



From SNP Mixtures to Separated Genotypes

Mark W. Perlin, PhD, MD, PhD
Cybergenetics, Pittsburgh, PA

Abstract

Single nucleotide polymorphism (SNP) sequencing is increasingly used in forensic genetics and forensic investigative genetic genealogy (FIGG). However, SNP mixture interpretation remains a challenging problem.

Unlike highly polymorphic STR markers, SNPs are usually bi-allelic, so contributors frequently overlap at the same two alleles. Consequently, many different contributor genotype combinations can appear very similar in a mixture, making contributor separation and evidential interpretation difficult.

We developed a hierarchical Bayesian method that separates contributor genotypes from SNP next-generation sequencing (NGS) mixture data. The method jointly analyzes all loci in a mixture, estimates contributor genotypes and mixture proportions, and produces posterior genotype probability distributions for each contributor.

Rather than giving a single definite genotype, these distributions describe the possible contributor genotypes and how much probability each receives after analyzing the sequencing data. They summarize the evidence contained in the data, quantify uncertainty, and can be used directly for likelihood-ratio (LR) calculations and forensic comparisons.

The SNP genotype separation method was applied to SNP sequence data generated using the Verogen ForenSeq DNA Signature Prep Kit (Primer Set B) processed through the MiSeq FGx sequencing platform. Analyses used only the panel's 94 identity SNP loci distributed across 23 chromosomes.

Despite the relatively small number of markers, contributor genotypes were successfully separated from mixed samples and produced substantial evidential support. In a two-person 25:75 mixture, the separated 25% contributor produced approximately 20 ban of log(LR) support. In a three-person 1:4:4 mixture, a separated 44% contributor yielded about 12 ban of support.

These results demonstrate that probabilistic separation and LR-based interpretation of forensic SNP mixtures are feasible using a small genome-wide SNP panel.

The approach extends probabilistic DNA interpretation beyond STR markers [1, 2] and provides a framework for using SNP evidence in forensic casework. Because STR and SNP loci represent independent sources of genetic information, available evidential support from both marker systems can be combined.

The SNP results reported here were obtained from just 94 identity SNP loci, whereas modern FIGG applications routinely analyze hundreds of thousands of SNP markers. These findings therefore suggest a path toward highly informative probabilistic interpretation of complex SNP mixtures in investigative genetic genealogy and other large-scale human identification sequencing applications.

Acknowledgement

The author thanks the Cybergenetics and Kern Regional Crime Laboratory NGS validation team for generating and organizing the sequencing data used in this study.

Next Generation Sequencing (NGS)

NGS, or massive parallel sequencing, is a newer methodology that can be used for forensic DNA applications. This technology produces identification information based on both DNA size (like traditional STR analysis) and DNA base pair sequence. NGS data comprises of traditional STR markers as well as single nucleotide polymorphisms (SNPs), mitochondrial DNA (mtDNA), and other forensically relevant markers.

Using NGS, DNA laboratories can analyze many DNA samples at one time (in parallel). Using this methodology, a lab can also increase the amount of DNA data produced for a sample, leading to a higher discriminatory power for DNA identification. Additionally, having the DNA sequence information allows labs to potentially distinguish between people having the same allele value at a DNA location (isoalleles) as the DNA sequences between the two people can differ. NGS analysis can also be used for better analysis of degraded DNA samples, genealogy applications, and high-throughput of samples in less time (due to parallel sequencing).

Both NGS and CE workflows start with DNA extraction, quantification, and amplification using polymerase chain reaction (PCR). The both also produce DNA data that an analyst can interpret. For the CE workflow, the STR DNA fragments are size separated and the relative fluorescent intensity for each DNA fragment (i.e., allele) is captured in an electropherogram. In the NGS workflow, there are some additional preparation steps (e.g., library preparation) after amplification. Once prepared, the sample is sequenced and the output includes both STR, SNP, and other forensic marker data. The NGS output also includes DNA sequence and read amount for each allele.

