Published in final edited form as:
Hum Genet. 2009 April ; 125(3): 295–303. doi:10.1007/s00439-009-0627-8.

Use of weighted reference panels based on empirical estimates of ancestry for capturing untyped variation

Matthew R. L. Egyud,
Program in Genomics and Division of Endocrinology, Children's Hospital, 300 Longwood Ave., Boston, MA 02115, USA

Boston University School of Medicine, 72 East Concord St., Boston, MA 02118, USA

Zofia K. Z. Gajdos,
Program in Genomics and Division of Endocrinology, Children's Hospital, 300 Longwood Ave., Boston, MA 02115, USA

Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

Department of Genetics, Harvard Medical School, Boston, MA 02115, USA

Johannah L. Butler,
Program in Genomics and Division of Endocrinology, Children's Hospital, 300 Longwood Ave., Boston, MA 02115, USA

Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

Sam Tischfield,
Program in Genomics and Division of Endocrinology, Children's Hospital, 300 Longwood Ave., Boston, MA 02115, USA

Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

Loic Le Marchand,
Cancer Research Center, University of Hawaii, Honolulu, HI 96813, USA

Laurence N. Kolonel,
Cancer Research Center, University of Hawaii, Honolulu, HI 96813, USA

Christopher A. Haiman,
Department of Preventative Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90089, USA

Brian E. Henderson, and
Department of Preventative Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90089, USA

Joel N. Hirschhorn

Correspondence to: Joel N. Hirschhorn.

Electronic supplementary material The online version of this article (doi:10.1007/s00439-009-0627-8) contains supplementary material, which is available to authorized users.

Note: All research was performed in compliance with current laws of the United States of America.

Program in Genomics and Division of Endocrinology, Children's Hospital, 300 Longwood Ave., Boston, MA 02115, USA

Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

Department of Genetics, Harvard Medical School, Boston, MA 02115, USA

Abstract

Many association methods use a subset of genotyped single nucleotide polymorphisms (SNPs) to capture or infer genotypes at other untyped SNPs. We and others previously showed that tag SNPs selected to capture common variation using data from The International HapMap Consortium (Nature 437:1299–1320, 2005), The International HapMap Consortium (Nature 449:851–861, 2007) could also capture variation in populations of similar ancestry to HapMap reference populations (de Bakker et al. in Nat Genet 38:1298–1303, 2006; González-Neira et al. in Genome Res 16:323–330, 2006; Montpetit et al. in PLoS Genet 2:282–290, 2006; Mueller et al. in Am J Hum Genet 76:387–398, 2005). To capture variation in admixed populations or populations less similar to HapMap panels, a "cosmopolitan approach," in which all samples from HapMap are used as a single reference panel, was proposed. Here we refine this suggestion and show that use of a "weighted reference panel," constructed based on empirical estimates of ancestry in the target population (relative to available reference panels), is more efficient than the cosmopolitan approach. Weighted reference panels capture, on average, only slightly fewer common variants (minor allele frequency > 5%) than the cosmopolitan approach (mean r^2 = 0.977 vs. 0.989, 94.5% variation captured vs. 96.8% at $r^2 > 0.8$), across the five populations of the Multiethnic Cohort, but entail approximately 25% fewer tag SNPs per panel (average 538 vs. 718). These results extend a recent study in two Indian populations (Pemberton et al. in Ann Hum Genet 72:535–546, 2008). Weighted reference panels are potentially useful for both the selection of tag SNPs in diverse populations and perhaps in the design of reference panels for imputation of untyped genotypes in genome-wide association studies in admixed populations.

Introduction

Association studies are powerful tools to assess genomic variation for disease risk variants. Currently, it is impractical to genotype all known genetic markers, either genome-wide or for candidate genes, in order to identify those variants associated with a disease or trait. Typically, a subset of single nucleotide polymorphisms (SNPs) is genotyped, and these SNPs can serve as proxies for other untyped variants. These variants can be either a fixed set of SNPs on commercially available platforms (Maresso and Broeckel 2008), or tag SNPs chosen specifically to cover a particular gene or region with minimal redundancy (Carlson et al. 2004; Johnson et al. 2001). In either case, the genotyped SNPs (or combinations of them) are used as proxies for untyped SNPs, but this approach requires a reference panel in which the linkage disequilibrium (LD) relationships between genotyped and untyped variants have been determined. Importantly, the reference panel should have some-what similar ancestry to the individuals in the association study, because LD relationships vary across populations, particularly those that have recent ancestry from different continents (The International HapMap Consortium 2005, 2007).

For populations that have well defined genetic ancestry from a single geographic region, data from the International HapMap Consortium (The International HapMap Consortium 2005, 2007) can probably serve as a reference panel, assuming that the population being studied has a similar ancestry to one of the populations represented in the HapMap. The use of HapMap as a reference panel has been validated for selection of "tag SNPs" that efficiently capture common (minor allele frequency >5%) variation over a region (de Bakker

et al. 2006; González-Neira et al. 2006; Montpetit et al. 2006; Mueller et al. 2005), and for imputation of untyped variants from dense genotype data from populations of European ancestry (Zeggini et al. 2008). However, the most efficient and appropriate choice of reference panels is less clear for populations where there has been significant admixture or where the ancestry does not closely match one of the HapMap panels.

