On forensic likelihood ratios from low-coverage sequencing

Feriel Ouerghi[1], Dan E. Krane[2], and Michael D. Edge*[1]

[1]Department of Quantitative and Computational Biology, University of Southern California
[1]Department of Biological Sciences, Wright State University

May 24, 2024

Abstract

With advances in sequencing technology, forensic workers can access genetic information from increasingly challenging samples. A recently published computational approach, `IBDGem`, analyzes sequencing reads, including from low-coverage samples, in order to arrive at likelihood ratios for tests of identity. Here, we show that likelihood ratios produced by `IBDGem` test a null hypothesis different from the traditional one used in a forensic genetics context. In particular, rather than testing the hypothesis that the sample comes from a person unrelated to the person of interest, `IBDGem` tests the hypothesis that the sample comes from an individual who is included in the reference sample used to run the method. This null hypothesis is not generally of forensic interest, because the defense hypothesis is not that the evidence comes from an individual included in a reference panel. Further, it does not take into account genetic variation outside the reference panel, and as a result, the computed likelihood ratios can be much larger than likelihood ratios computed for the standard forensic null hypothesis, often by many orders of magnitude, thus potentially creating an impression of stronger evidence for identity than is warranted. We lay out this result and illustrate it with examples, giving suggestions for directions that might lead to likelihood ratios that have the traditional interpretation.

Introduction

One crucial goal of forensic genetics is identity testing, in which a biological sample can be tested to see whether the genetic data extracted from it could have originated from a person of interest (Balding & Donnelly, 1995; Gill et al., 1985; Jobling & Gill, 2004). The results are often expressed as a likelihood ratio comparing two hypotheses, with the

*Correspondence: edgem@usc.edu

prosecution's hypothesis in the numerator and a defense hypothesis in the denominator (Balding & Steele, 2015). Most typically, in the numerator, the hypothesis is that the biological sample comes from a person of interest (e.g. a suspect), and in the denominator, the hypothesis is that the biological sample comes from a randomly chosen person from a population of alternative sources who is unrelated to the person of interest. A long forensic and judicial tradition has built up in the interpretation of these likelihood ratios. Historically, such tests are most often performed with short tandem repeat (STR) markers, also called microsatellites (Butler, 2005).

In the last 30 years, genotyping technologies have advanced rapidly, and it is now possible to extract genetic information from samples that would have been intractable previously (Haddrill, 2021). Alongside new ways of generating data from challenging samples, new methods of analysis have appeared. For example, probabilistic genotyping systems (PGS) are now frequently used to analyze STR data from low-template and/or mixed samples (Coble & Bright, 2019).

For very challenging samples, it may be possible to generate some sequencing reads even if PCR amplification, which is necessary for STR typing, fails. In such a scenario, the sequencing depth may be too low to allow accurate genotype calls (Nielsen et al., 2011), precluding the use of many existing methods and requiring special tools.

Recently, Nguyen and colleagues (2023) developed `IBDGem` to address this gap, facilitating identity inference from low-coverage sequence data. `IBDGem` was developed with forensics in mind as a primary application area—forensic applications are mentioned in the first lines of both the introduction and discussion of the paper and and many other times throughout the manuscript. Nguyen and colleagues developed a likelihood ratio statistic and evaluated its ability to distinguish identity from non-identity comparisons both within their reference data, with simulated lower coverage, and in real genotypes recovered from rootless hair and saliva. However, as we show here, the likelihood ratio statistic they highlight considers a hypothesis that is not generally of forensic interest. Specifically, it considers the hypothesis that the observed sequencing reads were derived from biological material from an individual who is included as a member of the reference panel used to run the method.

We show that likelihood ratios produced by `IBDGem` may be many orders of magnitude larger than ones that result from tests of the relevant defense hypothesis. If used in forensic work, these `IBDGem` likelihood ratios could potentially produce an impression of much stronger evidence for identity than is warranted. In the discussion, we suggest directions that might lead to likelihood ratios more in line with standard forensic use.

Results

The `IBDGem` approach

Typically, likelihood ratios in forensics take the form

$$LR = \frac{P(D|H_p)}{P(D|H_d)}, \tag{1}$$

where D represents the event of observing the data (often called the "evidence" in forensics), H_p is a hypothesis advanced by the prosecution, and H_d is an alternative explanation for the observed data friendly to the defense (Williams & Maskell, 2021). In identity testing scenarios, this general expression often appears as

$$LR = \frac{P(D|I)}{P(D|U)}, \tag{2}$$

where I represents the event that the biological sample comes from the person of interest, and U represents the event that the biological sample comes from a person drawn at random from the population of alternative sources, unrelated to the person of interest. (For brevity, we refer to the population of alternative sources as "the relevant population" or just "the population" below.)

`IBDGem` aims to arrive at a likelihood ratio of the form in equation 2, taking as input sequencing reads, and in particular, the alleles that map to sites that are known to be biallelic. At the ith biallelic site, a person of interest may have one of three genotypes: (0,0), (0,1), or (1,1), where "0" represents a reference allele and "1" an alternate allele. A biological sample may produce a set of reads that map to this site, and the data from those reads can be represented as a pair of integers (D_{i0}, D_{i1}), where D_{ij} represents the number of alleles of type j observed at biallelic site i.

`IBDGem` needs to compute two probabilities to form a likelihood ratio: the probability of observing the observed sequencing reads given that the source of the reads (i) is the person of interest, and (ii) is not. If the source is the person of interest, whose genotype is assumed known with certainty, then the probability of observing a set of reads can be understood as arising from randomness in which of the person's two allelic copies appear in the reads, and randomness from sequencing errors. If sequencing errors are symmetric and occur with probability ϵ per base pair, if each allele is equally likely to appear in a sequencing read, and if randomness in the total number of sequencing reads at a site is ignored, then (Nguyen et al., 2023),

$$P(D_i|G) = \begin{cases} (1-\epsilon)^{D_{i0}}\epsilon^{D_{i1}} & \text{if } G = (0,0) \\ (1-\epsilon)^{D_{i1}}\epsilon^{D_{i0}} & \text{if } G = (1,1) \\ \binom{D_{i0}+D_{i1}}{D_{i1}}\frac{1}{2^{D_{i0}+D_{i1}}} & \text{if } G = (0,1). \end{cases} \tag{3}$$

`IBDGem` uses a default value of 0.02 for the (symmetric) sequencing error rate ϵ. If the reads are independent (e.g. because variable sites are substantially distanced as to appear on different sequencing reads), then the probabilities produced by equation 3 for each of k variable sites can be multiplied to form a probability of observing the sequencing data at the set of sites:

$$P(D|I) = \prod_{i=1}^{k} P(D_i|G_i), \tag{4}$$

where G_i represents the genotype of the person of interest at site i, and I is the event that the biological sample comes from the individual of interest.

