frontiers in Genetics

METHODS
published: 07 August 2020
doi: 10.3389/fgene.2020.00781

MicroHapDB: A Portable and Extensible Database of All Published Microhaplotype Marker and Frequency Data

Daniel S. Standage and Rebecca N. Mitchell*

National Bioforensic Analysis Center, National Biodefense Analysis and Countermeasures Center (NBACC), Frederick, MD, United States

Microhaplotypes are the subject of significant interest in the forensics community as a promising multi-purpose forensic DNA marker for human identification. Microhaplotype markers are composed of multiple SNPs in close proximity, such that a single NGS read can simultaneously genotype the individual SNPs and phase them in aggregate to determine the associated donor haplotype. Abundant throughout the human genome, numerous recent studies have sought to discover and rank microhaplotype markers according to allelic diversity within and among populations. Microhaplotypes provide an appealing alternative to STR markers for human identification and mixture deconvolution, but can also be optimized for ancestry inference or combined with phenotype SNPs for prediction of externally visible characteristics in a multiplex NGS assay. Designing and evaluating panels of microhaplotypes is complicated by the lack of a convenient database of all published data, as well as the lack of population allele frequency data spanning disparate marker collections. We present MicroHapDB, a comprehensive database of published microhaplotype marker and frequency data, as a tool to advance the development of microhaplotype-based human forensics capabilities. We also present population allele frequencies derived from 26 global population samples for all microhaplotype markers published to date, facilitating the design and interpretation of custom multi-source panels. We submit MicroHapDB as a resource for community members engaged in marker discovery, population studies, assay development, and panel and kit design.

Keywords: microhaplotype, forensics, human identification, next generation sequencing, Python, database, bioinformatics

OPEN ACCESS

Edited by:
Kenneth K. Kidd,
Yale University, United States

Reviewed by:
Guanglin He,
Sichuan University, China
Yiping Hou,
Sichuan University, China
Dennis McNevin,
University of Technology Sydney, Australia

****Correspondence:***
Daniel S. Standage
daniel.standage@nbacc.dhs.gov

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *08 April 2020*
Accepted: *30 June 2020*
Published: *07 August 2020*

Citation:
Standage DS and Mitchell RN (2020) MicroHapDB: A Portable and Extensible Database of All Published Microhaplotype Marker and Frequency Data. Front. Genet. 11:781. doi: 10.3389/fgene.2020.00781

1. INTRODUCTION

Well-studied short tandem repeat (STR) markers have formed the basis of forensic human identification methods since the 1990s. The most common strategy in practice today utilizes several fluorescent dyes to type 20 or more STR markers in a single polymerase chain reaction (PCR) followed by capillary electrophoresis (CE) detection (Butler, 2010). The resulting DNA profiles, combined with STR allele frequency estimates, can then be used to calculate match statistics or evaluate the relative weight of evidence for competing propositions in a likelihood ratio framework

(Butler, 2015; Cowell et al., 2015; Bleka et al., 2016a,b). Statistics obtained via STR typing can provide high confidence given the number of independent markers in an assay and the multiallelic nature of each marker.

Despite impressive recent improvements in DNA sequencing technologies, next-generation sequencing (NGS) assays of single nucleotide polymorphism (SNP) markers have seen slow adoption for forensic human identification. The ability to genotype sufficient numbers of SNPs to achieve suitable statistical power remains beyond the scope of many forensics laboratories. A relevant factor is the forensics community's strong disinclination, on ethical and privacy grounds, to use DNA markers associated with human diseases or conditions, which limits the utilization of many commonly used microarrays and SNP chips. Also, because the majority of SNPs are bi-allelic, less population-level diversity is observed at each marker than at multi-allelic STRs, resulting in reduced discriminatory power when comparing reference and evidentiary samples. While this can be compensated for to some extent with a larger panel (SNPs are incredibly abundant in the human genome), the statistical requirement for markers that are inherited independently complicates panel design and places a practical limit on the resulting panel size.

Microhaplotypes (often abbreviated as *microhaps* or *MHs*) have recently prompted considerable interest in the forensics community as a promising alternative to independent SNPs and STRs for human identification (Kidd et al., 2018; Oldoni et al., 2019). A microhaplotype marker is defined by multiple SNPs[1] residing within a short genomic distance whose state is reported as the allelic combination of all its component SNPs–that is, the haplotype. Here, "short" simply means a few hundred base pairs or fewer, ensuring a low frequency of recombination within the marker, and that a single NGS read or read pair can span all of the marker's component SNPs. This length constraint enables each distinct read to both genotype and phase its target marker; that is, to determine (1) the individual allele of each component SNP, as well as (2) the haplotype. Even if a particular microhap is composed only of biallelic SNPs, the presence of multiple component SNPs makes it possible to observe several haplotypes at the marker, substantially increasing its discriminatory power over independent SNPs.

With a targeted NGS sequencing assay, a sufficient number of reads are collected to confidently genotype each marker, differentiating between true haplotypes and those arising from sequencing error. Microhap markers exhibit none of the stutter artifacts commonly observed in PCR-based STR assays, and the substitution and homopolymer errors common to some NGS platforms are easily resolved with sufficient depth of coverage. The restricted length of microhap markers makes them suitable for typing degraded samples, and the ability of NGS assays to capture additional rare SNPs within the microhaplotype can provide valuable information for mixture detection and analysis.