Several possible methods exist for selecting reference panels for a population whose ancestry does not correspond closely to that of an existing HapMap population. First, one can determine local LD structures in the region of interest—in essence creating a HapMap for the region of interest in the population. This is feasible across short regions, but requires extensive genotyping and is expensive. The planned future expansion of the HapMap populations to include seven additional populations worldwide (Coriell Institute for Medical Research 2008) should provide reference panels for some populations that currently do not have a clear analogue in HapMap. A second method would be to use all of the HapMap panels equally. For tag SNP selection, this “cosmopolitan tagging” method creates a comprehensive list of tag SNPs by first choosing tags using one population as a reference panel, and then generating sequentially larger supersets that also capture variation in additional HapMap reference panels (de Bakker et al. 2006). We previously showed that this method captures common variation well in a variety of populations, but at the cost of selecting a minimum of 22% additional tag SNPs compared with single reference panels (de Bakker et al. 2006).

We propose a third approach, employing a reference sample constructed from multiple HapMap samples, with different degrees of representation (weights), depending on the ancestry of the particular population being studied; the “target population.” In this approach, greater weight is placed on those HapMap panels that are more similar in ancestry to the target population. Specifically, we propose to use ancestry informative markers to estimate the ancestry of a target population relative to available HapMap panels, permitting the generation of a reference panel in which representation of HapMap populations is weighted according to these estimated ancestries. In theory, an ancestry-weighted panel should better mimic the target population than either an individual HapMap sample panel or a cosmopolitan panel containing equal representation of multiple HapMap samples. In this work, we assess how well and efficiently tag SNPs selected using each of these three approaches—single reference panel tagging, multiple reference panel tagging, and weighted multiple reference panel tagging—can capture common genetic variation across 25 genomic regions in five different racial/ethnic groups from the Multiethnic Cohort (MEC) (Kolonel et al. 2000).

Methods

DNA samples

The International HapMap Project aims to catalog common sequence variation in the human genome (The International HapMap Consortium 2005, 2007), (http://hapmap.org/). Currently, four populations have been well described: Utah residents of European ancestry (CEU), Han Chinese from Beijing, China (HCB), Japanese from Tokyo, Japan (JPT), and Yoruba from Ibadan, Nigeria (YRI). We used data and DNA samples from the CEU, HCB, JPT and YRI panels from the HapMap.

The Human Genome Diversity Panel (HGDP) includes 1,056 unrelated individuals from 52 worldwide populations (Cann et al. 2002). We used DNA samples from 25 individuals of African ancestry (HGDP-YRI) and 71 individuals from two populations of East Asian ancestry; 40 individuals from China (HGDP-HCB) and 31 individuals from Japan (HGDP-

JPT). Naming conventions for the HGDP populations were adopted from de Bakker et al. (2006).

The Multiethnic Cohort of Los Angeles and Hawai'i (MEC) is a collection of 215,251 adult men and women from Hawaii and Los Angeles County, California. The cohort was collected for the purpose of studying diet and cancer in the United States (Kolonel et al. 2000). The samples used in this study consist of samples from five self-reported racial/ethnic groups: African-Americans (MEC-AA, $N = 70$), Native Hawaiians (MEC-H, $N = 69$), Japanese-Americans (MEC-J, $N = 70$), US Latinos (MEC-L, $N = 70$), and Whites of presumed European ancestry (MEC-W, $N = 70$). The genotype data across 25 genomic regions generated in the MEC were described in de Bakker et al. (2006). Data from ancestry informative SNPs (see below) were generated for this current study.

Identification of ancestry informative markers

Ancestry informative markers (AIMs) were selected from autosomes using data from the HapMap release 21a: build 35 (The International HapMap Consortium 2007) data set and previously created admixture maps (Haiman et al. 2007; Smith et al. 2004; Tian et al. 2006). In the HapMap data, the 45 HCB and 45 JPT individuals were combined into a single population of East Asian ancestry (HCB + JPT). Our initial efforts to select ancestry informative markers from HapMap were enriched for markers with unrecognized allele flips in one or more populations, so we sought to select markers from regions where there was extended evidence of differentiation in allele frequency between populations. We focused on regions of at least 0.1 centimorgans in length with sustained $p_{\text{excess}} > 0.4$ (Bersaglieri et al. 2004). We calculated p_{excess} using the frequency in the HapMap population of interest and the average of the frequencies in the other two populations (for example, comparing the frequency in CEU to the average of the frequencies in YRI and HCB + JPT). Within these regions, markers with delta (the absolute difference in allele frequencies) values greater than 0.4 were identified. We chose a total of 50 SNPs, each with the highest delta in a given region, where genotyping success was at least 80%. To minimize false positives, we required the selected SNP to have a strongly correlated nearby SNP that was also a high delta SNP (pairwise $r^2 > 0.7$, algebraic difference between deltas for the two SNPs < 0.1).

