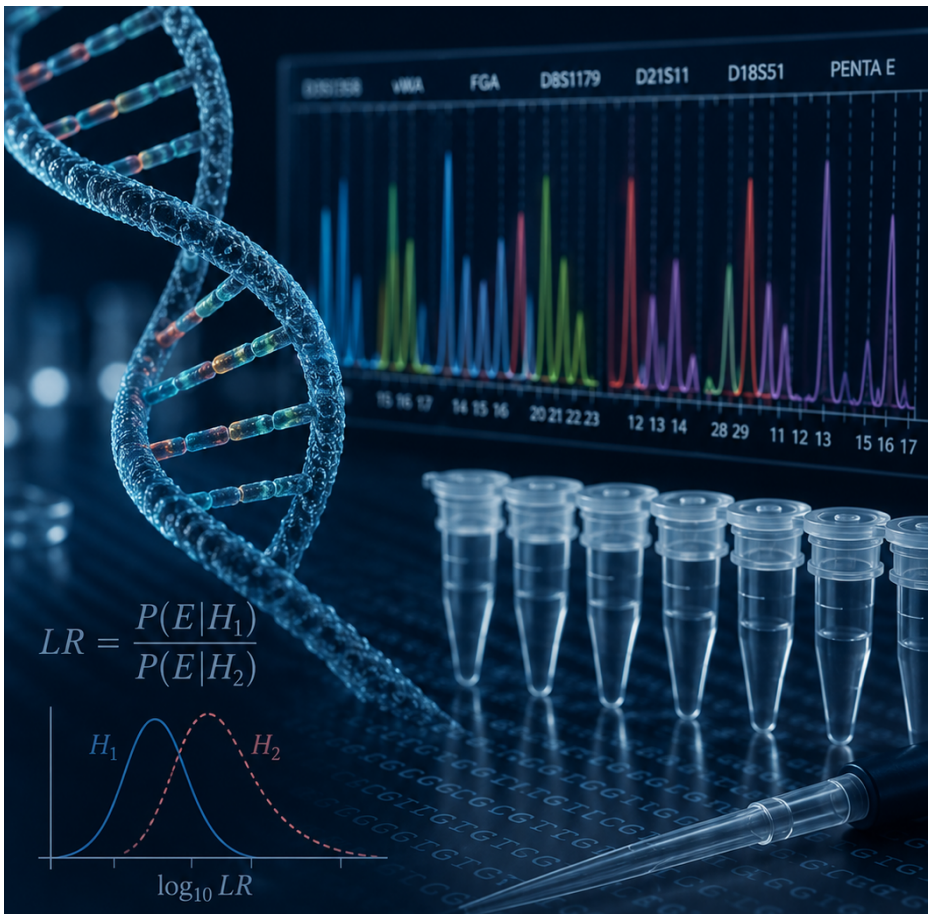
Statistical analysis of replicate measurements of DNA mixtures

Jort Koks



TU Delft

Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid

Statistical analysis of replicate measurements of DNA mixtures

by

Jort Koks

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Wednesday June 24, 2026.

Student number:	5121396
Project duration:	October, 2025 – June, 2026
Thesis committee:	dr. J. Söhl (TU Delft, supervisor)
	dr. D. Kurowicka (TU Delft)
	prof. dr. ir. R. J. F. Ypma (NFI, supervisor)

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid

Preface

This thesis marks the end of my master's programme in Applied Mathematics at Delft University of Technology. During this project, I had the opportunity to apply probability and statistics to a forensic problem at the Netherlands Forensic Institute. Working on DNA mixture interpretation showed me how mathematical modelling can contribute to questions with practical relevance.

I would like to thank my supervisors, Jakob Söhl and Rolf Ypma, for their guidance, feedback and support throughout the project. Their comments and discussions helped shape both the direction and the clarity of this thesis. I am also grateful to the Netherlands Forensic Institute for giving me the opportunity to work on this project and for providing an inspiring research environment.

Furthermore, I would like to thank my friends and family for their support during my studies and during the writing of this thesis.

Jort Koks
The Hague, 15 June 2026

Abstract

At the Netherlands Forensic Institute, additional replicate measurements of the same DNA trace, referred to as rework, can be performed to obtain more information from a DNA mixture profile. Rework may increase the evidential value, expressed by the likelihood ratio (LR), but it also costs laboratory time, resources and DNA sample material. This thesis investigates whether the LR after rework can be predicted from the original DNA mixture profile.

Two main contributions were made. First, a simulation framework was developed to construct predictive distributions for the rework LR. Starting from the deconvolution of the original profile, plausible contributor genotypes are sampled, additional replicate profiles are simulated, and the LR of the combined profile is calculated. Second, a Bayesian MCMC implementation was developed for the EuroForMix/DNAStatistX peak-height model, making it possible to propagate uncertainty in the nuisance parameters when computing LR values.

The framework was evaluated on cleaned two-person NFI research data, focusing on minor contributors. The frequentist plug-in simulation was not sufficiently calibrated: nominal 95% prediction intervals covered only 69.0% of the observed minor true-donor rework LRs. Including Bayesian parameter uncertainty improved the empirical coverage to 81.6% and reduced the mean interval score from 50.5 to 21.6. However, the predicted distributions remained insufficiently calibrated for casework use.

Overall, this thesis shows that predicting rework LRs is possible in principle, and that parameter uncertainty is important for such predictions. The current framework should be viewed as a mathematical proof of concept rather than an operational tool. Further work is needed on artefact modelling, computational scaling, full MCMC validation, extension to more complex mixtures and validation on casework-like data.

The Python code used for the simulations and analyses in this thesis is available at: `https://github.com/jkoks-svg/nfi_master_project`.

Contents

1. Introduction

Forensic evidence plays an important role in criminal investigations and in court, where it helps to assess whether a person of interest (usually a suspect) was involved in the committing of a crime. Forensic DNA analysis is one important example of this. It is used to evaluate whether biological material recovered in a criminal investigation may originate from the person of interest. A DNA trace can be processed and measured in the laboratory, resulting in an electropherogram in which alleles are represented by peaks. Each person has a different set of alleles which together form their genotype, and hence it becomes possible to link the DNA trace to a person of interest. However, recovered biological material can also contain a mixture of DNA of more than one person. Mixtures are more difficult to interpret because the observed alleles are not labelled by contributor, and therefore the genotypes of the individual contributors cannot be inferred directly. Furthermore, DNA measurement is subject to noise and several artefacts which further complicate deconvolution, which is the process of assigning probabilities to possible contributor genotype combinations given the observed mixture profile.

The strength of DNA evidence is expressed by a single number: the likelihood ratio (LR). In the setting considered in this thesis, the LR compares the likelihood of the observed DNA profile if the person of interest contributed to the trace with the likelihood of the same profile if the trace was produced by unknown contributors. The LR is therefore a comparison of two explanations for the evidence; it is not, by itself, the probability that the person of interest contributed.

At the Netherlands Forensic Institute (NFI), an analyst may decide to perform two additional measurements of the same DNA trace. This is referred to in this thesis as rework. Rework can strengthen the evidence because replicate measurements may reveal information that was missing from the first measurement, for example alleles that were not observed above the detection threshold. However, rework also costs laboratory time, resources and DNA sample material. Since forensic traces often contain limited biological material, it is useful to know beforehand whether rework is likely to add evidential value.

The difficulty is that the result of rework is unknown before it is performed. In some cases, rework may substantially increase the LR, while in others it may have no effect at all. Furthermore, it is possible that an analyst wants to know what the increase in LR after rework will be, when he does not yet have a person of interest. Based on the deconvolution of the DNA trace of the first measurement, it is possible to sample plausible contributors.

This thesis studies whether this deconvolution can be used to predict the LR distribution that may be obtained after rework. The target is not a separate LR distribution for each possible genotype. Instead, it is one predictive distribution in which plausible contributor genotypes are sampled according to their deconvolution probabilities, and additional uncertainty comes from the possible outcomes of the replicate measurements. The main research question is therefore:

> Can the original DNA mixture profile be used to construct a well-calibrated predictive distribution for the LR after rework?

If such a distribution is well-calibrated, it could help an analyst estimate the likely size and uncertainty of the LR after rework, and assess whether rework is expected to add evidential value for a specific mixture.

This thesis develops a simulation framework for predicting the LR after rework. Starting from the original mixture profile, the framework computes a deconvolution table, samples plausible genotype combinations, simulates additional replicate traces and computes the LR of the combined profile. Repeating this gives a predictive distribution for the LR after rework.

Two approaches are compared. The frequentist approach estimates the model parameters by maximum likelihood and treats them as fixed. The Bayesian approach propagates parameter uncertainty by sampling from a posterior distribution. The comparison between these approaches

is based on calibration: if the predictive distribution is well calibrated, then observed rework LRs from real replicate measurements should behave like draws from that distribution.

1.1 Thesis outline

The thesis is organised as follows.

- Chapters 2–4 introduce the forensic background, the statistical model and the NFI research dataset used in this thesis.
- Chapters 5 through 8 develop the frequentist approach. First, genotype sampling is validated on observed single and rework profiles. Then, a simulation framework is developed to predict the LR after rework, and its limitations are analysed.
- Chapters 9 through 11 develop the Bayesian extension. These chapters describe the MCMC implementation, the Bayesian LR and deconvolution calculations, and the Bayesian rework simulation algorithm.
- Chapter 12 compares the frequentist, Bayesian and combined simulation configurations. Chapters 13 and 14 discuss the main findings, limitations, practical implications and conclusions.

2. Background and related work

This chapter introduces the forensic background and related work needed for the rest of the thesis. It first describes DNA profiles, likelihood ratios and database searches, then discusses probabilistic genotyping, artefacts, replicate measurements and the role of deconvolution in constructing LR distributions.

2.1 DNA profiles and DNA mixtures

Forensic DNA profiles are based on short tandem repeat (STR) loci [1]. A locus is a specific position in the DNA that is examined in the profiling system. At an STR locus, the number of repeated DNA units may differ between individuals. The observed variant at a locus is called an allele. Because a person inherits one allele from each parent, a genotype at one locus consists of two alleles.

After amplification and electrophoresis, the laboratory output is displayed as an electropherogram. Peaks in the electropherogram indicate detected alleles, and their heights are measured in relative fluorescence units (RFU). A peak is treated as observed only if its height is above a certain detection threshold. Peak heights are informative because, in general, a contributor who contributed more DNA will produce larger peaks than a contributor who contributed less DNA.

For a single-source profile, the observed alleles can often be compared directly with a reference genotype. For a mixture, the profile is the combined result of two or more contributors. If contributors have the same allele at a locus, their DNA contributions are combined in one peak. The observed peaks are therefore not labelled by contributor, and stochastic variation in the laboratory process further complicates the interpretation. Consequently, many different genotype combinations may be compatible with the same profile. An example of a schematic two-person electropherogram at a single locus is shown in Figure 2.1.

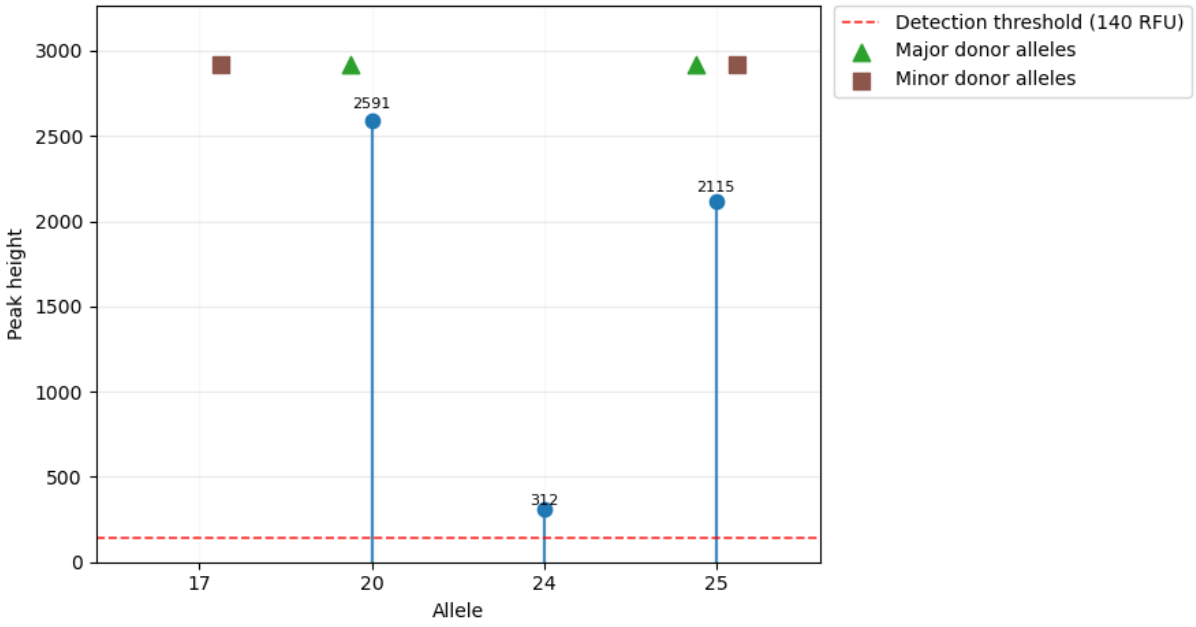


Figure 2.1: Example of a schematic two-person DNA mixture profile, seen at a single locus (mixture 1_5B2 at locus D2S1338). Peaks below the detection threshold of 140 RFU are treated as not observed. In this example, allele 17 is labelled as drop-out for the minor contributor because no peak is observed above the threshold. The small peak at allele 24 is not explained by the contributor genotypes in the schematic profile and is consistent with a possible drop-in or stutter-compatible peak.

2.2 Likelihood ratio

Forensic DNA evidence is evaluated by comparing hypotheses. In the two-person setting used throughout this thesis, the hypotheses are:

$$H_1 : \text{The person of interest and one unknown person contributed.} \tag{2.1}$$

$$H_2 : \text{Two unknown persons contributed.} \tag{2.2}$$

The likelihood ratio is

$$\text{LR} = \frac{P(E \mid H_1)}{P(E \mid H_2)}, \tag{2.3}$$

where E denotes the observed DNA profile. An LR larger than one means that the evidence is more probable under H_1 than under H_2, and vice versa.

2.3 Database searches

In forensic casework, a DNA trace is first recovered from an item or crime scene and processed in the laboratory to obtain a DNA profile. The profile may then be interpreted in different ways, depending on the case context. If there is already a person of interest, the trace profile can be compared directly with that person's reference profile and the strength of the evidence can be expressed by a likelihood ratio.

In other cases, there is not yet a person of interest. The DNA profile may then be used for a database search. In a database search, the trace profile is compared with reference profiles in a DNA database to identify possible candidate contributors. At the NFI, a database hit is only reportable if the LR for a database individual exceeds a reporting threshold, for example 10^9. For high-quality single-source profiles this threshold may be reached relatively easily. For mixtures, however, the individual contributor genotypes are not observed directly. Database searches with mixtures are therefore more difficult, especially when the mixture contains a low-level minor contributor, drop-out, or other artefacts.

This means that the true contributor may be present in the database, while the LR obtained from the original profile is still below the reporting threshold. In that situation, no reportable database hit is obtained, not necessarily because the contributor is absent from the database, but because the original profile contains too little information to support the match strongly enough.

2.4 Probabilistic genotyping

Probabilistic genotyping (PG) systems evaluate DNA profiles within a probability model and report the strength of evidence using an LR. Continuous models, which use the observed peak heights and model the measurement process quantitatively, are examples of PG systems. This thesis uses a continuous model because the objective is to predict the LR after future replicate measurements have been performed. Such prediction depends strongly on quantities such as the amount of DNA in the sample and the mixture proportions of the contributors. These parameters can only be estimated if peak heights are taken into account and therefore require a continuous model. This is also the type of model that is currently used in NFI casework.

The continuous model used in this thesis is based on the gamma peak-height model of Cowell et al. [1]. In this model, allele peak heights are modelled by gamma distributions. The model contains nuisance parameters describing, among other things, the total peak-height scale, peak-height variability, mixture proportions and degradation. EuroForMix is the name of a software implementation that builds on this model and extends it with modelling options for artefacts such as drop-out, drop-in, stutter and degradation [2].

DNAStatistX is the software module that is used at the NFI for deconvolution of DNA profiles and LR calculations [3]. DNAStatistX is closely related to EuroForMix and is used directly for the frequentist LR and deconvolution calculations in this work. Although stutter modelling is technically available, it is not part of the standard DNAStatistX workflow used here. Therefore, in the analyses in this thesis, stutter peaks are assumed to have been filtered from the input data rather than fitted directly by the model.

2.5 Artefacts in DNA mixture profiles

A DNA profile is not a perfect observation of the true contributor genotypes. Several artefacts are relevant for this thesis.

Degradation Degradation causes larger DNA fragments to amplify less efficiently than shorter fragments. In an electropherogram, larger fragments are displayed further to the right. Degradation therefore appears as a systematic decrease in expected peak height from left to right across the electropherogram.

Drop-out Drop-out occurs when an allele is present in the underlying DNA mixture but the corresponding peak is not observed above the detection threshold. Drop-out can occur whenever the amount of DNA is low, but it is especially common for low-level contributors because their alleles tend to have lower peak heights. In Figure 2.1, allele 17 is an example of drop-out for the minor donor.

Drop-in Drop-in occurs when a peak is observed that is not explained by the contributor genotypes. Drop-in may be caused by contamination, background noise or other laboratory effects. Such peaks can have a strong effect on an LR because the model must explain an additional observation.

Stutter Stutter is an amplification artefact. A stutter peak appears at a position related to a true allele, often one repeat unit below the parent allele. Stutter is not the same as random drop-in because its position and expected height are related to a parent allele. The small peak at allele 24 in Figure 2.1 is consistent with a possible stutter or drop-in peak.

2.6 Replicates and rework

A replicate measurement is an additional laboratory measurement of the same underlying DNA sample. In this thesis, rework means that two additional replicate measurements are performed and analysed together with the original measurement. Replicates can be useful because stochastic effects differ between measurements. An allele that drops out in one replicate may be observed in another. By revealing alleles that were missing in the original profile, rework can increase the LR for a true contributor.

2.7 Deconvolution and LR distributions

Deconvolution is the process of assigning posterior probabilities to possible genotype combinations of unknown contributors, given the observed DNA mixture. It combines two sources of information. The first is the prior probability of finding each genotype in the population, which can be obtained from a table used by the NFI that contains allele frequencies. The second is the likelihood of observing the measured peak heights if that genotype combination were the true

donor combination. The output of deconvolution can be viewed as a ranked table of plausible genotype combinations.

In this thesis, the deconvolution is also used as a proposal distribution for simulation. When predicting the LR of a true donor after rework, the simulated contributors should be plausible given the first measurement. These contributors, together with the estimated model parameters, can then be used to generate possible rework profiles.

This idea builds on earlier work on LR distributions for DNA donors, where sampling from the deconvolution was used to estimate LR distributions for observed mixtures [4]. This thesis extends that idea from predicting LR distributions for the current profile to predicting LR distributions after future rework has been performed.

This extension is relevant for database searches. If the LR distribution predicted from the original profile lies mostly below the database-search threshold, while the predicted distribution after rework lies mostly above this threshold, then rework has a clear potential benefit: it may turn a non-reportable database comparison into a reportable candidate hit. Conversely, if the predicted rework distribution remains far below the threshold, additional laboratory work may be less worthwhile for this purpose. In this way, a calibrated prediction of the rework LR distribution can help assess whether rework is likely to make a database search more informative.

3. Mathematical framework

This chapter defines the notation and statistical framework used in the remainder of the thesis. The continuous peak-height model follows the gamma-model formulation as described in [1], [2] and [5]. This is the setup that is currently being used in the DNAStatistX software of the NFI, but is restricted here to the case of two persons without co-ancestry.

3.1 Data, replicates and genotypes

Let E denote a DNA mixture profile. For one measurement, the profile consists of L loci,

$$E = (E_1, \ldots, E_L). \tag{3.1}$$

At locus ℓ, let $\mathcal{A}_\ell$ be the set of possible alleles. The observed data are the thresholded peak heights

$$E_\ell = \{Y^*_{\ell,a} : a \in \mathcal{A}_\ell\}, \tag{3.2}$$

where $Y^*_{\ell,a}$ is the observed peak height at allele a. If no peak is observed above the detection threshold, $Y^*_{\ell,a} = 0$.

For replicate data, write

$$E^{1:R} = (E^{(1)}, \ldots, E^{(R)}), \tag{3.3}$$

where $R = 1$ for a single measurement and $R = 3$ for the rework setting considered in this thesis. The genotype of contributor i at locus ℓ is

$$g_{i,\ell} = (g_{i,\ell,1}, g_{i,\ell,2}), \tag{3.4}$$

and the full genotype of contributor i is $g_i = (g_{i,1}, \ldots, g_{i,L})$. Let G_ℓ be the set of possible genotypes at locus ℓ, and let $G = G_1 \times \cdots \times G_L$ be the full genotype space.

The population probability of an unknown contributor genotype is computed from the allele frequency table that is used by the NFI. This table is shown in Appendix A.2.

$$P(g_{i,\ell}) = 2^{\mathbf{1}(g_{i,\ell,1} \neq g_{i,\ell,2})} P(g_{i,\ell,1}) P(g_{i,\ell,2}) \tag{3.5}$$

It is assumed in this thesis that there exists no shared co-ancestry between the two contributors. Furthermore, it is assumed that genotypes are independent between loci. It follows that:

$$P(g_i) = \prod_{\ell=1}^{L} P(g_{i,\ell}) \tag{3.6}$$

and, for two unknown contributors,

$$P(g_1, g_2) = P(g_1) P(g_2). \tag{3.7}$$

3.2 Gamma peak-height model

The gamma model for allele peak heights as described in [2] will be introduced here. For a two-person mixture, the model parameters are collected in

$$\boldsymbol{\theta} = (\mu, \sigma, \beta, \phi_1), \qquad \phi_2 = 1 - \phi_1. \tag{3.8}$$

Here μ is the expected peak-height scale, σ controls peak-height variability, β controls degradation, and ϕ_i is the mixture proportion of contributor i.

For contributor i, allele a at locus ℓ, define the allele-copy count

$$n_{i,\ell,a} = \sum_{r=1}^{2} \mathbf{1}(g_{i,\ell,r} = a). \tag{3.9}$$

The proportional DNA contribution from contributor i to allele a is

$$\alpha_{i,\ell,a} = \phi_i n_{i,\ell,a}, \tag{3.10}$$

and the total contribution (for the case with two contributors) is

$$\alpha_{\ell,a} = \sum_{i=1}^{2} \alpha_{i,\ell,a}. \tag{3.11}$$

Ignoring degradation for now, the latent peak height is modelled as

$$Y_{\ell,a} \mid g_{1,\ell}, g_{2,\ell}, \boldsymbol{\theta} \sim \Gamma\left(\sigma^{-2}\alpha_{\ell,a},\, \mu\sigma^2\right), \tag{3.12}$$

where the first argument is the shape parameter and the second is the scale parameter. Therefore,

$$\mathbb{E}(Y_{\ell,a} \mid g, \boldsymbol{\theta}) = \mu\alpha_{\ell,a}, \tag{3.13}$$

and

$$\mathrm{Var}(Y_{\ell,a} \mid g, \boldsymbol{\theta}) = \mu^2\sigma^2\alpha_{\ell,a}. \tag{3.14}$$

3.3 Modelling artefacts

The EuroForMix model [2] describes ways to model drop-out, degradation, drop-in, which will be discussed here. Stutter is not incorporated here, since in the NFI workflow it is assumed that this has already been filtered out of the mixture data and is therefore currently not used in DNAStatistX.

3.3.1 Drop-out

Let T_ℓ be the detection threshold at locus ℓ. The detection thresholds used in this thesis are described in table A.4 in the Appendix. A fractional threshold can be used in addition to the locus-specific detection threshold T_ℓ. Let M_ℓ denote the maximum observed peak height at locus ℓ, and let $f \in [0, 1]$ be the fractional-threshold parameter. The effective threshold is then defined as

$$T_\ell^{\text{eff}} = \max(T_\ell, fM_\ell). \tag{3.15}$$

Only allele peaks with height $Y_{\ell,a} \geq T_\ell^{\text{eff}}$ are treated as observed alleles. The observed peak height is

$$Y_{\ell,a}^* = \begin{cases} Y_{\ell,a}, & Y_{\ell,a} \geq T_\ell^{\text{eff}}, \\ 0, & Y_{\ell,a} < T_\ell^{\text{eff}}. \end{cases} \tag{3.16}$$

For an observed peak, the likelihood contribution is the gamma density evaluated at the observed height. For an unobserved allele, the likelihood contribution is the probability that the latent peak height falls below the effective threshold:

$$P(Y_{\ell,a}^* \mid g_\ell, \boldsymbol{\theta}) = \begin{cases} f_\Gamma(Y_{\ell,a}^*; g_\ell, \boldsymbol{\theta}), & Y_{\ell,a}^* \geq T_\ell^{\text{eff}}, \\ \int_0^{T_\ell^{\text{eff}}} f_\Gamma(y; g_\ell, \boldsymbol{\theta})\, dy, & Y_{\ell,a}^* = 0. \end{cases} \tag{3.17}$$

3.3.2 Degradation

Degradation is incorporated by scaling the expected peak height with a function based on the allele fragment length. If $f_{\ell,a}$ is the fragment length of allele a in basepairs, the expected peak height can be written schematically as

$$\mathbb{E}(Y_{\ell,a} \mid g, \boldsymbol{\theta}) = \mu\alpha_{\ell,a}\beta^{\frac{f_{\ell,a}-125}{100}}, \tag{3.18}$$

where $\beta^{\frac{f_{\ell,a}-125}{100}}$ decreases with fragment length when degradation is present. μ and σ must now be seen as the expected peak height and peak-height variability for an allele of length 125 basepairs. By dividing by 100 the domain of β can be restricted to $[0, 1]$. In section A.1 of the Appendix, the computation of the allele fragment length $f_{\ell,a}$ is discussed in more detail.

