

**msdup**  
(given time of the duplication)  
**readme file**

Kosuke M. Teshima and Hideki Innan

kmteshima@kyudai.jp & innan\_hideki@soken.ac.jp

The Graduate University for Advanced Studies, Hayama, Kanagawa 240-0193, Japan.  
Kyushu University, Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan

last update, April 1, 2015

## 1 msdup

**msdup** is a simulation program to generate patterns of single nucleotide polymorphism (SNP) data in a region that involves duplicates and their flanking regions (Teshima and Innan, 2012). The software was developed by modifying the commonly used Hudson's **ms** simulator (Hudson, 1990). Please note that our program was modified from the previous version of **ms**.

**msdup** generates polymorphism data conditional on the time of the ancestral duplication event. **msdup** reads the trajectory of the frequency of the duplicated chromosome and conducts the coalescent simulation given trajectory. The format of the trajectory file is described in section 1.2.1. **msdup** can incorporate the effect of selection acting on the duplicate chromosomes, as well as mutation, recombination and interlocus gene conversion.

### 1.1 Download and compilation

All files are included in the `source_msdup_given_dup_time.tar.gz` file, which can be downloaded at <http://www.sendou.soken.ac.jp/esb/innan/InnanLab>. The source code of the program is written in C and this program is intended to be run on UNIX, or a UNIX-like operating systems, such as Linux or Mac OS X.

Download the tar file onto your machine and extract it with:

```
tar xzvf source_msdup_given_dup_time.tar.gz
```

After extraction, change the directory by typing:

```
cd source_msdup_given_dup_time
```

Then, compile the program by typing:

```
gcc -o msdup ms_dup.c streec.c rand_mt32.c -lm
```



## 1.2 The basic command line

The following command line shows the usage of `msdup`.

```
./msdup nsam howmany
        -t 4N0u
        -r 4N0r nsites
        -g 4N0g ℓtract
        -d t0
        -l ℓleft ℓdup ℓmid ℓdup ℓright
```

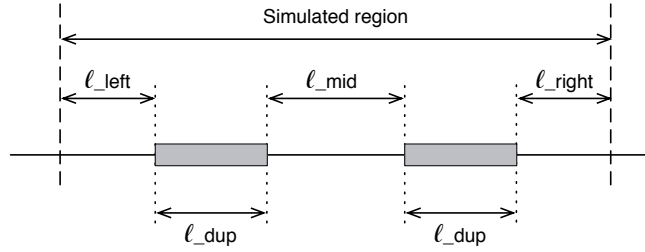


Figure 1: The length of each region. Note that (the length of the simulated region) =  $nsites = \ell_{left} + \ell_{dup} + \ell_{mid} + \ell_{dup} + \ell_{right}$ .

The following table is a summary of arguments that must be specified to run `msdup`.

switch	argument	
	<i>nsam</i>	number of chromosomes in the sample.
	<i>howmany</i>	number replication.
-t	4N <sub>0</sub> u	population mutation parameter per simulated region.
-r	4N <sub>0</sub> r	population recombination parameter per simulated region.
	<i>nsites</i>	the length of the simulated region (bp).
-g	4N <sub>0</sub> g	4N <sub>0</sub> g is the initiation rate of gene conversion per $\ell_{dup}$ .
	$\ell_{tract}$	average length of a conversion tract (bp).
-l	$\ell_{left}$	the length of the 5' flanking region.
	$\ell_{dup}$	the length of the duplicated region (original copy).
	$\ell_{mid}$	the length between the original and duplicated regions.
	$\ell_{dup}$	the length of the duplicated region (duplicate copy, this should be equal to the length of the original copy).
	$\ell_{right}$	the length of the 3' flanking region.
-d	t <sub>0</sub>	time of the ancestral duplication event in units of 4N.



