Int J Legal Med (2010) 124:353–361
DOI 10.1007/s00414-009-0400-5

ORIGINAL ARTICLE

Haplotype block: a new type of forensic DNA markers

Jianye Ge · Bruce Budowle · John V. Planz · Ranajit Chakraborty

Received: 23 September 2009 / Accepted: 30 November 2009 / Published online: 22 December 2009
© Springer-Verlag 2009

Abstract Forensic DNA analysis is currently performed using highly discriminating short tandem repeat (STR) markers. SNPs are being investigated as adjunct tools for human identity testing because of their abundance in the human genome, utility for genotyping degraded DNA samples, and amenability to automation. While SNPs can provide an alternative approach, on a per locus basis they have a lower power of discrimination (PD) than STRs. With the discovery of block structures in the human genome, a novel set of SNP markers are available for further exploration of forensic utility. Several neighboring, tightly linked SNPs are inherited together and form a haplotype block, which as a haploblock has a higher discrimination power than the individual SNPs within the block. Candidate haplotype blocks were selected from three major populations (Caucasian, East Asian, and African) using the following parameters: maximum match probability reduction=0.85, linkage disequilibrium (LD) $r^2 \geq 0.7$, maximum $F_{st}=0.06$, minimum number of SNPs=3, minimum heterozygosity=0.2, and minimum number of haplotypes=3. From the HapMap Phase II data, 253 haploblocks were identified on the 22 autosomal chromosomes. After removing haploblocks deviating from the Hardy–Weinberg equilibrium (HWE) or in LD with other haploblocks, 24 haploblocks remained as candidates for forensic consideration. The cumulative PD of these blocks can reach 10^{-12} in the populations studied. The data support within and between haplotype independence even when they are syntenic. We propose guidelines for evidence interpretation that address the application of haplotype blocks for transfer evidence, mixture, and kinship analyses.

Keywords Forensic DNA marker · Haplotype blocks · HapMap · Evidence interpretation · Single-nucleotide polymorphism

Electronic supplementary material The online version of this article (doi:10.1007/s00414-009-0400-5) contains supplementary material, which is available to authorized users.

J. Ge · B. Budowle · J. V. Planz
Institute of Investigative Genetics,
University of North Texas Health Science Center,
Fort Worth, TX 76107, USA

J. Ge (✉) · B. Budowle · J. V. Planz
Department of Forensic and Investigative Genetics,
University of North Texas Health Science Center,
Ft Worth, TX 76107, USA
e-mail: jge@hsc.unt.edu

R. Chakraborty
Center for Genome Information, Department of Environmental Health, University of Cincinnati College of Medicine,
Cincinnati, OH 45267, USA

Introduction

Sensitivity of detection and discriminatory power of short tandem repeats (STRs) and single-nucleotide polymorphisms (SNPs) have made DNA typing a routine and powerful tool in forensic investigations of both civil and criminal cases. Currently, a set of highly informative STR markers is used for forensic DNA typing [1, 2]. SNP panels are being investigated to serve as adjunct tools to the STRs because they are abundant in the human genome, may provide additional genetic information in kinship analyses, may be applicable to analysis of highly degraded specimens where STRs do not provide results, and are amenable to automation [1, 3–6]. Although SNPs can provide an alternative approach, on a per locus basis they have a lower power of discrimination (PD) compared with STRs [3]. Therefore, larger SNP panels would be needed for

forensic identification if the same level of discrimination is desired [7–12].

Through the Human Genome Project (HGP), block structures consisting of tightly linked loci were discovered in the genome [13–16]. The human genome is highly organized and recombination is not distributed evenly across the genome [14]. There are recombination 'hotspots' and 'coldspots' resulting in 'block-like' patterns in the genome. Patil et al. [17] defined haplotype blocks as a region with a large proportion of inferred common haplotypes. Gabriel et al. [15] described them based on linkage disequilibrium (LD), that is large pairwise $|D'|$ values between those SNP pairs within one haploblock. Wang et al. [18] further proposed explicit "no historical recombination" as a definition for haplotype blocks, which can be tested using a four-gamete test. Recombination occurs at a reduced rate within these regions, resulting in LD of SNPs within these haplotype blocks (or haploblocks) [15]. These structures provide a novel class of SNP markers to explore for forensic utility. A set of SNPs form a haploblock with several haplotypes observed at the population and as a result, the marker has a higher discriminating power than the individual SNPs comprising the block. In theory, a single haploblock would be equivalent in PD to that of an STR marker. The haplotypes observed would be analogous to alleles at individual STR loci.

From HapMap Phase II data [16], suitable candidate haploblocks can be identified for forensic applications. The discriminating power of haploblocks was compared with that of individual SNPs. Additionally, interpretation guidelines are proposed that are applicable to haploblock data for forensic transfer evidence, mixtures, and pairwise kinship analyses.