3.3.3 Drop-in

EuroForMix models drop-in alleles as observed peaks that are not explained by the contributor genotypes or by stutter [2]. In this thesis, a drop-in peak therefore means an unexplained peak above the detection threshold. A possible artefactual peak below the threshold is not observed in the electropherogram and is not classified as drop-in in the observed data.

Let C denote the probability of an observed drop-in event at a locus and let p_a be the population frequency of allele a. The probability that allele a appears as a drop-in is then given by $P(a \text{ is drop-in}) = Cp_a$. Conditional on a drop-in event, the peak height Y_a is assumed to follow an exponential distribution with rate parameter $\lambda > 0$:

$$Y_a \sim \text{Exp}(\lambda), \tag{3.19}$$

with density

$$h(y \mid \lambda) = \lambda e^{-\lambda y}, \qquad y > 0. \tag{3.20}$$

This choice ensures that small drop-in peaks are more probable than large peaks. Since only peaks above the detection threshold T are observed, the observed drop-in peak height is modelled as a truncated exponential random variable $Y_a^* = Y_a \mid Y_a \geq T$, such that

$$Y_a^* - T \sim \text{Exp}(\lambda). \tag{3.21}$$

The likelihood contribution of a drop-in allele is therefore

$$P(Y_a^* \mid a \text{ is drop-in}; \lambda) = Cp_a h(Y_a^* - T \mid \lambda). \tag{3.22}$$

In the full probabilistic model, the standard peak-height likelihood for allele a is replaced by this drop-in likelihood whenever the allele is classified as drop-in. For alleles without drop-in, the likelihood is scaled by the probability $1 - C$.

3.4 Likelihood over loci and replicates

Figure 3.1 shows the causal relationships between the mixture data, the donor genotypes and the model parameters for two loci ℓ and k. This setup easily generalizes to L loci by adding more $(E_\ell, g_{1,\ell}, g_{2,\ell})$ combinations.

Using this causal graph, it follows that:

1. Mixture data on one locus are independent of mixture data on all other loci given the model parameters:
$$E_\ell \perp E_k | \boldsymbol{\theta}, \forall k \neq \ell \tag{3.23}$$

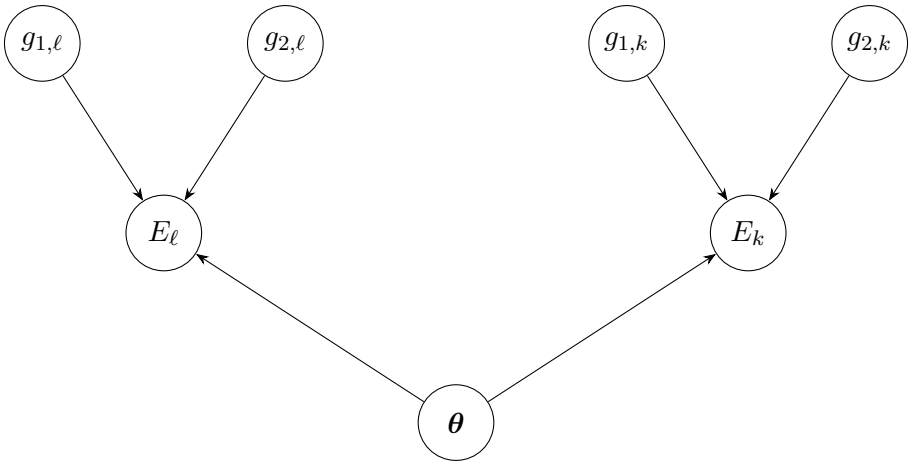


Figure 3.1: Causal graph for two loci, showing the dependence between the observed mixture data E_ℓ, E_k, the contributor genotypes $g_{1,\ell}, g_{2,\ell}, g_{1,k}, g_{2,k}$, and the model parameters $\boldsymbol{\theta}$. Conditional on $\boldsymbol{\theta}$, the observed data factorize over loci. This motivates the locus-wise likelihood factorisation used throughout the thesis.

2. Mixture data on one locus are independent of allele combinations of genotypes on all other loci:

$$E_\ell \perp (g_{1,k}, g_{2,k}), \forall k \neq \ell \tag{3.24}$$

3. Allele combinations of genotypes on one locus are independent of allele combinations of genotypes on all other loci:

$$(g_{1,\ell}, g_{2,\ell}) \perp (g_{1,k}, g_{2,k}), \forall k \neq \ell \tag{3.25}$$

4. Allele combinations of the genotype of the first donor are independent of allele combinations of the second donor, for all loci:

$$g_{1,\ell} \perp g_{2,\ell}, \forall \ell \tag{3.26}$$

5. Allele combinations of genotypes are independent of the model parameters, for all loci:

$$(g_{1,\ell}, g_{2,\ell}) \perp \boldsymbol{\theta}, \forall \ell \tag{3.27}$$

It now follows that:

$$\begin{aligned} P(E|g^1, g^2, \boldsymbol{\theta}) &= P(E_1, E_2, ..., E_L|g_1^1, g_2^1, ...g_L^1, g_1^2, g_2^2, ...g_L^2, \boldsymbol{\theta}) \\ &= \frac{P(E_1, E_2, ..., E_L, g_1^1, g_2^1, ...g_L^1, g_1^2, g_2^2, ...g_L^2|\boldsymbol{\theta})}{P(g_1^1, g_2^1, ...g_L^1, g_1^2, g_2^2, ...g_L^2|\boldsymbol{\theta})} \\ &= \frac{\prod_{\ell=1}^{L} P(E_\ell, g_{1,\ell}, g_{2,\ell}|\boldsymbol{\theta})}{\prod_{\ell=1}^{L} P(g_{1,\ell}, g_{2,\ell}|\boldsymbol{\theta})} \\ &= \prod_{\ell=1}^{L} \frac{P(E_\ell, g_{1,\ell}, g_{2,\ell}|\boldsymbol{\theta})}{P(g_{1,\ell}, g_{2,\ell}|\boldsymbol{\theta})} \\ &= \prod_{\ell=1}^{L} P(E_\ell|g_{1,\ell}, g_{2,\ell}, \boldsymbol{\theta}) \end{aligned} \tag{3.28}$$

We have thus described the likelihood of observing the full 2-person mixture data E given two genotypes g^1, g^2 and model parameters $\boldsymbol{\theta}$, i.e. for all loci together.

For R replicate measurements of the same underlying mixture, the contributor genotypes and model parameters are assumed to be the same across replicates. Conditional on the genotypes and model parameters, the replicate likelihood factorises as

$$P(E^{1:R} \mid g_1, g_2, \boldsymbol{\theta}) = \prod_{r=1}^{R} P(E^{(r)} \mid g_1, g_2, \boldsymbol{\theta}) \tag{3.29}$$

$$= \prod_{r=1}^{R} \prod_{\ell=1}^{L} P(E_\ell^{(r)} \mid g_{1,\ell}, g_{2,\ell}, \boldsymbol{\theta}) \tag{3.30}$$

$$= \prod_{\ell=1}^{L} \prod_{r=1}^{R} P(E_\ell^{(r)} \mid g_{1,\ell}, g_{2,\ell}, \boldsymbol{\theta}) \tag{3.31}$$

$$= \prod_{\ell=1}^{L} P(E_\ell^{1:R} \mid g_{1,\ell}, g_{2,\ell}, \boldsymbol{\theta}). \tag{3.32}$$

It follows that the likelihood of $E^{1:R}$ can be obtained by taking the product over loci of the likelihoods of $E_\ell^{1:R}$, and that the likelihood of $E_\ell^{1:R}$ can be obtained by taking the product over replicates of the likelihoods of $E_\ell^{(r)}$. In the following sections, E can refer to both a single replicate $E^{(r)}$ or a set of replicates $E^{1:R}$.

3.5 Hypotheses

Let g^p be the genotype of the person of interest. For a two-person mixture, the main hypotheses are

$$H_1 : g_1 = g^p, \quad g_2 \in G \tag{3.33}$$

$$H_2 : g_1 \in G, \quad g_2 \in G \tag{3.34}$$

Under H_1, one contributor is fixed to the person of interest and the other contributor is unknown. Under H_2, both contributors are unknown. The likelihoods therefore require marginalisation over the unknown genotypes.

3.6 Frequentist LR

In the frequentist approach, the model parameters in $\boldsymbol{\theta}$ are estimated by maximum likelihood. Separate estimates are obtained under the two hypotheses, because the likelihood surface changes when the person of interest is fixed as a contributor.

The likelihood under H_1 is

$$P(E \mid H_1, \boldsymbol{\theta}) = \prod_{\ell=1}^{L} P(E_\ell \mid H_1, \boldsymbol{\theta}) \tag{3.35}$$

$$= \prod_{\ell=1}^{L} \sum_{g_{2,\ell} \in G_\ell} P(E_\ell \mid g_\ell^p, g_{2,\ell}, \boldsymbol{\theta}) P(g_{2,\ell}), \tag{3.36}$$

where g^p denotes the genotype of the person of interest. The maximum-likelihood estimate under H_1 is

$$\hat{\boldsymbol{\theta}}_1 = \arg\max_{\boldsymbol{\theta}} P(E \mid H_1, \boldsymbol{\theta}). \tag{3.37}$$

Under H_2, both contributors are unknown. The likelihood is

$$P(E \mid H_2, \boldsymbol{\theta}) = \prod_{\ell=1}^{L} P(E_\ell \mid H_2, \boldsymbol{\theta}) \tag{3.38}$$

$$= \prod_{\ell=1}^{L} \sum_{g_{1,\ell} \in G_\ell} \sum_{g_{2,\ell} \in G_\ell} P(E_\ell \mid g_{1,\ell}, g_{2,\ell}, \boldsymbol{\theta}) P(g_{1,\ell}) P(g_{2,\ell}), \tag{3.39}$$

with maximum-likelihood estimate

$$\hat{\boldsymbol{\theta}}_2 = \arg\max_{\boldsymbol{\theta}} P(E \mid H_2, \boldsymbol{\theta}). \tag{3.40}$$

The frequentist likelihood ratio is obtained by plugging in these maximum-likelihood estimates:

$$\mathrm{LR}_{\mathrm{freq}} = \frac{P(E \mid H_1, \hat{\boldsymbol{\theta}}_1)}{P(E \mid H_2, \hat{\boldsymbol{\theta}}_2)}. \tag{3.41}$$

The frequentist likelihood ratio per locus is defined as:

$$\mathrm{LR}_{\mathrm{freq},\ell} = \frac{P(E_\ell \mid H_1, \hat{\boldsymbol{\theta}}_1)}{P(E_\ell \mid H_2, \hat{\boldsymbol{\theta}}_2)}. \tag{3.42}$$

Because the likelihood factorises over loci, the $\log_{10}$-LR can be written as a sum of locus-level contributions:

$$\log_{10} \mathrm{LR}_{\mathrm{freq}} = \sum_{\ell=1}^{L} \log_{10} \mathrm{LR}_{\mathrm{freq},\ell}. \tag{3.43}$$

For combined replicate profiles, the same definition is used with E replaced by $E^{1:R}$, using the replicate likelihood factorisation from Section 3.4. This is the maximum-likelihood strategy used in the frequentist analyses of this thesis.

3.7 Bayesian LR

In the Bayesian approach, the model parameters are treated as uncertain. Instead of estimating one parameter value and substituting it into the likelihood, the likelihood is averaged over possible parameter values. This requires a prior distribution for the parameters under each hypothesis.

The Bayesian likelihood under H_1 is

$$P(E \mid H_1) = \int \left[\prod_{\ell=1}^{L} \sum_{g_{2,\ell} \in G_\ell} P(E_\ell \mid g_\ell^p, g_{2,\ell}, \boldsymbol{\theta}_1) P(g_{2,\ell}) \right] P(\boldsymbol{\theta}_1 \mid H_1)\, d\boldsymbol{\theta}_1. \tag{3.44}$$

The Bayesian likelihood under H_2 is

$$P(E \mid H_2) = \int \left[\prod_{\ell=1}^{L} \sum_{g_{1,\ell} \in G_\ell} \sum_{g_{2,\ell} \in G_\ell} P(E_\ell \mid g_{1,\ell}, g_{2,\ell}, \boldsymbol{\theta}_2) P(g_{1,\ell}) P(g_{2,\ell}) \right] P(\boldsymbol{\theta}_2 \mid H_2)\, d\boldsymbol{\theta}_2. \tag{3.45}$$

The Bayesian likelihood ratio (or Bayes factor) is defined as

$$\mathrm{LR}_{\mathrm{Bayes}} = \frac{P(E \mid H_1)}{P(E \mid H_2)}. \tag{3.46}$$

The frequentist LR evaluates each hypothesis at its best-fitting parameter value, whereas the Bayesian LR averages over parameter uncertainty. The integrals in Eqs. (3.44) and (3.45) are not available in closed form. The approximations used for these integrals are described in Section 3.9.

3.8 Deconvolution and genotype sampling

The deconvolution of the DNA trace under H_2 is used to sample plausible contributor genotypes. This section defines the frequentist and Bayesian deconvolution distributions used later in the thesis.

Frequentist deconvolution. For a fixed parameter value $\boldsymbol{\theta}$, the joint posterior probability of a genotype pair (g_1, g_2) under H_2 is

$$P(g_1, g_2 \mid E, H_2, \boldsymbol{\theta}) = \prod_{\ell=1}^{L} P(g_{1,\ell}, g_{2,\ell} \mid E_\ell, H_2, \boldsymbol{\theta}), \tag{3.47}$$

where the locus-level joint posterior probability of the genotype pair $(g_{1,\ell}, g_{2,\ell})$ under H_2 is

$$P(g_{1,\ell}, g_{2,\ell} \mid E_\ell, H_2, \boldsymbol{\theta}) = \frac{P(E_\ell \mid g_{1,\ell}, g_{2,\ell}, \boldsymbol{\theta}) P(g_{1,\ell}) P(g_{2,\ell})}{\sum_{g'_{1,\ell} \in G_\ell} \sum_{g'_{2,\ell} \in G_\ell} P(E_\ell \mid g'_{1,\ell}, g'_{2,\ell}, \boldsymbol{\theta}) P(g'_{1,\ell}) P(g'_{2,\ell})}. \tag{3.48}$$

In the frequentist implementation, $\boldsymbol{\theta}$ is replaced by $\hat{\boldsymbol{\theta}}_2$, the maximum-likelihood estimate obtained under H_2. This gives the joint deconvolution

$$P(g_{1,\ell}, g_{2,\ell} \mid E_\ell, H_2, \hat{\boldsymbol{\theta}}_2). \tag{3.49}$$

The joint deconvolution describes possible genotype pairs for the two unknown contributors. The marginal deconvolution for the first contributor is

$$P(g_{1,\ell} \mid E_\ell, H_2, \hat{\boldsymbol{\theta}}_2) = \sum_{g_{2,\ell} \in G_\ell} P(g_{1,\ell}, g_{2,\ell} \mid E_\ell, H_2, \hat{\boldsymbol{\theta}}_2), \tag{3.50}$$

and for the second contributor

$$P(g_{2,\ell} \mid E_\ell, H_2, \hat{\boldsymbol{\theta}}_2) = \sum_{g_{1,\ell} \in G_\ell} P(g_{1,\ell}, g_{2,\ell} \mid E_\ell, H_2, \hat{\boldsymbol{\theta}}_2). \tag{3.51}$$

Bayesian deconvolution. In the Bayesian approach, uncertainty in $\boldsymbol{\theta}$ is integrated out. Conditional on a fixed parameter value, the locus-level deconvolution is still given by Eq. (3.48). The Bayesian deconvolution averages this conditional deconvolution over the posterior distribution of the parameters under H_2:

$$P(g_1, g_2 \mid E, H_2) = \int \left[\prod_{\ell=1}^{L} P(g_{1,\ell}, g_{2,\ell} \mid E_\ell, H_2, \boldsymbol{\theta}) \right] P(\boldsymbol{\theta} \mid E, H_2)\, d\boldsymbol{\theta}. \tag{3.52}$$

This distribution accounts for both genotype uncertainty and parameter uncertainty. The marginal deconvolution for the first contributor is

$$P(g_1 \mid E, H_2) = \int \left[\prod_{\ell=1}^{L} \sum_{g_2 \in G_\ell} P(g_{1,\ell}, g_{2,\ell} \mid E_\ell, H_2, \boldsymbol{\theta}) \right] P(\boldsymbol{\theta} \mid E, H_2)\, d\boldsymbol{\theta}. \tag{3.53}$$

and for the second contributor

$$P(g_2 \mid E, H_2) = \int \left[\prod_{\ell=1}^{L} \sum_{g_1 \in G_\ell} P(g_{1,\ell}, g_{2,\ell} \mid E_\ell, H_2, \boldsymbol{\theta}) \right] P(\boldsymbol{\theta} \mid E, H_2)\, d\boldsymbol{\theta}. \tag{3.54}$$

The integrals in Eqs. (3.52), (3.53) and (3.54) are not available in closed form. Their approximations will be described in Section 3.9 as well.

3.9 Approximations for LR and deconvolution

The full frequentist and Bayesian calculations described above are computationally expensive when repeated for many sampled genotypes and many simulated profiles. This section defines the approximations used in the simulation framework.

Frequentist LR approximation. In the full frequentist LR, the numerator is evaluated at $\hat{\boldsymbol{\theta}}_1$ and the denominator at $\hat{\boldsymbol{\theta}}_2$. Re-estimating $\hat{\boldsymbol{\theta}}_1$ for every sampled genotype would require a new maximum-likelihood optimisation for each sampled genotype. To reduce computation time, sampled genotype LRs are computed using the plug-in parameter estimate obtained under H_2.

The approximated frequentist LR is defined as

$$\mathrm{LR}'_{\mathrm{freq}} = \frac{P(E \mid H_1, \hat{\boldsymbol{\theta}}_2)}{P(E \mid H_2, \hat{\boldsymbol{\theta}}_2)} = \prod_{\ell=1}^{L} \frac{P(E_\ell \mid H_1, \hat{\boldsymbol{\theta}}_2)}{P(E_\ell \mid H_2, \hat{\boldsymbol{\theta}}_2)} = \prod_{\ell=1}^{L} \mathrm{LR}'_{\mathrm{freq},\ell} \tag{3.55}$$

The locus-level approximated LR associated with the genotype of the person of interest at locus ℓ g_ℓ^p is then

$$\mathrm{LR}'_{\mathrm{freq},\ell}(g_\ell^p) = \frac{\sum_{g_{2,\ell} \in G_\ell} P(E_\ell \mid g_\ell^p, g_{2,\ell}, \hat{\boldsymbol{\theta}}_2) P(g_{2,\ell})}{\sum_{g_{1,\ell} \in G_\ell} \sum_{g_{2,\ell} \in G_\ell} P(E_\ell \mid g_{1,\ell}, g_{2,\ell}, \hat{\boldsymbol{\theta}}_2) P(g_{1,\ell}) P(g_{2,\ell})} \tag{3.56}$$

$$= \frac{P(g_\ell^p \mid E_\ell, H_2, \hat{\boldsymbol{\theta}}_2)}{P(g_\ell^p)}. \tag{3.57}$$

Thus the approximated LR of a genotype can be interpreted as the ratio between the posterior probability under H_2 of the genotype after measuring the profile and the prior probability of the genotype in the population.

For a sampled donor genotype

$$g_i^{(b)} = (g_{i,1}^{(b)}, \ldots, g_{i,L}^{(b)}),$$

the corresponding approximate full-profile LR is

$$\mathrm{LR}'_{\mathrm{freq}}(g_i^{(b)}) = \prod_{\ell=1}^{L} \mathrm{LR}'_{\mathrm{freq},\ell}(g_{i,\ell}^{(b)}), \tag{3.58}$$

or, equivalently,

$$\log_{10} \mathrm{LR}'_{\mathrm{freq}}(g_i^{(b)}) = \sum_{\ell=1}^{L} \log_{10} \mathrm{LR}'_{\mathrm{freq},\ell}(g_{i,\ell}^{(b)}). \tag{3.59}$$

Throughout Chapters 5–6, the term "sampled LR" refers to this approximate donor-specific LR computed from the marginal deconvolution under H_2.

Bayesian LR approximation. The Bayesian LR can be approximated using posterior samples obtained under H_2, assuming that the same prior distribution for $\boldsymbol{\theta}$ is used under both hypotheses. For a fixed value of $\boldsymbol{\theta}$, define

$$\mathrm{LR}(\boldsymbol{\theta}) = \frac{P(E \mid H_1, \boldsymbol{\theta})}{P(E \mid H_2, \boldsymbol{\theta})}. \tag{3.60}$$

If

$$P(\boldsymbol{\theta} \mid H_1) = P(\boldsymbol{\theta} \mid H_2) = P(\boldsymbol{\theta}), \tag{3.61}$$

then the Bayesian LR can be written as an expectation under the posterior distribution of $\boldsymbol{\theta}$ under H_2:

$$\begin{aligned}
\mathrm{LR}_{\mathrm{Bayes}} &= \frac{P(E \mid H_1)}{P(E \mid H_2)} && (3.62)\\
&= \frac{\int P(E \mid H_1, \boldsymbol{\theta}) P(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{P(E \mid H_2)} && (3.63)\\
&= \int \frac{P(E \mid H_1, \boldsymbol{\theta})}{P(E \mid H_2, \boldsymbol{\theta})} \frac{P(E \mid H_2, \boldsymbol{\theta}) P(\boldsymbol{\theta})}{P(E \mid H_2)} d\boldsymbol{\theta} && (3.64)\\
&= \int \mathrm{LR}(\boldsymbol{\theta}) P(\boldsymbol{\theta} \mid E, H_2)\, d\boldsymbol{\theta} && (3.65)
\end{aligned}$$

If $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)}$ are samples from $P(\boldsymbol{\theta} \mid E, H_2)$, then

$$\mathrm{LR}_{\mathrm{Bayes}} \approx \frac{1}{M} \sum_{m=1}^{M} \frac{P(E \mid H_1, \boldsymbol{\theta}^{(m)})}{P(E \mid H_2, \boldsymbol{\theta}^{(m)})}. \tag{3.66}$$

Bayesian deconvolution approximation. The Bayesian deconvolution in Eqs. (3.52), (3.53) and (3.54) also contain an integral over $\boldsymbol{\theta}$. Using posterior samples $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)}$, these integrals are approximated by

$$P_{\mathrm{Bayes}}(g_1, g_2 \mid E, H_2) \approx \frac{1}{M} \sum_{m=1}^{M} \prod_{\ell=1}^{L} P(g_{1,\ell}, g_{2,\ell} \mid E_\ell, H_2, \boldsymbol{\theta}^{(m)}), \tag{3.67}$$

$$P_{\mathrm{Bayes}}(g_1 \mid E, H_2) \approx \frac{1}{M} \sum_{m=1}^{M} \prod_{\ell=1}^{L} \sum_{g_2 \in G_\ell} P(g_{1,\ell}, g_{2,\ell} \mid E_\ell, H_2, \boldsymbol{\theta}^{(m)}), \tag{3.68}$$

and

$$P_{\mathrm{Bayes}}(g_2 \mid E, H_2) \approx \frac{1}{M} \sum_{m=1}^{M} \prod_{\ell=1}^{L} \sum_{g_1 \in G_\ell} P(g_{1,\ell}, g_{2,\ell} \mid E_\ell, H_2, \boldsymbol{\theta}^{(m)}). \tag{3.69}$$

3.10 Calibration diagnostics

The validation analyses in this thesis compare a single observed LR value, computed for one of the true donors, with a distribution of sampled or simulated LR values. Since LRs are analysed on the $\log_{10}$-scale, let Z denote the observed log-LR value that should be compared with a predictive distribution F. If Z is a draw from F, and F is continuous, then

$$U = F(Z) \tag{3.70}$$

is uniformly distributed on $[0, 1]$. This is the percentile integral transform.

In practice, the predictive distribution F is represented by B sampled values

$$z_1, \ldots, z_B. \tag{3.71}$$

For an observed value z_{obs}, the empirical percentile integral transform is

$$\widehat{u} = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}\{z_b \leq z_{\mathrm{obs}}\}. \tag{3.72}$$

Equivalently, the percentile on the scale from 0 to 100 is

$$p = 100\widehat{u}. \tag{3.73}$$

Percentile histograms Across many validation cases, the empirical percentiles $\widehat{u}$ are visualized in a percentile histogram. If the sampled or simulated LR distributions are calibrated, the observed true-donor LR should behave like a draw from the corresponding predictive distribution. The percentile histogram should then be approximately uniform on $[0, 1]$. Systematic deviations from uniformity indicate miscalibration. For example, many percentiles near 0 or 1 indicate that the observed values often fall in the tails of the predictive distributions, while a concentration near 0.5 indicates that the predictive distributions may be too wide.

To summarize deviations from uniformity, the Kolmogorov–Smirnov statistic is used. Let

$$\widehat{u}_1, \ldots, \widehat{u}_n \tag{3.74}$$

be the empirical percentile values from n validation cases, and let

$$\widehat{u}_{(1)} \leq \cdots \leq \widehat{u}_{(n)} \tag{3.75}$$

denote the ordered values. The empirical distribution function is

$$\widehat{F}_n(u) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}\{\widehat{u}_j \leq u\}. \tag{3.76}$$

The Kolmogorov–Smirnov statistic against the uniform distribution is

$$D_n = \sup_{0 \leq u \leq 1} \left| \widehat{F}_n(u) - u \right|. \tag{3.77}$$

Equivalently, it can be computed from the ordered percentiles as

$$D_n = \max_{1 \leq j \leq n} \left\{ \frac{j}{n} - \widehat{u}_{(j)}, \widehat{u}_{(j)} - \frac{j-1}{n} \right\}. \tag{3.78}$$

A smaller value of D_n indicates that the percentile histogram is closer to uniform. In this thesis, D_n is used as a descriptive calibration summary, not as proof that the predictive distributions are calibrated. If the D_n of two predictive distributions show that both are not uniform, it is still interesting to know which prediction is less poor than the other. Therefore, additional summaries are used to describe other aspects of the predictive distributions.

