BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 23 2014, pages 3430–3431
doi:10.1093/bioinformatics/btu563

Genetics and population analysis

Advance Access publication August 21, 2014

diCal-IBD: demography-aware inference of identity-by-descent tracts in unrelated individuals

Paula Tataru[1,†], Jasmine A. Nirody[2,†] and Yun S. Song[3,4,5,*]

[1]Bioinformatics Research Centre, Department of Computer Science, Aarhus University, 8000 Aarhus C, Denmark, [2]Biophysics Graduate Group, [3]Computer Science Division, [4]Department of Statistics and [5]Department of Integrative Biology, University of California, Berkeley, CA 94720, USA

Associate Editor: Jeffrey Barrett

ABSTRACT

Summary: We present a tool, diCal-IBD, for detecting identity-by-descent (IBD) tracts between pairs of genomic sequences. Our method builds on a recent demographic inference method based on the coalescent with recombination, and is able to incorporate demographic information as a prior. Simulation study shows that diCal-IBD has significantly higher recall and precision than that of existing single-nucleotide polymorphism–based IBD detection methods, while retaining reasonable accuracy for IBD tracts as small as 0.1 cM.

Availability: http://sourceforge.net/projects/dical-ibd

Contact: yss@eecs.berkeley.edu

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

Received on May 13, 2014; revised on July 24, 2014; accepted on August 16, 2014

1 INTRODUCTION

The notion of identity-by-descent (IBD) between distantly related individuals is playing an increasing role in a variety of genetic analyses, including association mapping (Browning and Thompson, 2012), inferring past demographic history (Palamara *et al.*, 2012; Ralph and Coop, 2013) and detecting signals of natural selection (Albrechtsen *et al.*, 2010). Currently there exist several useful methods for detecting IBD tracts. These methods are based on characterizing similar haplotypes [e.g. GERMLINE (Gusev *et al.*, 2009)] or considering patterns of linkage disequilibrium [e.g. fastIBD, Refined IBD and IBDseq (Browning and Browning, 2011, 2013a, b)], but they do not explicitly model genealogical relationships between genomic sequences. Here, we present a new IBD detection tool, diCal-IBD, which is based on a well-used genealogical process in population genetics, namely, the coalescent with recombination. Another feature that distinguishes our method is that we can incorporate demographic information as a prior.

There seems to be no universally accepted definition of IBD. The definition we adopt is the same as that in Palamara *et al.* (2012) and Ralph and Coop (2013). Specifically, an IBD tract is defined as a maximally contiguous genomic region that is wholly descended from a common ancestor without any recombination occurring within the region. In contrast to other methods, we allow IBD tracts to contain point mutations, which are likely to occur in humans due to comparable mutation and recombination rates.

diCal-IBD is able to detect IBD tracts with high accuracy in unrelated individuals, between whom the vast majority of shared tracts are <1 cM. Single-nucleotide polymorphism (SNP)-based methods are successful in detecting tracts >2 cM, but have low power for shorter tracts, whereas sequence-based methdods, such as diCal-IBD and IBDseq, maintain reasonable accuracy for tracts as small as 0.1 cM.

2 METHOD

diCal-IBD uses a recently developed demographic inference method called diCal (Sheehan *et al.*, 2013). diCal is formulated as a hidden Markov model, a decoding of which returns the time to the most recent common ancestor (TMRCA) for each site when analyzing only a pair of sequences. A change in TMRCA requires a recombination event and diCal-IBD uses the posterior decoding of TMRCA to call IBD tracts above a user-specified length, optionally trimming the ends of the tracts that have low posterior probabilities.

diCal requires discretizing time by partitioning it into non-overlapping intervals. The user has the option of specifying any discretization scheme. The default setting implemented in diCal-IBD distributes the pairwise coalescence probability uniformly over the intervals, similarly as in PSMC (Li and Durbin, 2011), under a constant population size model. An alternative scheme concentrates the intervals in the period that is most likely to give rise to tracts that are long enough to be detected accurately. These schemes are detailed in the Supplementary Information. Given a variable population size history, we approximate it with a piecewise constant population size.

As an application of IBD prediction, we provide a framework for detecting natural selection. Using the average IBD sharing and posterior probability along the sequence, diCal-IBD identifies regions that exhibit high sharing relative to the background average, indicating possible influence of positive selection.

We refer the reader to the online Supplementary Information for details on data processing, options used in calling diCal, post-processing of posterior decoding and identification of selection.

3 IMPLEMENTATION

diCal-IBD is written in Python 2.7, is platform independent and has a command line interface that allows the user to completely specify its behavior. The implementation allows for parallel runs of diCal on different sequence pairs. diCal-IBD provides a

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

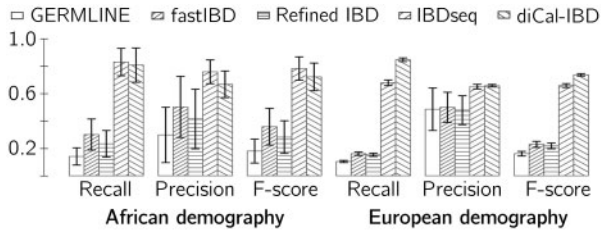

Fig. 1. diCal-IBD was run with bin size 100, constant population size, time discretized in 10 intervals according to unconditional coalescence probability, minimum tract length of 0.1 cM and trimming of tracts with a threshold of 0.2. Error bars show the variance

