CodeMonkeys meeting
Simulation tools for coalescences with recombination

Joe Zhu

2015-Jan-26

Background and Motivation

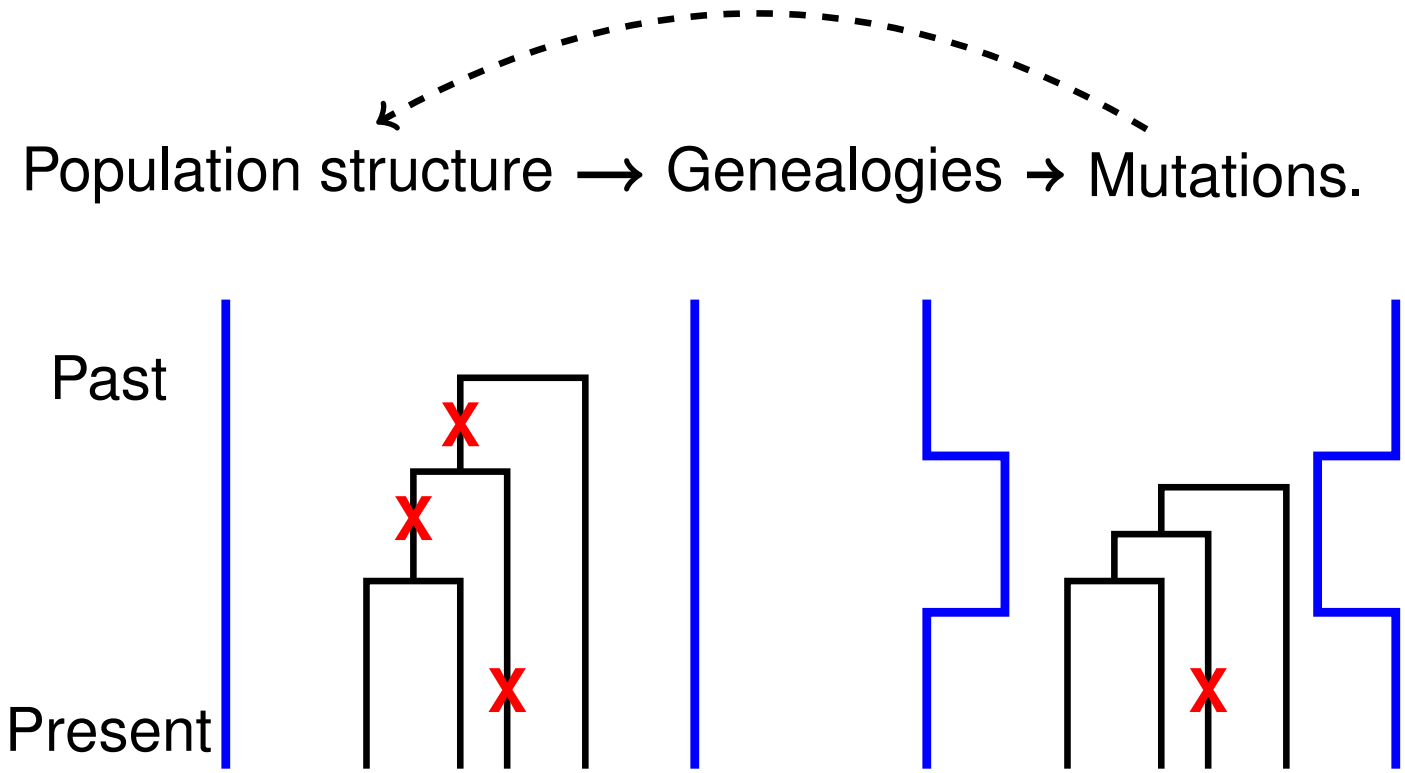
ms

scrm

Existing tools for simulating coalescences with recombination

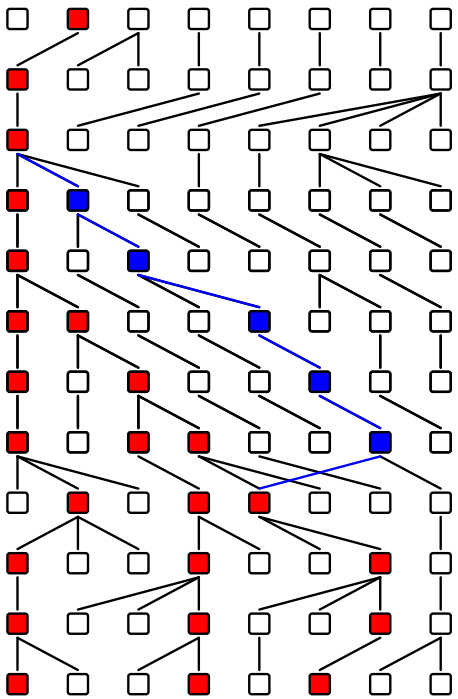
- ms (Hudson, 2002)
- MaCs (Chen et al., 2009)
- fastsimcoal (Excoffier and Foll, 2011)
- cosi2 (Shlyakhter et al., 2014)
- scrm (Staab et al., 2015)
- msprime (Jerome Kelleher)

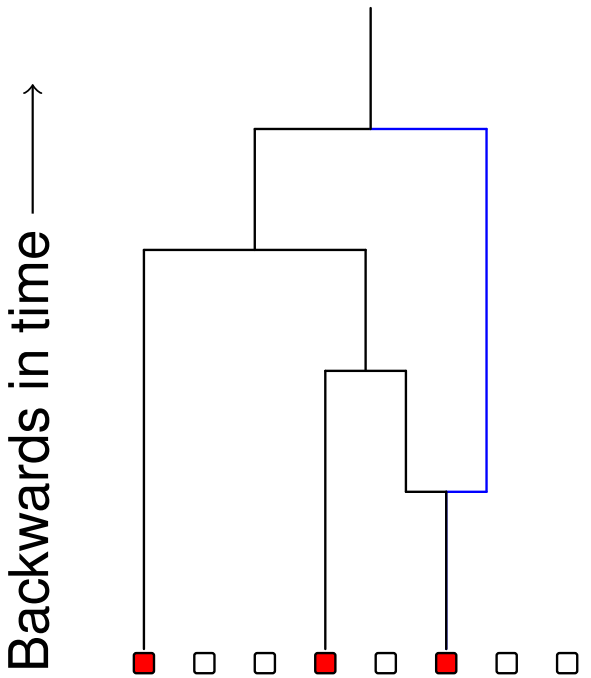
Background



Population structure: blue lines.
Genealogies: black lines.
Mutations: X.

