Documentation for mlehot
(Version 0.2; Distributed, but still being tested)

Jeff Wall <jeff.wall@ucsf.edu> and Laurie Stevison
(To accompany Wall and Stevison 2016, G3)

This software package allows for the estimation of recombination hotspots using the approach outlined in Wall and Stevison (2016) and the approaches of Auton et al. (2012 and 2014). Briefly, our method is an implementation of LDhot (cf. McVean et al. 2004) with improved power to detect recombination hotspots over previous implementations. LDhot uses a composite-likelihood approach based on the work of Hudson (2001). For a region of interest, LDhot tests whether the central 2 Kb sub-region has a higher underlying recombination rate than the rest of the region by forming a likelihood-ratio test statistic. Critical values for this test statistic are estimated using coalescent simulations with a constant recombination rate. Most of the improved power of our method is obtained by having a smaller window size (20 Kb versus 200 Kb in Auton et al. 2012), with some increased power due to a slightly different methodology for calling hotspots. See McVean et al. 2004; Myers et al. 2005, Auton et al. (2012, 2014) and Wall and Stevison (2016) for further details on these methods.

Lookup table

Our approach is to run null simulations (assuming a constant recombination rate) over a range of mutation and recombination parameter values and to store the results of these simulations in a big lookup table. Then, when analyzing actual data, the program scans this lookup table for simulations with roughly the same number of observed segregating sites and a similar estimated rho (assuming constant rates, cf. Hudson 2001) to estimate the critical values for a likelihood-ratio test. The file “ln1” gives an example of the programs used to construct these null simulations, with a haploid sample size of n=30, $\theta = 1$ / Kb and $\rho = 2$ / Kb. The file “out_n30new_large2” contains the lookup table we used for our simulations, which used $\theta = 0.4 – 4$ / Kb and $\rho = 0.1 – 50$ / Kb. Note that this workflow, as well as our program mlehot, use another lookup table of pairwise likelihoods (called “30eloutc”) using Hudson’s program “el”, as described in Hudson (2001). Similar lookup tables for other sample sizes are available from http://home.uchicago.edu/rhudson1/source/twolocus/tables.html

Data format

The preferred input format for the data is similar to the output format for the popular

coalescent simulator ms (Hudson 2002), but without the header. The first line should be the number of segregating sites. The second line should be the base positions of each segregating site, separated by spaces or tabs. Each subsequent line shows the haplotype for each chromosome, with alleles coded as 0's and 1's. For example, the following data set

```
8
11      57      204     378     590     841     859     991
00000001
00000011
00000111
00001111
00011111
00111111
01111111
11111111
```

shows genetic variation for 8 haplotypes (i.e., 4 diploid individuals) over a 1 Kb region, where there are 8 polymorphic sites at the positions given in line 2. The user should format his/her data into this format before using the program. We have included the programs used to simulate data (using a modification of an old version of Hudson's ms) and the commands used to filter this data as analyzed in Wall and Stevison (2016). Specifically, ms_varrho0 simulates data under the standard coalescent, filtscale2 converts the base positions (which originally lie between 0 and 1) to integer base positions, filtpositions checks to make sure that multiple polymorphisms do not occur at the same nucleotide site (i.e., the infinite-sites assumption), and filtfreq2 imposes a minor allele frequency cutoff of 0.05. The 5^{th} field in the command line for ms_varrho is the header file "hot_n30new_sims4.h", which displays the piecewise constant recombination rates for different sub-regions in the simulation. (The first line is the number of sub-regions, and each subsequent line gives the length in bp and the total ρ for the sub-region.) The first two command lines of the file "in1" show how to produce the input file sim1.hud, which is a simulation of a 1.1 Mb region from 30 chromosomes with variable recombination rates. This simulated region has a background recombination rate of ρ = 0.5 / Kb, and 2 Kb hotspots (with intensities ranging from ρ = 5 – 50 / Kb) interspersed every 100 Kb.

