Forensic Science International: Genetics 61 (2022) 102780

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

Eurasiaplex-2: Shifting the focus to SNPs with high population specificity increases the power of forensic ancestry marker sets

C. Phillips [a,*], M. de la Puente [a], J. Ruiz-Ramirez [a], A. Staniewska [a,b], A. Ambroa-Conde [a], A. Freire-Aradas [a], A. Mosquera-Miguel [a], A. Rodriguez [a], M.V. Lareu [a]

[a] *Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Spain*
[b] *Institute of Anthropology and Ethnology, Adam Mickiewicz University in Poznań, Poland*

ARTICLE INFO

Keywords:
SNPs
South Asia
Forensic ancestry analysis
Population-specific alleles
1000 Genomes
HGDP-CEPH

ABSTRACT

To compile a new South Asian-informative panel of forensic ancestry SNPs, we changed the strategy for selecting the most powerful markers for this purpose by targeting polymorphisms with near absolute specificity – when the South Asian-informative allele identified is absent from all other populations or present at frequencies below 0.001 (one in a thousand). More than 120 candidate SNPs were identified from 1000 Genomes datasets satisfying an allele frequency screen of ≥ 0.1 (10 % or more) allele frequency in South Asians, and ≤ 0.001 (0.1 % or less) in African, East Asian, and European populations. From the candidate pool of markers, a final panel of 36 SNPs, widely distributed across most autosomes, were selected that had allele frequencies in the five 1000 Genomes South Asian populations ranging from 0.4 to 0.15. Slightly lower average allele frequencies, but consistent patterns of informativeness were observed in gnomAD South Asian datasets used to validate the 1000 Genomes variant annotations. We named the panel of 36 South Asian-specific SNPs *Eurasiaplex-2*, and the informativeness of the panel was evaluated by compiling worldwide population data from 4097 samples in four genome variation databases that largely complement the global sampling of 1000 Genomes. Consistent patterns of allele frequency distribution, which were specific to South Asia, were observed in all populations in, or closely sited to, the Indian sub-continent. Pakistani populations from the HGDP-CEPH panel had markedly lower allele frequencies, highlighting the need to develop a statistical system to evaluate the ancestry inference value of counting the number of population-specific alleles present in an individual.

1. Introduction

We developed the original *Eurasiaplex* forensic single nucleotide polymorphism (SNP) ancestry panel in 2013 [1] specifically to enhance the distinction of European and South Asian ancestries. These population groups are more closely positioned geographically and lack physical barriers to migration, so consequently are genetically less well differentiated than other continentally defined population groups. For a sizeable proportion of their genomic variation, populations of the Indian sub-continent show allele frequencies with variability positioned in the middle of an allele frequency cline running between Europe and East Asia. Key additional patterns of variation are defined by differing ratios of variability from the inferred founding populations of Ancestral North Indians and Ancestral South Indians [2,3]. Such variation therefore tends to have limited informativeness for distinguishing South Asian individuals from Europeans or East Asians. Nevertheless, the *Eurasiaplex* SNPs have proved to be a useful set for supplementing other forensic ancestry panels that have a stronger emphasis on differentiating the five continentally based population groups of Africa, Europe, East Asia, America, and Oceania [4–6]. When considering the development of a new set of South Asian-informative SNPs to improve the power of massively parallel sequencing (MPS) ancestry tests for forensic use, we decided to revise the strategy for selecting optimum AIMs. This was accomplished by a change of focus away from allele frequency contrasts between South Asia and Europe and/or East Asia, towards selecting SNPs with near-absolute population specificity, defined in this respect as variants with zero or extremely low allele frequencies in population groups outside the targeted region. Therefore, despite the new variants identified having allele frequencies as low as 0.1 in South Asian populations, in all other regions the allele frequency of the specific allele is generally considerably lower, with values of 0.001 (1-in-1000) or less. Although an allele frequency of 0.1 is seemingly uncommon variation,

* Corresponding author.
E-mail address: c.phillips@mac.com (C. Phillips).

https://doi.org/10.1016/j.fsigen.2022.102780
Received 3 June 2022; Received in revised form 16 September 2022; Accepted 18 September 2022
Available online 20 September 2022

18 % of genotypes will be heterozygotes with the specific allele, and when specific allele frequencies are as high as 0.4, 64 % of genotypes in total carry that allele. These frequencies of a specific allele at any one locus contrast with those in individuals from other regions, including areas neighbouring South Asia, of less than 1 in 1000. As a result, when panels of thirty or more markers with near-absolute specificity are compiled, between 12 and 18 specific-allele genotypes are detected in individuals from the target population group, compared to a maximum of one, two or, much more rarely, three genotypes in individuals from elsewhere.

To develop a second-generation forensic AIMs panel for the analysis of South Asian ancestry, which we called *Eurasiaplex-2*, approximately 120 candidate SNPs were compiled with absolute or near-absolute specificity to South Asia, a region extending from the Indian subcontinent into Afghanistan in the northwest, and Myanmar along with closely sited parts of the SE Asian archipelago in the east. We used the five 1000 Genomes populations with South Asian origins to detect SNPs with specific alleles that had frequencies ranging from 10 % to 40 % in these populations but contrasting with zero frequencies or in the range of 1-in-200 to 1-in-1000 in the project's populations from East Asia, Europe, and Africa. Once identified, the best markers were compiled into a smaller panel of 36 AIMs suitable for inclusion in future globally applicable ancestry panels for SNP genotyping using MPS. The 36 SNPs selected for the *Eurasiaplex-2* panel were cross-checked for consistent South Asian-specific allele frequency patterns amongst the widely dispersed population sampling of five other whole-genome-sequence human diversity projects.

2. Methods and materials

2.1. Changing the concept of population informativeness when selecting forensic ancestry SNPs

Fig. 1 details three ancestry SNPs of potential interest for inclusion in a panel of South Asian-informative markers. SNP rs10008492 was chosen for the original *Eurasiaplex* panel as there is evident differentiation in the rs10008492-C allele frequencies (blue segment) between South Asian and European populations, although it is also evident that this allele only very weakly differentiates East Asian populations. This is reflected in the contrasting pairwise I_n values of 0.16 and 0.02, respectively, shown for each population comparison. The I_n divergence metric is widely used to gauge population differentiations [7] and was the basis for the first *Eurasiaplex* SNP selection process (Fig. 1 of [1]). Another simple metric often used and indicative of ancestry inference power, is the allele frequency differential (δ); the absolute difference of one population's allele frequency from another - in this case, the European-South Asian δ is 0.5, a comparatively high value. SNP rs6053171 provides very good differentiation between Europeans and East Asians, and South Asians have intermediate frequencies for both alleles, consequently giving high I_n and δ, comparing South Asians to both other populations. This SNP was chosen in a study by Pfaffelhuber et al. in 2020 [8] as part of a set of twelve loci to compile an optimum ancestry SNP set for distinguishing African, European, East Asian and South Asian populations, selected using a supervised feature selection system [8]. However, given a South Asian heterozygosity of 48 %, over half of genotypes are homozygous TT or GG and therefore these individuals are not distinguished from Europeans (69 % of TT homozygotes) or from East Asians (31 % of GG). The third SNP rs371763923 has the lowest I_n and δ values of all three loci, so would not be amongst the top selections as an ancestry marker. However, the power of this type of SNP's allele frequency distribution lies in the zero frequency for the rs371763923-G allele (yellow segment) in both Europeans and East Asians. In the 33 % of South Asians where the G allele is detected as a heterozygote, or the 4 % as a homozygote, these genotypes strongly signal origins from this population group as they are not observed elsewhere. Supplementary Fig. S1 details all the individual population allele frequencies in each SNP, indicating there are also zero rs371763923-G allele frequencies in African, Native American, Oceanian and Middle East populations, making this SNP universally specific for South Asia. Only one population, 1000 Genomes KHV (Kinh in Ho Chi Minh City, Vietnam) has two individuals with rs371763923-G alleles, representing < 1 % overall frequency. We selected multiple, well-spaced SNPs on each chromosome that display this kind of highly specific frequency distribution, by applying the strict criterium of the lowest possible specific allele frequency across all populations outside of South Asia. In all cases, the South Asian-specific allele was the Reference Sequence (*RefSeq*) alternative allele, not the reference allele (herein, Alt and Ref alleles, respectively).

2.2. Marker selection

BCFtools was used to make selections of suitable candidate SNPs from publicly available 1000 Genomes Phase III variant catalogues [9]. The chromosome based VCF data from the 1000 Genomes FTP site was searched using the simple allele frequency intersect of: '> 0.1 in South Asian, < 0.01 in African, East Asian, European and admixed American population sample' frequency cut-offs. Multiple-allele variants were excluded, and X-/Y-chromosome variants were identified but not

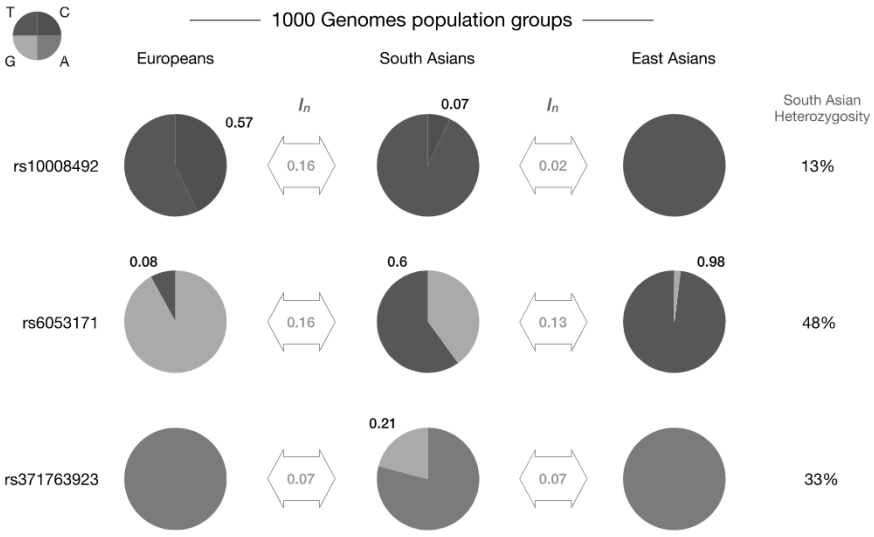

Fig. 1. Three different types of ancestry SNP with potentially South Asian-informative allele frequency distributions. The well-established I_n Divergence measurement of population diversity are shown for each pairwise comparison: European-South Asian left, and South Asian-East Asian right. SNP rs10008492 was part of the original *Eurasiaplex* panel and is informative for the differentiation of Europeans, but not for East Asians. SNP rs6053171 has high Divergence values, but South Asian homozygotes are uninformative for either Europeans or East Asians. Only SNP rs371763923 is equally informative for both comparisons and despite relatively low Divergence values, over 33 % of genotypes would be highly informative heterozygotes or GG homozygotes that are not found in the other populations.

