genes

MDPI

Article

US Population Data for 94 Identity-Informative SNP Loci

Kevin M. Kiesler *, Lisa A. Borsuk, Carolyn R. Steffen, Peter M. Vallone and Katherine B. Gettings

National Institute of Standards and Technology, 100 Bureau Drive, Mailstop 8314, Gaithersburg, MD 20899, USA
* Correspondence: kevin.kiesler@nist.gov

Abstract: The US National Institute of Standards and Technology (NIST) analyzed a set of 1036 samples representing four major US population groups (African American, Asian American, Caucasian, and Hispanic) with 94 single nucleotide polymorphisms (SNPs) used for individual identification (iiSNPs). The compact size of iiSNP amplicons compared to short tandem repeat (STR) markers increases the likelihood of successful amplification with degraded DNA samples. Allele frequencies and relevant forensic statistics were calculated for each population group as well as the aggregate population sample. Examination of sequence data in the regions flanking the targeted SNPs identified additional variants, which can be combined with the target SNPs to form microhaplotypes (multiple phased SNPs within a short-read sequence). Comparison of iiSNP performance with and without flanking SNP variation identified four amplicons containing microhaplotypes with observed heterozygosity increases of greater than 15% over the targeted SNP alone. For this set of 1036 samples, comparison of average match probabilities from iiSNPs with the 20 CODIS core STR markers yielded an estimate of 1.7×10^{-38} for iiSNPs (assuming independence between all 94 SNPs), which was four orders of magnitude lower (more discriminating) than STRs where internal sequence variation was considered, and 10 orders of magnitude lower than STRs using established capillary electrophoresis length-based genotypes.

Keywords: single nucleotide polymorphism; human identification; microhaplotype; next generation sequencing

Citation: Kiesler, K.M.; Borsuk, L.A.; Steffen, C.R.; Vallone, P.M.; Gettings, K.B. US Population Data for 94 Identity-Informative SNP Loci. *Genes* **2023**, *14*, 1071. https://doi.org/10.3390/genes14051071

Academic Editor: Chiara Turchi

Received: 18 April 2023
Revised: 5 May 2023
Accepted: 9 May 2023
Published: 12 May 2023

1. Introduction

When highly compromised biological samples are encountered, DNA template molecules may be too short for efficient amplification of any region beyond 100 to 200 bases in length using standard polymerase chain reaction (PCR) amplification. The largest PCR amplicons are most adversely affected by fragmented or inhibited samples, leading to a characteristic downward trend in detection intensity with increasing amplicon size. For the most highly compromised DNA samples, few or no STR markers will amplify in a typical human identification (HID) multiplex. In such cases, markers with smaller PCR amplicons are better suited to HID applications [1–4]. Single nucleotide polymorphisms (SNPs) can be assayed with short (<150 bp) PCR fragment lengths, increasing the likelihood of successful amplification of sufficient loci for comparison with reference samples. Modern DNA sequencing methods make analysis of large SNP panels an attractive method for HID testing challenging samples such as highly degraded DNA or complex kinship cases [5–13].

Prior to the introduction of next-generation sequencing methods at the turn of the millennium, obtaining DNA sequences was laborious. For forensic applications, DNA sequences could be analyzed individually using Sanger sequencing [14], or small multiplexes of target SNPs could be assayed using single-base extension [5]. These methods relied on relatively low-throughput techniques such as slab gels or capillary electrophoresis for detection. However, in 2010, Ion Torrent Systems, Inc. (San Francisco, CA, USA) introduced the Ion Personal Genome Machine (PGM) [15], and in 2011, Illumina Inc. (San Diego, CA, USA) marketed the MiSeq instrument [16]. Both benchtop systems offered scale and cost amenable to forensic applications. Importantly, the increased scale of sequencing

parallelized reactions and detection enabled the generation of large quantities of DNA sequence information. To complement these sequencing platforms, commercial vendors have developed and validated assays with both traditional (STR) and newer (SNP) marker types for HID purposes [11]. A key aspect of introducing new markers in the legal framework for HID is access to appropriate population genetic data to formulate weight-of-evidence estimates given the proposition of a one-to-one match with an investigated sample and an individual [17–25]. The ForenSeq DNA Signature Prep Kit (Verogen, San Diego, CA, USA) is a large multiplex PCR assay designed for use with the MiSeq FGx (Verogen) to produce DNA sequences from both STR and SNP markers. The focus of this paper is the 94 identity-informative SNP markers (iiSNPs) present in the ForenSeq DNA Signature Prep Kit (both DNA Primer Mix A (DPMA) and B (DPMB)). The 94 iiSNP markers in the ForenSeq DNA Signature Prep Kit were originally proposed as separate panels in the literature [5,26,27] for identity informativeness because they occur in relatively balanced allelic proportions in many worldwide populations.

We present the sequencing results and allele frequencies for these 94 iiSNPs, and additional variation contained in amplified regions flanking the targeted SNPs, in four sets of samples representative of the US population (African American, Asian American, Caucasian, and Hispanic). This work facilitates statistical calculations in investigative DNA identification efforts where the populations characterized here are relevant.