The Ancestral Recombination Graph





$$\delta = \frac{1}{2N_e}\frac{k(k-1)}{2}$$

$$\lambda = \frac{\rho}{4N_e}k$$

ms Input / Output

Samples size

```
ms 3 1 -seed 57659 62331 49571 -T
//
(3:0.819,(1:0.554,2:0.554):0.265);
```

ms Input / Output

Number of repeats

```
ms 3 1 -seed 57659 62331 49571 -T
//
(3:0.819,(1:0.554,2:0.554):0.265);
```

ms Input / Output

User defined randome seed

```
ms 3 1 -seed 57659 62331 49571 -T
//
(3:0.819,(1:0.554,2:0.554):0.265);
```

ms Input / Output

Print Newick formatted tree

```
ms 3 1 -seed 57659 62331 49571 -T
//
(3:0.819,(1:0.554,2:0.554):0.265);
```

ms Input / Output

Simulate segregating sites

```
ms 3 1 -seed 57659 62331 49571 -T -t 5
//
(3:0.819,(1:0.554,2:0.554):0.265);
segsites: 10
positions: 0.0641 0.2586 0.3100 0.4687 0.5998 \
0.6140 0.6364 0.8680 0.8799 0.9221
0110000011
0100110101
1001001000
```

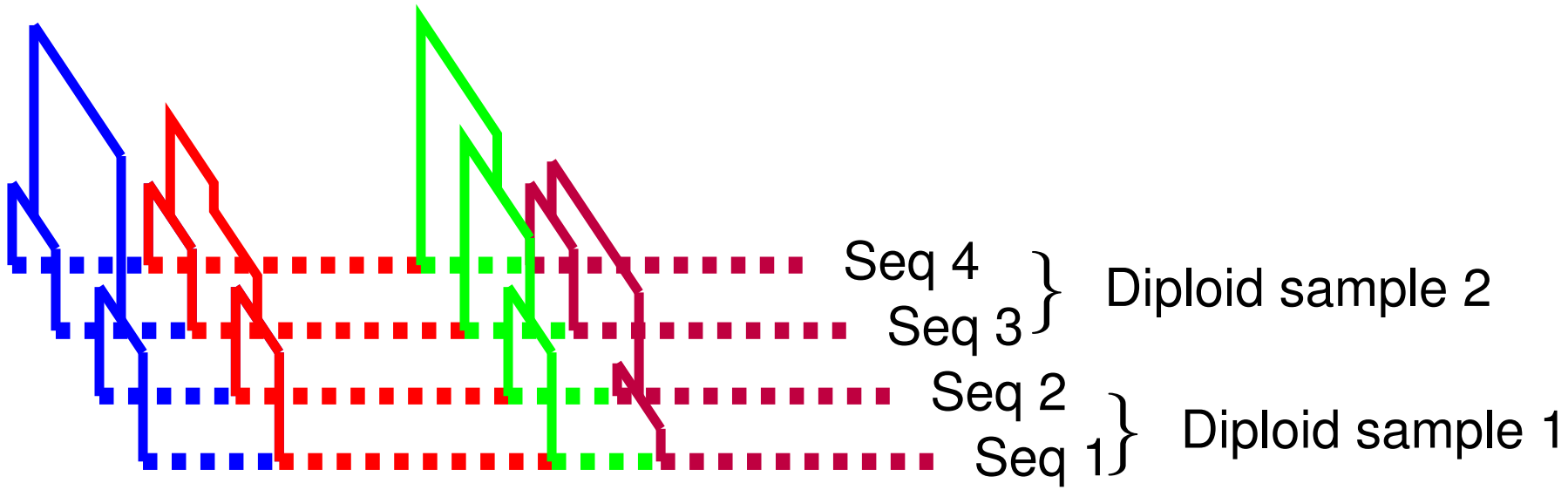
ms Input / Output

Recombination

```
ms 3 1 -seed 57659 62331 49571 -T -t 5 -r 5 1000
//
[441](3:0.211,(1:0.119,2:0.119):0.091);
[218](3:0.586,(1:0.119,2:0.119):0.467);
[71](3:0.749,(1:0.119,2:0.119):0.630);
[225](3:0.234,(1:0.119,2:0.119):0.115);
[9](3:0.749,(1:0.119,2:0.119):0.630);
[8](3:0.783,(1:0.119,2:0.119):0.664);
[10](3:0.905,(1:0.119,2:0.119):0.786);
[18](3:0.427,(1:0.119,2:0.119):0.307);
segsites: 4
positions: 0.0465 0.6674 0.6998 0.8613
0011
1011
0100
```

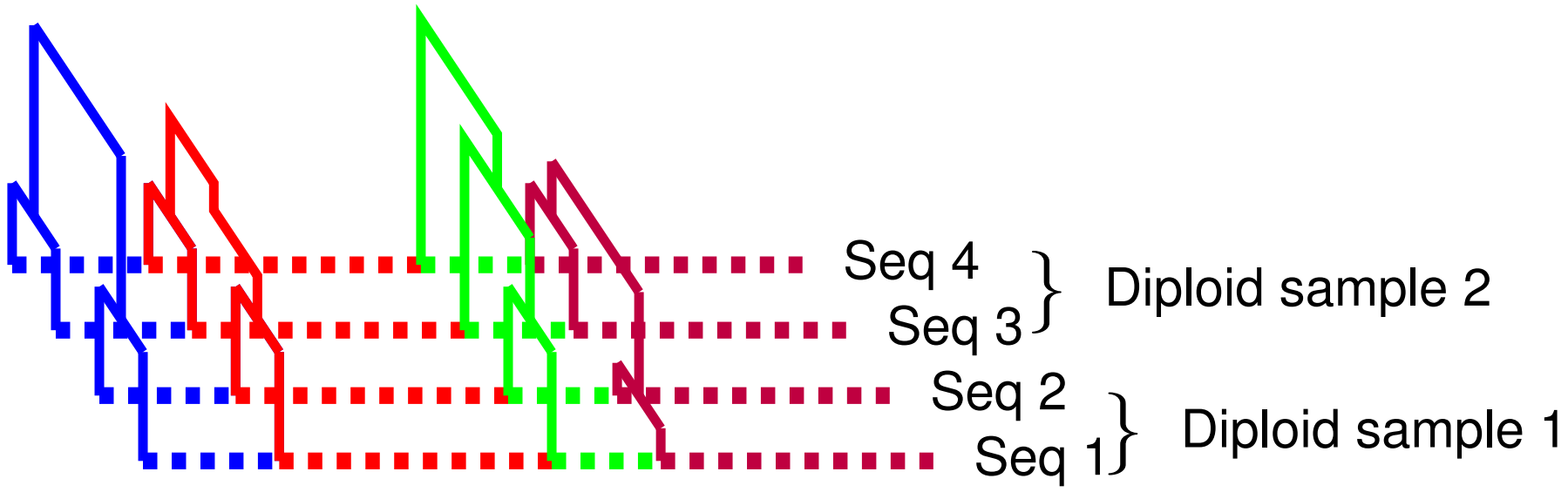
ms Input / Output

```
ms 3 1 -seed 57659 62331 49571 -T -t 5 -r 5 1000
```