Coverage Calibration is also evaluated using the empirical coverage of the 95% prediction interval. For validation case i, let z_i be the observed log-LR and let $z_{i,1}, \ldots, z_{i,B}$ be the sampled or simulated log-LRs. Let l_i and u_i be the lower and upper endpoints of the 95% prediction interval obtained from these sampled values. The empirical coverage is

$$\mathrm{cov}_{95} = \frac{100}{n} \sum_{i=1}^{n} \mathbf{1}\{l_i \leq z_i \leq u_i\}. \tag{3.79}$$

A calibrated 95% prediction interval should have coverage close to 95%. Coverage is useful because it directly measures how often the observed true-donor LR lies inside the predicted interval. However, coverage alone does not measure whether the intervals are unnecessarily wide. Very wide intervals can have high coverage while still being uninformative.

Interval score The 95% interval score is used to evaluate both the width and the calibration of the prediction interval. For validation case i, the interval score is

$$\mathrm{IS}_{95,i} = (u_i - l_i) + \frac{2}{0.05}(l_i - z_i)\mathbf{1}\{z_i < l_i\} + \frac{2}{0.05}(z_i - u_i)\mathbf{1}\{z_i > u_i\}. \tag{3.80}$$

The reported interval score is the average over all validation cases:

$$\overline{\mathrm{IS}}_{95} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{IS}_{95,i}. \tag{3.81}$$

Lower values are better. The first term rewards narrow prediction intervals, while the second and third terms penalize observations that fall below or above the interval. The absolute value of the interval score is not very meaningful on its own, but it does make it possible to compare different prediction methods or model settings to each other.

4. Research dataset

This chapter describes the dataset that was used in this thesis. The dataset is controlled, so the true contributors are known, and it contains replicate measurements of the same underlying mixtures. This makes it possible to compare simulated rework outcomes with observed laboratory rework outcomes. Further information on the dataset and the Python scripts that were used to analyse it are available on `https://github.com/jkoks-svg/nfi_master_project`.

4.1 Dataset overview

The dataset consists of DNA mixtures generated and analysed at the Netherlands Forensic Institute using the PowerPlex Fusion 6C kit. The full dataset contains mixtures with two to five contributors and varies in donor combination, mixture proportion, DNA input and allele sharing. Mixtures are evaluated with both high and low detection thresholds. This variation is useful because the value of rework is expected to depend on mixture balance and on how well contributors can be distinguished.

The analyses in this thesis use only the two-person subset evaluated at the high detection thresholds. This keeps the deconvolution and simulation steps computationally feasible while still retaining the main challenges of mixture interpretation, such as allele sharing, drop-out and uncertainty about contributor genotypes.

The two-person mixtures are constructed from five mixture types, shown in Table 4.1, and six donor-combination datasets, shown in Table 4.2. The mixture types determine the amount of DNA per contributor, and therefore also the mixture proportion. The donor-combination datasets determine which donor genotypes are used.

Table 4.1: Mixture types for the two-person PPF6C mixtures.

Mixture type	DNA input per contributor (pg)
A	300:150
B	300:30
C	150:150
D	150:30
E	600:30

Table 4.2: Donor combinations for the two-person PPF6C mixtures.

Dataset number	Type of dataset	Donor combination
1	High allele sharing	A:B
2	Low allele sharing	F:G
3	Random	K:L
4	Random	P:Q
5	Random	U:V
6	Random	Z:AA

Combining the five mixture types with the six donor combinations gives $5 \times 6 = 30$ two-person mixture samples. Since each sample was measured in three replicates, this corresponds to $30 \times 3 = 90$ two-person PPF6C profiles. The three profiles with mixture type B and dataset 3 were unavailable, so $90 - 3 = 87$ two-person profiles were used in the analyses.

Table 4.3 gives an example of one two-person mixture profile from the dataset. It shows the genotypes of the two known contributors and the alleles observed in the mixture profile. In the

mixture profile itself, the alleles are not labelled by contributor.

Table 4.3: Example of a two-person mixture profile from sample 1_1*A*2, together with the genotypes of donors A and B. Peak heights are given in RFU.

Locus	Donor A	Donor B	Observed mixture alleles and peak heights
D1S1656	13, 15.3	16, 18.3	13 (8682), 15.3 (6469), 16 (2751), 18.3 (3051)
TPOX	8, 8	8, 11	8 (10564), 11 (1182)
D2S441	11, 11	11, 11	11 (16330)
D2S1338	18, 20	17, 18	17 (2180), 18 (7284), 20 (5258)
D3S1358	15, 17	14, 15	14 (3950), 15 (8780), 17 (6486)
FGA	22, 24	21, 23	21 (1580), 22 (6183), 23 (2965), 24 (7445)
D5S818	11, 12	11, 12	11 (6539), 12 (8548)
CSF1PO	10, 11	11, 11	10 (5724), 11 (10454)
SE33	20, 29.2	15, 28.2	15 (2278), 20 (3086), 28.2 (2339), 29.2 (5763)
D7S820	8, 11	8, 11	8 (7550), 11 (6220)
D8S1179	14, 15	13, 13	13 (4874), 14 (5530), 15 (4196)
D10S1248	13, 13	15, 16	13 (15720), 15 (1885), 16 (1455)
TH01	6, 9.3	7, 9.3	6 (4147), 7 (1823), 9.3 (7340)
vWA	16, 17	16, 18	16 (8994), 17 (4650), 18 (1938)
D12S391	19.3, 21	21, 24	19.3 (8051), 21 (8364), 24 (2893)
D13S317	11, 12	11, 12	11 (6363), 12 (7488)
Penta E	7, 12	7, 7	7 (13284), 12 (4910)
D16S539	12, 13	11, 13	11 (3643), 12 (9117), 13 (7988)
D18S51	13, 15	13, 16	13 (8407), 15 (8959), 16 (2423)
D19S433	13, 15	14, 14	13 (6794), 14 (4217), 15 (4249)
Penta D	9, 9	10, 10	9 (11837), 10 (3565)
D21S11	30, 31	27, 30	27 (2022), 30 (8631), 31 (5890)
D22S1045	15, 16	11, 16	11 (2187), 15 (4679), 16 (6262)

4.2 Replicate structure

For each underlying mixture, three replicate profiles are available. One replicate can be treated as the original measurement, while the three-replicate combination is treated as the observed rework profile.

Let

$$E^{1:3} = \left(E^{(1)}, E^{(2)}, E^{(3)}\right)$$

denote the three replicate measurements of the same underlying mixture. In the validation setting, one replicate $E^{(r)}$, with $r \in \{1, 2, 3\}$, is treated as the original measurement. The combined profile $E^{1:3}$ is treated as the observed rework profile. The prediction framework uses only $E^{(r)}$ to simulate possible rework outcomes, while $E^{1:3}$ is used afterwards to evaluate the prediction.

4.3 Mixture types and donor roles

For two-person mixtures, the contributor with the larger DNA contribution is referred to as the major donor, and the other as the minor donor. Balanced mixtures are expected to behave differently from unbalanced mixtures. In unbalanced mixtures, minor-donor alleles are more likely to drop out, so rework may have greater potential to increase the LR for the minor donor.

The mixture types in Table 4.1 cover both balanced and unbalanced settings. Mixture type C is balanced, with equal DNA input for both contributors. Mixture types A, B, D and E are unbalanced. Among these, mixture types B, D and E contain a relatively small

minor contribution and are therefore especially relevant for studying whether additional replicate measurements improve the LR for minor donors.

5. Frequentist single profiles

This chapter studies the frequentist deconvolution output for single DNA mixture profiles. A single profile means one replicate measurement analysed on its own, denoted by $E^{(r)}$. The underlying mixture sample may have three replicate profiles, but only one profile is used at a time in this chapter.

For each profile, DNAStatistX is run with both contributors treated as unknown. In terms of the hypotheses from Section 2.2, this corresponds to the genotype model under H_2, where the evidence is assumed to originate from two unknown contributors. The known true donor is not used as input in this deconvolution step; it is only used afterwards for validation.

As described in Section 3.8, DNAStatistX estimates the model parameters under this two-unknown-contributor model, computes the joint deconvolution, and then marginalises this distribution to obtain donor-specific deconvolutions. The sampled LRs in this chapter are therefore donor-specific approximate LRs, as defined in Eq. (3.59).

The analysis answers the following question:

> If a single mixture profile is deconvolved while treating both contributors as unknown, does the LR of the known true donor behave like a draw from the LR distribution obtained by sampling donor genotypes from the marginal deconvolution?

The starting configuration is shown in Table 5.1. This was the baseline setting before introducing the fractional threshold and before removing peaks classified as drop-in.

Table 5.1: Starting DNAStatistX configuration values used in the baseline single-profile analysis. HT denotes the high-threshold input data.

Setting	Symbol	Value
Dataset	–	2-person HT mixtures with drop-ins
Fractional threshold	f	0.0
Drop-in probability	C	0.05
Drop-in rate	λ	0.01
Detection thresholds	T_ℓ	See Appendix
Population frequency table	–	See Appendix
Coancestry coefficient	–	0.0
Rare allele frequency	–	$3 \cdot 10^{-4}$

Algorithm 1 summarizes the single-profile analysis.

5.1 Sampled donor LR distributions

For each contributor position $i \in \{1, 2\}$, $B = 10,000$ donor genotypes are sampled from the marginal deconvolution distribution

$$P(g_{i,\ell} \mid E_\ell, H_2, \hat{\boldsymbol{\theta}}_2).$$

For each sampled donor genotype, the full-profile approximate LR is computed using Eq. (3.59). The same calculation is performed for the known true donor genotype assigned to that contributor position.

The position of the true-donor LR inside the sampled LR distribution is summarized by the percentile

$$p = 100 \cdot \frac{1}{B} \sum_{b=1}^{B} \mathbf{1} \left\{ \log_{10} \mathrm{LR}(g_i^{(b)}) \leq \log_{10} \mathrm{LR}(g_i^{\mathrm{true}}) \right\}.$$

Algorithm 1 Frequentist sampling method

Input: Mixture profile E
Run the frequentist H_2 deconvolution described in Section 3.8
Obtain the marginal deconvolution distributions

$$P(g_{1,\ell} \mid E_\ell, H_2, \hat{\boldsymbol{\theta}}_2) \quad \text{and} \quad P(g_{2,\ell} \mid E_\ell, H_2, \hat{\boldsymbol{\theta}}_2)$$

for each contributor position $i \in \{1, 2\}$ **do**
    for $b = 1, \ldots, B$ **do**
        Sample one genotype $g_{i,\ell}^{(b)}$ per locus from the marginal deconvolution
        Construct the sampled full donor genotype

$$g_i^{(b)} = (g_{i,1}^{(b)}, \ldots, g_{i,L}^{(b)})$$

        Compute its $\log_{10}$-LR using Eq. (3.59)
    end for
    Compute the same $\log_{10}$-LR for the known true donor assigned to contributor position i
    Compute the percentile of the true-donor LR among the sampled LRs
end for
Output: true-donor percentiles for the contributor positions

A percentile close to zero means that almost all sampled donor genotypes have a higher LR than the true donor. A percentile close to one hundred means that almost all sampled donor genotypes have a lower LR than the true donor.

The major donor is often reconstructed almost deterministically. The marginal deconvolution distribution for the major contributor frequently places nearly all probability mass on one genotype, so repeated sampling gives the same genotype and therefore the same LR. The minor donor is more informative for this diagnostic analysis, because its genotype is less certain and the sampled LR distribution has non-trivial variation. For this reason, the remaining analyses in this chapter focus on the minor donor LR.

Figure 5.1 shows one example. The histogram shows the sampled $\log_{10}$-LR distribution for the minor donor, while the vertical line shows the $\log_{10}$-LR of the known true donor.

Under H_2, the two unknown contributor positions are ordered by estimated mixture proportion. For unbalanced mixtures, the major true donor is naturally linked to the contributor position with the larger estimated mixture proportion, and the minor true donor to the contributor position with the smaller estimated mixture proportion. For balanced mixtures, especially mixture type C, this ordering can be ambiguous. Therefore, both possible assignments of the two unknown contributor positions to the two known donors are checked, and the assignment with the largest sum of the two true-donor $\log_{10}$-LRs is used.

5.2 Effect of the fractional threshold

The baseline analysis used $f = 0$, $C = 0.05$ and $\lambda = 0.01$, where f is the fractional-threshold parameter, C is the drop-in probability and λ is the drop-in peak-height rate parameter. In this setting, many true-donor percentiles were close to zero for the minor contributor, as shown in Figure 5.2. This means that sampled donor genotypes often obtained higher LRs than the true donor.

The fractional threshold from Section 3.3.1 was then set to $f = 0.03$, while keeping $C = 0.05$ and $\lambda = 0.01$. Figure 5.2 shows that this reduced the concentration of percentiles near zero. The value D shown in each panel is the Kolmogorov–Smirnov statistic from Eq. (3.77), computed against the uniform distribution. A smaller value of D indicates that the percentile histogram

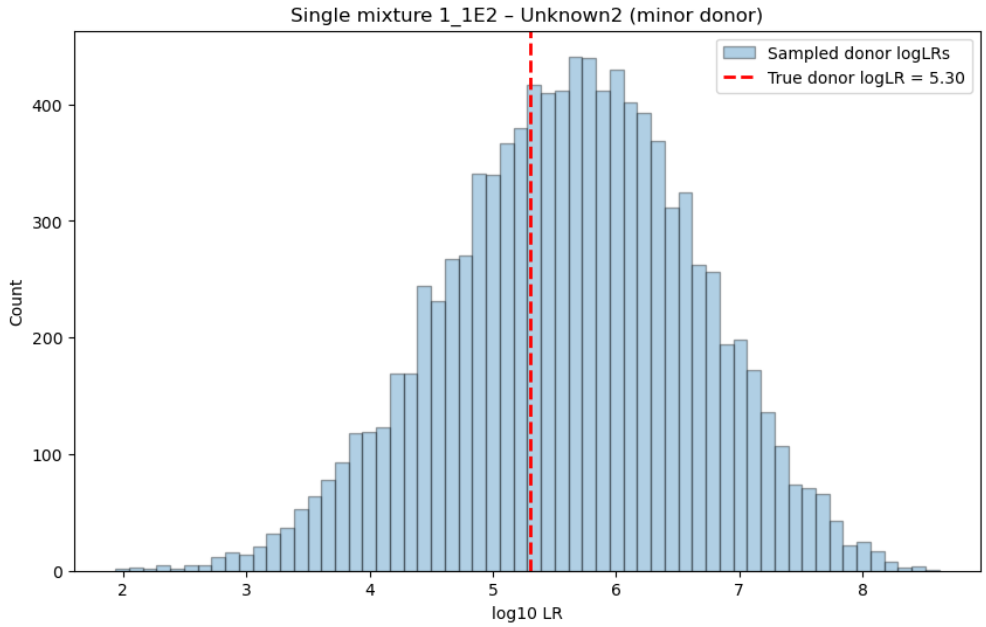


Figure 5.1: Example of a sampled $\log_{10}$-LR distribution for the minor contributor in one single profile. The histogram shows the LRs of donor genotypes sampled from the marginal deconvolution. The vertical line shows the $\log_{10}$-LR of the known true donor.

is closer to uniform.

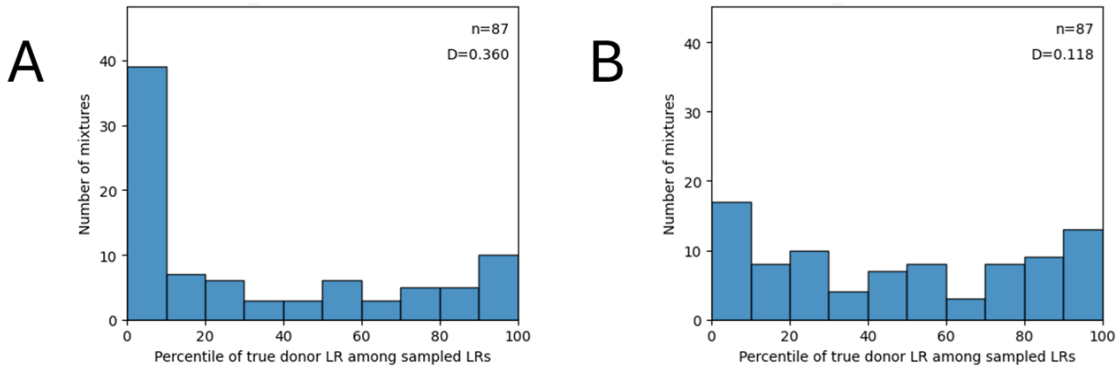


Figure 5.2: Effect of the fractional threshold for the minor contributor. The left panel (A) uses no fractional threshold. The right panel (B) uses $f = 0.03$. The reduction of percentiles near zero shows that weak relative peaks caused many sampled donor genotypes to obtain higher LRs than the true donor. The value D is the Kolmogorov–Smirnov statistic against the uniform distribution, as defined in Eq. (3.77). The 95%-coverage improved from 71.3% to 87.4%, while the mean interval score decreased (i.e. improved) from 50.0 to 13.1.

5.3 Locus-level diagnostics

The right panel of Figure 5.2 still contains several low true-donor percentiles. To identify which loci cause this behaviour, the log-LR decomposition in Eq. (3.59) is used.

For each locus, define

$$\Delta_\ell = \log_{10} \mathrm{LR}_\ell(g_{i,\ell}^{\mathrm{true}}) - \frac{1}{B}\sum_{b=1}^{B} \log_{10} \mathrm{LR}_\ell(g_{i,\ell}^{(b)}).$$

A negative value of Δ_ℓ means that the true donor has a lower LR than the average sampled donor genotype at that locus.

An example is locus D2S1338 in mixture 1_5B2. For the minor donor, this locus has

$$\Delta_\ell = -3.1.$$

The true-donor locus-level $\log_{10}$-LR is approximately -2.3, while the average sampled locus-level $\log_{10}$-LR is approximately $+0.7$. Thus, at this locus, the true donor LR is about three log units lower than the average sampled LR. The peak pattern for this locus was already shown schematically in Figure 2.1: the minor donor allele 17 is not observed above the threshold, while an unexplained peak occurs at allele 24. This example illustrates how a small number of loci can strongly lower the position of the true-donor LR within the sampled LR distribution.

5.4 Drop-out, drop-in and stutter-compatible peaks

The locus-level analysis was applied to 4002 locus instances. This number comes from the 87 two-person single profiles used in the analysis multiplied by the 23 autosomal loci considered per profile:

$$87 \times 23 = 4002.$$

For each locus instance, Δ_ℓ was computed for the minor donor. The loci were then classified by whether they contained drop-out and/or drop-in. Here, drop-out means that an allele of the known true donor is not observed above the effective threshold. Drop-in means that an observed allele is not present in any of the known true contributors.

Table 5.2 separates the locus instances by drop-out and drop-in status. For mildly negative values of Δ_ℓ, drop-out without drop-in is common. For the strongest negative deviations, loci with drop-in dominate. Among loci with $\Delta_\ell < -3$, half contain both drop-out and drop-in, and another 41.7% contain drop-in without drop-out. Thus, the most problematic loci are often associated with unexplained drop-in peaks.

Table 5.2: Classification of loci by drop-out and drop-in status. Percentages are relative to the number of loci in the corresponding row.

Dataset	# loci	No drop-out No drop-in	Drop-out No drop-in	No drop-out Drop-in	Drop-out Drop-in
Total dataset	4002	3194 (79.8%)	642 (16.0%)	152 (3.8%)	14 (0.4%)
$\Delta_\ell < 0$	1308	880 (67.3%)	353 (26.7%)	66 (5.1%)	9 (0.7%)
$\Delta_\ell < -1$	83	17 (20.5%)	35 (42.2%)	24 (28.9%)	7 (8.4%)
$\Delta_\ell < -2$	24	0 (0.0%)	6 (25.0%)	11 (45.8%)	7 (29.1%)
$\Delta_\ell < -3$	12	0 (0.0%)	1 (8.3%)	5 (41.7%)	6 (50.0%)

Table 5.3 shows where the drop-in peaks occur. A peak is counted as stutter-compatible if it occurs exactly one repeat unit or half a repeat unit before or after a parent allele. In the full set of 4002 loci, 166 drop-in peaks were found. Of these, 162 were at stutter-compatible positions. Thus, almost all drop-in peaks in this analysis occur at positions that could be related to stutter.

This explains why it is reasonable to remove peaks classified as drop-in before continuing. The removed peaks are not mainly arbitrary random peaks: almost all are at positions that could be related to stutter. If stutter were fully modelled, these peaks would contribute through the stutter component of the likelihood. In the present model, however, they remain unexplained and can lead to true-donor LR values that are lower than the average sampled LRs at the affected loci.

5.5 Final cleaned setting for validation

The previous sections show that the single-profile analysis is sensitive to weak unexplained peaks. The analysis was therefore adjusted in two steps. First, the fractional threshold was set

Table 5.3: Drop-in peaks at stutter-compatible and non-stutter-compatible positions. Percentages are relative to the number of loci in the corresponding row.

Dataset	# loci	# drop-in non-stutters	# drop-in stutters
Total dataset	4002	4 (0.1%)	162 (4.1%)
$\Delta_\ell < 0$	1308	2 (0.2%)	73 (5.6%)
$\Delta_\ell < -1$	83	0 (0.0%)	31 (37.4%)
$\Delta_\ell < -2$	24	0 (0.0%)	18 (75.0%)
$\Delta_\ell < -3$	12	0 (0.0%)	11 (91.7%)

to $f = 0.03$. Second, peaks classified as drop-in using the known donor genotypes were removed from the input data. After this cleaning step, the drop-in parameters were set close to zero, because the cleaned profiles should no longer rely on drop-in as an explanation for unexplained peaks.

Table 5.4 gives the final configuration used after these preprocessing steps. This is the setting used for the analyses in the remainder of the frequentist part of the thesis.

Table 5.4: Final DNAStatistX configuration values used after the single-profile preprocessing steps. HT denotes the high-threshold input data.

Setting	Symbol	Final value
Dataset	–	2-person HT mixtures without drop-ins
Fractional threshold	f	0.03
Drop-in probability	C	10^{-6}
Drop-in rate	λ	10^{-6}
Detection thresholds	T_ℓ	See Appendix
Population frequency table	–	See Appendix
Coancestry coefficient	–	0.0
Rare allele frequency	–	$3 \cdot 10^{-4}$

5.6 Validation on cleaned single profiles

The single-profile percentile analysis was repeated using the final cleaned setting. Figure 5.3 shows the resulting percentile histogram for the minor donor. Compared with the baseline analysis in Figure 5.2, the concentration of percentiles near zero is reduced. The smaller value of D indicates that the cleaned percentile histogram is closer to the uniform distribution. This suggests that the low true-donor percentiles in the baseline analysis were largely caused by weak unexplained peaks, many of which occurred at stutter-compatible positions.

This result supports the use of the cleaned setting for the frequentist analyses that follow. The next chapter applies the same genotype-sampling idea to observed rework profiles, where the three replicate measurements of the same mixture are analysed together.

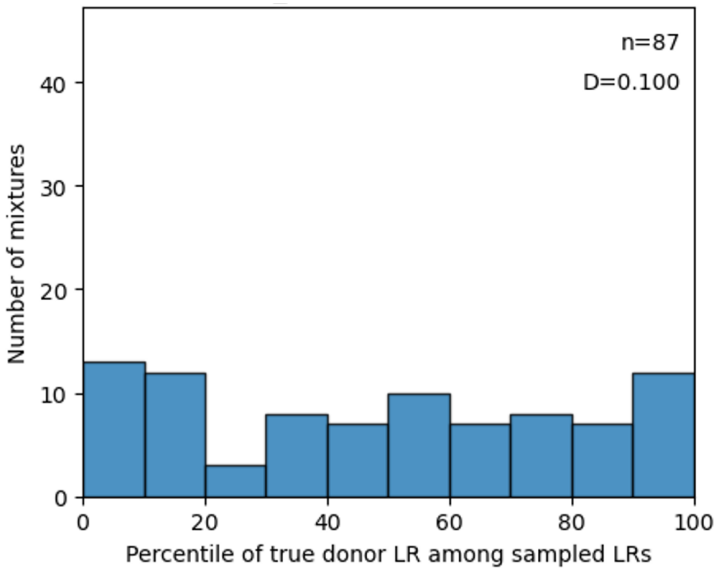


Figure 5.3: Single-profile percentile histogram after applying the final cleaned setting. Peaks classified as drop-in using the known donor genotypes were removed, and the drop-in parameters were set close to zero. The reduced concentration of percentiles near zero and the smaller Kolmogorov–Smirnov statistic D indicate improved agreement between the sampled LR distribution and the true-donor LR. The 95%-coverage was 92.0% and the mean interval score was 7.92.

6. Frequentist rework profiles

Chapter 5 analysed frequentist genotype sampling on cleaned single profiles. This chapter applies the same idea to observed rework profiles. The aim is to study what happens when the real replicate measurements from the laboratory are analysed together.

For each underlying mixture, the dataset contains three replicate measurements,

$$E^{1:3} = \left(E^{(1)}, E^{(2)}, E^{(3)}\right). \tag{6.1}$$

In this chapter, these three replicate files are analysed jointly as one combined replicate profile. This combined replicate profile is referred to as the observed rework profile. When a replicate $E^{(r)}$ is treated as the original profile, the observed rework profile is the combination of this original profile with the two remaining laboratory replicates. Thus, $E^{1:3}$ represents the result that would have been available after rework. The same final cleaned setting from Chapter 5 is used.

The analysis has two purposes. First, it quantifies the increase in true-donor $\log_{10}$-LR obtained by combining replicate measurements. Second, it checks whether the true-donor LR is consistent with the LR distribution obtained by sampling donor genotypes from the deconvolution of the observed rework profile.

6.1 Definition of the observed rework profile

The observed rework profile is analysed under H_2, where both contributors are unknown. The only difference from the single-profile analysis in Chapter 5 is that the deconvolution is computed from all three replicate measurements jointly, using the replicate likelihood from Section 3.4.

