BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 6 2010, pages 838–840
doi:10.1093/bioinformatics/btq029

Genome analysis Advance Access publication January 26, 2010

invertFREGENE: software for simulating inversions in population genetic data

Paul F. O'Reilly, Lachlan J. M. Coin* and Clive J. Hoggart*
Department of Epidemiology and Biostatistics, Imperial College
Associate Editor: Martin Bishop

ABSTRACT

Summary: Inversions are a common form of structural variation, which may have a marked effect on the genome and methods to infer quantities of interest such as those relating to population structure and natural selection. However, due to the challenge in detecting inversions, little is presently known about their impact. Software to simulate inversions could be used to provide a better understanding of how to detect and account for them; but while there are several software packages for simulating population genetic data, none incorporate inversion polymorphisms. Here, we describe a software package, modified from the forward-in-time simulator FREGENE, which simulates the evolution of an inversion polymorphism, of specified length, location, frequency and age, in a population of sequences. We describe previously unreported signatures of inversions in SNP data observed in invertFREGENE results and a known inversion in humans.

Availability: C++ source code and user manual are available for download from http://www.ebi.ac.uk/projects/BARGEN/ under the GPL licence.

Contact: l.coin@ic.ac.uk; c.hoggart@ic.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 28, 2009; revised and accepted on January 19, 2010

1 INTRODUCTION

An inversion mutation both reorientates and flips over the strands of a stretch of DNA sequence. Recombinations are typically only viable between sequences in the same orientation, leading to the separate evolution of inverted and non-inverted sequences at the locus (Navarro *et al*., 2000). While sequencing may distinguish inverted and non-inverted sequences directly, genotyping cannot because: (i) unlike copy number variants, the allele intensity is unaltered, and (ii) the alleles on the inverted sequence are mapped to their location on the reference genome rather than their actual location. However, their effect on recombination may leave a discernible pattern on local linkage disequilibrium (LD).

While most research on inversion polymorphisms has been performed in *Drosophila*, recent analyses of sequence (Kidd *et al*., 2008) and SNP data (Bansal *et al*., 2007) indicate that inversions are a common form of structural variation in the human genome. Studies have shown that inversions may effect genome-wide recombination rates and be associated with natural selection and disease (Antonaccil *et al*., 2007; Sharp *et al*., 2006, 2007). There has also been concern that inversions may bias statistical methods for inferring quantities of interest such as those relating to population structure and recent selection (Deng *et al*., 2008; Price *et al*., 2008). Therefore, statistical methods have been developed to detect inversion polymorphisms from population SNP data (Bansal *et al*., 2007; Sindi and Raphael, 2009), and methods for inferring population genetic parameters may need revising to account for their presence.

However, to test methods for detecting inversions and assess the robustness of population genetic methods to the presence of inversions, software to simulate genetic data incorporating inversions is required. To our knowledge, the only strategy adopted to simulate data with inversions consists of reorientating a proportion of sequences at a locus in the HapMap data (Bansal *et al*., 2007; Sindi and Raphael, 2009). While this strategy is simple and should produce data that reflects some features of inversion loci, reorientating contemporary genetic data does not model the evolution of the inversion through time. In particular, the suppression of recombination between inverted and non-inverted sequences, which results in their subsequent independent segregation, is not modelled but has specific effects on the data (Section 3).

FREGENE is a software package that simulates population sequence data forwards through time by simulating every meiotic and mutation event in a population of sequences over a fixed number of generations (Chadeau-Hyam *et al*., 2008; Hoggart *et al*., 2007). By simulating data forward-in-time, FREGENE allows complex scenarios of demography and recombination to be modelled simultaneously. This enables the generation of data closely matching the features of data from the major genotyping projects (Chadeau-Hyam *et al*., 2008). Here we modify the software to produce an affiliated program, invertFREGENE. This is the first software package to simulate genetic data with inversions and exploits the forward-in-time approach to model the fundamental features of an inversion as it segregates.

2 DESCRIPTION

invertFREGENE simulates genetic data in the same way as FREGENE, but can also simulate a single inversion polymorphism of a specified length, location, population frequency and age. Before simulating an inversion, invertFREGENE should be used to simulate an initial population, without an inversion, to equilibrium. A simulation of $10N$ generations (N = population size) should be adequate, ensuring that 98% of sites have a unique common ancestor (Hoggart *et al*., 2007). We use the recombination model developed by (Schaffner *et al*., 2005) to simulate the broad and fine-scale recombination variation observed in humans and

*To whom correspondence should be addressed.

apply genome-wide average mutation and recombination rates, so that at equilibrium the homozygosity levels, LD structure and allele frequency distribution reflect those observed in humans. invertFREGENE simulates under neutrality, applying a uniform rate finite-site mutation model and can model subpopulations with migration rates specific to each pair of subpopulations and instantaneous population expansions and contractions.

