Kenneth K. Kidd[1]
Andrew J. Pakstis[1]
William C. Speed[1]
Robert Lagace[2]
Sharon Wootton[2]
Joseph Chang[2]

[1]Department of Genetics, Yale University School of Medicine, New Haven, CT, USA
[2]Human Identification Group, ThermoFisher Scientific, South San Francisco, CA, USA

Received February 15, 2018
Revised June 13, 2018
Accepted June 14, 2018

Research Article

Selecting microhaplotypes optimized for different purposes

Massively parallel sequencing is transforming forensic work by allowing various useful forensic markers, such as STRPs and SNPs, to be multiplexed providing information on ancestry, individual and familial identification, phenotypes for eye/hair/skin pigmentation, and the deconvolution of mixtures. Microhaplotypes also become feasible with massively parallel sequencing, these are DNA segments (smaller than 300 nucleotides) that are selected to contain multiple SNPs unambiguously defining three or more haplotype alleles occurring at common frequencies. The physical extent of a microhaplotype can thus be covered by a single sequence read making these loci phase-known codominant genetic systems. Such microhaplotypes supply significantly more information than a single SNP can. Our efforts to develop useful sets of microhaplotypes have already identified 182 such loci that we have studied on a large number of human populations from around the world. We present various analyses on 83 populations in our ongoing study for a subset of the best microhaplotypes currently available illustrating their characteristics and potential utility for ancestry, identification, and mixture deconvolution.

Keywords:
Ancestry / Forensic / Identification / Microhaplotype / Sequencing / Single nucleotide polymorphism
DOI 10.1002/elps.201800092

Additional supporting information may be found online in the Supporting Information section at the end of the article.

1 Introduction

The proven capabilities of massively parallel sequencing (MPS) [1] have introduced new ways of thinking about studies of DNA polymorphisms. While large numbers of SNPs, on the order of millions, can be genotyped using chips, this methodology does not allow genotyping the short tandem repeat polymorphisms (STRPs) used in forensics. In contrast, MPS allows both SNPs and STRPs to be multiplexed as short PCR amplicons of about 300 nucleotides for the various loci. The sequencing process also allows direct determination of how the haplotype alleles are organized (referred to as phasing) along the chromosome. In contrast, when SNPs and STRPs are genotyped separately and then phased by statistical methods a degree of uncertainty is introduced that is avoided by MPS. Forensic databases currently contain only STRP profiles of selected loci; while this is good for individual matching of crime scene samples to suspects, those loci contain very limited information on ancestry [2] or externally visible characteristics of the sample donor. Selected sets of SNPs, in contrast, can provide quite refined ancestry information [3–6] or good estimates of eye, hair, and skin color [7–13]. Thus, SNPs are potentially useful for investigative leads and are also of great interest for anthropologists studying population relationships. The recognition of the advantages of MPS in diverse applications has been growing and the number of forensic labs that are beginning to add the ability to use MPS is increasing [14].

Many of our studies (e.g. [15–21]) have shown that the multiple alleles at haplotype loci can provide significant and detailed information on population relationships and the pattern of modern human diversification. The emergence of MPS with the ability to sequence small amplicons up to about 300 bases in length motivated the search [22, 23] for microhaplotypes–highly polymorphic haplotypes in very short DNA segments. The value of microhaplotypes is

Correspondence: Dr. Kenneth K. Kidd, Professor Emeritus, Senior Research Scientist, Department of Genetics, Yale University School of Medicine, 333 Cedar Street, New Haven 06520–8005, CT, USA
E-mail: Kenneth.Kidd@yale.edu

Abbreviations: AISNP, ancestry informative single nucleotide polymorphism; **ALFRED**, ALlele FREquency Database; **IISNP**, individual identification single nucleotide polymorphism; **MPS**, massively parallel sequencing; **PCA**, principal components analysis; **RMP**, random match probability; **STRP**, short tandem repeat polymorphism

enhanced with MPS because each of the two specific alleles at a locus is sequenced separately and thereby the phase is determined unambiguously for a single individual. We are interested in which microhaplotypes will be of the greatest benefit to forensics and worth the effort of designing and validating MPS multiplexes. Two different statistics can be used to identify the best microhaps for different purposes [24]. The effective number of alleles (A_e) is related to the level of variation within a population and will be positively correlated with the ability to distinguish individuals, i.e. the random match probability (RMP). Also, the larger the A_e value, the better a locus will be for identifying mixtures and resolving the genotypes and the number of genotypes contributing to the mixture. This contrasts with the informativeness for ancestry inference (I_n) which measures the variation in allele frequencies among a set of populations [25]. The higher the I_n value, the better the microhap locus will be for distinguishing among at least some of those populations.

As a part of our research to identify microhaplotypes as candidates for forensic and anthropologic studies, we have identified 182 different microhaplotypes thus far and characterized them in many different population samples. Most of these microhaplotypes have already been published and data are available in ALFRED (alfred.med.yale.edu), the ALlele FREquency Database. We are now transitioning from identifying potentially useful microhaps to identifying those most likely to be useful for specific forensic purposes now that extensive population data are available. Here we explore two sets of 50 microhaps identified at this stage of the project as ranking the highest by each of the two different statistics: effective allele number (A_e) and Rosenberg's measure of informativeness (I_n). Other potential microhaps are still being assessed, some may prove to be better than those included here. While there is some overlapping across the datasets in that 22 loci rank high by both statistics, the remainder are distinct and serve to illustrate some of the consequences of employing different criteria for selecting a set of microhaps.