If a variable site from which reads have been obtained can be assumed to be in Hardy–Weinberg equilibrium with known allele frequency p in the relevant population, then one can compute the probability that an unrelated individual would produce the observed reads at site i by summing $P(D_i|G)$ over the three possible genotypes, weighted by their population frequencies. That is, one can write

$$P(D_i|U) = \sum_{G} P(D_i|G)P(G|U), \tag{5}$$

where the probabilities of the genotypes $P(G|U)$ are the Hardy–Weinberg frequencies p^2, $2p(1-p)$, and $(1-p)^2$. If, further, genotypes at the variable sites are all independent (that is, they are in linkage equilibrium), and we retain the assumption of independence of reads as before, then the likelihood of the data can be computed by the product rule, giving

$$P(D|U) = \prod_{i=1}^{k} P(D_i|U) = \prod_{i=1}^{k} \sum_{G} P(D_i|G)P(G|U). \tag{6}$$

However, equation 6 will not give the correct probability if variable sites in linkage disequilibrium (i.e. sites with non-independent genotypes) are used. In particular, equation 6, by treating each site as independent, risks overstating the amount of identifying information available at each site. Rather than providing a likelihood, it provides a composite likelihood (Larribe & Fearnhead, 2011). In dealing with closely spaced markers derived from sequencing, linkage disequilibrium is very likely.

In response to the challenge presented by linkage disequilibrium, the makers of `IBDGem` developed and recommend an "LD mode" as a method to account for linkage disequilibrium. The LD mode approach is featured in Figure 1 of Nguyen and colleagues (2023) as a major advance of the method. Nguyen and colleagues suggest that the likelihood ratios arising from LD mode can be treated as testing the null hypothesis we label U, and they label "$IBD0$," writing "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." However, as we show below,

the `IBDGem` approach to LD correction tests a different null hypothesis, sometimes leading to statements that are more anticonservative even than those that follow from ignoring linkage disequilibrium.

`IBDGem`'s LD mode is based on comparing reads with the specific individuals composing the reference sample, rather than comparing them with population allele or haplotype frequencies estimated via the reference sample. The stated rationale is that in comparing to the specific individuals in the reference panel, `IBDGem` "asks how the sequence data look against random individuals who we know share zero IBD chromosomes" (Nguyen et al., 2023). If one has access to a reference sample of n individuals with known genotypes, then one can compute the probability in equation 4 with respect to each of these n individuals, simply replacing the person of interest's genotypes with the genotypes of each member of the reference panel in turn. Nguyen and colleagues (2023) then compute the average of this probability over individuals in the reference panel,

$$P(D|R) = \frac{1}{n}\sum_j P(D|R_j),$$

where R_j represents the hypothesis that the jth member of the reference panel is the source of the reads. Nguyen and colleagues refer to this quantity as an estimate of $P(D|IBD0)$ that accounts for LD, and they use it as the denominator in reported likelihood ratios. (They use $IBD0$ to denote the event that the source is an individual with no identity by descent segments with the person of interest in the region being examined.) We, however, label this quantity $P(D|R)$, the probability of the data arising from a (uniformly) randomly chosen member of reference panel. (R denotes the event that the source is an individual drawn uniformly at random from the reference panel.) This is because it has exactly that interpretation:

$$P(D|R) = \sum_j P(R_j)P(D|R_j) = \frac{1}{n}\sum_j P(D|R_j), \tag{7}$$

. This probability is not the one of forensic interest—that of observing the data from a random, unrelated individual in the population of alternative sources—unless all alternative sources are included in the reference panel. Nguyen and colleagues do not point out the difference.

Thus, whereas the likelihood ratio generally of forensic interest is, in our notation, $P(D|I)/P(D|U)$, `IBDGem`'s LD mode forms a likelihood ratio

$$LR = \frac{P(D|I)}{P(D|R)}, \tag{8}$$

replacing the probability of observing the data from an unrelated person in the relevant population, U, with the probability of observing the data from a random person in the reference panel, R. One way to express the difference between these likelihood ratios is

that in typical forensic practice, chance variation due to both genotyping error and genetic variation is considered. However, in the likelihood ratio produced by `IBDGem`'s LD mode, there is no model for genetic variation outside the reference panel.

To generalize from the reference panel to the population of potential sources of the sample requires a model of genetic variation, not just a model of sequencing reads and errors therein. A very low value of $P(D|R)$ can be informative about identity, particularly if the reference panel is large, but it cannot generally be interpreted as the probability of observing the data from a random unrelated person from the relevant population, $P(D|U)$.

One way in which $P(D|U)$ and $P(D|R)$ differ can be seen immediately: if the genotyping error rate is 0, and read data have at least one inconsistency with every member of the reference panel, then $P(D|R) = 0$. This is true even for arbitrarily small reference panels and arbitrarily few loci, even though a small reference panel with few loci represented leaves open the possibility that many individuals in the population carry genotypes consistent with the observed reads, meaning that $P(D|U) > 0$.

We illustrate that the values of $P(D|R)$ and $P(D|U)$ can diverge widely in examples below, often leading to likelihood ratios that differ by many orders of magnitude.

A simplified haploid example

For illustration, we focus on a simplified case. A haploid individual of interest carries "0" alleles at each of k biallelic loci. At each of these k loci, the frequency of the "1" allele is p, and alleles at distinct loci are independent. We have a reference panel of n randomly selected individuals with known alleles at each of the k loci. We also have sequencing reads from a biological source, exactly one read from each site, with symmetric error rate ϵ, and all of the reads are "0" alleles, meaning that they match the individual of interest.

In this scenario,

$$P(D|I) = (1-\epsilon)^k. \tag{9}$$

That is, if the haploid of interest is the source of the sequencing reads, we will observe sequencing reads that match the haploid of interest at all k sites if and only if there are no sequencing errors at any of the k sites.

We can also compute the probability of observing the reads from an individual with $L = l$ "1" alleles and $k-l$ "0" alleles as $P(D|L=l) = \epsilon^l(1-\epsilon)^{k-l}$. That is, if an individual does not match the sequencing reads at l positions, then to produce the sequencing reads, there must have been exactly l sequencing errors, occurring in exactly the positions of mismatch.