2. Materials and Methods

2.1. Samples and Sequencing

Anonymous liquid blood samples with self-reported ancestries were purchased from Interstate Blood Bank (Memphis, TN, USA) and Millennium Biotech, Inc. (Ft. Lauderdale, FL, USA) or provided by DNA Diagnostics Center (Fairfield, OH, USA) as buccal swabs from paternity testing samples anonymous to NIST. A total of 1036 samples were included in this study, the same samples previously reported by our group [28–30], divided among four US populations (population names as designated during sample collection): African American (N = 342), Asian American (N = 97), Caucasian (N = 361), and Hispanic (N = 236). Throughout this paper, N = 1036 is used to reference the number of samples, whereas 2072 is the implied number of chromosomes. All work presented has been reviewed and approved by the NIST Research Protections Office under protocol # MML-16-0080.

Samples were analyzed as previously described [29] with the ForenSeq DNA Signature Prep Kit (Verogen, San Diego, CA, USA) used for library construction and sequencing on a MiSeq FGx instrument (Verogen). Modifications were made to the manufacturer's recommended procedure to improve profile recovery (e.g., increasing DNA input, see [29] for an extensive description of methods).

2.2. Data Analysis and Interpretation

2.2.1. Universal Analysis Software (UAS)

Primary data analysis was performed using Universal Analysis Software (UAS) version 1.3 (Verogen) [11]. SNP genotypes and sequencing coverage data were exported from the UAS for downstream evaluation in Excel (Microsoft, Redmond, WA, USA). Analytical and stochastic thresholds were applied in evaluating genotype calls as follows: greater than 30 reads were required for homozygous genotypes, whereas for heterozygous calls greater than 10 reads for each of two alleles were required. Allele coverage ratio (ACR), defined here as the number of sequencing reads from the allele with lower coverage divided by that of the allele with higher coverage, was required to be greater than 20% for heterozygous calls.

2.2.2. In-House Analysis

As an alternate approach to the UAS, an in-house method incorporating STRait Razor v3.0 (SR3) [31] was used to analyze the fastq files, and the resulting target SNP genotypes were compared with UAS genotypes. Regions of DNA concomitantly sequenced with the target SNP ("flanking regions") were also evaluated for additional polymorphisms. STRait

Razor relies on a configuration ("config") file to define the analyzed sequence regions. The config file used in this analysis was the default ForenSeq config file provided with SR3, ForenSeqv1.27.config (https://github.com/Ahhgust/STRaitRazor, accessed on 26 August 2022). Allele calling thresholds used with UAS data (above) were also applied to the results of this in-house analysis.

2.3. Calculation of Allele Frequencies and Forensic Statistics

The finalized SNP call set was used as input for STRAF 2.0 [32] for calculation of allele frequencies and forensic statistics: power of discrimination (PD), gene diversity (GD), probability of matching (PM), polymorphism information content (PIC), power of exclusion (PE), and typical paternity index (TPI).

2.4. Testing for Linkage Disequilibrium and Hardy–Weinberg Equilibrium

Arlequin 3.5 [33] was used for testing linkage disequilibrium with 10,000 permutations and two initial conditions for Expectation Maximization. Testing for Hardy–Weinberg equilibrium was performed in Arlequin with 1×10^6 Markov chain steps and 1×10^5 dememorization steps on a locus-by-locus basis.

2.5. Calculation of Effective Alleles (A_e)

The number of effective alleles for each locus was calculated using the equation: $A_e = 1/(p^2 + q^2)$, where p and q are the frequencies of the two allelic states for a biallelic SNP. For each locus, A_e was calculated for each population tested using population-specific allele frequencies from the dataset.

2.6. Calculation of Random-Match Probabilities (RMP)

Random-Match Probabilities for individual SNP profiles were calculated using frequency data from the 1000 Genomes Project [34] for the 94 iiSNPs in the ForenSeq DNA Signature Prep Kit, assuming independence between all markers. Similarly, RMPs were calculated for STR loci using either CE-based frequencies [28] or sequence-based frequencies [29]. For purposes of comparison, only the 20 CODIS Core STR loci were included in the STR-based RMP calculations. Individual RMP, average RMP, and standard deviation for the full 1036 sample set were calculated in Excel.

3. Results and Discussion

3.1. Call Rate, Sequencing Coverage, and Heterozygote Balance

A total of 97,384 genotype calls can be made for a population of 1036 samples at 94 loci. In the current study, 96,889 genotypes are reported after manual curation of the UAS genotype calls and comparison with an orthogonal analysis method (Section 3.4) for an overall genotype call rate of 99.5%. The remaining 495 genotypes failed to meet the minimum criteria described above, primarily due to insufficient sequencing coverage. Two loci, rs1031825 and rs1736442, had uncharacteristically high no-call rates (i.e., genotype dropout) of 19.1% and 18.6%, respectively, accounting for 391 of the 495 UAS dropouts. These two amplicons had the lowest average sequencing coverage of the 94 iiSNPs in this data set (Figure S1). Recent literature also notes these loci as low performing [25]. The remaining no-calls were distributed among 19 SNP loci, each with dropout rates of less than 1% of the 1036 genotypes (except for rs7041158, which had a dropout rate of 1.4%). Median sequencing coverage per locus was 553× and ranged from nearly 5000× to 50×, spanning two orders of magnitude (Figure S1). Care should be taken in interpreting sequencing coverage outcome here because modification of the manufacturer's recommended procedure, in the form of increased DNA input for low-performing samples, may have led to variability from sample to sample.