2 Materials and methods

2.1 Data collection

A total of 490 SNPs defining 182 microhaplotypes in 57 different populations (2763 individuals) were typed by TaqMan (Applied BioSystems) on a 7900HT SDS scanner. Most of the individual SNP loci are well documented in the literature and also appear to be valid single copy sites in our data because they do not deviate significantly from Hardy–Weinberg ratios in all of our populations and showed Mendelian transmission in the several nuclear families tested. As noted in Table 1, for two of the previously published microhaplotypes recent analyses have led us to consider one of the SNPs at each to be unreliably genotyped by TaqMan and therefore they were omitted in the current study. The resulting data collected in our lab on 2763 individuals were then combined with the comparable SNP data from the 1000 Genomes project [26] and the entire dataset was phased [27, 28] into the 182 microhaplotypes we compared. The minimum probability cutoff for assigning haplotypes via the PHASE 2-1-1 software was 97.5%. Thus, the microhaplotypes are empirically validated and the frequency estimates are accurate to that level, at a minimum.

2.2 Selection of loci for analyses

We ranked all 182 microhaps by the average A_e for 83 different population samples and separately by the I_n value among the 83 different population samples. We then arbitrarily selected the top 50 microhaps by that 83-population average value of A_e and separately by the 83-population I_n value. The two sets of microhaplotypes analyzed, along with their A_e and I_n values and 83 population ranks, are listed in Table 1 using our naming convention [29]. The final combined data set consists of 78 microhaplotypes with 22 loci common to both 50 microhap sets (highlighted in italics in the Table 1). The SNPs involved in each and the haplotype frequencies in the various population samples are available in ALFRED under the microhap locus name, the entire list can be retrieved with the keyword–microhap. All 78 microhaps are part of the set previously published [30].

The 83 populations are listed in Supporting Information Table 1 along with the number of individuals in each population sample, the three character abbreviation employed in some figures, and the population sample unique identifier in the ALFRED database (https://alfred.med.yale.edu) for looking up descriptive information.

For two of the microhap loci, recent reanalyses have indicated that one of the SNPs in each microhap could not be genotyped in some populations with uniformly high accuracy using TaqMan. This probably happens because of rare, previously unidentified, SNPs in some populations that interfered with either a primer or the probe. After removing these problematic SNPs, the remaining SNPs in mh06KK-008-ALT include rs6921774/rs6605524/rs6605523, and, the SNPs in mh16KK-255-ALT include rs16956011/rs3934956/rs4073828. The haplotype frequencies can be extracted from ALFRED by pooling alleles ignoring the omitted SNP noted in Table 1.

For comparisons with the two sets of microhap loci illustrating different characteristics of specially selected marker sets, we have performed some of the same analyses on the same 83 populations using 55 ancestry informative single nucleotide polymorphism (AISNPs) [31, 32] and from a set of 45 unlinked individual identification single nucleotide polymorphism (IISNPs) [33].

2.3 Estimating individual genotypes and population allele frequencies

As we have not genotyped these populations by MPS, we necessarily estimated individual genotypes by PHASE

Table 1. The top 50 microhaplotypes by two different criteria. For each set of 50 microhaps the values and ranks of A_e and I_n are given with the corresponding mean values and variances. Microhaps that occur in both lists are italicized. The two microhaps in the A_e set with "ALT" indicated have one SNP each deleted from the definition given in the original publication and in ALFRED: rs6930377 was deleted from mh06KK-008 and rs3934955 was deleted from mh16KK-255. See text

