# An Ancestry Informative Marker Set Which Recapitulates the Known Fine Structure of Populations in South Asia

Ranajit Das[1,*,†] and Priyanka Upadhyai[2,†]

[1]Manipal Centre for Natural Sciences (MCNS), Manipal Academy of Higher Education, Manipal, Karnataka, India

[2]Department of Medical Genetics, Kasturba Medical College, Manipal Academy of Higher Education, Manipal, Karnataka, India

[†]These authors contributed equally to this work.

*Corresponding author: E-mail: ranajit.das@manipal.edu.

**Accepted:** August 19, 2018

## Abstract

The inference of genomic ancestry using ancestry informative markers (AIMs) can be useful for a range of studies in evolutionary genetics, biomedical research, and forensic analyses. However, the determination of AIMs for highly admixed populations with complex ancestries has remained a formidable challenge. Given the immense genetic heterogeneity and unique population structure of the Indian subcontinent, here we sought to derive AIMs that would yield a cohesive and faithful understanding of South Asian genetic origins. To discern the most optimal strategy for extracting AIMs for South Asians we compared three commonly used AIMs-determining methods namely, Infocalc, $F_{ST}$, and Smart Principal Component Analysis with ADMIXTURE, using previously published whole genome data from the Indian subcontinent. Our findings suggest that the Infocalc approach is likely most suitable for delineation of South Asian AIMs. In particular, Infocalc-2,000 ($N = 2,000$) appeared as the most informative South Asian AIMs panel that recapitulated the finer structure within South Asian genomes with high degree of sensitivity and precision, whereas a negative control with an equivalent number of randomly selected markers when used to interrogate the South Asian populations, failed to do so. We discuss the utility of all approaches under evaluation for AIMs derivation and interpreting South Asian genomic ancestries. Notably, this is the first report of an AIMs panel for South Asian ancestry inference. Overall these findings may aid in developing cost-effective resources for large-scale demographic analyses and foster expansion of our knowledge of human origins and disease, in the South Asian context.

**Key words:** ancestry informative markers, South Asian AIMs, Infocalc, SmartPCA, *FST*, admixture.

## Introduction

The human genome encapsulates a plethora of information that reflects our uniqueness as well as our proximity to contemporary and ancestral individuals (Paabo 2003); the availability of large-scale genomic data and the emergence of next generation sequencing (NGS) and genotyping approaches has facilitated our exploration of genetic structure and variation that is central to understanding our evolutionary history. Commonly genomic ancestry is inferred using a set of highly informative single nucleotide polymorphisms (SNPs) referred to as ancestry informative markers (AIMs) that exhibit large differences in allele frequencies between ancestral populations (Daya et al. 2013; Vongpaisarnsin et al. 2015; Santos et al. 2016; Mendoza et al. 2017; Santangelo et al. 2017; Wang et al. 2018). It is envisioned that genotyping a certain number of AIMs can facilitate the assignment of the likely ethnic and/or geographic origin for a query population of a given genomic profile and aid in the ascertainment of what proportion of ancestry in the query group is derived from a set of source populations or from distinct geographic regions.

Accordingly, the determination of AIMs have been carried out for uncovering population stratification and for the biogeographical localization of distinct continental or world-wide populations (Shriver et al. 2003; Shriver et al. 2005; Paschou et al. 2007; Halder et al. 2008; Taboada-Echalar et al. 2013; Barbosa et al. 2017). Delineation of population substructure and measuring ancestry are particularly vital for association studies to glean the genetic etiology for complex and multifactorial disorders. The increasing availability of large-scale association studies has manifested that population structure resulting from recent admixture or biased sampling can mask true correlations or increase the risk of false positives (Lander and Schork 1994; Ziv and Burchard 2003; Marchini et al. 2004; Campbell et al. 2005), thereby making correction for the corresponding errors highly recommended (Devlin and Roeder 1999; Pritchard et al. 2000; Reich and Goldstein 2001; Satten et al. 2001; Hoggart et al. 2003; Freedman et al. 2004; Tsai et al. 2005). Notably, the utilization of AIMs panels for ancestry determination has emerged as a cost effective and useful approach to control for population substructure in association studies (Pardo-Seco et al. 2014). Further AIMs panels have also been devised for ancestry inference by forensic geneticists (Sanchez et al. 2006; Phillips et al. 2007; Sanchez et al. 2008; Phillips et al. 2016; Sun et al. 2017).

It is noteworthy that the number of genetic markers necessary for deducing ancestry likely depends on their informativeness and the genetic heterogeneity, with the former being a function of allele frequency variability between the ancestral groups from which the query populations are derived. For highly admixed or multi-ethnic populations it may be envisioned that a highly dense panel of AIMs may be required to derive optimal ancestry information (Pardo-Seco et al. 2014). India and its neighboring regions in South Asia are characterized by immense genetic and ethno-linguistic diversity entwined with distinct sociocultural practices, resulting from a complex history of admixture events that populations from this region have experienced over long periods of time (Bamshad et al. 2001; Basu 2003; Reich et al. 2009; Moorjani et al. 2013; Basu et al. 2016; Sengupta et al. 2016; Das and Upadhyai 2017). Here, we have compared three strategies previously used for AIMs determination, namely the Infocalc algorithm (Paschou et al. 2007; Kosoy et al. 2009), Wright's $F_{ST}$ (Tian et al. 2007; Kidd et al. 2011; Nievergelt et al. 2013) and Smart Principal Component Analysis (Smart PCA; Patterson et al. 2006) with ADMIXTURE (Alexander et al. 2009), to interrogate previously published whole genome data from the Indian subcontinent (Nakatsuka et al. 2017), in order to delineate an AIMs panel that can reproducibly and efficiently capture the complex genomic history of South Asian populations.

## Data Sets

We utilized a data set composed of 1,648 South Asians corresponding to 79 unique ethnic groups, assessing 499,158 SNPs (Nakatsuka et al. 2017). File conversions and manipulations were performed using EIG v4.2 (Price et al. 2006), VCF tools (Danecek et al. 2011), and PLINK (Purcell et al. 2007).