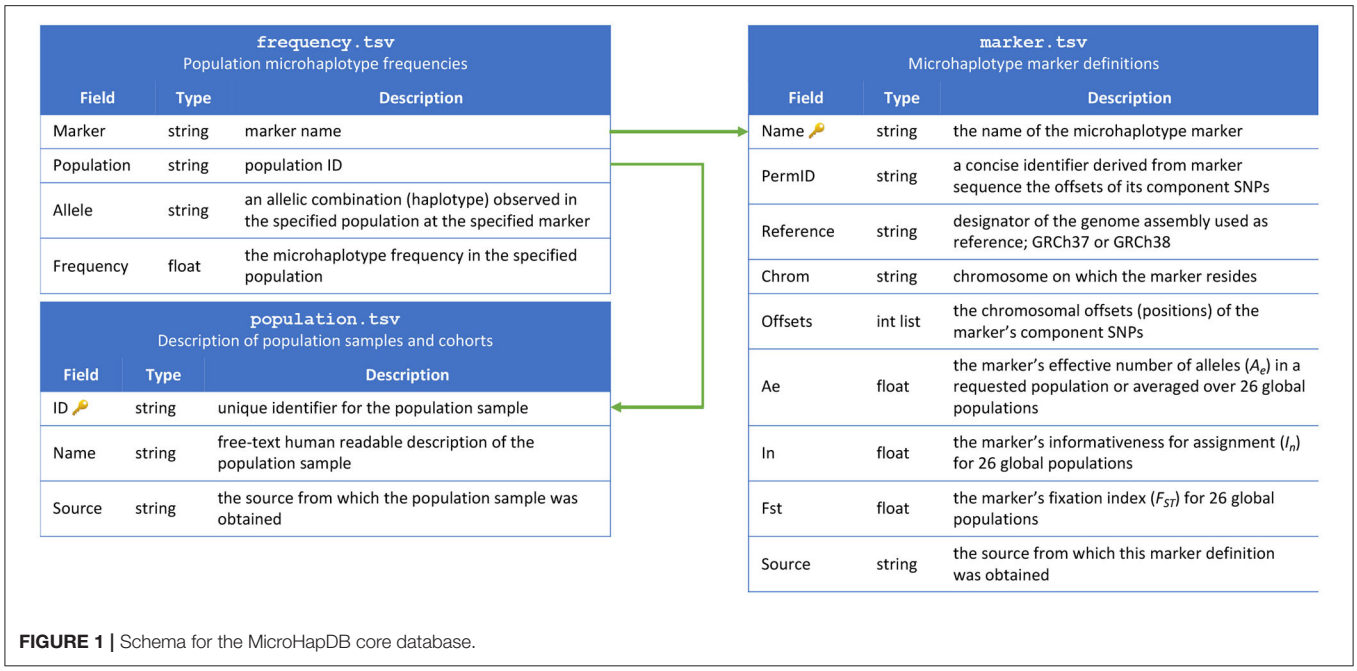
Another notable benefit of microhaps is that they can be selected not only for high *within*-population variation, but also for high *between*-population variation, facilitating prediction of biogeographic ancestry (Oldoni et al., 2017; Chen et al., 2019b; Zhu et al., 2019). It is thus possible to design a comprehensive forensic panel using a combination of microhap and SNP markers that will enable identification, mixture analysis (Bennett et al., 2019; Coble and Bright, 2019), ancestry inference, and prediction of externally visible characteristics (Ruiz et al., 2013; Walsh et al., 2013; Crawford et al., 2017) in a single NGS-based assay.

Microhaps are abundant in the human genome, and thus discovering and ranking them for different purposes is an area of active research interest in the forensics community. In just the last few years, numerous studies presenting new microhap marker collections have been published, together totaling more than 400 markers (Hiroaki et al., 2015; Kidd et al., 2018; van der Gaag et al., 2018; Voskoboinik et al., 2018; Chen et al., 2019a; Staadig and Tillmar, 2019; de la Puente et al., 2020). While two of these studies also include allele frequency data for multiple population samples—including 83 populations[2] in Kidd et al. (2018) and three populations in van der Gaag et al. (2018)—the others present either no frequency data or data for a single population sample, with little overlap between studies. The absence of a single point of access for published microhaps and the paucity of allele frequency data spanning disparate data sets are obstacles to developing and evaluating custom panels composed of microhap markers selected from different published collections.

To support the development of microhaplotype-based human forensics capabilities, we have compiled a database of all published microhap marker definitions and allele frequencies. MicroHapDB is a portable database that, once installed, can be accessed by the user without an Internet connection. The entire contents of the database are distributed with each copy of MicroHapDB, and instructions for adding private data to a local instance of the database are provided. The same instructions can alternatively be used by MicroHapDB maintainers or interested community contributors to submit new markers and allele frequencies for review and potential inclusion in the public database. MicroHapDB is designed to be user-friendly both for forensic practitioners and researchers, and supports a variety of access methods including browsing, simple or complex text queries, and programmatic database access via a Python application programming interface (API). Finally, to increase the value of the published microhaps in aggregate, we have used 2,504 fully phased genomes from the 1,000 Genomes Project (Auton et al., 2015) to estimate allele frequencies in 26 global populations for 412 microhap markers. MicroHapDB is a valuable resource for researchers, practitioners, and commercial entities engaged in marker discovery, population studies, assay development and validation, and design of custom panels and kits for forensic applications.

[1]While the vast majority of published microhaplotype markers are composed exclusively of SNPs, a small number include one or more insertion/deletion polymorphisms (indels).

[2]As of the latest December 2018 update, there are now 96 populations in the Allele Frequency Database (ALFRED) for which microhaplotype frequency data is available.



frequency.tsv
Population microhaplotype frequencies

Field	Type	Description
Marker	string	marker name
Population	string	population ID
Allele	string	an allelic combination (haplotype) observed in the specified population at the specified marker
Frequency	float	the microhaplotype frequency in the specified population

population.tsv
Description of population samples and cohorts

Field	Type	Description
ID 🔑	string	unique identifier for the population sample
Name	string	free-text human readable description of the population sample
Source	string	the source from which the population sample was obtained

marker.tsv
Microhaplotype marker definitions

Field	Type	Description
Name 🔑	string	the name of the microhaplotype marker
PermID	string	a concise identifier derived from marker sequence the offsets of its component SNPs
Reference	string	designator of the genome assembly used as reference; GRCh37 or GRCh38
Chrom	string	chromosome on which the marker resides
Offsets	int list	the chromosomal offsets (positions) of the marker's component SNPs
Ae	float	the marker's effective number of alleles (A_e) in a requested population or averaged over 26 global populations
In	float	the marker's informativeness for assignment (I_n) for 26 global populations
Fst	float	the marker's fixation index (F_{ST}) for 26 global populations
Source	string	the source from which this marker definition was obtained

FIGURE 1 | Schema for the MicroHapDB core database.

2. MATERIALS AND METHODS

2.1. Database Design

The contents of the MicroHapDB database are stored in nine tables, distributed as plain-text tab-delimited files. Three of the tables constitute the "core database," and include population sample descriptions, marker definitions, and microhaplotype frequencies (**Figure 1**). The remaining four tables contain ancillary metadata: a cross reference of third-party identifiers to MicroHapDB identifiers; data for a small number of indel variants present in the database; the genomic sequences spanning each marker (to facilitate amplicon design); a mapping of variant identifiers (rsIDs) to marker names; supplementary allelic variation statistics for specific populations; and marker coordinates for the GRCh37 reference genome assembly (GRCh38 is used by default).

2.2. User Interface

MicroHapDB is compatible with Windows and UNIX computers, and has been tested on Windows 10, Mac OS X, and Linux operating systems. In each case, the primary interface for querying MicroHapDB is the terminal or command line. The `microhapdb` command provides three operations corresponding to the three tables in the core database: `microhapdb marker`, `microhapdb population`, and `microhapdb frequency`. Executing any of these commands with no additional arguments will print the entire contents of the specified table to the terminal for browsing. Each command also enables a user to restrict the printed results to data matching a particular identifier, source, or genomic region.

