Journal of Forensic and Legal Medicine 103 (2024) 102678

Contents lists available at ScienceDirect

Journal of Forensic and Legal Medicine

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

Research Paper

Performance and characterization of 94 identity-informative SNPs in Northern Han Chinese using ForenSeq ™ DNA signature prep kit

Fei Guo [a,b,c,*], Ze Liu [d], Guannan Long [d], Biao Zhang [d], Dahua Liu [e], Shaobo Yu [d,**]

[a] Shenyang Medical College, Shenyang, Liaoning, 110034, PR China
[b] Key Laboratory of Human Ethnic Specificity and Phenomics of Critical Illness in Liaoning Province, Shenyang, Liaoning, 110034, PR China
[c] Key Laboratory of Phenomics in Shenyang City, Shenyang, Liaoning, 110034, PR China
[d] DNA Laboratory of Forensic Science Center, Shenyang Public Security Bureau, Shenyang, Liaoning, 110002, PR China
[e] Department of Forensic Medicine, Jinzhou Medical University, Jinzhou, Liaoning, 121001, PR China

ARTICLE INFO

Keywords:
Massively parallel sequencing (MPS)
Identity-informative single nucleotide polymorphism (iSNP)
ForenSeq ™ DNA Signature Prep Kit
MiSeq FGx ® Forensic Genomics System
Northern Han Chinese
Population genetics

ABSTRACT

Target and flanking region (FR) variation at 94 identity-informative SNPs (iSNPs) are investigated in 635 Northern Han Chinese using the ForenSeq DNA Signature Prep Kit on the MiSeq FGx Forensic Genomics System. The dataset presents the following performance characteristics (average values): ≥60% bases with a quality score of 20 or higher (%≥ Q20); >700 × of depth of coverage (DoC) from both Sample Details Reports and Flanking Region Reports; >80% of effective reads; ≥60% of allele coverage ratio (ACR); and ≥70% of inter-locus balance, while some stable low-performance characteristics are also observed: low DoC at rs1736442, rs1031825, rs7041158, rs338882, rs2920816, rs1493232, rs719366, and rs2342747; high noise at rs891700; and imbalanced ACR at rs6955448 and rs338882. The average amplicon length is 69 bp, suitable for detecting degraded samples. Bioinformatic concordance achieves 99.99% between the ForenSeq Universal Analysis Software (UAS) and the Integrative Genomic Viewer (IGV) inspection. Discordance results from flanking region deletions of rs10776839, rs8078417, rs2831700, and rs1454361. Due to FR variants within amplicons detected by massively parallel sequencing (MPS), the increases in the number of unique alleles, effective alleles (A_e), and observed heterozygosity (H_{obs}) are 46.81%, 4.51%, and 3.29%, respectively. Twelve FR variants are first reported to dbSNP, such as rs1252699848, rs1665500714, rs1771121532, rs2097285015, rs1851671415, rs2045669877, rs2046758811, rs2044248635, rs1251308240, rs1968822112, rs1981638299, and rs1341756746. All 94 iSNPs from target and amplicon data are in Hardy-Weinberg equilibrium (HWE) and independent within autosomes. As expected, forensic parameters from the amplicon data increase significantly on the combined power of discrimination (CPD = $1 - 3.9876 \times 10^{-38}$) and the combined power of exclusion (CPE = $1 - 6.6690 \times 10^{-8}$). Additionally, the power of the system effectiveness (CPD = $1 - 6.7054 \times 10^{-72}$ and CPE = $1 - 4.4719 \times 10^{-20}$) with sequence-based 27 autosomal STRs and 94 iSNP amplicons in combination is substantially improved compared to one type of marker alone. In conclusion, we have established a traditional length-based and current sequence-based reference database with 58 STRs and 94 iSNPs in the Northern Han Chinese population. We hope these data can serve as a solid reference and foundation for forensic practice.

1. Introduction

A single-base sequence variation between individuals at a specific target position in the genome is often called a single nucleotide polymorphism (SNP), which highly occurs in the *Homo sapiens* genome and commonly presents a bi-allelic nature (two alleles generating three genotypes) and a low mutation rate (~10^{-8}).[1] Due to their short amplicon length and high inheritance stability, identity-informative SNPs (iSNPs) can achieve a high detectable rate for degraded samples in personal identifications[2–4] and play an essential role in kinship analyses.[5–7] Besides, ancestry-informative SNPs (aSNPs) and phenotype-informative SNPs (pSNPs) can provide additional biogeographical ancestry and phenotype information predictions for crime scene investigations.[8,9] As the technology developed, massively parallel sequencing (MPS) has

* Corresponding author. Shenyang Medical College, Shenyang, Liaoning, 110034, PR China.
** Corresponding author. DNA Laboratory of Forensic Science Center, Shenyang Public Security Bureau, Shenyang, Liaoning, 110002, PR China.
E-mail addresses: nevergfaye@gmail.com (F. Guo), 243816450@qq.com (S. Yu).

https://doi.org/10.1016/j.jflm.2024.102678
Received 4 November 2023; Received in revised form 5 March 2024; Accepted 17 March 2024
Available online 21 March 2024

Table 1
Summary of characterization, allele frequencies, and forensic parameters of 94 iSNPs in the ForenSeq ™ DNA Signature Prep Kit (N = 635). Detailed information is listed in Tables S3, S5 and S8.

ForenSeq iSNP locus[a]	Amplicon length (bp)[b]	GRCh38 coordinates				Allele frequency at a target SNP				Forensic parameters of a target SNP		
		Chr	Amplicon start position	Amplicon end position	Target SNP position	Reference allele[c]	Alternative allele[c]	Reference frequency	Alternative frequency	PD	PE	*p*-HWE
rs1490413	55	1	4,307,219	4,307,273	4,307,263	G	A	0.4157	0.5843	0.6038	0.1995	0.1658
rs560681	48	1	160,816,871	160,816,918	160,816,880	A	G	0.6614	0.3386	0.6019	0.1300	0.2155
rs1294331	31	1	233,312,640	233,312,670	233,312,667	C (G)	T (A)	0.6811	0.3189	0.5821	0.1442	0.5777
rs10495407	67	1	238,275,955	238,276,021	238,276,008	G	A	0.6386	0.3614	0.6012	0.1608	0.7973
rs891700	72	1	239,718,572	239,718,643	239,718,626	A	G	0.4850	0.5150	0.6385	0.1620	0.1298
rs1413212	21	1	242,643,488	242,643,508	242,643,495	T (A)	C (G)	0.5630	0.4370	0.6205	0.1814	1.0000
rs876724	67	2	114,970	115,036	114,974	C	T	0.6047	0.3953	0.6027	0.1882	0.2443
rs1109037	78	2	9,945,582	9,945,659	9,945,593	G	A	0.5488	0.4512	0.6131	0.1995	0.3810
rs993934	70	2	123,351,571	123,351,640	123,351,637	A (T)	G (C)	0.5150	0.4850	0.6105	0.2113	0.2103
rs12997453	55	2	181,548,493	181,548,547	181,548,532	A	G	0.3882	0.6118	0.6106	0.1683	0.9330
rs907100	66	2	238,654,924	238,654,989	238,654,938	G	C	0.4551	0.5449	0.6252	0.1800	0.8074
rs1357617	69	3	920,039	920,107	920,099	A (T)	T (A)	0.2134	0.7866	0.5023	0.0822	0.7213
rs4364205	54	3	32,376,108	32,376,161	32,376,152	T	G	0.3693	0.6307	0.6102	0.1512	0.5510
rs2399332	110	3	110,582,177	110,582,286	110,582,279	T (A)	G (C)	0.3244	0.6756	0.5923	0.1300	0.4652
rs1355366	76	3	191,088,272	191,088,347	191,088,319	T (A)	C (G)	0.8496	0.1504	0.4154	0.0505	0.3511
rs6444724	71	3	193,489,543	193,489,613	193,489,591	T	C	0.6094	0.3906	0.6053	0.1800	0.4546
rs2046361	70	4	10,967,394	10,967,463	10,967,435	T (A)	A (T)	0.5701	0.4299	0.6132	0.1910	0.5134
rs279844	117	4	46,327,592	46,327,708	46,327,638	A	T	0.4488	0.5512	0.6145	0.1966	0.4746
rs6811238	64	4	168,742,417	168,742,480	168,742,464	T	G	0.3134	0.6866	0.5786	0.1431	0.4593
rs1979255	46	4	189,396,884	189,396,929	189,396,926	C (G)	G (C)	0.4858	0.5142	0.6265	0.1841	0.8746
rs717302	64	5	2,879,245	2,879,308	2,879,281	G	A	0.0945	0.9055	0.3008	0.0237	0.6416
rs159606	54	5	17,374,761	17,374,814	17,374,789	A	G	0.4197	0.5803	0.6292	0.1559	0.1895
rs13182883	126	5	137,297,586	137,297,711	137,297,649	G	A	0.5598	0.4402	0.6213	0.1814	1.0000
rs251934	55	5	175,351,637	175,351,691	175,351,675	A (T)	G (C)	0.8835	0.1165	0.3478	0.0311	0.8485
rs338882	116	5	179,263,618	179,263,733	179,263,724	G (C)	A (T)	0.4063	0.5937	0.6146	0.1747	0.9333
rs13218440	124	6	12,059,715	12,059,838	12,059,721	G	A	0.5961	0.4039	0.6211	0.1608	0.4606
rs1336071	59	6	93,827,491	93,827,549	93,827,537	T (A)	C (G)	0.4409	0.5591	0.6093	0.2024	0.2643
rs214955	72	6	152,376,519	152,376,590	152,376,571	C (G)	T (A)	0.4252	0.5748	0.6195	0.1774	1.0000
rs727811	65	6	164,624,247	164,624,311	164,624,301	G (C)	T (A)	0.3039	0.6961	0.5818	0.1178	0.3450
rs6955448	73	7	4,270,677	4,270,749	4,270,733	C	T	0.7016	0.2984	0.5723	0.1311	0.7057
rs917118	68	7	4,417,342	4,417,409	4,417,372	C	T	0.6984	0.3016	0.5781	0.1228	0.7111
rs321198	115	7	137,344,994	137,345,108	137,345,092	T	C	0.4480	0.5520	0.6101	0.2038	0.2633
rs737681	80	7	156,198,068	156,198,147	156,198,119	T	C	0.1591	0.8409	0.4283	0.0522	0.8823
rs763869	35	8	1,427,433	1,427,467	1,427,444	G (C)	A (T)	0.2969	0.7031	0.5727	0.1269	0.9241
rs10092491	65	8	28,553,546	28,553,610	28,553,555	T	C	0.3575	0.6425	0.5808	0.1938	0.0102[d]
rs2056277	57	8	138,386,821	138,386,877	138,386,873	C	T	0.8575	0.1425	0.4008	0.0451	0.6283
rs4606077	108	8	143,574,562	143,574,669	143,574,584	T	C	0.2819	0.7181	0.5628	0.1198	0.8450
rs1015250	68	9	1,823,726	1,823,793	1,823,774	G	C	0.5260	0.4740	0.6132	0.2053	0.3017
rs7041158	72	9	27,985,930	27,986,001	27,985,940	C	T	0.6252	0.3748	0.6101	0.1571	0.7994
rs1463729	58	9	124,119,138	124,119,195	124,119,169	C (G)	T (A)	0.5567	0.4433	0.6174	0.1896	0.6876
rs1360288	79	9	126,205,735	126,205,813	126,205,784	C	T	0.6472	0.3528	0.5968	0.1596	0.6659
rs10776839	63	9	134,525,445	134,525,507	134,525,462	G	T	0.5882	0.4118	0.6033	0.1980	0.1609
rs826472	102	10	2,364,342	2,364,443	2,364,437	T	C	0.2236	0.7764	0.5133	0.0877	0.7314
rs735155	128	10	3,331,961	3,332,088	3,331,986	C (G)	T (A)	0.2024	0.7976	0.4894	0.0748	0.9021
rs3780962	42	10	17,151,313	17,151,354	17,151,347	A (T)	G (C)	0.5693	0.4307	0.6126	0.1924	0.4640
rs740598	72	10	116,747,355	116,747,426	116,747,388	G	A	0.4339	0.5661	0.6255	0.1708	0.5748
rs964681	57	10	130,900,152	130,900,208	130,900,156	T	C	0.6724	0.3276	0.5898	0.1408	1.0000
rs1498553	64	11	5,687,776	5,687,839	5,687,798	C	T	0.4622	0.5378	0.6404	0.1535	0.0541
rs901398	45	11	11,074,649	11,074,693	11,074,674	C	T	0.2898	0.7102	0.5619	0.1386	0.1240
rs10488710	64	11	115,336,442	115,336,505	115,336,457	C (G)	G (C)	0.3087	0.6913	0.5833	0.1238	0.5212
rs2076848	77	11	134,797,628	134,797,704	134,797,652	A (T)	T (A)	0.6354	0.3646	0.6003	0.1658	0.6093
rs2107612	50	12	779,120	779,169	779,154	G	A	0.1346	0.8654	0.3832	0.0380	0.6086
rs2269355	23	12	6,836,737	6,836,759	6,836,750	C	G	0.4724	0.5276	0.6361	0.1645	0.1990
rs2920816	111	12	40,469,199	40,469,309	40,469,250	A (T)	G (C)	0.6614	0.3386	0.5992	0.1364	0.4788
rs2111980	52	12	105,934,430	105,934,481	105,934,476	T (A)	C (G)	0.6150	0.3850	0.6224	0.1420	0.1099
rs10773760	56	12	130,277,100	130,277,155	130,277,151	A	G	0.6276	0.3724	0.6235	0.1248	0.0076[d]
rs1335873	64	13	20,327,551	20,327,614	20,327,585	T (A)	A (T)	0.2811	0.7189	0.5594	0.1269	0.3325
rs1886510	70	13	21,800,531	21,800,600	21,800,561	G (C)	A (T)	0.8756	0.1244	0.3643	0.0347	0.8563
rs1058083	32	13	99,385,963	99,385,994	99,385,979	A	G	0.3976	0.6024	0.6211	0.1559	0.3582
rs354439	120	13	106,285,996	106,286,115	106,286,062	A (T)	T (A)	0.5724	0.4276	0.6014	0.2098	0.0850
rs1454361	75	14	25,381,580	25,381,654	25,381,626	T (A)	A (T)	0.5157	0.4843	0.6324	0.1734	0.4262
rs722290	56	14	52,749,993	52,750,048	52,750,005	G (C)	C (G)	0.4827	0.5173	0.6376	0.1633	0.1522
rs873196	71	14	98,379,189	98,379,259	98,379,194	C	T	0.1717	0.8283	0.4486	0.0607	0.5765
rs4530059	128	14	104,302,784	104,302,911	104,302,812	G	A	0.6921	0.3079	0.5787	0.1321	0.9259
rs1821380	76	15	39,021,164	39,021,239	39,021,201	C (G)	G (C)	0.6827	0.3173	0.5901	0.1228	0.2736
rs8037429	11	15	53,324,706	53,324,716	53,324,712	C	T	0.5283	0.4717	0.6173	0.1980	0.5231
rs1528460	65	15	54,918,491	54,918,555	54,918,507	C	T	0.4094	0.5906	0.6033	0.1966	0.1905
rs729172	51	16	5,556,181	5,556,231	5,556,196	G (C)	T (A)	0.8150	0.1850	0.4670	0.0652	1.0000
rs2342747	54	16	5,818,670	5,818,723	5,818,699	A	G	0.3110	0.6890	0.5794	0.1375	0.7102

