Bioinformatics, 32(15), 2016, 2389–2391
doi: 10.1093/bioinformatics/btw124
Advance Access Publication Date: 21 March 2016
Applications Note

OXFORD

Genetics and population analysis

XIBD: software for inferring pairwise identity by descent on the X chromosome

Lyndal Henden[1,2,*], David Wakeham[3] and Melanie Bahlo[1,2,4]

[1]Population Health and Immunity Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia, [2]Department of Medical Biology, [3]School of Physics and [4]School of Mathematics and Statistics, University of Melbourne, Melbourne, VIC, Australia

*To whom correspondence should be addressed.
Associate Editor: Oliver Stegle

Received on October 1, 2015; revised on February 4, 2016; accepted on March 1, 2016

Abstract

Summary: XIBD performs pairwise relatedness mapping on the X chromosome using dense single nucleotide polymorphism (SNP) data from either SNP chips or next generation sequencing data. It correctly accounts for the difference in chromosomal numbers between males and females and estimates global relatedness as well as regions of the genome that are identical by descent (IBD). XIBD also generates novel graphical summaries of all pairwise IBD tracts for a cohort making it very useful for disease locus mapping.
Availability and implementation: XIBD is written in R/Rcpp and executed from shell scripts that are freely available from http://bioinf.wehi.edu.au/software/XIBD along with accompanying reference datasets.
Contact: henden.l@wehi.edu.au
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genomic regions that have been inherited from a common ancestor are said to be identical by descent (IBD). Identification of IBD regions has proven useful in many applications, including discovery and quantification of unknown or misspecified familial relatedness (McPeek and Sun, 2000) and disease mapping where a region inherited in multiple affected individuals is indicative of a critical region containing disease susceptibility genes (Albrechtsen *et al.*, 2009).

Several methods have been proposed for inferring IBD, however few allow for analysis of the X chromosome. This is unfortunate as the X chromosome allows for inference of more distant relatedness than autosomes since fewer recombination events occur, resulting in IBD segments that are sustained over longer periods of time. Current IBD analysis of the X chromosome is based on identity by state sharing, which is non-probabilistic and does not necessarily imply IBD (Browning and Browning, 2013; Gusev *et al.*, 2009). Here we present XIBD; the only hidden Markov model (HMM) that infers IBD on the X chromosome in addition to the autosomes, where IBD is detectable between individuals with a recent common ancestor, within 25 generations, rather than more distant relatedness. Many IBD methodologies rely on large cohorts to estimate allele frequency data or to infer linkage disequilibrium structure. XIBD can be applied to as few as two samples with the option to use HapMap allele frequency data for 11 populations. Furthermore, it can be applied to either single nucleotide polymorphism (SNP) chip or next generation sequencing (NGS) data (exome or genome wide).

2 Methods

XIBD implements a first order continuous time HMM to infer IBD between pairs of individuals using unphased genotype data, where time is the genetic map distance in centimorgans (cM). While the model is continuous time, IBD is estimated at the genotyped positions only. The memoryless assumption of a Markov process is unlikely to hold due to recombination, however McPeek and Sun (2000) have shown it to be a good approximation and like Albrechtsen *et al.* (2009) and Epstein *et al.* (2000), we also make this assumption.

The hidden states in the Markov model are the number of alleles shared IBD between a pair of individuals, which depend on the genders of the individuals being compared. Assuming the pair are not inbred, if at least one individual is male, then 0 or 1 allele will be shared IBD and the state space is $Z = \{0, 1\}$. Alternatively, if both individuals are female then 0, 1 or 2 alleles will be shared IBD with $Z = \{0, 1, 2\}$.

The initial state probabilities of sharing 0, 1 or 2 alleles IBD are denoted ω_0, ω_1 and ω_2 respectively, where $\sum_{i=0}^{2} \omega_i = 1$. These probabilities can be calculated using identity coefficients if relationships are known. We use IdCoefs (Abney, 2009) for autosomes and have implemented the equivalent for the X chromosome. Individuals may be distantly related with unknown relationships, therefore these values must be calculated using an alternative approach. We use the method-of-moments approach described in Purcell *et al.* (2007) to estimate these probabilities. The estimated values are used in the analysis as they can be accurately calculated for known and unknown relationships and avoid misspecified pedigrees leading to incorrect global estimates.

