`hotspot`: Software to Support Sperm-Typing for Investigating Recombination Hotspots

Linda Odenthal-Hesse, Julien Y. Dutheil, Fabian Klötzl, Bernhard Haubold

Department of Evolutionary Genetics, Max-Planck-Institute for Evolutionary Biology, Plön, Germany

ABSTRACT

Motivation: In many organisms, including humans, recombination clusters within recombination hotspots. The standard method for *de novo* detection of recombinants at hotspots is sperm typing. This relies on allele-specific PCR at single nucleotide polymorphisms. Designing allele-specific primers by hand is time-consuming. We have therefore written a package to support hotspot detection and analysis.

Results: `hotspot` consists of four programs: `asp` looks up SNPs and designs allele-specific primers; `aso` constructs allele-specific oligos for mapping recombinants; `xov` implements a maximum-likelihood method for estimating the crossover rate; `six`, finally, simulates typing data.

Availability: `hotspot` is written in C. Sources are freely available under the GNU General Public License from
http://github.com/evolbioinf/hotspot/

Contact: haubold@evolbio.mpg.de

INTRODUCTION

Understanding recombination is one of the classic problems in biology (Maynard Smith, 1978). In the past decade the quantitative underpinnings of the debate about the function of recombination have dramatically improved as a result of population genomics based on genome sequencing. A major result of such studies is that in many organisms, including humans, recombination is localized in recombination hotspots (Myers *et al.*, 2005; Hinch *et al.*, 2011). Still, direct observation by typing of the products of recombination—usually sperm cells—is highly desirable to check and refine the results of population genetic inference (Jeffreys *et al.*, 2005; Odenthal-Hesse *et al.*, 2014).

Sperm typing is usually carried out after determining SNPs in a genomic region (Figure 1A). Given a set of SNPs, linkage disequilibrium (LD) can be computed between pairs of them leading to the discovery of LD blocks punctuated by hotspot candidate regions a few kb wide. The best-known method for measuring the frequency of crossover at a candidate hotspot is the full-crossover assay (Figure 1B), where allele-specific primers are designed against SNPs located upstream and downstream of the putative recombination hotspot (ASPs, Figure 1B). Nested sets of phased ASPs are then used in repulsion phase on pools of sperm DNA, such that only molecules affected by a crossover are amplified (Kauppi *et al.*, 2009). To map the crossover breakpoints, the SNPs of the amplification product are typed using allele-specific oligos (ASOs).

Recombination results in crossovers as well as non-crossovers, which are characterized by short, non-reciprocal gene conversion tracts. Both types of recombinants can be simultaneously detected by modifying the full-crossover assay such that one of the two sets of allele-specific primers is replaced by a set of universal primers (Figure 1C). The resulting half-crossover assay allows the selective amplification of either haplotype targeted. Recombinant molecules are isolated by sequential ASO hybridization to each SNP within the hotspot interval using ASOs complementary to the haplotype not amplified. Under this protocol, hybridization products may still be due to crossovers (Figure 1C, O_1). However, any stretch of the haplotype not amplified that is detected without the exchange of flanking markers must be the result of a non-crossover event (Figure 1C, O_2).

SNP annotation and the design of ASPs, ASOs, and universal primers is typically done by hand, which is slow and error-prone, especially when screening many hotspot candidates. We have written the software package `hotspot` to assist with this.

INDIVIDUAL PROGRAMS

`hotspot` consists of four programs: `asp` for designing allele-specific primers, `aso` for designing allele-specific oligos, `xov` for estimating the crossover rate, and `six` for simulating crossover data. All four programs are command-line driven and run under UNIX operating systems such as Linux and OS-X.

asp takes three kinds of input: a list of candidate regions, the SNPs of the target genome, and the genome sequence. Given these three inputs, `asp` returns candidate forward and reverse primers flanking both sides of the recombination hotspot. As explained above, these primers can be used for amplifying recombination products either in full-crossover assays (Figure 1B), or half-crossover assays (Figure 1C). The primers are constructed for every SNP in a 5 kb window adjacent to the candidate region. The 3'-end of each primer overlaps either of the two SNP alleles. `asp` optimizes primer length with respect to GC content. Apart from the primer sequences and the GC content, the program also prints out the melting temperature. Finally, `asp` can generate uniformly spaced "universal" primers that do not intersect known polymorphisms for haplotype amplification in the half-crossover assay.