(*continued on next page*)

Table 1 (*continued*)

ForenSeq iSNP locus[a]	Amplicon length (bp)[b]	GRCh38 coordinates				Allele frequency at a target SNP				Forensic parameters of a target SNP		
		Chr	Amplicon start position	Amplicon end position	Target SNP position	Reference allele[c]	Alternative allele[c]	Reference frequency	Alternative frequency	PD	PE	*p*-HWE
rs430046	68	16	77,983,107	77,983,174	77,983,154	C	T	0.6575	0.3425	0.6021	0.1353	0.3374
rs1382387	46	16	80,072,444	80,072,489	80,072,464	C (G)	A (T)	0.3205	0.6795	0.5932	0.1208	0.1659
rs9905977	119	17	3,016,058	3,016,176	3,016,099	A	G	0.3835	0.6165	0.6085	0.1683	0.8663
rs740910	64	17	5,803,260	5,803,323	5,803,303	A	G	0.9307	0.0693	0.2369	0.0146	0.3518
rs938283	53	17	79,472,374	79,472,426	79,472,416	T	C	0.8606	0.1394	0.3910	0.0389	0.4086
rs8078417	102	17	82,503,992	82,504,093	82,504,059	C	T	0.7094	0.2906	0.5706	0.1188	0.7735
rs1493232	28	18	1,127,969	1,127,996	1,127,985	C	A	0.6268	0.3732	0.6117	0.1523	0.5570
rs9951171	75	18	9,749,816	9,749,890	9,749,882	G	A	0.5150	0.4850	0.6040	0.2220	0.0670
rs1736442	103	18	57,558,486	57,558,588	57,558,545	T (A)	C (G)	0.3701	0.6299	0.6052	0.1620	0.9325
rs1024116	49	18	77,720,385	77,720,433	77,720,430	C (G)	T (A)	0.9157	0.0843	0.2770	0.0203	0.2989
rs719366	120	19	27,972,398	27,972,517	27,972,429	G (C)	A (T)	0.2142	0.7858	0.5033	0.0784	0.8156
rs576261	29	19	39,069,160	39,069,188	39,069,167	A	C	0.5827	0.4173	0.6093	0.1910	0.4116
rs1031825	77	20	4,466,793	4,466,869	4,466,836	A	C	0.4598	0.5402	0.6254	0.1814	0.8737
rs445251	68	20	15,144,244	15,144,311	15,144,287	G (C)	C (G)	0.3409	0.6591	0.5938	0.1512	0.7910
rs1005533	116	20	40,858,446	40,858,561	40,858,470	G	A	0.6575	0.3425	0.6021	0.1353	0.3347
rs1523537	70	20	52,679,563	52,679,632	52,679,623	T	C	0.5677	0.4323	0.6193	0.1814	0.9340
rs722098	49	21	15,313,267	15,313,315	15,313,279	A	G	0.4512	0.5488	0.6309	0.1683	0.3726
rs2830795	68	21	27,235,795	27,235,862	27,235,844	A	G	0.5205	0.4795	0.6129	0.2068	0.3014
rs2831700	34	21	28,307,343	28,307,376	28,307,368	A	G	0.4929	0.5071	0.6109	0.2113	0.2066
rs914165	108	21	41,043,962	41,044,069	41,044,003	G	A	0.6709	0.3291	0.5929	0.1364	0.7194
rs221956	52	21	42,186,845	42,186,896	42,186,887	T	C	0.4142	0.5858	0.6279	0.1547	0.1894
rs733164	80	22	27,420,770	27,420,849	27,420,823	G	A	0.8811	0.1189	0.3508	0.0306	0.4458
rs987640	75	22	33,163,486	33,163,560	33,163,522	T	A	0.4811	0.5189	0.6307	0.1761	0.5223
rs2040411	17	22	47,440,656	47,440,672	47,440,662	G	A	0.7740	0.2260	0.5163	0.0837	0.7335
rs1028528	36	22	47,966,528	47,966,563	47,966,541	A	G	0.6449	0.3551	0.5963	0.1633	0.5431

*Chr: chromosome; PD: power of discrimination; PE: power of exclusion; *p*-HWE: *p* value for Hardy-Weinberg equilibrium test.

[a] Underline: The target iSNP genotype and its amplicon string at a locus are named by reverse strands in ForenSeq ™ Universal Analysis Software (ForenSeq UAS) but changed to forward strands in this study.

[b] The amplicon length does not include primers and adaptors.

[c] (Brackets): The genotype is given by ForenSeq UAS on the reverse strand to aid current users.

[d] The locus shows a statistically significant departure from the HWE expectation ($p < 0.05$), but it met the HWE expectation after the Bonferroni correction ($\alpha = 0.000532$).

increasingly been applied in the forensic community.[10] At the same time, manufacturers have already launched commercial kits containing SNPs for forensic applications, such as the Precision ID Identity Panel and the Precision ID Ancestry Panel from Thermo Fisher Scientific,[3,11] the ForenSeq ™ DNA Signature Prep Kit, the ForenSeq ™ Kintelligence Kit, and the ForenSeq ™ Imagen Kit from Qiagen,[12–14] and the MGIEasy Signature Identification Library Prep Kit from MGI Tech.[15]

Among them, the ForenSeq ™ DNA Signature Prep Kit (hereafter referred to as "ForenSeq Kit") on the MiSeq FGx ® Forensic Genomics System (hereafter referred to as "MiSeq FGx") allows a targeted amplification of Amelogenin, 58 short tandem repeats (STRs), and 94 iSNPs using the DNA Primer Mix A (DPMA), with an option to add another 22 pSNPs and 56 aSNPs using the DNA Primer Mix B (DPMB).[12] With those iSNPs, some population genetic datasets have been established abroad up to now, such as Yavapai Native American,[16] U.S. African American, Caucasian, East Asian, and Hispanic,[17,18] Spanish,[19] Dane,[20] Saudi Arabian,[21] French,[22] Nigerian,[23] Peruvian,[24] El Salvadorian,[25] Mexican,[26] and UK White British, East Asian, South Asian, North-East African and West African,[27] and in China, such as Tibetan,[28] Hui,[29] Li,[30] and Uyghur.[31] However, such a dataset from Han Chinese is not available.

In the previous study, we reported the sequence variation, allele frequencies, and forensic parameters for 58 STRs included in ForenSeq Kit from 635 Northern Han Chinese (NHC 635).[32] For this study, the target and flanking region variation, allele frequencies, and population statistics were generated for 94 iSNPs for the same dataset. Additionally, the MPS performance of these iSNPs was evaluated, such as the depth of coverage, sequence coverage ratio, average coverage ratio, and inter-locus balance. Further, the power of the system effectiveness with the combination of 27 autosomal STRs (A-STRs) and 94 iSNPs was herein assessed for personal identifications and kinship analyses.

2. Materials and methods

2.1. Sample-to-Profile

All materials and methods, from samples to profiles, were described previously.[32] Briefly, quantified NHC 635 samples were amplified using the DPMA in the ForenSeq ™ DNA Signature Prep Kit (Qiagen, Hilden, Germany) and sequenced on the MiSeq FGx ® Forensic Genomics System (Qiagen). MPS raw data were processed by the ForenSeq ™ Universal Analysis Software (hereafter referred to as "ForenSeq UAS") v1.3 (Qiagen) at default analysis thresholds.

2.2. Data interpretation

The target iSNP variants were manually interpreted on the interface of ForenSeq UAS, mainly when quality control (QC) indicators were triggered at a specific locus, and then exported into a *Project Genotype Report*. The flanking region (FR) variants were found in a *Flanking Region Report*. In order to authenticate GRCh38 coordinates and decipher primer binding sites, FASTQ.GZ files exported from ForenSeq UAS had been remapped to the hg38 human reference genome and converted to binary alignment map (BAM) and binary alignment index (BAI) files in our previous study.[32] These target iSNP and their FR variants were parallelly investigated using the Integrative Genomic Viewer (IGV) package v.2.4.8[33] with an in-house hotspot BED file.[4]

In *Project Genotype Reports* and *Flanking Region Reports*, 33 target iSNPs were named by reverse strands (Table 1), such as rs1294331, rs1413212, rs993934, rs1357617, rs2399332, rs1355366, rs2046361, rs1979255, rs251934, rs338882, rs1336071, rs214955, rs727811, rs763869, rs1463729, rs735155, rs3780962, rs10488710, rs2076848, rs2920816, rs2111980, rs1335873, rs1886510, rs354439, rs1454361,

