Forensic Genomics
Volume 3, Number 4, 2023

DOI: 10.1089/forensic.2023.0013

Forensic Genomics

Open camera or QR reader and scan code to access this article and other resources online.

ORIGINAL ARTICLE

Biogeographical Ancestry Estimation from Autosomal Short Tandem Repeats in the Sequencing Era

Laurence Devesse,[1,*] Lucinda Davenport,[1] Denise Syndercombe Court,[1] and David Ballard[1]

Abstract

Autosomal short tandem repeats (STRs) are, and likely always will be, the first loci targeted for forensic DNA analysis as they offer the highest probability of individual identification. An ancestry-informative marker panel can then be used in "no hit, no suspect" cases, which requires additional time and cost investment, and relies on the presence of sufficient remaining sample. Traditionally this has relied on the use of specific ancestry-informative single nucleotide polymorphisms (SNPs), run as an additional test to STRs. STRs have largely been discounted for biogeographic ancestry determination due to their high mutation rate, which in turn makes them well suited for individual identification. Being able to obtain a DNA profile that can simultaneously be used both for biogeographical ancestry estimation and searching against offender databases would be of huge benefit to the field of forensic identification in terms of time, cost, and sample availability. As routine DNA testing of autosomal STRs progresses to next-generation/massively parallel sequencing, the opportunity presents itself to make use of observed sequence diversity in new ways. In particular, the presence of population-specific sequence variation raises the prospect of using STR profiles for population identification, both on their own and in combination with ancestry-informative SNPs. In this study, data were extracted from 989 samples from five global population groups prepared and sequenced using the ForenSeq DNA Signature Prep kit and the MiSeq FGx. Good differentiation between population was achieved using sequenced STR profiles, with 84% of samples classifying correctly using a conservative classification approach, and a general error rate of 3.5%—results that also showed a clear improvement over length-based data.

Keywords: biogeographical ancestry, STR, massively parallel sequencing

Introduction

Being able to accurately determine the genetic ancestry of an individual is of high interest in multiple areas, including the field of personal genomics, with individuals submitting samples to companies such as 23andme and Ancestry.com to gain insight into their biogeographic ancestry. In the field of forensics, a DNA profile obtained from a sample is typically compared with a reference profile (belonging to a victim, suspect, or from an elimination sample), or searched against a database. If no matches are obtained, a traditional profile is considered to provide no information more than the sex of the person who contributed the DNA.

[1]*King's Forensics, Department of Analytical, Environmental and Forensic Sciences, Faculty of Life Sciences and Medicine, King's College London, London, United Kingdom.*
A portion of this article was previously published in the thesis of Laurence Devesse, available at https://kclpure.kcl.ac.uk/ws/files/176389573/2022_Devesse_Laurence_1212813_ethesis.pdf

*Address correspondence to: *Laurence Devesse, Faculty of Life Sciences and Medicine—King's Forensics, King's College London, 150 Stamford Street, London SE1 9NH, United Kingdom,* Email: laurence.a.devesse@kcl.ac.uk

Additional tools, in the context of investigative intelligence, have been developed for the inference of biogeographical ancestry[1,2] and externally visible characteristics.[3,4] Given that eyewitness testimony is notoriously unreliable even when available,[5] techniques that allow the estimation of a person's biogeographic ancestry can be hugely beneficial for criminal investigations. It can also be used in the context of missing persons or mass disaster victim identification, in order to achieve more complete identifications.[6]

Markers used for the inference of biogeographical ancestry, often referred to as ancestry-informative markers (AIMs), are generally single nucleotide polymorphisms (SNPs),[7] but other markers such as microhaplotypes,[8,9] nucleotide insertions, or deletions[10] or short tandem repeats (STRs) have also been used.[10–13] Autosomal SNPs (aSNPs) have traditionally been the AIM of choice due to their stability, density of distribution, and range of allelic frequencies across global populations. Identifying candidate SNPs involves looking for the most pronounced allele frequency differences between populations.

A large number of SNP panels have been developed over the years, often designed to discern ancestry between specific populations,[14,15] or at the global level.[2,16,17] In 2013, the genetics department at Yale university published a highly discriminative 41-SNP panel that aimed to meet two forensically relevant criteria: first, the ability to distinguish ancestral origin at the continental level and second, an assay that reduced both the cost and quantity of DNA required to obtain useable results.

These SNPs have been leveraged by many laboratories in the forensic community, as well as by commercial companies. The 56 SNPs targeted by primer mix B in the Verogen ForenSeq DNA Signature Prep kit were selected from the since extended Yale university SNP panel,[18] and have been demonstrated to accurately estimate ancestry of individuals from European and East Asian origin.[19] In the past, SNP-AIMs for these purposes would have been detected using Sanger sequencing or the SNaPshot primer extension assay,[4,20] but the advent of massively parallel sequencing (MPS) has made the process considerably easier by allowing a greater number and type of markers to be sequenced simultaneously.

Traditionally, microsatellite markers such as autosomal STRs (STRs) have not been considered to any serious extent in the field of ancestry estimation due to the limited contrast in allelic frequencies between populations. Core STR loci were primarily selected for their high levels of polymorphism, enabling the discrimination of unrelated individuals. Prior to the application of MPS technologies to forensics, there was already a significant push to improve the discrimination power of these STRs.

The analysis of STR genotypes to infer genetic ancestry has been studied in the context of length-based allelic variation by a number of groups, but results have invariably highlighted limitations and inferior capabilities compared with SNP multiplexes designed for this reason. There are generally two broad approaches for consideration: either the adoption of specific STRs with strong population differentiation[10,11] or looking at traditional markers for the ability to distinguish populations.[12,13]

Rosenberg et al.[21] used 377 STRs to successfully differentiate the major global population groups, but an assay of this size would often be inappropriate in the context of a forensic scenario. Phillips et al.[22] used frequency data from this 377-marker data set to identify tetranucleotide makers with the highest level of population informativeness. They generated a 12-plex of AIM-STRs that could be used as a standalone test or combined with identity-informative STRs or even an AIM-SNP panel.

Moriot et al.[10] genotyped the CEPH Human Genome Diversity panel (CEPH-HGDP) for 23 deletion–insertion polymorphism (DIP)-STR markers, which combine an insertion/deletion and a closely linked STR, selected from a set developed for improved mixture deconvolution. The rationale behind this marker choice lays in the fact that they are not only forensically relevant, but the combination of fast- and slow-mutating markers would be beneficial for both individual and global population differentiation. Preliminary results from this study were promising, and clustering was considerably better when combining the data than when looking at each type of variation individually (STR or DIP); however, this set requires more markers in order to efficiently distinguish Eurasian populations.

Whilst these panels can be powerful tools when combined with standard STR typing, the fact remains that core forensic STRs are always typed first and foremost in routine profiling. If data from these STRs could be used effectively to distinguish global populations, this would be a huge advantage to the field of investigative intelligence. In 2001, Lowe et al.[23] suggested an approach for inferring ethnic origin using the six STR loci utilised by the UK Forensic Science Service at the time. They gathered allelic frequencies from the National DNA Database for five UK populations (Caucasian, AfroCarribean, Indian subcontinent, Southeast Asian, and Middle Eastern) to estimate the population proportion of a given profile to any of these populations.

Although promising from a research perspective, the authors emphasise the limitation of this work and the need for a larger number of more informative loci. Londin et al.[24] highlighted the need for small panels for accurate ancestral determination when studying diseases in populations, and initially looked at genotyping samples using the Identifiler and Coriell Identity mapping kits, which in combination target 19 STRs. They established that these markers were not sufficient in number or

ancestry informativeness for accurate ancestry determination and went on to develop a panel of specific AIM-STRs as the groups above have.

During their investigation into global variability of the 15 established and 5 new European Standard Set (ESS) STRs, Phillips et al. concluded that the CEPH populations of Europe, Middle East, and South Asia did not show sufficiently differentiated allele variation using these core STRs.[25] Using the program STRUCTURE,[26] they did, however, show clear differentiation between the African, European, and American populations in the CEPH group. When combining data from the 20 STR set with 34 AIM-SNPs, they were even able to differentiate the Oceanian population, which they were not able to previously do using AIM-SNPs alone.

Algee-Hewitt et al. conducted ancestry estimation using core STR sets, and found that a reduced number of markers limits the resolution of ancestry inference.[13] The same group later assessed the utility of these markers for ancestry estimation from postmortem blood cards.[27] Using a single cluster approach to ancestry inference, 17 of the 20 samples tested classified into an ancestry group corresponding to their self-reported ancestry.

