European Journal of Human Genetics (2008) 16, 1142–1150
© 2008 Macmillan Publishers Limited All rights reserved 1018-4813/08 $30.00
www.nature.com/ejhg

ARTICLE

Evaluation of HapMap data in six populations of European descent

Per E Lundmark[1], Ulrika Liljedahl[1], Dorret I Boomsma[2], Heikki Mannila[3], Nicholas G Martin[4], Aarno Palotie[5], Leena Peltonen[6,7], Markus Perola[6,7], Tim D Spector[8] and Ann-Christine Syvänen*,[1]

[1]*Molecular Medicine, Department of Medical Sciences, Uppsala University Hospital, Uppsala University, Uppsala, Sweden;* [2]*Department of Biological Psychology, Vrije Universiteit, Amsterdam, The Netherlands;* [3]*HIIT Basic Research Unit, Department of Computer Science, University of Helsinki, Helsinki, Finland;* [4]*Genetic Epidemiology Unit, Queensland Institute of Medical Research, Brisbane, Australia;* [5]*Finnish Genome Center, University of Helsinki, Helsinki, Finland;* [6]*Department of Molecular Medicine, National Public Health Institute, Helsinki, Finland;* [7]*Department of Medical Genetics, University of Helsinki, Helsinki, Finland;* [8]*Twin Research and Genetic Epidemiology Unit, St Thomas' Hospital Campus, Kings College London School of Medicine, London, UK*

We studied how well the European CEU samples used in the Haplotype Mapping Project (HapMap) represent five European populations by analyzing nuclear family samples from the Swedish, Finnish, Dutch, British and Australian (European ancestry) populations. The number of samples from each population (about 30 parent-offspring trios) was similar to that in the HapMap sample sets. A panel of 186 single nucleotide polymorphisms (SNPs) distributed over the 1.5 Mb region of the *GRID2* gene on chromosome 4 was genotyped. The genotype data were compared pair-wise between the HapMap sample and the other population samples. Principal component analysis (PCA) was used to cluster the data from different populations with respect to allele frequencies and to define the markers responsible for observed variance. The only sample with detectable differences in allele frequencies was that from Kuusamo, Finland. This sample also separated from the others, including the other Finnish sample, in the PCA analysis. A set of tagSNPs was defined based on the HapMap data and applied to the samples. The tagSNPs were found to capture the genetic variation in the analyzed region at $r^2 > 0.8$ at levels ranging from 95% in the Kuusamo sample to 87% in the Australian sample. To capture the maximal genetic variation in the region, the Kuusamo, HapMap and Australian samples required 58, 63 and 73 native tagSNPs, respectively. The HapMap CEU sample represents the European samples well for tagSNP selection, with some caution regarding estimation of allele frequencies in the Finnish Kuusamo sample, and a slight reduction in tagging efficiency in the Australian sample.

European Journal of Human Genetics (2008) **16,** 1142–1150; doi:10.1038/ejhg.2008.77; published online 9 April 2008

Keywords: linkage disequilibrium; tagSNPs; haplotype structure; *GRID2*; principal component analysis

Introduction

The Haplotype Mapping Project (HapMap) has created a significant resource for human genetic studies by mapping the patterns of linkage disequilibrium (LD) in the human genome, and by creating a resource for LD-based marker selection for genome-wide association studies.[1] In the HapMap project, samples from representative European,

*Correspondence: Professor A-C Syvänen, Molecular Medicine, Department of Medical Sciences, Uppsala University Hospital, Entrance 70, 3rd floor, Research Department 2, Uppsala S-751 85, Sweden.
Tel: +46 18 611 29 59; Fax: +46 18 55 36 01;
E-mail: ann-christine.syvanen@medsci.uu.se

Received 2 October 2007; revised 6 March 2008; accepted 6 March 2008; published online 9 April 2008

African and Asian populations were genotyped, and the genotype data were published online for free use by the scientific community. The latest HapMap release (no. 22 of March 2007), includes more than 3.8 million validated single nucleotide polymorphisms (SNPs). An important question for the application of the results from the HapMap project has been how representative the populations analyzed would be for other populations of interest worldwide. This question has been addressed in a number of studies. LD patterns and tagSNP transferability have been compared in samples from diverse geographical regions of the world,[2–4] or investigated more closely in one or more specific populations.[5,6] Gu *et al*[7] recently reported an investigation of 517 kb of the human genome distributed over 10 loci and found that popular definitions of haplotype blocks were highly variable between samples from different parts of the world, while tagSNPs were better transferable between populations. De Bakker *et al*[3] investigated 25 regions of approximately 100 kb length. They found that for samples from individuals self-described as 'white' from Hawaii and individuals from the Botnia region in Finland, loss of coverage in tagging when using HapMap-defined tagSNP sets was no larger than sampling variation within the HapMap samples. A less clear result was obtained when Mueller *et al*[5] analyzed four regions spanning 749 kb in total in populations across Europe from Italy to Estonia. Two of the regions were well represented by the HapMap CEU trios, whereas the data from the two other regions were more difficult to interpret.