## Data Analyses

### The Genetic Structure Canvas of Ancient South Asian Genomes

The population structure of the ancient genomes was described using Principal Component Analysis (PCA) implemented in PLINK v1.9 (https://www.cog-genomics.org/plink/1.9/; last accessed May 27, 2018) using –pca command. We also applied the model-based *unsupervised* clustering methods implemented in ADMIXTURE v1.3 (Alexander et al. 2009). The optimum number of ancestral components ($K$) was discerned by minimizing the cross-validation error (CVE; Alexander et al. 2009) implemented in ADMIXTURE v1.3 using a –cv flag to the ADMIXTURE command line. All PCA and Admixture plots were generated in R v3.2.3.

### Determination of AIMs for South Asian Populations

In order to deduce the South Asian AIMs or SNPs competent for inference of the genomic ancestry of South Asian samples with accuracy proximal to that of a complete SNP set (CSS) of 499,158 autosomal SNPs, we compared four methods enumerated below.

1 Infocalc

Infocalc algorithm (Rosenberg et al. 2003), implemented in Infocalc v1.1, determines the amount of ancestry information provided by multiallelic markers by calculating the informativeness ($I$) of each marker individually. It determines $I$ based on a mathematical expression described previously (Rosenberg et al. 2003):

$$I = \sum_{j=1}^{N} \left( -p_j \log \ p_j + \sum_{i=1}^{K} \frac{p_{ij}}{K} \log \ p_{ij} \right)$$

where $p_j$ is the mean frequency of allele $j$ over all populations, $p_{ij}$ is the relative frequency of allele $j$ in population $i$ and $K$ is the total number of populations.

We selected the top 10,000 most informative markers from the Infocalc v1.1 output file. Infocalc v1.1 compatible files were generated by using –structure modifier to the PLINK v1.9 command line. The top 10,000 most informative markers were selected based on the informativeness defining column ($I\_n$) of the output (supplementary fig. S1, Supplementary Material online).

2 Top Wright's $F_{ST}$ SNPs

$F_{ST}$ (Wright 1969) measures the degree of differentiation among populations likely arising due to genetic structure within them. Given a set of populations, PLINK estimated the fixation indices ($F_{ST}$) separately for all 499,158 markers under evaluation here, using –fst command alongside –within flag that defines population IDs of the genomes. Top 10,000 SNPs with the highest $F_{ST}$ values were selected for subsequent analyses (supplementary fig. S2, Supplementary Material online).

3 Admixture

We analyzed the ADMIXTURE output (P file) for $K$ of 10 to identify 9,816 SNPs with high $K$ (column to column) variance ($\geq$0.06).

4 SmartPCA

To determine the most informative markers, SNP weightings for each principal component (PC) were calculated using the "smartpca" algorithm implemented in EIG v7.2.1 (Patterson et al. 2006; Price et al. 2006). SmartPCA executes the PCA on input genotype data in "eigenstrat" format and outputs PCs (eigenvectors) and eigenvalues. In addition to these two files, SmartPCA also generates a "snpwt" file, depicting the weight of all 499,158 markers for each PC. The 10,000 SNPs with the highest "weights" for the first principal component (PC1) were selected for subsequent evaluation (supplementary fig. S3, Supplementary Material online).

### Estimation of Candidate AIMs Data Sets

To determine the most optimal AIMs-derivation tool for South Asian genomes, we first compared candidate data sets comprising of the top 10,000 SNPs obtained by Infocalc, $F_{ST}$ and SmartPCA with 9,816 SNPs detected using an Admixture based method, both qualitatively (via Admixture and PCA) and quantitatively (computing the Euclidean distances between the admixture components of the candidate data sets and CSS). We note that only five SNPs were found to be common to all approaches, while a consensus of three out of four AIMs-determining methods yielded 251 SNPs that were insufficient to detect the details of structure and variation within the South Asian ancestry (data not shown). Thereafter we generated a data set, comprising of 2,534 SNPs that were common to at least two of the four methods ($F_{ST}$, Infocalc, Admixture, and SmartPCA). We compared this pool of data with those comprising of the top 2,534 SNPs extracted solely via Infocalc, $F_{ST}$, Admixture, and SmartPCA based methods. To adjudge the predictive accuracy of the candidate AIMs data sets, a control data set was also generated by randomly sampling 2,534 SNPs from the CSS.

## Results

### Clustering of Populations

The ancestry of 1,648 samples was estimated using unsupervised clustering as implemented in ADMIXTURE v1.3 (Alexander et al. 2009). Model validation by optimum choice of the number of ancestral components ($K$) was achieved for all data sets by minimizing the cross-validation error (CVE). The lowest CVE was estimated at $K = 10$ (supplementary fig. S4, Supplementary Material online).

At $K = 10$, the CSS revealed a discernible degree of genetic admixture between the North and South Indian populations (supplementary fig. S5*A*, Supplementary Material online). Palliyar (violet), Pulliyar (blue), Kumhar (cyan), Juang (yellow), Ulladan (light-blue), Kalash, Shia Iranians from Hyderabad, and Kamboj (orange), Onge (red), Nysha (light-green), Malaikuravar and Narikuravar (green), and Vysya (magenta) populations were homogeneously assigned to distinct groups. Consistent with previous findings (Reich et al. 2009; Moorjani et al. 2013; Basu et al. 2016) the admixture plot revealed that most South Asians have variable fractions of orange (putative West Eurasian), magenta (South Indian), and yellow (Austro-Asiatic) genomic ancestral components. It indicated higher fractions of likely West Eurasian ancestry in North Indian and Pakistani genomes, while increased levels of South Indian and Austro-Asiatic ancestral components appeared to be present in South Indians and almost equal fractions of East Asian (light-green) and West Eurasian (orange) ancestral components were discerned in Hazara genomes.

At $K = 10$, the data set comprising of top 10,000 Infocalc SNPs (Infocalc-10,000) performed the best by precisely capturing the South Asian population structure and depicted a perceptible degree of admixture between the North and South Indian populations (supplementary fig. S5*B*, Supplementary Material online). Here, in concordance with the CSS, Palliyar (violet), Pulliyar (blue), Kumhar (cyan), Juang (yellow), Ulladan (light-blue), Kalash, Shia Iranians from Hyderabad, and Kamboj (orange), Onge (red), Nysha (light-green), Malaikuravar and Narikuravar (green), and Vysya (magenta) populations were homogeneously assigned to distinct groups and Hazara genomes appeared to contain almost equal fractions of East Asian (violet) and West Eurasian (orange) ancestral components.

