Genome analysis

HaploTypo: a variant-calling pipeline for phased genomes

Cinta Pegueroles¹,†,‡, Verónica Mixão¹,†,‡, Laia Carrete¹,†, Manu Molina¹,† and Toni Gabaldón ²,3,*†

¹Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, Barcelona 08003, Spain, ²Universitat Pompeu Fabra (UPF), Barcelona, Spain and ³ICREA, Barcelona 08010, Spain

*To whom correspondence should be addressed.
†Present address: Barcelona, Supercomputing Center (BSC-CNS) and Institute for Research in Biomedicine (IRB), Barcelona, Spain
‡The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.
Associate Editor: Inanc Birol

Received on July 16, 2019; revised on November 28, 2019; editorial decision on December 9, 2019; accepted on December 10, 2019

Abstract

Summary: An increasing number of phased (i.e. with resolved haplotypes) reference genomes are available. However, the most genetic variant calling tools do not explicitly account for haplotype structure. Here, we present HaploTypo, a pipeline tailored to resolve haplotypes in genetic variation analyses. HaploTypo infers the haplotype correspondence for each heterogeneous variant called on a phased reference genome.
Availability and implementation: HaploTypo is implemented in Python 2.7 and Python 3.5, and is freely available at https://github.com/gabaldonlab/haplotypo, and as a Docker image.
Contact: toni.gabaldon.bcn@gmail.com
Supplementary information: Supplementary data are available at Bioinformatics online.

1 Motivation

The heterozygosity (i.e. the presence of alternative alleles at the same locus) present in diploid organisms can complicate genome analyses, particularly when the levels of heterozygosity are high. Over the last years, several bioinformatics tools have been developed to account for this sequence complexity. These include pipelines and algorithms to assist during the genome assembly process (Pryszc and Gabaldón, 2016; Safonova et al., 2015), subsequent phasing of assembled genomes (Chin et al., 2016; Edge et al., 2017; Pan et al., 2014) or allele-specific transcriptomic analysis (Deonovic et al., 2017; Romanel et al., 2015). However, and to the best of our knowledge, available variant calling tools do not explicitly account for phased genomes. As a result, the user has to decide between using the combined phased haplotypes as reference and thereby losing heterozygosity information or, alternatively, using only one of the haplotypes as reference and sacrificing haplotype information. An illustrating example of such problem is studies on the heterozygous yeast pathogen Candida albicans. Although the diploid genome of this pathogen was phased in 2013 (Muzzey et al., 2013), subsequent studies have only used one of the haplotypes (Bensasson et al., 2019; Ropars et al., 2018), thereby losing the valuable haplotype information. Given the increasing amount of highly heterozygous genomes, including those from hybrids (Mixão and Gabaldón, 2018), and the relevance of phased information to reconstruct their population structures and evolutionary histories, there is an urgent need for solutions that allow the exploitation of phased genomes in genomic variation analysis. To fill in this gap, we developed HaploTypo, a python-based pipeline that, in the presence of a phased reference genome, provides detailed genome variation resolved at the haplotype level. HaploTypo is not a de novo genome phasing tool, but a tool to phase variants in re-sequencing analysis, using information of an already phased genome, resulting in a fast and accurate assessment of heterozygosity levels and reconstruction of haplotypes.

