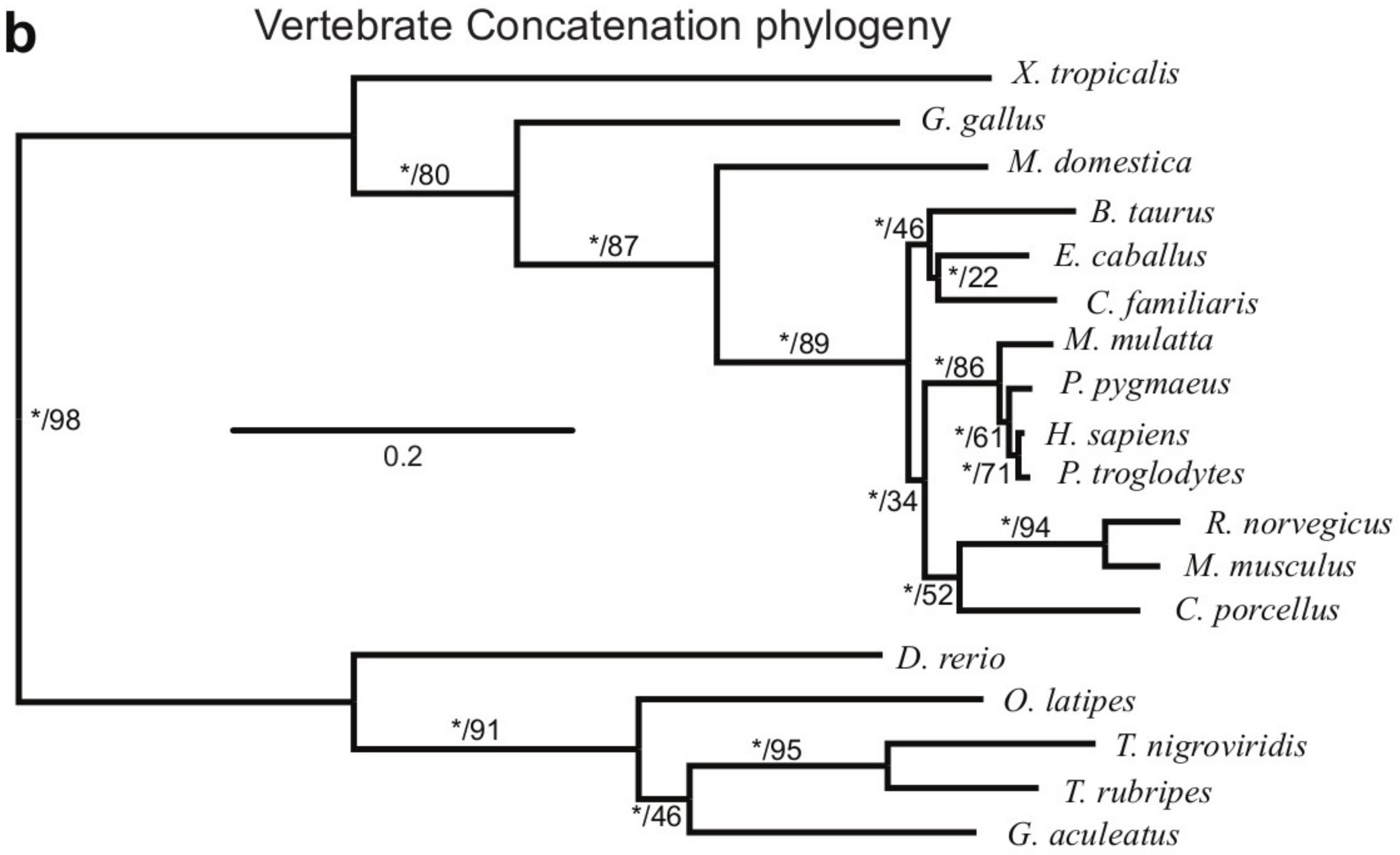

Yeast 2 ABC sample

Hybr = 0 has p=0.25

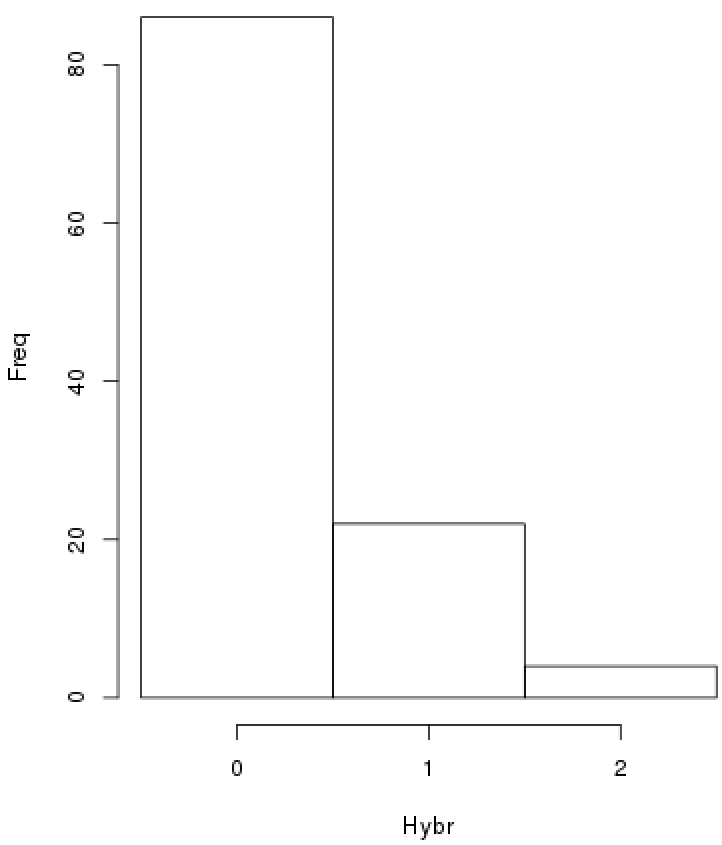
Data

