BIOINFORMATICS APPLICATIONS NOTE Vol. 19 no. 2 2003
Pages 289–290

Simulating haplotype blocks in the human genome

David Posada* and Carsten Wiuf

Variagenics Inc., 60 Hampshire St, Cambridge, MA 02139-1548, USA

Received on June 26, 2002; revised on August 7, 2002; accepted on August 10, 2002

ABSTRACT

Summary: A bioinformatic tool was written to simulate haplotypes and SNPs under a modified coalescent with recombination. The most important feature of this program is that it allows for the specification of non-homogeneous recombination rates, which results in the formation of the so-called 'haplotype blocks' of the human genome. The program also implements different mutation models and flexible demographic histories. The samples generated can be very useful to better understand the architecture of the human genome and to investigate its impact in association studies searching for disease genes.

Availability: The SNPsim package is available at http://www.evolgenics.com/software

Contact: dposada@evolgenics.com

GENOME BLOCKS

Recent studies suggest that most single nucleotide polymorphisms (SNPs) in the human genome are organized in regions with high levels of linkage disequilibrium and little haplotype diversity (Daly *et al.*, 2001; Gabriel *et al.*, 2002; Jeffreys *et al.*, 2001; Patil *et al.*, 2001). These regions, called haplotype blocks, are the result of population history and especially, non-homogeneous recombination. The interest in this structure is in that such blocks can be tagged with a few SNPs, enormously facilitating association studies searching for genes underlying complex diseases (Johnson *et al.*, 2001). Wiuf and Posada (unpublished data) have devised a population genetic model of recombination hotspots, based on the coalescent with recombination, that stochastically generates genetic samples with 'haplotype block' structure. Such models are expected to be very useful to understand and interpret genomic variation within populations (Hudson, 2002). Here we describe a program, SNPsim, written in C that implements this model. This computational tool should prove very useful to the study of the statistical properties of haplotype blocks and their relevance in our understanding of the human genome.

*To whom correspondence should be addressed.

