Genetic Epidemiology 30: 180–190 (2006)

Tag SNP Selection for Finnish Individuals Based on the CEPH Utah HapMap Database

Cristen J. Willer,[1]* Laura J. Scott,[1] Lori L. Bonnycastle,[2] Anne U. Jackson,[1] Peter Chines,[2] Randall Pruim,[1,3] Craig W. Bark,[4] Ya-Yu Tsai,[4] Elizabeth W. Pugh,[4] Kimberly F. Doheny,[4] Leena Kinnunen,[6] Karen L. Mohlke,[7] Timo T. Valle,[6] Richard N. Bergman,[5] Jaakko Tuomilehto,[6,8,9] Francis S. Collins[2] and Michael Boehnke[1]

[1]*Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan*
[2]*Genome Technology Branch, National Human Genome Research Institute, Bethesda, Maryland*
[3]*Department of Mathematics and Statistics, Calvin College, Grand Rapids, Michigan*
[4]*Center for Inherited Disease Research, Johns Hopkins University School of Medicine, Baltimore, Maryland*
[5]*Department of Physiology and Biophysics, Keck School of Medicine, University of Southern California, Los Angeles, California*
[6]*Diabetes and Genetic Epidemiology Unit, Department of Epidemiology and Health Promotion, National Public Health Institute, Helsinki, Finland*
[7]*Department of Genetics, University of North Carolina, Chapel Hill, North Carolina*
[8]*Department of Public Health, University of Helsinki, Helsinki, Finland*
[9]*South Ostrobothnia Central Hospital, Seinäjoki, Finland*

The pattern and nature of linkage disequilibrium in the human genome is being studied and catalogued as part of the International HapMap Project [2003: *Nature* 426:789–796]. A key goal of the HapMap Project is to enable identification of tag single nucleotide polymorphisms (SNPs) that capture a substantial portion of common human genetic variability while requiring only a small fraction of SNPs to be genotyped [International HapMap Consortium, 2005: *Nature* 437:1299–1320]. In the current study, we examined the effectiveness of using the CEU HapMap database to select tag SNPs for a Finnish sample. We selected SNPs in a 17.9-Mb region of chromosome 14 based on pairwise linkage disequilibrium (r^2) estimates from the HapMap CEU sample, and genotyped 956 of these SNPs in 1,425 Finnish individuals. An excess of SNPs showed significantly different allele frequencies between the HapMap CEU and the Finnish samples, consistent with population-specific differences. However, we observed strong correlations between the two samples for estimates of allele frequencies, r^2 values, and haplotype frequencies. Our results demonstrate that the HapMap CEU samples provide an adequate basis for tag SNP selection in Finnish individuals, without the need to create a map specifically for the Finnish population, and suggest that the four-population HapMap data will provide useful information for tag SNP selection beyond the specific populations from which they were sampled. *Genet. Epidemiol.* 30:180–190, 2006. © 2005 Wiley-Liss, Inc.

Key words: linkage disequilibrium; haplotype; HapMap; Finland

Contract grant sponsor: National Institutes of Health; Contract grant number: R01 DK62370; R01 HG00376; N01-HG-65403; Contract grant sponsor: National Human Genome Research Institute; Contract grant number: OH95-C-N030; Contract grant sponsor: Academy of Finland; Contract grant numbers: 38387; 46558; Contract grant sponsor: Burroughs Wellcome Fund.
*Correspondence to: Cristen J. Willer, PhD. School of Public Health, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109. E-mail: cristen@umich.edu
Received 26 August 2005; Accepted 2 October 2005
Published online 22 December 2005 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20131

INTRODUCTION

Genome-wide association studies provide a powerful means to detect genetic variants that predispose to complex human diseases [Risch and Merikangas, 1996]. Technologies that permit extremely high-throughput genotyping of SNPs [Gunderson et al., 2005; Matsuzaki et al., 2004] are now enabling such studies [Ozaki et al., 2002; Klein et al., 2005]. Despite increasing genotyping efficiency, it still is not feasible to type the millions of catalogued SNPs [Sherry et al., 2001] in large enough samples to identify disease-predisposing variants. Instead, genotyping a well-chosen subset of tag SNPs has the potential to capture most of the common human genetic variability [Carlson et al., 2004], since SNPs in close physical proximity are often correlated because of linkage disequilibrium (LD) [International HapMap Consortium, 2005].

The International HapMap Consortium [2003] is genotyping millions of SNPs to characterize patterns of LD across the genome and to generate haplotype maps in four reference populations: CEPH (Centre d'etude du polymorphisme humain) reference individuals from Utah, USA (CEU), Yoruba individuals from Ibadan, Nigeria (YRI), Japanese individuals from Tokyo (JPT), and Han Chinese individuals from Beijing (CHB) (www.hapmap.org). Patterns of LD are a result of the age of the genetic variants, population history, random genetic drift, recombination hotspots, and gene conversion [Tishkoff and Verrelli, 2003], and it is clear that patterns of LD differ between European-derived, African and Asian populations [Sawyer et al., 2005; International HapMap Consortium, 2005]. Thus, a key question is whether the resulting haplotype maps can be used to select SNPs in a broader set of populations.

In the current study, we addressed this question in a sample of 1,425 Finnish individuals in a 17.9-Mb region showing evidence of linkage to type 2 diabetes (T2D) [Silander et al., 2004]. We found strong correlations between the Finnish and HapMap CEU samples for allele frequency, haplotype frequency, and LD (r^2) estimates. Our results suggest that the HapMap CEU samples provide an adequate basis for tag SNP selection in Finnish individuals, and that the HapMap data will provide useful information for the design of association studies beyond the four populations from which they were sampled.

METHODS

SAMPLES

The 1,425 genotyped Finnish individuals were sampled as part of the Finland-United States Investigation of NIDDM Genetics (FUSION) study of T2D [Valle et al., 1998] or ascertained through the population-based Finrisk 2002 study [Saaristo et al., 2004]. They included 775 FUSION diabetic cases and 650 non-diabetic controls. Cases met WHO [1985] criteria for T2D and were selected from families ascertained for T2D sibling pairs. Controls included 187 normal glucose-tolerant spouses of FUSION diabetic individuals, 222 unrelated FUSION individuals who were normal glucose tolerant at ages 65 and 70 years, and 241 unrelated normal glucose-tolerant individuals from the Finrisk 2002 study [Saaristo et al., 2004]. Informed consent was obtained from all subjects and the study protocol was approved by local Institutional Review Boards or ethics committees at each of the participating centers. We obtained HapMap genotype data for 90 CEPH reference individuals from Utah, USA (CEU), 90 Yoruba individuals from Ibadan, Nigeria (YRI), 44 Japanese individuals from Tokyo (JPT), and 45 Han Chinese individuals from Beijing (CHB) from the HapMap database (www.hapmap.org). The CEU and YRI samples are comprised of 30 father-mother-offspring trios; the JPT and CHB individuals are unrelated.

