Supplementary Material for WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads*

Murray Patterson[†,1], Tobias Marschall[†,2,3],
Nadia Pisanti[4], Leo van Iersel[5],
Leen Stougie[5,6], Gunnar W. Klau[‡,5,6],
and Alexander Schönhuth [‡,5]

January 6, 2015

Here in Table 1 we include runtime tables of our method against three other methods (He et al., 2010; Chen et al., 2013; Deng et al., 2013) for chromosome 1 and 15 of Venter's genome in both the general and all-heterozygous case and for error profiles of 1% and 5% in the case of the artificial datasets. In Figure 1 give the same performance analsysis as Figure 2 of the main paper, but with an error profile of 5%.

*This work was done while all authors were affiliated with or visiting the Life Sciences Group at Centrum Wiskunde & Informatica (CWI).

[†]Joint first authorship.

[‡]Joint last authorship.

[1]Laboratoire de Biométrie et Biologie Évolutive (LBBE : UMR CNRS 5558), Université de Lyon 1, Villeurbanne, France
[2]Saarland University, Saarbrücken, Germany
[3]Max Planck Institute for Informatics, Saarbrücken, Germany
[4]Department of Computer Science, University of Pisa, Italy
[5]Life Sciences, Centrum Wiskunde & Informatica (CWI), Amsterdam, The Netherlands
[6]VU University Amsterdam, The Netherlands
murray.patterson@univ-lyon1.fr, t.marschall@mpi-inf.mpg.de,
{a.schoenhuth,gunnar.klau}@cwi.nl

Data set	Chen et al.	He et al.	Deng et al.	WhatsHap
chr1, **all-het**, **Cov. 5**				
2 x 100 (HiSeq)	445.8s	965.2s	0.3s	1.8s
2 x 150 (MiSeq)	679.9s	-	0.4s	2.5s
1 x 1000 (1%)	716.9s	-	0.5s	2.8s
1 x 1000 (5%)	739.6s	-	0.4s	2.7s
1 x 5000 (1%)	771.2s	-	0.6s	3.8s
1 x 5000 (5%)	778.0s	-	0.6s	3.7s
1 x 10000 (1%)	313.9s	-	0.5s	3.7s
1 x 10000 (5%)	324.0s	-	0.5s	3.7s
1 x 50000 (1%)	56.7s	-	0.4s	3.3s
1 x 50000 (5%)	60.6s	-	0.4s	3.3s
chr1, **all-het**, **Cov. 10**				
2 x 100 (HiSeq)	452.8s	-	3.2s	5.5s
2 x 150 (MiSeq)	646.2s	-	5.3s	8.1s
1 x 1000 (1%)	706.5s	-	9.6s	11.0s
1 x 1000 (5%)	708.6s	-	10.0s	11.0s
1 x 5000 (1%)	679.9s	-	10.3s	15.4s
1 x 5000 (5%)	744.9s	-	10.7s	15.4s
1 x 10000 (1%)	288.8s	-	9.7s	15.6s
1 x 10000 (5%)	290.7s	-	9.1s	15.5s
1 x 50000 (1%)	80.6s	-	7.1s	13.6s
1 x 50000 (5%)	94.8s	-	7.1s	13.7s
chr1, **all-het**, **Cov. 15**				
2 x 100 (HiSeq)	479.5s	-	377.8s	62.6s
2 x 150 (MiSeq)	629.1s	-	708.5s	101.7s
1 x 1000 (1%)	720.5s	-	3701.5s	192.6s
1 x 1000 (5%)	732.6s	-	4197.6s	192.0s
1 x 5000 (1%)	709.9s	-	2623.5s	271.9s
1 x 5000 (5%)	696.9s	-	2377.8s	271.5s
1 x 10000 (1%)	296.0s	-	1443.6s	276.9s
1 x 10000 (5%)	303.7s	-	1294.4s	274.9s
1 x 50000 (1%)	108.1s	-	440.5s	230.8s
1 x 50000 (5%)	139.9s	-	401.6s	233.8s

