BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 18 2009, pages 2436–2437
doi:10.1093/bioinformatics/btp412

Sequence analysis

HI: haplotype improver using paired-end short reads

Quan Long*, Daniel MacArthur, Zemin Ning and Chris Tyler-Smith

The Wellcome Trust Sanger Institute, Hinxton, Cambs, CB10 1SA, UK

Received on April 27, 2009; revised and accepted on June 25, 2009

Advance Access publication July 1, 2009

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: We present a program to improve haplotype reconstruction by incorporating information from paired-end reads, and demonstrate its utility on simulated data. We find that given a fixed coverage, longer reads (implying fewer of them) are preferable.

Availability: The executable and user manual can be freely downloaded from ftp://ftp.sanger.ac.uk/pub/zn1/HI.

Contact: ql2@sanger.ac.uk

1 INTRODUCTION

With recent advances in DNA sequencing technology, more and more ambitious population-scale sequencing projects have become feasible, e.g. the 1000 Genomes Project (http://www.1000genomes.org/page.php). Haplotype reconstruction is an important step in many genetic analyses. Currently, there are several successful population-genetic model-based haplotype inference tools using different methodologies, such as the MCMC-based PHASE (Stephens *et al*., 2001) or the HMM-based fastPHASE (Scheet and Stephens, 2006). However, these phasing algorithms assume the use of genotype data. To apply these tools to next-gen resequencing data, the typical procedure involves: (i) mapping the reads to the reference genome with a mapping tool; (ii) calling SNPs from the consensus sequence; and (iii) importing the SNP files into a phasing tool as if they were generated from a genotyping platform. An important source of information from the raw data is lost in this procedure if we start from paired-end reads. That is, if a pair of reads happens to carry a pair of heterozygous SNPs, it indicates the true chromosomal phase of these SNPs (Fig. 1). It is easy to imagine that the information linking two SNPs can be expanded to blocks, enabling us to phase a number of SNPs in a local region.

Making use of sequence read information to resolve haplotypes is not new. Actually, in traditional capillary sequencing projects (Kim *et al*., 2007), people have already used information from fosmid end assembly to infer haplotypic phase (Li *et al*., 2004). However, their approach was designed for traditional sequencing projects in which the reads are relatively long (500–800 bp) and the number of reads typically not very large, and therefore may not be suitable for high-throughput short-read sequencing data. Given the current read length of 35–75 bp from the short-read platforms, it is impossible to resolve individual haplotypes via *de novo* assembly (Kim *et al*., 2007). Reliable phasing must still rely on the population-genetic models that have been applied successfully to genotype data.

*To whom correspondence should be addressed.

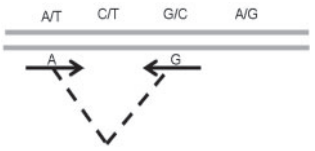

Fig. 1. If a pair of reads covers two heterozygous positions, the alleles carried (A and G in the example shown) must be on the same chromosome. Therefore we can phase the genotype (A/T, G/C) to (A, G) and (T, C).

In this article, we present our program, Haplotype Improver (HI) to improve haplotype reconstruction using paired-end short reads. Assuming that the users have run an existing phaser, HI processes the paired-end information in the raw data to form blocks of haplotypes and compares them with the output of a phasing tool (currently HI supports PHASE and fastPHASE). When inconsistencies are found, HI will decide whether or not, and at which loci, to change the haplotype reconstructions according to its calculations.

2 METHODS

First, we look for paired-end reads carrying two heterozygous SNPs in a tested sample. To facilitate the calculation, we designed two levels of hash tables. The first level is a hash table to store all the locations of heterozygous SNPs from a given individual, and we therefore scan all the reads from the alignment file to identify the relevant ones. Let m be the number of the heterozygous sites. The second level of hash tables initially consists of m hash tables similar to the first level of hash tables. If the locations of two SNPs, i and j, respectively, resulting from the paired reads are on the same chromosome, we mask i and j *in the second level hash tables,* and then record the corresponding haplotype in this combined region (note that it is not necessary for it to be a continuous region because we may have another SNP k located between i and j). Finally, after all the mapped reads have been scanned, there should be m' ($m' \leq m$) masked hash tables remaining in the dataset where each table stands for a haplotype block (again, the 'block' is not necessarily a continuous region).

Next, we check for inconsistencies between the information in blocks and the results provided by the phasing tool. By 'inconsistency', we mean that the alleles supposed to share the same haplotype are mistakenly distributed to two separate chromosomes by the phasing tool. We calculate a ratio based on a probability model to decide which segment to move to make the result consistent (see User Manual for details.) Finally, adjusted haplotypes of the sample are reported. When the data quality is low, the information from multiple paired-end reads themselves may be inconsistent. In this case, HI will not take any action.