As in Chapter 5, donor genotypes are sampled from the marginal deconvolution under H_2. The difference is that the deconvolution probabilities are now conditional on the combined replicate profile $E^{1:3}$, rather than on a single profile $E^{(r)}$. For each sampled genotype, the approximate donor-specific LR is computed using the same plug-in LR approximation as in Section 3.9.

For contributor position i, the observed rework analysis gives a true-donor rework LR,

$$\log_{10} \mathrm{LR}_{\mathrm{rework}}(g_i^{\mathrm{true}}). \tag{6.2}$$

This value can be compared with the true-donor LR obtained from a single original profile $E^{(r)}$,

$$\log_{10} \mathrm{LR}^{(r)}_{\mathrm{single}}(g_i^{\mathrm{true}}). \tag{6.3}$$

The true-donor LR increase (or gain) after rework is defined as

$$\Delta^{(r)}_{\mathrm{true}} = \log_{10} \mathrm{LR}_{\mathrm{rework}}(g_i^{\mathrm{true}}) - \log_{10} \mathrm{LR}^{(r)}_{\mathrm{single}}(g_i^{\mathrm{true}}). \tag{6.4}$$

A positive value of $\Delta^{(r)}_{\mathrm{true}}$ means that combining the replicate measurements increased the LR of the known true donor compared with the original single profile.

Figure 6.1 shows an example. The three single-profile panels correspond to the three separate replicate measurements. The combined-profile panel corresponds to the observed rework profile, where all three replicate measurements are analysed together.

6.2 Increase in true-donor log10-LR after rework

The next question is how much the true-donor LR changes after combining replicate measurements. The analysis focuses on the minor donor, since the major donor is often reconstructed

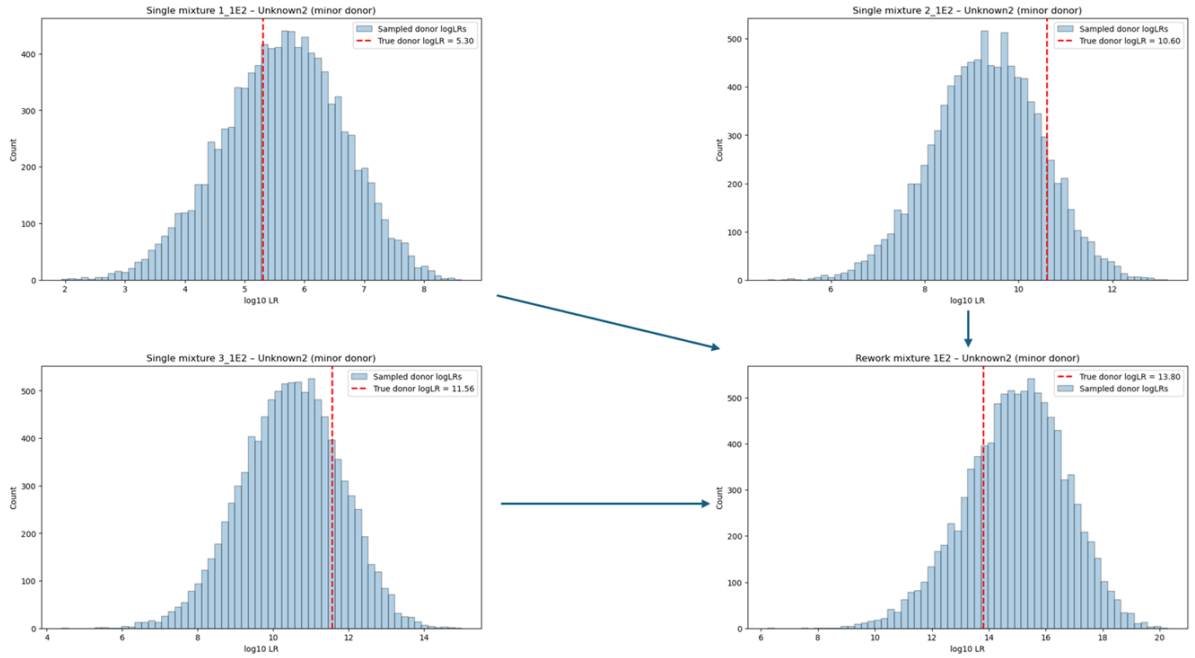


Figure 6.1: The three single-profile panels show the LR distributions for the minor donor obtained from the separate replicate measurements. The bottom-right panel shows the observed rework profile obtained by analysing the three replicates together. The red vertical line marks the LR of the minor true donor.

almost deterministically. Hence the true-donor LR mentioned in this chapter refers to the true-donor LR of the minor donor.

Figure 6.2 compares the true-donor $\log_{10}$-LRs for single profiles and observed rework profiles. All values in the figure refer to the minor donor. Because each available replicate can be treated as the original profile, the comparison is made for all available single-profile starting points. In general, the observed rework profiles give larger true-donor LRs than the single profiles, although the size of the increase differs between mixtures.

6.3 Drop-out and increase in true-donor log10-LR

The increase in true-donor $\log_{10}$-LR after rework is expected to be largest when the single profile is missing information that can be recovered in another replicate. This is especially relevant for the minor donor in unbalanced mixtures, because minor-donor alleles have lower expected peak heights and are therefore more likely to drop out in a single measurement. When replicate measurements are combined, an allele that dropped out in one replicate may be observed in another.

Table 6.1 summarizes the relation between mixture type, average minor-donor drop-out, and the average increase in true-donor $\log_{10}$-LR after rework. The table shows that drop-out is an important factor in the value of rework, especially for the most unbalanced mixtures. However, drop-out alone does not explain the increase in true-donor $\log_{10}$-LR after rework. This is most clearly seen for mixture type A: these mixtures have no average minor-donor drop-out by this definition, but still show a large average increase in true-donor LR after rework.

A possible explanation for the large increase in true-donor $\log_{10}$-LR after rework in A-type mixtures is ambiguity in the mixture-proportion estimate. In the peak-height model from Chapter 3, the expected contribution to a peak depends on the mixture proportion and the number of allele copies. For an A mixture, the configured major:minor proportion is 2:1, corresponding to mixture proportions 0.67:0.33. At this ratio, one major-donor allele copy has the same expected contribution as two minor-donor allele copies. As a result, several genotype assignments can

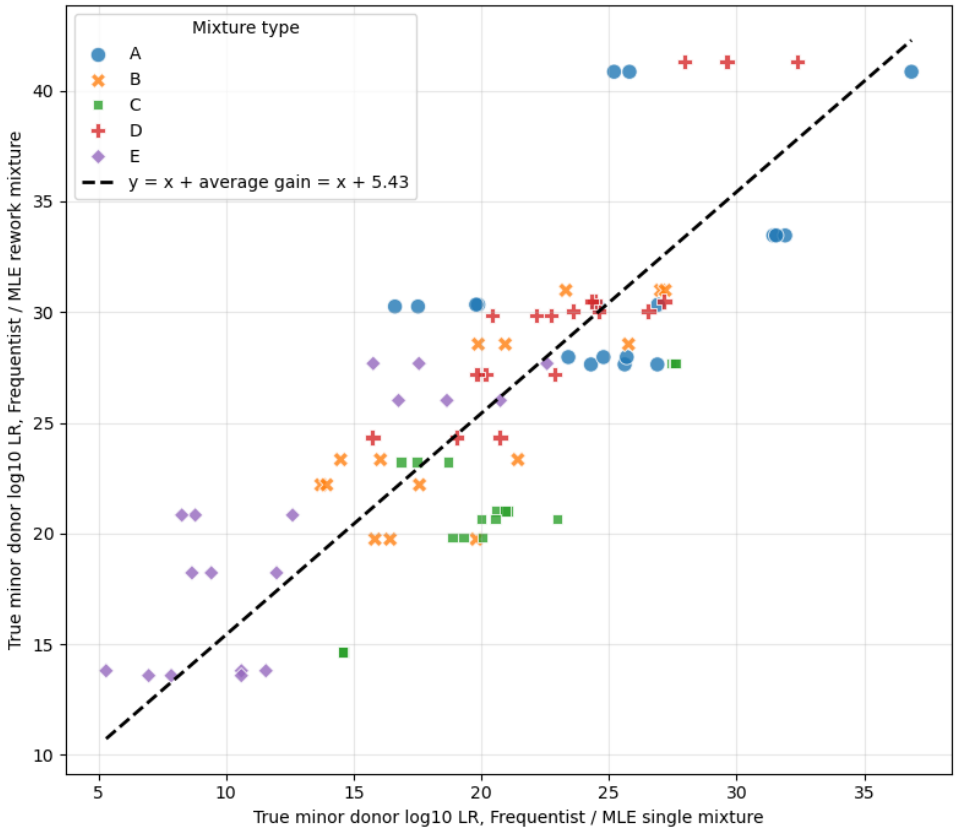


Figure 6.2: Comparison of the true minor-donor $\log_{10}$-LR for single-mixture and rework analyses using the frequentist/MLE approach. Points are coloured and shaped by mixture type; the dashed line shows the average increase in true-donor $\log_{10}$-LR after rework, $y = x + 5.43$

Table 6.1: Average minor-donor drop-out and increase in true-donor $\log_{10}$-LR after rework by mixture type. The drop-out column gives the average number of loci per single mixture profile with at least one dropped-out minor-donor allele.

Mixture type	Mixture proportion minor:major	Average number of drop-out loci	Average increase in true-donor $\log_{10}$-LR
C	1:1	0.00	+0.74
A	1:2	0.00	+6.75
D	1:5	2.83	+5.78
B	1:10	5.27	+6.40
E	1:20	11.00	+8.86

explain similar peak-height patterns reasonably well. A peak height that is compatible with one allele copy from the major donor may also be compatible with two allele copies from the minor donor. This makes it harder for the likelihood optimization to separate the contributors using a single profile.

This ambiguity affects the maximum-likelihood estimates. Instead of clearly estimating the configured value 0.67:0.33, the optimizer often finds mixture proportions closer to 0.50:0.50 for A-type mixtures, as shown in Figure 6.3. Such an estimate makes the profile appear more balanced than it was configured to be. In addition, the fitted peak-height variability can increase, allowing a wider range of genotype combinations to explain the observed peak heights. The single-profile LR for the true minor donor may therefore be lower not because alleles are missing, but because the contributor separation is uncertain.

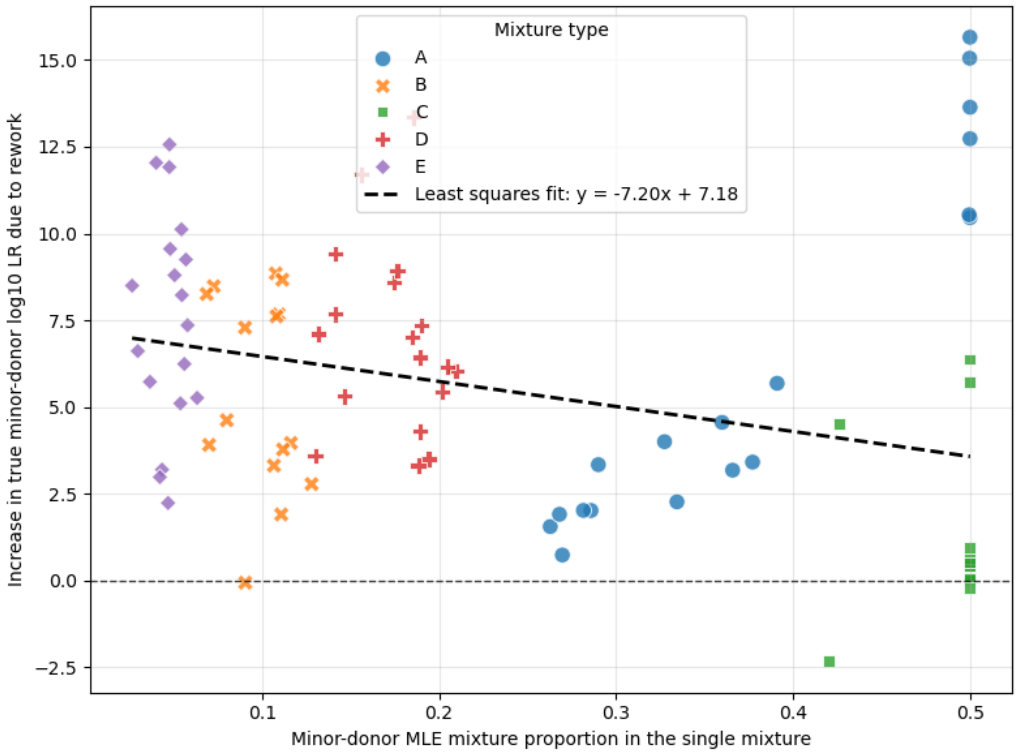


Figure 6.3: Maximum-likelihood estimates of the mixture proportions for the single profiles, grouped by mixture type. The A-type mixtures were configured with major:minor proportion 2:1, corresponding to 0.67:0.33, but many estimates are close to 0.50:0.50. This suggests that, for A-type mixtures, the single-profile likelihood often does not clearly separate the major and minor contributors.

Rework can reduce this uncertainty because the same contributor genotypes are observed through multiple independent replicate measurements. Even when no minor-donor alleles dropped out in the original profile, the additional peak-height information can make the true genotype assignment more distinguishable. This explains why A-type mixtures can show a large increase in true-donor $\log_{10}$-LR after rework despite having little or no minor-donor drop-out.

Figure 6.4 shows the relation between minor-donor drop-out and increase in true-donor $\log_{10}$-LR after rework at mixture level. The positive trend confirms that mixtures with more minor-donor drop-out generally benefit more from analysing replicate measurements together. However, the spread around the trend, together with the behaviour of the A-type mixtures, shows that the increase in true-donor $\log_{10}$-LR after rework is not determined by drop-out alone.

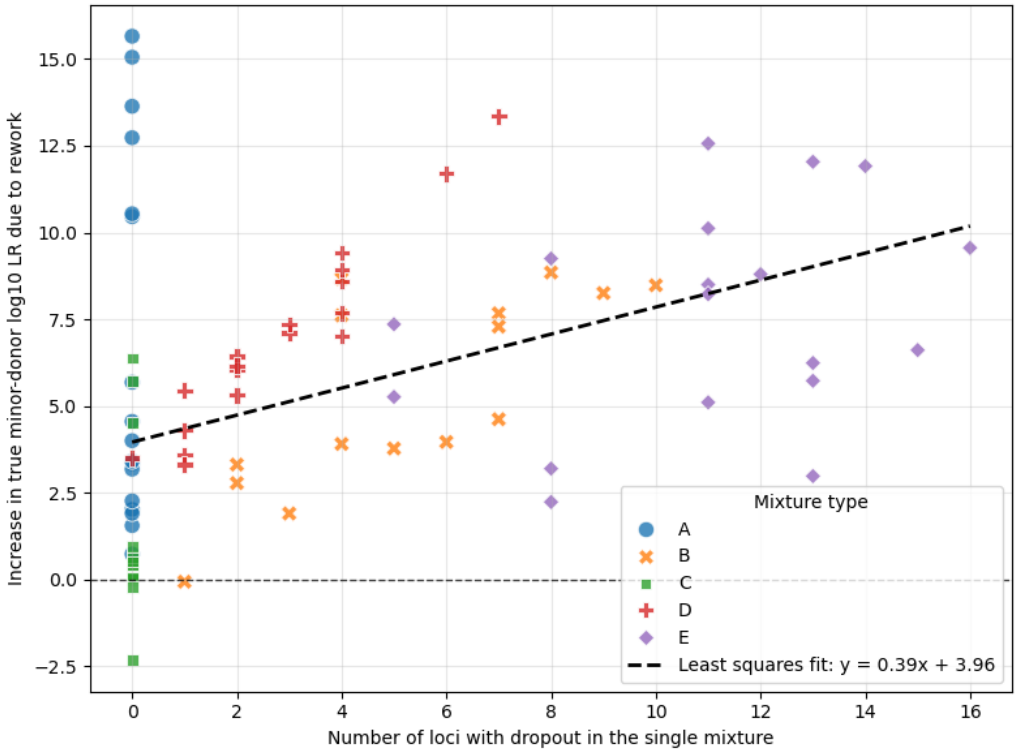


Figure 6.4: Relationship between the number of minor-donor drop-out loci in the single profile and the increase in true-donor $\log_{10}$-LR after rework. The positive trend shows that drop-out is an important factor in the value of rework, while the spread shows that other factors, such as peak-height information, genotype uncertainty and mixture-proportion ambiguity, also affect the increase in true-donor $\log_{10}$-LR after rework.

6.4 Frequentist sampling validation on rework profiles

This section assesses the calibration of the genotype-sampling method by checking whether the true-donor rework LR falls at approximately uniform percentiles of the sampled LR distributions. As in Chapter 5, we take $B = 10,000$ sampled donors from the deconvolution.

For contributor position i, the observed rework deconvolution gives sampled rework LRs

$$\log_{10} \mathrm{LR}_{\mathrm{rework}}(g_i^{(1)}), \ldots, \log_{10} \mathrm{LR}_{\mathrm{rework}}(g_i^{(B)}), \tag{6.5}$$

and a true-donor rework LR

$$\log_{10} \mathrm{LR}_{\mathrm{rework}}(g_i^{\mathrm{true}}). \tag{6.6}$$

The percentile of the true-donor LR is

$$p_{\mathrm{rework}} = 100 \cdot \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}\left\{ \log_{10} \mathrm{LR}_{\mathrm{rework}}(g_i^{(b)}) \leq \log_{10} \mathrm{LR}_{\mathrm{rework}}(g_i^{\mathrm{true}}) \right\}. \tag{6.7}$$

If the sampled donor LR distribution is calibrated for observed rework profiles, these percentiles should resemble a uniform distribution. If they systematically concentrate near zero or one hundred, this would indicate that the true-donor LR cannot be regarded as a random draw of the generated distribution.

Figure 6.5 shows the percentile histogram for the observed rework profiles. Each percentile gives the position of the true-donor rework LR within the LR distribution obtained by sampling donor genotypes from the deconvolution of the combined replicate profile $E^{1:3}$.

The rework percentile histogram should be interpreted as a diagnostic for the deconvolution sampling method after replicate measurements have already been combined. If the sampled

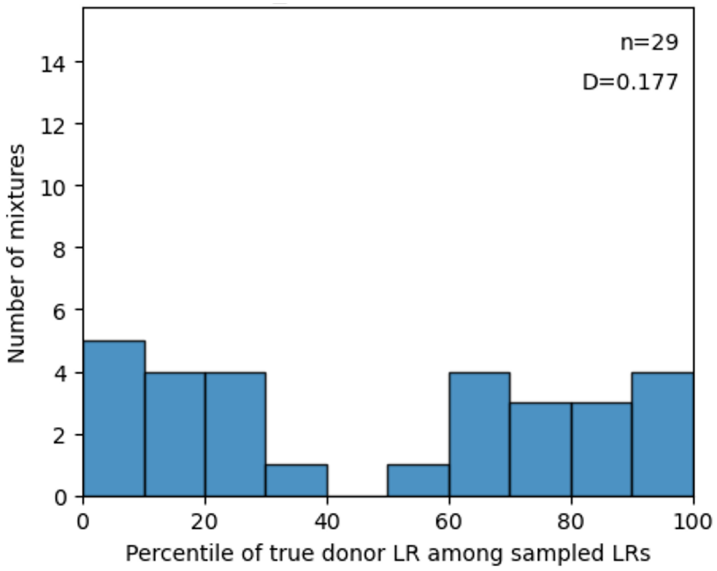


Figure 6.5: Percentile validation for observed rework profiles. Each percentile gives the position of the true-donor rework LR within the sampled LR distribution. The rework profile is obtained by analysing the three replicate measurements together. The 95%-coverage is 89.7% and the mean interval score is 8.50.

LR distributions are calibrated, the percentiles should be approximately uniformly distributed. However, this is difficult to assess visually because the histogram contains only 29 observations. Therefore, the figure should be read as a qualitative check rather than as strong evidence of calibration.

6.5 Implications for rework simulation

The observed rework analysis shows that combining replicate measurements can increase the true-donor $\log_{10}$-LR for the minor donor, especially in unbalanced mixtures with substantial drop-out. The percentile validation is important because the simulation algorithm in Chapter 7 aims to predict exactly this type of rework LR distribution. If the sampled LR distributions are not calibrated even after conditioning on the observed rework profile $E^{1:3}$, then it would not be meaningful to use similar distributions to predict the LR of the true donor before rework is performed. Good percentile behaviour in this chapter is therefore a necessary diagnostic for the deconvolution sampling method.

7. Frequentist rework simulation algorithm

Chapters 5 and 6 studied genotype sampling on observed profiles. The next step is to use the deconvolution of a single profile to predict what the LR could become after rework. This chapter describes the frequentist simulation algorithm used for this prediction.

The aim is to construct a predictive distribution for the rework LR using only the original single profile. The actual laboratory rework profile is used only afterwards for validation.

Let $E^{(r)}$ denote the original single profile, where $r \in \{1, 2, 3\}$ is the replicate treated as the starting profile. The observed laboratory rework profile is

$$E^{1:3} = \left(E^{(1)}, E^{(2)}, E^{(3)}\right). \tag{7.1}$$

In the simulation framework, the prediction is made from $E^{(r)}$. The observed profile $E^{1:3}$ is only used after the simulation to evaluate whether the prediction was calibrated.

7.1 Simulation target

The goal is to approximate the distribution of the LR that could be obtained after rework, conditional on the information available in the original profile. Since rework means two additional measurements, one simulated rework profile has the form

$$\widetilde{E}^{1:3}_{r,b} = \left(E^{(r)}, \widetilde{E}^{(r,1)}_{b}, \widetilde{E}^{(r,2)}_{b}\right), \tag{7.2}$$

where b denotes the simulation index. The first component is the observed original profile. The second and third components are artificial replicate profiles generated from the model.

For each original profile, the procedure is repeated B times. This gives simulated rework log-LRs

$$z_1, \ldots, z_B, \tag{7.3}$$

where

$$z_b = \log_{10} \mathrm{LR}_{\mathrm{sim},b}. \tag{7.4}$$

The empirical distribution

$$\mathcal{P}_{\mathrm{sim}} = \{z_1, \ldots, z_B\} \tag{7.5}$$

is the frequentist predictive distribution for the rework LR.

7.2 Original-profile deconvolution

The starting point is the original profile $E^{(r)}$. This profile is deconvolved with both contributors treated as unknown. In terms of the hypotheses from Section 2.2, this corresponds to the genotype model under H_2. The model parameters are estimated by maximum likelihood:

$$\hat{\boldsymbol{\theta}}_2^{(r)} = \arg\max_{\boldsymbol{\theta}} P(E^{(r)} \mid H_2, \boldsymbol{\theta}). \tag{7.6}$$

The same cleaned setting as in Chapters 5 and 6 is used: the fractional threshold is set to $f = 0.03$, peaks classified as drop-in have been removed, and the drop-in parameters are set close to zero.

Conditional on $\hat{\boldsymbol{\theta}}_2^{(r)}$, DNAStatistX computes the joint deconvolution distribution

$$P(g_{1,\ell}, g_{2,\ell} \mid E_\ell^{(r)}, H_2, \hat{\boldsymbol{\theta}}_2^{(r)}). \tag{7.7}$$

The simulation algorithm uses this joint deconvolution rather than the marginal deconvolutions. This is necessary because artificial profiles must be generated from a complete pair of contributor genotypes. Sampling the contributors independently from their marginal deconvolutions could break dependencies between the two contributors at a locus.

7.3 Sampling complete contributor genotypes

For simulation b, and for each locus ℓ, one joint genotype pair is sampled:

$$(g_{1,\ell}^{(b)}, g_{2,\ell}^{(b)}) \sim P(g_{1,\ell}, g_{2,\ell} \mid E_\ell^{(r)}, H_2, \hat{\boldsymbol{\theta}}_2^{(r)}). \tag{7.8}$$

Repeating this over loci gives a full sampled genotype pair

$$g^{(b)} = \left(g_1^{(b)}, g_2^{(b)}\right), \tag{7.9}$$

where

$$g_i^{(b)} = \left(g_{i,1}^{(b)}, \ldots, g_{i,L}^{(b)}\right). \tag{7.10}$$

The deconvolution output can contain the symbol $\emptyset$, representing an unobserved allele. Before generating artificial profiles, each $\emptyset$ is replaced by an actual allele. This replacement is sampled from the population allele frequencies restricted to alleles that were not observed at that locus. If both alleles are $\emptyset$, a complete genotype is sampled from the corresponding population genotype distribution over unobserved alleles.

After this step, each simulation has a complete genotype file for both unknown contributors.

7.4 Generating and analysing simulated rework profiles

For each sampled genotype pair $g^{(b)}$, two artificial replicate profiles are generated:

$$\widetilde{E}_b^{(r,1)}, \widetilde{E}_b^{(r,2)} \sim P(\, E \mid g_1^{(b)}, g_2^{(b)}, \hat{\boldsymbol{\theta}}_2^{(r)}). \tag{7.11}$$

The two artificial replicates are conditionally independent given the sampled genotypes and the fixed parameter estimate.

Together with the original profile, these artificial replicates form one simulated rework profile,

$$\widetilde{E}_{r,b}^{1:3} = \left(E^{(r)}, \widetilde{E}_b^{(r,1)}, \widetilde{E}_b^{(r,2)}\right). \tag{7.12}$$

This simulated rework profile is analysed in the same way as an observed rework profile. DNAStatistX is run on the three profiles together with both contributors treated as unknown. The model parameters for the simulated rework profile are estimated by

$$\hat{\boldsymbol{\theta}}_{2,b} = \arg\max_{\boldsymbol{\theta}} P(\widetilde{E}_{r,b}^{1:3} \mid H_2, \boldsymbol{\theta}). \tag{7.13}$$

Conditional on this parameter estimate, DNAStatistX returns the marginal deconvolution distribution

$$P(g_i \mid \widetilde{E}_{r,b}^{1:3}, H_2, \hat{\boldsymbol{\theta}}_{2,b}). \tag{7.14}$$

The simulated rework LR is computed using the same donor-specific LR approximation defined in Section 3.9, now applied to the simulated combined profile.

