Am. J. Hum. Genet. 76:387–398, 2005

Linkage Disequilibrium Patterns and tagSNP Transferability among European Populations

Jakob C. Mueller,[1,*] Elin Lõhmussaar,[1,3,*] Reedik Mägi,[3] Maido Remm,[3] Thomas Bettecken,[1] Peter Lichtner,[1] Saskia Biskup,[1] Thomas Illig,[2] Arne Pfeufer,[4] Jan Luedemann,[5] Stefan Schreiber,[6] Peter Pramstaller,[7] Irene Pichler,[7] Giovanni Romeo,[8] Anthony Gaddi,[9] Alessandra Testa,[10] Heinz-Erich Wichmann,[2] Andres Metspalu,[3] and Thomas Meitinger[1,4]

Institutes of [1]Human Genetics and [2]Epidemiology, GSF–National Research Centre for Environment and Health, Neuherberg, Germany; [3]Institute of Molecular and Cell Biology, University of Tartu, Estonian Biocentre, Tartu, Estonia; [4]Institute of Human Genetics, Technical University Munich, Munich; [5]Institute of Clinical Chemistry and Laboratory Medicine, University of Greifswald, Greifswald, Germany; [6]Institute for Clinical Molecular Biology, University Clinic Schleswig-Holstein, Christian-Albrechts-University, Kiel, Germany; [7]Department of Genetic Medicine, European Academy, Bolzano, Italy; [8]Department of Medical Genetics and [9]Center Artherosclerosi Giancarlo Descovich, Department of Clinical Medicine, University of Bologna, Bologna; and [10]Consiglio Nazionale delle Ricerche–Istituto di Biomedicina ed Immunologia Molecolare (CNR-IBIM), National Research Council–Institute of Biomedicine, Clinical Epidemiology and Physiopathology of Renal Diseases and Hypertension, Reggio Calabria, Italy

The pattern of linkage disequilibrium (LD) is critical for association studies, in which disease-causing variants are identified by allelic association with adjacent markers. The aim of this study is to compare the LD patterns in several distinct European populations. We analyzed four genomic regions (in total, 749 kb) containing candidate genes for complex traits. Individuals were genotyped for markers that are evenly distributed at an average spacing of ~2–4 kb in eight population-based samples from ongoing epidemiological studies across Europe. The Centre d'Etude du Polymorphisme Humain (CEPH) trios of the HapMap project were included and were used as a reference population. In general, we observed a conservation of the LD patterns across European samples. Nevertheless, shifts in the positions of the boundaries of high-LD regions can be demonstrated between populations, when assessed by a novel procedure based on bootstrapping. Transferability of LD information among populations was also tested. In two of the analyzed gene regions, sets of tagging single-nucleotide polymorphisms (tagSNPs) selected from the HapMap CEPH trios performed surprisingly well in all local European samples. However, significant variation in the other two gene regions predicts a restricted applicability of CEPH-derived tagging markers. Simulations based on our data set show the extent to which further gain in tagSNP efficiency and transferability can be achieved by increased SNP density.

Introduction

The efficiency of both candidate-gene and whole-genome approaches to identifying genetic loci associated with disease phenotypes relies on the minimization of SNP markers genotyped in a given population. For such a mapping approach, the selection process of markers to be genotyped is crucial (Chapman et al. 2003; Wang and Todd 2003). The observation that a significant fraction of the human genome is organized into a series of high–linkage disequilibrium (LD) regions that are separated by short segments in very low LD has led to the development of a number of algorithms that can be used to select informative markers for association studies (Cardon and Abecasis 2003). In Caucasians, approximately one-third to one-half of chromosomes are structured as high-LD regions, varying in length from a few kb to >300 kb (Gabriel et al. 2002; Phillips et al. 2003; Wall and Pritchard 2003; Ke et al. 2004). All marker-selection algorithms are based on the assumption that the complete set of sequence variants within a region of high background LD bears redundant information and can be significantly reduced to a selected subset of tagging markers. These markers can tag either neighboring markers or a set of common haplotypes within an LD block. There is an ongoing debate as to which tagging algorithm should be used, but little is known about the choice of reference populations to which such algorithms should be applied.

It has been suggested that the populations genotyped in the HapMap project may serve as reference populations for the selection of tagging markers in association studies (International HapMap Consortium 2003).

Received July 29, 2004; accepted for publication December 8, 2004; electronically published January 6, 2005.

Address for correspondence and reprints: Dr. Thomas Meitinger, Institute for Human Genetics, GSF–National Research Centre for Environment and Health, Ingolstädter Landstrasse 1, D-85764 Neuherberg, Germany. E-mail: meitinger@gsf.de

* These authors contributed equally to this work.