Materials and methods

Haploblock selection

For purposes of this study, we define a haploblock as a cluster of neighboring SNPs with tight LD for each pair of included SNPs. The LD was measured with r^2, which is more commonly used and relatively stable with sample size and allele frequencies as compared with D' [19]. For potential forensic utility, a minimum threshold value for r^2=0.7 was used to select for linked SNPs within the haploblocks. Haploblocks should exist in all major populations (Caucasian, East Asian, and African) that exhibit low measures of population substructure (F_{st}<0.06) [7]. The heterozygosity of haploblocks is defined as 1−sum of square of haplotype frequencies, and a threshold of 0.2 was used in screening candidate haploblocks. This level is equivalent to the heterozygosity of a single SNP locus with a minor allele frequency 0.1 (i.e., a heterozygosity of 0.18). By following these criteria and selecting haploblocks that contain at least three SNPs, with a minimum of three common haplotypes observed in each population (i.e., shared by all populations), a haploblock marker would behave similar to that of a STR. After selecting potential haploblocks based on these criteria, each haploblock was tested for conformity to HWE expectations and haploblock pairs were tested for potential LD [20]. In the tests, each block was considered as a pseudo STR marker, in which the haplotypes were similar to the "alleles" in a STR. Threshold p values of both HWE and LD tests were set at 0.05. These criteria were combined as a parameter set (PS). As a LD filter, a greedy algorithm was used when deciding which haploblock should be removed from a pair of haploblocks in LD; the haploblock which showed LD with more haploblocks was removed. If two haploblocks had the same number of detectable LDs with other haploblocks, the one with larger heterozygosity was retained.

Match probability reduction (*mpr*) per block is an important evaluation criterion in haploblock selection. This measure assesses the match probabilities (MP) of haploblocks compared to that of individual SNPs to obtain the degree of reduced MP. The MP of a single haploblock is defined as the sum of square of genotype frequencies (genotype is the combination of the two homologous haplotypes). Cumulative MP of multiple haploblocks, when shown to be independent markers, is the product of the MP of each haploblock. A lower *mpr* indicates better haploblock selection compared with the individual SNPs in the same haploblock.

Using the initial filter criteria outlined in the parameter set outlined above, Phase II data available from the HapMap project [16] on chromosome 1, one of the largest chromosomes in human genome, were used to identify candidate haploblocks for potential forensic applications. Table 1 lists the number of haploblocks that meet the criteria for candidate forensic haploblocks based on chromosome 1. It also evaluates the effects of five parameters on haploblock selection by calculating the cumulative MP of haploblocks (b) and the cumulative MP of SNPs with minimum MP within each block (s). *mpr* was calculated by

$$\mathrm{mpr} = \left(\frac{b}{s}\right)^{\frac{1}{n}} \quad (1)$$

where n represents number of haploblocks selected by the above criteria.

Generally, F_{st} and r^2 can effectively impact the number of candidate haploblocks selected, thus, relaxing the F_{st} and r^2 threshold criteria can increase the number of candidate haploblocks. However, the *mpr* is relatively stable with F_{st} and r^2, although there is some variation due to a limited number of haploblocks. Low values of minimum heterozygosity (MinHet) do not significantly change the number of

Table 1 Number of Haploblocks on chromosome 1 meeting the selection criteria and effects of parameters on haploblock selection: (a) maximum F_{st}; (b) minimum r^2; (c) minimum heterozygosity (MinHet); (d) minimum number of haplotypes in each population (MinHap); (e) minimum number of SNPs (MinSNP). PS represents the predefined parameter set

Parameter set (PS)	Num. blocks with PS	Num. blocks with PS & HWE filter	Num. blocks with PS & HWE & LD filters (*n*)	Avg. Cum. MP of Blocks (*b*)	Avg. MP perblock (amp)	Cum. Min. MP of SNPs (*s*)	MP reduction per block (*mpr*)
(a)							
F_{st}=0.01; r^2=0.6; MinHet=0.2; MinHap=4; MinSNP=4	5	1	1	0.3908	0.3908	0.4456	0.8770
F_{st}=0.03; r^2=0.6; MinHet=0.2; MinHap=4; MinSNP=4	14	4	3	3.839E−2	0.3373	6.913E−2	0.8220
F_{st}=0.06; r^2=0.6; MinHet=0.2; MinHap=4; MinSNP=4	25	9	5	3.599E−3	0.3245	1.214E−2	0.7841
F_{st}=0.1; r^2=0.6; MinHet=0.2; MinHap=4; MinSNP=4	33	13	8	2.734E−4	0.3586	1.317E−3	0.8216
(b)							
F_{st}=0.06; r^2=0.5; MinHet=0.2; MinHap=4; MinSNP=4	53	25	11	2.7585E−6	0.3123	4.075E−5	0.7829
F_{st}=0.06; r^2=0.6; MinHet=0.2; MinHap=4; MinSNP=4	25	9	5	3.699E−3	0.3245	1.214E−2	0.7841
F_{st}=0.06; r^2=0.7; MinHet=0.2; MinHap=4; MinSNP=4	14	3	3	3.950E−2	0.3406	6.372E−2	0.8526
F_{st}=0.06; r^2=0.8; MinHet=0.2; MinHap=4; MinSNP=4	6	0	0				
(c)							
F_{st}=0.06; r^2=0.6; MinHet=0.2; MinHap=4; MinSNP=4	25	9	5	3.599E−3	0.3245	1.214E−2	0.7841
F_{st}=0.06; r^2=0.6; MinHet=0.3; MinHap=4; MinSNP=4	21	9	5	3.599E−3	0.3245	1.214E−2	0.7841
F_{st}=0.06; r^2=0.6; MinHet=0.4; MinHap=4; MinSNP=4	20	7	4	9.162E−3	0.3094	2.753E−3	0.7595
F_{st}=0.06; r^2=0.6; MinHet=0.5; MinHap=4; MinSNP=4	7	2	1	0.2808	0.2808	0.3748	0.7491
(d)							
F_{st}=0.06; r^2=0.6; MinHet=0.2; MinHap=3; MinSNP=4	155	97	21	4.722E−10	0.3597	6.067E−9	0.8855
F_{st}=0.06; r^2=0.6; MinHet=0.2; MinHap=4; MinSNP=4	25	9	5	3.599E−3	0.3245	1.214E−2	0.7841
F_{st}=0.06; r^2=0.6; MinHet=0.2; MinHap=5; MinSNP=4	6	1	1	0.2728	0.2728	0.4332	0.6297
(e)							
F_{st}=0.06; r^2=0.6; MinHet=0.2; MinHap=4; MinSNP=3	41	19	9	9.533E−5	0.3575	4.346E−4	0.8449
F_{st}=0.06; r^2=0.6; MinHet=0.2; MinHap=4; MinSNP=4	25	9	5	3.598E−3	0.3245	1.214E−2	0.7841
F_{st}=0.06; r^2=0.6; MinHet=0.2; MinHap=4; MinSNP=5	11	4	4	1.062E−2	0.3210	2.765E−2	0.7871
F_{st}=0.06; r^2=0.6; MinHet=0.2; MinHap=4; MinSNP=6	6	3	3	2.714E−2	0.3005	6.240E−2	0.7576