SNP SELECTION

We selected a 17.9-Mb region of chromosome 14 (positions 51.00 to 68.88 Mb from UCSC genome build hg16, which is approximately equivalent to NCBI build 34) for SNP genotyping based on evidence for linkage in our Finnish T2D families [Silander et al., 2004]. There were 3,481 SNPs genotyped in this region in the HapMap CEU samples in the May 2004 release. Of these, 2,276 had minor allele frequency (MAF) $\geq 5\%$, Illumina design score > 0.4 (calculated in May 2004), and could be mapped on hg16. We estimated r^2, the squared correlation coefficient, for all pairs of these 2,276 SNPs, and selected 1,117 tag SNPs (average density 15.7 kb) using a greedy algorithm so that every unselected SNP had $r^2 \geq 0.80$ with one or more selected tag SNPs [Carlson et al., 2004]. Given a choice between multiple tag SNPs, we preferentially selected, first, nonsynonymous and splice site variants, and, second, SNPs with highest Illumina design scores (as calculated in May 2004). We added a second SNP from 14 regions with >10 SNPs in $r^2 \geq 0.8$. We also selected SNPs to fill gaps in the HapMap, 63 of which were genotyped by HapMap in the CEU sample by the time this analysis was performed (February 2005). We included an additional 10 nonsynonymous SNPs that had been validated by more than one submitter (''double-hit'') that were also subsequently genotyped by HapMap in the CEU sample. A total of 1,204 SNPs were genotyped in the CEU sample (as of February 2005), and had MAF $\geq 5\%$. These 1,204 SNPs were genotyped as part of a 1536-SNP Illumina panel, which included SNPs in the region that were not typed by HapMap as well as SNPs on other chromosomes, which were not used for comparisons between Finnish and HapMap samples. Quality assessment is provided for all 1,536 SNPs genotyped.

GENOTYPING

Genotyping was performed at the Center for Inherited Disease Research (CIDR) on a BeadLab 1000 system using the Illumina Golden Gate assay. Fan et al. [2003] and Gunderson et al. [2004] provide a detailed description of the Illumina genotyping platform and methods. Cluster definitions were determined for each locus using lllumina GenCall software. Genotypes were released on 100% of 2,208 attempted study samples. The sample consisted of affected sibling pair families that were used for identification of unlikely genotypes (see "Genotype Data" below) and from this, we selected 1,425 unrelated individuals for estimating allele frequencies, r^2 and haplotype frequencies. Of 1,536 attempted SNPs, 1,338 were released for a total of 2,954,304 study genotypes. The missing data rate for the 1,338 genotyped SNPs was 0.37%. In addition, we included CEPH control samples on each 96-well plate. The overall discrepancy rate for these controls was 0.09% and the overall parent-child discordance rate was 0.04%.

GENOTYPE DATA

Of the 1,204 chromosome 14 SNPs attempted at CIDR that were also genotyped in the HapMap CEU, 1,078 (89.5%) were successfully genotyped and released by CIDR. Of the 1,078 SNPs, we excluded 109 for poor clustering of genotypes. Twelve more SNPs were excluded for strong deviations from Hardy-Weinberg expectations ($p < 0.001$), ≥ 2 genotype discrepancies among 94 duplicate samples, or ≥ 3 close double recombinants identified using Merlin [Abecasis et al., 2002] in a sample of 615 sibling pairs in 270 families. Nine hundred fifty-seven SNPs (79.5%) of 1,204 attempted SNPs passed all quality control checks and were eligible for subsequent analysis. All of these SNPs were matched for the reference allele between the HapMap samples and our Finnish sample. Of the 957 SNPs, 897 were originally selected from HapMap information. The remaining 60 SNPs were originally non-HapMap SNPs but were subsequently genotyped by HapMap between May 2004 when SNP selection occurred and February 2005 when the analysis was performed. For these $897+60=957$ SNPs, the average genotype call rate was 99.7% in the Finnish sample; the minimum call rate was 91.0%. Three SNPs had a call rate $<95.0\%$.

We originally selected SNP rs8007267 for genotyping because it had no known tags in the HapMap CEU sample. After genotyping this SNP, however, it displayed the largest allele and haplotype frequency differences between the Finnish and CEU samples. The r^2 estimate of rs8007267 and its closest neighboring SNP rs943912 (6.4 kb apart) was 0.954 in the Finnish sample and 0.005 in the HapMap sample. We tested the genotypes for CEU individuals at SNP rs8007267 against all other SNPs on chromosome 14 (February 2005 release) and observed an r^2 value of 0.73 with SNP rs1467831 located 46 Mb away; this was significant even after correcting for the 21,160 SNPs on the chromosome. These discrepant results suggest that the CEU genotype data for this SNP were mislabeled, and we excluded SNP rs8007267 from all subsequent analyses.

Of the 956 SNPs included for Finnish and CEPH sample comparisons, 903 were genotyped in the HapMap YRI sample, of which 818 had MAF $\geq 5\%$ in YRI. Nine hundred and twenty-five of these 956 SNPs were genotyped in the HapMap JPT and CHB samples, of which 784 and 788 had MAF $\geq 5\%$, respectively.

GENOMIC DNA CHARACTERISTICS

We downloaded information on all known human gene transcripts that are mapped to specific chromosomal start and end positions (Ensembl, www.ensembl.org) [Birney et al., 2004], and selected one transcript per Ensembl Gene ID by first selecting known transcripts, and then by maximal cDNA length. We identified the minimum and maximum gene positions for each chromosome to approximate the amount of each chromosome that has reliable sequence information. This information provided estimates of the total number of genes and the total length of all cDNAs within our region and for the remainder of the genome. We estimated the ratio of centiMorgans (cM) per megabase (Mb) in the genome from physical and genetic map positions of proximal and distal short tandem repeats on each autosome [Kong et al., 2004].

