Title

A computational approach for positive genetic identification and relatedness detection from low-coverage shotgun sequencing data

Journal

Journal of Heredity, 114(5)

ISSN

0022-1503

Authors

Nguyen, Remy
Kapp, Joshua D
Sacco, Samuel
et al.

Publication Date

2023-08-23

DOI

10.1093/jhered/esad041

Journal of Heredity, 2023, **114**, 504–512
https://doi.org/10.1093/jhered/esad041
Advance access publication 29 June 2023
Original Article

American Genetic Association

OXFORD

Original Article

A computational approach for positive genetic identification and relatedness detection from low-coverage shotgun sequencing data

Remy Nguyen[1], Joshua D. Kapp[2], Samuel Sacco[2], Steven P. Myers[3] and Richard E. Green[1]

[1]Department of Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, CA, United States,
[2]Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, Santa Cruz, CA, United States,
[3]California Department of Justice Jan Bashinski DNA Laboratory, Richmond, CA, United States

Address correspondence to R.E. Green at the address above, or e-mail: ed@soe.ucsc.edu.

Corresponding Editor: Bridgett vonHoldt

Abstract

Several methods exist for detecting genetic relatedness or identity by comparing DNA information. These methods generally require genotype calls, either single-nucleotide polymorphisms or short tandem repeats, at the sites used for comparison. For some DNA samples, like those obtained from bone fragments or single rootless hairs, there is often not enough DNA present to generate genotype calls that are accurate and complete enough for these comparisons. Here, we describe IBDGem, a fast and robust computational procedure for detecting genomic regions of identity-by-descent by comparing low-coverage shotgun sequence data against genotype calls from a known query individual. At less than 1× genome coverage, IBDGem reliably detects segments of relatedness and can make high-confidence identity detections with as little as 0.01× genome coverage.

Key words: DNA identification, genome, identity-by-descent

Introduction

DNA-based identification in forensics is typically accomplished via genotyping allele length at a defined set of short tandem repeat (STR) loci via PCR (Kimpton et al. 1993). These PCR assays are robust, reliable, inexpensive (Jobling and Gill 2004), and amenable to samples with microbial contamination. Given the multiallelic nature of these loci, a small panel of STR markers can provide suitable discriminatory power for personal identification (Gill et al. 1985; Jeffreys et al. 1985). Since the markers in STR panels have little or no mutual information, i.e. linkage disequilibrium, between them, they provide independent information. This simplifies match probability calculation for DNA-based identity.

Massively parallel sequencing (MPS) technologies and genotype array technologies invite new approaches for DNA-based identification. Application of these technologies has provided catalogs of global human genetic variation at single-nucleotide polymorphic (SNP) sites and short insertion–deletion (INDEL) sites. For example, from the 1000 Genomes Project (Genomes Project et al. 2015), we now have a catalog of nearly all human SNP and INDEL variation down to 1% worldwide frequency. Large-scale population sequencing projects that will generate catalogs of segregating variation are also underway for many other species (Shaffer et al. 2022).

Genotype files, generated via MPS or genotype arrays, can be compared between individuals to find regions that are co-inherited or identical-by-descent (IBD) (Gusev et al. 2009; Browning and Browning 2013a, 2013b; Kling and Tillmar 2019; Kling et al. 2021). Finding IBD regions between 2 samples implies that the samples derived from individuals who are genetically related. These comparisons are the basis of the relative finder functions in many direct-to-consumer genetic testing products (Durand et al. 2014; Ball et al. 2016).

Relatedness estimation from genotype data can be performed in 2 general ways. In 1 style of approach, dense, genome-wide marker data are explicitly handled as *not* independent between nearby sites. Rather, these methods attempt to find genomic regions wherein markers indicate that at least one of each sample's chromosomes are IBD (Purcell et al. 2007; Gusev et al. 2009; Browning and Browning 2010). In this way, it is the aggregate signal from many linked markers in a region that signify IBD. The second category compares genotype data to measure an overall rate of genetic similarity (Conomos et al. 2016; Gorden et al. 2022). An elevated rate implies relatedness and can then be used to estimate a kinship

Received March 8, 2023; Accepted June 28, 2023

coefficient or degree of relatedness. This second approach does not attempt to identify which genomic regions are IBD. A special case of relative-finding is self-identification. This is a trivial comparison of genotype files as self-comparisons will be identical across all sites, minus the error rate of the assay.

For many forensic samples, however, the available DNA may not be suitable for PCR-based STR amplification (Alaeddini et al. 2010), genotype array analysis (de Vries et al. 2022), or MPS to the depth required for comprehensive, accurate genotype calling (Nielsen et al. 2011). In the case of PCR, one of the most common failure modes occurs when DNA is too fragmented for amplification. For these samples, it may be possible to directly observe the degree of DNA fragmentation from the decreased amplification efficiency of larger STR amplicons from a multiplex STR amplification (Swango et al. 2006). In the case of severely fragmented samples, where all DNA fragments are shorter than the shortest STR amplicon length, PCR simply fails with no product.

Here, we present a fast and straightforward computational approach for comparison of genotype data from one or more known individuals to limited amounts of DNA sequence data from an unknown sample. This approach, called IBDGem, does not attempt to call genotypes from the sequence data. Rather, IBDGem evaluates the likelihood of observing the sequence data if a test individual, whose genotype is known, was the source versus the likelihood of observing that same data if an unrelated individual was the source. We find that this approach can reliably identify samples with as little as 0.01× depth of coverage from the questioned sample. Consequently, IBDGem enables forensic identity using samples such as bone and single rootless hairs that typically yield sub-nanogram quantities of fragmented DNA and otherwise may not be amenable to DNA-based forensic analysis (Turner et al. 2022).

Material studied, methods, techniques

Data presented here are from: 1) The 1000 Genomes Project Phase 3 deep sequencing (Byrska-Bishop et al. 2022) and 2) a panel of 8 human volunteers from whom we derived DNA from a saliva sample and cut hairs (hair panel) under UCSC IRB protocol HS3382.

For each anonymous study participant, we collected saliva DNA using the OGR-500 collection device, head hair, and pubic hair. We extracted DNA from the saliva and submitted 1 µg to AKESOgen for genotype array processing using the Illumina Multi-Ethnic Global Array (MEGA). For each participant, we extracted DNA from 5 head and 3 pubic hairs and prepared single-stranded DNA Illumina sequencing libraries (Kapp et al. 2021) from the 2 highest concentration head and pubic hair extractions. We performed shotgun DNA sequencing of the libraries prepared from the hair extractions on an Illumina NovaSeq 6000 at UCSF. See the supplement for a detailed description of the wet lab methods.