Other data sets, comprising of 10,000 SNPs generated using SmartPCA and $F_{ST}$-based methods (SmartPCA-10,000 and $F_{ST}$-10,000, respectively) and 9,816 SNPs generated through an Admixture-based method (Admixture-9,816) performed moderately well (supplementary fig. S5*C*–*E*, Supplementary Material online). Exceptions included, SmartPCA-10,000 that failed to capture the population structure of Vysya population (magenta; supplementary fig. S5*C*, Supplementary Material online) and $F_{ST}$-10,000 that failed to demarcate the identity of Kamboj (orange), Kumhar (cyan), Ulladan (light-blue), and Palliyar (violet) populations (supplementary fig. S5*D*, Supplementary Material online). In contrast, Admixture-9,816 seemed to perform better and more efficiently reproduced the fine population structure for most South Asian genomes (supplementary fig. S5*E*, Supplementary Material online).

Among data sets comprising of 2,534 SNPs deduced via $F_{ST}$, Infocalc, Admixture, and SmartPCA, the candidate panel derived using Infocalc (Infocalc-2,534) performed superior to the rest and was most comparable to the CSS in recapitulating the population structure for South Asians

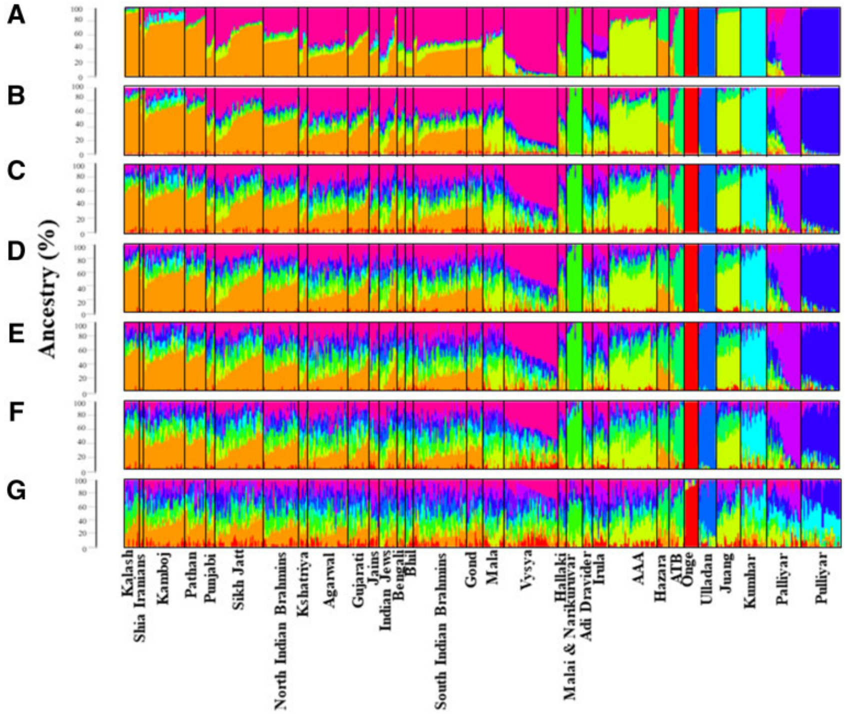

**Fig. 1.**—Admixture analyses of data sets generated using most informative SNPs detected by Infocalc algorithm. Admixture plots depicting the ancestry components of South Asian genomes. (*A*) Admixture analysis of the CSS ($N = 499{,}158$); (*B*) Admixture analysis of Infocalc-10,000; (*C*) Admixture analysis of Infocalc-2,534; (*D*) Admixture analysis of Infocalc-2,000; (*E*) Admixture analysis of Infocalc-1,500; (*F*) Admixture analysis of Infocalc-1,000; and (*G*) Admixture analysis of Infocalc-500. Admixture proportions were generated through an unsupervised admixture analyses at $K = 10$ using ADMIXTURE v1.3 and plotted in R v3.2.3. Each individual is represented by a vertical line partitioned into colored segments whose lengths are proportional to the contributions of the ancestral components to the genome of the individual. Note that Nyshas are included among the ATB group.

(supplementary fig. S6*B*, Supplementary Material online). It was followed by a panel of 2,534 SNPs obtained as a consensus of at least two of the four AIMs-determining methods (Consensus-2,534; supplementary fig. S6*F*, Supplementary Material online), and that deduced using Admixture (Admixture-2,534; supplementary fig. S6*E*, Supplementary Material online). Finally, data sets inferred using SmartPCA (SmartPCA-2,534) and $F_{ST}$ ($F_{ST}$-2,534) functioned poorly (supplementary fig. S6*C* and *D*, Supplementary Material online), with the former output being indistinguishable from that of the negative control comprising of 2,534 randomly sampled SNPs (Random-2,534; supplementary fig. S6*G*, Supplementary Material online).

Given that both Infocalc-10,000 and Infocalc-2,534 panels outperformed their SNP number matched counterpart data sets derived using alternative strategies, we sought to interrogate the minimum number of Infocalc derived SNPs that were most proximal to the CSS in capturing the detailed population structure and variation in South Asian ancestries. To this end we generated data sets consisting of 2,000, 1,500, and 1,000 Infocalc derived SNPs (Infocalc-2,000, Infocalc-1,500, and Infocalc-1,000, respectively) that appeared to be largely successful (fig. 1*D–F*). However, a set of 500 SNPs determined using Infocalc (Infocalc-500) was comparatively less efficient, it failed to demarcate the identity of the Vysya population (magenta) and depicted appreciably lower South Indian admixture component (magenta) among all South Asians (fig. 1*G*).

For comparing the data sets quantitatively, we computed Euclidean distances between the admixture components of all candidate panels and the CSS. The shortest Euclidean distance ($\mu = 0.14$) was discerned between Infocalc-10,000 and the CSS, followed by Admixture-10,000 and the CSS ($\mu = 0.20$; fig. 2). Among the 2,534 SNP panels, Infocalc-2,534 appeared the most sensitive ($\mu = 0.29$), followed by the Consensus-2,534 ($\mu = 0.47$); both Infocalc-2,534 and Consensus-2,534 performed significantly better than the Random 2,534 ($\mu = 0.61$; Tukey's *post hoc* test; *P*-value $<0.0001$). Congruent with our results from the Admixture analyses the $F_{ST}$-based data set appeared farthest from the CSS ($\mu = 0.88$) and functioned significantly worse than the Random-2,534 (Tukey's *post hoc* test; *P*-value $<0.0001$). Among the remainder, Admixture-2,534 performed moderately better than Random-2,534 ($\mu = 0.56$; Tukey's *post hoc* test; *P*-value $<0.001$), however there was no substantial difference between SmartPCA-2,534 ($\mu = 0.61$) and Random-2,534 (Tukey's *post hoc* test; *P*-value $= 0.99$).

