

Results for localisation accuracy and for association tests in phased and unphased data

Thomas Mailund

July 27, 2007

I present a few experiments with the localisation accuracy and with the significance test for association, comparing HapCluster² results with the MCMC on phased data to that of the MCMC on unphased data. I also present an analysis of the CYP2D6 dataset from Hosking et al.¹.

Simulation results

For the simulation setup, I simulated 10 data sets for each set of parameters, varying the number of individuals from 500 cases and 500 controls, to 500 cases and 2000 controls and 2000 cases and 2000 controls, and varying the genetic relative risk (in an additive model) between 1.5 and 2.0. Each data set contains 100 SNP markers on a recombination length of $\rho = 40$ (or about 100 Kbp assuming an effective population size of $N_e = 10,000$ and 1 cM per 1 Mbp).

For the simulated data, I built data sets both with the true known phase and with phase removed. I then ran the two MCMCs on the corresponding data sets, using windows options—the maximum number of markers to either side of the locus to include in the distance calculation—of 5 and 10, calculating the posterior distribution of the causative locus and the Bayes factor of association.

Localisation accuracy

Figures 1–6 shows localisation accuracy in each of the 10 data sets for the 6 choices of simulation parameters. The bullet indicates the posterior mode while the line indicates the 95 % highest density region (HDR). Blue is used to indicate that the true point lies in the HDR while red is used to indicate that it lies outside the HDR.

Association tests

Tables 1–6 shows log Bayes factors for association in each of the 10 data sets for the 6 choices of simulation parameters. For each parameter set, the sets are ordered according to the disease locus (to make it easier to compare the localisation figures with the Bayes factors). In general, the evidence found for association is roughly the same for the phased and unphased MCMC, but there seems to be a trend towards finding more evidence with the phased than with the unphased approach. This is, however, not unexpected as the unphased algorithm deals with more uncertainty having to integrate over the unknown phase.

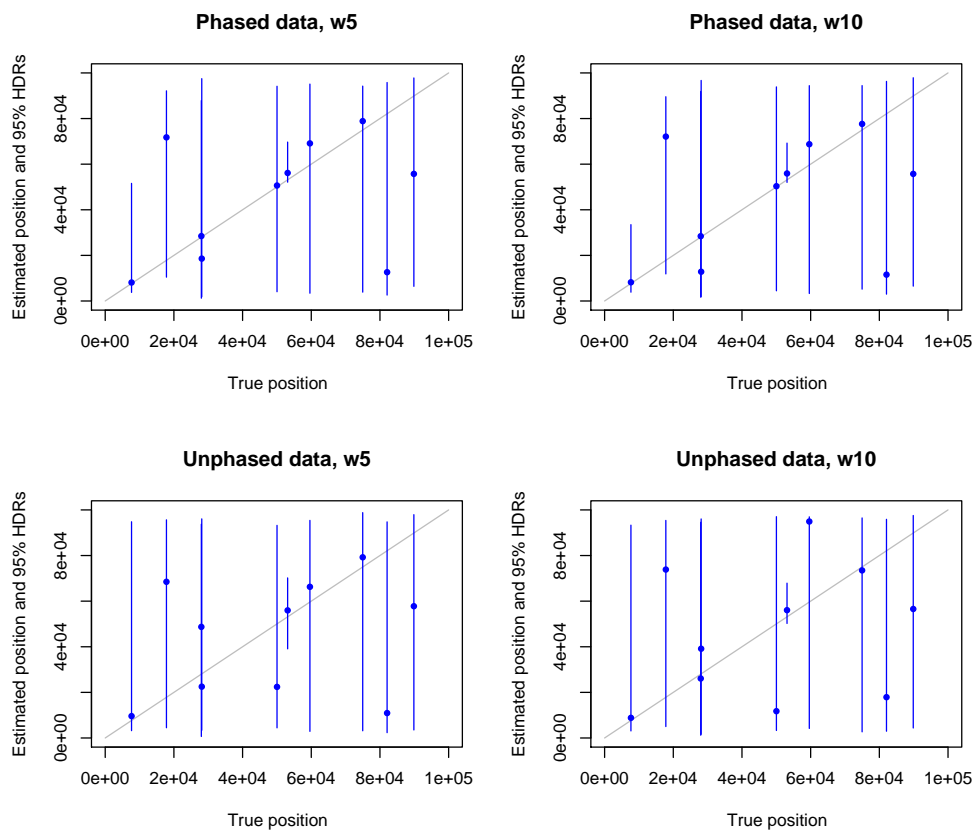


Fig. 1: Localisation accuracy in data with 500 cases and 500 controls with GRR 1.5.