These SNPs were validated by re-genotyping them in the HapMap samples; if the new genotypes yielded a delta greater than 5% different from that calculated from the HapMap data, the SNP was replaced. To the set of SNPs passing validation, additional SNPs from three existing sets of validated ancestry informative markers were added (Haiman et al. 2007; Smith et al. 2004; Tian et al. 2006). From this joint set of SNPs, 40 SNPs for each population (120 total) were selected so that each pair had an r^2 of <0.1 or were separated by at least 10 million base pairs. Successful SNPs also exhibited Hardy–Weinberg *p*-values greater than 0.01 in all four HapMap populations.

Genotyping and validation of AIMs

From the 120 AIMs, pools of SNPs for the iPLEX GOLD platform (Sequenom 2008) were designed, incorporating 99 of the SNPs. The 99 AIMs were genotyped in the MEC samples. Seventy-nine of these SNPs passed genotyping quality control checks, which consisted of >90% genotyping call rates in all of our samples and Hardy–Weinberg *p*-values greater than 0.01 in the MEC-W and MEC-J panels (Haiman et al. 2007). Failing markers were all due to low genotyping percentage, likely for technical reasons. To validate the AIMs, they were genotyped in samples closely resembling HapMap populations from Human Genome Diversity Panel, which contains individuals from 51 populations around the world (Cann et al. 2002). These samples were referred to as "HGDP-YRI," "HGDP-HCB," and "HGDP-JPT" by de Bakker et al. (2006).

Estimating ancestry relative to HapMap panels from AIM genotype data

To estimate the genetic ancestry of these populations relative to the HapMap panels, we used the program s*tructure* (Pritchard et al. 2000). We used the three HapMap reference panels (CEU, YRI, and combined HCB + JPT) as known populations and used *structure* to estimate the genetic ancestry of the test population relative to these reference points. The estimated percent contributions of each HapMap panel (by definition lying at the vertices on a triangle plot) were recorded for each population. The estimated ancestry relative to the HapMap panels was calculated as the average of percent contributions for each individual in the test population. All individuals, including apparent genetic outliers, were included in this average.

From the 79 passing SNPs, a random set of 27 SNPs (~1/3) was removed, and the *structure* analysis of the populations was run again using the remaining 52 SNPs. Percent contribution of HapMap was again recorded for each population to assess robustness of the previous estimates. This was performed five times for each MEC population.

Construction of weighted reference panels

We created weighted reference panels by using replicates (multiple identical copies) of HapMap panels selected to mimic the ancestry estimates of each target population from the *structure* analysis data. *Structure* estimates were rounded to the nearest 5%, and for each 5% we used 1 replicate of the appropriate HapMap panel (maximum number of replicates = 20; for example, the MEC-AA weighted reference panel utilized five copies of the entire CEU panel, 14 copies of YRI, and one copy of HCB + JPT). A different weighted reference panel was constructed for each target population.

Selection of tag SNPs

To analyze the 25 genomic regions studied in de Bakker et al. (2006), we filtered the dataset from this paper to include SNPs that genotyped successfully in all five MEC populations and in each HapMap population greater than 80% of the time. Genotype data for these SNPs in HapMap panels were obtained from the HapMap website (http://hapmap.org) (release 21a, build 35). SNPs with a minor allele frequency < 0.05 in all HapMap populations or in the test population were excluded from the analysis.

To select tag SNPs for a region, we used the *tagger* software package (de Bakker et al. 2005), implemented as part of the *haploview* software package (Barrett et al. 2005; Haploview 2008). We used *tagger* in pairwise mode, with a target $r^2 > 0.8$, using HapMap data from either a single reference panel (major ancestry tagging), multiple reference panels (cosmopolitan tagging), or weighted reference panels as input data. For the “cosmopolitan tagging” approach involving multiple reference panels, HapMap data from each of three panels–CEU, YRI, and combined HCB + JPT were used in succession in a three-stage approach. We first selected tags to capture common variation in the HapMap CEU panel. We then started with these tag SNPs and added additional tags to capture common variation in the HapMap YRI panel, and finally added additional tag SNPs to capture common variation in the combined HapMap HCB + JPT panel.

To evaluate the performance of a set of tag SNPs, we used the MEC data from de Bakker et al. (2006). For each test population, we ran *tagger* in evaluation mode, where the previously selected tags were force included and no additional tags were chosen. The max r^2 was recorded for each SNP in the region, and we recorded both the mean max r^2 and the fraction of targeted SNPs captured with $r^2 > 0.8$.

Results

We aimed to evaluate the utility of weighted reference panels for studying populations of mixed genetic ancestry. Specifically, we propose to create new artificial reference panels weighted according to ancestry estimates generated in the target population. We believe these new reference panels might more closely model the ancestry of the target population. In this paper, we projected the (unknown) ancestries of different target populations onto a 3-dimensional "HapMap space" (with the three axes in this space corresponding to the CEU, YRI and HCB + JPT samples). We then calculated the weights for each of the three main HapMap panels according to where the target populations fall in this space.