compiled into the final candidate lists for each chromosome. Lists of candidate SNPs were assembled in Excel per chromosome, for further allele frequency comparisons of allele frequencies in African, East Asian, European, and American population samples to maximise the level of South Asian specificity. It should be emphasised that we routinely compile 1000 Genomes variant data from the high sequence coverage datasets generated by the NYGC sequencing of the project's sample set, which has greatly improved the quality of SNP genotype calls across the human genome [10]. The SNP rs3857620 is illustrative of the problem of performing searches for ancestry markers using variant catalogues based on low coverage sequencing data, and which may continue to harbour incorrectly called genotypes. This SNP was identified by Zhao et al., in 2019 [11] as a South Asian-informative marker suitable for a small-scale forensic ancestry panel proposed to consist of 36 SNPs in total. SNP rs3857620 was also adopted for the VISAGE Enhanced Tool for Ancestry and Appearance in a much larger set of SNPs genotyped with MPS [12]. However, the currently listed genotypes in the 1000 Genomes Ensembl portal [13] indicating a 46 % frequency of the A-allele in South Asian populations vs. 0–13 % in the other population groups, contradicts an A-allele frequency estimate across all 1000 Genomes populations sequenced at high coverage, of less than 1 %. For this reason, we cross-checked all current human genome variant databases to ensure each SNP of interest showed consistent patterns across all datasets.

Once compiled, single SNPs were selected from each cluster of markers on any one chromosome segment showing near-identical allele frequency patterns, to optimise the genomic distribution of the final marker set. An exception was made to this rule for chromosome 16, which had a very large extended haplotype of markers with high South Asian specificity at 16p11.2-q11.2. In this case, multiple SNPs were chosen at widely dispersed positions which were clearly part of a large-scale chromosome segment where haplotypes of specific SNP variants had been preserved at very high frequencies in many South Asian populations.

2.3. Human genome variant databases accessed

The SNP variant catalogue of 1000 Genomes Phase III was interrogated with the South Asian-targeted allele frequency intersect described in Section 2.2. All genotypes were cross-checked for accuracy with the NYGC high sequence coverage dataset for each candidate SNP in turn. To further check accuracy of allele frequency estimates made from each set of 1000 Genomes genotypes, the gnomAD (Genome Aggregation Database [14]) v.3.1.2 dataset was checked for all *Eurasiaplex-2* candidate SNPs. The gnomAD database is the largest publicly available collection of population variation compiled from large-scale human genome sequencing projects. It only reports allele frequencies per population but compiles the most up-to-date data from 1000 Genomes (i.e., the NYGC high sequence coverage variant data); whole-genome sequence data for the widely used HGDP-CEPH diversity panel, as well as more than 152,000 samples from large-scale African, European (Finnish and non-Finnish compiled separately), East Asian, South Asian, Middle Eastern, Latino (admixed American), Ashkenazi Jewish and Amish population samples. Studies of patterns of human variation on this scale provide very accurate allele frequency estimates, and gnomAD data is particularly sensitive to very rare variation such as a tri-allelic polymorphism where a second alternative allele (Alt-2) at a SNP site is present in one population at a very low frequency. Genotypes from the HGDP-CEPH panel were obtained from this project's FTP site [15].

Human variant data from Simons Foundation human genome diversity project (herein, SGDP [16]) offers information on geographic areas outside of those covered by 1000 Genomes or the HGDP-CEPH sampling regimes; in particular, Northeast Asia (broadly, Siberia east towards the Bering Straits); Southeast Island Asia; and Central South Asia (broadly, the Caucasus, east towards the central Asian Steppe immediately north of the Hindu Kush, Kashmir, and Tibet). There are 278 samples with genome-wide SNP datasets, of which 22 overlap with 1000 Genomes and 133 overlap with the HGDP-CEPH panel samples, leaving 123 unique samples, from 67 populations, but each a very limited number of 1, 2 or 3 samples per population. The Estonian Biocentre genome diversity panel (herein, EGDP [17]) largely mirrors the sampling regimes of SGDP by being mainly samples of 1, 2 or up to 6 individuals per population or region. EGDP has 402 individual samples from 121 populations that almost all complement the SGDP samples by providing extensive geographic coverage of Siberian, Northeast Asian, Eastern European, and Central South Asian regions. We compiled the 123 SGDP-unique genotype datasets and 402 EGDP genotype datasets for the candidate *Eurasiaplex-2* SNPs.

Lastly, we collected genotypes for the *Eurasiaplex-2* final selection of 36 SNPs from the Singapore Genome Variation Project (herein SGVP), a whole-genome sequencing study from 2014 which included 36 individuals of Indian descent residing in Singapore, plus 96 Malays in Singapore [18,19].

2.4. Statistical considerations

To emphasise the power of specific variation to signal a particular population, Bayes analysis which generates a likelihood ratio (LR) between two possible populations-of-origin starts to build very high cumulative probabilities in SNPs with specific alleles at even moderate frequencies. The highest likelihood ratio for an rs10008492-TT or rs6053171-GG homozygote comparing South Asian and European allele frequencies (shown in Fig. 1) produces a probability of ~4.7 times more likely South Asian. The same likelihood test for an rs371763923-AG heterozygote, and applying a conservative 'global' G allele frequency of 1 % for all non-South Asian populations, produces a probability of 16.75 times more likely South Asian. Given most populations outside of South Asia have a zero frequency for the specific allele of most of the SNPs chosen for *Eurasiaplex-2*, Snipper avoids zero-value numerators in LR calculations by applying the default value of $1/n + 1$; where n is the sample size for that population.

We explored the application of Bayes analysis using the Snipper multiple profiles SNP classifier [20], which accepts multiple SNP profiles and generates principal component analysis (PCA) plots in the same test. Allowance was made for non-independence of linked SNPs when applying the chromosome 16 haplotype SNPs by choosing the 'Hardy--Weinberg principle need not apply' option in Snipper, which adjusts for association of allele frequencies amongst closely sited SNPs on the same chromosome. Additionally, estimations were made of likely recombination rates amongst the haplotype component SNPs by measuring the recombination fraction values between these markers using the HapMap genetic map for this chromosome, as previously described [21].

An alternative to Bayes LR tests and PCA is to run a genetic cluster algorithm such as STRUCTURE [22]. Since we were only compiling markers specific to one population group, there is limited information that can be obtained from STRUCTURE analyses seeking two genetic clusters (i.e., setting analysis runs for a K value of 2). Nevertheless, we performed a simple comparison of STRUCTURE analyses of 1000 genomes populations using *Eurasiaplex* and *Eurasiaplex-2* SNP sets to explore the extra power potentially gained from alleles with absolute specificity to a single population group.

A much simpler and potentially informative alternative to all three of the established population analysis systems, is to simply count the number of specific alleles found in any one individual and assess if these match the patterns observed across the whole region of interest. This was done in the current study on a large scale with the five South Asian 1000 Genomes sample sets and included HGDP-CEPH samples plus those on a much smaller scale, but geographically dispersed and covering a wide range of different populations from the Indian subcontinent samples of SGDP and EGDP.

3. Results

3.1. Screening South Asian-specific candidate SNPs

A total of 123 candidate South Asian-specific SNPs were compiled by applying the allele frequency intersect to the 1000 genomes Phase III variant catalogue. The full candidate SNP set is listed with summary genomic details and complete genotype data (including 1000 Genomes NYGC high coverage, HGDP-CEPH, SGDP and EGDP genotypes) in Supplementary Tables S1A and S1B, respectively. Genotype concordance was checked by comparing the currently published 1000 Genomes Phase III data used for the allele frequency screening, with the high coverage genotype calls from the NYGC re-sequenced 1000 Genomes samples. The 2,505 sample-by-sample comparisons for each of the 123 candidate SNPs are listed in Supplementary Table S1C.

An initial filter set was applied to exclude SNPs that had one of three characteristics: i. genotype discordancy rates higher than 20 incompatibilities (approximately 1 % or more differences in genotype calls); ii. SNPs that lost the expected specific allele pattern when the NYGC genotypes of South Asian samples were compared to those of the other populations; iii. SNP pairs which were physically well separated on the same chromosome segment, but which showed identical allele frequencies indicating they were in linkage disequilibrium - in such cases, one SNP was chosen. Note that the third screening rule was not applied to SNPs in the chromosome-16 extended haplotype. Fig. 2A shows the ten SNPs with more than 20 discordant genotypes, excluded from further consideration. No suitable SNPs were identified on chromosome 21, and the single SNP on chromosome 22, rs113693449, was excluded due to a high level of genotyping discordancy. Fig. 2B provides summary allele frequency charts for the SNPs that were expected to have South Asian-specific alleles in searches of the 1000 Genomes Phase III data, which were not detected in the high coverage NYGC data: rs3857620, rs199671447 and rs113693449. The NYGC data for SNP rs11103281 maintained the allele expected in South Asian samples, but it was also detected in the other population groups so lacked specificity. All four of the above SNPs had zero South Asian-specific allele frequencies in the Gujarati in Houston US (GIH), which suggested discrepancies in the way these SNPs had been genotyped - notably that the GIH population samples were originally studied by HapMap, and this data may simply have been merged with the South Asian populations added to 1000 Genomes Phase III studies. SNP rs9915709 had a much higher frequency in African samples than South Asians, therefore although this SNP was listed, in practice it would be less informative than other SNPs with zero, or near-zero allele frequencies across all non-South Asian populations. Finally, SNP rs371441513 had the highest level of genotype discordancy, which suggests it has sequencing issues, meaning it is unlikely to be reliably genotyped with any assay developed for *Eurasiaplex-2* SNPs.