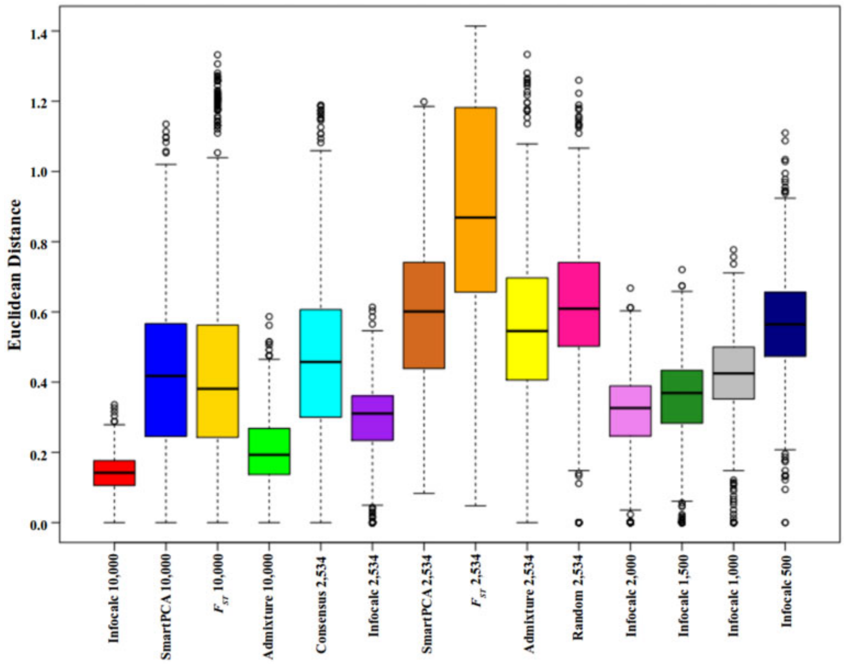

FIG. 2.—Box and whisker plots comparing the Euclidean distances between the admixture proportions of the South Asian genomes obtained using the CSS and candidate panels deduced using alternative AIMs determining approaches. The number of SNPs contained in each of the candidate panels illustrated has been indicated in the text. *Note*: Random-2,534 comprised of 2,534 randomly selected SNPs from the CSS and the Consensus-2,534 comprised of 2,534 SNPs that were detected by at least two out of the four AIMs-determining approaches under evaluation.

As observed with findings from the Admixture based studies, most Infocalc derived SNP panels, with the exception of Infocalc-500, appreciably outperformed the other 2,534 SNP containing data sets, inferred using alternative approaches, as well as the Random-2,534, negative control (Tukey's *post hoc* test; $P$-value $< 0.001$). Even though Infocalc-500 ($\mu = 0.57$) displayed significantly improved sensitivity as opposed to Random-2,534 (Tukey's *post hoc* test; $P$-value $<0.01$), it was only marginally better than the SmartPCA-2,534 (Tukey's *post hoc* test; $P$-value $= 0.04$) and seemed similar to Admixture-2,534 (Tukey's *post hoc* test; $P$-value $= 0.99$). Among the remaining Infocalc derived data sets, Infocalc-2,000 appeared most optimal, it was comparable to Infocalc-2,534 (Tukey's *post hoc* test; $P$-value $= 0.86$) and was significantly better than Infocalc-1,500 (Tukey's *post hoc* test; $P$-value $= 0.02$), Infocalc-1,000 and Infocalc-500 (Tukey's *post hoc* test; $P$-value $<0.0001$) in its capture sensitivity. Additionally, its performance was markedly better than both Consensus-2,534 and Random-2,534 (Tukey's *post hoc* test; $P$-value $<0.0001$). Whereas Infocalc-1,500 performed highly significantly better than Infocalc-1,000 and Infocalc-500 (Tukey's *post hoc* test; $P$-value $<0.0001$), Infocalc-1,000 was only superior to Infocalc-500 (Tukey's *post hoc* test; $P$-value $<0.0001$) in its efficacy.

We note that Infocalc-10,000 had the highest number of individuals with zero Euclidean distances from the CSS ($N = 41$), followed by Admixture-10,000 ($N = 25$), Infocalc-2,534 ($N = 23$), Infocalc-2,000 ($N = 22$), and Infocalc-1,500 ($N = 15$). Whereas the consensus and random data sets had eight and ten individuals, respectively with zero Euclidean distances, $F_{ST}$-2,534 and SmartPCA-2,534 had none. Overall, our results depict the pronounced informativeness of AIMs extracted via various ancestry determining approaches, underscoring their superiority over randomly selected markers in delineating South Asian ancestry information.

## PCA

PCA of South Asian genomes concurred with previously observed Ancestral North Indian (ANI)-Ancestral South Indian (ASI)-Ancestral Austro-Asiatic (AAA) contrast along the horizontal principal component (PC1; Basu et al. 2016; supplementary fig. S7*A*, Supplementary Material online). As surmised before Kalash, Shia Iranians from Hyderabad, and Kamboj populations with high West Eurasian admixture clustered at one extreme of the ANI-ASI cline, whereas the Juangs congregated at the other extreme, overlapping with the cluster consisting of Austro-Asiatic speakers. ASI-AAA-Ancestral Tibeto-Burman (ATB) contrast was observed along the vertical principal component (PC2), with Palliyar and Pulliyars (ASI) clustering at one end and Nyshi and Tibeto-Burman speakers clustering at the other end. Interestingly, Bengalis and Bhils formed discernible clusters along the ANI-ASI-AAA cline revealing their genomic distinctness. As observed in Admixture

