# Simulating the coalescent in practice

Matthieu Foll

21.11.2011

Population Genomics Course

Helsinki

# Coalescent simulators

BIOINFORMATICS APPLICATIONS NOTE Vol. 18 no. 2 2002 Pages 337–338

## Generating samples under a Wright–Fisher neutral model of genetic variation

Richard R. Hudson

Department of Ecology and Evolution, University of Chicago, 1101 E. 57th Street, Chicago, IL 60637, USA

Received on August 8, 2001; revised and accepted on September 13, 2001

ms: http://home.uchicago.edu/rhudson1/source/mksamples.html

Resource

## Fast and flexible simulation of DNA sequence data

Gary K. Chen,[1,2] Paul Marjoram,[1] and Jeffrey D. Wall[2,3]

[1]Department of Preventive Medicine, University of Southern California, Los Angeles, California 90033, USA; [2]Institute for Human Genetics and Department of Epidemiology and Biostatistics, University of California, San Francisco, California 94143, USA

macs: http://www-hsc.usc.edu/~garykche/

BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 9 2011, pages 1332–1334 doi:10.1093/bioinformatics/btr124

Genetics and population analysis Advance Access publication March 12, 2011

## fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios

Laurent Excoffier[1,2,*] and Matthieu Foll[1,2]

[1]Institute of Ecology and Evolution, University of Berne, 3012 Berne and [2]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

Associate Editor: Jeffrey Barrett

fastsimcoal: http://cmpg.unibe.ch/software/fastsimcoal/

# fastsimcoal

- N=500 diploids, 2N=1000
- 2 samples
- 1Mb sequence with mu=$5.10^{-8}$ / bp

```
//Parameters for the coalescence simulation program simcoal2
1 samples to simulate :
//Population effective sizes (number of genes)
1000
//Samples sizes
2
//Growth rates     : negative growth implies population expansion
0
//Number of migration matrices : 0 implies no migration between demes
0
//historical event: time, source, sink, migrants, new deme size, new growth rate,
0 historical event
//Number of independent loci [chromosome]
1 0
//Per chromosome: Number of contiguous linkage Block: a block is a set of contiguous
1
//per Block:data type, number of loci, per generation recombination and mutation rates
DNA    1000000    0.00000000    0.0000000500    0.33
```

# ./fastsimcoal -i setting_file.par -n 1

```
#Arlequin input file written by the simulation program fastsimcoal.exe

[Profile]
            Title="A series of simulated samples"
            NbSamples=1

            GenotypicData=0
            GameticPhase=0
            RecessiveData=0
            DataType=DNA
            LocusSeparator=NONE
            MissingData='?'

[Data]
            [[Samples]]

#Number of independent chromosomes: 1
# 26 polymorphic positions on chromosome 1
#1343, 29723, 60987, 172454, 217878, 250006, 256183, 314992, 345010, 380453, 405242, 429120, 461588,
462077, 480709, 482969, 501976, 524243, 608229, 626449, 707893, 734931, 807206, 819209, 833758, 974030

                        SampleName="Sample 1"
                        SampleSize=2
                        SampleData= {
1_1         1           GCTTCTGTGTAATCCCGATGTAACTA
1_2         1           AACAGGACCGGCAAAAACGAAGCGCC

}

[[Structure]]

            StructureName="Simulated data"
            NbGroups=1
            Group={
               "Sample 1"
            }
```

# ms

- 2 samples, N=500, mu=$5.10^{-8}$ , L=1Mb
- theta=4NmuL=100
- 1 replicate

```
./ms 2 1 -t 100

28240 55425 2020

//
segsites: 7
positions: 0.0754 0.1424 0.1590 0.1976 0.2468 0.5948 0.6898
1011100
0100011
```