NGS workflow

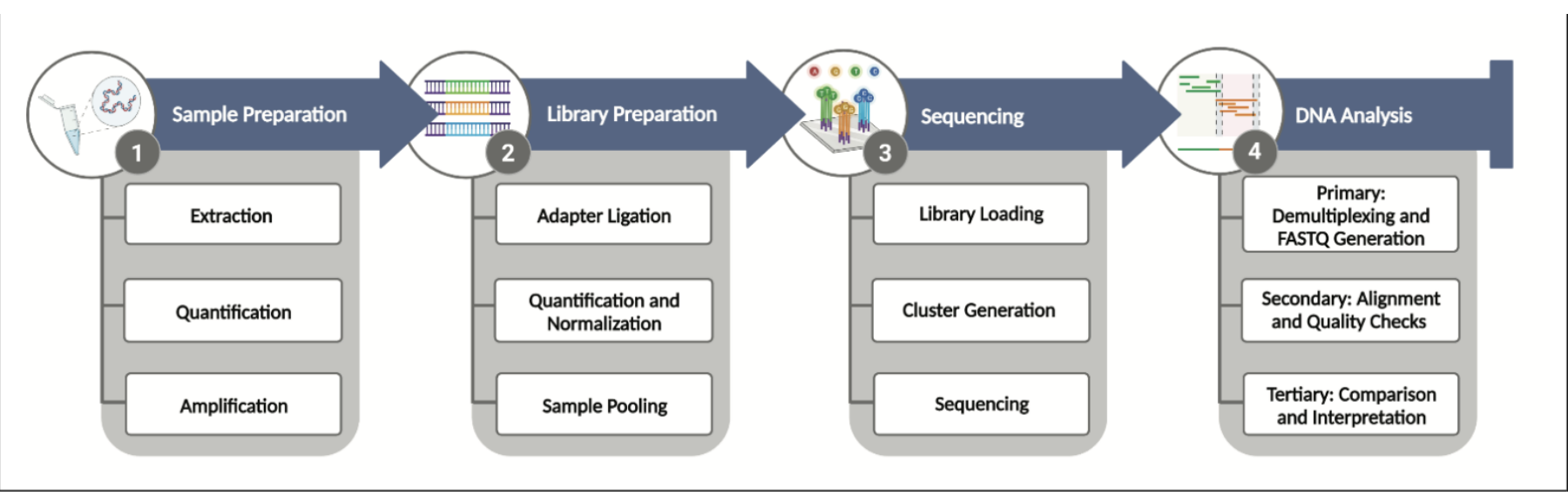


Exhibit 3. The four main phases and associated steps that comprise an NGS workflow.

NGS workflow image from Martin et al. [3]



Samples were prepared as with traditional DNA analysis but using specific NGS testing kits to amplify the DNA samples (left). The MiSeq FGx Sequencing System (right, image from [4]) was used to generate the NGS DNA data. NGS DNA data output included both autosomal STR data (select loci, bottom) and biallelic SNP markers.

SNP Mixture Study Design

Mixtures 1 (MX1)				
	709	710	711	712
501	M1:F1_1-1	M1:F1_1-39	M1:M2_1-9	M1:F1:M2_1-1-1
502	M1:F1_1-1	M1:F1_1-39	M1:M2_1-9	M1:F1:M2_1-4-4
503	M1:F1_1-3	M1:F1_1-99	M1:M2_1-19	M1:F1:M2_4-1-4
504	M1:F1_1-3	M1:F1_1-99	M1:M2_1-19	M1:F1:M2_4-4-1
505	M1:F1_1-9	M1:M2_1-1	M1:M2_1-39	M1:F1:M2:M3_1-1-1-1
506	M1:F1_1-9	M1:M2_1-1	M1:M2_1-39	M1:F1:M2:M3_1-4-4-4
507	M1:F1_1-19	M1:M2_1-3	M1:M2_1-99	POS1_MX1
508	M1:F1_1-19	M1:M2_1-3	M1:M2_1-99	NTC1_MX1

Mock 1 & Mixtures 2 (MO1_MX2)				
	701	702	703	704
501	POS1_MO1_MX2	Mock7	Mock15	M1:F1_1-1_500pg
502	NTC1_MO1_MX2	Mock8	Mock16	M1:F1_1-3_500pg
503	Mock1	Mock9	Mock17	M1:F1_1-9_500pg
504	Mock2	Mock10	Mock18	M1:F1_1-19_500pg
505	Mock3	Mock11	Mock19	M1:F1_1-1_250pg
506	Mock4	Mock12	Mock20	M1:F1_1-3_250pg
507	Mock5	Mock13	Mock21	M1:F1_1-9_250pg
508	Mock6	Mock14	NISTD	M1:F1_1-19_250pg

Table 1: ForenSeq DNA Signature Prep Kit—Forensic Loci

Feature	Number of Markers ^a	Amplicon Size Range (bp)	Included in DNA Primer Mix A	Included in DNA Primer Mix B ^b
Global Autosomal STRs	27	61–467	Yes	Yes
Y-STRs	24	119–390	Yes	Yes
X-STRs	7	157–462	Yes	Yes
Identity SNPs	94	63–231	Yes	Yes
Phenotypic SNPs	22	73–227	No	Yes
Biogeographical Ancestry SNPs	56	67–200	No	Yes

- SNP and STR chromosome locations can be found in the ForenSeq DNA Signature Prep Kit Reference Guide.
- Over 200 markers analyzed when running primer set B.

