BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 14 2007, pages 1851–1853
doi:10.1093/bioinformatics/btm253

Genetics and population analysis

Recombination-filtered genomic datasets by information maximization

August E. Woerner[1], Murray P. Cox[1,2,*] and Michael F. Hammer[1]

[1]Arizona Research Laboratories – Biotechnology, University of Arizona, Tucson, AZ 85721 and [2]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

Received on February 28, 2007; revised on May 3, 2007; accepted on May 4, 2007

Advance Access publication May 22, 2007

Associate Editor: Martin Bishop

ABSTRACT

Summary: With the increasing amount of DNA sequence data available from natural populations, new computational methods are needed to efficiently process raw sequences into formats that are applicable to a variety of analytical methods. One highly successful approach to inferring aspects of demographic history is grounded in coalescent theory. Many of these methods restrict themselves to perfectly tree-like genealogies (i.e. regions with no observed recombination), because theoretical difficulties prevent ready statistical evaluation of recombining regions. However, determining which recombination-filtered dataset to analyze from a larger recombination-rich genomic region is a non-trivial problem. Current applications primarily aim to quantify recombination rates (rather than produce optimal recombination-filtered blocks), require significant manual intervention, and are impractical for multiple genomic datasets in high-throughput, automated research environments. Here, we present a fast, simple and automatable command-line program that extracts optimal recombination-filtered blocks (no four-gamete violations) from recombination-rich genomic re-sequence data.

Availability: http://hammerlab.biosci.arizona.edu/software.html

Contact: mpcox@email.arizona.edu

1 INTRODUCTION

With the completion of the human genome sequence, there is now a focus on describing patterns of DNA sequence variability in many human populations. This endeavor has led to large DNA polymorphism and sequence databases, including re-sequencing of multiple unlinked regions across the human genome (e.g. National Institute of Environmental Health Sciences, 2007). These datasets are also useful for reconstructing human demographic history (Garrigan and Hammer, 2006). Several analytical approaches have been devised for evolutionary inference, some of the most elegant of which are based on Monte Carlo simulation with the *n*-coalescent approximation of Kingman (1982). One limitation of this approach is that most methods rely on perfectly tree-like genealogies; i.e. the fully resolved branching relationships of haplotypes from genetic data that show no four-gamete violations. Although coalescent approaches are equally applicable to recombining regions (e.g. through ancestral recombination graph theory; Griffiths, 1981; Hudson, 1983), these methods are not readily available as user-friendly software (Fearnhead and Donnelly, 2001; Griffiths and Marjoram, 1996; Kuhner *et al*. 2000).

However, several coalescent applications are available to help infer demographic parameters from genomic data; such as *Isolation-with-Migration* (*IM*, Hey, 2005), *Genetree* (Griffiths, 1994), *Batwing* (Wilson, *et al*., 2003), and *BEAST* (Drummond and Rambaut, 2003). These programs apply several mutation models, including infinite-sites implementations that require data containing no four-gamete violations. Although software has been developed to detect recombination events, such as *RecMin* (Myers, 2003), *LDhat* (McVean, 2004), *PHASE* (Stephens and Li, 2001) and programs described elsewhere (Posada, 2002), these only infer recombination rates or indicate likely recombination break points. To our knowledge, no existing program produces optimal recombination-filtered datasets from recombining input data. To fill this gap, we have developed *IMgc*—a simple, fast, stand-alone and automatable command-line program that extracts a maximally informative recombination-filtered block from recombination-rich re-sequence datasets.

2 ALGORITHM AND METHODS

IMgc is stand-alone command-line Perl code that uses a weighting heuristic (described below) to iterate over all possible blocks of non-recombining sequence within a larger dataset and determine the most data-rich recombination-filtered block. *IMgc* infers recombination from violations of the four-gamete rule, as defined by Hudson and Kaplan (1985), within a multiple sequence alignment. *IMgc* requires haplotyped data; phase should be inferred computationally or experimentally before analysis. The minimum number of recombination events is zero for data with no four-gamete violations, although strictly speaking, such data can never be proven to be recombination free. Homoplasy, sequencing miscalls and errors during the phasing process may also increase four-gamete violations, and therefore bias towards incorrect recombination-filtered subsets.