0002-9297/2005/7603-0003$15.00

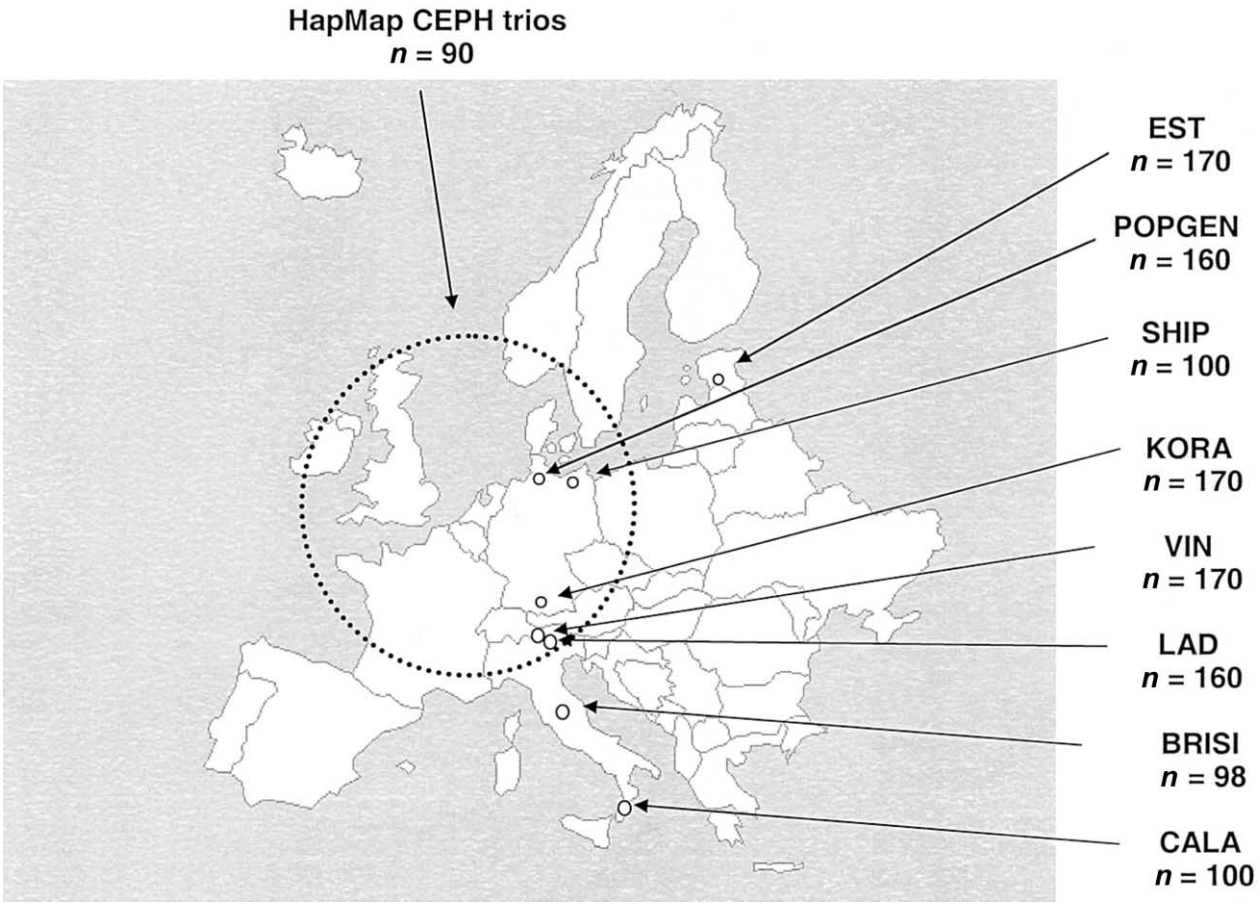


Figure 1 Study populations and sample sizes (*n*)

In its first round, the HapMap project aims to genotype 600,000 SNPs at an average distance of 5 kb across the whole genome in four populations with African, Asian, and European ancestry (see HapMap Homepage). The European patterns are represented by 30 trios from a U.S. (Utah) population of northern and western European ancestry (CEPH sample [Dausset et al. 1990]).

As stated by the International HapMap Consortium (2003), the general applicability of the HapMap data has to be confirmed by samples from several local populations. Our study aims to describe the SNP allelic variation within candidate-gene regions in eight local European populations selected along a line from north to south. All samples represent population-based samples of ongoing epidemiological collections. The dense marker spacing of 2–4 kb over four autosomal regions (total size 749 kb) and a novel robust method to assess the reliability of LD block boundaries enables us to compare LD block boundaries and LD block content among these European populations. Although there was general agreement in the majority of LD patterns, detectable differences among study populations were found. In the context of association studies, we tested the performance of tagging SNPs (tagSNPs) that were defined in local population samples in comparison with tagSNPs that were defined in the HapMap sample, and we simulated the effect of an increased marker density.

Subjects and Methods

Population Samples

All population samples came from ongoing cross-sectional epidemiological surveys. Figure 1 shows the locations and sample sizes of eight regional population surveys. Samples were chosen randomly from the entire population. A ninth sample, 30 CEPH trios (Coriell Cell Repositories) used in the HapMap project, represents an emigrant population of northern and western European origin. The 170 individuals of Estonian ethnicity (EST) represent a random selection from ~1.3 million Estonian inhabitants, excluding Russians. Randomly selected samples for the northern German population came from two epidemiological surveys, Study on Health in Pom-

erania (SHIP) (regional population size 212,000) and POPGEN (collected in Schleswig-Holstein [population size 1.15 million]) (see popgen Web site). The KORA samples were collected as part of a population-based, epidemiological project, KORA S2000 (Cooperative Health Research in the Region of Augsburg), and represent an urban region in the southern part of Germany with 610,000 inhabitants. Two Alpine populations were sampled: inhabitants of Vinschgau (VIN) in south Tyrol (population size 34,300), and members of the Ladin-speaking community (LAD) of Grödnertal and Gadertal (population size 16,800). The Brisighella sample (BRISI) represents a small town with 9,000 inhabitants from the region of Emilia-Romagna, Italy. A sample from Calabria (CALA) was sampled from a catchment area with 560,000 inhabitants. The sex ratio of all samples was ~0.5, except for that of CALA, with 70 males and 30 females. Mean age was 55 years for all population samples except CALA and EST (mean 30 years). Prior to collection, we obtained approval from the relevant ethical committees/institutional review boards and informed consent from all participating subjects. When necessary, approval from data privacy oversight committees was obtained.

SNP Selection and Genotyping

We selected four genomic regions, all containing candidate genes for different complex diseases. For each region, SNPs were evenly selected, covering the candidate gene and 76–174 kb of the upstream and downstream flanking regions (table 1). All information about the selected SNPs was extracted from the public dbSNP database. Genotyping of SNPs was achieved by primer extension of multiplex PCR products, with detection of the allele-specific extension products by matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF [Sequenom]) mass spectroscopy. The frequencies of genotypes from successfully typed SNPs (average call rate, 98%) were in Hardy-Weinberg equilibrium. The genotype data can be downloaded from our project Web site (see GSF European LD Pattern Project Web site).

Statistics and LD-Pattern Analyses

Population differentiation was tested by permutation tests (10,000 permutations) based on *F* statistics, by use of the software package ARLEQUIN. F_{ST} values were calculated on three levels: for each marker separately, for each gene region separately, and for all four gene regions combined. F_{ST} values based on haplotype frequencies in each block were also tested. The standard expectation-maximization algorithm was used to estimate the haplotype frequencies.

To compare haplotype block boundaries among populations, it is critical to apply a relatively robust method for the definition of haplotype blocks on a constant set of common markers (Cardon and Abecasis 2003; Schwartz et al. 2003). We developed a simple bootstrap approach based on the standard algorithm of Gabriel et al. (2002), which invokes confidence bounds of pairwise D' to define sequences of markers with little evidence of historical recombination. Bootstrapping, in the present context, means resampling the individual multilocus genotypes of a given population with replacement. The frequencies of block boundary positions across all 100 bootstrap runs represent confidence estimates for block borders. We plot boundary frequencies for the start and end of blocks separately, to be able to track individual blocks. In addition, we allowed blocks to overlap each other, which gives a more natural framework. Because most block definitions define haplotype blocks by block-internal characteristics, neighboring blocks may compete for the same intermediary markers, and there is no reason why one of the blocks (e.g., the larger of the two in a greedy algorithm) should win. Each block with at least one private marker is considered to be a block. Blocks completely nested within a larger block are not considered. An overall measure of similarity between two populations was calculated as the sum of cross-

Table 1

Selected Gene Regions and SNPs

			Size (in kb) of		No. of SNPs					Median Spacing (in kb) of	
Gene	Disease	Chromosome Region	Gene	Region	Selected[a]	Validated[b]	Common[c]	For HapMap Comparison[d]	Population Differentiating[e] (%)	Validated SNPs	HapMap SNPs
SNCA	Parkinson	4q21	112	188	97	78	73	33	5 (6)	2.1	4.5
LMNA	Cardiomyopathy	1q21.2	23	177	37	29	27	17	4 (14)	4.4	6.7
FKBP5	Depression	6p21.31	115	289	76	44	37	37	10 (23)	6.2[f]	6.3
PLAU	Alzheimer	10q22.2	6.3	95	53	34	32	13	27 (79)	2.2	6.0

[a] Selected from public SNP databases.
[b] Polymorphic in our sample set and in Hardy-Weinberg equilibrium.
[c] Minor-allele frequency >5%.
[d] For comparison, we selected a set of SNPs similar to the currently available set in the International HapMap Project (April 2004).
[e] $P < .001$.
[f] SNP spacing within the gene is 3.7 kb.