In 2010, Pereira et al.[28] published a new online calculator, designed to assign samples to one of three main population groups: Eurasian, East Asian, or sub-Saharan African based on length-based data from 17 autosomal STR loci. They tested this tool on 48 samples from the three ancestral groups, and obtained 86% accuracy for individual population affiliation. The results from these publications highlight two key points when it comes to using STR data for population differentiation:

(1) Length-based data from core STR loci data can be used to roughly distinguish between ancestrally different populations such as Europe, East Asia, and West Africa.
(2) A higher number of STRs, or specifically chosen AIMs, are necessary for more accurate and reproducible population differentiation from autosomal STRs.

MPS offers increased discrimination of STRs through access to sequence level information, as well as a higher multiplexing capability. The large number of alleles for forensic STR markers is an indication of their instability over population divergence time, whereas SNPs in the flanking regions are likely to be more stable and, therefore, offer better ancestry resolution.[10]

The added value of providing investigative leads such as biogeographic ancestry estimation is already a major well-documented advantage of using MPS, although it mostly relies on the targeting of specific ancestry-informative SNP markers typed as a standalone panel or in tandem with STRs. This research looks to address the possibility of using traditional autosomal STR markers for ancestry inference, based on sequencing data. Results are compared to those obtained using ancestry-informative SNPs.

Materials and Methods

Samples and library preparation

Data were collected from work performed for previous publications,[29,30] but is described here in brief. Approximately 200 samples from each of the following population groups were selected: White British, British Chinese, Northeast African, South Asian, and West African, from individuals who are resident in the United Kingdom. Ancestry information for each individual was self-declared at the time of sample collection. Individuals gave informed consent for their DNA to be used for research purposes, and ethical approval for this work was granted by the King's College London research ethics subcommittee (HR-16/17-2594).

The ForenSeq DNA Signature Prep Kit (Verogen, San Diego, CA) was used to prepare samples for sequencing.[31] Samples were prepared in batches of 96 (including a positive and negative control) as per manufacturer's guidelines when using DNA Primer Mix A. A minor modification to the volume of pooled libraries loaded for sequencing was made, as described in the previous publication by Devesse et al.[29]

In order to acquire ancestry-informative SNP genotypes for a subset of the samples, 47 samples from each group taken forward for concordance and frequency generation were selected from the samples described above. These were analysed using a custom primer mix provided by Verogen, containing only the primers for the 22 phenotype and 56 ancestry-informative SNPs usually found in DNA primer mix B (DPMB). Library preparation was performed in batches of 96 reactions (including a positive and negative amplification control) and sequenced in three runs on a MiSeq FGx using Verogen MiSeq FGx Micro Sequencing kits.

STR and SNP genotyping

STR genotypes for 989 samples were characterised and verified in a previous publication to generate sequence-based allelic frequencies from the following five population groups: White British ($n = 207$), British Chinese ($n = 193$), Northeast African ($n = 198$), South Asian ($n = 189$), and West African ($n = 202$). These same genotypes were used in this work, and the same designation was used to characterise alleles (length based = based on length of the repeat region alone; sequence based without flanking regions = based on the sequence of the repeat region provided by the Universal Analysis Software (UAS); sequence based with flanking regions = based on the entire sequence of the amplicon).

D22S1045 is known to have poor heterozygous balance in the ForenSeq DNA Signature Prep Kit,[31,32] and drop out was frequently observed at this locus. Because of this, genotype data for this locus were not carried forward for any ancestry analysis. The end file for SNP analysis is a variant call format (.VCF) file which is a standard output used in MPS genotyping and highlights the target SNPs and variants compared to a reference genome. To generate these files, a script was used to undertake the following steps:

(1) BWA: The sequences in the FASTQ files were aligned to a reference file containing the target sequences of interest (DPMB SNPs) using the MEM algorithm within BWA.[33] This reference file was created by searching for the target SNPs on dbSNP[34] and taking approximately 100 bases on either side of the SNP. For the HIrisPlex SNPs, the sequence between the published primers was used.[3] BWA creates sequence alignment map (SAM) files that are aligned to the reference sequences provided.
(2) SAMTools: SAM files are converted to BAM files, sorted indexed and intermediary files removed using SAMTools.[35]
(3) GATK: The Genome Analysis Toolkit[36] is then used to highlight all variants to the reference sequences in the initial reference file.

A final RStudio script was used to modify the files into a more user-friendly format, such as providing heterozygous allele balance and adding conditional formatting to the values in the spreadsheet for easier visualization. The files generated from this script were Excel files and were then collated to provide all genotypes for all samples in a run in a single workbook.

Results were manually verified using a set of genotyping ''rules,'' namely:

- Minimum number of reads to consider an allele genuine: 3.
- Minimum number of reads to consider a homozygous genotype genuine: 20.
- Heterozygous balance: anything below 0.85 was considered imbalanced.
- Samples with 15 or more poor genotypes (imbalanced, or where drop out has occurred) were removed from further analysis.

Based on these criteria, a final number of 219 samples were taken forward for SNP ancestry analysis (White British, $n=42$; South Asian, $n=39$; Northeast African, $n=45$; British Chinese, $n=47$ samples; and West African, $n=47$).

Ancestry analysis

STRUCTURE. GenAlEx (Genetic Analysis in Excel) was downloaded freely as an add-on to Microsoft Excel and used to manipulate STR and SNP genotype data into a format appropriate for downstream analysis.[37] Ancestry estimation was performed using STRUCTURE version 2.3.4,[26] which uses a model-based clustering algorithm to infer population structure from multilocus genotype data. Parameters were set at 100,000 for burn-in and 100,000 Markov Chain Monte Carlo repetitions, using the admixture model of analysis. A graphical display of the admixture model results shows each individual as a single vertical line, and the membership of proportion to each inferred K group is represented by splitting this line into different colours.

The model uses a Bayesian approach to discerning K genetic clusters within the data, using allelic frequencies. These frequencies are assumed to be in Hardy–Weinberg equilibrium, and genetic markers are assumed to be in linkage equilibrium. K values were set depending on the data being analysed. Results from STUCTURE were displayed graphically using the program CLUMPAK (Clustering Markov Packager Across K).[38]

Rosenberg informativeness for assignment measure. The informativeness for assignment (I_n) measure is used to determine the amount of information that multiallelic markers provide about individual ancestry.[39] Rosenberg et al. proposed this value with the purpose of reducing the genotyping required for ancestry inference, i.e., using a smaller subset of markers of highest informativeness will reduce the number of markers needing to be targeted whilst achieving the desired result in terms of ancestry estimation.

Infocalc is a script used to calculate statistics that measures the ancestry information content of genetic markers,[39] including the I_n measure. This perl script was downloaded from (https://rosenberglab.stanford.edu/infocalc.html). Analyses were performed without a weight file, meaning that each population is equally likely to be the source population. Given numbers of samples per population was practically identical, this was considered acceptable. The output file created for each run of Infocalc was copied into an excel spreadsheet.

Results and Discussion

Population-specific alleles

Due to population genetics processes, it is expected that certain alleles will be more widespread in some populations than others, such as the well-documented prevalence of a 9.3 allele at TH01 in European populations,[40] or of a 2.2 allele at Penta D which has a frequency of >11% in the West African population.[41] There has been limited research on the presence of sequence-based population-specific alleles, but it was assumed that this would also hold true for sequence-based alleles.

Gettings et al.[41] reported multiple examples of apparent population-specific enriched frequencies—where alleles with a specific motif had frequencies that were over 20% higher in one population compared to the others studied (such as the TCTA TCTG $[TCTA]_n$ motif at D3S1358 in the African American population). This phenomenon was observed during this work, where it also became apparent that certain alleles were seen only in one or two more closely related populations. Figure 1 shows the distribution of alleles at each locus according to how many alleles are seen in one, two, three, four, or even all five populations.

As expected, certain very common alleles are seen in all populations, such as allele 11 at TPOX which has a frequency of over 0.2 in the five population groups studied. At certain markers, there is a surprisingly high proportion of ''population-specific'' alleles, such as at D19S433 where over half the alleles observed are only seen in one population.

For certain rare alleles, the fact that they are only seen once means it is hard to draw any meaningful conclusions regarding their population specificity. At CSF1PO for example, all variation observed where the sequence diverged from the traditional $[AGAT]_n$ motif was population specific, and only seen once or twice across the entire data set, suggesting these could simply be one-off mutational events. The sequences were compared with alleles genotyped in the Caucasian, African American, and East Asian population samples published by NIST[41] and the University of North Texas (UNT).[42]

One CSF1PO allele, which was seen only once in a West African sample (allele $[AGAT]_8$ ACCT $[AGAT]_3$), was seen twice in the NIST African American population, and not in any other population. This suggests this sequence variant could be specific to the West African population, and demonstrates the utility of larger scale databases to properly capture all expected variation.[43]

Figure 2 illustrates the frequencies for all population-specific alleles observed across the five populations. The thicker the band, the more common the allele in the population it was recorded in. Certain alleles appear pointedly more common than others, despite the fact that they are only seen in one population. The thick band going from D19S433 in the South Asian population corresponds to a specific allele 13 sequence variant, which has a frequency of 0.016 in this population (i.e., 6/378 alleles).

