BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 17 2011, pages 2448–2450
doi:10.1093/bioinformatics/btr423

Genetics and population analysis

Advance Access publication July 15, 2011

IRiS: Construction of ARG networks at genomic scales

Asif Javed[1], Marc Pybus[2], Marta Melé[1,2,†], Filippo Utro[1], Jaume Bertranpetit[2], Francesc Calafell[2] and Laxmi Parida[1,*]

[1]Computational Biology Center, IBM T J Watson Research, Yorktown Heights, NY 10598, USA and [2]Biologia Evolutiva, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain

Associate Editor: Jeffrey Barrett

ABSTRACT

Summary: Given a set of extant haplotypes IRiS first detects high confidence recombination events in their shared genealogy. Next using the local sequence topology defined by each detected event, it integrates these recombinations into an ancestral recombination graph. While the current system has been calibrated for human population data, it is easily extendible to other species as well.

Availability: IRiS (Identification of Recombinations in Sequences) binary files are available for non-commercial use in both Linux and Microsoft Windows, 32 and 64 bit environments from https://researcher.ibm.com/researcher/view_project.php?id=2303

Contact: parida@us.ibm.com

Received on April 25, 2011; revised on July 4, 2011; accepted on July 12, 2011

1 INTRODUCTION

Genetic recombination is a key evolutionary force shaping the sequence variability of the recombining genome. Yet individual recombination events have played a limited role in studies on phylogenetics and population characteristics. Recombinational analysis generally relies on estimated population recombination rates which are statistical averages among individuals from the same population over many generations (Li and Stephens, 2003; McVean *et al.*, 2004). *In vitro* sperm typing experiments define individual recombination events (Jeffreys *et al.*, 2005) and so do familial datasets of related individuals across generations. However these analysis are restricted in time depth, both by the human generation time and the family sizes. The inherent difficulty in identifying past recombinations shared among unrelated individuals lies in the palimpsestic nature of the recombining genome. Over successive generations recombinations conjoin potentially divergent sequences, often overwriting the traces of older events, making the task of reconstructing the *true* recombining genealogy almost impossible. Even under simplifying assumptions, such as the Wright Fisher population model and the parsimony principle, the problem of reconstructing the ancestral recombination graph (ARG) from the haplotypes has been shown to be NP complete (Wang *et al.*, 2001). Thus attempts to solve the reconstruction problem exactly do not scale beyond a few sequences (Song and Hein, 2005).

IRiS (Identification of Recombinations in Sequences) is a suite of programs to study the recombinational variations in populations. It constructs the ARG from the haplotype data in two phases. In the first phase it detects the recombinations and only the ones with high confidence are used in the next phase. In the second phase these recombinations, along with local topology information, are reconciled into an ARG network.

To the best of our knowledge this is the first time that an ARG network is being constructed at a genomic scale for hundreds of samples. The ARG is not fully resolved, due to the difficulties discussed above, hence we call it a *subARG*.

2 APPROACH

The input to IRiS is n sample haplotypes, say H. The output is a directed network G where each sample is on a leaf node of G. The network has two types of nodes: recombination nodes and coalescent nodes. Also, IRiS assigns an estimated age to each internal node of G. We also suggest that a 'distance' estimate between two haplotypes, or nodes, be assigned the age of the most recent common ancestor of the two on G.

2.1 Phase 1: detecting recombination events

IRiS uses a model-based approach to detecting recombinations in H. The underlying combinatorial algorithm, called *dominant, subdominant* or *recombinant* (DSR), uses the parsimony principle; hence attempts to minimize the number of recombinations required to explain the data. This is done by iteratively classifying sets of lineages as DSR. The interested reader is directed to (Parida *et al.*, 2008) for details. The greedy DSR algorithm runs in polynomial time and guarantees the number of estimated recombinations to be within $\epsilon(H)$, a well-behaved function of H, of the optimal (Parida *et al.*, 2009).

However, in practice, many factors such as multiple co-located recombination events, back-mutation events, data and phasing errors and such others play confounding roles. Hence, we estimate the accuracy of the reconstructed events in very general settings through simulations. In particular, we have used the software COSI (Schaffner *et al.*, 2005): the technical details, along with the performance evaluation under different parameter regimes, are described in (Mele *et al.*, 2010). Thus the DSR algorithm has been calibrated for human population data and the optimal parameter combination, consolidating results from multiple runs of DSR, yields a false discovery rate of 5.8%, sensitivity 21% and the median distance between the inferred and true breakpoint position is 1.6 SNPs. The optimal parameters are the default settings of Phase 1 on the IRiS website. These error-rates may not be directly applicable to other species and it is best to estimate these values as done in (Mele *et al.*, 2010).