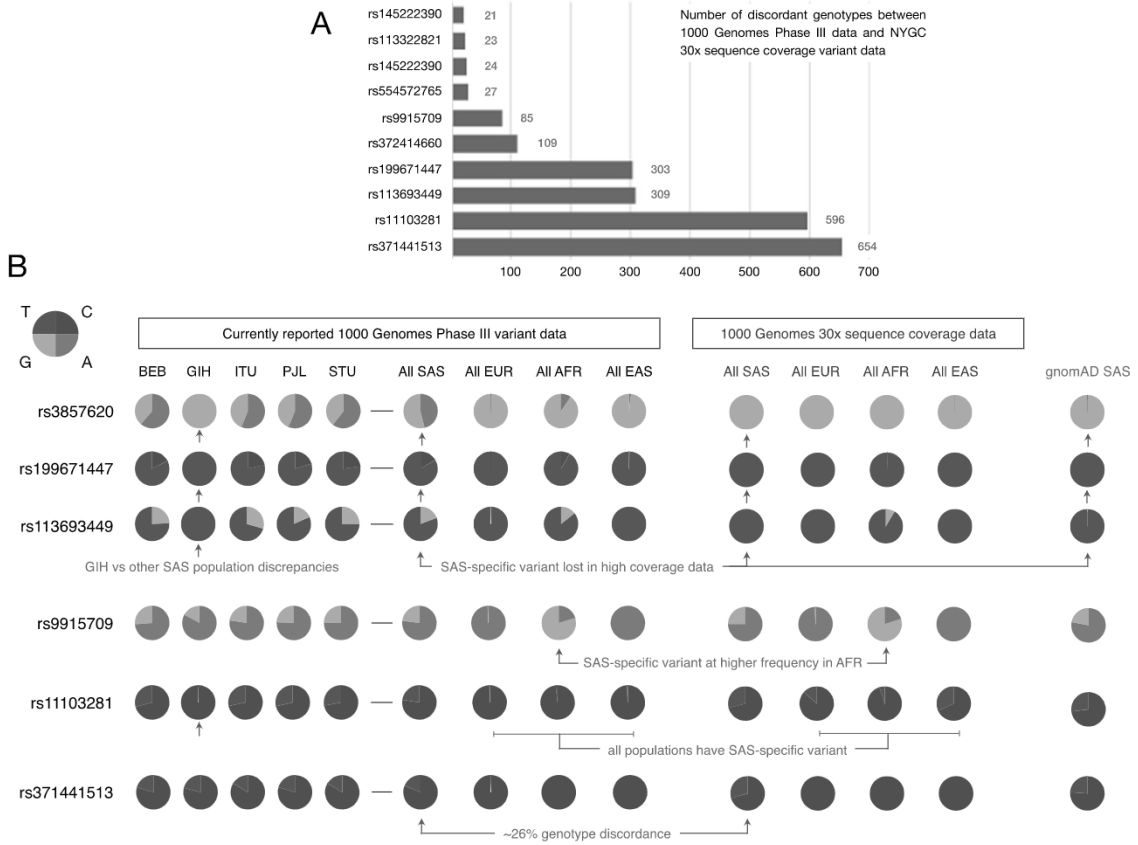

Fig. 2. (A) Ten *Eurasiaplex-2* candidate SNPs with discordant genotypes between the current 1000 genomes Phase III data (2–3x sequence coverage) and the high coverage (30x) NYGC re-sequenced samples. Although the top four have relatively low numbers of discordant genotypes, the other six indicate sequence alignment problems or complex sites (e.g., with closely positioned Indels) and all were rejected to minimise the risk of genotyping problems using MPS. (B) Three *Eurasiaplex-2* candidate SNPs with misleading allele frequency information in the current 1000 Genomes Phase III variant dataset (upper left-hand pie charts for South Asian populations and population group summaries), contrasting with the allele frequency information from the higher coverage sequence analysis data for the same samples, plus gnomAD South Asian data, where frequency estimates are based on ~4800 samples (upper right-hand charts). Lower pie charts show additional problems of insufficient specificity in rs9915709 (South Asian-specific allele at a higher frequency in African populations) and rs11103281 (all population groups have the South-Asian specific allele at >0.05 frequencies), plus SNP rs371441513 with the highest recorded genotype discordancy rate indicating a potentially complex variant site.

3.2. Selecting a core set of South Asian-specific SNPs for Eurasiaplex-2

3.2.1. Patterns of South Asian-specific genotype distributions

Table 1 outlines genomic details and summary allele frequencies of the 36 SNPs selected for *Eurasiaplex-2*. It is noteworthy that only the two SNPs rs77510889 and rs17158407 were previously identified in the 1000 Genomes Phase I variant catalogue (accessible in the SPSmart ENGINES genome browser [23]). This is mainly due to an absence of South Asian population samples in this first 1000 Genomes SNP genotyping project phase, so a SNP with the Alt allele only present in South Asian populations would consist entirely of Ref allele homozygotes in all Phase I populations and thus not be identified as a variant. SNPs were given internal codes based on their chromosome and *RefSeq* 5′ to 3′ locations, from 1A to 20 (i.e., one SNP on this chromosome). Generally, gnomAD South Asian-specific allele frequencies were slightly lower than most or all of those in 1000 Genomes. With the exception of rs374908464, the CEPH Pakistani allele frequencies averaged across the eight populations, are substantially lower than either 1000 Genomes or gnomAD South Asian samples. The overall average South Asian-specific allele frequency of 0.085 in CEPH Pakistani samples is two- to three-times lower, and for 27 of the 36 SNPs would not meet the selection criteria of > 0.1 allele frequencies.

The full list of genotypes and summary allele frequency estimates for the 36 SNPs in each population group are listed in Supplementary Table S1D. For each of the 4097 samples listed in this table, numbers of South Asian-specific genotypes and alleles were counted. The individual South Asian-specific genotype counts are plotted as blue bars in Fig. 3A (the four 1000 Genomes population groups), and Fig. 3B (all other samples from HGDP-CEPH, SGDP, EGDP and SGVP SNP databases), with the small CEPH Uyghur and EGDP Roma population samples highlighted as red bars for clarity. Aligned above each set of bar plots in Fig. 3A and B are graphic representations of South Asian-specific allele homozygotes (red lines) and heterozygotes (orange lines), with non-specific allele homozygotes in grey. All 36 of these non-specific alleles are the *RefSeq* reference allele. These graphics highlight the large number of informative genotypes recorded in the 1000 Genomes South Asians, with only rs77510889 (internal code 'SNP 10′) indicating a higher-than-average number of South Asian-specific genotypes in 1000 Genomes African and CEPH African populations. Note that there are four SNPs not genotyped in EGDP, five in SGVP Singapore Indian, and two in Singapore Malay genome datasets. In each of the other databases, South Asian samples are clearly indicated by blue bars with prominent heights and the corresponding dense patterns of orange and red lines. The average number of informative genotypes per individual is given above each South Asian population box (values for the individual CEPH Pakistani populations below the bar plots). The contrast is evident between the CEPH Pakistani populations and 1000 Genomes South Asian populations, with an average of ~14 specific genotypes per individual in BEB, ITU, GIH, STU populations, dropping to less than 12 in PJL, Punjabi from Lahore, Pakistan. The CEPH Pakistani populations have much lower average numbers of informative genotypes per individual, which range from ~8 in the Sindhi to ~2 in the Hazara samples. In the other datasets, the SGDP and EGDP South Asian, plus SGVP Singapore Indian samples have close to an average of 11 informative genotypes per individual (adjusted for the missing SNPs), which can be taken to represent an overall median number of informative genotypes for this target population group in the 36 SNPs selected. The other 1000 Genomes population groups have the expected very low average informative genotype value of less than 0.2, allowing a degree of differentiation to be made between East Asians and the South Asian-related samples of CEPH Uyghur, occupying regions to the northeast of the Indian sub-continent; and EGDP Roma, a trans-national cultural isolate suggested to have originated from a proto-Romani population living in northwest India [24]. The Uyghur have relatively low average informative genotype levels of 1.8, but the Roma, although only five individuals and based on 32/36 SNPs, show a high average value of ~6.7 informative genotypes.

3.2.2. South Asian-specific allele frequency estimates from Eurasiaplex-2 SNP genotypes

Although genotype frequencies provide indications of the population informativeness of the SNPs selected for *Eurasiaplex-2*, examination of allele frequencies allows a clearer overview of how the SNP variability is distributed across the Indian sub-continent. When population-wide average South Asian-specific allele frequencies are calculated for each of the 36 SNPs from 1000 Genomes combined South Asian populations, gnomAD samples (approximately 5000 South Asians, with no specific geographic data provided), and CEPH Pakistani datasets, the contrast between Pakistan and the rest of South Asia is further underlined. Fig. 4 gives bar plots for the South Asian-specific average allele frequencies per SNP in each dataset, ranking the markers from most informative to least, left to right. The Fig. 4 plots suggest a close match in frequency estimates between 1000 Genomes and gnomAD data, but much lower specific allele frequency estimates for Pakistanis that are between 10 % and 25 % of the other dataset values. The reciprocal bar plots in Fig. 4 use a 100-fold smaller scale and show most South Asian-specific allele frequencies in other parts of the world rarely exceed 0.001–0.003, or a maximum of 1 in 300. The boxed values for rs77510889 indicate high frequencies for the G-allele outside of South Asia, which exclude population data where this allele was at a particularly high frequency - notably the HGDP-CEPH hunter-gatherer populations of San, Mbuti Pygmies and Biaka Pygmies. South Asian-specific alleles in the rest of the world for rs186371551 and rs184748067 also had higher-than-average frequencies but were more widely dispersed.