ACR was generally consistent across loci (Figure S1), with a mean of 0.86 ± 0.07 (range 0.43 to 0.92). Two loci had mean ACR values below 0.6 (rs6955448 and rs338882). These loci had the most skewed ACR distributions and were observed to have similarly

low ACR values in recent literature [21,25,35,36]. Locus rs6955448 exhibited lower average coverage for the alternate allele [T] (245.33× ± 123.54×) versus the reference allele [C] (573.68× ± 289.43×), resulting in an average ACR of 0.43 ± 0.09 in this study. Recent work by Davenport et al. [25] identified sequence variation in the primer binding region of locus rs6955448 as a likely cause of imbalanced sequencing coverage.

Locus rs338882 exhibited lower average coverage for the reference allele [C] (73.45× ± 41.82×) versus the alternate allele [T] (146.82× ± 81.47×), with an average ACR of 0.51 ± 0.13 (please note: the UAS version 1.3 reports this SNP locus on the reverse strand of the GRCh38 genome [C/T]; here we use the genotype given by UAS on the reverse strand to aid current UAS operators, whereas the GRCh38 forward strand genotype is [G/A]; Table S1 includes the UAS v1.3 strand orientation of each iiSNP). Investigation into imbalanced sequence coverage at locus rs338882 by King et al. [21] was unable to conclusively determine a causal factor.

3.2. Allele Frequencies, Forensic Statistics, Linkage Disequilibrium, and Hardy–Weinberg Calculations

Allele frequencies, forensic statistics, and linkage disequilibrium calculations were performed overall and by population for each of the 94 iiSNPs (Table S1). Reported allelic frequency values for 94 iiSNPs are consistent with the strand orientation output by the UAS version 1.3 (as noted in Table S1). The maximum sample size for frequency calculations is 2072 alleles for the full 1036 sample set; some loci had fewer reported alleles (Table S1) when quality metrics did not meet minimum allele calling thresholds.

Forensic statistics (Gene Diversity (GD), Polymorphism Information Content (PIC), Probability of Matching (PM), Power of Discrimination (PD), Observed Heterozygosity (Hobs), Power of Exclusion (PE), and Typical Paternity Index (TPI)) calculated in STRAF [32] are presented in Table S2. Forensic parameters are displayed graphically as boxplots for each population in Figure S2, allowing the reader to explore trends in the data where we observe some dispersion in values for a small number of loci. Upon closer examination of dispersed data points, no clear patterns emerged in terms of any one marker performing differently across all populations or all categories of parameters.

Testing for Linkage Disequilibrium in Arlequin returned 913 SNP pairs with exact T-test values below Arlequin's default significance level ($p < 0.05$). When performing numerous pairwise comparisons, it is expected that some level of type-1 errors (i.e., false positives) will be observed due to random chance when using a significance threshold suitable for discovering linkage disequilibrium indicators. This is exemplified by the observation that, of the 913 SNP pairs with *t*-test *p*-values < 0.05, only 36 pairs are physically located on the same chromosome (Table S3). After applying the conservative Bonferroni correction, only one syntenic pair remains statistically significant, rs1355366 and rs1357617 ($p < 0.0001$ only in the Caucasian category); these loci are physically separated by 190 Mb of DNA sequence on Chromosome 3.

Hardy–Weinberg equilibrium testing in Arlequin produced a small number of loci (between one and four, depending on population) with *p*-values < 0.05 (Table S4). Again, after applying the Bonferroni correction method, no *p*-values remained statistically significant.

3.3. Effective Alleles per Locus (A_e)

The number of effective alleles per locus is shown in Figure 1, with each data point representing the value within a population. Loci rs938283, rs1357617, and rs2056277 produced lower A_e values for all populations. The original literature from the SNPforID Consortium recommending these three loci for identity-informative purposes [5] characterized performance in population samples collected in Denmark, Greenland, Somalia, Turkey, China, Germany, Taiwan, Thailand, and Japan, where allele frequencies may differ from those studied here. Fluctuations in A_e values reflect slight allele frequency variation among populations from different regions of the globe. This highlights the importance of using appropriate population frequency data for weight-of-evidence calculations.

