The Pharmacogenomics Journal (2009) 9, 362–372

www.nature.com/tpj

ORIGINAL ARTICLE

Application of principal component analysis to pharmacogenomic studies in Canada

H Visscher[1], CJD Ross[1], M-P Dubé[2], AMK Brown[2,3], MS Phillips[2,3], BC Carleton[4] and MR Hayden[1]

[1]*Centre for Molecular Medicine and Therapeutics, Child and Family Research Institute, Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada;* [2]*Montreal Heart Institute Research Centre and Université de Montreal, Montreal, Quebec, Canada;* [3]*Montreal Heart Institute and Genome Québec Pharmacogenomics Centre, Montreal, Quebec, Canada and* [4]*Pharmaceutical Outcomes and Policy Innovations Programme, Department of Paediatrics and Faculty of Pharmaceutical Sciences, University of British Columbia, Vancouver, British Columbia, Canada*

Correspondence:
Dr MR Hayden, Centre for Molecular Medicine and Therapeutics, Child and Family Research Institute, University of British Columbia, 950 West 28th Avenue, Vancouver, BC, Canada V5Z 4H4.
E-mail: mrh@cmmt.ubc.ca

Received 8 January 2009; revised 26 June 2009; accepted 30 June 2009; published online 4 August 2009

Ethnicity can confound results in pharmacogenomic studies. Allele frequencies of loci that influence drug metabolism can vary substantially between different ethnicities and underlying ancestral genetic differences can lead to spurious findings in pharmacogenomic association studies. We evaluated the application of principal component analysis (PCA) in a pharmacogenomic study in Canada to detect and correct for genetic ancestry differences using genotype data from 2094 loci in 220 key drug biotransformation genes. Using 89 Coriell worldwide reference samples, we observed a strong correlation between principal component values and geographic origin. We further applied PCA to accurately infer the genetic ancestry in our ethnically diverse Canadian cohort of 524 patients from the GATC study of severe adverse drug reactions. We show that PCA can be successfully applied in pharmacogenomic studies using a limited set of markers to detect underlying differences in genetic ancestry thereby maximizing power and minimizing false-positive findings.
The Pharmacogenomics Journal (2009) **9,** 362–372; doi:10.1038/tpj.2009.36; published online 4 August 2009

Keywords: ethnicity; principal components analysis; association study

Introduction

Individuals manifest considerable variability in their response to drug treatment influenced by differences in gender, age, environment as well as genetic determinants. The importance of ethnicity with regards to drug response, including efficacy and risk of toxicity, has been long recognized.[1] Geographic and ethnic differences in the frequency of allele variants in genes that influence drug response such as drug-metabolizing enzymes, transporters and drug targets have been studied since the 1950s and provide a mechanistic basis for at least part of the differences in drug response between populations.[2–4] In addition, there can be ancestry-specific variants associated with serious adverse drug reactions (ADRs) that occur in a specific ethnicity. An important example is the HLA-B*1502 allele and carbamazepine-induced Stevens–Johnson syndrome, which has only been found in the Asian populations.[5]

Recently, with advances in technology and availability of data from the HapMap project, there have been many publications of large genome-wide association studies for several common diseases (see ref. 6 for a current list). Furthermore, the first two genome-wide association studies that explore genetic variability of drug response have now been published.[7,8] Most of these studies are performed using a homogeneous study cohort with cases and controls of similar geographic, ethnic or racial descent and individuals from other origins are excluded.[9,10] The reason is that underlying differences in allele frequencies

between individuals of different genetic ancestry, called population stratification, can cause systematic differences in allele frequencies between cases and controls and lead to false-positive associations.[11] This confounding bias can occur when differences in phenotypic frequencies (e.g. disease or drug toxicity) exist between different genetic subpopulations included in a study. Genetic markers that happen to have a high allele frequency in a subpopulation with a high phenotype frequency, which would also be overrepresented in cases of a study, could lead to spurious associations.[11,12] Combining different ethnic (sub) populations in one study can also mask true effects leading to false negatives, especially if the populations are ethnically distant.[13] It is therefore necessary to detect and correct for these ancestral genetic differences.

In our national study of severe ADRs in children (genotype-specific approaches to therapy in children or GATC), we are collecting samples from many surveillance sites across Canada.[14] Canada is an ethnically diverse country with visible minorities representing 16% of the total population and almost half the population in large metropolitan areas such as Toronto and Vancouver.[15] Many marriages and common-law unions in Canada are inter-ethnic unions. In Vancouver, for example, up to 8.5% of all unions are between people of different ethnicities.[15]

To maximize the statistical power in our pharmacogenomic association study, we need to include as many samples as possible, while minimizing possible false-positive associations due to population stratification. It is vital to control for the genetic ancestry of each sample. However, commonly used ethnic labels are often insufficient and inaccurate proxies of genetic ancestry especially in populations with extensive admixture[16] and some individuals do not know or wrongly assume their ethnicity. It is therefore imperative to apply other methods to determine genetic ancestry.

Several methods have been developed to detect and correct for population stratification in genetic association studies and to estimate genetic ancestry.[17] In many pharmacogenomics studies to date, ancestry informative markers (AIMs) or neutral markers have been used to detect genetic ancestry differences and to assign individuals to different populations using a model-based clustering method.[18,19] However, this method can be computationally intensive when run on thousands of markers for many individuals at once and the assignment of individuals to ancestry clusters is limited by an *a priori* assumption of the number of clusters.[20]

Genomic control is an alternative adjustment method that relies on a quantitative estimate of the degree of population stratification at reference single nucleotide polymorphisms (SNPs) used to adjust for any stratification that might be present in the tested SNPs. The method relies on the use of unselected random SNPs, because AIMs could artificially inflate the population structure estimate. However, this method is conservative as the same correction factor is used for all investigated markers.[12]

Principal component analysis (PCA) is an alternative method to detect and correct for population stratification. Principal component analysis was first applied to genetic data more than 30 years ago,[21] but it was not until more recently that a solid statistical basis was provided.[20,22] Principal component analysis is a mathematical method that reduces complex multidimensional data (in this case genotype data) to a smaller number of dimensions by calculating the main axes or 'principal components' (PCs) of variation. These components are orthogonal vectors that capture the maximum variability present in the data. The first component explains the most variation in the data, and each subsequent component accounts for another, smaller part of the variability. When applied to genotype data, these axes of variation have been shown to have a striking relationship with geographic origin. For example, PCA of genotype data from hundreds of thousands of loci in 1387 individuals from across Europe, shows how a two-dimensional genetic map based on the first and second PCs closely mirrors the geographic map of Europe and can be used to accurately estimate ancestral origin of samples.[23] The PCA method is simple to use and is an efficient method even with large datasets. Compared with other clustering methods, PCA does not assume a predefined number of expected ancestry clusters,[20] and it has the advantage of being valid when used directly with the SNPs genotyped in a study, provided that those are present at a sufficiently high number.[24]

