Analysis of the Hamiltonian Monte Carlo genotyping algorithm on PROVEDIt mixtures including a novel precision benchmark

Mateusz Susik[1,2,*], Ivo F. Sbalzarini[2,3,4]

1 Biotype GmbH, Dresden, 01109, Germany
2 Technische Universität Dresden, Faculty of Computer Science, Dresden, 01187, Germany
3 Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, 01307, Germany
4 Center for Systems Biology Dresden, Dresden, 01307, Germany

*** m.susik [at] biotype.de**

Abstract

We provide an internal validation study of a recently published precise DNA mixture algorithm based on Hamiltonian Monte Carlo sampling [19]. We provide results for all 428 mixtures analysed by Riman et al. [17] and compare the results with two state-of-the-art software products: STRmix™ v2.6 and Euroformix v3.4.0. The comparison shows that the Hamiltonian Monte Carlo method provides reliable values of likelihood ratios (LRs) close to the other methods. We further propose a novel large-scale precision benchmark and quantify the precision of the Hamiltonian Monte Carlo method, indicating its improvements over existing solutions. Finally, we analyse the influence of the factors discussed by Buckleton et al. [5].

Keywords

Probabilistic genotyping, DNA mixtures, validation, HMC

1 Introduction

Robust interpretation of mixed DNA profiles is at the core of forensic practice. Algorithmic and computational support for this task has greatly evolved over time, and the most recent software generation of *fully continuous* methods can accurately resolve traces that were previously considered as unresolvable [6]. Given that every mistake of such algorithms could have drastic consequences (e.g. providing false evidence against a suspect), any software in practical use has to be thoroughly validated and trusted by the forensic community. Laboratories intending to adopt new technology should therefore have a way to assess the correctness and reproducibility of the results produced by a given genotyping software. A common way of validation is to check the results of the tested algorithms on a set of known DNA mixtures prepared under laboratory conditions [3, 7, 17]. Statistical analysis of the results can reveal the discriminatory power of the algorithms.

All laboratories are therefore recommended to validate the software they use before applying it in an investigation process [8, 10]. The effort required to adopt fully continuous methods is generally considered higher than with older techniques, especially

for software products that recommend estimation of laboratory-dependent hyper-parameters. Since most potential users do not have the possibility of preparing their own in-house validation mixtures, it should be possible to compare them without the need to perform a laboratory pipeline.

This requirement led to increasing efforts towards comparative evaluations of different methods, resulting in benchmark studies that offer convenient summaries of the discriminatory power of the available products [7, 17]. These studies are performed on standardized mixtures from the openly available PROVEDIt (Project Research Openness for Validation with Empirical Data) dataset [2], and thanks to the study by Riman et al. [17] it became possible to validate new inference methods by comparing them with previous solutions. Another reason for the importance of comparative studies is that they offer a possibility to estimate the effects of differences between solutions on the output likelihood ratios (LRs). This reason has been highlighted in the President's Council of Advisors on Science and Technology (PCAST) report [10].

Most comparative studies so far have focused on two software solutions: Euroformix [4] and STRmix™ [20]. The latter uses Random-Walk Markov Chain Monte Carlo (MCMC) to estimate the posterior probability distribution of a DNA mixture. The former was benchmarked using its maximum-likelihood estimations (MLE) method. Euroformix also offers an MCMC sampler (referred to as "Bayesian method" in that software), which, however, has never been considered in comparative benchmarks. The results of MCMC techniques are stochastic approximations, and the output LRs are expected to exhibit run-to-run variability. The magnitude of this variability, however, strongly depends on the design of the probabilistic genotyping model and on the sampling algorithm used, as it can be significantly reduced by appropriate choices [19]. This was achieved in conjunction with a Hamiltonian Monte Carlo (HMC) sampling algorithm.

So far, however, no comparative evaluation of HMC-based methods exists in the literature. Here, we therefore validate the HMC-based probabilistic genotyping method of Susik et al. [19] on the publicly available PROVEDIt mixtures [2], and we compare the results with those provided by STRmix™ v2.6 [17] and Euroformix v3.4.0. Doing so, we provide additional arguments for the robustness and correctness of the results obtained by the HMC method. We also provide a wider perspective for comparing of different software solutions. Given that the probabilistic model of the tested HMC method is shared with one of the comparison methods [20], the present results also shed light on the reproducibility of published studies. By comparing the results of our HMC method with those obtained by STRmix™ and Euroformix, we quantify the influence of the choice of the probabilistic model and of software implementation differences. We specifically analyse the effects of the choices mentioned by Buckleton et al. [5] when discussing the results of Riman et al. [17].

We introduce a new benchmark to test the precision of probabilistic genotyping algorithm on 23 randomly chosen mixtures from Riman et al. [17]. We hope that this enables quantitative comparisons and methodologically founded discussions about the precision of the available algorithms and the reproducibility of their results.

Finally, in Supplementary Material 3, we analyse the factors influencing the results of our model and how these factors correlate with the differences between our and alternative algorithms. Information content of the short tandem repeat mixtures influences the confidence of our algorithm corraborating results of other algorithms [7, 21].

2 Methods

We perform an internal validation study on publicly available mixtures that were also used in previous comparative studies [17]. This allows us to compare the results between studies. Riman et al. [17] reported results for one MCMC implementation of a log-normal model (STRmix™) and one MLE implementation of a gamma model (Euroformix).

2.1 PROVEDIt mixtures

The PROVEDIt database contains mixtures of varying contributors, numbers of contributors, contributor ratios, and mixture qualities [2]. In many cases the mixtures have been deteriorated by treatments with DNAse I, Fragmentase®, UV irradiation, sonification, or humic acid. The database contains mixtures analysed with different multiplexes, instrument models, and injection times.

We follow the study conducted by NIST [17] and perform our analysis on mixtures amplified with the GlobalFiler kit and analysed on a 3500 Genetic Analyzer with an injection time of 15 seconds. The filtered input files are used.

2.2 An update on Euroformix's results

Buckleton et al. [5] criticise the choices made by Riman et al. [17]. The critique is summarised in a list of 5 points. In order to be able to compare our solution with the only available larger STRmix™ benchmark, we follow the study of Riman et al., but take the following measures to provide a fair comparison:

- We re-analyse 2- and 3- contributor mixtures with Euroformix 3.4.0 with a set of hyperparameters we estimated. This responds to points 1 and 3 raised by Buckleton et al. [5].
- We compare different rare-allele models based on results obtained from the same likelihood ratio system [17]. This responds to point 4 raised by Buckleton et al. [5].
- Point 5 of Buckleton et al. [5] argues that diagnostics, such as convergence criteria, are of secondary interest. We respectfully disagree, as we have shown their importance in a previous study [19].

Point 2 raised by Buckleton et al. [5] criticises the use of a custom tool to pre-filter artefact peaks. We have independently verified (see Supplementary Material of Ref. [19]) that this tool has indeed filtered out some of the stutter peaks that are important for the analysis, so this is definitely an issue. However, the study of Riman et al. [17] is the only paper available that makes it possible to compare with STRmix results, which is why we nevertheless use it.

2.3 Hyperparameters settings

In order to reduce possible sources of differences between inference methods, we reuse the parameters calculated for STRmix™ in Ref. [17] with the exception of the parameters for the drop-in model. Since HMC uses the drop-in model from Euroformix, we reuse the published hyper-parameters from Euroformix for the drop-in model. The values of all hyperparameters are available in the original paper on the HMC algorithm [19]. We report sub-source LRs unless specified otherwise.

Euroformix v3.4.0 enables modelling of forward stutters. Still, no model for double-backward stutters is available yet. To follow these assumptions without biasing

the results, we use higher analytical thresholds (ATs), such that double-backward stutters are filtered out and will not distort the output. We determine these ATs by analysing the available single-source profiles (Supplementary Material 1) and removing the peaks of the contributors, backward, and forward stutters. We follow Riman et al. [17] and use the formula $\mathrm{AT} = \mu + k\sigma$, where μ and σ are the mean and standard deviation estimated from the remaining peak heights (including double-backward stutters). The resulting ATs are: 60, 80, 45, 75, and 100 RFU for the blue, green, yellow, red, and purple dyes, respectively. Under these ATs, no double-backward stutter was left in the 2- and 3-contributor mixtures. A single drop-in model has been estimated from these data with an estimated drop-in probability of 0.00073 and a lambda of 0.03846. An F_{ST} correction of 0.01 is used, and allele frequencies were normalised.