The top 50 microhaps by A_e						The top 50 microhaps by I_n					
Microhap name	83 pop average A_e	A_e rank 83 pops	I_n 83 pops	I_n rank 83 pops	Extent in bp	Microhap name	83 pop average A_e	A_e rank 83 pops	I_n 83 pops	I_n rank 83 pops	Extent in bp
mh13KK-218	6.010	1	0.367	20	146	mh20KK-016	2.296	84	0.682	1	259
mh05KK-170	5.244	2	0.374	19	137	mh17KK-014	1.936	136	0.660	2	37
mh10KK-169	4.656	3	0.354	21	292	mh05KK-120	2.002	127	0.539	3	276
mh21KK-320	4.605	4	0.217	77	186	mh02KK-139	1.575	161	0.522	4	54
mh10KK-163	4.494	5	0.335	25	260	mh05KK-121	2.549	62	0.468	5	281
mh02KK-134	4.445	6	0.389	14	104	*mh02KK-138*	3.338	27	0.462	6	249
mh16KK-049	4.206	7	0.314	33	174	*mh08KK-039*	3.937	12	0.454	7	213
mh11KK-180	3.982	8	0.268	50	194	mh05KK-058	2.067	118	0.450	8	303
mh21KK-315	3.972	9	0.187	105	146	mh09KK-020	2.489	66	0.432	9	128
mh13KK-217	3.960	10	0.245	62	193	mh09KK-161	1.818	149	0.410	10	104
mh04KK-030	3.955	11	0.335	26	115	mh14KK-101	1.666	158	0.403	11	96
mh08KK-039	3.937	12	0.454	7	213	mh16KK-062	2.296	85	0.403	12	298
mh01KK-117	3.889	13	0.293	40	187	mh02KK-003	1.972	130	0.393	13	125
mh21KK-324	3.857	14	0.382	17	159	*mh02KK-134*	4.445	6	0.389	14	104
mh19KK-299	3.834	15	0.318	29	154	mh16KK-053	1.403	172	0.387	15	12
mh06KK-008-ALT	3.808	16	0.255	54	170	*mh01KK-001*	3.013	33	0.384	16	260
mh13KK-223	3.793	17	0.218	76	154	*mh21KK-324*	3.857	14	0.382	17	159
mh01KK-205	3.777	18	0.136	137	155	*mh09KK-153*	2.980	34	0.374	18	113
mh02KK-136	3.746	19	0.188	102	71	*mh05KK-170*	5.244	2	0.374	19	137
mh04KK-013	3.671	20	0.277	48	201	*mh13KK-218*	6.010	1	0.367	20	146
mh13KK-213	3.645	21	0.209	87	141	*mh10KK-169*	4.656	3	0.354	21	292
mh13KK-225	3.466	22	0.196	97	97	mh15KK-104	2.196	99	0.348	22	138
mh18KK-293	3.402	23	0.307	36	83	mh22KK-069	1.933	137	0.341	23	79
mh20KK-307	3.401	24	0.215	81	141	*mh10KK-170*	2.864	43	0.338	24	190
mh09KK-157	3.361	25	0.251	59	154	*mh10KK-163*	4.494	5	0.335	25	260
mh16KK-255-ALT	3.361	26	0.209	85	143	*mh04KK-030*	3.955	11	0.335	26	115
mh02KK-138	3.338	27	0.462	6	249	mh06KK-031	1.463	167	0.329	27	159
mh22KK-061	3.320	28	0.169	114	147	mh11KK-103	2.323	79	0.324	28	262
mh05KK-020	3.255	29	0.198	96	169	*mh19KK-299*	3.834	15	0.318	29	154
mh03KK-150	3.194	30	0.177	108	185	*mh21KK-316*	2.949	37	0.318	30	135
mh01NK-001	3.026	31	0.268	51	280	mh02KK-131	1.893	140	0.317	31	87
mh11KK-191	3.017	32	0.228	72	190	mh15KK-068	1.791	153	0.316	32	217
mh01KK-001	3.013	33	0.384	16	260	*mh16KK-049*	4.206	7	0.314	33	174
mh09KK-153	2.980	34	0.374	18	113	mh05KK-122	1.924	138	0.313	34	50
mh12KK-202	2.979	35	0.154	126	154	*mh16KK-302*	2.949	36	0.312	35	114
mh16KK-302	2.949	36	0.312	35	114	*mh18KK-293*	3.402	23	0.307	36	83
mh21KK-316	2.949	37	0.318	30	135	mh19KK-301	1.817	150	0.303	37	64
mh09KK-152	2.938	38	0.204	89	142	mh03KK-216	2.067	119	0.303	38	6
mh11KK-092	2.937	39	0.153	128	296	*mh01KK-172*	2.683	50	0.300	39	226
mh21KK-313	2.899	40	0.284	45	173	*mh01KK-117*	3.889	13	0.293	40	187
mh17KK-272	2.892	41	0.184	106	131	mh03KK-020	2.008	126	0.293	41	220
mh12KK-046	2.867	42	0.159	122	72	mh01KK-135	2.010	125	0.287	42	289
mh10KK-170	2.864	43	0.338	24	190	mh01KK-002	2.538	64	0.286	43	18
mh09KK-033	2.814	44	0.166	117	78	mh05KK-124	2.210	96	0.285	44	67
mh15KK-066	2.765	45	0.201	93	75	*mh21KK-313*	2.899	40	0.284	45	173
mh04KK-010	2.746	46	0.174	110	35	mh12KK-042	1.946	134	0.283	46	175
mh14KK-048	2.716	47	0.202	90	159	mh01KK-106	2.511	65	0.279	47	171
mh15KK-067	2.711	48	0.252	58	122	*mh04KK-013*	3.671	20	0.277	48	201
mh20KK-058	2.702	49	0.169	113	106	mh15KK-069	2.090	114	0.273	49	88
mh01KK-172	2.683	50	0.300	39	226	*mh11KK-180*	3.982	8	0.268	50	194
Average	3.501		0.262		159.4	Average	2.801		0.363		158.8
Variance	0.490		0.007		3334	Variance	1.109		0.008		6697

bp, Base pairs

version 2-1-1 [27, 28]. As noted elsewhere [34, 35], the PHASE software yields a statistically accurate estimate of allele frequencies for populations and gives maximum likelihood estimates of the haplotype genotypes of individuals when each SNP is typed individually as in this exploratory phase of our studies. We note that for actual forensic data collection on samples, MPS will need to be used. For reference population data and the comparisons between different sets of loci the phased data are statistically suitable estimates.

2.4 Estimating RMP

For each set of microhaps the random match probability for each population sample was calculated assuming Hardy–Weinberg ratios. While RMP values will be different for each unique genotype, values have been calculated as the expected value based on the allele (haplotype) frequencies in each population. No correction for within-sample population structure was made since the focus is on distinct, well defined population samples representing global population structure.

2.5 Evaluating ancestry information

STRUCTURE v.2.3.4 [36] was used to evaluate and illustrate the clustering of individuals into predefined groups of genetic similarity based on the different sets of loci. The STRUCTURE analysis parameters employed include: 10, 000 burnins and 10, 000 Markov Chain Monte Carlo iterations, admixture model, correlated allele frequencies, 20 independent replicates per predefined number of clusters (K) from $K = 5$ to $K = 10$. The input data file for the STRUCTURE analyses contained the individual genotypes assigned by PHASE for each individual.

Principal components analysis (PCA) analyses were also used to compare the similarities and differences among the 83 populations as illustrated by the two different sets of microhap loci. The analyses used the population allele frequencies. PCA was calculated using XLSTAT 2017 (http://www.xlstat.com/en/about-us/company.html).

3 Results

3.1 The selected loci

Table 1 lists the two sets of microhaplotypes with their A_e and I_n values and ranks and the molecular span of the SNPs comprising each. The small size range (6–303 base pairs) and average size (159 base pairs) of the microhaplotypes provide the basis for predicting a high success rate on degraded samples. The two sets have 22 loci in common, indicated by the names in italics. Summary values show the expected differences based on the criteria for selection of the following two groups: the average A_e across all loci is significantly higher for the top 50 A_e loci and the I_n value significantly higher for the top I_n loci. Supporting Information Fig. 1 shows a scatterplot of the full set of 78 loci by A_e and I_n demonstrating the diversity of the individual loci.