The probability of observing the sequencing reads from an unrelated individual in the population can be obtained by summing $P(D|L=l)$ over the distribution of L for unrelated individuals,

$$P(D|U) = \sum_{l=0}^{k} P(D|L=l,U)P(L=l|U) = \sum_{l=0}^{k} \epsilon^l(1-\epsilon)^{k-l}\binom{k}{l}p^l(1-p)^{k-l}. \tag{10}$$

0	1	1	1	0	1	1	0	1	0
0	1	1	1	1	0	1	0	1	0
1	1	1	0	1	0	0	0	1	0
1	1	0	1	0	0	1	0	1	0
0	0	1	1	1	1	0	1	0	1

Table 1: Simulated data for the haploid model with $n = 5$ and $k = 10$. Individuals are in rows, loci in columns

The second step follows because under the described model, L has a Binomial(k, p) distribution among individuals unrelated to the haploid of interest, giving $P(L = l|U) = \binom{k}{l} p^l (1-p)^{k-l}$. The sum in equation 10 is clearly bounded from below by its $l = 0$ term, $(1-p)^k(1-\epsilon)^k$, the probability that an individual unrelated to the haploid of interest has a genotype exactly matching the haploid of interest and there are no sequencing errors. Thus, the likelihood ratio formed by dividing equation 9 by equation 10 is bounded from above by $1/(1-p)^k$. An alternative, equivalent expression for $P(D|U)$ closer to the spirit of equation 6 is

$$P(D|U) = [p\epsilon + (1-p)(1-\epsilon)]^k. \tag{11}$$

In contrast, the probability of observing the data from a randomly chosen member of the reference panel is sensitive to the genotypes of the specific individuals in the reference panel. For illustration, Table 1 shows a randomly drawn reference panel with $n = 5$ individuals, $k = 10$ loci, and $p = 1/2$ the allele frequency at each locus. For these parameters, $P(D|I) \approx 0.82$, and $P(D|U) \approx 9.8 \times 10^{-4}$. The typical forensic likelihood ratio comparing the probability of the observed data under the hypotheses that the sample source is the person of interest vs. a random individual from the population is thus $P(D|I)/P(D|U) \approx 837$. However, the probability of observing the data from a randomly chosen individual in the reference panel in Table 1 is only $P(D|R) \approx 1.2 \times 10^{-9}$. Thus, in this example, using $P(D|R)$ as a stand-in for $P(D|U)$, as `IBDGem` does in LD mode, gives a likelihood ratio of $\approx 6.9 \times 10^8$, a difference of over five orders of magnitude.

The numbers in this example are not unique, nor are they extreme. Figure 1 shows how the two likelihood ratios and their relationship change with the number of loci, size of the simulated reference panel, and error rate. For many sets of parameter values, likelihood ratios of the form `IBDGem` uses are many orders of magnitude larger than standard forensic likelihood ratios.

The difference between the likelihood ratios is determined by $P(D|U)$ and $P(D|R)$—if the probability of obtaining the observed reads from a random member of the reference panel is much lower than the probability of obtaining the observed reads from a random member of the relevant population, then the likelihood ratio produced by `IBDGem`'s LD mode will be much larger than the likelihood ratio testing the hypothesis generally of forensic interest. If no one in the reference panel closely matches the reads produced, then

the only way to accommodate the data with the reference panel is to posit sequencing errors. Thus, the `IBDGem` likelihood ratios will be especially large when the assumed sequencing error rate is low (or, similarly, when sequencing depth is high), when the observed reads contain alleles from rare haplotypes (less likely to be observed in the reference panel), or when the reference panel is small (meaning that it is unlikely to contain haplotypes similar to the observed reads).

LD-mode values of $P(D|R)$ roughly match $P(D|U)$ when $P(D|U) >> 1/n$

If the reference panel contains the entire population that is the target of inference, then $P(D|R)$ and $P(D|U)$ become equivalent, and `IBDGem`'s LD mode likelihood ratio addresses the forensic question of interest. We might expect, similarly, that if the reference panel is large enough to contain individuals with haplotypes matching the person of interest at approximately their frequency in the relevant population, then $P(D|R)$ and $P(D|U)$ should be similar. In simulations with small numbers of loci, it is tractable to simulate reference panels with $n >> 1/P(D|U)$, meaning that the individual-of-interest haplotype is likely to be observed in the reference.

Figure 2 shows simulations similar to Figure 1, but with smaller numbers of loci. In each of panels A-C, likelihood ratios with $P(D|U)$ and $P(D|R)$ appear to match (albeit with some variation) if the reference panel is large enough that we would expect approximately 5-10 copies of the individual-of-interest haplotype in the panel. With low sequencing error rates, this occurs when $P(D|U) > 5/n$. Higher sequencing error rates (in combination with low sequencing depth) cause the likelihood ratios to converge at lower reference sample sizes, because mismatching alleles can be readily explained as sequencing errors.

`IBDGem` LD-mode likelihood ratios under linkage disequilibrium

`IBDGem`'s LD mode was developed in order to account for correlations among loci, or linkage disequilibrium (Slatkin, 2008). Linkage disequilibrium is likely in sequence data, which may produce reads from closely spaced loci. In the above scenarios, loci were in linkage equilibrium, so the model assumed by `IBDGem`'s non-LD mode, ignoring linkage disequilibrium, is correct.

We ran simulations to explore the effect of linkage disequilibrum on likelihood ratios computed in the manner of `IBDGem`, extending the haploid model above. We use two different simplified representations of linkage disequilibrium, one that contains blocks of markers in perfect ($r^2 = 1$) linkage disequilibrium, and one in which all markers are in complete ($D' = 1$) linkage disequilibrium, but no markers are in perfect ($r^2 = 1$) linkage disequilibrium. Specifically, to simulate blocks of perfect linkage disequilibrium, we imagine that every variable locus has m exact copies in the same "block"—that is, if a "1" allele is observed in a block, then all other alleles will be "1" alleles, and similarly for "0" alleles. In this scenario, the no-LD model will count distinct markers in the same block as independent

1/5	0	0	0	0
1/5	1	0	0	0
1/5	1	1	0	0
1/5	1	1	1	0
1/5	1	1	1	1
p	4/5	3/5	2/5	1/5

Table 2: Five possible haplotypes in the haploid "caterpillar" simulation with $k = 5$ loci, along with the haplotype frequencies (1st column) and implied allele frequencies (bottom row).

pieces of information, when the genotypes are in fact completely redundant.

We also simulate a scenario in which k markers are in complete, but not perfect, linkage disequilibrium, forming $k + 1$ distinct haplotypes at equal frequency. The specific scheme we use is consistent with a "caterpillar" gene tree (Steel, 2016), and is exemplified in Table 2 for the case of $k = 4$ variable loci. In this scheme, all loci are in complete linkage disequilibrium ($D' = 1$), meaning that variation in the population is compatible with a history in which no recombination has occurred, but because the allele frequencies differ, they are not in perfect linkage disequilibrium (VanLiere & Rosenberg, 2008). For more, see Method Details.

Figure 3 shows results from these simulations. As expected, `IBDGem`'s non-LD approach produces biased likelihood ratios in the presence of linkage disequilibrium. But, as in the other scenarios, LD mode does not generally correct this problem if the reference panel size is not large enough to contain the exact haplotype of the individual of interest.