Estimating ancestry relative to HapMap reference panels using AIMs

To generate estimates of ancestry relative to the HapMap reference panels, we initially selected a set of appropriate ancestry informative markers (AIMs). We identified and evaluated AIMs from several sources (see "Methods" for more detail), and developed a working panel of 79 markers for estimating ancestry relative to the HapMap populations. To confirm the utility of these markers, we first genotyped them in the Human Genetic Diversity Panel, and confirmed that estimated ancestry conformed closely to the expected outcome (Supplementary Table 1). We also evaluated the markers on artificially generated populations created from HapMap data. Allele frequencies were determined for each HapMap population at each of the 79 loci, and weighted reference panels were created to artificially imitate genotype frequencies of a population of mixed HapMap ancestry. Estimated ancestry from *structure* closely conformed to expected values (Supplementary Table 1). Thus, this set of AIMs can be used to accurately estimate the proportion of ancestry from different continents.

We then used this panel of AIMs to provide weights for populations of different ancestry, including 5 self-reported racial/ethnic groups from the Multiethnic Cohort (MEC), (Table 1; Fig. 1a–e). MEC-J and MEC-W had population ancestries conforming closely to their expected related HapMap panels (97.3% HCB + JPT and 89.2% CEU respectively). Estimates of ancestry in MEC-H and MEC-L included contributions from CEU and HCB + JPT: MEC-L was ~2/3 (64.2%) CEU and 1/3 HCB + JPT (30.1%), while MEC-H was the inverse at ~1/3 CEU (35.6%) and 2/3 (62.1%) HCB + JPT. Estimates of ancestry in MEC-AA (MEC-AA) showed contributions from both CEU and YRI: ~3/4 (72.5%) YRI and 1/4 (23.2%) CEU. These estimates were then used to generate weighted reference panels for each population (see "Methods").

Comparing weighted reference panels with single reference panel and cosmopolitan tagging

We compared the efficacy and efficiency of tag SNP selection using weighted reference panels with other methods, including tagging using single existing reference panels, and cosmopolitan tagging (de Bakker et al. 2006).

In each of the five MEC populations tested, cosmopolitan tagging captured the most common variation (range = 94.9–98.2%), using a uniform set of 718 tags in each (Table 2). Weighted reference panel tagging captured the second most common variation in each of the five populations tested (range = 93.4–96.6%), but always required fewer tags than cosmopolitan tagging (range = 352–702 tags). Single reference panel tagging captured the least common variation of the three methods (range = 84.4–91.8%), using the fewest tags in every population (range = 346–611). Where the weighted reference panel method led to an increase in the number of variants captured, the increase was consistent across the range of minor allele frequencies from 10 to 50% (Supplementary Table 2). Complexity of

population linkage disequilibrium (LD) structure and degree of admixture play a role in both the number of tags needed and percentage of variation captured in each population. For example, in the MEC Japanese–American samples, the number of tags selected using a weighted reference panel was, as expected, close to the number selected using a single reference panel. Despite a strong correlation to the HapMap CEU panel, the MEC-W samples required all three HapMap populations for constructing the weighted reference panel (perhaps due to a small amount of admixture in this particular sample), thus increasing the number of tags selected compared to using a single reference panel but also increasing the number of variants captured. For the three more strongly admixed samples, weighted reference panels offer a compromise between the accuracy of cosmopolitan tagging and the economy of single reference panel tagging. The most noticeable gains in efficiency were seen in the Latino and Hawaiian samples, where most of the extra SNPs captured with cosmopolitan tagging were also captured with the weighted reference panel method, but with approximately 20% fewer tags. This result probably reflects the absence of African ancestry in these admixed populations, which makes equal inclusion of the YRI panel, with its increased genetic diversity, particularly inefficient.

Determining robustness of ancestry estimates from weighted reference panels

Because ancestry informative markers might not accurately estimate the ancestry of different populations, we explored the robustness of our ancestry estimates and the effect of misestimation of ancestry on the weighted reference panel method. We removed a random subset of 27 AIMs from the set of 79 AIMs used in this study and estimated population ancestry using data from the remaining 52 AIMs (Supplementary Table 3). This was repeated five times. We found that estimates of ancestry were relatively stable despite removal of approximately 1/3 of the AIMs from analysis, suggesting that 79 markers is suitable for estimating ancestry relative to the three HapMap populations.

We also deliberately skewed our estimates of ancestry relative to the three HapMap populations, by overestimating or underestimating the relative weights of the reference populations, to determine the effects of misestimation of ancestry on tagging (Table 3). At $r^2 > 0.8$, weighted reference panel tagging across the 5 MEC populations based on these skewed estimates still performed well, indicating a degree of tolerance in estimation of ancestry and creation of the weighted reference panel (Table 3).

