

## ARTICLE

# Global similarity with local differences in linkage disequilibrium between the Dutch and HapMap–CEU populations

Luba Pardo<sup>1</sup>, Zoltán Bochdanovits<sup>1</sup>, Eco de Geus<sup>2</sup>, Jouke J Hottenga<sup>2</sup>, Patrick Sullivan<sup>3</sup>, Danielle Posthuma<sup>1,2</sup>, Brenda WJH Penninx<sup>4,5,6</sup>, Dorret Boomsma<sup>2</sup> and Peter Heutink<sup>\*,1</sup>

<sup>1</sup>Medical Genomics Section, Department of Clinical Genetics, Vrije University Medical Center, Amsterdam, The Netherlands; <sup>2</sup>Department of Biological Psychology, VU University, Amsterdam, The Netherlands; <sup>3</sup>Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; <sup>4</sup>Department of Psychiatry, EMGO Institute, Institute of Neuroscience, Vrije University Medical Center, Amsterdam, The Netherlands; <sup>5</sup>Department of Psychiatry, Leiden University Medical Center, Leiden, The Netherlands; <sup>6</sup>Department of Psychiatry, University Medical Center Groningen, Groningen, The Netherlands

The HapMap project has facilitated the selection of tagging single nucleotide polymorphisms (tagSNPs) for genome-wide association studies (GWAS) under the assumption that linkage disequilibrium (LD) in the HapMap populations is similar to the populations under investigation. Earlier reports support this assumption, although in most of these studies only a few loci were evaluated. We compared pair-wise LD and LD block structure across autosomes between the Dutch population and the CEU–HapMap reference panel. The impact of sampling distribution on the estimation of LD blocks was studied by bootstrapping. A high Pearson correlation (genome-wide; 0.93) between pair-wise  $r^2$  for the Dutch and the CEU populations was found, indicating that tagSNPs from the CEU–HapMap panel capture common variation in the Dutch population. However, some genomic regions exhibited, significantly lower correlation than the genome-wide estimate. This might decrease the validity of HapMap tagSNPs in these regions and the power of GWAS. The LD block structure differed considerably between the Dutch and CEU–HapMap populations. This was not explained by demographic differences between the CEU and Dutch samples, as testing for population stratification was not significant. We also found that sampling variation had a large effect on the estimation of LD blocks, as shown by the bootstrapping analysis. Thus, in small samples, most of the observed differences in LD blocks between populations are most likely the result of sampling variation. This poor concordance in LD block structure suggests that large samples are required for robust estimations of local LD block structure in populations.

*European Journal of Human Genetics* (2009) 17, 802–810; doi:10.1038/ejhg.2008.248; published online 7 January 2009

**Keywords:** Dutch population; HapMap–CEU; pair-wise LD; LD blocks; bootstrapping

\*Correspondence: Dr P Heutink, Human Genetics, section Medical Genomics, VU Medical Center, Van der Boechorstraat 7, 1081 BT Amsterdam, Amsterdam, The Netherlands.

Tel: +31 205989962; Fax: +31 2059983596;

E-mail: p.heutink@vumc.nl

Received 27 August 2008; revised 18 October 2008; accepted 20 November 2008; published online 7 January 2009

## Introduction

Genome-wide association studies (GWAS) are now widely used to search for common variants underlying complex diseases and traits.<sup>1</sup> These studies rely on the presence of discrete blocks of linkage disequilibrium (LD) between polymorphic markers, most commonly single nucleotide

polymorphisms (SNPs).<sup>2–4</sup> By genotyping a set of SNPs (ie, tagging SNPs (tagSNPs)) that are in high LD with a large set of neighbouring (un-typed) variants, one can reduce the costs involved in a GWAS with a limited loss of power.<sup>5,6</sup> The International HapMap Project<sup>4,7</sup> has generated an increasingly dense map of the LD structure of the human genome, on the basis of selected reference panels, to optimize the selection of tagSNPs for association studies in other populations.<sup>4</sup> The main assumption is that the LD structure of the HapMap populations and the specific population under investigation is very similar. Deviations from this assumption may lead to a decrease in the power of GWAS based on tagSNPs.

Recent studies have assessed the similarity of the LD patterns in the three reference panels from the HapMap consortium (phase I and phase II)<sup>8–10</sup> with other populations. The general claim is that tagSNPs from one population can capture common variation in other populations of similar origin.<sup>9–13</sup> However, most of these studies have evaluated only a limited number of loci at varying degrees of SNP density,<sup>8,14</sup> and several of them have shown that the exchangeability of tagSNPs between populations may be locus and SNP-density specific.<sup>10,14</sup> Another issue is the variation in the local LD block structure observed between populations.<sup>15,16</sup> These differences have been attributed to differences in recombination rates,<sup>17</sup> differential demography,<sup>16</sup> stochastic genome variation or sampling variation.<sup>18</sup> The last issue is highly relevant when comparing local LD patterns within small sample sizes, which have been the case for most studies so far, including those for the generation of HapMap phase I and II.

In the past years, an increasing number of GWAS have been conducted using European samples.<sup>19–21</sup> Nonetheless, large scale studies that compare LD structure between European populations and the HapMap have not been performed. In this study, we estimated genome-wide measures of pair-wise LD and LD block structure in the Dutch population using a 600K SNP chip. We compared these LD estimates with data from the CEU–HapMap reference panel (individuals from Utah with northern and western European ancestry)<sup>4</sup> to assess similarities and differences in LD patterns between these two European population samples.

## Subjects and methods

The participants and genotypes we used were derived from a recent GWAS for major depression (MDD) in The Netherlands conducted as part of the Genetic Association Information Network (GAIN) (<http://www.fnih.org/>). The design of the study, description of the phenotypic data, as well as the genotyping procedures and quality control, are extensively detailed in a separate manuscript

(Genome-wide association for Major Depressive Disorder: a possible role for the protein PCLO 2008; Sullivan *et al*, 2008, in press), and in earlier reports.<sup>22</sup> Here, we briefly highlight the information that was relevant for our analysis.