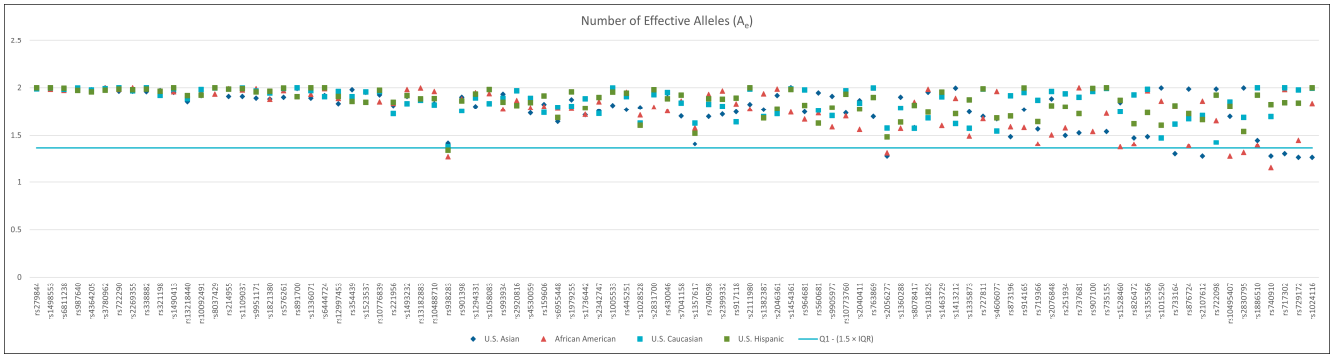

Figure 1. Number of effective alleles for 94 iiSNPs in four US populations. Each point on the plot represents the A_e value from one of four populations in the study. A reference line is drawn at the first quartile minus 1.5 times the interquartile range of A_e values.

3.4. Concordance with an Orthogonal Analysis Pipeline

Comparison of UAS genotype calls with those generated by an in-house analysis method incorporating SR3 yielded 97,222 concordant genotype calls from 97,384 possible genotypes (99.8% concordant). Genotypes that did not satisfy minimum thresholds by both analysis pipelines were considered concordant with the UAS calls (thus, the number of concordant genotypes is higher than the number of UAS genotype calls). The remaining 162 discordant genotypes were mostly no-calls by the SR3, where the presence of errors at various positions within sequencing reads excluded those reads from the count of reads corresponding to the 'correct' genotype call. This mechanism caused the read count to drop below allele calling thresholds in 154 cases. In eight instances, the SR3 analysis recovered genotypes that did not satisfy minimum thresholds in the UAS call set.

Close inspection of the genotypes from the two analyses revealed an incorrect UAS genotype call for one sample at SNP locus rs10092491. This sample was typed by the UAS as homozygous T, whereas the sequence analyzed by the in-house analysis resulted in a heterozygous T/C genotype. The UAS interpreted 109 of 196 reads (>50%) as deletions (Figure 2), which did not trigger any automatic flag to draw the attention of the reviewer. This target SNP (rs10092491 T > C) is adjacent to a homopolymer of up to four T bases containing a known low-frequency deletion (rs1306110296, frequency < 0.0001 in GnomAD [37]). The UAS produced a deletion call when the C allele of rs10092491 was present in combination with a deleted T from the homopolymer (Figure 2), presumably as an artifact of the alignment method used by the UAS. This genotype was adjusted in the final call set to reflect the correct heterozygote call (T/C) in this sample for downstream population genetic analysis.

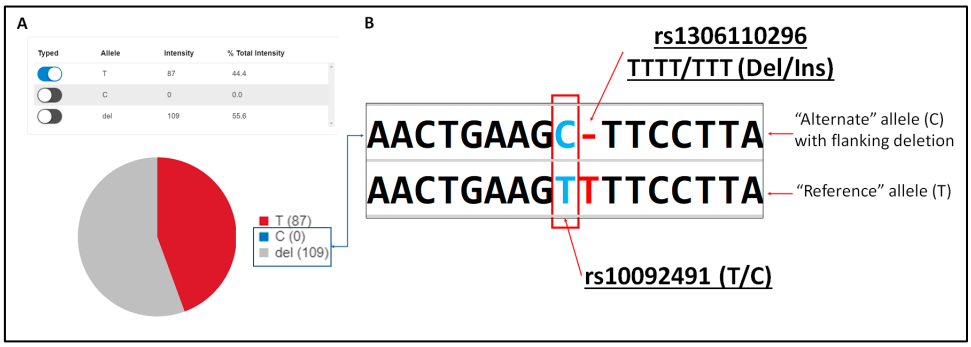

Figure 2. (**A**) Genotype of homozygous T called at rs10092491 produced by the UAS; 109 reads were interpreted as deletions while no reads were interpreted as C base calls, (**B**) sequence alignment showing a heterozygous C allele with deletion of one T base (rs1306110296) from the adjacent homopolymer stretch.

3.5. Microhaplotypes

The region containing SNP rs10092491 with an adjacent deletion represents a microhaplotype (MH, defined in the context of this study as any flanking sequence not matching the GRCH 38 reference genome, in combination with the UAS target SNP) obtained with the in-house analysis method. Further examination of the output of genotype calls from the in-house analysis indicated MHs of one or more flanking SNPs (plus the UAS target SNP) in 74 of the 94 iiSNP amplicons in the ForenSeq DNA Signature Prep Kit (Table S5). Of these 74 loci with MHs, 36 loci had one or more MH alleles labeled 'novel' (Table S6) in the output from the underlying SR3 database (downloaded 26 August 2022), indicating these MHs had not been observed in the original population study from which the SR3 database was established [19]. In total, 155 unique MH alleles accounted for 5198 flanking sequence variant observations, or 5.3% of 97,222 of the concordant genotype calls.

Of the 74 loci with microhaplotypes identified, four microhaplotypes associated with target SNP loci rs876724, rs1109037, rs10776839, and rs2830795 (Figure 3) exhibited the largest increases H_{obs} (15% to 26%) as compared to the target SNPs alone, consistent with the published literature [20,21]. Another eight microhaplotypes had increased H_{obs} from between 5% and 15%, while the remaining 62 loci with microhaplotypes exhibited low (<5%) or no increase in H_{obs}. Some microhaplotypes were observed at low frequency; 36 loci had singleton observations (accounting for 64 MH alleles) representing private variation which had negligible impact on H_{obs} in this study.