3.2 RMPs

Figure 1 plots the RMP values for all 83 populations based on the two microhap datasets with two SNP panels for comparison, the 45 IISNP panel [33] and the 55 AISNP panel. It is immediately obvious that the values for most of the populations are several orders of magnitude smaller based on the 50 top A_e loci than based on the 50 top I_n loci. The comparative AISNP data for these 83 populations are a subset of the AISNP data on 139 populations published previously [32]. We note these data only because this panel has been studied on a large number of populations and are part of some commercial kits. Clearly, these AISNPs are not intended to give very small RMP values. The data on 45 IISNPs are available on 73 of the 83 populations, the RMP values are plotted for corresponding populations. Since they were selected for uniformly high heterozygosity, the RMP values are significantly smaller than for the AISNP panel.

3.3 STRUCTURE analyses

Figure 2 shows the results of the STRUCTURE analyses for the two microhap datasets at $K = 9$. The results are plotted as averages for each population for the runs with the two highest likelihoods. For both datasets, the likelihoods continued to increase with increasing the K value through $K = 11$, but the results at $K = 9$ illustrate the differences and similarities most clearly. At lower values of K, there were only very minor differences in how individuals grouped into clusters (not shown). Only at $K = 8$ do differences begin to be evident, those are maintained at $K = 9$ but for both datasets the differences are not consistent among the replicates.

For the comparable populations the clustering shown by the 50 best A_e SNPs is very similar to that shown by the 55 AISNPs in an analysis of many more populations [33].

3.4 PCA

Figure 3 presents the results of the PCA analysis of the two datasets. There are some differences but, as with the STRUCTURE results, there is a general similarity for the two sets of microhaps. In Fig. 3 the plots for the I_n and A_e datasets present the 2D views based only on the first two principal components that account for the largest proportion of variation. Views (not shown) that include the third principal component show an even better separation of the Native American population cluster for both datasets. Inspecting the fourth principal component shows more separation of the Pacific populations in both datasets while the South Central Asian populations

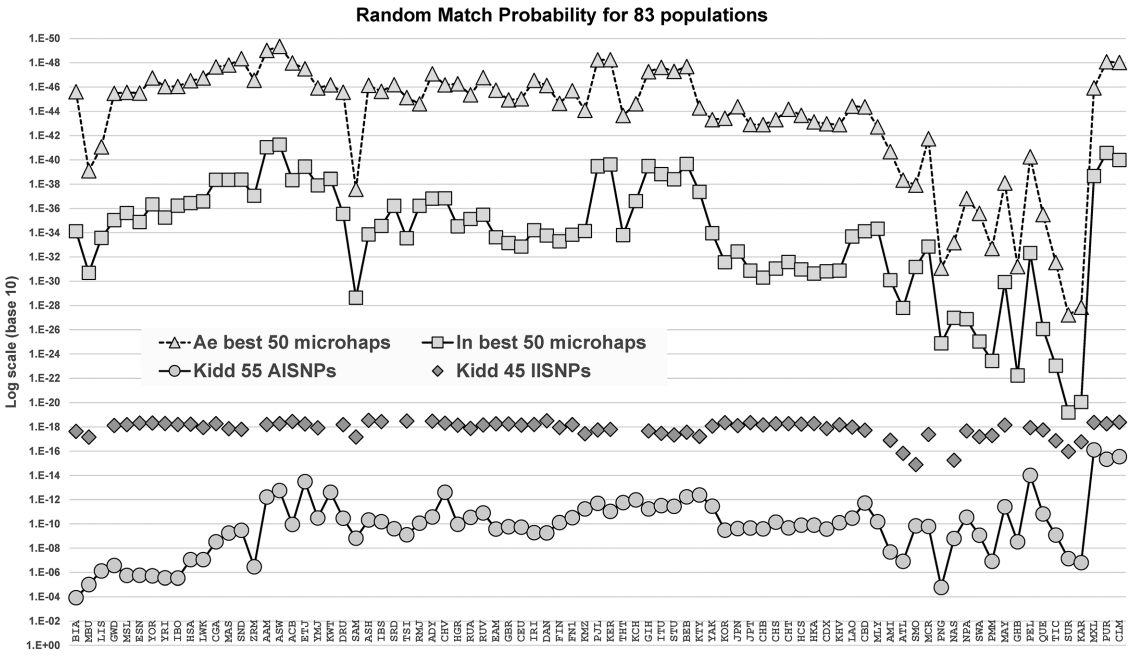

Figure 1. Random match probability values for each population in four datasets—the best 50 A_e microhaps, the best 50 I_n microhaps, the 45 Kidd IISNPs, and the 55 Kidd AISNPs.