Since there are two state spaces in the model, we require two transition probability matrices. These can be computed by solving Kolmogorov's forward (or backward) equation given the transition rate matrices (Supplementary information, Eqs. S1 and S2). The transition rate matrices require the number of meiosis m separating the pair of individuals, estimated as in Purcell *et al.* (2007), and the recombination rate θ estimated by Ott (1999).

The genotypic state space for an individual also depends on their gender. Let A and a denote the reference allele and alternative allele respectively. Since male X chromosomes are haploid, they cannot have heterozygous genotype calls. Therefore the male genotypic state space is $G = \{A, a\}$ while the female genotypic state space is $G = \{AA, Aa, aa\}$. A pair of genotypes makes up the observation for each marker in the model. Hence, the observation state space differs for each of the three pairwise combinations of genders and this difference leads to three sets of emission probabilities (Supplementary Information, Table S1). The emission probabilities are functions of the individuals' genders, the state space, the observed genotype pair and the population allele frequencies. The population allele frequencies are calculated from either a reference dataset such as HapMap (http://hapmap.ncbi.nlm.nih.gov/) or the input dataset itself. We include a genotyping error term into our calculation of the emission probabilities as implemented by Albrechtsen *et al.* (2009) (Supplementary Information, Tables S2 and S3) and accommodate missing data.

Dense marker datasets allow for better detection of small IBD tracts and hence more distant relatedness. However, the presence of LD can result in unwanted background sharing. To avoid this, we allow the user to select one of two models to accommodate for LD.

1. Like Purcell *et al.* (2007), model 1 assumes the markers are in linkage equilibrium, which may require thinning of datasets prior to use. However datasets with denser markers in LD can be used at the expense of false IBD segments being reported (Brown *et al.*, 2012).
2. Like Albrechtsen *et al.* (2009), LD is implicitly accounted for in model 2 using conditional emission probabilities (Supplementary Information, Tables S4–S6). Pairwise LD between markers is calculated using the squared correlation (R^2) of reference genotypes using PLINK (Purcell *et al.*, 2007).

Markers in high LD ($R^2 > 0.99$) and markers with low minor allele frequency (MAF < 0.01) are removed from the analysis.

Unlike Purcell *et al.* (2007) and Albrechtsen *et al.* (2009), reference datasets are provided with XIBD. These datasets are the combined HapMap Phase II and III genotypes and allele frequencies from build 19 (The International HapMap Consortium, 2003); allowing the user to choose between the 11 HapMap population. Furthermore, given a homogeneous population, we allow the user to calculate the necessary frequencies from the input dataset itself or to specify their own homogeneous reference dataset of matching population.

Global relatedness estimates (ω_0, ω_1, ω_2 and m) are reported for each pair of individuals analyzed, as well as inferred IBD tracts from the Viterbi algorithm and posterior probabilities from the forward–backward algorithm (Rabiner, 1989). IBD results are reported in spreadsheets. These can be cumbersome to investigate when many shared regions are inferred. However, XIBD also produces novel graphical summaries that allow the user to visualize shared regions in multiple individuals (Fig. 1).

Unlike other algorithms, XIBD does not require a large cohort for accurate IBD inference. In Shaw *et al.* (2015), XIBD was implemented on a single pair of male individuals with X-linked intellectual disability, to verify that a detected variation was contained within a small shared IBD tract on the X chromosome, thus leading to the conclusion that the variant was causal, rather than a technical artifact.