IMgc begins by parsing an aligned fasta record, determining segregating sites, and in the case of sites with more than two character

*To whom correspondence should be addressed.

states (infinite sites violations), treating as missing data all but the two highest frequency variants. For datasets containing four-gamete violations, smaller blocks of non-recombining data can be generated in two distinct ways: either by removing sites breaking the four-gamete rule (and all downstream or upstream sites), or by removing haplotypes that represent likely recombinant sequences. Initially, *IMgc* examines the entire dataset (segregating sites 1 through S), and infers the largest number of chromosome copies (1 through C) that form a dataset with no four-gamete violations. *IMgc* calculates an inclusiveness score for this dataset,

$$I = S_r C_r^a \quad (1)$$

where S_r is the number of sites retained, C_r is the number of chromosome copies retained and α is the chromosome copy weighting parameter. Given default settings, individuals and sites have equal weighting ($\alpha = 1$). Subsequently, *IMgc* decreases the number of segregating sites from S to 1. At each step, *IMgc* determines the largest number of chromosome copies that cause no four-gamete violations and recalculates the inclusiveness score. These iterations continue until further reductions in S must necessarily produce a lower inclusiveness score than the maximal value already obtained.

Finally, *IMgc* returns a genomic block representing the largest inclusiveness score of segregating sites versus weighted chromosome copies. Because users can define any value for the weighting parameter, *IMgc* can also generate non-recombining datasets that maximize either segregating sites or chromosome copies. Furthermore, *IMgc* can forcibly include an outgroup sequence in the recombination analysis, if no recombination between ingroup and outgroup sequences is desired (e.g. *Genetree*; Griffiths, 1994).

3 RESULTS AND DISCUSSION

IM*gc* is compatible with all Perl-enabled operating systems, including UNIX, OS X and Windows, and is easily run in batch mode. *IMgc* produces either fasta records containing the input dataset's largest non-recombining block, or file bodies compatible with *IM* (Hey, 2005).

We note that extracting blocks with no four-gamete violations ($\alpha = 1$) can result in datasets with different statistical properties than the originals (Fig. 1). In particular, *IMgc* introduces a downward bias in θ_w, a proxy for information content. This may have unintended effects for programs like *IM* and *Genetree*, which calculate effective population sizes from the θ observed in these subsets. This manipulation also affects the site frequency spectrum, as observed by a slight upward bias in *Tajima's D* ($\bar{\delta}_{TD} = 0.13$), also increasing its variance. These effects result from a more pronounced downward bias in θ_w than in θ_π (haplotype pairwise differences). It is important to note that this effect is exacerbated when heavily favouring recombination-filtered subsets with removal of sites ($\alpha >> 1$, $\bar{\delta}_{TD} = 0.36$) or removal of chromosome copies ($\alpha << 1$, $\bar{\delta}_{TD} = 0.39$). When recombination-filtered subsets are constructed manually, chromosome copies are often preferentially removed (e.g. Harding *et al.*, 1997), possibly because text editors make haplotype removal easier than site removal. However, this practice can strongly bias the site frequency spectra of resulting subsets. These simulations emphasize the need for software like *IMgc* to determine, and extract, optimal recombination-filtered subsets so as to minimize these effects.