IBDGem is a computer program implemented in C that compares genotype data (generated via genotype array or DNA sequence data) from a known individual to aligned sequence data from an unknown individual. For each variable site, it calculates the likelihood of the observed sequence data under 3 models of relatedness: 1) the compared samples share 2 chromosomes identical-by-descent, IBD2, 2) the compared samples share 1 chromosome identical-by-descent, IBD1, or 3) the compared samples share no chromosomes identical-by-descent, IBD0. Note that these 3 relationships are the only possible ways that 2 samples can be related to one another at a particular region in the genome. For the analyses presented here, the variable sites are all biallelic SNP sites from the 1000 Genomes panel or all biallelic sites from the Illumina Multi-Ethnic Global Array.

The likelihoods of the data under these 3 models can then be compared with find which best explains the data or to generate a log-likelihood ratio (LLR) between models. If the distances between variable sites are sufficiently large, i.e. longer than the length of a sequence read, then the observed alleles at each site can be treated as independent observations of the likelihood of each IBD state across a genomic region. Thus, we can aggregate these likelihoods across multiple sites to increase the discriminatory power between any 2 models. Additionally, to account for linkage disequilibrium among alleles in the calculation of the background model (IBD0), IBDGem uses genotypes from a panel of reference individuals. We calculate the likelihood of the data against each of these unrelated individuals and take the average to be likelihood of the data under the IBD0 model (Fig. 1).

In the special case of determining whether the sequence data derives from the same individual as the genotype data versus the model of it coming from an unrelated individual, which is analogous to the hypotheses tested in STR identification, we simply generate LLRs between the IBD2 and IBD0 models. Note that in the case that an individual in the reference panel has cryptic relatedness to the subject individual, the IBD0 model will be inflated, reducing the LLR (IBD2/IBD0). In this case, the genetic identity test is conservative. The calculations of likelihoods are described in detail in the Supplementary Note: IBDGem algorithm.

IBDGem software is available for noncommercial use via github: https://github.com/Paleogenomics/IBDGem

Results

IBDGem analyzes regions of the genome for which there is genotype data from a known sample and some amount of sequence data from an unknown sample. It implements 2 procedures. The first is a test of whether the sequence data (from a forensic sample, e.g.) is more likely if it is from an individual who is genetically identical to the known sample or if it is from an unrelated person. This result is expressed as an LLR. The second procedure is used to identify segments of relatedness, if any, between the samples. This second procedure calculates the likelihood of the 2 samples if related by 0, 1, or 2 shared chromosomes (IBD) regionally across the genome. Because humans are diploids, we carry 2 copies of each autosomal genomic locus. Thus, these 3 models (IBD0, IBD1, and IBD2) are the *only* ways that 2 individuals can be related at a particular autosomal genomic region. More closely related individuals have more IBD1 regions (genome segments inherited from common ancestors) than less closely related individuals.

Comparisons between unrelated individuals will be IBD0, i.e. not share either chromosomal region from a recent common ancestor, across all or nearly all regions of the genome. Conversely, comparisons between the same person will necessarily be IBD2 across every region of the genome. For

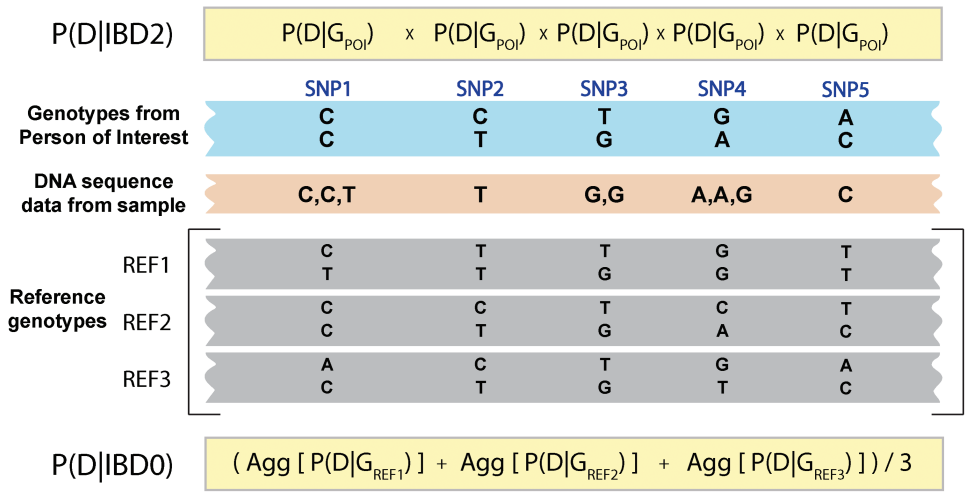

Fig. 1. IBDGem schematic. Comparisons are made between the known genotype of a subject person (Person of Interest; second band from top) and low-coverage sequence data from a DNA sample (third band from top). The probability of the observed data can be calculated under the model that the subject person carries the same 2 chromosomes as the person from whom the DNA sample is collected, i.e. is IBD2 (top). The probability of the observed data under the model that the subject person is genetically unrelated (IBD0; bottom) is the average aggregated likelihood that the unknown sample might have originated from a reference panel of individuals..

closely related individuals, some regions will be IBD1, where a segment of a chromosome is co-inherited from a recent common ancestor. Parent–offspring relatives are IBD1 across all chromosomes. Full siblings are roughly 25% IBD0, 50% IBD1, and 25% IBD2.

To test the ability of IBDGem to reliably compare samples, we first used data from the high-coverage 1000 Genomes panel (Byrska-Bishop et al. 2022). This panel provides both genotype calls and aligned DNA sequence data for each individual. In this analysis, self-versus-self comparisons represent positive controls wherein all segments of all chromosomes should be identifiable as IBD2. Further, self-versus-non-self comparisons represent negative controls wherein all segments of all chromosomes should be identifiable as IBD0, except in cases of cryptic relatedness.

We first analyzed the genotype and sequence data in the GBR (British) and LWK (Luhya) panels from the 1000 Genomes. We used all available genotypes at biallelic SNP sites for these comparisons. In this way, this experiment approximates the situation of having high-coverage DNA from 1 comparison individual with which to generate full genotype information for the known sample. After excluding known relatives from the panels, there are 91 and 94 samples in the GBR and LWK panels, respectively. We performed 1 self and 1 non-self comparison for each sample from each panel. In other words, we compared the genotype of each individual against either their own aligned sequence data (self comparison) or the sequence data of a different, random individual within the same panel (non-self comparison) that we down-sampled to specific target depths to simulate actual data from lower-coverage sequencing. Specifically, we down-sampled the sequence reads so that the final coverage is approximately 2×, 1×, 0.5×, 0.1×, and 0.01×. The SNP sites used for analysis are those where data are available after this down-sampling step, and as such the number of sites varies for each pairwise comparison.

