Population Data for 94 Human Identity SNPs in Four U.S. Population Groups

Kevin M. Kiesler, Lisa A. Borsuk, Katherine B. Gettings, Carolyn R. Steffen, and Peter M. Vallone[1]

U.S. National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899-8314, USA

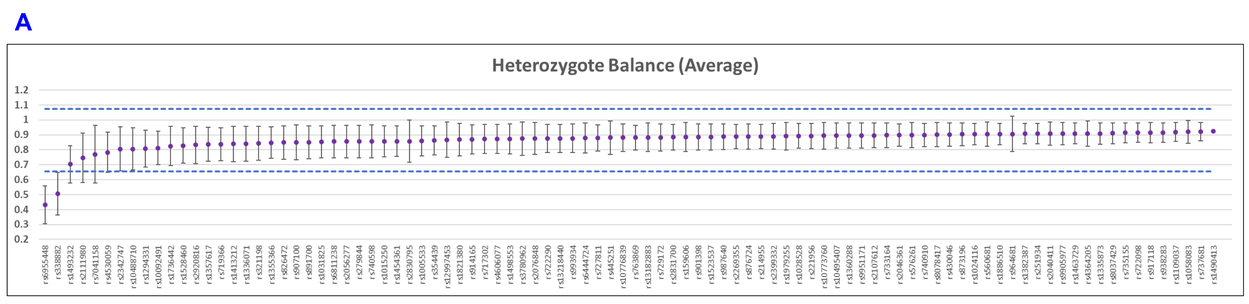
Email: Kevin.Kiesler@nist.gov

Poster # P113

The U.S. National Institute of Standards and Technology (NIST) sequenced 1036 human DNA samples from four United States population groups (African American, Asian, Hispanic, and Caucasian) using the ForenSeq DNA Signature Prep Kit (Verogen, San Diego CA, USA) with Primer Mix B (DPMB) as previously described [1]. DNA sequencing was performed on a MiSeq FGx instrument (Verogen). In addition to STR markers, DPMB includes amplification primers for single nucleotide polymorphisms (SNPs) used for individual identification (iiSNPs, n = 94), ancestry inference (aiSNPs, n = 56), and phenotype prediction (piSNPs, n = 22). Resulting sequencing coverage information was interpreted for the 94 iiSNP markers. Allele frequencies and relevant forensic statistics were calculated for each population group. Here we present match probabilities computed from the 94 iiSNPs compared to those derived from 27 autosomal STR loci using sequence-based and length-based allele frequencies [1,2]. Variations in forensic statistics by population group are also explored.

Technical Performance

A



B

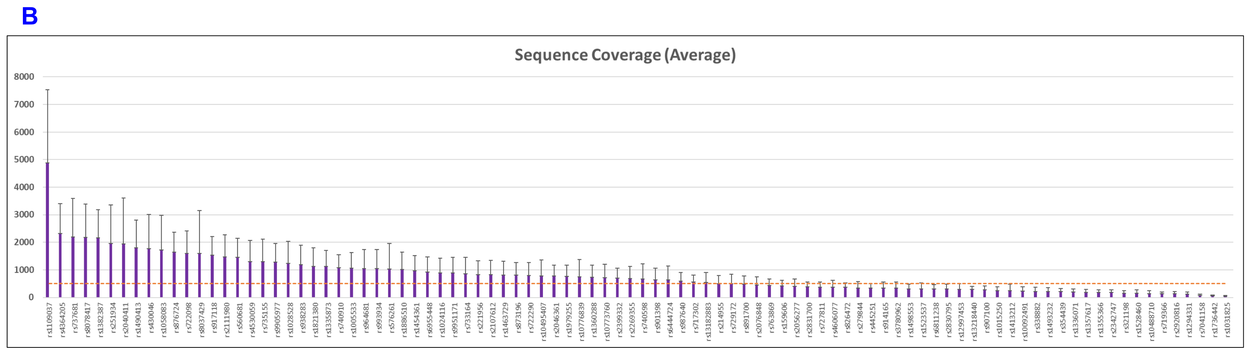


Figure 1: Sequencing metrics: A) heterozygote balance; the number of reads for the allele with lower coverage divided by the number of reads for the allele with higher coverage. Dashed blue lines represent three standard deviations from the mean value (mean = 0.86). Two loci (rs6955448 and rs338882) fell outside the region bounded by three standard deviations. B) sequencing coverage; the average number of reads reported by the Universal Analysis Software (Verogen, San Diego, USA) for each locus. Median coverage was 506 ×, represented by the red dashed line. Median coverage per locus spans two orders of magnitude.