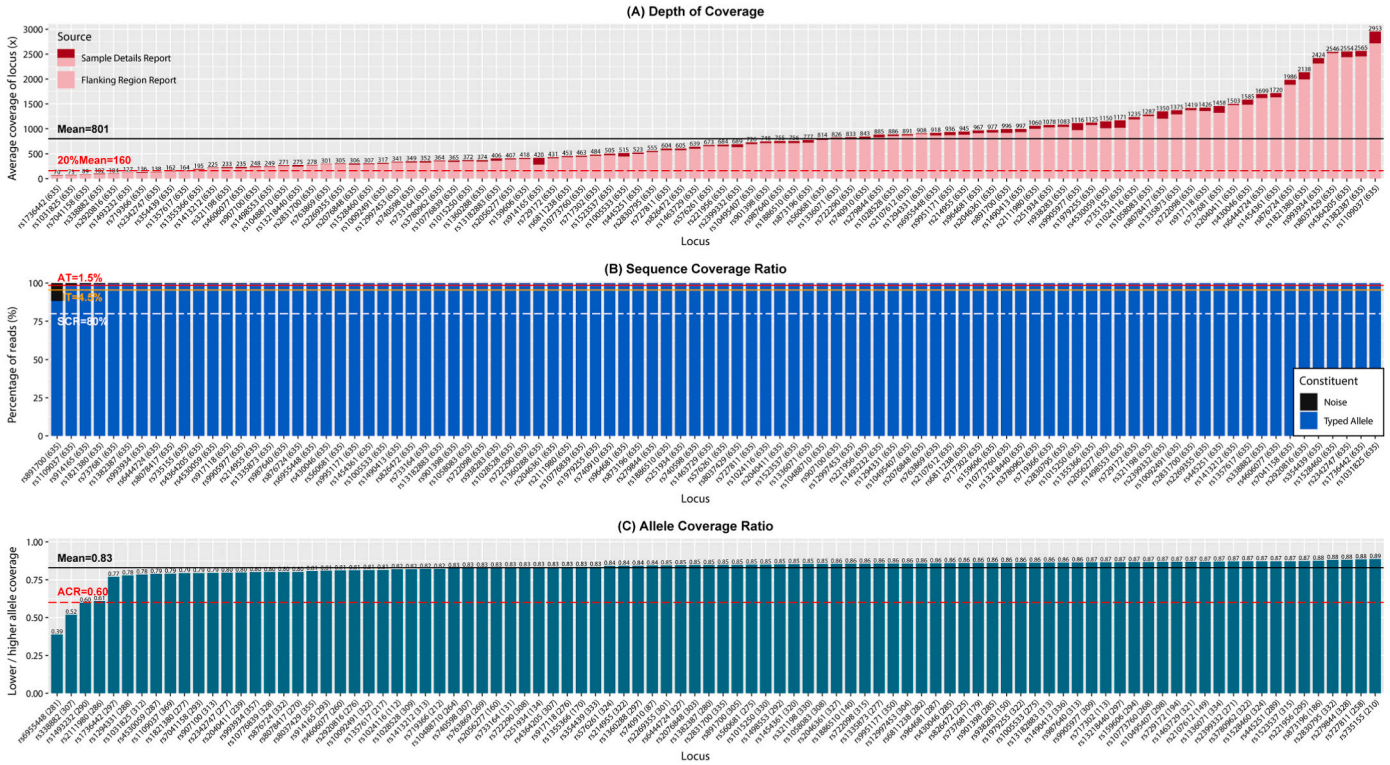

Fig. 1. MPS performance. **(A)** The depth of coverage (DoC) panel shows the average coverage from lowest to highest across 94 iSNPs from *Sample Details Reports* and *Flanking Region Reports*. The horizontal black solid line indicates the mean DoC = 801 × across 94 loci, and the horizontal red dashed line indicates 20% of the mean DoC = 160 × . The number in a bracket on the *X*-axis labels the size of the samples, where 635 samples are calculated at all loci. **(B)** The sequence coverage ratio (SCR) panel displays the average percentage of typed alleles and noise at each locus, where 94 iSNPs from *Flanking Region Reports* are arranged in order of the lowest to highest percentage of typed alleles. The horizontal red and orange solid lines indicate the analytical (1.5%) and interpretation (4.5%) thresholds, respectively, recommended in the ForenSeq Universal Analysis Software. The horizontal white dash line indicates SCR = 80%. The number in a bracket on the *X*-axis labels the size of the samples, where 635 samples are calculated at all loci. **(C)** The allele coverage ratio (ACR) panel provides the average ratio of lower allele coverage to higher allele coverage from lowest to highest across 94 iSNPs from *Sample Details Reports*. The horizontal black solid line indicates the mean ACR = 0.83, and the horizontal red dashed line indicates the recommended ACR threshold (0.60). The number in a bracket on the *X*-axis labels the number of heterozygotes. Details of DoC, SCR, and ACR are listed in Table S2. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

rs722290, rs1821380, rs729172, rs1382387, rs1736442, rs1024116, rs719366, and rs445251. According to the International Society for Forensic Genetics (ISFG) guidelines[34] and the Database of Single Nucleotide Polymorphisms (dbSNP) nomenclature,[35] target iSNP genotypes and their amplicon strings should be changed to forward strands carefully, especially at these markers with the complementary base transversion, e.g., reference allele [A] (reverse nomenclature) → reference allele [T] (forward nomenclature) and alternative allele [T] (reverse nomenclature) → alternative allele [A] (forward nomenclature) at rs354439, rs1357617, rs2046361, rs2076848, rs1335873, and rs1454361, and [C] (reverse) → [G] (forward) / [G] (reverse) → [C] (forward) at rs1979255, rs10488710, rs1821380, rs722290, and rs445251. Any FR variants that were not pre-determined in the *Flanking Region Report* were identified on the Ensembl Genome Browser (https://ensembl.org/Homo_sapiens/Info/Index). Further, any FR variants not found on the Ensembl were submitted to the dbSNP to assign a Submitted SNP (ss) number and then a Reference SNP (rs) number after authentication. Insertions and deletions (indels) positioned in a polymeric sequence tract were placed as insertions or deletions starting at the most 5' nucleotide coordinates.[36] The comprehensive nomenclature system was adopted as "Sample ID"_"Locus"_"Target Allele"_"Flanking Variant (rs number [Alternative Allele])" to capture the majority of genetic information in the amplicon string. The short nomenclature system was adjusted as a target allele followed by additional numerals for FR variants to facilitate population genetic and forensic parameters calculations. Finally, homozygous amplicons in the *Flanking Region Report* were filtered, and their single rows were replicated.

2.3. Statistical analysis

Depth of coverage (DoC) was defined by summing all reads within the locus, where the reads were extracted from the *Sample Details Report* generated by ForenSeq UAS. Sequence coverage ratio (SCR), including % allele and % noise, was calculated by dividing reads for typed alleles (also known as true alleles or effective reads) and noise (i.e., non-specific alleles or reads that belonged to neither reference alleles nor alternative alleles) by DoC. Allele coverage ratio (ACR), also known as heterozygote balance or intra-locus balance, was measured as a ratio of lower allele coverage to higher allele coverage at a locus. Inter-locus balance was assessed as the proportion of loci whose DoC exceeded 20% of the average DoC across all loci.

The observed heterozygosity (H_{obs}), expected heterozygosity (H_{exp}) or genetic diversity (GD), polymorphism information content (PIC), match probability (MP), power of discrimination (PD), power of exclusion (PE), and typical paternity index (TPI) were computed with the STRAF software v1.0.5.[37] Hardy-Weinberg equilibrium (HWE) and linkage disequilibrium (LD) were tested using the Arlequin ver 3.5.2.2.[38] The effective number of alleles (A_e) was calculated according to Kidd and Speed[39]: $A_e = 1/\sum p_i^2$, where p_i is the frequency of the *i*th allele. All parameters were calculated for target iSNPs and their amplicons.

One-way analysis of variance (ANOVA) was computed using R software version 4.0.5,[40] and figures were generated by Package "ggplot2" for R.

2.4. Sanger sequencing

Sanger sequencing was employed to confirm primer-binding site mutations observed in realignment results in Section 2.2. Primers are listed in Table S1 for PCR amplification and two-directional sequencing. Sanger sequencing was performed on the Applied Biosystems ® 3130*xl* Genetic Analyzer (Thermo Fisher Scientific, MA, USA) using the BigDye ® Terminator v3.1 Cycle Sequencing Kit (Thermo Fisher Scientific) following the manufacturer's instructions. Raw data were analyzed using the Sequencing Analysis Software v5.3.1 (Thermo Fisher Scientific).

3. Results and discussion

3.1. Run metrics

Run quality metrics measured by ForenSeq UAS and extracted from Sequencing Analysis Viewer Software (SAV) were reported in detail in the previous study (Table S2 in Ref. 32). Generally, the mean ± SD across 20 runs were calculated as (1367 ± 90) K/mm^2 for cluster density, (89.59 ± 1.62) % for clusters passing filter (PF), (0.128 ± 0.011) % for phasing, (0.090 ± 0.017) % for pre-phasing, (5.7 ± 0.3) G for yield total, (16.15 ± 1.03) M for reads, (14.46 ± 0.75) M for PF reads, (70.47 ± 1.28) % for %bases ≥ Q20 and (62.12 ± 1.30) % for %bases ≥ Q30, respectively. The outcome of run metrics showed that all runs yielded enough high-quality data for secondary analyses.

3.2. MPS performance

MPS performance of 94 targeted iSNPs in ForenSeq Kit was evaluated by examining the depth of coverage (DoC), sequence coverage ratio (SCR), allele coverage ratio (ACR), and inter-locus balance. DoC and ACR enrolled all reads exceeding the default analytical threshold (10 ×) reported in the *Sample Details Report*. SCR considered all reads of alleles and noise reported in the *Flanking Region Report* because noise reads are barely reported in the *Sample Details Report*. Fig. 1 displays the average DoC, SCR, and ACR at each locus from 635 samples. Details of performance metrics are listed in Table S2.

3.2.1. Depth of coverage and inter-locus balance

DoC was calculated as 801 × ± 632 × (mean ± SD) across 94 loci, ranging from the lowest (70 ×) at rs1736442 to the highest (2953 ×) at rs1109037 in Fig. 1A and Table S2. However, 47.87% (45/94) of iSNPs had an average DoC lower than the minimum requirement of 650 reads. This minimum requirement is used for determining the analytical (1.5% × 650 = 10 reads) and interpretation (4.5% × 650 = 30 reads) thresholds for the locus. Additionally, DoC extracted from the *Flanking Region Report* was calculated as 754 × ± 601 × . This lower DoC was expected given all bases on an amplicon string being analyzed rather than a single base on a target, which was also reported and well explained by King et al.[41] Nevertheless, DoCs had no significant difference between the *Flanking Region Report* and *Sample Details Report* (p = 0.6045, one-way ANOVA).

As a defined threshold of inter-locus balance, 20% of the average DoC (801 ×) was calculated as 160 × across 94 loci. A total of 86 iSNPs (91.49%) passed the threshold, while only eight loci did not meet the criteria in Fig. 1A. These eight loci were also identified as poorly performing markers in previous studies: rs1736442 (70 ×),[4,6,17–19,21,23,25,27,29,30] rs1031825 (73 ×),[4,6,17–19,21,23,25,27,29,30] rs7041158 (89 ×),[4,6,17–19,25,27,29,30] rs338882 (102 ×),[19,29] rs2920816 (104 ×),[6,17–19,21,25,27,29,30] rs1493232 (127 ×),[6,30] rs719366 (136 ×),[4,17–19,21,27,29] and rs2342747 (138 ×).[4,6] The inter-locus balance was generally within an acceptable range (≥70%).

In a large multiplex system, the cross-hybridization of primers may lead to a consistent reduction in DoCs at loci. The deciphered forward and reverse primer sequences of 58 STRs and 94 iSNPs were screened for potential primer-dimer and intramolecular hairpin formation using AutoDimer v1.0.[42] Screening results in Fig. S1 indicated significant complementarity or self-complementarity with the score of ≥7 at those eight loci, especially nearby the 3' end of a forward (F) or reverse (R) primer (a stretch of 5 or more uninterrupted Watson-Crick base pairs), such as **rs1736442_R** vs. DYS439_R, **rs1031825_F** vs. rs1979255_F, **rs7041158_F** vs. rs873196_F, **rs338882_R** vs. DYS392_F, **rs2920816_F** vs. DXS7423_F, **rs1493232_F** vs. rs7041158_R, **rs719366_F** vs. **rs719366_F** (self-complementarity), and **rs2342747_R** vs. rs722290_R.

Interestingly, remapping results revealed one mutation (rs6076682 [C > T]) was called at the 63rd nucleotide upstream from rs1031825 within the forward primer-binding region and two mutations (rs2342748 [G > C] and rs73514221 [G > A]) were at the 29th and 43rd nucleotides downstream from rs2342747 within the reverse primer-binding region. Further investigation indicated that the number ratios of the reference allele [C] vs. the alternative allele [T] at rs6076682 and [G] vs. [A] at rs73514221 were balanced (1.02:1 and 1.01:1, respectively). However, these ratios were calculated as 1:0 at rs6076682 in Asian (HapMap) and 11.95:1 at rs73514221 in East Asian (gnomAD) on dbSNP Build 156. At rs2342748, the alternative allele [C] dominated with 99.94%, but it accounted for 71.76% in East Asian and 65.74% in Global (gnomAD). King et al.[42] deduced this mutation might play a role in the reduced abundance of amplicon strings containing rs2342747 [G] that was phased with rs2342748 [C]. The primer-binding site mutation can destabilize primer annealing, leading to inefficient amplification that can result in underperformed DoC and/or imbalanced ACR. However, these absolutely balanced or complete mutations within the primer-binding region from realignment outcomes on a population level are more likely to be caused by designed degenerate and/or 1-mismatched primers rather than all actual mutations from each individual. Sanger sequencing results confirmed this assumption (see Section 3.3).

3.2.2. Sequence coverage ratio

A very small number of noise reads were only observed at rs1109037, rs6444724, rs1360288, rs964681, rs722290, rs1821380, rs8037429, rs1382387, rs8078417, rs9951171, and rs722098 in the *Sample Details Report* in this study. As shown in Fig. 1B and Table S2, % typed allele was averaged to 99.75%, with the lowest (88.20%) observed at rs891700 and the highest (100.00%) observed at 12 loci (rs1413212, rs1357617, rs338882, rs4606077, rs7041158, rs2920816, rs354439, rs1528460, rs2342747, rs1736442, rs1031825, and rs445251) from the Flanking R*egion Report*. The highest % noise (11.80%) was detected at rs891700, also reported by King et al.[41] and Li et al.,[43] which exceeded the recommended analytical and interpretation thresholds by ForenSeq UAS. The reason is due to the amplicon for rs891700 containing a 10-T homopolymer, where a small number of chemistry-related errors (9Ts, 11Ts, or 12Ts) accumulate, resulting in more noise reads (due to misinterpretations) and fewer typed allele reads within the amplicon. Empirically, typed alleles of all samples can be distinguished from noise in this study when % typed allele is more than 80% for ForenSeq iSNP markers.