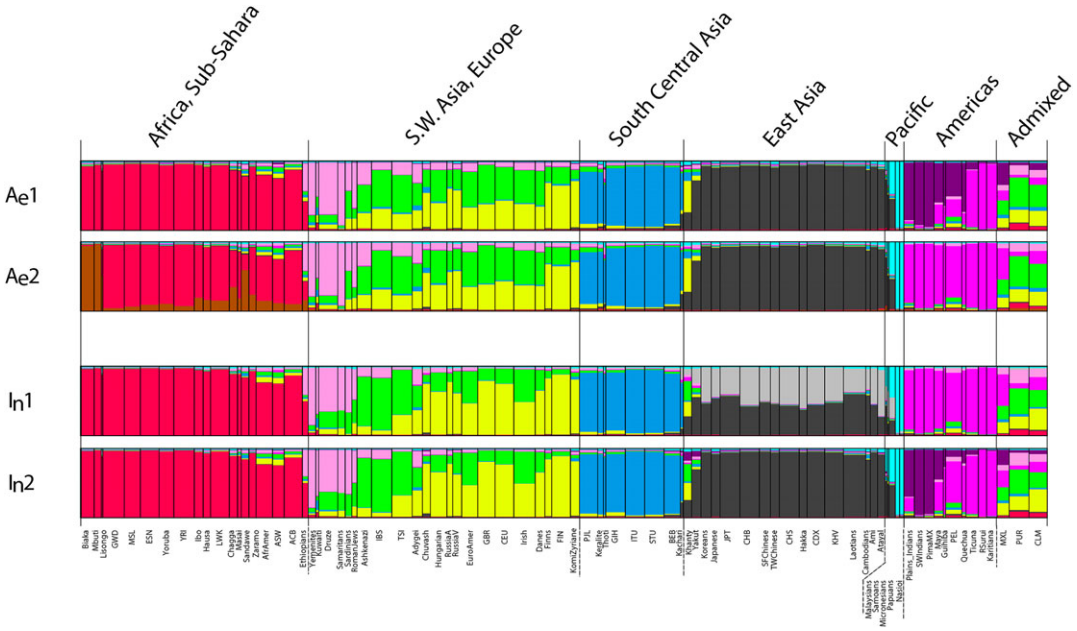

Figure 2. Estimated cluster membership values at $K = 9$ for each of 83 populations in the STRUCTURE analyses for the two microhaplotype datasets—the best 50 I_n and the best 50 A_e.

separate somewhat better from other world regions for the best I_n microhaps but not for the best A_e microhaps.

4 Discussion

The various uses of microhaps include individual identification, ancestry inference, familial identification, and mixture deconvolution. Phenotype inference is also possible to the extent that one of the associated SNPs is part of a microhap. As we demonstrated [30] many candidate microhaplotypes are poor at individual identification and other desirable characterizations when examined on multiple populations. Given the effort to develop and validate a multiplex for MPS use in forensics, we are continuing to consider different criteria for identifying a panel to convert to a MPS multiplex. Future

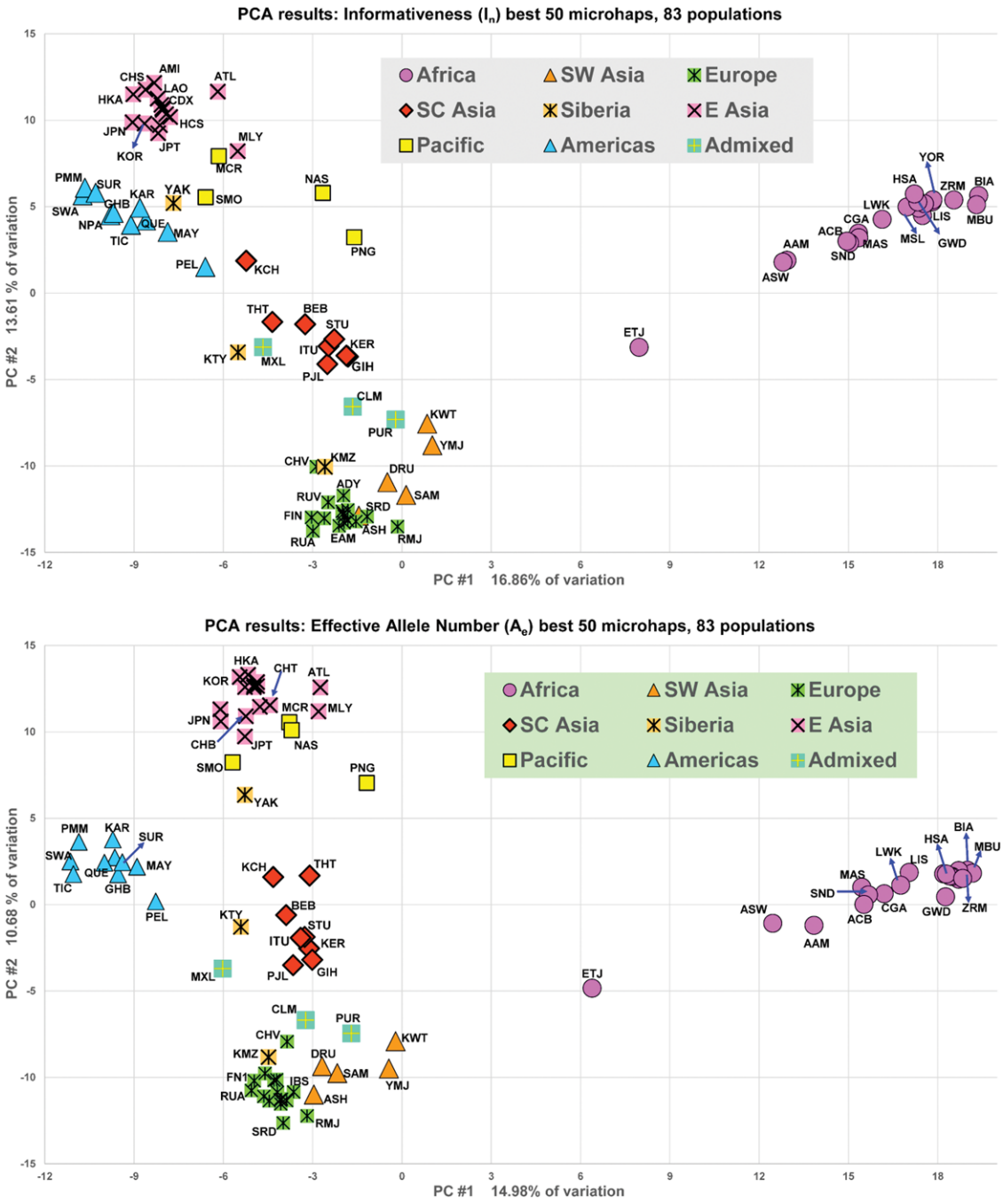

Figure 3. PCA plotting PC#1 by PC#2 for the best 50 I_n microhaps (Fig. 3A) and for the best 50 A_e microhaps (Fig. 3B).

empirical testing on population samples, on mixtures, on multi-generational pedigrees, and on casework type samples using MPS will provide critical information for identifying which microhaps should be converted to MPS multiplexes for forensic casework. Here we have focused on illustrating differences between microhaps selected for ancestry inference and those selected for usefulness for mixture deconvolution. While both groups of microhaps provide information on individual identification, those selected for A_e do better because, by definition, there are more haplotype alleles, on average, in each population and the process of selecting them aimed for the presence of multiple alleles at common frequencies.