Other notable common population-specific alleles include one sequence-based allele at D5S818, observed at

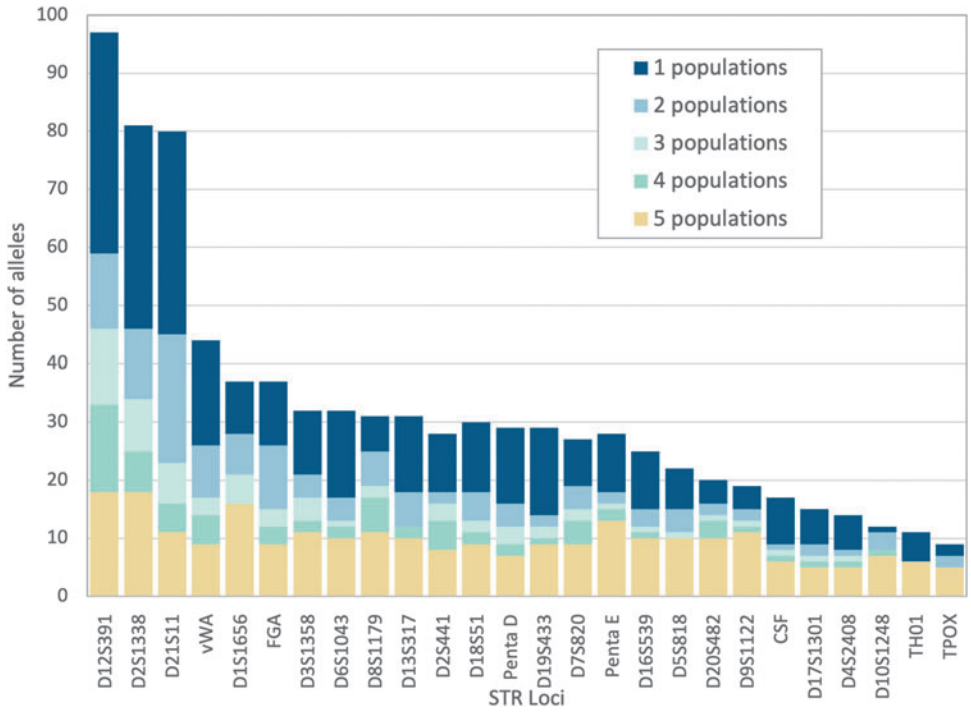

FIG. 1. Number of discernible alleles at each locus, split by number of populations in which they have been recorded. The number of discernible alleles at each aSTR locus studied is split into five categories, from those observed in all five populations (yellow) to those seen in just one population (dark blue). This graph does not account for the different characterisation of alleles (length based or sequence based), or their frequency within a population. aSTR, autosomal short tandem repeat.

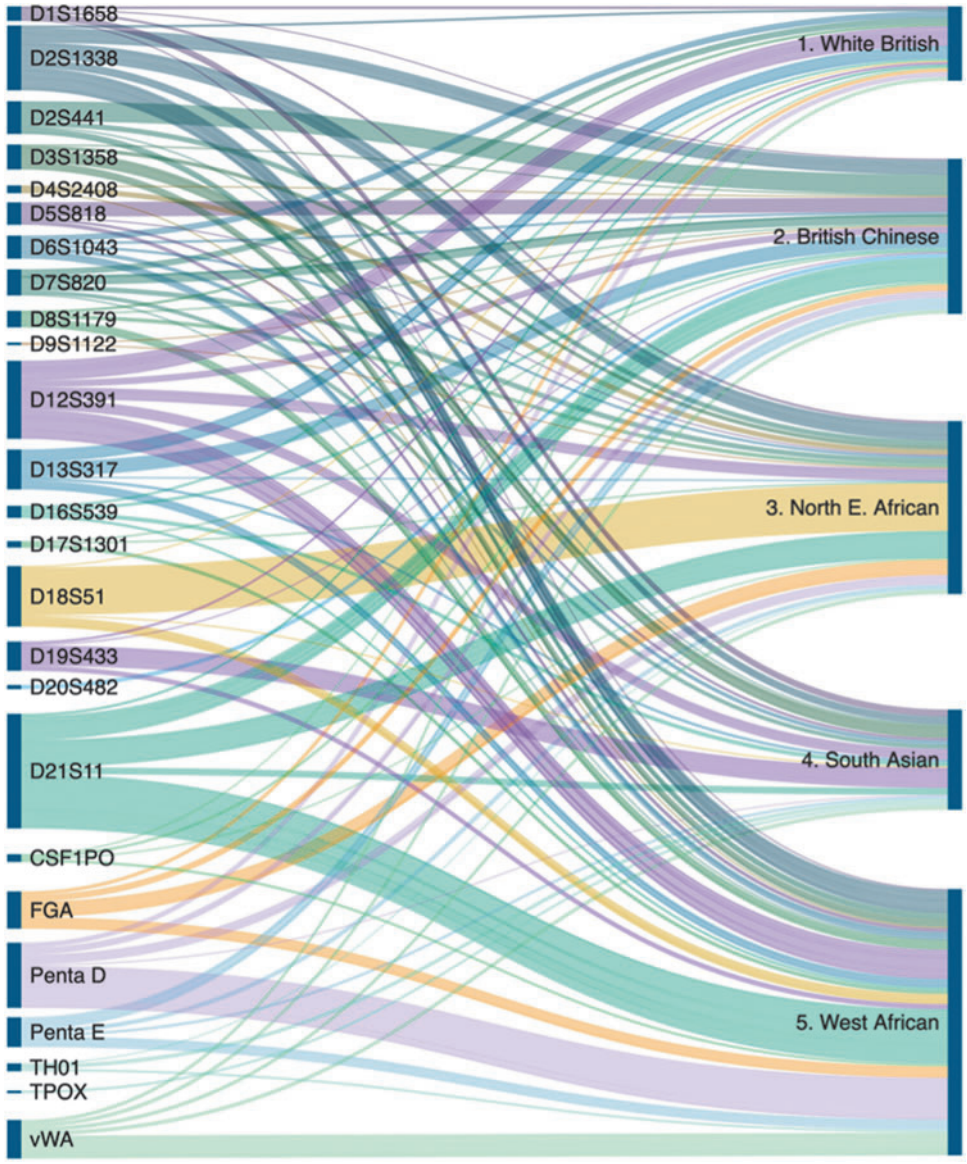

FIG. 2. Frequency of all observed population-specific alleles. The frequency for each population-specific allele (alleles seen in only one population) for the 26 aSTRs is represented by a single line. The thicker the line, the higher the frequency of the allele within the population. On the right-hand side of the graph, the corresponding population in which each allele is observed is provided.

a frequency of 0.029 in the British Chinese population (11/396) and a sequence-based allele at D21S11, observed at a frequency of 0.037 in the West African population ($n = 16/404$). The thick bands going from D18S51 to the two African populations correspond to the length-based .2 alleles at this marker.

Using autosomal STRs for population differentiation

STRUCTURE. To investigate whether autosomal STR data could be used to differentiate the 5 global populations studied in this work, genotypes for 26 autosomal markers (27 STRs in the ForenSeq DNA Signature Prep kit minus D22S1045) were run through STRUCTURE. This program was used to discern genetic clusters based on individuals' similarity or dissimilarity to others within the sample set. An initial aim was to check whether sequence-based alleles, particularly those including flanking region sequences, might be more useful for ancestry inference than length-based allelic data alone.

Moriot et al. stipulate that haplotypes composed of slow and fast-evolving loci might combine the advantages of identity and ancestry-informative marker types.[10] The instability of STR markers has led to a

large divergence in number of alleles in a population over time, whereas flanking region SNPs or insertions/deletions should be more stable through the course of evolution, possibly allowing for greater conservation within populations.

Analyses were run for all samples ($n = 989$) using data for length-based alleles, sequence-based without flanking regions alleles (Supplementary Figs. S1 and S2), and sequence-based with flanking regions alleles (Fig. 3). In these figures, each vertical line represents one sample, and the colour composition of that line reflects the proportion of membership for each calculated genetic cluster. Colour assignments correspond to the population group with the largest membership in that cluster.