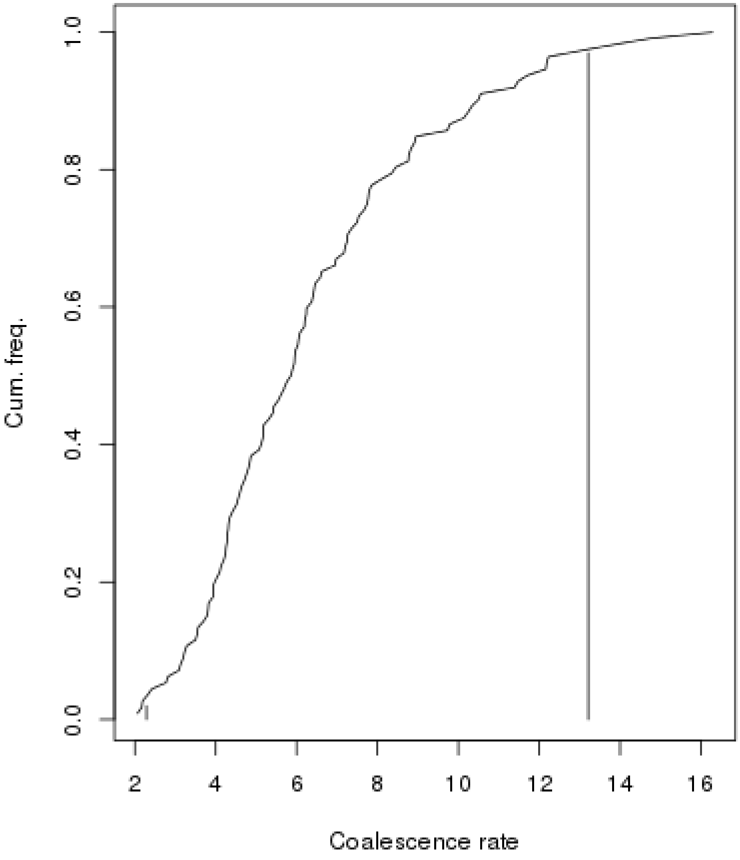
Inferring ancient divergences...: Salichos & Rokas, Nature, **497** 327-331 (2013)