`IBDGem` recapitulates these results with simulated diploid data

We next extend the haploid model illustrated in the first example to diploids and replicate the basic patterns seen in our haploid implementation in `IBDGem` itself. We simulate diploids by pairing together two haplotypes generated using the same method as in the unlinked haploid model. In these simulations, the individual of interest no longer carries "0" alleles at each of the k biallelic sites—they now carry a mix of heterozygous and homozygous genotypes at variable sites. Under this model, we simulate sequencing reads for the target individual based on their genotype while allowing for sequencing errors. We use the same error rate used to simulate sequencing reads to analyze the data in `IBDGem`, using the `--error-rate` flag. We also simulate a reference panel of n diploid individuals with known genotypes at each of the k loci. See Method Details for more description.

We use the simulated sequencing reads, the genotypes of the target individual and reference panel, and allele frequencies as input to `IBDGem`, running it in both LD and non-LD modes. Because we are simulating genotypes at linkage equilibrium, the non-LD-mode likelihood ratio is equal to the standard forensic likelihood ratio. Figure 4 shows how

IBDGem's LD-mode and non-LD-mode likelihood ratios and their relationship change with the size of the simulated reference panel, the sequencing depth (i.e. number of reads per variable site), the number of loci, and error rate. The basic patterns are the same as in the haploid case—for many sets of parameters, the likelihood ratios produced by IBDGem's LD mode are much larger than the likelihood ratios produced by the non-LD mode. Similar patterns are seen with more realistic allele-frequency distributions (not shown), though more loci are required to achieve comparable likelihood ratios when loci with low minor-allele frequency are included, since most individuals carry common alleles at most sites, and common alleles are less informative for purposes of identification.

Discussion

We showed that the likelihood ratios produced by IBDGem can be badly misleading about the strength of evidence in identity inference. Likelihood ratios produced by IBDGem's non-LD mode can be strongly anti-conservative in the presence of linkage disequilibrium, as expected. As such, the makers of IBDGem developed LD mode "to account for linkage disequilibrium" by using a reference panel and "compar[ing] the observed data against the genotypes of these unrelated individuals" (Nguyen et al., 2023). But this approach changes the meaning of the likelihood ratio, which Nguyen and colleagues do not note—in LD mode, the null hypothesis is that the data come from a random member of the reference sample, rather than a random, unrelated person from the population of potential sources of the sample.

In a forensic context, the defense's hypothesis for explaining sequencing reads from a biological sample is generally that the biological sample comes from an unknown person in the population of potential contributors of the sample, not from a member of a specific reference panel. Whether a member of the 1000 Genomes panel (Auton et al., 2015), for example, is the source of the biological sample is not generally the question of interest. Moreover, the likelihood ratios obtained from the LD-mode approach of IBDGem can differ from the ones calculated with respect to the standard forensic hypotheses by many orders of magnitude, potentially producing an impression of much stronger evidence for identity than is warranted.

The work of Nguyen and colleagues (2023) establishes that there is substantial information relevant for identity inference even in low-coverage sequencing reads. The issue is attaching a valid (or conservative) assessment of the strength of the evidence for identity. There are several possible routes by which approaches like that of IBDGem might pursue this goal. Perhaps the simplest is to prune variable sites such that only sites in linkage equilibrium are likely to be included, and then use the product rule. Though this would throw away information, a great deal would still be left over. (However, linkage disequilibrium arising from population structure would be a key consideration, and it is not clear whether sufficient genotype information is currently available to adequately as-

sess linkage disequilibrium for what are likely to be relevant pools of alternative sources of biological material.) A second approach would be to borrow methods from the analysis of rare "lineage markers," such as Y-STRs (Brenner, 2010; Buckleton et al., 2011)—when no haplotypes in a size-n reference panel of randomly selected individuals can readily explain the observed reads, then the probability of the underlying haplotype might be assigned a value based on $1/n$, perhaps with some adjustment (e.g. the use of the lower bound of a 95% confidence interval). Such an approach would be concordant with our result that $P(D|U)$ and $P(D|R)$ tend to be similar if the database is large enough to include multiple instances of the individual of interest's haplotype. A third approach might be to use the reference data to build a model of haplotype structure (Li & Stephens, 2003) and use the estimated haplotype frequencies as a basis for inference.

The issues we point out here regarding the interpretation of the likelihood ratio arise even in panmictic populations. More work is needed to understand the effects of population structure and representation in the database, as well as hypotheses involving biological relatives of the person of interest. Structure will lead to departures from Hardy–Weinberg equilibrium and from linkage equilibrium, even for unlinked markers, causing model violations for `IBDGem`'s non-LD mode. Further, if the source of the biological material in an evidence sample is from a population not well represented in the reference panel, then the evidence against a person of interest may be overstated, a recurrent issue in forensic genetics (Balding & Nichols, 1994; Lewontin & Hartl, 1991). Given the strong dependence of LD mode on the specific individuals included in the reference panel, issues of population structure and representation may be especially salient.

Our examination of `IBDGem` has been limited to the statistical issues regarding the likelihood ratios discussed here. We have not considered aspects of `IBDGem` downstream of the likelihood ratio for identity, such as likelihood ratios for IBD1 or detection of IBD segments.

We primarily used a haploid model and focused on the case in which exactly one read from each site is available, and all match the person of interest. We did so using our own implementation of haploid versions of the models in `IBDGem`. We made this choice so as to explain and demonstrate the basic issues as simply as possible. However, because the authors of `IBDGem` made their work open and available, we were able to explore some of our results in their software. This level of openness is admirable and not universal in forensic genetics (Chin et al., 2019; Edge & Matthews, 2022; Matthews et al., 2019).

We focused here on the values taken by the likelihood ratio when the hypothesis generally of interest to the prosecution—that a known person of interest is the source of the evidence—is true. It is also important to delimit the circumstances under which likelihood ratios computed by `IBDGem` could lead the evidence to appear to favor the prosecution hypothesis when the evidence should instead favor the defense hypothesis. Exploration of this question should entail both mathematical approaches and realistic simulations of genetic variation (Baumdicker et al., 2021; Haller & Messer, 2019).

The tremendous advances in DNA extraction from challenging samples holds promise

for forensic identification (Berglund et al., 2011; Kapp et al., 2021; Kayser & de Knijff, 2011; Vohr et al., 2015). At the same time, stakes in the courtroom are very high, and the statistical behavior of new methods needs careful validation. With respect to likelihood ratios used in the courtroom, it is important that they are constructed with respect to clear hypotheses of forensic interest, and that they are well calibrated or, if in error, conservative.

Code availability

Code used in this work can be found at https://github.com/fouerghi/IBDGem-LRs.

