Publisher : John Wiley & Sons, Inc.
Location : Hoboken, USA
DOI : 10.1002/(ISSN)2573-9468
ISSN (print) : 2573-9468
ISSN (electronic) : 2573-9468
ID (product) : WFS2
Title (main) : Wiley Interdisciplinary Reviews: Forensic Science
Title (short) : WIREs Forensic Sci

Numbering (journalVolume) : 9999
Numbering (journalIssue) : 9999
DOI : 10.1002/wfs2.1356
ID (unit) : WFS21356
ID (society) : FORSCI-067.R1
ID (eLocator) : e1356
Count (pageTotal) : 26
Title (articleCategory) : ADVANCED REVIEW
Title (tocHeading1) : ADVANCED REVIEWS

Event (manuscriptReceived) : 2019-04-25
Event (manuscriptRevised) : 2019-06-23
Event (manuscriptAccepted) : 2019-06-25
Event (xmlCreated) : 2019-07-03 (SPi Global)
Numbering (pageFirst) : n/a
Numbering (pageLast) : n/a
Link (toTypesetVersion) : file:wfs21356.pdf
Link (toAuthorManuscriptVersion) : file:wfs21356_am.pdf

Short Authors: McNevin

Forensic inference of biogeographical ancestry from genotype: The Genetic Ancestry Lab<<Query: Please check whether the edits made to the article title are appropriate. Ans: The title is fine>>

<<Query: Please confirm that given names (blue) and surnames/family names (vermilion) have been identified and spelled correctly. Ans: Yes, correct>>Dennis<<Query: Please check if link to ORCID is correct. Ans: Yes, correct>> McNevin*[1,2]

[1] Centre for Forensic Science, School of Mathematical & Physical Sciences, Faculty of Science, University of Technology Sydney, Ultimo, New South Wales, Australia

[2] Faculty of Science and Technology, University of Canberra, Bruce, Australian Capital Territory, Australia

Dennis McNevin: ✉ dennis.mcnevin@uts.edu.au

Correspondence to:

Correspondence
Dennis McNevin, Centre for Forensic Science, School of Mathematical & Physical Sciences, Faculty of Science, University of Technology Sydney, Broadway, Ultimo, New South Wales 2007, Australia.
Email: dennis.mcnevin@uts.edu.au

FundRef Name	FundRef Organization Name	Funding Number
AMP's Tomorrow Fund (https://www.ampstomorrowfund.com.au/)		2123
ANU Connect Ventures (http://www.anuconnectventures.com.au/)		DTF224

Abstract

Short tandem repeat (STR) profiling of DNA has become ubiquitous in forensic practice and is used to associate people, objects, and places with each other and with crimes. STRs can include or exclude a suspect or victim as the donor of biological evidence. In the absence of a matching profile, however, STRs have limited value. It is possible, then, to extract other information from the DNA that might lead forensic investigators to an offender. Examples include biogeographical ancestry (BGA) and externally visible characteristics (EVCs). These require alternative genetic markers including single nucleotide polymorphisms and microhaplotypes which can be genotyped on many different platforms including capillary electrophoresis, microarrays, and massively parallel sequencing (MPS). The Genetic Ancestry Lab (GAL) in Australia provides estimates of BGA and EVCs derived from DNA that is extracted from biological evidence and then subjected to targeted amplicon enrichment and subsequent MPS. This review will describe the process of BGA prediction employed by the GAL as well as describing alternative practices. Limitations are addressed and future directions highlighted, including resolution of genetic admixture. It is highly likely that inference of BGA will become standard forensic practice, performed simultaneously with or in addition to STR profiling, and it is hoped that this review might provide a road map.

This article is categorized under:

Forensic Anthropology > Ancestry Determination

Forensic Science in Action/Crime Scene Investigation > From Traces to Intelligence and Evidence

Forensic Biology > Ancestry Determination using DNA Methods

Forensic Biology > Forensic DNA Technologies

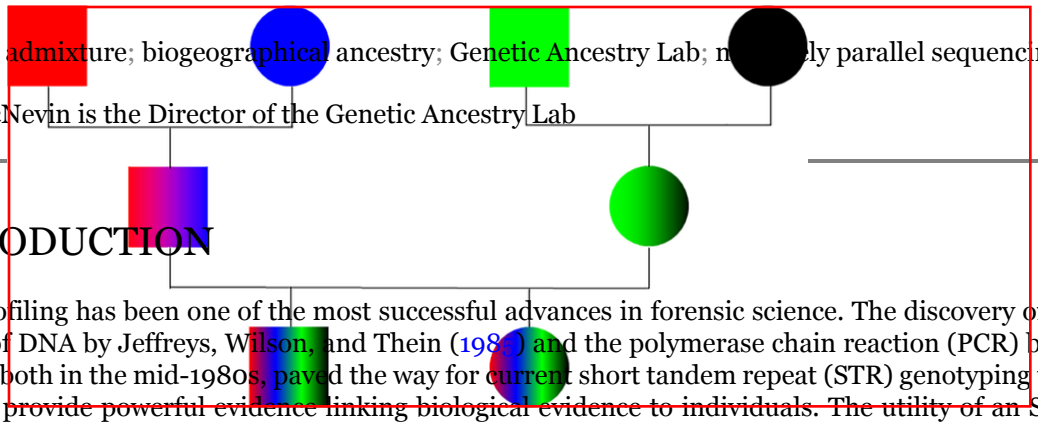
Graphical Abstract

Forensic inference of biogeographical ancestry (BGA) from genotype

This figure has been replaced by a file (image_n/Graphical abstract.tif) that is not supported to display in the browser. Thus the previous image is still being displayed.

Keywords: admixture; biogeographical ancestry; Genetic Ancestry Lab; massively parallel sequencing

Dennis McNevin is the Director of the Genetic Ancestry Lab



1 INTRODUCTION

DNA profiling has been one of the most successful advances in forensic science. The discovery of repetitive elements of DNA by Jeffreys, Wilson, and Thein (1985) and the polymerase chain reaction (PCR) by Mullis et al. (1986), both in the mid-1980s, paved the way for current short tandem repeat (STR) genotyping techniques which can provide powerful evidence linking biological evidence to individuals. The utility of an STR profile generated from biological evidence is diminished, however, if it does not match a suspect or a DNA database record. In this situation, we may be able to predict the biogeographical ancestry (BGA) of the DNA donor, as well other phenotypes including externally visible characteristics (EVCs) and biological age (Kayser & de Knijff, 2011). This information can be used to narrow a pool of suspects, saving valuable time and resources for forensic investigators (Phillips, 2015).

Moving beyond STR profiling, unveiling investigative information (intelligence) from DNA has required the use of different genetic markers and different genotyping technologies. While 20 STRs are sufficient for forensic identification, many more single nucleotide polymorphisms (SNPs), insertion/deletions (indels), and/or microhaplotypes may be required to predict BGA beyond the continental level (Kosoy et al., 2009; Phillips et al., 2014) and for the ultimate forensic goal of recreating the face of an offender through a molecular photofit (Claes et al., 2014, 2018; Kayser, 2015). Even a phenotype as simple as height is under the influence of at least several hundred SNPs, each with a very small effect (Marouli et al., 2017). Fortunately, genotyping at this scale is now possible as a result of new sequencing technologies (Alvarez-Cubero et al., 2017; Børsting & Morling, 2015) but forensic laboratories may need to invest in them in order to realize these new capabilities.

The path from genotype to phenotype can be divided into wet lab (chemistry) and dry lab (data). Forensic biology laboratories are less accustomed to the latter, which is also known as “bioinformatics” (Liu & Harbison, 2018). However, it should be remembered that STR profiling already involves significant data processing in that a fluorescent signal from a fluorescently labeled primer must be converted to an allele designation by way of deconvolution of spectral overlap and mathematical alignment with a size standard and allelic ladder (Shewale, Qi, & Calandro, 2013). This processing is often hidden from the forensic biologist but exists nevertheless. There should be no reluctance to accept bioinformatics as a component of ancestry prediction and it is likely that it will become incorporated into validated pipelines. There will, of course, be a requirement for training as well as significant increases in data handling and storage.

2 GENOTYPING TECHNOLOGY

The field of forensic phenotyping and ancestry prediction has been enabled by new genotyping technologies. However, it should be remembered that phenotyping has been possible ever since the availability of capillary electrophoresis (CE) and targeted SNP genotyping assays (Shewale et al., 2013). The first of these to be widely employed in a forensic context was the single base extension (SBE) assay which gained popularity because it could be employed using the equipment found in standard forensic biology laboratories. The first SBE assay was employed by DNAPrint Genomics, Inc. (Frudakis et al., 2003), in their “DNAWitness” product as early as 2002 when it was used to predict the BGA of a serial killer in Baton Rouge, Louisiana (Newsome, 2007).

2.1 SBE assays

As the name suggests, SBE involves the extension of an oligonucleotide primer by one nucleotide at the site to be genotyped after amplification of the target site by PCR (Sobrino, Brión, & Carracedo, 2005; Syvänen,

1999). The incorporated nucleotide is fluorescently labeled with each possible nucleotide (A, C, G, or T) having a different label. The most popular SBE assay is the SNaPshot™ Multiplex Kit (Thermo Fisher Scientific) which employs the following labeled nucleotides: ddATP-dR6G (green), ddCTP-dTAMRA™ (yellow), ddUTP-dROX™ (red), and ddGTP-dR110 (blue) (Applied Biosystems, 2010). The manufacturer recommends multiplexing up to 10 SNPs but it is possible to multiplex upward of 20 SNPs (de la Puente et al., 2016; Fondevila et al., 2013; Phillips et al., 2007, 2013; Phillips, Fondevila, & Lareau, 2012; Santos et al., 2016). Sensitivity (prior to the initial PCR) is generally less than 1 ng of template DNA. For a thorough review of forensically relevant SNaPshot™ assays for human DNA SNP analysis, see Mehta, Daniel, Phillips, and McNevin (2017).

2.2 Fragment length analysis of insertions/deletions (indels)

In the same way that CE can be used for fragment length analysis of STRs, it can be used for the same purpose to genotype insertions and deletions (indels). STR variants are simply insertions and deletions of repetitive DNA sequences. At least two indel panels with less than 50 loci have been developed to differentiate Africans, Europeans, Asians, and indigenous Americans (Pereira et al., 2012; Santos et al., 2010) and their applicability has been extended to other populations (Santos et al., 2015). These assays offer a more straightforward alternative to SBE assays that depend on complex, multistep protocols with many tube-to-tube transfers (and associated contamination risk). Both SBE and indel assays offer a simple and fast method of triaging samples before the more expensive options discussed later.

2.3 High-density genotyping (microarrays)

Until recently, microarrays have not featured heavily in the forensic landscape. This is chiefly because they lack the sensitivity expected of forensic genotyping assays currently (typically 0.5 ng DNA template input) which is one of the reasons that the Identitas v1 Forensic Chip was never used widely, even though it allowed parallel interrogation of 201,173 genome-wide autosomal, X-chromosomal, Y-chromosomal, and mitochondrial SNPs for inference of BGA, appearance, relatedness, and biological gender (Keating et al., 2013). Microarrays have undergone a renaissance with the advent of forensic genealogy, however, with the industry standard being the Infinium® BeadChip high-density arrays (Illumina) which can genotype 700,000 SNPs, insertion/deletions (indels) and copy number variants (CNVs) (Illumina, 2013). The manufacturer recommends 200 ng template DNA, thus restricting its use to large biological stains or pretreatment of the template with whole genome amplification. High-density genotypes can then be uploaded to third-party genealogy service providers like GEDmatch (https://www.gedmatch.com/) in order to find genetic relatives among other subscribers (Erlich, Shor, Pe'er, & Carmi, 2018; Henn et al., 2012; Phillips, 2018; Ram, Guerrini, & McGuire, 2018). There are a number of commercial providers that have entered the forensic genealogy market including Parabon® NanoLabs (https://snapshot.parabon-nanolabs.com/genealogy) (Armentrout, 2018), Family Tree DNA (https://www.familytreedna.com/) (Greenspan, 2019) and Bode Technology (https://bode-labs.com/pages/bode-forensic-genealogy-service) (Singer & Breakiron, 2019).

2.4 Massively parallel sequencing

So-called next generation sequencing (NGS) originally referred to the suite of DNA sequencing technologies which followed CE (Metzker, 2009; Pareek, Smoczynski, & Tretyn, 2011; Shendure & Ji, 2008; Zhang, Chiodini, Badr, & Zhang, 2011). They were alternatively referred to as massively parallel sequencing (MPS) and this is the term ~~we~~I shall use in order to distinguish them from newer, third generation sequencing technologies (Alvarez-Cubero et al., 2017; Berglund, Kiialainen, & Syvänen, 2011; Kircher & Kelso, 2010). The major difference between them is that MPS (now sometimes referred to as second generation sequencing) employs shorter read lengths, up to 400 base pairs (bp). Third generation sequencing, also referred to as single-molecule real-time (SMRT) sequencers or long read technologies, can sequence much larger tracts of DNA: greater than 100,000 bp on the MinION (Oxford Nanopore Technologies) (Jain et al., 2018).

