Population Genetic Analysis with the NimaGen IDseek OmniSNP Identity-Informative SNP Typing Kit

Sammed N. Mandape[a], Melissa Muenzler[a], Magdalena M. Bus[a,b], Jonathan L. King[a], Jennifer C. Cihlar[a,b]

[a]Center for Human Identification, University of North Texas Health Science Center, 3500 Camp Bowie Blvd., Fort Worth, TX 76107, USA

[b]Department of Microbiology, Immunology, and Genetics, University of North Texas Health Science Center; 3500 Camp Bowie Blvd., Fort Worth, TX 76107, USA

Introduction

DNA analysis to genotype identity-informative areas of the human genome is one of the most important tools in the field of human identification. Massively parallel sequencing (MPS) has improved upon capillary-electrophoresis-based technology by offering higher throughput and the ability to multiplex different samples and types of markers in larger numbers. While targeting short tandem repeats (STRs) is the mainstay of forensic casework due to their high power of discrimination, issues like degraded, low quality, or low copy number samples can make amplification of STRs inefficient. Forensic samples are often exposed to various insults such as UV light, oxidation, and microbial activity that can degrade DNA fragments.

Samples which are highly degraded or damaged often do not contain sufficient DNA for successful PCR amplification and produce either no or only a partial STR profile. In such cases, targeting markers in smaller amplicons, such as single nucleotide polymorphisms (SNPs) can overcome this limitation. A sufficient number of identity informative SNPs (e.g., >50) may produce an identification profile of similar specificity to those of STRs.

