Coalescent simulations

Aurelien Tellier & Karl Schmid

SYNBREED Winter School 2012 in Salzburg

November 28, 2012

Coalescent simulation tools and frameworks

Command-line based:

- ms:http://home.uchicago.edu/rhudson1
- msms: http://www.mabs.at/ewing/msms
- SIMCOAL2: http://cmpg.unibe.ch/software/simcoal2/

Framework and wrappers:

- CoaSim (Python) - http://www.daimi.au.dk/~mailund/CoaSim
- egglib (Python) - http://egglib.sourceforge.net/wrappers.html
- popGenome (R package)

GUI:

- msmsplay - part of msms

A more comprehensive list in http://en.wikipedia.org/wiki/Coalescent_theory

Example of a simulation with Hudson's ms

Simulation of a neutral population[1]

```
$ ms 10 20 -t 5
```

Meanings f the command line arguments:
Create 20 samples of 10 sequences each with a θ value of 5 (per gene; option: `-t 5`).
Run simulation on the command line and store in file:

```
$ ./msms 10 20 -t 5 > neutsim.dat
$ less neutsim.dat
```

[1] All simulations can be done with `msms` using the same arguments.

Output of the simulation

```
ms 10 20 -t 5  [3.2rc Build:74]
0x583c1949f06cbc38

//
segsites: 14
positions: 0.06851 0.17125 0.18167 0.19086 0.29312 0.33638 0.46071 0.56
01010001001110
01000010101011
01001010101011
00000000010000
01010001001110
01010001001110
01100000101010
01010001001110
01000000101010
11010100001110

//
segsites: 9
positions: 0.06684 0.14311 0.18304 0.43562 0.53559 0.66832 0.81876 0.89
110100000
110100000
```

Selection testing with coalescent simulations

A scenario:

1. You sequence 1000 randomly chosen genes in a sample of 100 individuals. The observed genome-wide average nucleotide diversity is $\pi = 0.001$ (per nucleotide)
2. You then sequence a single gene you suspect to have experienced a selective sweep and observe $\pi = 0.00002$

Selection testing with coalescent simulations

- Keep in mind, π is an unbiased estmator of $\theta = 4N_e\mu$
- Scale to the gene
- Expected diversity: $1000\text{nt} \times 0.001 = 1$
- Observed diversity: $1000\text{nt} \times 0.00002 = 0.02$

Simulate the neutral distribution based on genome-wide diversity (θ per gene $= 1$):

```
$ ms 100 1000 -t 1
```

We get 1000 samples with 100 individuals
rightarrow Very large, let's summarize it

Summarize the `msms` output

```
$ ms 100 1000 -t 1 | sample_stats
```

The resulting output is

```
pi: 1.741414 ss: 8 D: 0.308070 thetaH: 2.581818 H: -0.840404
pi: 1.550303 ss: 7 D: 0.343750 thetaH: 3.136566 H: -1.586263
pi: 0.872525 ss: 4 D: 0.256489 thetaH: 1.006263 H: -0.133737
pi: 2.125455 ss: 10 D: 0.256705 thetaH: 0.723030 H: 1.402424
pi: 0.506667 ss: 4 D: -0.682525 thetaH: 0.220606 H: 0.286061
```

- `pi`: π
- `ss`: Number of segregating sites
- `D`: Tajima's D
- `H`: Fay and Wu's H

Since we are only interested in π, we cut out the column (which is the second column in the output) with the UNIX `cut` program:

```
$ ./ms 100 1000 -t 1 | ./sample_stats | cut -f 2
```

The output is

```
1.347475
0.168687
0.785657
1.037576
1.723838
0.185455
0.801818
0.846263
...
```

We want to summarize the output. First, we use the program `stats` of `ms` to calculate the quantiles:

```
$ ms 100 1000 -t 1 | sample_stats | cut -f 2 | stats 0.025 0.5 0.095
```

and the output is

```
0.973357 sd: 0.741172 n: 1000 \
0.025 0.039798 0.500 0.823434 0.095 0.171313
```

The value of the 0.025 quantile is 0.039798, which means that the observed value of 0.02 is smaller than the 0.025 quantile.