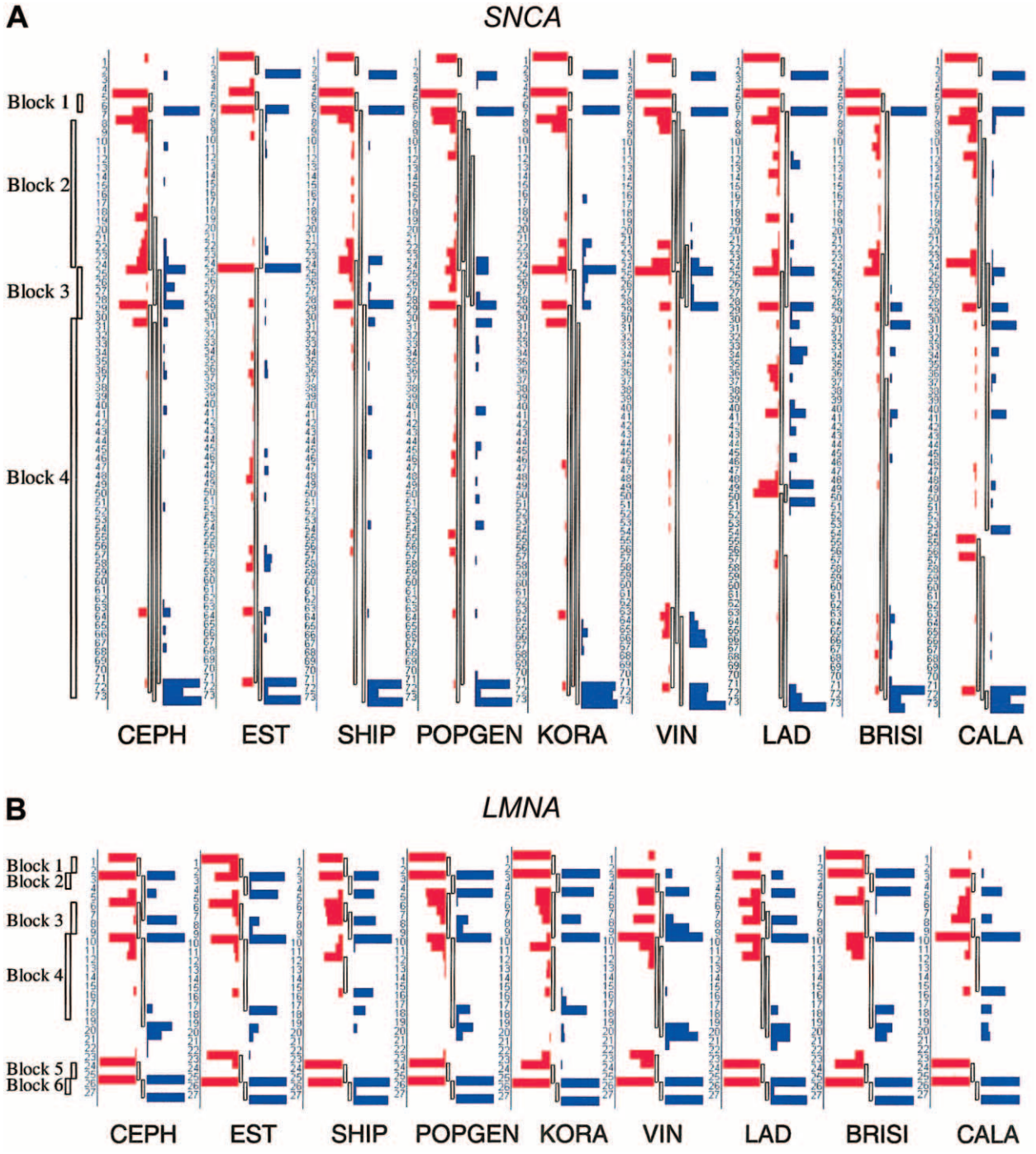
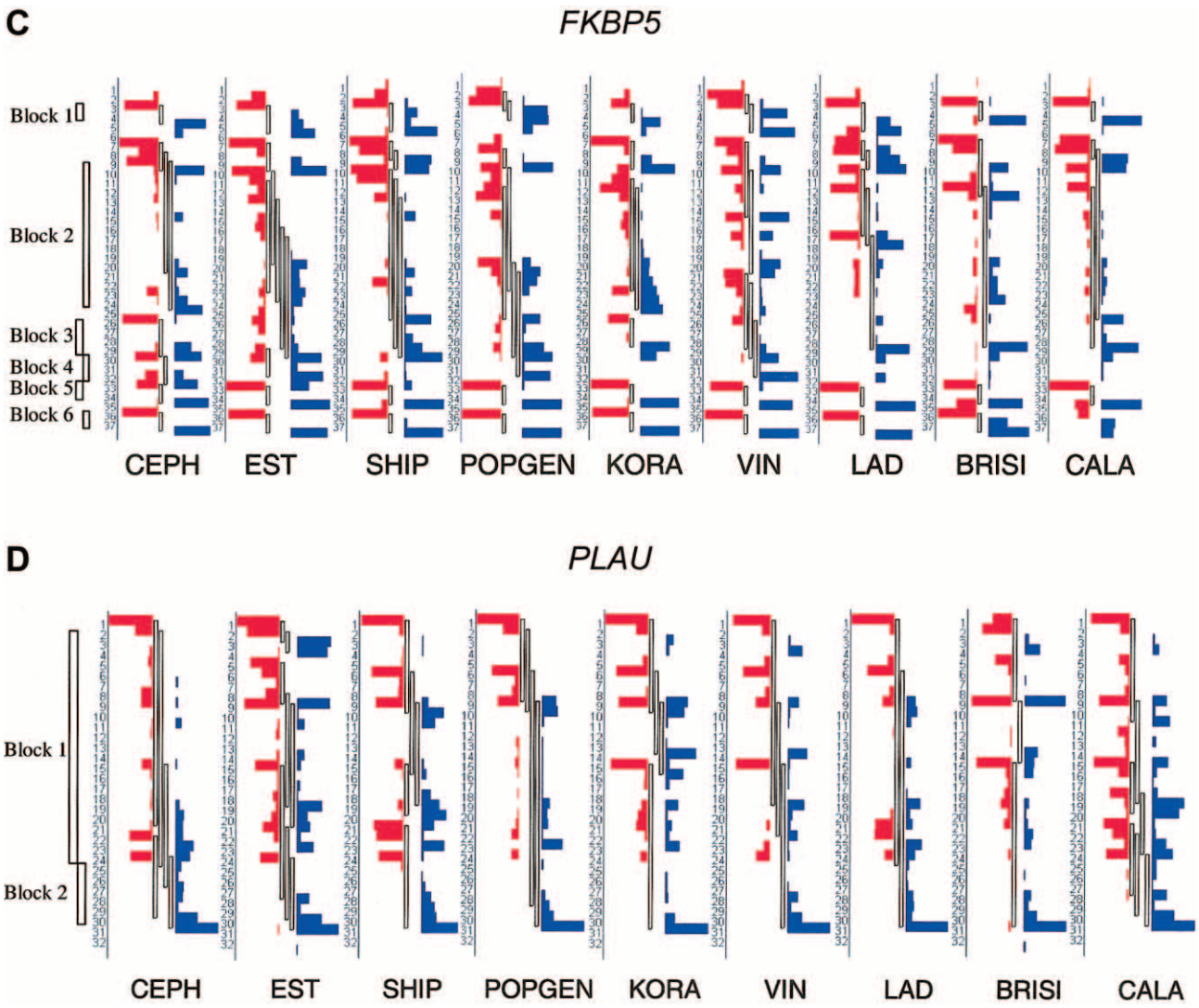


