Genetics and population analysis

msHOT: modifying Hudson's ms simulator to incorporate crossover and gene conversion hotspots

Garrett Hellenthal[1,*] and Matthew Stephens[2]

[1]Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK and [2]Department of Statistics, University of Chicago, 5734 S. University Avenue, Eckhart 126, Chicago, IL 60637, USA

ABSTRACT

Summary: We have incorporated both crossover and gene conversion hotspots into an existing coalescent-based program for simulating genetic variation data for a sample of chromosomes from a population.

Availability: The source code for msHOT is available at http://home.uchicago.edu/~rhudson1, along with accompanying instructions.

Contact: hellenth@stats.ox.ac.uk

1 INTRODUCTION

Richard R. Hudson's 'program for generating samples under neutral models' (ms simulator) (2002) is a widely-used program for simulating genetic variation data, in particular single nucleotide polymorphism (SNP) data, for randomly-sampled haplotypes from a population. The program allows the user to specify various aspects of population demography (e.g. population sizes and migration patterns) and factors governing evolution [e.g. mutation, crossover and gene conversion (gc) rates]. However, it presently does not allow for variation in recombination rates. In particular, 'hotspots,' or areas of the genome in which crossover and/or (allelic) gc occur at higher rates than the genome-wide average, appear to be common in humans (Jeffreys *et al.*, 2001; Jeffreys and May, 2004; Myers *et al.*, 2005). We have incorporated both crossover and gc hotspots into a freely available, updated simulator called msHOT. The output and usage is the same as in the ms program of Hudson (2002), but includes additional arguments for specifying hotspot features. Though other coalescent-based simulators have been written to incorporate variable crossover rates (Schaffner *et al.*, 2005), ours is the first to our knowledge to include the option for gc hotspots as well.

2 MODEL

The current implementation of ms allows the user to specify the (population-scaled) rate of crossing-over, ρ, and the relative rate of gc to crossover, f, for the genetic region to be simulated. Here $\rho = 4N_0 r$, where N_0 is the current diploid population size and r is the probability of a crossover occuring in the region in a single transmission from parent to offspring, and $f = g/r$, where g denotes the probability of a gc initiating in the region of interest in a single transmission from parent to offspring (Wiuf and Hein, 2000).

*To whom correspondence should be addressed.

Since r and g are typically small, dividing these parameters by the sequence length of the genetic region gives, respectively, the crossover probability per base pair, r_{bp}, and the gc probability per base pair, g_{bp}.

Our modification msHOT allows the user to insert as many (non-overlapping) crossover hotspots and (non-overlapping) gene conversion hotspots into the genetic region as they wish by specifying the locations and intensities for each. Specifically, incorporating H crossover hotspots requires the user to specify a left endpoint (a_h), right endpoint (b_h), and intensity (λ_h) for each, $h = 1, \ldots, H$. Inside hotspot h, the probability of a crossover occuring between two adjacent base pairs in a single transmission from parent to offspring is $\lambda_h r_{bp}$. Outside any hotspot, this probability is r_{bp}. Similarly, incorporating $\tilde{H}$ gc hotspots requires the user to specify a left endpoint ($\tilde{a}_h$), right endpoint ($\tilde{b}_h$) and intensity ($\tilde{\lambda}_h$) for each, $h = 1, \ldots, \tilde{H}$. Inside gc hotspot h, the probability of a gc intitation between two adjacent base pairs in a single transmission from parent to offspring is $\tilde{\lambda}_h g_{bp}$. Outside any hotspot, this probability is g_{bp} (see Figure 1a). Gene conversion hotspots may overlap with crossover hotspots.

In Hudson's ms, gc events initiate at some base pair, which is assumed to form the left-point of the region affected by the gc. The right-point is then determined by the length (in physical distance) of the region affected by the gc (i.e. the tract length), which is assumed to have a geometric distribution with user-specified mean. This difference in the treatment of the left and right endpoints causes some bothersome asymmetry when the rate of gc initiation is allowed to vary along the region. To deal with this, we changed the model to assume that gc events initiate at some point and then spread both right and left independently according to geometric distributions with user-specified mean t^*. Thus, in our model, the tract length is the sum of two independent geometric distributions. Incidentally, this may also better represent current knowledge of the biology underlying gc events (Szostak *et al.*, 1983).

3 IMPLEMENTATION

The basic algorithm of msHOT is as described in Hudson (1983). In brief, ms generates ancestral recombination graphs for a sample of chromosomes by stochastically determining 'events' to occur on the ancestral material of the chromosomes going back in time, until all the material has coalesced into a common ancestor. We refer to any individual segment of this ancestral material as an 'ancestral segment.' Potential 'events' include the coalesence of two such

(a) recombination variation in $[S,E]$ (b) the three distinct "types" of gc event
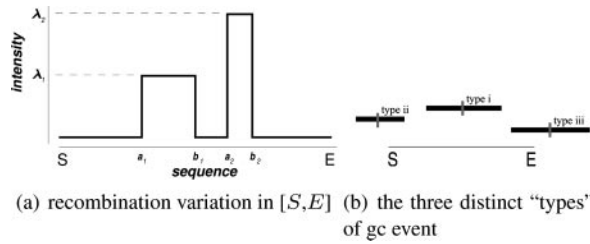
Fig. 1. (**a**) Illustration of varying crossover and/or gc intensities in a genetic region $[S, E]$. Here the crossover, (respectively gc) probability r_{bp}, (respectively g_{bp}) is increased by a multiple λ_1 in $[a_1,b_1]$ and by a multiple λ_2 in $[a_2,b_2]$. (**b**) Illustration of the three distinct gc types that can influence variation in the genetic region $[S, E]$. The grey vertical lines represent the initiation point of each gc event, and the black horizontal bars represent the tract length of each of the gc events.

segments or a recombination event (crossover or gc) occuring in a single segment. Incorporating hotspots involves changing the rates at which these recombination events occur, as described below. (The consequences of these events, which involve splitting ancestral segments, are not changed by the introduction of hotspots and are already dealt with in Hudson's code.)