## Subjects

A sample of 1860 participants who were selected as controls for a recent GWAS of MDD (GAIN) project<sup>23</sup> was available for our study. The participants were derived from two longitudinal studies, namely: The Netherlands Twin Registry (NTR) and The Netherlands Study of Depression and Anxiety (NESDA).<sup>23</sup> Participants were selected as controls on the basis of the following criteria: age between 18 and 65 years, place of birth for the individual and his/her parents being The Netherlands or northern Europe, and low liability for MDD.<sup>23</sup> The participants selected for the study were, on average, 42 years old and 60% of them were female. Details of the collection and sampling procedure, as well as other demographic characteristics are described elsewhere.<sup>23</sup> After excluding participants who did not fulfil the inclusion criteria (Sullivan *et al* 2008, in press), a total of 36 trios and 1766 unrelated participants were available for the study.

## DNA isolation, SNP genotyping and quality control

A description of the biological sample collection and DNA isolation is presented elsewhere.<sup>23</sup> The genotyping of approximately 600K SNPs in the Dutch samples was performed by Perlegen Sciences (Mountain View, CA, USA), as part of the GAIN project.<sup>22</sup> The genotyping platform consists of four proprietary, high-density oligonucleotide arrays that were chosen to tag common variation in the European and Asian panels from the HapMap consortium.<sup>2</sup> Besides the quality control of the SNP genotyping performed by Perlegen,<sup>22</sup> additional criteria were applied before the SNP genotypes were included in the final analysis. Briefly, SNPs were excluded if they had a gross mapping problem,<sup>24</sup>  $\geq 2$  genotype disagreements in 40 duplicated samples, a minor allele frequency (MAF)  $< 0.01$ , or  $> 0.05$  missing genotypes in either the controls, or in the cases from the GWAS of MDD. For our analysis, we also excluded SNPs with  $\geq 1$  Mendelian inheritance error in the 36 trios to improve the estimation of haplotypes. In addition, we also removed SNPs with deviations from Hardy–Weinberg Equilibrium proportions (HWE) ( $P < 0.0001$ ) in the trio sample. There were 2402 SNPs with significant deviations from HWE proportions, but that fulfilled the genotyping quality control (based on SNP missingness and/or low MAF).

## Statistical analysis

To assess the similarities and differences of pair-wise LD and LD block structure between the Dutch samples (trios and unrelated subjects) and those from the HapMap

consortium, we downloaded the genotypes available from the 30 trios from CEU–HapMap panel (phase II)<sup>25</sup> (<http://www.hapmap.org>). From the CEU panel, we retrieved the genotypes of the SNPs that were present in the Dutch sample. SNPs that were not genotyped as part of the HapMap project, but that were available from dbSNP<sup>26</sup> (eg, they were genotyped in the HapMap panels as quality control for the GAIN project) were also included in the analysis ( $n = \sim 7633$ ). The physical positions of the SNPs were updated using dbSNP build 36.<sup>26</sup>

The MAF distribution was estimated in the trio-founders (Dutch and CEU samples) and in the group of unrelated Dutch subjects. Pair-wise LD ( $r^2$  and  $D'$ ), and LD blocks were calculated per chromosome in both the Dutch (trios and unrelated subjects) as well as in the CEU samples. HAPLOVIEW v1.4<sup>27</sup> was used to estimate the above parameters with default settings except for the MAF of SNPs that was set to  $\geq 0.05$ . LD blocks were determined using the method of Gabriel *et al* (2000).<sup>3</sup>

We used Fisher's exact test to test for statistical differences in MAF between the Dutch and the CEU trio-founders for individual SNPs, adjusting for multiple testing. Next, to compare the similarities of pair-wise LD between the Dutch and the CEU samples (trios), Pearson correlation coefficients ( $r$ ) were calculated over non-overlapping 500 kb windows (with at least 50 pair-wise  $r^2$  values). Further, we used Fisher's z-transformation to convert  $r$  values to a normally distributed z distribution.<sup>28</sup> This allowed us to look for local variation in the correspondence of pair-wise LD between populations by testing whether the Pearson correlation in a specific window was significantly lower than the genome-wide average on all autosomes. The distributions of block-derived parameters (total number of blocks, median block size, number of markers per block, and number of singleton SNPs) were compared between the Dutch trios and unrelated subjects, and the CEU samples using non-parametric tests (Wilcoxon signed rank test). These analyses were carried-out using the R statistical package v 2.7.0.

The level of genetic differentiation between the CEU and Dutch samples was assessed by means of the identity-by-state (IBS) test.<sup>29</sup> This test was used to assess whether an individual is genetically more or less similar to a phenotypically discordant individual, when compared with a phenotypically concordant individual. Here, the phenotype is 1 if the individual is derived from the CEU sample, and 2 otherwise. For this analysis we restricted the comparison to the Dutch and CEU trios to have comparable sample sizes. These analyses were carried out using PLINK v1.03.<sup>29</sup>

To study the impact of the sampling variation on the estimation of the block structure, we performed bootstrapping. This means that we randomly sampled the 36 Dutch trios with replacement to create 1000 replicas, each

with 36 trios. Next, LD blocks were estimated using the Gabriel method for each replicated set of 36 trios. Therefore, we estimated LD blocks 1000 times per chromosome. For each LD block observed in the Dutch trios (real blocks), we derived a frequency distribution for the proportion ( $X$ ) of real blocks that were retrieved from the simulated data ( $X$  is the number of simulations in which the observed blocks were present out of 1000 simulations). We considered a real block robust if it was retrieved in at least 95% of the bootstraps. The analyses of bootstrapping and LD block estimation were carried out on the genetic cluster computer (<http://www.geneticcluster.org>).

## Results

The present analysis is on the basis of 427 853 autosomal SNPs that were genotyped in the CEU and the Dutch population samples. Table 1 presents the total number of markers analyzed per chromosome. The average inter-marker distance across all autosomes was 6.31 kb. The large variation in the estimates of the inter-marker distances shows that the markers are not evenly spaced on the physical map. This feature of our data allowed us to evaluate whether the differences in the LD parameters between populations and/or samples were dependent on the differential SNP density.