For contributor i, the simulated rework log-LR is denoted by

$$z_b = \log_{10} \mathrm{LR}_{\mathrm{sim},b}(g_i^{(b)}). \tag{7.15}$$

Repeating this for $b = 1, \ldots, B$ gives the predicted rework LR distribution $\mathcal{P}_{\mathrm{sim}}$ in Eq. (7.5).

7.5 Calibration percentile

The observed laboratory rework profile is used to evaluate the prediction. For the known true donor, the observed rework log-LR is written as

$$z_{\text{obs}} = \log_{10} \text{LR}_{\text{obs}}(g_i^{\text{true}}). \tag{7.16}$$

The observed value is placed inside the simulated predictive distribution. Its percentile is

$$p_{\text{sim}} = 100 \cdot \frac{1}{B} \sum_{b=1}^{B} \mathbf{1} \left\{ z_b \leq z_{\text{obs}} \right\}. \tag{7.17}$$

If the simulation framework is calibrated, the observed laboratory rework LR should behave like a draw from the predicted distribution. Across many mixtures, the values of p_{sim} should therefore be approximately uniformly distributed on $[0, 100]$. Algorithm 2 summarizes the complete frequentist simulation framework.

Algorithm 2 Frequentist rework simulation algorithm

Input: original profile $E^{(r)}$, number of simulations B
Estimate $\hat{\boldsymbol{\theta}}_2^{(r)}$ using Eq. (7.6)
Compute the joint deconvolution in Eq. (7.7) for all loci $\ell = 1, \ldots, L$
for $b = 1, \ldots, B$ **do**
    for each locus $\ell = 1, \ldots, L$ **do**
        Sample $(g_{1,\ell}^{(b)}, g_{2,\ell}^{(b)})$ using Eq. (7.8)
        Replace $\emptyset$ alleles by alleles sampled from the population distribution over unobserved alleles
    end for
    Construct $g_1^{(b)}$ and $g_2^{(b)}$ as in Eq. (7.10)
    Generate two artificial replicate profiles using Eq. (7.11)
    Define the simulated rework profile $\widetilde{E}_{r,b}^{1:3}$ as in Eq. (7.12)
    Estimate $\hat{\boldsymbol{\theta}}_{2,b}$ using Eq. (7.13)
    Obtain the marginal deconvolution distribution in Eq. (7.14)
    Compute z_b using the donor-specific LR approximation from Eq. (3.59)
end for
Form the predictive distribution $\mathcal{P}_{\text{sim}}$ in Eq. (7.5)
Compare $\mathcal{P}_{\text{sim}}$ with z_{obs} using Eq. (7.17)
Output: predicted rework LR distribution and calibration percentile

8. Frequentist simulation results

This chapter evaluates the frequentist rework simulation framework introduced in Chapter 7. The framework predicts, from one original single profile, a distribution for the minor-donor LR after rework. The predicted distribution is compared with the observed true-donor LR computed from the observed rework profile $E^{1:3}$. Only the minor donor is considered in this chapter. Since only 100 simulation samples are used, the same major-donor genotype is usually sampled repeatedly. The major-donor simulation results are therefore less informative for assessing calibration.

The results show that the frequentist simulation approach is not well calibrated for the minor donor. A possible explanation is that the method uses maximum likelihood estimates as fixed plug-in values. When the mixture proportion estimated from the original single profile differs strongly from the mixture proportion estimated from the observed rework profile, the prediction tends to be worse. This suggests that, for some profiles, the likelihood may not be sharp enough around the MLE to justify conditioning on one fixed estimate. The Bayesian approach in the next chapters is introduced as an attempt to address this issue by propagating parameter uncertainty.

8.1 Predicted rework LR distributions

For each original profile, the frequentist simulation algorithm generates 100 simulated rework profiles. Each simulated rework profile gives one simulated minor-donor rework LR. This gives a predicted rework LR distribution conditional on the original profile and on the maximum-likelihood parameter estimates obtained from that profile.

The figures in this section compare three quantities. The red vertical line is the observed true-donor rework LR, $z_{\text{true}} = \log_{10} LR_{\text{rework}}(g_i^{\text{true}})$. The blue distribution is obtained by sampling donor genotypes from the deconvolution of the actual observed rework profile $E^{1:3}$. This distribution is not a prediction from the original profile, because it uses the observed rework data. Instead, it is a diagnostic distribution: it shows whether the true-donor rework LR is typical among donor genotypes sampled from the deconvolution after rework has actually been observed. This is the diagnostic studied in Chapter 6.

The green or purple distribution is the predicted rework LR distribution produced by the frequentist simulation algorithm. This distribution is computed using only the original single profile $E^{(r)}$. It is therefore the distribution that is used to assess the predictive performance of the simulation algorithm.

Two different percentiles can be distinguished. Let

$$\mathcal{D}_{\text{obs}} = z_1^{\text{obs}}, \ldots, z_{B_{\text{obs}}}^{\text{obs}} \tag{8.1}$$

denote the blue observed-rework distribution, and let

$$\mathcal{D}_{\text{pred}}^{(r)} = z_{r,1}^{\text{pred}}, \ldots, z_{r,B}^{\text{pred}} \tag{8.2}$$

denote the predicted rework distribution from original profile $E^{(r)}$. The observed-rework percentile is

$$p_{\text{obs}} = 100 \cdot \frac{1}{B_{\text{obs}}} \sum_{b=1}^{B_{\text{obs}}} \mathbf{1}\{z_b^{\text{obs}} \leq z_{\text{true}}\}, \tag{8.3}$$

whereas the predicted-distribution percentile is

$$p_{\text{pred}}^{(r)} = 100 \cdot \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}\{z_{r,b}^{\text{pred}} \leq z_{\text{true}}\}. \tag{8.4}$$

The percentile histogram in Section 8.2 shows $p_{\text{pred}}^{(r)}$. Thus, Figure 8.4 evaluates whether the predicted rework LR distribution from the original profile contains the observed true-donor rework LR at approximately uniform percentiles.

Figure 8.1 shows an example where the observed true-donor rework LR is consistent with both distributions. The red line lies in the central part of the blue observed-rework distribution and also near the centre of the purple predicted distribution. In this case, the true-donor rework LR is typical after the actual rework profile has been observed, and the simulation from the original profile also predicts it well.

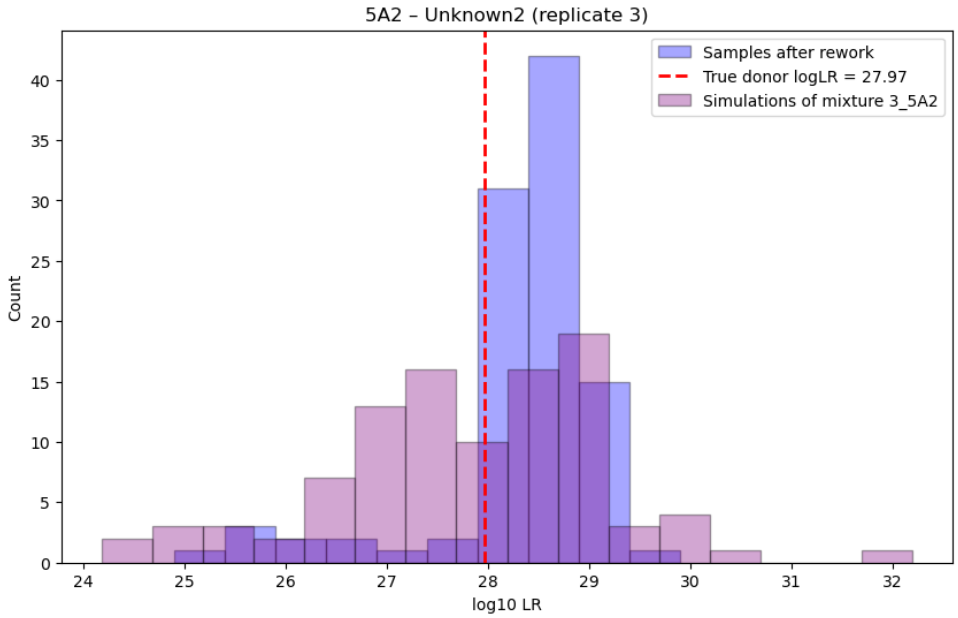


Figure 8.1: Example of a frequentist rework simulation for the minor donor where the observed true-donor rework LR is typical for both the observed-rework distribution and the predicted rework distribution. The blue distribution is obtained from the deconvolution of the actual observed rework profile. The purple distribution is the predicted rework LR distribution obtained by simulating rework from the original single profile. The red line has predicted-distribution percentile $p_{\text{pred}} = 50$.

Figures 8.2 and 8.3 show examples where the frequentist simulation prediction is poor, but for different reasons. In Figure 8.2, the observed true-donor rework LR lies below almost all values in the green predicted distribution, so $p_{\text{pred}} = 0$. However, the red line is also unusually low relative to the blue observed-rework distribution. This means that the true-donor rework LR is already atypical after conditioning on the observed rework profile itself. In such a case, an extreme predicted-distribution percentile cannot be attributed only to the simulation from the original profile: it is also related to the true donor being atypical in the observed-rework deconvolution distribution.

In Figure 8.3, the situation is different. The observed true-donor rework LR lies near the upper edge of the blue observed-rework distribution, but the green predicted distribution is shifted far to the left. Here, the true-donor LR is not as clearly atypical for the observed rework profile itself. The main problem is that the simulation from the original profile predicts rework LRs that are too low compared with the distribution obtained after observing the actual rework profile.

8.2 Calibration of the frequentist predictions

Calibration of the simulation algorithm is assessed using the predicted-distribution percentile $p_{\text{pred}}^{(r)}$. For each original profile $E^{(r)}$, the observed true-donor rework LR is placed within the

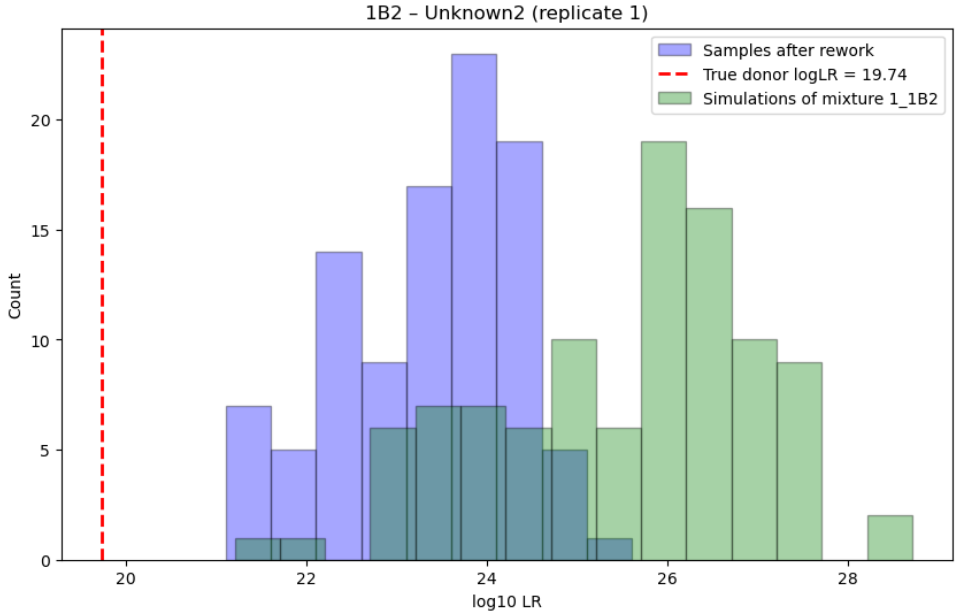


Figure 8.2: Example of a frequentist rework simulation for the minor donor where the observed true-donor rework LR lies below the predicted rework LR distribution. The blue distribution is obtained from the observed rework profile, while the green distribution is predicted from the original single profile. The red line has predicted-distribution percentile $p_{\text{pred}} = 0$. It is also low relative to the blue distribution, indicating that the true-donor LR is already atypical for the observed-rework deconvolution.

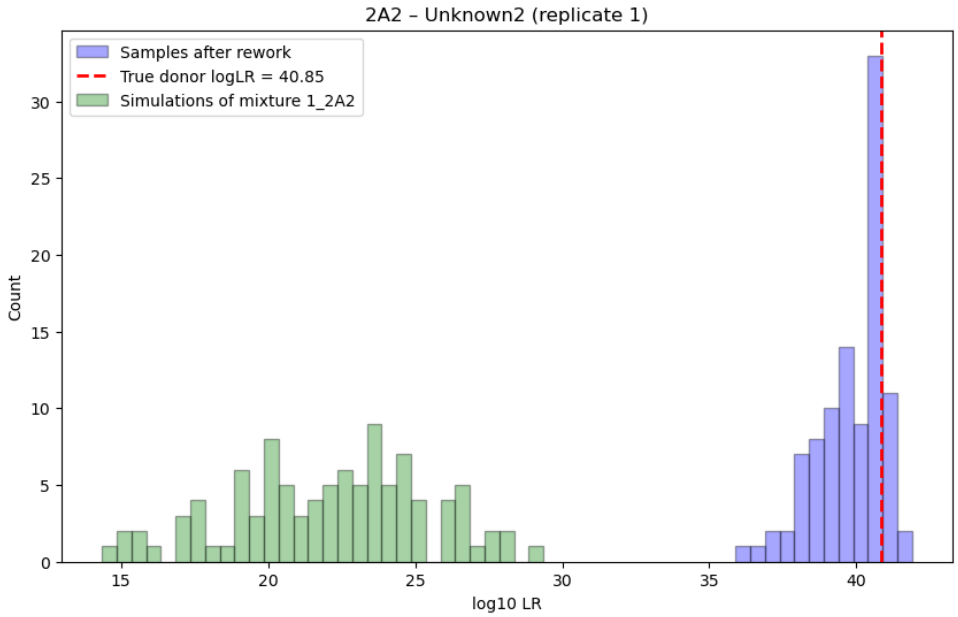


Figure 8.3: Example of a frequentist rework simulation for the minor donor where the predicted rework LR distribution is shifted away from the observed-rework distribution. The blue distribution is obtained from the observed rework profile, while the green distribution is predicted from the original single profile. The red line has predicted-distribution percentile $p_{\text{pred}} = 100$, because the predicted distribution assigns almost all simulated rework LRs below the observed true-donor rework LR.

corresponding predicted rework LR distribution obtained from simulated rework profiles. A uniform histogram of $p_{\text{pred}}^{(r)}$ values would indicate that the observed true-donor rework LR behaves like a typical draw from the predicted distribution.

Figure 8.4 shows that the predicted-distribution percentiles are not uniformly distributed. In particular, many observed true-donor rework LRs are located close to the lower or upper edge of

the predicted distributions. The examples in Figures 8.2 and 8.3 illustrate these two extremes: in Figure 8.2, the observed true-donor rework LR is below almost all predicted values, while in Figure 8.3, it is above almost all predicted values.

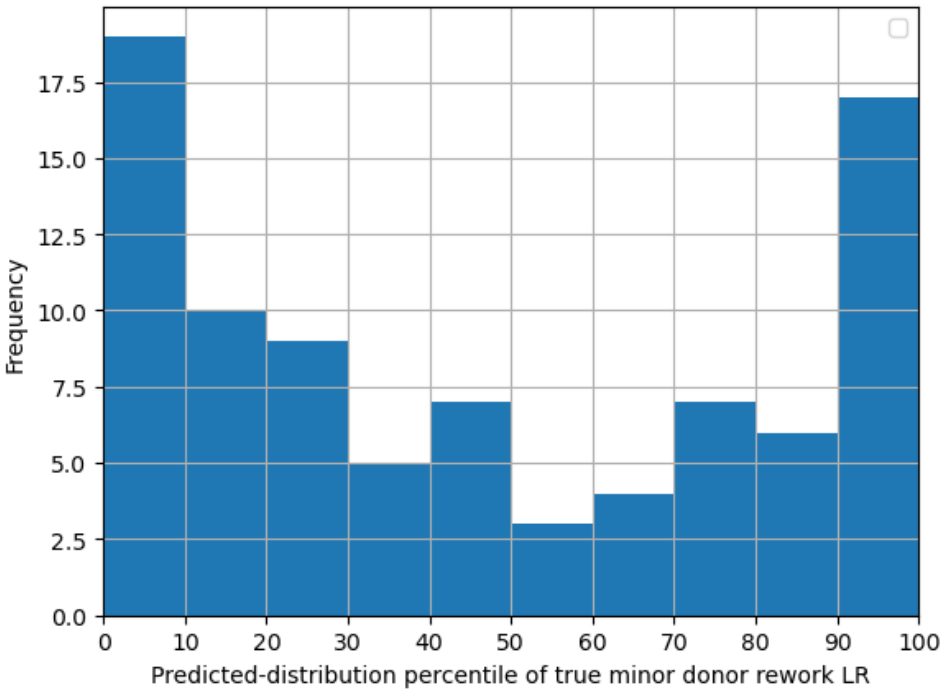


Figure 8.4: Percentile validation for frequentist rework simulations for the minor donor. For each original profile, the observed true-donor rework LR is placed within the corresponding predicted rework LR distribution obtained from simulated rework profiles. The histogram therefore shows p_{pred}. A calibrated prediction method would give an approximately uniform histogram. The Kolmogorov–Smirnov statistic of this histogram against the uniform distribution is 0.162. The 95%-coverage is 69.0% and the mean interval score is 50.5.

The empirical coverage of the 95% prediction interval gives the same indication. For the minor donor, the 95% interval covers on average 69.0% of the observed true-donor LRs after rework. This means that, among the 87 two-person profiles considered, 69.0% had the observed true-donor rework LR inside the corresponding 95% prediction interval. This is below the nominal 95% level and indicates that the observed true-donor rework LR is too often outside the predicted interval.

8.3 Mixture-proportion estimates

A possible cause of the poor calibration is the use of maximum likelihood estimates as fixed plug-in parameters. In the frequentist simulation, the nuisance parameters are estimated from the original single profile and then treated as known when simulating rework. However, the original profile is only one stochastic measurement of the underlying mixture. If the single-profile MLE is not representative of the profiles that would be obtained after replicate analysis, the simulated rework profiles may be generated under parameter values that are too specific to the original replicate.

This effect is especially relevant for the mixture proportion, because the LR calculation is strongly influenced by the estimated contributor proportions. Figure 8.5 compares the absolute difference between the single-profile and rework-profile MLE of the minor-donor mixture proportion with the 95% interval score for the minor-donor prediction. The vertical axis is shown as $\log_{10}(1 + \text{interval score})$, so that large interval scores remain visible without compressing the lower range too strongly. For two-person mixtures, the absolute difference is the same for the

minor-donor and major-donor mixture proportions, since the two proportions sum to one.

The figure shows a positive association: profiles for which the mixture-proportion MLE changes substantially after adding the replicate measurements tend to have larger interval scores. The fitted least-squares line is $y = 0.998 + 11.477x$, where x is the absolute difference between the single-profile and rework-profile MLE and $y = \log_{10}(1 + \text{interval score})$. This suggests that poor predictions are partly associated with instability of the mixture-proportion estimate. In such cases, the original single profile does not determine the mixture proportion reliably, but the frequentist simulation still conditions on one fixed MLE.

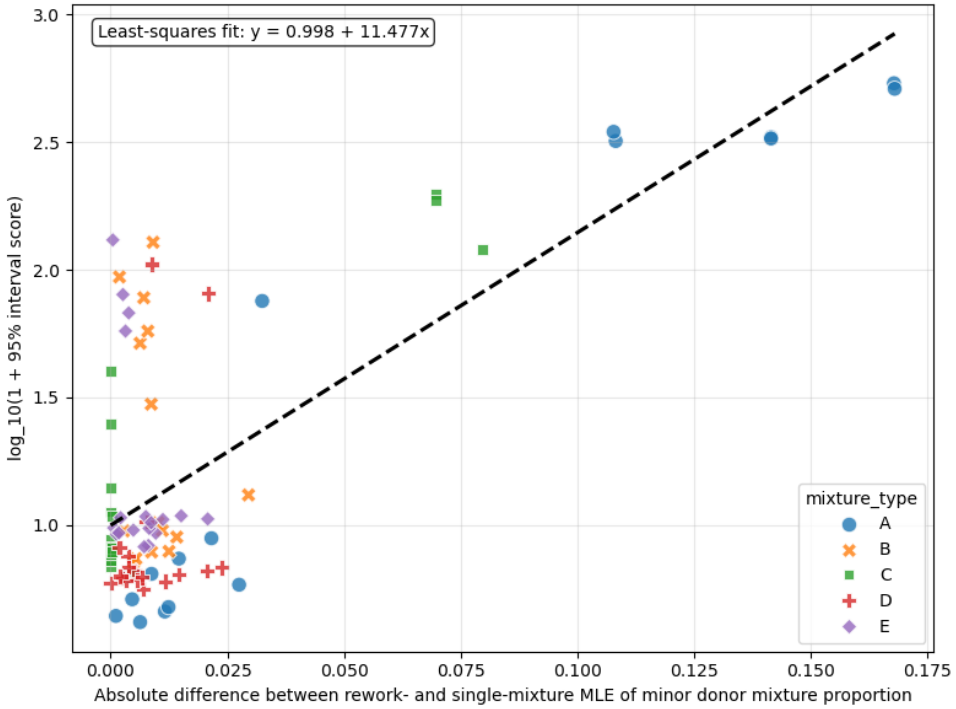


Figure 8.5: Relationship between the absolute difference between the single-profile and rework-profile MLE of the minor-donor mixture proportion, and the transformed 95% interval score for the minor-donor rework prediction. The vertical axis is $\log_{10}(1+\text{interval score})$. The dashed line shows the least-squares fit. For two-person mixtures, the absolute difference is the same for the major-donor mixture proportion.

This result is consistent with the interpretation that parameter uncertainty should not be ignored. If the mixture proportion estimated from the original profile is uncertain or unstable, then a predictive simulation based on only the point estimate can be shifted away from the observed rework result. A Bayesian approach can address this more naturally by sampling nuisance parameters from their posterior distribution instead of fixing them at a single MLE.

At the same time, Figure 8.5 also shows that mixture-proportion instability is not a complete explanation of the poor calibration. Some profiles have small changes in the mixture-proportion MLE but still have large interval scores. For these profiles, other sources of uncertainty, such as genotype uncertainty, peak-height variability, artefact modelling, or the approximation used in the LR calculation, may also contribute to the poor predictions.

9. MCMC implementation

The frequentist analyses in the previous chapters use a single maximum-likelihood estimate of the model parameters. This is computationally efficient, but it ignores uncertainty in these parameters. The aim of the MCMC implementation is to sample from the posterior distribution of the model parameters, so that this uncertainty can be propagated to the Bayesian LR and to the Bayesian rework simulation framework in later chapters.

The implementation in this thesis is written for two-person mixtures. Under H_2, both contributors are unknown. Following the notation from Chapter 3, the model parameter vector is

$$\boldsymbol{\theta} = (\mu, \sigma, \beta, \phi_1), \qquad \phi_2 = 1 - \phi_1. \tag{9.1}$$

Here μ is the expected peak-height scale, σ controls peak-height variability, β controls degradation, and ϕ_i is the mixture proportion of contributor i. The second mixture proportion ϕ_2 is not sampled as an independent variable, but is determined by ϕ_1. The drop-in probability C, the drop-in rate λ, and the locus-level detection thresholds T_ℓ are kept fixed at the values used in the DNAStatistX configuration.

9.1 Posterior target

For a fixed value of $\boldsymbol{\theta}$, the likelihood of a mixture profile is obtained by marginalising over the unknown contributor genotypes. For locus ℓ, the exact locus-level marginal likelihood under H_2 is

$$p(E_\ell \mid \boldsymbol{\theta}, H_2) = \sum_{g_{1,\ell} \in G_\ell} \sum_{g_{2,\ell} \in G_\ell} p(E_\ell \mid g_{1,\ell}, g_{2,\ell}, \boldsymbol{\theta}, H_2) P(g_{1,\ell}) P(g_{2,\ell}). \tag{9.2}$$

Assuming independence between loci, the full marginal likelihood is

$$p(E \mid \boldsymbol{\theta}, H_2) = \prod_{\ell=1}^{L} p(E_\ell \mid \boldsymbol{\theta}, H_2). \tag{9.3}$$

The posterior distribution targeted by the MCMC sampler is therefore

$$p(\boldsymbol{\theta} \mid E, H_2) \propto p(E \mid \boldsymbol{\theta}, H_2) \pi(\boldsymbol{\theta}), \tag{9.4}$$

where $\pi(\boldsymbol{\theta})$ denotes the prior density of the model parameters.

The peak-height likelihood $p(E_\ell \mid g_{1,\ell}, g_{2,\ell}, \boldsymbol{\theta}, H_2)$ uses the same gamma peak-height model as in Chapter 3. For allele a at locus ℓ, the total degraded contribution is

$$\alpha_{\ell,a}^{\deg} = (\phi_1 n_{1,\ell,a} + \phi_2 n_{2,\ell,a})\, \beta^{(f_{\ell,a} - 125)/100}, \tag{9.5}$$

where $n_{i,\ell,a} \in \{0, 1, 2\}$ is the allele-copy count of contributor i, and $f_{\ell,a}$ is the fragment length of allele a at locus ℓ. If $\alpha_{\ell,a}^{\deg} > 0$, the peak height is modelled using the gamma distribution from Chapter 3. If an observed allele is not explained by the contributor genotypes, it is handled through the fixed drop-in model.

9.2 Restricting the genotype space

A direct MCMC implementation would be too slow if Eq. 9.2 were evaluated over the full genotype-pair space at every MCMC iteration. The genotype summation must be recomputed for thousands of proposed parameter values and is therefore the main computational bottleneck.

To make the calculation feasible, the DNAStatistX deconvolution output under H_2 was used to restrict the genotype space. For each locus, the MLE deconvolution table from DNAStatistX

gives posterior probabilities for genotype pairs of the two unknown contributors. Genotype pairs with probability below a fixed threshold, for example 10^{-9}, were removed before running the MCMC sampler. In practice, this meant that only approximately the 10% most probable genotype pairs were retained. This substantially reduced computation time, because the likelihood no longer had to be evaluated over genotype pairs with negligible support under the initial deconvolution.