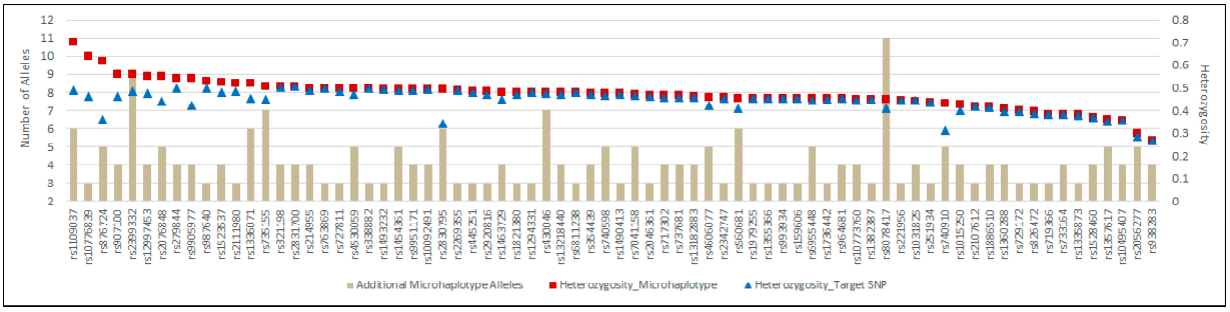

Figure 3. The additional number of alleles (left axis, bars) created by flanking sequence microhaplotypes and observed heterozygosity (right axis) when considering UAS target SNPs (triangles) versus microhaplotypes (squares) in 74 loci where flanking sequence variation was observed.

3.6. Calculation of Random-Match Probabilities

For the 1036 samples in the study (which were previously described for sequence-based STRs [29] and length-based STRs [28]), the RMP values calculated from the 94 iiSNPs (mean = 1.68×10^{-38}) were approximately four orders of magnitude lower (more discriminating) than the RMP values calculated from the 20 CODIS core sequence-based STRs (mean = 2.36×10^{-34}). Further, the 94 iiSNPs and sequence-based STRs outperformed length-based STR RMP calculations (mean = 1.02×10^{-28}) by ten and six orders of magnitude, respectively (Figure 4).

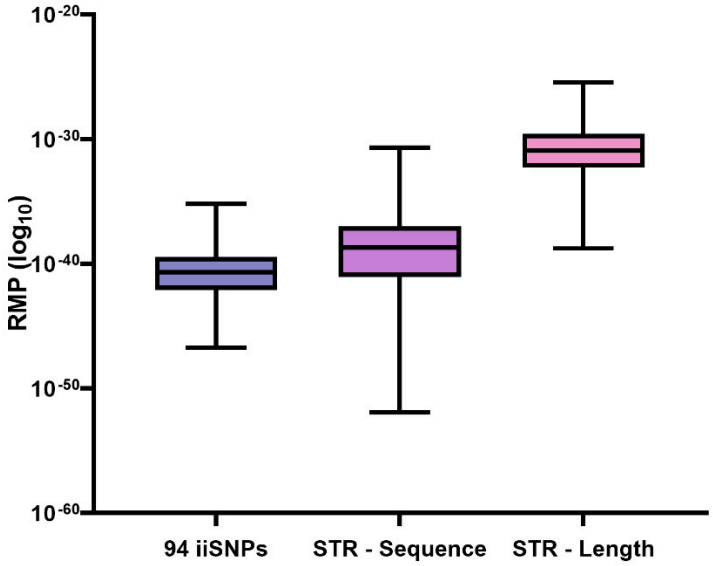

Figure 4. Boxplot of RMP values for the 1036 sample set with 94 iiSNPs, 20 CODIS sequence-based STRs, and 20 CODIS length-based STRs. The box represents the first and third quartiles with a line at the median value. Whiskers represent minimum and maximum values.

4. Conclusions

An overall call rate of 99.5% was observed for these 94 iiSNPs in this study of 1036 samples. While some variation in sequencing coverage was observed, modifications of the recommended procedure were made in the interest of improved profile recovery from underperforming DNA templates. Coverage metrics in this study may not be representative of routine use of the ForenSeq DNA Signature Prep Kit. The two SNP loci with the lowest average coverage, rs1031825 and rs1736442, produced most of the genotype dropouts (19.1% and 18.6% dropout rates, respectively). The assay exhibited mostly uniform coverage for heterozygous genotype calls (ACR). Two loci had frequently imbalanced allele coverage: rs6955448 and rs338882.

Forensic parameters were mainly consistent across all iiSNP markers in the multiplex, with some slight deviations observed in rare instances for specific populations. One locus, rs938283, exhibited minor skew in the number of effective alleles (A_e) in all populations in the study. Analysis of A_e values (Figure 1) across the 94 iiSNPs revealed some loci with allelic frequency variation among populations from different regions of the world.