3.2.3. Allele coverage ratio

ACR averaged 0.83 ± 0.07 for 94 loci from the lowest (0.39) at rs6955448 to the highest (0.89) at rs735155 in Fig. 1C and Table S2. All average ACRs were above 0.60 except rs6955448 and rs338882, which were also observed as the most imbalanced loci in recent literature.[4,6,17–19,27,41,43,44]

The locus rs6955448 had lower average coverage for the alternative allele [T] (220 × ± 90 ×) compared to the reference allele [C] (576 × ± 214 ×), leading to an average ACR of 0.39 ± 0.11 from 281 heterozygotes. Out of all the heterozygotes, only six had an ACR higher than 0.60. Sanger sequencing results showed that a mutation (rs6955464 [C > T]) was present at the 32nd nucleotide downstream from target rs6955448 within the reverse primer-binding region, which was

Table 2
Flanking region (FR) variants within iSNP amplicons observed in this study (N = 635). Detailed information is listed in Table S6, and IGV screenshots of variants submitted to dbSNP are shown in Fig. S3.

iSNP locus	Target reference (Ref) or alternative (Alt) allele	Target Allele	FR direction	FR variant rs number	GRCh38 position	Variant [Ref > Alt]	Alternative allele frequency	ss number on dbSNP (Submitter: GF)
rs1490413	Ref	G	Up	rs183248592[a]	chr1:4,307,228	C > G	0.0047	
	Alt	A	Up	rs1252699848[a]	chr1:4,307,247	A > T	0.0055	ss2137544118[e]
rs891700	Alt	G	Up	**rs12047255**[b]	chr1:239,718,578	G > A[d]	0.0992	
	Alt	G	Up	**rs12047255**	chr1:239,718,578	G > C[d]	0.0008	
	Ref/Alt	A/G	Up	rs543563536[a]	chr1:239,718,609[c]	insT	0.0016	
	Alt	G	Down	rs552859823	chr1:239,718,634	A > G	0.0024	
rs876724	Ref	C	Down	**rs77642176**	chr2:114,982	G > C	0.0677	
	Ref	C	Down	rs1665500714	chr2:115,020	T > A	0.0008	ss4035869449[e]
	Ref	C	Down	**rs300773**[b]	chr2:115,035	C > T	0.4591	
rs1109037	Ref	G	Down	rs550109468[a]	chr2:9,945,614	C > T	0.0008	
	Ref	G	Down	**rs183533496**[b]	chr2:9,945,624	C > T	0.0055	
	Ref/Alt	G/A	Down	**rs1109038**[b]	chr2:9,945,657	G > A	0.3961	
rs993934	Ref	A	Up	rs1573442732[a]	chr2:123,351,588[c]	insT	0.0008	ss4035869450
rs12997453	Ref	A	Up	**rs72883670**	chr2:181,548,511	C > T	0.2189	
rs907100	Alt	C	Down	**rs11689319**	chr2:238,654,956	G > A	0.1126	
rs2399332	Alt	G	Up	**rs2399334**[b]	chr3:110,582,178	C > T	0.6354	
	Alt	G	Up	**rs2399333**	chr3:110,582,215	G > T	0.6339	
rs279844	Alt	T	Down	**rs279845**[b]	chr4:46,327,706	T > A	0.5512	
rs6811238	Alt	G	Down	rs535583485[a]	chr4:168,742,474	G > A	0.0008	
rs1979255	Alt	G	Up	rs1259223242[a]	chr4:189,396,909	A > C	0.0024	ss4035869451
rs159606	Ref	A	Down	rs576005112	chr5:17,374,810	C > A	0.0008	
rs13182883	Alt	A	Up	rs1561496729	chr5:137,297,592	G > A	0.0008	ss4035869452[f]
	Ref	G	Up	rs1220236942[a]	chr5:137,297,605	T > C	0.0031	
rs251934	Ref	A	Down	rs1757388778	chr5:175,351,681	T > A	0.0008	ss4035869453[f]
rs338882	Ref	G	Up	**rs909522201**	chr5:179,263,639	C > T	0.0008	ss4035869454
rs1336071	Alt	C	Down	rs1771121532[a]	chr6:93,827,546	T > G	0.0008	ss4035869458[e]
rs214955	Ref	C	Up	rs1456697107	chr6:152,376,554	A > G	0.0008	
	Ref	C	Down	rs2097285015	chr6:152,376,572	G > A	0.0008	ss4035869459[e]
rs727811	Ref/Alt	G/T	Up	**rs1390470**	chr6:164,624,257	C > T	1.0000	
rs6955448	Ref/Alt	C/T	Up	**rs6950322**	chr7:4,270,685	G > A	0.3008	
rs737681	Alt	C	Up	rs1039895061	chr7:156,198,101	T > C	0.0008	ss4035869461[f]
rs4606077	Ref	T	Down	**rs58774517**[b]	chr8:143,574,594	C > T	0.1339	
	Alt	C	Down	**rs1869434**	chr8:143,574,595	G > A	0.7181	
	Alt	C	Down	rs534942109[a]	chr8:143,574,657	C > T	0.0071	
	Alt	C	Down	rs1160609366[a]	chr8:143,574,662	T > C	0.0008	
rs1015250	Alt	C	Up	**rs6475200**	chr9:1,823,749	A > G	0.4709	
rs10776839	Ref/Alt	G/T	Up	**rs7037930**[b]	chr9:134,525,459	A > G	0.7512	
	Alt	T	Down	rs542545139[a]	chr9:134,525,492[c]	delG	0.0008	
rs735155	Alt	T	Up	rs543475735[a]	chr10:3,331,975	T > G	0.0008	
rs740598	Ref	G	Up	rs1851671415[a]	chr10:116,747,373	T > A	0.0008	ss4035869462[e]
rs964681	Ref	T	Down	rs558108170	chr11:568,779	G > A	0.0008	
rs2076848	Ref	A	Up	**rs7947725**	chr11:134,797,630	C > T	0.0354	
rs2920816	Alt	G	Up	**rs142684512**	chr12:40,469,224	T > G	0.0016	
	Ref	A	Down	rs552576076	chr12:40,469,282	G > T	0.0008	
rs2111980	Alt	C	Up	rs1592813067[a]	chr12:105,934,431	G > T	0.0008	ss4035869464
rs1335873	Alt	A	Up	rs1011395106[b]	chr13:20,327,555	G > A	0.0008	ss4035869465
	Alt	A	Down	rs974633239[a]	chr13:20,327,590	G > A	0.0008	ss4035869466
rs1886510	Ref	G	Up	rs1593129468[a]	chr13:21,800,548	C > A	0.0008	ss4035869467
	Ref	G	Down	rs142721433[a]	chr13:21,800,567	T > G	0.0024	
rs1058083	Alt	G	Up	rs1157650381	chr13:99,385,968	T > C	0.0016	
rs1454361	Alt	A	Up	rs1878870026[a]	chr14:25,381,620–25,381,622[c]	delACA	0.0008	
rs722290	Alt	C	Up	rs564450658[a]	chr14:52,750,001	A > G	0.0008	
rs4530059	Ref/Alt	G/A	Down	**rs4450333**	chr14:104,302,886	C > T	0.2921	
rs1528460	Ref	C	Down	rs1412130030	chr15:54,918,520	C > T	0.0008	ss4035869469[f]
rs729172	Ref	G	Down	rs2045669877	chr16:5,556,205	A > G	0.0008	ss4035869471[e]
rs2342747	Ref	A	Up	rs897468214[a]	chr16:5,818,688	G > C	0.0016	ss4035869472
	Ref	A	Up	rs536580873[a]	chr16:5,818,697	G > A	0.0008	
rs430046	Alt	T	Up	**rs409820**	chr16:77,983,137	C > A	0.3425	
	Alt	T	Up	**rs430044**	chr16:77,983,148	C > T	0.3425	
	Ref	C	Up	rs534547615	chr16:77,983,149	G > A	0.0008	
rs9905977	Alt	G	Down	rs1407307549[a]	chr17:3,016,114	G > A	0.0008	
	Alt	G	Down	**rs28582109**	chr17:3,016,136	G > A	0.0764	
	Alt	G	Down	rs2046758811	chr17:3,016,141	G > A	0.0008	ss4035869473[e]
	Alt	G	Down	**rs73298992**	chr17:3,016,167	C > T	0.0024	
rs740910	Ref	A	Up	**rs60810599**	chr17:5,803,264	A > G	0.0110	
	Ref	A	Down	rs1248812772	chr17:5,803,319	A > G	0.0008	
rs8078417	Alt	T	Up	rs569140321	chr17:82,503,997	C > T	0.0016	
	Ref/Alt	C/T	Up	**rs182919351**[b]	chr17:82,504,004	C > T	0.0110	
	Ref	C	Up	rs752589755[b]	chr17:82,504,005	G > A	0.0008	ss4035869474
	Alt	T	Up	rs530011780	chr17:82,504,022	G > A	0.0008	
	Ref	C	Up	rs2044248635	chr17:82,504,026	C > T	0.0008	ss4035869475[e]

(continued on next page)

Table 2 (*continued*)

iSNP locus	Target reference (Ref) or alternative (Alt) allele	Target Allele	FR direction	FR variant rs number	GRCh38 position	Variant [Ref > Alt]	Alternative allele frequency	ss number on dbSNP (Submitter: GF)
	Ref	C	Down	rs1251308240[a]	chr17:82,504,060[c]	delG	0.0008	ss4035869476[e]
rs9951171	Ref	G	Up	rs1308337549	chr18:9,749,831	G > A	0.0008	
rs1736442	Ref	T	Up	**rs933726297**	chr18:57,558,487	G > A	0.0008	ss4035869477
	Ref	T	Up	rs935941264[a]	chr18:57,558,510	C > A	0.0024	ss4035869478
rs719366	Ref	G	Up	rs1968822112	chr19:27,972,398	G > C	0.0008	ss4035869479[e]
	Alt	A	Down	rs898461100	chr19:27,972,507	C > T	0.0008	ss4035869480
	Ref	G	Down	**rs719367**[b]	chr19:27,972,508	G > A	0.0252	
rs445251	Alt	C	Up	**rs369438**[b]	chr20:15,144,247	T > C	0.6591	
	Ref	G	Down	**rs117702247**	chr20:15,144,311	G > A	0.0008	
rs1005533	Alt	A	Down	rs1189113749[a]	chr20:40,858,540	G > C	0.0039	ss4035869481
rs1523537	Alt	C	Up	rs1981638299	chr20:52,679,576	A > G	0.0008	ss4035869482[e]
rs722098	Alt	G	Down	rs910933135	chr21:15,313,298	A > G	0.0008	ss4035869483
rs2830795	Ref/Alt	A/G	Up	**rs12626695**	chr21:27,235,806	T > C	0.1457	
rs2831700	Ref	A	Down	rs35270657[a]	chr21:28,307,375–28,307,377	delAAG	0.0024	
rs914165	Ref	G	Down	rs1341756746[b]	chr21:41,044,036	C > T	0.0008	ss4035869485[e]
rs987640	Alt	A	Up	**rs17793354**	chr22:33,163,488	A > C	0.0031	
rs2040411	Ref	G	Up	rs1043971565[a]	chr22:47,440,661	C > T	0.0016	ss4035869487
rs1028528	Ref	A	Up	rs976229315	chr22:47,966,534	C > T	0.0008	ss4035869488

Underline: The target iSNP and its flanking region variant(s) at a locus are named by reverse strands in ForenSeq ™ Universal Analysis Software (ForenSeq UAS) but changed to forward strands in this study.

Bold: The FR variant is detected by ForenSeq UAS using the pre-determined variant sites in the *Flanking Region Report*.

[a] Variant seems specific in East Asian (EAS), not reported in other populations on dbSNP Build 156.

[b] Reference SNP (RefSNP) is likely more prone to mutating in EAS, such as in Chinese, Japanese, and/or Korean populations.

[c] Indel positioned in a polymeric sequence tract is placed as an insertion or a deletion starting at the most 5′ nucleotide coordinate.

[d] Tri-allele SNP is observed in this study.

[e] Variant is newly found in this study, and its rs number has been authenticated by dbSNP.

[f] New alternative variant is first reported in this study and has been documented on dbSNP.

physically associated with rs6955448 [T]. Due to this primer-binding site mutation, the amplification of the DNA fragment containing the alternative allele [T] was less efficient. However, the average ACR at rs6955448 was well-balanced using the Precision ID Identity Panel,[3,43] with the reverse primer placed outside the rs6955464 mutation (Table S1). This mutation has been authenticated as the likely cause of ACR imbalance at rs6955448 in previous studies.[4,27,41]

The locus rs338882 presented lower average coverage for the reference allele [G] (33 × ± 12 ×) compared to the alternative allele [A] (65 × ± 22 ×). Out of 307 heterozygotes, 98.37% (302/307) exhibited an imbalance in favor of rs338882 [A]. The average ACR of this locus was 0.52 ± 0.17. As investigated in previous literature,[4,27,41] no mutation within the primer-binding regions at rs338882 was found to explain the observed ACR imbalance. Davenport et al.[27] found the only difference between the amplicons containing [G] and [A] was that the dinucleotide repeats $[GT]_4$ would be truncated to $[GT]_3$ when the reference allele [G] mutated to the alternative allele [A]. However, they were uncertain whether this was a causative factor on ACR imbalance at rs338882. In this study, we observed that target rs338882 presented a large flank deviation, a longer upstream flanking region (106 bp) and a shorter downstream flanking region (9 bp), and the potential primer-dimer (**rs338882_R** vs. DYS392_F) was detected in Section 3.2.1. We are also uncertain whether these observations can explain the compounding issue of imbalanced ACR and low DoC at this locus. However, the average ACR was balanced at rs338882 using the Precision ID Identity Panel,[3,43] with a 23-bp upstream flanking region and a 49-bp downstream flanking region (Fig. S2).