One obvious conclusion from the comparisons presented is that both sets of microhap loci and the 55 AISNPs provide very similar groupings of these 83 populations into six to seven broad biogeographic regions. There are some differences but the overall impression from the STRUCTURE analyses is similarity. However, at larger *K* values, the I_n

dataset gives clearer distinctions of some populations (e.g., Southwest Asian) than the A_e dataset. Conversely, for other groups of populations (e.g. African, Native American) the A_e dataset shows more distinctions. While it is expected that the different sets of loci would give different clusterings of individuals and populations, it is difficult to determine how much of the difference is attributable to the different loci and how much to chance. The two independent runs of the MCMC in STRUCTURE with the two highest likelihoods (Fig. 2) show that some seemingly stochastic element is involved, making it difficult to determine how much of the difference in the results from the two datasets is robust. Indeed, we find the results for the top A_e loci to be surprisingly good at biogeographic ancestry even up to $K = 9$ clusters. These results argue that a larger number of loci in each group will be necessary to show additional meaningful differences. In the meantime, new microhaps are being identified and the known rare variants in the regions of almost all microhaplotypes will clearly increase the values of both A_e and I_n at already identified loci. In our laboratory more populations are being tested for a larger number of microhap loci, including systematically the two sets of 50 loci studied here.

We have used high A_e as one criterion for selecting loci. As we showed [30] the probability of detecting a mixture of DNA is a function of A_e: the more alleles at a locus, the greater the probability of a mixture displaying more than two alleles. In many cases it will be possible to give a minimum number of samples in a mixture as the number of alleles detected divided by two with the answer rounded up to the next whole number. For example, if one locus in a multiplex has five distinguishable haplotypes, a minimum of three individuals constitute the mixture. For the same mixture other loci may have only two to four distinguishable alleles because of shared alleles among the individuals and homozygosity of individuals. In preliminary data (not shown, submitted for publication) for a multiplex of 36 microhaplotypes tested on an artificial mixture of DNA from four individuals, only one locus showed six alleles, but the reads for one allele were several fold higher than for the other five. The conclusion that one sample was a homozygote seemed reasonable, leading to the conclusion that the mixture involved (at least) four individual DNAs, as was the case for the artificial mixture.

All of this research on microhaplotypes is in flux. At the moment our efforts are focused on identifying the subset of the 182 microhap loci we are studying that work best for ancestry inference and mixture deconvolution with a much broader set of populations. As more populations are characterized for a common set of microhaps, the rankings of the loci will change. In a recent publication [37], we presented analyses of a subset of 65 of the 182 microhaps that have been characterized on a broader group of 96 populations, those 65 microhaps were not selected by either the I_n or A_e criterion and only 41 of the 78 different microhaps in the present combined I_n and A_e datasets were included in that 65 microhap dataset. Those 65 were part of the more general process aimed at identifying potentially useful microhaplotypes. Genotyping even more populations on the microhaps in order to represent better the diversity of human populations worldwide is an essential part of our study. It is also an important part of the process that will help to assure the utility and acceptance of microhaps in forensic laboratories.

Converting more of the loci to multiplexes for MPS will likely identify some that do not multiplex and type well with MPS. We chose the top 50 loci as a reasonable sample to examine and illustrate the differences in results attributable to the different selection criteria. The 50 based on I_n and A_e show a similar pattern to that shown by the 55 AISNP panel [32] for the populations in common.

The results for RMP are also quite relevant for optimizing microhaps for forensic applications. Obviously, with more loci the RMP values become smaller and we have not explored different numbers of loci but have compared the high ranking A_e loci with the high ranking I_n loci. On average the 50 A_e loci have 3.5 effective alleles for an average of 2.5 independent alleles at each locus. This translates to 150 independent variables. The comparable values for the 50 I_n loci are 2.8 effective alleles for an average of 1.8 independent alleles at each locus translating to 90 independent variables. Although the number of loci is the same, the difference in independent alleles is apparent in the two RMP values. Both the 55 AISNPs and the 45 IISNPs have only one independent allele per locus. The RMP and correlated forensic statistics from microhaps can be readily combined with the values from a standard forensic panel of STRPs. The take-home lesson from this is that microhaps with high A_e values provide excellent individual identification information for most parts of the world. Additional single-locus IISNPs are hardly necessary.

These RMP values also emphasize the need for additional highly informative microhap loci for the Pacific and Native American populations. The 55 AISNPs were selected from a large set of SNPs typed on over 3000 individuals from over 50 populations to give roughly balanced distinction of multiple biogeographic regions of the world, with special emphasis on Native American and Pacific Island populations. Unfortunately, there have not been large enough numbers of such populations tested for large numbers of microhap markers to allow comparable selection of loci with high A_e and/or I_n values in these populations. This is an area of need in our study for those populations.

MPS clearly offers important advantages for forensic work. The ability to obtain highly informative phased microhaplotype genotypes for the benefit of different forensic purposes in an efficient fashion has been emphasized in this presentation. The targeted amplification and genotyping of microhaplotype loci by sequencing incurs no greater expense than a single SNP, but the yield in useful information for forensic applications is considerably greater.