Acknowledgments

We thank N. Adams, G. Coop, V. Link, M. Przeworski, R. Rohlfs, N. Rosenberg, and J. Schraiber for helpful conversations. We acknowledge support from NIH grant R35GM137758 to MDE.

Declaration of interests

DEK is the president of Forensic Bioinformatic Services, Inc., a consulting company that provides assistance to parties in criminal trials involving DNA profiling evidence.

Method Details

Haploid models including linkage disequilibrium

Blocks of perfect linkage disequilibrium

To simulate blocks of m loci in perfect linkage disequilibrium, we made $m-1$ copies of the originally simulated haplotypes. $P(D|I)$ and $P(D|R)$ for data consisting of one "0" read from each locus were computed as before, only on sets of $m*k$ loci. To compute the likelihood ratio assuming linkage equilibrium, we used the product rule to obtain $P(D|U)$, including all loci. To compute the correct likelihood ratio accounting for linkage disequilibrium, we used

$$P(D|U) = \sum_{l=0}^{k} P(D|L=l,U)P(L=l|U) = \sum_{l=0}^{k} \epsilon^{ml}(1-\epsilon)^{m(k-l)}\binom{k}{l}p^l(1-p)^{k-l}. \quad (12)$$

This reduces to equation 10 if $m=1$. If $m>1$, then in any blocks that do not match the individual of interest, there must be m sequencing errors to explain the data, leading

to the $\epsilon^{ml}(1-\epsilon)^{m(k-l)}$, where l is now the number of blocks, rather than the number of variable loci, at which an unrelated individual does not match the individual of interest.

"Caterpillar" complete-linkage-disequilibrium haplotypes

To simulate "caterpillar-like" complete-linkage-disequilibrium haplotypes, as exemplified in Table 2, we drew integers between 0 and k uniformly at random. The value drawn determines the number of "1" alleles at the start of the haplotype; the rest are assigned "0".

$P(D|I)$ and $P(D|R)$ for data consisting of one "0" read from each locus were computed as before. To compute the likelihood ratio assuming linkage equilibrium, we used the product rule to obtain $P(D|U)$, including all loci, and using the allele frequencies $k/(k+1), (k-1)/(k+1), ..., 1/(k+1)$. To compute the correct likelihood ratios, we used

$$P(D|U) = \sum_{l=0}^{k} P(D|L=l,U)P(L=l|U) = \sum_{l=0}^{k} \epsilon^{l}(1-\epsilon)^{k-l}\frac{1}{k+1}. \tag{13}$$

Here, $P(L=l|U) = 1/(k+1)$ because there are $k+1$ equally frequent haplotypes, one with each possible value of L.

Diploid simulations using `IBDGem`

Diploid genotypes

Building on the simplified haploid example, we simulated a diploid target individual by simulating two haplotypes that are paired to form two genomes. In the Figure, we used $p = 1/2$ and constructed a target individual with genotypes close to Hardy–Weinberg frequencies—heterozygous at half the loci and homozygous for each of the alleles at a quarter of the loci each (getting as close to these numbers as possible given the number of loci). The target genotypes were kept constant across all figure panels except the panel varying the number of loci. Reference individuals were generated according to Hardy–Weinberg and linkage equilibrium.

Reads for diploid genotypes

In most panels of Figure 4, we generated exactly two reads per site, keeping the specific reads constant. In the panel exploring varied sequencing depth, the numbers of reads per site were drawn from a Poisson distribution parameterized by the sequencing depth. To obtain these reads, we used the genotypes of the individual of interest at variable sites and sampled from each allele randomly, picking the allele represented in the sequencing read uniformly at random if the site is heterozygous. We then drew sequencing reads from a biological source that matched the sampled allele at each locus with probability 1-ϵ; otherwise, reads did not match the sampled allele.

`IBDGem` Inputs

We stored the genotypes of the individuals in the reference panel in a `.hap` file required as input to `IBDGem`. We additionally generated the `.indv` file that has a list of the samples included in the reference panel. `IBDGem` also requires a `.legend` file that has the IDs of the variable sites, their positions along the specific chromosome and the reference and alternate bases. To generate this file, we randomly sampled from the four nucleotides A,T,C and G for the reference base and then sampled from the remaining three nucleotides for the alternate base.

We provided the simulated reads for the target genotype in the pileup format. This requires the "base quals" column that indicates the probability of an incorrect base call in a specific position. We adjusted the values in this column to match the error rate parameter. We simulate the column "mapping quality" to be around 0, to indicate our assumption that the reads have not been mis-mapped.

`IBDGem` Commands

To perform the `IBDGem` comparison between the DNA sequence and the individual of interest under linkage disequilibrium, we run the command with the following options:

```
ibdgem --LD \
    --hap reference_panel.hap \
    --legend reference_panel.legend \
    --indv reference_panel.indv \
    --pileup unknown_dna.pileup \
    --pileup-name sample100 \
    --sample sample100 \
    --window-size 50 \
    --error-rate 0.02
```

We remove the –LD flag to run the comparison under linkage equilibrium. We adjust the –error-rate and –window-size flags depending on the parameters defined in the simulation.

References

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., ... National Eye Institute, N. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Balding, D. J., & Donnelly, P. (1995). Inference in forensic identification. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *158*(1), 21–40. https://doi.org/https://doi.org/10.2307/2983402

Balding, D. J., & Nichols, R. A. (1994). DNA profile match probability calculation: How to allow for population stratification, relatedness, database selection and single bands. *Forensic Science International*, *64*(2), 125–140. https://doi.org/10.1016/0379-0738(94)90222-4

Balding, D. J., & Steele, C. D. (2015). *Weight-of-evidence for forensic dna profiles*. John Wiley & Sons.

Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A. P., Tsambos, G., Zhu, S., Eldon, B., Ellerman, E. C., Galloway, J. G., Gladstein, A. L., Gorjanc, G., Guo, B., Jeffery, B., Kretzschumar, W. W., Lohse, K., Matschiner, M., Nelson, D., Pope, N. S., ... Kelleher, J. (2021). Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, *220*(3), iyab229. https://doi.org/10.1093/genetics/iyab229

Berglund, E. C., Kiialainen, A., & Syvänen, A.-C. (2011). Next-generation sequencing technologies and applications for human genetic history and forensics. *Investigative Genetics*, *2*(1), 23. https://doi.org/10.1186/2041-2223-2-23

Brenner, C. H. (2010). Fundamental problem of forensic mathematics—The evidential value of a rare haplotype [Publisher: Elsevier]. *Forensic Science International: Genetics*, *4*(5), 281–291. https://doi.org/10.1016/j.fsigen.2009.10.013

Buckleton, J., Krawczak, M., & Weir, B. (2011). The interpretation of lineage markers in forensic DNA testing [Publisher: Elsevier]. *Forensic Science International: Genetics*, *5*(2), 78–83. https://doi.org/10.1016/j.fsigen.2011.01.010

Butler, J. M. (2005). *Forensic dna typing: Biology, technology, and genetics of str markers*. Elsevier.