Another two loci exhibited not well-balanced ACR, and they were also identified in previous studies: rs1493232 (0.60 ± 0.17)[6,18,19,27] and rs2111980 (0.61 ± 0.13).[6,18,19,27,43] Similarly, Davenport et al.[27] also reported no mutations within the primer-binding regions at these two loci. In cases where there is a less-pronounced ACR imbalance combined with low DoC, the lower coverage allele may drop out in heterozygotes. Thus, rs907100, rs1357617, and rs4606077 should also be cautiously evaluated during data interpretation.

3.3. Characterization of ForenSeq iSNPs

Characterization of each locus in ForenSeq Kit (length of amplicons and GRCh38 coordinates of target, amplicon start, and amplicon end positions) is listed in Table 1, and more details (upstream and downstream flanking region sequences, distance from target positions, genotypes from 11 standard control samples) in Table S3. Realignment results revealed that complete sequences between PCR primers were reported for all 94 iSNPs in the *Flanking Region Report* by ForenSeq UAS v1.3. The amplicon length without primers and adapters averaged to (69 ± 27) bp, ranging from the shortest (11 bp) at rs8037429 to the longest (128 bp) at rs735155 and rs4530059, which has the overwhelming superiority in detecting degraded samples as expected.[2,4,10,12]

Based on realignment outcomes, Table S1 shows another three loci showed extremely balanced mutations with the minor allele frequency (MAF) of approximately 0.50 on the population level in their forward primers, such as rs76875728 [G > A] from rs10776839, rs75460798 [G > G] from rs10488710, and rs112167443 [T > A] from rs221956. Similarly, another five loci exhibited mutations in reverse primers, such as rs1414119020 [C > T] from rs6955448, rs12437775 [C > G] from rs1821380, rs381840 [A > G] from rs430046, rs1005534 [G > A] from rs1005533, and rs16991914 [T > C] from rs987640. Most detected SNPs in primer-binding sites had MAF ≥0.01 on dbSNP, except for rs6076682 and rs1414119020. Davenport et al.[27] found that rs1414119020 was 2-bp away from the actual primer-binding site mutation. They speculated that the error resulted from the initial primer design because the intended target of the degenerate primer for rs6955448 was rs6955464. However, there is not any hotspot primer-binding site mutation nearby rs6076682. Sanger sequencing results indicated that primer-binding site polymorphisms observed in realignment results may have resulted from 1-mismatched and/or degenerate primers, masking actual mutations from individuals (Table S1). As discussed in Section 3.2.1, rs1031825 (with the degenerate forward primer) and rs2342747 (with the 1-mismatched and degenerate reverse primer) together with eight loci mentioned above were all found to have additional primers in ForenSeq Kit.

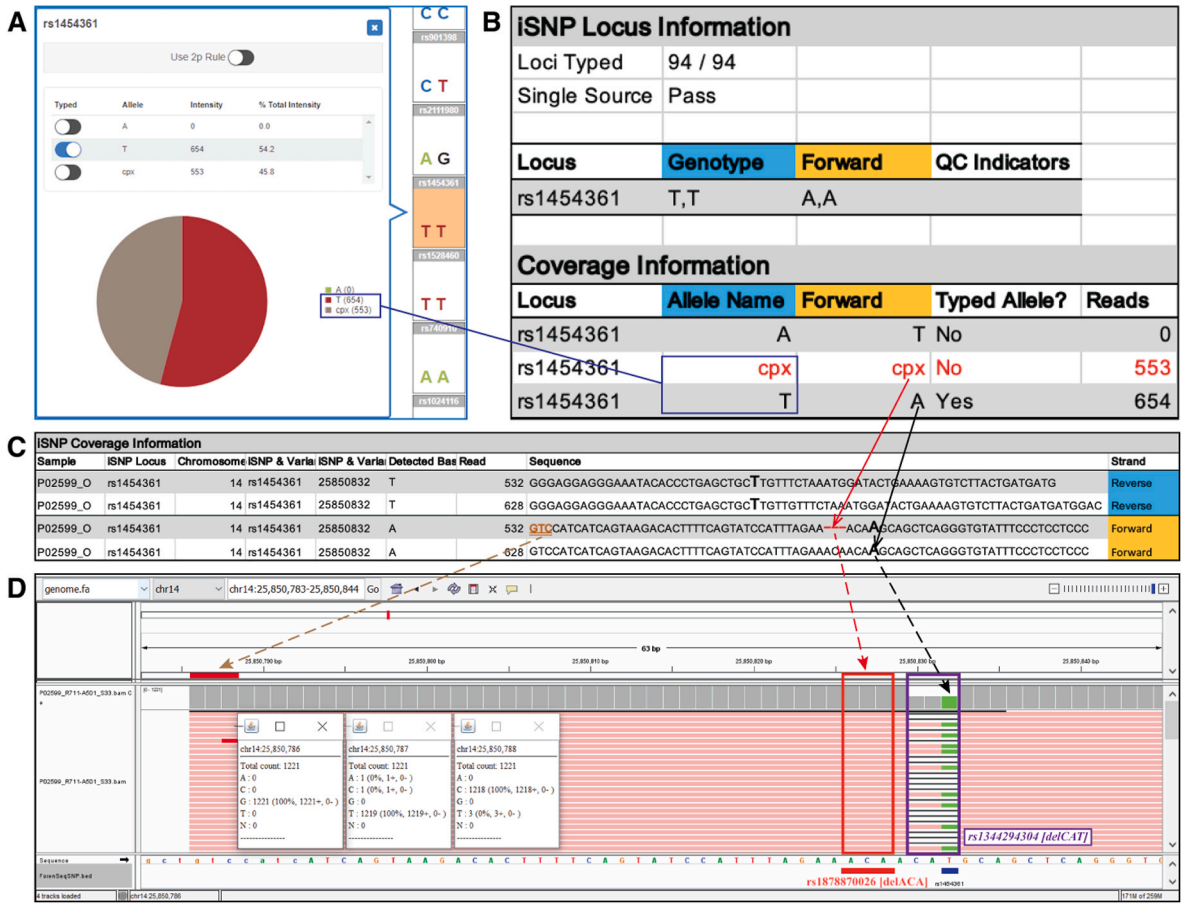

Fig. 2. Bioinformatic discordance at rs1454361. **(A)** The ForenSeq Universal Analysis Software (UAS) interface displays the target genotype of homozygous T at rs1454361, named by the reverse strand. However, it interprets 553 reads as a complex (cpx) but does not trigger any QC indicators. **(B)** The *Sample Details Report* generates the target genotype of homozygous T (reverse strand), which matches the ForenSeq UAS interface. If the strand is changed to forward, the target genotype becomes homozygous A. **(C)** The *Flanking Region Report* produces the amplicon genotype of heterozygous A/A (forward strand), one of which contains the first 3 bases not reported (marked in brown) and a 3-bp deletion (marked in red). **(D)** The Integrative Genomic Viewer (IGV) inspection confirms that the target rs1454361 is adjacent to a trinucleotide-repeat structure, $[ACA]_2$, that contains a rare deletion (rs1878870026, marked in a red box). This results in two isoalleles: one of [A] with rs1878870026 [delACA] in the upstream flanking region (UFR) and the other [A] without this deletion. Additionally, the first 3 bases not reported in one amplicon allele are caused by ForenSeq UAS reporting algorithm to erroneously truncate the flanking region sequence due to rs1878870026 [delACA]. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

3.4. Bioinformatic concordance

In this study, 99.99% (119374/119380) of alleles from *Project Genotype* and *Flanking Region Reports* were concordant with the IGV visual inspection, which was consistent with the findings of previous studies conducted by Davenport et al.[27] and King et al.[41] The remaining six discordant alleles (0.01%) resulted from flanking region deletions, causing the ForenSeq UAS reporting algorithm to erroneously truncate flanking region sequences in the *Flanking Region Report*. Details of these discordant alleles can be found in Table S4.

The first was an amplicon string containing an alternative allele [T] at rs10776839, where a 1-bp downstream flanking region (DFR) deletion (rs542545139 [delG]) caused the last base to be not reported in the *Flanking Region Report*. The second was an amplicon string containing a reference allele [C] at rs8078417, which also presented the last base not reported due to a 1-bp DFR deletion (rs1251308240 [delG]). The third was observed at rs1454361, where a 3-bp upstream flanking region (UFR) deletion (rs1878870026 [delACA]) led to the first 3 bases not reported in the manually converted forward amplicon (if the original reverse amplicon provided in a *Flanking Region Report* was investigated, then it was the last 3 bases not reported). The erroneous truncation also has its own mechanism in general: the deletion residing within the DFR of target iSNP causes the amplicon string to be truncated from the end; the deletion within the UFR causes the amplicon string to be truncated from the beginning; the number of bases in the deletion is equal to that not reported in the amplicon. A complex fourth discrepancy was observed in three samples with a reference allele [A] at rs2831700. The discrepancy involved a 3-bp deletion at rs35270657, where the same alternative allele [delAAG] reported on dbSNP was selected but not [delAGA] reported by Davenport et al.[27] The deletion spanned from the last 2 bases of the amplicon string to the first base of the reverse primer-binding region. This resulted in the last 6 bases being unreported (3 bases from the deletion plus 3 bases from the erroneous truncation). Moreover, these four flanking region deletions were identified as specific variants in East Asian (see Section 3.5.1). However, other flanking region deletions reported by Davenport et al.[27] were not found in Northern Han Chinese, such as rs575053109 [delT] in the UFR of rs2831700 in West African, [delTG] in the DFR of rs13218440 in North-East African, rs571330241 [delTTC] in the DFR of rs1454361 in West African, rs565694318 [delATCATA] in the UFR of rs740910 in West African, and rs1203010982 [delC] in the UFR of rs1355366 in East Asian. Additionally, although rs543563536 [insT] resided in the UFR of rs891700 and rs1573442732 [insT] in the UFR of rs993934 (Table 2), we did not observe discordant alleles resulting from flanking region insertions, causing additional bases from the primer-binding region merged into the ForenSeq UAS reported sequence, e.g., rs750429368

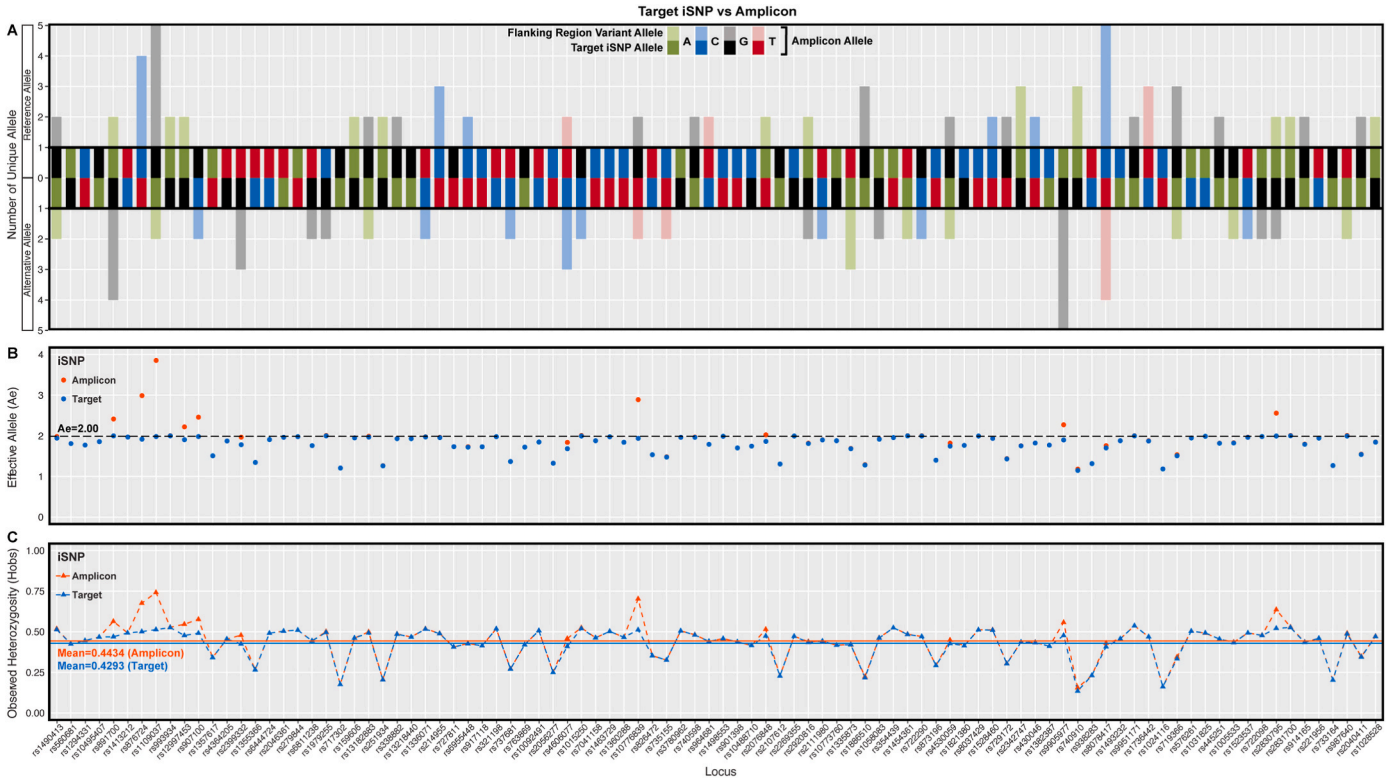

Fig. 3. Gains obtained by comparing the target data with the amplicon data. Herein, loci are arranged in order of those on the ForenSeq Universal Analysis Software interface and in the *Sample Details Report*. **(A)** The number of unique alleles is counted at each locus from 635 Northern Han Chinese. Herein, the amplicon alleles (also known as microhaplotypes) are composed of the target iSNP alleles (dark color bars) and flanking region variant alleles (light color bars). **(B)** The effective number of alleles (A_e) values are calculated from the target data (blue round dots) and the amplicon data (blue round dots) at 94 iSNP loci, using the formula $A_e = 1/\sum p_i^2$ where p_i is the frequency of the *i*th allele. The horizontal black dashed line indicates $A_e = 2.00$, which is the maximum value for a bi-allelic locus in the target data. **(C)** The observed heterozygosity (H_{obs}) values are calculated from the target data (the blue dashed line with triangles) and the amplicon data (the red dashed line with triangles) at 94 iSNP loci. The mean H_{obs} values for the target data ($H_{obs} = 0.4239$) and the amplicon data ($H_{obs} = 0.4434$) across 94 loci are represented by the horizontal blue and red solid lines, respectively. Details of unique alleles, A_e, and H_{obs} can be found in Tables S7 and S8. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