Figure 2 Bootstrap frequencies of block starts and block ends in all population samples. All samples have an equal population size of 100 individuals (except BRISI, with 98 individuals). SNP markers are ordered vertically by their physical sequence. The length of red and blue bars indicate the bootstrap frequency of block starts and block ends, respectively, at the given position. Between the bars, the observed block structure is shown, with blocks allowed to overlap. To the left of each CEPH graph, the block structure of CEPH is shown, in accordance with the standard algorithm of Gabriel et al. (2002), without allowance for overlapping blocks. An example of a boundary shift can be seen at the end of block 4 in *LMNA*, which shows clear differences between the populations tested.



products of bootstrap frequencies, standardized by the sums of within-products of bootstrap frequencies, similar to the genetic identity of Nei (1972). An appropriate distance was given by the negative logarithm of this measure and was used in a multidimensional scaling algorithm to map the overall block similarity.

Selection and Efficiency Testing of tagSNPs

Either the CEPH trios or the local European populations, with varying numbers of randomly selected subsamples (100 replicates), were used as reference samples. For these reference samples, two different tagSNP selection algorithms were applied. The first algorithm finds SNPs that best tag other typed SNPs (i.e., tagSNPs); the second algorithm finds SNPs that represent common haplotypes within predefined blocks (i.e., haplotype-tagging SNPs [htSNPs]).

A greedy algorithm for the selection of tagSNPs (Carlson et al. 2004) was employed. In the first step, a SNP exceeding an r^2 threshold of 0.8 with the maximum number of other SNP sites is identified. This SNP and all associated SNPs are grouped in one bin. A bin does not have to be a group of neighboring SNPs but rather can be split up in several regions. Any SNP exceeding the threshold r^2 with all other sites in the bin is specified as a tagSNP. There may be more than one tagSNP, and we used only the one with the maximum average r^2. This binning process is iterated, and all as-yet-unbinned SNPs are analyzed at each round, until all sites are binned or characterized as singleton bins. The efficiency of a given tagSNP set in other population samples is tested by the

following criteria: an average r^2 among all typed SNPs and the best SNP-specific tagSNP, a minimal r^2 among all typed SNPs and the best SNP-specific tagSNP, and a ratio of SNPs above the threshold r^2 to any tagSNP.

The htSNP selection method started with the definition of blocks, in accordance with the standard method of Gabriel et al. (2002). Because we wanted to compare the efficiency of htSNPs across several populations, we always used the block structure of the CEPH trios as the reference and forced the block structure of other populations to this reference structure. A single optimal set of htSNPs within each block was identified by sequential steps (Zhang and Jin 2003); these steps account for haplotype coverage (80% and 90% thresholds), optimal r^2 among SNPs, and even spacing. To evaluate the selected htSNP set in any population and to compare between populations, we defined two statistics (chromosomal coverage of tagged haplotypes and ratio of nontagged common haplotypes), whereby common haplotypes are defined by a frequency >5%.

Results

Single SNPs

Allele frequencies of most single markers did not differ significantly among population samples. In three gene regions, the proportion of population-differentiating markers (defined by signficance level $P < .001$) varied between 6% and 23% (table 1). An exception was *PLAU,* with 79% population-differentiating SNPs. Maximum values for allele-frequency differences were ~20% and were mostly seen between EST and CALA, thus indicating a geographical gradient between the northern and southern populations (table A1 [online only]). The pattern of population differentiation for each gene region is shown in table A2 (online only). Significant population differences in allele frequencies appeared mostly for the gene *PLAU* but also for *FKBP5* and *LMNA*. The CEPH founders were significantly different from the southern Italian populations BRISI and CALA. EST and the northern German collections SHIP and POPGEN showed significant differences from all Italian populations. The Alpine populations of VIN and LAD differed both from the southern Italian populations and from the northern European populations. The overall pattern of genetic differentiation reflects well the geographical localization of the population samples (table A3 and fig. A1 [online only]).

LD Structure

Standard plots of pairwise LD revealed similar patterns across samples (fig. A2 [online only]). To compare the LD structure across populations in a detailed and robust probability-based assessment, block overlaps were allowed and bootstrap frequencies of specific boundary positions were evaluated. The observed LD block structure and the bootstrap frequencies for each block start and block end are shown in figure 2. The calculations are based on a sample size of 100 individuals per population, to exclude variation in sample size as a potential confounder.

The general patterns of block structures are similar across samples, which is most prominent in the *LMNA* gene. Five of the six blocks in *LMNA* have nearly conserved block starts and ends across all study populations. Only the end of the largest block varied between positions 15 and 21, depending on the population studied. This represents a shift in the block extension in the range of 7–15 kb. Another example of differences in block boundaries is obvious for the *SNCA* region, where the largest block (between marker positions 30 and 73) has a tendency to break up into two pieces at different positions in the VIN (positions 63–67), LAD (positions 48–51), and CALA (positions 53–57) samples. Individual breakpoints of LD blocks for the Alpine populations VIN and LAD were also detected in the *FKBP5* region between positions 14 and 18. *PLAU* also exhibited variable block structure.

