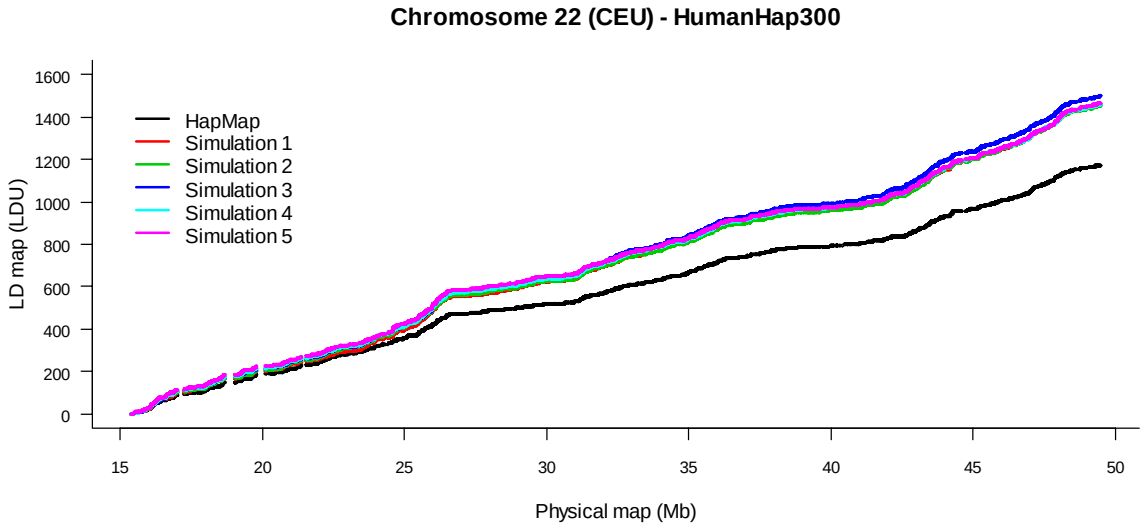
Supplementary materials for
Li C, Li M (2007) GWAsimulator: A rapid whole genome simulation program. Bioinformatics.

Supplementary Figure 1: **Comparison with the HapMap data using LDU maps**

To evaluate our simulation algorithm, we used HapMap phased data as input and compared with the simulated data. We obtained SNP names and positions for the Illumina HumanHap300 chip. After discarding SNPs that are not in the HapMap CEU phased data, 314,174 SNPs remained and were used for simulations. Figure 1 in the paper shows the similarity of short-range LD patterns between the two data sets. We also compared LD patterns using LD unit (LDU) maps (Maniatis et al. 2002). Using the SNPs on the HumanHap300, we constructed LDU maps for the HapMap CEU samples and for five simulated data sets of 60 unrelated individuals. The profiles of the LDU maps were very similar, although the LDU maps for the simulated data sets were longer in overall length (see Supplementary Figure 1).

Supplementary Figure 1: LDU maps of HapMap CEU samples and five simulated data sets on chromosome 22.