# macs

- 2 samples, N=500, mu=$5.10^{-8}$ , L=1Mb
- theta=4Nmu=$10^{-4}$

```
./macs 2 1000000 -t 0.0001


COMMAND:   ./macs 2 1000000 -t 0.0001
INPUT: Sample size is now 2
INPUT: Seq length is now 1e+06
INPUT: Scaled mutation rate is now 100
SEED:      1315226567
Debugging: 0

Graph Builder begin
Time for build prior tree: 0 seconds
Tree:0,pos:0,len:0.136958,TMRCA:0.0684789,ARG:,len:0.136958,TMRCA:0.0684789
SITE:      0              0.135060738       10
SITE:      1              0.308638343       10
SITE:      2              0.327420337       10
SITE:      3              0.382058725       01
SITE:      4              0.446238956       10
SITE:      5              0.602949569       01
SITE:      6              0.655999696       01
SITE:      7                0.6983443       10
SITE:      8              0.731537735       10
Completed the chromosome at position 1
gcstarts:0 gcends:0 xovers:0
TOTAL_SAMPLES:        2
TOTAL_SITES:          9
BEGIN_SELECTED_SITES
0          1          2          3          4          5          6          7          8
END_SELECTED_SITES
```

# ms / macs limitations

- Less flexible evolutionary scenarios
- Only infinite site model: need external programs to simulate DNA sequences
  - Add the option –T to output the tree
  - seq-gen program can use this tree to simulate sequences

**CABIOS**

Vol. 13 no. 3 1997
Pages 235–238

***Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees***

*Andrew Rambaut[1] and Nicholas C.Grassly*

# ms + seq-gen example

```
./ms 2 1 -T | tail -1 >treefile
./seq-gen -mHKY -l100 <treefile


Simulations of 2 taxa, 100 nucleotides
  for 1 tree(s) with 1 dataset(s) per tree

Branch lengths assumed to be number of substitutions per site

Rate homogeneity of sites.
Model = HKY: Hasegawa, Kishino & Yano (1985)
  Rate of transitions and transversions equal:
  transition/transversion ratio = 0.5 (K=1)
  with nucleotide frequencies equal.
 2 100
1
AATCCCGCCTTTGTGCGAAGTGTTCGACAAACGGATTACCTTGTGAGTACGTTCTTCCGCTGAGCAAGTATGCTTT
GATTTGCAATGCTTGTTTAACCGT
2
TACGCGAAGGTGCAGTCACCCCGACGTTTATTAGCATAGAAGCCCATATGGCGCGAAGGCATCTATAGTGATGTTT
CACAATCTTTGTCAGAATCGGTAT
Time taken: 0.000422 seconds
```

# fastsimcoal

- http://cmpg.unibe.ch/software/fastsimcoal/
- Complex evolutionary scenarios (migration, population resize, population fusion and fission, admixture, population growth…).
- Serial sampling.
- DNA sequences, SNP, STR (microsatellite).
- Arbitrary recombination rates, at any position. SMC' algorithm.
- Very fast.

# Performance

| Data set | No. of replicates | Sequence length | Program | | |
|---|---|---|---|---|---|
| | | | *ms* | *MaCS* | *fastsimcoal* |
| 1popNoRec | 1000 | 1 Mb | 1.1 | 11.1 | 9.5 |
| | 100 | 10 Mb | 9.6 | 107.0 | 72.9 |
| | 100 | 100 Mb | 147.9 | 1319.5 | 1038.1 |
| 2popNoRec | 1000 | 1 Mb | 1.2 | 12.5 | 9.3 |
| | 100 | 10 Mb | 8.9 | 128.1 | 71.5 |
| | 100 | 100 Mb | 161.2 | 1513.2 | 1099.9 |

| Data set | No. of replicates | Sequence length | Program | | |
|---|---|---|---|---|---|
| | | | *ms* | *MaCS* | *fastsimcoal* |
| 1popSmallSample | 1000 | 1 Mb | 0.344 | 0.242 | 0.095 |
| | 100 | 10 Mb | 159.246 | 2.618 | 0.460 |
| | 100 | 100 Mb | x | 26.124 | 4.364 |
| 2popSmallSample | 1000 | 1 Mb | 0.378 | 0.907 | 0.152 |
| | 100 | 10 Mb | 165.507 | 9.094 | 1.080 |
| | 100 | 100 Mb | x | 97.876 | 10.559 |