Identity Informative SNPs

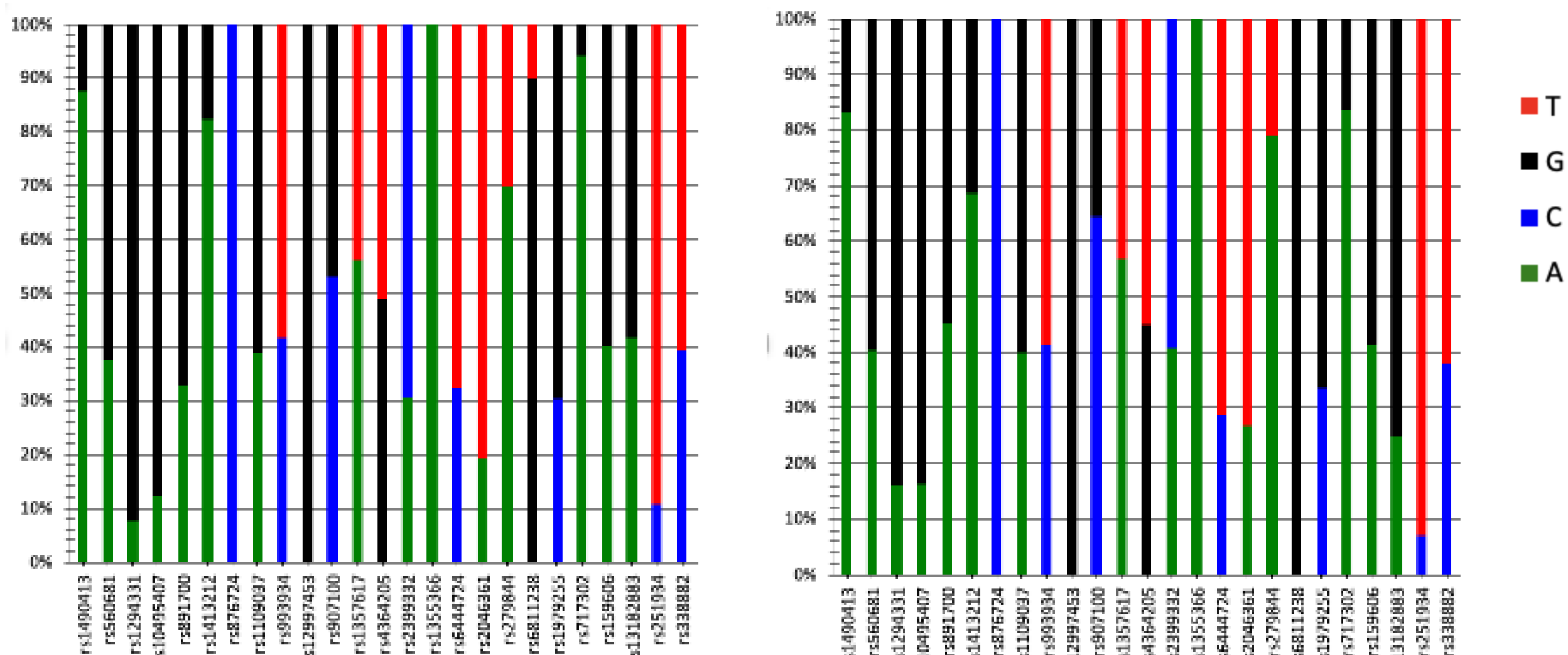
Locus	Amplicon Length (bp)	Chromosome	Amplicon Start Position	Amplicon End Position	2800M Control Alleles
rs10495407	109	1	238439234	238439342	GG
rs1294331	85	1	233448359	233448443	GA
rs1413212	64	1	242806767	242806830	GG
rs1490413	98	1	4367256	4367353	AA
rs560681	90	1	160786641	160786730	AG
rs891700	115	1	239881850	239881964	AG
rs1109037	118	2	10085691	10085808	GG
rs12997453	100	2	182413195	182413294	AA