To explore the benefits of PCA in pharmacogenomic studies in Canada, we applied PCA to genotype data from 2094 loci in 220 key drug biotransformation genes to detect and correct for genetic ancestry differences. Using Coriell reference samples we observed a strong correlation between PC values and geographic origin. In addition, we were able to infer the genetic ancestry of samples from the GATC study with high accuracy and to estimate the genetic ancestry of samples of unknown origin. This shows that PCA can successfully be applied in pharmacogenomics to correct for population stratification using a limited set of markers.

Results

We genotyped 89 Coriell worldwide reference samples using a customized pharmacogenomics SNP panel, which was designed to capture the genetic variation of 220 key drug biotransformation genes (see Material and methods section and Supplemental Table 1). These samples were chosen to reflect as much genetic variation as possible across the world. We then genotyped 524 patient samples from our GATC study (see Supplemental Table 2 for detailed patient clinical characteristics). To identify underlying genetic ancestry differences we performed PCA on the genotype data from both Coriell reference samples and GATC samples. The first two principal components (PC1 and PC2) represent the main axes of variation within this data and explained 4.62 and 3.46% of variation, respectively. We created scatter plots of these components to visualize these data.

Initially, we plotted PC1 and PC2 of the Coriell reference samples to assess the worldwide pattern of genetic variation

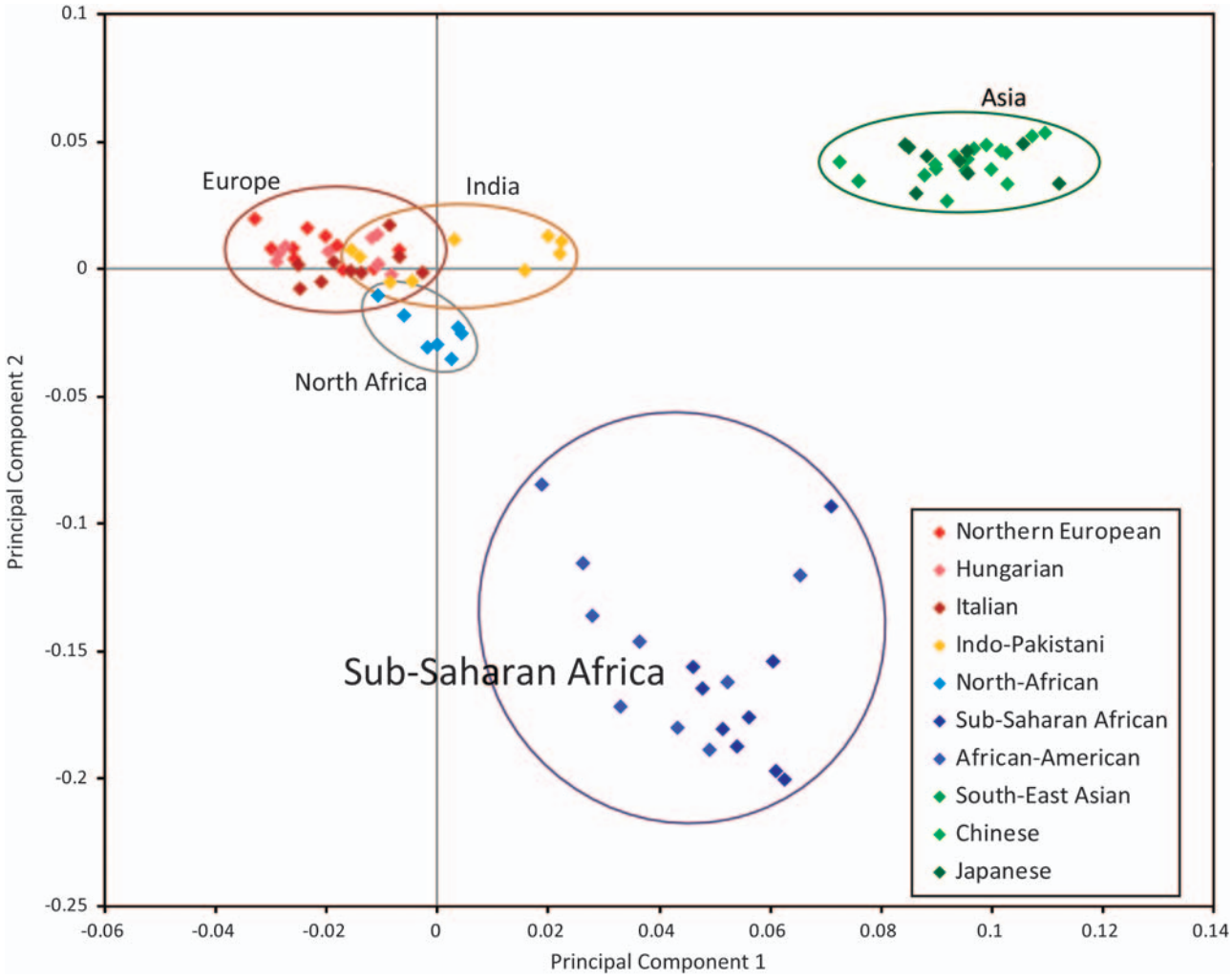

Figure 1 Scatter plot of principal component axis one (PC1) and axis two (PC2) based on genotype data of Coriell reference samples shows the pattern of genetic structure to resemble the worldwide geographic map. Individual data points are colored similarly by continental or ethnic origin (see legend).

for this pharmacogenomics panel (Figure 1). As expected, the cluster pattern resembled a geographic map of the world with the three continents Europe, Asia and Africa each on different points of the 'triangle', consistent with other reports.[25,26] The first PC distinguished between Europeans and East-Asians, with samples from the Indian subcontinent at intermediate values.