Data set	Chen et al.	He et al.	Deng et al.	WhatsHap
chr1, **general**, **Cov. 5**				
2 x 100 (HiSeq)	445.3s	-	0.3s	1.8s
2 x 150 (MiSeq)	544.0s	-	0.4s	2.5s
1 x 1000 (1%)	566.4s	-	0.5s	2.8s
1 x 1000 (5%)	581.9s	-	0.4s	2.8s
1 x 5000 (1%)	607.1s	-	0.6s	3.7s
1 x 5000 (5%)	643.1s	-	0.6s	3.8s
1 x 10000 (1%)	262.4s	-	0.5s	3.7s
1 x 10000 (5%)	271.3s	-	0.6s	3.7s
1 x 50000 (1%)	62.8s	-	0.4s	3.3s
1 x 50000 (5%)	68.2s	-	0.4s	3.3s
chr1, **general**, **Cov. 10**				
2 x 100 (HiSeq)	362.9s	-	3.1s	5.6s
2 x 150 (MiSeq)	408.8s	-	5.2s	8.2s
1 x 1000 (1%)	388.0s	-	9.6s	11.1s
1 x 1000 (5%)	389.3s	-	9.6s	11.1s
1 x 5000 (1%)	375.9s	-	10.9s	15.5s
1 x 5000 (5%)	386.3s	-	11.0s	15.4s
1 x 10000 (1%)	206.1s	-	9.8s	15.7s
1 x 10000 (5%)	216.6s	-	9.2s	15.6s
1 x 50000 (1%)	101.1s	-	7.2s	13.7s
1 x 50000 (5%)	121.0s	-	7.3s	13.8s
chr1, **general**, **Cov. 15**				
2 x 100 (HiSeq)	358.5s	-	380.4s	63.4s
2 x 150 (MiSeq)	427.2s	-	698.9s	103.0s
1 x 1000 (1%)	439.9s	-	4248.1s	194.8s
1 x 1000 (5%)	458.2s	-	4169.5s	195.9s
1 x 5000 (1%)	417.3s	-	2583.2s	274.8s
1 x 5000 (5%)	458.0s	-	2628.3s	274.8s
1 x 10000 (1%)	242.7s	-	1486.5s	279.6s
1 x 10000 (5%)	261.5s	-	1500.2s	278.9s
1 x 50000 (1%)	154.1s	-	443.2s	233.9s
1 x 50000 (5%)	210.4s	-	431.9s	234.5s

Data set	Chen et al.	He et al.	Deng et al.	WhatsHap
chr15, all-het, Cov. 5				
2 x 100 (HiSeq)	133.9s	2.2s	0.1s	0.7s
2 x 150 (MiSeq)	172.4s	65.4s	0.1s	0.9s
1 x 1000 (1%)	175.8s	54.5s	0.2s	1.0s
1 x 1000 (5%)	179.1s	53.0s	0.1s	1.0s
1 x 5000 (1%)	140.2s	-	0.2s	1.4s
1 x 5000 (5%)	168.4s	-	0.2s	1.4s
1 x 10000 (1%)	84.6s	-	0.2s	1.4s
1 x 10000 (5%)	88.6s	-	0.2s	1.4s
1 x 50000 (1%)	30.0s	-	0.1s	1.2s
1 x 50000 (5%)	31.3s	-	0.1s	1.2s
chr15, all-het, Cov. 10				
2 x 100 (HiSeq)	135.5s	24.6s	1.1s	2.1s
2 x 150 (MiSeq)	177.2s	47.3s	2.0s	3.1s
1 x 1000 (1%)	181.6s	46.0s	3.8s	4.3s
1 x 1000 (5%)	185.2s	66.1s	4.0s	4.3s
1 x 5000 (1%)	163.1s	-	4.1s	5.8s
1 x 5000 (5%)	168.8s	-	4.1s	5.8s
1 x 10000 (1%)	85.0s	-	3.4s	5.8s
1 x 10000 (5%)	88.8s	-	3.6s	5.8s
1 x 50000 (1%)	37.7s	-	2.7s	5.1s
1 x 50000 (5%)	40.8s	-	2.6s	5.1s
chr15, all-het, Cov. 15				
2 x 100 (HiSeq)	150.8s	34.7s	137.4s	24.4s
2 x 150 (MiSeq)	173.1s	81.0s	251.9s	39.9s
1 x 1000 (1%)	183.6s	85.7s	1534.8s	75.5s
1 x 1000 (5%)	189.8s	88.5s	1635.3s	75.4s
1 x 5000 (1%)	163.8s	-	962.8s	103.6s
1 x 5000 (5%)	174.7s	-	871.2s	103.5s
1 x 10000 (1%)	89.4s	-	503.3s	105.3s
1 x 10000 (5%)	94.4s	-	505.4s	104.4s
1 x 50000 (1%)	58.5s	-	183.8s	86.8s
1 x 50000 (5%)	56.3s	-	163.1s	87.7s