Vertebrates
18 taxa
1087 genes



Results


Vertebrate 2 ABC sample


Vertebrate 2 ABC sample

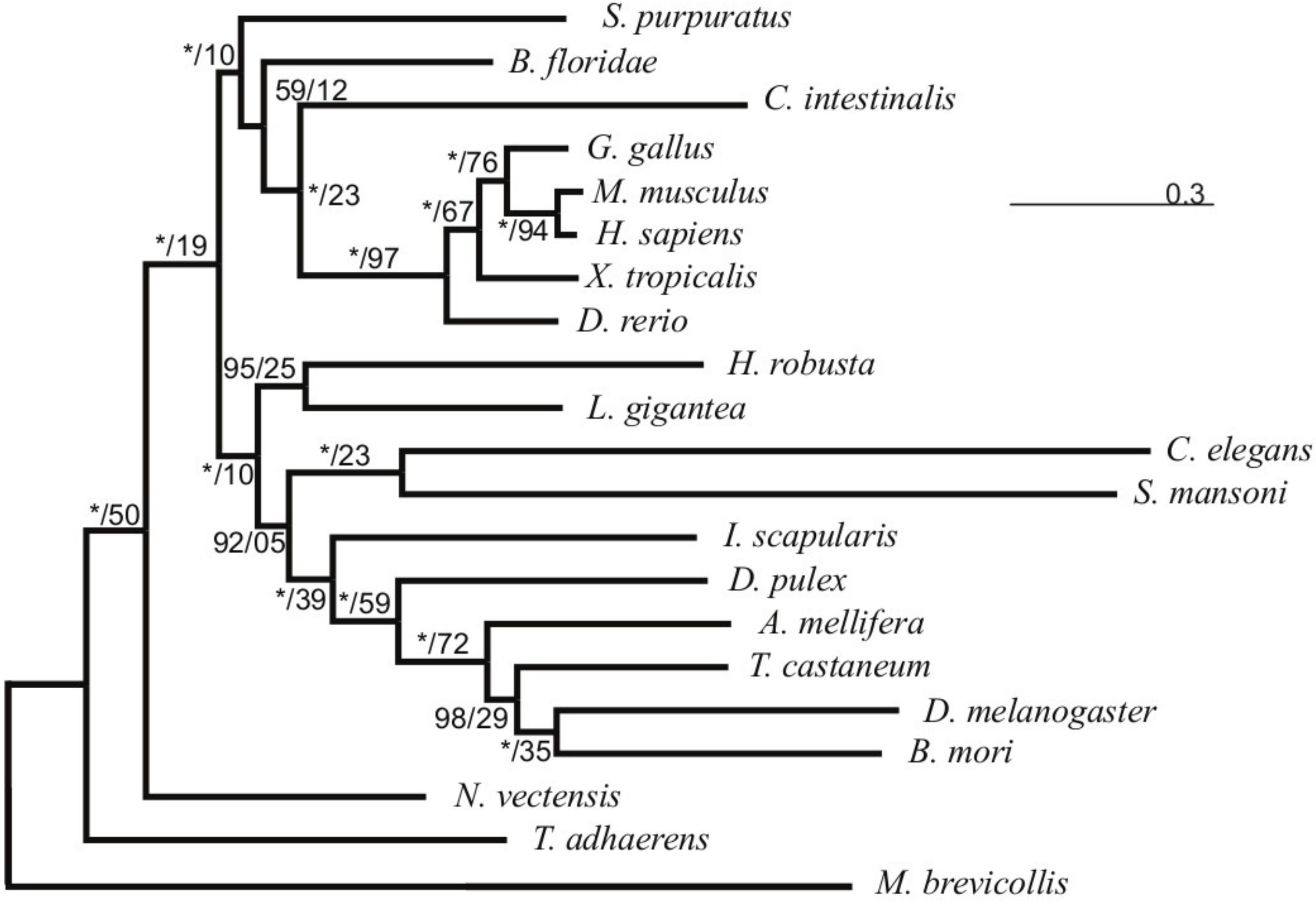
Hybr > 0 has p=0.23

Data

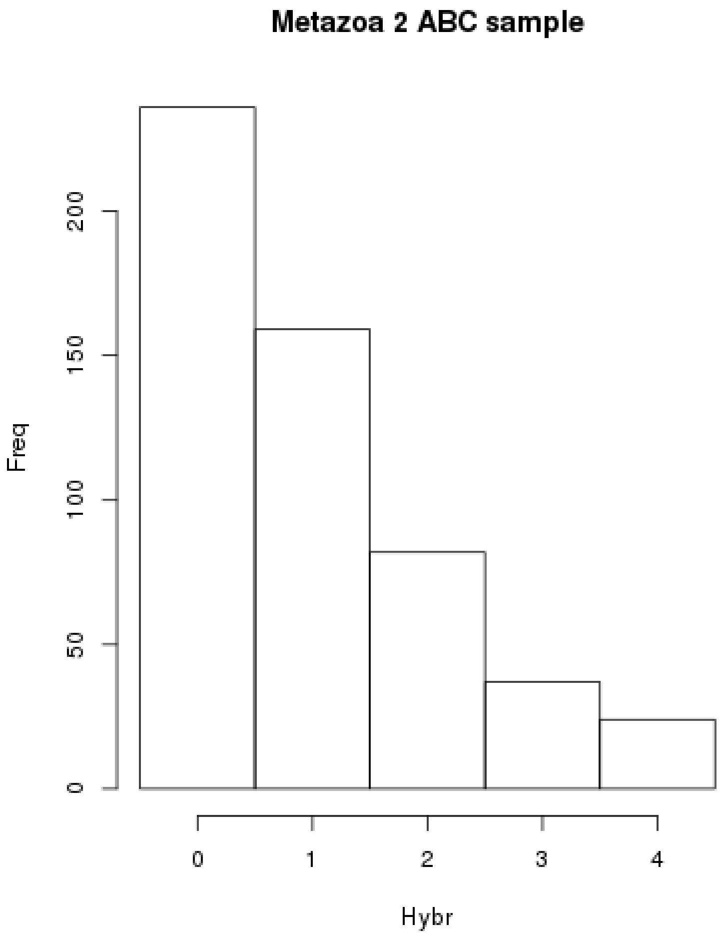