Table 1: Log Bayes factors for association in data with 500 case and 500 controls with GRR 1.5.

| Data set | Locus | Known phase | | Unknown phase | |
|----------|-------|-------------|----------|---------------|----------|
| | | w5 | w10 | w5 | w10 |
| 1 | 7669 | 2.9610 | 3.2350 | -0.09552 | 0.1105 |
| 2 | 17830 | 0.7258 | 1.0450 | 0.1213 | 0.2632 |
| 3 | 27992 | 0.8205 | 0.8226 | 0.5156 | 0.8157 |
| 4 | 28116 | -0.4708 | -0.5289 | -0.3259 | -0.1201 |
| 5 | 50026 | 0.1570 | -0.08542 | 0.4921 | -0.1280 |
| 6 | 53120 | 3.6210 | 4.7050 | 2.0900 | 3.0460 |
| 7 | 59608 | 0.06047 | 0.1495 | -0.1676 | 0.5341 |
| 8 | 74979 | 0.2459 | 0.3125 | -0.2127 | -0.04848 |
| 9 | 82083 | -0.4168 | -0.4107 | -0.1867 | -0.5177 |
| 10 | 89876 | 0.4586 | 0.5563 | 0.1777 | 0.05627 |

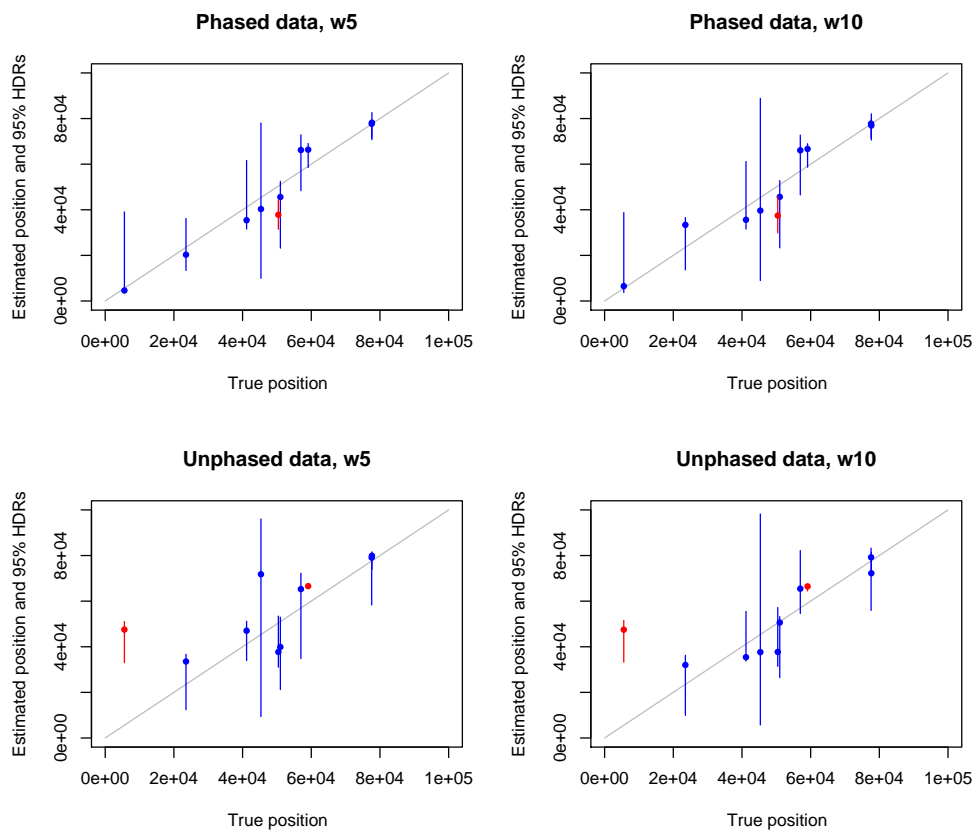


Fig. 2: Localisation accuracy in data with 500 cases and 500 controls with GRR 2.0.

