Bioinformatics, 32(18), 2016, 2861–2862
doi: 10.1093/bioinformatics/btw356
Advance Access Publication Date: 9 June 2016
Applications Note

OXFORD

Genetics and population analysis

GHap: an R package for genome-wide haplotyping

Yuri T. Utsunomiya[1,2,*], Marco Milanesi[2,3], Adam T. H. Utsunomiya[2,3], Paolo Ajmone-Marsan[4] and José F. Garcia[1,2,3]

[1]Departamento de Medicina Veterinária Preventiva e Reprodução Animal, UNESP – Univ. Estadual Paulista, Faculdade de Ciências Agrárias e Veterinárias, 14884-900 Jaboticabal, São Paulo, Brazil, [2]International Atomic Energy Agency (IAEA) Collaborating Centre on Animal Genomics and Bioinformatics, 16050-680 Araçatuba, São Paulo, Brazil, [3]Departamento de Apoio, Produção e Saúde Animal, UNESP – Univ. Estadual Paulista, Faculdade de Medicina Veterinária de Araçatuba, 16050-680 Araçatuba, São Paulo, Brazil and [4]Istituto di Zootecnica, Università Cattolica del Sacro Cuore, 29122 Piacenza, Italy

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

Received on February 12, 2016; revised on May 3, 2016; accepted on May 31, 2016

Abstract

The GHap R package was designed to call haplotypes from phased marker data. Given user-defined haplotype blocks (HapBlock), the package identifies the different haplotype alleles (HapAllele) present in the data and scores sample haplotype allele genotypes (HapGenotype) based on HapAllele dose (i.e. 0, 1 or 2 copies). The output is not only useful for analyses that can handle multi-allelic markers, but is also conveniently formatted for existing pipelines intended for bi-allelic markers.

Availability and implementation: https://cran.r-project.org/package=GHap

Contact: ytutsunomiya@gmail.com

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

1 Introduction

The use of high-density marker panels in genomics relies on the concept of linkage disequilibrium (LD) and tagging, such that information from unobserved variants can be indirectly captured by correlation with nearby markers (Bush and Moore, 2012). Methods for genomic analysis are usually based on single markers, ignoring that unobserved variants may be better modeled by the use of phase information (Browning and Browning, 2008). Moreover, the need for tools to perform haplotype calls from phased data has been underserved in spite of the growing interest in haplotype-based analyses in the last years. Here we describe GHap, an R package designed to call haplotypes from phased data. The goal of GHap is to compute summary statistics for haplotype blocks (HapBlock) and haplotype alleles (HapAllele), as well as to construct a matrix of haplotype genotypes (HapGenotype). As a general framework, each HapAllele can be treated as a pseudo-marker in downstream analyses, facilitating the incorporation of phase information in existing pipelines. This approach differs from competing methods as it uses HapAllele as markers, instead of hidden haplotype states generated by expectation maximization or hidden Markov models.

2 Implementation

2.1 Loading and manipulating data

The GHap input format is described in the online supplementary information and can be derived from popular phasing programs such as SHAPEIT2 (O'Connell *et al.*, 2014). GHap assumes that family information, if applicable, has been taken into account during phasing. As GHap assumes known phase, low quality phasing will directly impact the reliability of haplotype calls. The phased data is loaded using *ghap.loadphase()*. The *ghap.maf()* function can be used to identify markers with low polymorphic information content. The *ghap.subsetphase()* function subsets the data by inactivating

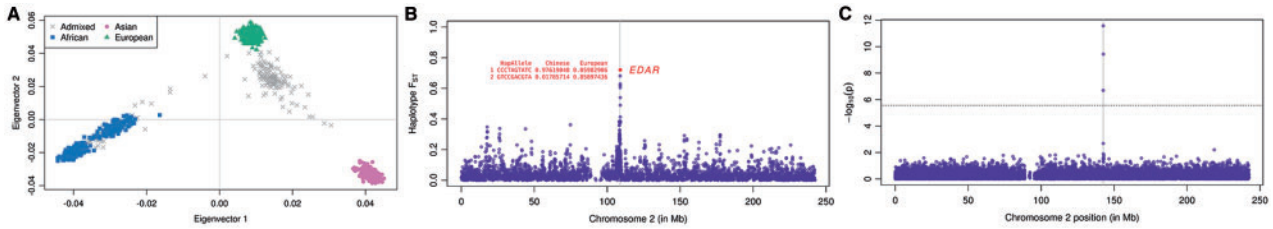

Fig. 1. Examples of applications of the GHap package with Human HapMap Project Phase 3 data. Analyses were based on HapBlocks of 10 markers with overlaps of 5 markers between consecutive blocks. (**A**) Clustering of subjects based on a PCA of the HapAllele relationship matrix. (**B**) Haplotype-based F_{ST} for Chinese x European. (**C**) Mixed model association analysis using simulated phenotypes. The vertical grey line marks the position of the simulated causal nucleotide, and the horizontal dashed line marks the Bonferroni significance threshold ($P < 2.73 \times 10^{-6}$) (Color version of this figure is available at *Bioinformatics* online.)

specified samples and markers, while *ghap.mergephase()* combines different phased data. The data can be exported using *ghap.outphase()*.