We aggregated the likelihood results into nonoverlapping regions/bins containing 200 SNPs across the genome. Within each region, the LLRs between models IBD2 and IBD0 for self comparisons were strongly identifiable from non-self comparisons (Supplementary Figs. S1 and S2). Overall, this experiment demonstrates that IBDGem can make accurate genetic identifications for all individuals in the GBR and LWK panels even at the ultra-low coverage of 0.01× (Fig. 2, Supplementary Figs. S3 and S4, and Supplementary Table S1). Because the likelihood calculations are straightforward and use a fixed panel of genotypes, IBDGem executes quickly, taking about 90 s per chromosome on a 2 socket Intel Xeon Silver 4216 CPU server.

The LLR for self comparisons (positive controls) showed an interesting behavior at the lowest coverages, particularly at 0.01× and 0.1×. The mean LLR (IBD2/IBD0) were *higher* at these lower coverages than at higher coverages. We speculated that this may be due to the fact that the genomic regions spanned within each 200 SNP bin is necessarily longer at lower coverages since fewer variable sites will have sampled data. Wider regions might reduce IBD0 aggregate likelihood values since they will be less prone to fluctuations in the degree of cryptic or distant relatedness. Consistent with our prediction, at 0.01× the peak of the IBD0 likelihood distribution is more negative than that of the same distribution at 1× (Supplementary Fig. S5A). On the other hand, the IBD2 likelihood distribution at 1× is more negative than at 0.01×, suggesting that at this higher coverage, the higher number of observations per site leads to a necessary decrease in the likelihood of IBD2. We note that as the coverage of the sequence data increases beyond 1×, the LLRs increase again. Despite the effects, all positive and negative controls were correctly and strongly identified at all depths of sequencing coverage from 0.01× to 2×.

The probability model for IBD0 relies on comparisons to a reference panel of unrelated individuals to model the likelihood of the observed sequence data under the scenario that it derives from an unrelated individual. Human populations, in general, have low cross-population F_{st} values (Rosenberg et al. 2002). Thus, one might expect that the background population used to define IBD0 has little impact on the IBD0 calculation. However, population differences in haplotype

frequencies or other phenomena may cause mis-specification of the background population to impact the power to determine self versus non-self using IBDGem.

To test the sensitivity of IBDGem to the background reference panel, we reran the previous comparisons at 1× average genome coverage on individuals in the GBR and LWK panels. For this experiment, we did not use the most general model, i.e. using all unrelated individuals from 1000 Genomes as our reference panel. Instead, we used individuals from specific continental or subcontinental subsets (Fig. 3, Supplementary Figs. S6 and S7, and Supplementary Table S2). In each case, the comparisons correctly identified data from the same individual, regardless of the population used as the background, IBD0 model. For example, the genotypes of all GBR individuals were identifiable from 1× random genome coverage with LLR means of greater than 50 across genomic regions even when the background population was the SAS (South Asian) or AFR (African) superpopulations in the 1000 Genomes data. Non-self comparisons were similarly identifiable as such, despite using a population to which the individual does not belong to model the background population.

IBDGem with genotype array data

The 1000 Genomes project pipeline calls variants from shotgun sequencing data across the genome. Each individual has genotype calls at nearly all sites that are found to be variable in the panel. Therefore, for each IBDGem analysis, the number of sites available for comparison is limited chiefly by the data available from the questioned sample. High-coverage genomic data can be used to generate nearly complete call sets at all of the sites known to be variable within humans in, for example, the 1000 Genomes panel. Thus, the genotype call set will include tens of millions of sites, although any specific individual will be homozygous for the reference allele at most of these sites.

In contrast, commercially available genotype arrays provide highly accurate genotype calls at about 1 million sites of known variation—those on the array—but no information at other sites. Genotype arrays are an accurate and less expensive approach for generating genotype data. To test the sensitivity of IBDGem when limited to genotype array sites for the subject individual, we specified the program to perform comparisons on only biallelic sites found on the Illumina Global Screening Array (GSA). In both the GBR and LWK panels we found that for all self comparisons the IBD2/IBD0 LLRs remain higher than 100 and for all non-self comparisons, these ratios are typically less than –100 (Fig. 4). That is, IBDGem can compare data at only GSA array sites against 1× genome coverage DNA data and confidently discriminate self from non-self comparisons. Note that LLRs for self comparisons at GSA-only sites were also much higher than at all sites from 1000 Genomes for the same coverage (1×) (Figs. 2 and 4). This is due to the phenomenon described above: fewer sites result in wider genomic bins and a reduction in IBD0 likelihoods for self-comparisons (Supplementary Fig. S5B).

IBDGem comparison with data from rootless hairs

The sequencing libraries that generated the 1000 Genomes data were predominantly made from cell line derived, high-molecular-weight DNA. Thus, the data quality is superior to what is possible from many forensic samples. To test the power of IBDGem using data derived from a more realistic