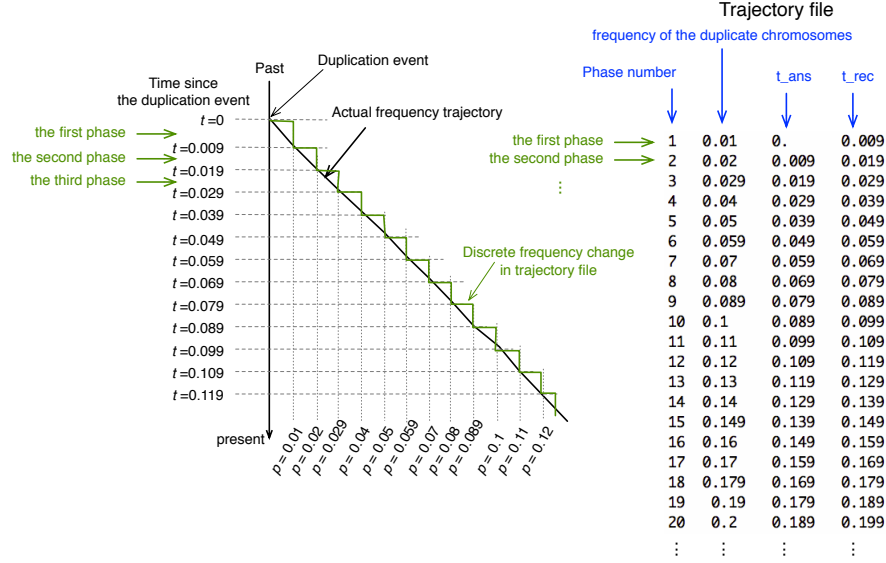


Figure 2: Format of the trajectory file.

### 1.2.1 Format of the trajectory file

In `msdup`, the frequency change was approximated by the discrete frequency change. The trajectory is considered as a consecutive time phases. Within each time phase, the frequency of the duplicated chromosomes stays constant. The illustration of the frequency change and the format of the trajectory file was shown in Figure 2.

The trajectory file consists of four columns: phase number, frequency, beginning time of the phase, and ending time of the phase. The phase number should start from 1. The beginning time of the first phase should start from 0. The ending time of the phase and the beginning time of the next phase should be the same. An example trajectory file is involved in `source_msdup_given_dup_time` directory.

### 1.2.2 example command line

The following is an example command line:

```
./msdup 20 100 -t 13 -r 13 13000 -g 1 100
-1 1000 1000 9000 1000 1000 -d 10
```

Note the trajectory file should be in the same directory. In this setting, the program will output polymorphism data in a 13,000 bp region. The lengths of regions are: (left, original copy, middle, duplicated copy, right) = (1000, 1000, 9000, 1000, 1000). The sample size is  $nsam = 20$ , and  $4N_0u = 4N_0r = 13$  (per simulated region). 100 replications are performed. The ancestral duplication event was  $10 \times 4N$  generations ago.





Figure 3: An example output of msdup.

Let  $c$  be the per site interlocus conversion rete and  $C = 4Nc$ .  $C$  is given by  $C = G \times \ell$ , where  $G = 4N_0g$  is the initiation rate of gene conversion and  $\ell$  is the average tract length. In the above command line, the ratio of the conversion rate to mutation rate can be calculated as

$$c/\mu = \frac{1 \times 100}{13 \times (1000/13000)} = 100$$

### 1.3 The output

An output of the example command line in Section 1.2.2,

```

./msdup 20 100 -t 13 -r 13 13000 -g 1 100
-l 1000 1000 9000 1000 1000 -d 10

```

is shown in Figure 3.

The first line shows the command line you have just typed. The output of each replication starts with a line with ‘//’, followed by “replication:” and



the replication number. In the second line, the number of segregating sites are shown. The first number is the total number of segregating sites across the simulated region. The following five numbers mean the numbers of segregating sites in 5', original copy, middle, duplicated copy and 3' regions, respectively. The third and the following lines show the relative positions of segregating sites. The state of SNP at each segregating site is show with 0 and 1. 0 means ancestral and 1 means derived allele. Spaces are inserted to show boundaries between regions. For single-copy chromosomes, the region of the duplicated copy is filled with "-", which means there is no duplicated copy on the chromosome. Note that at least one site is printed out even when there is no segregating sites in the region. In such cases, all samples have the same allele 0 at the site (eg. 3' region of Figure 3).

## 1.4 Some notes

### 1.4.1 Sample configuration

When the age of the duplicate is young and the duplicated chromosomes are not fixed in a population, the number of single-copy chromosomes and duplicate chromosomes are randomly assigned based on the frequency of the single-copy and duplicate chromosomes. Let  $p$  be the frequency of the duplicated chromosome at  $t = 0$  (present). For each sample, random number is generated. A chromosome is assigned to single-copy if ( $\text{rand} < 1 - p$ ), otherwise assigned to duplicate chromosome.

You can change the assignment scheme by modifying the code (l.170-) in `segtre_mig()` function of `streec.c`.

### 1.4.2 Random number generator

Mersenne Twister (32bit version) is used to generate random numbers (Matsumoto and Nishimura, 1998).

## References

- Hudson, R. (1990). Gene genealogies and the coalescent process. In D. Futuyma and J. Antonovics, editors, *Oxford Surveys in Evolutionary Biology*, volume 7, pages 1–44. Oxford University Press.
- Matsumoto, M. and Nishimura, T. (1998). Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, **8**(1), 3–30.
- Teshima, K. M. and Innan, H. (2012). The coalescent with selection on copy number variants. *Genetics*, **190**(3), 1077–1086.



## Release note

- March, 2015:
  - Readme file updated.
- Jan, 2012:
  - The first release.