Table 2: Log Bayes factors for association in data with 500 case and 500 controls with GRR 2.0.

| Data set | Locus | Known phase | | Unknown phase | |
|----------|-------|-------------|-------|---------------|--------|
| | | w5 | w10 | w5 | w10 |
| 1 | 5587 | 14.97 | 14.72 | 6.823 | 10.54 |
| 2 | 23519 | 8.961 | 9.186 | 4.963 | 3.895 |
| 3 | 41167 | 13.96 | 13.88 | 13.93 | 12.85 |
| 4 | 45348 | 2.162 | 1.254 | 0.8135 | 0.5319 |
| 5 | 50419 | 6.783 | 9.414 | 5.633 | 4.165 |
| 6 | 51001 | 20.09 | 20.08 | 18.45 | 19.28 |
| 7 | 56985 | 11.79 | 11.74 | 8.475 | 8.575 |
| 8 | 59103 | 21.14 | 21.40 | 19.17 | 18.09 |
| 9 | 77596 | 12.86 | 13.17 | 6.788 | 10.13 |
| 10 | 77668 | 20.09 | 20.41 | 22.09 | 20.62 |

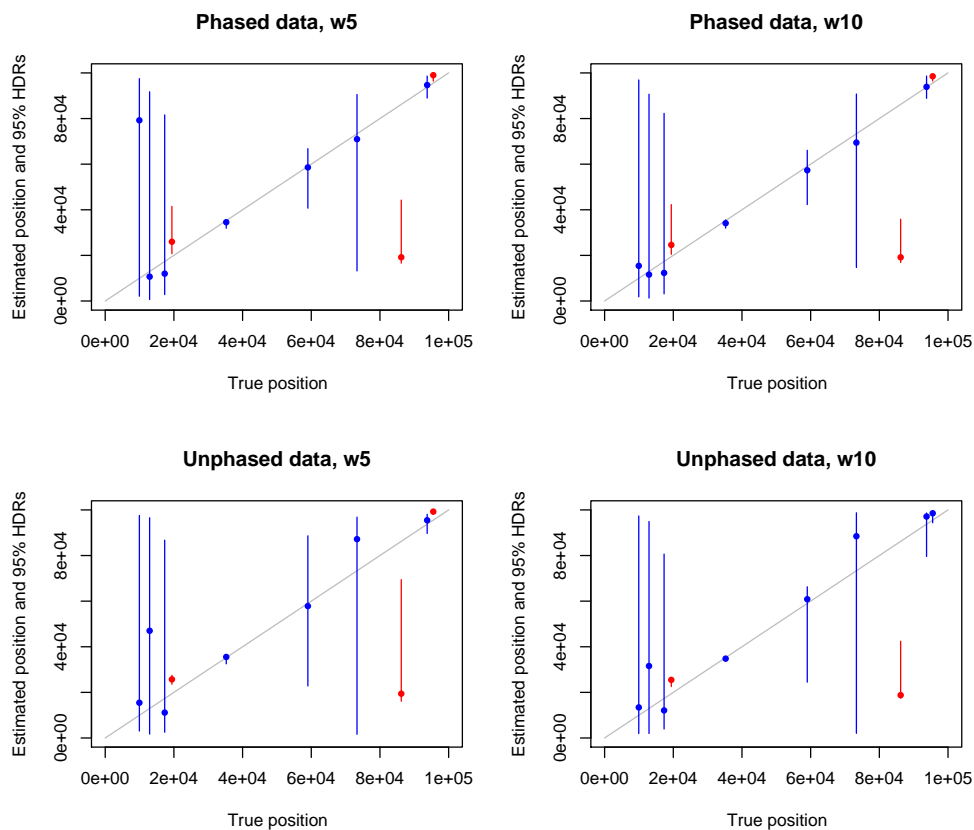


Fig. 3: Localisation accuracy in data with 500 cases and 2000 controls with GRR 1.5.

Table 3: Log Bayes factors for association in data with 500 case and 2000 controls with GRR 1.5.

| Data set | Locus | Known phase | | Unknown phase | |
|----------|-------|-------------|---------|---------------|---------|
| | | w5 | w10 | w5 | w10 |
| 1 | 9949 | -0.5816 | -0.4608 | -0.5109 | -0.3231 |
| 2 | 12933 | 1.044 | 1.077 | -0.3394 | 0.1632 |
| 3 | 17322 | 1.955 | 2.294 | 1.834 | 2.636 |
| 4 | 19423 | 9.441 | 6.852 | 6.869 | 4.051 |
| 5 | 35244 | 16.62 | 15.50 | 13.52 | 12.98 |
| 6 | 59011 | 2.744 | 3.282 | 1.294 | 1.908 |
| 7 | 73321 | 1.279 | 0.9044 | -0.4532 | -0.4835 |
| 8 | 86205 | 2.860 | 3.134 | 2.116 | 2.469 |
| 9 | 93750 | 8.329 | 5.876 | 4.151 | 3.217 |
| 10 | 95529 | 5.104 | 8.057 | 4.658 | 10.18 |