Let

$$\widetilde{G}_\ell = \{(g_{1,\ell}, g_{2,\ell}) \in G_\ell \times G_\ell : P_{\mathrm{MLE}}(g_{1,\ell}, g_{2,\ell} \mid E_\ell, H_2) \geq \tau_{\mathrm{geno}}\} \tag{9.6}$$

denote the retained set of genotype pairs for locus ℓ, where τ_{geno} is the filtering threshold. In the MCMC implementation, Eq. 9.2 is approximated by

$$p(E_\ell \mid \boldsymbol{\theta}, H_2) \approx \sum_{(g_{1,\ell}, g_{2,\ell}) \in \widetilde{G}_\ell} p(E_\ell \mid g_{1,\ell}, g_{2,\ell}, \boldsymbol{\theta}, H_2) P(g_{1,\ell}) P(g_{2,\ell}). \tag{9.7}$$

This is an approximation, because genotype pairs outside $\widetilde{G}_\ell$ are ignored. The choice of τ_{geno} is somewhat arbitrary and reflects a trade-off between accuracy and runtime. A lower threshold retains more genotype pairs, but increases computation time. A higher threshold speeds up the MCMC run, but may remove genotype pairs that still have some posterior support. The same retained genotype-pair sets are also used later when constructing Bayesian deconvolution tables from the MCMC output.

The unobserved allele category $\emptyset$ is included in the genotype representation, as in the deconvolution calculations from the previous chapters. This allows genotype pairs to represent alleles that are not observed above the detection threshold.

9.3 Priors

The MCMC implementation uses weakly informative bounded priors for the model parameters. In the runs shown in this chapter, the bounds are

$$\mu_{\max} = 40000, \qquad \sigma_{\max} = 1. \tag{9.8}$$

The prior density is specified, up to a proportionality constant, as

$$\pi(\boldsymbol{\theta}) \propto \mathbf{1}\{0 < \mu < \mu_{\max}\}\mathbf{1}\{0 < \sigma < \sigma_{\max}\}\mathbf{1}\{0 < \beta < 1\}\phi_1 \mathbf{1}\{0.5 < \phi_1 < 1\}. \tag{9.9}$$

Thus, μ, σ and β have uniform priors on their allowed intervals. The prior density of a Beta$(2, 1)$ random variable is

$$p(x) = 2x, \qquad 0 < x < 1. \tag{9.10}$$

After restricting this prior to the interval $(0.5, 1)$, the density is still proportional to x on this interval. Since constants do not affect the Metropolis–Hastings acceptance ratio, this gives the factor

$$\phi_1 \mathbf{1}\{0.5 < \phi_1 < 1\} \tag{9.11}$$

in Eq. 9.9.

The restriction $\phi_1 > 0.5$ is a practical way to handle the symmetry between the two unknown contributors. Without this restriction, the likelihood has two symmetric solutions: one where the first contributor has mixture proportion ϕ_1, and one where the contributor labels are exchanged. Sampling both symmetric regions is unnecessary for the simulation framework. Therefore, only the part of the parameter space with $\phi_1 > 0.5$ is sampled. In this chapter, ϕ_1 should be interpreted as the larger mixture proportion, not as the mixture proportion of a named contributor. The second mixture proportion is then determined by $\phi_2 = 1 - \phi_1$.

The priors in Eq. 9.9 were chosen to obtain a workable MCMC implementation for the simulation framework. No systematic study was performed to determine optimal priors. The sensitivity of the Bayesian LR and the Bayesian rework simulations to these prior choices is therefore an important topic for future research.

9.4 Transformation to unrestricted variables

The model parameters are restricted to fixed intervals: $\mu \in (0, \mu_{\max})$, $\sigma \in (0, \sigma_{\max})$, $\beta \in (0, 1)$, and $\phi_1 \in (0.5, 1)$. A random-walk proposal directly on these parameters could propose values outside the allowed range. To avoid this, the MCMC sampler operates on unrestricted variables and transforms them to valid model-parameter values.

Let

$$s(z) = \frac{1}{1 + \exp(-z)} \tag{9.12}$$

be the logistic function. The unrestricted variables are

$$\mathbf{z} = (z_\mu, z_\sigma, z_\phi, z_\beta), \tag{9.13}$$

and these are transformed to the model parameters by

$$\begin{aligned} \mu &= \mu_{\max} s(z_\mu), \\ \sigma &= \sigma_{\max} s(z_\sigma), \\ \phi_1 &= 0.5 + 0.5 s(z_\phi), \\ \beta &= s(z_\beta), \\ \phi_2 &= 1 - \phi_1. \end{aligned} \tag{9.14}$$

This transformation maps every value of $\mathbf{z} \in \mathbb{R}^4$ to an admissible parameter vector $\boldsymbol{\theta}$.

Since the Metropolis–Hastings algorithm is run in $\mathbf{z}$-space, the target density must include the Jacobian of the transformation. The Jacobian corrects for the change in volume when transforming from the unrestricted variables $\mathbf{z}$ to the bounded model parameters $\boldsymbol{\theta}$. In one dimension, this correction is the absolute value of the derivative. In multiple dimensions, it is the absolute value of the determinant of the matrix of first derivatives:

$$\left| \frac{\partial \boldsymbol{\theta}}{\partial \mathbf{z}} \right| = \left| \det \left(\frac{\partial \theta_i}{\partial z_j} \right)_{i,j} \right|. \tag{9.15}$$

In this implementation, each transformed parameter depends only on its own unrestricted variable. Therefore, the derivative matrix is diagonal and the Jacobian is the product of the four one-dimensional derivative terms.

The target density in $\mathbf{z}$-space is

$$p(\mathbf{z} \mid E, H_2) \propto p(E \mid \boldsymbol{\theta}(\mathbf{z}), H_2) \pi(\boldsymbol{\theta}(\mathbf{z})) \left| \frac{\partial \boldsymbol{\theta}}{\partial \mathbf{z}} \right|. \tag{9.16}$$

The derivatives of the transformations in Eq. 9.14 are

$$\frac{\partial \mu}{\partial z_\mu} = \mu_{\max} s(z_\mu)\{1 - s(z_\mu)\}, \tag{9.17}$$

$$\frac{\partial \sigma}{\partial z_\sigma} = \sigma_{\max} s(z_\sigma)\{1 - s(z_\sigma)\}, \tag{9.18}$$

$$\frac{\partial \phi_1}{\partial z_\phi} = 0.5 s(z_\phi)\{1 - s(z_\phi)\}, \tag{9.19}$$

and

$$\frac{\partial \beta}{\partial z_\beta} = s(z_\beta)\{1 - s(z_\beta)\}. \tag{9.20}$$

Because each transformed parameter depends only on its own unrestricted variable, the Jacobian matrix is diagonal, up to the ordering of the parameters. Its determinant is therefore the product of these four derivatives:

$$\begin{aligned}\left|\frac{\partial \boldsymbol{\theta}}{\partial \mathbf{z}}\right| = \mu_{\max} s(z_\mu)\{1 - s(z_\mu)\} \cdot \sigma_{\max} s(z_\sigma)\{1 - s(z_\sigma)\} \\ \cdot 0.5 s(z_\phi)\{1 - s(z_\phi)\} \cdot s(z_\beta)\{1 - s(z_\beta)\}.\end{aligned} \tag{9.21}$$

Taking the logarithm of this product gives the log-Jacobian contribution

$$\begin{aligned}\log\left|\frac{\partial \boldsymbol{\theta}}{\partial \mathbf{z}}\right| = \; & \log(\mu_{\max}) + \log s(z_\mu) + \log\{1 - s(z_\mu)\} \\ & + \log(\sigma_{\max}) + \log s(z_\sigma) + \log\{1 - s(z_\sigma)\} \\ & + \log(0.5) + \log s(z_\phi) + \log\{1 - s(z_\phi)\} \\ & + \log s(z_\beta) + \log\{1 - s(z_\beta)\}.\end{aligned} \tag{9.22}$$

The log target used by the sampler is therefore the log marginal likelihood, plus the log prior from Eq. 9.9, plus the log-Jacobian term from Eq. 9.22.

9.5 Metropolis–Hastings sampler

A random-walk Metropolis–Hastings sampler was used. At iteration t, a proposal is generated on the unrestricted scale:

$$\mathbf{z}^* = \mathbf{z}^{(t)} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \Sigma_{\text{prop}}), \tag{9.23}$$

where Σ_{prop} is diagonal and contains separate proposal variances for the unrestricted variables. The proposed value is accepted with probability

$$\alpha = \min\left\{1, \frac{p(\mathbf{z}^* \mid E, H_2)}{p(\mathbf{z}^{(t)} \mid E, H_2)}\right\}. \tag{9.24}$$

After burn-in, the retained values are transformed back to the original parameter scale using Eq. 9.14. These retained samples are used as approximate draws from $p(\boldsymbol{\theta} \mid E, H_2)$.

Table 9.1 summarizes the MCMC settings used for the runs shown in this chapter. The initial values were taken from the DNAStatistX MLE under H_2, with the contributor labels chosen such that $\phi_1 > 0.5$.

Setting	Value
Sampler	Random-walk Metropolis–Hastings
Sampling scale	Unrestricted scale $\mathbf{z} = (z_\mu, z_\sigma, z_\phi, z_\beta)$
Number of iterations	15000
Burn-in period	5000 iterations
Number of retained samples	10000
Proposal distribution	$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \Sigma_{\text{prop}})$
Proposal scales	$(0.12, 0.10, 0.10, 0.10)$ for $(z_\mu, z_\sigma, z_\phi, z_\beta)$
Initial values	DNAStatistX MLE under H_2
Parameter bounds	$\mu_{\max} = 40000$, $\sigma_{\max} = 1$, $0 < \beta < 1$, $0.5 < \phi_1 < 1$
Genotype-pair filtering threshold	$\tau_{\text{geno}} = 10^{-9}$
Fixed quantities	Drop-in probability C, drop-in rate λ, and locus-level detection thresholds T_ℓ
Number of contributors	2

Table 9.1: MCMC settings used for the posterior parameter sampling in Chapter 9.

The same implementation can also be used for observed rework profiles. In that case, the three replicate profiles are assumed to share the same model parameter vector $\boldsymbol{\theta}$. The fixed

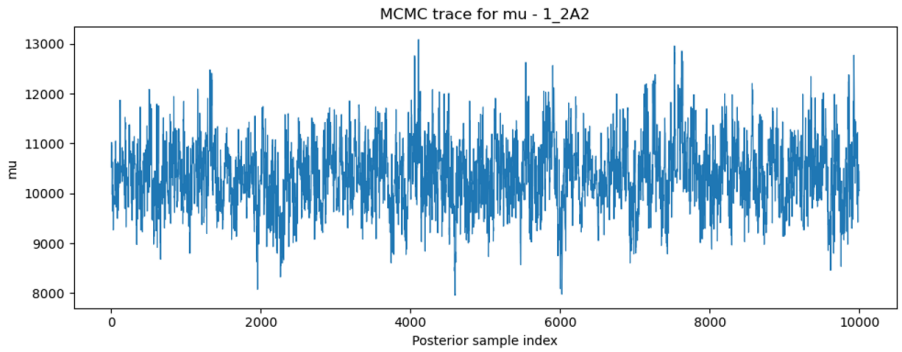


Figure 9.1: Trace plot after burn-in of the retained posterior samples for μ for mixture 1_2A2.

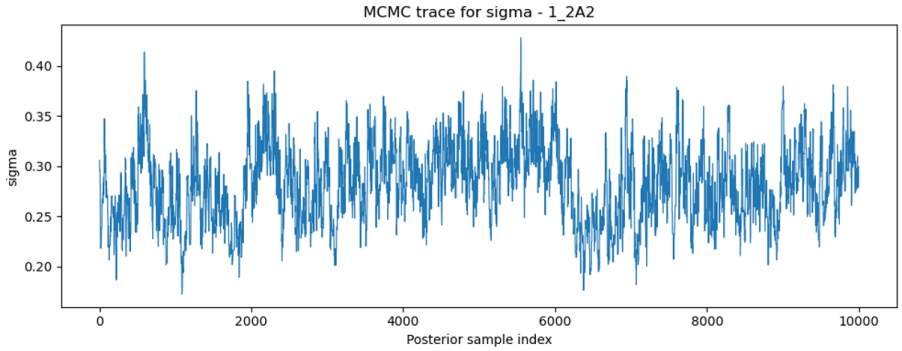


Figure 9.2: Trace plot after burn-in of the retained posterior samples for σ for mixture 1_2A2.

detection thresholds may differ between loci and replicates, but are not sampled. The log posterior becomes

$$\log p(\boldsymbol{\theta} \mid E^{1:3}, H_2) = \text{constant} + \log \pi(\boldsymbol{\theta}) + \sum_{r=1}^{3} \log p(E^{(r)} \mid \boldsymbol{\theta}, H_2). \tag{9.25}$$

Thus, the rework MCMC implementation differs only in the likelihood term: the marginal log likelihoods of the three replicate profiles are added.

9.6 Example: mixture 1_2A2

Figures 9.1–9.4 show trace plots for mixture 1_2A2. These plots are used as diagnostics for the posterior sampling and to illustrate what information is gained by using a Bayesian treatment of the model parameters.

For μ, σ and β, the posterior samples are concentrated around a central value and the marginal distributions are roughly symmetric.

The most important difference is visible for the mixture proportion. The frequentist approach gives a single MLE of approximately 0.50, suggesting an almost balanced mixture. The posterior samples for ϕ_1, however, show that mixture proportions between roughly 0.50 and 0.70 are plausible for this profile. This uncertainty is relevant, because downstream deconvolution and LR calculations can depend strongly on whether a contributor is treated as an equal contributor or as a moderate major contributor. A point estimate at 0.50 would miss this uncertainty.

Figures 9.5 and 9.6 show the dependence between the sampled parameters. The deterministic relation $\phi_2 = 1 - \phi_1$ gives a correlation of -1 between ϕ_1 and ϕ_2. The correlation matrix also shows a clear negative correlation between μ and β, and between σ and ϕ_1. The marginal histograms in Figure 9.6 are included only as visual summaries of the sampled posterior dis-

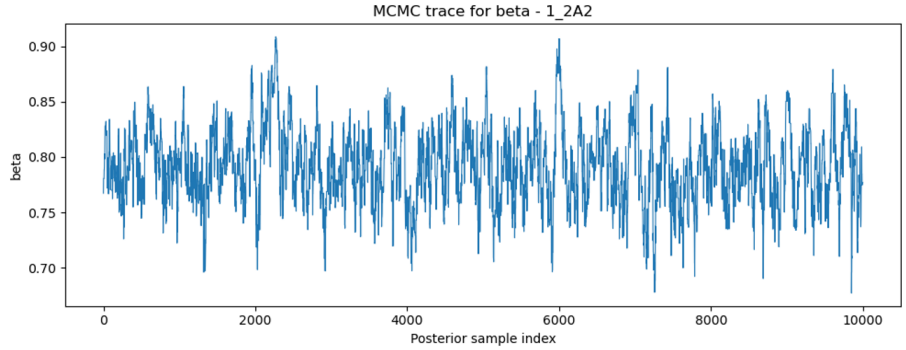


Figure 9.3: Trace plot after burn-in of the retained posterior samples for the degradation parameter β for mixture 1_2A2.

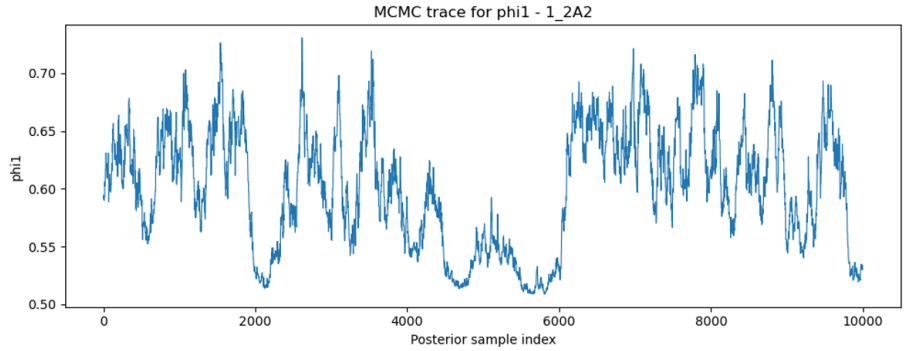


Figure 9.4: Trace plot after burn-in of the retained posterior samples for the larger mixture proportion ϕ_1 for mixture 1_2A2.

tribution. The actual Bayesian calculations in later chapters use the retained joint MCMC samples.

The trace plots show that the random-walk Metropolis–Hastings sampler explores a plausible posterior range for this mixture. The MCMC implementation should be interpreted as a proof of concept for propagating model-parameter uncertainty in the simulation framework, not as an optimized MCMC method. No systematic comparison of priors, proposal scales, burn-in choices or convergence diagnostics was performed. Further research is needed to determine which MCMC settings are most appropriate for routine use and how sensitive the final Bayesian predictions are to these implementation choices.

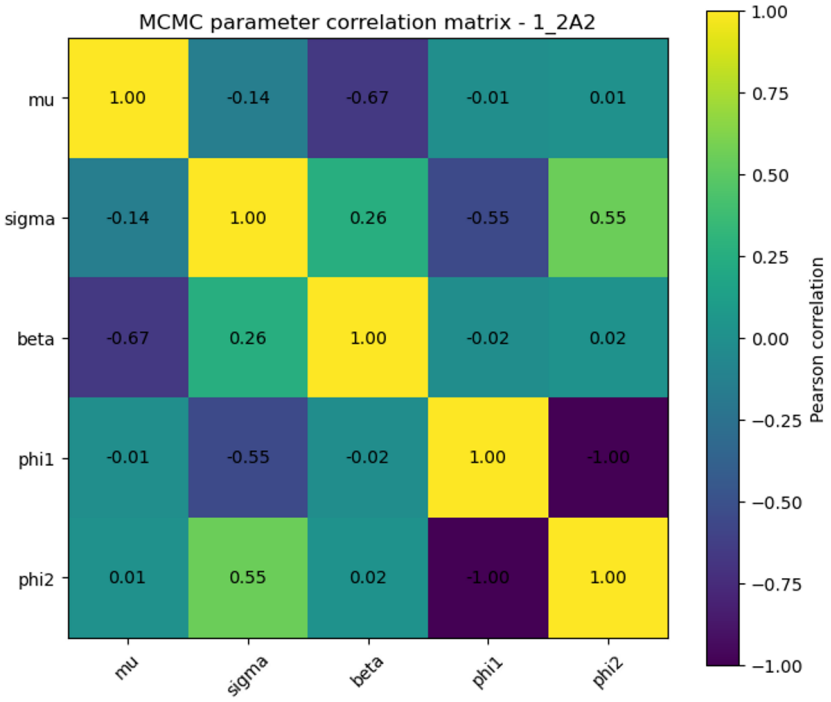


Figure 9.5: Pearson correlation matrix of the retained posterior parameter samples for mixture 1_2A2. The correlation of -1 between ϕ_1 and ϕ_2 is deterministic because $\phi_2 = 1 - \phi_1$.

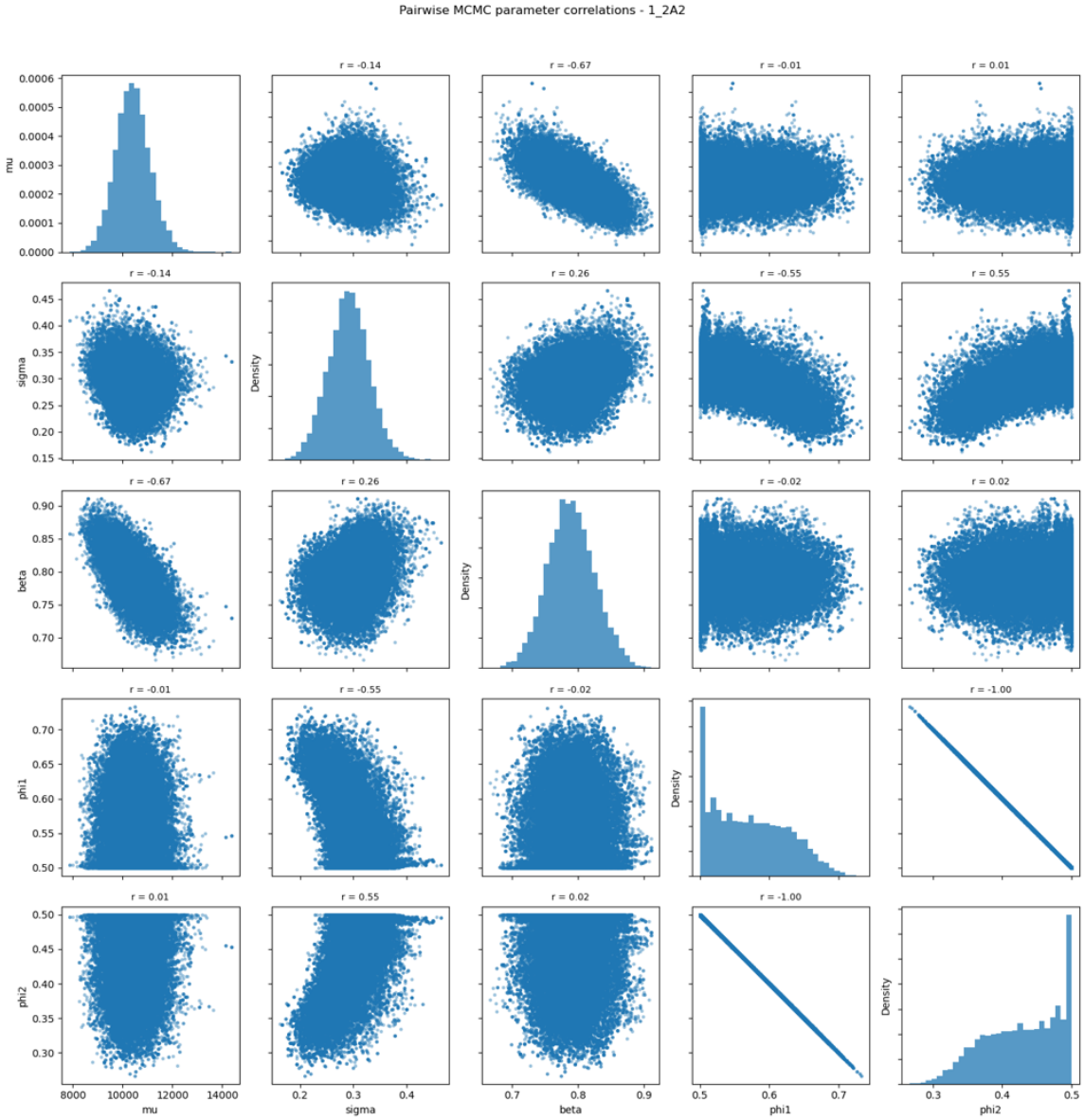


Figure 9.6: Pairwise scatter plots and marginal histograms of the retained posterior parameter samples for mixture 1_2A2.

10. Bayesian single and rework profiles

The previous chapters used a frequentist implementation of the peak-height model, where nuisance parameters were estimated by maximum likelihood and then treated as fixed. This chapter studies the corresponding Bayesian implementation, in which uncertainty in the nuisance parameters is propagated through the LR calculation. As in the previous validation chapters, all true-donor and sampled-donor results in this chapter refer only to the minor donors in the two-person mixtures. Major donors are not included in the comparisons or validation results.

We compare Bayesian and frequentist LRs for observed single and rework profiles, and validate the Bayesian deconvolution sampling method for both profile types.

10.1 True-donor LR comparisons

Figure 10.1 compares the Bayesian and frequentist LRs for cleaned single profiles, using only the minor true donors. For most minor donors in the two-person research dataset, the two methods give very similar LRs. Most points lie close to the diagonal, showing that the Bayesian implementation usually does not strongly change the single-profile LR.

The main exceptions are several A- and C-mixtures. These are the same mixtures for which Figure 8.5 showed atypical single-profile MLEs relative to the rework-profile MLEs. The larger changes in Figure 10.1 are therefore expected. They occur in cases where the frequentist point estimate is least representative, so that integrating over parameter uncertainty can have a visible effect.

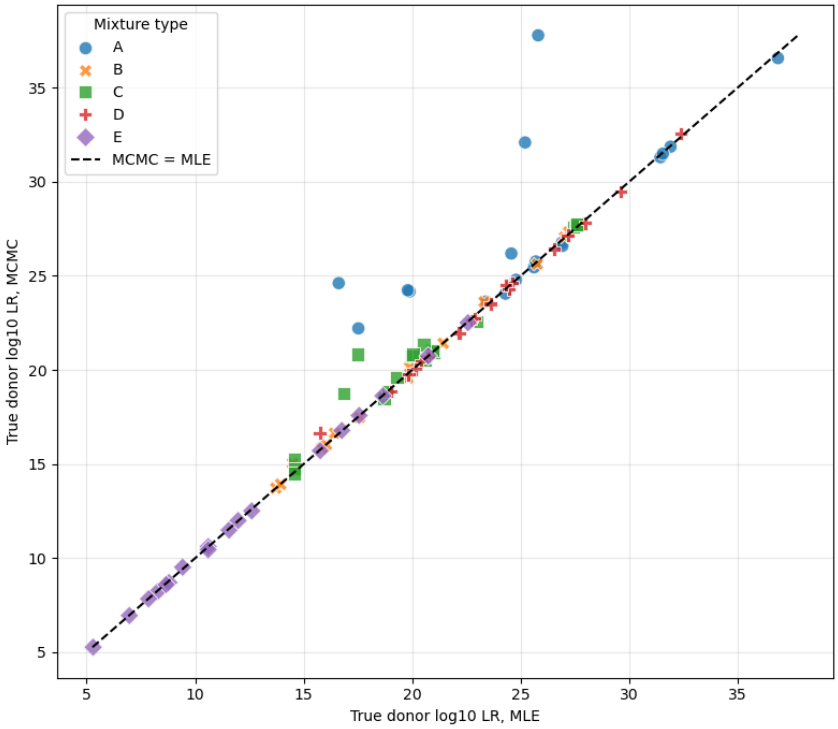


Figure 10.1: Comparison of Bayesian and frequentist single-profile $\log_{10}$-LRs for the minor true donors. Each point represents one cleaned single profile. Points close to the diagonal indicate similar Bayesian and frequentist LR values.