Likelihood calculations

Subsequent commands in "in1" calculate a background and hotspot estimate of ρ for a 20 Kb sliding window and an increment of 1 Kb, after trimming 40 Kb from each side of

the simulated region. The output, shown in the file “temp1”, lists the number of SNPs (field 1), background ρ estimate (field 2), central/hotspot ρ estimate (field 3) and p-value (field 7, compared with null simulations with constant recombination rate found in the lookup table) for each 20 Kb window. Note that the pairwise likelihood file (e.g., “30eloutc”), the grid of recombination rates considered (“grid6c.h”) and the lookup table described before (“out_n30new_large2”) are all necessary command line arguments. The file “grid6c.h” gives a list of total scaled recombination rates (e.g., ρ for the full 20 Kb window) and hotspot multiplier factors over which the pairwise composite likelihoods are calculated. So, for “grid6c.h”, we consider 151 background ρ values, ranging from 0 – 2000 (i.e., 0 – 0.1 / bp). For each background recombination rate, we consider a ‘hotspot’ recombination rate for the central 2 Kb that is 1 – 400 times higher. The format of the recombination rate file consists of the number of background recombination rates, followed by each value in ascending order, followed by the number of hotspot multiplier factors, followed by the values in ascending order.

The program “filtslidwin2” parses the original simulation file into overlapping subsets, with command line arguments (haploid) sample size, left endpoint of first window, left endpoint of last window, window size and increment. Next, “seqhot” processes each window into the site configurations for each pair of SNPs, with command line arguments (haploid) sample size, scaled (between 0 and 1) hotspot start location and scaled hotspot end location (assuming the hotspot is centered across the test region). For a window size of 20 Kb and a putative hotspot length of 2 Kb, this corresponds to the values 0.45 and 0.55 shown in “in1”. Then, “mlehot” calculates the composite likelihood of the data for each recombination parameter combination in the file “grid6c.h”, and “hotspotliks3” calculates p-values for the data using the likelihoods from “out_n30new_large2” (or other lookup table). Finally, we use the program “filthotspotliks” to convert the output to potential hotspot locations with p-values less than a certain cutoff (e.g., “out_.01” gives the locaations of all windows with p-value less than 0.01).

LDhat analysis

The hotspot definitions used by Auton et al. (2012) and Wall and Stevison (2016) both use recombination rate estimates obtained from LDhat (McVean et al. 2004). For completeness, we have included the programs “interval” and “stat” from the LDhat suite and the likelihood file “lk_n30_t0.001”. We suggest that users download and use the most up-to-date versions of these programs/files. We use perl scripts (brief description in “in1”) to convert the output to a list of the average estimated value of ρ for each non-overlapping 1 Kb window. An example of the proper format of the LDhat file is “sim1_averaged.txt”, which shows the LDhat estimates for the data set “sim1.hud”.

LDhot analysis

The final lines of “in1” show the commands used to estimate the locations of recombination hotspots using the methods of Auton et al. (2012, 2014) and Wall and Stevison (2016) and a window size of 20 Kb. Note that changing the window size (from 20 Kb) and/or the potential hotspot length (from 2 Kb) requires a different lookup table, different command line arguments for “filtslidwin2” and “seqhot”, and potentially a different window size for the LDhat output. The file “LDhatn30_10_1” lists those 1 Kb regions where the LDhat estimate is 10 times the background rate, while the file “LDhat_lmax_1” outputs those 1 Kb intervals whose peak rho is a local maximum.

The simulations described in Wall and Stevison (2016) analyzed the power and false positive rates for the region lying between bases 150,000 – 950,000. The output files, “outauton12_20kb”, outauton14_20kb” and “outwall15_20kb” show the start and stop locations of each called hotspot in the 2nd and 3rd fields. The first field for each line is 1, as specified in “in1”, and can be used to specify different simulation replicates.