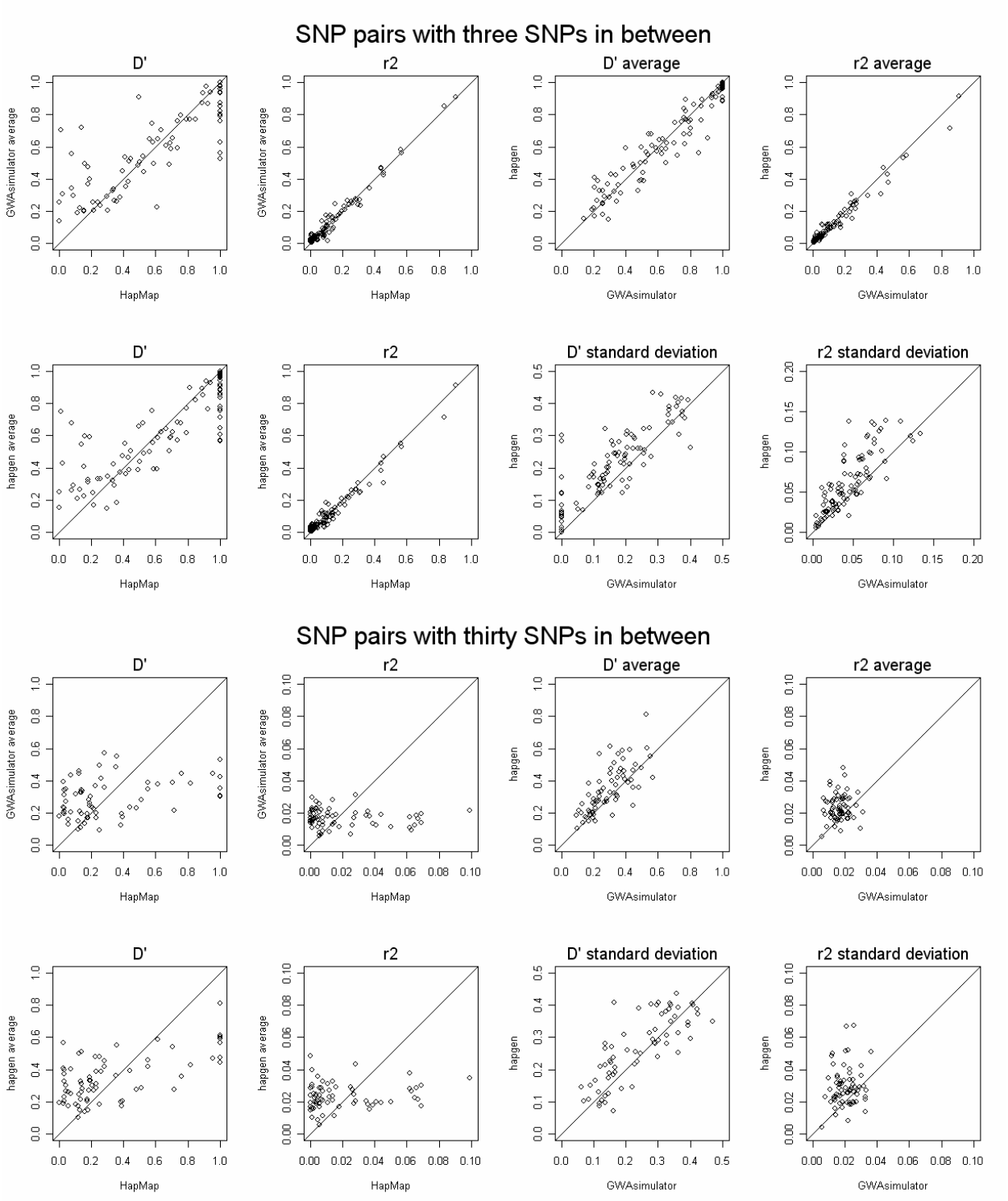


Maniatis N. et al. (2002) The first linkage disequilibrium (LD) maps: Delineation of hot and cold blocks by diplotype analysis. *Proc Natl Acad Sci USA* 99:2228-2233.

Results comparison with the program hapgen

During the review process, a reviewer asked to compare the results from GWAsimulator and those from a program implementing the coalescent approach. We compared the simulated data from GWAsimulator and those from hapgen (http://www.stats.ox.ac.uk/~marchini/software/gwas/hapgen.html), which is the only GWA simulation program that we found and which also uses HapMap phased data as input.

For each program, we simulated 20 replicates of 60 subjects on chromosome 22 for the SNPs on the Illumina's HumanHap300 SNP chip, and evaluated the LD estimates for all SNP pairs in a region (SNPs 1001-1100 of the input file) with 3 SNPs in between and with 30 SNPs in between. The results (on next page) show that GWAsimulator performs similarly as hapgen, especially for SNP pairs with 3 SNPs in between. For SNP pairs that had 30 SNPs in between, both programs tend to overestimate weak LD and underestimate strong LD compared that of the input data. GWAsimulator seems to have a slightly smaller magnitude of overestimation and a slightly bigger magnitude of underestimation. GWAsimulator also tends to achieve slightly smaller variation in LD, which is expected because it relies on empirical LD patterns while hapgen relies on recombination fractions derived from empirical data.


SNP pairs with three SNPs in between
D'
r2
D' average
r2 average
GWAsimulator average
HapMap
happgen
GWAsimulator
D'
r2
D' standard deviation
r2 standard deviation
happgen average
HapMap
happgen
GWAsimulator
SNP pairs with thirty SNPs in between
D'
r2
D' average
r2 average
GWAsimulator average
HapMap
happgen
GWAsimulator
D'
r2
D' standard deviation
r2 standard deviation
happgen average
HapMap
happgen
GWAsimulator