The second component (PC2) distinguished between Africans and non-Africans, with North Africans clustered between sub-Saharan Africans and Europeans. Within the African cluster there was more variability, which reflects the greater genetic diversity in samples of African ethnicity.[27]

Next, we plotted PCs of GATC samples for which the self-reported geographic origins of all four grandparents were from the same continental cluster (Europe, Asia or Africa) or India (Figure 2). Of these samples, almost all of the individuals fell within or close to their expected cluster with one notable exception; one of the East-Asian samples fell very close to the Indian cluster. This individual was of Singaporean origin and was therefore labeled as East-Asian. However, according to the census data, 8.9% of residents in Singapore are ethnic Indians.[28]

In large studies, participants may not know their ancestry or simply list themselves as 'Canadian'. For example, the Canadian census data list more than 30% of the total population responded as having 'Canadian' origins.[15] Of all GATC samples, parents of 107 individuals identified their family origin as from Canada. When we plotted the PCs of these samples, most individuals fell within the European cluster (Figure 3), which is not surprising given the large number of descendants from immigrants from Europe in Canada.[15]

Next, PCA was used to estimate the genetic ancestry of individuals from other geographic or ethnic origins that could not be easily placed in one of the earlier clusters (Figure 4a–e). For each of these groups only a small number of samples were available in our cohort. Of the six individuals of Caribbean origin, four showed a genetic

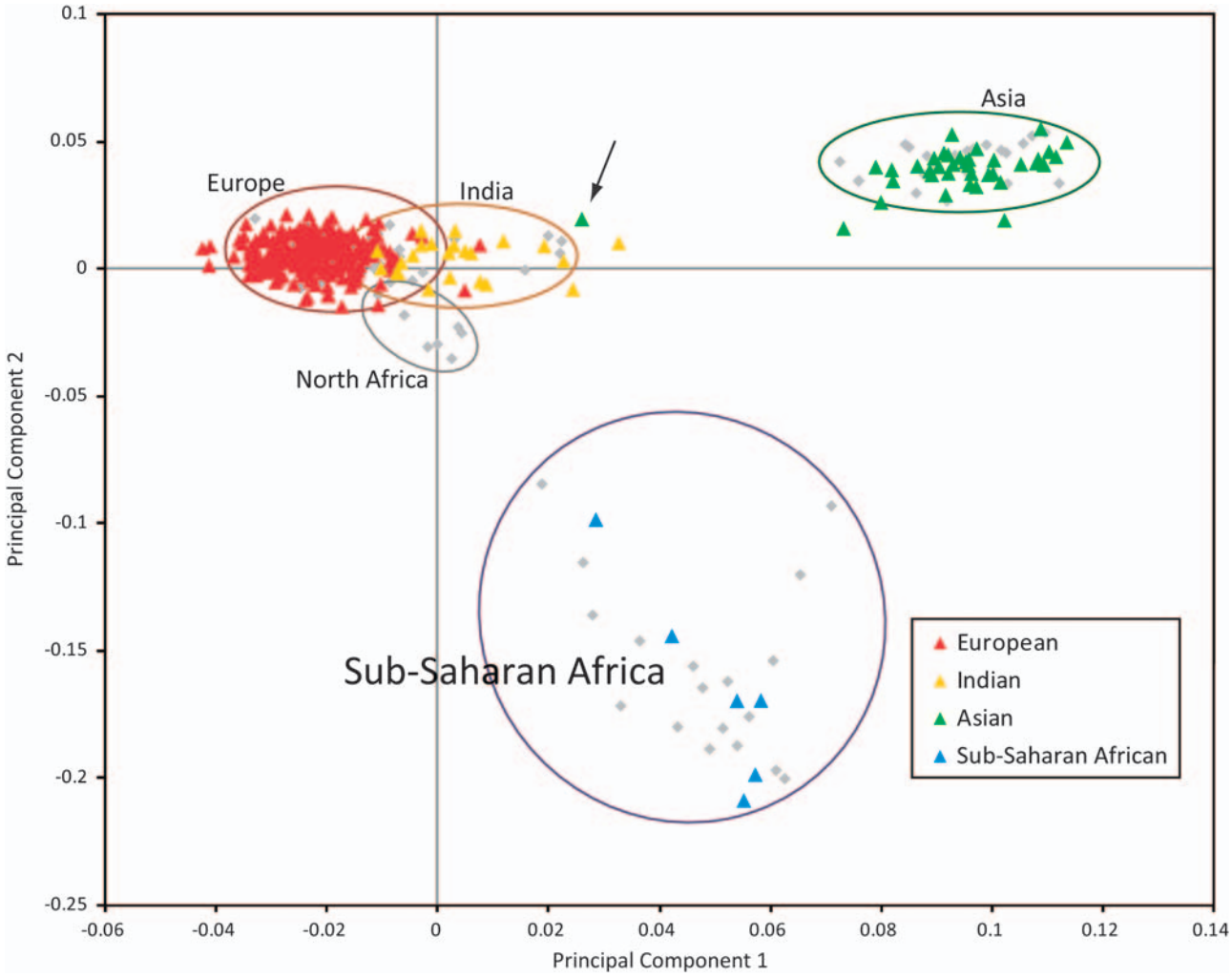

Figure 2 Scatter plot of PC1 and PC2 of GATC samples with self-reported geographic origin of all four grandparents from same continental cluster (Europe, Asia or Africa) or India. For reference, the Coriell samples are also plotted in light gray. Almost all individuals have values within or close to their expected clusters. One exception is one individual labeled as Asian with origins from Singapore that falls close to the Indian cluster (see arrow). This individual could well be a descendant from one of the many ethnic Indians in Singapore.

structure similar to sub-Saharan Africans and African-Americans and are most likely of Afro-Caribbean descent (Figure 4a). Two others had values close to the Indian and European clusters; again not unexpected given the history of immigration of Europeans and Indians to the Caribbean. The Canadian First Nations samples showed intermediate values between the European and East-Asian clusters. (Figure 4b) The native inhabitants of the Americas are thought to have originated from East-Asia,[29] so one might have expected values closer to East-Asians. Middle-Eastern individuals clustered near the border between the European and North-African clusters, which is consistent with what has been shown before.[26,30] (Figure 4c) Two Latin American individuals clustered between the three continental clusters, which likely reflects the admixture of Europeans, Africans and Native Americans over the last five centuries in South and Central America (Figure 4d). Interestingly, three individuals originating from the Fijian Islands in the South Pacific clustered within the Indian ancestry cluster. Several studies have shown that people originating from the South Pacific or Oceania are genetically closest to East-Asians.[31,32] However, demographic information shows that 37.1% of Fijians are of Indian descent,[33] because of immigration in colonial times, so these three samples are likely of Indian descent.