In our study we compared how representative the HapMap data were for six samples from five European populations. We found that tagSNPs defined in the HapMap CEU sample represented our European samples well, with the proportion of tagged SNPs ranging from 87% in the Australian sample to 95% in the Kuusamo sample. Allele frequencies were fairly similar in the samples, with the exception of the Finnish sample from the Kuusamo region, where we observed overall differences in allele frequencies compared to the other populations.

Materials and methods

Study populations/DNA samples

Nuclear family samples with one child and two parents from Australia, Finland (two samples), the Netherlands, United Kingdom and Sweden were genotyped. The Australian trios were selected from an adolescent twin family study.[8] The Dutch sample consisted of families who are registered with the Netherlands Twin Register.[9] The UK adult twin registry (http://www.TwinsUK.ac.uk) consists of volunteer Caucasian twins from all over the United Kingdom and born in the United Kingdom. This is shown to be broadly representative of the UK population.[10] The Finnish samples were from the general Finnish population, derived largely from early settlement Finland, and a sample from Kuusamo, a small subisolate within late settlement Finland.[11] The Swedish samples were from the Uppsala Family Study[12] and originate from the central region of Sweden. The samples were selected at random from the larger population cohorts, but with the practical limitation that samples were avoided where the amount of extracted DNA available for study was limited. The DNA samples had been extracted using standard methods, for Australia, using the salt extraction method.[13] Eleven of the Dutch DNA samples were amplified by multiple displacement amplification using the GenomiPhi DNA amplification kit (GE Healthcare, Uppsala, Sweden) prior to genotyping. The genotyped samples are summarized in Table 1.

Genomic region and SNP selection

The genomic region considered in this study was a 1.47 Mbp region on chromosome 4q22, containing the glutamate receptor, ionotropic, delta 2 gene (*GRID2*). Out of the available SNPs genotyped in the HapMap project, 197 SNPs with an average spacing of 8 kb were genotyped. One of these SNPs was located in the coding regions of *GRID2*, but was synonymous, and the remaining SNPs were located outside exon-coding regions of *GRID2*. The allele frequency distributions for the populations are graphed in Figure 1.

Genotyping

Primers for PCR and minisequencing (single-base primer extension) were designed using the Autoprimer software (http://www.autoprimer.com, Beckman Coulter Inc., Fullerton, CA, USA). The SNPs were genotyped using the GenomeLab SNPstream system[14] using reagents provided by the manufacturer (Beckman Coulter). In brief, the genotyping protocol involved 12-plex PCR amplification and genotyping of SNPs with the same nucleotide variation. PCR products were cleaned up using exonuclease I and shrimp alkaline phosphatase, and genotyped using 12-plex minisequencing with fluorescent terminating nucleotides, followed by capture of the extension products by hybridization to complementary tag-oligonucleotides immobilized in 384-well microtiter plate wells. Incorporated fluorescence was detected using a CCD camera of the SNPstream instrument and the genotypes were called by cluster analysis using the GetGenos software (Beckman Coulter). Of the 197 SNPs selected, 11 failed in genotyping

Table 1 The genotyped samples in the study

Origin of DNA samples	*Number of samples*	*Number of trios*
Australia	99	33
Finland, Kuusamo	93	16
Finland	92	30
Netherlands	72	24
UK	90	30
Sweden	96	32