Samples are grouped together by self-declared ancestry in the diagrams for simplicity, with the five populations separated by black lines. The K value for each STRUCTURE analysis refers to the number and patterns of genetic clusters found, and is user defined. STRUCTURE plots were run using $K = 2$ to $K = 5$ to see how

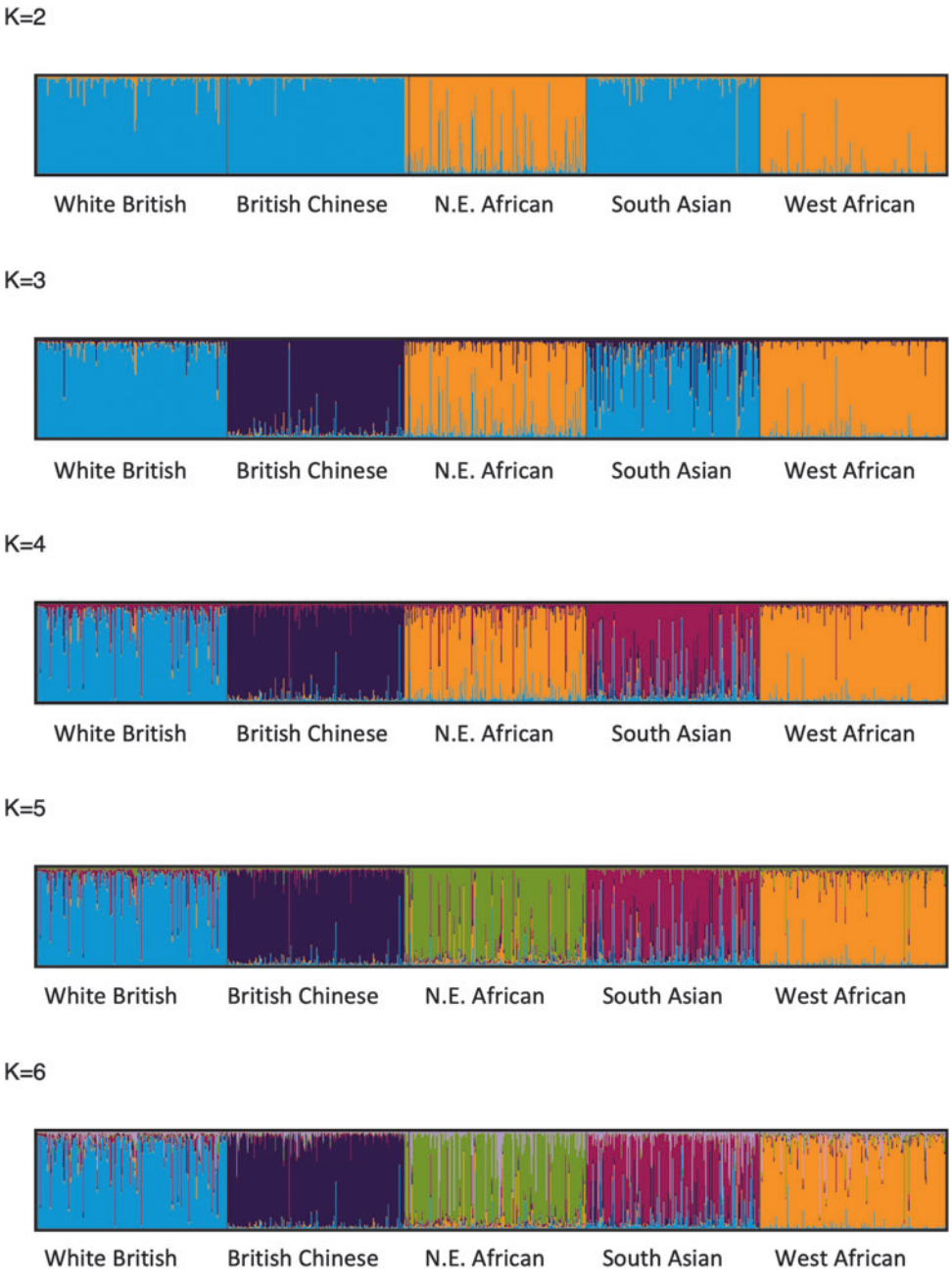

FIG. 3. STRUCTURE plots for the 5 populations, generated using sequence-based allelic data that include flanking region sequences for 26 aSTRs in the ForenSeq DNA Signature Prep Kit.

many groups (/populations) the program could separate by genetic stratification of the data. $K=6$ was also run, to ensure no additional substructure was being picked up, which would possibly indicate subpopulations. Although STRUCTURE is primarily a way of clustering individual samples rather than classifying them, looking at the number of samples that cluster incorrectly may provide a measure of the STR set for ancestry inference.

When applying $K=2$, it is clear that the software is able to distinctly separate the African populations from the others, highlighting the genetic dissimilarity between African and non-African ancestral populations. STRUCTURE defines genetic clusters without prior knowledge of population affiliation, hence the fact that the African populations have been successfully separated confirms that the STRs provide good African: non-African differentiation. The next cluster to be distinguished with $K=3$ is the British Chinese.

With $K=4$, the trend differs between the length-based and sequence-based plots, with the former separating a cluster for the Northeast African group whilst the two sequence-based plots differentiate the South Asian group first. Whether using length-based or full sequence-based allelic data (including flanking regions), STRUCTURE appears to be able to distinguish five (i.e., $K=5$) distinct genetic clusters, which correspond to the five ancestral population groups. In order to confirm which value of K best represented the data in a more objective way, results from STRUCTURE were uploaded to STRUCTURE Harvester.[44]

One of the plots produced by this program provides an indication of how likely each K value tested is. The results from STRUCTURE Harvester (data not shown) confirm what could be seen in the STRUCTURE plots for $K=2$ to $K=6$, which is that a partition of the data into five genetic clusters is the most likely scenario. From here on, STRUCTURE data focus on plots generated using a $K=5$ assumption.

Proportion of membership. For all STRUCTURE analyses run with $K=5$, the average proportion of membership of each predefined population in each of the five clusters was extracted from the results file and collated in Figure 4. This provides the average proportion of membership coefficient for the samples in each population group assigned to each of the five clustered inferred with $K=5$. A proportion of membership of 1 would indicate 100% assignment to one cluster. For example, the samples in the White British group have an average proportion of membership of 0.73 to cluster 1 in the STRUCTURE plot run with length-based allelic data, whereas this goes up to 0.83 when taking full sequence-based allelic data into account.

Overall, the strongest average proportion of membership for each group (highest value assigned to one cluster) for all groups occurs when including flanking

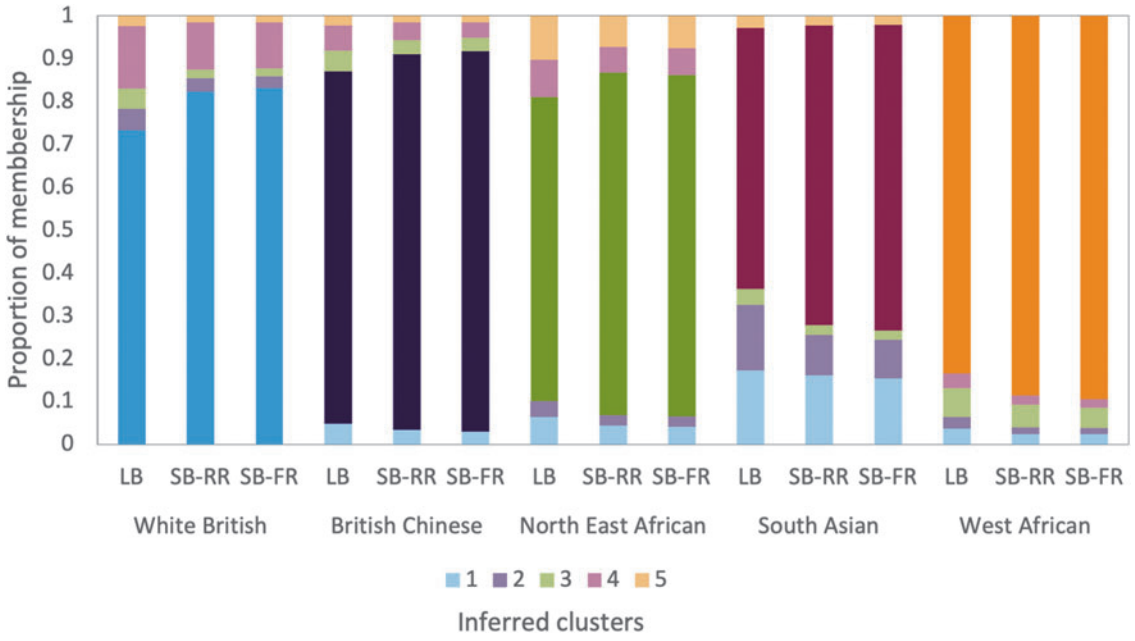

FIG. 4. Proportion of membership of each cluster for each of the five pre-defined populations. The main inferred cluster for each pre-defined population group is colour coded, with all other cluster assignments shown in a more transparent colour. LB, length-based alleles; SB-RR, repeat region sequence-based alleles; SB-FR, sequence-based alleles including flanking regions.

region variation, apart from the Northeast African group where membership of population essentially remains the same when going from repeat region alleles to flanking region data.