For Euroformix, we only provide results for 2- and 3-contributor mixtures. This is for reasons of analysis runtime, which were too large (regularly exceeding an hour per scenario) to be practical for the required numbers of repetitions when using Euroformix on 4-contributor mixtures. There are at least two reasons for this: (1) In order to provide results for different POIs (persons of interest) under the same hypothesis of the prosecutor, the inference process has to be run separately for each POI in Euroformix. In the case of HMC, the inference can be performed once, as the LRs for different POIs can re-use the same estimation of the posterior probability distribution. (2) In cases with few or no stutter peaks, the stutter model used by Euroformix might not converge. The inference might then never provide an answer. In such cases where the Euroformix model does not converge, we switch off the stutter model that caused the issue (firstly forward, and then if it did not help, also the backward model). Such a strategy causes an interesting situation in which we report results with different settings of stutter models within the same mixture. Still, we believe this is the approach that would imitate the best an approach adopted in a laboratory, as often the laboratories will test only one hypothesis of prosecutor. The list of the resulting stutter model settings is available in Supplementary Material 1.

2.4 Benchmark framework

All compared software products compute likelihood ratios [14]. This ratio expresses the strength of the evidence and is defined as:

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

where V is the observed electropherogram (EPG), and H_p and H_d are the hypotheses of the prosecutor and the defendant, respectively. All hypotheses in the benchmark assume that the contributors are of Caucasian ethnicity and come from the U.S. population genetic background. Therefore, NIST 1036-Caucasian allele frequencies are used. The hypotheses include the information about the number of contributors. In all cases, H_p assumes the inclusion of a certain person of interest (POI), while H_d assumes that all contributors are unknown. For each mixture, all possible scenarios with a true contributor in H_p are considered against the same number of scenarios with false POIs (i.e. non-contributors). For example, a 3-contributor mixture will be analysed with 6 different hypotheses: 3 with a true contributor and 3 with a false one. The false POIs has been chosen randomly from the genotypes of the NIST 1036-Caucasian sample [17]. The benchmark consists of 154 two-contributor mixtures, 147 three-contributor mixtures, and 127 four-contributor mixtures. We define:

- Scenario: A scenario is a setting where a mixture, H_p, and H_d are defined. For each mixture in the benchmarks, multiple scenarios are provided.

Table 1. Misclassification rate (MR), number of false positives (FP), and number of false negatives (FN) for the tested analysis methods in scenarios with different numbers of contributors.

	HMC			STRmix™			Euroformix		
Number of contributors	MR	FP	FN	MR	FP	FN	MR	FP	FN
2	0.002	0	1	0.008	1	4	0.023	13	1
3	0.014	2	10	0.016	2	12	0.066	57	1
4	0.021	11	10	0.024	14	10	-	-	-

- Person of interest (POI): The POI is the person (and his/her genotype) assumed present in the mixture under the prosecutor's hypothesis, but not under the defendant's hypothesis.
- True/false hypothesis: A true hypothesis is a prosecutor hypothesis H_p that includes a true POI. A H_p that does not include the true POI is a false hypothesis.
- Misclassification rate: The misclassification rate is the proportion of incorrectly classified scenarios. A scenario is correctly classified if the software output an LR larger than 1 for a true hypothesis or an LR lower than 1 for a false hypothesis.
- Precision: Precision measures reproducibility of the results when the same analysis run is repeated multiple times. The lower the run-to-run variation, the more precise the analysis. Precision depends on the scenario, hyperparameter values, background population frequencies and is quantified by the LR variability.
- Ban: A ban is a logarithmic unit for base 10 logarithms. 1 ban = 10, 2 bans = 100, 3 bans = 1000, etc.

3 Results

We first analyse the classification performance of the HMC method in comparison with the established approaches. Then, we introduce a novel large-scale benchmark for quantifying and comparing the precision of the different methods. Finally, we quantify differences that result from different rare-allele models used by the different methods.

3.1 Classification performance

The HMC method is characterised by a small misclassification rate, comparable to that of STRmix™ and lower than that of Euroformix (Table 1). For the 4-contributor mixtures the misclassification rates of the MCMC-based methods is under 2.5% (2.1% for HMC and 2.4% for STRmix™), while Euroformix misclassified 16.1% in the original study [17]. For Euroformix, the false positives dominate all misclassified cases with a total of 70 false positives and 2 false negatives across all 2- and 3-contributor mixtures. The MCMC-based methods more often misclassified a true hypothesis (21 false negatives vs. 13 false positives for HMC, 26 false negatives vs. 17 false positives for STRmix™), rendering them more conservative for usage in court cases.

We further analyse the misclassified scenarios in Figure 1. Most of the misclassified scenarios do not provide strong evidence, with the majority of $\log_{10}$ LR between -1 and 1: 67.6% (23/34) for HMC, 69.8% (30/43) for STRmix™, and 93.1% (67/72) for Euroformix.

Another popular metric for evaluating the quality of a classification method is the area under the Receiver Operating Characteristic (ROC) [11]. We therefore plot the ROC curves for the tested methods in Figure 2. All tested methods achieve

Figure 1. Histogram of misclassified examples. For each analysis method, results with $\log_{10}\mathrm{LR} < 0$ are false exclusions and results with $\log_{10}\mathrm{LR} > 0$ are false inclusions.

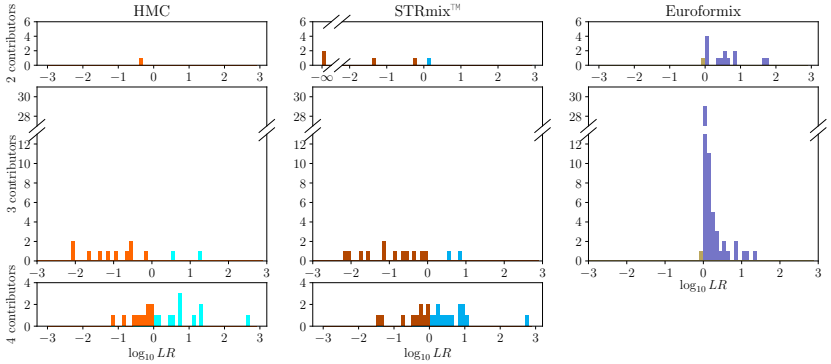
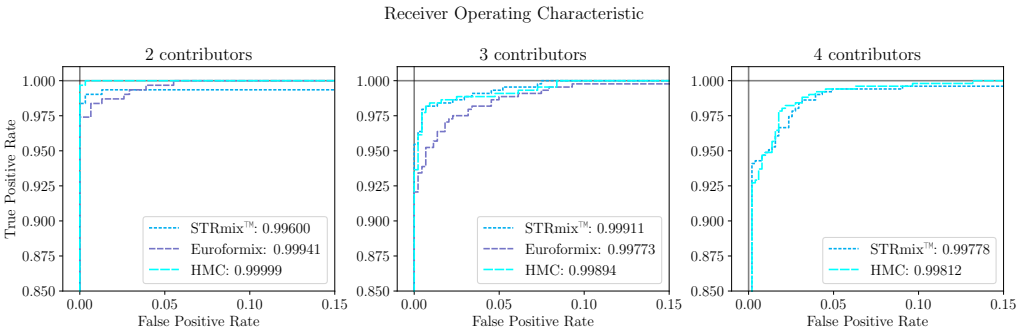


Figure 2. Comparison of Receiver Operating Characteristic (ROC) curves between the tested methods. The normalized areas under the ROC curves are given in the inset legends. They measure classification performance with a perfect classifier achieving 1.0.



performances with normalized areas under the curves close to the maximum possible value of 1.0, see inset legends. The slight difference between STRmix™ and HMC is mainly caused by two false negatives of STRmix™. Increasing the number of STRmix™ iterations in these scenarios increases the area under the ROC curve for STRmix™ from 0.99600 to 0.99994 for the 2-contributor mixtures. This is then perfectly comparable to the performance achieved by HMC.