Although there are very different sample sizes between the populations sampled by 1000 Genomes, CEPH, SGDP and EGDP (approximately 100, 25, 2–3, and 2–6, respectively), it is instructive to map the distribution of South Asian-specific allele frequencies in all the populations studied which are located in, or near, the Indian sub-continent. Fig. 5 provides pie charts of the average percentage of South Asian-specific alleles in each population sample in 1000 Genomes, CEPH, SGDP and EGDP (percentage values adjusted for four missing SNPs in EGDP) from in or near the Indian sub-continent, mapped to their approximate geographic locations. The UK-resident 1000 Genomes Indian Telugu (ITU) and Sri Lankan Tamil (STU); and US-resident Gujarati (GIH) populations are placed in their approximate locations of origin, and the EGDP Roma have no sampling location described. The CEPH Cambodian and EGDP Aeta from the Philippines are also too distant from the centre of South Asia to be easily placed on this map. Note that the average percentage of South Asian-specific alleles in all other population samples not shown in Fig. 5 was less than 0.5 %, except the two SGVP samples: Indians in Singapore = 20 %; Malays in Singapore = 0.76 %. Patterns of average allele frequency distributions in the 31 population samples shown in Fig. 5 indicate a high percentage value in most populations within India and Bangladesh, but a sharp drop in these values in populations from regions northwest and east of this country. A notable exception is the 1000 Genomes Punjabi from Lahore, Pakistan, with this population group occupying the most easterly part of Pakistan at the northwest corner of India. The smallest recorded average percentage of South Asian-specific alleles were observed in the most geographically distant Iranians in the West (1.38 %), Uyghur in the north (2.5 %) and Cambodians in the east (2 %).

3.3. Analysis of the six Eurasiaplex-2 SNPs on chromosome 16

Six SNPs chosen from chromosome 16 (C16) had the highest levels of South Asian specificity and three were clustered around the centromere, potentially meaning they could show full or high levels of allelic association due to reduced recombination, precluding their use as independent loci. Table 2 summarises the recombination rates between the six SNPs using the HapMap genetic map database to estimate map distances in Centimorgans (cM) and Kosambi-adjusted recombination fractions (Rc). The Rc estimates in Table 2 indicate the three 5′ SNPs rs368479296-rs376893831-rs370130302 show minimal linkage with

Table 1
Genomic details and South Asian-specific allele frequencies from 1000 Genomes, HGDP-CEPH and gnomAD databases of an optimum set of 36 SNPs.

Genomic details									Frequency of the alternative (South Asian-specific) allele in 1000 Genomes groups/South Asian populations								HGDP-CEPH	gnomAD
No.	Code	Chr.	GRCh37	GRCh38	rs-number	Ref.	Alt.	Gene	African	European	East Asian	BEB	GIH	ITU	PJL	STU	Pakistani	South Asian
1	1A	1	27588988	27262497	rs191008849	T	C	WDTC1	0	0.0010	0.0010	0.2267	0.2500	0.2353	0.1927	0.2353	0.2171	0.1866
2	1B	1	207023473	206850128	rs370300597	C	G	–	0	0	0	0.2093	0.1977	0.1961	0.1302	0.2549	0.2072	0.1459
3	2A	2	18702265	18520999	rs373262633	A	G	LOC105373454	0	0	0	0.1802	0.1977	0.2451	0.1354	0.1765	0.1836	0.1336
4	2B	2	98816440	98199977	rs183145214	A	G	VWA3B	0	0	0.0050	0.2151	0.1744	0.1520	0.2031	0.2157	0.1873	0.1861
5	3A	3	44250057	44208565	rs578118259	T	G	–	0	0	0.0010	0.2558	0.2791	0.2843	0.1875	0.2353	0.2444	0.1817
6	3B	3	159603038	159885249	rs369609492	C	A	SCHIP1	0	0	0	0.2500	0.2209	0.2157	0.0938	0.1912	0.1836	0.1572
7	3C	3	167730032	168012244	rs375081853	A	G	GOLIM4	0	0	0.0010	0.1570	0.1512	0.2059	0.1615	0.1912	0.1873	0.1588
8	4A	4	115827968	114906812	rs182767282	T	C	NDST4	0	0.0010	0.0010	0.1453	0.1919	0.2157	0.2031	0.2598	0.2270	0.1703
9	4B	4	152007237	151086085	rs146398591	A	G	–	0	0.0010	0.0030	0.2733	0.2500	0.3235	0.1875	0.2794	0.2457	0.2192
10	4C	4	167677008	166755857	rs554572765	A	C	SPOCK3	0	0	0	0.1860	0.1919	0.1912	0.1458	0.2010	0.1811	0.1068
11	5	5	33704125	33704020	rs375710694	T	C	ADAMTS12	0	0	0.0020	0.1512	0.1628	0.2206	0.1615	0.2059	0.1923	0.1468
12	6A	6	117206569	116885406	rs186371551	G	A	RFX6	0	0.0050	0.0050	0.1279	0.1686	0.2304	0.1823	0.1716	0.2084	0.1703
13	6B	6	130116672	129795527	rs368650154	C	T	–	0	0	0.0010	0.1686	0.1395	0.2255	0.1615	0.2402	0.1935	0.1578
14	6C	6	154459950	154138815	rs368661757	C	A	OPRM1	0	0	0.0010	0.2384	0.2500	0.2451	0.2135	0.2549	0.2444	0.1946
15	7	7	50238881	50199285	rs368444091	C	T	–	0	0	0	0.1686	0.1860	0.1814	0.1667	0.1912	0.1774	0.1396
16	9	9	97560517	94798235	rs187619767	C	T	AOPEP	0	0	0.0020	0.2442	0.2849	0.1618	0.1458	0.2255	0.1849	0.1740
17	10	10	122095086	120335574	rs77510889*	A	G	–	0.0714	0	0.0020	0.2267	0.2093	0.2304	0.2031	0.2206	0.2171	0.1593
18	11A	11	29262859	29241312	rs370097977	T	C	–	0	0	0.0050	0.1628	0.1686	0.1912	0.2031	0.2451	0.2258	0.1529
19	11B	11	59462759	59695286	rs375766368	G	A	–	0	0	0.0010	0.1453	0.1512	0.2157	0.1771	0.1520	0.1824	0.1222
20	11C	11	72175159	72464115	rs377589165	G	A	–	0	0	0	0.1512	0.1860	0.2206	0.1615	0.1863	0.1873	0.1514
21	12A	12	4268703	4159537	rs376263717	C	T	–	0	0.0020	0.0010	0.2267	0.2035	0.2598	0.1510	0.2010	0.1911	0.1655
22	12B	12	22570119	22417185	rs371763923	A	G	–	0	0	0.0020	0.2500	0.2442	0.2059	0.1510	0.2304	0.2022	0.1636
23	12C	12	50428379	50034596	rs368764180	A	G	RACGAP1	0	0	0.0020	0.1570	0.1395	0.1961	0.0885	0.1863	0.1787	0.1388
24	13	13	56057499	55483364	rs184748067	G	A	–	0	0.0060	0.0069	0.1744	0.1802	0.2353	0.1615	0.2304	0.1998	0.1600
25	14	14	65712298	65245580	rs189013802	G	A	–	0	0	0.0079	0.1744	0.1744	0.2402	0.1823	0.2059	0.1873	0.1372
26	15	15	83236825	82568075	rs17158407*	C	T	CPEB1	0	0.0010	0.0020	0.2558	0.2558	0.2990	0.3125	0.3333	0.2990	0.2681
27	16A	16	3178971	3128970	rs368479296	C	T	ZNF213-AS1	0	0	0.0020	0.1802	0.1802	0.1422	0.1667	0.2108	0.1836	0.1645
28	16B	16	23053815	23042494	rs376893831	G	T	–	0	0	0.0010	0.2849	0.2674	0.1569	0.1823	0.1961	0.2146	0.1991
29	16C	16	28588059	28576738	rs370130302	C	G	SGF29	0	0	0.0010	0.2907	0.2674	0.1961	0.1719	0.2206	0.2134	0.2024
30	16D	16	33921593	34119126	rs368738705	C	T	–	0	0	0	0.4186	0.4302	0.4853	0.3385	0.4412	0.4007	0.3329
31	16E	16	46499858	46465946	rs368538881	C	T	–	0	0	0	0.4128	0.4070	0.4951	0.3021	0.4608	0.4020	0.3549
32	16F	16	48327788	48293877	rs377323011	A	G	LONP2	0	0	0.0010	0.3314	0.2965	0.4314	0.2760	0.3775	0.3362	0.2742
33	17A	17	43964966	45887600	rs369091847	A	T	MAPT-AS1	0	0	0	0.1337	0.1512	0.2402	0.1250	0.2451	0.2035	0.1604
34	17B	17	80660204	82702328	rs376153825	G	C	LOC105376791	0	0	0.0010	0.1279	0.1279	0.1765	0.1771	0.1716	0.1762	0.1273
35	19	19	8371240	8306356	rs374908464	A	G	CD320	0	0	0.0020	0.1802	0.2093	0.2500	0.1719	0.1078	0.1985	0.1740
36	20	20	4987550	5006904	rs186201674	C	T	SLC23A2	0.001	0.0040	0.0050	0.2093	0.2209	0.2255	0.1875	0.2451	0.2109	0.1545
					* SNP also identified in 1000 Genomes Phase I			Average:	0.002	0.001	0.002	0.214	0.216	0.240	0.182	0.233	0.219	0.178