By default, all results are printed in tabular format. Population data can optionally be printed in a "detail" format summarizing the number of markers for which microhaplotype frequency is available and the total number of microhaplotypes (microhaps) observed in the population sample. Marker data can optionally be printed in FASTA format for use with third-party programs, or in a "detail" format showing the genomic location of the marker and its component SNPs, the core marker sequence (spanning only the microhap's most distal component SNPs), all haplotypes observed at the marker, and a candidate amplicon sequence for the marker (for which the amount of flanking sequence can be configured using the `--delta` and `--min-length` parameters) (see **Figure 2**).

Once MicroHapDB is installed on a computer, all database list and search operations access only the data files resident on that machine. MicroHapDB doesn't require or permit requests to transfer data to or from databases residing on remote machines.

Installed as a Python package, MicroHapDB also supports programmatic access to the core and ancillary database tables. After invoking `import microhapdb`, users can write custom code to query and analyze marker, population, or frequency data resident in the database tables, pre-loaded into memory as pandas dataframes (McKinney, 2011). Alternatively, users can execute the `microhapdb --files` command from the UNIX shell to show the location of the database table files, which can be imported directly into R, Excel, or any other data analytics environment preferred by the user.

2.3. Implementation and Availability

At its core, MicroHapDB is composed of a small number of tabular plain-text data files containing marker, population, and frequency information for published microhaps. These files enclose the entire contents of the database. In contrast to many databases of genetic variation, each instance of MicroHapDB stores the entire contents of the database locally. MicroHapDB does not communicate with any central database server, and

```
$ microhapdb marker --format=detail mh04CP-002
------------------------------------------------------------[ MicroHapDB ]----
mh04CP-002    a.k.a MHDBM-342c1521, SI664878N

Marker Definition (GRCh38)
    Marker extent
        - chr4:24304952-24304972 (20 bp)
    Target amplicon
        - chr4:24304922-24305002 (80 bp)
    Constituent variants
        - chromosome offsets: 24304952,24304955,24304971
        - marker offsets: 0,3,19
        - amplicon offsets: 30,33,49
        - cross-references: rs35619595,rs34017818,rs6814654
    Observed haplotypes
        - C,A,A
        - C,A,T
        - C,G,A
        - C,G,T
        - G,A,A


--[ Core Marker Sequence ]--
>mh04CP-002
CGCGCCAGGTATGAAGTTAT


--[ Target Amplicon Sequence with Haplotypes ]--
                              *  *              *
ATTGCTAAGCATCTACTATGTGGCAAACCCCGCGCCAGGTATGAAGTTATTATGGCTGAAGATGGATAAGTCAGACAAAG
.............................C..A..............A.............................
.............................C..A..............T.............................
.............................C..G..............A.............................
.............................C..G..............T.............................
.............................G..A..............A.............................
--------------------------------------------------------------------------------
```

FIGURE 2 | The "detail" view for a 3-SNP microhaplotype marker, displayed using the MicroHapDB command-line interface.

network connections are only used to install the database or upgrade to a newer version.

The user interface described in section 2.2 is implemented in a Python package that can be installed and upgraded using the Bioconda software manager (Grüning et al., 2018). Source code for MicroHapDB is published open-access on the GitHub platform at https://github.com/bioforensics/MicroHapDB/ and is free for commercial and non-commercial use under a permissive open source license. The authors operate the MicroHapDB project under an open governance model that facilitates and encourages contributions from the community.

The software and procedure used by the authors to build the database is also published on the MicroHapDB GitHub repository. During the build process, data from several sources is independently pre-processed and standardized, and then all sources are aggregated and sorted to compile the final database. This strategy, described further in section 2.4, serves several purposes. First, it provides a clear mechanism for the authors or other community members to extend the database in the future as additional marker and frequency data is published in the literature. Second, the same mechanism enables interested users to supplement the public MicroHapDB database with private data in a safe and secure way. By following the guidelines in the database build instructions provided in the MicroHapDB repository, a user can rebuild their local copy of MicroHapDB with additional sources of marker and/or frequency data. Because MicroHapDB doesn't communicate with any central database, changes made to a user's local copy of MicroHapDB do not propagate to GitHub or any other location. Third, it permits careful scrutiny of the entire database construction process by any interested party in case errors in the database contents are ever discovered.

2.4. Data Collection and Pre-processing

MicroHapDB was compiled from seven distinct sources, each of which organized and reported data in a unique format. Extracting the relevant data and cross-referencing with public databases of genomic variation required a combination of manual and automated strategies uniquely designed for each distinct source. The result of this preliminary data acquisition and pre-processing was a collection of seven data sets with consistently formatted population descriptions, marker definitions, and allele frequencies. Once data from each distinct source was collected, cross-referenced with the GRCh37 and GRCh38 human reference genomes, and standardized, the final database compilation was performed by aggregating and sorting all data sources.

Source code and corresponding technical documentation describing data collection, pre-processing, and aggregation of the final database is available at https://github.com/bioforensics/MicroHapDB/tree/0.6/dbbuild.

2.5. Estimation of Haplotype Frequencies for 26 Global Populations

MicroHapDB includes data from several distinct sources, but the availability of population frequencies for published microhaps

is inconsistent. For some markers, frequencies are reported for dozens of population samples. Other markers have frequencies reported only for a single cohort, and yet other markers have no frequencies reported whatsoever. Designing and testing panels composed of markers from multiple distinct sources is possible, but prior to the release of MicroHapDB, interpretation of any sample assayed using such a panel would require the development of appropriate frequency data. We used 2,504 fully phased genotypes from a publicly available large-cohort study (Auton et al., 2015) to estimate population frequencies for all published microhaps across a set of 26 global population samples. These frequencies were first published in MicroHapDB version 0.5.

In the most recent version, MicroHapDB 0.6 contains definitions for 417 microhap markers. Five of these markers[3] are defined by rare variants not genotyped in the 1,000 Genomes Project Phase 3 data, and were thus excluded from this analysis. For each of the remaining 412 markers, population frequencies for 26 global populations were estimated using the following procedure. First, phased genotype records for each of the marker's component variants were retrieved using the variants' rsIDs. Next, the phased genotypes were aggregated to determine the two haplotypes for each individual at the marker (or the single haplotype observed at X chromosome markers in males). Then, noting the population sample with which each individual was associated, a tally of haplotypes was compiled for each population. Finally, the haplotype tallies for each population sample were normalized by the corresponding number of alleles to compute the final frequency estimates.

2.6. Calculation of Measures of Variation Within and Among Populations

Microhaps are suitable for numerous forensic applications. Two common statistics used for ranking microhaps are the effective number of alleles A_e and the informativeness for assignment I_n (Crow and Kimura, 1970; Rosenberg et al., 2003; Kidd and Speed, 2015; Kidd et al., 2018). The A_e statistic is the reciprocal of a marker's homozygosity. For a marker with N alleles, A_e is computed as

$$A_e = \frac{1}{\sum p_i^2}$$

where p_i is the frequency of allele $i \in N$ and summation is over all alleles. It is a measure of allelic variation within a population, and corresponds to a marker's power for individual identification. For a microhaplotype with N SNPs, the maximal A_e value of 4^N occurs if and only if every possible allelic combination is observed at equal frequencies in the population. In reality, only a subset of possible allelic combinations are generally present in a population, and typically at unequal frequencies, resulting in A_e values that most commonly fall between 1.5 and 4.5 for previously reported microhap markers. The minimal A_e value of 1 occurs when only a single haplotype is observed at the locus in a population.