The MAF distribution of SNPs in both the Dutch and CEU trio-founders, and in the unrelated subjects was estimated. There was an excess of rare variants ( $0 < \text{MAF} < 5\%$ ; approximately 5% of SNPs per chromosome) in the Dutch trio-founders, but for SNPs with  $\text{MAF} \geq 10\%$ , the pattern was very similar between the three samples (Figure 1). Despite large differences in the density of markers per chromosome, the distribution of MAF was very similar across all autosomes (Supplementary Figure 1). There were 31 SNPs with significant MAF differences between the Dutch and the CEU trio-founders (Fisher test; Bonferroni-adjusted  $P$ -value:  $< 0.002$ ). We inspected those SNPs in SNPdb (latest release) and found that the MAF for these SNPs in the CEU sample differed according to the submitter group, and for most of them there were discordant genotypes. Thus, the differences we observed are not due to population differences.

Figure 2a depicts the median pair-wise  $r^2$  calculated over non-overlapping 500 kb windows in the Dutch trios. The median pair-wise  $r^2$  averaged across all chromosomes was 0.015 and 0.016 for the Dutch and CEU trios, respectively. These low pair-wise  $r^2$  estimates are expected as the Perlegen SNP-chips were designed to analyze tagging markers. More interesting is the high correlation of pair-wise  $r^2$  between the two trio samples (Figure 2b). The back-transformed mean Pearson correlation was 0.925 with 95% CI from 0.921 to 0.997. We tested whether the correlation estimates per window were significantly lower than the

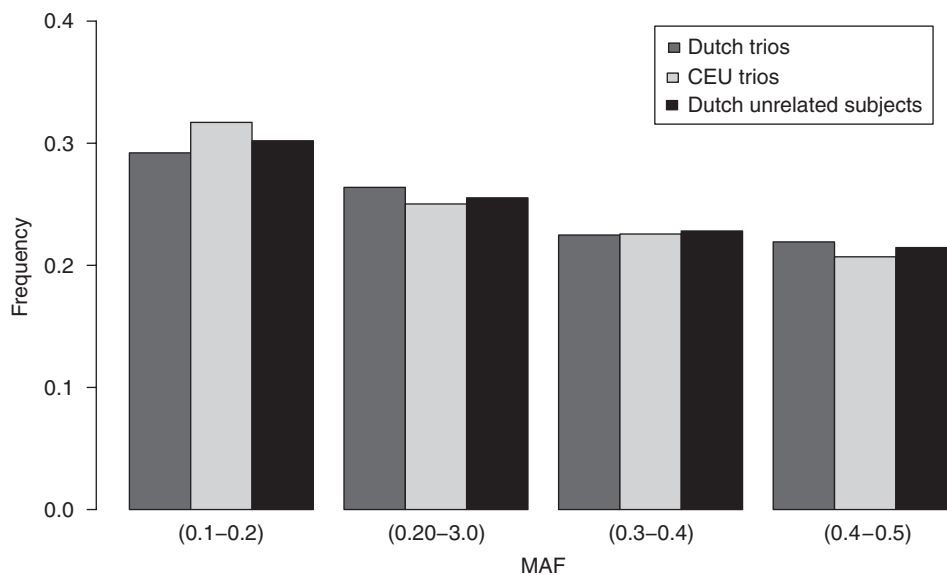
**Table 1** Median block length (interquartile range) and number of haplotypes per block in the Dutch and CEU populations

Chromosome	No. markers	Average IM <sup>a</sup> distance (kb)	Dutch sample (trios)			CEU–HapMap		
			Total no. blocks <sup>b</sup>	% DNA in blocks <sup>b</sup>	No. SNP singletons <sup>b</sup>	Total no. blocks	% DNA in blocks	No. SNP singletons
1	31 108	7.92 (129.62)	4652	84.71	13 537	4055	67.92	16 487
2	34 766	6.98 (29.05)	5304	91.74	14 914	4777	75.63	17 659
3	28 925	6.89 (29.57)	4414	78.41	12 579	3964	63.59	14 942
4	26 170	7.30 (25.24)	4062	73.96	10 935	3623	60.06	13 203
5	27 515	6.6 (25.5)	4142	72.2	11 610	3757	58.84	13 911
6	28 582	5.97 (21.06)	4230	72.27	11 655	3841	59.94	14 005
7	23 471	6.76 (25.38)	3625	56.27	10 171	3212	45.1	12 215
8	23 421	6.24 (26.36)	3611	55.1	10 035	3141	43.72	12 128
9	20 341	6.88 (175.44)	3064	40.19	9240	2724	32.2	10 996
10	22 655	5.97 (25.26)	3441	51.42	9877	2975	40.22	12 143
11	21 146	6.35 (25.70)	3181	54.16	8976	2858	42.28	10 908
12	20 648	6.39 (15.49)	3066	50.94	9045	2657	42.03	10 963
13	17 013	5.64 (9.22)	2565	40.41	7177	2230	31.95	8870
14	14 346	6.07 (11.13)	2123	31.92	6546	1879	26.2	7738
15	13 467	6.08 (22.28)	2086	28.36	6157	1839	23.41	7271
16	13 869	6.39 (88.88)	2088	20.81	6753	1809	17.07	7940
17	10 503	7.5 (17.1)	1582	23.32	5138	1356	18.96	6090
18	12 967	5.87 (17.04)	1989	28.13	5832	1752	22.34	6910
19	6081	10.44 (108.51)	884	11.77	3285	749	9.75	3752
20	10 856	5.75 (22.90)	1686	22.34	4882	1483	17.66	5856
21	6103	5.46 (8.51)	938	11.63	2768	826	9.65	3235
22	5522	6.17 (13.01)	817	10.01	2858	680	8.04	3325

SNP, single nucleotide polymorphism.

<sup>a</sup>IM, intermarker distance (kilobases).

<sup>b</sup>Estimates were statistically significant (Wilcoxon Rank test; *P*-value <0.001).