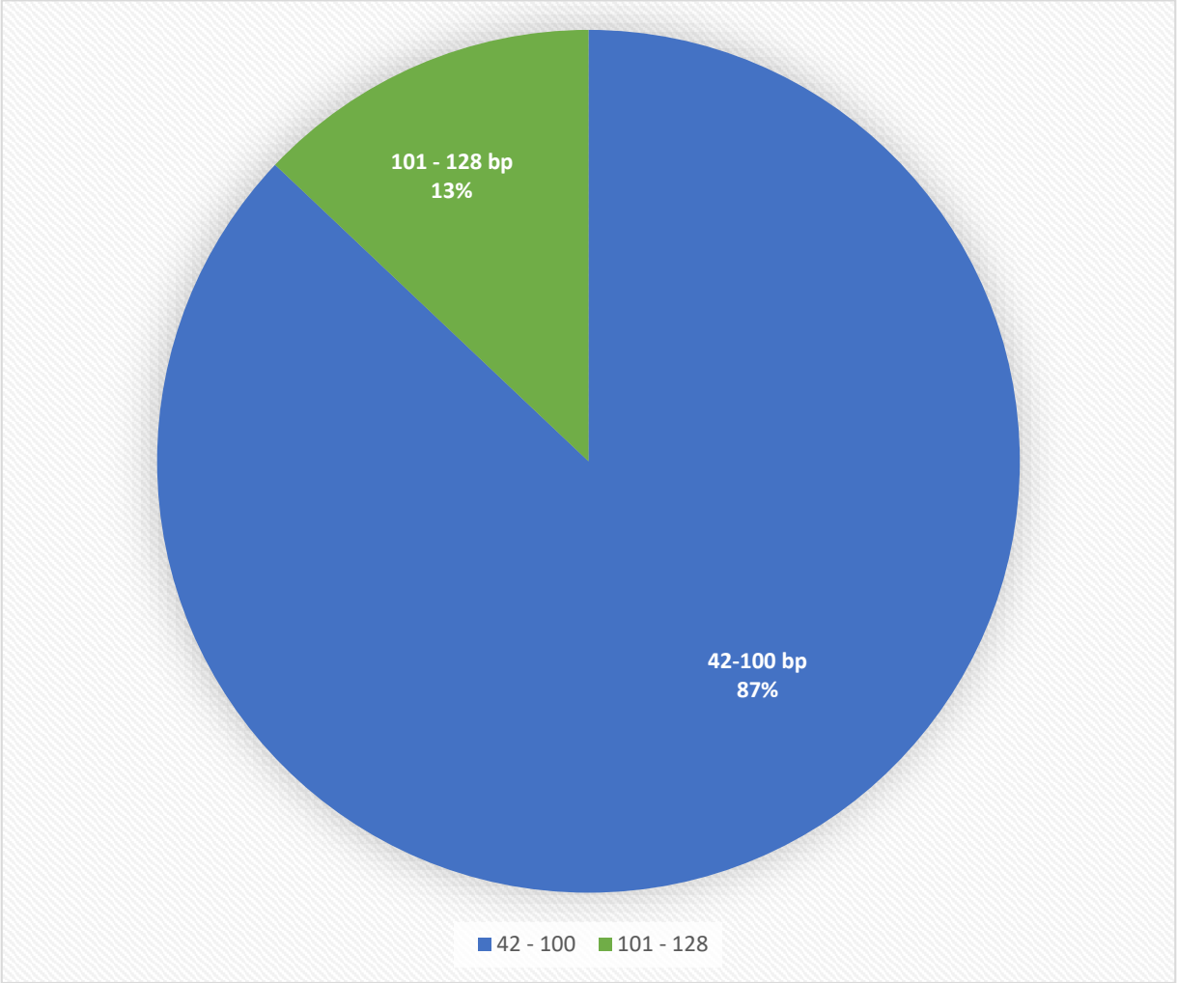

Figure 1. The majority of the SNP amplicons in the new 85-plex RC-PCR panel are ≤ 100 bp in length.

The IDseek OmniSNP Identity Informative SNP Typing Kit targets 85 identity-informative SNPs (iiSNPs) in amplicons no longer than 128 base pairs, with 87% of the SNPs found in amplicons less than 100 base pairs (Figure 1). Additionally, the entire process of sample indexing, adapter ligation, and amplification occurs in a single-tube, reverse complement PCR reaction, minimizing the number of handling steps and thus opportunities for sample loss or contamination.

Materials and Methods

Samples

Whole blood samples were obtained from 144 presumably unrelated individuals representing three major population groups (US Caucasian, CAU, N = 48; Southwest Hispanic, HIS, N = 48; African American, AFA, N = 48).

RC-PCR Library Preparation and Illumina Sequencing

A total of 144 libraries were prepared from the three population groups using one nanogram of DNA per sample. Indexing, Illumina adapter ligation and amplification were performed according to the manufacturer's recommendations. After amplification, samples were pooled in equal volumes and underwent clean up with AMPure XP size-selection beads.