By contrast, I_n measures the extent of population-specific allelic variation among a set of populations, and corresponds to a microhap's power for predicting an individual's biogeographic ancestry. The I_n statistic for a marker with N alleles across K populations is calculated as

$$I_n = \sum_{j=1}^{N} \left(-p_j \log p_j + \sum_{i=1}^{K} \frac{p_{ij}}{K} \log p_{ij} \right)$$

where p_{ij} is the frequency of allele j in population i. This statistic measures the difference in information content when allele frequencies are aggregated across all populations versus when they are collated within populations. The minimal I_n of 0 occurs when all alleles have equal frequencies in all populations, and the maximal value $\log K$ occurs when $N \geq K$ and no allele is found in more than one population (Rosenberg et al., 2003).

A third statistic, the *fixation index* (F_{ST}), is another measure of allelic variation that considers *coancestry*, and is commonly used in forensic analysis to correct for population substructure (Butler, 2015). High F_{ST} values indicate that allele frequencies differ substantially among subpopulations, while low F_{ST} values indicate higher similarity among subpopulations.

As a final post-processing step in the MicroHapDB database build procedure, A_e and I_n statistics were computed for all markers in MicroHapDB[4]. Using population microhaplotype frequencies computed from the 1,000 Genomes Project Phase 3 genotypes (Auton et al., 2015), MicroHapDB scripts computed per-marker A_e values individually for each population. By default, the `Ae` column of MicroHapDB's `markers` table displays the arithmetic mean of A_e over all 26 populations, but the command line interface and Python API both provide an option for the user to choose a specific population for which to display A_e values. I_n statistics over 26 populations were calculated with the same frequency data using INFOCALC (https://rosenberglab.stanford.edu/infocalc.html), and are listed in the `In` column of the `markers` table. The same frequency data were also used to calculate F_{ST} statistics using the Weir and Cockerham formulation (Weir and Cockerham, 1984), as implemented in the `scikit-allel` package version 1.21 (Miles et al., 2019). The F_{ST} statistics reported in MicroHapDB were averaged across all alleles for each marker.

3. RESULTS

3.1. MicroHapDB Aggregates Data for More Than 400 Microhaplotypes

MicroHapDB version 0.6, released in June 2020, includes descriptions of 102 cohorts and population samples from six sources, 417 marker definitions from seven sources, and numerous population frequencies for 5,373 observed haplotypes from six sources, all together comprising a total of 113,995

[3] mh06PK-24844, mh0XUSC-XqD, mh11PK-63643, mh15PK-75170, and mh22PK-104638.

[4] With the exception of the 5 markers discussed in footnote 3.

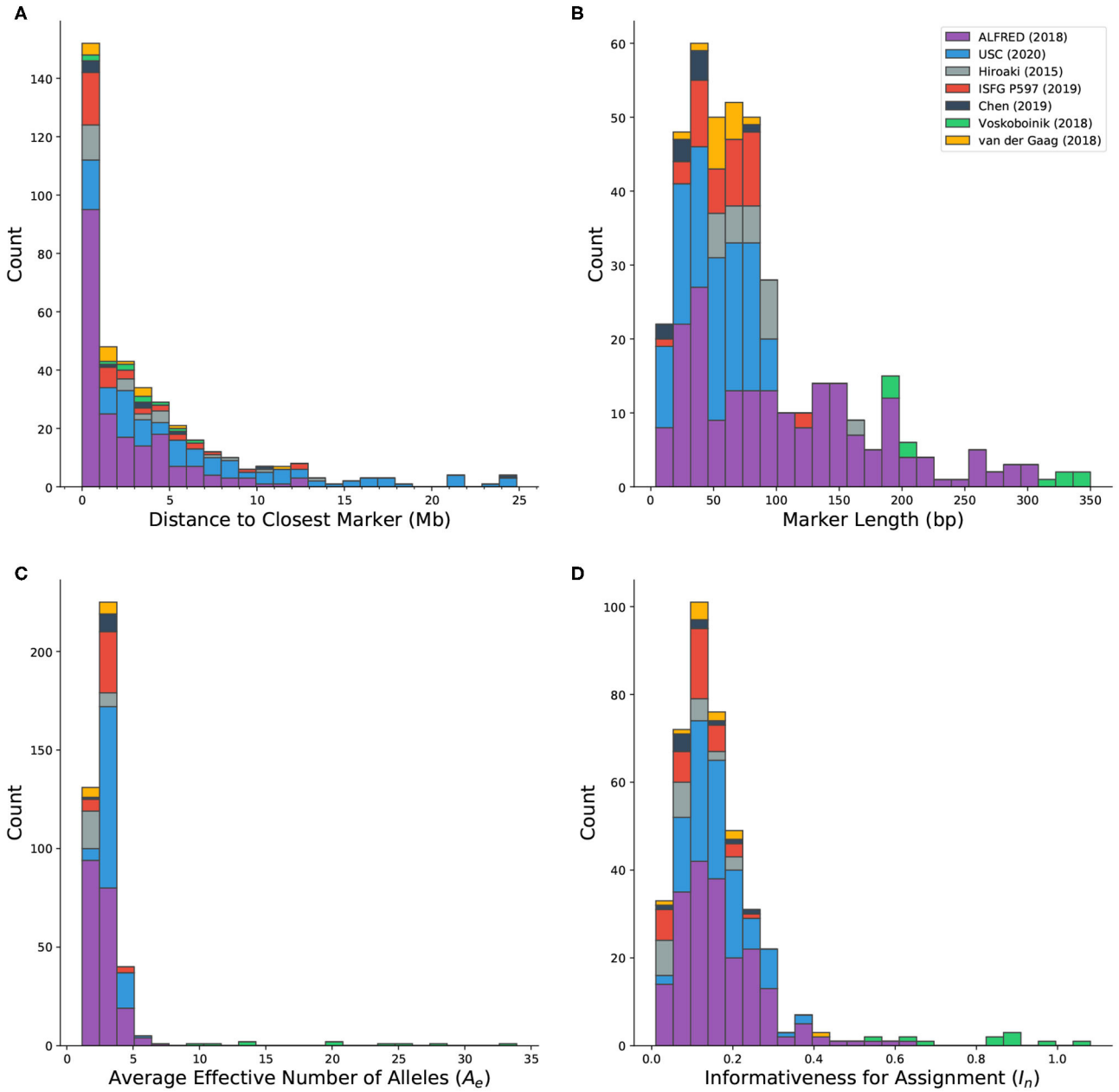


FIGURE 3 | Histograms showing the distribution of various characteristics of microhaplotype markers in MicroHapDB, shaded by publication source: **(A)** distance between a marker and its closest non-overlapping marker (outliers not shown: 42.8 and 56.1 Mb); **(B)** marker length; **(C)** *effective number of alleles* (A_e), averaged over 26 global populations; **(D)** *informativeness for assignment* (I_n) to 26 global populations.

records. This database represents a comprehensive collection of all microhaplotype (microhap) data published to date.