Cum. cumulative, *avg* average, *Cum. MP* is the average of cumulative match probability over three populations, *Cum. Min. MP of SNPs* is the average of cumulative match probabilities of SNPs with minimum MP in each block. *b* and *s* were calculated based on haploblock selection after both HWE and LD filters. MP reduction per block (*mpr*) was calculated in terms of Eq. 1

[a] HapMap Phase II data

haploblocks that meet the criteria. Most haploblocks have a heterozygosity that is less than 0.4. Minimum number of haplotypes (MinHap) is a more significant parameter affecting *mpr*, since high number of haplotypes decreases match probability of a haploblock. This is correlated with the number of SNPs within the haploblock. As an example, haploblocks with five haplotypes in each population generated a *mpr* less than 0.7. The average MP per block (amp) is generally from 0.3 to 0.4, and can be less than 0.3 with a high MinHap.

Refining haploblock selection with *mpr*

To further refine the candidate list, the *mpr* was incorporated in to the parameter set as a primary parameter for haploblock selection criteria and several alternate parameter sets were evaluated to determine optimal selection and retention criteria. To reach a high discrimination power, approximately ten haploblocks selected based on the above criteria and the lowest maximum mpr (*mmpr*) were needed before applying the LD filter for chromosome 1. Assuming these same criteria, about 100~200 haploblocks across the 22 autosomal chromosomes would be expected. Additionally, a relatively high r^2 with a low F_{st} is desirable for forensic applications. Table 2 lists 24 parameter sets for an optimal parameter set searching. By selecting low MinHet, MinHap, and MinSNP values (e.g., MinHet=0.2; MinHap=3; MinSNP=3), a low r^2 (e.g., 0.6) may provide a reasonable number of haploblocks even with low F_{st} (e.g., 0.01 or 0.03). With a high r^2=0.8,

Table 2 Parameter sets for appropriate haploblock selection with Maximum mpr (*mmpr*), F_{st}, r^2, MinHet, MinHap, and MinSNP