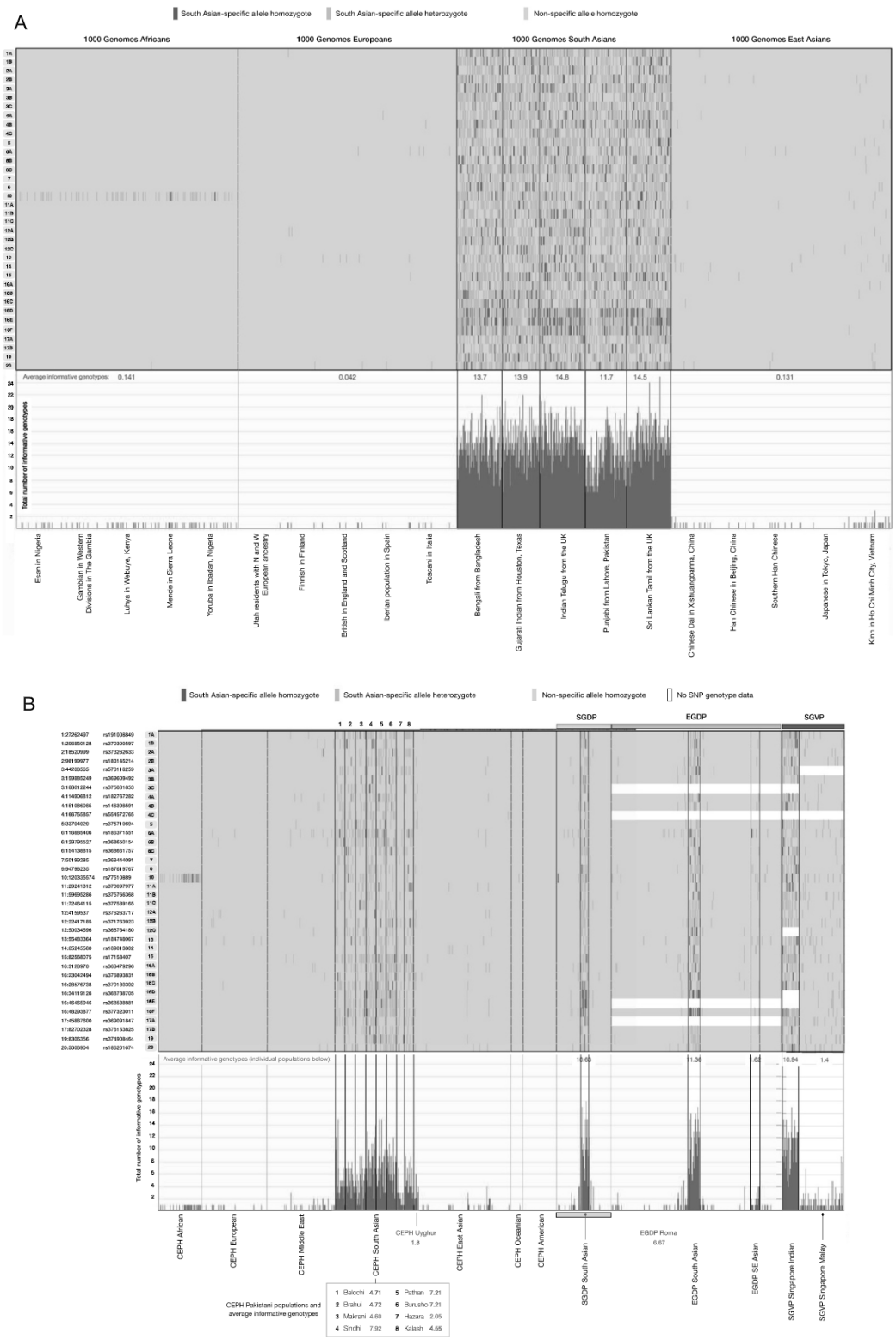

Fig. 3. (A) South Asian-specific genotype distributions and total number of informative genotypes in each 1000 Genomes sample; specific allele homozygotes marked in red and heterozygotes in orange (both genotypes counted singly) of 36 *Eurasiaplex-2* SNPs in twenty 1000 Genomes populations. The average number of informative genotypes are shown as group-wide values for Africans, Europeans, and East Asians, and for individual populations for South Asians. Internal codes are used to label each SNP, which are detailed in B. (B) South Asian-specific genotype distributions and total number of informative genotypes in four whole-genome-sequencing human diversity projects, additional to 1000 Genomes: HGDP-CEPH diversity panel; Simons Foundation genome diversity project (SGDP); Estonian Biocentre diversity project (EGDP); Singapore genome variation project. HGDP-CEPH Uyghur and EGDP Roma geographic outlier samples are marked in red. The average number of informative genotypes are shown individually for eight HGDP-CEPH Pakistani populations and HGDP-CEPH Uyghur; SGDP South Asian samples; EGDP South Asian, SE Asian and Roma samples; SGVP Singapore Indian and Singapore Malay samples. Note EGDP lacks data for four SNPs, SGVP lacks data for five in Singapore Indian and two in Singapore Malay samples. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

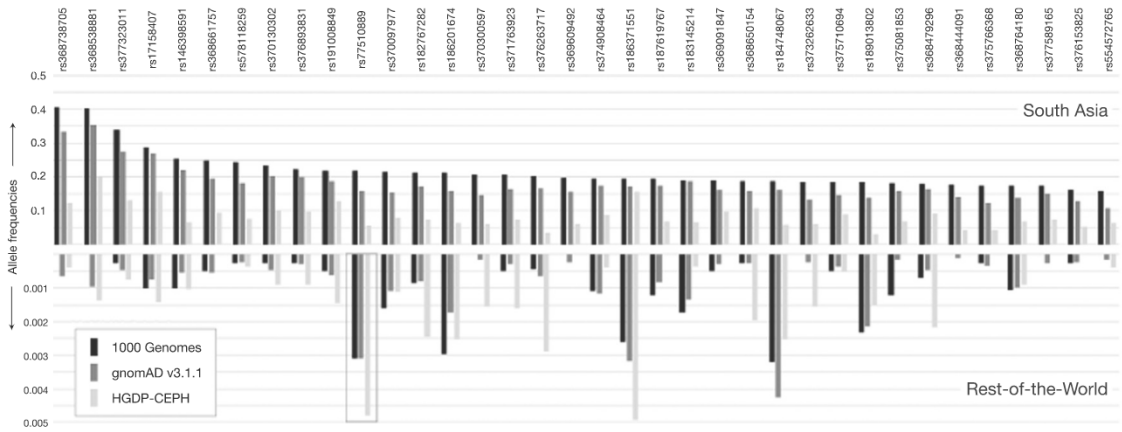

Fig. 4. Allele frequency spectra of the 36 *Eurasiaplex-2* SNPs. Markers are arranged in descending 1000 Genomes South Asian-specific average allele frequency and show bars for 1000 Genomes South Asians, gnomAD South Asians and CEPH Pakistanis. Rest-of-the-World average frequencies are compiled from 1000 Genomes African, European, and East Asian average frequencies, plus gnomAD and HGDP-CEPH average non-South Asian population data. The Rest-of-the-World axis shows 1/100th allele frequency values compared to those of South Asian populations. The boxed values of rs77510889 exclude high frequencies for the South Asian-specific allele amongst HGDP-CEPH African Khoisan, Biaka Pygmies and Mbuti Pygmies (rs77510889-G allele frequency of 0.23 in HGDP-CEPH African hunter-gatherer populations vs. 0.06 in HGDP-CEPH Pakistani populations).

32 % and 11 % recombination fractions, but the other three 3′ centromeric SNPs rs368738705-rs368538881-rs377323011 have a much lower level of recombination of ~0.35 % across the 3-SNP span. The top half of Fig. 6 outlines the structural landscape of these three C16 centromeric SNPs and the common genotype combinations they form in South Asian vs. other populations.

To explore the possibility of association between the rs368479296-rs376893831-rs370130302 SNPs, we decided to treat them as a haplotype and gauge haplotype diversity in the 1000 Genomes samples (excluding admixed samples). Although the Rc values between these SNPs are very low, the physical distances in megabases (Mb) are much bigger, with the 12.35 Mb span between rs368738705 and rs368538881 alone representing nearly 14 % of the total C16 length. Therefore, it would not be possible for 1000 Genomes to accomplish accurate phasing of alleles for this series of SNPs over such distances. We decided to convert any localised phasing (i.e., a SNP allele's phase with reference to its immediate neighbours) made by 1000 Genomes of rs368479296-rs376893831-rs370130302 heterozygotes, into alphabetic order, respectively: TC>CT, TC>CT, GA>AG. This created the root haplotype of CCA – universally present in all 1000 Genomes non-South Asian samples, plus the South Asian-specific TTG haplotype, exclusively confined to these populations in 1000 Genomes. In this way, any 3-SNP genotypes which are not CC-CC-AA; TT-TT-GG; or, CT-CT-AG, represent disruptions to the South Asian-specific haplotypes which signify reduced allelic association. We counted the derived haplotypes amongst South Asian individuals as homozygous genotypes in each SNP, specifically (in bold): **CC**-CT-AG and **TT**-CT-AG in rs368479296; CT-**CC**-AG and CT-**TT**-AG in rs376893831; CT-CT-**AA** and CT-CT-**GG** in rs370130302. These patterns can be interpreted to indicate disrupting recombination between rs368479296-rs376893831, double recombination between rs368479296-rs376893831 and rs376893831-rs370130302, or recombination between rs376893831-rs370130302, respectively. Although true phasing cannot be achieved, the extent to which the above six derived haplotypes occur in 1000 Genomes populations will indicate the level of disruption of allelic association amongst the three SNPs. The rs368479296-rs376893831-rs370130302 inferred haplotypes for all population samples are listed in Supplementary File S1E and the numbers of each haplotype in 1000 Genomes populations are summarised in the lower half of Fig. 6. Except for a singleton CCG haplotype (present as CC-CC-AG genotypes in a Vietnamese KHV sample), all 3,021 non-South Asian samples from 1000 Genomes had the CCA root haplotype. Amongst the South Asian samples, 206 were inferred to have specific TTG haplotypes and 464 non-specific CCA haplotypes, but a significant number of South Asian specific haplotypes, a total of 308, were derived, i.e., inferred to be different combinations of alleles to either CCA or TTG. This would suggest there is very little association between the centromeric C16 SNPs. Furthermore, levels of recombination between these three SNPs are likely to be higher, given the South Asian individuals with CT-CT-AG genotypes (88 of 489, 18 %) cannot be reliably phased.

3.4. Statistical analyses

3.4.1. Conventional Bayes analysis of South Asian population variability

The results of the Bayes analysis likelihood assessments and PCA patterns generated by Snipper, are summarised in Supplementary File S1. First, evaluations were made of pairwise cumulative Divergence ($\boldsymbol{I}_{\mathbf{n}}$) values calculated for South Asian vs. European, and vs. East Asian populations, using 1000 Genomes data and comparing the 23 SNPs of the original *Eurasiaplex* with the 36 of *Eurasiaplex-2* panel. Supplementary File S1.1 shows the cumulative $\boldsymbol{I}_{\mathbf{n}}$ for South Asians vs. Europeans at the point of 23 SNPs have similar values in *Eurasiaplex* (1.87) compared to *Eurasiaplex-2* (2.01), but for South Asian vs. East Asian variation *Eurasiaplex* only reaches 0.77, compared to 1.99 from 23 *Eurasiaplex-2* SNPs. This highlights how frequencies for the specific alleles of zero, or near zero outside of the targeted population, produce almost identical $\boldsymbol{I}_{\mathbf{n}}$ values in all populations comparisons and for each SNP added. This is expressed in the cumulative $\boldsymbol{I}_{\mathbf{n}}$ chart in Supplementary File S1 as two diverging and flattening curves in *Eurasiaplex*, compared with the two straight lines with identical trajectories in *Eurasiaplex-2*. Therefore, when the final cumulative values are calculated for 36 *Eurasiaplex-2* SNPs, each population comparison has almost identical values of 2.84 (vs. Europe) and 2.79 (vs. East Asia). When population specific SNPs are combined in the future to differentiate all the main population groups, it will be straightforward to balance the $\boldsymbol{I}_{\mathbf{n}}$ values for each comparison as this will just entail adjustment of the number of SNPs needed to reach a final cumulative value that can be matched across all populations.