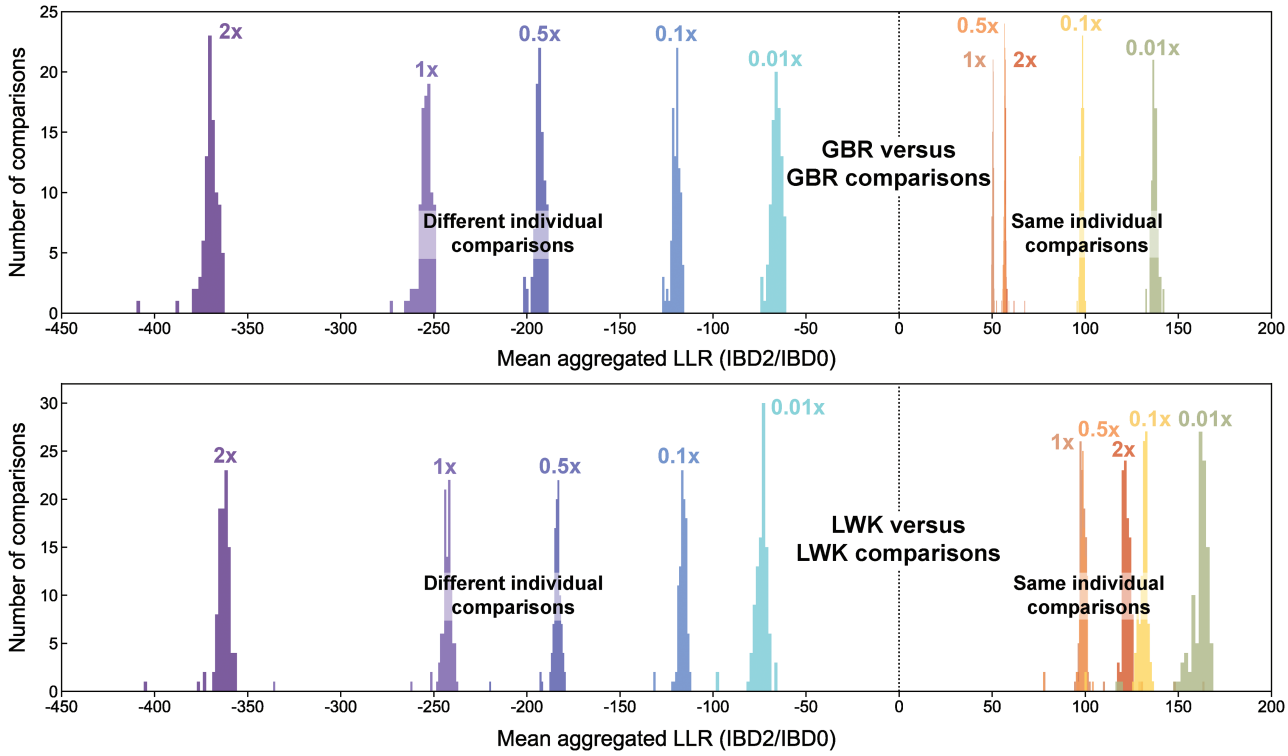

Fig. 2. IBDGem performance at various levels of genome sequence coverage. LLRs are aggregated across regions of 200 SNPs. A histogram of the means, across bins, is shown. Top panel: each individual from the GBR panel was compared against itself (same individual comparisons) or a random non-self GBR individual (different individual comparisons) following down-sampling of sequence data to 2×, 1×, 0.5×, 0.1×, and 0.01× genome coverages. Bottom panel: analogous comparisons among individuals in the LWK panel.

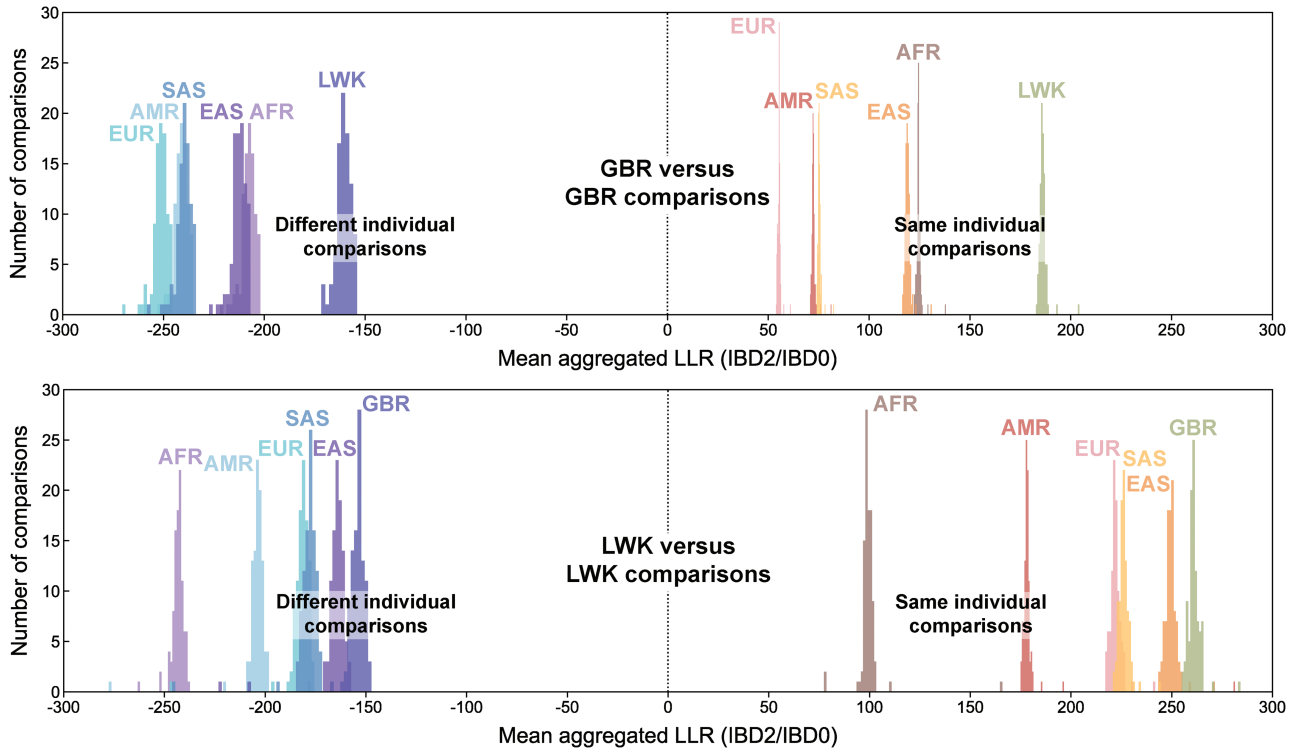

Fig. 3. IBDGem performance using various population background genotype frequency models. LLRs are aggregated across 200 SNPs. A histogram of the means, across bins, is shown. Top panel: each individual from the GBR panel was compared against itself (same individual comparisons) or a random non-self GBR individual (different individual comparisons) using samples from the indicated superpopulation as the background reference panel. AFR = African, AMR = American, EAS = East Asian, EUR = European, GBR = British, LWK = Luhya, SAS = South Asian. Bottom panel: analogous comparisons among individuals in the LWK panel.

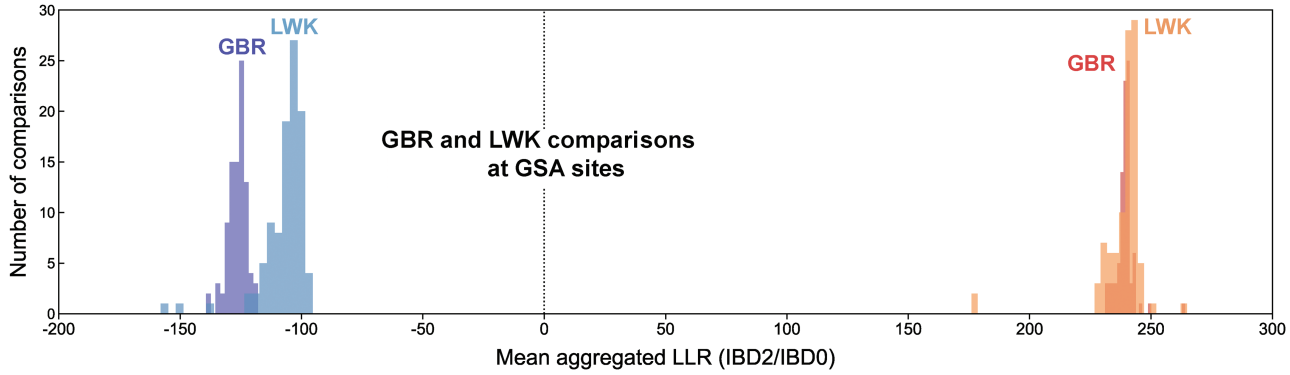

Fig. 4. IBDGem self and non-self comparisons of GBR and LWK individuals at GSA genotype array sites, with sequence data down-sampled to a depth of 1×.

forensic DNA source, we extracted and sequenced DNA from the rootless hairs of a panel of 8 individuals. Separately, we collected DNA from the saliva of these same 8 individuals for genotype analysis using the Illumina Multi-Ethnic Global array.