The first MPS technology to emerge onto the market was also the first casualty. Pyrosequencing (Ahmadian, Ehn, & Hober, 2006; Margulies et al., 2005; Ronaghi, Karamohamed, Pettersson, Uhlén, & Nyrén, 1996), otherwise known as 454 sequencing, was developed by 454 Life Sciences and later acquired by Roche in 2007. It was the basis of the GS20, the first commercial next generation sequencer, but was discontinued in 2013 after it became noncompetitive. Polony sequencing (Mitra, Shendure, Olejnik, Edyta Krzymanska, & Church, 2003) formed the basis of the Sequencing by Oligonucleotide Ligation and Detection (SOLiD) system (<<Query: Please provide manufacturer location for Life Technologies, Thermo Fisher Scientific, Roche, Illumina, Verogen, Applied Biosystems Ans: Applied Biosystems and Life Technologies are subsidiaries of Thermo Fisher Scientific which has Headquarters in Waltham, MA 02451, USA.Roche (Hoffman-La Roche) has Headquarters in Basel, CH-4070, Switzerland.Illumina has Headquarters in San Diego, CA 92122, USA.Verogen has Headquarters in San Diego, CA 92121, USA.>>Life Technologies, later Thermo Fisher Scientific) but it was sequencing-by-synthesis, employing reversible terminator chemistry (Bentley et al., 2008), that quickly dominated the market. This was the basis of the Solexa sequencing technology acquired by Illumina. By 2008, there were three competing high-throughput next-generation sequencers: the GS FLX (Roche), the SOLiD system and the Genome Analyzer (Illumina).

In 2010, the Ion Torrent semi-conductor sequencing technology (Rothberg et al., 2011) was acquired by Life Technologies. This represented one of two developments that initiated the uptake of MPS by the forensic community. Ion Torrent technology facilitated production of bench scale sequencers, the first being the Ion Personal Genome Machine (PGM™; Thermo Fisher Scientific) (Churchill et al., 2015). Illumina quickly followed suit with the MiSeq. A forensic version of the MiSeq called the MiSeq FGx was also made available (Jäger et al., 2017), specifically for use with the ForenSeq™ DNA Signature Prep Kit (Verogen, a spin off from Illumina) which includes 27 autosomal STRs, 24 Y STRs, 7 X STRs, 94 identity SNPs, 22 phenotype SNPs and 56 BGA SNPs in the one assay (Churchill, Schmedes, King, & Budowle, 2016; Silvia, Shugarts, & Smith, 2017). The ForenSeq™ kit cannot be used on the standard MiSeq: it is confined to the FGx. Likewise, standard Illumina chemistry cannot be used on the FGx.

The other development (preceding bench top sequencing) to initiate forensic usage of MPS was oligonucleotide barcoding for library preparation (Binladen et al., 2007; Hoffmann et al., 2007; Parameswaran et al., 2007) by either ligation (Meyer, Stenzel, & Hofreiter, 2008; Meyer, Stenzel, Myles, Prüfer, & Hofreiter, 2007) or nested PCR (Guo & Milewicz, 2003). This enabled targeted amplicon sequencing which then allowed forensic application to STR and SNP genotyping (Børsting & Morling, 2015). Illumina have signed an agreement to utilize the Ion AmpliSeq (Thermo Fisher Scientific) targeted sequencing (ligation) chemistry which is increasingly being used for multiplex PCR-based target enrichment prior to MPS (Minotta & Endicott, 2018). The forensic reach of MPS has recently been extended to molecular autopsy, microbial forensics, and differentiation of monozygotic twins (Budowle, Schmedes, & Wendt, 2017).

The forensic MPS workflow to emerge can be summarized as follows, depending on whether the Ion Torrent (Applied Biosystems, 2017b) or MiSeq (Illumina, 2015) platforms are employed:

1. **Target enrichment**, involving the amplification of target loci by highly multiplexed PCR

2. **Library preparation,** involving oligonucleotide barcoding to allocate amplicons to sample of origin by ligation (Ion Torrent, MiSeq) or nested PCR (ForenSeq™ DNA Signature Prep Kit on the MiSeq FGx) and subsequent sample pooling

3. **Template preparation**, involving:

   (a) **Immobilization** of single-stranded DNA (ssDNA) to Ion Sphere™ Particles (ISPs: Ion Torrent) or flow cells (MiSeq)

(b)
Clonal amplification (in situ PCR) employing emulsion PCR in nanoliter wells (Ion Torrent) or bridge PCR on a flow cell surface (MiSeq)

4. **Sequencing**, involving:

(a)
Sequential addition of deoxynucleoside triphosphates (dNTPs) to growing DNA strands complementary to clonally amplified DNA

(b)
Detection of dNTP incorporation as a result of electrical signal proportional to pH change on a semiconductor chip (Ion Torrent) or fluorescently labeled "chain terminators" (MiSeq)

3 BIOGEOGRAPHICAL ANCESTRY

While not being the only technology to enable BGA prediction, MPS has certainly facilitated the use of phenotyping and BGA inference in the forensic community, mainly because it tolerates very large PCR multiplexes. While SNaPsot™ is restricted to a few dozen SNPs at most in a single multiplex, MPS multiplexes can include hundreds of targets, including SNPs, insertion/deletions (indels) and microhaplotypes. It is possible to combine the genotypes obtained from multiple SNaPshot™ assays but this means providing enough evidentiary material to these multiple PCR multiplexes. In fact, a hybrid approach where the PCR products from established SNaPshot™ assays (before SBE) are combined into a library for MPS sequencing has been shown to be effective on both the Ion PGM™ (Daniel et al., 2015) and the MiSeq (Mehta et al., 2016).

Regardless of the genotyping technology, in order to predict BGA, the following elements are required:

- A panel of **ancestry informative markers** (AIMs) that are known to be associated with BGA
- **Reference populations** consisting of genotypes at the selected AIMs from individuals with known BGA
- **Prediction algorithms** or classifiers that are able to infer BGA from the genotype of an unknown DNA donor by comparing it to those in reference populations.

3.1 Ancestry informative genetic markers

Because of recombination and mutation events, inherited, identical sequences of DNA (haplotypes) become shorter with increasing numbers of generations. These identical-by-decent (IBD) segments can be used to infer BGA and this is the method used by most commercial genealogy service providers (Erlich, Shor, Carmi, & Pe'er, 2018; Erlich, Shor, Pe'er, & Carmi, 2018). The forensic community has instead focused on shorter genetic sequences, most often preferring the shortest possible markers of all: SNPs. However, interest has been recently invested in microhaplotypes, sequences of DNA (with SNP variants) up to 200 bp in length, free of recombination hotspots (Kidd et al., 2013; Kidd et al., 2017; Kidd et al., 2018; Kidd & Speed, 2015).

Historically, the term “AIMs” has been reserved in the forensic community for autosomal markers which account for both maternal and paternal genetic contributions to an unknown genotype (Rosenberg, Li, Ward, & Pritchard, 2003). However, the use of “lineage markers” predates the use of AIMs in forensic genetics (Shriver & Kittles, 2004). This term has been reserved for markers that are uniparentally inherited. Y chromosome markers on the nonrecombining portion of the Y chromosome (NRY) are only found in human males and are only inherited paternally (Jobling & Tyler-Smith, 2003). Mitochondrial DNA (mtDNA) markers are found in both males and females but are only inherited maternally (Behar et al., 2012). While AIMs are diploid, lineage markers are haploid.

Because they are not subject to recombination during meiosis, lineage markers produce genotypes which are stable over multiple generations. As such they are indicative of ancestral genetic affiliation. However, because of their uniparental inheritance, the genetic contributions of many of the ancestors of a DNA donor are not represented. In Figure 1, for example, the male grandchild represented in the third generation only inherited a Y chromosome from his paternal grandfather and mtDNA from his maternal grandmother while the female grandchild only inherited mtDNA from her maternal grandmother. Neither of them inherited any lineage markers from their paternal grandmother or their maternal grandfather. Both of them, however, inherited equal proportions of autosomal DNA from each of their grandparents. Autosomal AIMs represent proportional representation from all ancestors, with the strongest representation from more recent ancestors (Phillips, 2015). It is the individual allele in an autosomal genotype that is indicative of ancestral genetic affiliation. For this reason, AIMs can provide a more accurate picture of recent genetic history while lineage markers can illuminate more ancient affiliations. It is lineage markers (and longer autosomal haplotypes) that have been of most use in determining human genetic origins (Behar et al., 2012; Jobling & Tyler-Smith, 2003; Underhill & Kivisild, 2007).

This figure has been replaced by a file (image_n/Figure 1.tif) that is not supported to display in the browser. Thus the previous image is still being displayed.

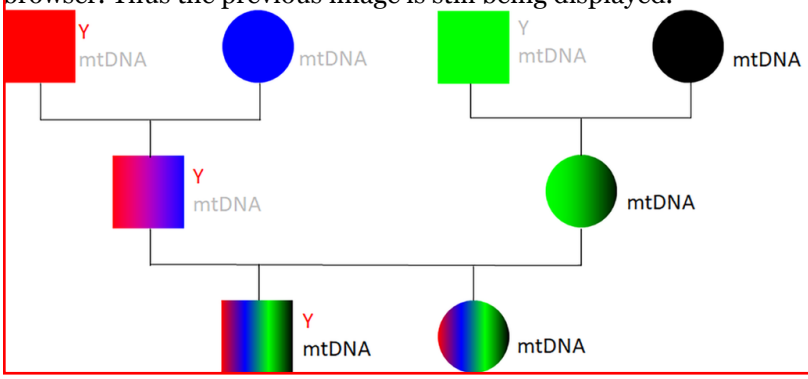


Figure 1 Pedigree over two generations showing males (squares, with both a Y chromosome and mtDNA) and females (circles, with no Y chromosome). Colors of lineage markers indicate ancestral origin (gray represents a lineage marker that is not passed on to grandchildren). Colors inside squares and circles represent the proportions of autosomal DNA inherited from each of the four grandparents

The first forensically relevant AIM panels were SNaPshot™ assays (Mehta et al., 2017). In order of publication, they included the SNP*for*ID 34-plex (Fondevila et al., 2013; Phillips et al., 2007, 2012) which can differentiate African, European, and East Asian populations; Eurasiaplex (Phillips et al., 2013) which was designed to be used in conjunction with the 34-plex to further differentiate European and South Asian populations; Pacifiplex (Santos et al., 2016) which was again designed to be used in conjunction with the 34-plex to further differentiate Oceanian populations; EurEAs_Gplex (Daca-Roszak et al., 2016) for differentiating European and East Asian populations; and Global AIMs Nano (de la Puente et al., 2016) which can differentiate African, European, East Asian, Oceanian, and American populations.

The increasing popularity of MPS has seen two global ancestry panels, originally developed as TaqMan® assays, incorporated into commercial MPS ancestry panels. The Kidd lab (Yale University) panel of 55 SNPs

(Kidd et al., 2014) comprise the AIMs in the ForenSeq™ DNA Signature Prep Kit (Verogen) (Churchill et al., 2016) and they have been combined with the Seldin Lab (University of California Davis) panel of 128 SNPs (Kidd et al., 2011; Kosoy et al., 2009) to comprise the Applied Biosystems Precision ID Ancestry Panel (Thermo Fisher Scientific) (Al-Asfi et al., 2018; Pereira, Mogensen, Børsting, & Morling, 2017).

3.2 Reference populations

There is an ever-increasing number of publically accessible reference human genotypes available as either high-density SNP genotypes or whole genome sequences. The most useful for forensic purposes have BGA metadata associated with them and include:

- International Genome Sample Resource (IGSR: http://www.internationalgenome.org/: formerly the 1000 Genomes Project and incorporating samples previously included in the International HapMap Project) (Sudmant et al., 2015; The 1000 Genomes Project Consortium et al., 2015)
- HGDP-CEPH Human Genome Diversity Cell Line Panel (http://www.cephb.fr/en/hgdp_panel.php) (Cann et al., 2002; Cavalli-Sforza, 2005; Dausset et al., 1990)
- Simons Genome Diversity Project (SGDP: https://www.simonsfoundation.org/simons-genome-diversity-project/) (Mallick et al., 2016)
- Estonian Biocentre Human Genome Diversity Panel (EGDP: http://evolbio.ut.ee/) (Pagani et al., 2016)
- HUGO Pan Asian SNP database (PanSNPdb: http://www4a.biotec.or.th/PASNP) (Ngamphiw et al., 2011; The HUGO Pan-Asian SNP Consortium, 2009)

The Forensic Resource/Reference on Genetics knowledge base (FROG-kb: http://frog.med.yale.edu/FrogKB/functionality.jsp), a part of the ALlele FREquency Database (ALFRED: https://alfred.med.yale.edu), can supply allele and genotype frequencies for global reference populations but does not provide individual genotypes (Kidd et al., 2018). Table 1 shows the database holdings. Databases with greater geographic coverage (e.g., SGDP, EGDP) have smaller numbers of individuals in each subpopulation (sometimes only two or three). PanSNPdb covers only East Asia and South East Asia. EGDP is concentrated in Eastern Europe and Asia. It should be noted that continental groupings are somewhat arbitrary, especially in the landmass bounded by Europe and Asia. Bioinformatics tools such as BCFtools (https://samtools.github.io/bcftools/) can be used to mine these databases in order to obtain reference genotypes for selected AIMs from variant call format (VCF) files (Danecek et al., 2011). SPSmart (SNPs for Population Studies: Amigo, Salas, Phillips, & Carracedo, 2008) provides an easily-accessible web-based portal for downloading SNP genotypes and associated metadata from 1000 Genomes Phase I and HGDP-CEPH, as well as HapMap and Perlegen (http://spsmart.cesga.es/). It allows different databases and populations to be combined into user-defined groups and gives graphical summaries of SNP population variability.