STATISTICAL ANALYSIS

We estimated allele frequencies for HapMap CEU founders and Finnish cases and controls by gene counting, and used a large-sample comparison of two proportions to test for allele frequency differences in the two samples. Contingency tables of allele counts for the two samples using Fisher's exact test gave essentially the same

results (data not shown). We repeated the allele frequency analysis using genotype data for the four HapMap populations from HapMap release 16c.1 and obtained nearly identical results (data not shown). We estimated haplotype frequencies and r^2 using an EM algorithm that properly accounts for family relationships, as programmed in Fugue (Gonçalo Abecasis, personal communication). We estimated confidence intervals for haplotype frequency and r^2 estimates by assuming haplotypes were known and ignoring error associated with phasing of haplotypes, resulting in confidence intervals with somewhat lower coverage than normal. We estimated the standard error for estimates of r^2 as $4 \cdot r^2 \cdot (1-r^2)/2N$, where N is the number of individuals genotyped [Alf and Graf, 1999].

For haplotype frequency comparisons, we used r^2 estimates in our Finnish sample to define sets of SNPs in which each SNP had $r^2 > c$ ($c = 0.40$, 0.60, or 0.80) with at least one other SNP in the set; singleton SNPs falling between the start and end position of a set were also included and sets were allowed to overlap. This approach resulted in sets of <20 SNPs, allowing for efficient haplotype frequency estimation. We present results for $r^2 > 0.40$; we obtained similar results for $r^2 > 0.60$ or $r^2 > 0.80$, or using the HapMap CEU sample to form the SNP sets. When 180 sets were defined based on $r^2 > 0.40$, 736 of 957 SNPs (76.9%) were assigned to a set and 540 SNPs (56.4%) were assigned to exactly one set. Haplotype diversity for each SNP bin was estimated from estimated haplotype frequencies as $H = (n/(n-1))[1 - \sum_{i=1}^{k} p_i^2]$ for each of k haplotypes with frequency p_i and n representing the number of chromosomes examined [Nei, 1987]. We estimated recombination rates in 667 Finnish non-diabetic controls and 60 HapMap CEU founders using Phase v2.0.2 [Stephens and Donnelly, 2003]. We generated 1,000 replicates, thinned to every 3rd replicate to reduce the correlation between replicates, and determined the median value for each interval.

RESULTS

ALLELE FREQUENCIES

We compared allele frequency estimates in the 1,425 Finnish combined cases and controls and the 60 HapMap CEU samples for 956 chromosome 14 SNPs with MAF >0.05 in the CEPH sample. Overall, absolute allele frequency differences between the Finnish combined cases and controls and HapMap CEU samples tended to be small, and the Pearson product moment correlation for the allele frequency estimates between the two samples was 0.98 (Fig. 1a). In sum, 60.0, 90.2, and 98.8% of SNPs had an allele frequency difference <0.05, <0.10, and <0.15, respectively (Fig. 1b). Still, there was an excess of SNPs that showed significantly different allele frequencies: 200 (20.9%) were significant at the 0.05 level, 72 (7.5%) at the 0.01 level, and 29 (3.0%) at the 0.001 level. Allele frequency estimates for the Finnish samples were very different from those in the HapMap YRI, JPT, and CHB samples (Fig. 1a), and Pearson correlations between the allele frequencies in the Finnish samples and these HapMap samples were 0.37, 0.63, and 0.67, respectively (Fig. 1). The extent of allele frequency differences were very similar when the analysis was restricted to Finnish non-diabetic controls (data not shown) as were r^2 and haplotype estimate differences (results shown below for combined Finnish case and control sample).

For 956 SNPs with a minor allele frequency $>5\%$ in the HapMap CEU sample, the observed heterozygosity in the Finnish sample was slightly less than that observed in the CEU sample (Finnish = 33.5%, CEU = 34.7%); with the large number of samples, this difference is highly statistically significant ($p < 0.0001$).

r^2 estimates

We estimated pairwise r^2 for all pairs of SNPs, revealing highly similar patterns of LD across our 17.9-Mb region in the HapMap CEU and Finnish samples (Fig. 2). Estimates of r^2 for the 955 pairs of adjacent SNPs along the chromosome were also quite similar between the Finnish and HapMap CEU samples (Fig. 3a). The Pearson correlation of the r^2 estimates for adjacent pairs was 0.91. Of adjacent SNP pairs, 56.2, 75.8, and 87.1% had absolute r^2 difference estimates <0.05, <0.10, and <0.15, respectively (Fig. 3b). Consistent with the results for allele frequency estimates, and likely partly due to the anti-conservative nature of our confidence intervals, there was an excess of SNP pairs that showed significantly different r^2 estimates between the two samples. Of 946 pairs of SNPs, 155 (16.4%) were significant at the 0.05 level, 83 (8.7%) at the 0.01 level, and 44 (4.7%) at the 0.001 level.

Almost all of the SNPs genotyped in our Finnish samples were tag SNPs selected for having $r^2 < 0.8$ in the HapMap CEU sample. However, because