Data set	Chen et al.	He et al.	Deng et al.	WhatsHap
chr15, general, Cov. 5				
2 x 100 (HiSeq)	134.0s	-	0.1s	0.7s
2 x 150 (MiSeq)	152.0s	-	0.1s	0.9s
1 x 1000 (1%)	157.1s	-	0.1s	1.0s
1 x 1000 (5%)	152.9s	-	0.2s	1.1s
1 x 5000 (1%)	147.8s	-	0.2s	1.4s
1 x 5000 (5%)	141.2s	-	0.2s	1.4s
1 x 10000 (1%)	78.1s	-	0.2s	1.4s
1 x 10000 (5%)	80.5s	-	0.2s	1.4s
1 x 50000 (1%)	26.3s	-	0.1s	1.2s
1 x 50000 (5%)	32.6s	-	0.1s	1.2s
chr15, general, Cov. 10				
2 x 100 (HiSeq)	129.0s	-	1.2s	2.1s
2 x 150 (MiSeq)	143.5s	-	1.9s	3.2s
1 x 1000 (1%)	127.6s	-	4.0s	4.3s
1 x 1000 (5%)	129.9s	-	4.0s	4.3s
1 x 5000 (1%)	116.0s	-	4.1s	5.8s
1 x 5000 (5%)	115.4s	-	3.9s	5.8s
1 x 10000 (1%)	77.1s	-	3.6s	5.8s
1 x 10000 (5%)	81.2s	-	3.4s	5.8s
1 x 50000 (1%)	48.9s	-	2.7s	5.1s
1 x 50000 (5%)	47.7s	-	2.8s	5.1s
chr15, general, Cov. 15				
2 x 100 (HiSeq)	131.0s	-	126.0s	24.7s
2 x 150 (MiSeq)	139.5s	-	255.3s	40.4s
1 x 1000 (1%)	136.7s	-	1516.5s	76.4s
1 x 1000 (5%)	139.2s	-	1522.9s	76.3s
1 x 5000 (1%)	128.0s	-	965.7s	104.8s
1 x 5000 (5%)	123.6s	-	925.0s	104.7s
1 x 10000 (1%)	86.9s	-	495.6s	106.6s
1 x 10000 (5%)	94.3s	-	496.9s	105.6s
1 x 50000 (1%)	80.6s	-	163.7s	87.9s
1 x 50000 (5%)	88.2s	-	139.1s	88.8s

Table 1: Runtimes in CPU seconds for haplotype assembly approaches in the unweighted (both allhet and general) case on chromosomes 1 and 15 ofr J. Craig Venter's genome. A '-' stands for an unsuccessful run, either due to an exceeded time limit of 5 CPU h for an out of memory exception.

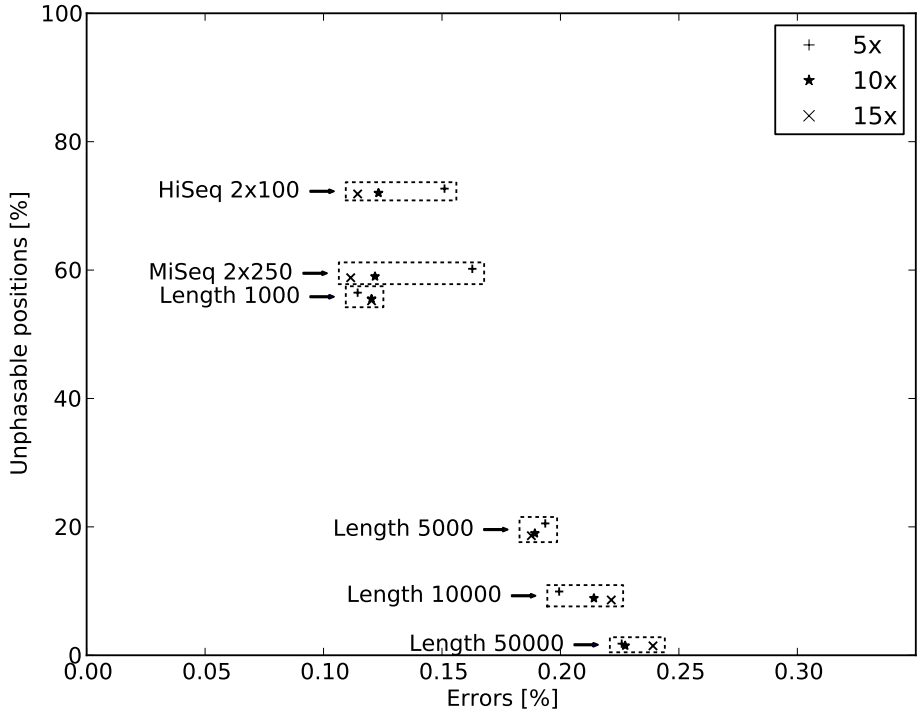

Figure 1: Performance of phasing human chromosome 1 with 68 184 heterozygous SNPs in total using different simulated data sets and different coverages. The *unphasable positions* percentage (y-axis) gives the fraction of the SNP positions that could not be phased due to not being covered by reads that span more than one SNP position. The x-axis shows the percentage of all SNPs that were not unphasable but wrongly phased by the algorithm, either because of a flip error, a switch error, or due to being reported as ambiguous position by WhatsHap. Length 1 000, 5 000, 10 000, and 50 000 refer to reads of this length from a hypothetical sequencer with an error rate of 5%. HiSeq/MiSeq refers to using error profiles specific to these instruments during read sampling. Data sets are pruned to three different target coverages (5x, 10x, 15x) encoded by different symbols in the plot (see legend).

References

Z.-Z. Chen, F. Deng, and L. Wang. Exact algorithms for haplotype assembly from whole-genome sequence data. *Bioinformatics*, 29(16):1938–45, 2013.

F. Deng, W. Cui, and L.-S. Wang. A highly accurate heuristic algorithm for the haplotype assembly problem. *BMC Genomics*, 14(Suppl 2):S2, 2013.

D. He, A. Choi, K. Pipatsrisawat, A. Darwiche, and E. Eskin. Optimal algorithms for haplotype assembly from whole-genome sequence data. *Bioinformatics*, 26(12):i183–i190, 2010.