Discussion

We attempted to determine how well tag SNPs selected using weighted reference panels—panels comprised of HapMap samples and designed to mimic the estimated ancestry of the target population—capture common genetic variation (minor allele frequency > 5%) across populations of mixed genetic ancestry. We compared this method with two other methods: single reference panel (major ancestry) tagging and multiple reference panel (cosmopolitan) tagging. We found that weighted reference panel tagging is only slightly less effective at capturing common genetic variation than cosmopolitan tagging in populations of mixed ancestry (93.4–96.6% of all variation captured vs. 94.9–98.2% captured by cosmopolitan tagging). However, the weighted reference panel method utilized an average of ~180 fewer tag SNPs in each panel, meaning that weighted reference panels are substantially more efficient with only a slight loss in SNP coverage. Both weighted reference panels and cosmopolitan tagging methods capture more common variation than tags selected using a single reference panel.

A recent publication by Pemberton et al. (2008), published while our work was in progress, utilized a similar method to ours to select tag SNPs for 30 individuals of Indian ancestry.

Our study extends these results to additional admixed populations and larger sample sizes, and strongly confirms the utility of weighted reference panel in tag SNP selection.

Tags from weighted reference panels constructed by deliberately skewing the estimated ancestry relative to the HapMap populations still capture a similar amount of variation as properly constructed weighted reference panels. Thus, this method is robust to slight misestimation of ancestry in the target population. Furthermore, based on these estimates of robustness, it is likely that weighted reference panels designed to mimic predicted ancestral contributions to closer than the nearest 5% would not substantially improve the performance of tag SNPs, and any gains are probably not worth the extra computational cost incurred by using the larger reference panels required to achieve a greater level of precision. Finally, at least for the YRI, CEU and HCB + JPT HapMap samples, a modest set of AIMs is sufficient to adequately estimate ancestry. More AIMs may be required to distinguish between more closely related populations, but extremely accurate estimation of relative contributions to ancestry by closely related populations may be less critical for tag SNP selection, so a modest number of AIMs may suffice even when weighted reference panels are constructed from more similar groups of samples.

The expansion of the HapMap panels to include populations of additional and/or admixed ancestries will lessen the need for new weighted reference panels for particular populations that have close correlates in the new HapMap samples. However, the weighted reference approach may also be strengthened by the addition of new samples, since this will effectively expand the number of dimensions in “HapMap space” and thereby permit the construction of weighted reference panels that even more closely mimic a diverse range of populations from around the world. Despite the highly imperfect proxies for ancestral populations of Latinos and Hawaiians represented in the original HapMap (HCB + JPT utilized as one ancestral population for both samples), the weighted reference panel performed equally well in these and the other three samples: Whites, Japanese, and African–Americans. This result suggests that LD patterns around common variation in the HCB + JPT samples may not be dramatically different than in the actual ancestral populations.

A limitation to our study was that we only considered pairwise tagging. Multimarker tagging is a more efficient method for capturing additional variation and reducing the size of the tag set (Chapman et al. 2003; Clayton et al. 2004; de Bakker et al. 2006), and the impact of different reference panels on multimarker tagging would need to be assessed, but weighted reference panels might be expected to provide similar benefits given their closer estimation of the underlying LD patterns in the populations being studied. Indeed, even more sophisticated methods have been developed to impute genotypes at untyped markers, and these also require reference panels. These methods combine genotype data at markers typed in the target population with data at untyped markers in a reference panel to infer the likely genotypes at untyped markers in the target population (Marchini et al. 2007; Li et al. 2008). These methods essentially depend on finding the most likely match between an extended haplotype in the target population and in the reference panel. Using a weighted reference panel should in theory increase the likelihood of selecting the correct haplotype, which could increase the accuracy of imputation in populations that do not closely match an existing HapMap population. Although this possibility would need to be verified empirically, the utility of weighted reference panels in tag SNP selection suggests that they may also be useful in imputation, thereby increasing the reach of genome-wide association studies in populations of diverse ancestries.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank P. de Bakker of the Broad Institute for support in analyzing the generated data, and members of the Hirschhorn lab for helpful discussions. Finally, we extend our deepest thanks to the participants of the International HapMap Project, Multiethnic Cohort, and Human Genome Diversity Panel, without whom none of this work would have been possible. This work was supported by grant R01DK075787 to JNH and a Strategic Program for Asthma Research award to JNH from the American Asthma Foundation.