The overall variability in block structure among the populations is shown in figure 3. In a combined multidimensional scaling for all four gene regions, the most extreme and individual block structures were indicated

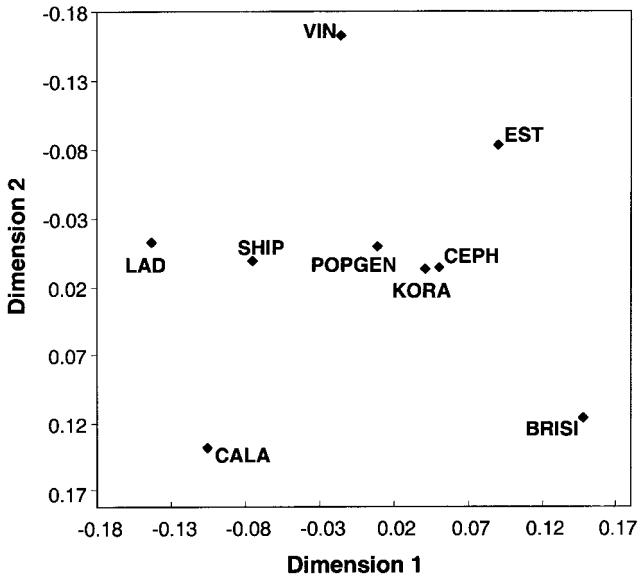


Figure 3 Overall similarity of block boundaries across all four gene regions. The first two dimensions, after a multidimensional scaling of the dissimilarity measure of block boundaries, are shown. Sample sizes are adjusted to a size of 100 individuals. The Alpine and geographically peripheral populations (EST, LAD, VIN, BRISI, and CALA) differ the most from all other population samples.

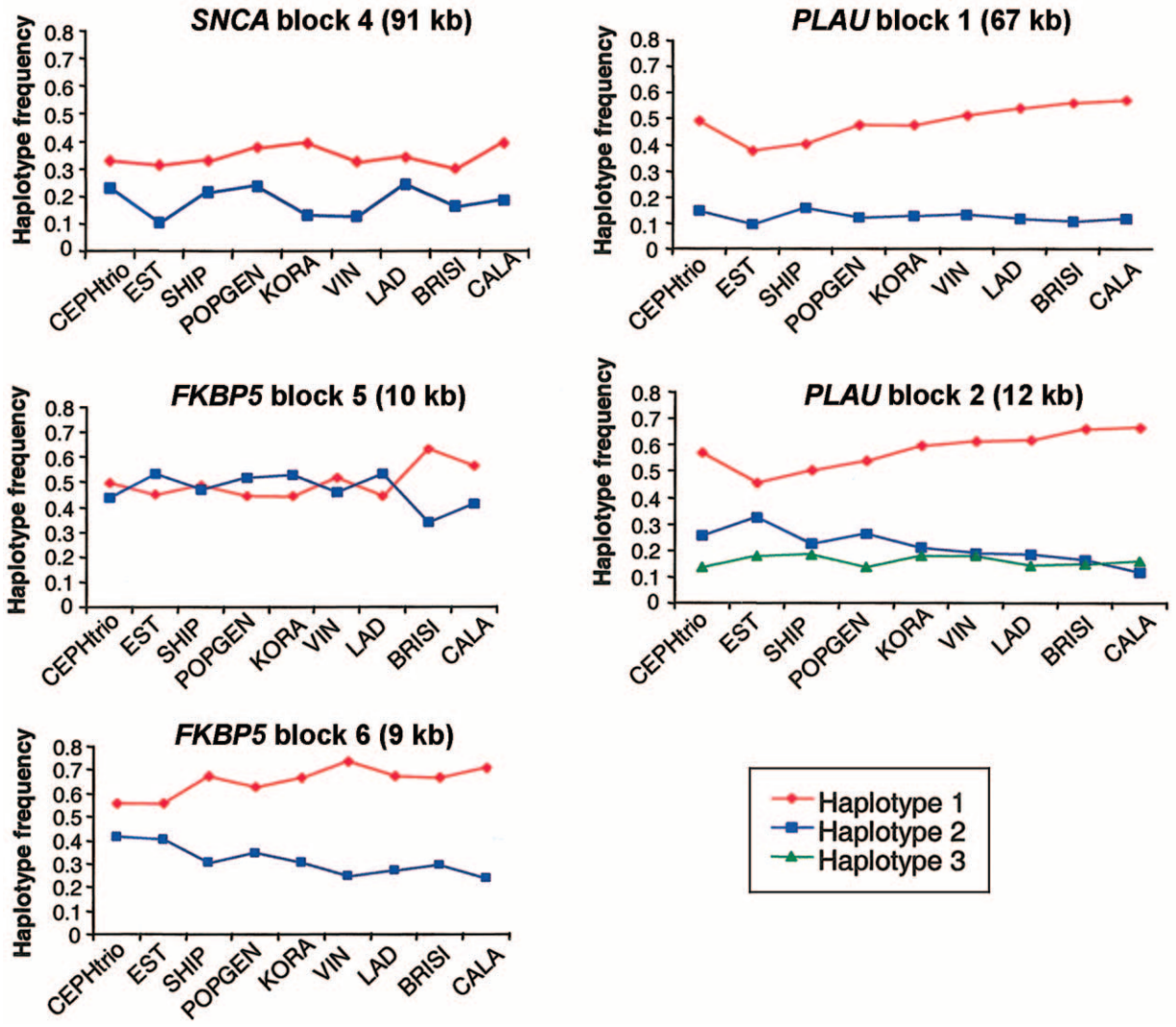


Figure 4 Frequencies of common haplotypes (>10%) in all populations for the five haplotype blocks with significant population differentiation. For block numbers, see figure 2. Populations are arranged on the *X*-axis in a north-to-south localization. Geographical frequency gradients are prominent in blocks 1 and 2 of *PLAU* and in block 6 of *FKBP5*.

for EST, LAD, VIN, BRISI, and CALA. The German populations, SHIP, POPGEN, and KORA, and the reference population CEPH appear in the center, indicating an intermediate block structure. Similar patterns were found when each gene was analyzed separately.