To look up SNPs, `asp` calls the program `tabix` (Li, 2010), which is part of the samtools package (Li *et al.*, 2009). This makes
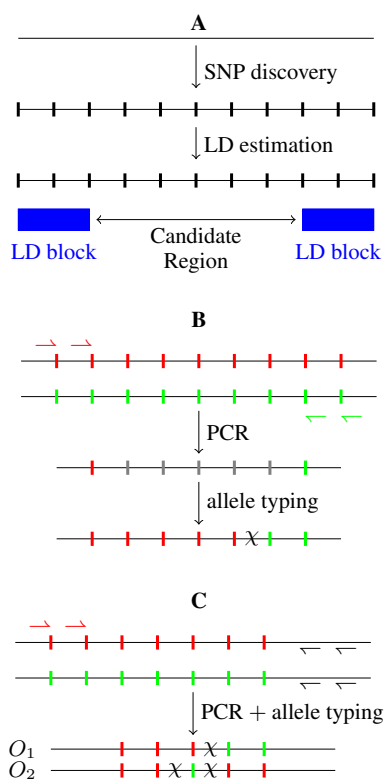
Figure 1. Crossover detection by sperm typing. **A**: Pre-PCR steps consisting of sequencing, SNP discovery, and linkage disequilibrium (LD) estimation for identifying LD blocks flanking hotspot candidates. **B**: Full-crossover assay. Nested PCR is carried out across a hotspot candidate with allele-specific primers constructed using our program `asp`. This leads to positive PCR reactions indicating crossovers, and to negative PCR reactions indicating no crossover. The cartoon shows an example positive reaction, where the crossover region (χ) is identified using allele-specific oligos constructed with our program `aso`. **C**: Half-crossover assay. Allele-specific primers (red) are paired with universal primers (black). PCR and SNP typing can then lead to the detection of crossover (O_1) and non-crossover (O_2) events. Adapted from Figures 20.1 and 20.5 by Kauppi *et al.* (2009).

`asp` fast; for example, it takes only 10.8 s on a Xeon 2.60 GHz system to construct forward primers for all 98,692 SNPs contained in the 5 kb upstream regions of 753 hotspot candidates identified on mouse chromosome 1 by Smagulova *et al.* (2011).

`aso` also takes as input a list of hotspot candidates, SNPs, and genome sequences. Using `tabix` (Li, 2010), it then looks up each SNP in the candidate region and returns two short (e. g. 17 bp) oligos that straddle both alleles of the SNP. These are used to map crossover as well as non-crossover breakpoints (χ in Figures 1B and 1C, respectively). `aso` needs just 9.6 s to construct allele-specific oligos for the same 98,692 SNPs analyzed with `asp` in the previous paragraph.

Again, like `asp`, `aso` can generate evenly spaced oligos that do not overlap SNPs. These "universals" tend to be longer than the allele-specific oligos, say 100 bp. These can be used as capture probes for in-solution enrichment. However, such oligos may come from low complexity regions in the genome, which would cause unspecific enrichment. To allow filtering for low complexity, `aso` also computes a measure based on the Lempel-Ziv complexity (Lempel and Ziv, 1976).

`xov` implements a maximum likelihood procedure for estimating the crossover probability from results of allele-specific PCR (Figure 1B). In a typical screen, m experiments with different numbers of molecules, d_i, are carried out with n_i PCR reactions leading to the observation of k_i blanks. `xov` comes with example data to demonstrate how these quantities are entered.

`six` was written to check the accuracy of `xov` by simulating sperm typing results. It allows the user to modify all three relevant variables in the design of sperm typing studies: the number of experiments (m), including the number of molecules per experiment (d_i); the number of replicates per experiment (n_i); and the crossover rate (x). Apart from checking `xov`, `six` can be used to test the efficacy of experimental setups to detect a given recombination rate.

In conclusion, `hotspot` enables large-scale design of allele-specific oligos and accurate estimation of the rate of crossing over. In addition, it allows exploration of typing designs through simulation.

REFERENCES

Hinch, A. G., Tandon, A., *et al.*, 2011. The landscape of recombination in African Americans. *Nature* 476, 170–175.

Jeffreys, A. J., Neumann, R., *et al.*, 2005. Human recombination hot spots hidden in regions of strong marker association. *Nature Genetics* 37, 601–606.

Kauppi, L., May, C. A., *et al.*, 2009. Analysis of meiotic recombination products from human sperm. In Keeney, S., ed., *Meiosis, Volume 1, Molecular and Genetic Methods*, volume 557, chapter 20, 323–354. Humana Press.

Lempel, A. and Ziv, J., 1976. On the complexity of finite sequences. *IEE Transactions on Information Theory* IT-22, 75–81.

Li, H., 2010. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* 27, 718–719.

Li, H., Handsaker, B., *et al.*, 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079.

Maynard Smith, J., 1978. *The Evolution of Sex*. Cambridge University Press.

Myers, S., Bottolo, L., *et al.*, 2005. A fine-scale map of recombination hotsports across the human genome. *Science* 310, 321–324.

Odenthal-Hesse, L., Berg, I. L., *et al.*, 2014. Transmission distortion affecting human noncrossover but not crossover recombination: a hidden source of meiotic drive. *PlosGenetics* 10, e104106.

Smagulova, F., Gregoretti, I. V., *et al.*, 2011. Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. *Nature* 472, 375–378.