The time required by this algorithm is linearly proportional to the number of reads, and the space required is linearly proportional to the number of heterozygous sites in all individuals. Compared with the time-consuming

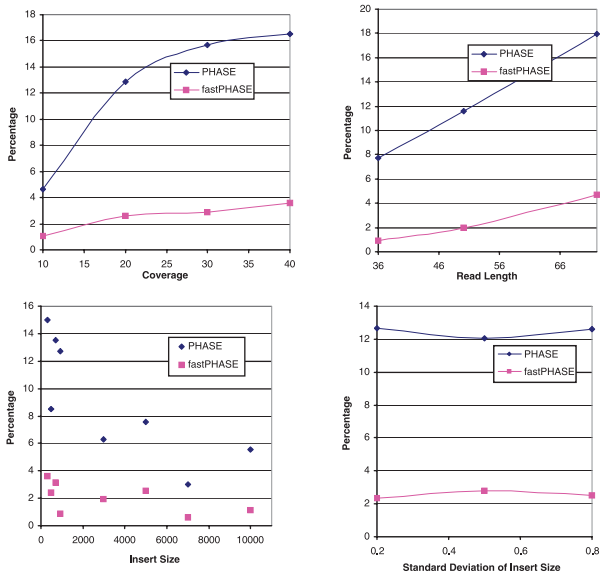

Fig. 2. The *y*-axis shows the percentage of phasing errors eliminated by HI. The *x*-axis shows the values of the parameter indicated. In each plot, the performance value varied with the parameter of interest; while the other three parameters were marginalized. (The data for insert size longer than 3 kb are only used in the insert size plot.)

sampling process of phasing tools, the time added by HI is small. Any user who can run phasing tools can afford the RAM for HI.

3 SIMULATIONS AND EXPERIMENTAL DESIGN

To validate the performance and explore the best experimental design strategy, we tested HI on simulated data.

In each simulation, we use MS (Hudson, 2002) to generate a population of 100 000 haploid SNP sequences under the standard neutral model and sample 60 sequences from it to form 30 diploid individuals. We embed them within non-repetitive regions of the human genome. The average SNP density is 3.3 SNP/kb and the heterozygous SNP density is 0.72 het/kb. We then simulate Illumina reads and map them back to the reference genome to call SNPs. In the simulation, we use SSAHA (Ning *et al*., 2001) for read mapping and SNP calling. We have four parameters: coverage per base (with values 10, 20, 30 and 40), mean insert size (with values 300, 500, 700, 900, 3k, 5k, 7k, 10k), standard deviation of insert size (with values 0.2, 0.5 and 0.8), and read length (with values 36, 50 and 72). We tested each combination of the values of the four parameters to see how many errors caused by phasing could be improved. The marginalized results are shown in Figure 2. When the insert size is moderate, one can see that for PHASE, usually more than 10% errors can be eliminated, whereas for fastPHASE, this proportion is around 1–5%.

The analysis of simulated data indicates that longer read length is preferred. Note that this conclusion is not trivial because, given the same coverage, longer read length means fewer reads. From the simulations, we observed that very long insert size significantly larger than the mean heterozygous SNP spacing looks not preferable for HI itself. But long insert size may be an advantage for other purpose, e.g. read mapping, and may therefore also impact the precision of haplotype reconstruction. Finally, high sequence coverage is preferred, a requirement that will become easier to satisfy as the throughput of new sequencing technologies continues to increase.

4 DISCUSSION

Researchers familiar with the statistical framework for haplotype reconstruction may have the following concerns: (i) Why not integrate paired-end information with population-genetic models by modifying the MCMC sampling schema to improve phasing precision? We are currently developing such an algorithm, but the method presented here represents a simple and robust initial approach to improve phasing using paired-end information. (ii) Modifying the individual haplotype will change the haplotype distribution in a non-statistically sound manner, therefore causing problems in downstream analyses. In fact, our simulation shows that the haplotype distribution is only slightly altered or unchanged by HI (data not shown).

There are many alignment formats available. In its current form, HI uses SSAHA's CIGAR alignment format. We will soon switch to support the SAM (http://samtools.sourceforge.net/) format which may be accepted as the uniform standard by the community in the near future. Because SAM files are usually sorted by chromosomal coordinate, following this transition it will be straightforward to reduce the time requirement to $O(m\log(n))$, where n, m are the number of reads and heterozygous positions, respectively.

Funding: The Wellcome Trust.

Conflict of Interest: none declared.

REFERENCES

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

Kim,J.H. *et al*. (2007) Diploid genome reconstruction of *Ciona intestinalis* and comparative analysis with *Ciona savignyi*. *Genome Res.*, **17**, 1101–1110.

Li,L.M. *et al*. (2004) Haplotype reconstruction from SNP alignment. *J. Comput. Biol.*, **11**, 505–516.

Ning,Z. *et al*. (2001) SSAHA: a fast search method for large DNA databases. *Genome Res.*, **11**, 1725–1729.

Scheet,P. and Stephens,M. (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.*, **78**, 629–644.

Stephens,M. *et al*. (2001) A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.*, **68**, 978–989.