Second, the distribution of likelihood ratios (LR) from the comparison of 1000 Genomes South Asian and African likelihoods (Africans produced the second highest likelihoods in five population comparisons) using Bayes analyses in Snipper, was compiled in a chart of ranked LRs shown in Supplementary File S1.2. These values are generally much higher than those observed with alternative ancestry SNPs [4,6], with the bulk of samples producing LR values in excess of 1E+ 12 or '1 in a

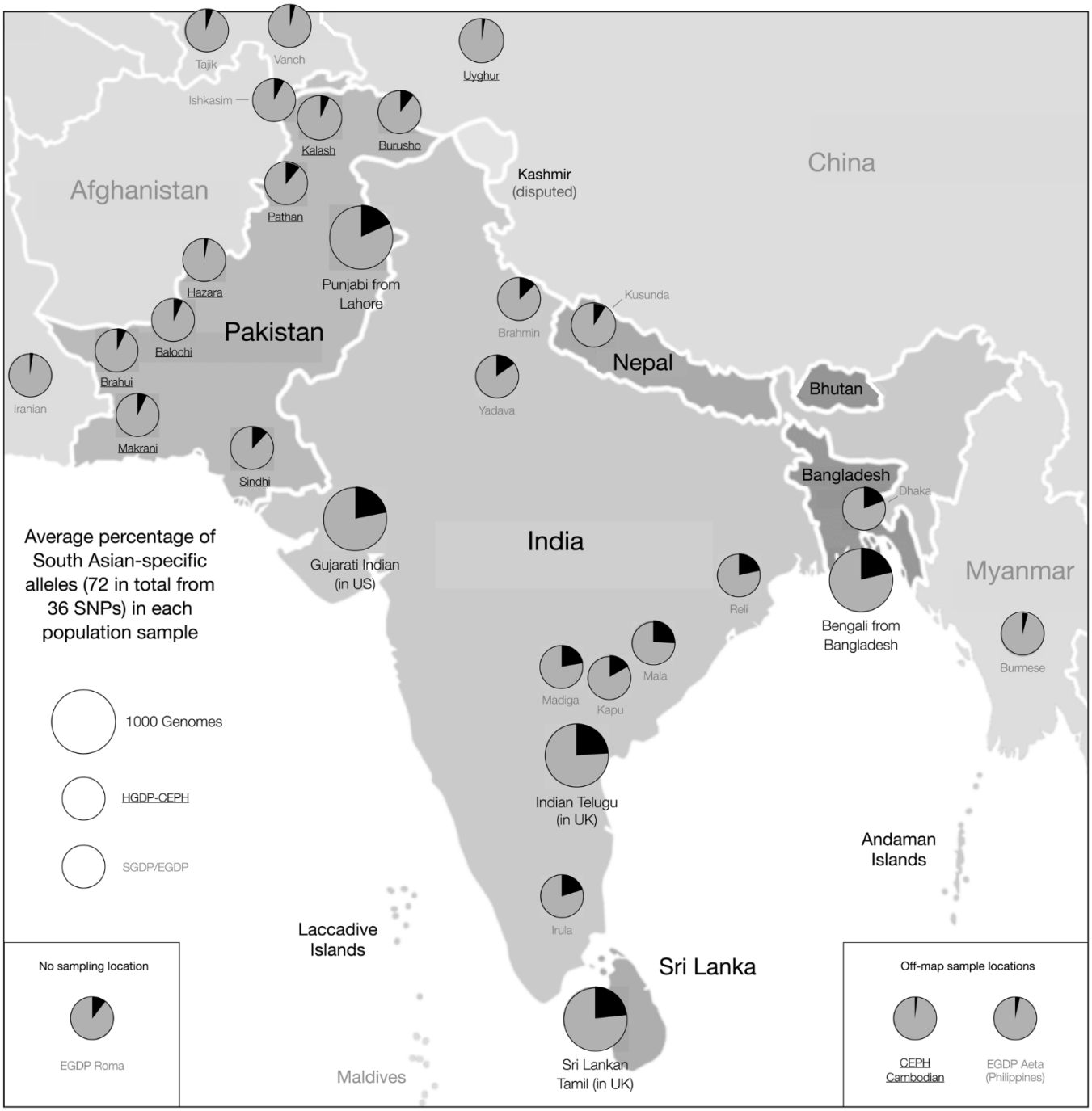

Fig. 5. The average percentage of South Asian-specific alleles (from a total of 72) in each population sample with a recorded value higher than 0.5 %. All such populations are located within or closely sited to the Indian sub-continent apart from the CEPH Cambodian and EGDP Aeta of the Philippines. Populations to the west, east and north of India tend to show much lower average percent specific alleles compared to the maximum value of 24 % seen in the 1000 Genomes STU and ITU population samples. These population samples based in the UK, and the GIH in the US, are positioned in their approximate geographic location. EGDP Roma samples have no stated sampling location.

trillion times more likely to be from South Asia than Africa'. Nevertheless, five individuals had LR values below '2000 times more likely', that all corresponded to samples with the lowest number (6) of South Asian-specific alleles. Third, two PCA plots are shown in Supplementary File S1.3 from analyses in Snipper of 1000 Genomes YRI, CEU, CHB, and CEPH Native American, Oceanian populations compared with CEPH Pakistanis (Plot 1), and compared with 1000 Genomes GIH (plot 2). These plots illustrate the very tightly distributed set of PCA points in populations outside of South Asia obtained with *Eurasiaplex-2* SNPs. The target population PCA point distributions are different between Pakistanis and GIH, with Pakistanis equally diffuse in distribution, but overlapping with the much smaller area of the 2D plot occupied by populations outside of South Asia. While PCA itself would not be a system of choice for assigning ancestry and this represents analysis with a single set of population-specific markers, it is interesting that zero or near-zero allele frequencies in most populations causes samples, even in large numbers, to occupy a very small area of the PCA plot.

Table 2
HapMap genetic map analysis of the six chromosome 16 South Asian-specific SNPs in *Eurasiaplex-2*. The map distance was estimated in Centimorgans (cM) and the recombination fraction (Rc) calculated from the cM values using Kosambi adjusted data. The 5′ SNPs rs368479296-rs376893831-rs370130302 show very little linkage with 32 % and 11 % recombination fractions, but the other three 3′ SNPs rs368738705-rs368538881-rs377323011 (in bold) have a much lower level of recombination of less than 0.3 %.

Internal ID	SNP	GRCh37 position	GRCh38 position	cM inter-SNP distance	Kosambi-adjusted Rc
16A	rs368479296	3178971	3128970		
16B	rs376893831	23053815	23042494	38.681	0.324515
16C	rs370130302	28588059	28576738	11.3898	0.111968
16D	**rs368738705**	33921593	34119126	1.793	0.017922
16E	**rs368538881**	46499858	46465946	0.0422	0.000422
16F	**rs377323011**	48327788	48293877	0.3118	0.003118

3.4.2. Genetic cluster analysis with STRUCTURE comparing Eurasiaplex and Eurasiaplex-2 SNPs

Supplementary File S1.4 shows the cluster plots from separate STRUCTURE analysis of 1000 Genomes populations using the original 23 *Eurasiaplex* SNPs and the 36 *Eurasiaplex-2* SNPs. Patterns show that the *Eurasiaplex-2* SNP set clearly distinguishes South Asians from all other 1000 Genomes populations at K:2, with a clean set of columns and some minor mixed cluster proportions in the PJL. As there is almost no genetic variation present in non-South Asian samples for *Eurasiaplex-2* SNPs, no other genetic clusters are identified at K:3, K:4, or higher K values (data not shown). It is noteworthy that the South Asian-specific rs77510889-G allele detected in African populations did not produce an African cluster for any higher K values analysed. In contrast, the 23 *Eurasiaplex* SNPs distinguish Europeans as the first major genetic cluster at K:2, then Africans at K:3, with South Asians only emerging as a differentiated population group at K:4. To some degree, these patterns are likely to reflect the selection of the original *Eurasiaplex* SNPs that had strongly contrasting allele frequencies between Europeans and other population groups, including small, but above-average allele frequency differences between Europeans and South Asians.