Plotting PC1 and PC2 for individuals of mixed origin showed that the majority of these individuals had intermediate values between their two continental clusters of origin (Figure 5a). In some cases, there was a clear genetic dose effect. For example, the three individuals with two grandparents from Europe and two from the Caribbean clustered directly between the African and European clusters, whereas one individual with one grandparent from the Caribbean and three from Europe fell closer to the European continental cluster.

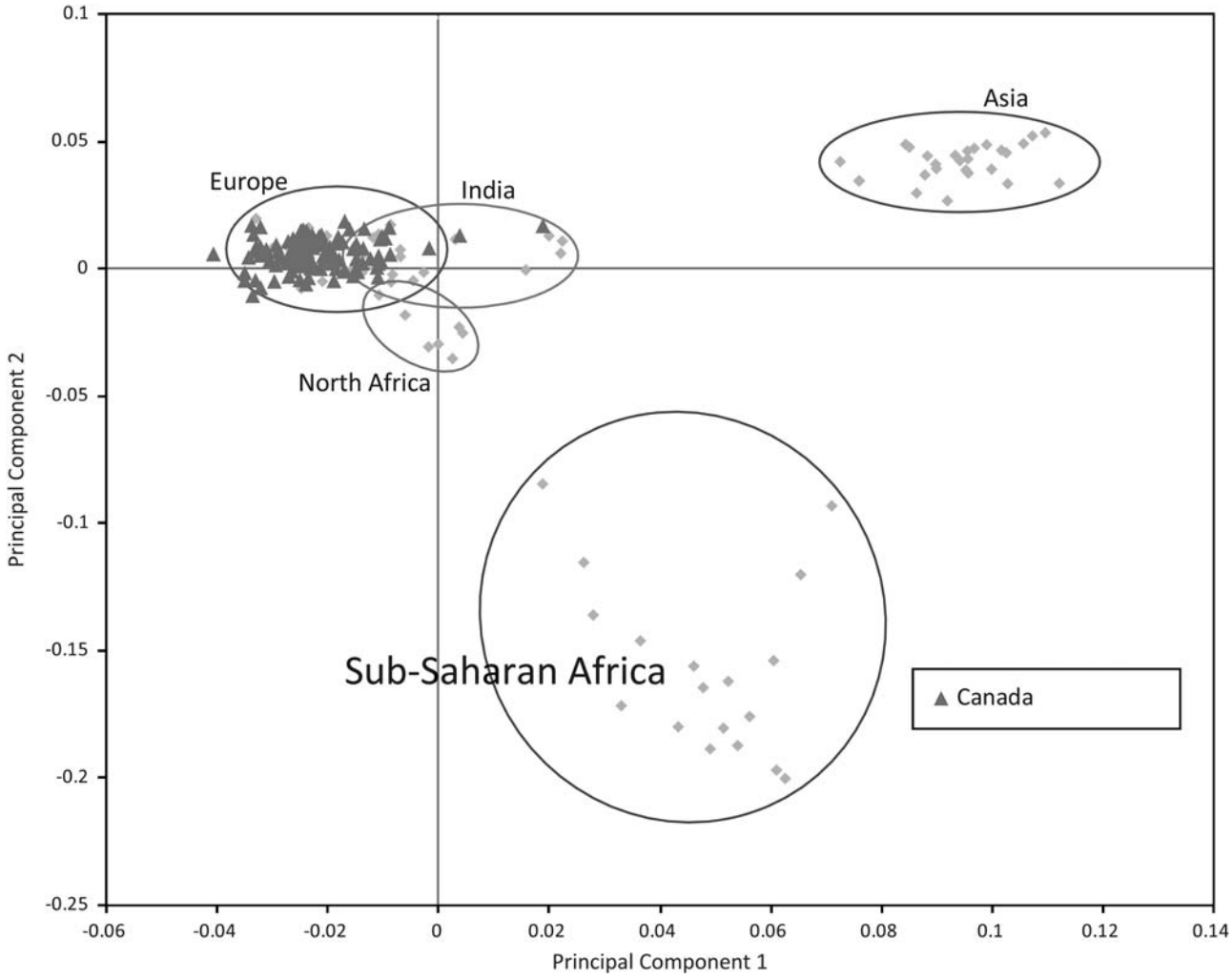

Figure 3 Scatter plot of PC1 and PC2 of GATC samples with self-reported geographic origin from Canada. PCA shows that the genetic structure of most of these individuals is similar to Europeans.

The PCA results of individuals of unknown or unreported geographic origin showed that the majority of these samples were similar to individuals of European ancestry (Figure 5b) with some being similar to Asians and Africans. The individuals with intermediate values between the European and Asian continental clusters could be of mixed or First Nations origin; however, with the current data it is not possible to distinguish them.

Discussion

Ethnicity has an important function in pharmacogenomics because allele frequencies of genetic variants with significant effects on the biotransformation of drugs can vary considerably between different ethnicities. Mixing populations of different ancestries in an association study can lead to spurious associations. Therefore, it is critically important to determine genetic ancestry of samples to correct for these differences in a pharmacogenomic association study. We describe here the implementation of PCA to easily and reliably ascertain the genetic ancestry of study samples using a limited set of pharmacogenomic markers to assess population stratification within a Canadian study cohort and to correct for these ancestry differences, thereby maximizing the number of samples and power, while minimizing false-positive associations.

When studies are conducted in ethnically diverse cohorts with mixed and sometimes unknown ancestry, such as the GATC pharmacogenomics study of ADRs in children, determining underlying genetic ancestry can be difficult because self-reported ethnic or racial labels are often insufficient proxies for genetic ancestry, or may be undisclosed or unknown by the study participant. Methods that determine genetic ancestry using patient genotype data reflect differences in allele frequencies significantly more accurately.[4,16,34] Preferably, these methods should be fast and easy to use even on large datasets without the need for