2 Implementation

HaploTypo requires as input the phased haplotypes of a diploid genome, and filtered genomic paired-end sequencing reads or, alternatively, their alignment to each of the phased haplotypes. The pipeline is divided in four modules, which can be run in block or separately (Fig. 1). The first module aligns the genomic paired-end reads independently to each of the phased haplotypes using BWA-MEM (Li, 2013). The second module performs variant calling on the two generated alignments using GATK (McKenna et al., 2010), BCFtools (Liu, 2011) or FreeBayes (Garrison and Marth, 2012) followed by variant filtration. From here, variability information is obtained for each reference haplotype independently. The third module of HaploTypo implements a variant phasing algorithm that, based on the comparison of reference haplotypes and previously called variants, infers which variants correspond to each haplotype.
Phase results, specially at low divergence levels (Supplementary Table S1). It is worth noting that accuracy and specificity remained high and stable (>99% and 100% respectively) independently of the levels of divergence and the variant caller used. When using HaploTypo, reads are mapped independently to the two haploid references of a phased genome (approach with the best results, as shown above) and outputs the two haplotypes, correctly phasing >99% of the positions independently of the variant caller used, with few exceptions (Supplementary Table S2). The unphased cases always represented ambiguous situations that cannot possibly be resolved with this type of data (see manual for details), and the user can decide whether to include them in the VCF or not. In addition, HaploTypo also reports positions that have incomplete results in the two haplotypes and therefore are likely to be mapping or variant calling errors (unresolved positions, see Table 1 from the manual for details). HaploTypo benchmarking was performed on a workstation [Intel(R) Xeon(R) CPU E5-1650 v3] and 64 GB of RAM with default number of threads. The total running time ranged from 1 to 15 h, depending on the level of heterozygosity of the dataset and the variant caller (Supplementary Table S3). Hence HaploTypo is a user-friendly tool that eases variant analyses and allows to incorporate haplotype-specific information when a phased reference genome is available.

Acknowledgements
We thank all the members of Gabaldón’s lab for the relevant discussions.

Funding
This work was supported by the European Union’s Horizon 2020 research and innovation programme under the grant agreement ERC-2016-724173 and Marie Skłodowska-Curie grant agreements N° 642095, and 747607; the Spanish Ministry of Economy, Industry, and Competitiveness (MEIC) for the EMBL partnership, and grants ‘Centro de Excelencia Severo Ochoa’ SEV-2012-0208, and BFU2015-67107 co-founded by European Regional Development Fund (ERDF); the CERCA Programme/Generalitat de Catalunya; from the Catalan Research Agency (AGAUR) SGR857; and INB Grant (PT17/0009/0023—ISCIII-SGFU/ERDF).

Conflict of Interest: none declared.

References
Bensasson,D. et al. (2019) Diverse lineages of live on old oaks. Genetics, 211, 277–288.
Chin,C.-S. et al. (2016) Phased diploid genome assembly with single-molecule real-time sequencing. Nat. Methods, 13, 1050–1054.
Deonovic,B. et al. (2017) IDP-ASE: haplotyping and quantifying allele-specific expression at the gene and gene isoform level by hybrid sequencing. Nucleic Acids Res., 45, e32.
Edge,P. et al. (2017) HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. Genome Res., 27, 801–812.
Garrison,E. and Marth,G. (2012) Haplotype-based variant detection from short-read sequencing. arXiv, arXiv:1207.3907.
Li,H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv, arXiv: 1303.3997
Li,H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics, 27, 2987–2993.
McKenna,A. et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res., 20, 1297–1303.
Mínxio,V. and Gabaldón,T. (2018) Hybridization and emergence of virulence in opportunistic human yeast pathogens. Yeast, 35, 5–20.
Muzzey,D. et al. (2013) Assembly of a phased diploid Candida albicans genome facilitates allele-specific measurements and provides a simple model for repeat and indel structure. Genome Biol., 14, R97.
Pan, W. et al. (2014) WinHAP2: an extremely fast haplotype phasing program for long genotype sequences. *BMC Bioinformatics*, 15, 164.

Pryszcz, L.P. et al. (2015) The genomic aftermath of hybridization in the opportunistic pathogen *Candida metapsilosis*. *PLoS Genet.*, 11, e1005626.

Pryszcz, L.P. and Gabaldón, T. (2016) Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res.*, 44, e113.

Romanel, A. et al. (2015) ASEQ: fast allele-specific studies from next-generation sequencing data. *BMC Med. Genomics*, 8, 9.

Ropars, J. et al. (2018) Gene flow contributes to diversification of the major fungal pathogen *Candida albicans*. *Nat. Commun.*, 9, 2253.

Safonova, Y. et al. (2015) dipSPAdes: assembler for highly polymorphic diploid genomes. *J. Comput. Biol.*, 22, 528–545.