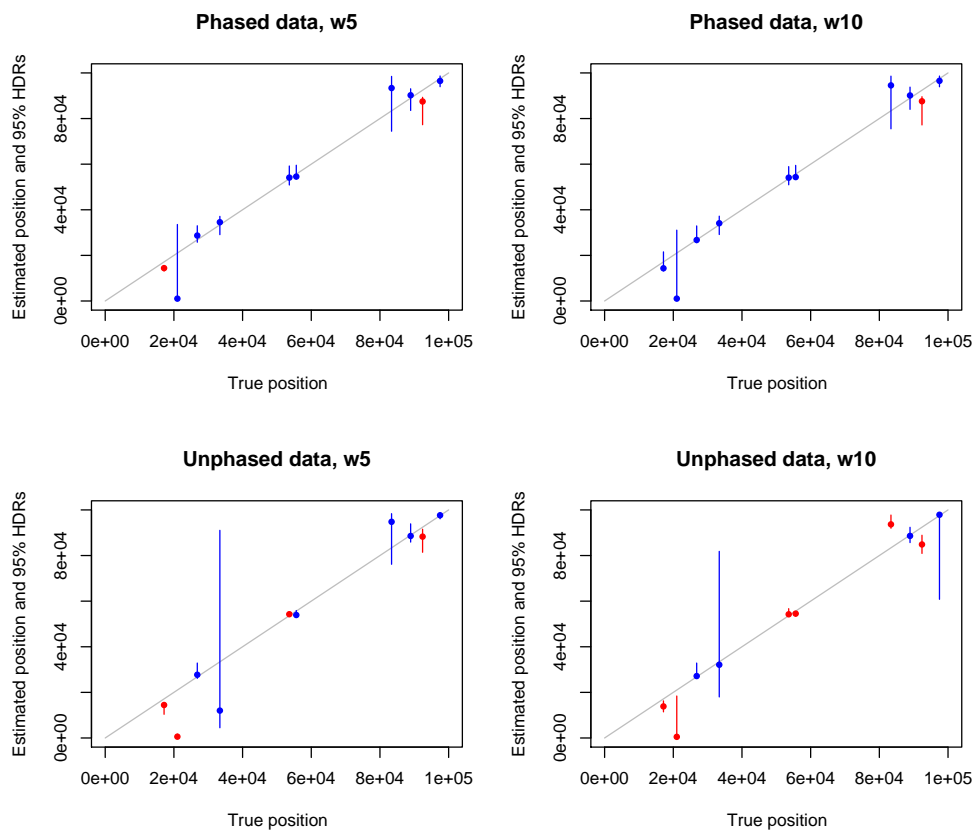


Fig. 4: Localisation accuracy in data with 500 cases and 2000 controls with GRR 2.0.

Table 4: Log Bayes factors for association in data with 500 case and 2000 controls with GRR 2.0.

| Data set | Locus | Known phase | | Unknown phase | |
|----------|-------|-------------|-------|---------------|-------|
| | | w5 | w10 | w5 | w10 |
| 1 | 17136 | 15.70 | 14.38 | 15.75 | 12.05 |
| 2 | 21016 | 27.53 | 29.00 | 28.49 | 28.33 |
| 3 | 26814 | 35.25 | 17.41 | 18.65 | 20.03 |
| 4 | 33389 | 18.84 | 18.87 | 0.1688 | 1.264 |
| 5 | 53608 | 19.90 | 21.08 | 17.05 | 18.96 |
| 6 | 55640 | 23.50 | 25.54 | 24.96 | 31.89 |
| 7 | 83383 | 34.49 | 34.59 | 34.44 | 34.39 |
| 8 | 88939 | 40.92 | 40.84 | 40.52 | 41.98 |
| 9 | 92418 | 28.91 | 29.19 | 28.31 | 27.92 |
| 10 | 97509 | 28.49 | 28.38 | 24.53 | 24.58 |