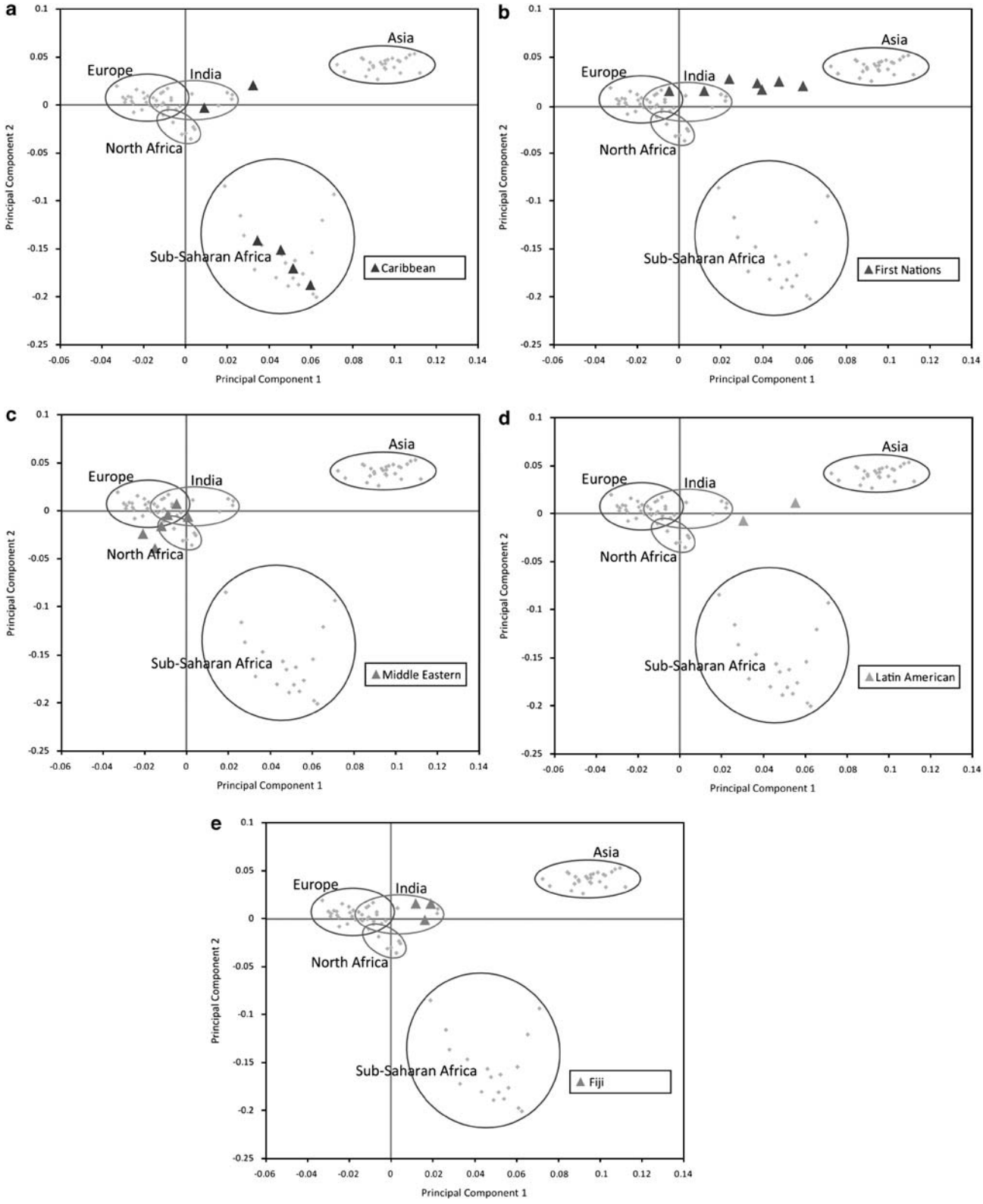

Figure 4 Scatter plots of PC1 and PC2 showing the genetic structure of GATC samples with origins from other geographic regions. (**a**) Caribbean. Four individuals have a genetic structure similar to sub-Saharan Africans, whereas the two others are closer to India and Europe. (**b**) First Nations. The individuals with First Nations origin show intermediate values between Europe and Asia. (**c**) Middle-Eastern. PC values of individuals with Middle-Eastern origin are on the border of Europe and North Africa. (**d**) Latin American. The two individuals from Latin America have PC values between the three continental clusters. (**e**) Fijian. Three individuals with origins from the Fijian Islands show genetic structure similar to Indians.