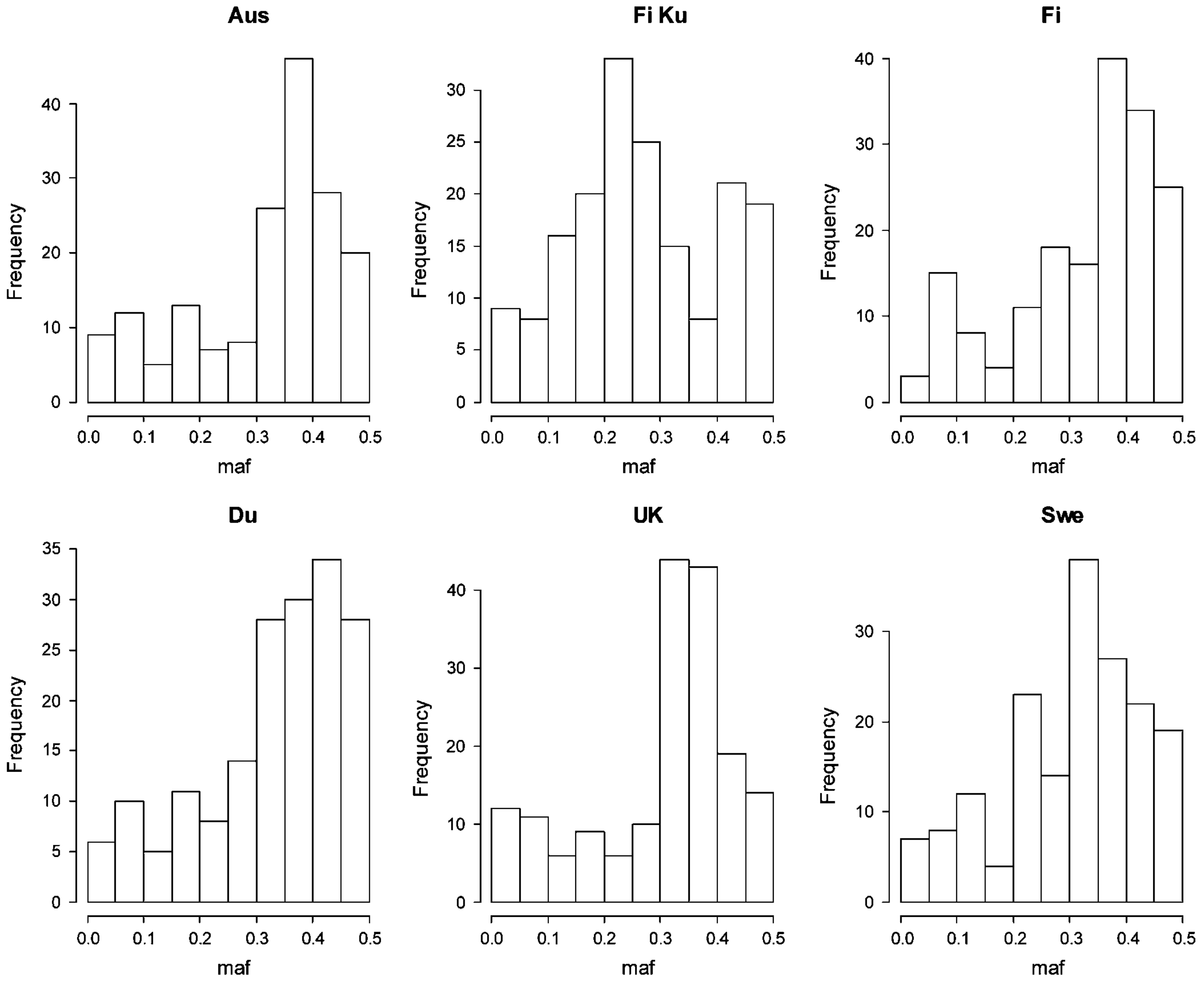

Figure 1 Minor allele frequency distributions for the genotyped population samples.

due to technical reasons and 13 were monomorphic in the whole sample set. Thus, 173 SNPs were included in the subsequent analyses (Supplementary Table 1). Checks for Hardy–Weinberg equilibrium and Mendelian inheritance were used as quality controls using an in-house developed software and database system (SNP Quality Analysis Tool). The average sample success rate was 96.4% and the reproducibility was 99.98% according to duplicate determination of 21% of the genotypes in two independent experiments.

Data analysis

The Haploview[15] v3.32. software was used for D' and r^2 calculations and for analysis of the LD patterns in each of the genotyped populations. The tagger[16] implementation in Haploview was used for tagSNP selection using pair-wise tagging and an r^2 cutoff value of 0.8. TagSNPs were defined in the HapMap CEU data set and their performance was assessed in the six study population samples. To test for differences in allele frequencies, Fisher's exact test was used to compare the genotype data from the parents in the trio families using the R software package version 2.2.1 (www.r-project.org). The population assignment of each individual was altered in a permutation test with 7000 permutations to determine a global 5% error level for the allele frequency comparisons in the HapMap against all others comparison.

Principal component analysis (PCA) was used to project SNP allele frequencies in the population samples into two dimensions to summarize similarities and differences. In the PCA, populations were used as observations and the SNP allele frequencies were used as variables. The data were centered but not scaled.

The Gabriel *et al* blocking algorithm[17] was used to annotate areas of high D' across the 1.47 Mbp GRID2 genomic region. Four areas with different patterns of r^2 in the blocks defined by this algorithm were selected for haplotype reconstruction and comparison of haplotype distributions in the GRID2 region. The PHASE version 2.2.1 software was used for haplotype reconstruction and testing for differences in the haplotype distributions between the HapMap sample and the other population samples.[18,19] A thousand permutations were used in the PHASE software to calculate the *P*-values.

Results

To compare the LD pattern, haplotype structure and tagSNP efficiency in a 1.47 Mbp genomic region containing the *GRID2* gene between five European populations and the HapMap CEU data, we analyzed 173 SNPs in this region. Figure 2 is a schematic view of the GRID2 genomic region, the positions of the SNPs and the four selected regions with high LD used for haplotype reconstruction. The investigated genomic region contains both regions with a strong block-like LD structure and more fragmented areas (Supplementary Figures 1 and 2).

Allele frequencies

The minor allele frequency (MAF) distributions in each of the genotyped samples are graphed in Figure 1. When calculated across all six populations, the first quartile of the MAFs was 0.23, the median and mean were 0.34 and 0.31, respectively, and the third quartile of the MAFs was 0.41. The minimum and maximum values for the MAFs were 0.018 and 0.49, respectively. To determine whether the allele frequencies of the SNPs differed between the population samples, the allele frequencies of the parents in the trio families were subjected to pair-wise comparisons between populations and the HapMap CEU data.