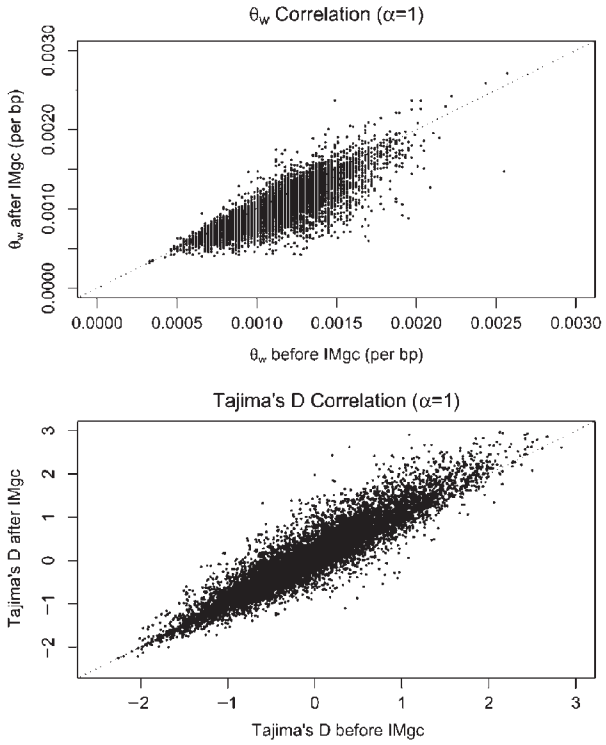

Fig. 1. Comparing the site frequency spectrum between coalescent datasets and their optimal recombination-filtered subsets. Data for 96 autosomal chromosomes were simulated using $N_e = 10^4$, $\mu = 1.1 \times 10^{-9}$, and a recombination rate of 1 cM/Mb over a 10 kb region (10^4 replicates).

Finally, *IMgc* runs extremely quickly, even for datasets with high recombination rates. For instance, a dataset of 96 chromosomes with 96 segregating sites in a 10 kb stretch of DNA sequence with a recombination rate 10-fold the human genome average takes substantially < 1 min to process on a 3 GHz Xeon processor. Furthermore, *IMgc* sets no a priori limitations on the size or nature of input datasets. This fast runtime is particularly useful for downstream applications in which multiple genomic loci must be analyzed (e.g. *IM*; Hey, 2005), and it is vital for high-throughput, automated genomic data processing environments.

ACKNOWLEDGEMENTS

We thank F. Mendez for his invaluable insight on this topic. This program was developed as part of the *Hominid* project, a genomic resequencing study funded by the National Science Foundation grant BCS-0423670.

Conflict of Interest: none declared.

REFERENCES

Drummond,A.J. and Rambaut,A. (2003) *BEAST v1.0*. http://beast.bio.ed.ac.uk/Main_Page.

Fearnhead,P. and Donnelly,P. (2001) Estimating recombination rates from population genetic data. *Genetics*, **159**, 1299–1318.

Garrigan,D. and Hammer,M.F. (2006) Reconstructing human origins in the genomic era. *Nat. Rev. Genet.*, **7**, 669–680.

Griffiths,R.C. (1981) Neutral two-locus multiple allele models with recombination. *Theor. Popul. Biol.*, **19**, 169–186.

Griffiths,R.C. (1994) *Genetree v. 9.0*. http://www.stats.ox.ac.uk/~griff/software.html.

Griffiths,R.C. and Marjoram,P. (1996) Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.*, **3**, 479–502.

Harding,R.M. *et al.* (1997) Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am. J. Hum. Genet.*, **60**, 772–789.

Hey,J. (2005) *IM*. http://lifesci.rutgers.edu/~heylab/HeylabSoftware.htm#IM

Hey,J. and Nielsen,R. (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*, **167**, 747–760.

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

Hudson,R.R. and Kaplan,N.L. (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, **111**, 147–164.

Kingman,J.F.C. (1982) *On the Genealogy of Large Populations*. Applied Probability Trust, Sheffield.

Kuhner,M.K. *et al.* (2000) Maximum likelihood estimation of recombination rates from population data. *Genetics*, **156**, 1393–1401.

McVean,G. (2004) *LDhat v.2.0*. http://www.stats.ox.ac.uk/~mcvean/LDhat/

Myers,S.R. (2003) *RecMin*. http://www.stats.ox.ac.uk/~myers/

National Institute of Environmental Health Sciences (2007) NIEHS Environmental Genome Project. http://egp.gs.washington.edu/

Posada,D. (2002) Evaluation of methods for detecting recombination from DNA sequences: empirical data. *Mol. Biol. Evol.*, **19**, 708–717.

Stephens,M. and Li,N. (2001) *PHASE v.2.1*. http://www.stat.washington.edu/stephens/software.html

Wilson,I.J. *et al.* (2003) *Batwing*. http://www.mas.ncl.ac.uk/~nijw/