The number of single nucleotide polymorphisms (SNPs) used to define microhaps ranges from 2 to 49, with an average of 3.77 SNPs per marker. MicroHapDB includes 115 markers defined by two SNPs, 171 defined by three SNPs, 87 defined by four SNPs, 20 defined by five SNPs, 5 defined by six SNPs, and 19 defined by seven or more SNPs.

Microhap markers are defined on all autosomes as well as the X chromosome. Forty-five marker definitions overlap with other markers. Most of these (40/45) were defined by Staadig and Tillmar (2019), which includes 11 exact duplicates and 29 probable adjustments to markers from other sources. The distance between each marker and the closest non-overlapping marker ranges between 143 bp and more than 56 Mb. Out of 417 markers in MicroHapDB, 152 (36.5%) reside within 1 Mb of their closest neighbor, and 53 (12.7%) reside within 100 kb of their closest neighbor (**Figure 3A**). Any set of markers separated by a small physical distance is likely in high linkage disequilibrium (LD) and the individual markers would therefore

not be independent. DNA profiles containing information for linked markers further complicates interpretation. This requires either sophisticated statistical modeling to account for the dependencies between markers or adopting an either/or strategy in which one of the loci is discarded when both produce reliable data.

Published microhaps occupy a wide range of lengths, with core marker length (the number of nucleotides spanning the most distal SNPs that define the marker, inclusive) ranging from 4 to 350 bp (**Figure 3B**). The majority of the microhaps in MicroHapDB (307/417; 73.6%) span <100 bp, and a substantial minority (145/417; 34.8%) span <50 bp.

3.2. Microhaplotype Frequencies for 26 Global Populations Enable Interpretation of Multi-Source Panels

Interpretation of any microhap typing result requires the use of appropriate microhaplotype frequency data. Prior to the release of MicroHapDB, availability of frequency data was inconsistent for published microhaps, with some sources providing frequencies for numerous population samples, while other sources providing frequencies for only a single population sample, or no frequency data at all.

MicroHapDB provides a comprehensive set of frequencies for all microhap markers to date. Estimated using 2,504 fully phased genotypes from Phase 3 of the 1,000 Genomes Project (Auton et al., 2015), the MicroHapDB database contains frequencies for 412 microhaplotype markers across 26 global population samples. A total of 113,477 frequencies for 5,373 alleles furnish a broad view of marker variation amongst Africans, admixed Americans, East Asians, Europeans, and South Asians.

3.3. MicroHapDB Provides Three Measures of Allelic Variation Within and Among Populations

The availability of microhaplotype frequency estimates across a standard set of 26 global population samples for all published microhaps provides a consistent means of comparing, ranking, and evaluating microhap markers for different applications. The per-marker effective number of alleles (A_e) was computed independently for each population, and then averaged across all 26 populations (section Materials and Methods). This statistic serves as a measure of within-population allelic diversity observed at a particular marker, and corresponds to the marker's power for individual identification. A previous study (Kidd and Speed, 2015) proposed an A_e threshold of 3, above which a microhaplotype can be considered "exceedingly useful" for forensic purposes. Average A_e values for microhaps in MicroHapDB range from 1.16 to 33.92, with a mean of 3.28 (**Figure 3C**). Most markers in MicroHapDB (238/412, 57.8%) have an average A_e below 3, and only 17 markers (4.1%) have an average A_e above 5.

Marker informativeness for assignment (I_n) was computed for the same 26 global populations (section Materials and Methods). This statistic serves as a measure of variation among populations, and corresponds to the marker's power for predicting an individual's biogeographic ancestry. I_n values for markers in MicroHapDB fall between 0.01 and 1.08, with a mean of 0.17 (**Figure 3D**). Eight markers have an I_n value >0.682, the highest I_n value previously reported for a microhap (Kidd et al., 2018)—we note however that these I_n values were computed for a different set of populations and are therefore not directly comparable.

Figure 4 shows allele frequency distributions across the 26 populations for six representative microhap markers. The markers were selected from a range of A_e and I_n values to demonstrate how differences in these statistics are reflected in allele frequencies.

F_{ST} values were also computed for all markers. **Figure 5** shows the correlation between F_{ST}, A_e, and I_n for 391 markers. A weak positive correlation exists between A_e and I_n, while a weak negative correlation exists between A_e and F_{ST}. The latter trend suggests that while it's *possible* for an increase in allelic diversity to coincide with population-specific patterns of allele frequency, this is not generally the case for the microhap markers here considered. The strongest correlation is between I_n and F_{ST}, reflecting the sensitivity of these two statistics to population-specific allele distributions.

Ten highly polymorphic microhaps reported by Voskoboinik et al. (2018) stand out from microhaps published in other sources in several ways. Originally defined by locus boundaries rather than by explicit lists of SNP identifiers, these 10 microhaps include more SNPs (14–49 per marker) than any other marker in MicroHapDB. They have the highest average A_e values in MicroHapDB by a significant margin, and nine of these markers are included in MicroHapDB's top 10 microhaps ranked by I_n. It is worth noting that these microhaps are also among the longest, reflecting the study's distinct selection criteria and sequencing and evaluation strategy. The longest five markers in this set are also the five longest markers in MicroHapDB, and the remaining five are above the 89th percentile in length with respect to all markers in MicroHapDB.

4. DISCUSSION

In this study, we report the development of a comprehensive database of published microhaplotype (microhap) marker and frequency data. We describe the estimation of microhap frequencies in 26 global population samples, and the use of these frequencies to compute measures of allelic variation, enabling the ranking of microhaps for different forensic applications. This extensive collection of allele frequencies and ranking statistics will facilitate the design and interpretation of forensic panels that include markers from distinct sources without the need for extensive development of frequency data up front. MicroHapDB is a free open-access resource that contains information for all microhaps published as of February 2020. It requires minimal computing resources to install and maintain, and is designed to be easily extended with additional sources of public or private microhap data in the future. We hope MicroHapDB will democratize and accelerate advances in microhap-based forensics capabilities by enabling