Permutation testing indicated that the global 5% error level for a single comparison was $P<1.30\times10^{-4}$, when the allele frequencies in the HapMap sample were tested against the allele frequencies in the other six population samples. This threshold was only reached for the SNPs rs1912718 and rs6854968 in the comparisons between the HapMap and the Finnish Kuusamo samples. The allele frequencies for the SNP rs1912718 were 0.48 in the HapMap sample and 0.76 in the Kuusamo sample. The corresponding numbers for the SNP rs6854968 were 0.34 and 0.62, respectively. An all-against-all comparison was also performed between the populations, including the HapMap sample, to search for overall differences in allele frequencies between populations. The numbers of differences between populations at a level of $P<0.05$ are tabulated in Table 2. All samples except that from Kuusamo, Finland had a similar number of allele frequency differences. The Kuusamo sample had a larger number of allele frequency differences when compared to the other samples, and for every population, the comparison with Kuusamo was the one that yielded the highest number of differences. Given the number of SNPs tested, the expected number of random results below 5% would be between 8 and 9 for each combination of samples. The sample from Kuusamo was the only one where these numbers differed clearly from those expected. For all comparisons with Kuusamo, the number of differences was more than threefold larger than expected by chance.

PCA was used to project the allele frequencies of the 173 SNPs into two dimensions. This can be visualized as positioning a plane (the first and second component) in 173-dimensional space to describe as much of the variance in the data as possible. When projected onto the plane, the coordinates of the data points are converted to the new two-dimensional coordinates. Again, the Kuusamo sample separated to the largest extent from the rest of the populations along the first principal component, that is the *x* axis in Figure 3, while the other populations clustered together. The second principal component, that is the *y* axis in Figure 3, separates the Finnish sample from the other populations, indicating that in part a different set of SNPs were responsible for this separation than for that of the Kuusamo sample. To determine which of the SNPs were responsible for the separation of the samples, the loadings plot (Supplementary Figure 3) was used. This plot shows which of the original variables (SNP allele frequencies) influenced the principal components. For example, the SNPs plotted to the extremes in the *x* axis in the loadings plot had the strongest influence on the *x* axis (first component) in the score plot of the populations. In this case the SNPs 104–110 (Supplementary Table 1) together with SNP 46 had the strongest effect on the position along

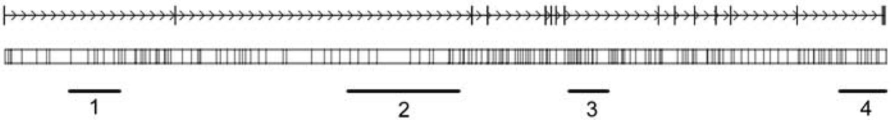

Figure 2 A schematic illustration of the chromosomal region containing the *GRID2* gene on chromosome 4 (top), the single nucleotide polymorphism (SNP) positions in this region (middle) and the example regions 1–4 used for haplotype reconstruction (bottom).

Table 2 The number of pair-wise allele frequency differences between populations at $P<0.05$

	Australia	*Finland Kuusamo*	*Finland*	*Dutch*	*UK*	*Sweden*	*HapMap*
Australia		50	8	1	8	10	6
Finland, Kuusamo	50		39	76	46	32	55
Finland	8	39		12	11	11	10
Dutch	1	76	12		5	18	2
UK	8	46	11	5		16	5
Sweden	10	32	11	18	16		5
HapMap	6	55	10	2	5	5	
Average	13.8	49.6	15.2	19.0	15.2	15.3	13.8

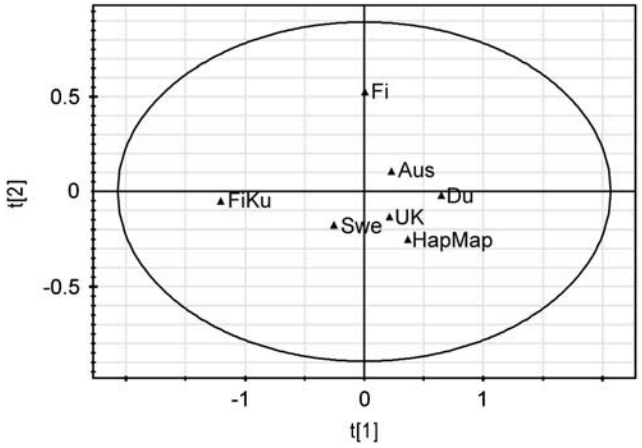

Figure 3 Principal component score plot of the first two principal components of the allele frequencies in the populations. The plot is generated by positioning a plane in the cloud of data points representing the single nucleotide polymorphism (SNP) allele frequencies so that the maximum variation is captured when the data points are projected onto the plane. In effect, the plot gives a reduction in dimensions from 173 representing the number of allele frequencies from different SNPs to two, while maintaining as much information as possible.

the first component in the samples. The SNPs 104–110 are located at the end of a region of high LD adjacent to a recombination hot spot described by the HapMap project, and the position of SNP 46 is in a region of low LD between two regions with high D' values.