We collected multiple head hairs from each individual. We then extracted and isolated DNA from individual hairs and chose the highest and lowest DNA concentration extracts for sequencing. Note that for many of the hair extracts, the DNA concentration was below the level of detection with qubit fluorimetry. We used a single-stranded library preparation approach (Kapp et al. 2021) to generate Illumina sequencing libraries from 50% (20 µL) of each extract. We pooled these libraries and generated roughly 60 million read-pairs per library (Supplementary Table S3). See the supplement for a detailed description of the wet-lab methods.

After mapping these shotgun hair DNA sequence data to the reference human genome, we found that the amount of usable human DNA for each hair sample was variable (Fig. 5, top). This is likely due to the variability of the amount of DNA present per unit length of hair among people (Szabo et al. 2012).

After appropriate filtering of the hair sequence data (Supplementary Material), we ran IBDGem, comparing each hair DNA dataset to each saliva genotype dataset. For this comparison, we used the whole-panel 1000 Genomes individuals as our reference set for IBD0 since nothing about the donors was known and, as shown above, the method is

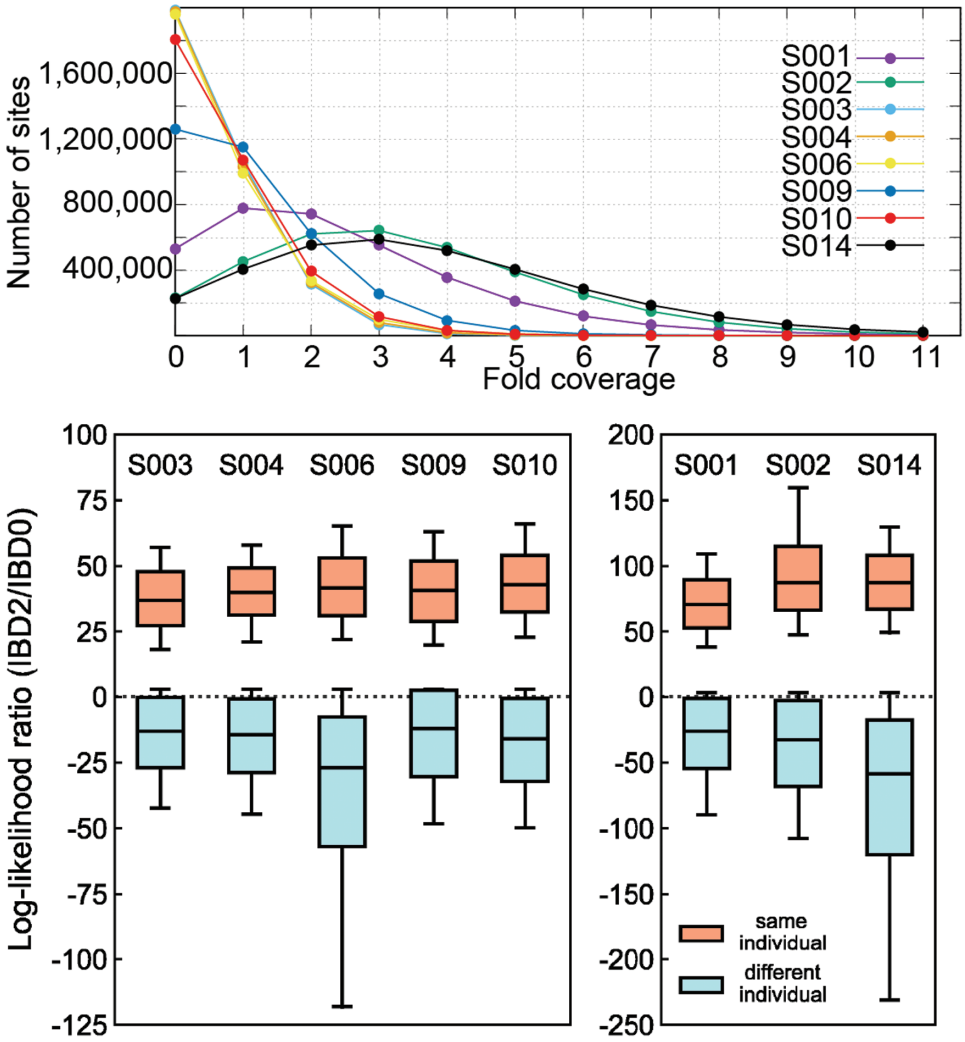

Fig. 5. IBDGem comparisons using DNA from hair. Top: sequence coverage distribution at known variable sites on chromosome 1 of hair samples. Illumina libraries were sequenced to similar depths. Variation in coverage represents the variability of DNA presence and recovery in human hairs. Bottom: IBDGem self (same individual) and non-self (different individual) comparisons using DNA data from hair and corresponding high-quality, saliva-derived genotype array data. LLRs are aggregated across 50 SNPs for which there was sample data. Plot shows distributions of LLRs within these 50-SNP regions. Left panel is lower-coverage samples (<1×). Right panel is higher-coverage samples (>1×).

largely insensitive to the use of a specific population background panel. All 8 self comparisons and all 56 non-self comparisons were correctly identified (Fig. 5, bottom).

Relatedness detection using IBDGem

Determining self versus non-self using this framework is straightforward as self comparisons are IBD2 across every region of the genome and non-self comparisons are IBD0 across nearly every region. Closely related individuals, however, will share genomic regions where 1 chromosome is identical-by-descent (IBD1). For example, parent/offspring relationships will share the whole-genome IBD1, baring mutations, and full siblings will also share some regions of IBD2.