The two false negatives responsible for this behaviour are two of the true hypotheses among the 2-contributor mixtures for which STRmix™ falsely outputs LR=0. Riman et al. [17] therefore decided to re-run the analysis on these cases with a larger number of iterations. Those repeated runs provided strong evidence for the inclusion of the respective POIs ($\log_{10}$ LR of 24.88 for the mixture marked as C02 and 19.66 for case H06). Closer analysis of those two scenarios reveals that the initial results provided support for the inclusion of the POI in all loci but one (D1S1656 in C02 and D3S1358 in H06). In those loci, the calculated LRs were 0. A biased analysis excluding those loci would provide strong evidence for including the POI (sub-sub-source $\log_{10}$ LR of 27.32 and 23.71, respectively). A skilled user should notice this imbalance between the estimated results for different loci and act accordingly, but the question remains why this happens. Algorithms that consider the genotype set as a parameter of the posterior (like STRmix™) will always struggle if the genotype of the POI is only included in genotype sets that are unlikely under the considered model. It is then possible that no Markov chain will ever explore those genotype sets when an insufficient number of iterations is performed. This does not seem to be specific to the present benchmark,

since an independent example of a mixture leading to this problem has been described by Lin et al. [13]. Buckleton et al. [5] noted:

> "We would ask the question: 'Does one want an *LR* system to allow 15,18 [POI genotype – our note] in this case?' If you answer 'yes' then the inevitable result is that a much higher false inclusion rate will occur.

We here show that this is actually avoidable, since the HMC method does not display a higher false inclusion rate than STRmix™ (Table 1), yet provides the correct answer in those two scenarios. Indeed, HMC provides strong inclusion for both scenarios with $\log_{10}$ LR of 24.94 for scenario C02 and 20.07 for H06. The problematic loci, however, correctly resulted in low sub-sub-source LRs: 0.0044 for D1S1656 in C02 and 0.000041 for D3S1358 in H06.

This suggests that the behaviour of STRmix™ is caused by a combination of the modelling decision to treat the genotype set as a modelled parameter and running an insufficient number of MCMC iterations. Those decisions were made due to the computational complexity of the inference and can be avoided by, for example, using GPU acceleration [19].

An interesting question is what happens when the POI's genotype is represented in the posterior estimate of one of the chains, but absent in the others, e.g.:

$$\sum_{S_j \in S_{\mathrm{POI}}} \hat{P}(V|S_j)_1 > 0 \;\; \wedge \sum_{S_j \in S_{\mathrm{POI}}} \hat{P}(V|S_j)_k = 0 \quad \forall_{k \in 2,\ldots,c}\,, \tag{2}$$

where S_{POI} is the set of all genotype sets S_j that match the POI, and $\hat{P}(V|S_j)_k$ denotes the deconvolution computed by the k-th Markov chain. In this case, intra-chain estimate for LR is 0 for all chains except one, but the final answer to the user will still indicate a positive LR. This is an example of the nuances one has to deal with if convergence of the estimated distribution over the discrete genotype sets is not rigorously checked. For this reason, the original study [17] should have repeated the calculations with an extended number of iterations whenever the initial analysis output an LR of 0 and only one (or few) LR estimation for single loci was 0.

3.2 Results on the verbal scale

Some sources recommend reporting results on a verbal scale that reflects the strength of the evidence [9, 16, 18]. This removes small, likely technical, differences between analysis methods and is easier to understand in court. However, it also introduces arbitrary thresholds by binning the possible values of LR. We analyse the effect this binning has on the results from the tested analysis methods on two verbal scales: the official SWGDAM scale [18] and an alternative scale suggested by ENFSI [9].

Both scales are compared in Table 2.

We present 10 contingency tables to compare the results on the verbal scale between HMC and the other softwares (STRmix™ or Euroformix). For each number of contributors, we present two tables: one where the hypothesis of the prosecutor is tested and one where the support for the defendant (i.e. support for negative evidence) is tested. As HMC and STRmix™ are based on similar models, the verbal scale classification of their results is similar. On the higher-resolution scale suggested by ENFSI for true hypotheses over 2-contributor mixtures, these two methods produce the same verbal classification in 96.43% of scenarios (297/308, Table 3). For true hypotheses over 3-contributor mixtures, they agree in 92.74% of scenarios (409/441, Table 6) and for 4-contributor mixtures in 88.19% of scenarios (448/508, Table 11). Thus, consensus becomes rarer with larger increasing number of contributors. It is also rarer when the false hypotheses are considered: 93.18% for 2-contributor scenarios

SWGDAM scale	Grade	$\log_{10}$LR	Scale suggested by ENFSI
Supports alternative	A	$< -\log_{10}(1.5)$	Supports alternative
Uninformative	B	$[-\log_{10}(1.5), \log_{10}(1.5))$	Does not support any proposition
Limited support	C	$[\log_{10}(1.5), 1)$	Weak support
	D	$[1, 2)$	Moderate support
Moderate support	E	$[2, 3)$	Moderately strong support
	F	$[3, 4)$	Strong support
Strong support	G	$[4, 6)$	Very strong support
Very strong support	H	≥ 6	Extremely strong support

Table 2. A comparison between the SWGDAM verbal scale and the ENFSI recommendation for reporting LRs. The provided $\log_{10}$LR ranges present the scenario of reporting the hypothesis of the prosecutor. As both of the scales do not define precisely how to deal with LRs between below 1, we set an LR of 1.5 as the border between grades A, B, and C.

STRmix™		HMC A	B	C	D	E	F	G	H
supports alternative proposition	A			1		1			2
does not support any proposition	B		1						
weak support	C			1					
moderate support	D				1				
moderately strong support	E				1	5	1		
strong support	F						2		
very strong support	G						1	8	1
extremely strong support	H							3	279

Table 3. Comparison of the results of HMC and STRmix™ when a verbal scale is used. True hypotheses, 2-contributor scenarios.

Euroformix	HMC A	B	C	D	E	F	G	H
A								
B	1		2					
C				1	1			
D					2	1	1	1
E					2	1	1	
F				1			2	2
G					1	1	8	14
H						1		264

Table 4. Comparison of the results of HMC and Euroformix when a verbal scale is used. True hypotheses, 2-contributor scenarios.

STRmix™ \ HMC	A	B	C	D	E	F	G	H
A								
B		1						
C			1					
D				4	2			
E			1		1			
F					2			
G				1	2	2	4	
H				1		5	5	276

Table 5. Comparison of results of HMC and STRmix™ when a verbal scale is used. False hypotheses, 2-contributor scenarios.

STRmix™ \ HMC	A	B	C	D	E	F	G	H
A	2							
B								
C		1	6	1				
D			10	11	1			
E			1	3	15			
F					13	8	1	
G				1	4	14	18	1
H					1	2	25	302

Table 7. Comparison of results of HMC and STRmix™ when a verbal scale is used. False hypotheses, 3-contributor scenarios.

STRmix™ \ HMC	A	B	C	D	E	F	G	H
A	7	2	1					
B	1		1					
C	1		8	2				
D			3	10	1	1		
E				3	10	1		
F						11	3	
G						3	30	3
H							6	333

Table 6. Comparison of results of HMC and STRmix™ when a verbal scale is used. True hypotheses, 3-contributor scenarios.

Euroformix \ HMC	A	B	C	D	E	F	G	H
A			1	1		1	1	5
B			1	3	2		3	4
C					4	1	1	3
D					2	3	1	16
E				1	1	1	3	16
F								15
G							1	10
H								207