In the first generation of an inversion simulation, one sequence is subject to an inversion mutation from which all subsequent inversions derive. Each 'sequence' is stored as a vector of nucleotide locations where the mutant allele is present. An inversion mutation is modelled by inverting the locations of mutants in a section of one vector. If the inversion mutation is between the 100th and 200th base on a sequence with mutant alleles at locations 105, 148 and 185, then the inverted sequence has mutants at locations 115, 152 and 195. In the simulation of subsequent generations, recombinations proposed within the inversion between inverted and non-inverted sequences are rejected and new recombination locations are proposed. Recombinations proposed between sequences in the same orientation occur as usual. This will result in the correct partioning of DNA sequence after recombination between two inverted sequences because the location of SNPs within the inverted sequence has been flipped.

If the inverted sequences are lost from the population then a new inversion mutation occurs in the next generation. Simulations in which the inversion frequency exceeds a user-defined threshold (default 10%) but are then lost without reaching the specified target frequency or age, are restarted with a new seed and the initial input population. Setting a low threshold ensures that the final data are not affected by previous common inversions. The user can set either or both of the frequency and age of the inversion, specifically the minimum and maximum age of the inversion in generations, since both parameters are likely to contribute to the characteristics of the resulting data. The simulation stops at the first generation where these criteria are met, and is restarted if the target frequency is not reached before the maximum age of the inversion.

The data can be output to reflect either genotyped data, produced by reorientating all inverted sequences to their original position (mimicking mapping to a reference genome), or sequenced data, where the data is output directly.

3 RESULTS

Using invertFREGENE we simulated a 500 kb inversion at the centre of a 2 Mb region, with a target inverted sequence frequency of 40%. Figure 1a displays a plot of minor allele frequency (MAF) by physical position for the simulated region. The inversion is spanned by a set of SNPs with MAF close to 40%, forming a visible line at the inversion locus. This feature occurs because mutant alleles present on the original inverted sequence, that were otherwise absent or very rare in the population, should be present in every subsequent inverted sequence but not in non-inverted sequences because there are no recombinations between sequences in opposing orientation. These alleles therefore tag the inversion and have population frequency equal to that of the inversion. The corresponding SNPs are perfectly correlated with each other. The well-established inversion at the *MAPT* locus on chromosome 17 (Kidd *et al*., 2008; Stefansson *et al*., 2005) has an estimated frequency of 20% in Europeans (Stefansson *et al*., 2005). Figure 1b displays the MAF plot at this locus for

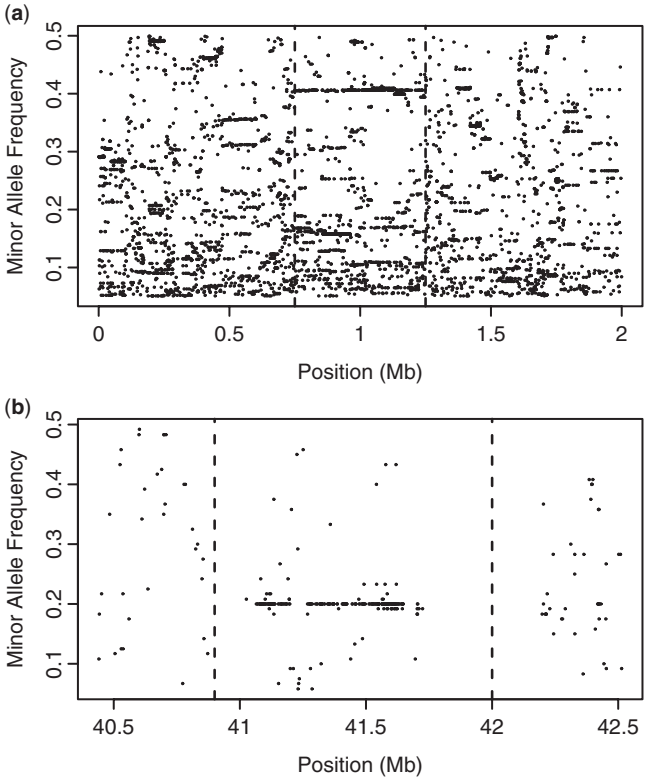

Fig. 1. MAF plot from (**a**) an invertFREGENE simulation with a 500 kb inversion, at frequency 40%, at the centre of a 2 Mb region (dashed vertical lines show location of breakpoints) and (**b**) HapMap data from 120 CEPH chromosomes in the region of the *MAPT* inversion on chromosome 17, with estimated breakpoints at 40.9 Mb and 42.0 Mb (dashed lines).