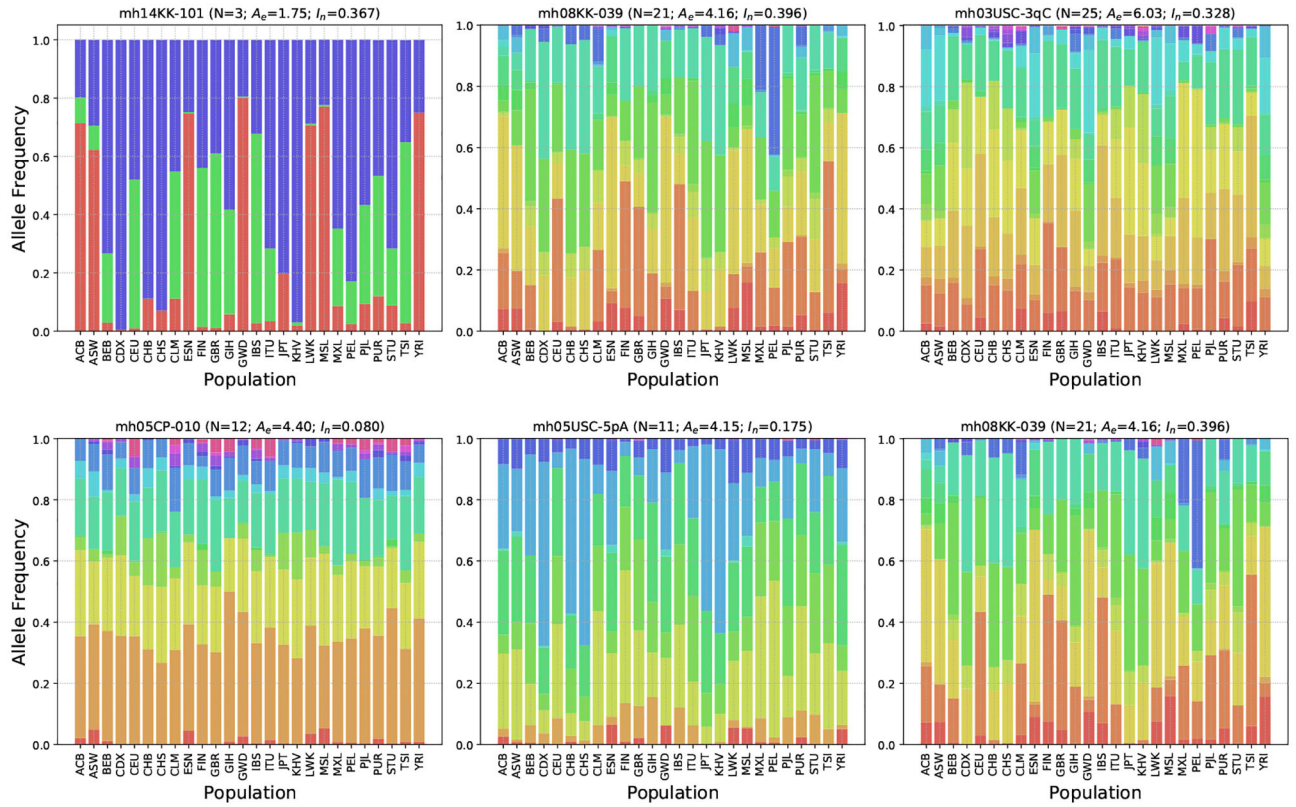


FIGURE 4 | Allele frequency distributions for six representative microhaplotype markers. Shown above each plot is the marker's name, the number of observed haplotypes (*N*), the average *effective number of alleles* (A_e), and the *informativeness for assignment* (I_n). Different colors correspond to distinct haplotypes. Holding I_n roughly constant, the top row shows relative allele frequencies for three microhap markers with low, medium and high A_e values. Holding A_e roughly constant, the bottom row shows relative allele frequencies for three microhap markers with low, medium, and high I_n values.

researchers and companies to focus solely on marker discovery, or solely on population surveys, or solely on panel and kit design, without the need to invest in development of all of the above.

MicroHapDB provides A_e and I_n values for ranking microhaps for different forensic purposes. The genomic coordinates of each marker are also stored in MicroHapDB, enabling convenient calculation of physical distances between markers residing on the same chromosome. However, in addition to normal considerations that must always be addressed when designing a forensic DNA panel (e.g., amplicon sizes, primer kinetics, off-target amplification), correct interpretation of DNA profiles requires researchers to determine the independence of markers in a proposed panel based on the extent of linkage between the markers in the population(s) of interest. The length of candidate markers is also an important consideration depending on the sequencing technology utilized and the priority of recovering profiles from low input or low quality DNA samples.

Unlike conventional short tandem repeat (STR) markers, no comprehensive database analogous to the FBI's Combined DNA Index System (CODIS) databases (https://www.fbi.gov/services/laboratory/biometric-analysis/codis) yet exists for microhap profiles. Constructing, populating, and performing requisite quality control for such a database will require substantial time and investment, which likely could only be pursued in earnest as the forensic community approaches consensus regarding optimal targets and assays. We expect that MicroHapDB can play a useful role in that process.

A major challenge in establishing a database like MicroHapDB, or a shared national database of microhap profiles, or indeed in communicating clearly about microhaps in the scientific literature, is the lack of consistency in nomenclature and in the way that markers are defined. A few papers describing microhap markers have used the nomenclature proposed by Kidd (2016), with marker names such as `mh01KK-172`, `mh01CP-007`, and `mh06PK-25713`. Other papers used a variety of *ad hoc* marker designators, such as `1` and `MH02`. MicroHapDB has adopted the Kidd nomenclature since its inception in 2018, and for sake of consistency has applied it to microhap collections where it was not previously used (e.g., `mh01NH-01` for `1` and `mh01AT-02` for `MH02`).

The question of how microhap markers are *defined* is at least as consequential as how they are *labeled*. A small number of published microhaps have been defined as a specific (but undisclosed) set of single nucleotide polymorphisms (SNPs)