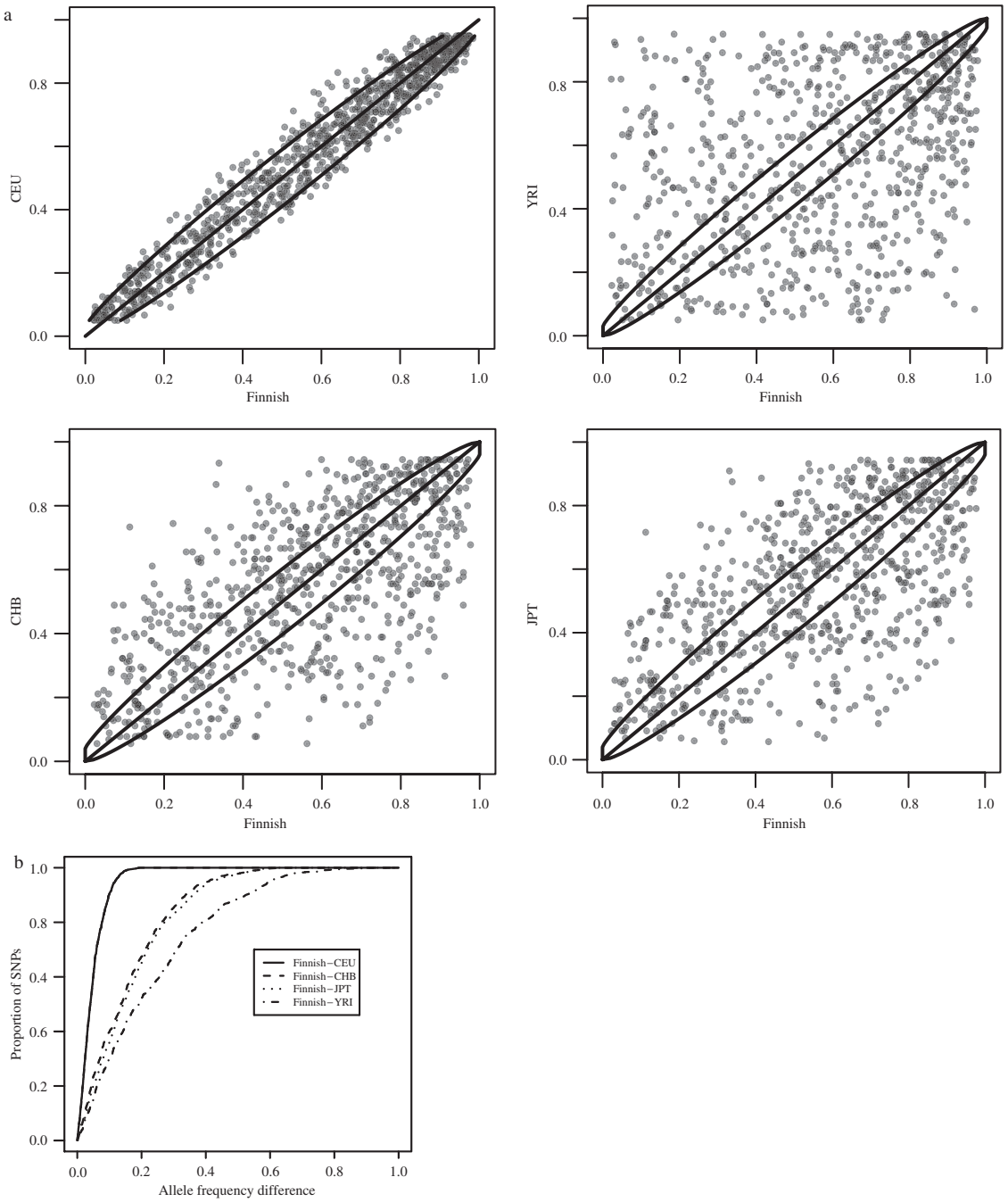

Fig. 1. a: Comparison of allele frequency estimates in the Finnish and HapMap samples for N SNPs with MAF >5% in the HapMap sample indicated: CEU (N = 957), YRI (N = 819), CHB (N = 789), and JPT (N = 785). The slope of one through the origin and the 95% confidence intervals for the allele frequency differences with respect to the HapMap samples are indicated. b: The cumulative distribution of the absolute allele frequency differences between the Finnish sample and each of the four HapMap samples. The allele frequency difference between the Finnish sample and the CEU (solid line), CHB (dashed line), JPT (dotted line), and YRI (alternating dot and dash line) are indicated.

some SNPs were selected without LD information or to duplicate tag a large set of SNPs, 22 SNP pairs with $r^2 \geq 0.8$ in the HapMap CEU sample were available for further examination. We randomly selected one SNP from each of the 22 pairs to represent a SNP that might remain untyped, and determined the maximum r^2 estimate for the selected SNP with any other SNP in the Finnish