94 Identity SNPs

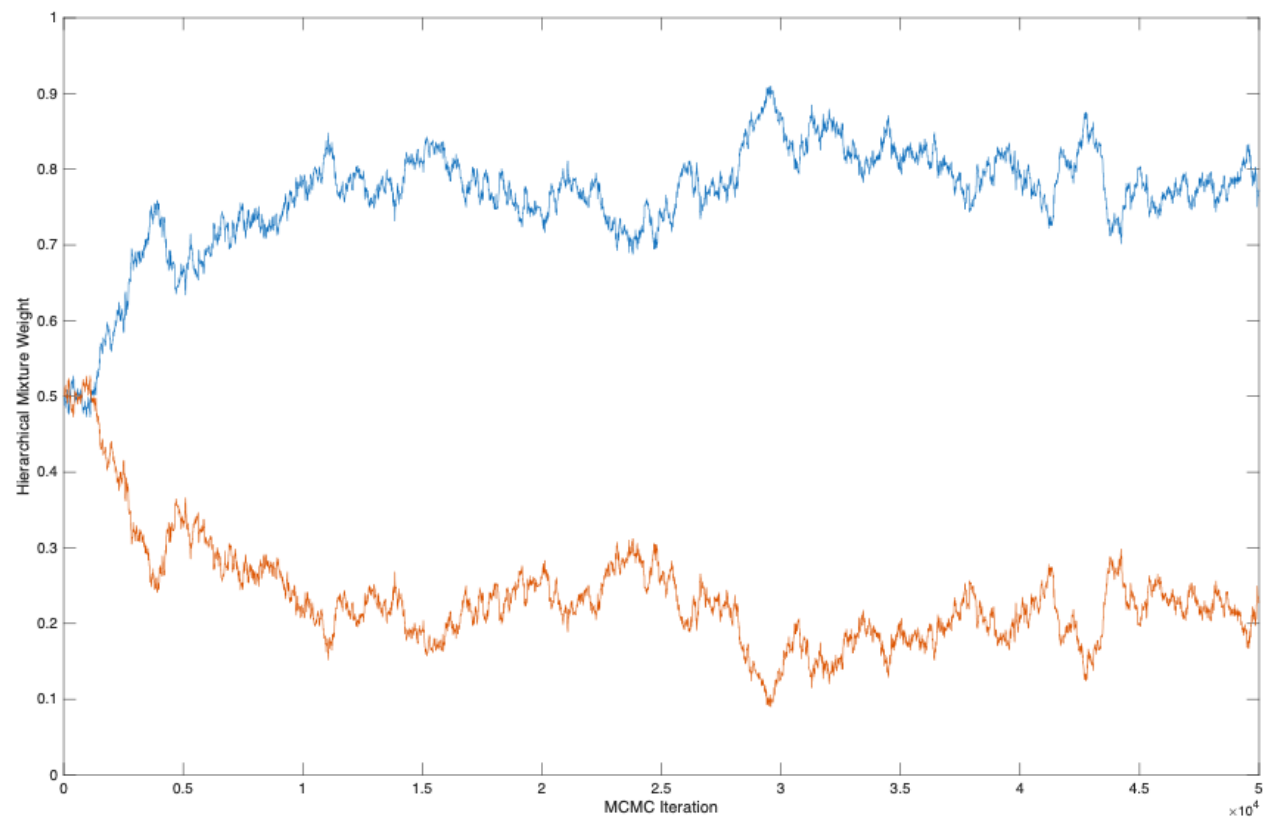
rs1523537	117	20	51296076	51296192	CC
rs445251	119	20	15124865	15124983	CG
rs221956	97	21	43606933	43607029	CC
rs2830795	114	21	28608089	28608202	AA
rs2831700	79	21	29679639	29679717	AA
rs722098	101	21	16685561	16685661	AG
rs914165	156	21	42415865	42416020	AG
rs1028528	78	22	48362256	48362333	AG
rs2040411	68	22	47836378	47836445	AA
rs731614	120	22	27816711	27816830	AG
rs987640	120	22	33559450	33559569	AT