Index	Parameter set (*PS*)	Num. blocks with PS	Num. blocks with PS & HWE filter	Num. blocks with PS & HWE & LD filters (*n*)	Avg. Cum. MP of Blocks (*b*)	Avg. MP per block (*amp*)	Cum. Min. MP of SNPs (*s*)	MP reduction per block (*mpr*)
1	mmpr=0.8; F_{st}=0.01; r^2=0.6; MinHet=0.2; MinHap=3; MinSNP=3	4	3	3	3.111E−2	0.3145	6.563E−2	0.7797
2[a]	mmpr=0.85; F_{st}=0.01; r^2=0.6; MinHet=0.2; MinHap=3; MinSNP=3	11	9	6	9.207E−4	0.3119	3.332E−3	0.8070
3	mmpr=0.9; F_{st}=0.01; r^2=0.6; MinHet=0.2; MinHap=3; MinSNP=3	26	19	9	5.137E−5	0.3337	2.723E−4	0.8308
4	mmpr=0.85; F_{st}=0.01; r^2=0.7; MinHet=0.2; MinHap=3; MinSNP=3	3	3	3	3.293E−2	0.3205	6.172E−2	0.8111
5[a]	mmpr=0.8; F_{st}=0.03; r^2=0.6; MinHet=0.2; MinHap=3; MinSNP=3	12	10	6	7.690E−4	0.3027	3.790E−3	0.7666
6	mmpr=0.8; F_{st}=0.03; r^2=0.7; MinHet=0.2; MinHap=3; MinSNP=3	2	2	1	0.3027	0.3027	0.3847	0.7868
7	mmpr=0.85; F_{st}=0.03; r^2=0.7; MinHet=0.2; MinHap=3; MinSNP=3	6	6	3	3.293E−2	0.3205	6.172E−2	0.8111
8	mmpr=0.85; F_{st}=0.03; r^2=0.8; MinHet=0.2; MinHap=3; MinSNP=3	0	0	0				
9	mmpr=0.8; F_{st}=0.06; r^2=0.6; MinHet=0.2; MinHap=3; MinSNP=3	22	17	12	5.297E−7	0.2999	1.726E−5	0.7480
10	mmpr=0.8; F_{st}=0.06; r^2=0.7; MinHet=0.2; MinHap=3; MinSNP=3	2	2	1	0.3027	0.3027	0.3847	0.7868
11	mmpr=0.8; F_{st}=0.06; r^2=0.8; MinHet=0.2; MinHap=3; MinSNP=3	0	0	0				
12	mmpr=0.85; F_{st}=0.06; r^2=0.6; MinHet=0.2; MinHap=3; MinSNP=3	58	45	15	1.779E−8	0.3043	8.806E−7	0.7710
13[b]	mmpr=0.85; F_{st}=0.06; r^2=0.7; MinHet=0.2; MinHap=3; MinSNP=3	9	9	6	1.113E−3	0.3219	3.749E−3	0.8168
14	mmpr=0.85; F_{st}=0.06; r^2=0.8; MinHet=0.2; MinHap=3; MinSNP=3	0	0	0				
15	mmpr=0.9; F_{st}=0.06; r^2=0.8; MinHet=0.2; MinHap=3; MinSNP=3	8	4	4	1.416E−2	0.3450	2.274E−2	0.8883
16	mmpr=0.85; F_{st}=0.06; r^2=0.6; MinHet=0.2; MinHap=3; MinSNP=4	32	24	10	7.037E−6	0.3053	9.131E−5	0.7739
17[a]	mmpr=0.8; F_{st}=0.06; r^2=0.6; MinHet=0.2; MinHap=3; MinSNP=4	10	8	6	5.707E−4	0.2880	3.632E−3	0.7345
18	mmpr=0.8; F_{st}=0.06; r^2=0.7; MinHet=0.2; MinHap=3; MinSNP=4	0	0	0				
19	mmpr=0.85; F_{st}=0.06; r^2=0.7; MinHet=0.2; MinHap=3; MinSNP=4	4	4	4	1.127E−2	0.3258	2.447E−2	0.8238
20	mmpr=0.8; F_{st}=0.06; r^2=0.7; MinHet=0.2; MinHap=4; MinSNP=3	0	0	0				
21	mmpr=0.85; F_{st}=0.06; r^2=0.7; MinHet=0.2; MinHap=4; MinSNP=3	1	1	1	0.3093	0.3093	0.3826	0.8085
22[a]	mmpr=0.85; F_{st}=0.06; r^2=0.7; MinHet=0.3; MinHap=3; MinSNP=3	9	9	6	1.113E−3	0.3219	3.749E−3	0.8168
23[a]	mmpr=0.85; F_{st}=0.06; r^2=0.7; MinHet=0.4; MinHap=3; MinSNP=3	8	8	5	3.325E−3	0.3194	9.406E−3	0.8123
24	mmpr=0.85; F_{st}=0.06; r^2=0.7; MinHet=0.5; MinHap=3; MinSNP=3	4	4	4	1.075E−2	0.3220	2.458E−3	0.8132

b, ,*s* and *mpr* were calculated by the same method as in Table 1

[a] Presents the sub-optimal parameter sets for haploblock selection

[b] Presents the optimal parameter set for haploblock selection

no candidate haploblocks were discovered, unless the *mmpr* was relaxed to 0.9. Reducing r^2 to 0.7, with F_{st}= 0.06 and *mmpr*=0.85, nine haploblocks were found after HWE filtering, suggesting that this appears to be a reasonable parameter set. Increasing the MinHet from 0.2 to 0.4 did not significantly change the haploblock selection, which is consistent with the results in Table 1. Since a high r^2 is crucial in haploblock definition, and *mmpr* and F_{st} are secondary criteria, we chose parameter set No. 13 (*mmpr*=0.85; F_{st}=0.06; r^2=0.7; MinHet=0.2; MinHap=3; MinSNP=3) in Table 2 as the optimal parameter set for further interrogation of the HapMap Phase II data.

The numbers of haploblocks returned from the HapMap Phase II data with optimal *PS* were detailed in Table 3. A total of 253 haploblocks across the 22 autosomal chromosomes met the selection criteria. Only 170 haploblocks remained after the dataset was filtered for HWE, and after the LD filter was applied, only 24 haploblocks containing 138 SNPs were retained. The largest haploblock contains 27 SNPs resulting in 13 haplotypes across the set of all populations. The average heterozygosity of the haploblocks was 0.5293, and average F_{st} was 0.024. The average match probability of the haploblocks was 0.3178, and the cumulative power of discrimination of these 24 haploblocks was approximately 10^{-12}, which is lower than that of SNPs with minimal match probability in each haploblock and close to the MP of about 10 STR loci [3]. The rs# of the SNPs are listed in the Supplementary materials.

The observed and expected distributions (under the assumption of independence of haplotypes within and across loci) of haplotypes (i.e., pseudo alleles in haploblocks) and genotype sharing within each population were compared by the computational methods described by Chakraborty and Jin [21] (Table 4 and Fig. 1). The means and standard deviations of the distributions were not significantly different ($p>0.05$). These results are similar to that reported in STRs [1] and indicate that the haplotypes within or across haploblocks are independent in the populations examined. In addition, Yorubans have relatively lower average genotype and haplotype sharing compared with the other populations, which is consistent with the higher diversities observed within African populations [22].

