WhatsHap: Haplotype Assembly for Future-Generation Sequencing Reads

Murray Patterson[1,2,⋆], Tobias Marschall[1,⋆], Nadia Pisanti[3,4], Leo van Iersel[1], Leen Stougie[1,5], Gunnar W. Klau[1,5,⋆⋆], and Alexander Schönhuth[1,⋆⋆]

[1] Life Sciences, CWI Amsterdam, The Netherlands
{m.d.patterson,t.marschall,gunnar.klau,a.schoenhuth}@cwi.nl
[2] LBBE, CNRS and Université de Lyon 1, Villeurbanne, France
[3] Department of Computer Science, University of Pisa, Italy
[4] LIACS, Leiden University, The Netherlands
[5] VU University Amsterdam, The Netherlands

Abstract. The human genome is diploid, that is each of its chromosomes comes in two copies. This requires to *phase* the *single nucleotide polymorphisms (SNPs)*, that is, to assign them to the two copies, beyond just detecting them. The resulting haplotypes, lists of SNPs belonging to each copy, are crucial for downstream analyses in population genetics. Currently, statistical approaches, which avoid making use of direct read information, constitute the state-of-the-art. *Haplotype assembly*, which addresses phasing directly from sequencing reads, suffers from the fact that sequencing reads of the current generation are too short to serve the purposes of genome-wide phasing.

Future sequencing technologies, however, bear the promise to generate reads of lengths and error rates that allow to bridge all SNP positions in the genome at sufficient amounts of SNPs per read. Existing haplotype assembly approaches, however, profit precisely, in terms of computational complexity, from the limited length of current-generation reads, because their runtime is usually exponential in the number of SNPs per sequencing read. This implies that such approaches will not be able to exploit the benefits of long enough, future-generation reads.

Here, we suggest WhatsHap, a novel dynamic programming approach to haplotype assembly. It is the first approach that yields provably optimal solutions to the *weighted minimum error correction (wMEC)* problem in runtime linear in the number of SNPs per sequencing read, making it suitable for future-generation reads. WhatsHap is a *fixed parameter tractable (FPT)* approach with coverage as the parameter. We demonstrate that WhatsHap can handle datasets of coverage up to 20x, processing chromosomes on standard workstations in only 1-2 hours. Our simulation study shows that the quality of haplotypes assembled by WhatsHap significantly improves with increasing read length, both in terms of genome coverage as well as in terms of switch errors. The switch error rates we achieve in our simulations are superior to those obtained by state-of-the-art statistical phasers.

⋆ Joint first authorship.
⋆⋆ Joint last authorship.

R. Sharan (Ed.): RECOMB 2014, LNBI 8394, pp. 237–249, 2014.

1 Introduction

The human genome is *diploid*, that is, each of its chromosomes comes in two copies (except for sex chromosomes in males), one from the mother and one from the father. These parental copies are affected by different *single nucleotide polymorphisms (SNPs)*, and assigning the variants to the copies is an important step towards the full characterization of an individual genome. The corresponding assignment process is referred to as *phasing* and the resulting groups of SNPs are called *haplotypes*. Phasing SNPs in population studies allows to, for example, identify selective pressures and subpopulations, and to link possibly disease-causing SNPs with one another [13]. This explains that phasing SNPs has been an instrumental step in many human whole-genome projects [5,28]. In the meantime, globally concerted efforts have generated *reference panels* of haplotypes, for various populations, which may serve corresponding downstream analyses [29,30].

There are two major approaches to phasing variants. The first class of approaches relies on *genotypes* as input, which are lists of SNP alleles, together with their zygosity status. While *homozygous* alleles show on both chromosomal copies, and obviously apply for both haplotypes, *heterozygous* alleles show on only one of the copies, and have to be partitioned into two groups. If m is the number of heterozygous SNP positions, there are 2^m many possible haplotypes. This illustrates that directly phasing from genotype data is a hard computational problem. The corresponding approaches are usually statistical in nature, and they integrate existing reference panels. The underlying assumption is that the haplotypes to be computed are a mosaic of reference haplotype blocks that arises from recombination during meiosis. The output is the statistically most likely mosaic, given the observed genotypes. Most prevalent approaches are based on latent variable modeling [17,21,26]. Other approaches use Markov chain Monte Carlo techniques [23].

The other class of approaches makes direct usage of sequencing read data. Such approaches virtually assemble reads from identical chromosomal copies and are referred to as *haplotype assembly* approaches. Following the parsimony principle, the goal is to compute two haplotypes to which one can assign all reads with the least amount of sequencing errors to be corrected and/or erroneous reads to be removed. Among such formulations, the *minimum error correction (MEC)* problem has gained most of the recent attention. The MEC problem, which we will formally define in Section 2, consists of finding the minimum number of corrections to the SNP values to be made to the input in order to be able to arrange the reads into two haplotypes without conflicts. A major advantage of MEC is that it can be easily adapted to a weighted version (wMEC), in order to deal with phred-based error probabilities. Such error schemes are common in particular for *next-generation sequencing (NGS)* data. An optimal solution for the wMEC problem then translates to a maximum likelihood scenario relative to the errors to be corrected.

In tera-scale sequencing projects, e.g., [5,28], ever increasing read length and decreasing sequencing cost make it clearly desirable to phase directly from read