2.2 Genome-wide haplotyping procedure

Let a HapBlock be a user-defined set of adjacent markers and the haplotype library (HapLibrary) be the collection of observed HapAlleles for that HapBlock. The haplotyping procedure implemented in the *ghap.haplotyping()* function is straightforward: each HapAllele in the HapLibrary is treated as a marker, and HapGenotypes are scored as 0, 1 or 2 allele copies (for more information see **online supplementary information**). HapGenotypes can then be loaded into R using *ghap.loadhaplo()* and manipulated with *ghap.subsethaplo()* and *ghap.mergehaplo()*, or exported to text file and the transposed format from PLINK (Purcell *et al.*, 2007) with *ghap.outhaplo()* and *ghap.hap2tped()*, respectively. Typically, the user may want to target a specific HapBlock based on prior information (e.g. pre-computed LD blocks). We also provide an alternative approach through the *ghap.blockgen()* function, which allows the user to specify arbitrary windows and step size based on markers or segments.

2.3 Haplotype statistics and auxiliary functions

The *ghap.hapstats()* function computes a series of summary statistics for HapAlleles, which includes: number of observations, frequency, observed number of homozygotes and heterozygotes, expected number of homozygotes and three different measures of deviations from Hardy–Weinberg equilibrium. We also implemented a series of auxiliary functions, namely *ghap.blockstats()*, *ghap.fst()*, *ghap.ancestral()*, *ghap.kinship()*, *ghap.pca()*, *ghap.blmm()*, *ghap.assoc()*, *ghap.blup()* and *ghap.simpheno*, to estimate block expected heterozygosity and number of HapAlleles, haplotype-based F_{ST}, HapAllele origin, relationship matrices, principal components, linear mixed models, association analyses, Best Linear Unbiased Predictor (BLUP) and simulate phenotypes, respectively. Additionally, given a vector of arbitrary scores for HapAlleles, the *ghap.profile()* function allows for computing individual profiles as *sum*(HapGenotypes*scores).

3 Examples

We tested GHap using reference phased data available at the IMPUTE2 (Howie *et al.*, 2009) software website (https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#reference). These data derive from the HapMap Project Phase 3 (Altshuler *et al.*, 2010), and comprise 1011 subjects from 11 human populations and 20 000 random SNPs mapping to chromosome 2. This dataset is available through the *ghap.makefile()* function. Benchmarking of the main tasks showed that elapse time scaled linearly with increasing number of samples and markers in a trial up to 100 000 SNPs and 50 000 subjects (online supplementary information). We performed three analyses: (i) Principal Components Analysis (PCA); (ii) detection of divergent loci between Chinese and Europeans and (iii) mixed model association analysis with phenotypes simulated based on the real genotypes. Similar analyses can be done following the package documentation. Although based on a small example set of markers in a single chromosome, the haplotype calls generated by GHap resolved the known genetic structure in the HapMap dataset (Fig. 1A) . The haplotype-based F_{ST} analysis (Fig. 1B) identified a previously reported signature of selection in Chinese encompassing *EDAR* (Sabeti *et al.*, 2007), with the top scoring HapBlock mapping to its intragenic region. Finally, the association analysis (Fig. 1C) efficiently pinpointed the HapAllele segregating with the simulated causal variant.

4 Conclusion

The GHap package provides means for haplotype calling, facilitating the incorporation of phase information in genome-wide analyses. The package is available at: https://cran.r-project.org/package=GHap.

Funding

This research received financial support from: São Paulo Research Foundation (FAPESP – http://www.fapesp.br/) (process 2014/01095-8); The National Council for Scientific and Technological Development (CNPq – http://www.cnpq.br/) (process 407502/2013-0).

Conflict of Interest: none declared.

References

Altshuler,D.M. *et al.* (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**, 52–58.

Browning,B.L. and Browning,S.R. (2008) Haplotypic analysis of Wellcome Trust Case Control Consortium data. *Hum. Genet.*, **123**, 273–280.

Bush,W.S. and Moore,J.H. (2012) Chapter 11: Genome-wide association studies. *PLoS Comput. Biol.*, **8**, e1002822.

Howie,B.N. *et al.* (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.*, **5**, e1000529.

O'Connell,J. *et al.* (2014) A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet*, **10**, e1004234.

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.

Sabeti,P.C. *et al.* (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature*, **449**, 913–918.