We look forward to the identification of additional microhaps by the research community and their characterization on multiple populations. We hope that soon the forensics and anthropology communities can agree on a suitable set(s) of loci that will allow researchers around the world to collaborate on building a global set of reference population data. We think that enough empirical work has accumulated on

microhaps in published reports that their value for criminal applications has been demonstrated. The urgent issue now is to agree on the set of markers to be used in such forensic applications to supplement the current STRPs so that a forensic database can be started. Part of that decision will involve knowing that the microhaps will work well in MPS multiplexes. So far, two multiplexes have been designed based on earlier evaluations of the 130 microhaps then available [30] and are now in beta tests. We are not convinced that a final selection is yet warranted but we are sufficiently convinced that the top 50 A_e microhaplotypes warrant examination by MPS. We are currently in the process of designing another multiplex for MPS that includes these 50 microhaplotypes.

This work was funded primarily by NIJ grants 2013-DN-BX-K023, 2015-DN-BX-K023, and 2014-DN-BX-K030 awarded to KKK by the National Institute of Justice, Office of Justice Programs of the United States Department of Justice. Points of view in this presentation are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice. The authors also thank Dr. Francoise Friedlaender for her expert help in preparing the Structure figure. Special thanks are due to the many hundreds of individuals who volunteered to give blood or saliva samples for studies of gene frequency variation and to the many colleagues who helped collect the samples. In addition, some of the cell lines were obtained from the National Laboratory for the Genetics of Israeli Populations at Tel Aviv University, and African American samples were obtained from the Coriell Institute for Medical Research, Camden, New Jersey.

The authors declare no conflict of interest.

5 References

[1] Hinds, D. A., McMahon, G., Kiefer, A. K., Do, C. B., Eriksson, N., Evans, D. M., St Pourcain, B., Ring, S. M., Mountain, J. L., Francke, U., Davey-Smith, G., Timpson, N. J., Tung, J. Y., *Nat. Genet.* 2013, *45*, 907–911.

[2] Algee-Hewitt, B. F., Edge, M. D., Kim, J., Li, J. Z., Rosenberg, N. A., *Curr. Biol. CB* 2016, *26*, 935–942.

[3] Kosoy, R., Nassir, R., Tian, C., White, P. A., Butler, L. M., Silva, G., Kittles, R., Alarcon-Riquelme, M. E., Gregersen, P. K., Belmont, J. W., De La Vega, F. M., Seldin, M. F., *Human Mutat.* 2009, *30*, 69–78.

[4] Kidd, J. R., Friedlaender, F. R., Speed, W. C., Pakstis, A. J., De La Vega, F. M., Kidd, K. K., *Investig. Genet.* 2011, *2*, 1.

[5] Phillips, C., Freire Aradas, A., Kriegel, A. K., Fondevila, M., Bulbul, O., Santos, C., Serrulla Rech, F., Perez Carceles, M. D., Carracedo, A., Schneider, P. M., Lareu, M. V., *Forensic Sci. Int. Genet.* 2013, *7*, 359–366.

[6] Kidd, K. K., Speed, W. C., Pakstis, A. J., Furtado, M. R., Fang, R., Madbouly, A., Maiers, M., Middha, M., Friedlaender, F. R., Kidd, J. R., *Forensic Sci. Int. Genet.* 2014, *10*, 23–32.

[7] Ruiz, Y., Phillips, C., Gomez-Tato, A., Alvarez-Dios, J., Casares de Cal, M., Cruz, R., Maronas, O., Sochtig, J., Fondevila, M., Rodriguez-Cid, M. J., Carracedo, A., Lareu, M. V., *Forensic Sci. Int. Genet.* 2013, *7*, 28–40.

[8] Gettings, K. B., Lai, R., Johnson, J. L., Peck, M. A., Hart, J. A., Gordish-Dressman, H., Schanfield, M. S., Podini, D. S., *Forensic Sci. Int. Genet.* 2014, *8*, 101–108.

[9] Walsh, S., Kayser, M., *Methods Mol. Biol.* 2016, *1420*, 213–231.

[10] Andersen, J. D., Pietroni, C., Johansen, P., Andersen, M. M., Pereira, V., Borsting, C., Morling, N., *Mol. Genet. Genomic Med.* 2016, *4*, 420–430.

[11] Walsh, S., Chaitanya, L., Breslin, K., Muralidharan, C., Bronikowska, A., Pospiech, E., Koller, J., Kovatsi, L., Wollstein, A., Branicki, W., Liu, F., Kayser, M., *Human Genet.* 2017, *136*, 847–863.

[12] Crawford, N. G., Kelly, D. E., Hansen, M. E. B., Beltrame, M. H., Fan, S., Bowman, S. L., Jewett, E., Ranciaro, A., Thompson, S., Lo, Y., Pfeifer, S. P., Jensen, J. D., Campbell, M. C., Beggs, W., Hormozdiari, F., Mpoloka, S. W., Mokone, G. G., Nyambo, T., Meskel, D. W., Belay, G., Haut, J., Rothschild, H., Zon, L., Zhou, Y., Kovacs, M. A., Xu, M., Zhang, T., Bishop, K., Sinclair, J., Rivas, C., Elliot, E., Choi, J., Li, S. A., Hicks, B., Burgess, S., Abnet, C., Watkins-Chow, D. E., Oceana, E., Song, Y. S., Eskin, E., Brown, K. M., Marks, M. S., Loftus, S. K., Pavan, W. J., Yeager, M., Chanock, S., Tishkoff, S. A., *Science* 2017, *358*.

[13] Martin, A. R., Lin, M., Granka, J. M., Myrick, J. W., Liu, X., Sockell, A., Atkinson, E. G., Werely, C. J., Moller, M., Sandhu, M. S., Kingsley, D. M., Hoal, E. G., Liu, X., Daly, M. J., Feldman, M. W., Gignoux, C. R., Bustamante, C. D., Henn, B. M., *Cell* 2017, *171*, 1340–1353, e1314.