To assess the power of IBDGem to detect regions of IBD1 and, more generally, to assess the degree of relatedness between compared samples, we implemented a module (HiddenGem) that finds the most likely path of the 3 IBD states through the genome using regional likelihood values of each state (Supplementary Section 4 and Supplementary Fig. S13). We used the family pedigrees present within the 1000 Genomes Phase 3 panel as relationships between these individuals are provided. While the IBD state (0, 1, or 2) is not known for any particular region of the genome, the total amount of each state is a simple function of the type of relatedness. For example, parent–child relatives must be IBD1 across the whole genome as the child inherits exactly one of their 2 chromosomes from each parent. On the other hand, full siblings are expected to share both parental chromosomes at one-quarter of the genome, neither parental chromosome at one-quarter of the genome, and 1 parental chromosome at one-half of the genome.

We ran IBDGem followed by the maximum-likelihood IBD-state caller HiddenGem, comparing genotypes at only GSA sites for the known relatives of 2 individuals, NA19662 and NA19686, from the MXL (Mexican-American) population. In this experiment, we assumed site independence and calculated likelihoods for the IBD models (including IBD0) on a per-site basis, using the alternate allele frequency learned from all unrelated 1000 Genomes individuals at each site. Then, we simply multiplied over sites to get regional IBD

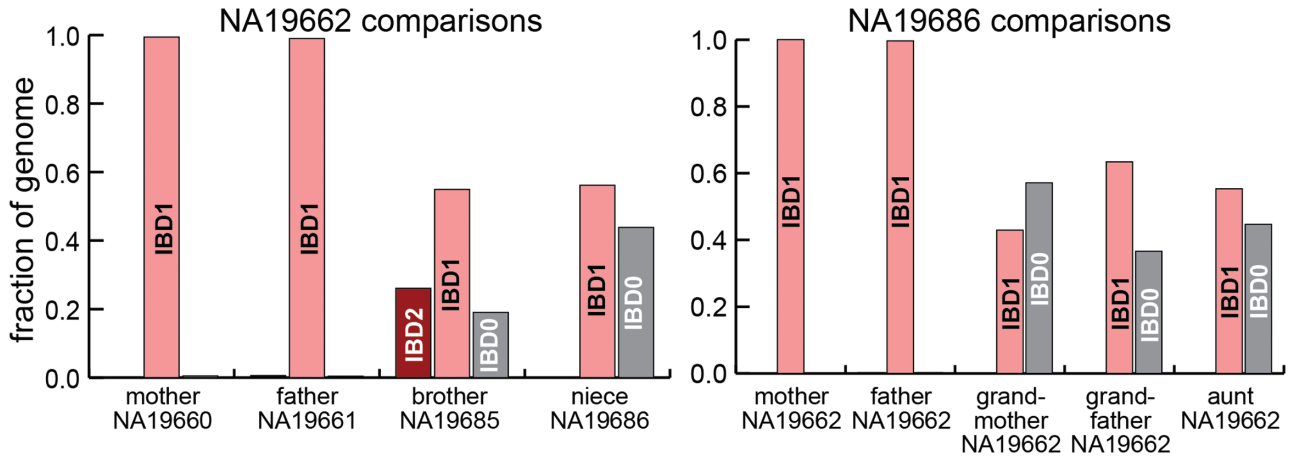

Fig. 6. IBDGem comparisons between related individuals in the MXL panel. Results of IBDGem at 1× down-sampled coverage followed by HiddenGem to apportion each genomic segment into IBD0, IBD1, or IBD2 states among annotated pedigrees.

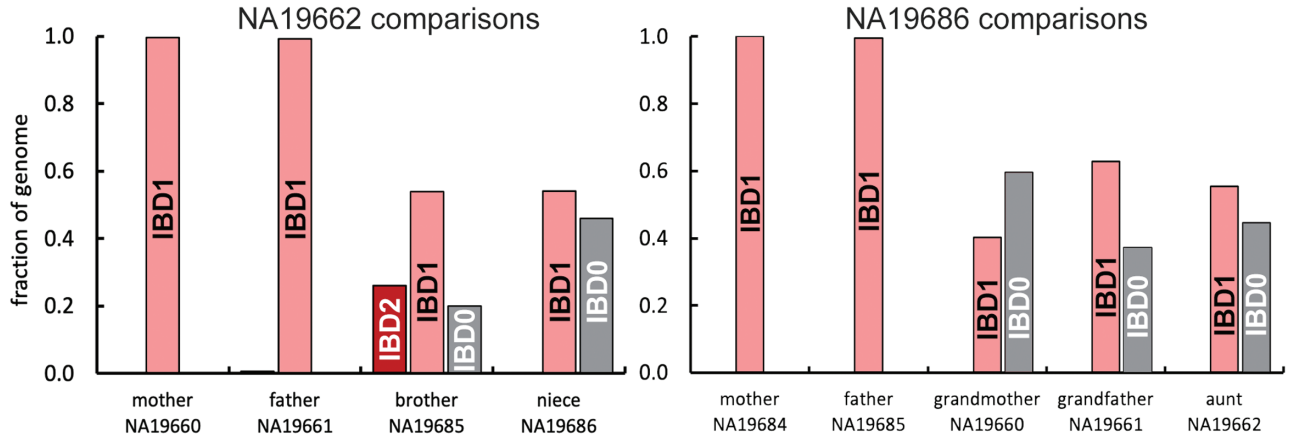

Fig. 7. IBDGem comparisons between related individuals in the MXL panel, using allele frequencies from a 50-individual subset for the background model. Results of IBDGem at 1× down-sampled coverage followed by HiddenGem to apportion each genomic segment into IBD0, IBD1, or IBD2 states among annotated pedigrees.

likelihoods. We down-sampled the sequence data from each individual to 1× average genome coverage. For each known relative, there is general concordance between the observed proportion of each IBD state and the expected values given the degree of relatedness (Fig. 6 and Supplementary Fig. S8). As expected, only the full-sibling comparison generates more than 1% of the genome assigned to IBD2. All parent–child comparisons assign all or nearly all of the genome to IBD1.

Aside from human-related applications, our method can be readily extended to other organisms for which identity and kinship are also of interest, in particular wildlife species in the context of conservation and population monitoring. The minimal sequence input requirements make IBDGem an appropriate tool for the analysis of challenging specimens such as hair and feathers that can be collected noninvasively for conservation purposes.

However, unlike humans, for most nonmodel species there is no reference panel similar in scale to the 1000 Genomes dataset from which allele frequencies can be inferred. Therefore, as a surrogate for evaluating the program's performance on a non-human organism with a smaller sample size, we subset the global 1000 Genomes dataset to only include 50 random individuals and used allele frequencies derived from this panel to perform kinship estimation. As before, we down-sampled the sequence data of individuals NA19662 and NA19686 from the MXL population to 1× genome coverage and compared these samples to their known relatives at only GSA sites, using the allele frequencies calculated from our 50-individual subset for the background model (IBD0). We then estimated the proportion of the genome shared IBD0, IBD1, and IBD2 between these relatives with the IBD-state caller HiddenGem (Fig. 7).