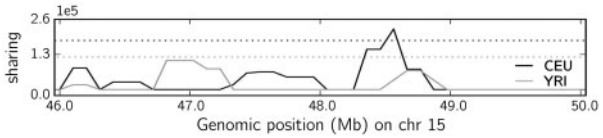

Fig. 2. Detection of high sharing using diCal-IBD, on a 4 Mb genomic segment (46.0–50.0 Mb) on chromosome 15 from Complete Genomics data (Drmanac *et al.*, 2010). This region contains a gene thought to be under positive selection in the European population (CEU), but not in the African population (YRI), located at 48.41–48.43 Mb, corresponding with the observed peak in the plot. Dotted lines indicate the thresholds for considering that a region exhibits high sharing

visualization of the predicted tracts, their posterior probabilities and the corresponding TMRCAs and sequence-wide average IBD sharing and posterior probability. Accuracy information is also provided if the true IBD tracts are known.

4 PERFORMANCE

We carried out a simulation study to compare diCal-IBD with the state-of-the-art IBD detection methods. We used ms (Hudson, 2002) to simulate full ancestral recombination graphs (ARGs) for 50 sequences of 10 Mb each. We used a constant recombination rate of 10^{-8}, and the African and European demographic histories inferred by Tennessen *et al.* (2012). We simulated perfectly phased sequence data on the ARGs with a constant mutation rate of 1.25×10^{-8} per base per generation. From the simulated ARGs, we reconstructed the true pairwise IBD tracts by finding maximally consecutive sites that have the same TMRCA for the pair in question. We only considered tracts of length >0.1 cM.

To run SNP-based methods, we generated SNP data with ~1 marker per 0.2 kb. For further details on running existing tools, see Supplementary Information.

Figure 1 shows the recall (percentage of true tracts that were correctly recovered), precision (percentage of predicted tracts that were correctly predicted) and F-score (harmonic mean of recall and precision) for each method. See Supplementary Information for other measures of accuracy as a function of the true tract length, as well as the effects of errors in the data, demography, discretization and trimming based on posterior probabilities. As the figure shows, diCal-IBD was able to recall significantly more tracts with greater precision than could SNP-based methods, leading to a much higher F-score. diCal-IBD was run assuming a constant population size, but its accuracy performance for the examples considered did not seem to be affected much by using this incorrect prior. This suggests that the posterior distribution inferred by diCal is robust to mis-specification of population sizes; whether this trend persists for more complex demographies deserves further investigation.

The precision and recall performance of diCal-IBD was comparable with that of IBDseq, with neither one strictly dominating the other on all population size histories. See Figure 1. A strength of diCal-IBD is its ability to explicitly incorporate demographic information. diCal, on which diCal-IBD is based, was originally developed for inferring variable effective population sizes, but it is being extended to handle more complex demographic models, incorporating multiple populations, population splits, migration and admixture. diCal-IBD will be updated in parallel with diCal and hence will be able to use a complex demographic model as a prior.

Figure 2 illustrates the potential of applying diCal-IBD to identify regions under selection (Albrechtsen *et al.*, 2010). We refer the reader to the Supplementary Information for further details.

ACKNOWLEDGEMENTS

We thank Sara Sheehan, Jack Kamm, Matthias Steinrücken and other members of the Song group for helpful discussions.

Funding: This research is supported in part by an NSF IGERT grant (J.A.N.) from CiBER at UC Berkeley, an NIH grant R01-GM094402 (Y.S.S.) and a Packard Fellowship for Science and Engineering (Y.S.S.).

Conflict of interest: none declared.

REFERENCES

Albrechtsen,A. *et al.* (2010) Natural selection and the distribution of identity-by-descent in the human genome. *Genetics*, **186**, 295–308.

Browning,B.L. and Browning,S.R. (2011) A fast, powerful method for detecting identity by descent. *Am. J. Hum. Genet.*, **88**, 173–182.

Browning,B.L. and Browning,S.R. (2013a) Detecting identity by descent and estimating genotype error rates in sequence data. *Am. J. Hum. Genet.*, **93**, 840–851.

Browning,B.L. and Browning,S.R. (2013b) Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, **194**, 459–471.

Browning,S.R. and Thompson,E.A. (2012) Detecting rare variant associations by identity-by-descent mapping in case-control studies. *Genetics*, **190**, 1521–1531.

Drmanac,R. *et al.* (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, **327**, 78–81.

Gusev,A. *et al.* (2009) Whole population, genome-wide mapping of hidden relatedness. *Genome Res.*, **19**, 318–326.

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

Li,H. and Durbin,R. (2011) Inference of human population history from individual whole-genome sequences. *Nature*, **475**, 493–496.

Palamara,P.F. *et al.* (2012) Length distributions of identity by descent reveal fine-scale demographic history. *Am. J. Hum. Genet.*, **91**, 809–822.

Ralph,P. and Coop,G. (2013) The geography of recent genetic ancestry across Europe. *PLoS Biol.*, **11**, e1001555.

Sheehan,S. *et al.* (2013) Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. *Genetics*, **194**, 647–662.

Tennessen,J.A. *et al.* (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, **337**, 64–69.