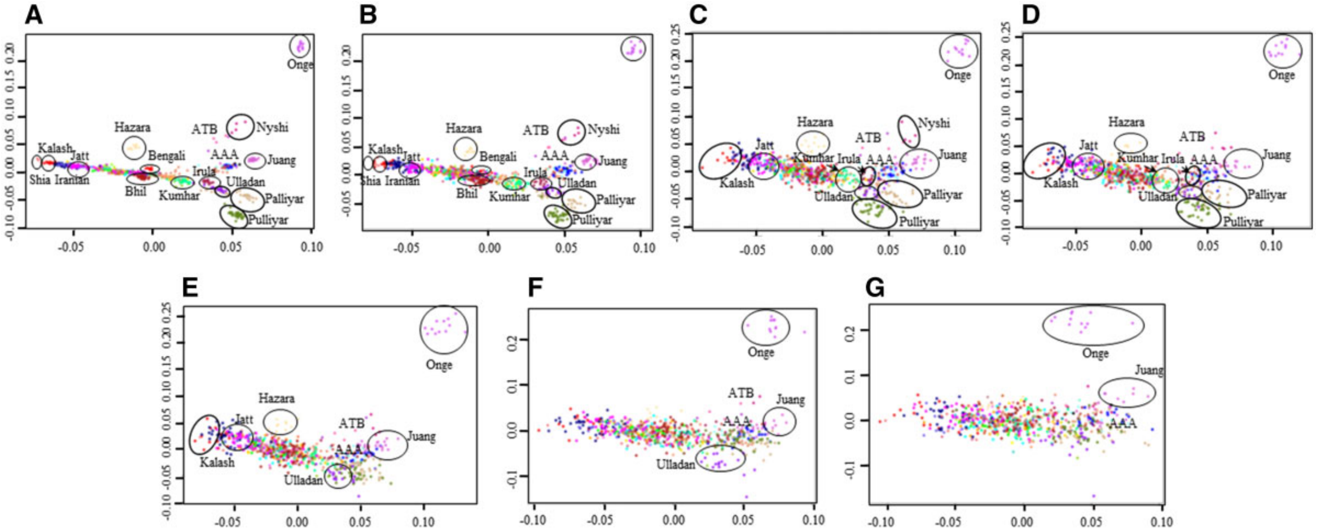

Fig. 3.—PCA of South Asian genomes. PCA plots showing genetic differentiation among South Asian genomes. The candidate panels were generated using highly informative SNPs detected through the Infocalc algorithm. (*A*) PCA of the CSS (*N* = 499,158), where the *X*-axis (PC1) explained 39.7% variance, whereas the *Y*-axis (PC2) explained 24.2% variance of the data. (*B*) PCA of Infocalc-10,000, where the *X*-axis (PC1) explained 39.8% variance, whereas the *Y*-axis (PC2) explained 23.9% variance of the data. (*C*) PCA of Infocalc-2,534, where the *X*-axis (PC1) explained 39.8% variance, whereas the *Y*-axis (PC2) explained 23.8% variance of the data. (*D*) PCA of Infocalc-2,000, where the *X*-axis (PC1) explained 39.3% variance, whereas the *Y*-axis (PC2) explained 24.2% variance of the data. (*E*) PCA of Infocalc-1,500, where the *X*-axis (PC1) explained 39.6% variance, whereas the *Y*-axis (PC2) explained 24.3% variance of the data. (*F*) PCA of Infocalc-1,000, where the *X*-axis (PC1) explained 38.3% variance, whereas the *Y*-axis (PC2) explained 23.2% variance of the data. (*G*) PCA of Infocalc-500, where the *X*-axis (PC1) explained 36.7% variance, whereas the *Y*-axis (PC2) explained 23.1% variance of the data. Notable populations are marked with circles. In all four cases illustrated here, PCA was performed in PLINK v1.9 and the top four principal components (PCs) were extracted. Top two PCs (PC1 and PC2), explaining the highest variance of the data were plotted in R v3.2.3. **X-axis designates PC1 and Y-axis designates PC2.

analyses (fig. 1*A*), Hazara, Onge, and Kumhar populations were found as distinct clusters along the ANI-ASI-AAA cline.

Moreover, as noted using Admixture, Infocalc-10,000, Infocalc-2,534, and Infocalc-2,000 data sets recapitulated the population clusters depicted by CSS with high degree of accuracy (fig. 3). However, the accuracy diminished discernably for Infocalc-1,500, Infocalc-1,000, and Infocalc-500. The latter failed to recapitulate the clustering of most South Asian genomes with the exception of Onge and Juang (fig. 3*G*). Among the remaining strategies, the $F_{ST}$-based derivation turned out to be the poorest, $F_{ST}$-10,000 could only depict clustering for a limited number of South Asian populations (Onge, Juang, Nyshi, Ulladan and Sikh Jatt; supplementary fig. S7*D*, Supplementary Material online), $F_{ST}$-2,534 completely failed to depict any contrast among the query genomes (supplementary fig. S8*D*, Supplementary Material online). We note that SmartPCA-10,000 (supplementary fig. S7*C*, Supplementary Material online), Admixture-10,000 (supplementary fig. S7*E*, Supplementary Material online), SmartPCA-2,534 (supplementary fig. S8*C*, Supplementary Material online), and Consensus-2,534 (supplementary fig. S8*F*, Supplementary Material online) performed reasonably well in clustering distinct South Asian genomes but were outperformed by Infocalc-10,000, Infocalc-2,534, and Infocalc-2,000. Notably even Random-2,534 (supplementary fig. S8*G*, Supplementary Material online) performed better than $F_{ST}$-2,534 (supplementary fig. S8*D*, Supplementary Material online); however, the former was distinctly worse than the remaining candidate AIMs panels, reflecting overall the ineffectiveness of randomly sampled SNPs in capturing population substructures in the current context.

Taking together all qualitative and quantitative findings, we discerned that the Infocalc approach as the best tool for delineating South Asian AIMs. Further, we interpreted Infocalc-2,000 as the most optimal AIMs panel for South Asians as it emerges as a moderately small set of markers that not only surpasses other Infocalc derived panels, except Infocalc-10,000, but also outperforms most 10,000 SNP containing candidate AIMs panels deduced through alternative approaches.