[14] Alonso, A., Muller, P., Roewer, L., Willuweit, S., Budowle, B., Parson, W., *Forensic Sci. Int. Genet.* 2017, *29*, e23-e25.

[15] Li, H., Gu, S., Han, Y., Xu, Z., Pakstis, A. J., Jin, L., Kidd, J. R., Kidd, K. K., *Ann. Human Genet.* 2011, *75*, 497–507.

[16] Pakstis, A. J., Fang, R., Furtado, M. R., Kidd, J. R., Kidd, K. K., *Eur. J. Human Genet. EJHG* 2012, *20*, 1148–1154.

[17] Tishkoff, S. A., Dietzsch, E., Speed, W., Pakstis, A. J., Kidd, J. R., Cheung, K., Bonne-Tamir, B., Santachiara-Benerecetti, A. S., Moral, P., Krings, M., *Science* 1996, *271*, 1380–1387.

[18] Tishkoff, S. A., Goldman, A., Calafell, F., Speed, W. C., Deinard, A. S., Bonne-Tamir, B., Kidd, J. R., Pakstis, A. J., Jenkins, T., Kidd, K. K., *Am. J. Human Genet.* 1998, *62*, 1389–1402.

[19] Kidd, J. R., Friedlaender, F., Pakstis, A. J., Furtado, M., Fang, R., Wang, X., Nievergelt, C. M., Kidd, K. K., *Am. J. Phys. Anthropol.* 2011, *146*, 495–502.

[20] Brissenden, J. E., Kidd, J. R., Evsanaa, B., Togtokh, A. J., Pakstis, A. J., Friedlaender, F., Kidd, K. K., Roscoe, J. M., *Human Biol.* 2015, *87*, 73–91.

[21] Cherni, L., Pakstis, A. J., Boussetta, S., Elkamel, S., Frigi, S., Khodjet-El-Khil, H., Barton, A., Haigh, E., Speed, W. C., Ben Ammar Elgaaied, A., Kidd, J. R., Kidd, K. K., *Am. J. Phys. Anthropol.* 2016, *161*, 62–71.

[22] Kidd, K. K., Pakstis, A. J., Speed, W. C., Lagace, R., Chang, J., Wootton, S., Ihuegbu, N., *Forensic Sci. Int. Genet. Suppl. Series* 2013, *4*, e123-e124.

[23] Kidd, K. K., Pakstis, A. J., Speed, W. C., Lagace, R., Chang, J., Wootton, S., Haigh, E., Kidd, J. R., *Forensic Sci. Int. Genet.* 2014, *12*, 215–224.

[24] Kidd, K. K., Speed, W. C., *Investig. Genet.* 2015, *6*, 1.

[25] Rosenberg, N. A., Li, L. M., Ward, R., Pritchard, J. K., *Am. J. Human Genet.* 2003, *73*, 1402–1422.

[26] 1000 Genomes Consortium, 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., Abecasis, G. R., *Nature* 2015, *526*, 68–74.

[27] Scheet, P., Stephens, M., *Am. J. Human Genet.* 2006, *78*, 629–644.

[28] Stephens, M., Scheet, P., *Am. J. Human Genet.* 2005, *76*, 449–462.

[29] Kidd, K. K., *Human Genomics* 2016, *10*, 16.

[30] Kidd, K. K., Speed, W. C., Pakstis, A. J., Podini, D. S., Lagace, R., Chang, J., Wootton, S., Haigh, E., Soundararajan, U., *Forensic Sci. Int. Genet.* 2017, *29*, 29–37.

[31] Pakstis, A. J., Haigh, E., Cherni, L., ElGaaied, A. B. A., Barton, A., Evsanaa, B., Togtokh, A., Brissenden, J., Roscoe, J., Bulbul, O., Filoglu, G., Gurkan, C., Meiklejohn, K. A., Robertson, J. M., Li, C. X., Wei, Y. L., Li, H., Soundararajan, U., Rajeevan, H., Kidd, J. R., Kidd, K. K., *Forensic Sci. Int. Genet.* 2015, *19*, 269–271.

[32] Pakstis, A. J., Kang, L., Liu, L., Zhang, Z., Jin, T., Grigorenko, E. L., Wendt, F. R., Budowle, B., Hadi, S., Al Qahtani, M. S., Morling, N., Mogensen, H. S., Themudo, G. E., Soundararajan, U., Rajeevan, H., Kidd, J. R., Kidd, K. K., *Int. J. Legal Med.* 2017, *131*, 913–917

[33] Pakstis, A. J., Speed, W. C., Fang, R., Hyland, F. C., Furtado, M. R., Kidd, J. R., Kidd, K. K., *Human Genet.* 2010, *127*, 315–324.

[34] Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E., Lin, S., Qin, Z. S., Munro, H. M., Abecasis, G. R., Donnelly, P., *Am. J. Human Genet.* 2006, *78*, 437–450.

[35] Tishkoff, S. A., Pakstis, A. J., Ruano, G., Kidd, K. K., *Am. J. Human Genet.* 2000, *67*, 518–522.

[36] Pritchard, J. K., Stephens, M., Donnelly, P., *Genetics* 2000, *155*, 945–959.

[37] Bulbul, O., Pakstis, A. J., Soundararajan, U., Gurkan, C., Brissenden, J. E., Roscoe, J. M., Evsanaa, B., Togtokh, A., Paschou, P., Grigorenko, E. I., Gurwitz, D., Wootton, S., Lagace, R., Chang, J., Speed, W. C., Kidd, K. K., *Int. J. Legal Med.* 2017, *132*, 703–711.