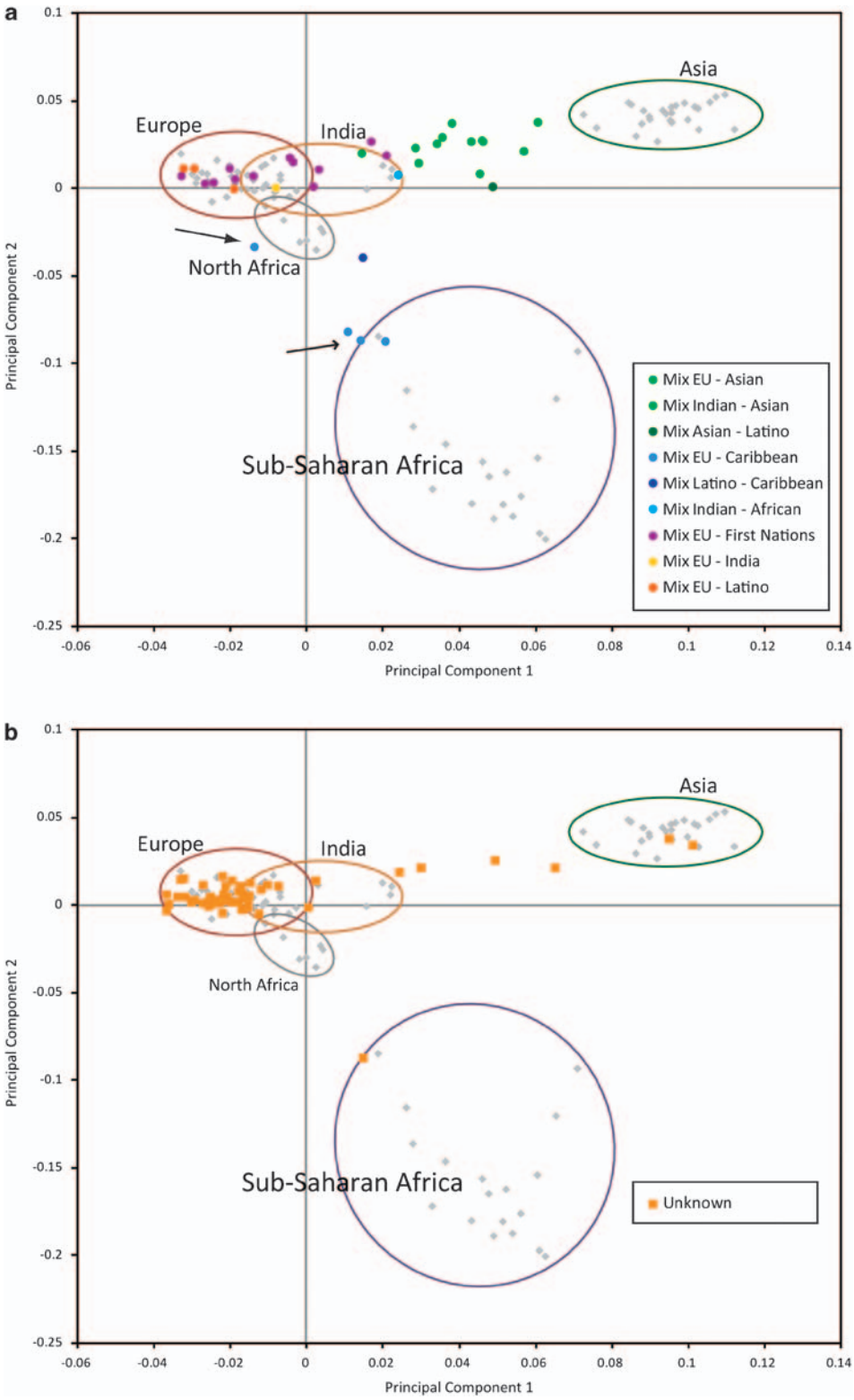

Figure 5 (**a**) Scatter plot of PC1 and PC2 of GATC samples with mixed origins. The majority has PC values between the two continental or geographic clusters they are originating from. There seems to be a genetic dose effect, for example, three individuals with two European and two Caribbean grandparents fall on the border of the African cluster (open arrow), whereas one individual with three European and one Caribbean grandparent falls closer to the European cluster (closed arrow); (**b**) Scatter plot of PC1 and PC2 of GATC with unknown geographic origin. On the basis of PC values most samples seem to be of European origin.

genotyping additional AIMs. Unlike some other methods, PCA of genotype data is fast and easy, can be used on large datasets and does not require any additional genotyping.[20] Other groups have shown that PCA can accurately detect underlying differences in genetic ancestry and estimate an individuals' ancestral origin with high accuracy.[20,23] We show here that PCA using a limited set of important pharmacogenomic markers in an ethnically diverse cohort can reliably detect underlying differences in genetic ancestry.

We have applied PCA to detect differences in genetic ancestry using 2094 SNPs in 220 key drug biotransformation genes in a varied cohort of samples from the GATC study in patients with geographic origins from around the world. Initially, worldwide Coriell reference samples of predetermined ancestry were used to assess worldwide genetic variation using PCA, and then the PCs of GATC samples were plotted. We found that the first two PCs were highly informative of genetic ancestry and strongly correlated with the known geographic origin of samples. We were also able to determine the genetic ancestry of samples with mixed or unknown origins. Principal component analysis could also detect outliers, such as an individual from Singapore that otherwise may have been misclassified. Furthermore, PCA is a powerful approach to classify groups of patients that do not belong to one of the three continental groups. The PC values could be used to stratify samples into different, genetically similar groups for analysis. However, PCs are continuous axis of variation,[20] and therefore, in most cases, it will be more appropriate to use them as continuous covariates in regression analyses. This strategy will increase the power of a study because all participants can be analyzed together, while at the same time correct for population stratification.

Our ability to discriminate between samples from different ethnic or geographic origins and to detect further population stratification will likely increase when more SNPs are investigated in our pharmacogenomics panel.[22] Using PCA on hundreds of thousands of markers, several groups were able to show detailed underlying genetic differences with high correlation between the genetic and geographic map in Europe.[23,35]

Increasing the number of samples might further increase the ability to detect additional population structure. Currently, samples with genetic ancestry from Europe were relatively overrepresented in our cohort compared with the number of samples from Asian and especially African ancestry. Genetic diversity in certain regions such as Africa or India is extensive[27,36] and including more samples from these origins and from other geographic locations will likely further improve this method.

Increasing the number of SNPs or including more samples, however, will not likely change the overall worldwide map when plotting PC1 and PC2, but investigating plots of subsequent PCs might show additional geographic structure in subpopulations.[25,26] In this study, visual inspection of plots of subsequent components (PC3–PC10) did not show further discrimination possible between subpopulations (data not shown).

To test the robustness of PCA and to estimate the minimum number of SNPs needed in our SNP panel to detect population structure, we performed PCA using sets of different numbers of randomly selected SNPs. Plotting the first two components, using 1000 or 500 randomly selected SNPs, shows a similar cluster pattern as using the full dataset (Supplemental Figure 1a and b). Using less SNPs (250, 100 or even as low as 50) still shows some evidence of the continental clusters (Supplemental Figure 1c–e), however, the clusters are close together and there is substantial variability within clusters. The plots suggest that at least 250–500 SNPs are needed to discriminate clearly between clusters.

The pharmacogenomic SNP genotyping panel consisted of different classes of SNPs (i.e. non-synonymous, synonymous, intronic, etc.). To evaluate whether the results only depended on a particular class, we applied PCA to the dataset using only intronic ($n = 972$) or only non-synonymous ($n = 332$) SNPs. Again, plotting the first two components using either group of SNPs shows a similar cluster pattern as using the full dataset (Supplemental Figure 2). This suggests that the observed structure does not depend on a particular class of SNP, as long as the dataset is sufficiently large. Removing the AIMs, which were included in the design of the SNP panel, from the full dataset also did not change the results (data not shown).

Principal component analysis to detect population structure works well if the genetic markers are independent. However, in extreme cases of linkage disequilibrium (LD) the results can be difficult to interpret. Components can be correlated with genotype patterns in large blocks in which all markers are in LD and distort the eigenvalue structure if not corrected for.[22] In this study, many genes have been densely genotyped with varying patterns of LD. To assess whether LD between markers might influence our results, we removed a marker from every pair of markers that were in tight LD, so that only unlinked markers were included in the genotype dataset ($n = 778$). We then applied PCA to this reduced dataset and plotted the first two components (PC1 and PC2) of the Coriell reference samples (Supplemental Figure 3). The cluster pattern showed in this analysis is very similar to the initial PCA using the full set of markers, which suggests that LD between markers in our dataset did not influence our results, which is consistent with what has been shown earlier.[25]

Several other factors such as inter-assay variability or cryptic relatedness can potentially lead to evidence for structure in genotyping data and be mistaken for population stratification.[20,37,38] However, inter-assay variability for our genotyping panel was low with a concordance in genotyping results for replicates of greater than 99.9% ($n = 132$). We also assessed cryptic relatedness, or hidden distant kinship, by calculating the average identity by state, but found no cryptic related individuals. These factors will therefore not have influenced the results substantially.