## Discussion

Deducing genome ancestry plays a central role in understanding human evolution, the underlying molecular mechanism of human diseases and in forensic analyses. Genetic variants with small differences in frequencies between populations when genotyped in sufficiently large numbers can be utilized for making ancestry inferences. The increasing availability of high resolution NGS data has enabled the utilization of large-scale genome wide SNP data for ancestry inference

(Novembre et al. 2008; Reich et al. 2012). However, this approach is not cost-effective nor amenable for use in specific situations, such as when insufficient and/or poor quality genomic material is obtained, as in forensic investigations. In contrast, a panel of SNPs with large differences between populations or the AIMs can be useful in gleaning ancestry information in varied scenarios, including association studies, forensic and population genetic evaluations.

The deduction of AIMs for understanding the genetic origins of highly admixed populations with complex ancestries has been a challenging prospect. India lies at the crossroads for the ancient migration of anatomically modern humans (Cann 2001; Misra 2001; Metspalu et al. 2004; Thangaraj et al. 2006; Singh et al. 2016). Consequently, its demographic history has been shaped by large-scale population migration, admixture, existence of varied geographical niches, linguistic groups and stringent enforcement of sociocultural practices like endogamy (Bamshad et al. 2001; Sengupta et al. 2006; Chaubey et al. 2007; Moorjani et al. 2013; Basu et al. 2016). Given their intricate population structure and genetic heterogeneity here we sought to determine an AIMs panel for South Asian populations that will sensitively and robustly capture their genetic history.

Previous studies have delineated AIMs using strategies such as the Infocalc (Paschou et al. 2007; Kosoy et al. 2009), Wright's $F_{ST}$ (Tian et al. 2007; Kidd et al. 2011; Nievergelt et al. 2013) and Smart PCA (Patterson et al. 2006). Here, we compared the aforesaid approaches together with ADMIXTURE (Alexander et al. 2009) to determine the most optimal strategy for delineating an AIMs panel that will accurately recapitulate the intricate structure within the highly admixed South Asian genomes, using previously published whole genome data (Nakatsuka et al. 2017).

Overall our qualitative and quantitative analyses concur that Infocalc was significantly superior to other ancestry determining strategies, in the South Asian context. Results from our Admixture analyses reflected that Infocalc-2,000 functioned largely equivalent to Infocalc-10,000, the latter outperformed all other candidate SNP panels and was most proximal to the CSS in capturing population fine structure in South Asia (fig. 1*B* and supplementary fig. S5*B*, Supplementary Material online). While Infocalc-1,500 and Infocalc-1,000 seemed somewhat similar to Infocalc-2,000, Infocalc-500 was markedly less efficient. It was unsuccessful in demarcating the identity of the Vysya population and depicted discernibly lower South Indian admixture component among South Asians (fig. 1*G*). Consistent with these findings, Infocalc-2,000 depicted South Asian population clusters with high precision and was indistinguishable from Infocalc-10,000, Infocalc-2,534 and the CSS to this end using PCA (fig. 3). However, the accuracy declined perceptibly for Infocalc-1,500, Infocalc-1,000 and was the least for Infocalc-500 that failed to capture the clustering of most South Asian genomes, except Onge and Juang (fig. 3*G*). Quantitative assessment suggested that while Infocalc-2,000 was significantly worse than Infocalc-10,000, it was indistinct from Infocalc-2,534 and considerably surpassed the smaller Infocalc derived panels, as well as other 10,000 SNP panels deduced via alternate approaches (fig. 2). Infocalc-10,000 had the highest number of individuals with zero Euclidean distances from the CSS ($N = 41$), followed by Admixture-10,000 ($N = 25$), Infocalc-2,534 ($N = 23$), Infocalc-2,000 ($N = 22$), and Infocalc-1,500 ($N = 15$). Among the remainder, Infocalc-1,000 and Random-2,534 had ten and Consensus-2,534 had eight individuals, respectively with zero Euclidean distances, while, $F_{ST}$-2,534 and SmartPCA-2,534 had none. Amidst the remaining approaches, we note that $F_{ST}$ derived candidate AIMs panels performed the poorest in capturing fine-scale population structure (fig. 2, supplementary figs. S7*D* and S8*D*, Supplementary Material online), while Admixture based ancestry delineation appeared moderately competent, in this context (fig. 2, supplementary figs. S5*E*, S6*E*, Supplementary Material online). When attempting to prune candidate SNP panels so as to obtain those containing fewer SNPs capable of adequate discrimination of population fine structure and variability, Consensus-2,534 appeared as the second-most sensitive, falling short of only most Infocalc based panels, including, Infocalc-2,534, Infocalc-2,000, and Infocalc-1,000 (fig. 2, supplementary fig. S8, Supplementary Material online). However, regardless of the approach, we note that <1,000 SNPs do not suffice to reliably capture the intricacies of population structure in South Asians, owing to their high genomic complexity. A case in point being the genomic distinctness of the Vysya population, a Telugu speaking community from South-East India that was discerned by all Infocalc-based candidate panels, except Infocalc-500 (supplementary fig. 2, Supplementary Material online). This could be attributed to smaller Infocalc panels (Infocalc-2,534, Infocalc-2,000, Infocalc-1,500 and Infocalc-1,000) comprising of markers competent of fine-scale population structure depiction within the same language (Telugu) group that is likely lost while paring them down further to Infocalc-500. A previous study had suggested that 500–1,000 SNPs chosen at random performed as well as an AIMs panel of similar size (Pardo-Seco et al. 2014). Contrary to this, except $F_{ST}$ based candidate AIMs panels, all others presently inferred have captured the South Asian population structure reasonably well (figs. 1 and 2). Taken together we interpreted Infocalc-2,000 as the most optimal AIMs panel for South Asians, as it is a collection of moderately small number of markers that displayed high sensitivity and accuracy in recapitulating population structure and diversity within the South Asian genomes.