Table 1 Human reference populations held by some of the forensically relevant databases, organized into continental populations and subpopulations

Continental population	Subpopulation	IGSR	HGDP-CEPH	SGDP	EGDP	PanSNPdb
Africa	Sub-Saharan Africa	Esan	Bantu	Bantu	Congo	
		Gambian	Biaka	Biaka		
		Luhya	Mbuti	Dinka		
		Sierra Leone	Mandenka	Esan		
		Yoruba	San	Gambian		
			Yoruba	Ju'hoan		
				Luhya		
				Luo		
				Masai		
				Mandenka		
				Mbuti		
				Mende		
				San		
				Yoruba		
	North Africa		Mozabite	Mozabite		
				Saharawi		
Europe	British Isles	English	Orcadian	English		
				Orcadian		
	Scandinavia	Finnish		Finnish	Finnish	
				Icelandic	Saami	
					Swede	
					Vepsas	
	Mediterranean	Iberian	Bergamo	Bergamo		
		Italian	Tuscan	Greek		
				Sardinian		
				Spanish		
				Tuscan		
	Western Europe		Basque	Basque	German	
			French	French		
	Eastern Europe		Russian	Albanian	Albanian	
				Bulgarian	Bashkir	
				Czech	Belarusian	
				Estonian	Chuvash	
				Hungarian	Cossack	
				Polish	Croat	
				Russian	Estonian	
					Hungarian	
					Ingrian	
					Karelian	
					Komis	
					Tatar	
					Latvian	
					Lithuanian	
					Maris	
					Moldavian	
					Polish	
					Roma	
					Russian	
					Udmurd	
					Ukranian	
Middle East			Bedouin	Bedouin	Arab	
			Druze	Druze	Assyrian	
			Palestinian	Iraqi	Druze	
				Jordanian	Jordanian	
				Palestinian	Lebanese	
				Samaritan	Saudi	
				Turkish		
				Yemenite		
Asia	Caucasus		Adygei	Abkhasian	Abkhazian	
				Adygei	Armenian	
				Armenian	Avars	
				Chechen	Azerbaijan	
				Georgian	Balkars	
				Lezgin	Circassian	
					Georgian	
					Kabardin	

Continental population	Subpopulation	IGSR	HGDP-CEPH	SGDP	EGDP	PanSNPdb
					Kumyk	
					Lezgin	
					Ossetian	
					Tabasaran	
	Central Asia		Uygur	Kyrgyz	Ishkasim	Uygur
				Tajik	Kazakh	
				Uygur	Kyrgyz	
					Rushan-Vanch	
					Shugnan	
					Tajik	
					Turkmen	
					Uygur	
					Uzbek	
					Yaghnobi	
					Ishkasim	
	South West Asia	Punjabi	Balochi	Balochi	Iranian	
			Brahui	Brahui		
			Burusho	Burusho		
			Hazara	Hazara		
			Kalash	Iranian		
			Makrani	Kalash		
			Pathan	Makrani		
			Sindhi	Pathan		
				Punjabi		
				Sindhi		
	South Asia	Bengali		Bengali	Asur	
		Gujarati		Brahmin	Balija	
		Tamil		Irula	Bengali	
		Telugu		Kapu	Brahmin	
				Khonda-Dora	Dhaka	
				Kusunda	Gond	
				Madiga	Gupta	
				Mala	Ho	
				Relli	Kapu	
				Yadava	Kol	
				Bengali	Kshatriya	
					Kurmi	
					Malayan	
					Marwadi	
					Orissa	
					Punjab	
					Santhl	
					Tamang	
					Thakur	
	North Asia		Mongolian	Aleut	Altaian	
			Yakut	Altaian	Buryat	
				Chaplin	Chukchi	
				Chukchi	Evenk	
				Even	Nenet	
				Itelman	Ket	
				Mansi	Khanty	
				Mongola	Koryak	
				Naukan	Mansi	
				Sireniki	Mongolian	
				Tlingit	Nganasan	
				Tubalar	Sakha	
				Ulchi	Selkup	
				Yakut	Shor	
					Tuvinian	
					Yakut	
	East Asia	Dai	Dai	Dai		Han
		Han	Daur	Daur		Hmong
		Japanese	Han	Han		Japanese
			Hezhen	Hezhen		Jiamao
			Japanese	Japanese		Jinuo
			Lahu	Korean		Korean
			Miaozu	Lahu		Ryukyuan

Continental population	Subpopulation	IGSR	HGDP-CEPH	SGDP	EGDP	PanSNPdb
			Naxi	Miao		Wa
			Oroqen	Naxi		Zhuang
			She	Oroqen		
			Tu	She		
			Tujia	Tu		
			Xibu	Tujia		
			Yizu	Xibo		
				Yi		
	South East Asia	Vietnamese	Cambodian	Ami	Aeta	Alorese
				Atayal	Agta	Ati
				Burmese	Bajo	Agta
				Cambodian	Batak	Ayta
				Dusun	Burmese	Batak
				Igorot	Dusun	Dayak
					Igorot	Filipino
					Lebbo	H'tin
					Luzon	Iraya
					Murut	Javanese
					Vietnamese	Kambera
					Vizayan	Karen
						Lamahalot
						Lawa
						Lembata
						Malay
						Mamanwa
						Manggarai
						Mentawai
						Minanubu
						Mlabri
						Mon
						Paluang
						Plang
						Tai
						Sunda
						Toraja
						Yao
Oceania			Papuan	Australian	Koinanbe	
			Melanesian	Bougainville	Kosipe	
				Hawaiian		
				Maori		
				Papuan		
America	North America		Pima	Pima		
	Central America	Mexican	Mayan	Mayan		
		Puerto Rican		Mixe		
				Mixtec		
				Zapotec		
	South America	Columbian	Curripaco	Chane	Cachi	
		Peruvian	Karitiana	Karitiana	Colla	
			Piapoco	Piapoco	Wichi	
			Surui	Quechua		
				Surui		

3.3 Prediction algorithms (classifiers)

There are a number of algorithms or classifiers for inferring BGA from autosomal genotype. For a review, see Wollstein and Lao (2015). The performances of some of them have been compared where no admixture is assumed (Cheung, Gahan, & McNevin, 2017) and for individuals with mixed parentage (Cheung, Gahan, & McNevin, 2018a). The most popular algorithms for forensic use can be categorized into two types:

- Multidimensional scaling (MDS)
- Model-based likelihood estimators

3.3.1 Multidimensional scaling

This class of algorithm could also be referred to as reduced dimensionality spatial representation. It includes principal components analysis (PCA) (Abdi & Williams, 2010) and principal coordinates analysis (PCoA) (McVean, 2009; Patterson, Price, & Reich, 2006). The ForenSeq™ Universal Analysis Software (Verogen) (Illumina, 2016), for use with the ForenSeq™ DNA Signature Prep Kit, employs a two-dimensional (2D) PCA plot to analyze BGA.

PCA takes as input a matrix, **G**, of numerically represented genotypes where each element $g_{i,j}$ is the genotype of the *i*th individual at the *j*th locus. It is only biallelic genotypes that can be represented numerically such that genetic distances between genotypes are preserved. This can be achieved, for example, by coding heterozygotes as 0 and alternate homozygotes as −1 and +1. This preserves a distance of 2 between alternate homozygotes and a distance of 1 between homozygotes and heterozygotes, a reflection of actual genetic distances. However, for tri-allelic SNPs, tetra-allelic SNPs and microhaplotypes, genetic distances will be biased according to the choice of numerical code. For example, consider the six possible genotypes for a triallelic SNP (A/C/G): AA, AC, AG, CC, CG, and GG. If these are coded as 0, 1, 2, 3, 4, and 5, respectively, the distance between the homozygous genotypes AA and GG (5) is artificially greater than the distance between AA and CC (3).

To avoid this limitation, PCoA takes as input a matrix of genetic distances, **D**, where each element $d_{i,j}$ is the genetic distance of the *i*th individual from the *j*th individual, across all loci. There are a number of methods for calculating the genetic distances in **D** including Manhattan distance, Euclidian distance, chord distance (Cavalli-Sforza & Edwards, 1967), Nei's distance (Nei, 1972), the τ distance of Kidd and Cavalli-Sforza (1974), the coancestry coefficient (Reynolds, Weir, & Cockerham, 1983) and pairwise F_{ST} (Boca & Rosenberg, 2011) but in all cases, **D** is derived from **G**. Hence, PCoA can be used for genotypes consisting of three or more possible allele variants whereas PCA is limited to biallelic markers only (unless dummy variables are used to preserve genetic distances between variants).

The matrix **G** is reduced to orthogonal principal components or coordinates (PCs) by eigenvector decomposition. The matrix **D** is reduced to a specified number of PCs (usually 2 or 3) by eigenvalue decomposition. The PCs are then ordered so that the first PC accounts for the greatest amount of variance (and the greatest genetic distances) between genotypes, the second PC accounts for the second greatest amount of variance, etc. A ~~2D~~two- or three-dimensional plot with axes consisting of the first two or three PCs forms a spatial representation of genetic distances between individual genotypes. Individuals with genetic similarity will cluster together (Figure 2). If clusters have some correspondence with reference populations, then any individual's genetic relationship to BGAs can be framed in terms of those reference populations.

This figure has been replaced by a file (image_n/Figure 2.tif) that is not supported to display in the browser. Thus the previous image is still being displayed.

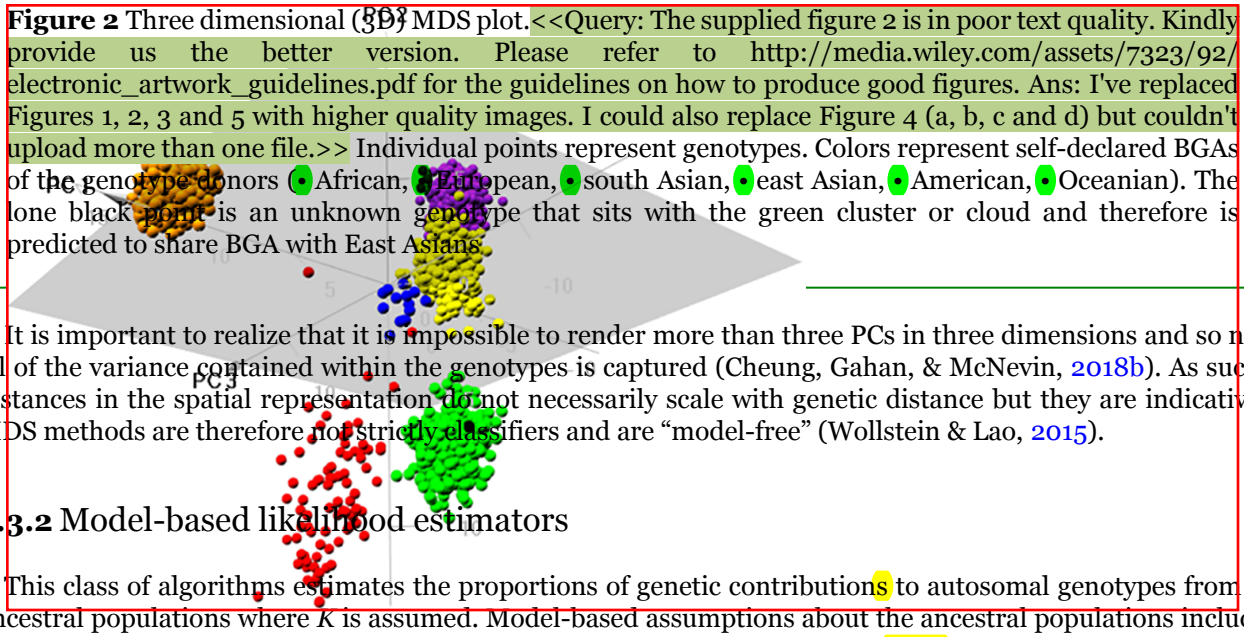


Figure 2 Three dimensional (3D) MDS plot.<<Query: The supplied figure 2 is in poor text quality. Kindly provide us the better version. Please refer to http://media.wiley.com/assets/7323/92/electronic_artwork_guidelines.pdf for the guidelines on how to produce good figures. Ans: I've replaced Figures 1, 2, 3 and 5 with higher quality images. I could also replace Figure 4 (a, b, c and d) but couldn't upload more than one file.>> Individual points represent genotypes. Colors represent self-declared BGAs of the genotype donors (• African, • European, • south Asian, • east Asian, • American, • Oceanian). The lone black point is an unknown genotype that sits with the green cluster or cloud and therefore is predicted to share BGA with East Asians.

It is important to realize that it is impossible to render more than three PCs in three dimensions and so not all of the variance contained within the genotypes is captured (Cheung, Gahan, & McNevin, 2018b). As such, distances in the spatial representation do not necessarily scale with genetic distance but they are indicative. MDS methods are therefore not strictly classifiers and are "model-free" (Wollstein & Lao, 2015).