Pooled libraries were diluted to 9 pM and sequenced on the Illumina MiSeq FGx desktop sequencer using the MiSeq FGx Standard Reagent Kit with a read length of 2 x 121. Samples were sequenced over three runs with 48 samples each and a fourth run with 40 samples which included libraries for resequencing and controls.

Data Analysis

The FASTQ files were analyzed with STRait Razor Online v.0.1.7. Forensic parameters such as discrimination power (PD), polymorphism information content (PIC), and heterozygosity (expected, He and observed, Ho) were computed using in-house developed R script (v4.1.0). Additionally, allele frequencies, effective number of alleles (Ae), heterozygote balance (HB), and combined Random Match Probability (cRMP) were estimated. RStudio (v2023.3.0.386) was used for bioinformatics analyses with plots rendered using ggplot2. Genetic Data Analysis 1.1 was used to calculate Fishers exact tests for pairwise linkage disequilibrium (LD) and Hardy-Weinberg equilibrium (HWE) at individual loci. The estimates were obtained using 3200 shuffling tests of the exact significance levels. P-values were corrected by Bonferroni. To further investigate any flanking region variants, raw FASTQ files were aligned to the human reference genome build 38 (GRCh38) using BWA-MEM algorithm. The BAM files were visualized with Integrative Genomics Viewer.