**Figure 1** Distribution of minor allele frequency (MAF) of single nucleotide polymorphisms (SNPs) across 22 autosomes in the Dutch trios (red), CEU trios (light blue) and Dutch unrelated subjects (black). The frequency distribution of MAF for SNPs was categorized in bins. A full colour version of this figure is available at the European Journal of Human Genetics online.

mean value estimated for all autosomes, and found that 1527 (30%) of the 5092, 500 kb-windows had significantly lower correlation estimates. Nonetheless, the lowest observed correlation was reasonably high ( $r=0.6$ ). We also

estimated pair-wise  $r^2$  in the Dutch unrelated subjects. The median  $r^2$  derived from this sample was lower than those estimated from the Dutch trios. In addition, the Pearson correlation of  $r^2$  between the Dutch samples (trios vs

unrelated subjects) was not 1, despite the fact that we were comparing samples from the same population. This reflects the inaccuracy in estimating haplotypes in unrelated subjects in the presence of low  $r^2$  values with certain algorithms (eg Expectation–Maximization as implemented in Haploview)<sup>30</sup> or an overestimation of low  $r^2$  values in small samples.<sup>31</sup>

The LD block structure between the CEU and Dutch trios was also compared. We observed a higher number of blocks per chromosome and longer blocks in the Dutch trios, while the number of SNPs that were not in blocks was reduced (Table 1). The number of SNP singletons (SNPs that were not present in blocks) was significantly larger in the HapMap sample than in the Dutch trio sample for all chromosomes.

In addition, the median length of the blocks was higher for the Dutch than for the CEU trios (genome-wide Wilcoxon test;  $P$ -value  $<0.0001$ ), although the interquartile distribution of this parameter overlaps in the two population samples (Table 2). Differences in other parameters, such as the median number of SNPs per block and the median number of haplotypes (for haplotypes with population frequencies  $\geq 0.05$ ) per block, were different (Table 2). LD blocks were also calculated in the unrelated Dutch individuals. Owing to computational limitations, we restricted the analysis of LD blocks to 400 randomly chosen subjects and observed a significantly larger number of blocks when compared with the Dutch trios. However, the estimates per block, such as median number of

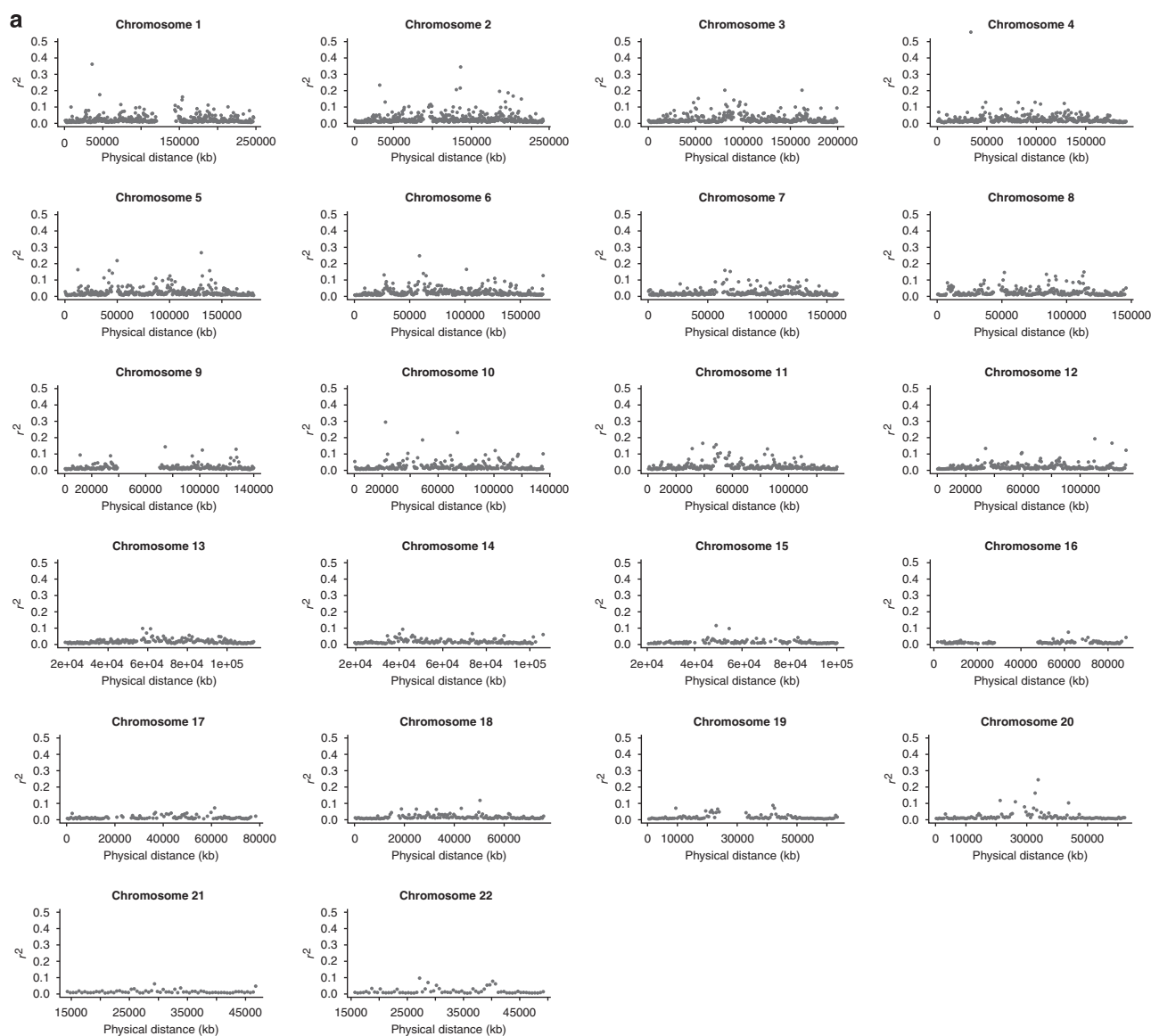
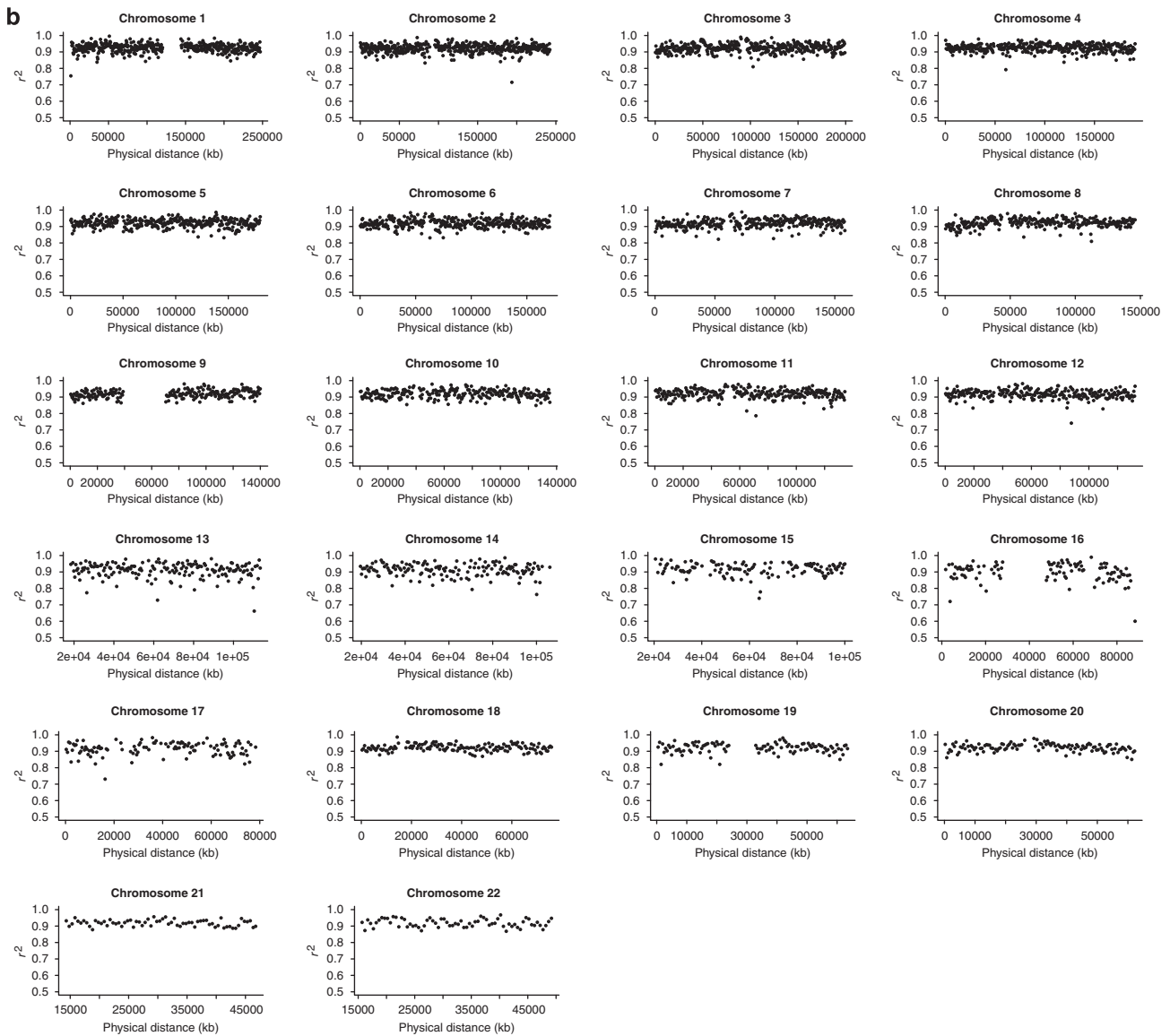