SNP Data and Results

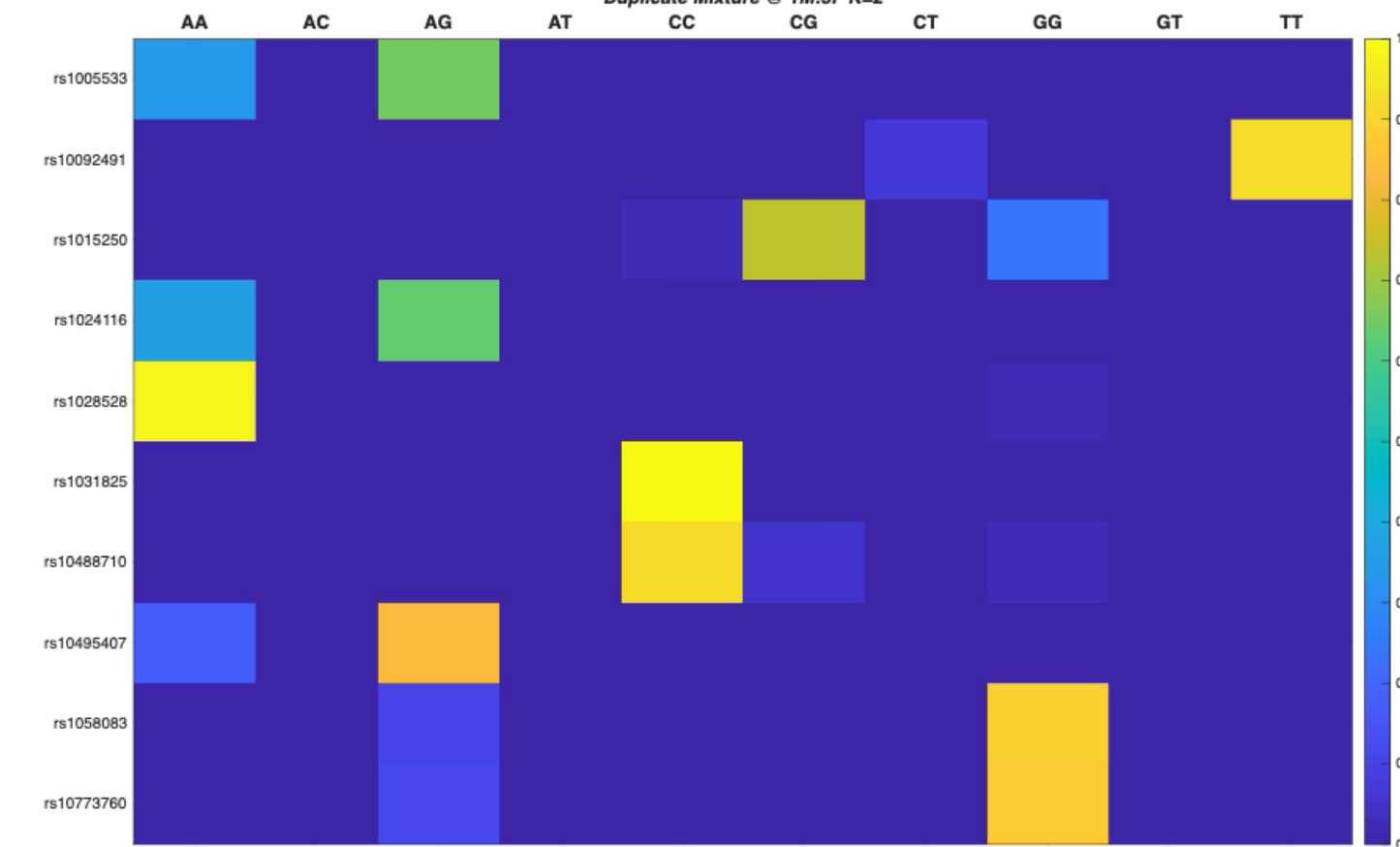
Two unknown DNA mixture contributors



Quantitative NGS data generated in duplicate from 94 biallelic SNP markers for a two-person 1:3 male:female DNA mixture.