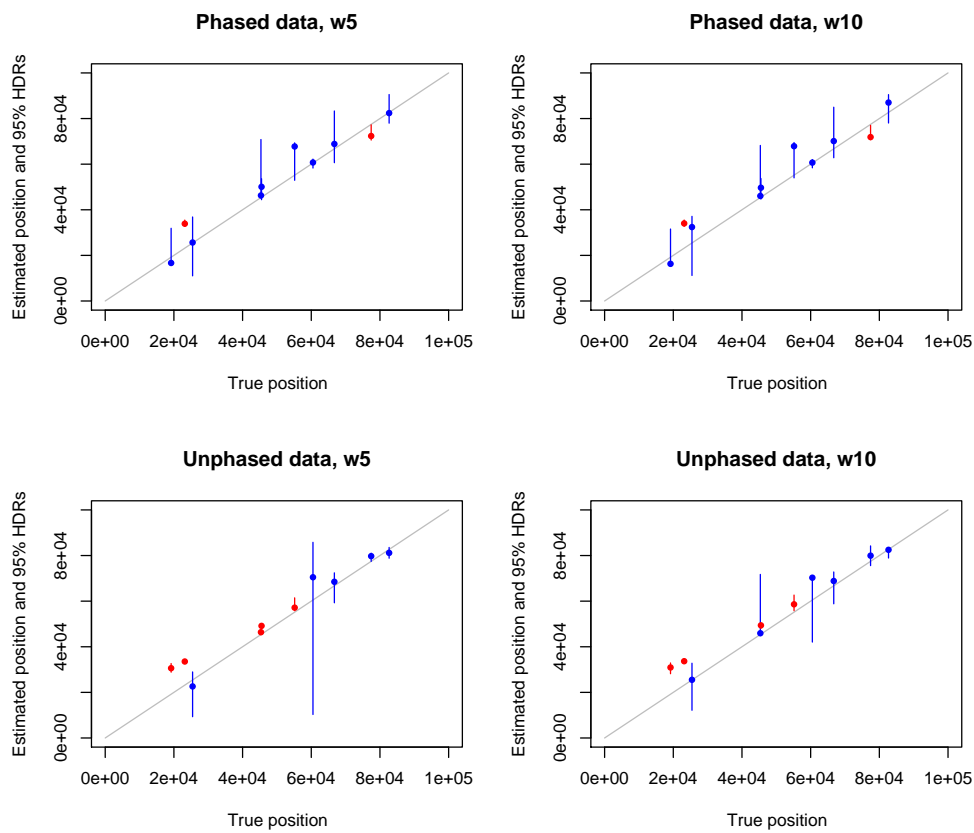


Fig. 5: Localisation accuracy in data with 2000 cases and 2000 controls with GRR 1.5.

Table 5: Log Bayes factors for association in data with 2000 case and 2000 controls with GRR 1.5.

| Data set | Locus | Known phase | | Unknown phase | |
|----------|-------|-------------|-------|---------------|-------|
| | | w5 | w10 | w5 | w10 |
| 1 | 19192 | 21.65 | 20.77 | 18.94 | 19.78 |
| 2 | 23171 | 48.98 | 43.76 | 48.42 | 48.70 |
| 3 | 25444 | 14.09 | 14.08 | 13.79 | 13.80 |
| 4 | 45350 | 13.92 | 13.55 | 12.30 | 10.08 |
| 5 | 45518 | 31.68 | 29.37 | 26.27 | 28.78 |
| 6 | 55162 | 17.92 | 15.90 | 14.54 | 16.37 |
| 7 | 60495 | 15.03 | 19.56 | 0.6433 | 2.509 |
| 8 | 66728 | 6.843 | 10.32 | 5.517 | 6.372 |
| 9 | 77444 | 21.73 | 22.43 | 14.45 | 12.01 |
| 10 | 82670 | 24.64 | 24.64 | 23.96 | 23.05 |