Haplotypes

The standard algorithm of Gabriel et al. (2002), applied to the CEPH trios, allowed us to define four blocks in *SNCA*, six blocks in *LMNA*, six blocks in *FKBP5*, and two blocks in *PLAU* (see fig. 2, for reference of block positions, and fig. A3 [online only], for haplotype estimations). Significant haplotype frequency differences among populations were found only in block 4 of *SNCA* ($P = .001$), blocks 5 and 6 of *FKBP5* ($P = .03$ and $P = .02$, respectively), and blocks 1 and 2 of *PLAU* ($P = .001$ and $P = .01$, respectively). Figure 4 shows the frequencies of all common haplotypes (frequency >10%) within the blocks that showed significant differences between populations. After Bonferroni correction for multiple testing, only the haplotype distributions of *SNCA* block 4 and *PLAU* block 1 remained significant. A clear geographic variation was evident with *FKBP5* and *PLAU* but not with *SNCA*. In the *PLAU* gene region, haplotype 1 in block 1 showed the most extreme frequency values in EST (40%) and CALA (57%) and showed a gradient in between these two values in the remaining six populations. CEPH trios showed intermediate frequencies. A similar pattern was found for the haplotypes in block 2 of *PLAU*. There was also a gra-

dient between EST and CALA in block 6 of *FKBP5,* but here CEPH trios are most similar to the EST sample. In block 5 of *FKBP5,* BRISI and CALA diverge from all other samples.

tagSNPs

We first tested the efficiency of tagSNPs, which were defined to represent untagged SNPs with a high correlation coefficient ($r^2 > 0.8$ [Carlson et al. 2004]). Figure 5 shows the performance of CEPH trios, as a reference for this tagSNP selection, in comparison with local population samples of different sizes. For each sample size, average values across 100 replicates are given. Only the criterion of a ratio of tagged SNPs above the threshold, which is the relative portion of SNPs correlated with any tagSNP by an r^2 value >0.8, is shown. Other evaluation criteria (see the "Subjects and Methods" section) gave similar results (see table A4 [online only]). A reduced SNP set, which is comparable to the HapMap set, was used for the tagSNP selection (table 1). This reduced SNP set comprised ~40% of SNPs identical to HapMap data. The selected tag SNPs were then tested on the full SNP set in all populations. The LD patterns appeared to be similar across all different SNP sets (see fig. A4 [online only]). There was no difference between the tagSNP set defined from CEPH trios and the tagSNP set defined from CEPH founders only, indicating that the additional phase information does not change the outcome.

The tagSNPs identified in the CEPH trios performed well for the genes *SNCA, FKBP5,* and *LMNA*. For >70% of typed SNPs, the r^2 value was >0.8 with the best tagSNP. SNP allelic variation in KORA, for example, is well represented by CEPH tagSNPs for the genes *SNCA* and *LMNA*. In *LMNA* and *SNCA,* a local sample size of 20 individuals as a reference performs mostly worse than the 30 CEPH trios. Only a sample size of 40 or 60 individuals is comparable to CEPH trios. In *FKBP5,* most local samples of 20 individuals performed better as a reference than the CEPH sample, except for VIN and CALA. A different situation was seen in *PLAU,* where six populations showed a ratio of tagged SNPs of <70%, when the CEPH sample was used as reference—the worst ratio was from the CALA sample, with only 53%. For the same gene, data from 20 random individuals of most populations performed better as a reference than CEPH trios.

With the current HapMap SNP density as a reference, minimal r^2 values between tagged and tagSNPs were as low as 0.036, even for the most conserved gene regions around *SNCA* and *LMNA*. We therefore tested the performance of tagSNP sets, when selected from the full SNP set in *SNCA* and *PLAU,* for which we had a >2-fold SNP density compared with that of the HapMap. The general pattern—a local sample with 20 individuals performs mostly better as a reference than the CEPH sample in *PLAU,* and CEPH performs better than local samples in *SNCA*—did not change, but the differences were less pronounced, and, even for *PLAU,* the ratio of tagged SNPs was >70% in all tested populations (table A4 [online only]).

Performance patterns were similar when the haplotype-based tagSNP selection method (i.e., the htSNP selection method) was used (Zhang and Jin 2003), but differences between CEPH and local references were weaker than those measured by the r^2 method (table A5 [online only]). When CEPH trios were used as reference, chromosomal coverage of tagged haplotypes was below the intended threshold (80% and 90%, respectively) only for the *PLAU* gene (see population samples EST, SHIP, and KORA).

tagSNPs may also be seen as a set of relatively independent markers. To assess the probability of recruiting population-differentiating SNPs in a genomic approach, we plotted a histogram of the *P* values of tests for population differentiation for all tagSNPs defined by the method of Carlson et al. (2004) in CEPH trios (fig. 6). The majority of tagSNPs did not show strong population differences, underlining their universality. Most highly significant markers were found in the gene regions *PLAU* and *FKBP5*.

Discussion

Individual population history and geographic variation may challenge the usefulness of a single European reference population for the selection of tagSNPs in association studies. Allele and haplotype frequencies show a clear geographic variation. Most dramatic frequency shifts lie in the upper range of values found in the survey of random loci done by Cavalli-Sforza et al. (1994), but these are still low compared with the values in strongly selected loci (e.g., cystic fibrosis variants [Lao et al. 2003]). The pattern of genetic differentiation corresponds relatively well to the European genetic variation described by Barbujani and Sokal (1990) and Cavalli-Sforza et al. (1994). The relatively strong genetic divergence of the Italian populations, CALA and BRISI, and the Alpine populations, LAD and VIN, from all other populations can be attributed to isolation resulting from linguistic differences (Germanic-Romance) and physical boundaries (the Alps). The two Italian populations, BRISI and CALA, also show significant differences from the CEPH sample and are therefore less well represented by this reference population.

Large-scale association studies—probably across different ethnic groups—are needed to detect small genetic effects on complex traits, and it is well known that even small amounts of cryptic population stratification can