Linkage disequilibrium and tagSNP efficiency

The average levels of LD were similar between the population samples, with the highest average D' value for the Kuusamo sample and the highest average r^2 value in the UK sample (Table 3). The patterns of D' and r^2 across the region were similar in all the samples, with smaller differences in the general level of LD in some regions (Supplementary Figure 1 and 2, HapMap D' and r^2 plots). To evaluate the suitability of using a set of tagSNPs defined based on the HapMap data in our study populations, a set of tagSNPs was selected using a pair-wise r^2 cutoff value of 0.8, which resulted in a set of 63 SNPs. This set of tagSNPs was then applied to the other populations, and its ability to capture the genetic variation was determined. About 90% of the SNPs in the five European populations were represented at $r^2>0.8$ using the HapMap tagSNP set. In the Kuusamo sample we observed the best capture of variation, with 95% of markers represented. The lowest efficiency of the tagSNPs was observed in the Australian sample, where 87% of the SNPs were captured (Table 4). A similar pattern was seen for the number of native tagSNPs defined specifically in each of the populations to capture the variation of all other SNPs. The Kuusamo sample required only 58 SNPs to tag all 173 SNPs, whereas in the Australian sample, 73 SNPs were required.

The distribution of r^2 between SNPs and the HapMap tagSNPs in Figure 4 illustrates the slight differences between populations. In the Australian sample with a slightly lower tagging efficiency, this difference was due to a wider distribution of values bringing a larger number of markers below the $r^2=0.8$ threshold (22 SNPs for the Australian sample *versus* 9 for the Finnish Kuusamo sample). No increase in SNPs with very low LD to the HapMap tagSNPs was observed in the less effectively tagged samples.

Haplotypes

To illustrate the range of variation in haplotype frequencies between different population samples, four regions defined by the Garbriel *et al* algorithm[17] within the 1.47 Mbp region were selected for haplotype reconstruction. They were all regions of high D' to allow the haplotype reconstruction software to perform optimally to increase the confidence of detecting real differences in haplotype frequencies, rather than observing shortcomings of the software used.[19] In Figure 5, the haplotypes with frequencies > 1% in the study populations are shown in a separate panel for each region. In most cases differences in haplotype frequencies were smaller than 10% between the HapMap and other populations, but in some cases they differed as much as 20% (Figure 5).

Differences in the haplotype distribution between the HapMap sample and the other six population samples were tested pair wise using the case–control functionality in PHASE.[18,19] This software compares the haplotype

Table 3 Summary statistics for pair-wise r^2- and D′-values between SNPs in the full sets of SNPs and between SNPs within 50 kb of each other

	HapMap	*Australia*	*Finland, Kuusamo*	*Finland*	*Dutch*	*UK*	*Sweden*
Average r^2	0.11	0.10	0.12	0.12	0.11	0.12	0.10
Median r^2 (IQR)	0.023 (0.09)	0,02 (0.08)	0.04 (0.12)	0.04 (0.10)	0.03 (0.09)	0.02 (0.1)	0.02 (0.1)
Average r^2 <50 kb	0.47	0.44	0.48	0.46	0.46	0.50	0.46
Median r^2 <50 kb (IQR)	0.44 (0.72)	0.38 (0.74)	0.47 (0.75)	0.45 (0.73)	0.42 (0.75)	0.51 (0.82)	0.40 (0.71)
Average *D*′	0.41	0.39	0.53	0.46	0.42	0.44	0.40
Median *D*′ (IQR)	0.33 (0.54)	0.3 (0.46)	0.47 (0.72)	0.37 (0.56)	0.34 (0.52)	0.32 (0.65)	0.31 (0.49)
Average *D*′ <50 kb	0.86	0.84	0.89	0.87	0.85	0.86	0.85
Median *D*′ <50 kb (IQR)	1 (0.20)	0.97 (0.25)	1 (0.09)	1 (0.13)	1 (0.21)	1 (0.19)	1 (0.22)

Abbreviation: IQR, interquartile range.

Table 4 Performance of HapMap tagSNPs in the HapMap samples and in the study population samples

	Hap-Map	*Australia*	*Finland, Kuusamo*	*Finland*	*Dutch*	*UK*	*Sweden*
Average r^2 (using HapMap tagSNPs)	0.95	0.93	0.95	0.93	0.94	0.95	0.93
Minimum r^2 (using HapMap tags)	0.81	0.28	0.27	0.25	0.24	0.27	0.22
% of SNPs with r^2>0.8 to HapMap tags	100	87	95	91	90	92	90
Native SNPs needed to tag at r^2>0.8	63	73	58	67	70	61	68

Abbreviations: IQR, interquartile range; SNP, single nucleotide polymorphism.