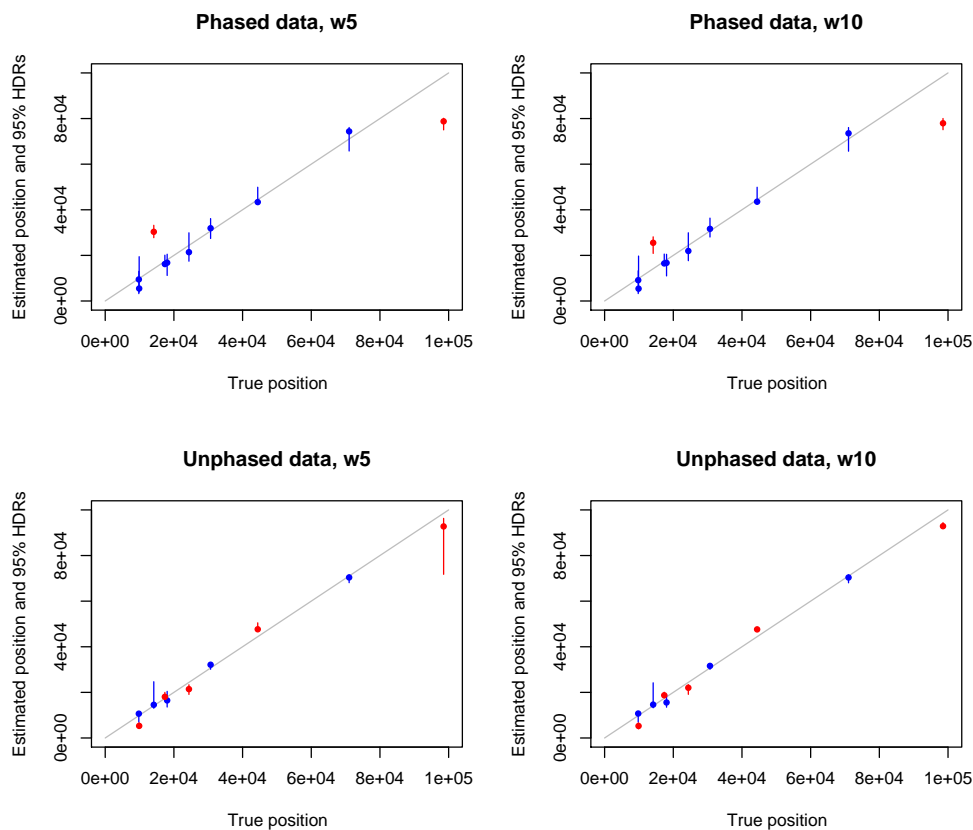


Fig. 6: Localisation accuracy in data with 2000 cases and 2000 controls with GRR 2.0.

Table 6: Log Bayes factors for association in data with 2000 case and 2000 controls with GRR 2.0.

| Data set | Locus | Known phase | | Unknown phase | |
|----------|-------|-------------|-------|---------------|-------|
| | | w5 | w10 | w5 | w10 |
| 1 | 9786 | 87.67 | 87.67 | 85.84 | 87.18 |
| 2 | 9886 | 83.86 | 83.82 | 82.47 | 83.14 |
| 3 | 14142 | 53.02 | 46.91 | 43.67 | 44.50 |
| 4 | 17370 | 126.0 | 117.7 | 104.7 | 95.09 |
| 5 | 18037 | 78.79 | 78.57 | 78.71 | 78.09 |
| 6 | 24380 | 70.06 | 70.34 | 69.84 | 68.62 |
| 7 | 30685 | 95.59 | 95.71 | 95.78 | 95.91 |
| 8 | 44415 | 75.27 | 75.27 | 55.57 | 70.18 |
| 9 | 71030 | 68.81 | 68.63 | 67.37 | 67.12 |
| 10 | 98541 | 52.65 | 36.48 | 24.50 | 25.05 |

Analysis of CYP2D6

I analysed the CYP2D6 data set from Hosking et al.¹. The data contains 1018 individuals (41 cases and 977 controls) typed at 32 SNP markers. The true phase is not know, but taking the approach of Waldron et al.² I compare with phase inferred by PHASE, comparing with five different inferences to—to a certain extent—deal with the uncertainty there.

Posterior densities from these runs are shown in figures 7 (for $w5$) and 8 (for $w10$). Clearly, for $w5$, all data sets—phased or not—performs poorly in localisation accuracy. For $w10$, all the phased data sets include the disease locus in their 95 % HDR, but sadly the unphased MCMC does not. Expanding the window to include all markers (see figure 9) does not improve this.

The Bayes factors for association is shown in table 7. As was the case for the simulated data sets, the Bayes factor resulting from the unphased MCMC is smaller than those from the phased data, but still highly significant.

Table 7: Log Bayes factors for association the CYP2D6 data for five different inferred phases and for the unphased MCMC.

| Data set | $w5$ | $w10$ | $full$ |
|---------------|-------|-------|--------|
| PHASE 1 | 111.9 | 112.9 | 107.9 |
| PHASE 2 | 110.8 | 110.5 | 110.7 |
| PHASE 3 | 110.5 | 113.2 | 111.7 |
| PHASE 4 | 111.4 | 112.6 | 109.8 |
| PHASE 5 | 107.0 | 109.8 | 112.3 |
| Unknown phase | 87.32 | 88.00 | 104.6 |