Inferring ancient divergences...: Salichos & Rokas, Nature, **497** 327-331 (2013)

Metazoa
21 taxa
225 genes

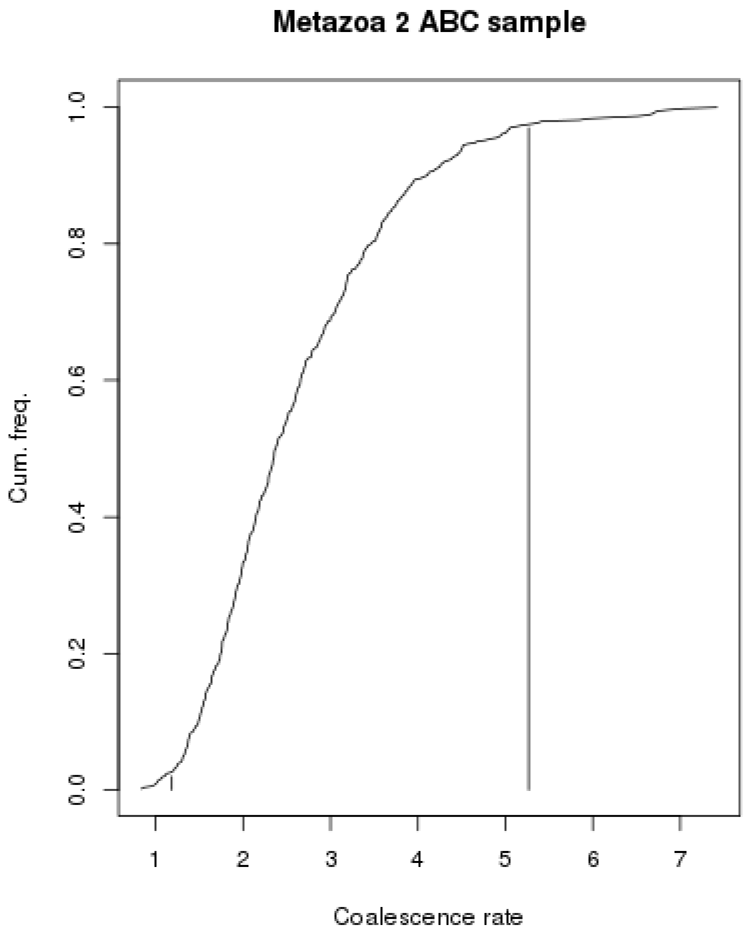
e Metazoan Concatenation phylogeny



Results



Metazoa 2 ABC sample



Metazoa 2 ABC sample