Results

Genotype calls generated by the OmniSNP panel compared to those previously published were 99.9% concordant. Fourteen loci that were heterozygous in OmniSNP panel were found to be homozygous in the original ForenSeq DNA Signature Prep Kit data. Additionally, there was one instance each of locus drop out and drop in. Average sequencing read depth by locus ranged from 182.8x (SD = 139.2x) to 12091.2x (SD = 5351.9x) with a mean value of 2664.8x (green dashed line) and a median value of 2137.1x (Figure 2).

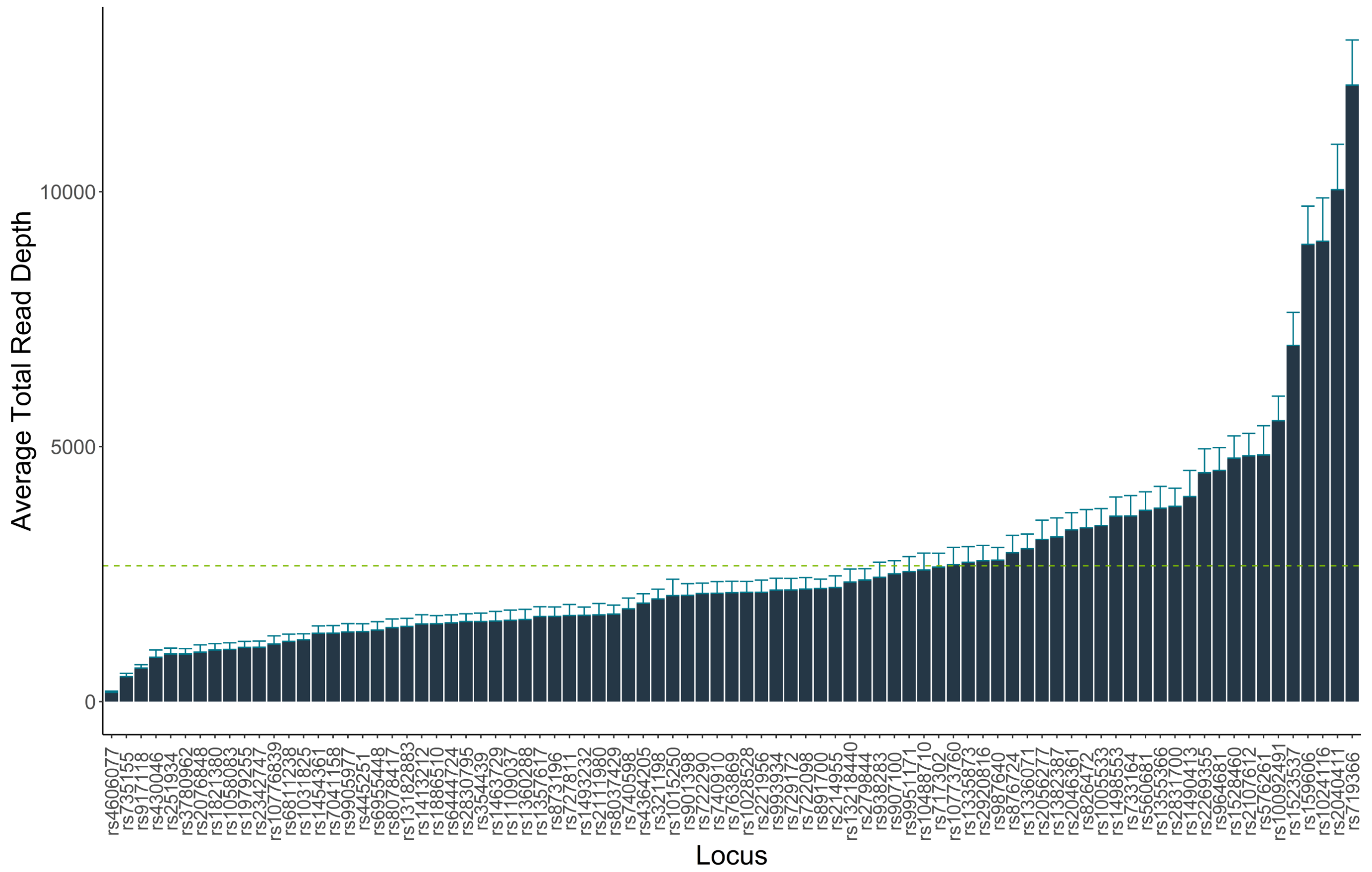

Figure 2. Average read depth by locus. Green dashed line indicates mean read depth.

The minor allele frequencies for all loci by population were greater than 0.1, except locus rs740910 that had a value of 0.08 in the AFA population. All the other loci showed relatively uniform distribution, highlighting their highly polymorphic nature (Figure 3).