3.4.3. Exploration of a simple South Asian-specific allele counting system

Arguably, a much more straightforward system of population assignment than Bayes analysis can be achieved by simply counting the number of South Asian-specific *Eurasiaplex-2* alleles in an individual. Fig. 3 illustrating the worldwide distribution of genotype counts, and Fig. 5, those of allele counts greater than 0.5 %, show they are both almost completely confined to the Indian sub-continent and adjoining areas. These patterns show that highly contrasted allele counts in individuals from South Asia vs. individuals from other regions will give strong indications of origins from the regions targeted by *Eurasiaplex-2,* provided there is no overlap between minimum and maximum counts from each set of populations. Fig. 7 plots the distribution of South Asian-

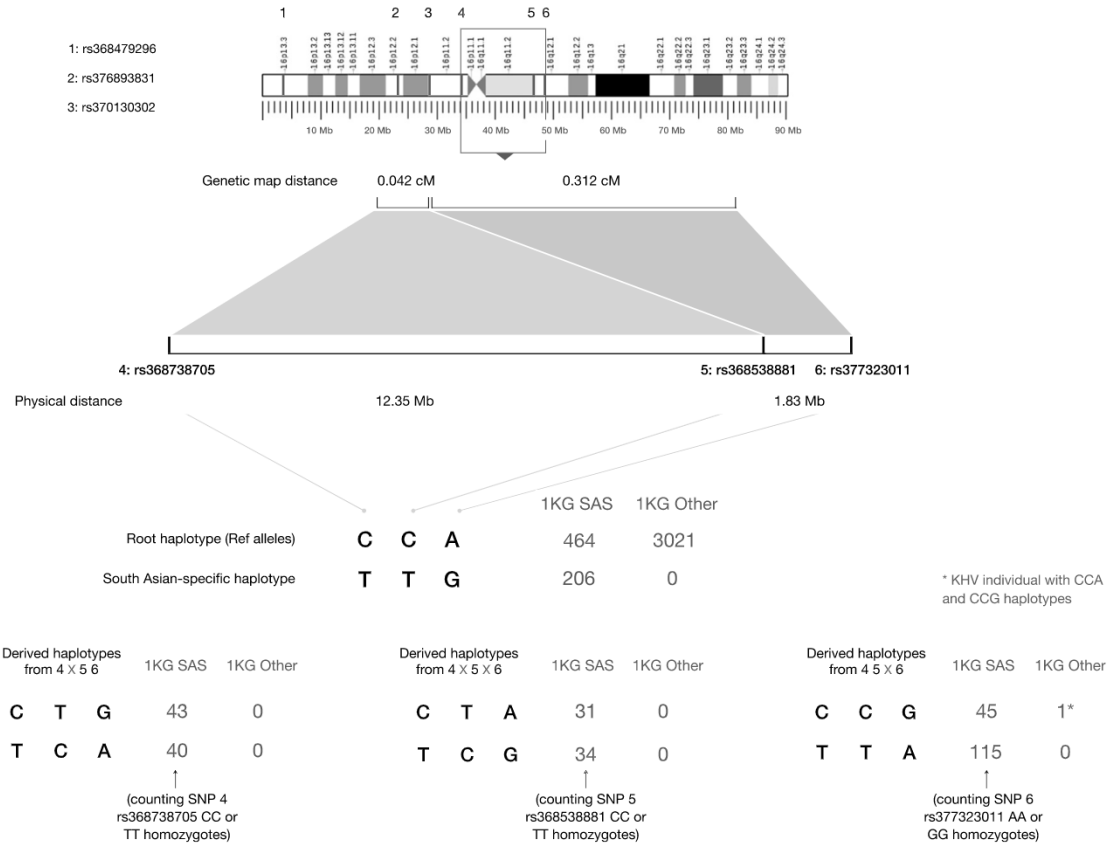

Fig. 6. The six SNPs of Chromosome 16 (red bars) and the haplotype landscape around SNPs 4–6 [rs368738705-rs368538881-rs377323011] closest to the centromere, where Centimorgan (cM) genetic map distances are particularly small (red box). As all three SNPs have alphabetic ordering for the Ref and Alt (South Asian-specific) alleles - i.e., CT, CT, AG, respectively, all 1000 Genomes heterozygous genotypes were alphabetised, and homozygote genotypes counted in order to record recombination between the root haplotype CCA and the South Asian-specific haplotype TTG. In this way rs368738705 CC or TT homozygote genotypes indicated recombination between rs368738705-rs368538881 (‘4 × 5 × 6′); rs377323011 AA or GG indicated recombination between rs368538881-rs377323011 (‘4 5 × 6′); and double sequential recombination events (‘4 × 5 × 6′) were recorded as rs368538881 AA or CC homozygotes. Counts are given for the 1000 Genomes South Asians (1KG SAS) vs. all other 1000 Genomes populations (1KG Other, excluding admixed individuals), indicating widespread disruption of CCA and TTG haplotypes and likely minor levels of association amongst the South Asian-specific alleles of these three SNPs. The single CCG haplotype recorded in a non-South Asian individual was inferred from CC-CC-AG genotypes. Mb: megabase; KHV: Kinh in Ho Chi Minh City, Vietnam.

specific allele counts in the four 1000 Genomes population groups plus CEPH Pakistanis. In the 1000 Genomes Africans, Europeans and East Asians, the majority of samples, some 80–90 %, have no specific alleles present in any genotypes. Africans have 16 % of genotypes with a single specific allele, mainly due to the rs77510889-G allele in these populations, but that represents the upper limit of specific allele counts in this population group. Only East Asians have a few individuals with more than one specific allele, with five samples having two alleles and a singleton with a maximum of three. The lower limit of specific alleles in South Asians is 5, with a singleton sample with this number, and then ten with 6 alleles, meaning there is no overlap in counts between the South Asians of 1000 Genomes and all individuals from the other population groups. Although a count of three, four or five specific alleles might be considered ambiguous, such a small number of individuals could stay unassigned. Therefore, there would be a very low probability of incorrect assignment of individuals using a lower limit of six specific alleles to signal South Asian origins.

The CEPH Pakistanis show a distribution with a greater degree of overlap with populations outside of South Asia, which is not unexpected, but strict adherence to a minimum six specific alleles means just over half of Pakistanis (88 of 168 with five or less South Asian-Specific alleles) are not assigned as South Asian. We intend to develop a statistical system for the handling of population-specific allele counts, based on hypothesis testing where the null hypothesis represents origins in the specific-allele target population. This will be more easily accomplished when a globally applicable set of population-specific SNPs has been compiled for all population groups.

4. Discussion

By constructing a SNP panel composed of a completely new type of ancestry marker focused on variation that is specific to the single targeted South Asia population group, rather than using SNPs with highly contrasting but shared variation across multiple populations, we have identified a characteristic signature of South Asian origin in almost all individuals from the Indian sub-continent. This specific-allele signature, illustrated by the dense pattern of orange and red bars in Fig. 3, is clearly observed across a large, broadly-based collection of South Asian samples taken from across the world. Apart from rs77510889, that showed a relatively high frequency of the South Asian-specific G-allele in Africans, all other SNP alleles chosen to be specific to South Asia had frequencies below 0.005 (0.5 %) in populations outside this region. Applying the same allele frequency intersect to other 1000 Genomes population groups plus Oceanian, American, and Middle East populations represented in the HGDP-CEPH panel, will allow the compilation of a global population-specific ancestry marker set, with marker numbers appropriate for MPS-scale SNP multiplexes. The process of building a large MPS multiplex requires some adjustments of targeted SNPs with poor context sequence or flanking region variation that interferes with sequence alignments or primer binding. For this reason, we chose to report a full list of suitable South Asian-specific candidate loci, rather than build a small-scale multiplex of 30–40 markers, as we have done for many forensic SNP panels previously [1,25,26]. The near-equal cumulative Divergence values between South Asians vs. East Asians and vs. Europeans shown by the 36 SNPs in *Eurasiaplex-2*, will make the process of balancing SNPs specific for each population group relatively straightforward, as the number of markers can be adjusted to produce a comparable average number of population-specific alleles per individual from that population.

The variant data and its statistical treatment that we have briefly explored in this study, requires a rethink of how best to adapt highly population-specific SNP allele patterns of variation into a forensic ancestry prediction framework. We do not feel that Bayes analysis or PCA will provide the necessary detailed assessments of the number of alleles specific to a given population that are detected in an individual. We expect STRUCTURE to be more sensitive to specific-allele patterns as well as being able to efficiently analyse co-ancestry in individuals with admixed backgrounds [6,12]. Until a global panel of population-specific SNPs is constructed this will need to be explored *in-silico* once enough candidate SNPs for each population group have been compiled. It is also useful to begin to develop a statistical framework that centres on formal testing of hypotheses for H1: originating from the target population vs. H2: originating from another.

An ancestry SNP selection process that enriches for markers with near-absolute specificity is more likely to detect variation present in the targeted population due to less common evolutionary genetic processes than those that underlie traditionally compiled ancestry sets (i.e., natural selection and genetic drift). These processes might include recent

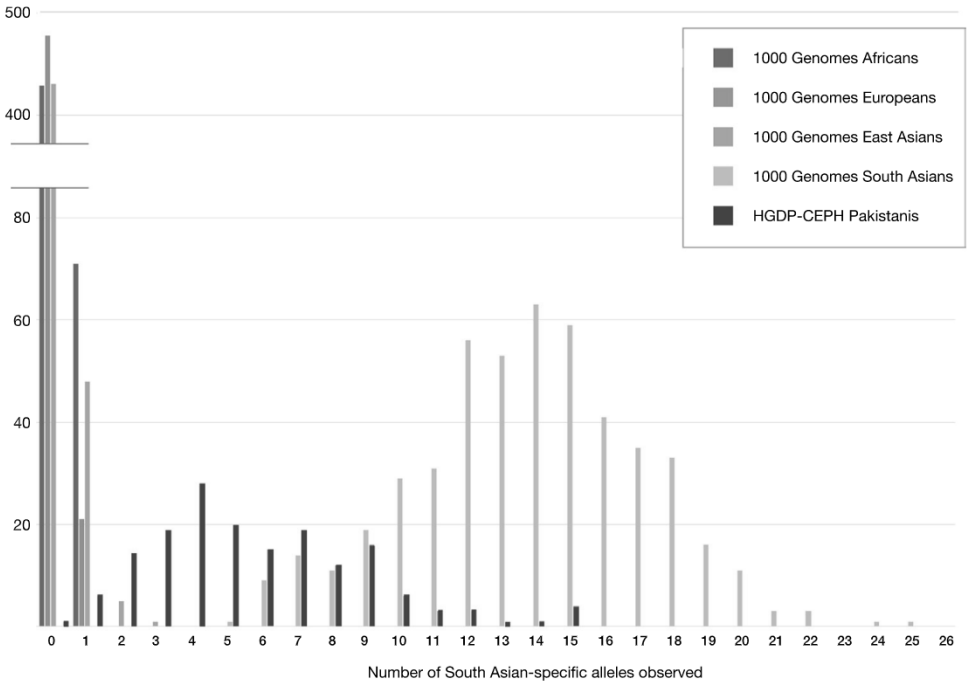

Fig. 7. Distribution of South Asian-specific allele counts in the four 1000 Genomes population groups, showing a bell-shaped distribution of specific alleles in the South Asian populations, and the lowest value represented by a single South Asian individual with five specific alleles. There is no overlap with the distribution of specific allele counts in the other 1000 Genomes populations with a single East Asian individual with a maximum three specific alleles. CEPH Pakistani samples show a degree of overlap with a single individual having no specific alleles, but more than half of Pakistani samples having less than a nominal six specific alleles lower limit to signify South Asian origins.

mutation events creating new SNP variants confined to a specific geographic region [27]; gene flow from Hominin introgression taking place in a particular locality [28,29]; localised selective sweeps, which might favour certain low frequency variants which then become region-specific [30]. Any of these processes could have occurred after the South Asian root populations of Ancestral North Indians and Ancestral South Indians [2,3] separated from other Eurasian groups of populations [31]. There is also the additional complexity of the caste system in Indian populations creating highly stratified distributions of variability across the sub-continent (e.g., the Brahmin caste has higher Iranian ancestry than other Indian castes and this differentiation would be maintained by reduced outbreeding across castes [32]). The silk roads provided a strong driver of East-West gene flow across the central parts of Eurasia, but these were largely routed to the north of the Himalayas which acted as a lengthy barrier to mass movements into the Indian sub-continent. Overall, the South Asian SNP variation we have identified and compiled represents only a very small proportion of total genomic variability, but the markers have maintained their high levels of specificity by consistently showing zero, or near zero allele frequencies in every region outside of the Indian sub-continent studied so far.