[insT] in the DFR of rs1821380 in South Asian from.[27]

Fig. 2 shows that the most complex discrepancy was observed at rs1454361 in this study. A sample (P02599) was typed as homozygous A on the ForenSeq UAS interface and in the *Sample Details Report*, whereas the amplicon strings were analyzed as heterozygous A/A in the *Flanking Region Report*. The ForenSeq UAS interpreted 45.8% (553/1207) of DoC as a complex (cpx) but did not trigger any QC indicators. Visual inspection revealed there were two ways to explain this discrepancy. First, the target rs1454361 [T > A] is adjacent to trinucleotide repeats $[ACA]_2$ containing a rare EAS-specific deletion (rs1878870026, delACA = 0.00004 in 14KJPN on dbSNP), resulting in isoalleles of [A] with rs1878870026 [delACA] in the UFR and [A] without this deletion, as shown in the *Flanking Region Report*. This way was similar to the explanation at rs10092491 by Kiesler et al.[18] Second, the target rs1454361 is located within another rare African-specific deletion (rs1344294304, delCAT $_{African}$ = 0.00005 and delCAT $_{Global}$ = 0.000014 in gnomAD on dbSNP), thus causing one allele to be as [delT] and the other as [A], as shown in the original IGV graph. This was the explanation at the same rs1454361 by Li et al.[43] For this locus, we chose the first way to interpret its genotype (i.e., heterozygous A/A) according to indels starting at the most 5' nucleotide coordinates[36] and the allele frequency in a specific population.

3.5. Target alleles and amplicon alleles

In this study, a target allele refers to an allele that is specifically found at a target iSNP locus. This information can be easily accessed through the ForenSeq UAS interface, *Sample Details Report*, and *Project Genotype Report*. On the other hand, an amplicon allele, also known as a microhaplotype, refers to a haplotype that includes the target iSNP with or without flanking region variants from the full amplicon sequence. These are always included in the *Flanking Region Report*. Details for observed target and amplicon alleles by locus are listed in Table S5, including short nomenclature, flanking region (FR) variants, full amplicon sequences, microhaplotype categories, allele counts, and allele frequencies.

3.5.1. Flanking region variants

Table 2 shows that 88 FR variants, comprising 82 SNPs and 6 indels, were identified at 55 iSNPs. Of these, 32 FR variants were detected by ForenSeq UAS using the 120 pre-determined variant sites in the *Flanking Region Report* (Table S6), while 56 additional FR variants were manually identified using IGV. It was discovered that all additional FR variants had an alternative allele frequency of less than 1%. However, out of the 32 pre-determined variants, 25 (78.13%) were observed with a frequency of more than 1%. This suggests that ForenSeq UAS has selected the most advantageous polymorphic sites at a population level to be reported.[27]

Of these FR variants, 12 were found for the first time in this study (IGV screenshots in Fig. S3) and have been validated by dbSNP and then released in Build 155, such as rs1252699848 [A > T] from rs1490413, rs1665500714 [T > A] from rs876724, rs1771121532 [T > G] from rs1336071, rs2097285015 [G > A] from rs214955, rs1851671415 [T > A] from rs740598, rs2045669877 [A > G] from rs729172,

rs2046758811 [G > A] from rs9905977, rs2044248635 [C > T] and rs1251308240 [delG] from rs8078417, rs1968822112 [G > C] from rs719366, rs1981638299 [A > G] from rs1523537, and rs1341756746 [C > T] from rs914165. New alternative alleles (marked in bold) were first observed at four FR SNPs (IGV screenshots in Fig. S3) and also recorded on dbSNP, such as rs1561496729 [**G > A** / G > T] from rs13182883, rs1757388778 [T > A / T > G] from rs251934, rs1039895061 [T > A / T > C] from rs737681, and rs1412130030 [C > A / C > T] from rs1528460. Additionally, a tri-allele SNP (G = 0.9000, A = 0.0992, C = 0.0008) was observed at rs12047255 from rs891700, also reported by King et al.[41]

Compared with alternative allele frequencies from 14KJPN (Japanese), KRGDB (Korean), gnomAD (East Asian and Global), and 1000Genomes_30x (East Asian and Global) on dbSNP Build 156, twenty-eight FR variants (23 SNPs, 4 deletions, and 1 insertion) seemed to be specific in East Asian (EAS) and not reported in other populations. Another 14 FR SNPs were likely more prone to mutating in EAS, where frequencies were higher than those in other populations at a global level (Table S6). On the premise of excluding individual or family private mutations, forensic biogeographical ancestry estimation may be aided by studying specific, highly mutable EAS variants.

3.5.2. Unique alleles

As shown in Fig. 3A and Table S7, the number of target and amplicon unique alleles from 635 Northern Han Chinese. During MPS typing, 276 amplicon alleles were detected across 94 iSNPs. The ForenSeq target iSNPs accounted for 188 alleles (i.e., target alleles), while the remaining 88 alleles (i.e., flanking region variant alleles) were identified when including variants in the flanking regions adjacent to target iSNPs. In comparison to target alleles, the amplicon alleles showed a significant increase of 46.81% in the total number of unique alleles. Among these, the top four with the highest number and largest increase of FR variant alleles were observed at rs8078417 (7 FR variant alleles and 350% growth from target alleles to amplicon alleles), rs1109037 (5 alleles and 250% growth), rs891700 and rs9905977 (4 alleles and 200% growth for both), and rs876724, rs4606077, and rs719366 (3 alleles and 150% growth for all). Recent literature also observed the highest number of amplicon alleles at rs8078417.[18,27,41] Notably, three target iSNPs (marked in bold) with their FR variants showed perfect linkages (p = 0.0000 for tests of LD between all pairs) in Northern Han Chinese, such as **rs279844**-rs279845 (reference combination: **A**-T; alternative combination: **T**-A), rs409820-rs430044-**rs430046** (A-T-**T**; C–C–**C**), and **rs4606077**-rs1869434 (**T**-G; **C**-A), the first two of which was also observed in Saudi Arabia.[21] Also, the alternative allele frequency of rs1390470 [T] from rs727811 was 1.0000 in Northern Han Chinese. Thus, these five FR variants did not contribute to any polymorphism in this study.

As those amplicons are designed by the manufacturer to target a single nucleotide, any FR variants may be considered adventitious.[24] In Northern Han Chinese, 25.72% (71/276) of amplicon unique alleles had a low frequency (<0.01). Of these 71 low-frequency alleles, 48 were observed only once (=0.0008). As per the findings of King et al.,[41] amplicons have been classified into four categories, which are summarized in Table S7 and elaborated in Table S5. The categories consisted of microhaplotypes at 15 loci (with ≥2 SNPs that can generate ≥3 haplotypes with a minor haplotype frequency (MHF) ≥ 0.01), minor microhaplotypes at 38 loci (with ≥2 SNPs that can generate ≥3 haplotypes and only 2 haplotypes with MHF ≥0.01), single-SNP haplotypes at 39 loci (target iSNPs without FR variants within amplicons), and as non-variable haplotypes introduced in this study at 2 loci (with ≥2 SNPs that can only generate 2 haplotypes like single-SNP haplotypes, e.g., rs279844 and rs727811). Out of all the presented loci, around 16% exhibited microhaplotypes, which could be especially valuable for mixture analysis.

3.5.3. Effective alleles

The effective number of alleles (A_e) is an important index used for evaluating the selection of microhaplotypes for mixture detection. In the amplicon data, 17 loci showed an A_e greater than 2.00, as indicated in Fig. 3B and Table S7. Among them, rs1109037 met the necessary criterion for microhaplotypes to be used in mixture detection,[39] with an A_e > 3.00. On the other hand, none of the loci in the target data had an $A_e \geq$ 2.00, which is the maximum value for a bi-allelic locus. Two loci, rs740910 (as microhaplotype) and rs1024116 (as single-SNP haplotype), displayed the lowest values (A_e < 1.20) in both amplicon and target data, which were also observed in East Asian populations.[27,41] However, the lowest values slightly differed from those in other populations[18,28] due to variation in allele frequency among populations. When compared to the target data, the amplicon data showed an average increase of 4.51% in A_e, with the largest increases (>15%) observed at rs1109037, rs876724, rs10776839, rs2830795, rs907100, rs891700, rs9905977, and rs12997453, sorted in descending order. These eight loci were all included in microhaplotypes identified in Section 3.5.2. Taking into account the highest number and largest increase of the amplicon unique alleles and A_e values, four loci (rs1109037, rs891700, rs9905977, and rs876724) performed the best in this study.

3.6. Population genetic and forensic parameters

Population genetic and forensic parameters were calculated for both target and amplicon data across 94 iSNPs, including observed heterozygosity (H_{obs}), expected heterozygosity (H_{exp}), polymorphism information content (PIC), match probability (MP), power of discrimination (PD), power of exclusion (PE), typical paternity index (TPI), and tests of Hardy-Weinberg equilibrium (HWE) and linkage disequilibrium (LD) listed in Tables S8 and S9. Although the Bonferroni correction was utilized in this study to correct the p-value for multiple HWE and LD comparisons, it is well known that this is too conservative and will result in under-reporting of deviations from HWE and linkage equilibrium (i.e., it will result in false negatives). To address this issue, alternative methods such as Zaykin's truncated product method[45] and Buckleton's p-p plot method[46] were also introduced to mitigate against the multiple comparison problems.

Fig. 3C and Table S8A showed that the average H_{obs} was calculated as 0.4293 for the target data, with a range of 0.1354 (rs740910) to 0.5370 (rs9951171), and 0.4434 for the amplicon data, with a range of 0.1559 (rs740910) to 0.7417 (rs1109037). As expected, H_{obs} increased at most loci with FR variants. A comparison of H_{obs} by target and amplicon showed the average percentage of increase was 3.29%, similar to a 3.75% increase by length and sequence across 27 A-STRs.[32] The top five largest increases were observed at rs1109037 (44.48%), rs10776839 (37.23%), rs876724 (34.90%), rs2830795 (22.05%), and rs891700 (20.13%), and 20%–5% increases at rs907100, rs9905977, rs740910, rs12997453, rs2399332, rs4606077, rs2076848, and rs8078417, sorted in descending order. However, the remaining 42 loci with FR variants exhibited low (<5%) or no increase in H_{obs}. Meanwhile, the average H_{exp} was 0.4284 from 0.1291 (rs740910) to 0.5003 (rs2831700) for the target data and 0.4428 from 0.1508 (rs740910) to 0.7411 (rs1109037) for the amplicon data. The distribution pattern of H_{exp} for the target and amplicon data was similar to that in East Asian.[41] At each locus from the target and amplicon data, the absolute value of H_{exp} minus H_{obs} ($|H_{exp} - H_{obs}|$) was less than 0.40. In the study, rs1109037 demonstrated the highest H_{obs} and H_{exp} in the amplicon data, surpassing heterozygosity of the poorly performing length-based (LB) and sequence-based (SB) A-STRs from ForenSeq Kit (i.e., TPOX, TH01, and D17S1301) in Northern Han Chinese (Table S14 in Ref. 32). Notably, this particular locus has also been reported in previous studies.[24,28,41] Tests of HWE showed statistically significant departures from expectations (p < 0.05) at rs10092491 and rs10773760 for the target data and rs2399332, rs10092491, rs10776839, and rs10773760