European HapMap individuals. As expected, a line of SNPs at frequency 20% resides within the inversion locus. There is a gap in the genotyped SNPs at the right breakpoint, possibly due to repeat sequence. If such gaps are found to be common at inversion loci then they should be modelled in simulated data by removing SNPs close to breakpoints.

Supplementary Figures 1 and 2 show the pairwise LD (Barrett *et al*., 2005) between SNPs from the invertFREGENE simulation described above. The LD block at the centre of the region corresponds to the inversion locus, which occurs because of the fewer recombinations within the inversion and the set of perfectly correlated SNPs ($r^2 = 1$) tagging the inversion. The breakdown of LD close to the inversion breakpoints is a consequence of the flipped SNP positions in the inverted sequences. Although the recorded separation between SNPs on either side of the breakpoints is small, the actual distance in the inverted sequences is large. A particular feature of the LD block reflecting the presence of an inversion is that while there are many pairs of SNPs in perfect correlation, generally indicative of a paucity of recombination, there are also many pairs of SNPs with $D' < 1$, due to recombination events at the locus between sequences in the same orientation. Supplementary Figures 3 and 4 show LD plots for the region of the *MAPT* inversion locus in the HapMap data of Figure 1b. Again, an LD block corresponds to the inversion locus, within which there are many perfectly correlated SNPs but also pairs of SNPs with $D' < 1$.

We have shown that invertFREGENE produces some of the fundamental features expected of an inversion locus, namely that perfectly correlated SNPs span the inversion and that, despite this, D' is often below 1. We have also shown that these features are observed in data from a known inversion in the human genome. The previous strategy of simulating inversions by flipping contemporary sequences would not produce these features, and therefore analyses based on data simulated in this way may be inaccurate. This highlights the necessity for a software tool that accurately models inversion polymorphisms.

ACKNOWLEDGEMENTS

We would like to thank Katerina Seich al Basatena for testing the code and David Balding for conceiving the FREGENE project.

Funding: ENGAGE consortium (grant P12892_DFHM to P.F.O.); Research Council UK Fellowhsip (to L.J.M.C.); Europenan Union (grant HEALTH-F4-2007-201550 HyperGenes to C.J.H.).

Conflict of Interest: None declared.

REFERENCES

Antonaccil,F. *et al.* (2007) Characterization of six human disease-associated inversion polymorphisms. *Hum. Mol. Genet.*, **18**, 2555–2566.

Bansal,V. *et al*. (2007) Evidence for large inversion polymorphisms in the human genome from HapMap data. *Genome Res.*, **17**, 219–230.

Barrett,J.C. *et al.* (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265

Chadeau-Hyam,M. *et al.* (2008) Fregene: simulation of realistic sequence-level data in populations and ascertained samples. *BMC Bioinformatics*, **9**, 364.

Deng,L. *et al.* (2008) An unusual haplotype structure on human chromosome 8p23 derived from the inversion polymorphism. *Hum. Mut.*, **29**, 1209–1216.

Hoggart,C.J. *et al.* (2007) Sequence-level population simulations over large genomic regions. *Genetics*, **177**, 1725–1731.

Kidd,J.M. *et al.* (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature*, **453**, 56–64.

Navarro,A. *et al*. (2000) Effect of inversion polymorphism on the neutral nucleotide variability of linked chromosomal regions in Drosophila. *Genetics*, **155**, 685–698.

Price,A.L. *et al.* (2008) Long-range LD can confound genome scans in admixed populations. *Am. J. Hum. Genet.*, **83**, 132–135.

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

Sharp,A.J. *et al*. (2006) Structural variation of the human genome. *Ann. Rev. Genomics Hum. Genet.*, **7**, 407–442.

Sharp,A.J. *et al.* (2007) Characterization of a recurrent 15q24 microdeletion syndrome. *Hum. Mol. Genet.*, **16**, 567–572.

Sindi,S. and Raphael,B. (2009) Identification and frequency estimation of inversion polymorphisms from haplotype data. *RECOMB*, **5541**, 418–433.

Stefansson,H. *et al.* (2005) A common inversion under selection in Europeans. *Nat. Genet.*, **37**, 129–137.

The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.