ms Input / Output

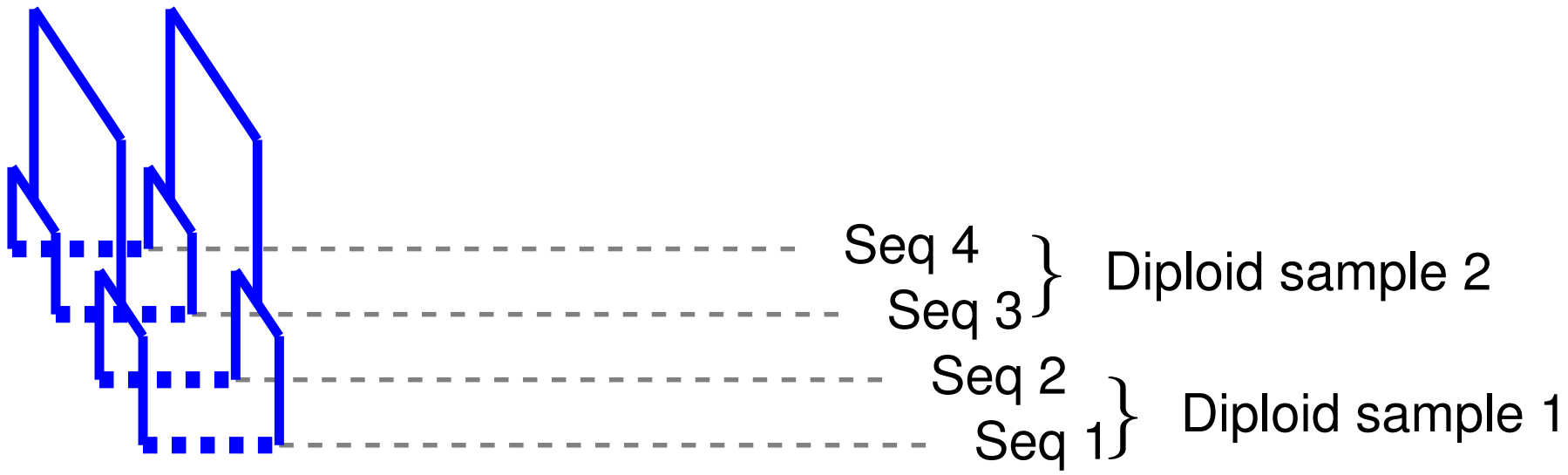
```
ms 3 1 -seed 57659 62331 49571 -T -t 5 -r 5 1000
```




```
scrm 3 1 -seed 57659 62331 49571 -T -t 5 -r 5 1000
```

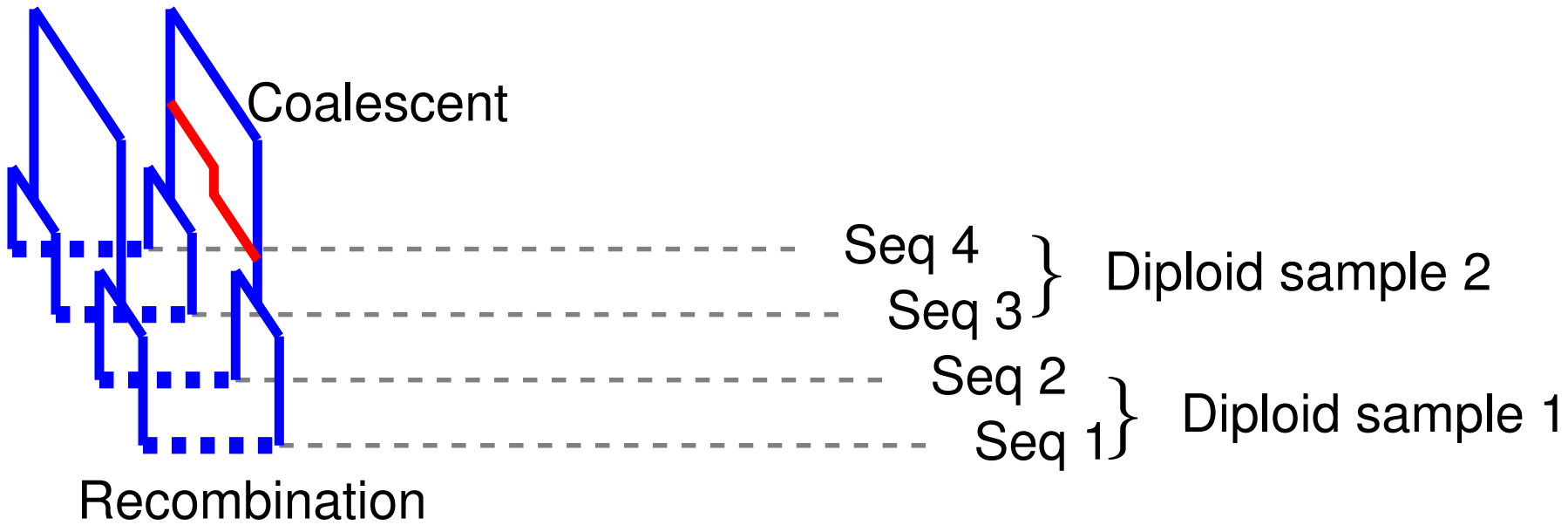
Sequentially Markovian Coalescent (SMC)

```
scrm 3 1 -seed 57659 62331 49571 -T -t 5 -r 5 1000 -l 0
```



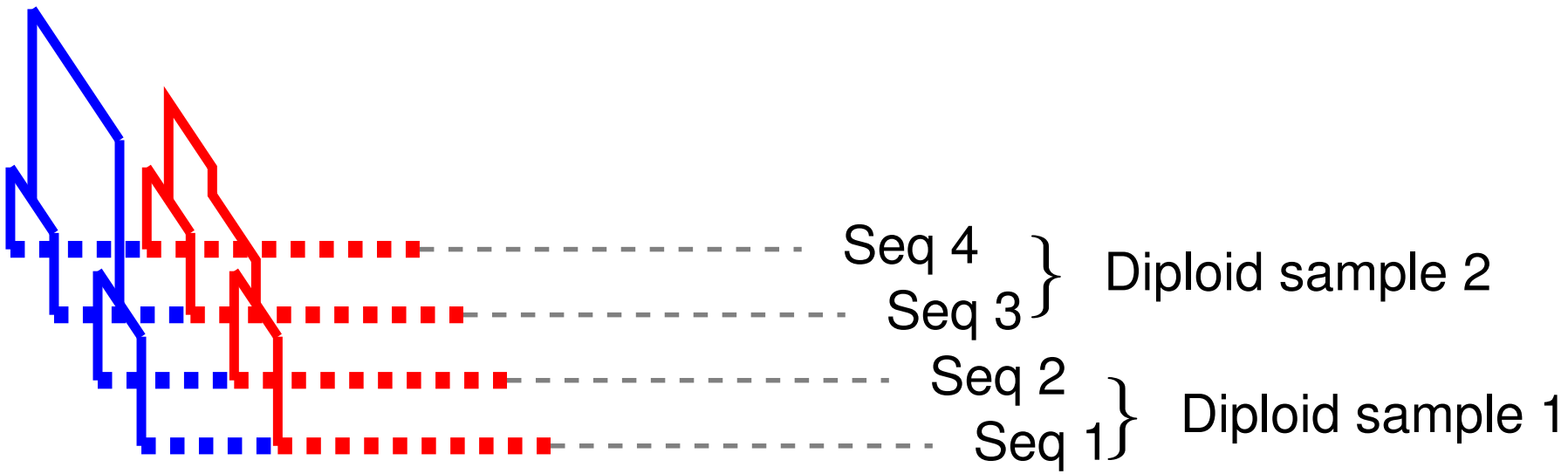
Sequentially Markovian Coalescent (SMC)

```
scrm 3 1 -seed 57659 62331 49571 -T -t 5 -r 5 1000 -l 0
```