Match Probabilities for SNPs and STRs

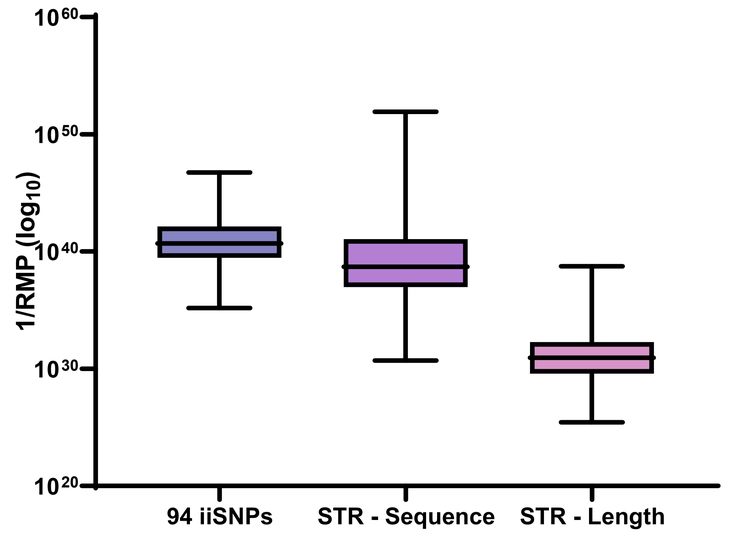


Figure 2: Boxplot of Random Match Probability (RMP, presented as inverse 1/RMP with higher values denoting lower probability of a random match) showing values calculated without theta correction for 1036 samples using three marker systems: 94 iiSNPs found in the ForenSeq Signature Kit* (SNP population frequencies from 1000 genomes project [3]), 20 CODIS core STR markers using sequence-based frequencies [1] and 20 CODIS core STR markers using allele frequencies from capillary electrophoresis (CE) amplicon length measurements [2]. Boxplot outline represents the first and third quartiles with the central line depicting the median datapoint with whiskers showing the minimum and maximum values in the data set. The RMPs for the 94 iiSNPs exhibited a similar range of values as the 20 CODIS core sequence-based STRs. These values contrast with RMPs from length-based STRs that differ by approximately 10 orders of magnitude.

*Note that the 94 iiSNPs in the ForenSeq Signature Prep Kit may exhibit signals of pairwise linkage disequilibrium. Excluding a small number of SNPs is expected to have a minimal impact on these 1/RMP estimates.

Population Genetic Analysis

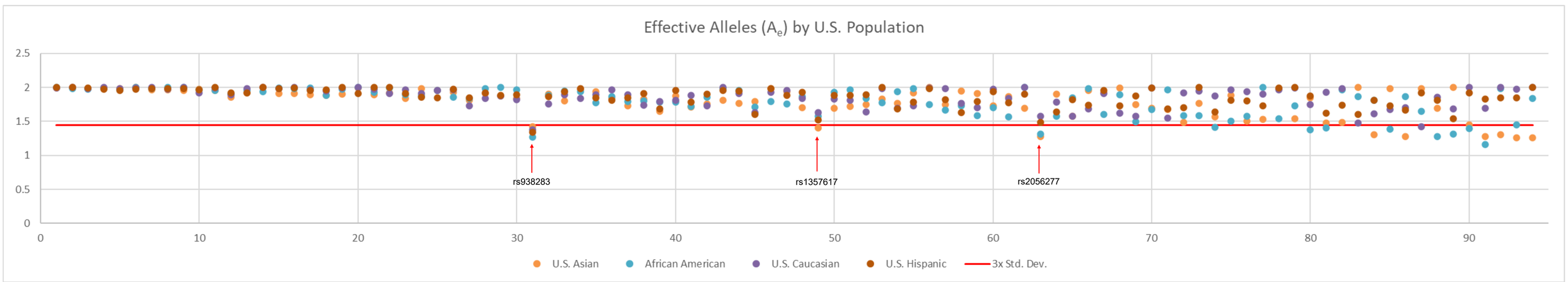
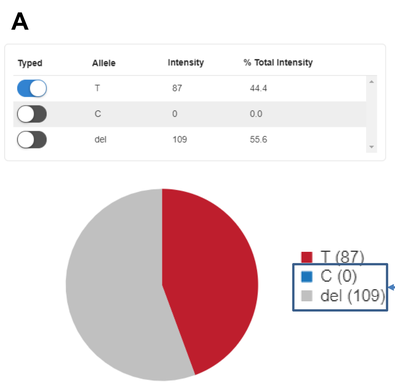


Figure 3: Effective Number of Alleles (A_e), given by the equation $A_e = 1/(p^2 + q^2)$, for a biallelic identity informative SNP has an ideal value of two; meaning that each allele has equal incidence within the population measured, e.g. $1/(0.5^2 + 0.5^2) = 2$. SNP loci at the left of the histogram are performing as expected for identity informative markers. Loci towards the right side of the figure are skewed in allele frequency for at least one population. A reference line is drawn (red) at three standard deviations of the A_e data. Three loci, rs938283, rs1357617, and rs2056277 (marked with arrows) exhibited consistently skewed allele frequency across all populations.