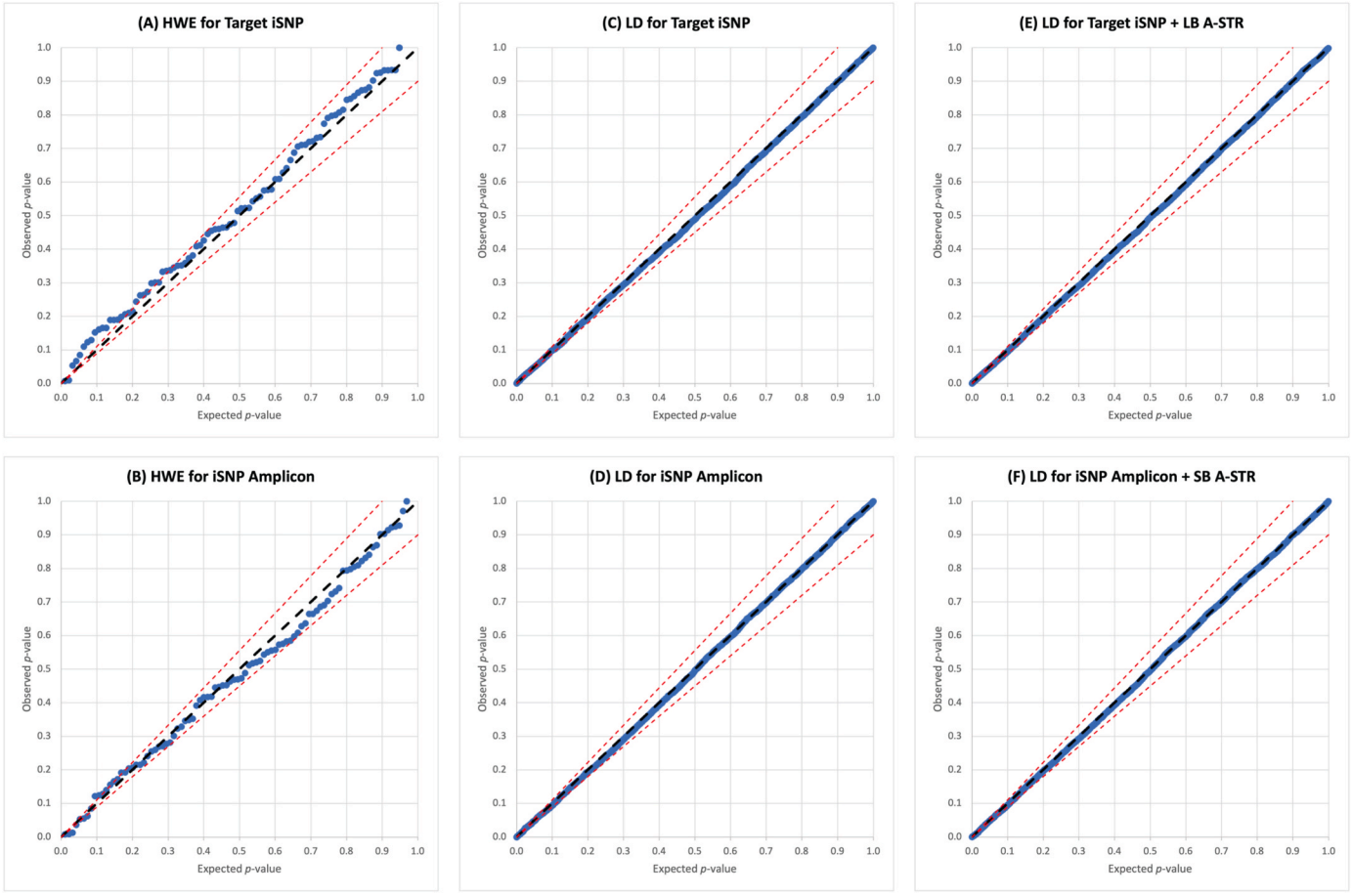

Fig. 4. Observed and expected *p*-values (Buckleton's *p-p* plots) in Northern Han Chinese (N = 635). The black dashed line represents no deviation between the observed and expected *p*-values. The red dashed line represents a 0.9 ratio between the observed and expected *p*-values. **(A)** and **(B)**: 94 Hardy-Weinberg equilibrium (HWE) tests for iSNPs; **(C)** and **(D)**: 4371 pairwise linkage disequilibrium (LD) tests for iSNPs; **(E)** and **(F)**: 7260 pairwise LD tests for iSNPs and A-STR. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

for the amplicon data, while all deviation data met HWE expectations after the Bonferroni correction ($\alpha = 0.000532$). Additionally, the study found that *p*-values for the overall null hypothesis, which suggests that all 94 iSNPs are in HWE, were 0.8590 for the target data and 0.4485 for the amplicon data when using the truncated product method. This was determined by performing a chi-squared test, which involved the sum of $-2\ln(p)$ of 94 HWE tests with a degree of freedom (*df*) of 188 (2×94 tests). The *p-p* plots in Fig. 4A and B shows very little deviation from HWE for 94 iSNPs, with *p*-values uniformly distributed between 0 and 1 (i.e., observed *p*-values are equal to expected *p*-values and, therefore, should lie along the diagonal).

Table S9A summarizes tests of LD between iSNPs as well as between iSNPs and A-STRs in Northern Han Chinese. Table S9B contains p-values associated with LD tests between 4371 pairwise comparisons, out of which 227 pairwise comparisons for the target data and 230 for the amplicon data presented statistically significant LD ($p < 0.05$). Furthermore, 7 pairwise comparisons for the target data and 5 for the amplicon data were located in identical chromosomes, while those LD could not be detected after the Bonferroni correction ($\alpha = 0.000011$). The truncated product method (TPM) analysis revealed that all pairwise comparisons met linkage equilibrium (LE) with $2 \times 4371 = 8742$ *df* ($p = 0.2543$ for the target data and $p = 0.1652$ for the amplicon data). Also, there was no deviation from the diagonal line observed in the *p-p* plots (Fig. 4C and D). Thus, it can be concluded that 94 iSNPs are independent in Northern Han Chinese. Moreover, LD was tested with the combined 27 A-STRs and 94 iSNPs (7260 pairwise comparisons). At the marker level, 138 pairwise comparisons demonstrated statistically significant LD ($p < 0.05$) for LB A-STRs and target iSNPs analysis and 136 pairwise comparisons for SB A-STRs and iSNP amplicons analysis. At the chromosomal level, the detectable LD was observed in 9 pairs between LB A-STRs and target iSNPs (*p*-values in Table S9C), such as TPOX-rs12997453, D2S441-rs907100, CSF1PO-rs13182883, D6S1043-rs214955, D9S1122-rs1015250, TH01-rs1498553, D18S51-rs1024116, D20S482-rs445251, and D21S11-rs2830795, and 12 pairs between SB A-STRs and iSNP amplicons (Table S9C), such as D2S441-rs907100, D3S1358-rs1357617, D3S1358-rs2399332, D5S818-rs13182883, CSF1PO-rs13182883, D6S1043-rs214955, D8S1179-rs2056277, D9S1122-rs1015250, TH01-rs1498553, vWA-rs10773760, D20S482-rs1031825, and D21S11-rs2830795. However, none of the above-mentioned LD was detected after the Bonferroni correction ($\alpha = 0.000007$). All pairwise LE tests with $2 \times 7260 = 14520$ *df* were found to be true using TPM. The *p*-values were 0.1098 for the target data and 0.1063 for the amplicon data, respectively. Fig. 4E and F shows no deviation between the observed and expected *p*-values in the *p-p* plots. The outcome demonstrates that 27 A-STR and 94 iSNPs are independent within autosomes in Northern Han Chinese, which means the "product rule" can be used to estimate forensic parameters.

Table S8B lists forensic parameters calculated for both target and amplicon data by locus. MP ranged from 0.3596 (rs1498553) to 0.7631 (rs740910) for the target data and from 0.1165 (rs1109037) to 0.7277 (rs740910) for the amplicon data. The combined match probability (CMP) of 94 iSNPs decreased to 3.9876×10^{-38} when analyzed by amplicon, which was approximately three orders of magnitude (304 times) lower than 1.2120×10^{-35} by target and even three to seven

Table 3
Forensic parameters with the combined A-STRs and iSNPs in Northern Han Chinese.

Parameters	Combination					
	27 LB A-STRs	27 SB A-STRs	94 target iSNPs	94 iSNP amplicons	27 LB A-STRs + 94 target iSNPs	27 SB A-STRs + 94 iSNP amplicons
CMP	1.5190×10^{-31}	3.2908×10^{-35}	1.2120×10^{-35}	3.9876×10^{-38}	1.8410×10^{-66}	1.3122×10^{-72}
CPD	$1 - 1.5190 \times 10^{-31}$	$1 - 3.2908 \times 10^{-35}$	$1 - 1.2120 \times 10^{-35}$	$1 - 3.9876 \times 10^{-38}$	$1 - 1.8410 \times 10^{-66}$	$1 - 1.3122 \times 10^{-72}$
CPE	$1 - 4.8096 \times 10^{-11}$	$1 - 6.7054 \times 10^{-13}$	$1 - 4.4577 \times 10^{-7}$	$1 - 6.6690 \times 10^{-8}$	$1 - 2.1439 \times 10^{-17}$	$1 - 4.4719 \times 10^{-20}$

CMP: combined match probability; CPD: combined power of discrimination; CPE: combined power of exclusion; LB: length-based; SB: sequence-based.

orders of magnitude lower than that of 27 A-STRs (CMP = 3.2908×10^{-35} for the SB data and 1.5190×10^{-31} for the LB data calculated in Ref. 32). PD ranged from 0.2369 (rs740910) to 0.6404 (rs1498553) for the target data with the combined power of discrimination (CPD) of $1 - 1.2120 \times 10^{-35}$ and from 0.2723 (rs740910) to 0.8835 (rs1109037) for the amplicon data with the CPD of $1 - 3.9876 \times 10^{-38}$. PE ranged from 0.0146 (rs740910) to 0.2220 (rs9951171) for the target data with the combined power of exclusion (CPE) of $1 - 4.4577 \times 10^{-7}$ and from 0.0189 (rs740910) to 0.4957 (rs1109037) for the amplicon data of $1 - 6.6690 \times 10^{-8}$. As shown in Table 3, the power of personal identification of 94 iSNPs was found to be higher than that of 27 A-STRs (CPD = $1 - 3.2908 \times 10^{-35}$ for the SB data and $1 - 1.5190 \times 10^{-31}$ for the LB data in Ref. 32), but the power of parentage testing was lower than that of 27 A-STRs (CPE = $1 - 6.7054 \times 10^{-13}$ for the SB data and $1 - 4.8096 \times 10^{-11}$ for the LB data in Ref. 32), which has been proved in Chinese Tibetan with the same kit[28] and Northern Han Chinese with the HID-Ion AmpliSeq ™ Identity Panel.[3] Additionally, the power of the system effectiveness with the combined 27 A-STRs and 94 iSNPs was added more substantially than with one type of marker alone, e.g., CPD = $1 - 6.7054 \times 10^{-72}$ and CPE = $1 - 4.4719 \times 10^{-20}$ using the SB A-STRs and iSNP amplicons in combination.

4. Conclusion

The article outlines target and flanking region (FR) variants at 94 identity-informative SNPs (iSNPs) are investigated in 635 Northern Han Chinese using the ForenSeq DNA Signature Prep Kit on the MiSeq FGx Forensic Genomics System. MiSeq FGx quality metrics and massively parallel sequencing (MPS) performance may reflect the quality of the dataset in this study. The dataset has the following characteristics (average values): ≥60% bases with a quality score of 20 or higher (%≥ Q20); >700 × of depth of coverage (DoC) from both *Sample Details Reports* and *Flanking Region Reports*; >80% of effective reads; ≥60% of allele coverage ratio (ACR); and ≥70% of inter-locus balance. Meanwhile, some stable low-performance characteristics reported in previous studies have also been observed: low DoC at rs1736442, rs1031825, rs7041158, rs338882, rs2920816, rs1493232, rs719366, and rs2342747; high noise at rs891700; and imbalanced ACR at rs6955448 and rs338882. We have identified the start and end positions of iSNP amplicons in ForenSeq Kit using GRCh38 coordinates. As a result, we can confirm 100% of the flanking sequences reported in the Flanking Region Report exported from the ForenSeq Universal Analysis Software v1.3. The average amplicon length is 69 bp, which is very suitable for detecting degraded samples. Also, additional primers designed in ForenSeq Kit are confirmed for rs10776839, rs10488710, rs1031825, rs221956, rs6955448, rs1821380, rs2342747, rs430046, rs1005533, and rs987640 using Sanger sequencing. ForenSeq UAS and IGV inspection achieve 99.99% bioinformatic concordance, with discordance due to flanking region deletions of rs10776839, rs8078417, rs2831700, and rs1454361.

The increase in the number of unique alleles and the average of gene diversity is 46.81% and 3.29% between amplicons and target iSNPs, respectively, attributing to FR variants within amplicons detected by MPS. In this study, 12 FR variants are first found and authenticated by dbSNP. Microhaplotypes associated with target rs1109037, rs891700, rs9905977, and rs876724 present a higher number and larger increase in all aspects of unique alleles, effective alleles (A_e), and observed heterozygosity (H_{obs}), which may provide additional genetic information. All 94 iSNPs from both target and amplicon data meet Hardy-Weinberg equilibrium (HWE) expectations and are independent within autosomes. As expected, forensic parameters from the amplicon data increase significantly on the combined power of discrimination (CPD) and the combined power of exclusion (CPE). Additionally, the power of the system effectiveness with the combined 27 autosomal STRs and 94 iSNPs is substantially improved compared to only one type of marker.