Chin, J. M., Ribeiro, G., & Rairden, A. (2019). Open forensic science*. *Journal of Law and the Biosciences*, *6*(1), 255–288. https://doi.org/10.1093/jlb/lsz009

Coble, M. D., & Bright, J.-A. (2019). Probabilistic genotyping software: An overview. *Forensic Science International: Genetics*, *38*, 219–224. https://doi.org/10.1016/j.fsigen.2018.11.009

Edge, M. D., & Matthews, J. N. (2022). Open practices in our science and our courtrooms [Publisher: Elsevier]. *Trends in Genetics*, *38*(2), 113–115. https://doi.org/10.1016/j.tig.2021.09.010

Gill, P., Jeffreys, A. J., & Werrett, D. J. (1985). Forensic application of DNA 'fingerprints'. *Nature*, *318*(6046), 577–579. https://doi.org/10.1038/318577a0

Haddrill, P. R. (2021). Developments in forensic DNA analysis. *Emerging Topics in Life Sciences*, *5*(3), 381–393. https://doi.org/10.1042/ETLS20200304

Haller, B. C., & Messer, P. W. (2019). SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. *Molecular Biology and Evolution*, *36*(3), 632–637. https://doi.org/10.1093/molbev/msy228

Jobling, M. A., & Gill, P. (2004). Encoded evidence: DNA in forensic analysis. *Nature Reviews Genetics*, *5*(10), 739–751. https://doi.org/10.1038/nrg1455

Kapp, J. D., Green, R. E., & Shapiro, B. (2021). A Fast and Efficient Single-stranded Genomic Library Preparation Method Optimized for Ancient DNA. *Journal of Heredity*, *112*(3), 241–249. https://doi.org/10.1093/jhered/esab012

Kayser, M., & de Knijff, P. (2011). Improving human forensics through advances in genetics, genomics and molecular biology. *Nature Reviews Genetics*, *12*(3), 179–192. https://doi.org/10.1038/nrg2952

Larribe, F., & Fearnhead, P. (2011). On composite likelihoods in statistical genetics. *Statistica Sinica*, *21*(1), 43–69. Retrieved May 6, 2024, from http://www.jstor.org/stable/24309262

Lewontin, R. C., & Hartl, D. L. (1991). Population Genetics in Forensic DNA Typing [Publisher: American Association for the Advancement of Science]. *Science*, *254*(5039), 1745–1750. https://doi.org/10.1126/science.1845040

Li, N., & Stephens, M. (2003). Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics*, *165*(4), 2213–2233. https://doi.org/10.1093/genetics/165.4.2213

Matthews, J., Babaeianjelodar, M., Lorenz, S., Matthews, A., Njie, M., Adams, N., Krane, D., Goldthwaite, J., & Hughes, C. (2019). The right to confront your accusers: Opening the black box of forensic dna software. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 321–327. https://doi.org/10.1145/3306618.3314279

Nguyen, R., Kapp, J. D., Sacco, S., Myers, S. P., & Green, R. E. (2023). A computational approach for positive genetic identification and relatedness detection from low-coverage shotgun sequencing data. *Journal of Heredity*, *114*(5), 504–512. https://doi.org/10.1093/jhered/esad041

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, *12*(6), 443–451. https://doi.org/10.1038/nrg2986

Slatkin, M. (2008). Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, *9*(6), 477–485. https://doi.org/10.1038/nrg2361

Steel, M. (2016). Chapter 3: Tree shape and random discrete phylogenies. In *Phylogeny* (pp. 41–62). https://doi.org/10.1137/1.9781611974485.ch3

VanLiere, J. M., & Rosenberg, N. A. (2008). Mathematical properties of the r2 measure of linkage disequilibrium. *Theoretical Population Biology*, *74*(1), 130–137. https://doi.org/https://doi.org/10.1016/j.tpb.2008.05.006

Vohr, S. H., Buen Abad Najar, C. F., Shapiro, B., & Green, R. E. (2015). A method for positive forensic identification of samples from extremely low-coverage sequence data. *BMC Genomics*, *16*(1), 1034. https://doi.org/10.1186/s12864-015-2241-6

Williams, G. A., & Maskell, P. D. (2021). Embracing likelihood ratios and highlighting the principles of forensic interpretation. *Forensic Science International: Reports*, *3*, 100209. https://doi.org/https://doi.org/10.1016/j.fsir.2021.100209

