



Georg-August-Universität
Göttingen



Haplotype reconstruction for dense marker panels using PHASE

Eduardo Pimentel



Georg-August-Universität
Göttingen



Outline:

- overview of PHASE;
- application to real data;
- application to simulated data

Overview of PHASE

- MCMC approach for haplotype reconstruction;
[Stephens M, Smith NJ and Donnelly P (2001) Am J Hum Genet 68:978-989]
- regarded as an accurate algorithm;
[Marchini et al. (2006) Am J Hum Genet 78:437-450]
- can handle data on unrelated individuals.

Overview of PHASE

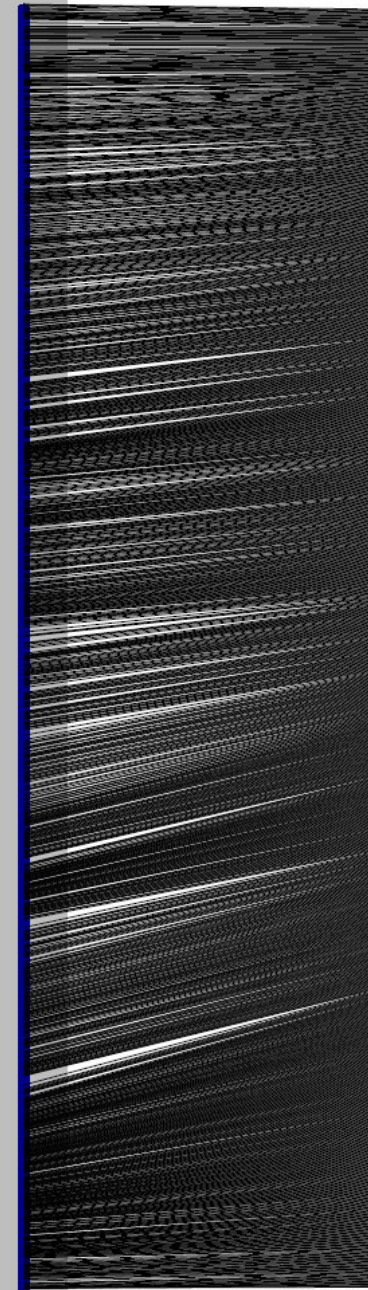
Using the program:

```
gwdul01> PHASE ✓ ✓ ? ? ?  
usage is PHASE <filename.inp> <filename.out><number of iterations> <thinning interval> <burn-in>  
gwdul01> █
```

Application to real data

- 95 animals;
- 1338 markers on BTA14;

81



0



81.323942

71.515858

57.99353

46.080034

30.907284

18.376289

7.834896

0.001182

Application to real data

- 1338 markers on BTA14;



13 segments with 100 markers + 1 with 38 markers

Application to real data

- For each segment, 5 runs with PHASE:

<u># of iterations</u>	<u>thinning interval</u>	<u>burn in</u>
100	2	10
200	5	20
300	6	30
400	8	40
500	10	50

Application to real data

- The output from PHASE contains something like:

```

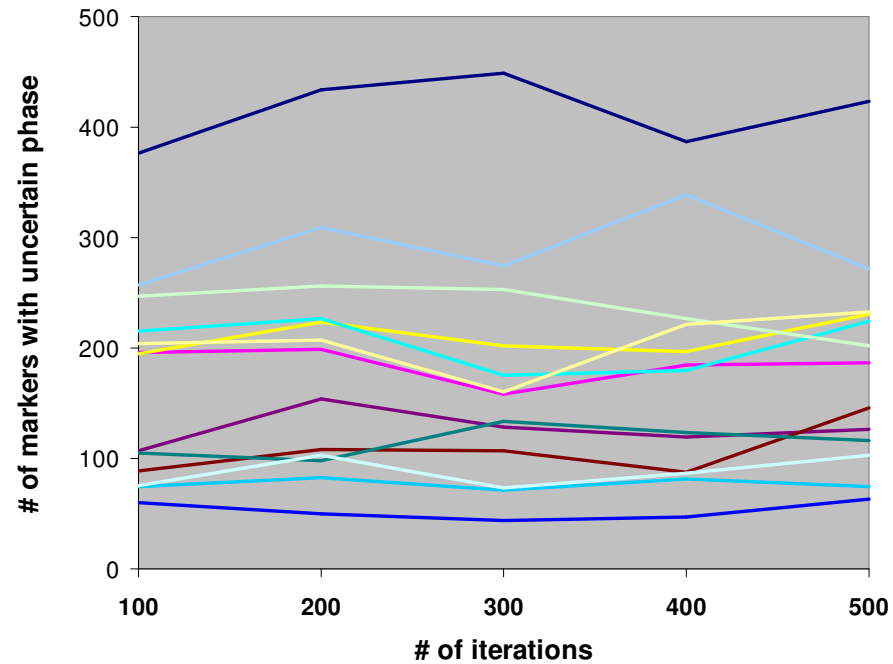
.
.
.
Haplotype estimates for each individual, with uncertain phases enclosed in ()
and uncertain genotypes enclosed in []:

BEGIN BESTPAIRS1
0 an_67
A (G) (A) (G) G A G A A C A G G C G G A G G G A A G C G G G C G G A (A) A G G A A A G
A (A) (G) (A) G A G A G C A A G C A A A A A A G G A C G G A C G A G (G) A A A A A A G
0 an_68
A A A G (G) (A) G (G) (G) (A) C A A A A A A A A A A G A A G A A A G A A A G [G] G A G
A A A G (A) (C) G (A) (A) (C) C A A A A A A A A G G G G A A G G A A G A G A A [G] A A A
0 an_69
C G A G G (A) G A A C C A A A A A A G A A A G A A G A A [A] A A A A G G A A G G G G G
C G G G G (C) G A G A A A A A A A A A G G A A G C G G A [C] G G A A A G G A A A G A A
0 an_72
A G A G G A G A A C A G G C G G [A] G G G A A G C G G G C G G A A A G G A A A G A A A
A G A G G C G A A C A G G C G G [A] G G G A A G C G G G C G G A A A G G A A A G A A A
.
.
.

