Advanced Algorithms in Computational Biology

DM845

Martin Østergaard Villumsen

mvill11@student.sdu.dk

University of Southern Denmark

May 14, 2015

1 Introduction

The human genome is diploid which means that each cell has two homologous copies of each chromosome, one from the mother and one from the father. In order to understand the genetic basis for different diseases (e.g. cancer) it is not sufficient to detect the SNPs, we also need to assign each SNP to the two copies of the chromosome and current methods for doing so suffers from the fact that the reads are too short [1].

In this project we will develop a simple read simulator, which generates reads that are much longer than those obtained from e.g. Next-Gen sequencing machines. We will simulate reads with different parameters such as read length and sequencing error probability and assign SNPs to each chromosome from these reads using WhatsHap, a novel dynamic programming approach to haplotype assembly described in [1]. We will then evaluate the results and compare them with those presented in [1].

1.1 Biological Problem

The task of assigning SNPs to a concrete chromosome is called phasing and the resulting groups of SNPs are called haplotypes. We need to discover and phase all these SNPs in order to gain a better understanding for e.g. some diseases by linking possibly disease-causing SNPs with one another. By doing so we can reconstruct haplotypes from a collection of sequenced reads which is known as haplotype assembly.

1.2 Computational Problem

There are two groups of single nucleotide variants: Those that are homozygous and those that are heterozygous. Individuals that are homozygous at every locus or heterozygous at just one locus can easily be phased, however, if we have m heterozygous SNPs, there are 2^m possible haplotypes which illustrates that it is a hard computational problem. Therefore we are only concerned with heterozygous SNPs when doing haplotype assembly.

What also makes this a computational hard problem is the fact that we want to phase an reconstruct the haplotypes directly from sequencing reads.

2 Haplotype Assembly with WhatsHap

There are two major approaches to phasing variants: One approach relies on genotypes as input along with the zygosity status of the SNPs, and the other approach phases directly from sequencing read data [1]. WhatsHap belongs to the latter.

WhatsHap has been developed with the parsimony principle in mind, i.e. computing two haplotypes to which we can assign all reads while minimizing the amount of sequencing errors to be corrected or removed [1]. This resembles the minimum error correction (MEC) problem which 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 [1]. This can easily be adapted to a weighted version of the problem, namely the weighted minimum error correction problem (wMEC). The weight in this case reflects the relative confidence that a single nucleotide is correctly sequenced.

Even though the wMEC problem is NP-hard WhatsHap creates an exact solution to the problem in linear time. This is done by the use of dynamic programming and assuming a bounded coverage, i.e. the implementation can solve the problem in linear time on datasets of maximum coverage up to 20x [1]. The algorithm is a fixed parameter tractable approach to the wMEC where the running time is only depending on the coverage, i.e. the number of reads that cover a SNP position.

The input to the wMEC problem is a matrix with entries $\in \{0, 1, -\}$ where each row corresponds to a read and each column corresponds to a SNP position. Each entry is associated with a weight telling how likely it is that the entry is correctly sequenced. We want to find a minimum weight solution and when these weights are log-likelihoods, summing them up corresponds to multiplying probabilities, thus finding the minimum weight solution corresponds to finding a maximum likelihood bipartition of the reads [1]. And this is basically what the authors of WhatsHap is using a dynamic programming approach to do. By using this approach they find an optimal solution to the wMEC problem for each column of the matrix and then an optimal bipartition of the reads can be obtained by backtracking along the columns of the dynamic programming table. For further details see [1].