Genetic admixture poses enormous challenges and has significant implications in biomedical research (Cooper et al. 2008). AIMs panels can be useful to control for population substructure in association studies for multifactorial disorders (Pardo-Seco et al. 2014). The delineation of AIMs for highly mixed populations may also facilitate uncovering loci that contribute to ethnic variation in complex disease risk and aid in understanding the evolutionary mechanisms underlying disease biology. Previous studies have revealed how AIMs panels pertaining to admixed populations can apprise regarding correlations between individual ancestry and specific traits, including hypertension (Zhu et al. 2005; Mukhtar et al. 2018), type II Diabetes (Parra 2004), breast cancer (Serrano-Gomez et al. 2017), and skin pigmentation (Shriver et al. 2003). Here, we note that the top 20 SNPs with the highest "informativeness" (Infocalc *I_n* scores) in the current analyses included those correlated with heart rate response to β-blockers (*rs11931264*; Shahin et al. 2018), high blood pressure (*rs225555*; Morrison et al. 2008), and blood glucose levels (*rs1516510*; Liu et al. 2009), in concordance with increased predisposition to complex, heterogeneous disorders such as type II Diabetes and cardiovascular ailments in the Indian subcontinent (Goyal and Yusuf 2006; Wells et al. 2016). Further the Infocalc-2,000 AIMs panel also included SNPs associated with triglyceride levels in type II Diabetes (*rs2240466*; Kong et al. 2015), abdominal fat in women (*rs7927727*; Sung et al. 2016), and cognitive processing (*rs2839627*; Luciano et al. 2011). In addition, four SNPs (*rs242105*, *rs931885*, *rs4377353*, and *rs4858613*) present in our AIMs panel have already been reported as ancestry informative for South-Central Asian populations, in a study by Dr Petros Drineas's group from Purdue University that did not investigate any Indian populations (https://www.cs.purdue.edu/homes/pdrineas/documents/HGDPAIMS/; last accessed May 27, 2018). While the functional relevance of these findings would entail intensive molecular exploration, nevertheless this underscores the utility of a South Asian AIMs panel for expanding our understanding of the etiology of corresponding diseases, in the realms of medical genetics in South Asia.

To the best of our knowledge this is the first study to deduce an AIMs panel capturing the intricate population structure and diversity of South Asian genomes with high sensitivity and precision. Utilization of these results while exercising adequate caution and in combination with detailed functional investigation will potentially afford cost-effective alternatives to whole genome sequencing for large-scale demographic analyses, extending our knowledge of human history and disease, in the South Asian context.

## Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

## Literature Cited

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19(9): 1655–1664.

Bamshad M, et al. 2001. Genetic evidence on the origins of Indian caste populations. Genome Res. 11(6):994–1004.

Barbosa FB, et al. 2017. Ancestry informative marker panel to estimate population stratification using genome-wide human array. Ann Hum Genet. 81(6):225–233.

Basu A. 2003. Ethnic India: a genomic view, with special reference to peopling and structure. Genome Res. 13(10):2277–2290.

Basu A, Sarkar-Roy N, Majumder PP. 2016. Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure. Proc Natl Acad Sci U S A. 113(6):1594–1599.

Campbell CD, et al. 2005. Demonstrating stratification in a European American population. Nat Genet. 37(8):868–872.

Cann RL. 2001. Genetic clues to dispersal in human populations: retracing the past from the present. Science 291(5509):1742–1748.

Chaubey G, Metspalu M, Kivisild T, Villems R. 2007. Peopling of South Asia: investigating the caste-tribe continuum in India. Bioessays 29(1):91–100.

Cooper RS, Tayo B, Zhu X. 2008. Genome-wide association studies: implications for multiethnic samples. Hum Mol Genet. 17(R2):R151–R155.

Danecek P, et al. 2011. The variant call format and VCFtools. Bioinformatics 27(15):2156–2158.

Das R, Upadhyai P. 2017. Application of geographic population structure (GPS) algorithm for biogeographical analyses of populations with complex ancestries: a case study of South Asians from 1000 genomes project. BMC Genet. 18(S1):109.

Daya M, et al. 2013. A panel of ancestry informative markers for the complex five-way admixed South African coloured population. PLoS One. 8(12):e82224.

Devlin B, Roeder K. 1999. Genomic control for association studies. Biometrics 55(4):997–1004.

Freedman ML, et al. 2004. Assessing the impact of population stratification on genetic association studies. Nat Genet. 36(4):388–393.

Goyal A, Yusuf S. 2006. The burden of cardiovascular disease in the Indian subcontinent. Indian J Med Res. 124(3):235–244.

Halder I, Shriver M, Thomas M, Fernandez JR, Frudakis T. 2008. A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. Hum Mutat. 29(5):648–658.

Hoggart CJ, et al. 2003. Control of confounding of genetic associations in stratified populations. Am J Hum Genet. 72(6):1492–1504.

Kidd JR, et al. 2011. Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples. Investig Genet. 2(1):1.

Kong X, et al. 2015. Genetic variants associated with lipid profiles in Chinese patients with type 2 diabetes. PLoS ONE. 10(8):e0135145.

Kosoy R, et al. 2009. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. Hum Mutat. 30(1):69–78.

Lander ES, Schork NJ. 1994. Genetic dissection of complex traits. Science 265(5181):2037–2048.

Liu C, Yang Q, Cupples LA, Meigs JB, Dupuis J. 2009. Selection of the most informative individuals from families with multiple siblings for association studies. Genetic Epidemiol. 33(4):299–307.

Luciano M, et al. 2011. Whole genome association scan for genetic polymorphisms influencing information processing speed. Biol Psychol. 86(3):193–202.

Marchini J, Cardon LR, Phillips MS, Donnelly P. 2004. The effects of human population structure on large genetic association studies. Nat Genet. 36(5):512–517.

Mendoza L, Aguirre DP, Builes JJ. 2017. Ancestry evaluation of an Afro-descendant population sample of the department of Chocó-Colombia. Forensic Sci Int: Genet Suppl Ser. 6:e292–e293.

Metspalu M, et al. 2004. Most of the extant mtDNA boundaries in south and southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. BMC Genet. 5:26.

Misra VN. 2001. Prehistoric human colonization of India. J Biosci. 26(4 Suppl):491–531.

Moorjani P, et al. 2013. Genetic evidence for recent population mixture in India. Am J Hum Genet. 93(3):422–438.

Morrison AC, Boerwinkle E, Turner ST, Ferrell RE. 2008. Regional association-based fine-mapping for sodium-lithium countertransport on chromosome 10. Am J Hypertens. 21(1):117–121.

Mukhtar O, et al. 2018. A randomized controlled crossover trial evaluating differential responses to antihypertensive drugs (used as mono- or dual therapy) on the basis of ethnicity: the comparIsoN oF Optimal Hypertension RegiMens; part of the Ancestry Informative Markers in HYpertension program-AIM-HY INFORM trial. Am Heart J. 204:102–108.