Linkage disequilibrium analysis indicated signals of pairwise disequilibrium in 36 pairs of SNPs co-located on the same chromosome. Applying the Bonferroni correction eliminates all but one syntenic pair from statistical significance. This pair is physically separated by nearly the entirety of chromosome 3; therefore, it is unlikely this LD signal arises from a lack of recombination between these two loci. However, further exploration of the LD *p*-values reported here and comparison with results from other studies is encouraged prior to statistically combining these 94 iiSNPs, which arose from independently characterized panels.

Analysis with an alternate bioinformatic method brought to light a rare deletion (rs1306110296) in the short homopolymer adjacent to the target SNP locus rs10092491. This ostensibly caused misalignment by the UAS analysis algorithm, resulting in an incorrect homozygous genotype call. While this low-frequency deletion was observed in only one sample in this population study of 1036 individuals, the presence of a high number of reads interpreted as deletions could serve as an indicator of this or similar artifacts, which may not be automatically flagged by the UAS.

Flanking region variation comprising microhaplotypes within the iiSNP amplicons in the ForenSeq DNA Signature Prep Kit suggests the possibility of extracting more information from the kit's SNP content by analysis of the full sequence string. Microhaplotypes

associated with target SNP loci rs876724, rs1109037, rs10776839, and rs2830795, are likely to routinely provide additional information. Some flanking variation seen at low frequency may be of benefit to kinship analysis. The full complement of 94 biallelic iiSNPs produces extremely low random-match probabilities, making complex additional analyses beyond the UAS optional. The iiSNP random-match probabilities are comparable to values calculated for sequence-based alleles of STRs, both of which outperformed length-based STRs by several orders of magnitude in this study.

The NIST 1036 sample set was used here to characterize performance of the iiSNP content of the ForenSeq DNA Signature Prep Kit in terms of technical performance, population genetic parameters, and direct comparison with RMP values from STR markers by length and sequence. In the absence of corroborating genotype data for the 94 iiSNP loci, we used a parallel analysis method to improve the accuracy of genotype curation and to provide frequency information for sequence variation in flanking regions of the amplicons. The use of the NIST 1036 sample set in this study adds to the extensive forensic marker information available for these samples and maintains their relevance in facilitating technology transition.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/genes14051071/s1 and at the NIST Public Data Repository: US population data for human identification markers (doi:10.18434/t4/1500024); Figure S1: Sequencing metrics for 94 ForenSeq Signature Prep Kit identity-informative SNP loci in a population of 1036 samples; Figure S2: Forensic parameters (Genetic Diversity (GD), Polymorphism Information Content (PIC), Random-Match Probability (PM), Power of Discrimination (PD), Observed Heterozygosity (Hobs), Power of Exclusion (PE), and Typical Paternity Index (TPI)) shown in boxplot format for each of four populations studied ((A)African American, (B) Asian American, (C) Caucasian, and (D) Hispanic) [38,39]; Table S1: Allele Frequencies; Table S2: Forensic Parameters; Table S3: Linkage Disequilibrium; Table S4: Hardy–Weinberg Equilibrium; Table S5: Microhaplotype Frequencies; Table S6: Novel Microhaplotypes.

Author Contributions: Conceptualization, all authors; methodology, all authors; software, L.A.B. and K.M.K.; investigation, K.M.K., C.R.S., L.A.B. and K.B.G.; resources, P.M.V.; data curation, L.A.B., K.M.K. and K.B.G.; writing—original draft preparation, K.M.K.; writing—review and editing, K.B.G., P.M.V., L.A.B. and C.R.S.; visualization, K.M.K., L.A.B. and K.B.G.; supervision, P.M.V. and K.B.G.; project administration, P.M.V. and K.B.G.; funding acquisition, P.M.V. All authors have read and agreed to the published version of the manuscript.

Funding: NIST received funding to support this work through an interagency agreement with the Federal Bureau of Investigation: NIST IAA # DJF-19-1200-R000221.

Institutional Review Board Statement: All work presented has been reviewed and approved by the NIST Research Protections Office under protocol # MML-16-0080.

Informed Consent Statement: The National Institute of Standards and Technology Research Protections Office reviewed the protocol for this project and determined it is "not human subjects research" as defined in 15 CFR 27, the Common Rule for the Protection of Human Subjects. As part of the review, NIST ensured that participant consent was provided at the time of sample collection by the specimen provider.

Data Availability Statement: All SNP genotypes are available from the NIST Public Data Repository: US population data for human identification markers (doi:10.18434/t4/1500024).