# Modeler for Simcoal2 (and fastsimcoal…)

BIOINFORMATICS **APPLICATIONS NOTE** *Vol. 23 no. 14 2007, pages 1848–1850*
*doi:10.1093/bioinformatics/btm243*

*Genetics and population analysis*

**MODELER4SIMCOAL2: A user-friendly, extensible modeler of demography and linked loci for coalescent simulations**

T. Antao[1,3,*], A. Beja-Pereira[1] and G. Luikart[1,2]

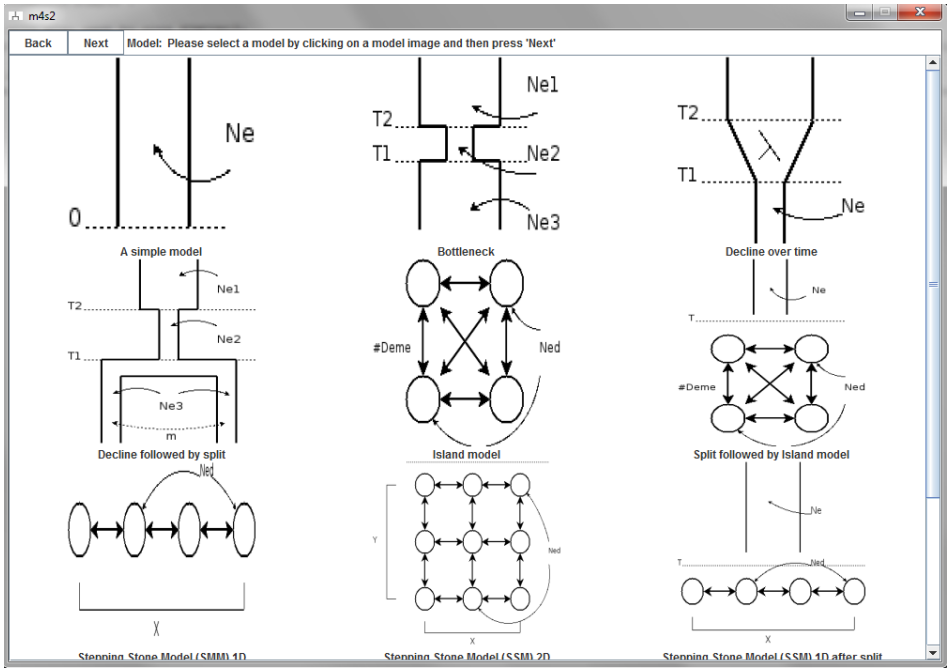
[1]CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, Campus Agrário de Vairão, Universidade do Porto, Portugal, [2]Division of Biological Sciences, University of Montana, Missoula, MT 59812, USA and [3]Departamento de Zoologia e Antropologia, Faculdade de Ciências do Porto, Portugal

Received on March 12, 2007; revised on April 11, 2007; accepted on April 27, 2007

Advance Access publication May 8, 2007

Associate Editor: Martin Bishop

- GUI for modeling demography
- http://popgen.eu/soft/m4s2/



# arlsumstat

arlsumstat is a command line version of Arlequin to compute summary statistics from arlequin input file, as e.g. generated by fastsimcoal.

List of summary statistics to computed are listed in a file with name ssdefs.txt

(ssdefs.txt can be either written by hand or generated by WinArl35)

arlsumstat also requires an arlequin settings file that tells arlsumstat which computations to perform. This *.ars file can also be generated with WinArl35.