We found that the program is robust to a panel size reduced to 50 individuals to define the background, IBD0 model. All pairs of relatives showed the expected proportions of IBD0, IBD1, and IBD2 predicted by their degrees of relatedness. The observed proportions also show only small deviations from the previous experiment using the whole 1000 Genomes dataset containing thousands of individuals. We have also found that minor allele frequencies for a large proportion of biallelic SNP sites, when inferred from a random 50-individual subset, are still largely consistent with those estimated from the global 1000 Genomes panel (Supplementary Fig. S9). Overall, these results follow the expectation that allele frequencies, especially at common SNPs (Chakraborty 1992), can be reliably estimated from a small panel.

Discussion

Because it can reliably detect identity with minute quantities of input data, IBDGem addresses a specific problem in DNA-based forensics and other applications: extremely limited input DNA. We show that for the special case of identity, IBDGem reliably discriminates self versus non-self data with as little as 1% genome coverage. We also show that at 1× genome coverage, IBDGem and HiddenGem can detect IBD segments. Using DNA sequence data derived from rootless hairs, we show that IBDGem can distinguish self from non-self comparisons using input DNA samples that are not amenable to PCR-based forensic analysis. IBDGem does not require any genotype calling or genotype likelihood inference, imputation, or phasing of the query sample.

The central task in DNA-based forensics is determining if 2 samples derive from the same individual. Typically, the data from the 2 samples are symmetrically generated, i.e. 2 PCR-derived STR profiles, although the samples themselves may be different in DNA quality and quantity. IBDGem provides a framework for comparing nonsymmetric data types, i.e. a genotype file from a DNA sample for which there is abundant DNA and low-coverage sequence data from a limited forensic sample.

The statistical model within IBDGem works regionally across the genome. One pertinent question for forensic use is how the LLRs, across the genome, should be aggregated to generate a single statistic for distinguishing between self and non-self. For this, we propose aggregating LLRs across each arm of each autosome. For identification purposes, the chromosome arm with the minimum IBD2/IBD0 ratio represents a conservative and straightforward metric for discriminating self versus non-self.

For all GBR and LWK comparisons at 1× genome coverage, this approach generates LLR values of >3,576 for self-comparisons and <−12,410 for non-self comparisons (Supplementary Fig. S10A and B). For the hair panel comparisons, the minimum self-comparison result is 267.1. Using this metric, in the least confident self-comparison the hair data are 2.54×10^{80} times more likely under the model that they derive from a genetically identical person than under the model that they derive from an unrelated person. For the non-self comparisons using the hair data, the maximal LLR (IBD2/IBD0) is −35.26 (Supplementary Fig. S10C).

We also calculated this chromosome arm statistic for several first-degree relationships using known pedigrees from 1000 Genomes. We found that the smallest aggregated LLR between models IBD2 and IBD0 for every first-degree comparison is consistently negative (Supplementary Figs. S11 and S12). As described, lower-coverage data can result in higher LLR values (Fig. 2) across bins. When we aggregate likelihoods over a full chromosome arm, the distance spanned is the same regardless of the coverage of sequence data or sites chosen, and hence our summary statistics are less affected by this phenomenon.

We envision that the IBDGem and HiddenGem framework described here could be extended in several ways. First, HiddenGem currently has a single penalty value for switching between IBD states in its maximum-likelihood path calculation. Incorporation of known human recombination map data could help refine transition states, improving the fine-scale accuracy of IBD-state determination for comparisons between related individuals. This may be particularly beneficial for detecting distantly related individuals who share few, small IBD segments.

Second, HiddenGem does not explicitly estimate or report a level of relatedness. The amount and distribution of IBD states between 2 individuals necessarily fall within non-continuous categories (parent/child, full sibling, cousin, etc.) each with a characteristic mean and variance of expected IBD0, IBD1, and IBD2. Specifically, for any degree of relatedness, the amount of shared (IBD) DNA is reduced by a factor of 2. A simple future extension of IBDGem and HiddenGem could convert its results to the most likely level of relatedness. Because IBD2 and IBD1 are distinguishable using this framework, disambiguating between parent/child and full-sibling pairs (both first-degree genetic relationships) should be trivial even though both relationships are 50% genetically identical.

Third, the IBDGem algorithm does not currently analyze sex chromosome data. Simple modifications of the underlying probability equations could be made to handle the special case of the hemizygous sex chromosome in genetic males. This is a particularly easy extension given that identification of genetic sex is possible from minute amounts of genome DNA sequence data since genetic males have half the amount of chromosome X data as females.

Fourth, many older DNA samples will contain some amount of cytosine deamination (Briggs et al. 2007). This manifests as C to T errors in the sequencing data from these samples, with a specific profile along the DNA strand. IBDGem could therefore be extended to measure and model cytosine deamination.

Finally, the likelihood framework described here compares sequence data to a known genotype. The motivation for this approach is that it addresses a real-world scenario. However, in some instances it may be useful to directly compare 2 lower quality or lower-coverage datasets to one another. For example, it may be useful to compare low-coverage sequence data from 2 rootless hairs directly to one another, without calling genotypes. The likelihood framework described here could be extended, for example, to calculate the likelihood of the sequence data from 1 sample to a probabilistic genotype called from the other limited sample.

IBDGem makes comparisons between a known sample with genotype calls and a sample with limited amounts of DNA sequence data. In this asymmetric framework, it is assumed that the genotype error in the known sample is near zero and thus negligible. For the results presented here, these genotype calls come from either the 1000 Genomes Project analysis or from genotype array data from high-quality saliva-derived DNA samples. In other instances, it will be important to evaluate the quality of the genotype calls for the known sample, in whatever manner they are generated.

IBDGem further assumes that the known sample genotype and the low-input comparison DNA sequence are free of mixtures or contamination. Appropriate analysis of the input data should be done to rule out the presence of DNA from multiple contributors. For the low-input DNA sequence, analysis of the haploid mitochondrial genome or use of a program like tilde can detect the presence of sample mixtures or contamination (Vohr et al. 2015, 2017).

After more than a decade of advances in high-throughput DNA sequencing technology, it is now possible to recover and sequence DNA from sources that were once considered intractable for forensic purposes. Coupled with comprehensive catalogs of existing DNA variation, these technologies open powerful avenues for personal identification. Given the

ubiquitous nature of shed hair, however, it will be important for the criminal justice community to pay even greater attention to what has been termed activity level propositions (e.g. how the hair came to be in that location) rather than just source level propositions (e.g. whose hair it is) (Cook et al. 1998).