Although promising in terms of population differentiation, certain samples in each population group are still being assigned incorrect cluster membership. Phillips et al. measured the relative ability of a forensic STR set to differentiate ancestries by measuring group misclassification, defined as the number of samples with less than 0.5 group membership proportion for their true population of origin.[25] In the White British population, 18 samples out of 207 did not achieve a proportion of membership higher than 0.5 for the correct cluster (sequence data including flanking regions). A similar phenomenon was seen in the British Chinese (4/193), Northeast African (27/198), South Asian (37/189), and West African (8/202) population groups.

These numbers, and the more general assignment results presented in Figure 4, also highlight the fact that certain populations are considerably easier to genetically separate from the others, namely the West African and British Chinese. Although some of these samples do not show a proportion of membership coefficient higher than 0.5 for any of the groups, 67 samples in total (6.8%) appear to show assignment to the wrong cluster. These values effectively correspond to the samples represented by the ''wrong'' colour in the STRUCTURE plots.

This is an important improvement over the fact that with the length-based data, 185 samples do not have a proportion of membership higher than 0.5 for the correct cluster, with 108 of these having a coefficient of over 0.5 for the wrong cluster. An improvement in error rate and nonclassification rate is also observed when comparing these results to those obtained in publications focusing on 20 and 32 length-based STRs for ancestry estimation[25,45]

A more realistic indication of the rate of incorrect assignment would be to use a more conservative proportion of membership coefficient as the cutoff. If a proportion of membership to the correct cluster of over 0.7 is taken as a ''correct'' population assignment, with anything below classifying as ''inconclusive,'' 84% of samples (832 out of 989) are assigned to the correct group. This is a noteworthy improvement to the length-based data and a very slight improvement over the repeat-region sequence-based data, where 72% (719 out of 989) and 83% (818 out of 989) of samples are assigned correctly, respectively.

The number of samples that are assigned to the incorrect group with over 0.7 proportion of membership is inversely related, with 56 samples being assigned the wrong group for length-based data, 38 samples with the repeat-region sequence-based data, and 35 samples being assigned incorrectly for the full flanking region sequence-based data. This is equivalent to a general error rate of 3.5% for this STR set's ability to assign correct group membership.

With 84% of samples assigned to the correct cluster according to the five predetermined population groups, and over 90% of samples assigned for the correct cluster with a proportion of 0.5, it is fair to state that there is ancestry-informative data within the autosomal STR results for this data set.

Ancestry informativeness, I_n. It has been stipulated that the use of highly informative markers can reduce the amount of genotyping required for ancestry inference, as using markers with the highest level of informativeness can reduce the number of overall markers needed.[39] The number of markers analysed in this data set forms a commercial panel, and are, therefore, always amplified together, so it is not a case of needing to reduce the number of markers. However, the use of highly uninformative markers (expected given the loci type, autosomal STRs) could be adding noise to the STRUCTURE plots presented above.

To assess the ancestry informativeness of the autosomal STR loci studied, and, therefore, identify which contribute more meaningfully to the population clustering in the STRUCTURE plots, the Informativeness for Assignment value, I_n, was calculated for each locus. The I_n metric is highly correlated to the fixation index (F_{ST}), which is often used to measure population differentiation of SNPs,[6] but is better suited for multilocus data according to Rosenberg et al.[39] Table 1 shows the I_n values for all loci when comparing the five population groups at once.

This table shows that the ancestry informativeness of all markers increases when taking sequence variation into account. Xu et al. state that I_n values of over 0.2 can be considered as ''the signal of very great genetic difference between populations,''[46] and the results presented here show that no length-based locus data gave an I_n of above 0.2.

Penta D is the marker with most ancestry informativeness when only accounting for length-based data, which is unsurprising given the known prevalence of specific length variants such as the 2.2 and 3.2 alleles in the African populations. Interestingly, these do not reflect true changes to the allelic repeat structure and are in fact caused by a 13 base pair deletion in the flanking region. This makes these alleles particular in the sense that they combine a slowly mutating one (the deletion) and a faster mutating one (the STR), which had been suggested to be beneficial in the context of population-specific allelic enrichment.[10]

When adding sequence-based data, Penta D is far outstripped in terms of ancestry informativeness by other markers. The marker showing the most pronounced increase in I_n is D21S11, which is also one of the markers

Table 1. I_n values for all markers

Five pops LB		*Five pops RR*		*Five pops FR*	
Locus	I_n	*Locus*	I_n	*Locus*	I_n
All 26 markers					
Six most informative					
Penta D	0.168	D21S11	0.320	D21S11	0.320
Penta E	0.163	D2S1338	0.300	D2S1338	0.300
D1S1656	0.147	D12S391	0.248	D12S391	0.261
D18S51	0.145	D1S1656	0.220	D1S1656	0.220
TH01	0.117	D13S317	0.203	D13S317	0.216
D2S441	0.113	vWA	0.199	vWA	0.199
Twelve most informative					
D13S317	0.110	D3S1358	0.184	Penta D	0.195
D2S1338	0.103	Penta D	0.173	D3S1358	0.185
D19S433	0.100	Penta E	0.173	D2S441	0.177
D6S1043	0.092	D8S1179	0.161	Penta E	0.173
FGA	0.091	D2S441	0.156	D8S1179	0.161
D12S391	0.090	D18S51	0.147	D18S51	0.149
Twenty-four most informative					
D21S11	0.087	D4S2408	0.145	D4S2408	0.145
D4S2408	0.080	D5S818	0.125	D5S818	0.125
TPOX	0.079	TH01	0.117	TH01	0.118
D5S818	0.078	D19S433	0.111	D7S820	0.116
vWA	0.074	D6S1043	0.109	D6S1043	0.112
D8S1179	0.074	FGA	0.107	D19S433	0.112
D10S1248	0.061	D9S1122	0.088	FGA	0.107
CSF1PO	0.061	TPOX	0.079	D9S1122	0.088
D7S820	0.051	CSF1PO	0.065	D16S539	0.084
D3S1358	0.034	D10S1248	0.061	TPOX	0.079
D16S539	0.030	D7S820	0.058	CSF1PO	0.065
D20S482	0.030	D17S1301	0.034	D10S1248	0.061
D17S1301	0.029	D20S482	0.032	D20S482	0.051
D9S1122	0.028	D16S539	0.031	D17S1301	0.034

I_n values are given for 26 loci, across the 5 population groups studied, using LB, SB-RR, and SB-FR data.

LB, length based; SB-FR, sequence-based flanking region; SB-RR, sequence-based repeat region.

showing the highest increase in heterozygosity when accounting for sequence variation.[30,47] There is little difference when adding the flanking regions in, although the markers showing the most increase in I_n do correlate with those showing the most variation in the flanking regions: D7S820, D16S539, D20S482, and Penta D. The improvement in I_n at D7S820 due to the addition of flanking region data even pushes it out of the six least informative markers in the set.

The three most informative markers for differentiating the five population groups using full sequence-based allelic data are D21S11, D2S338, and D12S391, which corresponds to the three markers showing the highest frequency of population-specific alleles discussed earlier.

STRUCTURE plots were run using data for the 6, 12, and 24 most informative STR markers (highest I_n), as well as for the 6, 12, and 24 least informative STR markers (lowest I_n), but there appeared to be little difference between the 24 most and 24 least informative markers (data not shown). The average proportion of membership, generated by STRUCTURE for each population group with varying number of markers, confirmed that there is little difference between using data for the 22 or 24 most informative markers and all 26 markers. The total number of incorrectly classified samples also showed very limited improvement by reducing marker numbers (Supplementary Fig. S3).

The rationale for reducing the number of loci used was that some noise may be removed from the STRUCTURE plots by removing markers that are not useful in terms of ancestry inference but could be adding in unnecessary variation unrelated to population-specific enrichment. Whilst an interesting concept, reducing the number of loci according to ancestry informativeness does not seem to greatly improve population inference in this data set. Calculating I_n values helped confirm how much more useful sequence-level data are in terms of using the autosomal STR data for ancestry estimation, but the extent of what can be achieved using these STR data alone appears to reach its limit when using all data available: complete sequence-level allelic data for 26 autosomal STRs.

Using ancestry-informative SNPs for population differentiation

DPMB SNPs. The ForenSeq DNA Signature Prep kit contains a primer mix (DPMB) that consists of the same primers as those in DNA Primer Mix A, with the addition of primers for the amplification of 22 phenotype-informative SNPs, and 56 biogeographical ancestry-informative SNPs. Given the coamplification of these SNPs with the previously discussed STRs when using DPMB, and the promising results from the sequence-based STR alleles, the combined value of these markers for ancestry estimation comes into question.