We provide you here with a bash script LaunchArlSumStatDir*.sh, which

- takes as a first argument the name of a target directory where *.arp files are located. e.g. 1PopDNA_sta
- takes as a second argument the name of the settings file to use e.g. SettingsDNAStats.ars
- Takes as a third argument the name of the file where to ouput SSs. e.g. ssout.txt
- Launches arlsumstat as many times as there are *.arp files in the target directory.
- Output all summary statistics in the output file in a tabulated way

# Computing summary statistics on output from fastsimcoal

1. Copy arlsumstat*, ssdefs.txt, SettingsDNAStats.ars, LaunchArlSumstatDir*.sh to the directory containing the fastsimcoal par files
2. Open cmd prompt
3. launch LaunchArlSumstatDir*.sh with the following arguments

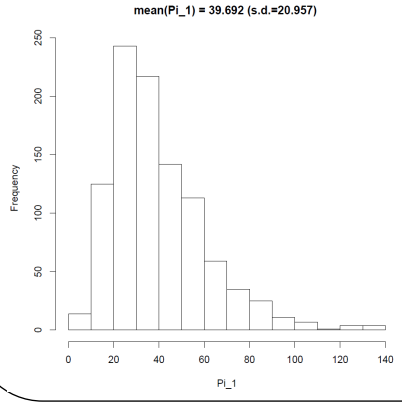
   (bash) ./LaunchArlsumstatDirPC.sh 1PopDNA_sta SettingsDNAStats.ars ss1000.txt

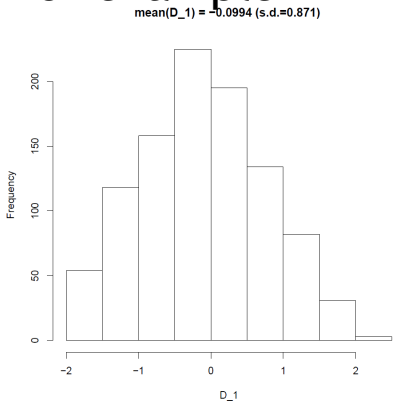
target directory

settings file

output file for summary statistics

4. Look at results in files ss1000.txt
5. Plot results with R for example




| K_1 | tot_K | S_1 | prS_1 | tot_S | D_1 | Pi_1 |
|---|---|---|---|---|---|---|
| 10 | 10 | 156 | 156 | 156 | −0.326022 | 51.5333 |
| 8 | 8 | 85 | 85 | 85 | 0.0216295 | 30.1778 |
| 9 | 9 | 54 | 54 | 54 | −1.43653 | 13.4889 |
| 9 | 9 | 99 | 99 | 99 | 0.705354 | 39.9778 |
| 6 | 6 | 34 | 34 | 34 | −1.38175 | 8.57778 |
| 7 | 7 | 147 | 147 | 147 | 0.801731 | 60.3333 |
| 10 | 10 | 184 | 184 | 184 | −1.65399 | 43.4667 |
| 9 | 9 | 67 | 67 | 67 | −0.502197 | 21.2667 |

# Simulate the coalescent in a stationary population

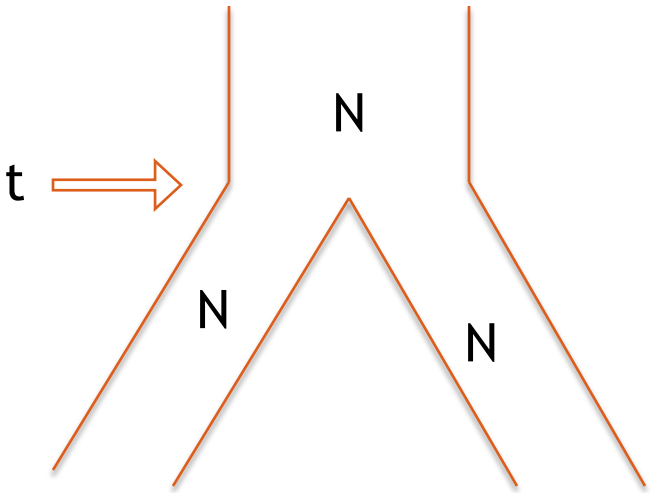
- Examine and use project file 1PopDNA_sta.arp
- Generate 10 coalescent trees with fastsimcoal and output trees (-T option)
- Use FigTree to visualize the resulting coalescent trees (1PopDNA_sta_1_true_trees.trees)
  What can you say about these trees?
- Have a look at the mutation trees (1PopDNA_sta_1_mut_trees.trees)
  How do the compare to the coalescent trees?
- Examine also the resulting Arlequin output files
- Try to see the effect of using different population sizes