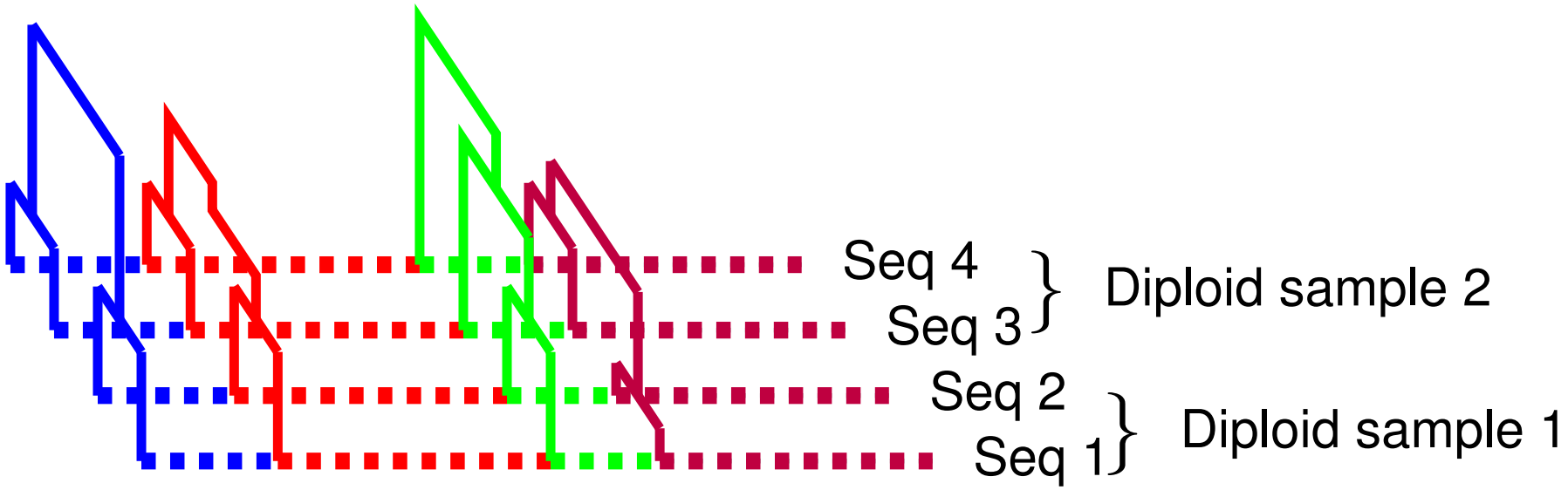
Sequentially Markovian Coalescent (SMC)

```
scrm 3 1 -seed 57659 62331 49571 -T -t 5 -r 5 1000 -l 0
```



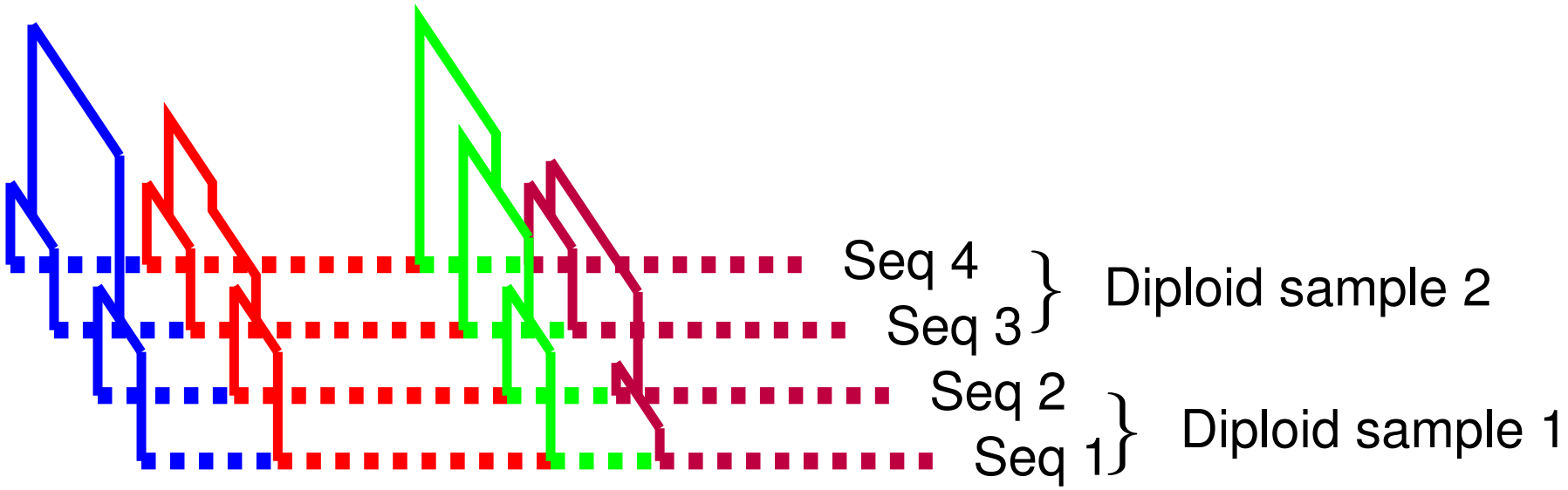
Sequentially Markovian Coalescent (SMC)

```
scrm 3 1 -seed 57659 62331 49571 -T -t 5 -r 5 1000 -l 0
```



Sequentially Markovian Coalescent (SMC)

```
scrm 3 1 -seed 57659 62331 49571 -T -t 5 -r 5 1000 -l 0
```




```
scrm 3 1 -seed 57659 62331 49571 -T -t 5 -r 5 1000 -l 300
```

More command options [1]

- A split model of two populations (A and B) at generation $3 \times 4N_e$. Global population size change at generation $0.4 \times 4N_e$ and $1 \times 4N_e$ to $10.01 \times N_e$ and $0.01 \times N_e$ respectively. A population bottleneck is presented in population B at time $0.25 \times 4N_e$ with the size of $0.2 \times N_e$.
- Sequence length of 100kb, with recombination and mutation rates both equal to 10 ($10/4N_e/1000000$ per generation per base pair).
- Ten haplotypes are sampled from two populations: two from population A and eight individuals from population B.
- Sample 100000 times.

```
 scrm 10 100000 -t 10 -r 10 100000 -I 2 2 8 \
-eN 0.4 10.01 -eN 1 0.01 -en 0.25 2 0.2 -ej 3 2 1 -T
```