3.3.2 Model-based likelihood estimators

This class of algorithms estimates the proportions of genetic contributions to autosomal genotypes from K ancestral populations where K is assumed. Model-based assumptions about the ancestral populations include that they are in Hardy Weinberg equilibrium (HWE) and linkage equilibrium (LE). In essence, the algorithms apply the following equality:

$$\mathbf{G} = \mathbf{QP} \tag{1}$$

where **G** is the matrix of (known) genotypes (represented as numbers of a particular allele: 0, 1, or 2) such that $g_{i,j}$ is the genotype of the ith individual at the jth locus, **P** is the matrix of (unknown) genotype frequencies in the K ancestral populations such that $p_{j,k}$ is the genotype frequency at the jth locus in the kth population. **Q** is the matrix of (unknown) genetic contributions such that $q_{i,k}$ is the contribution to the ith individual from the kth population.

The (unknown) elements of **P** and **Q** are estimated by different methods, depending on the algorithm. Arguably the most popular algorithm is *structure* (Porras-Hurtado et al., 2013) which uses a Bayesian framework to update prior estimates of **P** and **Q** given **G** according to the posterior probability distribution (Pritchard, Stephens, & Donnelly, 2000) given by:

$$P(\mathbf{Q}, \mathbf{P} \mid \mathbf{G}) \propto P(\mathbf{Q})P(\mathbf{P})P(\mathbf{G} \mid \mathbf{Q}, \mathbf{P}) \tag{2}$$

Markov chain Monte Carlo (MCMC) simulations of P(**P**), P(**Q**), and P(**G**|**Q**,**P**) enable sampling from the posterior probability distribution and a log-likelihood estimation is maximized until convergence. Initial values are that **P** is modeled by the Dirichlet distribution (Balding & Nichols, 1995; Foreman, Smith, & Evett, 1997; Rannala & Mountain, 1997) and **Q** is defined by equal contributions from each of the K populations. The ADMIXTURE algorithm avoids MCMC simulations by directly maximizing the log-likelihood estimation for **P** and **Q** rather than sampling from the posterior distribution (Alexander, Novembre, & Lange, 2009). This results in faster run times than *structure*.

The HID SNP Genotyper Plugin (Applied Biosystems, 2017a), which supports the Precision ID Ancestry Panel, defines **P** from seven root populations (Africa, America, Southwest Europe or Middle East, Europe, Oceania, East Asia, and South Asia) which are derived from ALFRED. This is different to *structure* where **P** is inferred. The posterior probability distribution is now:

$$P(\mathbf{Q} \mid \mathbf{P}, \mathbf{G}) = P(\mathbf{Q} \mid \mathbf{G}) \propto P(\mathbf{Q})P(\mathbf{G} \mid \mathbf{Q}) \tag{3}$$

HID SNP Genotyper then simulates **Q** by generating combinations of $q_{i,k}$ in 5% increments and converges on the matrix which maximizes P(**G**| **Q**). A confidence value for each q_i is reported by comparing the log-likelihood P($g_i|q_i$) with the same log-likelihood for 10,000 randomly simulated individuals with the same q_i. High confidence is reported if log-likelihood P($g_i|q_i$) lies within the 95% confidence interval for the 10,000 simulations. Low confidence is reported if it lies outside the 95% confidence interval.

HID SNP Genotyper also calculates a population likelihood for each of 65 subpopulations (of the seven root populations) according to:

$$L_k = \prod_j p_{j,\,k} \tag{4}$$

Subpopulations can be ranked by L such that the highest L represents the highest probability of membership for an unknown genotype (which is assumed not to be admixed for this calculation).

It is the elements of **Q** that provide the proportion of each ancestral population or cluster to each individual autosomal genotype. As for MDS, if inferred ancestral clusters have some correspondence with reference populations, then any individual's genetic relationship to ancestral clusters can be framed in terms of those reference populations.

Figure 3 shows a summary (bar) plot of an inferred **Q** matrix derived from *structure* analysis. Reference populations 1, 2, and 3 correspond with the green, blue and red ancestral clusters, respectively, even though there was no prior population membership assumed by the model. However, there are some individuals who appear misplaced. For example, there is one individual in population 1 with greater than 50% genetic contribution from the blue ancestral cluster and two individuals in population 2 with greater than 50% genetic contribution from the green ancestral cluster. The contributions of the ancestral clusters (and, by association, the reference populations) to any unknown genotype will be represented by the proportions of each color for that genotype.

This figure has been replaced by a file (image_n/Figure 3.tif) that is not supported to display in the browser. Thus the previous image is still being displayed.

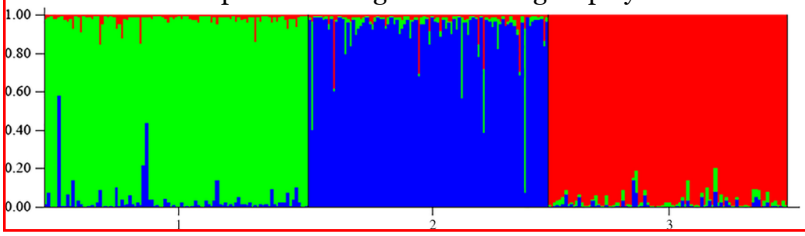


Figure 3 Summary plot of estimates of **Q**. each individual genotype is represented by a single vertical line broken into K colored segments, with lengths proportional to genetic contributions from each of the K inferred clusters. The numbers (1, 2, and 3) correspond to reference populations (<<Query: Please note that reference Pritchard, Wen & Falush (2010) has been cited in Figure 3 but not provided in list. Please provide in list or delete the citation. Ans: The following reference has been added (but I can't seem to include carriage returns):Pritchard, J. K., Wen, X., & Falush, D. (2010). Documentation for structure software:Version 2.3. Retrieved from https://web.stanford.edu/group/pritchardlab/structure_software/release_versions/v2.3.4/structure_doc.pdf>>Pritchard, Wen & Falush, 2010)

3.3.3 Online tools for inference of BGA

There are at least two online portals for ancestry inference:

- FROG-kb (previously mentioned)
- The Snipper app suite (http://mathgene.usc.es/snipper/forensic_mps_aims.html) (Phillips et al., 2007)

FROG-kb allows use of numerous ancestry informative SNP panels including the SNP*for*ID 34-plex (Fondevila et al., 2013; Phillips et al., 2007, 2012), the Seldin Lab panel of 128 SNPs (Kidd et al., 2011; Kosoy et al., 2009), the Kidd lab panel of 55 SNPs (Kidd et al., 2014), Eurasiaplex (Phillips et al., 2013), Pacifiplex (Santos et al., 2016), and Precision ID Ancestry Panel (Al-Asfi et al., 2018; Pereira et al., 2017). The user can paste delimited genotype data as text into a window and the portal returns populations ranked by likelihood calculated under the assumption of HWE. This approach is only applicable for nonadmixed individuals, however, as noted earlier and discussed later.

The Snipper app suite allows the user to upload custom reference genotypes as well as providing reference data for the SNP*for*ID 34plex, Eurasiaplex, Pacifiplex, the Kidd lab panel, the Seldin lab panel, the Precision ID Ancestry Panel, the ForenSeq™ DNA Signature Prep Kit 55 AIMs (Churchill et al., 2016) and the EUROFORGEN Global AIM-SNP set (Phillips et al., 2014). Reference genotypes in formatted Microsoft Excel files are uploaded and the portal returns a selection of data exploration and ancestry inference algorithms including MDS, a naïve Bayesian classifier and a genetic distance algorithm.

3.4 Admixture

Predicting BGA from autosomal genotype can be likened to predicting the contributions to a paint mixture from its color. Just as any individual has genetic contributions from their parents who in turn have genetic contributions from their parents, so too, any paint color can be made up of primary colors, assuming an additive model (e.g., RGB color space). This analogy does not extend to represent segregation of alleles during meiosis whereby chromosomes are randomly assorted and are, in turn, crossed-over as a result of recombination but it is adequate to illustrate that someone with mixed parentage may display maximum likelihood for a population from which neither parent is derived. Primary colors red, green, and blue can be combined as shown in Figure 4a to produce cyan, magenta, and yellow. These in turn can be combined in three generations as shown in Figure 4b–d which all produce the same paint color in the third generation. The first generations in Figure 4b and c both consist of the same primary paint color proportions (2 × red, 1 × blue, 1 × green) but different second generation paint colors. All three third generation paints in Figure 4b–d have primary colors red, green, and blue in the proportion 2:1:1, but each has different parentage. The third generation paint in Figure 4d has all paint parents and grandparents of the same color.

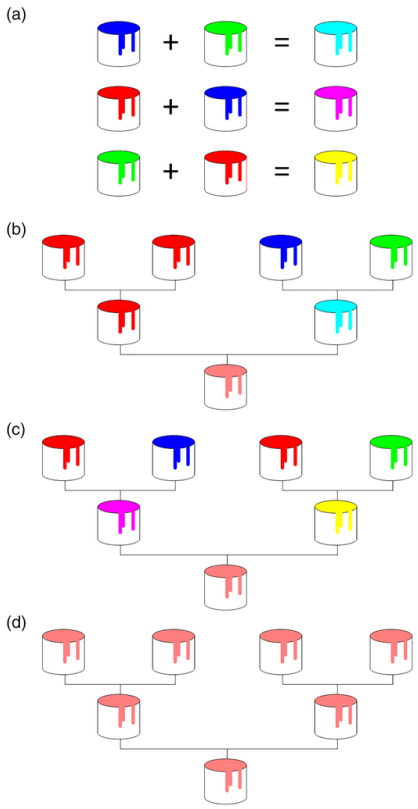


Figure 4 (a) Combinations of red, green, and blue that produce cyan, magenta, and yellow. (b) A third generation paint produced from red and cyan paint parents. (c) The same paint produced from magenta and yellow paint parents. (d) The same paint produced from paint parents and grandparents of the same color

3.4.1 Admixture due to mixed recent parentage

The third generation paints in Figure 4b and c are analogous to individuals who have grandparents from three different BGAs. For example, two grandparents with European ancestry (red), one with African ancestry (blue) and one with South Asian ancestry (green). In Figure 4b, this corresponds with a European (red) parent and another parent with mixed African/South Asian (blue/green = cyan) ancestry. In Figure 4c, it corresponds with one European/African (red/blue = magenta) parent and one European/South Asian (red/green = yellow) parent. We say that these second and third generation individuals are admixed due to mixed recent parentage.

3.4.2 Apparent admixture due to unavailable reference populations

The third generation paint in Figure 4d, on the other hand, is analogous to an individual whose parents and grandparents have the same BGA but this BGA is not represented by an existing reference population. Just as the color of this paint is comprised of a mixture of primary colors (2 × red, 1 × blue, 1 × green), so too, the individual, their parents and grandparents all belong to a BGA that is a mixture of BGAs represented by existing reference populations. For example, if our primary colors (red, blue, and green) represent European, African and South Asian ancestry, then all the individuals in Figure 4d belong to a BGA that is genetically intermediate between these but for which there may not be an existing reference population (e.g., Middle Eastern). These individuals have apparent admixture due to unavailable reference populations.

3.4.3 Detecting admixture

How can we differentiate between admixture due to mixed recent parentage and apparent admixture due to unavailable reference populations? There are two supplemental analyses that can be added to autosomal genotyping.

Haplotyping of lineage markers allows paternal (Y chromosome) and maternal (mtDNA) lineages to be established (Figure 1). Consider, for example, an unknown DNA donor with an autosomal genotype that was found to have ancestral contributions from European, African, and South Asian BGAs. No Middle Eastern reference population is available. The donor also has NRY and mtDNA haplotypes that are relatively common in the Middle East but relatively absent in Europe, Africa, and South Asia. It is more likely that the individual and their maternal and paternal ancestors are derived from a Middle Eastern population.

Another helpful analysis is comparison of heterozygosity in the unknown genotype with heterozygosities in reference populations. Figure 5 demonstrates that any offspring of parents from different populations with different allele frequencies is expected to have a higher heterozygosity than if both parents were from the same population, assuming HWE in those populations.This is just a restatement of the well-known Wahlund effect (Wahlund, 1928). The difference in expected heterozygosity between offspring and parents from different populations will be exacerbated for AIMs that have been selected exactly because they have large differences in allele frequencies between populations, with a maximum difference observed for AIMs that are fixed for one allele in one parent's population and fixed for an alternate allele in the other parent's population (Figure 5). Conversely, if both parents are from the same population, then the expected heterozygosity in the offspring will be no different, assuming the population is in HWE.

This figure has been replaced by a file (image_n/Figure 5.tif) that is not supported to display in the browser. Thus the previous image is still being displayed.