3 Building a DNA Simulator

4 Results and Discussion

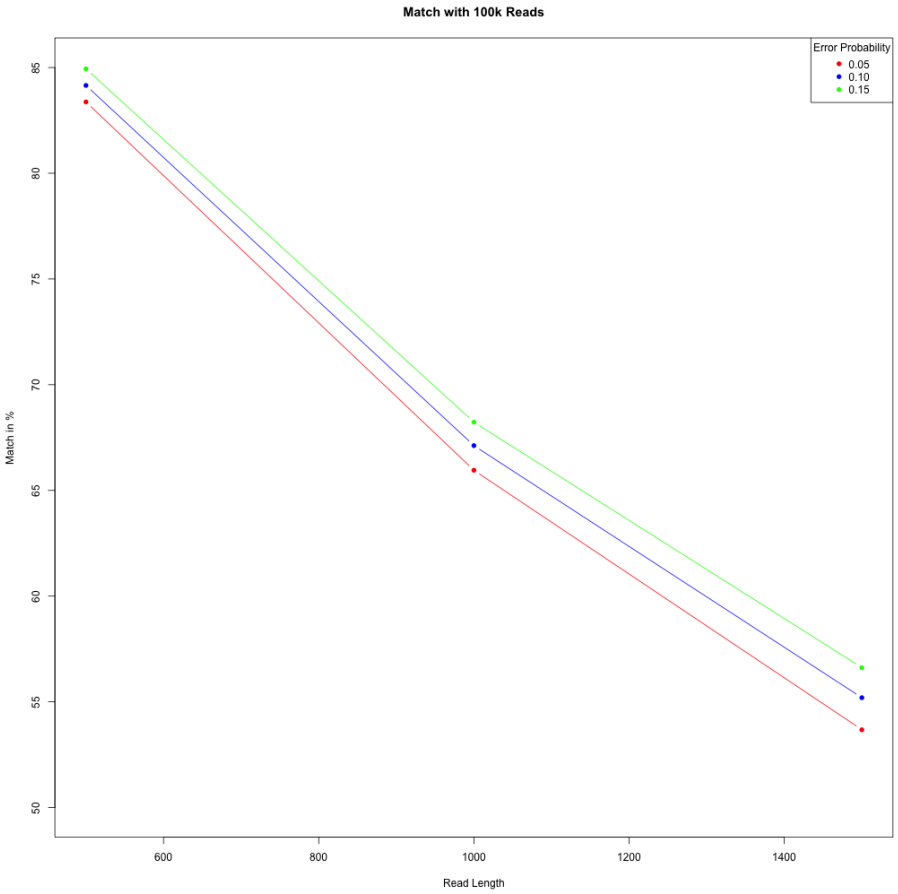

Figure 1: Results of haplotype assembly of 100.000 reads

5 Conclusion

References

[1] M. Patterson, T. Marschall, N. Pisanti, L. van Iersel, L. Stougie, G. W. Klau, and A. Schönhuth. Whatshap: Haplotype assembly for future-generation sequencing reads. In *Research in Computational Molecular Biology - 18th*