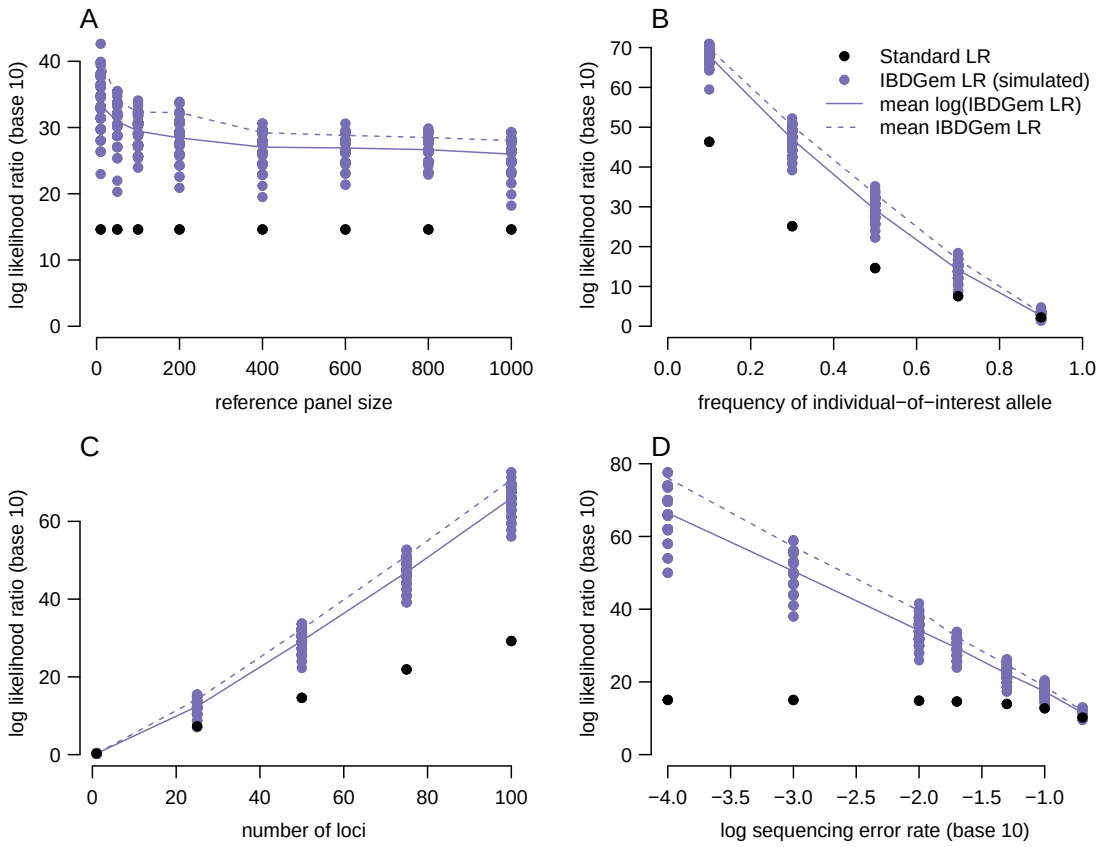

Figure 1: Dependence of the standard likelihood ratio ($P(D|U)$ in the denominator) and `IBDGem` likelihood ratio ($P(D|R)$ in the denominator) on parameters in the haploid model. Baseline parameters are reference panel size $n = 100$, number of loci $k = 50$, allele frequency $p = 1/2$, and sequencing error rate $\epsilon = 0.02$. One parameter is varied per panel: (A) reference panel size, (B) allele frequency (axis label is $1 - p$), (C) number of loci, (D) sequencing error rate. All likelihood ratios are displayed on a log scale (base 10). Standard forensic likelihood ratios are displayed in black, those from the `IBDGem` LD-mode approach in purple. There are multiple `IBDGem` dots per parameter value because LD-mode likelihood ratios depend on the specific individuals included in the reference panel. For `IBDGem` likelihood ratios, dots represent values from simulated reference panels (100 per parameter configuration), solid lines give the mean of the log likelihood ratios, and dashed lines give the log of the mean likelihood ratios.

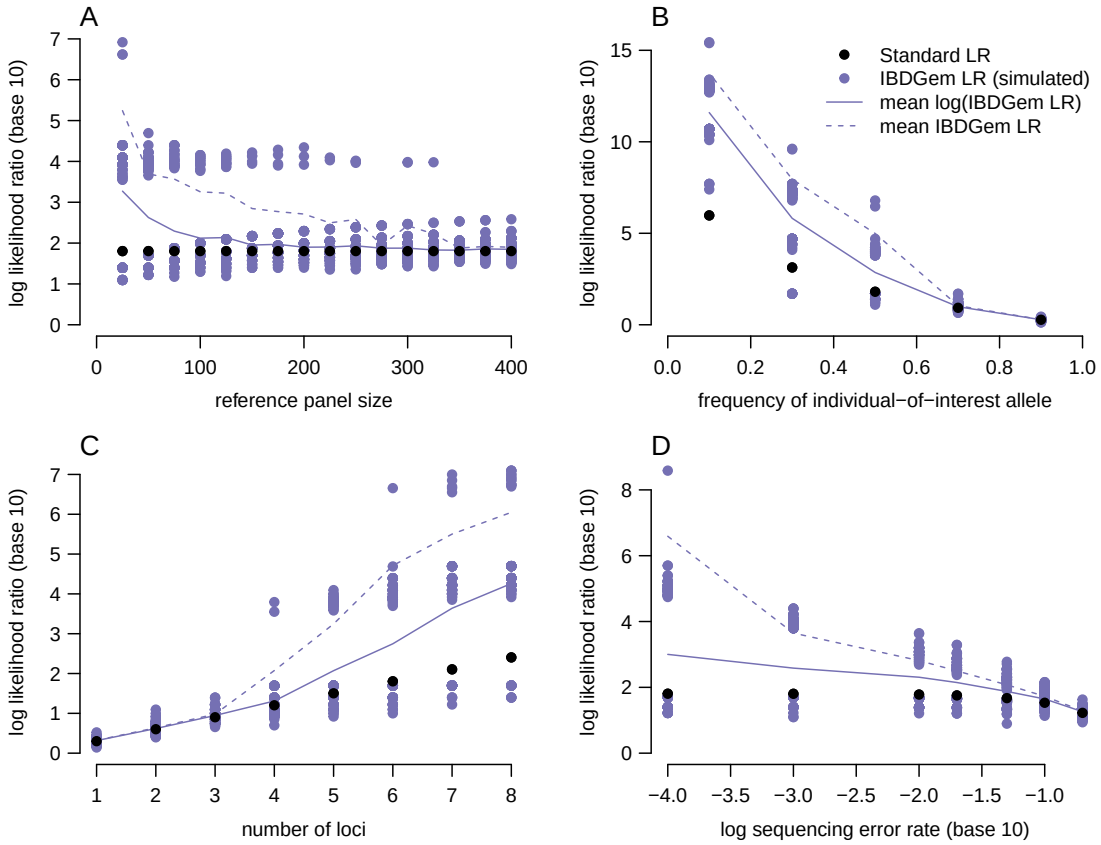

Figure 2: When the reference sample is large enough to include the individual-of-interest haplotype, standard and `IBDGem` likelihood ratios roughly match. Conventions are the same as in Figure 1, but here, the values of k are smaller. Where not otherwise specified on the horizontal axis, the number of loci is $k = 5$, the reference panel size is $n = 50$, the sequencing error probability is $\epsilon = .001$, and the allele frequency is $p = 0.5$. At low sequencing error rates, the log likelihood ratios become comparable approximately when $P(D|U) > 5/n$.