[1] Please refer to ms manual:
http://home.uchicago.edu/rhudson1/source/mksamples.html,
https://github.com/scrm/scrm/wiki/Command-Line-Options

More command options [1]

- A split model of two populations (A and B) at generation $3 \times 4N_e$. Global population size change at generation $0.4 \times 4N_e$ and $1 \times 4N_e$ to $10.01 \times N_e$ and $0.01 \times N_e$ respectively. A population bottleneck is presented in population B at time $0.25 \times 4N_e$ with the size of $0.2 \times N_e$.
- Sequence length of 100kb, with recombination and mutation rates both equal to 10 ($10/4N_e/1000000$ per generation per base pair).
- Ten haplotypes are sampled from two populations: two from population A and eight individuals from population B.
- Sample 100000 times.

```
 scrm 10 100000 -t 10 -r 10 100000 -I 2 2 8 \
-eN 0.4 10.01 -eN 1 0.01 -en 0.25 2 0.2 -ej 3 2 1 -T
```

[1] Please refer to ms manual:
http://home.uchicago.edu/rhudson1/source/mksamples.html,
https://github.com/scrm/scrm/wiki/Command-Line-Options

More command options [1]

- A split model of two populations (A and B) at generation $3 \times 4N_e$. Global population size change at generation $0.4 \times 4N_e$ and $1 \times 4N_e$ to $10.01 \times N_e$ and $0.01 \times N_e$ respectively. A population bottleneck is presented in population B at time $0.25 \times 4N_e$ with the size of $0.2 \times N_e$.
- Sequence length of 100kb, with recombination and mutation rates both equal to 10 ($10/4N_e/1000000$ per generation per base pair).
- Ten haplotypes are sampled from two populations: two from population A and eight individuals from population B.
- Sample 100000 times.

```
 scrm 10 100000 -t 10 -r 10 100000 -I 2 2 8 \
-eN 0.4 10.01 -eN 1 0.01 -en 0.25 2 0.2 -ej 3 2 1 -T
```

[1] Please refer to ms manual:
http://home.uchicago.edu/rhudson1/source/mksamples.html,
https://github.com/scrm/scrm/wiki/Command-Line-Options

More command options [1]

- A split model of two populations (A and B) at generation $3 \times 4N_e$. Global population size change at generation $0.4 \times 4N_e$ and $1 \times 4N_e$ to $10.01 \times N_e$ and $0.01 \times N_e$ respectively. A population bottleneck is presented in population B at time $0.25 \times 4N_e$ with the size of $0.2 \times N_e$.
- Sequence length of 100kb, with recombination and mutation rates both equal to 10 ($10/4N_e/1000000$ per generation per base pair).
- Ten haplotypes are sampled from two populations: two from population A and eight individuals from population B.
- Sample 100000 times.

```
 scrm 10 100000 -t 10 -r 10 100000 -I 2 2 8 \
-eN 0.4 10.01 -eN 1 0.01 -en 0.25 2 0.2 -ej 3 2 1 -T
```

[1] Please refer to ms manual:
http://home.uchicago.edu/rhudson1/source/mksamples.html,
https://github.com/scrm/scrm/wiki/Command-Line-Options

More command options [1]

- A split model of two populations (A and B) at generation $3 \times 4N_e$. Global population size change at generation $0.4 \times 4N_e$ and $1 \times 4N_e$ to $10.01 \times N_e$ and $0.01 \times N_e$ respectively. A population bottleneck is presented in population B at time $0.25 \times 4N_e$ with the size of $0.2 \times N_e$.
- Sequence length of 100kb, with recombination and mutation rates both equal to 10 ($10/4N_e/1000000$ per generation per base pair).
- Ten haplotypes are sampled from two populations: two from population A and eight individuals from population B.
- Sample 100000 times.

```
 scrm 10 100000 -t 10 -r 10 100000 -I 2 2 8 \
-eN 0.4 10.01 -eN 1 0.01 -en 0.25 2 0.2 -ej 3 2 1 -T
```

[1] Please refer to ms manual:
http://home.uchicago.edu/rhudson1/source/mksamples.html,
https://github.com/scrm/scrm/wiki/Command-Line-Options

More command options [1]

- A split model of two populations (A and B) at generation $3 \times 4N_e$. Global population size change at generation $0.4 \times 4N_e$ and $1 \times 4N_e$ to $10.01 \times N_e$ and $0.01 \times N_e$ respectively. A population bottleneck is presented in population B at time $0.25 \times 4N_e$ with the size of $0.2 \times N_e$.
- Sequence length of 100kb, with recombination and mutation rates both equal to 10 ($10/4N_e/1000000$ per generation per base pair).
- Ten haplotypes are sampled from two populations: two from population A and eight individuals from population B.
- Sample 100000 times.

```
 scrm 10 100000 -t 10 -r 10 100000 -I 2 2 8 \
-eN 0.4 10.01 -eN 1 0.01 -en 0.25 2 0.2 -ej 3 2 1 -T
```

[1] Please refer to ms manual:
http://home.uchicago.edu/rhudson1/source/mksamples.html,
https://github.com/scrm/scrm/wiki/Command-Line-Options

More output options[2]

- Option "-L" reports TMRCA and tree length:

```
scrm 4 1 -r 1 100 -L
```

- Option "-oSFS" reports the site frequency spectrum :