```

Application to real data

- number of markers with uncertain phase:



Application to real data

- some results:

Proportion of disagreements (%)

	200_4_20	300_6_30	400_8_40	500_10_50
100_2_10	0.78	0.88	0.77	0.58
200_4_20		0.81	0.62	0.78
300_6_30			0.7	0.83
400_8_40				0.68

	200_4_20	300_6_30	400_8_40	500_10_50
100_2_10	1.22	1.32	1.22	1.32
200_4_20		1.35	0.77	1.1
300_6_30			1.4	1.29
400_8_40				1.21

	200_4_20	300_6_30	400_8_40	500_10_50
100_2_10	1.27	1	1.09	1.21
200_4_20		1.75	1.42	1.46
300_6_30			1.32	1.19
400_8_40				1.27

	200_4_20	300_6_30	400_8_40	500_10_50
100_2_10	0.59	0.55	0.75	0.53
200_4_20		0.74	0.75	0.72
300_6_30			0.72	0.5
400_8_40				0.6

Application to real data

- looking at the averages:

Proportion of disagreements (%)

averages

<u>Segment 1:</u>	1.79
<u>Segment 2:</u>	1.06
<u>Segment 3:</u>	0.93
<u>Segment 4:</u>	0.63
<u>Segment 5:</u>	0.99
<u>Segment 6:</u>	1.34
<u>Segment 7:</u>	0.65
<u>Segment 8:</u>	1.30
<u>Segment 9:</u>	0.74
<u>Segment 10:</u>	0.94
<u>Segment 11:</u>	1.22
<u>Segment 12:</u>	1.23
<u>Segment 13:</u>	1.18
<u>overall:</u>	1.08

Application to real data

For most of the animals: < 1% missing genotypes...

... but for 3 of them: there were 55%, 43% and 34%

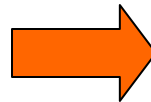
Application to real data

- excluding those 3 animals:

Proportion of disagreements (%)

averages

<u>Segment 1:</u>	1.79
<u>Segment 2:</u>	1.06
<u>Segment 3:</u>	0.93
<u>Segment 4:</u>	0.63
<u>Segment 5:</u>	0.99
<u>Segment 6:</u>	1.34
<u>Segment 7:</u>	0.65
<u>Segment 8:</u>	1.30
<u>Segment 9:</u>	0.74
<u>Segment 10:</u>	0.94
<u>Segment 11:</u>	1.22
<u>Segment 12:</u>	1.23
<u>Segment 13:</u>	1.18
overall:	1.08