*To whom correspondence should be addressed.

†Work done during internship at IBM Research.

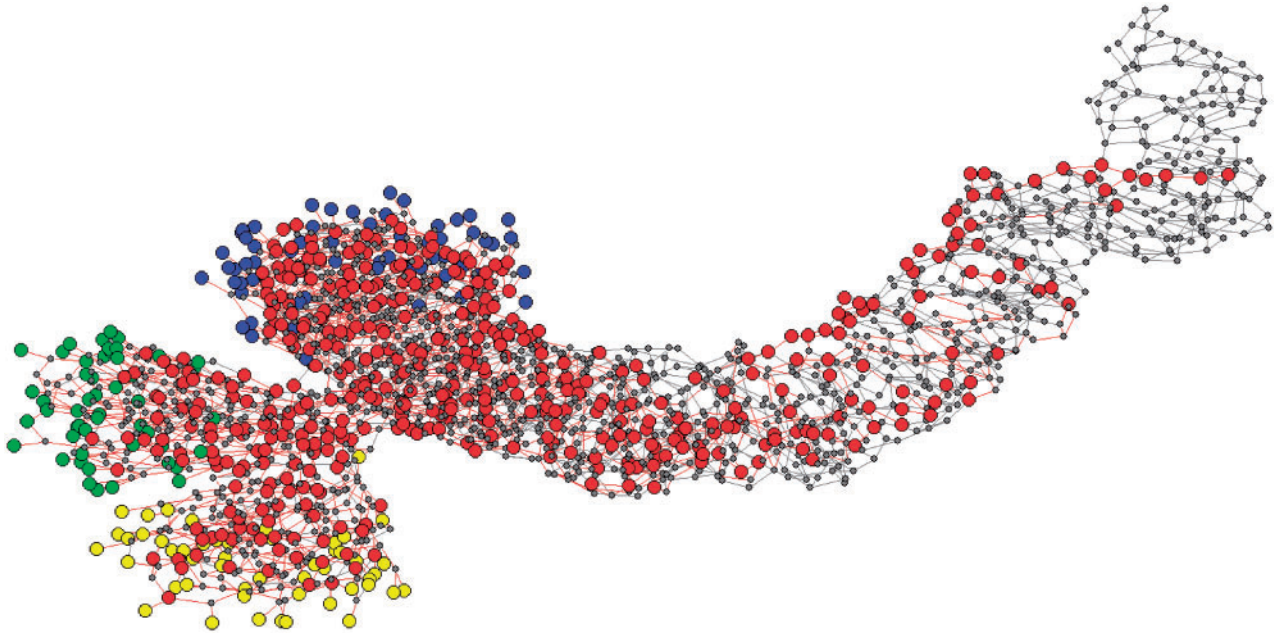

Fig. 1. The fidelity of the reconstructed subARG to the true ARG is shown here. COSI (Schaffner *et al.*, 2005) simulates a demographic scenario with African (blue), Asian (yellow) and European (green) extant samples and denotes the true ARG. The MRCA of all the samples is toward the top right corner. The internal nodes and edges of the reconstructed subARG are shown in red. Note that all the leaf nodes, i.e. the blue, yellow and green colored nodes are (trivially) reconstructed in the subARG. The remaining nodes and edges of the true ARG are shown in gray. The network is rendered using Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek/).

The additional topology information for each recombination event is obtained as follows. Note that each detected recombination is defined by a breakpoint location and a set of extant samples which are descendants of the recombination node. Furthermore, often subsequences surrounding the breakpoint bear evidence of the pair of ancestral haplotypes conjoining at the recombination. These *donor haplotypes* are ancestral to the recombination event and the extant samples bearing them share a common lineage either upstream or downstream of the breakpoint. Thus each recombination defines a trichotomy of related extant samples: those bearing the recombination junction, and the two sets of descendants sharing parentage one either side of the breakpoint. This information is captured as a *recomatrix* for the next phase.

2.2 Phase 2: reconstructing the subARG

In the second phase the local topologies defined by individual recombinations are reconciled to construct the partial ARG. The term partial indicates that the complete genealogy, among all the samples, may not be defined for every segment. Thus the subARG comprises of only a subset of the true ARG nodes and edges, i.e. only the nodes and edges that can be ascertained with high confidence. See Figure 1 for an example of a reconstructed ARG juxtaposed to the true ARG.