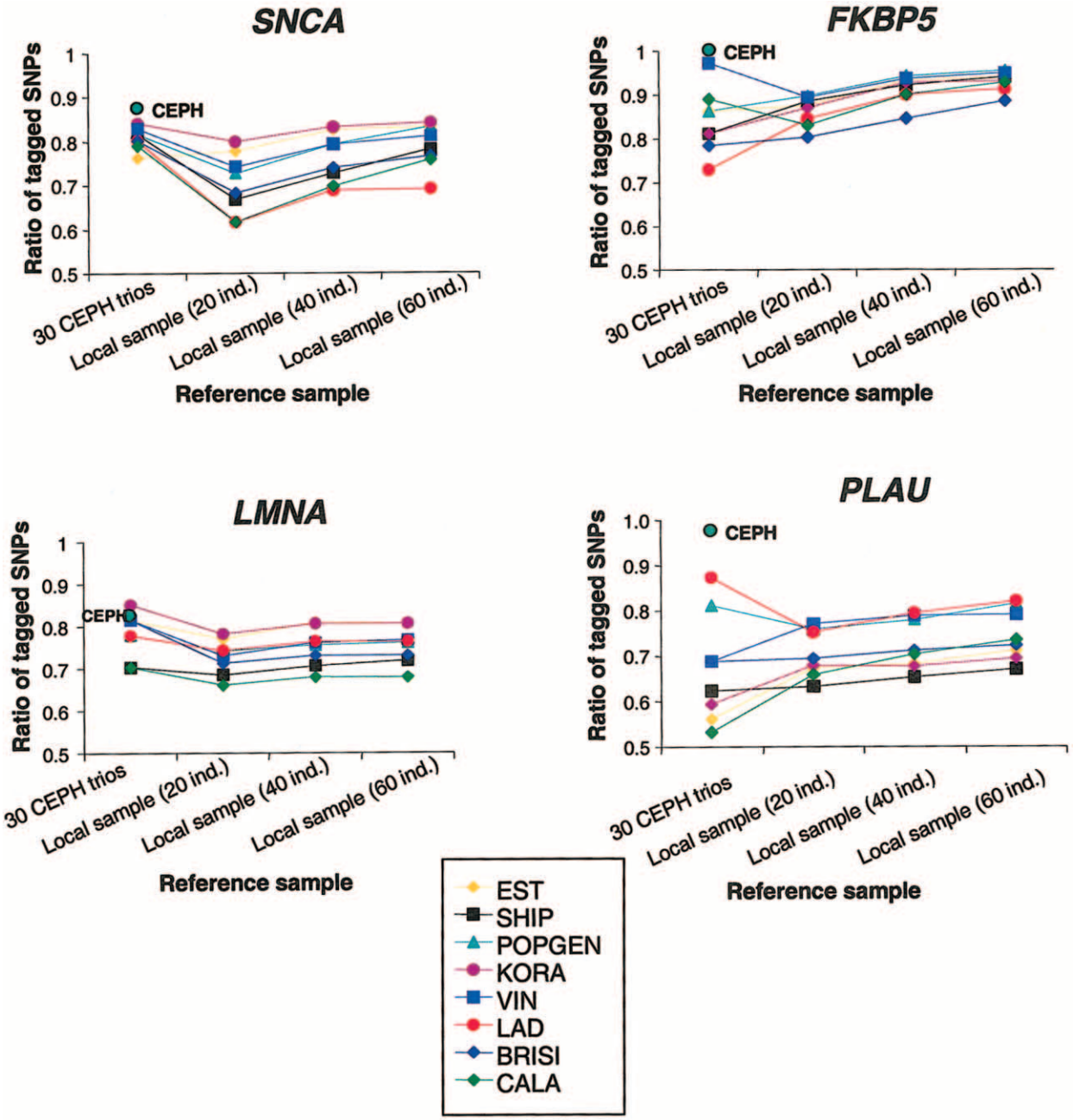


Figure 5 Performance of CEPH trios and local samples with different sample sizes used as references for tagSNP definition (by use of the method of Carlson et al. [2004]). The performance criterion shown is the ratio of tagged SNPs above the r^2 threshold of 0.8. The tagSNP sets that were defined in the CEPH trios were tested on all populations, whereas the tagSNP sets of local samples were tested only on the same local population. CEPH trios performed relatively well as reference (ratio of tagged SNPs >0.7), except for the *PLAU* gene region.

undermine such association studies (Marchini et al. 2004). However, high numbers of markers—in the range of several hundred microsatellite loci or a multitude of SNP loci—are required to detect genetic clusters of different ethnic origins in Europe (Rosenberg et al. 2002). Our results indicate that, in our total region of 749 kb, ~28% (16/57) of the tagSNPs showed highly significant differences ($P < .001$) among our set of study populations. This rate indicates that the recruitment of population-differentiating SNPs for the purpose of genetic matching strategies in case-control studies is feasible (Hoggart et al. 2004).

Comparative analyses of the haplotype block structure revealed a high degree of concordance among Eu-

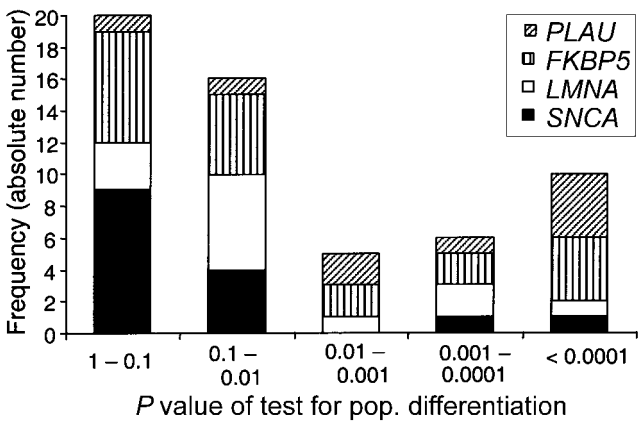


Figure 6 Histogram of *P* values of tests for population differentiation, on the basis of 57 tagSNPs from all gene regions in the CEPH trios. The allele frequencies of most tagSNPs were similar across populations ($P > .01$). Exceptions were the tagSNPs of the *PLAU* gene.

ropean populations (Nejentsev et al. 2004; Ng et al. 2004; Stenzel et al. 2004), as well as among populations from different continents, such as Asia, Africa, and Europe (Gabriel et al. 2002; Wall and Pritchard 2003). This presumably reflects, in part, the shared ancestry of human populations or common variation patterns of recombination rates, but, to some extent, it also reflects the effect of uneven marker spacing in these studies. However, small, well-defined differences for block boundaries have been reported among Finnish subpopulations (Mannila et al. 2003). All the above-mentioned studies (except Mannila et al. 2003) compare the positions of LD block boundaries by use of a greedy algorithm—or just by inspection of pairwise LD measures—but do not account for the relative probabilities of specific boundary positions. With our method of evaluating the strength of block boundaries, which was applied to exactly the same set of common markers in each population, we were able to show clear examples of block boundary shifts and block fragmentation among European samples. The values of our similarity measure for block structure, which estimate the average probability that boundaries coincide, ranged from 0.72 among LAD and BRISI to 0.87 among SHIP and POPGEN. With the exception of the Alpine populations, the overall variation appeared in a pattern that was concordant with geography, indicating the usefulness of our similarity measure for population-genetic comparisons. The observed pattern suggests that demographic and/or biological factors shaping block boundaries vary in a geographical sense and differentiate in accordance with the level of presumed genetic isolation of populations.

The observed population differences in haplotype frequencies and LD structure may affect the power to detect phenotype-genotype associations. Association signals at markers, which are correlated with a true causal variant, may appear at different positions in populations with an individual LD structure, such as the VIN, LAD, and CALA populations. Repeated studies among such populations are likely to present different results and are problematic for finding positive replications. In contrast, population-specific fragmented LD blocks are useful for the fine-mapping of causal variants within the region.