Rare Deletion Creates Genotyping Artifact

A



B

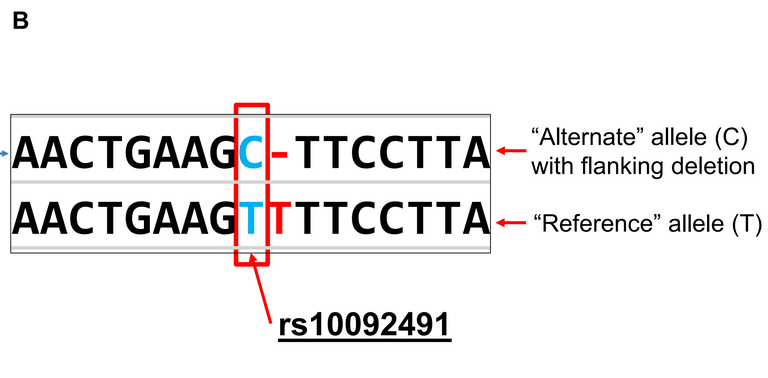
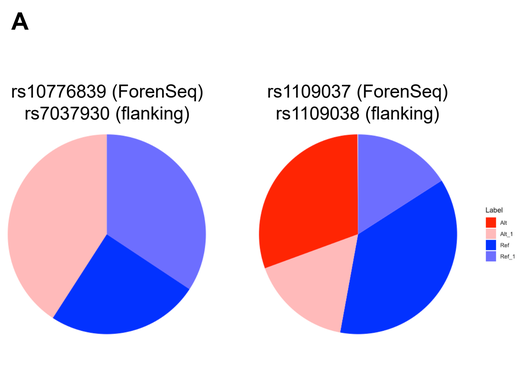


Figure 4: Rare deletion observed once in the 1036 data set, adjacent to locus rs10092491, causes misassignment of the correct ‘alternate’ allele. (A) Universal Analysis Software (UAS) produces incorrect genotype (del) due to (B) deletion of T residue in an adjacent homopolymer stretch. The correct “Alternate” allele, C, is present in the sequence, suggesting that the UAS genotyping algorithm could be the root cause of the mis-called base.

SNPs in flanking regions

A



B

SNP			Microhaplotype			
Label	Frequency	A_e	Label	Frequency	A_e	% Change
Locus rs10776839 G>T			rs10776839 G>T, rs7037930 A>G			
Alt T	0.41	1.93	Alt T+A	0.00	2.89	49%
			Alt_1 T+G	0.41		
Ref G	0.59		Ref G+A	0.25		
			Ref_1 G+G	0.34		
Locus rs1109037 G>A			rs1109037 G>A, rs1109038 G>A			
Alt A	0.47	1.99	Alt A+G	0.30	3.54	78%
			Alt_1 A+A	0.17		
Ref G	0.53		Ref G+G	0.37		
			Ref_1 G+A	0.16		

Figure 5: Sequence data was analyzed using STRaitRazor 3.0 [4], a freely-available software, to characterize additional polymorphisms in regions flanking the target iiSNP in ForenSeq amplicons. Several (n > 20) loci exhibited microhaplotypes which could increase the performance of the locus. To illustrate, two examples (A) are shown where a flanking SNP creates a microhaplotype. (B) Locus rs10776839 contains a flanking SNP which increases the measure of effective alleles (A_e) by 49 %, while rs1109037 has a flanking SNP which increases A_e by 78 % relative to using the UAS genotypes alone.

References

[1] Gettings K.B., Borsuk L.A., Steffen C.R., Kiesler K.M., Vallone P.M. (2018) Sequence-based US population data for 27 autosomal STR loci. Forensic Sci Int Genet 37:106-115.

[2] Steffen C.R., Coble M.D., Gettings K.B., Vallone P.M. (2017) Corrigendum to 'U.S. Population Data for 29 Autosomal STR Loci' [Forensic Sci. Int. Genet. 7 (2013) e82-e83]. Forensic Sci Int Genet. 31:e36-e40.

[3] The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature 526, 68–74 (2015). https://doi.org/10.1038/nature15393

[4] Woerner AE, King JL, Budowle B. Fast STR allele identification with STRait Razor 3.0. Forensic Sci Int Genet. 2017 Sep;30:18-23. doi: 10.1016/j.fsigen.2017.05.008. Epub 2017 Jun 1. PMID: 28605651.

Download a PDF copy of this poster:

Dancing bear credit: www.Telegraph.co.uk

Disclaimer

Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Commerce. Certain commercial software, instruments, and materials are identified in order 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.

Research Protections

All work has been reviewed and approved by the National Institute of Standards and Technology Research Protections Office. This study was determined to be “not human subjects research” (often referred to as research not involving human subjects) as defined in U. S. Department of Commerce Regulations, 15 CFR 27, also known as the Common Rule (45 CFR 46, Subpart A), for the Protection of Human Subjects by the NIST Human Research Protections Office and therefore not subject to oversight by the NIST Institutional Review Board.

Funding

This work was in part supported by the NIST Special Programs Office: Forensic Genetics.

This work was in part supported by an interagency agreement with the FBI: DNA as a Biometric.