Table 3 Number of haploblocks with optimal parameter set (e.g., *mmpr*=0.85; F_{st}=0.06; r^2=0.7; MinHet=0.2; MinHap=3; MinSNP=3)

Chromosome	Num. blocks with PS	Num. blocks with PS & HWE filter	Num. blocks with PS & HWE & LD filters (*n*)	Avg. Cum. MP of blocks (*b*)	Cum. Min. MP of SNPs (*s*)	MP reduction per block (*mpr*)	Num. Of SNPs
1	9	9	0				
2	23	14	1	0.3287	0.4050	0.8117	6
3	12	10	2	0.1144	0.1617	0.8412	9
4	21	15	1	0.2926	0.3765	0.7773	6
5	16	12	3	0.02633	0.05480	0.7833	25
6	15	10	0				
7	16	9	2	0.1035	0.1465	0.8403	30
8	18	12	2	0.1025	0.1518	0.8215	7
9	8	6	0				
10	15	8	1	0.3527	0.4169	0.8460	4
11	14	12	3	0.03872	0.06700	0.8209	13
12	12	5	1	0.3036	0.3890	0.7806	5
13	17	14	3	0.0344	0.06409	0.8123	14
14	10	6	3	0.02339	0.04789	0.7876	11
15	9	4	0				
16	7	4	1	0.3310	0.4053	0.8167	3
17	5	4	0				
18	8	7	1	0.3123	0.3689	0.8465	5
19	5	4	0				
20	6	1	0				
21	6	3	0				
22	1	1	0				
Total	253	170	24	1.059E−12	1.566E−10	0.8121	138

b, *s* and *mpr* were calculated by the same method as Table 1. LDs were tested across all candidate haploblocks

Table 4 Mean and standard deviations of expected and observed haplotype and genotype sharing of 24 haploblocks with optimal parameter set (e.g., *mmpr*=0.85; F_{st}=0.06; r^2=0.7; MinHet=0.2; MinHap=3; MinSNP=3)

Population	Genotype sharing				Haplotype sharing			
	Expected		Observed		Expected		Observed	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
CEU	7.68	2.28	7.39	2.14	28.34	3.18	27.71	3.14
JPT+CHB	8.21	2.32	8.13	2.28	29.11	3.19	29.06	3.20
YRI	6.63	2.18	6.62	2.13	26.71	3.18	26.60	3.07

CEU Caucasian, Utah USA; *JPT* Japanese, Tokyo; *CHB* Han Chinese, Beijing; *YRI* Yoruba, Nigeria

Evidence interpretation based on haploblocks

Traditionally, forensic DNA evidence has been tested using sets of individual with independent SNPs or STR loci. To date, little effort has focused on the application of haploblocks on the autosomal chromosomes. Haploblocks offer the possibility to extend the features of lineage-based markers, such as mitochondrial DNA and Y chromosome markers, to markers residing on the autosomal chromosomes. But there are some considerations regarding haploblock SNPs that must be addressed to optimize their implementation. In most samples, the specific haplotypes that comprise a heterozygous individual may not be unambiguously determined. Multiple haplotype combinations are possible; but because of strong LD, some haplotypes will be more frequent than others in a given reference population. To address this phenomenon, we suggest some notations for interpretation.

Notations and definitions

Suppose there are n haplotypes in a haploblock, let $H_1, H_2, \ldots, H_n$ be the haplotypes with frequencies $p_1, p_2, \ldots, p_n$, respectively. A genotype (G) is *comprised* of a set of haplotypes if this set of haplotypes can compose the exact G *for a diploid individual*. For example, a genotype with two heterozygous loci (A/T; A/T) can be comprised of the two haplotypes AT and TA, or alternatively by the haplotypes AA and TT.

Transfer evidence

Transfer evidence analysis can involve the comparison of a single-source DNA profile from crime scene evidence with a DNA profile derived from a known sample (e.g. suspect or other potential contributor). This type of analysis is straightforward and similar to that routinely performed for STRs [23]. An exclusion or inclusion can be determined by simply comparing the genotypes of haploblocks. If the comparison results in an inclusion, then the random match probability can be calculated as the sum of the frequencies of all possible pairwise haplotype combinations Eq. 2.

$$\Pr(G) = \sum_{\substack{\text{Haplotype combination} \\ (Hi,\, Hj)\text{composes G}}} p_i p_j \qquad (2)$$

For example, consider a two-locus haploblock with four haplotypes TT, TA, AA, and AT and frequencies 0.4, 0.3, 0.2, and 0.1, respectively. Both the known sample and evidentiary sample have genotype (AT) at both loci. Then, the random match probability is 0.11, which is sum of two possible haplotype combinations (i.e., haplotypes TT and AA with frequency $0.4\times0.2=0.08$; haplotypes TA and AT with frequency $0.3\times0.1=0.03$).

However, there may be some situations where an apparent single-source sample may be a mixed sample (e.g., a heterozygote may not be distinguishable from two homozygotes mixed together or two heterozygotes sharing the same alleles are mixed together). These scenarios are more likely if SNP alleles are common, if the contributors of the mixture are in relatively equal proportions, and if the assay is not quantitative. Evaluating all the haploblocks by the mixture detection method described below for assuming a single-source sample is strongly recommended.