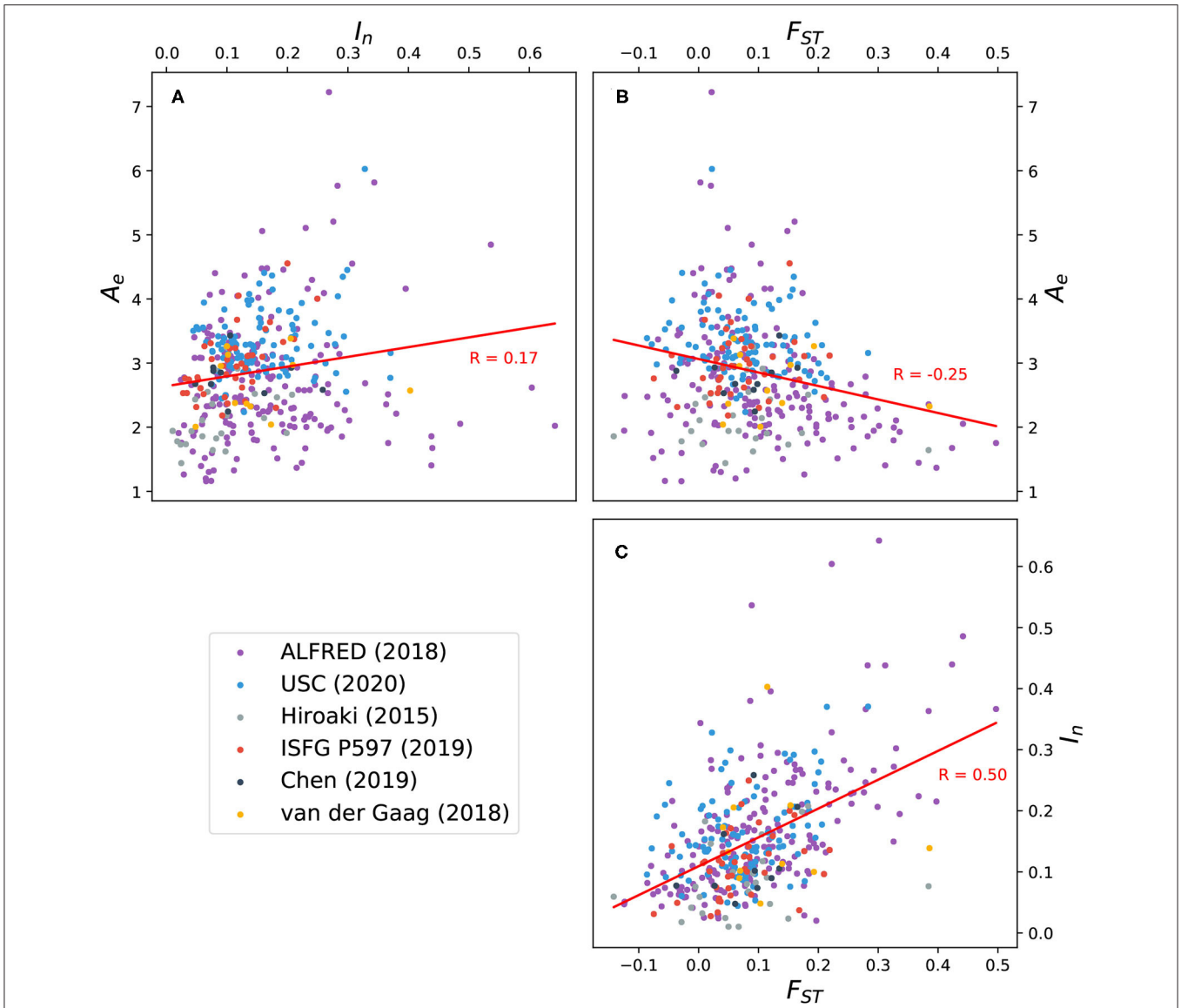


FIGURE 5 | Scatter plots and trend lines showing pairwise correlation between three measures of allelic variation for 391 microhaplotypes: *effective number of alleles* (A_e) averaged over 26 global populations; *informativeness for assignment* (I_n); and *fixation index* (F_{ST}). **(A)** Correlation between A_e and I_n; **(B)** correlation between A_e and F_{ST}; **(C)** correlation between F_{ST} and I_n. Ten microhaps described by Voskoboinik et al. (2018) exhibit uncharacteristically high levels of polymorphism (see **Figure 3**) and were excluded from the analysis. Eleven microhaps defined on the X chromosome were also excluded.

at a genomic locus, the endpoints of which are indicated using coordinates on a reference genome assembly. Other microhaps are defined by a designator that refers to a set of SNPs at a particular locus, but whose specific component SNPs have been adjusted over time to improve the marker's performance. Ambiguous marker definitions of these kinds create substantial challenges for reproducibility and establishing provenance, and should be avoided. de la Puente et al. (2020) propose that microhaps should forgo marker names altogether (e.g., `mh01USC-1pA` or `1pA`) in favor of an explicit list of SNP variants, as designated by dbSNP rsIDs (e.g., `rs28503881,rs4648788,rs72634811,rs28689700`).

We strongly endorse the sentiment behind this recommendation, although we concede the convenience of concise marker designators, especially when marker definitions are composed of dozens of SNP variants. What is most critical is the need for marker definitions to be *unambiguous* and *invariant over time*.

This discussion highlights the tension that has been provoked by the emergence of NGS technologies in forensics. Conventional assays have required the design of probes for specific SNP targets, which are often genotyped independently and then phased statistically. In contrast, NGS assays permit recovery of the entire sequence at a microhap locus and the simultaneous genotyping

and phasing of all its component SNPs, and indeed any additional intermediate (and often rare) SNPs. We anticipate that as NGS forensic assays become more routine, typing results for microhap assays will include all of the variants occurring in the sequenced genomic segment. This kind of typing result would have full "backwards compatibility" in that it could be used to determine the haplotype of *any* microhap marker explicitly defined at the locus. At the same time, full-coverage sequences of microhap loci will enable significant improvements in, e.g., mixture detection and deconvolution.

The question remains as to whether microhap designators should refer to *markers* (i.e., explicit sets of variants) or to *loci*. We suggest that minor addenda to the nomenclature proposed by Kidd (2016), such as the use of version numbers or other suffixes, would enable support for both. In the mean time, MicroHapDB searches based on genomic coordinates provide a convenient way to resolve spatial relationships between distinct marker definitions.

DATA AVAILABILITY STATEMENT

MicroHapDB version 0.6 has been archived in the Open Science Framework repository and is available at https://osf.io/gr7h6. The MicroHapDB database and software can be installed and updated from the Bioconda repository: https://bioconda.github.io/ for more details. The publicly available datasets analyzed in this study can be found here: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/.

AUTHOR CONTRIBUTIONS

DS conceived the study, constructed the database, implemented the software interface, and wrote the manuscript. DS and RM collected data, performed quality control, and edited and approved the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was funded under Agreement No. HSHQDC-15-C-00064 awarded to Battelle National Biodefense Institute by the Department of Homeland Security Science and Technology Directorate (DHS S&T) for the management and operation of the National Biodefense Analysis and Countermeasures Center, a Federally Funded Research and Development Center. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security or the U.S. Government. The Department of Homeland Security does not endorse any products or commercial services mentioned in this presentation. In no event shall the DHS, BNBI, or NBACC have any responsibility or liability for any use, misuse, inability to use, or reliance upon the information contained herein. In addition, no warranty of fitness for a particular purpose, merchantability, accuracy, or adequacy is provided regarding the contents of this document.

ACKNOWLEDGMENTS

We want to thank three reviewers whose thoughtful feedback improved this article. We also express gratitude to Kenneth Kidd, Christopher Phillips, and Lev Voskoboinik for responding to inquiries about published data sets, and for insightful discussions about microhaplotypes. Finally, we thank Rebecca Just, Tim Stockwell, M. J. Rosovitz, and Adam Bazinet for their valuable feedback on drafts of this manuscript and early versions of the database.