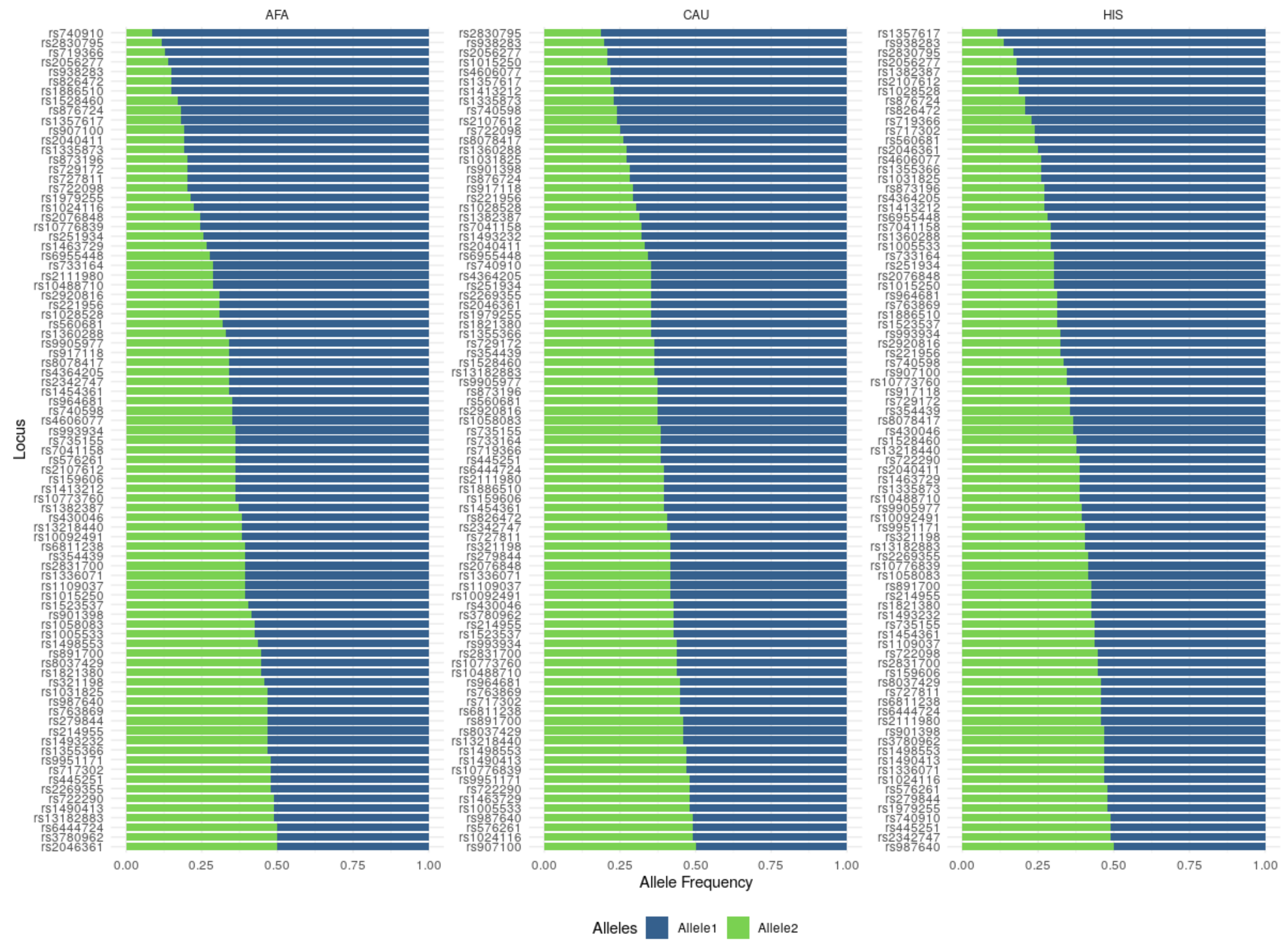

Figure 3. Allele frequency by locus.

Average heterozygosities, including both observed (Ho) and expected (He), were similar across all SNPs in the three populations: Ho = 0.44, 0.45, 0.43; He = 0.43, 0.45, 0.44; for AFA, CAU, and HIS, respectively (Figure 4). The lowest Ho (0.17) and He (0.15) was detected in locus rs740910 in the AFA population, whereas CAU and HIS population groups exhibited: Ho = 0.53, 0.40 and He = 0.45, 0.50, respectively, at the same locus. The highest Ho and He were detected in locus rs10776839 and were at similar levels in all populations: AFA Ho = 0.66, He = 0.62; CAU Ho = 0.68, He = 0.63; and HIS Ho = 0.70, He = 0.65.