References

1. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of ld and haplotype maps. Bioinformatics. 2005; 21:263–265. [PubMed: 15297300]
2. Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet. 2004; 74:1111–1120. [PubMed: 15114531]
3. Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, Chen Z, Chu J, Carcassi C, Contu L, Du R, Excoffier L, Ferrara GB, Friedlaender JS, Groot H, Gurwitz D, Jenkins T, Herrera RJ, Huang X, Kidd J, Kidd KK, Langaney A, Lin AA, Mehdi SQ, Parham P, Piazza A, Pistillo MP, Qian Y, Shu Q, Xu J, Zhu S, Weber JL, Greely HT, Feldman MW, Thomas G, Dausset J, Cavalli-Sforza LL. A human genome diversity cell line panel. Science. 2002; 296:261–262. [PubMed: 11954565]
4. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet. 2004; 74:106–120. [PubMed: 14681826]
5. Chapman JM, Cooper JD, Todd JA, Clayton DG. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and determinants of statistical power. Hum Hered. 2003; 56:18–31. [PubMed: 14614235]
6. Clayton D, Chapman J, Cooper J. Use of unphased multilocus genotype data in indirect association studies. Genet Epidemiol. 2004; 27:415–428. [PubMed: 15481099]
7. Coriell Institute for Medical Research. Available at http://ccr.coriell.org/Sections/Collections/NHGRI/hapmap.aspx?PgId=266
8. de Bakker PI, Yelenski R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. Efficiency and power in genetic association studies. Nat Genet. 2005; 37:1217–1223. [PubMed: 16244653]
9. de Bakker PI, Burtt N, Graham RR, Guiducci C, Yelensky R, Drake JA, Bersaglieri T, Penney KL, Butler J, Young S, Onofrio RC, Lyon HN, Stram DO, Haiman CA, Freedman ML, Zhu X, Cooper R, Groop L, Kolonel LN, Henderson BE, Daly MJ, Hirschhorn JN, Altshuler D. Transferability of tag SNPs in genetic association studies in multiple populations. Nat Genet. 2006; 38:1298–1303. [PubMed: 17057720]
10. González-Neira A, Ke X, Lao O, Calafell F, Navarro A, Comas D, Cann H, Bumpstead S, Ghori J, Hunt S, Deloukas P, Dunham I, Cardon LR, Bertranpetit J. The portability of tag SNPs across populations: a worldwide survey. Genome Res. 2006; 16:323–330. [PubMed: 16467560]
11. Haiman CA, Patterson N, Freedman ML, Myers SR, Pike MC, Wal-iszewska A, Neubauer J, Tandon A, Schirmer C, McDonald GJ, Greenway SC, Stram DO, Le Marchand L, Kolonel LN, Frasco M, Wong D, Pooler LC, Ardlie K, Oakley-Girvan I, Whittemore AS, Cooney KA, John EM, Ingles SA, Altshuler D, Henderson BE, Reich D. Multiple regions within 8q24 independently affect risk for prostate cancer. Nat Genet. 2007; 39:638–644. [PubMed: 17401364]
12. Haploview. 2008. Available at http://www.broad.mit.edu/mpg/haploview/
13. Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RC, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SC, Clayton DG, Todd JA. Haplotype tagging for the identiWcation of common disease genes. Nat Genet. 2001; 29:233–237. [PubMed: 11586306]
14. Kolonel LN, Henderson B, Hankin JH, Nomura AM, Wilkens LR, Pike MC, Stram DO, Monroe KR, Earle ME, Nagamine FS. A Multiethnic Cohort in Hawaii and Los Angeles: baseline characteristics. Am J Epidemiol. 2000; 151:346–357. [PubMed: 10695593]

15. Li Y, Willer CJ, Ding J, Sheet P, Abecasis GR. Rapid Markov chain haplotyping and genotype inference (Submitted). 2008

16. Marchini J, Howie B, Myers S, Mcffean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet. 2007; 39:906–913. [PubMed: 17572673]

17. Maresso K, Broeckel U. Genotyping platforms for mass-throughput genotyping with SNPs, including human genome-wide scans. Adv Genet. 2008; 60:107–139. [PubMed: 18358318]

18. Montpetit A, Nelis M, Laflamme P, Magi R, Ke X, Remm M, Cardon L, Hudson TJ, Metspalu A. An evaluation of the performance of tag SNPs derived from HapMap in a Caucasian population. PLoS Genet. 2006; 2:282–290.

19. Mueller JC, Lõhmussaar E, Mägi 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. Linkage disequilibrium patterns and tag SNP transferability among European populations. Am J Hum Genet. 2005; 76:387–398. [PubMed: 15637659]

20. Pemberton TJ, Jakobsson M, Conrad DF, Coop G, Wall JD, Pritchard JK, Patel PI, Rosenberg NA. Using population mixtures to optimize the utility of genomic databases: linkage disequilibrium and association study design in India. Ann Hum Genet. 2008; 72:535–546. [PubMed: 18513279]

21. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000; 155:945–959. [PubMed: 10835412]

22. Sequenom. 2008. Available at http://www.sequenom.com

23. Smith MW, Patterson N, Lautenberger JA, Truelove AL, McDonald GJ, Waliszewska A, Kessing BD, Malasky MJ, Scafe C, Le E, De Jager PL, Mignault AA, Yi Z, De The G, Essex M, Sankale JL, Moore JH, Poku K, Phair JP, Goedert JJ, Vlahov D, Williams SM, TishkoV SA, Winkler CA, De La Vega FM, Woodage T, Sninsky JJ, Hafler DA, Altshuler D, Gilbert DA, O'Brien SJ, Reich D. A high-density admixture map for disease gene discovery in African Americans. Am J Hum Genet. 2004; 74:1001–1013. [PubMed: 15088270]

24. The International HapMap Consortium. A haplotype map of the human genome. Nature. 2005; 437:1299–1320. [PubMed: 16255080]

25. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007; 449:851–861. [PubMed: 17943122]

26. Tian C, Hinds D, Shigeta R, Kittles R, Ballinger D, Seldin M. A genome-wide single-nucleotide-polymorphism panel with high ancestry information for African American admixture mapping. Am J Hum Genet. 2006; 79:640–649. [PubMed: 16960800]