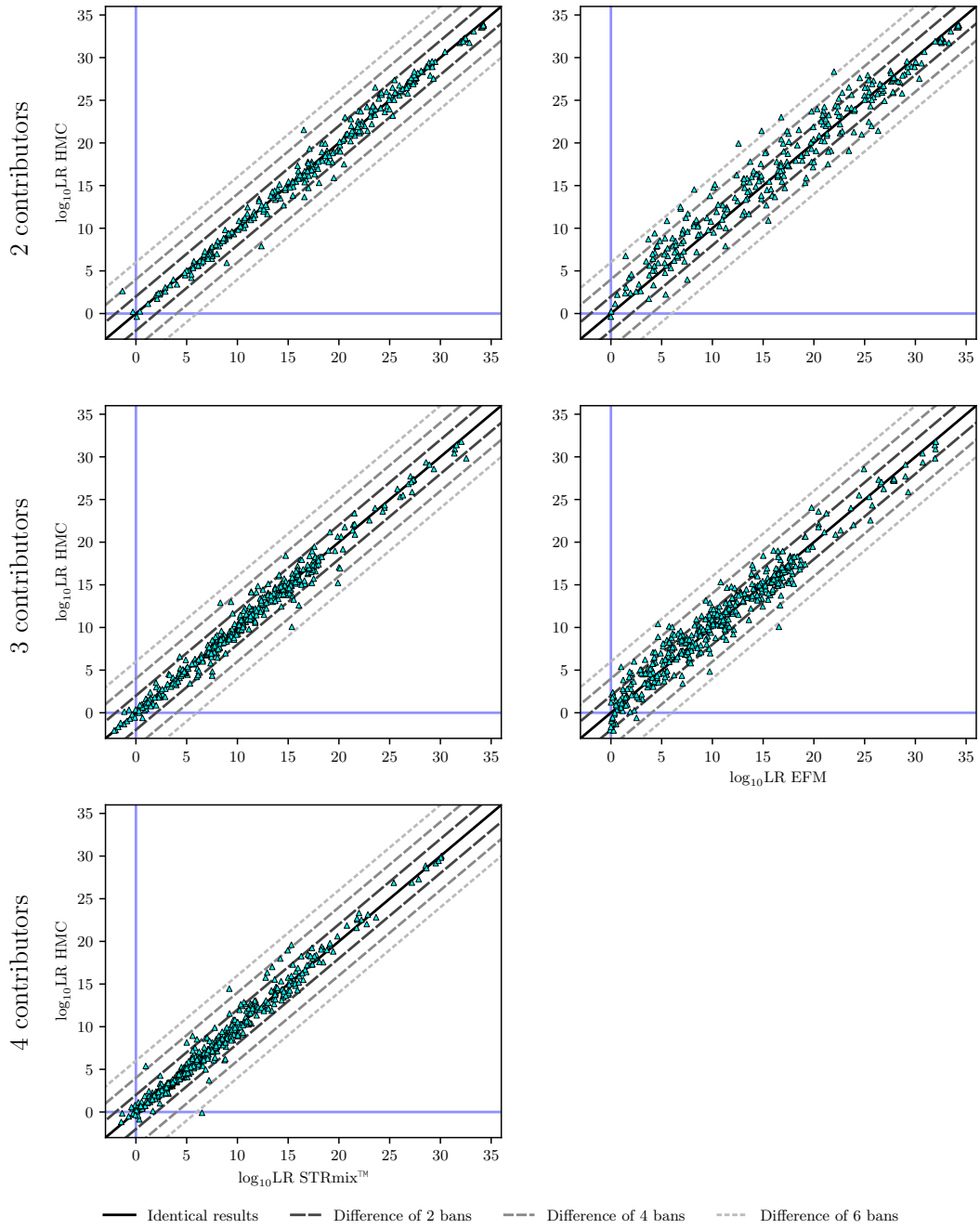
Table 8. Comparison of results of HMC and Euroformix when a verbal scale is used. False hypotheses, 2-contributor scenarios.

(287/308, Table 5), 82.09% (362/441, Table 7) for 3-contributor scenarios, and 70.67% (359/508, Table 12) for 4-contributor scenarios. This is likely because both methods become less precise when LRs are low (cf. Fig. 7) and the genotype set of the suspect is unlikely in at least some of the loci.

The discrepancy between HMC and Euroformix is larger, which may be explained by the different probabilistic models, ATs, and optimization formulations (Euroformix has been run in MLE mode, though an MCMC implementation is available, in order to be able to directly compare the results with Riman et al. [17] who also used the MLE mode). In cases where there is a difference in the reported grades, HMC tends to provide stronger support in scenarios with true POIs. This is the case for 9.42% (29/308, Table 4) of all 2-contributor scenarios and 12.47% (55/441, Table 9) of 3-contributor scenarios. In the 3-contributor cases, Euroformix provides a stronger verbal equivalent in 33 (7.48%) of cases.

The largest discrepancies can be observed between Euroformix and HMC under the negative hypotheses. In many of these cases, HMC provides stronger negative evidence. This is the case in 31.82% (98/308, Table 8) of 2-contributor scenarios and 81.86% (361/441, Table 10) of 3-contributor scenarios. When compared with the original benchmark [17], HMC provides stronger negative evidence in 94.29% (479/508) of the 4-contributor scenarios. The opposite situation, where Euroformix provides stronger negative verbal equivalent, is rare, with only one scenario among 2-contributor ones and no scenarios among 3-contributors ones.

Figure 3. Comparison of the computed LRs for all tested true hypotheses between HMC (y axes) and the other tested methods (x axes, STRmix™ and Euroformix (EFM)). The two 2-contributor scenarios for which STRmix™ outputs an LR of 0 are omitted from the plot.



Euroformix \ HMC	A	B	C	D	E	F	G	H
A								
B	5			2				
C	4	1	8	5	1			
D		1	4	6	5	2	3	
E	1		1	2	2	1	3	1
F				1	1	6	10	1
G						6	19	21
H						1	5	312

Table 9. Comparison of results of HMC and Euroformix when a verbal scale is used. True hypotheses, 3-contributor scenarios.

Euroformix \ HMC	A	B	C	D	E	F	G	H
A	2		3	4	1	3		7
B		1	6	11	26	12	14	27
C			2	6	7	5	12	44
D						3	13	50
E						1	3	46
F							1	26
G								30
H								75

Table 10. Comparison of results of HMC and Euroformix when a verbal scale is used. False hypotheses, 3-contributor scenarios.

STRmix™ \ HMC	A	B	C	D	E	F	G	H
A	3	2	2					
B	1	2	3					
C	2		11	3			1	
D		1		20	4			
E				7	27	2	1	
F					7	30	1	
G						5	63	9
H		1				1	7	292

Table 11. Comparison of results of HMC and STRmix™ when a verbal scale is used. True hypotheses, 4-contributor scenarios.

STRmix™ \ HMC	A	B	C	D	E	F	G	H
A	9	2		1				
B		4						
C		2	34	7				
D			8	32	15	1	3	1
E				12	34	13	2	1
F				2	19	20	3	2
G					4	19	27	16
H							16	199

Table 12. Comparison of results of HMC and STRmix™ when a verbal scale is used. False hypotheses, 4-contributor scenarios.

3.3 Results on the numerical scale

In order to provide a refined comparison, we plot in Figure 3 all of the computed LRs when true hypotheses were considered. This again highlights that HMC and STRmix™ provide similar results. In 95.54% (1201/1257) of all scenarios the difference between the LRs computed by these two methods is smaller than 2 bans. In 98.97% (1244/1257) of cases the difference is smaller than 4 bans, and only in 3 scenarios the difference is larger than 6 bans. These extreme cases are CO2 and H06, which were already discussed in Ref. [17] and in Subsection 3.1, as well as E05_RD14−0003−33_34_35_36−1;4;4;1−M2a−0.63GF (E05). In scenario E05 with Contributor 33 the POI, HMC computes $\log_{10} \mathrm{LR} = -0.0852$, while STRmix™ outputs 6.5106. We analyse this case in detail in Supplementary Material 2.

Larger discrepancies are observed between HMC and Euroformix, with 22.43% (168/749) of scenarios with a true contributor in H_p resulting in LRs that differ by more than 2 bans, 4.01% (30/749) by more than four bans, and 0.8% (6/749) by more than six bans. These differences are visualised again as $\log_{10} \mathrm{LR}$ histograms in Figure 4.

For the false hypotheses, a slight trend toward stronger negative evidence can be observed in the results of STRmix™ (Figure 5). There are 61 2-contributor scenarios where $\log_{10} \mathrm{LR}$ is lower when STRmix™ is used (and 11 scenarios with the opposite being true), 244 3-contributor scenarios (57 otherwise), and 262 4-contributor scenarios (203 otherwise). We note, however, that the number of scenarios where STRmix™

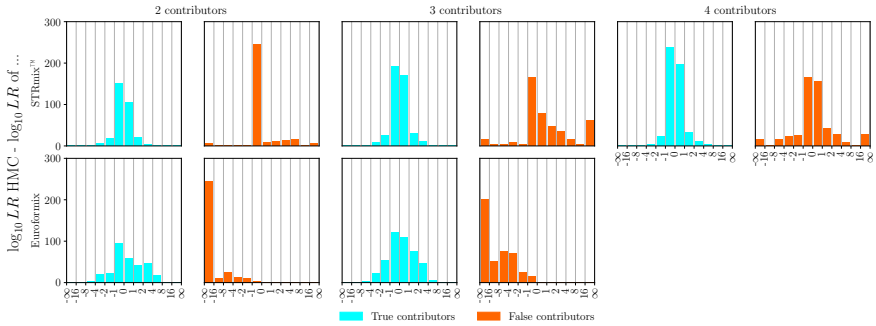


Figure 4. Histograms of the differences in $\log_{10}$ LR between HMC, STRmix™, and Euroformix for both true (cyan) and false (orange) hypotheses.

resorted to trivial rejection of the hypothesis (i.e. LR=0) is larger than for HMC: 202 3-contributor scenarios for STRmix™ vs. 155 for HMC; 71 4-contributor scenarios for STRmix™ vs. 54 for HMC. Two algorithmic factors influence the occurrence of inference results with LR=0. First, the way the considered genotype sets are pruned by removing unlikely ones may lead the algorithm to ignore the genotype of the POI. The second factor is the general behaviour of MCMC methods that consider a single genotype set at a time, as discussed in Subsection 3.1.