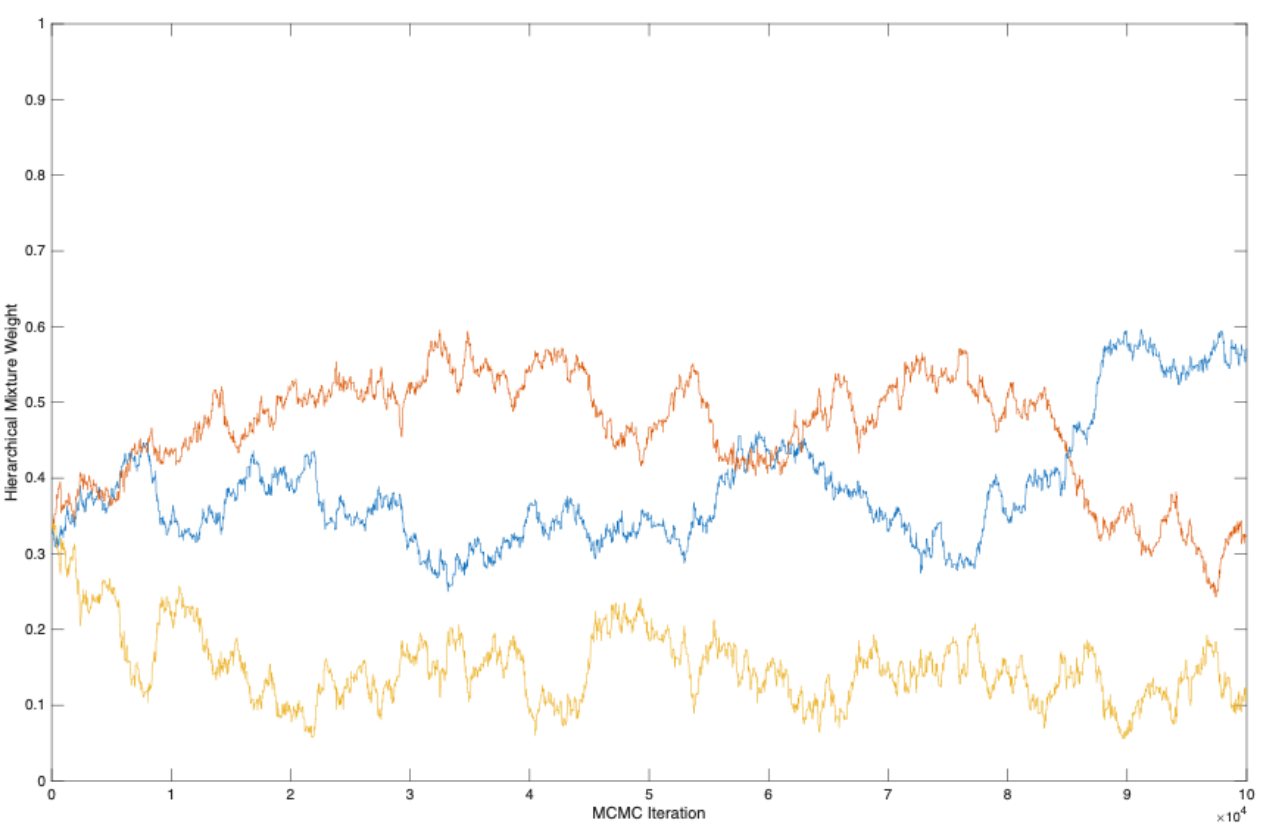


Mixture weight Markov chain from a 15-second 50,000-cycle MCMC that separated the minor 25% male component (orange) using an optimized TrueAllele® hierarchical Bayesian model.