The rate of each possible recombination event, backwards in time, is determined by computing the probability of the event occurring in a single generation forwards in time, and multiplying this by $4N_0$. We therefore focus on computing the relevant probabilities forwards in time. In the following we use $[S, E]$ to denote an ancestral segment beginning at S and ending at E.

Crossover. Assume the entire simulated region contains H crossover hotspots, each with left endpoint a_h, right endpoint b_h, and intensity λ_h, $h = 1, \ldots, H$. Under the model described above, the probability of a crossover initiating at any particular location $z \in [S, E]$ is:

$$\Pr(\text{crossover at } z) = \left[1 + \sum_{h=1}^{H} I_{z \in [a_h, b_h]}(\lambda_h - 1)\right] r_{bp}. \quad (1)$$

Here $I_{z \in [a_h, b_h]}$ is an indicator function, taking the value 1 if z is in crossover hotspot h and 0 otherwise. The total probability of a crossover occurring in $[S, E]$ is found by summing over z in Equation (1). If a crossover is to occur in $[S, E]$, the location z is selected with probability proportional to Equation (1).

Gene conversion. Each gc event can be thought of as having an 'initiation point' and 'right' and 'left' endpoints. We distinguish three types of gc event that can influence patterns of genetic variation in $[S, E]$ (see Figure 1b):

(1) Type i: a gc event initiates within $[S, E]$ and has endpoints that may be either inside or outside this region.

(2) Type ii: a gc event initiates to the left of S and has a right endpoint within $[S, E]$.

(3) Type iii: a gc event initiates to the right of E and has a left endpoint within $[S, E]$.

The following subsections give the relative probabilities, and describe how to determine the endpoints, for each of these types of event. Assume the entire simulated region contains $\tilde{H}$ gene conversion hotspots, each with left endpoint $\tilde{a}_h$, right endpoint $\tilde{b}_h$, and intensity $\tilde{\lambda}_h$, $h = 1, \ldots, \tilde{H}$.

'Type i': The probability of a type i event initiating at $z \in [S, E]$:

$$\Pr(\text{type i gc at } z) = \left[1 + \sum_{h=1}^{\tilde{H}} I_{z \in [\tilde{a}_h, \tilde{b}_h]}(\tilde{\lambda}_h - 1)\right] g_{bp}, \quad (2)$$

where $I_{z \in [\tilde{a}_h, \tilde{b}_h]}$ is an indicator denoting whether location z is in gc hotspot h. If a type i event occurs, its endpoints are determined by first selecting the initiation point, z, with probabilities proportional to Equation (2), and then simulating the left and right endpoints as $z - T_1$ and $z + T_2$, where T_i are randomly sampled from a geometric(t^*). (These endpoints may fall outside $[S, E]$.)

'Type ii': The probability that a type ii event initiates at a location y to the left of S (thus y is outside $[S, E]$, in contrast to z above) and has a right endpoint at $x \in [S, E]$ is given by:

$$\Pr(\text{type ii gc, } y \to x) = q^{x-y}(1-q)\left[1 + \sum_{h=1}^{\tilde{H}} I_{y \in [\tilde{a}_h, \tilde{b}_h]}(\tilde{\lambda}_h - 1)\right] g_{bp}, \quad (3)$$

where $q \equiv \frac{t^*}{1+t^*}$, y ranges from $-\infty$ to S (for simplicity, we have assumed, as in ms, that the chromosome has infinite length), and $I_{y \in [\tilde{a}_h, \tilde{b}_h]}$ is an indicator for whether y is in hotspot h. The total probability of a type ii event occurring in $[S, E]$ is obtained by summing Equation (3) over possible values of x and y. (We deal with the infinite sum over y by use of standard geometric series results. All locations outside the simulated region are assumed to have the background probability of a gene conversion, g_{bp}.) If a type ii event occurs, its right endpoint, x^*, is chosen via a truncated geometric distribution [i.e. $\Pr(X^* = x^*) \propto q^{x^*-S}(1-q)$, for $x^* = S+1, \ldots, E$].

'Type iii': The type iii gc events are similar to the type ii events above, but with locations starting from the end of an ancestral segment and counting from right to left.

REFERENCES

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

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

Jeffreys,A. and May,C. (2004) Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat. Genet.*, **36**, 151–156.

Jeffreys,A. *et al.* (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.*, **29**, 217–222.

Myers,S. *et al.* (2005) A fine-scale map of recombination rates and hotspots across the human genome. *Science*, **310**, 321–324.

Schaffner,S. *et al.* (2005) Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.*, **15**, 1576–1583.

Szostak,J. *et al.* (1983) The double-strand-break repair model for recombination. *Cell*, **33**, 25–35.

Wiuf,C. and Hein,J. (2000) The coalescent with gene conversion. *Genetics*, **155**, 451–462.