Results from simulation studies can be found in the Supplementary Information, Figures S1 and S2. We note that pseudo-autosomal regions are excluded from analysis. Finally,

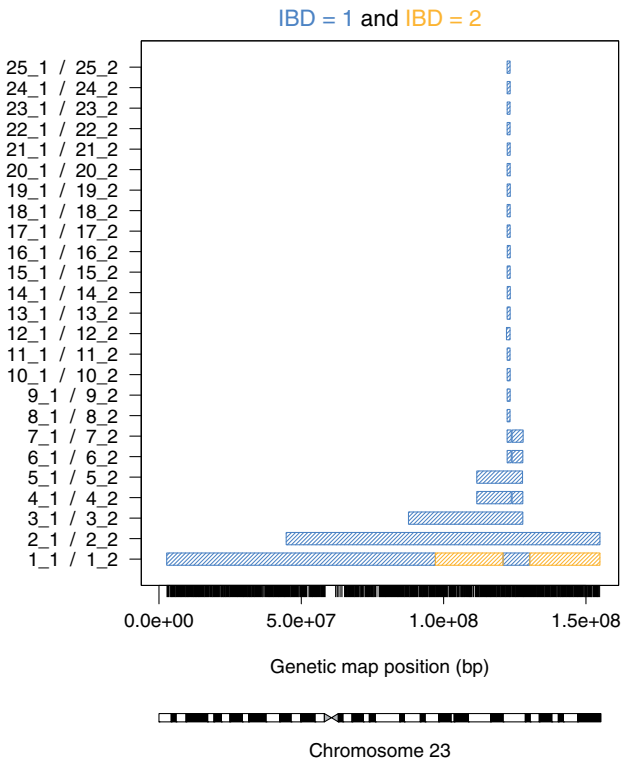

Fig. 1. Summary figure produced by XIBD of simulated IBD segments inherited up to 25 generations from the common ancestor. Pair identifiers are given on the *y*-axis. i.e. 23_1/23_2 are individuals 1 and 2 from generation 23, respectively. The *x*-axis displays the genetic map position in base-pairs with tick markers indicating the SNP positions. The ideogram for the X chromosome is below. Blue rectangles are regions where pairs share one allele IBD while yellow rectangles represent two alleles shared IBD

XIBD can also be extended for use on non-human haploid organisms to identify shared IBD tracts. We encourage feedback from users.

Funding

This work was supported by the Victorian Government's Operational Infrastructure Support Program and Australian Government NHMRC IRIIS [grant numbers 1002098 to M.B., 1054618 to M.B.]; The John and Patricia Farrant Scholarship [to L.H.]; and an Australian Postgraduate Award Scholarship [to L.H.].

Conflict of Interest: none declared.

References

Abney,M. (2009) A graphical algorithm for fast computation of identity coefficients and generalized kinship coefficients. *Bioinformatics*, **25**, 1561–1563.

Albrechtsen,A. *et al*. (2009) Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genet. Epidemiol*., **33**, 266–274.

Brown,M.D. *et al*. (2012) Inferring coancestry in population samples in the presence of linkage disequilibrium. *Genetics*, **190**, 1447–1460.

Browning,B.L. and Browning,S.R. (2013) Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, **194**, 459471.

Epstein,M.P. *et al*. (2000) Improved inference of relationship for pairs of individuals. *Am. J. Hum. Genet*., **67**, 1219–1231.

Gusev,A. *et al*. (2009) Whole population, genome-wide mapping of hidden relatedness. *Genome Res*., **12**, 318326.

McPeek,M.S. and Sun,L. (2000) Statistical test for detection of misspecified relationships by use of genome-screen data. *Am. J. Hum. Genet*., **66**, 1076–1094.

Ott,J. (1999) Introduction and basic genetic principles. *Analysis of Human Genetic Linkage*. 3rd edn. Baltimore, London: Johns Hopkins University Press.

Purcell,S. *et al*. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet*., **81**, 559–575.

Rabiner,L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–286.

Shaw,M. *et al*. (2015) Identical by descent L1CAM mutation in two apparently unrelated families with intellectual disability without L1 syndrome. *Eur. J. Med. Genet*., **58**, 364–368.

The International HapMap Consortium. (2003) The International HapMap Project. (2003) *Nature*, **426**, 789–796.