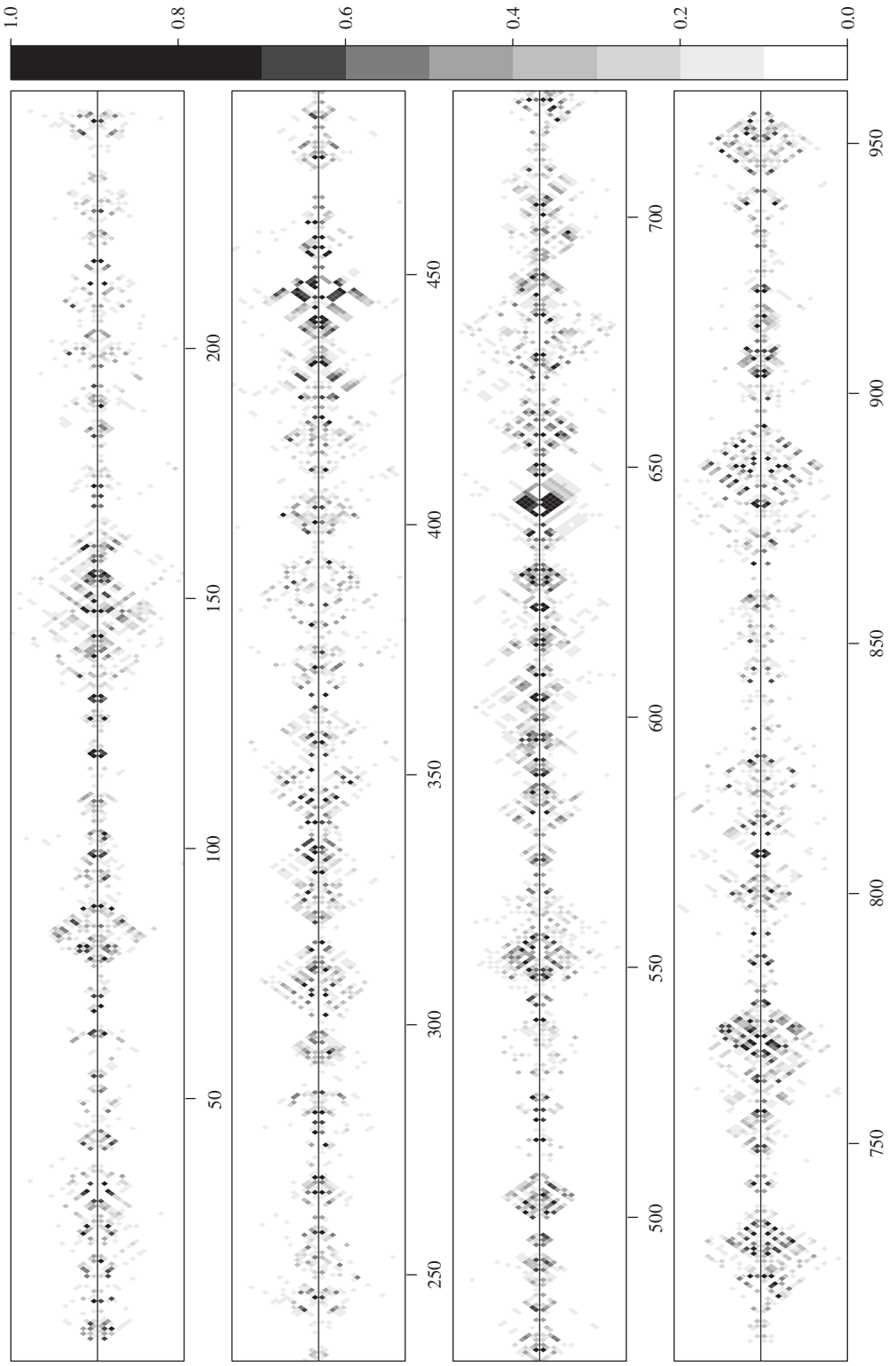

Fig. 2. Plot of estimated r^2 values for SNPs paired with each of the neighboring 20 SNPs in the Finnish (top segment of each row) and HapMap CEU (bottom segment of each row) samples. SNPs were selected based on estimated r^2 values in the HapMap CEU samples from a 17.9-Mb region on chromosome 14q.

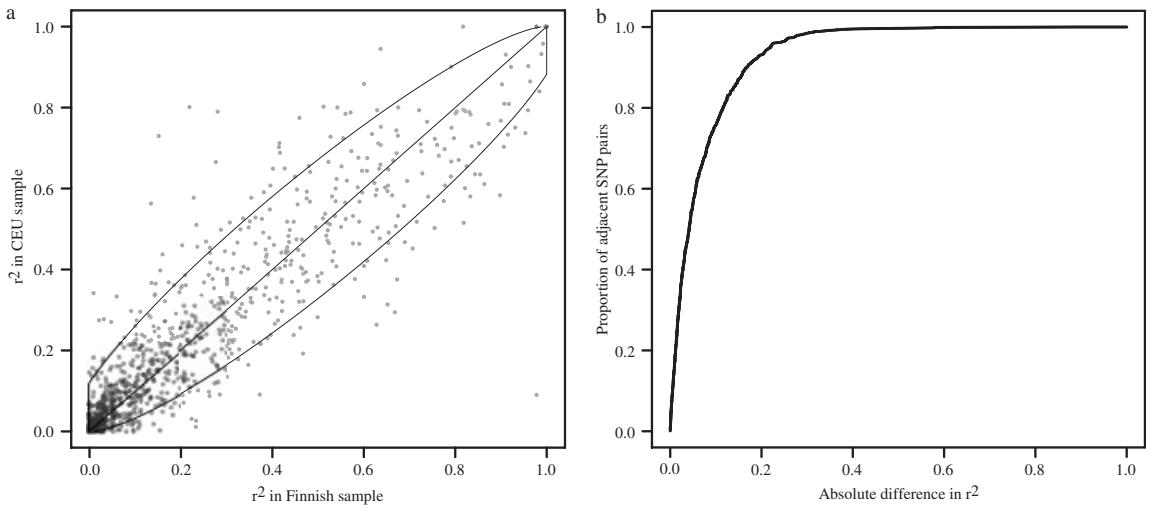

Fig. 3. a: Comparison of r^2 estimates in the Finnish and HapMap CEU samples for 956 pairs of adjacent SNPs. b: Cumulative distribution of the absolute difference in r^2 estimates between the HapMap CEU and Finnish samples.

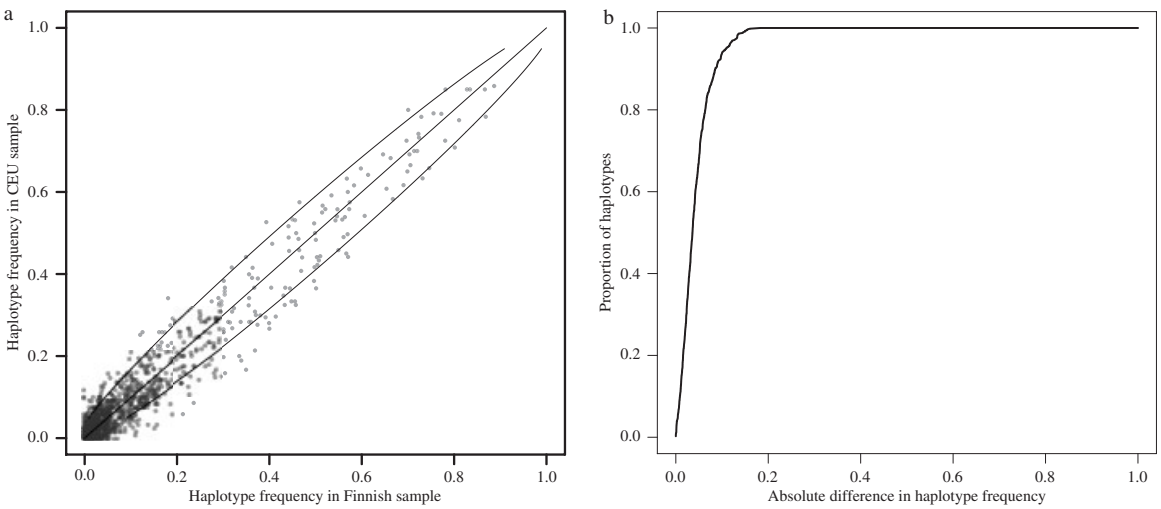

Fig. 4. a: Comparison of haplotype frequency estimates in the Finnish and HapMap CEU samples for 583 haplotypes with frequency estimates >5% in the HapMap CEU sample. The slope of one through the origin and approximate 95% confidence intervals for the allele frequency differences with respect to the HapMap samples are indicated. b: Cumulative distribution of the absolute difference in haplotype frequency estimates in the HapMap CEU and Finnish samples.

sample. Seventeen SNPs (77.3%) had $r^2 > 0.8$ with another ''typed'' SNP and 21 (95.5%) had $r^2 > 0.6$ with another ''typed'' SNP.