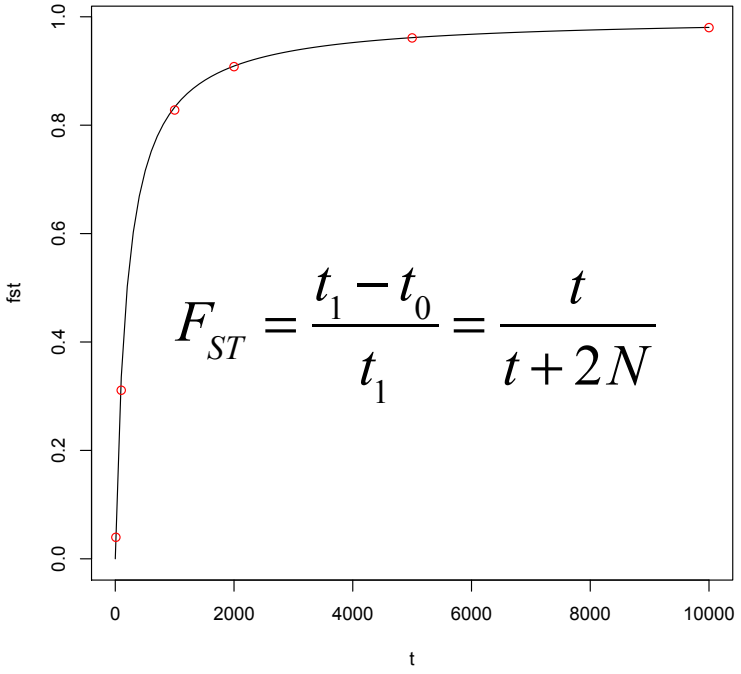
# Simulate the coalescent in expanding and contracting populations

- Examine and use project file 1PopDNA_exp.arp and 1PopDNA_bot.arp
- Generate 10 coalescent trees with fastsimcoal and output trees (-T option)
- Use FigTree to visualize the resulting coalescent trees (*_true_trees.trees)
  What can you say about these trees, as compared to those of a stationary population?

# Population structure exercise

- Simulate (several replicates) a population split using fastsimcoal for different split times (2PopDNAN100TXX.par).
- Calculate Fst for each of your replicate using arlsumstat.
- Calculate the average Fst for each split time (use R)
- Plot the graph of Fst vs t you obtained and compare it with what we expect from the theory.



# Use fastsimcoal to simulate coalescent with recombinations

- Try to simulate the genetic diversity in a short DNA fragment (e.g. 10 Kb) and see the effect of single recombination events on coalescent trees (option -T).
- Try to infer where recombination events occurred.
- Do all recombination events affect the topology, the branch lengths, or the TMRCA?
- If enough time, try to get , by simulation, the distribution of S and $\pi$ in a single population, by studying a segment of 100 Kb with $q=4Nm=40$, and $R=4Nr=40$, and contrast it with what you obtain for the same segment without recombinatio

# Free additional exercises

- Look at the differences in parameter specification between fastsimcoal, ms and macs. Try to use ms.
- Simulate an Isolation with Migration (IM) model and look at the tree depending on the migration rate (2PopIMDNA_sta.par).
- Simulate an island model and verify the Fst vs Nm theoretical result (model_20_0.005_10_100.par).
- Try SPLATCHE.
- Have a look at m4s2 ("modeler for simcoal2") on the web.
- …