When comparing HMC with Euroformix, we observe that Euroformix tends to not provide strong negative evidence (Figure 6). The largest differences (if HMC did not result in LR=0) are observed for the scenarios:

- 2-contributor: mixture B04_RD14—0003—42_43—1;9—M2U60—0.15GF, POI is C06C_Cauc - Euroformix $\log_{10}$ LR = −1.19, HMC −23.9.
- 3-contributor: mixture B01_RD14—0003—36_37_38—1;2;1—M2a—0.5GF, POI is PT84183_AA - Euroformix $\log_{10}$ LR = −8.21, HMC −164.73.

However, this would be important only in cases in which the expert has to report the strength of the negative evidence.

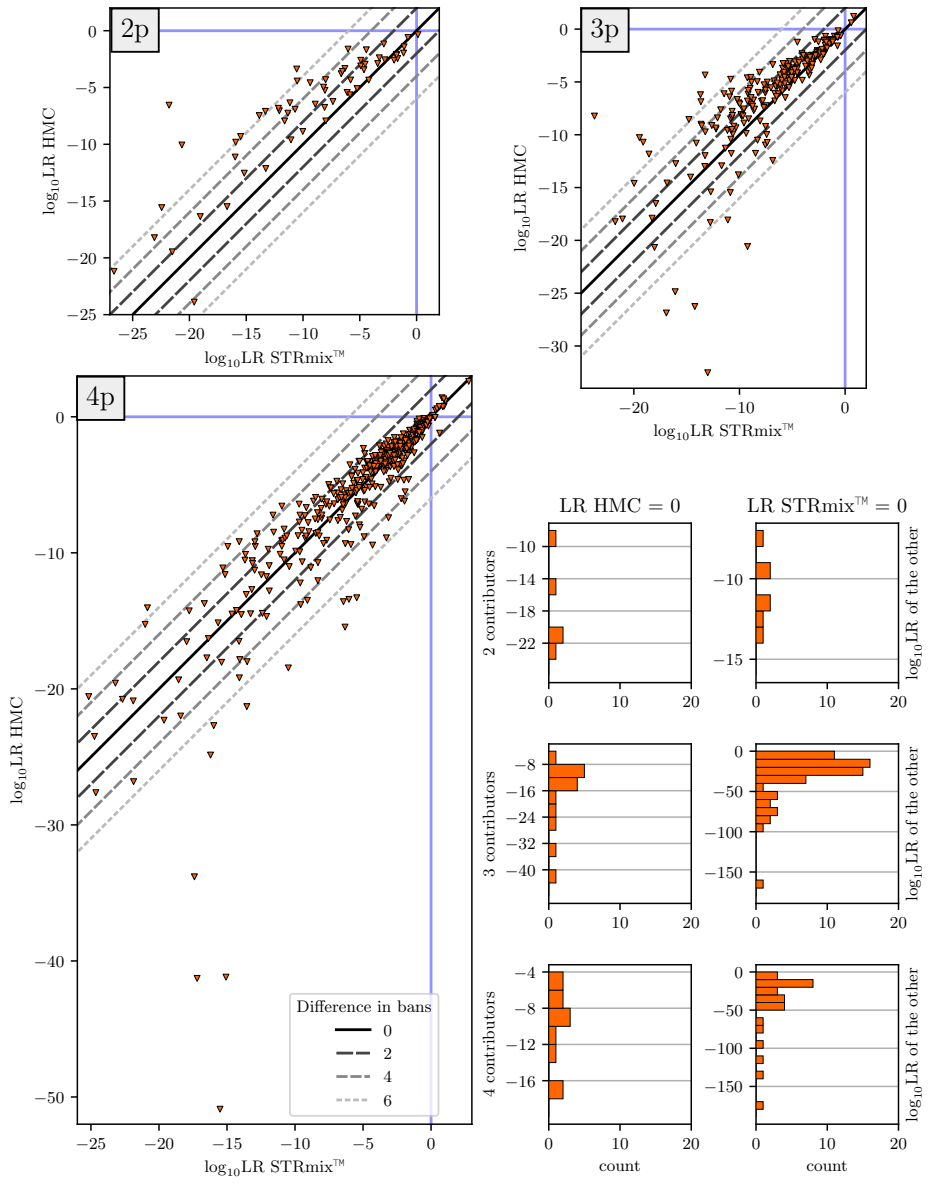
3.4 A new precision benchmark

Previously published studies [7, 17] compared methods by running them once on every scenario. An exception was the study by Riman et al. [17], where the computations for 8 scenarios were repeated twice. These were the two scenarios described in Subsecion 3.1 and 6 scenarios in which the convergence metric of STRmix™ remained above the threshold of 1.2 suggested by the authors of the algorithm. In those 6 scenarios the average difference between the two runs was 0.282 bans [17]. The inverse of the run-to-run variability over multiple runs with identical hyper-parametrisation on the same EPG data for the same hypotheses defines the *precision* of a probabilistic genotyping method. Precision has to be considered in addition to accuracy [15].

In order to quantify the precision of the HMC method, we here introduce a new benchmark that expands upon earlier results [19]. Because precision studies of other methods are not available, we cannot provide a comparison. In order to keep the benchmark tractable, we select 23 mixtures uniformly at random (7 2-contributor, 6 3-contributor, and 10 4-contributor mixtures)[1]. For each mixture, we run HMC inference for all scenarios provided in Ref. [17], resulting in a total of 120 scenarios. We run each scenario 11 times, resulting in 1320 runs, and display the resulting LRs in

[1]The list of mixtures and the detailed results are available in Supplementary Material 1.

Figure 5. Comparison of HMC and STRmix™ for all false hypotheses. The plots marked with 2p, 3p, and 4p present scenarios with two, three, and four contributors, respectively, for which both methods output non-zero LR. The histograms at the bottom-right show the distributions of LRs from the respective other method in scenarios for which exactly one of the methods outputs LR=0. In addition, there are 236 2-contributor, 140 3-contributor, and 43 4-contributor scenarios were both methods output LR=0.



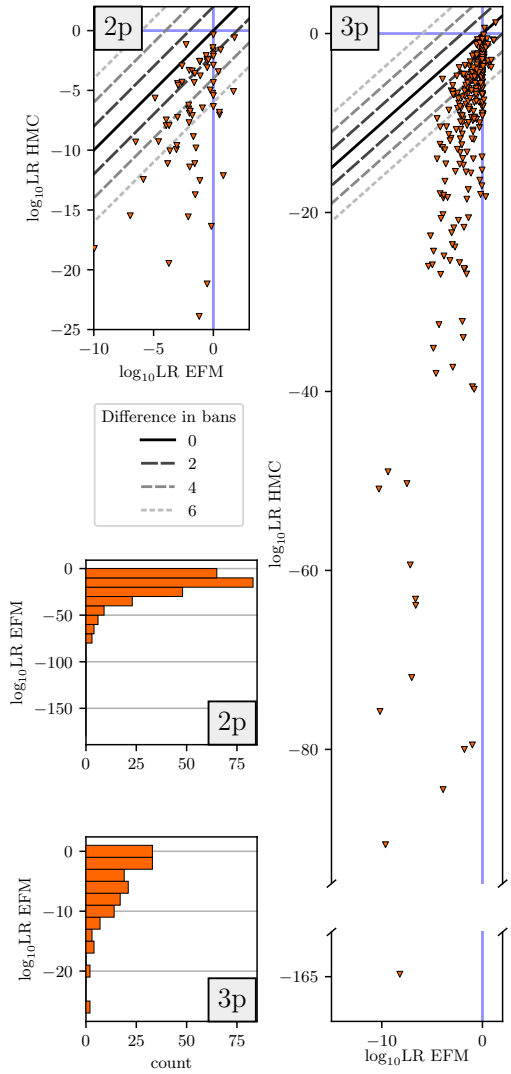


Figure 6. Comparison of HMC and Euroformix (EFM) for all false hypotheses. The plots marked with 2p and 3p present scenarios with two and three contributors, respectively, for which HMC outputs non-zero LR. The histograms at the bottom-left show the distributions of LRs from Euroformix in scenarios for which HMC outputs LR=0. Euroformix outputs non-zero LR for all scenarios in the benchmark.

Figure 7. In one of the scenarios, HMC misclassifies the hypothesis: the 4-contributor scenario G06 results in $\log_{10}$ LR between -0.937 and -0.637 when Contributor 33 is the POI.