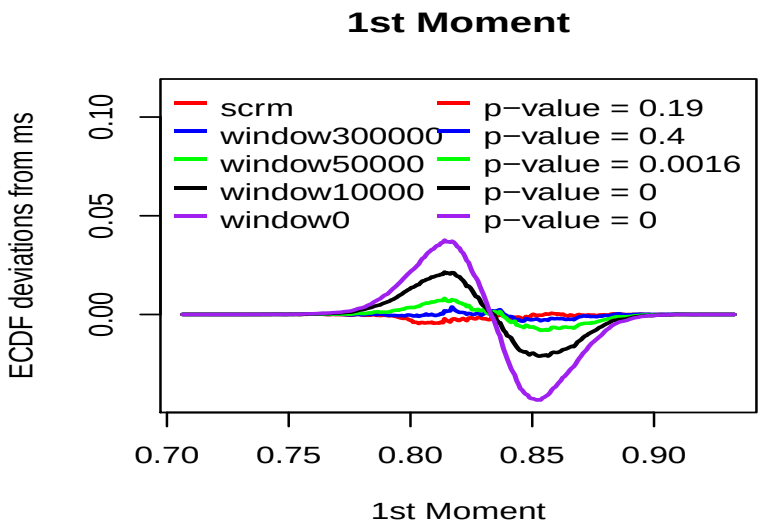
```
scrm 4 1 -t 5 -oSFS
```

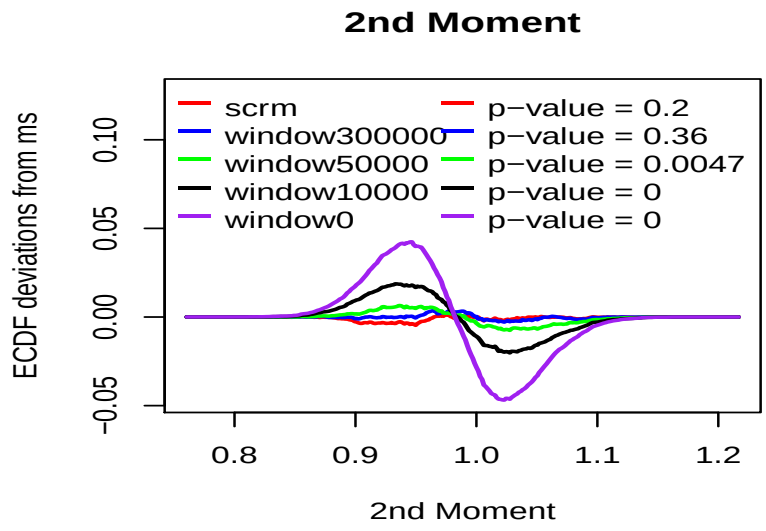
- Option "-O" reports local trees in JSON formatted strings:

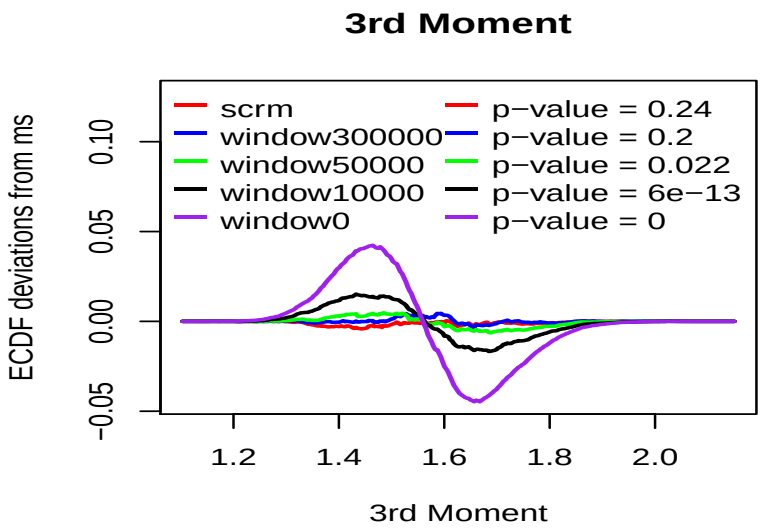
```
scrm 4 1 -r 1 100 -seed 16 -O
{"length":70, "parents":[5,6,7,5,6,7,0], \
"node_times":[0,0,0,0,0.328647,0.559424,0.88458]}
{"length":30, "parents":[5,7,5,6,6,7,0], \
"node_times":[0,0,0,0,0.0700593,0.328647,0.559424]}
```

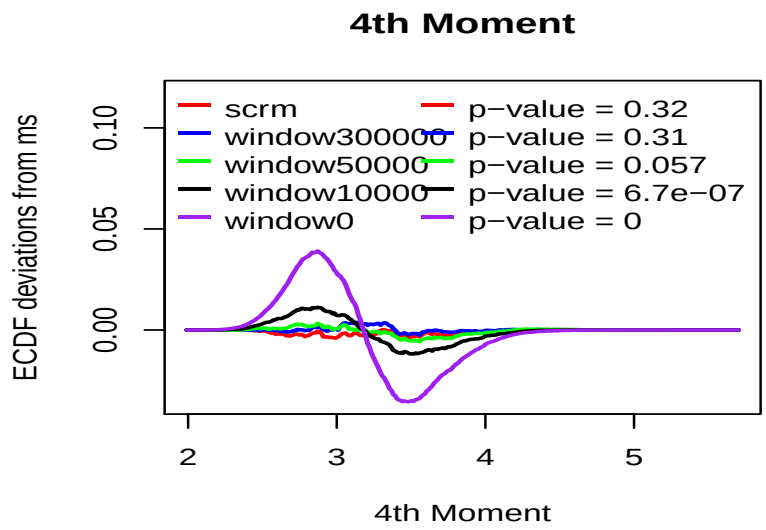
[2] Please refer to https://github.com/scrm/scrm/wiki/Output

scrm vs ms

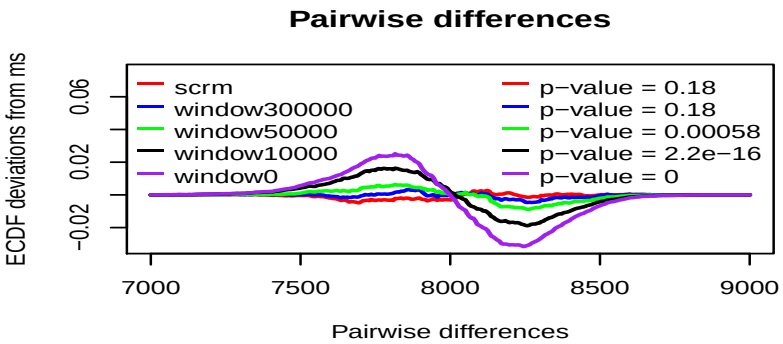

1st Moment
ECDF deviations from ms
0.00
0.05
0.10
scrm
window300000
window50000
window10000
window0
p−value = 0.19
p−value = 0.4
p−value = 0.0016
p−value = 0
p−value = 0
0.70
0.75
0.80
0.85
0.90
1st Moment

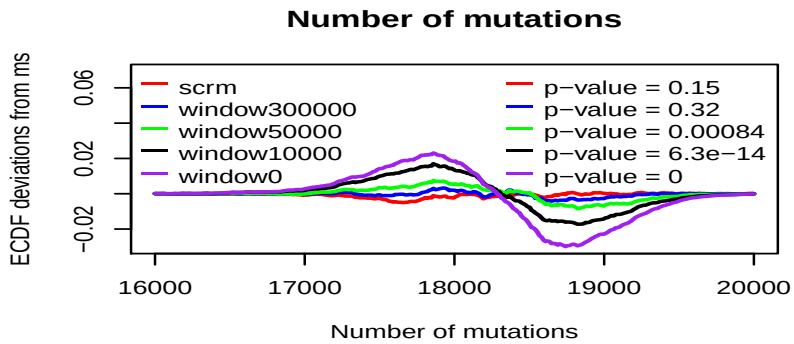

2nd Moment
ECDF deviations from ms
−0.05
0.00
0.05
0.10
scrm
window300000
window50000
window10000
window0
p−value = 0.2
p−value = 0.36
p−value = 0.0047
p−value = 0
p−value = 0
0.8
0.9
1.0
1.1
1.2
2nd Moment