Mixture evaluation

Since the haplotypes of a haploblock cannot be directly observed, all possible haplotype combinations that comprise a mixture need to be inferred and frequencies of these haplotypes are based on the observed genotype and haplotype distributions in a reference population(s). The probability of the genotypes of potential contributors can be calculated by summing the genotype frequencies of all

possible even numbers (e.g., 2, 4, 6, … , etc.) of haplotype combinations Eq. 3.

$$\Pr(G) = \sum_{\substack{\text{Haplotype combination} \\ \left(H_{k(1)},\ldots,H_{k(2n)}\right)\text{composes } G}} \prod_{i=k(1)}^{k(2n)} p_i \qquad (3)$$

This essentially is the same as the probability of inclusion used in forensic DNA mixture calculations or the random man not excluded calculation used in paternity testing [23].

To estimate the most likely number of contributors (N) in a mixture, the likelihood of a mixed profile can be computed given a different numbers of contributors.

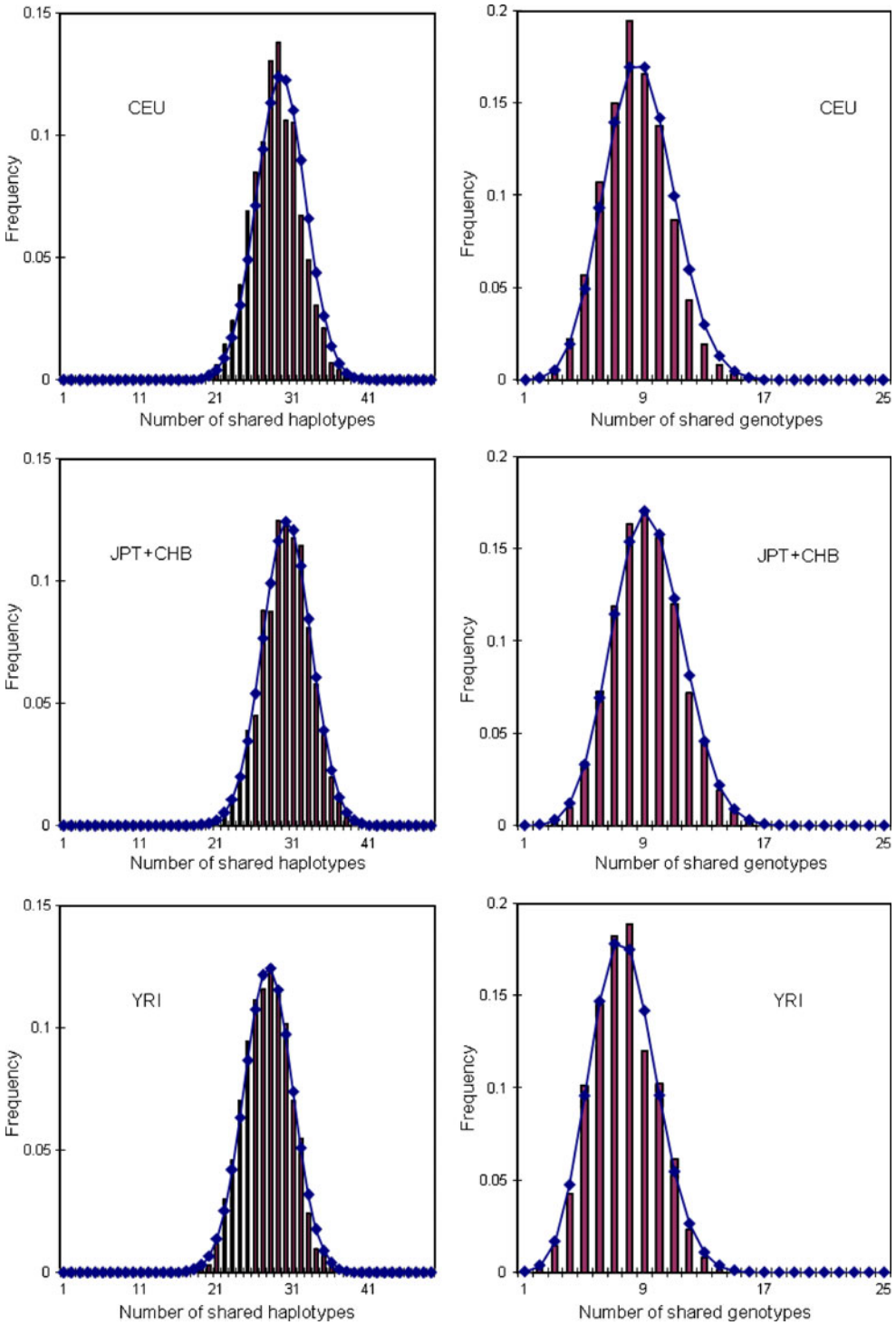

Fig. 1 Distributions of observed and expected (under the assumption of haplotype independence within and across loci) number of haplotypes and genotypes shared in 24 haploblocks of pairs of individuals within three main populations. Histograms represent the observed distributions and the line diagrams correspond to the expected distributions (see [21] for computational methods)

Equation 4 calculates the probability of a profile given 1, 2, and 3 unrelated contributors, and a greater number of contributors can be calculated in a similar fashion. The number of contributors is estimated by maximizing the likelihoods of haplotypes. Assuming independence between haploblocks, the likelihood of multiple haploblocks is the product of the likelihood of each haploblock.