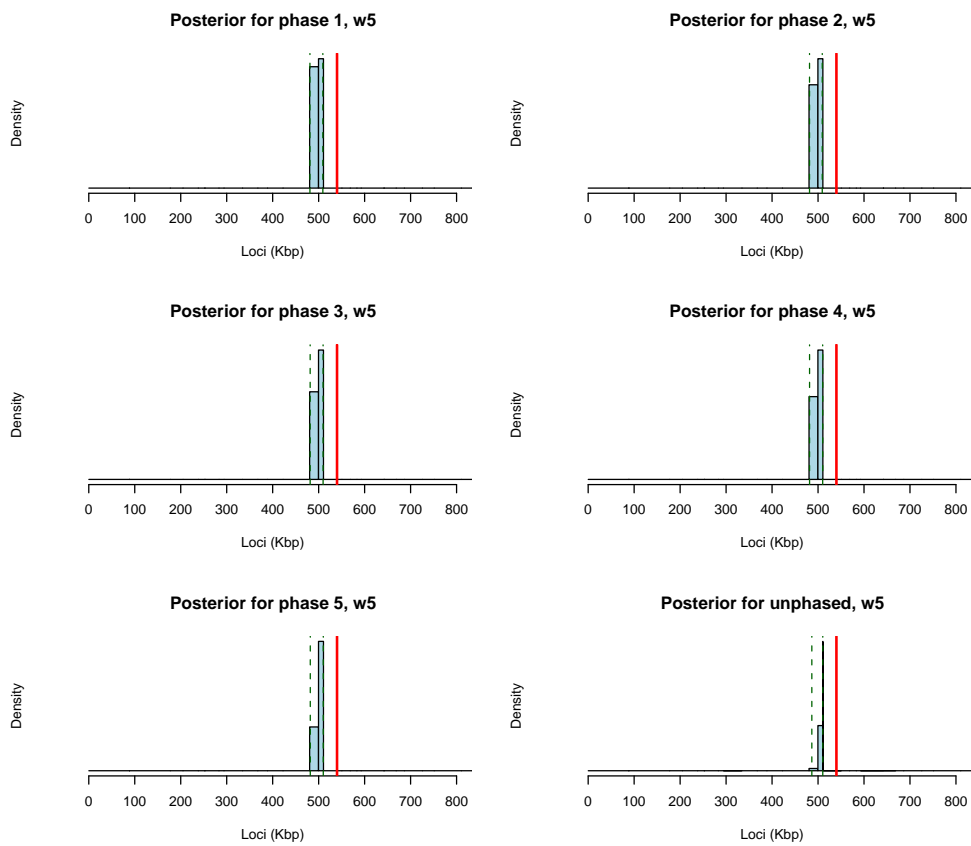


Fig. 7: Posterior densities for CYP2D6 with $w=5$. The blue histograms show the posterior density. The red line indicates the disease locus and the dashed green lines the 95% HDRs.