HAPLOTYPE FREQUENCIES AND RECOMBINATION RATES

We selected sets of 3–20 SNPs that had $r^2 > 0.4$ with at least one other SNP in the bin. In the resulting 180 SNP sets, we identified 583 haplotypes with frequency estimates >0.05 in the HapMap CEU sample. Differences in these 583 haplotype frequency estimates between the Finnish and HapMap CEU samples were modest (Fig. 4a), and the Pearson correlation between the two samples was 0.96. The similarity in haplotype frequencies was consistent with our results for allele frequencies (Fig. 1): 67.8, 93.6, and 98.0% of the 583 haplotypes had absolute frequency difference <0.05, <0.10, and <0.15, respectively, in the two samples (Fig. 4b). Incorrectly assuming all haplotypes were phase-known, among 583 observed haplotypes, 120 (20.5%) showed significant differences in frequency estimates in the two

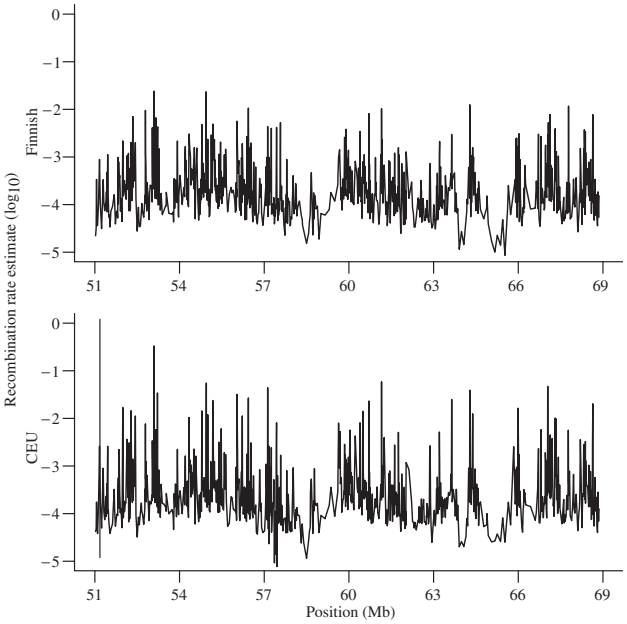

Fig. 5. Estimated recombination rates from the Finnish (top) and HapMap CEU samples (bottom) calculated using Phase [Stephens and Donnelly, 2003].

populations at the 0.05 level, 49 (8.4%) at the 0.01 level, and 19 (3.3%) at the 0.001 level. We obtained similar results using the Gabriel D′ block definitions [Gabriel et al., 2002] implemented in HaploView [Barrett et al., 2005] to define haplotype blocks using the larger Finnish sample. We also calculated the haplotype diversity for each SNP bin [Nei, 1987] and found extremely similar measures of haplotype diversity in the two samples (Pearson correlation 0.96).

We examined haplotypes that were observed in one sample but likely unobserved in the other sample, defined as a frequency estimate <0.001. Of 583 haplotypes present in the HapMap CEU sample with estimated frequency $>5\%$, 4 (0.69%) were unobserved in the Finnish sample, and, similarly, of 563 haplotypes with frequency estimates in the Finnish sample $>5\%$, only 6 (1.1%) were unobserved in the CEU sample. We also examined whether there were instances of haplotypes with frequency $<5\%$ in one sample and $>10\%$ in the other. Of 357 haplotypes with frequency estimates $>10\%$ in the CEU sample, we observed 5 haplotypes with frequencies ranging from 10.8–13.3% in the CEU and 3.3–4.1% in the Finnish sample. Conversely, of 386 haplotypes with frequency estimates $>10\%$ in the Finnish sample, we observed 11 haplotypes with frequency estimates ranging from 10.6–16.0% in the Finnish sample and 0.8–4.6% in the CEU sample.

Estimated recombination rates showed strong correlations between Finnish non-diabetic individuals and the CEU individuals (Fig. 5). The Pearson correlation was 0.82.

DISCUSSION

We examined the use of the HapMap samples for selecting tag SNPs for a case-control association study of T2D in Finnish individuals. Allele frequencies for most SNPs were similar for the HapMap CEU and Finnish samples; 90% of SNPs had a frequency in Finnish individuals within 10% of the HapMap CEU frequency. Although we observed more SNPs with significantly different allele frequencies between the CEU and Finnish samples than expected by chance, the magnitude of the allele frequency differences generally were small and did not reduce the utility of tag SNPs to represent untyped SNPs with r^2 estimates, at least in our limited data of SNP pairs with $r^2>0.8$ in the CEU. This was demonstrated by the r^2 values estimated for pairs of adjacent SNPs from the HapMap CEU sample and the Finnish sample, which were similar; 75.8% of adjacent marker pairs had an absolute r^2 difference less than 0.10. r^2 estimates for pairs of non-adjacent SNPs at further distances also showed remarkable similarity between the CEU and Finnish samples (Fig. 2). We also identified sets of three to 20 SNPs in moderate to high LD and estimated frequencies for the resulting haplotypes. We found that the haplotype frequencies were quite similar between the two samples regardless of the LD level at which the haplotype sets were constructed, consistent with the similarities in the estimated allele frequencies. Haplotype diversity can show differences between populations that may interfere with the ability of SNPs selected based on one population to adequately distinguish haplotypes in another population [Beaty et al., 2005]. However, in the Finnish and CEU samples examined here, haplotype diversity based on estimated haplotype frequencies within each marker bin showed strong similarity between the two samples (Pearson correlation 0.96), as were estimates of recombination rate (Fig. 5), which had a Pearson correlation of 0.82. In contrast, allele frequency (Fig. 1), r^2 (data not shown), and haplotype frequency estimates (data not shown) in our Finnish samples were very different from those

for the HapMap YRI, CHB, and JPT samples; this was expected, as other studies have shown differences between European, Chinese, and African samples [Hinds et al., 2005]. However, the finding that both common and rare haplotypes are often shared between the four HapMap samples [International HapMap Consortium, 2005] and evidence of strong similarities in allele frequencies from European samples [Rosenberg et al., 2002] supports our observation that nearly all haplotypes observed in the Finnish sample were also observed in the CEU sample.