Results obtained from the STRs were compared to those of this specifically chosen ancestry-informative SNP set on the same samples, before ascertaining if the combination of both marker types provides a better population differentiation than either alone.

The 56 SNPs used for ancestry inference in DPMB were selected from two publications that focused on identifying a small panel of highly informative SNP markers for ancestry estimation.[48,49] In 2014, Kidd et al. stated that a very large number of ancestry-informative markers could provide accurate discrimination of six to seven geographic regions, but a small, efficient, and robust panel is more relevant for forensic applications. They went on to identify a small panel of SNPs which would be useful for global population differentiation.[9]

The resulting panel of 55 SNPs is well characterised, has been broadly applied as a standalone panel, and is commonly referred to as the ''Kidd SNPs.'' The 56 SNPs amplified by the ForenSeq DNA Signature Prep DMPB correspond to these 55 Kidd SNPs, as well as SNP rs1919550 which appears to have limited global variability but is useful at distinguishing Native American individuals from other populations.[50] Presumably,

the decision was made to incorporate this SNP to enhance the primer set's capability to separate American populations, although the fact that it is in full linkage disequilibrium with another SNP (rs12498138) in DPMB makes it redundant (Personal communication; C. Phillips).

Universal Analysis Software estimation. In the UAS, the 56 aSNPs are analysed using principal component analysis (PCA). The model in UAS was trained on the European, East Asian, and African (except for the ASW, ''African Ancestry in Southwest US'' group) super populations of the 1000 Genomes Phase I data.[51] The sample being tested is then projected based on its aSNP genotype calls onto the pretrained components of the PCA plot, alongside data for the Admixed American super population for context.

Although the estimation feature of UAS can be useful for a sample whose biogeographical ancestry aligns with one of the three reference super populations, it is not effective for predicting the ancestry of an unrepresented group. Supplementary Figure S4 shows an example of PCA plot generated by UAS for a sample from each population group studied.

This shortcoming of the UAS software for biogeographical ancestry estimation has already been highlighted by several publications, including Ramani et al. who genotyped 1030 unrelated individuals living in Singapore of Chinese, Malay, and Indian origin.[52] Whilst UAS was able to accurately place the Chinese samples with the East Asian cluster on the PCA plot, the Malay and Indian samples clustered in between reference populations. Hussing et al. found that 22 out of 23 European samples projected close to the correct cluster, whereas the handful of Middle Eastern and North African samples they sequenced did not return a useable prediction.[19]

Wendt et al. noted that their Yavapei Native American population samples clustered either with the East Asian cluster or in between reference populations.[53] Despite the presence of a SNP specifically chosen for differentiating American populations, the lack of reference data for Native American populations once again hinders any interpretation.

Although the UAS has limited use for ancestry determination due to lacking reference data for certain populations such as South Asian, Northeast African, and Middle Eastern, it is expected that genotype results for the SNPs targeted in DPMB would still be useful for distinguishing these populations using other software.

STRUCTURE analysis. As with the autosomal STR data, aSNP genotypes were used to generate STRUCTURE plots. Figure 5 shows that the software was able to distinguish five distinct clusters under $K=5$, corresponding to the five population groups.

Proportion of membership and incorrectly assigned samples. If a proportion of membership to the correct cluster of over 0.7 is taken as a correct population assignment, 94% of samples (205/219) are assigned to the correct group, and 98% of samples (214/219) if using a value of 0.5 or higher. Given that the samples reanalysed for the autosomal SNPs were chosen at random from the 989 samples analysed for the autosomal STRs, aSNP data were not available for all the samples that were incorrectly assigned with aSTRs. Of the 67 samples that had a proportion of membership of over 0.5 for the wrong cluster using aSTR data, 17 happened to be reanalysed using the primers for the aSNPs in DPMB.

Of these, only four have a proportion of membership of over 0.5 for the wrong cluster using aSNP results, and only one of those had a coefficient of over 0.7. Table 2 shows the proportion of membership for these four samples, and shows the only sample that did not have a proportion of membership of over 0.5 for any of the groups.

The four samples that had a proportion of membership of over 0.5 for the wrong cluster were all Northeast African. Samples NEA404 and NEA465 have a proportion of membership of 0.66 and 0.505, respectively, for the West African cluster, despite self-declared Northeast African ancestry. In the aSTR STRUCTURE analysis, NEA404 and NEA465 also clustered with the West African population cluster, with a proportion of membership of 0.9 and 0.8 assignment, respectively, for this population.

The reason for these samples misclassifying is uncertain, although both are known to have identity-informative SNP alleles only seen in the West African population.[54] This could indicate incorrect self-declared ancestry, or even recent admixture.

Samples NEA438 and NEA439 both showed a proportion of membership of just over 0.5 for the South Asian cluster using aSTR data, but this increases to 0.85 for NEA439 when using aSNP genotypes, causing it to be misclassified. Although very little is known about the donors of the samples, they do appear to have come from the same region of Somalia according to their self-declared ancestry. These results may indicate that STRUCTURE is picking up on population substructure, or a population that is more closely related genetically to the samples in the South Asian cluster than those in the Northeast African cluster.

The four samples listed in Table 2, which returned a proportion of assignment of over 0.5 for the incorrect population, were also run through FROG-kb.[55] This database contains data from populations not included in the UAS reference set, including from the Middle East, South and Central Asian, and Oceania. Sample NEA438 returns probabilities of genotypes for multiple populations that are all within an order of magnitude of the highest probability of genotype, which is for the Makrani

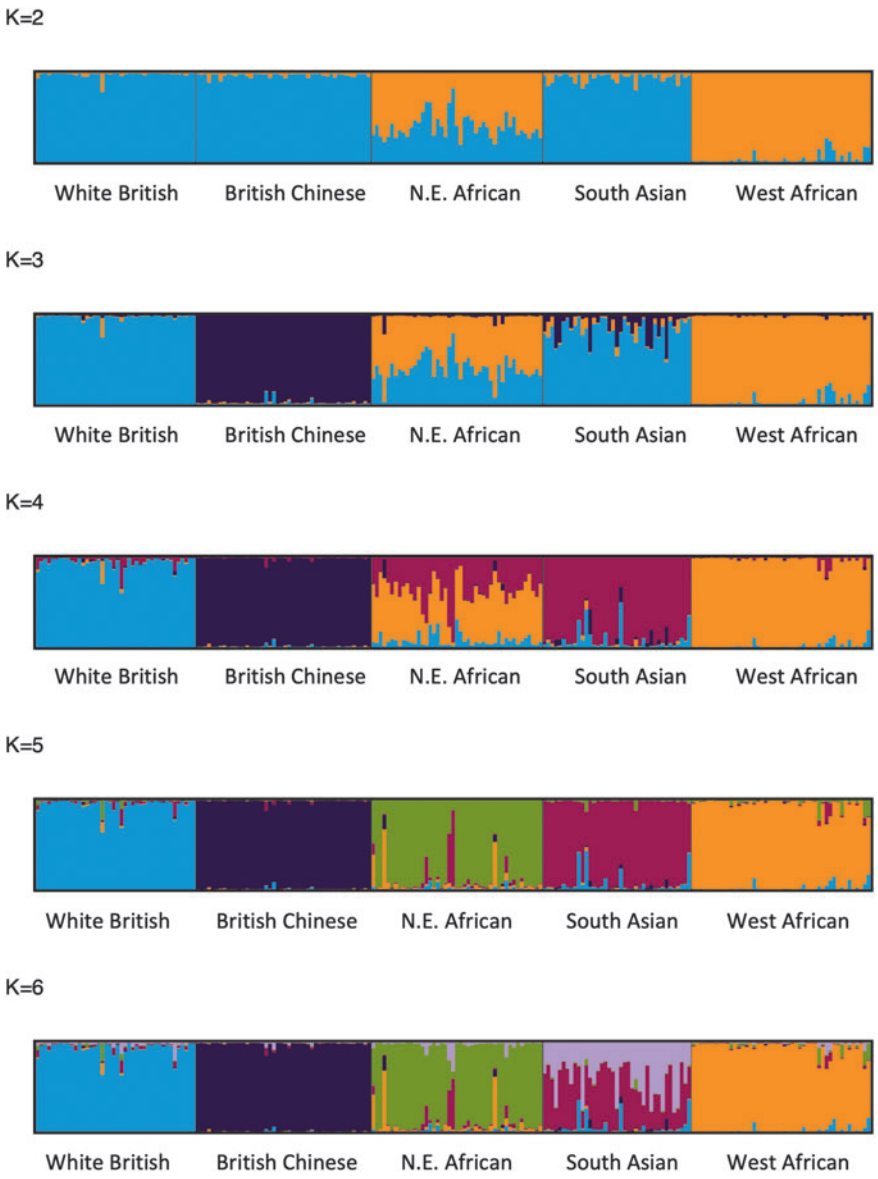

FIG. 5. STRUCTURE plots for the 5 populations studied, generated using genotypes for the 56 aSNP markers targeted by DNA Primer Mix B in the ForenSeq DNA Signature Prep Kit.