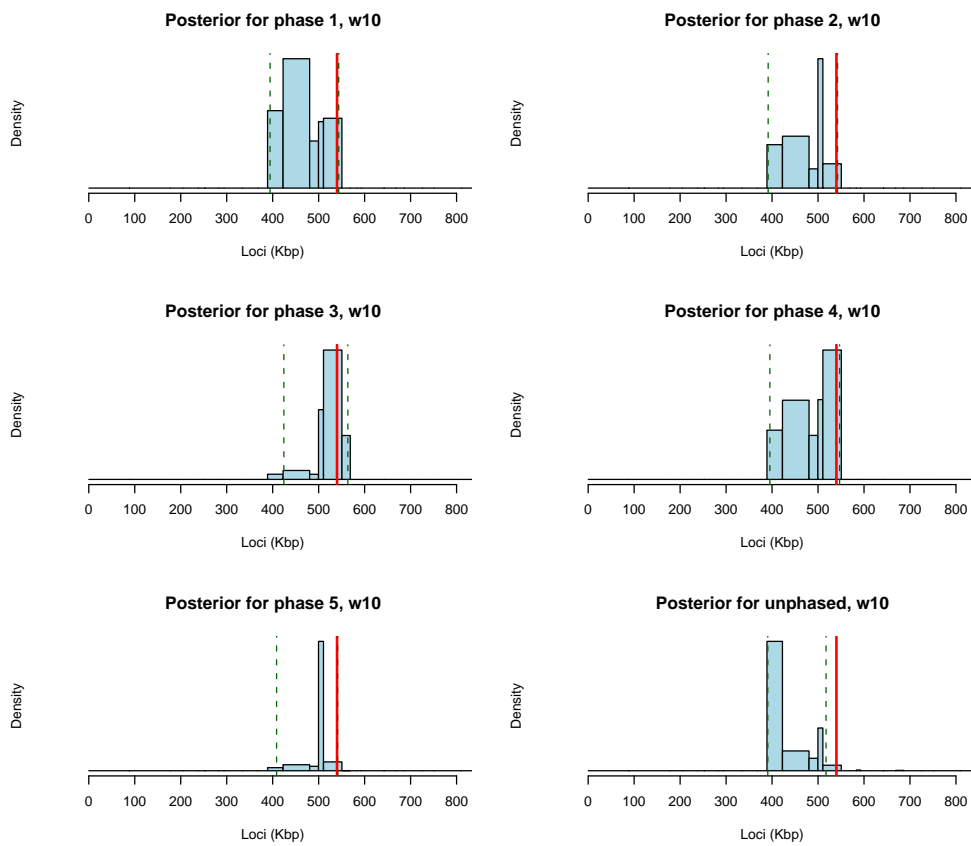


Fig. 8: Posterior densities for CYP2D6 with $w=10$. The blue histograms show the posterior density. The red line indicates the disease locus and the dashed green lines the 95% HDRs.

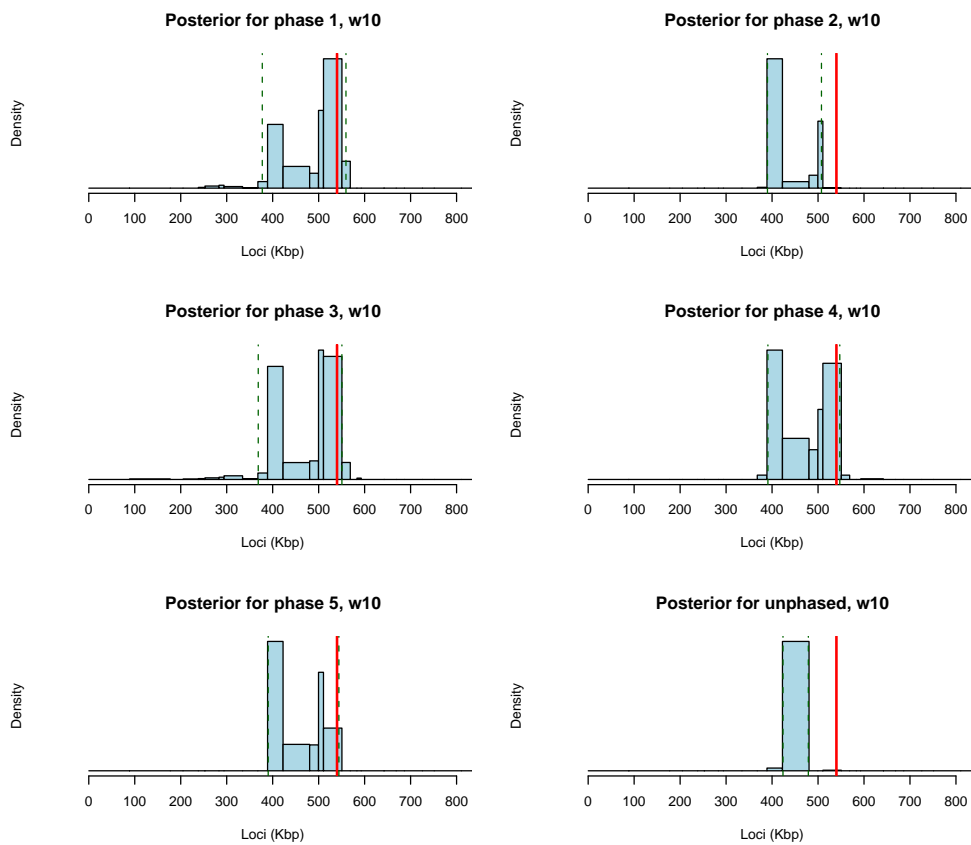


Fig. 9: Posterior densities for CYP2D6 with no windows limitation (so in theory all markers are used when calculating the distance). The blue histograms show the posterior density. The red line indicates the disease locus and the dashed green lines the 95% HDRs.

References

1. L. K. Hosking, P. R. Boyd, C. F. Xu, M. Nisum, K. Cantone, I. J. Purvis, R. Khakhar, M. R. Barnes, U. Liberwirth, K. Hagen-Mann, M. G. Ehm, and J. H. Riley. Linkage disequilibrium mapping identifies a 390 kb region associated with CYP2D6 poor drug metabolising activity. *Pharmacogenomics J*, 2(3):165–75, 2002.
2. E. R. B. Waldron, J. C. Whittaker, and D. J. Balding. Fine mapping of disease genes via haplotype clustering. *Genet Epidemiol*, 30(2):170–179, 2006.