Figure 2 Continued



**Figure 2** (a) Pair-wise linkage disequilibrium (LD) ( $r^2$ ) per chromosome for the Dutch trio sample. The median  $r^2$  per 500 kb window is plotted against physical distance (kb) in each chromosome. (b) Pearson correlation between  $r^2$  estimates for the CEU and the Dutch trios. The Pearson correlation between  $r^2$  calculated over 500 kb windows is plotted against physical distance (kb).

haplotypes per blocks and median length, were not different between these Dutch samples (Table 2).

We investigated whether the differences observed in the LD block structure were due to population structure. We used permutation tests for observing between-group IBS differences<sup>29</sup> (IBS test as implemented in PLINK), which allows testing whether, on average, individuals from one population (CEU) are more similar/dissimilar than individuals from another population (Dutch). The nominal  $P$ -values of the IBS test were significant for chromosomes 14, 15, and 17 (chromosome 14;  $P$ -value = 0.021, chromosome 15;  $P$ -value = 0.03, and chromosome 17;  $P$ -value = 0.045),

although these  $P$ -values were not significant after adjusting for multiple testing (0.05/22 chromosomes). This suggests that large differences in population structure between the Dutch and the CEU samples are unlikely.

To assess the effect of sampling variation on the LD structure, we performed a bootstrapping analysis as described in the section Subjects and methods. We derived a frequency distribution of the proportion of observed blocks in the Dutch population (real blocks) that were present in the replicas (Figure 3). By defining a Dutch block as robust if it was identified in 95% or more of the simulations, we found that only between 7.14% (chromosome 8)