Our 17.9-megabase (Mb) region of chromosome 14 is in many ways a typical region of the genome. It contains 134 unique transcripts (7.55 genes per Mb) per Ensembl [Birney et al., 2004], similar to the estimate of 7.49 genes per Mb in the remainder of the autosomal genome (20,844 genes in 2,782 Mb). The percentage of nucleotides that are observed in cDNAs (transcribed) is 1.7% in our region and 1.6% in the rest of the genome (range of individual autosomes: 0.89–4.3%). The percentage of nucleotides defined as translated into protein by Ensembl is 0.96% in our region and 0.97% in the rest of the genome (range for individual autosomes: 0.53–2.9%). The ratio of centiMorgans (cM) to megabases (Mb) for this region of chromosome 14 was 1.16, and the autosomal genome rate estimated from published linkage maps [Kong et al., 2004] was 1.14 cM/Mb (range of individual chromosomes (0.98–2.1). These observations suggest there is nothing unusual about this region of chromosome 14 that would prevent generalization of our results to the rest of the genome.

The primary limitation of our study is that we did not genotype all 3,481 SNPs in both the HapMap CEU and Finnish samples to determine the similarity of allele frequency, r^2, and haplotype frequency estimates, and so can make comparisons only for the 957 SNPs selected based on the HapMap CEU samples for genotyping in the Finns. In particular, our study provides limited information on SNP pairs with $r^2 > 0.80$. Still, the data we have in that range suggest a strong similarity between the two samples. It also must be acknowledged that any SNP selection based on HapMap samples, even if for one of the four genotyped populations, is limited to the variability captured by the modest numbers of samples genotyped in HapMap. SNPs that are untyped in the reference population with no surrogate in HapMap because of modest LD will not be selected for genotyping and will remain untested for potential disease association.

Several previous studies have examined similarities and differences between LD measures and tag SNPs in various sets of populations. As expected, closely related populations generally gave similar results, while more distantly related populations gave more divergent results. Nejentsev et al. [2004] reported the utility of tag SNP selection in the region of the vitamin D receptor gene in a sample comprised mainly of European families with at least one individual affected with type 1 diabetes. They genotyped 68 SNPs with MAF >10% in 1,635 individuals sampled from Africans from the Gambia and from four European populations: British, Finnish, Norwegian, and Romanian. They concluded that tag SNP selection based on one European population was effective at predicting tag SNPs in the other Europeans population, but that using European samples to choose tag SNPs for Gambians or Gambians to choose tag SNPs for Europeans was not as effective.

Mueller et al. [2005] genotyped 1,218 individuals sampled from nine European populations for 169 SNPs with MAF >5% in four genomic regions totaling ~750 kb. They compared the proportion of population-specific markers, location of haplotype block boundaries, haplotype frequencies, and tag SNP transferability. Using either pairwise r^2 or haplotype-based tag SNP selection, they concluded that most tag SNPs did not show strong population differences.

Evans and Cardon [2005] examined data on 4,107 SNPs in a 10-Mb region of chromosome 20 in samples from four populations: 42 Asians from Japan and China, 97 African-Americans, 46 CEPH individuals, and 96 UK individuals of western European ancestry. They saw a strong Spearman correlation of 0.95 between r^2 estimates for the CEPH and UK samples. Dawson et al. [2002] also observed highly similar patterns of LD between CEPH trios and UK individuals on chromosome 22, while Tapper et al. [2003] observed similarities in linkage disequilibrium unit (LDU) maps based on these same chromosome 22 data.

LD patterns differed more substantially when more diverse populations were compared. Sawyer et al. [2005] concluded that patterns of LD in three small genomic regions differed substantially among the sixteen populations they studied: four European populations (Druze, Danes, Irish, and European-American), seven African populations (Biaka, Mbuti, Yoruba, Ibo, Hausa, Ethiopian, and

African-American), and five Asian populations (Japanese, Nasioi, Yakut, PimaMX, and R. surui). However, the authors reported similarity of haplotype frequencies for samples from the same geographic region. Examination of one of the regions suggested strong similarity of haplotype-block boundaries between the European samples [Liu et al., 2004].

In summary, we observed strong correlations in allele frequency, r^2, haplotype frequency, and recombination rate estimates between the HapMap CEU sample and our Finnish sample for 957 SNPs in a 17.9-Mb region of chromosome 14. Our results suggest that the HapMap CEU samples provide an adequate basis for tag SNP selection in Finnish individuals. This finding is consistent with several previous studies and suggests that HapMap data will prove useful for the design of association studies in populations beyond the four that were genotyped.

ACKNOWLEDGMENTS

We thank the participants in the FUSION, Finrisk 2002, and HapMap studies for generously donating samples for research. We gratefully acknowledge our FUSION colleagues Karen Conneely, Mike Erdos, Terry Gliedt, Andrew Skol, and Peggy White for helpful discussions and assistance with data preparation, our CIDR colleagues for carrying out the genotyping studies on our Finnish samples, and the International HapMap Consortium for making available genotypes for the CEU, YRI, JPT, and CHB samples. This research was supported by intramural funds from the National Human Genome Research Institute (OH95-C-N030). CIDR is fully funded through federal contract N01-HG-65403 from the National Institutes of Health to The Johns Hopkins University. K.L.M. is supported in part by a Career Award in the Biomedical Sciences from the Burroughs Wellcome Fund.

REFERENCES

Abecasis GR, Cherny SS, Cookson WO, Cardon LR. 2002. Merlin: rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30:97–101.

Alf EF Jr, Graf RG. 1999. Asymptotic confidence limits for the difference between two squared multiple correlations: a simplified approach. Psych Methods 4:70–75.

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

Beaty TH, Fallin MD, Hetmanski JB, McIntosh I, Chong SS, Ingersoll R, Sheng X, Chakraborty R, Scott AF. 2005. Haplotype diversity in 11 candidate genes across four populations. Genetics 171:259–267.

Birney E, Andrews TD, Bevan P, Caccamo M, Chen Y, Clarke L, Coates G, Cuff J, Curwen V, Cutts T, Down T, Eyras E, Fernandez-Suarez XM, Gane P, Gibbins B, Gilbert J, Hammond M, Hotz HR, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Keenan S, Lehvaslaiho H, McVicker G, Melsopp C, Meidl P, Mongin E, Pettett R, Potter S, Proctor G, Rae M, Searle S, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Ureta-Vidal A, Woodwark KC, Cameron G, Durbin R, Cox A, Hubbard T, Clamp M. 2004. An overview of Ensembl. Genome Res 14:925–928.