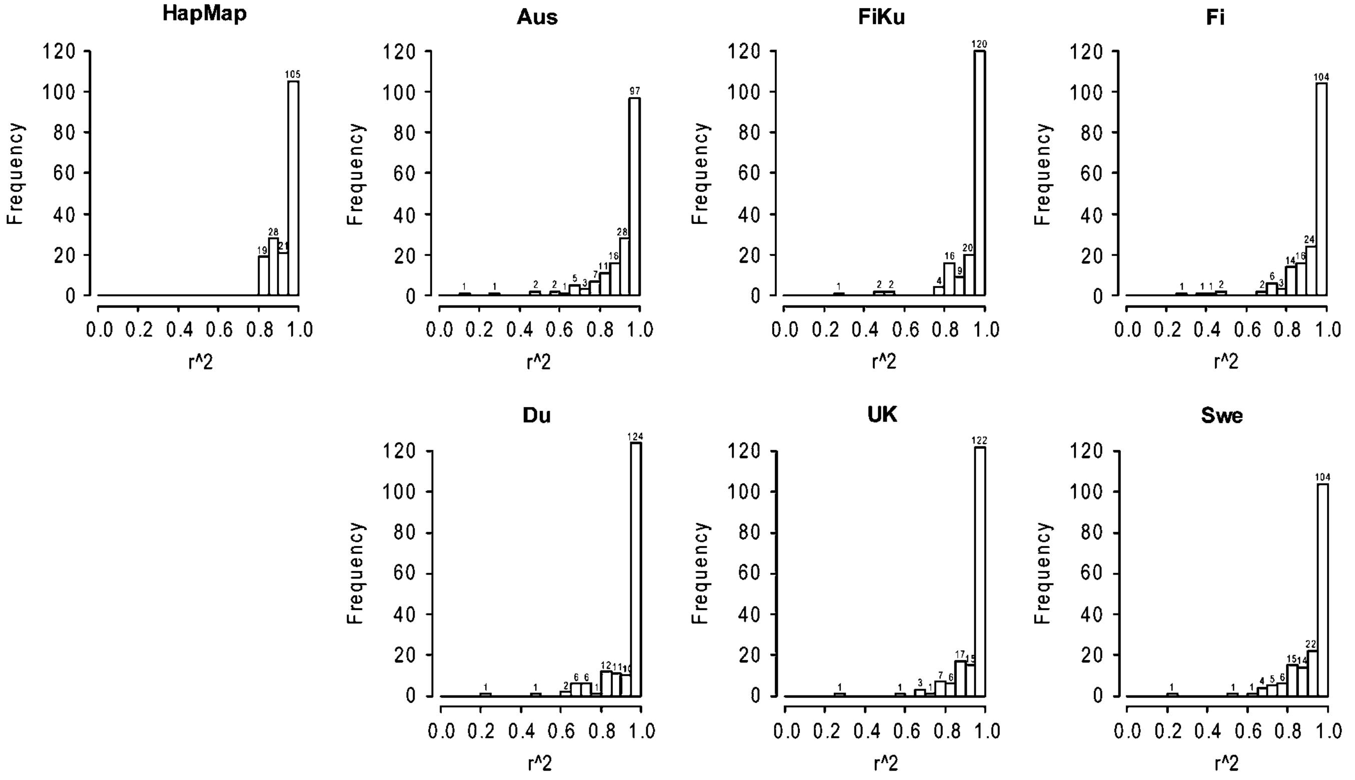

Figure 4 Distributions of pair-wise r^2-values between best HapMap tagSNPs and all single nucleotide polymorphisms (SNPs) in the genotyped region.

distributions in two population samples, and also takes into account the degree of similarity between the observed haplotypes.

The null hypothesis is that the two haplotype distributions in the tested populations are random selections from the same set of haplotype frequencies. This is compared to