Acknowledgments: This work was supported by NIST Special Programs Office: Forensic Genetics. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the US Department of Commerce. Certain commercial software, instruments, and materials are identified to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by NIST, nor does it imply that any of the materials, instruments, or equipment identified are necessarily the best available for the purpose.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Borsting, C.; Mogensen, H.S.; Morling, N. Forensic genetic SNP typing of low-template DNA and highly degraded DNA from crime case samples. *Forensic Sci. Int. Genet.* **2013**, *7*, 345–352. [CrossRef] [PubMed]
2. Gettings, K.B.; Kiesler, K.M.; Vallone, P.M. Performance of a next generation sequencing SNP assay on degraded DNA. *Forensic Sci. Int. Genet.* **2015**, *19*, 1–9. [CrossRef] [PubMed]
3. Zavala, E.I.; Rajagopal, S.; Perry, G.H.; Kruzic, I.; Basic, Z.; Parsons, T.J.; Holland, M.M. Impact of DNA degradation on massively parallel sequencing-based autosomal STR, iiSNP, and mitochondrial DNA typing systems. *Int. J. Legal. Med.* **2019**, *133*, 1369–1380. [CrossRef] [PubMed]
4. Elwick, K.; Bus, M.M.; King, J.L.; Chang, J.; Hughes-Stamm, S.; Budowle, B. Utility of the Ion S5 and MiSeq FGx sequencing platforms to characterize challenging human remains. *Leg. Med.* **2019**, *41*, 101623. [CrossRef] [PubMed]
5. Sanchez, J.J.; Phillips, C.; Borsting, C.; Balogh, K.; Bogus, M.; Fondevila, M.; Harrison, C.D.; Musgrave-Brown, E.; Salas, A.; Syndercombe-Court, D.; et al. A multiplex assay with 52 single nucleotide polymorphisms for human identification. *Electrophoresis* **2006**, *27*, 1713–1724. [CrossRef]
6. Pakstis, A.J.; Speed, W.C.; Kidd, J.R.; Kidd, K.K. Candidate SNPs for a universal individual identification panel. *Hum. Genet.* **2007**, *121*, 305–317. [CrossRef]
7. Boonyarit, H.; Mahasirimongkol, S.; Chavalvechakul, N.; Aoki, M.; Amitani, H.; Hosono, N.; Kamatani, N.; Kubo, M.; Lertrit, P. Development of a SNP set for human identification: A set with high powers of discrimination which yields high genetic information from naturally degraded DNA samples in the Thai population. *Forensic Sci. Int. Genet.* **2014**, *11*, 166–173. [CrossRef]
8. Warshauer, D.H.; Davis, C.P.; Holt, C.; Han, Y.; Walichiewicz, P.; Richardson, T.; Stephens, K.; Jager, A.; King, J.; Budowle, B. Massively parallel sequencing of forensically relevant single nucleotide polymorphisms using TruSeq forensic amplicon. *Int. J. Legal. Med.* **2015**, *129*, 31–36. [CrossRef]
9. Churchill, J.D.; Chang, J.; Ge, J.; Rajagopalan, N.; Wootton, S.C.; Chang, C.W.; Lagace, R.; Liao, W.; King, J.L.; Budowle, B. Blind study evaluation illustrates utility of the Ion PGM system for use in human identity DNA typing. *Croat. Med. J.* **2015**, *56*, 218–229. [CrossRef]
10. Grandell, I.; Samara, R.; Tillmar, A.O. A SNP panel for identity and kinship testing using massive parallel sequencing. *Int. J. Legal. Med.* **2016**, *130*, 905–914. [CrossRef]
11. Jager, A.C.; Alvarez, M.L.; Davis, C.P.; Guzman, E.; Han, Y.; Way, L.; Walichiewicz, P.; Silva, D.; Pham, N.; Caves, G.; 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. [CrossRef]
12. Fattorini, P.; Previdere, C.; Carboni, I.; Marrubini, G.; Sorcaburu-Cigliero, S.; Grignani, P.; Bertoglio, B.; Vatta, P.; Ricci, U. Performance of the ForenSeq(TM) DNA Signature Prep kit on highly degraded samples. *Electrophoresis* **2017**, *38*, 1163–1174. [CrossRef]
13. Ballard, D.; Winkler-Galicki, J.; Wesoly, J. Massive parallel sequencing in forensics: Advantages, issues, technicalities, and prospects. *Int. J. Legal. Med.* **2020**, *134*, 1291–1303. [CrossRef]
14. Sanger, F.; Donelson, J.E.; Coulson, A.R.; Kossel, H.; Fischer, D. Use of DNA polymerase I primed by a synthetic oligonucleotide to determine a nucleotide sequence in phage fl DNA. *Proc. Natl. Acad. Sci. USA* **1973**, *70*, 1209–1213. [CrossRef]
15. Rothberg, J.M.; Hinz, W.; Rearick, T.M.; Schultz, J.; Mileski, W.; Davey, M.; Leamon, J.H.; Johnson, K.; Milgrew, M.J.; Edwards, M.; et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **2011**, *475*, 348–352. [CrossRef]
16. Bentley, D.R.; Balasubramanian, S.; Swerdlow, H.P.; Smith, G.P.; Milton, J.; Brown, C.G.; Hall, K.P.; Evers, D.J.; Barnes, C.L.; Bignell, H.R.; et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **2008**, *456*, 53–59. [CrossRef]
17. Calafell, F.; Anglada, R.; Bonet, N.; Gonzalez-Ruiz, M.; Prats-Munoz, G.; Rasal, R.; Lalueza-Fox, C.; Bertranpetit, J.; Malgosa, A.; Casals, F. An assessment of a massively parallel sequencing approach for the identification of individuals from mass graves of the Spanish Civil War (1936–1939). *Electrophoresis* **2016**, *37*, 2841–2847. [CrossRef]
18. Casals, F.; Anglada, R.; Bonet, N.; Rasal, R.; van der Gaag, K.J.; Hoogenboom, J.; Sole-Morata, N.; Comas, D.; Calafell, F. Length and repeat-sequence variation in 58 STRs and 94 SNPs in two Spanish populations. *Forensic Sci. Int. Genet.* **2017**, *30*, 66–70. [CrossRef]
19. Churchill, J.D.; Novroski, N.M.M.; King, J.L.; Seah, L.H.; 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. [CrossRef]
20. Wendt, F.R.; King, J.L.; Novroski, N.M.M.; Churchill, J.D.; Ng, J.; Oldt, R.F.; McCulloh, K.L.; Weise, J.A.; Smith, D.G.; Kanthaswamy, S.; 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. [CrossRef]
21. King, J.L.; Churchill, J.D.; Novroski, N.M.M.; Zeng, X.; Warshauer, D.H.; Seah, L.H.; Budowle, B. 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. [CrossRef] [PubMed]
22. Khubrani, Y.M.; Hallast, P.; Jobling, M.A.; Wetton, J.H. Massively parallel sequencing of autosomal STRs and identity-informative SNPs highlights consanguinity in Saudi Arabia. *Forensic Sci. Int. Genet.* **2019**, *43*, 102164. [CrossRef]