**Table 2** Number of SNP analyzed and intermarker distance

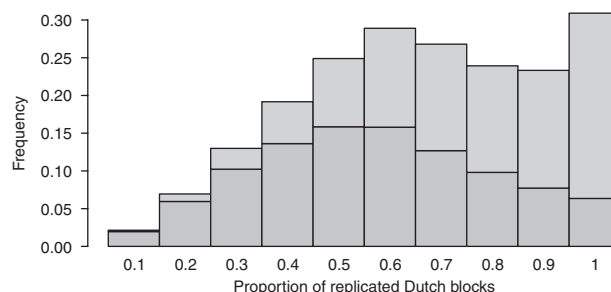
Chromosome	Dutch sample				CEU–HapMap	
	Unrelated subjects		No. haplotypes	Trios	Median length <sup>a</sup>	No. haplotypes
No. haplotypes	Median length	No. haplotypes				
1	3 (3–4)	7.23 (2.38–17.82)	3 (3–4)	7.11 (2.51–18.41)	3 (3–4)	6.54 (2.28–17.15)
2	3 (3–4)	7.19 (2.32–17.61)	3 (3–4)	7.49 (2.34–18.39)	3 (3–4)	6.64 (2.21–16.68)
3	3 (3–4)	6.64 (2.22–16.84)	3 (3–4)	7.01 (2.15–18.22)	3 (3–4)	6.11 (1.92–16.45)
4	3 (3–4)	7.80 (2.61–18.91)	3 (3–4)	7.67 (2.62–19.47)	3 (3–4)	6.77 (2.33–18.09)
5	3 (3–4)	6.95 (2.48–17.67)	3 (3–4)	7.51 (2.52–18.74)	3 (3–4)	6.59 (2.22–17.07)
6	3 (3–4)	7.13 (2.26–17.86)	3 (3–4)	7.25 (2.44–18.36)	3 (3–4)	6.53 (2.12–17.23)
7	3 (3–4)	6.59 (2.23–16.52)	3 (3–4)	6.54 (2.16–17.08)	3 (3–4)	5.65 (1.90–15.50)
8	3 (3–4)	6.24 (2.07–15.57)	3 (3–4)	6.19 (2.02–16.6)	3 (3–4)	5.63 (1.89–15.36)
9	3 (3–4)	5.27 (1.85–13.72)	3 (3–4)	5.56 (1.93–14.11)	3 (3–4)	5.02 (1.69–12.75)
10	3 (3–4)	6.47 (2.22–15.43)	3 (3–4)	6.70 (2.307–15.64)	3 (3–4)	5.90 (1.96–14.38)
11	3 (3–4)	6.50 (2.18–16.82)	3 (3–4)	6.81 (2.25–17.42)	3 (3–4)	5.92 (1.98–15.71)
12	3 (3–4)	6.47 (2.20–16.42)	3 (3–4)	6.78 (2.27–17.98)	3 (3–4)	6.41 (2.12–16.37)
13	3 (3–4)	6.71 (2.27–17.16)	3 (3–4)	6.78 (2.25–17.73)	3 (3–4)	5.93 (2.01–15.61)
14	3 (3–4)	6.23 (2.07–14.90)	3 (3–4)	6.34 (2.07–15.68)	3 (3–4)	5.80 (1.87–14.22)
15	3 (3–4)	5.51 (1.84–13.13)	3 (3–4)	5.63 (1.67–13.52)	3 (3–4)	4.96 (1.59–12.32)
16	3 (3–4)	4.15 (1.34–10.44)	3 (3–4)	4.24 (1.37–10.93)	3 (3–4)	3.75 (1.23–10.29)
17	3 (3–4)	5.54 (1.94–13.43)	3 (3–4)	5.40 (1.82–13.52)	3 (3–4)	5.17 (1.71–12.74)
18	3 (3–4)	6.00 (2.09–14.70)	3 (3–4)	6.03 (2.11–15.22)	3 (3–4)	5.68 (1.97–14.16)
19	3 (3–4)	6.15 (1.99–13.86)	3 (3–4)	5.59 (1.88–13.37)	3 (3–4)	5.10 (1.68–13.34)
20	3 (3–4)	5.40 (1.87–12.99)	3 (3–4)	5.55 (1.76–14.22)	3 (3–4)	4.86 (1.60–12.40)
21	3 (3–4)	6.07 (1.96–14.11)	3 (3–4)	5.96 (2.05–14.68)	3 (3–4)	5.43 (2.04–13.40)
22	3 (3–4)	4.85 (1.61–11.10)	3 (3–4)	4.21 (1.63–10.55)	3 (3–4)	4.10 (1.32–10.04)

<sup>a</sup>Estimates were statistically significant (Wilcoxon rank test; *P*-values < 0.001).

and 12.44% (chromosome 19) of the real blocks were robust. Figure 3 shows that when the real blocks were robust, there is a 70–75% probability that the real block will be identified in the CEU population as well (blue bars).

## Discussion

In this study, we estimated pair-wise  $r^2$  over 427K SNP markers in a Dutch population and compared these estimates with data from the CEU–HapMap project. A high correlation (0.93) between pair-wise  $r^2$  across all autosomes was found, which suggests that tagSNPs from the HapMap project can capture common variation in the Dutch population. Our findings extend earlier studies of high similarity of pair-wise LD between populations with shared ancestry based on fewer loci. We also observed that approximately 30%, 500-kb windows showed local variation of the Pearson correlation estimates, which were significantly lower than the genome-wide estimate. This suggests that in a considerable part of the genome, the efficiency of tagSNPs from HapMap to capture common variants may decrease as proposed in earlier studies.<sup>14,15,32</sup> The LD block structure between the CEU and the Dutch populations was also compared. The average statistics per block, such as the number of common haplotypes and the number of SNPs per block, were not statistically different, but there was a large variation in the block boundaries between these two populations.



**Figure 3** Proportion of Dutch blocks retrieved in 1000 bootstraps. Frequency distribution of the proportion of observed Dutch blocks that were retrieved from the simulated data. The proportion of Dutch blocks that were observed in the CEU sample are depicted in light blue and the proportion of Dutch blocks that were not present in the CEU sample are depicted in grey. A full colour version of this figure is available at the European Journal of Human Genetics online.