$$\begin{aligned}
\Pr(G|N=1) &= \sum_{\substack{\text{Haplotype combination}\\(Hi,Hj)\text{composes } G}} p_i p_j \\
\Pr(G|N=2) &= \sum_{\substack{\text{Haplotype combination}\\(Hi,Hj,Hk,Hl)\text{composes } G}} p_i p_j p_k p_l \\
\Pr(G|N=3) &= \sum_{\substack{\text{Haplotype combination}\\(Hi,Hj,Hk,Hl,Hm,Hn)\text{composes } G}} p_i p_j p_k p_l p_m p_n
\end{aligned} \tag{4}$$

Two approaches have been used to convey the weight of the evidence: The probability of exclusion (PE) and the likelihood ratio (LR). The PE is the sum of the probabilities of all possible haplotypes that cannot be part contributors of the mixture evidence and does not require an assessment of the number of contributors comprising the mixture. However, if the number of contributors can be estimated with high confidence, a LR can be calculated by comparing two probabilities of observing the DNA mixture evidence given the alternative hypotheses. For instance, in a simple mixture with two contributors (S is suspect; V is victim; UN is a possible unknown contributor), one possible LR is

$$\mathrm{LR} = \frac{\Pr(V+S)}{\Pr(V+\mathrm{UN})} \tag{5}$$

Pairwise kinship estimation

Barring mutation, let $X_i=(H_{i1}, H_{i2})$ be one haplotype combination comprising genotype (G), and there are total K combinations which can comprise G. The likelihood of X_i, $P(X_i)$, equals to ($p_{i1}\times p_{i2}$). We define w_i as the weight of X_i. The sum of all w_i is 1.

$$w_i = P(X_i) / \sum_{i=1}^{K} P(X_i) \tag{6}$$

Suppose genotypes of two individuals can be comprised of two sets of haplotype combinations, X_{i1} ($i=1, \ldots, k_1$) and X_{j2} ($j=1, \ldots, k_2$), with weight w_{i1} and w_{j2}, respectively, k_1 and k_2 are the number of possible haplotype combinations for the two individuals. The likelihood of these two individuals given relationship (R) is the summation of ($k_1\times k_2$) pairs of likelihoods with specific haplotype combinations Eq. 7. Thompson [24] described the details for calculating the likelihood of two genotypes for a given relationship. Other statistics or scores for a single locus (e.g., IBS scores) can be calculated in a similar manner. The time complexity of the calculation is quadratic, depending on k_1 and k_2.

$$L_{\mathrm{Block}} = \sum_{i=1}^{k_1}\sum_{j=1}^{k_2} w_{i1} w_{j2} L\left(X_{i1}, X_{j2}|R\right) \tag{7}$$

Discussion

This is the first report attempting to select haploblock structures for identity testing and forensic purposes. Haploblock selection criteria were established to identify candidate haploblocks for forensic individualization purposes. These include strong LD, a minimum number of SNPs to define a haploblock, a high heterozygosity, low F_{st} (i.e., low population heterogeneity), and meeting HWE. Under these criteria fewer haploblocks will be needed to reach high levels of PD and fewer reference population data sets will be required for statistical assessments for forensic casework.

With the best parameter set, 253 haploblocks were found. After HWE and LD filtering, 24 haploblocks remained as good candidates for forensic utility. The cumulative match probability of these haploblocks is approximately 10^{-12} for the populations studied. If criteria are further relaxed, more candidate haploblocks can be identified. For most haploblocks, the match probability only slightly decreased (<10%) compared with the SNPs with minimum MP within haploblocks. This is due to the fact that most of the haploblocks have two major haplotypes and the sum of frequencies of all minor haplotypes is relatively small [25]. The overall PD can be increased by selecting haploblocks with a greater number of haplotypes in each population. This may be a challenge requiring haploblocks with more SNPs and possibly relaxing the r^2 criterion. The greedy algorithm may not yield the optimal set of candidate haploblocks. The greedy algorithm can bring instability in haploblock selection; one haploblock selected by one parameter set may be filtered out for another parameter set. However, considering the computation complexity with a large number of haploblocks, a greedy algorithm is a practical solution. Therefore, the LD filter was invoked as the last step for haploblock selection.

Another consideration for forensic utility is the differences in haplotype frequencies among reference populations. Most haploblocks have an F_{st} greater than 0.06, which is relatively higher than that for STR markers [26]. Low F_{st} haploblocks may provide more consistent haplotype frequencies across populations and are better candidates for identity testing. Alternatively, high F_{st} haploblocks can be used for assessing ethnic origin and admixture, but require more population data to be generated.

The mixture interpretation of haploblock profiles is more complicated than that for STRs, because of ambiguous haplotype composition of haploblock genotypes. However, it is relatively easier than individual SNPs because of the number of haplotypes (i.e., alleles) and their structure. Moreover, the number or the range of number of contributors of a mixed profile can be more readily estimated with haploblocks than with unlinked individual SNP panels. Because the mutation rate of a SNP is very low ($\sim 10^{-8}$) [27] and the haploblocks by definition have strong LD, mutation events within haploblocks generally will not impact kinship investigations. Recombination within haploblocks can also be ignored because of the limited number of generations in kinship analysis. Although the haploblocks only have a small decrease in MP than a SNP with 50% heterozygosity for single-source identity testing, they can be used for IBD analyses where their power is much higher than individual SNPs. Currently, haploblock interpretation is based on genotype since there may exist multiple haplotype combinations which can be combined and present as the same genotype. Genotype-based interpretation somewhat decreases the PD of the haploblock. Hence, those haploblocks in which the haplotypes can be unambiguously determined in each population or in the total population are preferred.