We have shown that PCA applied to a limited set of pharmacogenomic data can be used to detect important underlying differences in genetic ancestry. Principal

component analysis can be used in pharmacogenomic association studies to maximally use available patient samples, which are often very valuable and in limited supply, while at the same time minimizing the likelihood of finding false-positive associations.

Material and methods

Populations and samples

Initially, DNA samples from different reference populations were selected and obtained from the Human Variation Panel from the Coriell Cell Repositories (Coriell Institute for Medical Research, Camden, NJ, USA): 10 Northern-European, 9 Italian, 9 Hungarian, 9 Indo-Pakistani, 9 Chinese, 9 Japanese, 9 Southeast Asians, 9 African-American, 7 Africans North of the Sahara and 9 Africans South of the Sahara.

A total of 524 patient DNA samples were then analyzed as part of the GATC project, a national project established in Canada to identify novel predictive genomic markers of severe ADRs in children[14]. Patients who suffered any serious ADR were enrolled as well as drug-matched controls. Patients were treated for a variety of diseases with many different drugs. There were slightly more males than females (57% vs 43%). No siblings or relatives were included in this analysis. As part of this study, parents were asked for the geographic origin of the four grandparents of their child. See Supplemental Table 2 for more details on patient characteristics. Written informed consent was obtained from each study participant or parent and the study was approved by the ethics committees of all the participating universities and hospitals. DNA was extracted from blood, saliva or buccal swabs using the QIAamp 96 DNA Blood kit (Qiagen, ON, Canada).

Genotyping

All DNA samples were genotyped for 2977 SNPs using a customized Illumina GoldenGate SNP genotyping assay (Illumina, San Diego, CA, USA), which was designed to capture the genetic variation of 220 key drug biotransformation genes (i.e. phases I and II drug-metabolism enzymes, drug transporters, drug targets, drug receptors, transcription factors, ion channels and other disease-specific genes related to the physiological pathway of ADRs). The panel consisted of 1536 tagSNPs identified using the LD Select algorithm to select a maximally informative set of tag SNPs to assay in the candidate genes.[39] The tag SNP selection was performed using data from the International HapMap project that included all four populations (CEU, CHB, JPT, YRI) with a threshold for the LD statistic r^2 of 0.8, and a minor allele frequency of more than 0.05. Furthermore, 1536 functional SNPs were included that had been identified primarily by literature review or from public databases that cause non-synonymous amino-acid changes or have been or could be associated with changes in enzyme activity or function. Ninety-five SNPs were both tag and functional SNP, so a total of 2977 unique SNPs were included. The initial design also included 50 AIMs. All SNPs were manually clustered in the BeadStudio software suite. SNPs that could not be clustered or were non-polymorphic were excluded from further analyses (883 SNPs). In total 2094 SNPs were available for analysis. See Supplemental Table 1 for more details on the SNP panel. Researchers who wish to obtain more information can contact the corresponding author.

Data quality control

Samples with a call rate of less than 95% were not included in the analysis. The average genotyping call rate for all samples was 99.3%.

To evaluate our genotyping platform for inter-assay variability or batch-to-batch variation, we genotyped all Coriell samples as well as a subset of patient samples in replicate. In addition, a positive control sample with known genotypes was included in each genotyping batch. The concordance of genotype calls between these replicate genotyped samples was greater than 99.9% ($n=132$).

The average identity by state was computed for each subject pair, as implemented in PLINK,[40] but no duplicates (>99% identity) or (cryptic) related individuals (86–98% identity) were found.

Principal component analysis

Principal components of the SNP genotype data from the Coriell reference samples and the GATC patient samples were calculated using the Eigenstrat method[20] implemented in HelixTree 6.4.2 (Golden Helix, Bozeman, MT, USA) using default settings. Applying PCA to the genotype data consisting of 2094 SNPs reduced this multidimensional dataset into a smaller number of PCs, which each explains subsequent parts of variation. The first and second PCs that explain the largest portion of variation, 4.62 and 3.46%, respectively, were plotted using Excel (Microsoft, Redmond, WA, USA).

A second PCA was performed using only markers not in LD ($n=778$). Linkage disequilibrium between all pairs of markers was calculated in HelixTree. A marker was removed from every pair of markers that were in LD, so that only unlinked markers ($r^2<0.2$) were included.

Abbreviations

PCA	*principal component analysis*
GWAS	*genome-wide association study*
ADR	*adverse drug reaction*
SNP	*single nucleotide polymorphism*
GATC	*genotype-specific approaches to therapy in children*

Acknowledgments

We thank the patients and their families for their participation in the GATC project. We acknowledge the support of the GATC active ADR surveillance network, particularly the site investigators Cheri Nijssen-Jordan, David Johnson, Kevin Hall, Michael Rieder, Shinya Ito, Gideon Koren, Regis Vaillancourt, Pat Elliott-Miller, Jean-Francois Bussières, Denis Lebel, Margaret Murray, Darlene Boliver, Carol Portwine; site surveillance clinicians Linda Verbeek, Rick Kaczowka, Shanna Chan, Becky Malkin, Facundo Garcia, Miho Inoue, Sachi Sakaguchi, Toshihiro

Tanaka, Elaine Wong, Brenda Wilson, Pierre Barret, Carol-anne Osborne, Amy Cranston; and research staff at POPI and the CMMT: Anne Smith, Claudette Hildebrand; Lucila Castro, Reza Ghannadan, Catherine Carter, Fudan Miao, Terry Pape and Graeme Honeyman. The study was funded by the Genome Canada, and additional funding was also provided by the Genome British Columbia; the Child & Family Research Institute (Vancouver, BC); the Canadian Institutes of Health Research; Faculties of Pharmaceutical Sciences and Medicine, University of British Columbia; University of Western Ontario; the Canada Gene Cure Foundation; the Canadian Society of Clinical Pharmacology; C^{17} Research Network: the Childhood Cancer Foundation, Candlelighters Canada; the Canadian Paediatric Society, Merck Frosst; Janssen-Ortho; Illumina; IBM; Eli Lilly; and Pfizer. HV is supported by a postdoctoral research fellowship award from the Michael Smith Foundation for Health Research and the Child and Family Research Institute.