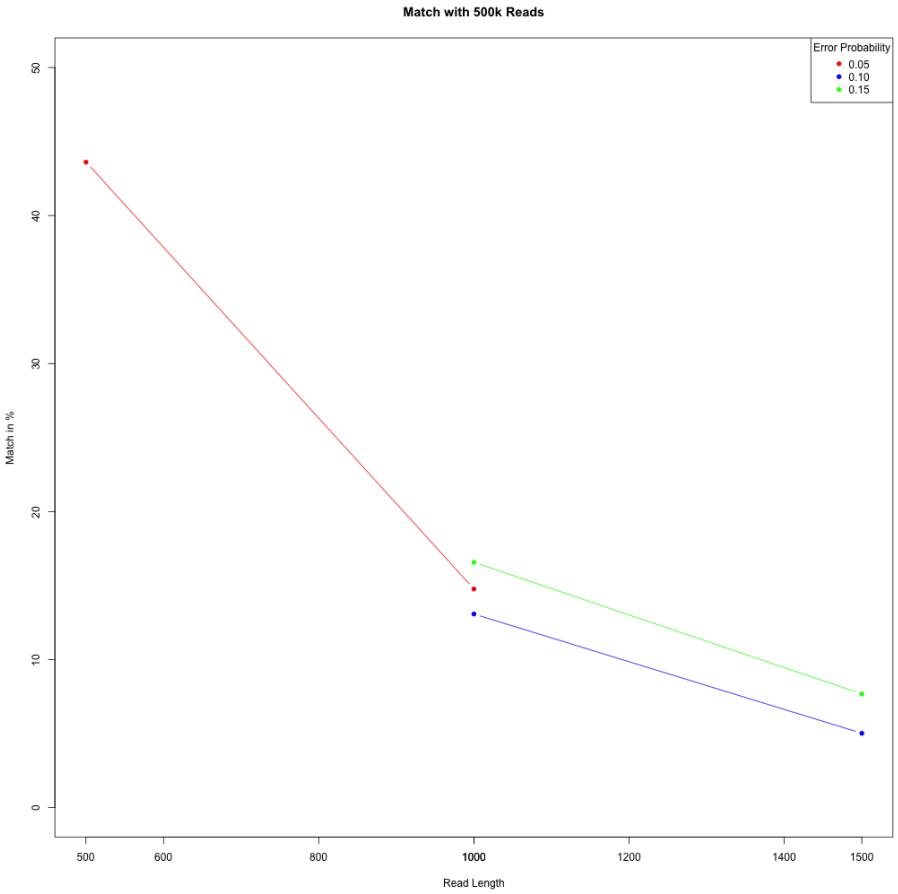

Figure 2: Results of haplotype assembly of 500.000 reads

Annual International Conference, RECOMB 2014, Pittsburgh, PA, USA, April 2-5, 2014, Proceedings, pages 237–249, 2014.

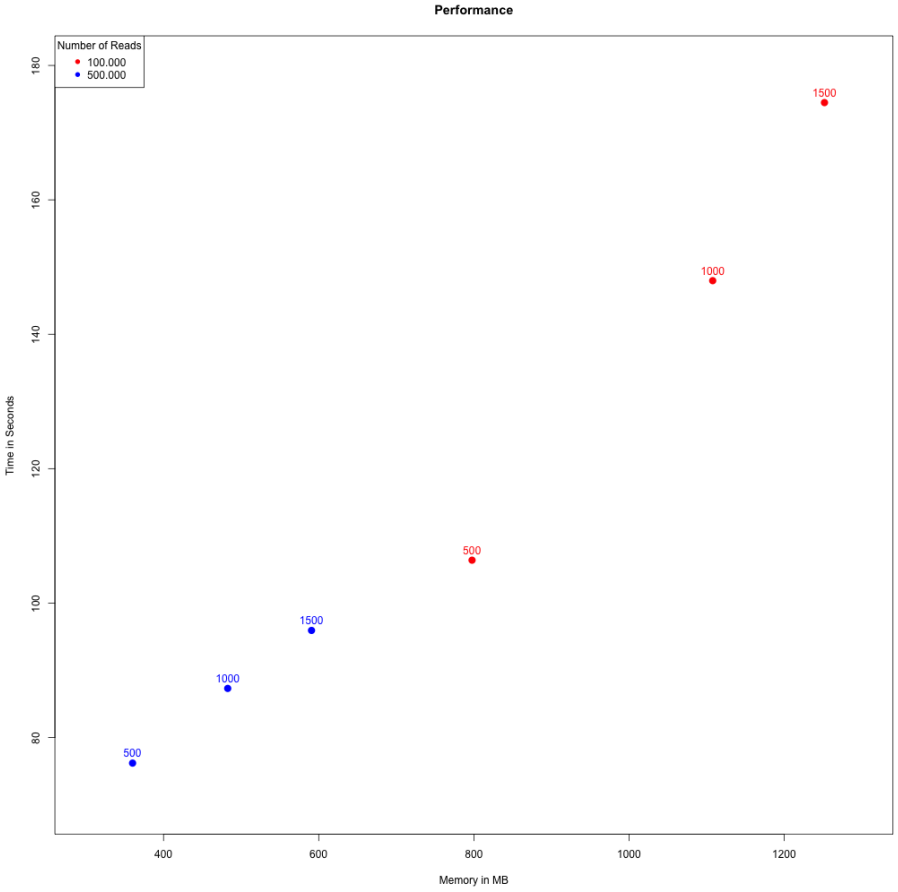

Figure 3: Performance