We quantify the precision of HMC inference by the standard deviation of the $\log_{10}$ LR and their min–max span (Figure 8). Supporting the hypothesis set out in the Supplementary Material of Ref. [19], we generally observe lower precision when considering false hypotheses. The average $\pm$ standard deviation of the standard deviations of the $\log_{10}$ LR across all 2-contributor scenarios was 0.0415 ± 0.0372, 0.0335 ± 0.0293 for 3-contributor scenarios, and 0.0213 ± 0.0156 for 4-contributor scenarios. The outlier among 3-contributor scenarios with false hypotheses is the case with the lowest estimated $\log_{10}$ LR (excluding cases for which all 11 runs result in LR=0): H04 3p with ZT80028_Hisp as the POI. In this case, all per-run $\log_{10}$ LR < -60. Taken together, these results confirm the high precision of the HMC method [19]. In all considered cases, the min–max span of the LRs in scenarios with true hypotheses was smaller than 0.5 bans with standard deviation below 0.1 bans.

Next, we count how often the result on the ENFSI verbal scale (cf. Subsection 3.2) changed between repeated runs on the same scenario. Considering the grades assigned to the prosecutor's hypotheses, 48 scenarios are classified as *extreme strong support*, 9 scenarios as *very strong support*, and 6 as *strong support*, 4 as *moderately strong support*, and 3 as *moderate support* unanimously in all 11 runs. The only case where the grades differed between runs of the HMC method was mixture D03 4p with Contributor 47 as POI: 9 runs classified it as *strong support*, but 2 runs as *moderately strong support*. None of the benchmark scenarios was classified as *weak support*. One scenario (A02 4p with false contributor ZT80070_Hisp) unanimously resulted in LRs in all 11 runs that do not support any hypothesis. All other scenarios, including the false negative in one scenario from the G06 4p mixture, unanimously provided support for the alternative hypothesis in all 11 runs. We conclude that a situation in which HMC provides different verbal equivalents when run multiple times is rare, with only one (out of 120) case observed in our benchmark.

Finally, we also measure the computer time spent by the HMC algorithm on each of the considered scenarios (Figure 9). Every 3-contributor case was solved in under 11 minutes, and every 4-contributor case in under 1 hour and 3 minutes. These times include the compilation of the computational graph by the Accelerated Linear Algebra library [1] used in HMC.

3.5 Differences between rare-allele models

Buckleton et al. [5] suggested that different rare allele treatment might distort the results of method comparisons (*"There was no compensation for the different rare allele models used by each software."* [5]). We therefore quantify the effect of the rare-allele model by recalculating all LRs of HMC with the posterior mean allele frequency model [22], reusing the deconvolution provided by the inference model. We then compare the results with LRs obtained using the 5/2N rare-allele model [23].

The results in Figure 10 show that different rare-allele models lead to differences in $\log_{10}$ LR that are larger that the run-to-run variability and therefore should be considered significant. The data points for true contributors can be visually clustered into three categories: POIs with no rare alleles present in their genotypes, POIs with one rare allele, and POIs with two rare alleles. The majority of points from these clusters lie around the horizontal line at 0 for cases with no rare alleles, 0.22 for cases with one rare allele, and 0.44 for cases with two rare alleles. Points away from the respective line of their cluster are often caused by the rare allele(s) dropping out from the mixture. This was often the case especially for Contributor 40's rare allele in the FGA locus and for Contributor 50's rare allele in the TPOX locus. The corresponding

Figure 7. Run-to-run variability of HMC. Each small plot shows the $\log_{10}$ LR computed across 11 runs for one of the 120 scenarios (see Supplementary Material 1 for the list of scenarios). True hypotheses are plotted in cyan, false ones in orange.

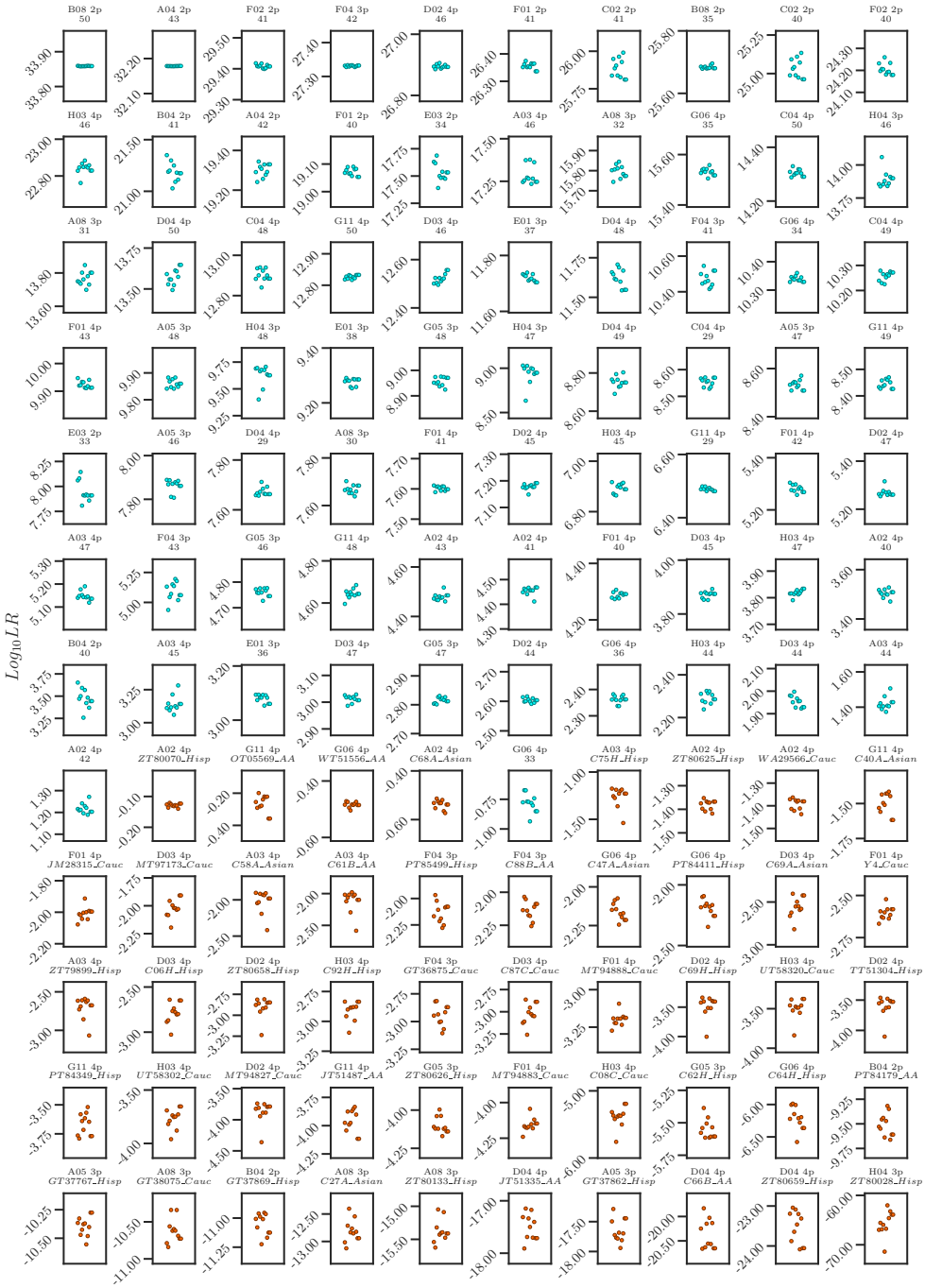


Figure 8. Standard deviations and min–max spans of the $\log_{10}$ LR values in all scenarios of the precision benchmark across 11 repeated runs of the HMC method for each scenario. The presented results also include Samples 1b and 2b from Ref. [19]. Scenarios with LR=0 are omitted.