The depth, in generations, of an allele correlates with the frequency of its extant descendants. IRiS uses the following age estimate to approximate the time depth of the nodes (Kimura and Ohta, 1969). Let $E(v)$ be the age estimate of node v, then

$$E(v) = \frac{-2p}{1-p} \times ln(p),$$

where p is fraction of samples reachable from v. The reconstructed subARG is validated in extensive coalescent simulations using COSI. In each simulation, every subARG node is mapped to a node in the true ARG based on the genomic segment borne by it, and the set of extant descendants reachable from it within this segment. On an average, the reconstructed subARGs exhibited 92% precision and 81% recall; 17% of the non-leaf true ARG nodes are recovered in each subARG (A.Javed *et al.*, submitted for publication).

3 PERFORMANCE

The running time and memory requirements for IRiS are summarized in Table 1. For these computations we use phased data from the HapMap project (HapMap, 2003). For each experiment a subset of samples and a contiguous set of markers were randomly chosen. Each row of the table indicates the average computed over 20 independent experiments. Notice that in most cases IRiS takes only minutes to complete both the phases. The experiments were performed on a commodity desktop with Intel i7-920 processor and 6GB of total RAM.

4 CONCLUSION

We present a software suite to study genetic recombination. While most currently available software systems estimate only the recombination rate, we undertake the orthogonal task of identifying individual events and reconstructing the ARG network. The system is currently calibrated for human population data based on realistic coalescent (COSI) simulations. To the best of our knowledge this is the first time that an ARG network has been constructed at a genomic

Table 1. Running time and memory requirements of IRiS

Chr	SNPs	Phase 1		Phase 2		Combined	
		Time (mins)	Memory (MB)	Time (mins)	Memory (MB)	Time (mins)	Memory (MB)
200	500	0.8	85.1	0.1	4.3	0.9	85.1
200	1000	1.8	172	0.1	5.8	1.9	172.0
200	1500	2.8	268.5	0.2	7.1	3.0	268.5
200	2000	3.8	363.2	0.3	8.4	4.2	363.2
200	2500	4.7	443.6	0.4	9.4	5.1	443.6
400	500	2.3	222.1	0.4	7.9	2.7	222.1
400	1000	4.9	477.6	1.3	15.0	6.3	477.6
400	1500	7.0	644.6	2.3	16.3	9.3	644.6
400	2000	9.0	835.5	2.5	18.4	11.5	835.5
400	2500	12.9	1149.5	3.9	23.5	16.8	1149.5
600	500	4.0	371.2	1.3	26.2	5.3	371.2
600	1000	8.7	738.7	2.9	42.9	11.6	738.7
600	1500	16.0	1313.8	10.6	67.3	26.6	1313.8
800	500	6.9	618.4	12.9	64.5	19.8	618.4
800	1000	12.6	1148.2	56.2	111.2	68.8	1148.2

Chr denotes the number of chromosome samples.

scale for hundreds of samples. We believe that IRiS is an invaluable tool to comprehensively understand the dynamics of recombining genomes.

Conflict of Interest: none declared.

REFERENCES

Coop,G. *et al*. (2008) High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science*, **319**, 1395–1398.

Jeffreys,A.J. *et al*. (2005) Human recombination hot spots hidden in regions of strong marker association. *Nat. Genet.*, **37**, 601–606.

Kimura,M. and Ohta,T. (1969) The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, **61**, 763–771.

Li,N. and Stephens,M. (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, **165**, 2213–2233.

Liang,L. *et al*. (2007) GENOME: a rapid coalescent-based whole genome simulator. *Bioinformatics*, **23**, 1565–1567.

McVean,G.A. *et al*. (2004) The fine-scale structure of recombination rate variation in the human genome. *Science*, **304**, 581–584.

Melé,M. *et al*. (2010) A new method to reconstruct recombination events at a genomic scale. *PLoS Comput. Biol.*, **6**, e1001010.

Parida,L. *et al*. (2008) Estimating the ancestral recombinations graph (ARG) as compatible networks of SNP patterns. *J. Comput. Biol.*, **15**, 1133–1154.

Parida,L. *et al*. (2009) Minimizing recombinations in consensus networks for phylogeographic studies. *BMC Bioinformatics*, **10** (Suppl. 1), S72.

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

Song,Y.S. and Hein,J. (2005) Constructing minimal ancestral recombination graphs. *J. Comput. Biol.*, **12**, 147–169.

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

Wang,L. *et al*. (2001) Perfect phylogenetic networks with recombination. *J. Comput. Biol.*, **8**, 69–78.