Table 2. Samples with a proportion of membership of >0.5 for the incorrect cluster using aSNP data

Sample	*Population*	*WB*	*SA*	*BC*	*WA*	*NEA*
NEA404	NEA	0.007	0.012	0.161	0.666	0.155
NEA438	NEA	0.025	0.55	0.004	0.046	0.375
NEA439	NEA	0.008	0.855	0.002	0.023	0.111
NEA465	NEA	0.03	0.014	0.075	0.505	0.376
SA654	SA	0.432	0.384	0.013	0.045	0.126

For each sample, the value in black corresponds to the proportion of membership for the correct cluster, and the value in red corresponds to the highest proportion of membership. All other values are shown in grey.

BC, British Chinese; NEA, Northeast African; SA, South Asian; WA, West African; WB, White British.

population, an ethnic group of Pakistan and Indian with African heritage (Supplementary Fig. S5).

Sample NEA439 returned the highest probability of genotype for the Saudi and Quatari populations. As stated above, these individuals are of Somali self-declared ancestry, but analysis may be picking up on population substructure, or highlighting a small population with genetic similarities to Middle Eastern or South Asian groups, for which reference data are perhaps not yet available.

Results obtained by targeting the 56 aSNPs in DPMB confirm that this panel can reliably be used for estimating ancestry. The lack of reference populations in UAS makes its utility limited, but genotypes can easily be extracted, and with the appropriate training data, good ancestry inference was obtained. These results show how samples can be separated into five distinct clusters

with STRUCTURE, due to the genetic similarities within a population, and the genetic differences between separate global populations. Individual sample results can also be extracted from UAS for upload to third-party tools such as FROG-kb or *Snipper*,[56] which have their own reference populations for ancestry estimation, enabling the probability of any given profile within each population to be calculated.

Combining ancestry-informative SNP and aSTR data

In their 2011 study, Phillips et al. concluded that the highest classification success could be obtained by combining the genotypes from forensic STRs with ancestry-informative SNPs.[25] Given that the ForenSeq DNA Signature Prep kit DPMB is used to amplify autosomal STRs and aSNPs simultaneously, combining data for both marker types for population differentiation was explored. As with the aSTR and aSNP data, STRUCTURE plots were run for $K=2$ to $K=6$ for the combined genotypes for the 219 samples analysed using both marker types. These plots are shown in Supplementary Figure S6.

Average proportion of membership was extracted from the STRUCTURE analyses ($K=5$) run with autosomal STR data alone, ancestry-informative SNP data alone, and combined aSTR and aSNP data. Figure 6 shows the proportion of membership to the different clusters for each population. Overall, the aSNP and combined runs appear to provide the highest average proportion of membership to the correct population for all groups. This improvement compared to the aSTR run is more visible for the White British and South Asian populations. The difference between the combined run and the aSNP data alone run appears negligible, except for the Northeast African population where the combined run is more similar to the aSTR run.

The proportion of assignment for the individual samples that were analysed with $K=5$ for just autosomal STR data, and for the combined STR and ancestry-informative SNP data is shown in Supplementary Figure S7. All British Chinese samples had been correctly assigned using STR data alone, so although the addition of SNP genotypes did improve the average assignment coefficient for the correct cluster, the difference on an individual sample level is negligible. A similar observation can be made for the West African population, where most samples classified correctly using aSTR data alone.

Figure 6 and Supplementary Figure S7 highlight that the benefit of combining the two sets of markers is most apparent in the White British and South Asian populations. These two populations were the hardest to separate when looking at aSTR data alone, and the addition of 56 ancestry-informative SNPs here clearly helps to push the individual proportion of membership to the correct cluster. In the Northeast African population data, the two samples that consistently appear to be assigned a higher proportion of membership for the South Asian cluster are again seen.

The $k=6$ run shows that STRUCTURE does appear to be picking up a ''new'' group for those two samples amongst the other Northeast African samples (shown in Fig. 7), supporting the theory that these samples may be more genetically related to a different under-represented population in this study. These two

FIG. 6. Proportion of membership obtained from STRUCTURE for runs generated using aSTR, aSNP, and combined aSTR and aSNP genotypes.

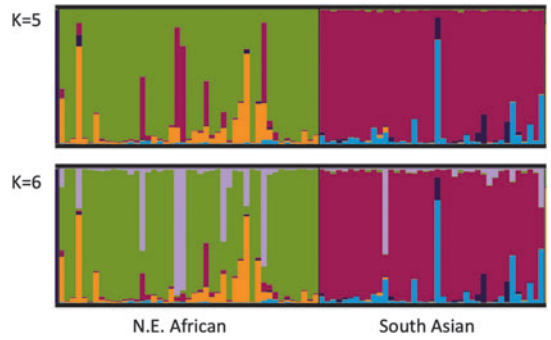

FIG. 7. STRUCTURE results for $K=5$ and $K=6$ for the Northeast African and South Asian samples with combined aSNP and aSTR data. Zoomed in view from the STRUCTURE plots in Supplementary Figure S6.

individuals may self-report as of Somali origin but in fact have recent Middle Eastern ancestry for example. Future work could look at adding Middle Eastern sample data in the STRUCTURE analyses to see whether these particular samples cluster with a higher proportion of membership to this population group.

Conclusion

As routine DNA testing of autosomal STRs progresses to MPS, the results from this work show that there is now the very real possibility of getting both an individual's DNA profile and an estimation of their biogeographic origin from one test. Combining aSNP and aSTR data did not show any improvement on using a dedicated aSNP panel alone, but the ancestry inference potential of the STRs in the ForenSeq DNA Signature Prep kit is almost as good as an aSNP panel. It is likely that this ancestry prediction may also improve with wider population data sets, as intimated by the two Northeast African samples that showed membership to a sixth cluster, suggesting subpopulations within the data.

The results presented from STRUCTURE analyses and group membership proportions imply that data from the aSTRs present in the ForenSeq DNA Signature Prep kit have the potential to be used for ancestry estimation for five global populations. With a strict group membership proportion of 0.7 or above, 84% of samples were grouped correctly, with a general error rate of 3.5%. This is a major improvement on a previous publication looking at length-based data for 20 autosomal markers, which had error averages of 12–15%.[25] It is also an advance on results obtained using length-based data for the 26 ForenSeq STRs, where 72% of samples were classified correctly with an error rate of 5.7%. These results once again highlight the value added by massively parallel sequencing.

Acknowledgments

The laboratory work performed at King's College London was undertaken as part of a PhD project, and the authors would like to thank Verogen and the Royal Commission for the Exhibition of 1851 for their support with this project. The authors would also like to thank Kathy Stevens and Cydne Holt for the custom SNP primer mix.

Authors' Contributions

Laboratory work was carried out by La.D. and Lu.D.; data analysis was done by La.D.; writing—original draft preparation was carried out by La.D.; writing—review and editing was carried out by D.B.; supervision was done by D.B. and D.S.C.

Author Disclosure Statement

No competing financial interests exist.

Funding Information

No funding was received for this article.

Supplementary Material

Supplementary Figure S1
Supplementary Figure S2
Supplementary Figure S3
Supplementary Figure S4
Supplementary Figure S5
Supplementary Figure S6
Supplementary Figure S7

References

1. Kidd JR, Friedlaender FR, Speed WC, et al. Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples. Investig Genet 2011;2(1):1.
2. Phillips C, Salas A, Sanchez JJ, et al. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. Forensic Sci Int Genet 2007;1(3–4):273–280.
3. Walsh S, Liu F, Wollstein A, et al. The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. Forensic Sci Int Genet 2013;7(1):98–115.
4. Walsh S, Liu F, Ballantyne KN, et al. IrisPlex: A sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. Forensic Sci Int Genet 2011;5(3):170–180.
5. Albright TD. Why eyewitnesses fail. Proc Natl Acad Sci U S A 2017;114(30):7758–7764.
6. Phillips C. Forensic genetic analysis of bio-geographical ancestry. Forensic Sci Int Genet 2015;18:49–65.
7. Bulbul O, Filoglu G. Development of a SNP panel for predicting biogeographical ancestry and phenotype using massively parallel sequencing. Electrophoresis 2018;39(21):2743–2751.
8. Bulbul O, Pakstis AJ, Soundararajan U, et al. Ancestry inference of 96 population samples using microhaplotypes. Int J Legal Med 2018;132(3):703–711.
9. Kidd KK, Pakstis AJ, Speed WC, et al. Current sequencing technology makes microhaplotypes a powerful new type of genetic marker for forensics. Forensic Sci Int Genet 2014;12:215–224.
10. Moriot A, Santos C, Freire-Aradas A, et al. Inferring biogeographic ancestry with compound markers of slow and fast evolving polymorphisms. Eur J Hum Genet 2018;26(11):1697–1707.
11. Phillips C, Fernandez-Formoso L, Gelabert-Besada M, et al. Development of a novel forensic STR multiplex for ancestry analysis and extended identity testing. Electrophoresis 2013;34(8):1151–1162.
12. Messina F, Di Corcia T, Ragazzo M, et al. Signs of continental ancestry in urban populations of Peru through autosomal STR loci and mitochondrial DNA typing. PLoS One 2018;13(7):e0200796.