Supplementary material

Supplementary material is available at *Journal of Heredity* online.

Funding

R.N. is a Koerner Family Foundation Fellow and Phi Beta Kappa Northern California Association Fellow. This work was supported by the National Human Genome Research Institute at the National Institutes of Health (T32 HG012344); and the National Institute of Justice (2020-DQ-BX-0014 to R.E.G.).

Conflict of Interest

R.E.G. and R.N. are listed as co-inventors in patent applications filed by the University of California, Santa Cruz on the method described in this paper. R.E.G. is founder of and consultant for Astrea Forensics.

Data availability

We have deposited the raw sequence data as well as the genotype data for the rootless hair panel (Hair panel 1.0) underlying our analyses in dbGaP (Study Accession Number phs002979.v2.p1).

References

Alaeddini R, Walsh SJ, Abbas A. Forensic implications of genetic analyses from degraded DNA—a review. *Forensic Sci Int Genet*. 2010;4(3):148–157.

Ball CA, Barber MJ, Byrnes J, Carbonetto P, Chahine KG, Curtis RE, Granka JM, Han E, Hong EL, Kermany AR, et al. *AncestryDNA matching white paper*. AncestryDNA; 2016. Available from: https://www.ancestry.com/dna/resource/whitePaper/AncestryDNA-Matching-White-Paper.pdf.

Briggs AW, Stenzel U, Johnson PLF, Green RE, Kelso J, Prüfer K, Meyer M, Krause J, Ronan MT, Lachmann M, et al. Patterns of damage in genomic DNA sequences from a Neandertal. *Proc Natl Acad Sci U S A*. 2007;104(37):14616–14621.

Browning SR, Browning BL. High-resolution detection of identity by descent in unrelated individuals. *Am J Hum Genet*. 2010;86(4):526–539.

Browning BL, Browning SR. Detecting identity by descent and estimating genotype error rates in sequence data. *Am J Hum Genet*. 2013a;93(5):840–851.

Browning BL, Browning SR. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*. 2013b;194(2):459–471.

Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, Corvelo A, Clarke WE, Musunuri R, Nagulapalli K et al. High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios, *Cell*. 2022;185(18):3426–3440.e19.

Chakraborty R. Sample size requirements for addressing the population genetic issues of forensic use of DNA typing. *Hum Biol*. 1992;64(2):141–159.

Conomos MP, Reiner AP, Weir BS, Thornton TA. Model-free estimation of recent genetic relatedness. *Am J Hum Genet*. 2016;98(1):127–148.

Cook R, Evett IW, Jackson G, Jones PJ, Lambert JA. A hierarchy of propositions: deciding which level to address in casework. *Sci Justice*. 1998;38(4):231–239.

de Vries JH, Kling D, Vidaki A, Arp P, Kalamara V, Verbiest MMPJ, Piniewska-Róg D, Parsons TJ, Uitterlinden AG, Kayser M. Impact of SNP microarray analysis of compromised DNA on kinship classification success in the context of investigative genetic genealogy. *Forensic Sci Int Genet*. 2022;56:102625.

Durand EY, Eriksson N, McLean CY. Reducing pervasive false-positive identical-by-descent segments detected by large-scale pedigree analysis. *Mol Biol Evol*. 2014;31(8):2212–2222.

The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015;526:68–74.

Gill P, Jeffreys AJ, Werrett DJ. Forensic application of DNA 'fingerprints'. *Nature*. 1985;318:577–579.

Gorden EM, Greytak EM, Sturk-Andreaggi K, Cady J, McMahon TP, Armentrout S, Marshall C. Extended kinship analysis of historical remains using SNP capture. *Forensic Sci Int Genet*. 2022;57:102636.

Gusev A, Lowe JK, Stoffel M, Daly MJ, Altshuler D, Breslow JL, Friedman JM, Pe'er I. Whole population, genome-wide mapping of hidden relatedness. *Genome Res*. 2009;19(2):318–326.

Jeffreys AJ, Wilson V, Thein SL. Individual-specific 'fingerprints' of human DNA. *Nature*. 1985;316:76–79.

Jobling MA, Gill P. Encoded evidence: DNA in forensic analysis. *Nat Rev Genet*. 2004;5:739–751.

Kapp JD, Green RE, Shapiro B. A fast and efficient single-stranded genomic library preparation method optimized for ancient DNA. *J Hered*. 2021;112(3):241–249.

Kimpton CP, Gill P, Walton A, Urquhart A, Millican ES, Adams M. Automated DNA profiling employing multiplex amplification of short tandem repeat loci. *PCR Methods Appl*. 1993;3(1):13–22.

Kling D, Phillips C, Kennett D, Tillmar A. Investigative genetic genealogy: current methods, knowledge and practice. *Forensic Sci Int Genet*. 2021;52:102474.

Kling D, Tillmar A. Forensic genealogy—a comparison of methods to infer distant relationships based on dense SNP data. *Forensic Sci Int Genet*. 2019;42:113–124.

Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet*. 2011;12:443–451.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559–575.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. Genetic structure of human populations. *Science*. 2002;298(5602):2381–2385.

Shaffer HB, Toffelmier E, Corbett-Detig RB, Escalona M, Erickson B, Fiedler P, Gold M, Harrigan RJ, Hodges S, Luckau TK, et al. Landscape genomics to enable conservation actions: the California Conservation Genomics Project. *J Hered*. 2022;113(6):577–588.

Swango KL, Timken MD, Chong MD, Buoncristiani MR. A quantitative PCR assay for the assessment of DNA degradation in forensic samples. *Forensic Sci Int*. 2006;158(1):14–26.

Szabo S, Jaeger K, Fischer H, Tschachler E, Parson W, Eckhart L. In situ labeling of DNA reveals interindividual variation in nuclear DNA breakdown in hair and may be useful to predict success of forensic genotyping of hair. *Int J Legal Med*. 2012;126(1):63–70.

Turner SD, Nagraj VP, Scholz M, Jessa S, Acevedo C, Ge J, Woerner AE, Budowle B. Evaluating the impact of dropout and genotyping error on SNP-based Kinship analysis with forensic samples. *Front Genet*. 2022;13:882268.

Vohr SH, Buen Abad Najar CF, Shapiro B, Green RE. A method for positive forensic identification of samples from extremely low-coverage sequence data. *BMC Genomics*. 2015;16:1034–1034.

Vohr SH, Gordon R, Eizenga JM, Erlich HA, Calloway CD, Green RE. A phylogenetic approach for haplotype analysis of sequence data from complex mitochondrial mixtures. *Forensic Sci Int Genet*. 2017;30:93–105.