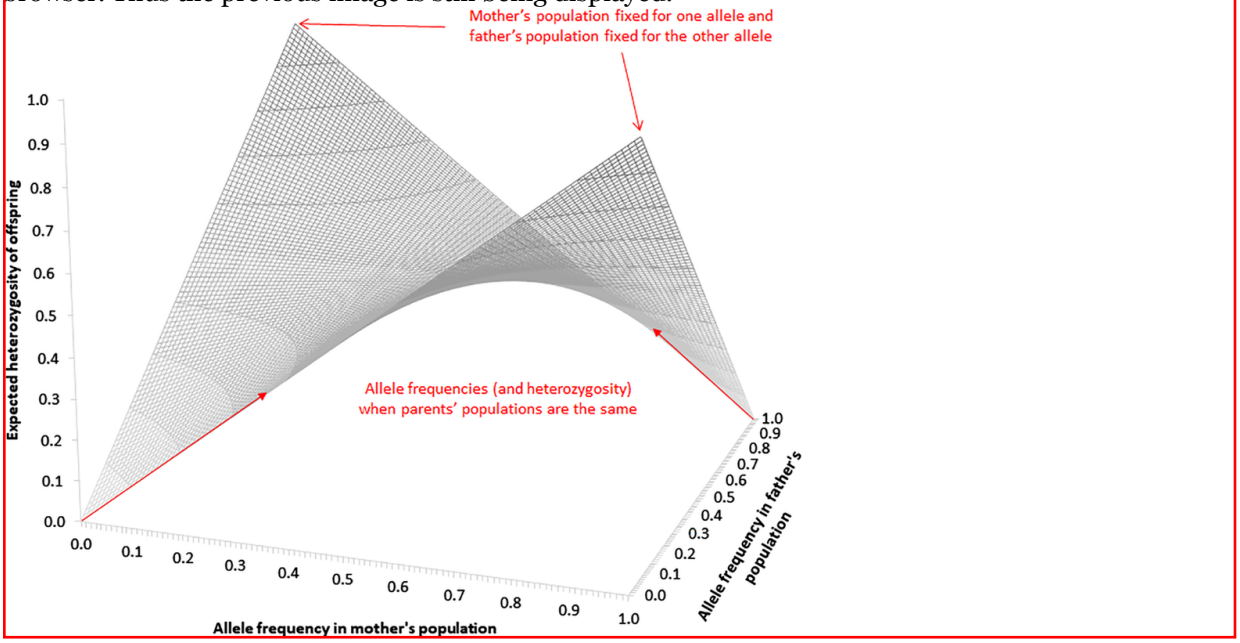


Figure 5 Expected heterozygosity of admixed offspring as a function of allele frequencies in parents' populations. Offspring heterozygosity will always be greater than heterozygosities in parents' populations when allele frequencies in parents' populations are not equal

Hence, we can differentiate between admixture due to mixed recent parentage and apparent admixture by testing for statistically significant differences in heterozygosity between the unknown genotype (averaged over all loci) and the reference populations which are found to contribute, genetically. Let H^0 be the heterozygosity

observed over all loci in an unknown genotype and let $H^{o}_{i,k}$ be the heterozygosity observed over all loci for the *i*th individual in the *k*th reference population found to contribute to the unknown genotype. We can use an appropriate statistical test for the null hypothesis that H^{o} lies in the range of the following distribution, with a prescribed probability of type I error (e.g., $p < .05$):

$$\sum_{k}\left(q_k \sum_{i} H^{o}_{i,k}\right) \tag{5}$$

where q_k is the proportion of the individual genotype contributed by the *k*th population ~~contributing to the individual genotype~~. If the reference populations are in HWE then allele frequencies can be used to calculate expected heterozygosities and:

$$\sum_{k}\left(q_k \sum_{i} H^{o}_{i,k}\right) = \sum_{k}\left(q_k \sum_{j} H^{e}_{j,k}\right) \tag{6}$$

where $H^{e}_{j,k}$ is the expected heterozygosity at the *j*th locus in the *k*th population.

3.4.4 Reporting

The potential for admixture must be carefully reported. In the absence of inbreeding, each (biological) parent contributes 50% of the alleles to any of their offspring, each grandparent contributes 25%, each great grandparent contributes 12.5%, and so on. As such, any genetic contribution less than 20% will represent the equivalent of a single great grandparent at most and any contribution less than 10% will represent the equivalent of a single great-great grandparent at most. The genetic influence of any ancestor therefore diminishes with the number of generations that separate them from any individual.

The Centre for Forensic Sciences in Toronto, Canada, reports admixture from the seven root populations used by HID SNP Genotyper (Jin et al., 2018). They report four possible results:

- Single inclusion (one root population ≥80%, others ≤15%)
- Single mixed inclusion (one root population in the range 55–75%, others ≤15%)
- Double inclusion (two root populations ≥20%)
- Uninformative (at least three root populations ≥20%)

This system provided correspondence of single inclusion with individuals who self-declared ancestry from one root population (99% of 648 individuals). For potentially admixed individuals, however, correspondence with double inclusion was weak (15 of 33 individuals).

4 THE GENETIC ANCESTRY LAB

The Genetic Ancestry Lab (GAL) is a joint venture between the University of Canberra (UC) and the University of Technology Sydney (UTS). It received seed funding from ANU Connect Ventures (DTF224) (http://www.anuconnectventures.com.au/) and the AMP's Tomorrow Fund (2123) (https://www.ampstomorrowfund.com.au/) and provides predictions of BGA and EVCs from biological evidence received as either extracted DNA or original tissue. The operations of the GAL can be divided into wet lab and dry lab.

4.1 Wet lab

Items are generally received by courier and refrigerated at 4°C in a secure laboratory. Chain of custody is documented. All subsamples are stored at 4°C or −20°C, as appropriate, until they are consumed in the process of analysis.

4.1.1 DNA quantitation

DNA is quantified using the Quantifiler™ Human DNA Quantification Kit (Applied Biosystems) according to the manufacturer's recommended protocol (Applied Biosystems, 2014) in a 7500 Real Time PCR System (Applied Biosystems). The amplification of the target is compared with the amplification of a dilution series of standard (control) DNA and the concentration of the extracted DNA is calculated from a standard curve. Amplification of internal PCR controls (IPCs) in questioned samples is compared with amplification of IPCs in the standards and any relative delay in amplification is indicative of inhibition. Quantifiler™ Human is used in preference to later kits (e.g., Quantifiler™ Trio DNA Quantification Kit) because it is more sensitive to inhibitors, in keeping with the PCR used for MPS target enrichment.

4.1.2 Target enrichment

The extracted DNA is diluted to 0.067 ng/μL and PCR is performed on one nanogram (1 ng = 15 μL) of DNA from each sample using the Precision ID Ancestry Panel (Applied Biosystems) as described earlier and according to the manufacturer's recommended protocols (Applied Biosystems, 2017b). A total of 22 amplification cycles are employed with a 4-min anneal and extension time. More cycles can be used if less than 1 ng of DNA is available but it is important that they are not included in the same library as more concentrated samples which will dominate sequence coverage.

4.1.3 Library preparation

Library preparation is performed using the Precision ID Ancestry Panel library preparation procedure (Applied Biosystems, 2017b) on an Ion Chef™ automated library preparation and templating instrument (Applied Biosystems). Sample-specific IonCode™ DNA barcodes (Applied Biosystems) are ligated to the DNA amplicons generated by PCR and these amplicons are then pooled with the amplicons from other samples which have their own sample-specific barcodes ligated. Each amplicon in the pool can be identified by genetic locus of origin (from alignment to a reference genome) and sample of origin (from DNA barcode). There are 32 IonCode™ barcodes currently available in four Precision ID DL8 Kits, each accommodating eight samples, which means that the minimum batch size for processing is eight. Each batch of eight contains a negative library preparation control (NLPC) consisting of autoclaved, deionized water and a positive library preparation control (PLPC) consisting of AmpFℓSTR™ DNA Control 007 (Applied Biosystems).

4.1.4 Template preparation

Individual barcoded DNA amplicons in the pooled library are attached to individual ISPs by the Ion Chef™ instrument according to the manufacturer's recommended protocol (Applied Biosystems, 2017b). They are

then clonally amplified so that each ISP has multiple copies of each barcoded amplicon. Excess ISPs are removed so that only ISPs with clonally amplified amplicons remain.

4.1.5 Sequencing

Individual enriched ISPs with clonally amplified, barcoded DNA amplicons are loaded into individual wells on an Ion 520™ chip (Ion Torrent) according to the manufacturer's recommended protocol (Applied Biosystems, 2017b). This chip can accommodate 3–6 million reads. With 165 loci in the Precision ID panel, this equates to over 18,000 reads per locus which is over 500 reads per locus when distributed over 32 samples. The clonally amplified amplicons in each well are sequenced on an Ion GeneStudio™ S5 System massively parallel sequencer (Applied Biosystems) using Ion S5™ sequencing chemistry (Ion Torrent) according to the manufacturer's recommended protocol (Applied Biosystems, 2017b) with 200 bp, single-end reads.

4.2 Dry lab

Individual sequences are aligned to a human reference genome (GRCh37/hg19) and then combined into a BAM (binary alignment map) file for each sample using Torrent Suite software on an Ion Server (Ion Torrent). Target region variants defined by .bed files for the Precision ID Ancestry Panel are downloaded in the following formats:

- VCF file as .cov.xls from the variantCaller plugin
- HID SNP Genotyper Report from the HID_SNP_Genotyper plugin

4.2.1 Sequence output

For each sample at each of the 165 genetic loci in the Precision ID Ancestry Panel, the number of reads for each nucleotide (A, C, G, T) is extracted from the VCF file. The PLPC is checked for the following quality metrics:

- Coverage (the number of reads) for each SNP is >100×.
- No allele frequencies in the range 0.1–0.3 and 0.7–0.9 (this represents the range where the distinction between homozygote and heterozygote is ambiguous).
- SNP genotypes are concordant with consensus genotypes

Total coverage (number of reads) for the NLPC should be a negligible fraction (e.g., <1%) of the total coverage for the PLPC. Finally, any locus in any sample for which allele frequencies are in the range 0.1–0.3 and 0.7–0.9 are excluded from analysis.

4.2.2 Genotyping

Allele frequency windows are used to define genotypes according to Table 2. Coverage thresholds are not applied as it has been shown that setting appropriate allele frequency windows is more effective for reducing erroneous genotypes than coverage thresholds (Avent et al., 2018).

Table 2 Genotyping decisions derived from allele frequencies

Allele frequency window	Decision	Rationale
0–0.1	Allele ignored	Potential sequencing error
0.1–0.3	Genotype ignored	Distinction between homozygote and heterozygote is ambiguous
0.3–0.7	Heterozygous genotype	Two alleles with similar relative frequencies
0.7–0.9	Genotype ignored	Distinction between homozygote and heterozygote is ambiguous
0.9–1.0	Homozygous genotype	Allele with relative frequency >0.9

4.2.3 BGA prediction

All but one[1] of the 165 ancestry informative genotypes for each sample are analyzed using *structure* and PCoA using the *pcoa* function (*ape* package) in R (https://www.r-project.org/) (Paradis, Claude, & Strimmer, 2004; R Core Team, This should be on a new line (I can't seem to enter a carriage return).2015). Unknown genotypes are analyzed together with reference data consisting of genotypes at the same 164 loci for 2,262 individuals drawn from the IGSR, HGDP-CEPH, and SGDP databases. Most (2,099) of these are included in the "Applied Biosystems Precision ID Ancestry Panel 165" reference data downloaded from the "Forensic MPS AIMs Panel Reference Sets" webpage from The Snipper 2.5 app suite. The remainder has been drawn from other sources, including SPSmart.

In addition, only 151[2] of the 165 ancestry informative genotypes for each sample are analyzed using the HID SNP Genotyper Plugin within the Torrent Suite software on an Ion Server (Applied Biosystems). As described earlier, this algorithm produces two assignments:

- Continental-level admixture proportions, that is, genetic contributions from seven major continental root populations and
- Subpopulation likelihoods, that is, relative probability that the DNA donor is derived genetically from each of 65 subpopulations

5 INTERPRETATION AND REPORTING

The results of analyses by *structure*, PCoA and HID SNP Genotyper are compared in order to provide BGA predictions for each sample. An example of such a comparison for three samples is shown in Table 3. The process of interpretation can then be summarized as follows:

1. Identify continental-level BGAs that are unambiguously excluded by all analyses.

2. Identify continental-level BGAs that are included by any analyses.

3. If more than one continental-level BGA is included, document the possibility of admixture or apparent admixture due to the unavailability of reference populations.

4. Test the hypothesis that the unknown individual has mixed recent parentage by testing for statistically significant differences in heterozygosity between the unknown genotype (averaged over all loci) and the reference populations which are found to contribute, genetically.

5. Provide examples of subpopulations with high likelihoods. For samples that appear admixed, these only apply if apparent admixture due to the unavailability of reference populations cannot be excluded as a possibility.

Table 3 Comparison of classifications from HID SNP Genotyper, structure and PCoA for three samples

	HID SNP Genotyper			
Sample	**Continental BGA**	**Subpopulations**	**Structure**	**PCoA**
1	100% European	European	80% Nth European 18% Sth European	European
2	95% East Asian	East Asian	80% East Asian 11% Sth East Asian	East Asian
3	40% Sth West Asian 35% Sth Asian 25% East Asian	Asian	46% Mid Eastern 44% Sth Asian	Sth Asian

Table 4 shows this process applied to the three samples in Table 3. Samples 1 and 2 are relatively unadmixed while sample 3 is apparently admixed where two possibilities exist: the donor is truly admixed (with mixed recent parentage) or the donor and their ancestors are derived from a population for which a reference does not exist (apparent admixture). For the latter possibility, examples of subpopulations with high likelihoods (as estimated by HID SNP Genotyper: Table 5) are suggested. Currently, the GAL does not haplotype lineage markers although this would provide further information about the potential for admixture or apparent admixture. Haplotyping of lineage markers is a future direction for the GAL.