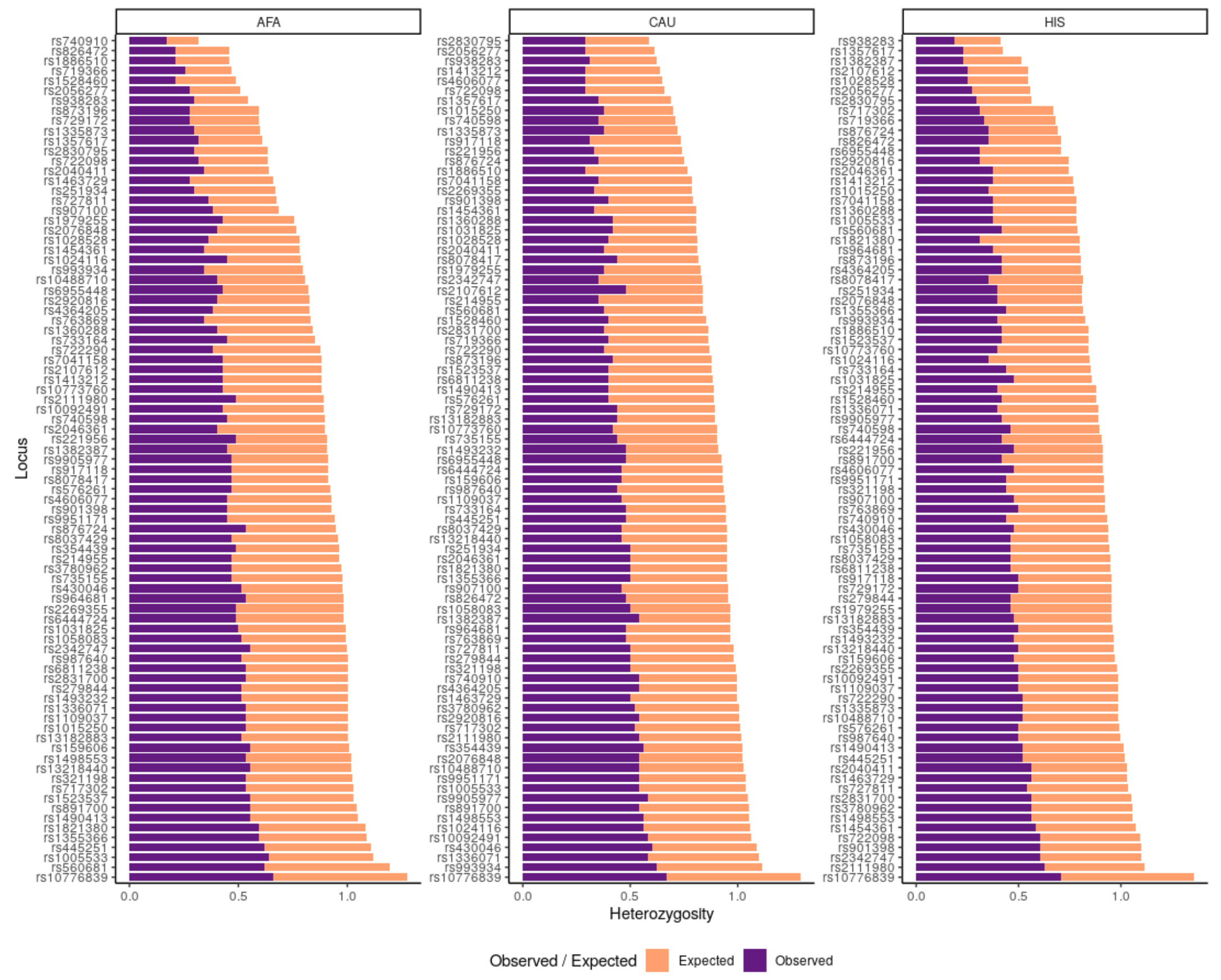

Figure 4. Average heterozygosities by locus and population

The results of sequence data analyses in population samples suggest that 81 SNPs display robust heterozygote balance and four SNP loci - rs430046, rs1490413, rs1015250, and rs9071100 showed allele imbalance (see Figure 5). Heterozygote imbalance was defined operationally as HB <0.5.