We also tested the transferability of tagSNP sets among populations. It is often stated that tagSNPs are population specific and should be newly assessed in each local population or geographic area in which an association study is planned (Thompson et al. 2003; Weale et al. 2003; Carlson et al. 2004). On the other hand, the HapMap project claims that its data may be able to be used to define tagSNPs for related populations (International HapMap Consortium 2003). It has also been reported that tagSNPs can be effectively transferred among British, Norwegian, Finnish, and Romanian populations (Nejentsev et al. 2004). It is, however, not clear to what level of population differentiation tagSNPs are transferable between the HapMap data and local European populations. Our results indicate that tagSNPs defined in the HapMap CEPH trios perform relatively well for two of four candidate-gene regions, particularly in central European populations. For *SNCA* and *LMNA*, the data from CEPH trios perform even better as a reference than data from 20 local individuals. For two of the tested candidate genes (*PLAU* and *FKBP5*), CEPH is not such a good reference. A local sample size of only 20 individuals in most populations is more appropriate for determination of tagSNPs than the standard sample of 30 CEPH trios. By genotyping larger sample sizes (>20 individuals) in the population being studied, the advantage of a local reference will be stronger, but it appears that an increase in sample size beyond 40 individuals is not very effective. A substantial increase in tagSNP efficiency and transferability, however, is achieved by increasing the density of genotyped SNPs in the reference sample.

The surprisingly high performance of CEPH as a reference for tagSNP design in two gene regions was not due to an increased number of selected tagSNPs in CEPH or the additional phase information available from the trios (see table A4 [online only]). The special characteristic of CEPH being a multilocalized but panmictic European population probably confers the advantage to this sample collection. Our results suggest that future HapMap releases with a denser genotype data set will allow the sufficient selection of tagSNPs in the majority of gene regions in central European populations. However, for an as-yet-unknown proportion of genes, and especially for isolated and peripheral populations within Europe, the HapMap reference may not

perform optimally, making it necessary to establish the LD pattern from a local sample.

Acknowledgments

This work was supported by the National Genome Research Network and the Bioinformatics for the Functional Analysis of Mammalian Genomes project from the German Federal Ministry of Education and Research. A.M. and E.L. were partially supported by Targeted Funding EMRE 0182582s03, and E.L. had a fellowship from the E.U. grants Mol Tools 503155 and "Genera" to Estonian Biocentre. M.R. and R.M. were supported by a core grant from the Estonian Ministry of Education and Research. The recruitment of the south Tyrolian samples VIN and LAD was supported by a grant from the Autonomous Province Bolzano and from the Südtiroler Sparkasse, Bolzano. The project POPGEN is supported by the Deutsche Forschungsgemeinschaft research group FOR 423 ("Polygenic Disorders"). The SHIP studies are funded by the German Federal Ministry for Education and Research (grant 01ZZ96030), by the Ministry for Education, Research, and Cultural Affairs, and by the Ministry for Social Affairs of the State of Mecklenburg-West Pomerania. We gratefully acknowledge the participation of all probands, as well as the review of the manuscript by Jack Favor.

Electronic-Database Information

The URLs for data presented herein are as follows:

GSF European LD Pattern Project, http://ihg.gsf.de/LD/ (for a downloadable version of the genotype data presented in this study)
HapMap Homepage, http://www.hapmap.org/ (for the International HapMap Project)
popgen, http://www.popgen.de/

References

Barbujani G, Sokal RR (1990) Zones of sharp genetic change in Europe are also linguistic boundaries. Proc Natl Acad Sci USA 87:1816–1819

Cardon LR, Abecasis GR (2003) Using haplotype blocks to map human complex trait loci. Trends Genet 19:135–140

Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 74:106–120

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, Princeton, NJ

Chapman JM, Cooper JD, Todd JA, Clayton DG (2003) Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. Hum Hered 56:18–31

Dausset J, Cann H, Cohen D, Lathrop M, Lalouel JM, White R (1990) Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. Genomics 6:575–577

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. Science 296: 2225–2229

Hoggart CJ, Shriver MD, Kittles RA, Clayton DG, McKeigue PM (2004) Design and analysis of admixture mapping studies. Am J Hum Genet 74:965–978

International HapMap Consortium (2003) The International HapMap Project. Nature 426:789–796

Ke X, Hunt S, Tapper W, Lawrence R, Stavrides G, Ghori J, Whittaker P, Collins A, Morris AP, Bentley D, Cardon LR, Deloukas P (2004) The impact of SNP density on fine-scale patterns of linkage disequilibrium. Hum Mol Genet 13:577–588

Lao O, Andres AM, Mateu E, Bertranpetit J, Calafell F (2003) Spatial patterns of cystic fibrosis mutation spectra in European populations. Eur J Hum Genet 11:385–394

Mannila H, Koivisto M, Perola M, Varilo T, Hennah W, Ekelund J, Lukk M, Peltonen L, Ukkonen E (2003) Minimum description length block finder, a method to identify haplotype blocks and to compare the strength of block boundaries. Am J Hum Genet 73:86–94

Marchini J, Cardon LR, Phillips MS, Donnelly P (2004) The effects of human population structure on large genetic association studies. Nat Genet 36:512–517

Nei M (1972) Genetic distance between populations. Am Nat 106:283–292

Nejentsev S, Godfrey L, Snook H, Rance H, Nutland S, Walker NM, Lam AC, Guja C, Ionescu-Tirgoviste C, Undlien DE, Ronningen KS, Tuomilehto-Wolf E, Tuomilehto J, Newport MJ, Clayton DG, Todd JA (2004) Comparative high-resolution analysis of linkage disequilibrium and tag single nucleotide polymorphisms between populations in the vitamin D receptor gene. Hum Mol Genet 13:1633–1639

Ng MCY, Wang Y, So WY, Cheng S, Visvikis S, Zee RYL, Fernandez-Cruz A, Lindpaintner K, Chan JCN (2004) Ethnic differences in the linkage disequilibrium and distribution of single-nucleotide polymorphisms in 35 candidate genes for cardiovascular diseases. Genomics 83:559–565

Phillips MS, Lawrence R, Sachidanandam R, Morris AP, Balding DJ, Donaldson MA, Studebaker JF, et al (2003) Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. Nat Genet 33:382–387

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298:2381–2385

Schwartz R, Halldorsson BV, Bafna V, Clark AG, Istrail S (2003) Robustness of inference of haplotype block structure. J Comput Biol 10:13–19

Stenzel A, Lu T, Koch WA, Hampe J, Guenther SM, De La Vega FM, Krawczak M, Schreiber S (2004) Patterns of linkage disequilibrium in the MHC region on human chromosome 6p. Hum Genet 114:377–385

Thompson D, Stram D, Goldgar D, Witte JS (2003) Haplotype tagging single nucleotide polymorphisms and association studies. Hum Hered 56:48–55

Wall JD, Pritchard JK (2003) Haplotype blocks and linkage

disequilibrium in the human genome. Nat Rev Genet 4:587–597

Wang WYS, Todd JA (2003) The usefulness of different density SNP maps for disease association studies of common variants. Hum Mol Genet 12:3145–3149

Weale ME, Depondt C, Macdonald SJ, Smith A, Lai PS, Shorvon SD, Wood NW, Goldstein DB (2003) Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene *SCN1A*: implications for linkage-disequilibrium gene mapping. Am J Hum Genet 73:551–565

Zhang K, Jin L (2003) HaploBlockFinder: haplotype block analyses. Bioinformatics 19:1300–1301