Table 4 Interpretation of the samples in Table 3

Sample	Interpretation
1	The donor of this DNA does not have significant African, South Asian, East Asian, Oceanian or indigenous American BGA. They have a majority ancestral genetic contribution from Europe. Examples include Irish, Hungarians, and Danes. They are more likely to have ***European*** ancestry than any other continental BGA[1]. They are likely to have a majority of ancestors (e.g., Parents, grandparents) from Europe.
2	The donor of this DNA does not have significant African, European, South West Asian (Middle Eastern), South Asian, Oceanian, or indigenous American BGA. They have a majority ancestral genetic contribution from East Asia. Examples include Taiwanese, Han, Hakka, Koreans, or Japanese. They are more likely to have ***East Asian*** ancestry than any other continental BGA[1]. They are likely to have a majority of ancestors (e.g., Parents, grandparents) from East Asia.
3	The donor of this DNA does not have significant African, European, Oceanian, or indigenous American BGA. They have major ancestral genetic contributions from South West Asia (Middle East) and South Asia and a minor contribution from East Asia. There are two possibilities: • The donor has ancestors from a region genetically intermediate between the Middle East, South Asia and East Asia. Examples include Kachari, Pashtun, Keralite, Hazara, and Kuwaiti. • The donor has ancestors from the Middle East, South Asia and East Asia (e.g., a Middle Eastern grandparent, a South Asian grandparent and an East Asian grandparent).

[1]BGAs include African, Middle Eastern, European, South Asian, East Asian, Oceanian, and indigenous A merican.

Table 5 The five subpopulations with the highest likelihoods for samples in Table 3 as estimated by HID SNP Genotyper

Sample 1			Sample 2			Sample 3		
Population	**Geo-region**	**Likelihood**	**Population**	**Geo-region**	**Likelihood**	**Population**	**Geo-region**	**Likelihood**
Irish	Europe	1.076×10^{-37}	Taiwanese Han	East Asia	8.911×10^{-49}	Kachari	Asia	1.185×10^{-50}
Hungarian	Europe	6.565×10^{-38}	Han—HapMap	East Asia	5.090×10^{-49}	Pashtun	Asia	6.945×10^{-51}
Europeans-HapMap	Europe	1.254×10^{-38}	Hakka	East Asia	5.071×10^{-49}	Keralite	Asia	3.499×10^{-51}
Danes	Europe	1.220×10^{-38}	Koreans	East Asia	1.754×10^{-49}	Hazara	Asia	1.139×10^{-51}
European Americans	Europe	1.011×10^{-38}	Japanese HapMap	East Asia	5.037×10^{-50}	Kuwaiti	Asia	1.064×10^{-52}

5.1 Conclusions

To the author's knowledge, the GAL is the first forensic phenotyping service to operate in Australia. It makes use of the Ion Torrent platform including the Ion Chef™ for automated library and template preparation as well as the Ion GeneStudio™ S5 System for MPS of 165 SNPs included in the Precision ID Ancestry Panel. These AIMs are then used to provide estimates of BGA using three different algorithms: PCoA, *structure* and the HID SNP Genotyper plugin for Ion Torrent applications. By analyzing the data in these three different ways, a degree of cross-verification is possible that provides added confidence in predictions.

Samples are processed in batches of eight which is the number of samples that can be accommodated in Precision ID DL8 cartridges. Every cartridge includes a PLPC and NLPC. This means that there are six noncontrol samples processed in each cartridge. The libraries from up to four DL8 cartridges can be pooled for sequencing on a single Ion chip resulting in 24 noncontrol samples, four NLPCs and four PLPCs. This limitation is a result of only 32 available IonCode™ barcodes for the Precision ID DL8 Kits. With each DL8 library (eight samples) taking about 7 hr to prepare on the Ion Chef™ (Applied Biosystems, 2017b), library preparation represents a bottle neck in the GAL. However, this is counterbalanced by the cost savings that can be achieved by processing multiple samples. A single sample requires the use of a DL8 cartridge (for eight samples) and Ion S5™ sequencing reagents for two chips (even if only one chip is filled). This means that the reagent usage (and cost) for sequencing one sample is about half that for 24 samples and there are definite economies of scale to be achieved with the cost per sample decreasing as more samples are processed together.

There are other considerations. Pooling of libraries (to achieve economies of scale) requires equimolar concentrations of barcoded amplicons from each sample in order to ensure equal access to nanolitre wells on the Ion Chip. This is achieved on the Ion Chef™ using magnetic bead purification which acts to remove DNA beyond a concentration threshold. However, DNA concentrations below the threshold will remain low. It is important, therefore, to avoid processing low template amounts (<1 ng) of DNA with high template amounts (> 1 ng). It is also good practice to rotate barcodes to avoid any possibility of carryover between samples.

MPS sequencing is error prone, regardless of platform, although some are more error prone than others (Liu et al., 2012; Ratan et al., 2013). With minimum error rates in the order of about 1% of base calls, confidence is increased with the number of reads (or depth of coverage). The greater the depth of coverage, the more accurate the genotype. There is a point where a sequence variant (e.g., SNP) must be distinguished from an erroneous genotype and this is why careful delineation of allele frequency windows for genotype designations is important (e.g., Table 2). Avent et al. (2018) were able to show that any alleles with a frequency less than 15% could be regarded as potential sequencing error and removed from analysis when using the GeneRead DNAseq panel (QIAGEN) to genotype identity SNPs on the Ion PGM™ (Applied Biosystems). They also demonstrated that high coverage thresholds (below which some alleles were ignored) led to allele drop out and resultant genotyping errors. The mean coverage should be at least five times greater than any coverage threshold applied (Avent et al., 2018).

Finally, the legal, ethical, and privacy implications of deriving personal information from DNA have not been considered in this review. They are, nevertheless, important: see Scudder, McNevin, Kelty, Walsh, and Robertson (2018b) for a discussion. Privacy concerns can be addressed by the implication of a privacy impact assessment (Scudder, McNevin, Kelty, Walsh, & Robertson, 2018a). It is also possible for intelligence information to mislead an investigation if not properly integrated into a general law enforcement intelligence framework (Scudder, Robertson, Kelty, Walsh, & McNevin, 2019).

ACKNOWLEDGMENTS

The GAL is hosted by the Faculty of Science and the Faculty of Engineering and Information Technology (FEIT) at the University of Technology Sydney (UTS) and the Faculty of Science and Technology at the University of Canberra. Access to the Ion Chef™and Ion GeneStudio™ S5 System in the Biomedical and

Tissue Engineering Laboratory at FEIT has been kindly facilitated by the School of Biomedical Engineering at UTS.The GAL depends on its dedicated staff:

- Dr Michelle Nelson (Manager, University of Canberra)
- Ms Sumaiya Quasim (University of Canberra)
- Dr Greg Adcock (University of Canberra)
- ~~Associate~~ Professor Dianne Gleeson (University of Canberra)

Valuable technical support for the GAL is supplied by Thermo Fisher Scientific, especially Dr Lucy Dagostino and Dr Daniel Power. The author thanks two anonymous reviewers for their constructive feedback which greatly improved this review. ~~The GAL is hosted by the Faculty of Science and the Faculty of Engineering and Information Technology (FEIT) at the UTS and the Faculty of Science and Technology at the UC. Access to the Ion Chef™ and Ion GeneStudio™ S5 System in the Biomedical and Tissue Engineering Laboratory at FEIT has been kindly facilitated by the School of Biomedical Engineering at UTS.~~

CONFLICT OF INTEREST

The<<Query: We have provided this standard statement as per journal style. Please modify only if you have any conflict to declare, else retain the same. Ans: The author is the Director of the Genetic Ancestry Lab (GAL).>> author ~~has declared no conflicts of interest for this article~~is the Director of the Genetic Ancestry Lab (GAL).

Endnotes

[1]rs3811801 is not available in the reference databases

[2]Only 151 of the 165 available SNPs are used by the HID SNP Genotyper Plugin

REFERENCES

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *WIREs Computational Statistics*, **2**(4), 433–459. https://doi.org/10.1002/wics.101

Ahmadian, A., Ehn, M., & Hober, S. (2006). Pyrosequencing: History, biochemistry and future. *Clinica Chimica Acta*, **363**(1), 83–94. https://doi.org/10.1016/j.cccn.2005.04.038

Al-Asfi, M., McNevin, D., Mehta, B., Power, D., Gahan, M. E., & Daniel, R. (2018). Assessment of the precision ID ancestry panel. *International Journal of Legal Medicine*, **132**(6), 1581–1594. https://doi.org/10.1007/s00414-018-1785-9

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, **19**(1), 655–664. https://doi.org/10.1101/gr.094052.109

Alvarez-Cubero, M. J., Saiz, M., Martínez-García, B., Sayalero, S. M., Entrala, C., Lorente, J. A., & Martinez-Gonzalez, L. J. (2017). Next generation sequencing: An application in forensic sciences? *Annals of Human Biology*, **44**(7), 581–592. https://doi.org/10.1080/03014460.2017.1375155

Amigo, J., Salas, A., Phillips, C., & Carracedo, Á. (2008). SPSmart: Adapting population based SNP genotype databases for fast and comprehensive web access. *BMC Bioinformatics*, **9**, 428. https://doi.org/10.1186/1471-2105-9-428

Applied Biosystems. (2010). *ABI PRISM® SNaPshot™ Multiplex Kit*. Foster City, CA. Retrieved from https://assets.thermofisher.com/TFS-Assets/LSG/manuals/cms_041203.pdf

Applied Biosystems. (2014). *Quantifiler® Human and Y human male DNA quantification kits*. Retrieved from http://tools.thermofisher.com/content/sfs/manuals/cms_041395.pdf

Applied Biosystems. (2017a). *HID SNP genotyper plugin USER GUIDE v5.2.2*. Carlsbad, CA. Retrieved from https://assets.thermofisher.com/TFS-Assets/LSG/manuals/MAN0010641_HIDSNP_Genotyper_Plugin.pdf

Applied Biosystems. (2017b). Precision ID panels with the ion S5™ system: Application guide. Carlsbad, CA: Retrieved from http://tools.thermofisher.com/content/sfs/manuals/MAN0015831_PrecisionID_Panels_IonS5_UG.pdf

Armentrout, P. (2018). Parabon® Announces Snapshot® genetic genealogy service for law enforcement: Forensic DNA samples screened for nearly 100 agencies and hits abound (Press release). Retrieved from http://parabon-nanolabs.com/news-events/2018/05/parabon-snapshot-genetic-genealogy-dna-analysis-service.html

Avent, I., Kinnane, A. G., Jones, N., Petermann, I., Daniel, R., Gahan, M. E., & McNevin, D. (2018). The QIAGEN 140-locus single-nucleotide polymorphism (SNP) panel for forensic identification using massively parallel sequencing (MPS): An evaluation and a direct-to-PCR trial. *International Journal of Legal Medicine*, **133**, 677–688. https://doi.org/10.1007/s00414-018-1975-5

Balding, D. J., & Nichols, R. A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, **96**(1), 3–12. https://doi.org/10.1007/bf01441146

Behar, D. M., van Oven, M., Rosset, S., Metspalu, M., Loogväli, E.-L., Silva, N. M., … Villems, R. (2012). A "Copernican" reassessment of the human mitochondrial DNA tree from its root. *The American Journal of Human Genetics*, **90**(4), 675–684. https://doi.org/10.1016/j.ajhg.2012.03.002

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59. https://doi.org/10.1038/nature07517

Berglund, E. C., Kiialainen, A., & Syvänen, A.-C. (2011). Next-generation sequencing technologies and applications for human genetic history and forensics. *Investigative Genetics*, **2**(23), 23. https://doi.org/10.1186/2041-2223-2-23

Binladen, J., Gilbert, M. T. P., Bollback, J. P., Panitz, F., Bendixen, C., Nielsen, R., & Willerslev, E. (2007). The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS One*, **2**, e197. https://doi.org/10.1371/journal.pone.0000197

Boca, S. M., & Rosenberg, N. A. (2011). Mathematical properties of *Fst* between admixed populations and their parental source populations. *Theoretical Population Biology*, **80**(3), 208–216. https://doi.org/10.1016/j.tpb.2011.05.003

Børsting, C., & Morling, N. (2015). Next generation sequencing and its applications in forensic genetics. *Forensic Science International: Genetics*, **18**, 78–89. https://doi.org/10.1016/j.fsigen.2015.02.002

Budowle, B., Schmedes, S. E., & Wendt, F. R. (2017). Increasing the reach of forensic genetics with massively parallel sequencing. *Forensic Science, Medicine and Pathology*, **13**(3), 342–349. https://doi.org/10.1007/s12024-017-9882-5

Cann, H. M., de Toma, C., Cazes, L., Legrand, M.-F., Morel, V., Piouffre, L., … Cavalli-Sforza, L. L. (2002). A human genome diversity cell line panel. *Science*, **296**(5566), 261–262. https://doi.org/10.1126/science.296.5566.261b

Cavalli-Sforza, L. L. (2005). The human genome diversity project: Past, present and future. *Nature Reviews Genetics*, **6**, 333–340. https://doi.org/10.1038/nrg1596