In a recent study, large variation in the LD block structure, even in populations of similar geographical background, was reported.<sup>16</sup> Gu *et al*<sup>32</sup> suggested that individual population history and genetic factors, including genetic drift, may account for these differences. Here, we found that sampling variation has a strong effect on the local variation of LD. The bootstrapping analysis showed that even in the same population sample, the observed blocks are retrieved only in 60–70% of the replicas. Moreover, we did not find substantial population stratification between the CEU–HapMap and Dutch samples. Furthermore, it has been shown that genetic drift has large

effects only for variants with low population frequencies,<sup>33</sup> which in our analyses were not included to estimate LD blocks. Thus, although differential recombination and other genetic stochastic factors are likely to be involved in the differential pattern of local LD structure between populations, we found that sampling variation has a major role, which is in agreement with other studies.<sup>34</sup>

The discrepancy between the high overall correlation of pair-wise  $r^2$  and the variation in the LD blocks between the Dutch and the CEU samples can partly be explained by the fact that  $|D'|$  statistics are sensitive to sampling distribution, whereas  $r^2$  is more robust.<sup>35,36</sup> We also estimated pair-wise LD using  $|D'|$  and observed larger variation in the correlation estimates between the CEU and Dutch population (Supplementary Figure 2). This further supports the finding that small samples largely account for the observed variation of the block structure, as the Gabriel method is based on  $D'$ . It is not clear what the optimal sample size for a robust estimation of LD blocks should be. Samples of 60–100 individuals have been shown to be enough for a robust reconstruction of LD block structure,<sup>14,16,37</sup> which may hold when assessing the transferability of HapMap tagSNPs to other populations. We and others<sup>16,18</sup> have shown that for a reliable estimation of local LD structure, larger sample sizes are needed. This has consequences for interpreting association studies performed using tagSNPs derived from small samples, as in association studies regions with different LD blocks in cases compared with controls are assumed to harbour putative functional variants. This might partly explain the lack of replication of genetic association studies.<sup>38</sup>

A limitation of this study is that the participants of this study were selected as controls for a GWAS of MDD on the basis of their low liability for MDD. As a result, this sample may not be representative of the entire Dutch population. Nevertheless, the demographic characteristics of this Dutch sample were comparable to other European (or with European ancestry) samples used as replication controls (Sullivan *et al*, 2008, submitted). This means that the Dutch individuals used for our analysis do not largely deviate from other European population-based samples. Another issue was the use of tagSNPs to compare LD patterns, which results in a relatively large spacing between markers. As a result of this, we were not able to relate our findings to genomic features, such as differential recombination rates, that are an important cause of local differences in LD block structure between populations.<sup>4</sup> However, our aim was to assess whether estimates of pair-wise LD vary substantially at the genomic level. At the marker density used in this study, we observed a consistent pattern of MAF distribution (Supplementary Figure 1) and a high correlation (above 0.6) between pair-wise  $r^2$  (Figure 2a) across all autosomes. This suggests that the marker density does not have a large effect on the patterns of LD we observed.

The HapMap project has been a milestone for the design of GWAS, as it provides the most complete human-SNP catalogue to the scientific community. The SNP-chip platforms that are currently used for these studies have been designed using the HapMap data. In the past years, GWAS have identified common variants associated with a wide range of disorders and other human traits.<sup>39</sup> Nonetheless, many GWAS have been negative, and others have not been replicated.<sup>38</sup> Low power, selection bias and population stratification<sup>39</sup> might explain these negative GWAS. Given our results, as well as other recent findings,<sup>32</sup> it can be argued that local differences in LD structure between the HapMap populations and the target population may also lead to a reduced power to detect candidate variants in regions of the genome where the correlation between pair-wise  $r^2$  in the HapMap panels and the target population is decreased. This is likely to affect the efficiency of HapMap tagSNPs. Of note, we only analyzed LD patterns between Dutch and CEU–HapMap reference panel, and therefore the patterns of LD we observed between these two European samples may not extend to other settings. Nonetheless, our analysis highlights the importance of local variations of LD between populations with shared ancestry that should be taken into account in both the design of GWAS and in the interpretation of results derived from these studies.

In summary, in this large-scale comparison of LD patterns between a Dutch population and the CEU–HapMap sample, we observed a high overall correlation of pair-wise  $r^2$  across 22 autosomes, with local variation in some genomic regions. Both a denser SNP-chip platform covering a larger number of SNPs derived from the HapMap panels, as well as a higher threshold to select HapMap tagSNPs, may reduce this problem. The second message of our work is that the reliable estimation of LD blocks is still an unsolved issue that is largely affected by sample size. Given the current definition of LD blocks, there is too much stochastic variability in the estimates of the blocks, because of the small samples that are used. Hence the observed blocks are often not robust. The current phase of the HapMap project (phase III) has been extended to include more populations as well as more individuals from the three earlier reference panels, which should provide an interesting opportunity to evaluate this issue.

#### Acknowledgements

We would like to acknowledge support from NWO: genetic basis of anxiety and depression (904-61-090); resolving cause and effect in the association between exercise and well-being (904-61-193); twin-family database for behavior genomics studies (480-04-004); twin research focusing on behavior (400-05-717); Center for Medical Systems Biology (NWO Genomics); Spinozapremie (SPI 56-464-14192); NWO-VI016-065-318; Centre for Neurogenomics and Cognitive Research (CNCR-VU); genomewide analyses of European twin and population cohorts (EU/QLRT-2001-01254); genome scan



for neuroticism (NIMH R01 MH059160); Geestkracht program of ZonMW (10-000-1002); matching funds from universities and mental health care institutes involved in NESDA (GGZ Buitenamstel-Geestgron84den, Rivierduinen, University Medical Center Groningen, GGZ Lentis, GGZ Friesland, GGZ Drenthe). Major funding for this project is from the Genetic Association Information Network of the Foundation for the US National Institutes of Health, a public–private partnership between the NIH and Pfizer Inc., Affymetrix Inc. and Abbott Laboratories. Genetic Cluster Computer is financially supported by the Netherlands Scientific Organization (NWO 480-05-003). Stichting Nationale Computerfaciliteiten –NCF (SH-104-08 Grant) is also acknowledged.