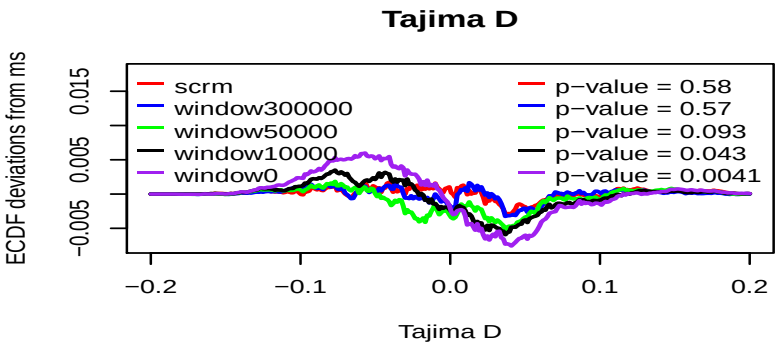

3rd Moment
ECDF deviations from ms
−0.05
0.00
0.05
0.10
scrm
window300000
window50000
window10000
window0
p−value = 0.24
p−value = 0.2
p−value = 0.022
p−value = 6e−13
p−value = 0
1.2
1.4
1.6
1.8
2.0
3rd Moment

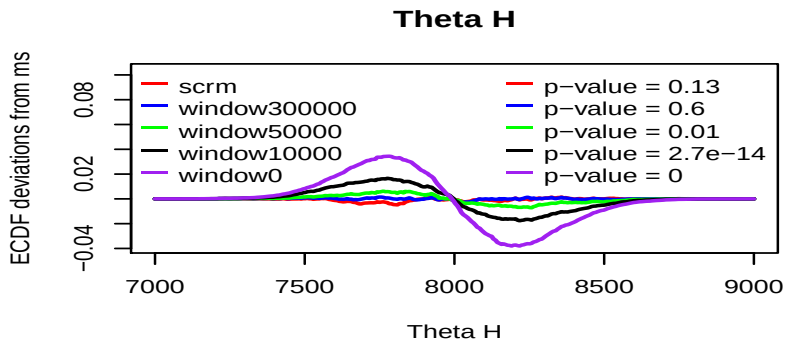

4th Moment
ECDF deviations from ms
0.00
0.05
0.10
scrm
window300000
window50000
window10000
window0
p−value = 0.32
p−value = 0.31
p−value = 0.057
p−value = 6.7e−07
p−value = 0
2
3
4
5
4th Moment

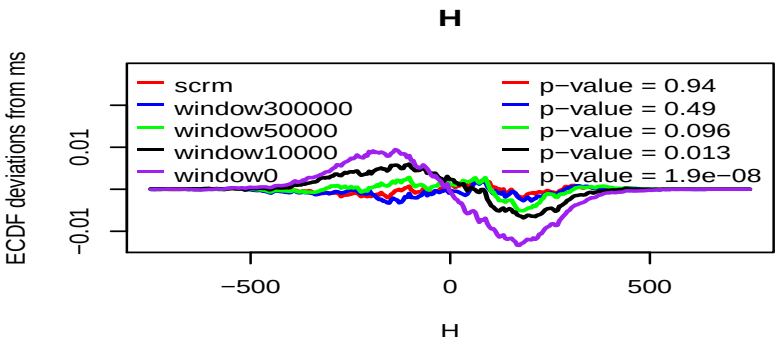
scrm vs ms


Pairwise differences
ECDF deviations from ms
scrm
window300000
window50000
window10000
window0
p−value = 0.18
p−value = 0.18
p−value = 0.00058
p−value = 2.2e−16
p−value = 0
−0.02
0.02
0.06
7000
7500
8000
8500
9000
Pairwise differences

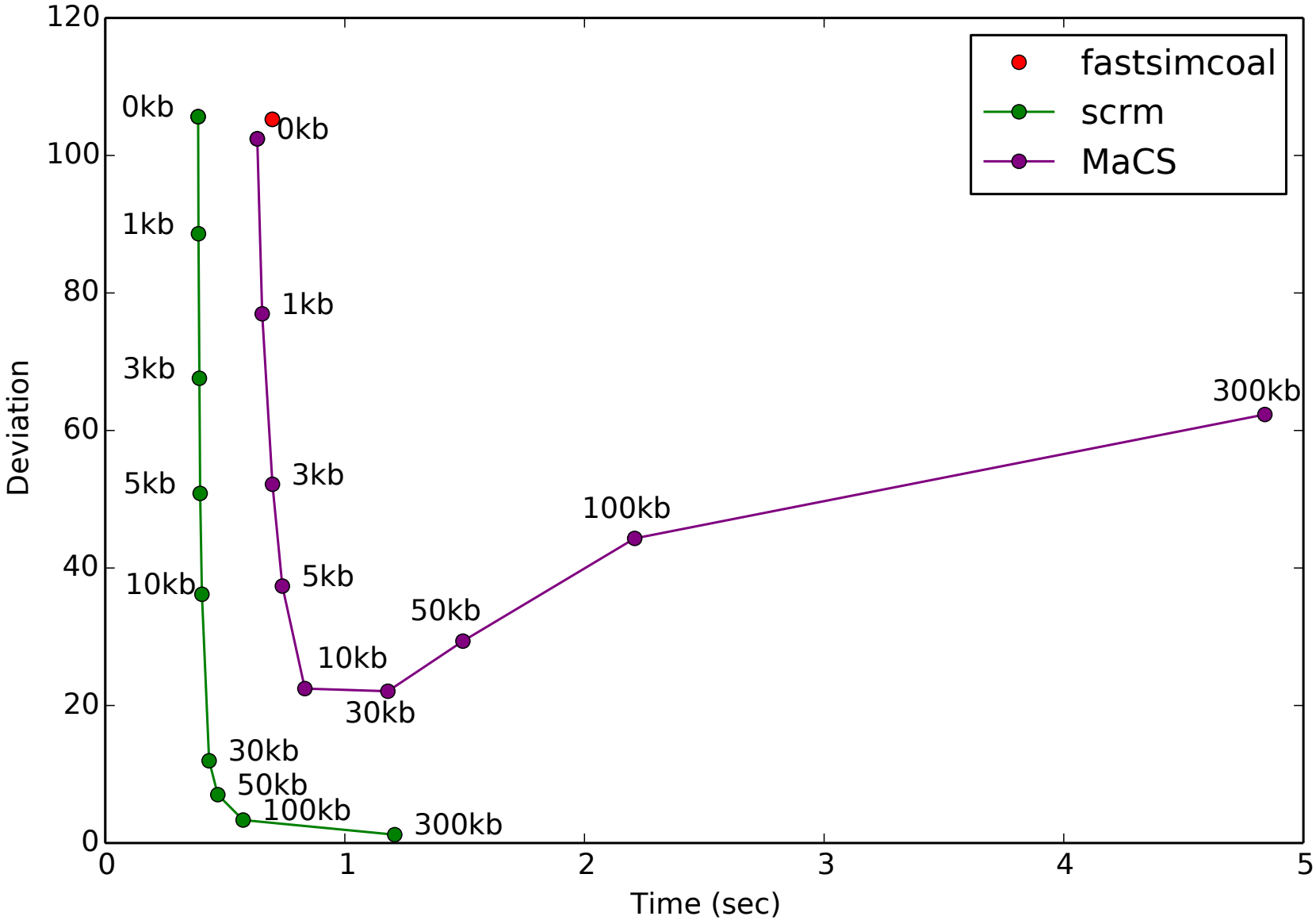

Number of mutations
ECDF deviations from ms
scrm
window300000
window50000
window10000
window0
p−value = 0.15
p−value = 0.32
p−value = 0.00084
p−value = 6.3e−14
p−value = 0
−0.02
0.02
0.06
16000
17000
18000
19000
20000
Number of mutations