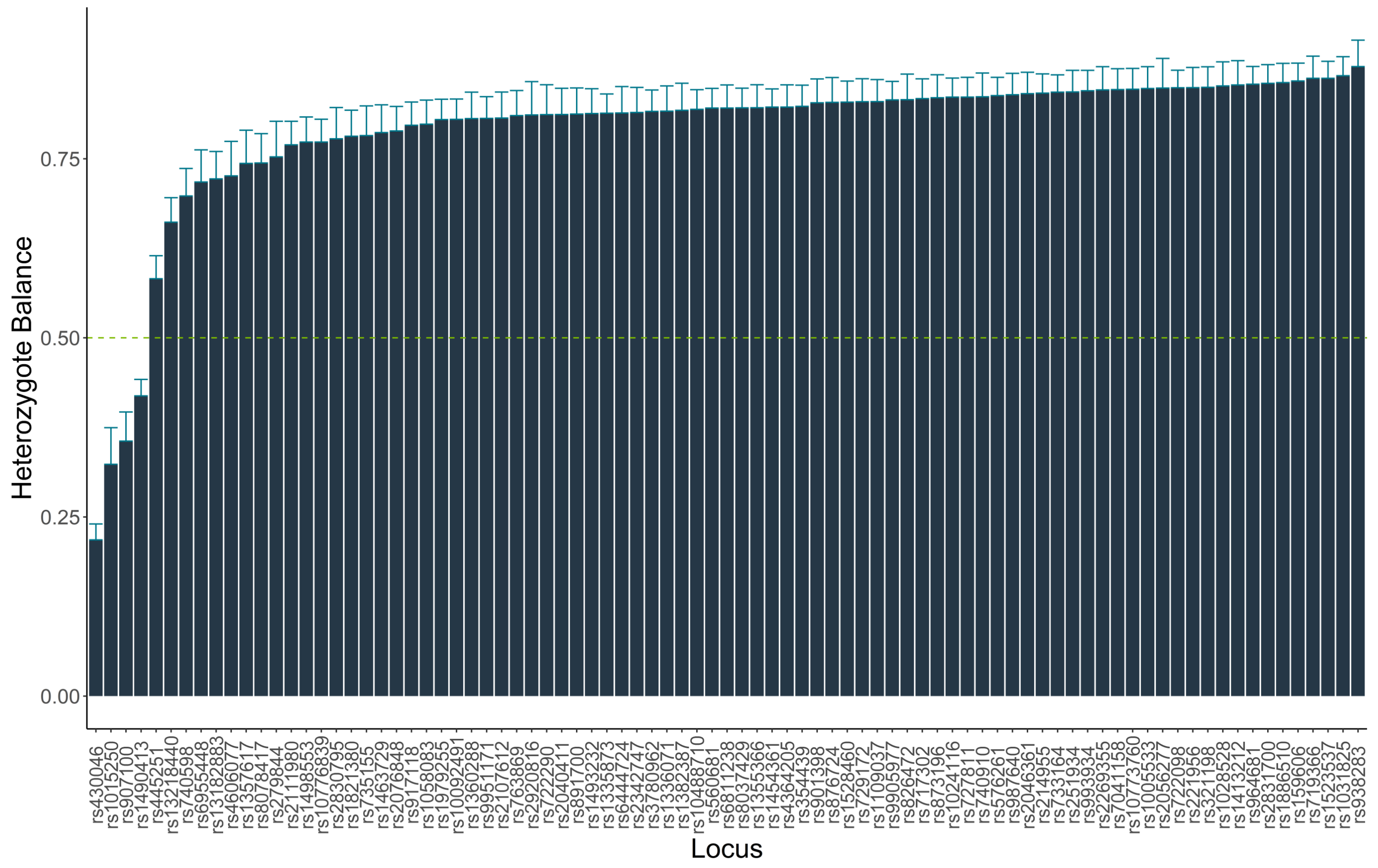

Figure 5. Mean heterozygote balance loci included in the RC-PCR 85-plex panel.

The cRMPs for each of the three population groups were 3.48 x 10^{-33}, 1.13 x 10^{-34}, and 7.80 x 10^{-34} for the AFA, CAU, and HIS groups, respectively. These 85 iiSNPs exceed discrimination power of the CE-STR profiles determined by various STR kits, approximately 1 x 10^{-26}, by about seven and eight orders of magnitude in AFA population and CAU and HIS populations, respectively.

No significant deviations from Hardy Weinberg Equilibrium were observed in any of the three populations (p<0.05). After calculating linkage disequilibrium, 258 locus pairs in the Caucasian population deviated from expectations (p<0.05). No pairwise comparisons in the African American or Hispanic populations were significant. After a Bonferonni correction adjusting for multiple tests, two pairwise comparisons in the Caucasian population remained significant (rs1886510/rs214955 and rs221956/rs2269355).

Patterns observed in PD and PIC values were in line with trends observed in heterozygosity across all population groups.