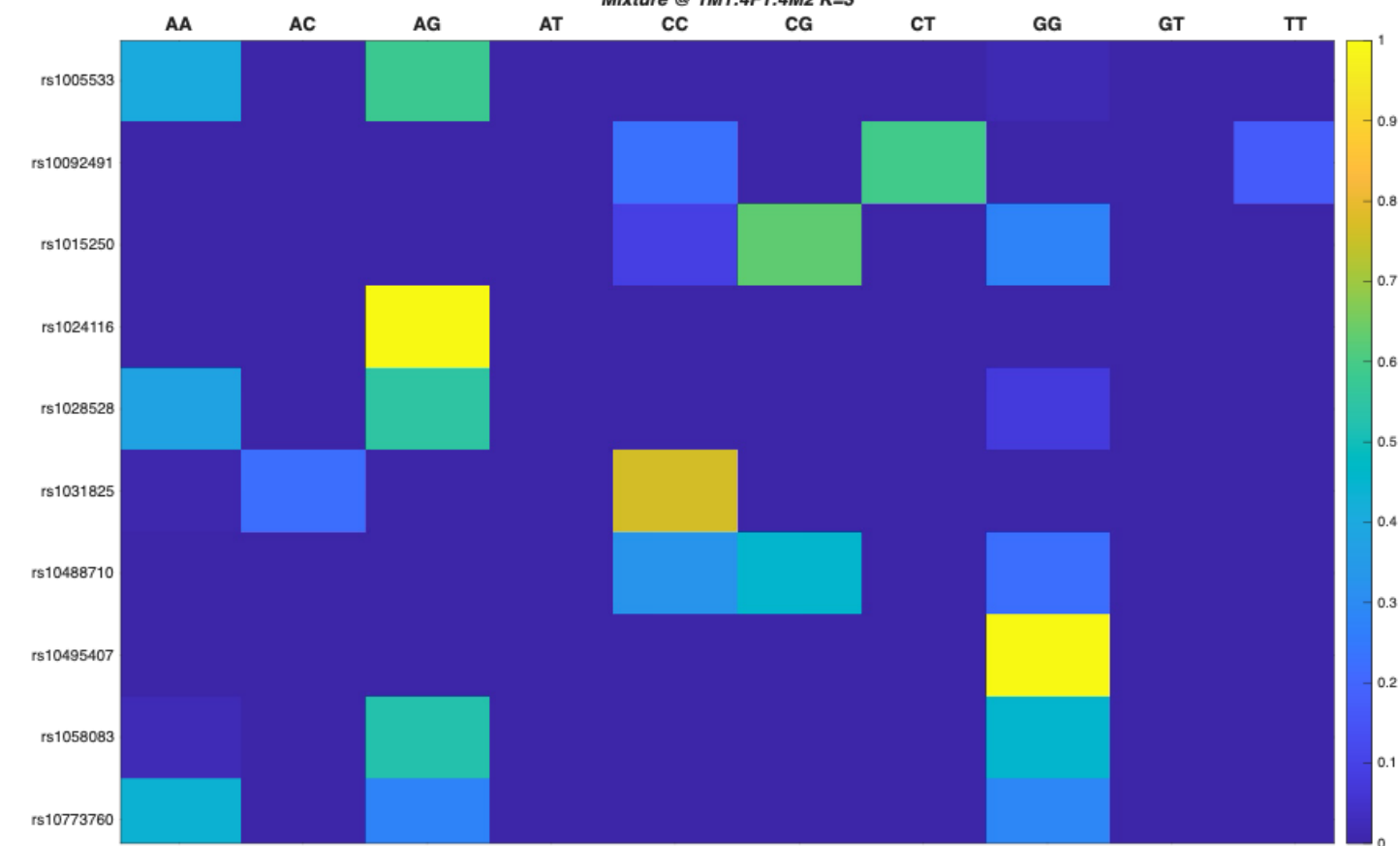


A heat map showing the resulting TrueAllele® biallelic genotype probabilities at 10 of 94 SNP loci for the 25% minor male contributor. The comparison LR was 10²⁰.

Three unknown DNA mixture contributors



Mixture weight Markov chain from a 30-second 100,000-cycle MCMC that separated a minor 44% male component (orange) from a 1:4:4 DNA mixture using a biallelic TrueAllele® hierarchical Bayesian probability model.



A heat map showing the resulting TrueAllele® biallelic genotype probabilities at 10 of 94 SNP loci for a minor male contributor. The comparison LR score was 10¹².

Conclusions

Mixtures of biallelic SNP loci repeatedly stack the same two alleles at a locus, presenting less allelic polymorphism than multi-allelic STR loci. Such allele redundancy can make mixture analysis challenging. This work showed that a hierarchical Bayesian model can rapidly and accurately separate quantitative NGS data from SNP mixtures into its component contributors. The resulting probabilistic genotypes preserve DNA information that can readily identify mixture contributors.

References

- Perlin, MW, and Szabady, B. Linear mixture analysis: a mathematical approach to resolving mixed DNA samples. *Journal of Forensic Sciences*, 46(6), pp. 1372-77, 2001.
- Perlin, MW and Sinelnikov, A. An information gap in DNA evidence interpretation. *PLoS ONE*, 4(12):e8327, 2009.
- Martin, M, Estus, E, Krauss, S, Sorrell, C, Shute, R, Ferrara, L, et al. Landscape Study of NGS Technologies for Forensic Applications. *National Institute of Justice*, 2023.
- Verogen. MiSeq FGx Sequencing System.