Human genome block structures can differ among populations [13–15]. Our results have shown that Africans have significantly less haploblocks than Caucasians and East Asians. There were 11,741, 12,456, and 12,237 haploblocks given r^2=0.7 and MinSNP=3 on chromosome 1 for Caucasian, Chinese, and Japanese, respectively. However, only 7,318 haploblocks were found in Yoruba with the same parameters. It confirms that there are different haploblock structures among Africans, Caucasian, and Asians [13–15]. To increase the PD when using haploblocks, population-specific haploblocks instead of across populations, may provide more candidate haploblocks.

Lastly, an assay will need to be developed to type the haploblocks selected. There may be specific molecular constraints on some SNPs that may make them more difficult to analyze than others; thus reducing some of candidate haploblocks. As always, proper validation is required.

References

1. Chakraborty R, Stivers DN, Su B, Zhong Y, Budowle B (2000) The utility of STR loci beyond human identification: implications for the development of new DNA typing systems. Electrophoresis 20:1682–1696
2. Budowle B, Shea B, Niezgoda S, Chakraborty R (2001) CODIS STR loci data from 41 sample populations. J Forensic Sci 46 (3):453–489
3. Gill P (2001) An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes. Int J Legal Med 114:204–210
4. Amorim A, Pereira L (2005) Pros and cons in the use of SNPs in forensic kinship investigation: a comparative analysis with STRs. Foren Sci Int 150:17–21
5. Budowle B, van Daal A (2008) Forensically relevant SNP classes. Biotechniques 44:603–610
6. Phillips C, Fondevila M, García-Magarinos M, Rodriguez A, Salas A, Carracedo A, Lareu MV (2008) Resolving relationship tests that show ambiguous STR results using autosomal SNPs as supplementary markers. Forensic Sci Int Genet 2:198–204
7. Kidd KK, Pakstis AJ et al (2006) Developing a SNP panel for forensic identification of individuals. Forensic Sci Int 164:20–32
8. Vallone PM, Decker AE, Butler JM (2005) Allele frequencies for 70 autosomal SNP loci with U.S. Caucasian, African American, and Hispanic samples. Foren Sci Int 149:279–286
9. Dixon LA, Murray CM, Archer EJ, Dobbins AE, Koumi P, Gill P (2005) Validation of a 21-locus autosomal SNP multiplex for forensic identification purposes. Forensic Sci Int 154:62–77
10. Sanchez JJ, Phillips C, Borsting C et al (2006) A multiplex assay with 52 single nucleotide polymorphisms for human identification. Electrophoresis 27(9):1713–1724
11. Pakstis AJ, Speed WC, Kidd JR, Kidd KK (2007) Candidate SNPs for a universal individual identification panel. Hum Genet 121:305–317
12. Amigo J, Phillips C, Lareu M, Carracedo Á (2008) The SNP *for* ID browser: an online tool for query and display of frequency data from the SNP *for* ID project. Int J Legal Med 122:435–440
13. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nat Genet 29:229–232
14. McVean GA, Myers SR, Hunt S, Deloukas P et al (2004) The fine-scale structure of recombination rate variation in the human genome. Science 23(304):581–584
15. Gabriel SB, Schaffner SF, Nguyen H et al (2002) The structure of haplotype blocks in the human genome. Science 296:2225–2229
16. International HapMap Consortium (2003) The international HapMap project. Nature 406:789–796
17. Patil N, Berno AJ, Hinds DA, Barrett WA et al (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294:1719–1723
18. Wang N, Akey JM, Zhang K, Chakraborty R, Jin J (2002) Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. Am J Hum Genet 71:1227–1234
19. Devlin B, Risch N (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29:311–322
20. Weir BS (1996) Genetic data analysis II: methods for discrete population genetic data. Sinauer, Sunderland
21. Chakraborty R, Jin L (1993) Determination of relatedness between individuals using DNA fingerprinting. Hum Biol 65:875–895
22. Papiha SS, Deka R, Chakraborty R (1999) Genomic diversity: applications in human population genetics. Kluwer/Plenum, New York
23. Evett IW, Weir BS (1998) Interpreting DNA evidence. Sinauer, Sunderland, MA
24. Thompson EA (1975) The estimation of pairwise relationships. Ann Hum Genet 39:173–188
25. Zhang J, Rowe WL, Clark AG, Buetow KH (2003) Genome wide distribution of high-frequency, completely mismatching SNP haplotype pairs observed to be common across human populations. Am J Hum Genet 73:1073–1081
26. NRC II, National Research Council Committee on DNA Forensic Science (1996) The evaluation of forensic DNA evidence. National Academy Press, Washington, DC
27. Reich DE, Schaffner SF, Daly MJ, McVean G, Mullikin JC, Higgins JM, Richter DJ, Lander ES, Altshuler D (2002) Human genome sequence variation and the influence of gene history, mutation and recombination. Nat Genet 32:135–140