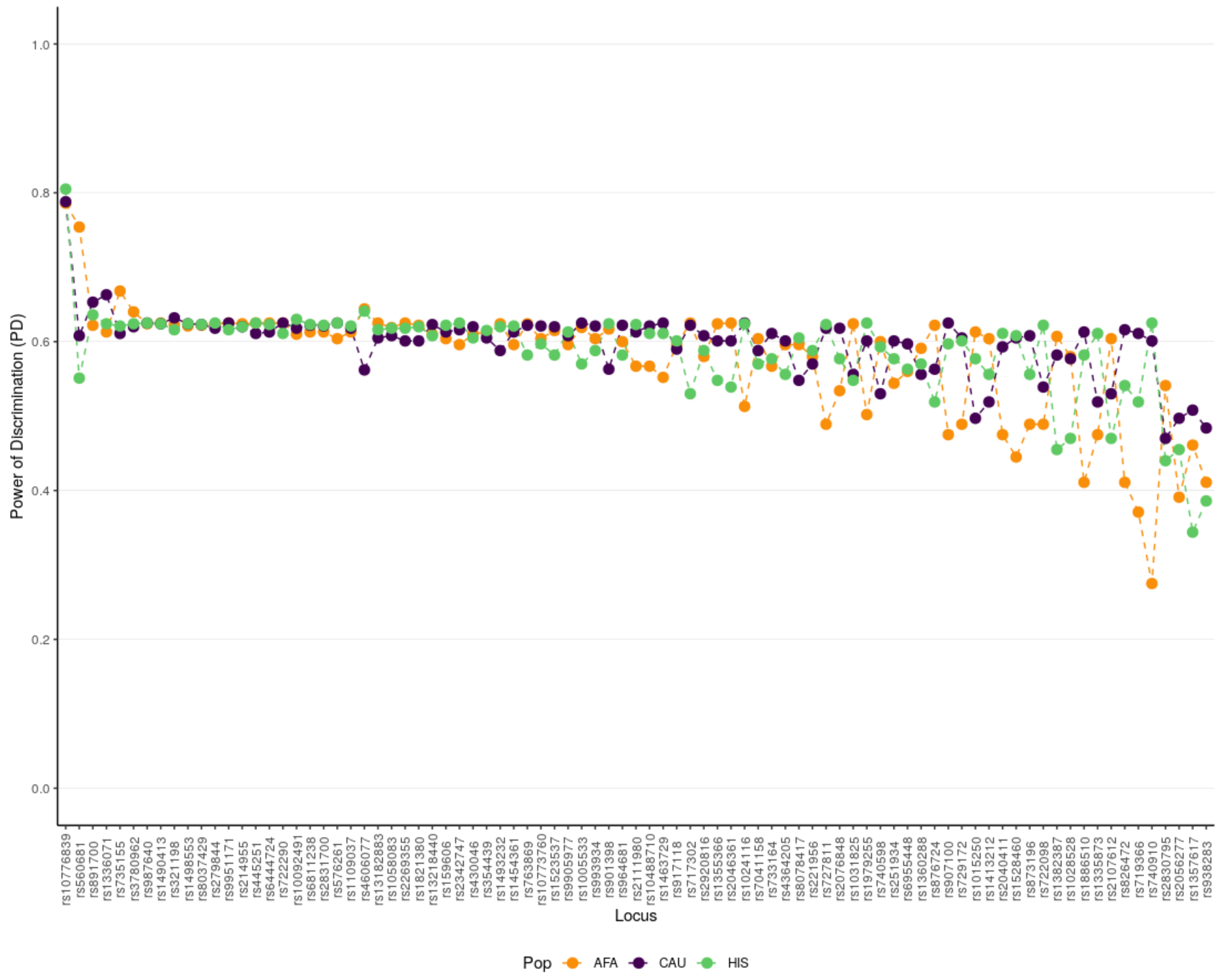

Figure 6. Power of discrimination for each population by locus.

Conclusions

In this study, we evaluated the performance of the IDseek OmniSNP Identity Informative SNP Typing Kit on 143 samples from three major populations. The kit produced high quality, robust data across all populations tested, supporting the future use of RC-PCR chemistry in forensic genetics and human identification. The library preparation can be completed in one day and produced combined random match probabilities ranging from 3.48 x 10^{-33} to 7.80 x 10^{-34} for the three populations evaluated in this study, which is comparable to or exceeds the discrimination power of current CE-STR profiles. The minor allele frequencies for all loci by population were greater than 0.1 with the exception of locus rs740910 which was also observed in the original literature.

Acknowledgements

This work was funded by award no. 2020-DQ-BX-0005, awarded by the National Institute of Justice (NIJ), Office of Justice Programs, US Department of Justice. The opinions, findings, and conclusions expressed within this publication are those of the authors and do not necessarily reflect those of the US Department of Justice. The authors would like to thank Nimagen for providing library preparation and sequencing reagents.