Tajima D
ECDF deviations from ms
scrm
window300000
window50000
window10000
window0
p−value = 0.58
p−value = 0.57
p−value = 0.093
p−value = 0.043
p−value = 0.0041
−0.005
0.005
0.015
−0.2
−0.1
0.0
0.1
0.2
Tajima D

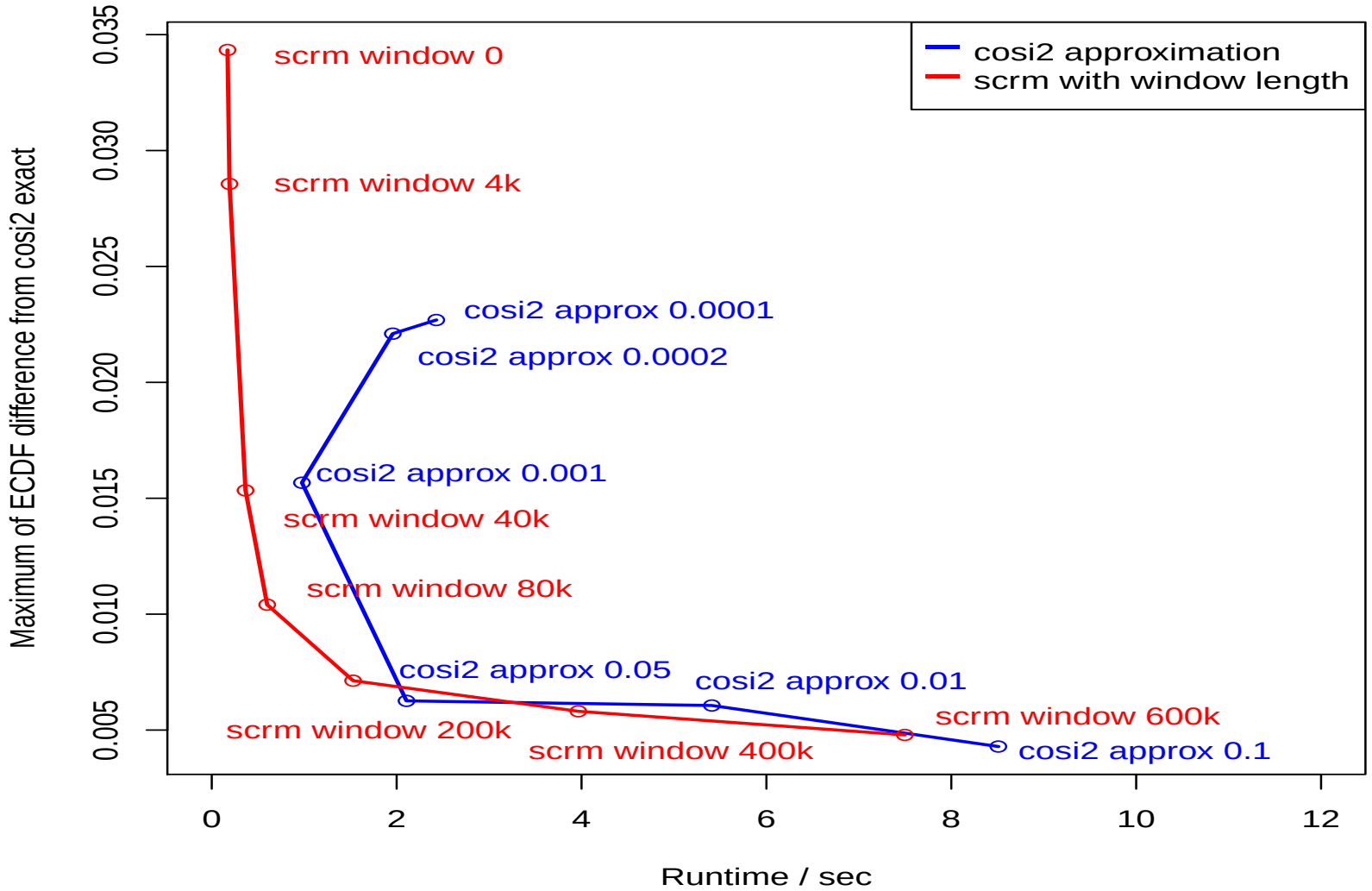

Theta H
ECDF deviations from ms
scrm
window300000
window50000
window10000
window0
p−value = 0.13
p−value = 0.6
p−value = 0.01
p−value = 2.7e−14
p−value = 0
−0.04
0.02
0.08
7000
7500
8000
8500
9000
Theta H


H
ECDF deviations from ms
scrm
window300000
window50000
window10000
window0
p−value = 0.94
p−value = 0.49
p−value = 0.096
p−value = 0.013
p−value = 1.9e−08
−0.01
0.01
−500
0
500
H

scrm vs fastsimcoal vs MaCs



scrm vs cosi2



Download and Install[3]

- Download
  - `git clone https://github.com/scrm/scrm.git`
  - `wget https://scrm.github.io/releases/scrm-1.3.2.tar.gz`
    `tar -xf scrm-1.3.2.tar.gz`
- Install

```
./bootstrap
make install
```

- Open `R`, then type

```
install.packages('scrm')
```

[3] Please refer to `https://scrm.github.io/` and `https://github.com/scrm/scrm/wiki/Installation`.

Use scrm as part of your own project[4]

- Include scrm as a submodule.
- Doxygen documentation.
- Unittest and continuous integration.
- A testing openmp version of scrm.

[4] Please refer to `https://github.com/shajoezhu/CodeMonkey_scrm`.

Reference

Chen, G. K., P. Marjoram, and J. D. Wall (2009). Fast and flexible simulation of dna sequence data. *Genome Res. 19*, 136–142.

Excoffier, L. and M. Foll (2011). fastsimcoal: A continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics 27*(9), 1332–1334.

Hudson, R. R. (2002). Generating samples under a Wright–Fisher neutral model. *Bioinformatics 18*, 337–338.

Lunter, G. (2012). A new Markovian approximation to the coalescent with recombination. Society for Molecular Biology & Evolution. http://imgpublic.mci-group.com/ie/PCO/AllAbstracts_FINAL.pdf. Aug 14, 2013.

Shlyakhter, I., P. C. Sabeti, and S. F. Schaffner (2014). Cosi2: an efficient simulator of exact and approximate coalescent with selection. *Bioinformatics 30*(23), 3427–3429.

Staab, P. R., S. Zhu, D. Metzler, and G. Lunter (2015). scrm: efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics*.