HAPCOL: Accurate and Memory-efficient Haplotype Assembly from Long Reads

Paola Bonizzoni[1], Riccardo Dondi[2], Gunnar W. Klau[3,5], Yuri Pirola[1,*], Nadia Pisanti[4,5] and Simone Zaccaria[1,*]

[1] DISCo, Univ. degli Studi di Milano-Bicocca, Milan, Italy
[2] Dipartimento di Scienze Umane e Sociali, Univ. degli Studi di Bergamo, Bergamo, Italy
[3] Life Sciences, Centrum Wiskunde & Informatica (CWI), Amsterdam, The Netherlands
[4] Dipartimento di Informatica, Univ. degli Studi di Pisa, Italy
[5] Erable Team, INRIA, France

Contact: {pirola,simone.zaccaria}@disco.unimib.it

Motivation

Haplotype assembly is the computational problem of reconstructing haplotypes in diploid organisms and is of fundamental importance for characterizing the effects of Single Nucleotide Polymorphisms (SNPs) on the expression of phenotypic traits. Minimum Error Correction (MEC) is one of the prominent combinatorial approaches for haplotype assembly. It aims at correcting the input data with the minimum number of corrections to the SNP values, such that the resulting reads can be unambiguously partitioned into two sets, each one identifying a haplotype.

Haplotype assembly highly benefits from the advent of "future generation" sequencing technologies and their capability to produce long reads at increasing coverage. Existing methods are not able to deal with such data in a fully satisfactory way, either because accuracy or performances degrade as read length and sequencing coverage increase, or because they are based on restrictive assumptions. In particular, three current state-of-the-art approaches are:

REFHAP (Duitama *et al.*, 2012): a heuristic approach, offering a good accuracy with good performance under the *all-heterozygous* assumption.

PROBHAP (Kuleshov, 2014): a probabilistic dynamic programming algorithm, that is slower than REFHAP, but improving its accuracy.

WHATSHAP (Patterson *et al.*, 2014): an exact algorithm, properly designed for long reads offering good accuracy, but with coverage only up to 20x.

Methods

In this paper, we exploit a novel characteristic of future-generation technologies, namely the uniform distribution of sequencing errors, for introducing an exact fixed-parameter tractable algorithm for a new constrained variant, called ***k*-cMEC**, of the MEC problem where the parameters are (i) the maximum number k of corrections that are allowed on each SNP position and (ii) the coverage. The designed algorithm, called **HAPCOL**, is able to work with or without the all-heterozygous assumption and can solve the weighted variant of the problem, exploiting the confidence degrees assigned to the SNP values (e.g. *phred scores*) in order to improve the accuracy of the reconstructed haplotypes.

We have experimentally compared accuracy, in terms of switch error rate (Browning and Browning, 2011) and number of phased positions, and performance, in terms of running time and peak memory usage, of HAPCOL on real and realistically simulated datasets of **long reads** with three state-of-the-art approaches for haplotype assembly – REFHAP, PROBHAP, and WHATSHAP:

- **NA12878 dataset**: on a real standard benchmark of long reads produced using a fosmid-based technology from the HapMap sample NA12878 by Duitama *et al.*, 2012, we executed each tool under the all-heterozygous assumption, since this dataset has low coverage (~3x on average) and since the covered positions are heterozygous with high confidence.
- **Simulated datasets**: we also assessed accuracy and performance of HAPCOL on a large collection of realistically-simulated datasets reflecting the characteristics of "future-generation" sequencing technologies that are currently (or soon) available (coverage up to 25x, read length from 10 000 to 50 000 bases, error rate up to 5%, and indel rate equal to 10%) (Carneiro *et al.*, 2012). At higher coverage, interesting applications such as SNP calling or heterozygous SNPs validation become feasible and reliable. Since these applications require that haplotypes are reconstructed without the all-heterozygous assumption, on the simulated datasets we only considered the tools that do not rely on this assumption – WHATSHAP and HAPCOL.

References

1. Duitama, J. *et al.* (2012). Fosmid-based whole genome haplotyping of a HapMap trio child: evaluation of single individual haplotyping techniques. Nucleic Acids Res, 40, 2041–2053.
2. Kuleshov, V. *et al.* (2014). Whole-genome haplotyping using long reads and statistical methods. Nat. Biotechnol, 32(3), 261, 266.
3. Patterson, M. *et al.* (2014). WhatsHap: Haplotype assembly for future-generation sequencing reads. In RECOMB, volume 8394 of LNCS, pages 237–249.
4. Browning, S. and Browning, B. (2011). Haplotype phasing: existing methods and new developments. Nat Rev Genet, 12(10), 703–714.
5. Carneiro *et al.* (2012). Pacific Biosciences sequencing technology for genotyping and variation discovery in human data. BMC Genomics, 13(1), 375.