Figure 10.2 gives the same comparison for observed rework profiles, again using only the minor true donors. Here the differences are smaller. Most points again lie close to the diagonal, and there are fewer large deviations than for the single profiles. This suggests that the combined

replicate profile gives more stable parameter estimates, so that Bayesian integration has less influence on the final LR.

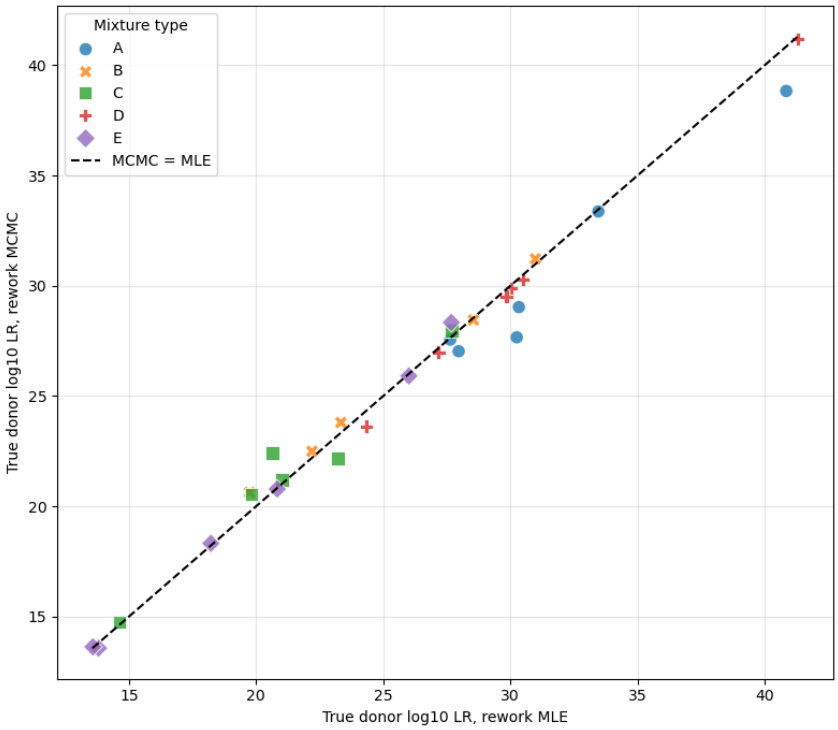


Figure 10.2: Comparison of Bayesian and frequentist observed-rework $\log_{10}$-LRs for the minor true donors. Each point represents one observed rework profile formed by combining three replicate measurements. Points close to the diagonal indicate similar Bayesian and frequentist LR values.

Figure 10.3 compares the Bayesian LR for the original single profile with the Bayesian LR for the combined rework profile, using only the minor true donors. Rework usually increases the LR. On average, the increase is 4.71 $\log_{10}$-units:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\log_{10}\mathrm{LR}^{\mathrm{Bayes}}_{i,\mathrm{rework}} - \log_{10}\mathrm{LR}^{\mathrm{Bayes}}_{i,\mathrm{single}}\right) = 4.71. \tag{10.1}$$

This is roughly the same as the average frequentist true-donor log10-LR increase shown in Figure 6.2. Thus, the Bayesian implementation leads to the same main conclusion as the frequentist analysis: adding rework can substantially increase the LR for the minor true donor.

10.2 Validation of Bayesian sampling

The second part of this chapter validates the Bayesian deconvolution sampling method, using the same percentile diagnostic as in Chapters 5 and 6. The sampled donors are drawn from the minor-donor deconvolution distribution only, and the observed LR used for validation is the LR of the minor true donor.

Figure 10.4 shows the validation result for single profiles. The histogram still deviates from uniformity, with a Kolmogorov–Smirnov test statistic of 0.144. The minor true-donor LR is too often located high in the sampled LR distribution, meaning that the Bayesian sampled LR distribution relatively often underestimates the LR of the minor true donor.

Figure 10.5 shows the corresponding result for observed rework profiles. The deviation is stronger than for the single profiles, with a Kolmogorov–Smirnov test statistic of 0.229. Thus, even for rework profiles that were actually observed in the laboratory, the Bayesian minor-donor

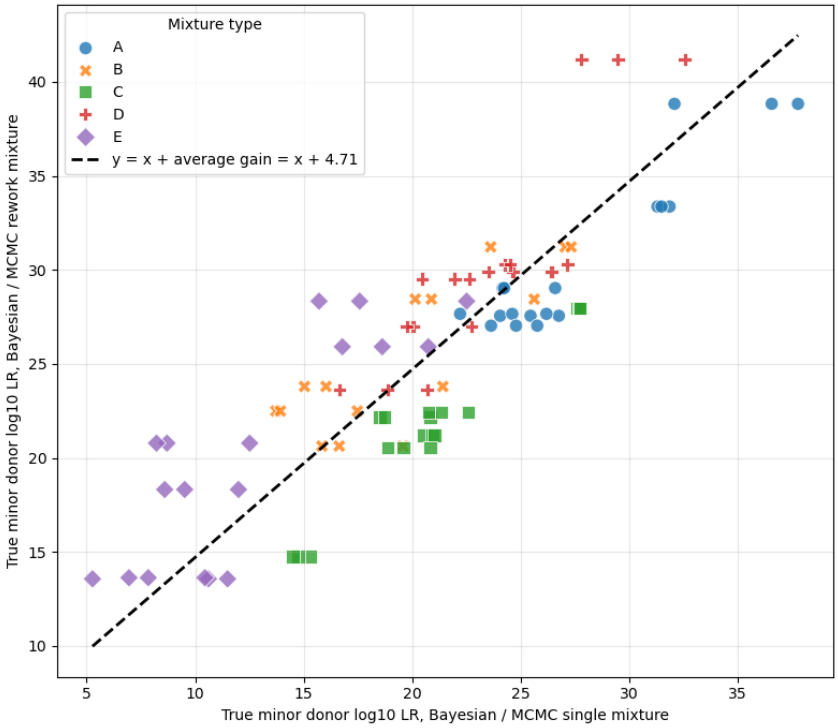


Figure 10.3: Bayesian rework-profile LR compared with Bayesian single-profile LR for the minor true donors. The difference between the two axes represents the Bayesian increase in true-donor $\log_{10}$-LR obtained by performing rework.

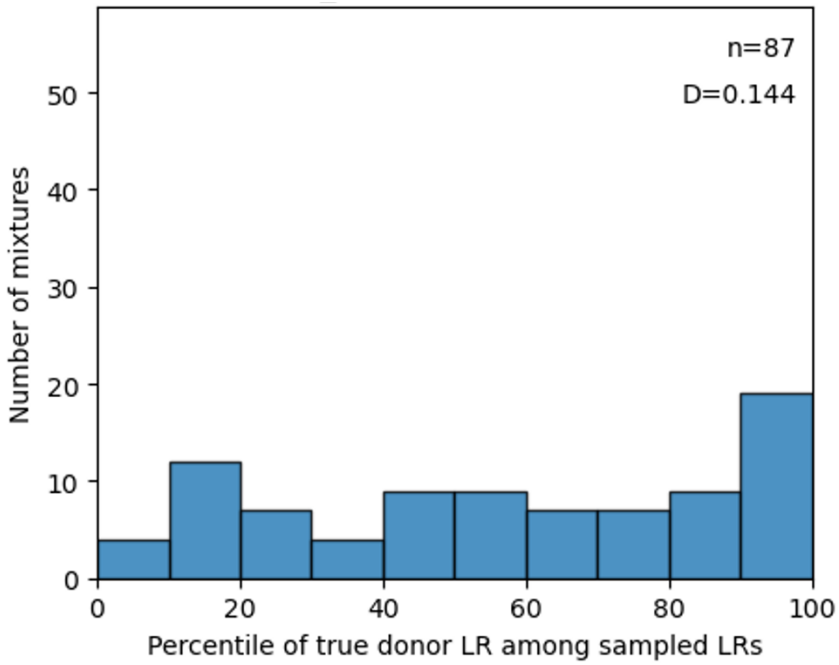


Figure 10.4: Bayesian sampling validation for single profiles. The histogram shows the percentile of the minor true-donor LR within the Bayesian sampled minor-donor LR distribution. The 95%-coverage is 85.1% and the mean interval score is 13.6.

sampling distribution does not make the minor true-donor LR behave as a calibrated draw from the sampled LR distribution.

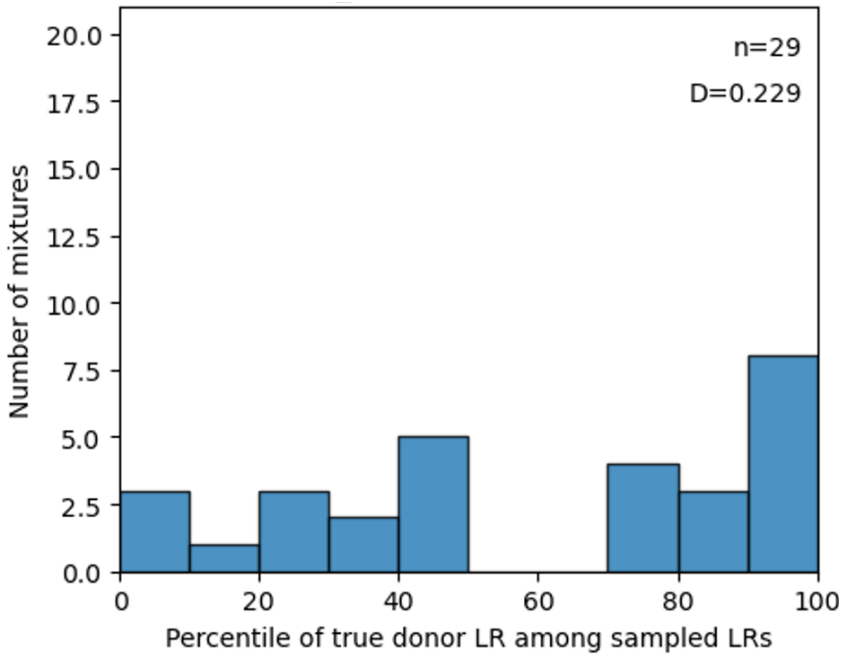


Figure 10.5: Bayesian sampling validation for observed rework profiles. The histogram shows the percentile of the minor true-donor rework LR within the Bayesian sampled minor-donor LR distribution for the combined replicate profile. The 95%-coverage is 86.2% and the mean interval score is 9.86.

These results are important for the rework simulation algorithm. The algorithm predicts a rework LR distribution from only the original single profile. However, Figures 10.4 and 10.5 show that the Bayesian minor-donor deconvolution sampling is not fully calibrated even when the target profile has already been observed. This makes it unlikely that the predicted rework LR distribution will be perfectly calibrated when the rework profile is not observed, but simulated.

The Bayesian implementation therefore improves the treatment of parameter uncertainty, but does not automatically solve the calibration problem in the minor-donor deconvolution sampling. This limitation should be kept in mind when interpreting the Bayesian rework simulation results in the next chapters.

11. Bayesian rework simulation algorithm

The frequentist rework simulation algorithm in Chapter 7 uses maximum-likelihood estimates as plug-in parameter values at several stages of the pipeline. First, the original single profile $E^{(r)}$ is analysed under H_2, giving the single-profile maximum-likelihood estimate $\hat{\boldsymbol{\theta}}_2^{(r)}$. This estimate is used to construct the deconvolution table from which contributor genotypes are sampled. The same estimate is then used to generate two artificial replicate profiles. Finally, after the artificial profiles are combined with the original profile, a new MLE is computed for each simulated rework profile. The simulated rework LR is then obtained from the corresponding simulated-rework deconvolution.

Chapter 8 showed that the frequentist rework predictions were not well calibrated for the minor donor. This motivates studying what happens when uncertainty in the peak-height model parameters is propagated through the simulation pipeline. The Bayesian algorithm in this chapter keeps the same overall structure as the frequentist algorithm, but replaces several fixed plug-in choices by draws from the posterior distribution obtained from an MCMC analysis of the original profile.

11.1 Simulation target

The simulation target is the same as in Section 7.1. Starting from one original profile $E^{(r)}$, the aim is to predict the LR that could be obtained after two additional replicate measurements have been performed. The actual laboratory rework profile $E^{1:3}$ is not used during simulation, but only afterwards for validation.

For each simulation b, the algorithm generates two artificial replicate profiles and combines them with the original profile:

$$\widetilde{E}_{r,b}^{1:3} = \left(E^{(r)}, \widetilde{E}_b^{(r,1)}, \widetilde{E}_b^{(r,2)}\right). \tag{11.1}$$

The corresponding simulated rework log-LR is denoted by

$$z_b = \log_{10} \mathrm{LR}_{\mathrm{sim},b}. \tag{11.2}$$

Repeating the procedure B times gives the Bayesian predictive rework LR distribution

$$\mathcal{P}_{\mathrm{sim}} = z_1, \ldots, z_B. \tag{11.3}$$

11.2 Bayesian analysis of the original profile

The starting point is again the original profile $E^{(r)}$. This profile is analysed under H_2, where both contributors are unknown. Instead of estimating a single parameter vector by maximum likelihood, an MCMC analysis is performed under H_2. This gives samples from the posterior distribution

$$p(\boldsymbol{\theta} \mid E^{(r)}, H_2). \tag{11.4}$$

The parameter vector $\boldsymbol{\theta}$ contains the peak-height model parameters, including the expected peak-height scale, peak-height variability, degradation and mixture proportion.

For simulation b, one posterior draw is sampled:

$$\boldsymbol{\theta}_{\mathrm{sample}}^{(b)} \sim p(\boldsymbol{\theta} \mid E^{(r)}, H_2). \tag{11.5}$$

This sampled parameter value is then used for the deconvolution step:

$$\boldsymbol{\theta}_{\mathrm{deconv}}^{(b)} = \boldsymbol{\theta}_{\mathrm{sample}}^{(b)}. \tag{11.6}$$

It is also used for the generation step:

$$\boldsymbol{\theta}^{(b)}_{\text{generate}} = \boldsymbol{\theta}^{(b)}_{\text{sample}}. \tag{11.7}$$

Thus, within one simulation run, the same sampled parameter vector is used to sample contributor genotypes and to generate the artificial replicate profiles. Across simulation runs, different plausible parameter vectors are used.

11.3 Sampling complete contributor genotypes

Conditional on $\boldsymbol{\theta}^{(b)}_{\text{deconv}}$, the original-profile deconvolution is computed under H_2. For each locus ℓ, this gives the joint deconvolution distribution

$$P(g_{1,\ell}, g_{2,\ell} \mid E^{(r)}_{\ell}, H_2, \boldsymbol{\theta}^{(b)}_{\text{deconv}}). \tag{11.8}$$

For simulation b, one joint genotype pair is sampled at each locus:

$$(g^{(b)}_{1,\ell}, g^{(b)}_{2,\ell}) \sim P(g_{1,\ell}, g_{2,\ell} \mid E^{(r)}_{\ell}, H_2, \boldsymbol{\theta}^{(b)}_{\text{deconv}}). \tag{11.9}$$

Repeating this over loci gives a full sampled genotype pair

$$g^{(b)} = \left(g^{(b)}_1, g^{(b)}_2\right), \tag{11.10}$$

where

$$g^{(b)}_i = \left(g^{(b)}_{i,1}, \ldots, g^{(b)}_{i,L}\right). \tag{11.11}$$

As in Chapter 7, the deconvolution output can contain the symbol $\emptyset$, representing an unobserved allele. Before artificial profiles are generated, each $\emptyset$ is replaced by an actual allele. This replacement is sampled from the population allele frequencies restricted to alleles that were not observed at that locus. If both alleles are $\emptyset$, a complete genotype is sampled from the corresponding population genotype distribution over unobserved alleles.

After this step, simulation b has a complete genotype file for both unknown contributors.

11.4 Generating simulated rework profiles

Given the sampled genotype pair $g^{(b)}$ and the sampled parameter value $\boldsymbol{\theta}^{(b)}_{\text{generate}}$, two artificial replicate profiles are generated:

$$\widetilde{E}^{(r,1)}_b, \widetilde{E}^{(r,2)}_b \sim P(E \mid g^{(b)}_1, g^{(b)}_2, \boldsymbol{\theta}^{(b)}_{\text{generate}}). \tag{11.12}$$

The two artificial replicates are conditionally independent given the sampled genotypes and the sampled model parameters.

Together with the original profile, these artificial replicates form one simulated rework profile:

$$\widetilde{E}^{1:3}_{r,b} = \left(E^{(r)}, \widetilde{E}^{(r,1)}_b, \widetilde{E}^{(r,2)}_b\right). \tag{11.13}$$

The generated profile therefore represents one possible outcome of performing two additional replicate measurements, conditional on the original profile and on one sampled parameter vector from the posterior distribution.

11.5 Approximate LR calculation for simulated rework profiles

In principle, the Bayesian analogue of the frequentist algorithm would run a full MCMC analysis for every simulated rework profile $\widetilde{E}_{r,b}^{1:3}$. This would propagate parameter uncertainty after observing the simulated rework profile. However, it would also be computationally expensive, because a new MCMC run would be required for every simulation b.

An approximation will be used in the implementation of the Bayesian pipeline. Instead of running a full MCMC analysis for each simulated rework profile, the simulated rework LR is computed conditionally on the same sampled parameter vector that was used to obtain the original-profile deconvolution and to generate the artificial profiles. Thus, for simulation b, the LR calculation uses

$$\boldsymbol{\theta}_{\mathrm{LR}}^{(b)} = \boldsymbol{\theta}_{\mathrm{sample}}^{(b)}. \tag{11.14}$$

For contributor i, the simulated rework LR is then computed as

$$\mathrm{LR}_{\mathrm{sim},b} = \frac{P(\widetilde{E}_{r,b}^{1:3} \mid H_1(g_i^{(b)}), \boldsymbol{\theta}_{\mathrm{LR}}^{(b)})}{P(\widetilde{E}_{r,b}^{1:3} \mid H_2, \boldsymbol{\theta}_{\mathrm{LR}}^{(b)})}. \tag{11.15}$$

Here $H_1(g_i^{(b)})$ denotes the hypothesis that the sampled genotype $g_i^{(b)}$ is a contributor together with one unknown contributor. The simulated rework log-LR is

$$z_b = \log_{10} \mathrm{LR}_{\mathrm{sim},b}. \tag{11.16}$$

Repeating this for $b = 1, \ldots, B$ gives the Bayesian predictive rework LR distribution $\mathcal{P}_{\mathrm{sim}}$.

This approximation should be taken into account when interpreting the results. The simulated profiles are generated with parameter uncertainty, but the LR of each simulated profile is computed conditionally on one sampled parameter value. The approximation avoids running a separate MCMC analysis for every simulated rework profile.

11.6 Observed laboratory rework comparison

The observed laboratory rework profile $E^{1:3}$ is analysed separately. This gives the target against which the simulated predictive distribution is compared. Two choices are possible. The first is to use a plug-in MLE deconvolution of the observed rework profile under H_2. The second is to use an MCMC analysis of the observed rework profile under H_2.

For the Bayesian comparison in this thesis, the observed rework target is based on the MCMC analysis of $E^{1:3}$ under H_2. This gives LRs obtained after Bayesian analysis of the actual laboratory rework profile. These LRs form the observed rework reference distribution, denoted by

$$\mathcal{P}_{\mathrm{obs}}^{\mathrm{MCMC}}. \tag{11.17}$$

For the true-donor percentile diagnostic, the corresponding observed true-donor rework log-LR is denoted by

$$z_{\mathrm{obs}}^{\mathrm{MCMC}} = \log_{10} \mathrm{LR}_{\mathrm{obs}}^{\mathrm{MCMC}} \left(g_i^{\mathrm{true}}\right). \tag{11.18}$$

The percentile of this observed value within the simulated predictive distribution is

$$p_{\mathrm{sim}} = 100 \cdot \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}\left\{z_b \leq z_{\mathrm{obs}}^{\mathrm{MCMC}}\right\}. \tag{11.19}$$

If the Bayesian simulation framework is calibrated, the observed laboratory rework LR should behave like a draw from the predicted distribution. Across many mixtures, the values of p_{sim} should therefore be approximately uniformly distributed on $[0, 100]$.

11.7 Bayesian algorithm summary

Algorithm 3 summarizes the Bayesian rework simulation algorithm. The algorithm is written for two-person mixtures, as used throughout the main analyses of this thesis.

The key difference from Algorithm 2 is that the parameter vector used for deconvolution and generation is not fixed at the original-profile MLE. Instead, each simulation run uses a sampled value from the posterior distribution under H_2. This allows uncertainty in the original-profile model parameters to influence the predicted rework LR distribution.

11.8 Relation to the frequentist algorithm

The Bayesian and frequentist algorithms differ at four points in the simulation pipeline. These choices are introduced here because Chapter 12 compares their effect on the calibration of the predicted rework LR distributions.

The original frequentist algorithm from Chapter 7 corresponds to choices 1A, 2A, 3A and 4A. The Bayesian implementation described in Algorithm 3 corresponds to choices 1B, 2B, 3B and 4B. Table 11.1 summarizes the four choices.

Step	**Frequentist choice**	**Bayesian choice**
Original-profile deconvolution	1A: use the single-profile MLE, $\boldsymbol{\theta}_{\text{deconv}} = \hat{\boldsymbol{\theta}}_2^{(r)}$.	1B: sample $\boldsymbol{\theta}_{\text{sample}}$ from the MCMC output and set $\boldsymbol{\theta}_{\text{deconv}} = \boldsymbol{\theta}_{\text{sample}}$.
Generation of artificial replicate profiles	2A: use the single-profile MLE, $\boldsymbol{\theta}_{\text{generate}} = \hat{\boldsymbol{\theta}}_2^{(r)}$.	2B: use $\boldsymbol{\theta}_{\text{generate}} = \boldsymbol{\theta}_{\text{sample}}$, or sample a new posterior draw if no previous draw is available.
LR calculation for simulated rework profiles	3A: compute a new MLE over $\widetilde{E}_{r,b}^{1:3}$ and compute the LR conditionally on this value.	3B: compute the LR conditionally on $\boldsymbol{\theta}_{\text{sample}}$, without running a full MCMC or MLE optimisation over $\widetilde{E}_{r,b}^{1:3}$.
Observed rework target for comparison	4A: compare with the MLE-based analysis of the observed rework profile $E^{1:3}$.	4B: compare with the MCMC-based analysis of the observed rework profile $E^{1:3}$.

Table 11.1: Four algorithmic choices distinguishing the frequentist and Bayesian rework simulation algorithms. The frequentist algorithm uses choices 1A, 2A, 3A and 4A. The Bayesian algorithm in this chapter uses choices 1B, 2B, 3B and 4B.

The purpose of separating the algorithm into these choices is that the Bayesian and frequentist methods differ in more than one place. Sampling $\boldsymbol{\theta}_{\text{sample}}$ affects the original-profile deconvolution, the generation of artificial replicate profiles, and the LR calculation for simulated rework profiles. In addition, the observed laboratory rework target can be defined either through an MLE-based analysis or through an MCMC-based analysis. Chapter 12 uses these choices to study which changes have the largest effect on the calibration of the predicted rework LR distributions.

Algorithm 3 Bayesian rework simulation algorithm

Input: original profile $E^{(r)}$, number of simulations B
Run an MCMC analysis under H_2 for $E^{(r)}$
for $b = 1, \ldots, B$ **do**
    Sample $\boldsymbol{\theta}^{(b)}_{\text{sample}}$ from the MCMC output
    Set $\boldsymbol{\theta}^{(b)}_{\text{deconv}} = \boldsymbol{\theta}^{(b)}_{\text{sample}}$
    Set $\boldsymbol{\theta}^{(b)}_{\text{generate}} = \boldsymbol{\theta}^{(b)}_{\text{sample}}$
    Compute the joint H_2 deconvolution using $\boldsymbol{\theta}^{(b)}_{\text{deconv}}$
    for each locus $\ell = 1, \ldots, L$ **do**
        Sample $(g^{(b)}_{1,\ell}, g^{(b)}_{2,\ell})$ from the joint deconvolution
        Replace $\emptyset$ alleles using the population distribution
    end for
    Construct $g^{(b)}_1$ and $g^{(b)}_2$
    Generate $\widetilde{E}^{(r,1)}_b$ and $\widetilde{E}^{(r,2)}_b$
    Use $\boldsymbol{\theta}^{(b)}_{\text{generate}}$ for the generation step
    Define $\widetilde{E}^{1:3}_{r,b} = \left(E^{(r)}, \widetilde{E}^{(r,1)}_b, \widetilde{E}^{(r,2)}_b\right)$
    Compute $z_b = \log_{10} \text{LR}_{\text{sim},b}$ for $g^{(b)}_i$
    Use $\boldsymbol{\theta}^{(b)}_{\text{sample}}$ for this LR calculation
end for
Form $\mathcal{P}_{\text{sim}} = z_1, \ldots, z_B$
Compare $\mathcal{P}_{\text{sim}}$ with $z^{\text{MCMC}}_{\text{obs}}$
Use Eq. (11.19) for the percentile
Output: Bayesian predictive rework LR distribution and percentile

12. Bayesian and combined simulation results

This chapter compares the frequentist and Bayesian rework simulation algorithms for the minor donor in the two-person mixtures. The comparison is based on the four algorithmic choices introduced in Table 11.1. The fully frequentist algorithm corresponds to configuration AAAA, while the fully Bayesian algorithm corresponds to configuration BBBB.