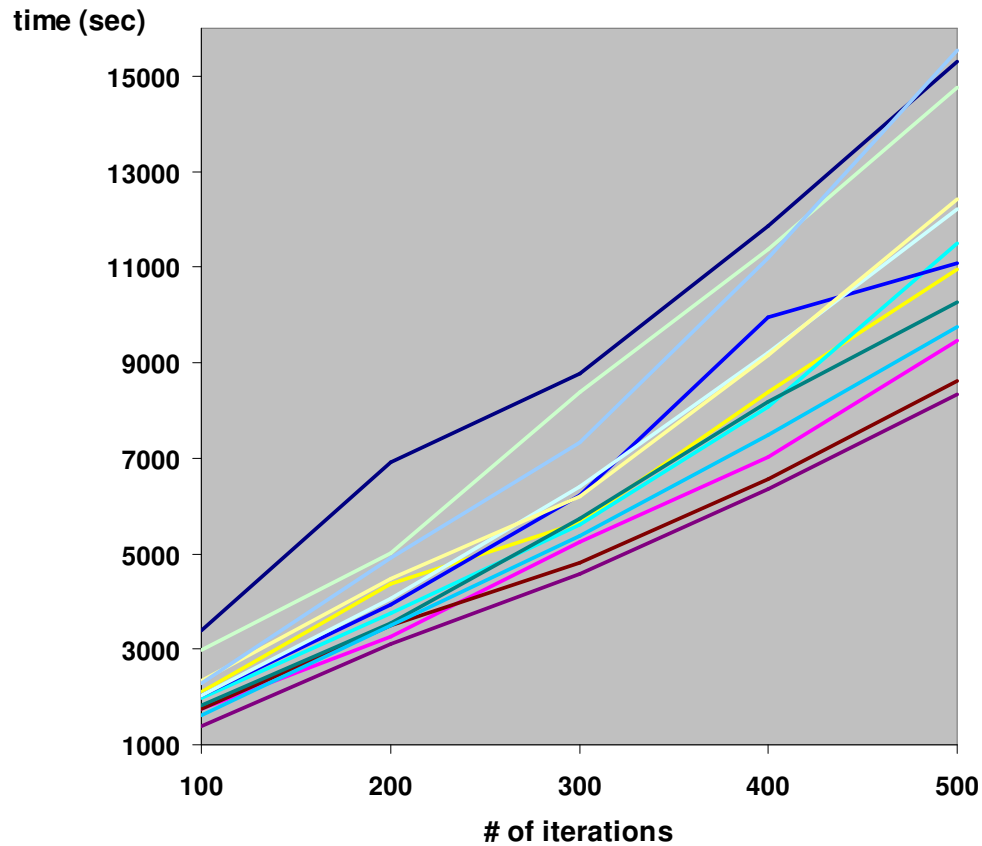


averages

<u>Segment 1:</u>	1.10
<u>Segment 2:</u>	0.43
<u>Segment 3:</u>	0.39
<u>Segment 4:</u>	0.11
<u>Segment 5:</u>	0.29
<u>Segment 6:</u>	0.77
<u>Segment 7:</u>	0.07
<u>Segment 8:</u>	0.64
<u>Segment 9:</u>	0.09
<u>Segment 10:</u>	0.32
<u>Segment 11:</u>	0.63
<u>Segment 12:</u>	0.60
<u>Segment 13:</u>	0.54
overall:	0.46

Application to real data

- processing time:



100 iter.
2110

on average
vs.

500 iter.
11556

Application to simulated data

- Data from the XII QTL-MAS Workshop, Uppsala
- subset of 500 animals and 100 markers;
- un-phased at random (20 replicates);
- phased inferred with: $\left\{ \begin{array}{l} 100 / 2 / 10 \\ 500 / 10 / 50 \end{array} \right.$

Application to simulated data

Proportion of disagreements (%)

<u>rep</u>	<u>T vs. 100</u>	<u>T vs. 500</u>
1	0.160	0.158
2	0.154	0.200
3	0.182	0.152
4	0.156	0.156
5	0.190	0.182
6	0.192	0.156
7	0.184	0.184
8	0.160	0.158
9	0.198	0.156
10	0.158	0.156
11	0.178	0.158
12	0.184	0.156
13	0.162	0.158
14	0.192	0.156
15	0.154	0.158
16	0.178	0.156
17	0.154	0.176
18	0.158	0.158
19	0.250	0.180
20	0.190	0.152
average	0.177	0.163
Reduction (%):		7.58

Application to simulated data

Processing time (sec)

<u>rep</u>	<u>100 iter.</u>	<u>500 iter.</u>
1	3550.71	14857.80
2	3526.66	14235.59
3	3730.88	14096.88
4	3364.49	13881.01
5	3062.30	15515.95
6	3763.87	14411.01
7	3253.37	14834.63
8	3165.81	14173.17
9	3389.84	15035.39
10	3368.66	14788.23
11	3934.69	15877.54
12	3503.84	14727.12
13	3344.25	14232.49
14	3379.08	15987.73
15	3322.71	14227.24
16	3492.98	14651.98
17	3897.51	15209.04
18	3435.91	13994.29
19	3328.50	15707.99
20	3442.36	14774.57
average	3462.92	14760.98
Increase (%):		326.26

Summarizing:

- Setting larger numbers of iterations did not provide too much different results in both simulated and real data;
- Therefore, the increase in processing time, with larger chains, did not seem to be compensated.

Acknowledgment

This study is part of the project FUGATO-plus GenoTrack and was financially supported by the German Ministry of Education and Research, BMBF, the Förderverein Biotechnologieforschung e.V. (FBF), Bonn, and Lohmann Tierzucht GmbH, Cuxhaven.



Bundesministerium
für Bildung
und Forschung



LOHMANN
TIERZUCHT