27. Zeggini E, Scott L, Saxena R, Voight BF, Marchini JL, Hu T, de Bakker PI, Abecasis GR, Almgren P, Andersen G, Ardlie K, Boström KB, Bergman RN, Bonnycastle LL, Borch-Johnsen K, Burtt NP, Chen H, Chines PS, Daly MJ, Deodhar P, Ding CJ, Doney AS, Duren WL, Elliott KS, Erdos MR, Frayling TM, Freathy RM, Gianniny L, Grallert H, Grarup N, Groves CJ, Guiducci C, Hansen T, Herder C, Hitman GA, Hughes TE, Isomaa B, Jackson AU, Jørgensen T, Kong A, Kubalanza K, Kuruvilla FG, Kuusisto J, Langenberg C, Lango H, Lauritzen T, Li Y, Lindgren CM, Lyssenko V, Marvelle AF, Meisinger C, Midthjell K, Mohlke KL, Morken MA, Morris AD, Narisu N, Nilsson P, Owen KR, Palmer CN, Payne F, Perry JR, Pettersen E, Platou C, Prokopenko I, Qi L, Qin L, Rayner NW, Rees M, Roix JJ, Sandbaek A, Shields B, Sjögren M, Steinthorsdottir V, Stringham HM, Swift AJ, Thorleifsson G, Thorsteinsdottir U, Timpson NJ, Tuomi T, Tuomilehto J, Walker M, Watanabe RM, Weedon MN, Willer CJ, Illig T, Hveem K, Hu FB, Laakso M, Stefansson K, Pedersen O, Wareham NJ, Barroso I, Hattersley AT, Collins FS, Groop L, McCarthy MI, Boehnke M, Altshuler D. Wellcome Trust Case Control Consortium. Meta-analysis of genome-wide association data and large-scale replication identiWes additional susceptibility loci for type 2 diabetes. Nat Genet. 2008; 40:638–645. [PubMed: 18372903]

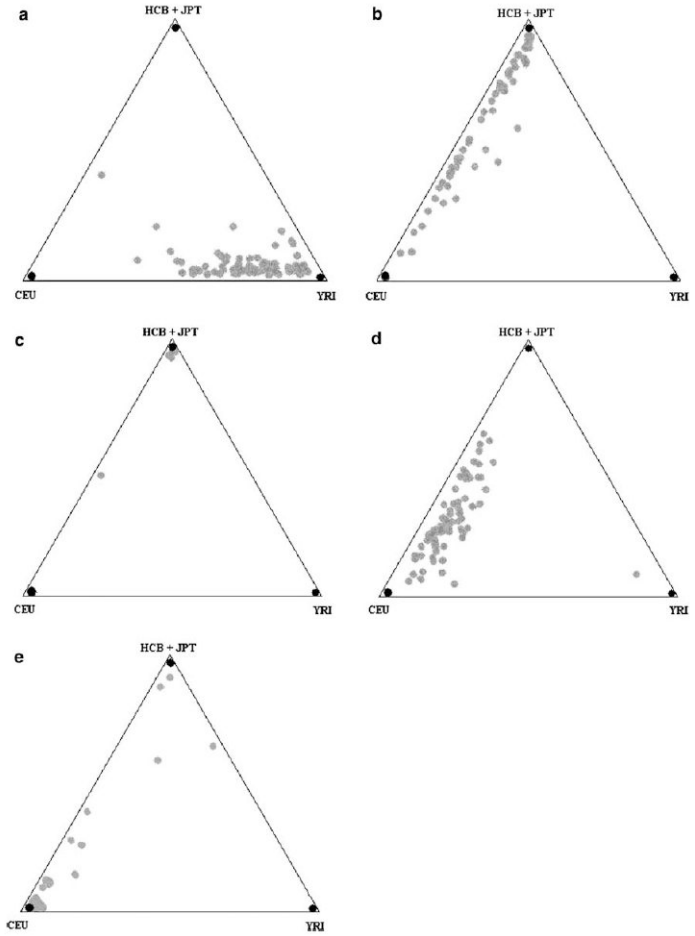

Fig. 1.

Estimated ancestry of five samples from the Multiethnic Cohort (MEC) relative to 3 HapMap populations. Ancestry for samples from self-described African–Americans (**a**), structure analysis of HapMap versus MEC-AA. Native Hawaiians (**b**), structure analysis of HapMap versus MEC-H. Japanese (**c**), structure analysis of HapMap versus MEC-J. Latinos (**d**), structure analysis of HapMap versus MEC-L and 'whites' (**e**), structure analysis of HapMap versus MEC-W, as described in de Bakker et al. (2006) are shown. Vertices are the three HapMap populations, which were treated as known samples in *structure* (Pritchard et al. 2000). *Each grey spot* represents one individual from the selected MEC panel

Table 1

Estimated ancestry of Multiethnic Cohort (MEC) panels relative to three HapMap populations