23. Delest, A.; Godfrin, D.; Chantrel, Y.; Ulus, A.; Vannier, J.; Faivre, M.; Hollard, C.; Laurent, F.X. Sequenced-based French population data from 169 unrelated individuals with Verogen's ForenSeq DNA signature prep kit. *Forensic Sci. Int. Genet.* **2020**, *47*, 102304. [CrossRef] [PubMed]
24. Guevara, E.K.; Palo, J.U.; King, J.L.; Bus, M.M.; Guillen, S.; Budowle, B.; Sajantila, A. 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. [CrossRef] [PubMed]
25. 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. [CrossRef]
26. Pakstis, A.J.; Speed, W.C.; Fang, R.; Hyland, F.C.; Furtado, M.R.; Kidd, J.R.; Kidd, K.K. SNPs for a universal individual identification panel. *Hum. Genet.* **2010**, *127*, 315–324. [CrossRef]
27. Kidd, K.K.; Kidd, J.R.; Speed, W.C.; Fang, R.; Furtado, M.R.; Hyland, F.C.; Pakstis, A.J. Expanding data and resources for forensic use of SNPs in individual identification. *Forensic Sci. Int. Genet.* **2012**, *6*, 646–652. [CrossRef]
28. Steffen, C.R.; Coble, M.D.; Gettings, K.B.; Vallone, P.M. Corrigendum to 'U.S. Population Data for 29 Autosomal STR Loci'. *Forensic Sci. Int. Genet.* **2017**, *31*, e36–e40. [CrossRef]
29. Gettings, K.B.; Borsuk, L.A.; Steffen, C.R.; Kiesler, K.M.; Vallone, P.M. Sequence-based U.S. population data for 27 autosomal STR loci. *Forensic Sci. Int. Genet.* **2018**, *37*, 106–115. [CrossRef]
30. Steffen, C.R.; Huszar, T.I.; Borsuk, L.A.; Vallone, P.M.; Gettings, K.B. A multi-dimensional evaluation of the 'NIST 1032' sample set across four forensic Y-STR multiplexes. *Forensic Sci. Int. Genet.* **2022**, *57*, 102655. [CrossRef]
31. Woerner, A.E.; King, J.L.; Budowle, B. Fast STR allele identification with STRait Razor 3.0. *Forensic Sci. Int. Genet.* **2017**, *30*, 18–23. [CrossRef]
32. Gouy, A.; Zieger, M. STRAF-A convenient online tool for STR data evaluation in forensic genetics. *Forensic Sci. Int. Genet.* **2017**, *30*, 148–151. [CrossRef]
33. Excoffier, L.; Lischer, H.E. 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. [CrossRef]
34. Genomes Project, C.; Auton, A.; Brooks, L.D.; Durbin, R.M.; Garrison, E.P.; Kang, H.M.; Korbel, J.O.; Marchini, J.L.; McCarthy, S.; McVean, G.A.; et al. A global reference for human genetic variation. *Nature* **2015**, *526*, 68–74. [CrossRef]
35. Apaga, D.L.; Dennis, S.E.; Salvador, J.M.; Calacal, G.C.; De Ungria, M.C. Comparison of Two Massively Parallel Sequencing Platforms using 83 Single Nucleotide Polymorphisms for Human Identification. *Sci. Rep.* **2017**, *7*, 398. [CrossRef]
36. Guo, F.; Yu, J.; Zhang, L.; Li, J. Massively parallel sequencing of forensic STRs and SNPs using the Illumina((R)) ForenSeq DNA Signature Prep Kit on the MiSeq FGx Forensic Genomics System. *Forensic Sci. Int. Genet.* **2017**, *31*, 135–148. [CrossRef]
37. Karczewski, K.J.; Francioli, L.C.; Tiao, G.; Cummings, B.B.; Alfoldi, J.; Wang, Q.; Collins, R.L.; Laricchia, K.M.; Ganna, A.; Birnbaum, D.P.; et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **2020**, *581*, 434–443. [CrossRef]
38. Tukey, J.W. *Exploratory Data Analysis*; Addison-Wesley: Reading, MA, USA, 1977.
39. McGill, R.; Tukey, J.W.; Larsen, W.A. Variations of box plots. *Am. Stat.* **1978**, *32*, 12–16.