Simulated distribution of π

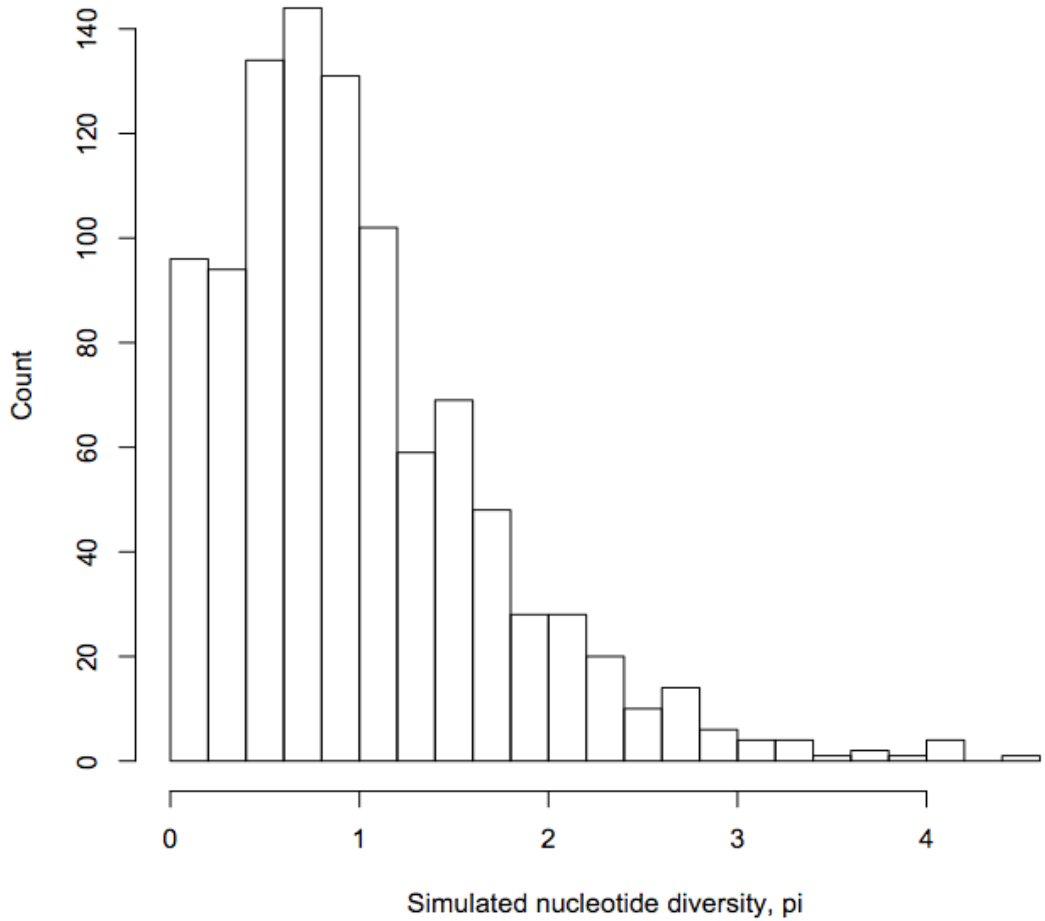


Figure : Distribution of nucleotide diversity values in 1000 simulations of 100 individuals with an value of $\theta = 4N_e\mu = 1$ per gene.