Population	%CEU	%YRI	%HCB + JPT	*n*
MEC-AA	23.2	72.5	4.4	70
MEC-H	35.7	2.2	62.1	69
MEC-J	1.9	0.8	97.3	70
MEC-L	64.2	5.7	30.1	70
MEC-W	89.2	2.5	8.4	70

%CEU, %YRI, and %HCB + JPT indicate estimated genetic ancestry in each MEC population relative to each of the corresponding HapMap population, as determined by *structure* (Pritchard et al. 2000). MEC-AA, MEC-H, MEC-J, MEC-L, and MEC-W correspond to the self-described African–American, Native Hawaiian, Japanese, Latino, and ‘white’ samples from the MEC described in de Bakker et al. (2006). Weighted reference panels were designed to approximate these ancestries by rounding any population with over 1% estimated contribution to the nearest 5%, starting with the least-represented population, followed by the most represented population, and within the constraint of the contributions summing to 100%. Populations with an estimated contribution between 1 and 5% were rounded up to 5%. These weights are given in Table 3

Table 2

Performance of tag SNPs selected using different reference panels in samples from the Multiethnic Cohort (MEC)

Population	Reference panel	Average % captured	Average max r^2	No. of tags selected	No. of SNPs captured	No. of SNPs total
MEC-AA	Single	91.8	0.981	611	890	976
	Multiple	94.9	0.983	718	922	976
	Weighted	94.1	0.982	702	921	976
MEC-H	Single	86.4	0.973	362	773	937
	Multiple	97.7	0.991	718	914	937
	Weighted	94.6	0.979	531	883	937
MEC-J	Single	91.4	0.975	346	822	893
	Multiple	98.2	0.993	718	880	893
	Weighted	93.5	0.982	352	845	893
MEC-L	Single	84.4	0.967	375	760	924
	Multiple	96.2	0.988	718	886	924
	Weighted	93.4	0.972	566	852	924
MEC-W	Single	91.5	0.967	372	826	901
	Multiple	97.2	0.990	718	878	901
	Weighted	96.6	0.973	538	862	901
All	Single	89.1	0.973	413	814	N/A
	Multiple	96.8	0.989	718	896	N/A
	Weighted	94.5	0.977	538	873	N/A

Single reference panel (major ancestry tagging) uses genotype data from one HapMap population to select tag SNPs. Multiple reference panel (cosmopolitan tagging) uses genotype data from all three HapMap populations. Weighted reference panel uses genotype data from all three HapMap populations and is constructed based on empirical estimates of ancestry in the target population. Average % captured is the percent of HapMap SNPs in de Bakker et al. (2006) captured at $r^2 > 0.8$ using tag SNPs selected with the reference panel shown. Average max r^2 is the average of the maximum r^2 between a SNP and any of the tag SNPs. Number of tags selected is the number of tag SNPs chosen by *tagger* (de Bakker et al. 2005) using an r^2 threshold of 0.8 and pairwise tagging. Number of SNPs captured is the total number across the 25 genes analyzed. Number of SNPs total varies between populations due to differences in population linkage disequilibrium (LD) structure. Percent captured and max r^2 values shown are the averages of results from each of the 25 genes studied. The last set of rows gives the average across all five populations

Table 3

Robustness of tagging performance to misestimation of ancestry

Population	Reference panel	%CEU	%YRI	%HCB + JPT	Average % captured	Average max (r^2)	Number of tags selected	Number of SNPs captured	Number of SNPs total
MEC-AA	Weighted	20	75	5	94.1	0.982	702	921	976
	AA1 skewed	10	85	5	94.1	0.981	710	902	976
	AA2 skewed	30	65	5	93.0	0.982	703	898	976
MEC-H	Weighted	35	5	60	94.6	0.979	531	883	937
	H1 skewed	25	5	70	93.9	0.978	547	872	937
	H2 skewed	45	5	50	95.3	0.982	553	885	937
MEC-J	Weighted	5	0	95	93.5	0.982	352	845	893
	J1 skewed	15	0	85	92.8	0.981	536	832	893
	J2 skewed	0	15	85	96.8	0.986	569	870	893
MEC-L	Weighted	65	5	30	93.4	0.972	566	852	924
	L1 skewed	55	5	40	92.2	0.973	563	851	924
	L2 skewed	75	5	20	92	0.971	554	833	924
MEC-W	Weighted	90	5	5	96.6	0.973	538	862	901
	W1 skewed	80	15	5	95.9	0.980	595	863	901
	W2 skewed	95	0	5	92.4	0.971	532	821	901
All	Weighted	N/A	N/A	N/A	94.5	0.977	538	873	N/A
	Skewed	N/A	N/A	N/A	93.9	0.978	586	863	N/A

Weighted reference panels were compared with two panels in which the estimates of ancestry were deliberately skewed. %CEU, %YRI, and %HCB + JPT are the genetic contributions of each HapMap panel used in constructing the selected reference panel. Average % captured, average max r^2, number of tags selected, number of SNPs captured, and number of SNPs total are as in Table 2. Values shown are the averages of results from each of the 25 genes studied. The last set of rows gives the average across all 5 populations