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.

Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, Pabial J, Dibling T, Tinsley E, Kirby S, Carter D, Papaspyridonos M, Livingstone S, Ganske R, Lohmussaar E, Zernant J, Tonisson N, Remm M, Magi R, Puurand T, Vilo J, Kurg A, Rice K, Deloukas P, Mott R, Metspalu A, Bentley DR, Cardon LR, Dunham I. 2002. A first-generation linkage disequilibrium map of human chromosome 22. Nature 418:544–548.

Evans DM, Cardon LR. 2005. A comparison of linkage disequilibrium patterns and estimated population recombination rates across multiple populations. Am J Hum Genet 76: 681–687.

Fan JB, Oliphant A, Shen R, Kermani BG, Garcia F, Gunderson KL, Hansen M, Steemers F, Butler SL, Deloukas P, Galver L, Hunt S, McBride C, Bibikova M, Rubano T, Chen J, Wickham E, Doucet D, Chang W, Campbell D, Zhang B, Kruglyak S, Bentley D, Haas J, Rigault P, Zhou L, Stuelpnagel J, Chee MS. 2003. Highly parallel SNP genotyping. Cold Spring Harb Symp Quant Biol 68:69–78.

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.

Gunderson KL, Kruglyak S, Graige MS, Garcia F, Kermani BG, Zhao C, Che D, Dickinson T, Wickham E, Bierle J, Doucet D, Milewski M, Yang R, Siegmund C, Haas J, Zhou L, Oliphant A, Fan JB, Barnard S, Chee MS. 2004. Decoding randomly ordered DNA arrays. Genome Res 14:870–877.

Gunderson KL, Steemers FJ, Lee G, Mendoza LG, Chee MS. 2005. A genome-wide scalable SNP genotyping assay using microarray technology. Nat Genet 37:549–554.

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR. 2005. Whole-genome patterns of common DNA variation in three human populations. Science 307:1072–1079.

International HapMap Project. 2003. Nature 426:789–796.

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, Sangiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J. 2005. Complement factor H polymorphism in age-related macular degeneration. Science 308:385–389.

Kong X, Murphy K, Raj T, He C, White PS, Matise TC. 2004. A combined linkage-physical map of the human genome. Am J Hum Genet 75:1143–1148.

Liu N, Sawyer SL, Mukherjee N, Pakstis AJ, Kidd JR, Kidd KK, Brookes AJ, Zhao H. 2004. Haplotype block structures show

significant variation among populations. Genet Epidemiol 27: 385–400.

Matsuzaki H, Dong S, Loi H, Di X, Liu G, Hubbell E, Law J, Berntsen T, Chadha M, Hui H, Yang G, Kennedy GC, Webster TA, Cawley S, Walsh PS, Jones KW, Fodor SP, Mei R. 2004. Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. Nat Methods 1:109–111.

Mueller JC, Lohmussaar E, Magi R, Remm M, Bettecken T, Lichtner P, Biskup S, Illig T, Pfeufer A, Luedemann J, Schreiber S, Pramstaller P, Pichler I, Romeo G, Gaddi A, Testa A, Wichmann HE, Metspalu A, Meitinger T. 2005. Linkage disequilibrium patterns and tagSNP transferability among European populations. Am J Hum Genet 76:387–398.

Nei M. 1987. Molecular evolutionary genetics. New York: Columbia University Press.

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.

Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, Sato H, Hori M, Nakamura Y, Tanaka T. 2002. Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nat Genet 32:650–654.

Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. Science 273:1516–1517.

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

Saaristo T, Peltonen M, Lindstrom J, Saarikoski L, Eriksson J, Tuomilehto J. 2005. Cross-sectional evaluation of the Finnish Diabetes Risk Score: a tool to identify undetected type 2 diabetes, abnormal glucose tolerance and metabolic syndrome. Diabet Vasc Dis Res 2:67–72.

Sawyer SL, Mukherjee N, Pakstis AJ, Feuk L, Kidd JR, Brookes AJ, Kidd KK. 2005. Linkage disequilibrium patterns vary substantially among populations. Eur J Hum Genet 13: 677–686.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308–311.

Silander K, Scott LJ, Valle TT, Mohlke KL, Stringham HM, Wiles KR, Duren WL, Doheny KF, Pugh EW, Chines P, Narisu N, White PP, Fingerlin TE, Jackson AU, Li C, Ghosh S, Magnuson VL, Colby K, Erdos MR, Hill JE, Hollstein P, Humphreys KM, Kasad RA, Lambert J, Lazaridis KN, Lin G, Morales-Mena A, Patzkowski K, Pfahl C, Porter R, Rha D, Segal L, Suh YD, Tovar J, Unni A, Welch C, Douglas JA, Epstein MP, Hauser ER, Hagopian W, Buchanan TA, Watanabe RM, Bergman RN, Tuomilehto J, Collins FS, Boehnke M. 2004. A large set of Finnish affected sibling pair families with type 2 diabetes suggests susceptibility loci on chromosomes 6, 11, and 14. Diabetes 53:821–829.

Stephens M, Donnelly P. 2003. A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73:1162–1169.

Tapper WJ, Maniatis N, Morton NE, Collins A. 2003. A metric linkage disequilibrium map of a human chromosome. Ann Hum Genet 67:487–494.

The International HapMap Consortium. 2005. A haplotype map of the human genome. Nature 437:1299–1320.

Tishkoff SA, Verrelli BC. 2003. Role of evolutionary history on haplotype block structure in the human genome: implications for disease mapping. Curr Opin Genet Dev 13:569–575.

Valle T, Tuomilehto J, Bergman RN, Ghosh S, Hauser ER, Eriksson J, Nylund SJ, Kohtamaki K, Toivanen L, Vidgren G, Tuomilehto-Wolf E, Ehnholm C, Blaschak J, Langefeld CD, Watanabe RM, Magnuson V, Ally DS, Hagopian WA, Ross E, Buchanan TA, Collins F, Boehnke M. 1998. Mapping genes for NIDDM. Design of the Finland-United States Investigation of NIDDM Genetics (FUSION) Study. Diabetes Care 21:949–958.

WHO. 1985. Diabetes mellitus: Report of a WHO Study Group. Geneva: World Health Organization.