Cavalli-Sforza, L. L., & Edwards, A. W. F. (1967). Phylogentic analysis: Models and estimation procedures. *Evolution*, **21**(3), 550–570. https://doi.org/10.1111/j.1558-5646.1967.tb03411.x

Cheung, E. Y. Y., Gahan, M. E., & McNevin, D. (2017). Prediction of biogeographical ancestry from genotype: A comparison of classifiers. *International Journal of Legal Medicine*, **131**(4), 901–912. https://doi.org/10.1007/s00414-016-1504-3

Cheung, E. Y. Y., Gahan, M. E., & McNevin, D. (2018a). Prediction of biogeographical ancestry in admixed individuals. *Forensic Science International: Genetics*, **36**, 104–111. https://doi.org/10.1016/j.fsigen.2018.06.013

Cheung, E. Y. Y., Gahan, M. E., & McNevin, D. (2018b). Predictive DNA analysis for biogeographical ancestry. *Australian Journal of Forensic Sciences*, **50**(6), 651–658. https://doi.org/10.1080/00450618.2017.1422021

Churchill, J. D., Chang, J., Ge, J., Rajagopalan, N., Wootton, S. C., Chang, C. W., … B., B. (2015). Blind study evaluation illustrates utility of the ion PGM™ system for use in human identity DNA typing. *Croatian Medical Journal*, **56**, 218–229. https://doi.org/10.3325/cmj.2015.56.218

Churchill, J. D., Schmedes, S. E., King, J. L., & Budowle, B. (2016). Evaluation of the Illumina® Beta version ForenSeq™ DNA signature prep kit for use in genetic profiling. *Forensic Science International: Genetics*, **20**, 20–29. https://doi.org/10.1016/j.fsigen.2015.09.009

Claes, P., Liberton, D. K., Daniels, K., Rosana, K. M., Quillen, E. E., Pearson, L. N., … Shriver, M. D. (2014). Modeling 3D facial shape from DNA. *PLoS Genetics*, **10**(3), e1004224. https://doi.org/10.1371/journal.pgen.1004224

Claes, P., Roosenboom, J., White, J. D., Swigut, T., Sero, D., Li, J., … Weinberg, S. M. (2018). Genome-wide mapping of global-to-local genetic effects on human facial shape. *Nature Genetics*, **50**(3), 414–423. https://doi.org/10.1038/s41588-018-0057-4

Daca-Roszak, P., Pfeifer, A., Żebracka-Gala, J., Jarząb, B., Witt, M., & Ziętkiewicz, E. (2016). EurEAs_Gplex—A new SNaPshot assay for continental population discrimination and gender identification. *Forensic Science International: Genetics*, **20**, 89–100. https://doi.org/10.1016/j.fsigen.2015.10.004

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Group, G. P. A. (2011). The variant call format and VCFtools. *Bioinformatics*, **27**(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Daniel, R., Santos, C., Phillips, C., Fondevila, M., van Oorschot, R. A. H., Carracedo, Á., … McNevin, D. (2015). A SNaPshot of next generation sequencing for forensic SNP analysis. *Forensic Science International: Genetics*, **14**, 50–60. https://doi.org/10.1016/j.fsigen.2014.08.013

Dausset, J., Cann, H., Cohen, D., Lathrop, M., Lalouel, J.-M., & White, R. (1990). Centre d'Etude du polymorphisme humain (CEPH): Collaborative genetic mapping of the human genome. *Genomics*, **6**(3), 575–577. https://doi.org/10.1016/0888-7543(90)90491-C

de la Puente, M., Santos, C., Fondevila, M., Manzo, L., Carracedo, Á., Lareu, M. V., & Phillips, C. (2016). The global AIMs Nano set: A 31-plex SNaPshot assay of ancestry-informative SNPs. *Forensic Science International: Genetics*, **22**, 81–88. https://doi.org/10.1016/j.fsigen.2016.01.015

<<Query: Please provide the “volume number” for reference Erlich et al., 2018. Ans: There are no volumes for bioRxiv, only DOIs.>>Erlich, Y., Shor, T., Carmi, S., & Pe'er, I. (2018). Re-identification of genomic data using long range familial searches. *bioRxiv*, 350231. https://doi.org/10.1101/350231

Erlich, Y., Shor, T., Pe'er, I., & Carmi, S. (2018). Identity inference of genomic data using long-range fa milial searches. *Science*, **362**(6415), 690–694. https://doi.org/10.1126/science.aau4832

Fondevila, M., Phillips, C., Santos, C., Freire Aradas, A., Vallone, P. M., Butler, J. M., … Carracedo, Á. (2013). Revision of the SNPforID 34-plex forensic ancestry test: Assay enhancements, standard referenc e sample genotypes and extended population studies. *Forensic Science International: Genetics*, **7**(1), 63–7 4. https://doi.org/10.1016/j.fsigen.2012.06.007

Foreman, L. A., Smith, A. F. M., & Evett, I. W. (1997). Bayesian analysis of DNA profiling data in forensi c identification applications. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **16 0**(3), 429–459. https://doi.org/10.1111/j.1467-985X.1997.00074.x

Frudakis, T., Venkateswarlu, K., Thomas, M. J., Gaskin, Z., Ginjupalli, S., Gunturi, S., … Nachimuthu, P. K. (2003). A classifier for the SNP-based inference of ancestry [published erratum appears in J Forensic Sci 2004, 49(5)]. *Journal of Forensic Science*, **48**(4), 771–782.

Greenspan, B. (2019). *Connecting families and saving lives (Press release)*. Retrieved from https://blog.fa milytreedna.com/press-release-connecting-families-and-saving-lives/

Guo, D.-C., & Milewicz, D. M. (2003). Methodology for using a universal primer to label amplified DNA s egments for molecular analysis. *Biotechnology Letters*, **25**(24), 2079–2083. https://doi.org/10.1023/B:BI LE.0000007075.24434.5e

Henn, B. M., Hon, L., Macpherson, J. M., Eriksson, N., Saxonov, S., Pe'er, I., & Mountain, J. L. (201 2). Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. *PLoS One*, **7**(4), e34267. https://doi.org/10.1371/journal.pone.0034267

Hoffmann, C., Minkah, N., Leipzig, J., Wang, G., Arens, M. Q., Tebas, P., & Bushman, F. D. (2007). DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. *Nucleic Acids Researc h*, **35**(13), e91–e91. https://doi.org/10.1093/nar/gkm435

Illumina. (2013). *Infinium® HTS assay protocol guide*. San Diego, CA. Retrieved from https://support.illu mina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/humano mniexpress-24/infinium_hts_assay_protocol_user_guide_15045738_a.pdf

Illumina. (2015). *ForenSeq™ DNA signature prep: Reference guide*. San Diego, CA. Retrieved from http s://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_docume ntation/forenseq/forenseq-dna-signature-prep-guide-15049528-01.pdf

Illumina. (2016). *ForenSeq™ universal analysis software guide*. Retrieved from https://support.illumina.c om/content/dam/illumina-support/documents/documentation/software_documentation/forenseq-univer sal-analysis-software/forenseq-universal-analysis-software-guide-15053876-01.pdf

Jäger, A. C., Alvarez, M. L., Davis, C. P., Guzmán, E., Han, Y., Way, L., … Holt, C. L. (2017). Developmental validation of the MiSeq FGx forensic genomics system for targeted next generation sequencing in forensic DNA casework and database laboratories. *Forensic Science International: Genetics*, **28**, 52–70. https://doi.org/10.1016/j.fsigen.2017.01.011

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., … Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, **36**, 338–345. https://doi.org/10.1038/nbt.4060

Jeffreys, A. J., Wilson, V., & Thein, S. L. (1985). Hypervariable 'minisatellite' regions in human DNA. *Nature*, **314**(6006), 67–73. https://doi.org/10.1038/314067a0

Jin, S., Chase, M., Henry, M., Alderson, G., Morrow, J. M., Malik, S., … Laird, J. (2018). Implementing a biogeographic ancestry inference service for forensic casework. *Electrophoresis*, **39**(21), 2757–2765. https://doi.org/10.1002/elps.201800171

Jobling, M. A., & Tyler-Smith, C. (2003). The human Y chromosome: An evolutionary marker comes of age. *Nature Reviews Genetics*, **4**, 598–612. https://doi.org/10.1038/nrg1124

Kayser, M. (2015). Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes. *Forensic Science International: Genetics*, **18**, 33–48. https://doi.org/10.1016/j.fsigen.2015.02.003

Kayser, M., & de Knijff, P. (2011). Improving human forensics through advances in genetics, genomics and molecular biology. *Nature Reviews Genetics*, **12**, 179–192. https://doi.org/10.1038/nrg2952

Keating, B., Bansal, A. T., Walsh, S., Millman, J., Newman, J., Kidd, K., … Consortium, o. b. o. t. I. V. T. G. (2013). First all-in-one diagnostic tool for DNA intelligence: Genome-wide inference of biogeographic ancestry, appearance, relatedness, and sex with the Identitas v1 forensic chip. *International Journal of Legal Medicine*, **127**(3), 559–572. https://doi.org/10.1007/s00414-012-0788-1

Kidd, J. R., Friedlaender, F. R., Speed, W. C., Pakstis, A. J., De La Vega, F. M., & Kidd, K. K. (2011). Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples. *Investigative Genetics*, **2**(1), 1. https://doi.org/10.1186/2041-2223-2-1

Kidd, K. K., & Cavalli-Sforza, L. L. (1974). The role of genetic drift in the differentiation of Icelandic and Norwegian cattle. *Evolution*, **28**(3), 381–395. https://doi.org/10.2307/2407159

Kidd, K. K., Pakstis, A. J., Speed, W. C., Lagace, R., Chang, J., Wootton, S., & Ihuegbu, N. (2013). Microhaplotype loci are a powerful new type of forensic marker. *Forensic Science International: Genetics Supplement Series*, **4**(1), e123–e124. https://doi.org/10.1016/j.fsigss.2013.10.063

Kidd, K. K., Pakstis, A. J., Speed, W. C., Lagace, R., Wootton, S., & Chang, J. (2018). Selecting microh aplotypes optimized for different purposes. *Electrophoresis*, **39**(21), 2815–2823. https://doi.org/10.1002/elps.201800092

Kidd, K. K., Soundararajan, U., Rajeevan, H., Pakstis, A. J., Moore, K. N., & Ropero-Miller, J. D. (201 8). The redesigned forensic research/reference on genetics-knowledge base, FROG-kb. *Forensic Science In ternational: Genetics*, **33**, 33–37. https://doi.org/10.1016/j.fsigen.2017.11.009

Kidd, K. K., & Speed, W. C. (2015). Criteria for selecting microhaplotypes: Mixture detection and deconvo lution. *Investigative Genetics*, **6**(1), 1. https://doi.org/10.1186/s13323-014-0018-3

Kidd, K. K., Speed, W. C., Pakstis, A. J., Furtado, M. R., Fang, R., Madbouly, A., … Kidd, J. R. (2014). Progress toward an efficient panel of SNPs for ancestry inference. *Forensic Science International: Genetic s*, **10**, 23–32. https://doi.org/10.1016/j.fsigen.2014.01.002

Kidd, K. K., Speed, W. C., Pakstis, A. J., Podini, D. S., Lagacé, R., Chang, J., … Soundararajan, U. (20 17). Evaluating 130 microhaplotypes across a global set of 83 populations. *Forensic Science International: Genetics*, **29**, 29–37. https://doi.org/10.1016/j.fsigen.2017.03.014

Kircher, M., & Kelso, J. (2010). High-throughput DNA sequencing – Concepts and limitations. *BioEssay s*, **32**(6), 524–536. https://doi.org/10.1002/bies.200900181

Kosoy, R., Nassir, R., Tian, C., White, P. A., Butler, L. M., Silva, G., … Seldin, M. F. (2009). Ancestry informative marker sets for determining continental origin and admixture proportions in common populati ons in America. *Human Mutation*, **30**(1), 69–78. https://doi.org/10.1002/humu.20822

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., … Law, M. (2012). Comparison of next-generation sequ encing systems. *Journal of Biomedicine & Biotechnology*, **2012**, 1–11. https://doi.org/10.1155/2012/251364

Liu, Y.-Y., & Harbison, S. (2018). A review of bioinformatic methods for forensic DNA analyses. *Forensic Science International: Genetics*, **33**, 117–128. https://doi.org/10.1016/j.fsigen.2017.12.005

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., … Reich, D. (2016). The Simo ns genome diversity project: 300 genomes from 142 diverse populations. *Nature*, **538**, 201–206. https://doi.org/10.1038/nature18964

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., … Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380. https://doi.org/10.1038/nature03959

Marouli, E., Graff, M., Medina-Gomez, C., Lo, K. S., Wood, A. R., Kjaer, T. R., … Lettre, G. (2017). Rare and low-frequency coding variants alter human adult height. *Nature*, **542**, 186–190. https://doi.org/10.1038/nature21039

McVean, G. (2009). A genealogical interpretation of principal components analysis. *PLoS Genetics*, **5**(10), e1000686. https://doi.org/10.1371/journal.pgen.1000686

Mehta, B., Daniel, R., Phillips, C., Doyle, S., Elvidge, G., & McNevin, D. (2016). Massively parallel sequencing of customised forensically informative SNP panels on the MiSeq. *Electrophoresis*, **37**(21), 2832–2840. https://doi.org/10.1002/elps.201600190

Mehta, B., Daniel, R., Phillips, C., & McNevin, D. (2017). Forensically relevant SNaPshot® assays for human DNA SNP analysis: A review. *International Journal of Legal Medicine*, **131**(1), 21–37. https://doi.org/10.1007/s00414-016-1490-5

Metzker, M. L. (2009). Sequencing technologies — The next generation. *Nature Reviews Genetics*, **11**, 31–46. https://doi.org/10.1038/nrg2626

Meyer, M., Stenzel, U., & Hofreiter, M. (2008). Parallel tagged sequencing on the 454 platform. *Nature Protocols*, **3**, 267–278. https://doi.org/10.1038/nprot.2007.520

Meyer, M., Stenzel, U., Myles, S., Prüfer, K., & Hofreiter, M. (2007). Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Research*, **35**(15), e97–e97. https://doi.org/10.1093/nar/gkm566

Minotta, M., & Endicott, E. (2018). *Thermo Fisher Scientific and Illumina sign agreement to provide research market broader access to Ion AmpliSeq technology (Press release)*. Retrieved from https://www.businesswire.com/news/home/20180108006946/en/

Mitra, R. D., Shendure, J., Olejnik, J., Edyta Krzymanska, O., & Church, G. M. (2003). Fluorescent in situ sequencing on polymerase colonies. *Analytical Biochemistry*, **320**(1), 55–65. https://doi.org/10.1016/S0003-2697(03)00291-4

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., & Erlich, H. (1986). *Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction*. ~~Paper presented at the~~ *Cold Sping Harbor Symposia on Quantitative Biology*, **51**, 263-273.<<Query: Please provide conference location for ref. Mullis et al. (1986). Ans: The reference has been corrected.>>

Nei, M. (1972). Genetic distance between populations. *The American Naturalist*, **106**(949), 283–292. https://doi.org/10.1086/282771

Newsome, M. (2007). A new DNA test can ID a suspect's race, but police won't touch it. *Wired*.