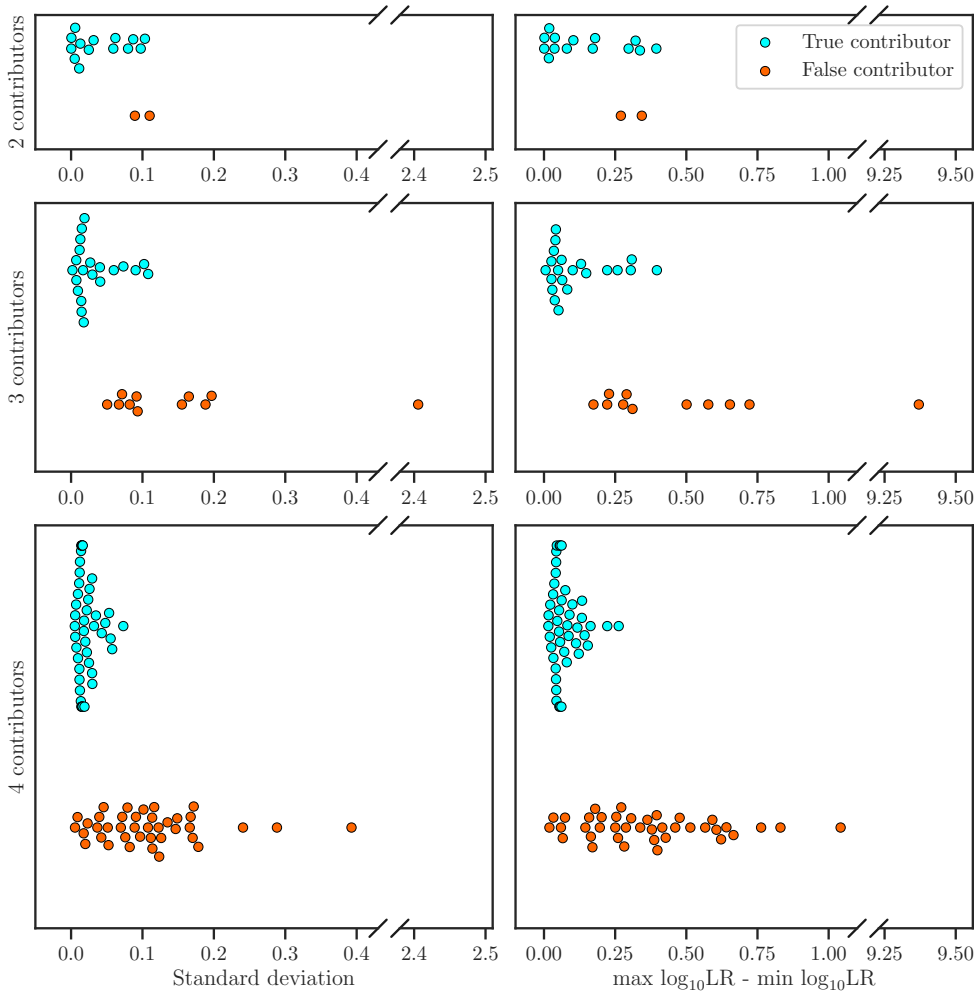


Figure 9. Execution times of the HMC method. Each dot represents a single run of one of the scenarios from the present benchmark. All runs were performed on a single Nvidia Tesla T4 GPU.

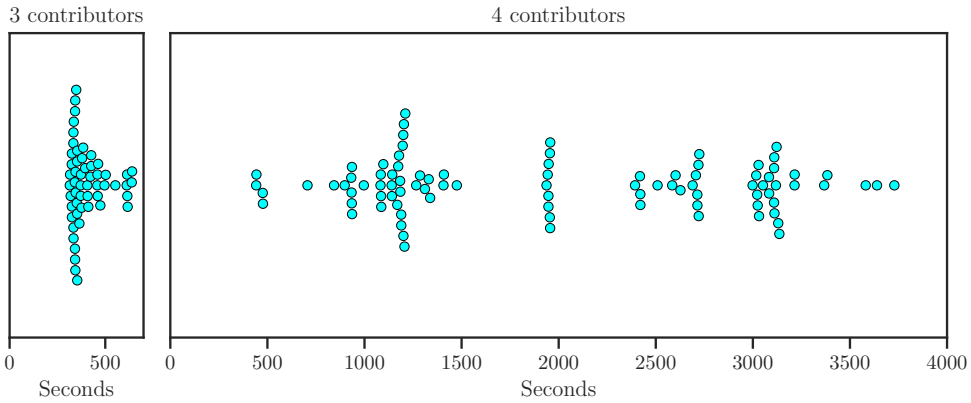
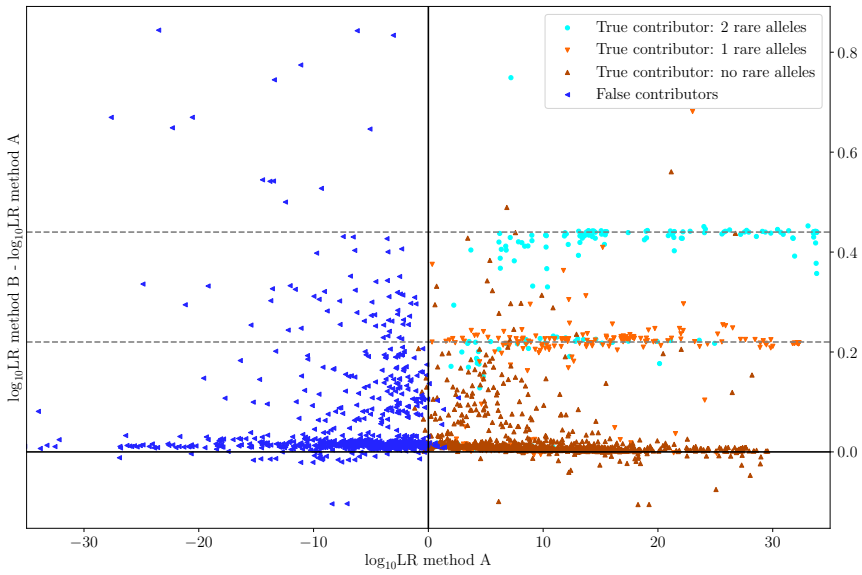


Figure 10. Difference between the $\log_{10}$ LR obtained by HMC with two different rare-allele models: the 5/2N model (Method A) and the posterior mean allele frequency model (Method B). Colours denote whether the POI is a true contributor and, if so, how many rare alleles are present in his/her genotype. The dashed horizontal lines are at differences of 0.22 and 0.44. The true POIs with rare alleles cluster around those lines.



markers are located towards the right ends of the dyes, such that these peaks risk to drop out in degraded mixtures.

When there are rare alleles present in both the POI's genotype and the mixture, the $\log_{10}$ LR difference between the 5/2N model and the posterior allele frequency model will depend on the value of N in the 5/2N formula and therefore on the population allele frequencies. For different values of N, the data points will cluster around horizontal lines at different values.

3.6 Differences between two settings of Euroformix

As discussed in Subsection 2.3, our benchmark setting required raising the analytical thresholds. This must be done in order to avoid double-backward stutters that are not modelled by Euroformix. In case any double-backward stutter was present in a profile, it could significantly distort the result. When the analytical threshold is raised, we expect the results to express lower confidence (i.e. lower LRs for true contributors, higher LRs for false contributors). We therefore compare the results obtained by Riman et al. [17] with Euroformix version 2.1.0 at lower ATs with the results we obtained here with Euroformix version 3.4.0 and higher ATs. The comparison is presented in Figure 11. As expected, higher ATs mostly lead to lower strengths of evidence. The exceptions are the 2-contributor mixtures with forward stutter peaks. Since these stutters are not modelled in Euroformix 2.1.0, the resulting deconvolution must be sub-optimal. This behaviour has independently been observed by Cheng et al. [7]. While the use of higher ATs might be perceived as a disadvantage for Euroformix, we note that this was required because of the limitations of the model. An alternative benchmark design would involve all the methods tested with higher ATs.

4 Conclusions

We provided a large-scale validation study of the HMC method for probabilistic genotyping [19] and compared it with the current state of the art. The results revealed that the HMC method achieves state-of-the-art classification performance at high precision (i.e., low run-to-run variability). In misclassified scenarios, HMC did not provide strong false evidence, but demonstrated uncertainty.