Conflict of interest

The authors declare no conflict of interest. This work was funded as part of the peer-reviewed Genome Canada Applied Health Research Program.

References

1 Gurwitz D, Motulsky AG. Drug reactions, enzymes, and biochemical genetics 50 years later. *Pharmacogenomics* 2007; **8**: 1479–1484.
2 Ameyaw MM, Regateiro F, Li T, Liu X, Tariq M, Mobarek A *et al.* MDR1 pharmacogenetics: frequency of the C3435T mutation in exon 26 is significantly influenced by ethnicity. *Pharmacogenetics* 2001; **11**: 217–221.
3 McLeod HL, Pritchard SC, Githang'a J, Indalo A, Ameyaw MM, Powrie RH *et al.* Ethnic differences in thiopurine methyltransferase pharmacogenetics: evidence for allele specificity in Caucasian and Kenyan individuals. *Pharmacogenetics* 1999; **9**: 773–776.
4 Wilson JF, Weale ME, Smith AC, Gratrix F, Fletcher B, Thomas MG *et al.* Population genetic structure of variable drug response. *Nat Genet* 2001; **29**: 265–269.
5 Lonjou C, Thomas L, Borot N, Ledger N, de Toma C, LeLouet H *et al.* A marker for Stevens-Johnson syndrome: ethnicity matters. *Pharmacogenomics J* 2006; **6**: 265–268.
6 Hindorff L, Junkins H, Mehta J, Manolio TA. *Catalog of Published Genome-Wide Association Studies*, Available at www.genome.gov/26525384|.
7 Cooper GM, Johnson JA, Langaee TY, Feng H, Stanaway IB, Schwarz UI *et al.* A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. *Blood* 2008; **112**: 1022–1027.
8 Link E, Parish S, Armitage J, Bowman L, Heath S, Matsuda F *et al.* SLCO1B1 variants and statin-induced myopathy—a genomewide study. *N Engl J Med* 2008; **359**: 789–799.
9 Wellcome Trust Case Control Consortium. Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature* 2007; **447**: 661–678.
10 Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* 2007; **445**: 881–885.
11 Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. *Am J Hum Genet* 2000; **67**: 170–181.
12 Balding DJ. A tutorial on statistical methods for population association studies. *Nat Rev* 2006; **7**: 781–791.
13 Huang RS, Duan S, Shukla SJ, Kistner EO, Clark TA, Chen TX *et al.* Identification of genetic variants contributing to cisplatin-induced cytotoxicity by use of a genomewide approach. *Am J Hum Genet* 2007; **81**: 427–437.
14 Ross CJ, Carleton B, Warn DG, Stenton SB, Rassekh SR, Hayden MR. Genotypic approaches to therapy in children: a national active surveillance network (GATC) to study the pharmacogenomics of severe adverse drug reactions in children. *Ann N Y Acad Sci* 2007; **1110**: 177–192.
15 Statistics Canada. *Ethnic Origin and Visible Minorities 2006 Census* (2008). Statistics Canada Demography Division: Ottawa.
16 Suarez-Kurtz G, Vargens DD, Struchiner CJ, Bastos-Rodrigues L, Pena SD. Self-reported skin color, genomic ancestry and the distribution of GST polymorphisms. *Pharmacogenet Genomics* 2007; **17**: 765–771.
17 Barnholtz-Sloan JS, McEvoy B, Shriver MD, Rebbeck TR. Ancestry estimation and correction for population stratification in molecular epidemiologic association studies. *Cancer Epidemiol Biomarkers Prev* 2008; **17**: 471–477.
18 Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics* 2000; **155**: 945–959.
19 Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 2003; **164**: 1567–1587.
20 Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006; **38**: 904–909.
21 Menozzi P, Piazza A, Cavalli-Sforza L. Synthetic maps of human gene frequencies in Europeans. *Science (NY)* 1978; **201**: 786–792.
22 Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet* 2006; **2**: e190.
23 Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A *et al.* Genes mirror geography within Europe. *Nature* 2008; **456**: 98–101, Epub 31 August 2008.
24 Tian C, Gregersen PK, Seldin MF. Accounting for ancestry: population substructure and genome-wide association studies. *Hum Mol Genet* 2008; **17**: R143–R150.
25 Nelson MR, Bryc K, King KS, Indap A, Boyko AR, Novembre J *et al.* The population reference sample, POPRES: a resource for population, disease, and pharmacological genetics research. *Am J Hum Genet* 2008; **83**: 347–358.
26 Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC *et al.* Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 2008; **451**: 998–1003.
27 Campbell MC, Tishkoff SA. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu Rev Genomics Hum Genet* 2008; **9**: 403–433.
28 Singapore Department of Statistics. *Population Trends* 2008. Department of Statistics: Singapore.
29 Wang S, Lewis CM, Jakobsson M, Ramachandran S, Ray N, Bedoya G *et al.* Genetic variation and population structure in native Americans. *PLoS Genet* 2007; **3**: e185.
30 Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science (NY)* 2008; **319**: 1100–1104.
31 Friedlaender JS, Friedlaender FR, Reed FA, Kidd KK, Kidd JR, Chambers GK *et al.* The genetic structure of Pacific Islanders. *PLoS Genet* 2008; **4**: e19.
32 Kimura R, Ohashi J, Matsumura Y, Nakazawa M, Inaoka T, Ohtsuka R *et al.* Gene flow and natural selection in oceanic human populations inferred from genome-wide SNP typing. *Mol Biol Evol* 2008; **25**: 1750–1761.
33 Fiji Islands Bureau of Statistics (2008) Census 2007.
34 Estrela RC, Ribeiro FS, Carvalho RS, Gregorio SP, Dias-Neto E, Struchiner CJ *et al.* Distribution of ABCB1 polymorphisms among Brazilians: impact of population admixture. *Pharmacogenomics* 2008; **9**: 267–276.
35 Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A *et al.* Correlation between genetic and geographic structure in Europe. *Curr Biol* 2008; **18**: 1241–1248.
36 Indian Genome Variation Consortium. Genetic landscape of the people of India: a canvas for disease gene exploration. *J Genet* 2008; **87**: 3–20.
37 Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM *et al.* Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat Genet* 2005; **37**: 1243–1246.
38 Voight BF, Pritchard JK. Confounding from cryptic relatedness in case-control association studies. *PLoS Genet* 2005; **1**: e32.

39 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.

40 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–575.

Supplementary Information accompanies the paper on the *The Pharmacogenomics Journal* website (http://www.nature.com/tpj)