Nakatsuka N, et al. 2017. The promise of discovering population-specific disease-associated genes in South Asia. Nat Genet. 49(9):1403–1407.

Nievergelt CM, et al. 2013. Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel. Investig Genet. 4(1):13.

Novembre J, et al. 2008. Genes mirror geography within Europe. Nature 456(7218):98–101.

Paabo S. 2003. The mosaic that is our genome. Nature 421(6921):409–412.

Pardo-Seco J, Martinon-Torres F, Salas A. 2014. Evaluating the accuracy of AIM panels at quantifying genome ancestry. BMC Genomics. 15(1):543.

Parra EJ. 2004. Relation of type 2 diabetes to individual admixture and candidate gene polymorphisms in the Hispanic American population of San Luis Valley, Colorado. J Med Genet. 41(11):e116.

Paschou P, et al. 2007. PCA-correlated SNPs for structure identification in worldwide human populations. PLoS Genet. 3(9):1672–1686.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. PLoS Genet. 2(12):e190.

Phillips C, et al. 2007. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. Forensic Sci Int Genet. 1(3–4):273–280.

Phillips C, Santos C, Fondevila M, Carracedo A, Lareu MV. 2016. Inference of ancestry in forensic analysis I: autosomal ancestry-informative marker sets. Methods Mol Biol. 1420:233–253.

Price AL, et al. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 38(8): 904–909.

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. 2000. Association mapping in structured populations. Am J Hum Genet. 67(1):170–181.

Purcell S, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 81(3):559–575.

Reich D, et al. 2012. Reconstructing native American population history. Nature 488(7411):370–374.

Reich D, Thangaraj K, Patterson N, Price AL, Singh L. 2009. Reconstructing Indian population history. Nature 461(7263):489–494.

Reich DE, Goldstein DB. 2001. Detecting association in a case-control study while correcting for population stratification. Genet Epidemiol. 20(1):4–16.

Rosenberg NA, Li LM, Ward R, Pritchard JK. 2003. Informativeness of genetic markers for inference of ancestry. Am J Hum Genet. 73(6):1402–1422.

Sanchez JJ, et al. 2006. A multiplex assay with 52 single nucleotide polymorphisms for human identification. Electrophoresis 27(9):1713–1724.

Sanchez JJ, et al. 2008. Forensic typing of autosomal SNPs with a 29 SNP-multiplex – results of a collaborative EDNAP exercise. Forensic Sci Int Genet. 2(3):176–183.

Santangelo R, et al. 2017. Analysis of ancestry informative markers in three main ethnic groups from Ecuador supports a trihybrid origin of Ecuadorians. Forensic Sci Int Genet. 31:29–33.

Santos HC, et al. 2016. A minimum set of ancestry informative markers for determining admixture proportions in a mixed American population: the Brazilian set. Eur J Hum Genet. 24(5):725–731.

Satten GA, Flanders WD, Yang Q. 2001. Accounting for unmeasured population substructure in case–control studies of genetic association using a novel latent-class model. Am J Hum Genet. 68(2):466–477.

Sengupta D, Choudhury A, Basu A, Ramsay M. 2016. Population stratification and underrepresentation of indian subcontinent genetic diversity in the 1000 genomes project dataset. Genome Biol Evol. 8(11):3460–3470.

Sengupta S, et al. 2006. Polarity and temporality of high-resolution y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. Am J Hum Genet. 78(2):202–221.

Serrano-Gomez SJ, et al. 2017. Ancestry as a potential modifier of gene expression in breast tumors from Colombian women. PLoS One. 12(8):e0183179.

Shahin MH, et al. 2018. Genome-wide association approach identified novel genetic predictors of heart rate response to beta-blockers. J Am Heart Assoc. 7(5):e006463.

Shriver MD, et al. 2003. Skin pigmentation, biogeographical ancestry and admixture mapping. Hum Genet. 112(4):387–399.

Shriver MD, et al. 2005. Large-scale SNP analysis reveals clustered and continuous patterns of human genetic variation. Hum Genomics. 2(2):81–89.

Singh S, et al. 2016. Dissecting the influence of Neolithic demic diffusion on Indian Y-chromosome pool through J2-M172 haplogroup. Sci Rep. 6(1):19157.

Sun Q, et al. 2017. Twenty-seven continental ancestry-informative SNP analysis of bone remains to resolve a forensic case. Forensic Sciences Res. 1–3.

Sung YJ, et al. 2016. Genome-wide association studies suggest sex-specific loci associated with abdominal and visceral fat. Int J Obes (Lond) 40(4):662–674.

Taboada-Echalar P, et al. 2013. The genetic legacy of the pre-colonial period in contemporary Bolivians. PLoS ONE. 8:e58980.

Thangaraj K, et al. 2006. In situ origin of deep rooting lineages of mitochondrial Macrohaplogroup`M' in India. BMC Genomics. 7:151.

Tian C, et al. 2007. A genomewide single-nucleotide-polymorphism panel for Mexican American admixture mapping. Am J Hum Genet. 80(6):1014–1023.

Tsai HJ, et al. 2005. Comparison of three methods to estimate genetic ancestry and control for stratification in genetic association studies among admixed populations. Hum Genet. 118(3–4):424–433.

Vongpaisarnsin K, Listman JB, Malison RT, Gelernter J. 2015. Ancestry informative markers for distinguishing between Thai populations based on genome-wide association datasets. Leg Med (Tokyo). 17(4):245–250.

Wang Y, Lu D, Chung YJ, Xu S. 2018. Genetic structure, divergence and admixture of Han Chinese, Japanese and Korean populations. Hereditas 155(1):19.

Wells JC, Pomeroy E, Walimbe SR, Popkin BM, Yajnik CS. 2016. The elevated susceptibility to diabetes in india: an evolutionary perspective. Front Public Health. 4:145.

Wright S. 1969. Evolution and the genetics of populations. Chicago, Illinois, USA: University of Chicago.

Zhu X, et al. 2005. Admixture mapping for hypertension loci with genome-scan markers. Nat Genet. 37(2):177–181.

Ziv E, Burchard EG. 2003. Human population structure and genetic association studies. Pharmacogenomics 4(4):431–441.

**Associate Editor:** Partha Majumder