Acknowledgements

M.d.l.P. is supported by a post-doctorate grant funded by the Consellería de Cultura, Educación e Ordenación Universitaria e da Consellería de Economía, Emprego e Industria from Xunta de Galicia, Spain (ED481D-2021-008). J.R. is supported by the "Programa de axudas á etapa predoutoral" funded by the Consellería de Cultura, Educación e Ordenación Universitaria e da Consellería de Economía, Emprego e Industria from Xunta de Galicia, Spain (ED481A-2020-039).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fsigen.2022.102780.

References

[1] C. Phillips, A. Freire Aradas, A.K. Kriegel, M. Fondevila, O. Bulbul, C. Santos, F. Serrulla Rech, M.D. Perez Carceles, Á. Carracedo, P.M. Schneider, M.V. Lareu, *Eurasiaplex*: a forensic SNP assay for differentiating European and South Asian ancestries, Forensic Sci. Int. Genet. 7 (2013) 359–366.
[2] D. Reich, K. Thangaraj, N. Patterson, A.L. Price, L. Singh, Reconstructing Indian population history, Nature 461 (2009) 489–494.
[3] P.P. Majumder, The human genetic history of South Asia, Curr. Biol. 20 (2010) R184–187.
[4] C. Phillips, W. Parson, B. Lundsberg, C. Santos, A. Freire-Aradas, M. Torres, M. Eduardoff, C. Børsting, P. Johansen, M. Fondevila, et al., Building a forensic ancestry panel from the ground up: the EUROFORGEN Global AIM-SNP set, Forensic Sci. Int. Genet. 11 (2014) 13–25.
[5] J.R. Kidd, F.R. Friedlaender, W.C. Speed, A.J. Pakstis, F.M. De La Vega, K.K. Kidd, Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples, Invest. Genet. 2 (2011) 1.
[6] M. de la Puente, J. Ruiz-Ramírez, A. Ambroa-Conde, C. Xavier, J. Pardo-Seco, J. Álvarez-Dios, A. Freire-Aradas, A. Mosquera-Miguel, T.E. Gross, E.Y.Y. Cheung, et al., Development and evaluation of the ancestry informative marker panel of the VISAGE basic tool, Genes 12 (2021) 1284.
[7] N.A. Rosenberg, L.M. Li, R. Ward, J.K. Pritchard, Informativeness of genetic markers for inference of ancestry, Am. J. Hum. Genet. 73 (2003) 1402–1422.
[8] P. Pfaffelhuber, F. Grundner-Culemann, V. Lipphardt, F. Baumdicker, How to choose sets of ancestry informative markers: a supervised feature selection approach, Forensic Sci. Int. Genet. 46 (2020), 102259.
[9] The 1000 Genomes Project Consortium, A. Auton, L.D. Brooks, R.M. Durbin, E. P. Garrison, H.M. Kang, J.O. Korbel, J.L. Marchini, S. McCarthy, G.A. McVean, et al., A global reference for human genetic variation, Nature 526 (2015) 68–74.
[10] M. Byrska-Bishop, U.S. Evani, X. Zhao, A.O. Basile, H.J. Abel, A.A. Regier, A. André Corvelo, W.E. Clarke, R. Musunuri, K. Nagulapalli, et al., High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios, bioRxiv preprint, posted February 7 2021 doi: ⟨https://doi.org/10.1101/2021.02.06.430068⟩.
[11] S. Zhao, C.-M. Shi, L. Ma, Q. Liu, Y. Liu, F. Wu, L. Chi, H. Chen, AIM-SNPtag: a computationally efficient approach for developing ancestry-informative SNP panels, Forensic Sci. Int. Genet. 38 (2019) 245–253.
[12] J. Ruiz-Ramírez, M. de la Puente, C. Xavier, A. Ambroa-Conde, J. Álvarez-Dios, A. Freire-Aradas, A. Mosquera-Miguel, A. Ralf, C. Amory, M.A. Katsara, et al., Development and evaluations of the ancestry informative markers of the VISAGE enhanced tool for appearance and ancestry, Forensic Sc. Int. Genet. (2022).
[13] ⟨http://www.ensembl.org/Homo_sapiens/Variation/Explore?r=6:60527829–60528829;v=rs3857620;vdb=variation;vf=169483878⟩ (Accessed June 2022).
[14] M. Lek, K.J. Karczewski, E.V. Minikel, K.E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J.S. Ware, A.J. Hill, B.B. Cummings, et al., Analysis of protein-coding genetic variation in 60,706 humans, Nature 536 (2016) 285–291.
[15] A. Bergström, S.A. McCarthy, R. Hui, M.A. Almarri, Q. Ayub, P. Danecek, Y. Chen, S. Felkel, P. Hallast, J. Kamm, J, et al., Insights into human genetic variation and population history from 929 diverse genomes, Science 367 (2020) 1339–1349.
[16] S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo, M. Zhao, N. Chennagiri, S. Nordenfelt, A. Tandon, et al., The Simons Genome Diversity Project: 300 genomes from 142 diverse populations, Nature 538 (2016) 201–206.
[17] L. Pagani, D.J. Lawson, E. Jagoda, A. Mörseburg, A. Eriksson, M. Mitt, F. Clemente, G. Hudjashov, M. DeGiorgio, L. Saag, et al., Genomic analyses inform on migration events during the peopling of Eurasia, Nature 538 (2016) 238–242.
[18] L.-P. Wong, R.T. Ong, W.T. Poh, X. Liu, P. Chen, R. Li, K. Koi-Yau Lam, N. Esakimuthu Pillai, K.-S. Sim, H. Xu, et al., Deep whole-genome sequencing of 100 southeast Asian Malays, Am. J. Hum. Genet. 92 (2013) 52–66.
[19] L.-P. Wong, J. Kuan-Han Lai, W.-Y. Saw, R.T. Ong, A. Youzhi Cheng, N. Esakimuthu Pillai, X. Liu, W. Xu, P. Chen, J.-N. Foo, et al., Insights into the genetic structure and diversity of 38 South Asian Indians from deep whole-genome sequencing, PLoS Genet. 10 (2014), e1004377.
[20] ⟨http://mathgene.usc.es/snipper/analysismultipleprofiles.html⟩.
[21] C. Phillips, D. Ballard, P. Gill, D. Syndercombe Court, A. Carracedo, M.V. Lareu, The recombination landscape around forensic STRs: accurate measurement of genetic distances between syntenic STR pairs using HapMap high density SNP data, Forensic Sci. Int. Genet. 6 (2012) 354–365.
[22] L. Porras-Hurtado, Y. Ruiz, C. Santos, C. Phillips, Á. Carracedo, M.V. Lareu, An overview of STRUCTURE: applications, parameter settings, and supporting software, Front. Genet. 4 (2013) 98.
[23] J. Amigo, A. Salas, C. Phillips, ENGINES: exploring single nucleotide variation in entire human genomes, BMC Bioinf. 12 (2011) 105.
[24] A. Gómez-Carballa, J. Pardo-Seco, L. Fachal, A. Vega, M. Cebey, N. Martinón-Torres, F. Martinón-Torres, A. Salas, Indian signatures in the westernmost edge of the European Romani diaspora: New insight from mitogenomes, PLoS One 8 (2013), e75397.
[25] M. de la Puente, C. Santos, M. Fondevila, L. Manzo, EUROFORGEN-NoE Consortium, A. Carracedo, M.V. Lareu, C. Phillips, The Global AIMs Nano set: a 31-plex SNaPshot assay of ancestry-informative SNPs, Forensic Sci. Int. Genet. 22 (2016) 81–88.
[26] C. Phillips, L. Manzo, M. de la Puente, M. Fondevila, M.V. Lareu, The MASTiFF panel - a versatile multiple-allele SNP test for forensics, Int. J. Leg. Med. 134 (2020) 441–450.
[27] L.M. Williams, M.F. Oleksiak, Ecologically and evolutionarily important SNPs identified in natural populations, Mol. Biol. Evol. 28 (2011) 1817–1826.
[28] S. Sankararaman, N. Patterson, H. Li, S. Pääbo, D. Reich, The date of interbreeding between Neandertals and modern humans, PLoS Genet. 8 (2012), e1002947.
[29] E. Huerta-Sánchez, F.P. Casey, Archaic inheritance: supporting high-altitude life in Tibet, J. Appl. Physiol. 119 (1985) 1129–1134.
[30] H. Chen, N. Patterson, D. Reich, Population differentiation as a test for selective sweeps, Genome Res. 20 (2010) 393–402.
[31] V.M. Narasimhan, N. Patterson, P. Moorjani, N. Rohland, R. Bernardos, S. Mallick, I. Lazaridis, N. Nakatsuka, I. Olalde, M. Lipson, et al., The formation of human populations in South and Central Asia, Science 365 (2019) eaat7487.
[32] G. Debortoli, C. Abbatangelo, F. Ceballos, C. Fortes-Lima, H.L. Norton, S. Ozarkar, E.J. Parra, M. Jonnalagadda, Novel insights on demographic history of tribal and caste groups from West Maharashtra (India) using genome-wide data, Sci. Rep. 10 (2020) 10075.