Acknowledgements

The authors acknowledge the support of the MIUR PRIN 2010-2011 grant 2010LYA9RH (Automi e Linguaggi Formali: Aspetti Matematici e Applicativi), of the Cariplo Foundation grant 2013-0955 (Modulation of anti cancer immune response by regulatory non-coding RNAs) of the FA 2013 grant (Metodi algoritmici e modelli: aspetti teorici e applicazioni in bioinformatica).

Results

A prototypical implementation of HAPCOL is available under the terms of the GPL at: http://hapcol.algolab.eu/.

Since coverage varies across columns, HAPCOL adaptively adopts a different k for each column depending on the estimated error rate (ε) and significance level (α), given in input by the user.

Table 1 reports, for each tool, the switch error rate and the percentage of phased positions over all the phasable positions, the total running time, and the peak of memory for the whole dataset with one combination of ε and α (5% and 10^{-3}).

Table 1.

Tool	error [%]	phased [%]	time [*sec*]	mem. [GB]
HAPCOL	**1.91**	**99.88**	332	2.1
WHATSHAP	2.02	99.73	172	23.9
PROBHAP	3.36	98.02	1205	0.6
REFHAP	3.68	97.75	**43**	**0.5**

HAPCOL reconstructed the most accurate haplotypes and phased the largest number of positions compared with the other tools. To the contrary, REFHAP was the fastest and most memory efficient tool among the four considered. Overall, all the tools can be run with modest/medium computing resources. However, PROBHAP was significantly slower than the others (~20 minutes) and WHATSHAP required significantly more memory than REFHAP (44 times).

Table 2 reports, for one combination of the input parameters ε and α (5% and 10^{-3}), the average error of the reconstructed haplotypes, the average running time, and the average memory usage over all the simulated instances of a given coverage (15x and 20x), and error rate e (1% and 5%).

Table 2.

		error [%]		time [*sec*]		mem. [GB]	
cov	*e*	WH	HC	WH	HC	WH	HC
15x	1	2.35	2.36	**48**	64	4.6	**0.8**
	5	2.35	2.35	**49**	56	4.7	**0.7**
20x	1	1.95	1.94	1306	**586**	138.0	**5.6**
	5	2.07	2.08	1347	**526**	138.5	**5.1**

In terms of accuracy, on all the instances HAPCOL (HC) obtained the same phasing error rate of WHATSHAP (WH). However, in terms of performances HAPCOL is both faster and significantly more memory-efficient than WHATSHAP. In particular, on average, HAPCOL is at least twice faster than WHATSHAP when the coverage is 20x even for large values of k. Concerning memory usage, we observe the same general trend, but even more evident. In fact, the average memory usage of WHATSHAP at coverage 20x is ~138GB, while HAPCOL requires only ~5GB.

Table 3 reports the same results of Table 2 for simulated instances with coverage 25x, read length of 50 000 bases and one combination of ε and α (5% and 10^{-3}).

Table 3.

e	*cov*	error [%]	time [*sec*]	mem. [GB]
1	20x	1.66	832	9.5
	25x	1.52	4272	40.7
5	20x	1.71	737	8.5
	25x	1.55	4357	39.2

In this case, WHATSHAP was not able to successfully conclude the execution on these instances since it exhausted the available memory (256GB). Hence, we evaluated how accuracy and performances of HAPCOL vary between instances with coverage 20x and 25x. In particular, we observe that increasing coverage allows to improve accuracy (~9%) of the reconstructed haplotypes (as we already observed for coverage 15x and 20x).

Conclusions

On a real benchmark dataset of long reads, we showed that HAPCOL is competitive with state-of-the-art methods, improving the accuracy and the number of phased positions. Moreover, HAPCOL is able to overcome the traditional all heterozygous assumption and to process datasets of long reads with coverage 25x on standard workstations/small servers, while the current state-of-the art methods either rely on this assumption or become unfeasible on coverages over 20x. Thanks to these results, HAPCOL is potentially able to directly perform SNP calling or heterozygous SNPs validation that become feasible and reliable on coverage around 25x and above.