While ForenSeq UAS has the capability to provide complete flanking sequences and detect the most advantageous FR variants in a *Flanking Region Report*, it can still be challenging to interpret the data in practice. Nevertheless, the system's effectiveness in terms of CPD and CPE increases considerably when using amplicon data as opposed to target data. As Davenport et al.[27] discussed, the decision on whether or not to include flanking region variants in a data interpretation workflow will depend on the trade-off between analytical simplicity and evidentiary gain. The use of STR markers involves the same decision as well. So far, we have established a comprehensive reference database of 58 STRs and 94 iSNPs in the Northern Han Chinese population, including both traditional length-based and current sequence-based data. We hope that these data can provide some reference and foundation for decisions.

Compliance with ethical standards

Informed consent was written by all individuals, and the study was approved by the Ethical Committee of Jinzhou Medical University (No. 2020016). This study follows the ethical principles stated in the 1964 Helsinki Declaration and its later amendments.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors thank Sheng Peng and Huiyuan Niu at QIAGEN and Ziming Deng at TopScience for their technical support.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jflm.2024.102678.

References

1. Butler JM. *Advanced Topics in Forensic DNA Typing: Methodology*. Boston: Academic Press; 2011.
2. Gettings KB, Kiesler KM, Vallone PM. Performance of a next generation sequencing SNP assay on degraded DNA. *Forensic Sci. Int. Genet.* 2015;19:1–9. https://10.1016/j.fsigen.2015.04.010.
3. Guo F, Zhou Y, Song H, et al. Next generation sequencing of SNPs using the HID-Ion AmpliSeq™ identity panel on the Ion Torrent PGM™ platform. *Forensic Sci. Int. Genet.* 2016;25:73–84. https://doi.org/10.1016/j.fsigen.2016.07.021.
4. Guo F, Yu J, Zhang L, Li J. Massively parallel sequencing of forensic STRs and SNPs using the Illumina ® ForenSeqTM DNA signature prep kit on the MiSeq FGxTM

forensic genomics system. *Forensic Sci. Int. Genet.* 2017;31:135–148. https://doi.org/10.1016/j.fsigen.2017.09.003.
5. Li R, Li H, Peng D, et al. Improved pairwise kinship analysis using massively parallel sequencing. *Forensic Sci. Int. Genet.* 2019;38:77–85. https://doi.org/10.1016/j.fsigen.2018.10.006.
6. Tao R, Xu Q, Wang S, et al. Pairwise kinship analysis of 17 pedigrees using massively parallel sequencing. *Forensic Sci. Int. Genet.* 2022;57, 102647. https://doi.org/10.1016/j.fsigen.2021.102647.
7. Kling D, Phillips C, Kennett D, Tillmar A. Investigative genetic genealogy: current methods, knowledge and practice. *Forensic Sci. Int. Genet.* 2021;52, 102474. https://doi.org/10.1016/j.fsigen.2021.102474.
8. Phillips C. Forensic genetic analysis of bio-geographical ancestry. *Forensic Sci. Int. Genet.* 2015;18:49–65. https://doi.org/10.1016/j.fsigen.2015.05.012.
9. Kayser M. Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes. *Forensic Sci. Int. Genet.* 2015;18:33–48. https://doi.org/10.1016/j.fsigen.2015.02.003.
10. Butler JM. Recent advances in forensic biology and forensic DNA typing: INTERPOL review 2019-2022. *Forensic Sci. Int. Synerg.* 2022;6, 100311. https://doi.org/10.1016/j.fsisyn.2022.100311.
11. Pereira V, Mogensen HS, Børsting C, Morling N. Evaluation of the Precision ID Ancestry Panel for crime case work: a SNP typing assay developed for typing of 165 ancestral informative markers. *Forensic Sci. Int. Genet.* 2017;28:138–145. https://doi.org/10.1016/j.fsigen.2017.02.013.
12. Jäger AC, Alvarez ML, Davis CP, et al. Developmental validation of the MiSeq FGx forensic genomics system for targeted next generation sequencing in forensic DNA casework and database laboratories. *Forensic Sci. Int. Genet.* 2017;28:52–70. https://doi.org/10.1016/j.fsigen.2017.01.011.
13. Verogen. *ForenSeq Kintelligence Kit Reference Guide, Document # VD2020053 Rev. B*. San Diego: Verogen; 2021 (Available from: https://verogen.com/wp-content/uploads/2021/03/forenseq-kintelligence-reference-guide-vd2020053-b.pdf.
14. Verogen. *ForenSeq Imagen Kit Reference Guide, Document # VD2022008 Rev. A*. San Diego: Verogen; 2022 (Available from: https://verogen.com/wp-content/uploads/2022/08/forenseq-imagen-reference-guide-PCR1-vd2022008-a.pdf.
15. Li R, Shen X, Chen H, Peng D, Wu R, Sun H. Developmental validation of the MGIEasy Signature Identification Library Prep Kit, an all-in-one multiplex system for forensic applications. *Int J Leg Med*. 2021;135:739–753. https://doi.org/10.1007/s00414-021-02507-0.
16. Wendt FR, King JL, Novroski NMM, et al. Flanking region variation of ForenSeq™ DNA signature prep kit STR and SNP loci in Yavapai native Americans. *Forensic Sci. Int. Genet.* 2017;28:146–154. https://doi.org/10.1016/j.fsigen.2017.02.014.
17. Churchill JD, Novroski NMM, King JL, Seah LH, Budowle B. Population and performance analyses of four major populations with Illumina's FGx Forensic Genomics System. *Forensic Sci. Int. Genet.* 2017;30:81–92. https://doi.org/10.1016/j.fsigen.2017.06.004.
18. Kiesler KM, Borsuk LA, Steffen CR, Vallone PM, Gettings KB. US population data for 94 identity-informative SNP loci. *Genes*. 2023;14:1071. https://doi.org/10.3390/genes1405107.
19. Casals F, Anglada R, Bonet N, et al. Length and repeat-sequence variation in 58 STRs and 94 SNPs in two Spanish populations. *Forensic Sci. Int. Genet.* 2017;30:66–70. https://doi.org/10.1016/j.fsigen.2017.06.006.
20. Hussing C, Bytyci R, Huber C, Morling N, Børsting C. The Danish STR sequence database: duplicate typing of 363 Danes with the ForenSeq™ DNA Signature Prep Kit. *Int J Leg Med*. 2019;133:325–334. https://doi.org/10.1007/s00414-018-1854-0.
21. Khubrani YM, Hallast P, Jobling MA, Wetton JH. Massively parallel sequencing of autosomal STRs and identity-informative SNPs highlights consanguinity in Saudi Arabia. *Forensic Sci. Int. Genet.* 2019;43, 102164. https://doi.org/10.1016/j.fsigen.2019.102164.
22. Delest A, Godfrin D, Chantrel Y, et al. Sequenced-based French population data from 169 unrelated individuals with Verogen's ForenSeq DNA signature prep kit. *Forensic Sci. Int. Genet.* 2020;47, 102304. https://doi.org/10.1016/j.fsigen.2020.102304.
23. Wu R, Peng D, Ren H, et al. Characterization of genetic polymorphisms in Nigerians residing in Guangzhou using massively parallel sequencing. *Forensic Sci. Int. Genet.* 2020;48, 102323. https://doi.org/10.1016/j.fsigen.2020.102323.
24. Guevara EK, Palo JU, King JL, et al. Autosomal STR and SNP characterization of populations from the northeastern Peruvian andes with the ForenSeq™ DNA signature prep kit. *Forensic Sci. Int. Genet.* 2021;52, 102487. https://doi.org/10.1016/j.fsigen.2021.102487.
25. Casals F, Rasal R, Anglada R, et al. A forensic population database in El Salvador: 58 STRs and 94 SNPs. *Forensic Sci. Int. Genet.* 2022;57, 102646. https://doi.org/10.1016/j.fsigen.2021.102646.
26. Aguilar-Velázquez JA, Á M, Duran-Salazar, et al. Characterization of 58 STRs and 94 SNPs with the ForenSeq™ DNA signature prep kit in Mexican-Mestizos from the Monterrey city (Northeast, Mexico). *Mol Biol Rep*. 2022;49:7601–7609. https://doi.org/10.1007/s11033-022-07575-y.
27. Davenport L, Devesse L, Syndercombe Court D, Ballard D. Forensic identity SNPs: characterisation of flanking region variation using massively parallel sequencing. *Forensic Sci. Int. Genet.* 2023;64, 102847. https://doi.org/10.1016/j.fsigen.2023.102847.
28. Peng D, Zhang Y, Ren H, et al. Identification of sequence polymorphisms at 58 STRs and 94 iiSNPs in a Tibetan population using massively parallel sequencing. *Sci Rep*. 2020;10, 12225. https://doi.org/10.1038/s41598-020-69137-1.
29. Chen C, Jin X, Zhang X, et al. Comprehensive insights into forensic features and genetic background of Chinese Northwest Hui group using six distinct categories of 231 molecular markers. *Front Genet*. 2021;12, 705753. https://doi.org/10.3389/fgene.2021.705753.
30. Fan H, Du Z, Wang F, et al. The forensic landscape and the population genetic analyses of Hainan Li based on massively parallel sequencing DNA profiling. *Int J Leg Med*. 2021;134:1295–1317. https://doi.org/10.1007/s00414-021-02590-3.
31. Simayijiang H, Morling N, Børsting C. Sequencing of human identification markers in an Uyghur population using the MiSeq FGx™ Forensic Genomics System. *Forensic Sci. Res.* 2022;7:154–162. https://doi.org/10.1080/20961790.2020.1779967.
32. Guo F, Liu Z, Long G, et al. High-resolution genotyping of 58 STRs in 635 northern han Chinese with MiSeq FGx ® forensic genomics system. *Forensic Sci. Int. Genet.* 2023;65, 102879. https://doi.org/10.1016/j.fsigen.2023.102879.
33. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol*. 2011;29:24–26. https://doi.org/10.1038/nbt.1754.
34. Parson W, Ballard D, Budowle B, et al. Massively parallel sequencing of forensic STRs: considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements. *Forensic Sci. Int. Genet.* 2016;22:54–63. https://doi.org/10.1016/j.fsigen.2016.01.009.
35. National Center for Biotechnology Information. *Database of Single Nucleotide Polymorphisms (dbSNP), dbSNP Build ID: 155*. Bethesda: National Library of Medicine; 2021. Available from: http://www.ncbi.nlm.nih.gov/SNP/.
36. Phillips C, Gettings KB, King JL, et al. The devil's in the detail": release of an expanded, enhanced and dynamically revised forensic STR Sequence Guide. *Forensic Sci. Int. Genet.* 2018;34:162–169. https://doi.org/10.1016/j.fsigen.2018.02.017.
37. Gouy A, Zieger M. STRAF-A convenient online tool for STR data evaluation in forensic genetics. *Forensic Sci. Int. Genet.* 2017;30:148–151. https://doi.org/10.1016/j.fsigen.2017.07.007.
38. Excoffier L, Lischer HEL. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour*. 2010;10:564–567. https://doi.org/10.1111/j.1755-0998.2010.02847.x.
39. Kidd KK, Speed WC. Criteria for selecting microhaplotypes: mixture detection and deconvolution. *Invest Genet*. 2015;6:1. https://doi.org/10.1186/s13323-014-0018-3.
40. R Core Team, R. *A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing; 2015 (Available from: https://www.r-project.org.
41. King JL, Churchill JD, Novroski NMM, et al. Increasing the discrimination power of ancestry- and identity-informative SNP loci within the ForenSeq™ DNA Signature Prep Kit. *Forensic Sci. Int. Genet.* 2018;36:60–76. https://doi.org/10.1016/j.fsigen.2018.06.005.
42. Vallone PM, Butler JM. AutoDimer: a screening tool for primer-dimer and hairpin structures. *Biotechniques*. 2004;37:226–231. https://doi.org/10.2144/04372ST03.
43. Li R, Wang Q, Yang J, et al. Comparison of three massively parallel sequencing platforms for single nucleotide polymorphism (SNP) genotyping in forensic genetics. *Int J Leg Med*. 2023;137:1361–1372. https://doi.org/10.1007/s00414-023-03035-9.
44. Apaga DL, Dennis SE, Salvador JM, Calacal GC, De Ungria MC. Comparison of two massively parallel sequencing platforms using 83 single nucleotide polymorphisms for human identification. *Sci Rep*. 2017;7:398. https://doi.org/10.1038/s41598-017-00510-3.
45. Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS. Truncated product method for combining P-values. *Genet Epidemiol*. 2002;22:170–185. https://doi.org/10.1002/gepi.0042.
46. Buckleton JS, Bright J, Curran JM, Taylor D. Validating databases. In: Buckleton JS, Bright J, Taylor D, eds. *Forensic DNA Evidence Interpretation*. second ed. Boca Raton: CRC Press; 2016:133–180.