REFERENCES

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

Bennett, L., Oldoni, F., Long, K., Cisana, S., Madella, K., Wootton, S., et al. (2019). Mixture deconvolution by massively parallel sequencing of microhaplotypes. *Int. J. Legal Med.* 133, 719–729. doi: 10.1007/s00414-019-02010-7

Bleka, Ø., Benschop, C. C., Storvik, G., and Gill, P. (2016a). A comparative study of qualitative and quantitative models used to interpret complex STR DNA profiles. *Forens. Sci. Int.* 25, 85–96. doi: 10.1016/j.fsigen.2016.07.016

Bleka, Ø., Storvik, G., and Gill, P. (2016b). EuroForMix: An open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts. *Forens. Sci. Int.* 21, 35–44. doi: 10.1016/j.fsigen.2015.11.008

Butler, J. M. (2010). *Fundamentals of Forensic DNA Typing*. San Diego, CA: Academic Press.

Butler, J. M. (2015). *Advanced Topics in Forensic DNA Typing: Interpretation*. Oxford: Academic Press.

Chen, P., Deng, C., Li, Z., Pu, Y., Yang, J., Yu, Y., Li, K., et al. (2019a). A microhaplotypes panel for massively parallel sequencing analysis of DNA mixtures. *Forens. Sci. Int.* 40, 140–149. doi: 10.1016/j.fsigen.2019.02.018

Chen, P., Zhu, W., Tong, F., Pu, Y., Yu, Y., Huang, S., et al. (2019b). Identifying novel microhaplotypes for ancestry inference. *Int. J. Legal Med.* 133, 983–988. doi: 10.1007/s00414-018-1881-x

Coble, M. D., and Bright, J.-A. (2019). Probabilistic genotyping software: an overview. *Forens. Sci. Int.* 38, 219–224. doi: 10.1016/j.fsigen.2018.11.009

Cowell, R. G., Graversen, T., Lauritzen, S. L., and Mortera, J. (2015). Analysis of forensic DNA mixtures with artefacts. *J. R. Stat. Soc.* 64, 1–48. doi: 10.1111/rssc.12071

Crawford, N. G., Kelly, D. E., Hansen, M. E. B., Beltrame, M. H., Fan, S., Bowman, S. L., et al. (2017). Loci associated with skin pigmentation identified in African populations. *Science* 358:6365. doi: 10.1126/science.aan8433

Crow, J. F., and Kimura, M. (1970). *An Introduction to Population Genetics Theory*. New York, NY: Harper & Row.

de la Puente, M., Phillips, C., Xavier, C., Amigo, J., Carracedo, A., Parson, W., et al. (2020). Building a custom large-scale panel of novel microhaplotypes for forensic identification using MiSeq and Ion S5 massively parallel sequencing systems. *Forens. Sci. Int.* 45. doi: 10.1016/j.fsigen.2019.102213

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., et al. (2018). Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat. Methods* 15, 475–476. doi: 10.1038/s41592-018-0046-7

Hiroaki, N., Koji, F., Tetsushi, K., Kazumasa, S., Hiroaki, N., and Kazuyuki, S. (2015). Approaches for identifying multiple-SNP haplotype blocks for use in human identification. *Legal Med.* 17, 415–420. doi: 10.1016/j.legalmed.2015.06.003

Kidd, K. K. (2016). Proposed nomenclature for microhaplotypes. *Hum. Genomics* 10:16. doi: 10.1186/s40246-016-0078-y

Kidd, K. K., Pakstis, A. J., Speed, W. C., Lagace, R., Wootton, S., and Chang, J. (2018). Selecting microhaplotypes optimized for different purposes. *Electrophoresis* 39, 2815–2823. doi: 10.1002/elps.201800092

Kidd, K. K., and Speed, W. C. (2015). Criteria for selecting microhaplotypes: mixture detection and deconvolution. *Invest. Genet.* 6:1. doi: 10.1186/s13323-014-0018-3

McKinney, W. (2011). pandas: a foundational python library for data analysis and statistics. Python for High Performance and Scientific Computing, 14.

Miles, A., Ralph, P., Rae, S., and Pisupati, R. (2019). *scikit-allel v1.2.1: A Python Package for Exploring and Analysing Genetic Variation Data*.

Oldoni, F., Hart, R., Long, K., Maddela, K., Cisana, S., Schanfield, M., et al. (2017). Microhaplotypes for ancestry prediction. *Forens. Sci. Int*. 6, e513–e515. doi: 10.1016/j.fsigss.2017.09.209

Oldoni, F., Kidd, K. K., and Podini, D. (2019). Microhaplotypes in forensic genetics. *Forens. Sci. Int*. 38, 54–69. doi: 10.1016/j.fsigen.2018.09.009

Rosenberg, N. A., Li, L. M., Ward, R., and Pritchard, J. K. (2003). Informativeness of genetic markers for inference of ancestry. *Am. J. Hum. Genet*. 73, 1402–1422. doi: 10.1086/380416

Ruiz, Y., Phillips, C., Gomez-Tato, A., Alvarez-Dios, J., de Cal, M. C., Cruz, R., et al. (2013). Further development of forensic eye color predictive tests. *Forens. Sci. Int*. 7, 28–40. doi: 10.1016/j.fsigen.2012.05.009

Staadig, A., and Tillmar, A. (2019). "Evaluation of microhaplotypes–a promising new type of forensic marker," in *The 28th Congress of the International Society for Forensic Genetics* (Prague), P 597.

van der Gaag, K. J., de Leeuw, R. H., Laros, J. F., den Dunnen, J. T., and de Knijff, P. (2018). Short hypervariable microhaplotypes: a novel set of very short high discriminating power loci without stutter artefacts. *Forens. Sci. Int*. 35, 169–175. doi: 10.1016/j.fsigen.2018.05.008

Voskoboinik, L., Motro, U., and Darvasi, A. (2018). Facilitating complex DNA mixture interpretation by sequencing highly polymorphic haplotypes. *Forens. Sci. Int*. 35, 136–140. doi: 10.1016/j.fsigen.2018.05.001

Walsh, S., Liu, F., Wollstein, A., Kovatsi, L., Ralf, A., Kosiniak-Kamysz, A., et al. (2013). The hirisplex system for simultaneous prediction of hair and eye colour from DNA. *Forens. Sci. Int*. 7, 98–115. doi: 10.1016/j.fsigen.2012.07.005

Weir, B. S., and Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution* 38, 1358–1370. doi: 10.1111/j.1558-5646.1984.tb05657.x

Zhu, J., Lv, M., Zhou, N., Chen, D., Jiang, Y., Wang, L., et al. (2019). Genotyping polymorphic microhaplotype markers through the Illumina® MiSeq platform for forensics. *Forens. Sci. Int*. 39, 1–7. doi: 10.1016/j.fsigen.2018.11.005

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.