Analysis of simulated gene genealogies

It is also possible to run the same simulations and show only the trees. The command

```
$ ./ms 10 5 -T
```

results in:

```
./ms 10 5 -T
64768 5784 45784
//
((5:0.003,10:0.003):0.514,(7:0.343,(1:0.342,(((6:0.013,8:0.013):0.120,(
//
(2:0.523,((1:0.059,10:0.059):0.319,((3:0.025,6:0.025):0.189,((4:0.054,8
//
(3:1.192,((8:0.006,9:0.006):0.655,(6:0.354,((10:0.039,(7:0.020,(2:0.011
//
((2:0.162,(4:0.056,(8:0.035,9:0.035):0.021):0.106):0.232,(5:0.241,((3:0
//
(4:0.750,(9:0.448,(8:0.192,(3:0.107,(7:0.050,((2:0.001,6:0.001):0.017,(
```

Plot the genealogies

Store the genealogies in a file:

```
$ ./ms 10 5 -T > treefile.tre
```

We can use the ape package of R to read these trees (which are stored in the file treefile.tre) and to plot them.

```
library(ape)
trees <- read.tree("~/src/ms_hudson/msdir/treefile.tre",\
    skip = 4, comment.char="/")
par(mfrow=c(2,2),mai=c(0.2,0.2,0.2,0.2))
for (tree in trees) {
  plot.phylo(tree)
}
```