Ngamphiw, C., Assawamakin, A., Xu, S., Shaw, P. J., Yang, J. O., Ghang, H., … the HUGO Pan-Asian SNP Consortium. (2011). PanSNPdb: The pan-Asian SNP genotyping database. *PLoS One*, **6**(6), e21451. https://doi.org/10.1371/journal.pone.0021451

Pagani, L., Lawson, D. J., Jagoda, E., Mörseburg, A., Eriksson, A., Mitt, M., … Metspalu, M. (2016). Genomic analyses inform on migration events during the peopling of Eurasia. *Nature*, **538**, 238–242. https://doi.org/10.1038/nature19792

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of Phylogenetics and evolution in R language. *Bioinformatics*, **20**(2), 289–290. https://doi.org/10.1093/bioinformatics/btg412

Parameswaran, P., Jalili, R., Tao, L., Shokralla, S., Gharizadeh, B., Ronaghi, M., & Fire, A. Z. (2007). A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Research*, **35**(19), e130–e130. https://doi.org/10.1093/nar/gkm760

Pareek, C. S., Smoczynski, R., & Tretyn, A. (2011). Sequencing technologies and genome sequencing. *Journal of Applied Genetics*, **52**(4), 413–435. https://doi.org/10.1007/s13353-011-0057-x

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and Eigenanalysis. *PLoS Genetics*, **2**(12), e190. https://doi.org/10.1371/journal.pgen.0020190

Pereira, R., Phillips, C., Pinto, N., Santos, C., Santos, S. E. B. d., Amorim, A., … Gusmão, L. (2012). Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing. *PLoS One*, **7**(1), e29684. https://doi.org/10.1371/journal.pone.0029684

Pereira, V., Mogensen, H. S., Børsting, C., & Morling, N. (2017). Evaluation of the precision ID ancestry panel for crime case work: A SNP typing assay developed for typing of 165 ancestral informative markers. *Forensic Science International: Genetics*, **28**, 138–145. https://doi.org/10.1016/j.fsigen.2017.02.013

Phillips, C. (2015). Forensic genetic analysis of bio-geographical ancestry. *Forensic Science International: Genetics*, **18**, 49–65. https://doi.org/10.1016/j.fsigen.2015.05.012

Phillips, C. (2018). The Golden state killer investigation and the nascent field of forensic genealogy. *Forensic Science International: Genetics*, **36**, 186–188. https://doi.org/10.1016/j.fsigen.2018.07.010

Phillips, C., Aradas, A. F., Kriegel, A. K., Fondevila, M., Bulbul, O., Santos, C., … Lareu, M. V. (2013). Eurasiaplex: A forensic SNP assay for differentiating European and south Asian ancestries. *Forensic Science International: Genetics*, **7**(3), 359–366. https://doi.org/10.1016/j.fsigen.2013.02.010

<<Query: Please provide the “location of publisher” for reference Phillips et al., 2012. Ans: Humana Press is now a subsidiary of Springer which has Headquarters in New York, USA.>>Phillips, C., Fondevila, M., & Lareau, M. V. (2012). A 34-plex autosomal SNP single base extension assay for ancestry investigations. In

A. Alonso (Ed.), *DNA electrophoresis protocols for forensic genetics. Methods in molecular biology (meth ods and protocols)* (Vol. **830**). Humana Press.

Phillips, C., Parson, W., Lundsberg, B., Santos, C., Freire-Aradas, A., Torres, M., … Lareu, M. V. (201 4). Building a forensic ancestry panel from the ground up: The EUROFORGEN global AIM-SNP set. *Foren sic Science International: Genetics*, **11**, 13–25. https://doi.org/10.1016/j.fsigen.2014.02.012

Phillips, C., Salas, A., Sánchez, J. J., Fondevila, M., Gómez-Tato, A., Álvarez-Dios, J., … Carracedo, Á. (2007). Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *F orensic Science International: Genetics*, **1**(3), 273–280. https://doi.org/10.1016/j.fsigen.2007.06.008

<<Query: Please provide the "page range" for reference Porras-Hurtado et al., 2013. Ans: Pages 1-13. The re ference has been corrected.>>Porras-Hurtado, L., Ruiz, Y., Santos, C., Phillips, C., Carracedo, Á., & La reu, M. (2013). An overview of STRUCTURE: Applications, parameter settings, and supporting software. *F rontiers in Genetics*, **4**(98) ,1-13. https://doi.org/10.3389/fgene.2013.00098

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, **155**(2), 945–959.

Pritchard, J. K., Wen, X., & Falush, D. (2010). Documentation for structure software:Version 2.3. Retrieved from https://web.stanford.edu/group/pritchardlab/structure_software/release_versions/v2.3.4/structur e_doc.pdfR Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ram, N., Guerrini, C. J., & McGuire, A. L. (2018). Genealogy databases and the future of criminal investi gation. *Science*, **360**(6393), 1078–1079. https://doi.org/10.1126/science.aau1083

Rannala, B., & Mountain, J. L. (1997). Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences*, **94**(17), 9197–9201. https://doi.org/10.1073/pnas.94.17.9197

Ratan, A., Miller, W., Guillory, J., Stinson, J., Seshagiri, S., & Schuster, S. C. (2013). Comparison of se quencing platforms for single nucleotide variant calls in a human sample. *PLoS One*, **8**(2), e55089. http s://doi.org/10.1371/journal.pone.0055089

Reynolds, J., Weir, B. S., & Cockerham, C. C. (1983). Estimation of the coancestry coefficient: Basis for a short-term genetic distance. *Genetics*, **105**(3), 767–779.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M., & Nyrén, P. (1996). Real-time DNA sequen cing using detection of pyrophosphate release. *Analytical Biochemistry*, **242**(1), 84–89. https://doi.org/1 0.1006/abio.1996.0432

Rosenberg, N. A., Li, L. M., Ward, R., & Pritchard, J. K. (2003). Informativeness of genetic markers for inference of ancestry*. *The American Journal of Human Genetics*, **73**(6), 1402–1422. https://doi.org/10.1086/380416

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., … Bustillo, J. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, **475**, 348–352. https://doi.org/10.1038/nature10242

Santos, C., Phillips, C., Fondevila, M., Daniel, R., van Oorschot, R. A. H., Burchard, E. G., … Lareu, M. V. (2016). Pacifiplex: An ancestry-informative SNP panel centred on Australia and the Pacific region. *Forensic Science International: Genetics*, **20**, 71–80. https://doi.org/10.1016/j.fsigen.2015.10.003

Santos, C., Phillips, C., Oldoni, F., Amigo, J., Fondevila, M., Pereira, R., … Lareu, M. V. (2015). Completion of a worldwide reference panel of samples for an ancestry informative Indel assay. *Forensic Science International: Genetics*, **17**, 75–80. https://doi.org/10.1016/j.fsigen.2015.03.011

Santos, N. P. C., Ribeiro-Rodrigues, E. M., Ribeiro-dos-Santos, Â. K. C., Pereira, R., Gusmão, L., Amorim, A., … Santos, S. E. B. (2010). Assessing individual interethnic admixture and population substructure using a 48–insertion-deletion (INSEL) ancestry-informative marker (AIM) panel. *Human Mutation*, **31**(2), 184–190. https://doi.org/10.1002/humu.21159

Scudder, N., McNevin, D., Kelty, S. F., Walsh, S. J., & Robertson, J. (2018a). Forensic DNA phenotyping: Developing a model privacy impact assessment. *Forensic Science International: Genetics*, **34**, 222–230. https://doi.org/10.1016/j.fsigen.2018.03.005

Scudder, N., McNevin, D., Kelty, S. F., Walsh, S. J., & Robertson, J. (2018b). Massively parallel sequencing and the emergence of forensic genomics: Defining the policy and legal issues for law enforcement. *Science & Justice*, **58**(2), 153–158. https://doi.org/10.1016/j.scijus.2017.10.001

<<Query: Please provide the “volume number” for reference Scudder et al., 2019. Ans: There is no print version. This paper has only been published online (with a DOI).>>Scudder, N., Robertson, J., Kelty, S. F., Walsh, S. J., & McNevin, D. (2019). A law enforcement intelligence framework for use in predictive DNA phenotyping. *Australian Journal of Forensic Sciences*, 1–4. https://doi.org/10.1080/00450618.2019.1569132

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, **26**, 1135–1145. https://doi.org/10.1038/nbt1486

Shewale, J. G., Qi, L., & Calandro, L. M. (2013). Principles, practice, and evolution of capillary electrophoresis as a tool for forensic DNA analysis. In J. G. Shewale & R. H. Liu (Eds.), *Forensic DNA analysis: Current practices and emerging technologies* (1st ed., pp. 131–162). Boca Raton, FL: CRC Press.

Shriver, M. D., & Kittles, R. A. (2004). Genetic ancestry and the search for personalized genetic histories. *Nature Reviews Genetics*, **5**, 611–618. https://doi.org/10.1038/nrg1405

Silvia, A. L., Shugarts, N., & Smith, J. (2017). A preliminary assessment of the ForenSeq™ FGx system: Next generation sequencing of an STR and SNP multiplex. *International Journal of Legal Medicine*, **131**(1), 73–86. https://doi.org/10.1007/s00414-016-1457-6

Singer, A., & Breakiron, C. (2019). *Bode Technology announces forensic genealogy service to law enforcement agencies and crime laboratories: Using DNA testing and genealogical research provides investigators with more investigative leads in violent crimes and unidentified remains cases (Press release)*. Retrieved from https://www.prweb.com/releases/bode_technology_announces_forensic_genealogy_service_to_law_enforcement_agencies_and_crime_laboratories/prweb16091796.htm

Sobrino, B., Brión, M., & Carracedo, A. (2005). SNPs in forensic genetics: A review on SNP typing methodologies. *Forensic Science International*, **154**(2), 181–194. https://doi.org/10.1016/j.forsciint.2004.10.020

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., … Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**, 75–81. https://doi.org/10.1038/nature15394

Syvänen, A.-C. (1999). From gels to chips: "Minisequencing" primer extension for analysis of point mutations and single nucleotide polymorphisms. *Human Mutation*, **13**(1), 1–10. https://doi.org/10.1002/(SICI)1098-1004(1999)13:1<1::AID-HUMU1>3.0.CO;2-I

The 1000 Genomes Project Consortium, Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, **526**, 68–74. https://doi.org/10.1038/nature15393

The HUGO Pan-Asian SNP Consortium. (2009). Mapping human genetic diversity in Asia. *Science*, **326**(5959), 1541–1545. https://doi.org/10.1126/science.1177074

Underhill, P. A., & Kivisild, T. (2007). Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annual Review of Genetics*, **41**(1), 539–564. https://doi.org/10.1146/annurev.genet.41.110306.130407

Wahlund, S. (1928). Zusammensetzung von Populationen und Korrelationserscheinungen vom Standpunkt der Vererbungslehre aus Betrachtet. *Hereditas*, **11**(1), 65–106. https://doi.org/10.1111/j.1601-5223.1928.tb02483.x

Wollstein, A., & Lao, O. (2015). Detecting individual ancestry in the human genome. *Investigative Genetics*, **6**(7), 7. https://doi.org/10.1186/s13323-015-0019-x

Zhang, J., Chiodini, R., Badr, A., & Zhang, G. (2011). The impact of next-generation sequencing on genomics. *Journal of Genetics and Genomics*, **38**(3), 95–109. https://doi.org/10.1016/j.jgg.2011.02.003