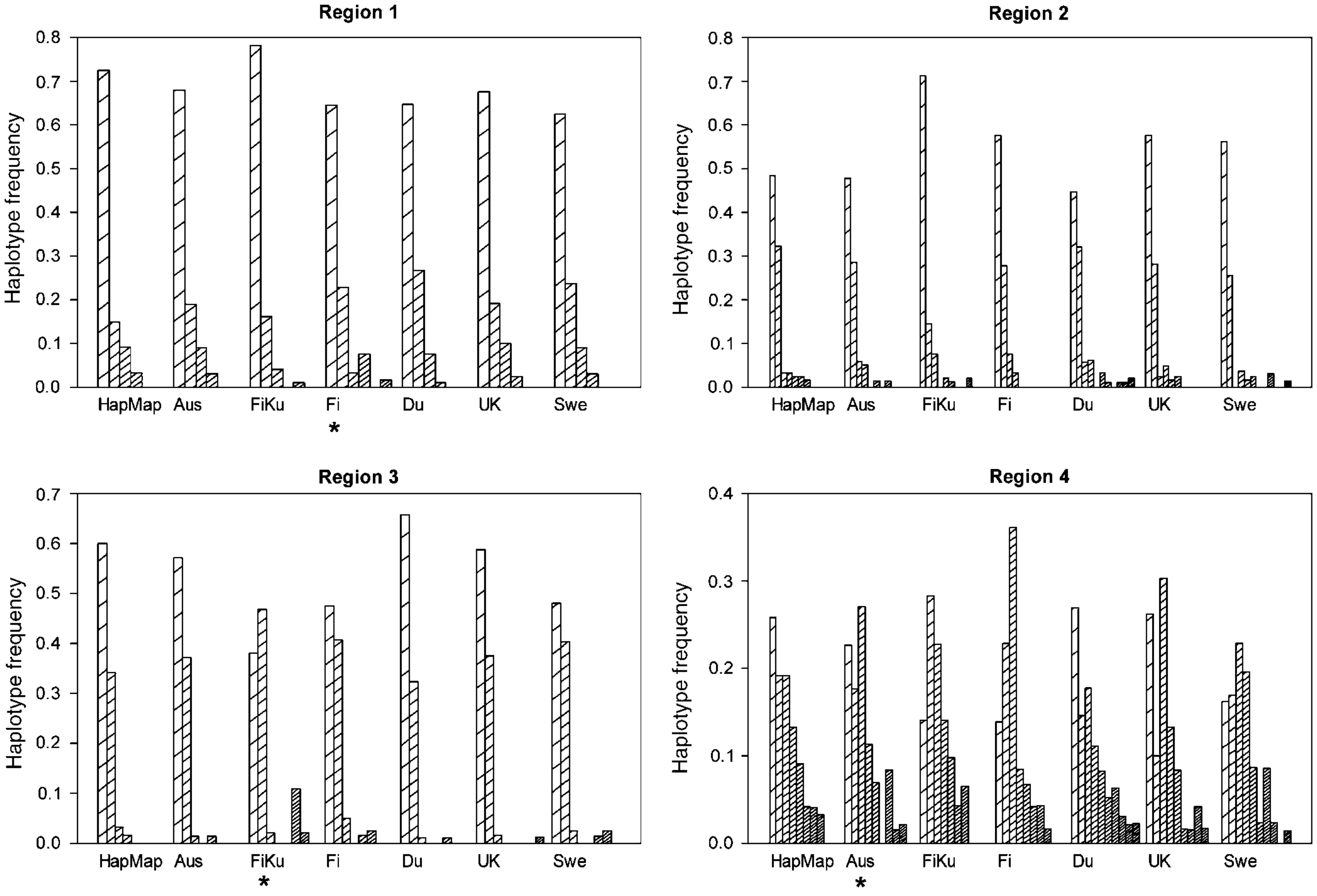

Figure 5 Haplotype frequencies>0.01 in the example regions, ordered by their frequency in the HapMap sample. Haplotype distributions differing from the HapMap data at $P<0.05$ are marked with an asterisk. Region 1, single nucleotide polymorphisms (SNPs) rs9993173–rs7683035;region 2, SNPs rs1456359–rs6811974; region 3, SNPs rs1433654–rs1433648; region 4, SNPs rs4692986–rs9993013.

the alternative hypothesis, that the second population is more similar to itself, than the first population, which in our case is the HapMap data. Three of these comparisons were significant at a $P<0.05$ level. These were the comparisons of the HapMap sample with the Australian sample in region 4 ($P=0.039$), the Kuusamo sample in region 3 ($P=0.002$) and the Finnish sample from the general population in region 1 ($P=0.032$). The haplotype distributions in the population samples that differed at $P<0.05$ from the HapMap sample are marked with an asterisk in Figure 5. Only the result in the Kuusamo sample would be significant after a Bonferroni correction for the multiple (6×4) tests with a critical P-value of 0.0021. This result appears to have been mainly driven by a haplotype present at a frequency of 0.11 in the Kuusamo sample, while absent in the HapMap sample. This haplotype also contained several sequence differences compared to the haplotypes present in the HapMap sample.

Discussion

Genome-wide association studies with SNP markers have recently become a standard tool in the search for genes underlying complex diseases.[20] The data from the HapMap project have served as the key resource for selecting the SNPs for the panels used in genome-wide genotyping studies.[21,22] Information on LD is also required to use a tagSNP strategy for custom design of SNP panels that have maximal power to capture the genetic variation in a specific genomic region of interest. The aim of our study was to investigate how well the HapMap European data represent five specific European populations, from which SNP genotype data will be subjected to combined/pooled association analysis. Our results from a careful investigation of SNPs in a 1.47 Mbp genomic region using a set of SNPs located at an average spacing of 8 kb are reassuring.

We observed no indications of differences in allele frequencies between the studied populations and the HapMap data, with the single exception of the Kuusamo sample. This is somewhat in contrast to the result by Willer *et al*,[23] who found evidence for population differences in SNPs allele frequencies between the HapMap sample and their Finnish sample. They investigated a 17.9 Mb region on chromosome 14 in a more general Finnish population. However, their sample size was larger than ours, and thus smaller differences would be detectable than in our study. The Kuusamo isolate has an exceptionally well-documented history, and was founded by just 78 settlers 310 years ago. Extended LD has been demonstrated in the Kuusamo isolate, using microsatellite markers.[24] Our study was designed to include a similar number of samples as the HapMap project. Permutation testing indicated that only two of the comparisons were significantly different at a global $P<0.05$, as a result of the many tests involved in testing the HapMap samples against the six other population samples for each SNP, as well as the limited sample sizes. When examining the marker set as a whole, the Kuusamo sample had more differences in SNP allele frequencies than expected by chance. Our conclusion is that allele frequency estimates based on the HapMap data are as good as can be expected from the sample sizes in all populations, with the exception of the Kuusamo sample where extra caution would be advisable based on the observed differences.