Plot of neutral genealogies

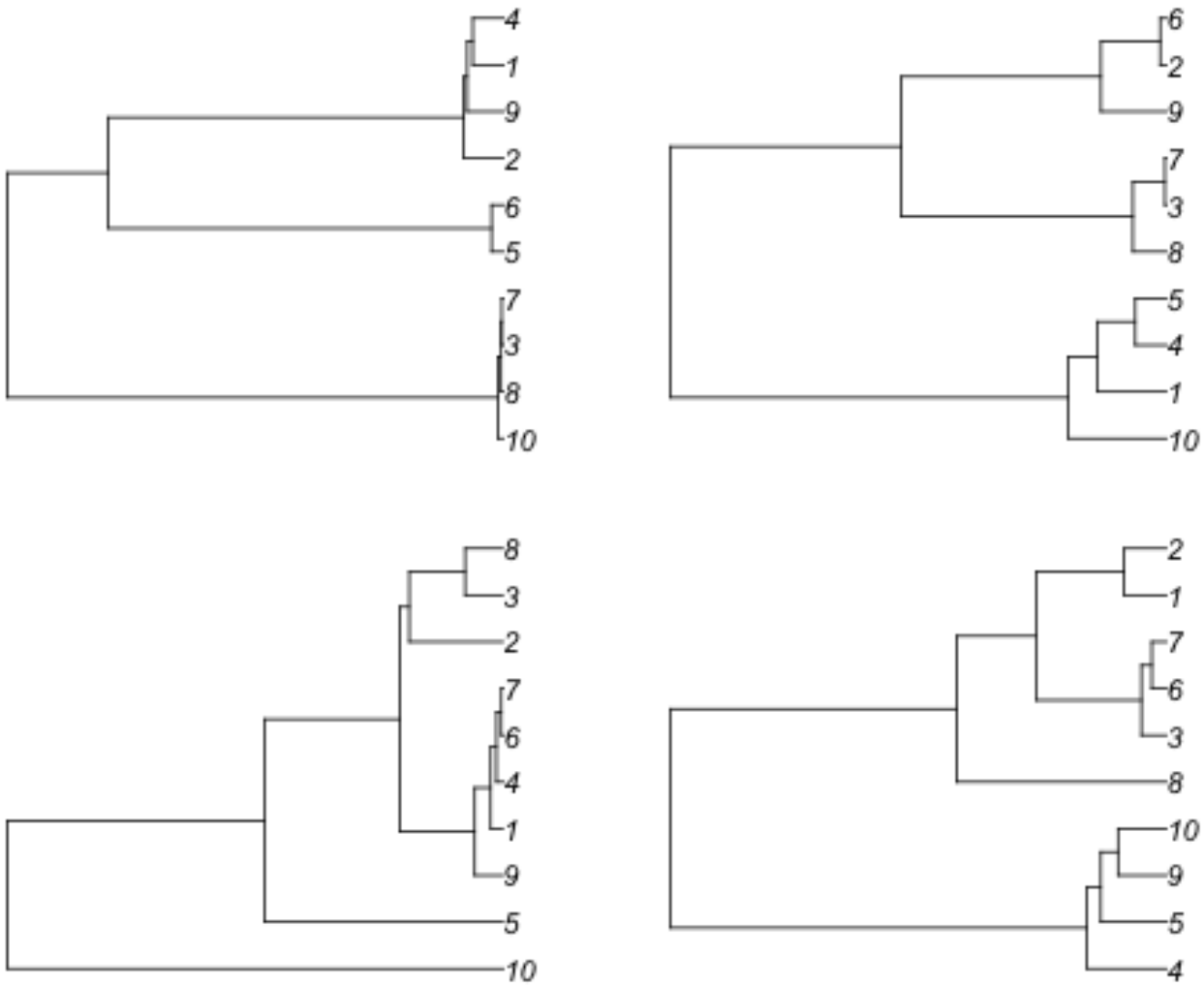
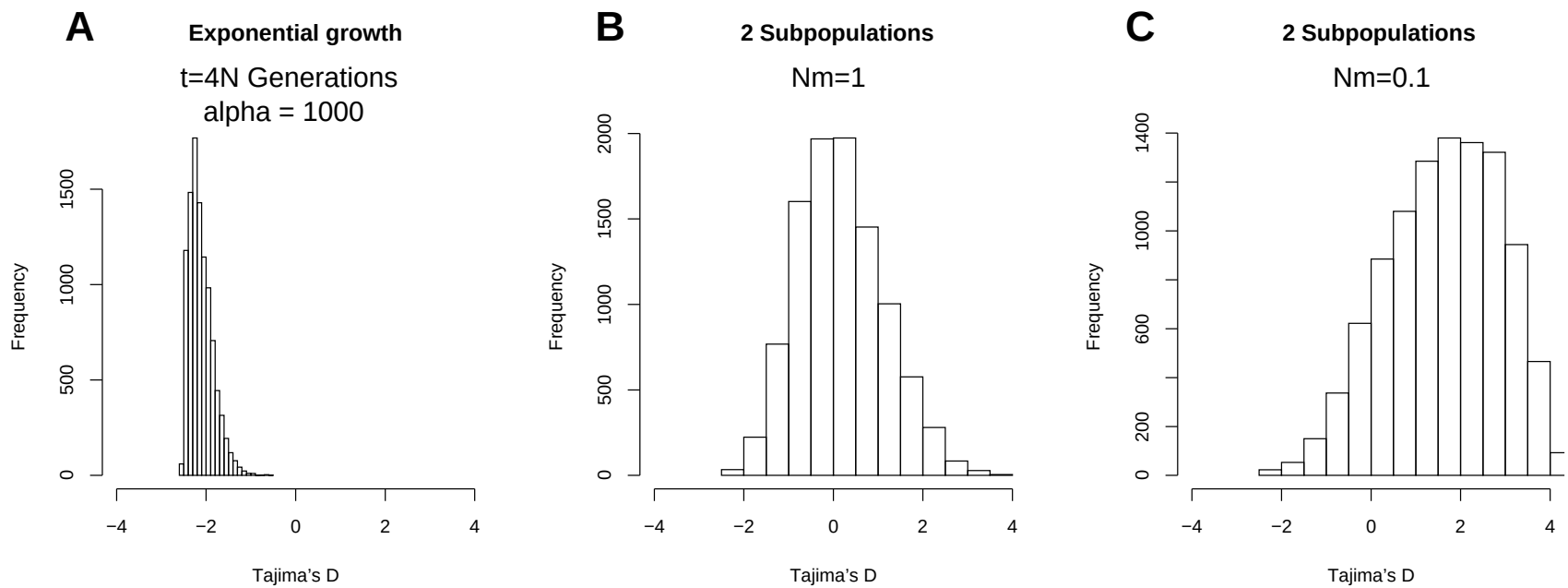
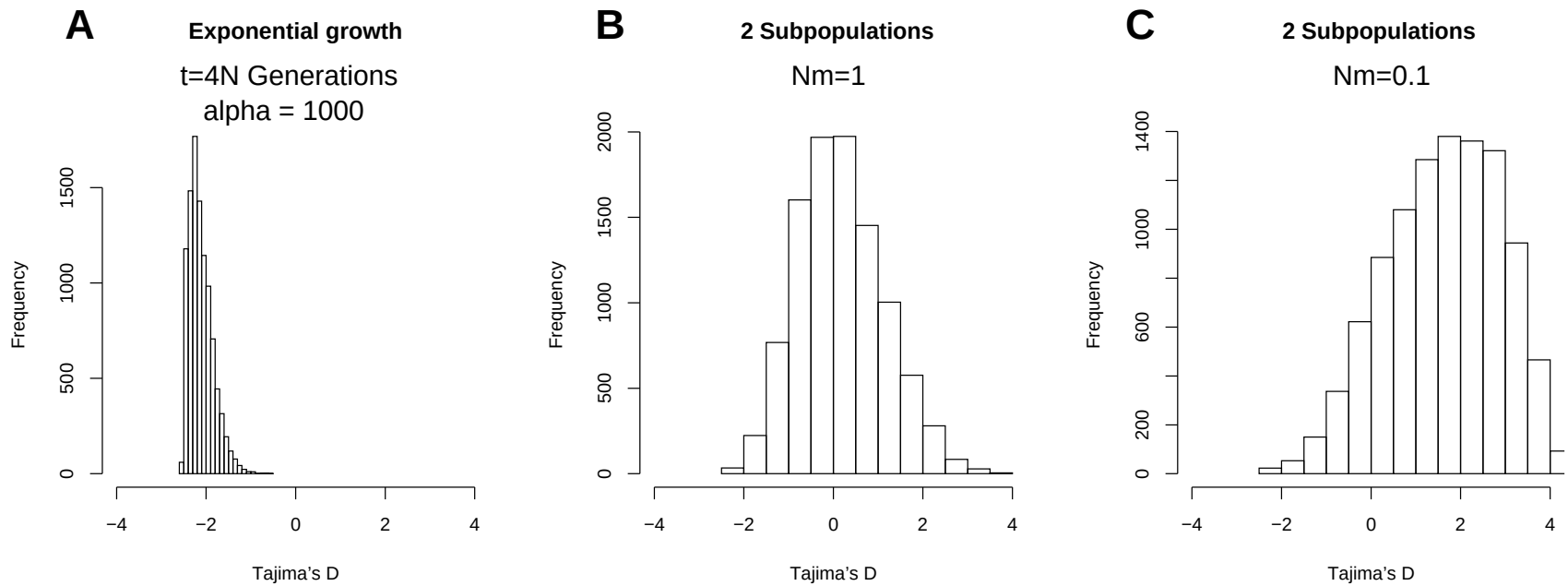


Figure : Gene genealogies of four random trees simulated under an exponentially growing population model.

Simulation of Tajima's D with different demographic models



Simulation of Tajima's D with different demographic models



Growth model:

- Growth rate: $N(t) = N_0 \exp^{-\alpha t}$
- $N(t)$: Population size in the past
- t is the time *before present*, measured in units of $4N_0$ Generations
- N_0 is current population size

Further reading

- The manuals of the simulation tools are usually very good
- Excoffier and Heckel (2006) Computer programs for population genetics data analysis: a survival guide. Nature Reviews Genetics. doi:10.1038/nrg1904
- Hoban, Bertorelle and Gaggiotti (2012) Computer simulations: tools for population and evolutionary genetics. Nature Reviews Genetics. doi:10.1038/nrg3130