13. Algee-Hewitt BF, Edge MD, Kim J, et al. Individual identifiability predicts population identifiability in forensic microsatellite markers. Curr Biol 2016;26(7):935–942.
14. Tian C, Kosoy R, Nassir R, et al. European population genetic substructure: Further definition of ancestry informative markers for distinguishing among diverse European ethnic groups. Mol Med 2009;15(11–12):371–383.
15. Kidd JR, Friedlaender F, Pakstis AJ, et al. Single nucleotide polymorphisms and haplotypes in Native American populations. Am J Phys Anthropol 2011;146(4):495–502.
16. Paschou P, Lewis J, Javed A, et al. Ancestry informative markers for fine-scale individual assignment to worldwide populations. J Med Genet 2010;47(12):835–847.
17. Nassir R, Kosoy R, Tian C, et al. An ancestry informative marker set for determining continental origin: Validation and extension using human genome diversity panels. BMC Genet 2009;10:39.
18. Kidd KK, Speed WC, Pakstis AJ, et al. Progress toward an efficient panel of SNPs for ancestry inference. Forensic Sci Int Genet 2014;10:23–32.
19. Hussing C, Borsting C, Mogensen HS, et al. Testing of the Illumina (R) ForenSeq (TM) kit. Forensic Sci Int Genet Suppl Ser 2015;5:E449–E450.
20. Phillips C, Freire Aradas A, Kriegel AK, et al. Eurasiaplex: A forensic SNP assay for differentiating European and South Asian ancestries. Forensic Sci Int Genet 2013;7(3):359–366.
21. Rosenberg NA, Pritchard JK, Weber JL, et al. Genetic structure of human populations. Science 2002;298(5602):2381–2385.
22. Zhao Z, Wen W, Michailidou K, et al. Association of genetic susceptibility variants for type 2 diabetes with breast cancer risk in women of European ancestry. Cancer Causes Control 2016;27(5):679–693.
23. Lowe AL, Urquhart A, Foreman LA, et al. Inferring ethnic origin by means of an STR profile. Forensic Sci Int 2001;119(1):17–22.
24. Londin ER, Keller MA, Maista C, et al. CoAIMs: A cost-effective panel of ancestry informative markers for determining continental origins. PLoS One 2010;5(10):e13443.
25. Phillips C, Fernandez-Formoso L, Garcia-Magarinos M, et al. Analysis of global variability in 15 established and 5 new European Standard Set (ESS) STRs using the CEPH Human Genome Diversity panel. Forensic Sci Int Genet 2011;5(3):155–169.
26. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics 2000;155(2):945–959.
27. West FL, Algee-Hewitt BFB. Cadaveric blood cards: Assessing DNA quality and quantity and the utility of STRs for the individual estimation of trihybrid ancestry and admixture proportions. Forensic Sci Int 2020;2:114–122.
28. Pereira L, Alshamali F, Andreassen R, et al. PopAffiliator: Online calculator for individual affiliation to a major population group based on 17 autosomal short tandem repeat genotype profile. Int J Legal Med 2011;125(5):629–636.
29. Devesse L, Ballard D, Davenport L, et al. Concordance of the ForenSeq™ system and characterisation of sequence-specific autosomal STR alleles across two major population groups. Forensic Sci Int Genet 2018;34:57–61.
30. Devesse L, Davenport L, Borsuk L, et al. Classification of STR allelic variation using massively parallel sequencing and assessment of flanking region power. Forensic Sci Int Genet 2020;48:102356.
31. Verogen, ForenSeq™ DNA Signature Prep Reference Guide. Document #VD2018005 Rev. A, 2018.
32. Churchill JD, Novroski NMM, King JL, et al. Population and performance analyses of four major populations with Illumina's FGx Forensic Genomics System. Forensic Sci Int Genet 2017;30:81–92.
33. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25(14):1754–1760.
34. Database of Single Nucleotide Polymorphisms (dbSNP). Available from: https://www.ncbi.nlm.nih.gov/snp/rs11642858 [Last accessed: October 9, 2019]
35. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25(16):2078–2079.
36. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20(9):1297–1303.
37. Peakall R, Smouse PE. GenAlEx 6.5: Genetic Analysis in Excel. Population genetic software for teaching and research—An update. Bioinformatics 2012;28(19):2537–2539.
38. Kopelman NM, Mayzel J, Jakobsson M, et al. Clumpak: A program for identifying clustering modes and packaging population structure inferences across K. Mol Ecol Resour 2015;15(5):1179–1191.
39. Rosenberg NA, Li LM, Ward R, et al. Informativeness of genetic markers for inference of ancestry. Am J Hum Genet 2003;73(6):1402–1422.
40. Butler JM, Butler JM. Fundamentals of Forensic DNA Typing. Academic Press/Elsevier: Amsterdam; Boston, MA, 2010; p. 1 online resource (xviii, 500 p.).
41. Gettings KB, Borsuk LA, Steffen CR, et al. Sequence-based U.S. population data for 27 autosomal STR loci. Forensic Sci Int Genet 2018;37:106–115.
42. Novroski NM, King JL, Churchill JD, et al. Characterization of genetic sequence variation of 58 STR loci in four major population groups. Forensic Sci Int Genet 2016;25:214–226.
43. Devesse LA, Ballard DJ, Davenport LB, et al. The tao of MPS: Common novel variants. Forensic Sci Int Genet Suppl Ser 2017;6:E579–E581.
44. Earl DA, vonHoldt BM. Structure Harvester: A website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genet Resour 2012;4:359–361.
45. Phillips C, Gelabert-Besada M, Fernandez-Formoso L, et al. "New turns from old STaRs": Enhancing the capabilities of forensic short tandem repeat analysis. Electrophoresis 2014;35(21–22):3173–3187.
46. Xu S, Huang W, Qian J, et al. Analysis of genomic admixture in Uyghur and its implication in mapping strategy. Am J Hum Genet 2008;82(4):883–894.
47. Phillips C, Devesse L, Ballard D, et al. Global patterns of STR sequence variation: Sequencing the CEPH Human Genome Diversity panel for 58 forensic STRs using the Illumina ForenSeq DNA Signature Prep Kit. Electrophoresis 2018;39(21):2708–2724.
48. Sampson JN, Kidd KK, Kidd JR, et al. Selecting SNPs to identify ancestry. Ann Hum Genet 2011;75(4):539–553.
49. Nievergelt CM, Maihofer AX, Shekhtman T, et al. Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel. Investig Genet 2013;4(1):13.
50. Yaeger R, Avila-Bront A, Abdul K, et al. Comparing genetic ancestry and self-described race in african americans born in the United States and in Africa. Cancer Epidemiol Biomarkers Prev 2008;17(6):1329–1338.
51. 1000 Genomes Project Consortium; Abecasis GR, Altshuler D, et al. A map of human genome variation from population-scale sequencing. Nature 2010;467(7319):1061–1073.
52. Ramani A, Wong Y, Tan SZ, et al. Ancestry prediction in Singapore population samples using the Illumina ForenSeq kit. Forensic Sci Int Genet 2017;31:171–179.
53. Wendt FR, Churchill JD, Novroski NMM, et al. Genetic analysis of the Yavapai Native Americans from West-Central Arizona using the Illumina MiSeq FGx forensic genomics system. Forensic Sci Int Genet 2016;24:18–23.
54. Davenport L, Devesse L, Syndercombe Court D, et al. Forensic identity SNPs: Characterisation of flanking region variation using massively parallel sequencing. Forensic Sci Int Genet 2023;64:102847.
55. Rajeevan H, Soundararajan U, Pakstis AJ, et al. Introducing the Forensic Research/Reference on Genetics knowledge base, FROG-kb. Investig Genet 2012;3(1):18.
56. Phillips C. Online resources for SNP analysis: A review and route map. Mol Biotechnol 2007;35(1):65–97.