The performance tagSNP defined by the HapMap project in the studied European populations was good. Between 87 (Australia) and 95% (Finland, Kuusamo) of the SNPs were captured at $r^2>0.8$. This is slightly higher than the 89% reported by de Bakker *et al*[3] as the sampling variation of the CEPH parent-offspring trios from Utah (HapMap CEU) and the 78 and 83% for the self-described white Hawaiian and the Finnish, Botnia samples, respectively. One limitation in our study is that we do not sample other SNPs than those genotyped by the HapMap study, where especially rarer SNP variants could be less well represented. Despite this limitation we conclude that the HapMap CEU sample represents our samples of interest well and is of value for tagSNP selection. Although we observed differences in allele frequencies between the Finnish Kuusamo sample and the other populations, the Kuusamo population was the one best captured by the HapMap tagSNP set. This surprising finding seems to be the result of the fact that the Kuusamo sample is the easiest to tag, as only 58 native Kuusamo tagSNPs were needed. For comparison, 63 native HapMap tagSNPs were required to tag the HapMap set, which was applied to the other populations. These results are similar to those recently reported by Service *et al.*[25] They investigated the performance of HapMap tagSNPs in several population isolates, including the Kuusamo isolate, which was also included in our study presented here. Despite factors such as genetic drift that could have a strong effect in population isolates, they found that HapMap tagSNPs worked well in several population isolates, including the one from Kuusamo, Finland, for 2215 markers on chromosome 22. In their panel 84% of non-tags were captured at $r^2>0.8$.[25] Our higher representation of Kuusamo SNPs by HapMap tagSNPs could be an effect of sampling as well as a larger proportion of more common SNPs in the panel. Overall, our work agrees with other previous studies such as the one by Willer *et al*[23] that concluded that HapMap tagSNPs represented their Finnish sample well for tagSNP selection despite indications of allele frequency differences, and the study by Stankovich *et al*[26] that analyzed regions on chromosomes 6q and 10q in Australian multiple sclerosis cases, and found that their Australian samples were similarly well represented.

Acknowledgements

Genotyping was performed at the SNP technology platform in Uppsala (www.genotyping.se). The study was supported by the EC-funded GenomEUtwin project (QLG2-CT-2002-01254), the Swedish Research Council and the Knut and Alice Wallenberg Foundation. We thank Megan Campbell for preparation of the Australian samples.

References

1 The International HapMap Consortium: A haplotype map of the human genome. *Nature* 2005; **437**: 1299–1320.
2 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.
3 de Bakker PI, Burtt NP, Graham RR *et al*: Transferability of tag SNPs in genetic association studies in multiple populations. *Nat Genet* 2006; **38**: 1298–1303.
4 Sawyer SL, Mukherjee N, Pakstis AJ *et al*: Linkage disequilibrium patterns vary substantially among populations. *Eur J Hum Genet* 2005; **13**: 677–686.
5 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.
6 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.
7 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.
8 Zhu G, Montgomery GW, James MR *et al*: A genome-wide scan for naevus count: linkage to CDKN2A and to other chromosome regions. *Eur J Hum Genet* 2007; **15**: 94–102.
9 Boomsma DI, de Geus EJ, Vink JM *et al*: Netherlands Twin Register: from twins to twin families. *Twin Res Hum Genet* 2006; **9**: 849–857.
10 Spector TD, Williams FM: The UK Adult Twin Registry (TwinsUK). *Twin Res Hum Genet* 2006; **9**: 899–906.
11 Service S, DeYoung J, Karayiorgou M *et al*: Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. *Nat Genet* 2006; **38**: 556–560.
12 Leon DA, Koupil I, Mann V *et al*: Fetal, developmental, and parental influences on childhood systolic blood pressure in 600 sib pairs: the Uppsala Family study. *Circulation* 2005; **112**: 3478–3485.
13 Miller SA, Dykes DD, Polesky HF: A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res* 1988; **16**: 1215.
14 Bell PA, Chaturvedi S, Gelfand CA *et al*: SNPstream UHT: ultra-high throughput SNP genotyping for pharmacogenomics

and drug discovery. *Biotechniques* 2002; **32** (Suppl): 70–72,74, 76–77.

15 Barrett JC, Fry B, Maller J, Daly MJ: Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005; **21**: 263–265.

16 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.

17 Gabriel SB, Schaffner SF, Nguyen H *et al*: The structure of haplotype blocks in the human genome. *Science* 2002; **296**: 2225–2229.

18 Stephens M, Scheet P: Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet* 2005; **76**: 449–462.

19 Stephens M, Smith NJ, Donnelly P: A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 2001; **68**: 978–989.

20 The Wellcome Trust Case Control Consortium: Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature* 2007; **447**: 661–678.

21 Duerr RH, Taylor KD, Brant SR *et al*: A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* 2006; **314**: 1461–1463.

22 Sladek R, Rocheleau G, Rung J *et al*: A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* 2007; **445**: 881–885.

23 Willer CJ, Scott LJ, Bonnycastle LL *et al*: Tag SNP selection for Finnish individuals based on the CEPH Utah HapMap database. *Genet Epidemiol* 2006; **30**: 180–190.

24 Varilo T: The age of the mutations in the Finnish disease heritage a genealogical and linkage disequilibrium study. PhD thesis, University of Helsinki, Finland; on-line publications, Helsinki, 1999, pp 1–98.

25 Service S, Sabatti C, Freimer N: Tag SNPs chosen from HapMap perform well in several population isolates. *Genet Epidemiol* 2007; **31**: 189–194.

26 Stankovich J, Cox CJ, Tan RB *et al*: On the utility of data from the International HapMap Project for Australian association studies. *Hum Genet* 2006; **119**: 220–222.

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