SIMULATING HAPLOTYPE BLOCKS

The model implemented here is an extension of the coalescent with recombination (Hudson, 1983). The program simulates any number of samples. Given an expected number of recombination hotspots, a background homogeneous recombination rate and a hotspot recombination rate, the program starts by choosing the exact position and number of recombination hotspots for a particular sample from a Poisson distribution. Next, the recombination rates for each site along the region of interest are specified by simulating recombination events according to a Normal distribution around the hotspot center. Other recombination events coming from recombinational hotspots centered outside the region of interest are also considered. This fast simulation results in different recombination rates for different sites along the region (hotspots and coldspots). Given these recombination rates and other parameters like the effective population size (N) and growth rate, random genealogies are produced that describe the history of different portions of the region of interest. Complex demographic histories can be implemented by defining demographic periods in which population sizes augment, reduce, or remain constant. Time is scaled in units of $2N$ generations. Mutations can be placed upon the genealogies under a biallelic infinite-sites mutation model or under a biallelic Jukes–Cantor mutation model (JC; Jukes and Cantor, 1969) that allows for recurrent mutations). In a given sample, different sites can evolve under one of these models by setting up a parameter that represents the expected proportion of sites under each model. Different mutation rates can be specified for each model. In addition, there is the possibility of conditioning the simulations to produce samples with a fixed number of SNPs. The result is a sample of haplotypes, which are then randomly combined to form diploid genotypes. SNPsim offers very flexible simulation schemes, and the user can specify many options in the command line or in a parameter file. Detailed documentation is included in the SNPsim package.

```
SNPsim  -  Tue Jun 18 12:50:17 2002
-n5 -s20 -l1000 -e1000 -d0 -b0.0e+00 -r1.0e-06 -k1.0e-04 -h2 -v1.0e+00
-u1.0e-06 -p0.0e+00 -w1.0e+00 -y1 -g1000 -m1024404616
4
10 1000
10
 5 275 461 492 500 530 599 724 845 997
00 00 10 10 10 00 00 00 00 00
00 00 11 01 01 01 00 00 01 00
00 00 11 11 11 10 00 00 10 10
00 00 01 01 01 01 00 01 01 00
00 00 01 01 00 00 00 00 00 00
00 00 11 00 00 00 00 00 00 00
01 01 11 11 01 01 01 01 01 00
00 00 10 00 00 00 00 00 00 00
00 00 10 10 10 00 00 00 10 00
00 00 11 11 00 00 00 00 00 00
19
266 288 301 329 332 383 397 454 470 524 581 582 608 615 646 658 703 840 985
01 00 00 01 01 00 01 00 01 00 01 00 01 01 00 01 01 01 01
00 00 00 00 00 00 11 00 11 00 00 00 00 00 00 00 00 00 00
11 11 00 00 00 00 11 00 11 00 00 10 00 00 00 00 00 00 00
11 01 10 10 10 00 11 00 11 00 00 00 10 10 10 10 10 10 10
00 00 00 00 00 00 11 00 11 00 00 00 00 00 00 00 00 00 00
01 00 00 00 01 00 11 00 11 01 00 01 00 00 00 01 01 01 01
00 00 00 00 00 00 01 00 01 00 00 00 00 00 00 00 00 00 00
10 10 00 00 10 00 11 00 11 00 00 10 00 00 00 10 00 00 00
00 00 00 00 00 00 10 00 10 00 00 00 00 00 00 00 00 00 00
11 00 00 01 11 01 11 01 11 10 00 00 11 00 00 11 11 11 11
12
20 142 181 345 347 404 456 469 495 747 789 887
00 01 01 00 00 00 00 00 11 01 00 11
11 00 00 00 00 00 00 00 00 00 00 00
10 00 00 00 00 00 00 00 00 00 00 11
11 00 00 00 00 00 00 00 00 00 00 00
01 10 10 10 00 00 10 00 10 10 10 00
11 00 00 00 00 00 00 00 00 00 00 00
11 00 00 00 00 00 00 00 00 00 00 00
00 11 11 00 00 00 11 10 11 11 11 00
10 00 00 00 01 01 01 00 00 01 01 00
10 00 00 00 00 00 00 00 00 00 00 01
18
13 44 81 251 329 355 477 568 584 594 610 644 755 766 793 953 966 984
11 11 11 01 11 01 00 00 11 00 11 11 01 01 01 01 10 11
10 10 10 00 10 00 00 10 10 01 10 10 00 00 00 00 10 10
01 01 00 00 00 00 00 00 00 11 00 00 00 00 00 00 01 01
00 00 00 00 00 00 00 00 00 11 00 00 00 00 00 00 00 00
01 01 01 00 00 00 00 00 00 11 00 00 00 00 00 00 01 01
10 10 00 00 00 00 00 00 00 11 00 00 00 00 00 00 11 11
11 11 10 10 10 10 10 00 10 01 10 10 00 00 10 00 11 11
00 00 00 00 00 00 00 00 00 11 00 00 00 00 00 00 01 01
00 00 00 00 00 00 00 00 00 11 00 00 00 00 00 00 01 01
```

Fig. 1. Output from SNPsim. Four samples of 1000 bp from 10 individuals were simulated. For each sample individuals are in rows and SNPs in columns.

OUTPUT

The user can control the output, which consists of different files and information on the screen. By default, the program produces a file with diploid SNP genotypes that includes a header with the date and time of the simulation, the value of the parameters in command line style, the number of samples and individuals per sample, and the length of the region simulated. For each sample, the number and position of SNPs is indicated before the description of the individual genotypes (Figure 1). The data is biallelic, 0 being the ancestral and 1 the derived states. The program can also print different files with the complete haplotypes for each individual, all genotypes including monomorphic sites, times and branches for each genealogy, the genealogies, and the mutation model assigned to each site. The user also has control of the output on the screen, with the option of obtaining information for each sample along the simulation. At the end of the simulation, a summary is printed to the screen describing the values specified for the different parameters, the mean and number of recombination and mutation events observed, and some expectations for these quantities.

OPERATIVE SYSTEMS

SNPsim is written in ANSI C, and it has been compiled without problems in Macintosh (OS 9 and OS X), Windows XP and Linux.

REFERENCES

Daly,M.J. *et al.* (2001) High-resolution haplotype structure in the human genome. *Nature Genet.*, **29**, 229–232.

Gabriel,S.B. *et al.* (2002) The structure of haplotype blocks in the human genome. *Science*, **23**, 23.

Hudson,R.R. (1983) Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.*, **23**, 183–201.

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

Jeffreys,A.J., Kauppi,L. and Neumann,R. (2001) Intensively punctuated meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genet.*, **29**, 217–222.

Johnson,G.C. *et al.* (2001) Haplotype tagging for the identification of common disease genes. *Nature Genet.*, **29**, 233–237.

Jukes,T.H. and Cantor,C.R. (1969) In Munro,H.M. (ed.), *Evolution of protein molecules*, Mammalian Protein Metabolism, Academic Press, New York, pp. 21–132.

Patil,N. *et al.* (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, **294**, 1719–1723.