## References

- Zeggini E, Weedon MN, Lindgren CM *et al*: Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* 2007; **316**: 1336–1341.
- Hinds DA, Stuve LL, Nilsen GB *et al*: Whole-genome patterns of common DNA variation in three human populations. *Science* 2005; **307**: 1072–1079.
- Gabriel SB, Schaffner SF, Nguyen H *et al*: The structure of haplotype blocks in the human genome. *Science* 2002; **296**: 2225–2229.
- International HapMap Consortium: A haplotype map of the human genome. *Nature* 2005; **437**: 1299–1320.
- Collins FS, Guyer MS, Chakravarti A: Variations on a theme: cataloging human DNA sequence variation. *Science* 1997; **278**: 1580–1581.
- Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES: High-resolution haplotype structure in the human genome. *Nat Genet* 2001; **29**: 229–232.
- The International HapMap Consortium: The International HapMap Project. *Nature* 2003; **426**: 789–796.
- Xing J, Witherspoon DJ, Watkins WS, Zhang Y, Tolpinrud W, Jorde LB: HapMap tagSNP transferability in multiple populations: general guidelines. *Genomics* 2008; **92**: 41–51.
- Ribas G, Gonzalez-Neira A, Salas A *et al*: Evaluating HapMap SNP data transferability in a large-scale genotyping project involving 175 cancer-associated genes. *Hum Genet* 2006; **118**: 669–679.
- Mueller JC, Lohmussaar E, Magi R *et al*: Linkage disequilibrium patterns and tagSNP transferability among European populations. *Am J Hum Genet* 2005; **76**: 387–398.
- de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D: Efficiency and power in genetic association studies. *Nat Genet* 2005; **37**: 1217–1223.
- Gonzalez-Neira A, Ke X, Lao O *et al*: The portability of tagSNPs across populations: a worldwide survey. *Genome Res* 2006; **16**: 323–330.
- Conrad DF, Jakobsson M, Coop G *et al*: A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet* 2006; **38**: 1251–1260.
- Montpetit A, Nelis M, Laflamme P *et al*: An evaluation of the performance of tag SNPs derived from HapMap in a caucasian population. *PLoS Genet* 2006; **2**: e27.
- Liu N, Sawyer SL, Mukherjee N *et al*: Haplotype block structures show significant variation among populations. *Genet Epidemiol* 2004; **27**: 385–400.
- Gu S, Pakstis AJ, Li H, Speed WC, Kidd JR, Kidd KK: Significant variation in haplotype block structure but conservation in tagSNP patterns among global populations. *Eur J Hum Genet* 2007; **15**: 302–312.
- Greenwood TA, Rana BK, Schork NJ: Human haplotype block sizes are negatively correlated with recombination rates. *Genome Res* 2004; **14**: 1358–1361.
- Osabe D, Tanahashi T, Nomura K *et al*: Evaluation of sample size effect on the identification of haplotype blocks. *BMC Bioinformatics* 2007; **8**: 200.
- Kayser M, Liu F, Janssens ACJW *et al*: Three genome-wide association studies and a linkage analysis identify HERC2 as a human iris color gene. *Am J Hum Genet* 2008; **82**: 411–423.
- van Es MA, van Vught PWJ, Blauw HM *et al*: Genetic variation in DPP6 is associated with susceptibility to amyotrophic lateral sclerosis. *Nat Genet* 2008; **40**: 29–31.
- Sulem P, Gudbjartsson DF, Stacey SN *et al*: Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat Genet* 2007; **39**: 1443–1452.
- GAIN Collaborative Research GroupManolio TA, Rodriguez LL, Brooks L *et al*: New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nat Genet* 2007; **39**: 1045–1051.
- Boomsma DI, Willemsen G, Sullivan PF *et al*: Genome-wide association of major depression: description of samples for the GAIN Major Depressive Disorder Study: NTR and NESDA biobank projects. *Eur J Hum Genet* 2008; **16**: 335–342.
- Hemminger BM, Saelim B, Sullivan PF: TAMAL: an integrated approach to choosing SNPs for genetic studies of human complex traits. *Bioinformatics* 2006; **22**: 626–627.
- International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR *et al*: A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007; **449**: 851–861.
- Wheeler DL, Barrett T, Benson DA *et al*: Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2008; **36**: D13–D21.
- Barrett JC, Fry B, Maller J, Daly MJ: Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005; **21**: 263–265.
- Sokal R, Rohlf F: Correlation; in Company WHFa (ed): *Biometry*. NewYork, 2003, pp 555–608.
- Purcell S, Neale B, Todd-Brown K *et al*: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–575.
- Marchini J, Cutler D, Patterson N *et al*: A comparison of phasing algorithms for trios and unrelated individuals. *Am J Hum Genet* 2006; **78**: 437–450.
- Terwilliger JD, Hiekkalinna T: An utter refutation of the 'Fundamental Theorem of the HapMap'. *Eur J Hum Genet* 2006; **14**: 426–437.
- Gu CC, Yu K, Rao DC: Characterization of LD structures and the utility of HapMap in genetic association studies. *AdvGenet* 2008; **60**: 407–435.
- Pardo LM, MacKay I, Oostra B, van Duijn CM, Aulchenko YS: The effect of genetic drift in a young genetically isolated population. *Ann Hum Genet* 2005; **69**: 288–295.
- Gu CC, Yu K, Boerwinkle E: Measuring marker information content by the ambiguity of block boundaries observed in dense SNP data. *Ann Hum Genet* 2007; **71**: 127–140.
- Teare MD, Dunning AM, Durocher F, Rennart G, Easton DF: Sampling distribution of summary linkage disequilibrium measures. *Ann Hum Genet* 2002; **66**: 223–233.
- Weiss KM, Clark AG: Linkage disequilibrium and the mapping of complex human traits. *Trends Genet* 2002; **18**: 19–24.
- Zeggini E, Rayner W, Morris AP *et al*: An evaluation of HapMap sample size and tagging SNP performance in large-scale empirical and simulated data sets. *Nat Genet* 2005; **37**: 1320–1322.
- Ioannidis JPA: Non-replication and inconsistency in the genome-wide association setting. *Hum Hered* 2007; **64**: 203–213.
- McCarthy MI, Abecasis GR, Cardon LR *et al*: Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 2008; **9**: 356–369.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (<http://www.nature.com/ejhg>)