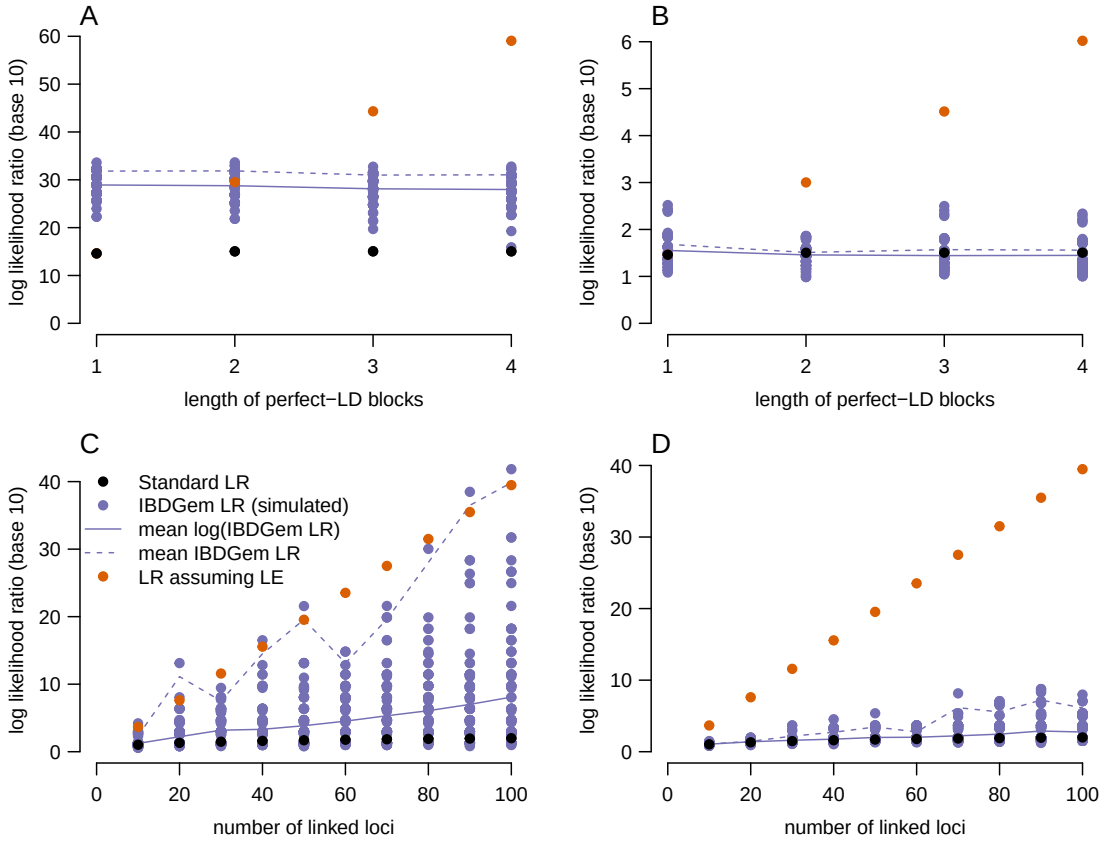

Figure 3: In the presence of linkage disequilibrium, likelihood ratios following the model for `IBDGem`'s non-LD mode will overstate evidence for identity. However, as in the other examples, LD mode only gives roughly correct likelihood ratios reliably when the reference panel size is considerably larger than one over the frequency of the haplotype of interest. Conventions are the same as in Figure 1, but the orange dots indicate likelihood ratios computed using the method of `IBDGem`'s non-LD mode. Panels A and B show results from perfect blocks of LD, as described in the text, and panels C and D show results from the caterpillar haplotype model exemplified in Table 2. In all panels, the sequencing error rate is $\epsilon = .02$, and in panels A and B, $p = 0.5$ and $n = 100$. In Panel A, $k = 50$ and is now interpreted as the number of sets of mutually independent loci, whereas in Panel B, $k = 5$. In Panel C, $n = 20$, whereas in Panel D, $n = 100$.

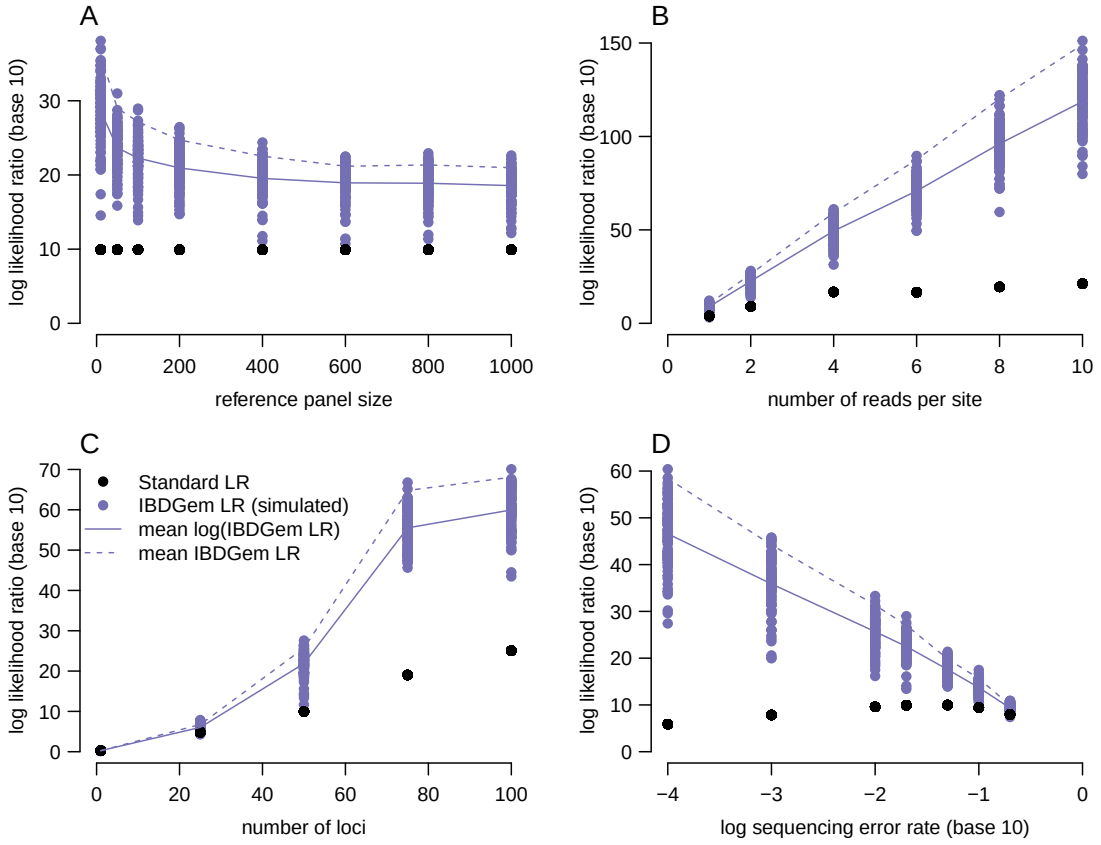

Figure 4: Dependence of the standard likelihood ratio ($P(D|U)$ in the denominator) and `IBDGem` likelihood ratio ($P(D|R)$ in the denominator) on parameters in a diploid setting with no linkage disequilibrium. These likelihood ratios were computed using `IBDGem` itself. Baseline parameters are reference panel size $n = 100$, number of loci $k = 50$, allele frequency $p = 1/2$, and sequencing error rate $\epsilon = 0.02$. One parameter is varied per panel: (A) reference panel size, (B) coverage depth, (C) number of loci, (D) sequencing error rate. All likelihood ratios are displayed on a log scale (base 10). Likelihood ratios computed using `IBDGem` non-LD mode are displayed in black, those from the `IBDGem` LD mode approach in purple. For LD-mode likelihood ratios, dots represent values from simulated reference panels (100 per parameter configuration), solid lines give the mean of the log likelihood ratios, and dashed lines give the log of the mean likelihood ratios.