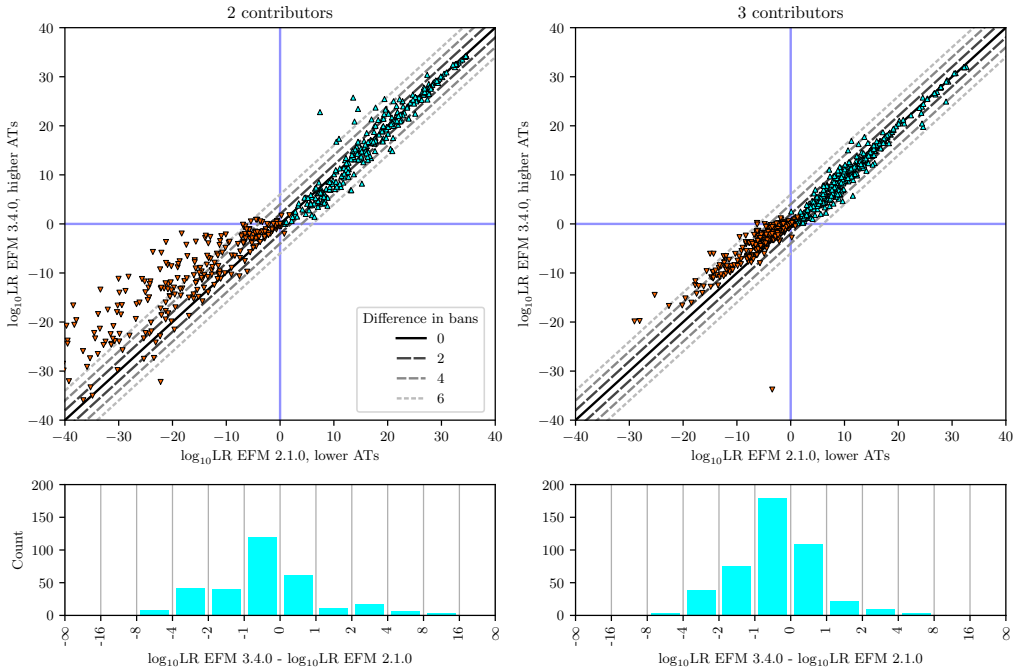
The classification results obtained by HMC were similar to those obtained using the popular software STRmix™. The similarity can be explained by the fact that the probabilistic genotyping model used by HMC is based on the work by Taylor et al. [20]. Compared to Euroformix, HMC exhibited a lower misclassification rate, a significantly better ability to provide negative evidence, and a slightly higher area under the ROC curve for 3-contributor mixtures.

We further introduced a new precision benchmark consisting of 120 scenarios that can serve as a baseline for future methods development and comparison. This benchmark provided additional results corroborating the high precision of the HMC method.

Finally, in Supplementary Material 3, we analysed the factors influencing the strength of the evidence provided by the HMC method and quantified the correlations between those factors and the differences of the results from different methods. This provided additional arguments that the strength of the evidence decreases with:

- decreasing total amount of DNA material in the mixture,
- decreasing amount of DNA material coming from the POI,
- decreased weight of the contributor, and

Figure 11. Difference between the results obtained with versions 2.1.0 and 3.4.0 of Euroformix (EFM) at different analytical thresholds (AT). **Top row:** Scatter plots of the scenarios for 2-contributor (left) and 3-contributor (right) mixtures. True hypotheses are in cyan, false ones in orange. Scenarios where any of the software versions computed $\log_{10}$ LR below -40 are omitted. **Bottom row:** Histograms of the differences for all scenarios with true POI in 2-contributor (left) and 3-contributor (right) mixtures.



- decreased sample quality (i.e., increased "quality index").

While this study focused on understanding the differences between software products, all of the compared methods are characterised by high accuracy and a healthy ability to provide a lower strength of evidence when the information content in the electropherogram is lower.

An interesting extension of the benchmarks would be to scenarios in which the (simulated) relatives of the true contributors are considered as POIs. Such experiments would be more challenging for probabilistic genotyping methods. Initial analyses, though without comparing different models, have already been performed [12].

Acknowledgments

We would like to thank Dr. João Nunes and Dr. Suzanne Walz (both Biotype GmbH, Dresden) for their insights on the influence of the treatment type on the compromised whole blood mixtures and Dr. Øyvind Bleka (Oslo University Hospital, Oslo) for his advice on proper usage of Euroformix software and input on the manuscript content.

Declaration of competing interest

Mateusz Susik is a research scientist employed at Biotype GmbH. He is the author of the HMC method validated in this study. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. XLA: Optimizing compiler for machine learning. `https://www.tensorflow.org/xla`. Accessed: 2022-07-13.

2. L. E. Alfonse, A. D. Garrett, D. S. Lun, K. R. Duffy, and C. M. Grgicak. A large-scale dataset of single and mixed-source short tandem repeat profiles to inform human identification strategies: PROVEDIt. *Forensic Science International: Genetics*, 32:62–70, Jan. 2018.

3. D. W. Bauer, N. Butt, J. M. Hornyak, and M. W. Perlin. Validating TrueAllele ® interpretation of DNA mixtures containing up to ten unknown contributors. *Journal of Forensic Sciences*, 65(2):380–398, Oct. 2019.

4. Ø. Bleka, G. Storvik, and P. Gill. EuroForMix: An open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts. *Forensic Science International: Genetics*, 21:35–44, Mar. 2016.

5. J. Buckleton, J.-A. Bright, D. Taylor, R. Wivell, Ø. Bleka, P. Gill, C. Benschop, B. Budowle, and M. Coble. Re: Riman et al. examining performance and likelihood ratios for two likelihood ratio systems using the PROVEDIt dataset. *Forensic Science International: Genetics*, page 102709, Apr. 2022.

6. J. S. Buckleton, J.-A. Bright, and D. Taylor, editors. *Forensic DNA Evidence Interpretation*. CRC Press, paperback edition, 3 2021.

7. K. Cheng, Ø. Bleka, P. Gill, J. Curran, J.-A. Bright, D. Taylor, and J. Buckleton. A comparison of likelihood ratios obtained from EuroForMix and STRmix™. *Journal of Forensic Sciences*, Sept. 2021.

8. M. Coble, J. Buckleton, J. Butler, T. Egeland, R. Fimmers, P. Gill, L. Gusmão, B. Guttman, M. Krawczak, N. Morling, W. Parson, N. Pinto, P. Schneider, S. Sherry, S. Willuweit, and M. Prinz. DNA commission of the international society for forensic genetics: Recommendations on the validation of software programs performing biostatistical calculations for forensic genetics applications. *Forensic Science International: Genetics*, 25:191–197, Nov. 2016.

9. European Network of Forensic Science Institutes. Enfsi guideline for evaluative reporting in forensic science: Strengthening the evaluation of forensic results across europe (steofrae. 2015.

10. Executive Office of the President, President's Council of Advisors on Science and Technology. *Report to the President - Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-comparison Methods*. CreateSpace Independent Publishing Platform, 2016.

11. J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, Apr. 1982.

12. H. Kelly, M. Coble, M. Kruijver, R. Wivell, and J.-A. Bright. Exploring likelihood ratios assigned for siblings of the true mixture contributor as an alternate contributor. *Journal of Forensic Sciences*, 67(3):1167–1175, Feb. 2022.

13. M.-H. Lin, J.-A. Bright, S. N. Pugh, and J. S. Buckleton. The interpretation of mixed DNA profiles from a mother, father, and child trio. *Forensic Science International: Genetics*, 44:102175, Jan. 2020.

14. D. V. Lindley. A problem in forensic science. *Biometrika*, 64(2):207–213, Aug. 1977.

15. G. S. Morrison. Special issue on measuring and reporting the precision of forensic likelihood ratios: Introduction to the debate. *Science & Justice*, 56(5):371–373, sep 2016.

16. A. Nordgaard, R. Ansell, W. Drotz, and L. Jaeger. Scale of conclusions for the value of evidence. *Law, Probability and Risk*, 11(1):1–24, Oct. 2011.

17. S. Riman, H. Iyer, and P. M. Vallone. Examining performance and likelihood ratios for two likelihood ratio systems using the PROVEDIt dataset. *PLOS ONE*, 16(9):e0256714, Sept. 2021.

18. Scientific Working Group - DNA Analysis Methods. Recommendations of the swgdam ad hoc working group on genotyping results reported as likelihood ratios. 2018.

19. M. Susik, H. Schönborn, and I. F. Sbalzarini. Hamiltonian monte carlo with strict convergence criteria reduces run-to-run variability in forensic DNA mixture deconvolution. *Forensic Science International: Genetics*, 60:102744, sep 2022.

20. D. Taylor, J.-A. Bright, and J. Buckleton. The interpretation of single source and mixed DNA profiles. *Forensic Science International: Genetics*, 7(5):516–528, Sept. 2013.

21. D. Taylor and J. Buckleton. Do low template DNA profiles have useful quantitative data? *Forensic Science International: Genetics*, 16:13–16, may 2015.

22. C. M. Triggs and J. M. Curran. The sensitivity of the bayesian HPD method to the choice of prior. *Sci. Justice*, 46(3):169–178, July 2006.

23. B. S. Weir. The second national research council report on forensic DNA evidence. *Am. J. Hum. Genet.*, 59(3):497–500, Sept. 1996.