12.1 Fully Bayesian simulation result

The evaluation uses the calibration diagnostics defined in Section 3.10. For each validation case, the observed rework log-LR of the minor true donor is compared with the corresponding simulated rework log-LR distribution. The resulting percentile histogram, empirical 95% coverage and mean 95% interval score are used to compare the algorithmic configurations. Figure 12.1 shows the percentile histogram for the fully Bayesian simulation algorithm. Each percentile gives the position of the observed rework LR of the minor true donor within the corresponding simulated rework LR distribution.

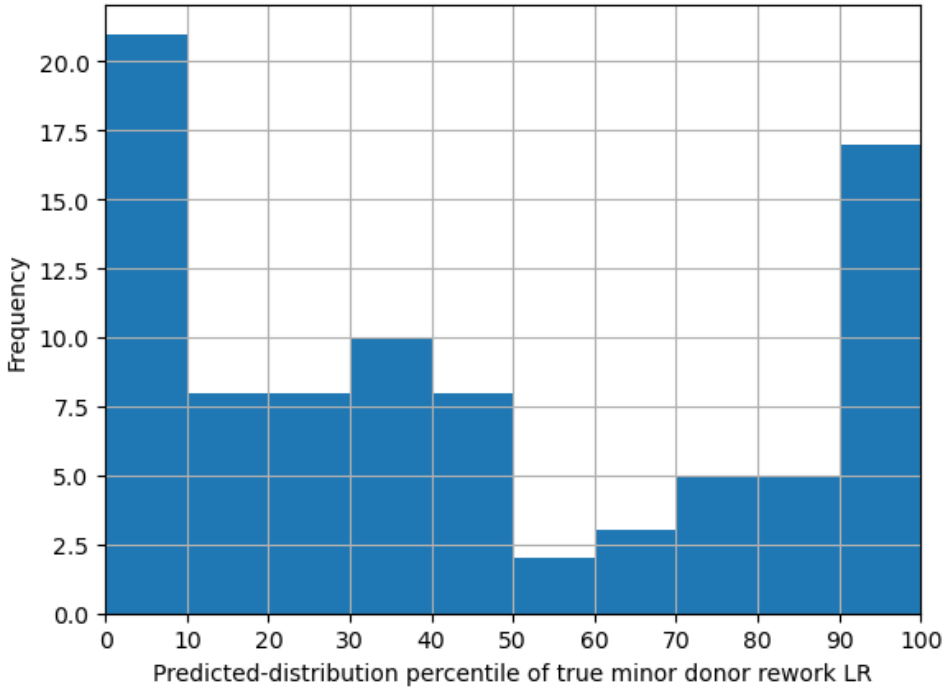


Figure 12.1: Calibration histogram for the fully Bayesian rework simulation algorithm (configuration BBBB). For each minor true donor, the percentile of the observed rework LR within the predicted Bayesian rework LR distribution is plotted. A calibrated prediction method would give an approximately uniform histogram. The Kolmogorov–Smirnov statistic is $D = 0.167$, the empirical 95% coverage is 81.6%, and the mean interval score is 21.6.

The percentile histogram is still not approximately uniform. Many observed rework LRs are located in the lowest or highest percentile bins, whereas a calibrated predictive distribution would give a flatter histogram. The Kolmogorov–Smirnov statistic is $D = 0.167$. This is slightly larger than the value $D = 0.162$ obtained for the fully frequentist simulation in Section 8.2. Therefore, the percentile histogram does not improve when moving from the fully frequentist to the fully Bayesian algorithm.

The coverage and interval score give a more favourable picture. The empirical 95% cover-

age increases from 69.0% for the fully frequentist configuration to 81.6% for the fully Bayesian configuration. The mean interval score decreases from 50.5 to 21.6. Thus, although the percentile histogram is still not uniform and the coverage remains below the nominal 95% level, the Bayesian prediction is less poor than the frequentist prediction. The observed rework LR is more often contained in the 95% prediction interval, and the interval score is substantially lower.

12.2 Results for the algorithmic combinations

To study which parts of the Bayesian algorithm are responsible for the improvement, all combinations of the four algorithmic choices were compared. The percentile histogram is shown only for the fully Bayesian configuration in Figure 12.1. For the full set of 16 configurations, the comparison is summarized using coverage and interval score.

Each configuration is represented by a four-letter code. The first letter gives the choice for step 1, the second letter gives the choice for step 2, the third letter gives the choice for step 3, and the fourth letter gives the choice for step 4. For example, configuration BABA means that step 1 uses choice B, step 2 uses choice A, step 3 uses choice B, and step 4 uses choice A. Thus, AAAA is the fully frequentist configuration and BBBB is the fully Bayesian configuration.

Table 12.1 shows the results. The fully frequentist configuration AAAA has a coverage of 69.0% and a mean interval score of 50.5. The fully Bayesian configuration BBBB has a coverage of 81.6% and a mean interval score of 21.6.

Configuration	Step 1	Step 2	Step 3	Step 4	cov95 (%)	Interval score
AAAA	A	A	A	A	69.0	50.5
AABA	A	A	B	A	73.6	44.7
ABAA	A	B	A	A	77.0	30.2
ABBA	A	B	B	A	73.6	34.3
BAAA	B	A	A	A	70.1	47.1
BABA	B	A	B	A	72.4	48.0
BBAA	B	B	A	A	78.2	28.0
BBBA	B	B	B	A	79.3	29.0
AAAB	A	A	A	B	70.1	41.7
AABB	A	A	B	B	74.7	37.4
ABAB	A	B	A	B	80.5	23.7
ABBB	A	B	B	B	74.7	26.4
BAAB	B	A	A	B	73.6	39.1
BABB	B	A	B	B	73.6	39.4
BBAB	B	B	A	B	80.5	**21.0**
BBBB	B	B	B	B	**81.6**	21.6

Table 12.1: Coverage and mean interval score for all 16 combinations of the four algorithmic choices. The four-letter configuration code indicates, in order, the choices for steps 1 to 4. For example, BABA means choice B for step 1, choice A for step 2, choice B for step 3, and choice A for step 4. Lower interval scores indicate better predictive performance.

The best coverage is obtained by the fully Bayesian configuration BBBB. The lowest interval score is obtained by BBAB, although its interval score is only slightly lower than that of BBBB. Moving from AAAA to BBBB increases the empirical coverage by 12.6 percentage points and decreases the mean interval score by 28.9 points. The Bayesian choices therefore improve the prediction, but do not make it fully calibrated.

12.3 Contribution of the individual choices

To summarize the contribution of each step, an additive ordinary least squares model was fitted to the 16 configurations:

$$y_i = \beta_0 + \sum_{j=1}^{4} \beta_j x_{i,j} + \varepsilon_i,$$

where $x_{i,j} = 0$ if step j uses option A and $x_{i,j} = 1$ if it uses option B. Because the design is balanced, the fitted coefficients can be interpreted as average main effects of switching one step from A to B, averaged over all settings of the other steps.

The resulting OLS effects and Shapley contributions are shown in Table 12.2. For coverage, the largest average OLS effect is obtained for step 2: switching step 2 from A to B increases coverage by approximately +6.04 percentage points. Steps 1 and 4 each increase coverage by approximately +2.01 percentage points, while step 3 has only a small average effect of +0.56 percentage points.

For the interval score, Table 12.2 again shows that step 2 has the largest average OLS effect. Switching step 2 from A to B changes the interval score by −16.71, whereas step 4 changes it by −7.69, step 1 by −1.96, and step 3 by only −0.06. Because lower interval scores are better, these negative values correspond to improvement.

Because the additive OLS model only gives average main effects, a Shapley decomposition was also computed. The Shapley decomposition starts from configuration AAAA and considers all possible orders in which the four steps can be changed from A to B. For each order, the marginal contribution of a step is the change in the performance measure when that step is added. The Shapley value is the average of these marginal contributions over all possible orders. In this way, the total change from AAAA to BBBB is distributed over the four steps.

For coverage, the Shapley effects in Table 12.2 sum to the total coverage increase of 12.6 percentage points. For the interval score, the Shapley effects sum to the total interval-score change of −28.9. The percentages in Table 12.2 show the share of the total coverage increase or interval-score decrease assigned to each step.

	Coverage		Interval score	
Step	OLS effect	Shapley effect (share)	OLS effect	Shapley effect (share)
1	+2.01	+2.68 (21.2%)	−1.96	−2.68 (9.3%)
2	+6.04	+6.69 (53.1%)	−16.71	−17.49 (60.5%)
3	+0.56	+1.33 (10.5%)	−0.06	−0.91 (3.1%)
4	+2.01	+1.91 (15.1%)	−7.69	−7.83 (27.1%)

Table 12.2: Estimated contribution of each algorithmic choice. The OLS effects are average main effects of switching a step from A to B. The Shapley effects decompose the total change from AAAA to BBBB over the four steps; the percentages give the share of the total coverage increase or total interval-score decrease. For the interval score, negative values denote improvement, because they correspond to a reduction in interval score.

Table 12.2 shows that step 2 has the largest Shapley contribution for both evaluation criteria. It accounts for +6.69 of the total coverage increase of 12.6 percentage points, corresponding to 53.1% of the total gain. For the interval score, it accounts for −17.49 of the total decrease of 28.9 points, corresponding to 60.5% of the total improvement. This makes step 2 the dominant source of improvement.

Step 4 has a smaller contribution to coverage, with a Shapley effect of +1.91, corresponding to 15.1% of the total coverage increase. Its contribution to the interval score is larger: Table 12.2 shows a Shapley effect of −7.83, corresponding to 27.1% of the total decrease. Step 1 has a moderate contribution to coverage and a smaller contribution to the interval score. Step 3 contributes little, especially for the interval score.

Overall, Table 12.2 shows that the main improvement comes from step 2, where artificial replicate profiles are generated using a posterior parameter draw instead of a fixed maximum-likelihood estimate. This suggests that propagating parameter uncertainty during the generation of future replicate profiles is the most important part of the Bayesian simulation framework. Step 4 also improves the interval score, while steps 1 and 3 have smaller effects.

13. Discussion

This thesis has two closely related contributions. The first is the development of a simulation framework for predicting the LR after rework from the original DNA mixture profile. The second is the introduction of a Bayesian MCMC implementation for the peak-height model used in DNAStatistX. DNAStatistX currently estimates the model parameters by maximum likelihood and then treats these estimates as fixed. The MCMC implementation developed in Chapter 9 instead samples from the posterior distribution of these parameters, so that parameter uncertainty can be propagated into deconvolution and LR calculation.

The Bayesian part of this thesis therefore has implications beyond rework prediction. A Bayesian LR integrates over uncertainty in the nuisance parameters. When the likelihood is sharply concentrated around the MLE, Bayesian and frequentist LRs may be similar. However, examples were found of mixtures with flatter likelihood surfaces, and where the MLE was not in the centre of the posterior distribution. In such cases, integrating over this uncertainty can change the assessment of the evidential value. The results in this thesis are not sufficient to recommend operational changes, but they show that Bayesian LR calculations are a relevant direction for further development and validation at the NFI.

13.1 Interpretation of the main findings

Predicting the LR after rework. The rework simulation framework was developed to answer the main research question of this thesis: whether the LR after rework can be predicted from the original DNA mixture profile. The original profile contains information about plausible contributor genotypes and about the model parameters. By sampling plausible genotypes, simulating additional replicate profiles and calculating the LR of the combined replicate profile, a predictive distribution for the rework LR can be obtained. The observed laboratory rework profile can then be used afterwards to assess whether the realised LR behaves like a draw from this predictive distribution.

The results in Section 6.2 and Figure 6.2 show that rework can substantially increase the LR, especially for minor contributors in unbalanced mixtures. Section 6.3 and Table 6.1 show that this increase is related to minor-donor drop-out, although drop-out does not fully explain the increase in true-donor $\log_{10}$-LR after rework. Additional replicate measurements may reveal alleles that were not observed above the detection threshold in the first measurement. When these replicate profiles are combined, the LR for the true donor can increase strongly.

The frequentist simulation results in Section 8.2 show that the predicted rework distributions were not calibrated. In the frequentist simulation framework, the model parameters are estimated from the original profile by maximum likelihood and then treated as fixed. If the parameter estimate from the original profile differs substantially from the parameter values supported by the combined rework profile, the simulated rework profiles can be generated from a parameter setting that is too specific to the original replicate. This helps explain why the observed rework LR of the true donor was outside the predicted distribution.

The Bayesian framework was introduced as an experimental approach to address this limitation. Instead of simulating all rework profiles from one fixed parameter estimate, the Bayesian method samples parameter values from the posterior distribution. The results in Section 12.2 show that this improved the empirical coverage and interval score compared with the fully frequentist configuration. However, the percentile histogram in Figure 12.1 was still not uniform. The Bayesian framework therefore improved the treatment of parameter uncertainty, but did not fully solve the calibration problem.

Differences between Bayesian and frequentist LRs. The comparison between Bayesian and frequentist LRs in Section 10.1 is also important. Figures 10.1 and 10.2 show that the

Bayesian and frequentist LRs were similar for most profiles, but differed more clearly for some A- and C-mixtures. These are also the mixtures for which Figure 8.5 showed larger differences between the single-profile and rework-profile MLEs. This suggests that the Bayesian approach may have the largest effect when the likelihood is not sharply concentrated around the MLE.

13.2 Limitations and future research

A practical limitation throughout this thesis is that all calculations were performed on a personal laptop. Many approximations were therefore made for runtime reasons, not because they are fundamental limitations of the method. Future research should repeat the analyses on dedicated servers.

Simplified frequentist LR. The frequentist LR calculations used for sampled donors in this thesis use the simplified quantity $\mathrm{LR}'_{\mathrm{freq}}$, defined in Section 3.9, rather than the full frequentist LR from Section 3.6. In the full frequentist LR, the MLE under H_1 is recomputed for the person of interest. In $\mathrm{LR}'_{\mathrm{freq}}$, the same parameter value estimated under H_2 is used in both numerator and denominator. This approximation made repeated sampled-donor LR calculations manageable, but it also means that the results describe the behaviour of this computational framework rather than the final performance of the full operational LR calculation. A direct next step is therefore to repeat the simulation framework using the full frequentist LR, including recomputation of the MLE under H_1 for each sampled donor.

Genotype filtering in the MCMC implementation. The MCMC implementation in Section 9.2 still depends on the MLE-based deconvolution output of DNAStatistX. Approximately the 10% most probable genotype pairs were retained, so about 90% of the genotype pairs were filtered out. This made the MCMC computation feasible, but the filtering threshold is somewhat arbitrary. A cleaner Bayesian solution would include the genotypes in the MCMC state space and sample genotype configurations and model parameters jointly. This would increase runtime, but it would avoid both the DNAStatistX pre-filtering step and the fixed genotype-filtering threshold.

MCMC validation. The MCMC implementation itself was only validated to a limited extent. Section 9.6 used visual inspection of the trace plots and posterior distributions of the sampled parameters to conclude that the implementation worked well enough for use in the simulation algorithm. These checks were sufficient for a first implementation, but they are not a complete convergence study. Future work should include multiple chains, convergence diagnostics, effective sample size checks, longer runs where necessary, numerical stability checks and sensitivity analyses for the prior distributions. The priors from Section 9.3 were not systematically varied, so the influence of the prior choices remains unknown.

MCMC for simulated rework profiles. The Bayesian simulation algorithm in Section 11.5 did not perform a new MCMC run for every simulated rework profile. For each simulated rework profile, the LR was computed conditionally on the same sampled parameter value that was used for deconvolution and profile generation. A full MCMC run for every simulated rework profile would be the most direct Bayesian analogue of the frequentist procedure, where each simulated profile is analysed separately. This was not feasible on a personal laptop. The algorithmic comparison in Table 12.2 now suggests that the Bayesian choice for step 3 had only a limited effect compared with the frequentist choice for step 3 on the coverage and interval score in the tested configurations. However, it could still be possible that another implementation, for example a full MCMC run of each simulated mixture, does have a large effect on coverage or interval score.

Validation of deconvolution sampling. For frequentist single profiles, Figure 5.3 was close to uniform. For frequentist rework profiles, Bayesian single profiles and Bayesian rework profiles, shown in Figures 6.5, 10.4 and 10.5, it was unclear based on the available number of data points whether the percentiles were uniform or not. This matters because the rework simulation framework builds on the deconvolution distribution. Any remaining miscalibration in the sampled genotype distribution may also affect the simulated rework LR distributions in Chapter 12. Future research should therefore improve the validation of the deconvolution step before using it as input for rework prediction. Locus-level analysis of true donor LRs that are higher or lower than expected could be a way to achieve this.

Calibration diagnostics for rework prediction. The calibration diagnostics for rework prediction should also be extended. In this thesis, the predicted rework distribution was mainly compared with the observed true-donor LR after rework. However, Chapter 6 showed that the true-donor LR is not always typical within the observed rework deconvolution distribution. It may therefore be useful to compare the predicted rework distribution directly with the observed rework distribution obtained after actual rework. For example, the Wasserstein distance could be used to quantify how far these two distributions are apart. This would separate two questions: whether the simulated rework distribution resembles the observed rework deconvolution distribution, and whether the true-donor LR is itself typical within the observed rework distribution.

Comparison with a simple baseline. The rework simulation framework should also be compared with a simpler prediction rule based on the average increase in true-donor $\log_{10}$-LR after rework observed in Chapter 6. Table 6.1 already gives a first approximation of the expected increase in true-donor $\log_{10}$-LR after rework for each mixture type. A baseline predictive distribution could therefore be constructed by shifting the sampled single-profile LR distribution by the average observed increase in true-donor $\log_{10}$-LR after rework from Table 6.1. This baseline would not simulate additional replicate peak heights, but it would test whether the full simulation algorithm adds predictive value beyond the average $\log_{10}$-LR increase already visible in the observed rework data. Future validation should therefore compare the calibration, coverage and interval score of this shifted-distribution baseline with those of the full simulation framework. Only if the simulation framework performs clearly better is the additional computational complexity justified.

Scope of the dataset and model. Finally, the scope of the analysis was limited. The dataset was cleaned by removing peaks classified as drop-in using the known contributor genotypes, which was done to assess the simulation algorithm with one unmodelled artefact removed. The analyses were restricted to two-person mixtures, co-ancestry was ignored, and validation was performed on a controlled NFI research dataset. These choices made the project computationally feasible, but they also mean that the current results should be viewed as a mathematical proof of concept. Future research should extend the framework to stutter modelling or improved stutter filtering, mixtures with three or more contributors, co-ancestry correction and validation on data that better represent casework. It would also be interesting to study how the LR of non-contributors behaves, since this thesis only looked at the LR of contributors that are plausible based on the deconvolution.

13.3 Practical implications for the NFI

The current rework prediction framework is not yet usable in casework because it is not sufficiently calibrated. The percentile validation in Figure 12.1 and the coverage results in Table 12.1 show that the method cannot yet provide a reliable 95%-predictive interval for where the LR

after rework should be expected. The main practical implication is therefore not an immediate decision rule, but a direction for development.

Possible decision support for rework. If future research succeeds in calibrating the predictive distributions, the framework could become a decision-support tool for rework. After the first DNA measurement, the analyst could use the deconvolution of the original profile to simulate a distribution of possible rework LRs. It should also be shown that this simulated distribution performs better than simpler baselines, such as shifting the single-profile LR distribution by the average observed increase in true-donor $\log_{10}$-LR after rework. This could then help assess whether rework is likely to add substantial evidential value, whether the uncertainty is too large to draw a useful conclusion, or whether rework is unlikely to change the interpretation.

This is relevant for the case when there is not (yet) a person of interest: for database searches, rework could improve the chance that a true contributor obtains a sufficiently informative LR to become a useful lead. Conversely, if the predicted distribution shows little probability of a relevant LR increase, the analyst may decide that rework is not worth the effort of going back to the laboratory and performing the additional measurements.

Bayesian LR instead of frequentist LR. If future validation shows that Bayesian LR calculations are reliable, this could have a second practical implication for the NFI. Bayesian calculations could be used as a sensitivity analysis for profiles where the frequentist MLE is unstable, or where the mixture proportion is close to $1/N$. These are precisely the profiles for which Bayesian and frequentist LRs may differ. In such cases, reporting officers could be alerted that the evidential value depends on how parameter uncertainty is treated. Before this could become part of casework, the NFI would need clear validation results, reporting guidelines and computational infrastructure. It could then be assessed whether the Bayesian LR should be used only as a sensitivity analysis or whether it should eventually replace the frequentist LR completely.

14. Conclusion

This thesis investigated whether the LR after rework can be predicted from the original DNA mixture profile, before additional replicate measurements are performed. The goal was to construct a predictive distribution for the rework LR by sampling plausible contributors from the deconvolution of the original profile and simulating possible additional replicate profiles. This was studied with a frequentist plug-in simulation framework and a Bayesian extension based on an MCMC implementation for the peak-height model used in DNAStatistX. The main conclusion is that the framework can produce informative predicted rework LR distributions, but is not yet sufficiently calibrated for operational use. The observed rework LR for the minor true donor too often fell near the lower or upper edge of the predicted distribution. Therefore, the current method cannot provide a reliable 95% predictive interval for the LR after rework in casework.

Rework and artefacts. The locus-level analysis showed that artefacts can strongly affect LR calculations. Loci where the true-donor LR was much lower than the sampled-donor LRs were often associated with drop-out, drop-in or stutter-compatible peaks. Because this thesis did not include a full stutter model, peaks classified as drop-in were removed using the known contributor genotypes. This created a cleaner validation setting, but it is not a casework solution. Practical implementation would require better artefact modelling, in particular a full stutter model or a validated stutter-filtering procedure. The observed rework profiles showed that rework can substantially increase the LR, especially for minor contributors in unbalanced mixtures. This increase is positively correlated with minor-donor drop-out: additional replicate measurements can recover alleles that were not observed in the original profile.

Frequentist and Bayesian prediction. The frequentist simulation framework estimated the model parameters from the original profile by maximum likelihood and then treated them as fixed. This plug-in approach was insufficiently calibrated: the nominal 95% prediction intervals covered only 69.0% of the observed minor true-donor rework LRs. A likely reason is that a single original-profile MLE can be too specific to one stochastic replicate and may not represent the parameter values supported by the combined rework profile. The Bayesian framework improved the prediction by propagating parameter uncertainty. In the fully Bayesian configuration, the empirical 95% coverage increased to 81.6%, and the mean interval score decreased from 50.5 to 21.6. The largest improvement came from generating simulated replicate profiles using posterior parameter draws instead of one fixed MLE. Nevertheless, the percentile histogram remained non-uniform, and Bayesian deconvolution sampling was not fully calibrated even for observed profiles. The Bayesian framework should therefore be interpreted as an improvement over the frequentist plug-in approach, not as a calibrated prediction method.

MCMC implementation. A second contribution of this thesis is the MCMC implementation for the DNAStatistX peak-height model. It makes it possible to sample from the posterior distribution of the nuisance parameters and to propagate parameter uncertainty into deconvolution and LR calculations. This is relevant not only for rework prediction, but also for ordinary LR reporting. The Bayesian and frequentist LRs were often similar, especially for combined rework profiles, but they differed for some mixtures where the frequentist point estimate was less representative. This suggests that Bayesian LR calculations may be useful as a sensitivity analysis for profiles with unstable MLEs, for example when mixture proportions are close to $1/N$.

Scope and next steps. The results should be viewed as a mathematical proof of concept. The validation was restricted to cleaned two-person mixtures and mainly focused on minor donors.

Co-ancestry was ignored, the dataset was a controlled NFI research dataset rather than casework data, and several approximations were used for runtime reasons. These included the simplified frequentist sampled-donor LR, genotype-space filtering in the MCMC implementation, and the absence of a full MCMC run for every simulated rework profile.

Overall, this thesis shows that predicting the effect of rework is a meaningful problem for DNA mixture interpretation. Rework can add substantial evidential value for minor contributors in unbalanced mixtures, and Bayesian parameter uncertainty improves the predicted rework LR distributions. Before the framework can be used in practice, the most important next steps are computational scaling, better artefact modelling, full MCMC validation, extension to mixtures with more contributors, inclusion of co-ancestry, and validation on casework-like data. If these steps lead to calibrated predictions, the framework could support NFI decisions about rework, including cases where rework may make database searches more informative.

A. Appendix

This appendix collects implementation details and input values that are used throughout the thesis. Section A.1 describes how allele fragment lengths were approximated for the degradation model. Section A.2 gives the allele population frequencies used for genotype probabilities. Section A.3 lists the detection thresholds used to decide whether peaks were treated as observed.

Appendix A.1 : Allele fragment lengths

In Section 3.3.2, degradation was included by multiplying the expected peak height by

$$\beta^{(f_{\ell,a}-125)/100},$$

where $f_{\ell,a}$ is the fragment length of allele a at locus ℓ. This appendix describes how the fragment lengths $f_{\ell,a}$ were approximated in the implementation.

For each locus, the PPF6C kit definition gives a slope s_ℓ and an offset o_ℓ. For an integer allele a, the fragment length was calculated as

$$f_{\ell,a} = o_\ell + s_\ell a. \tag{A.1}$$

For a microvariant allele written as $a.p$, the part after the decimal point was added directly:

$$f_{\ell,a.p} = o_\ell + s_\ell a + p. \tag{A.2}$$

For example, for allele 31.2 at locus D21S11,

$$f_{\text{D21S11},31.2} = 105.83 + 4.08 \cdot 31 + 2 = 234.31.$$

The grouped unobserved allele is denoted by $\varnothing$. Since $\varnothing$ does not correspond to one specific allele, its repeat length was approximated by the frequency-weighted average repeat length of all alleles that were not observed at that locus. Let $\mathcal{A}_\ell$ be the set of alleles in the allele-frequency table for locus ℓ, and let $\mathcal{O}_\ell$ be the set of alleles observed in the mixture. Then

$$\bar{a}_{\ell,\varnothing} = \frac{\sum_{a\in\mathcal{A}_\ell\setminus\mathcal{O}_\ell} r(a)p_{\ell,a}}{\sum_{a\in\mathcal{A}_\ell\setminus\mathcal{O}_\ell} p_{\ell,a}}, \tag{A.3}$$

where $p_{\ell,a}$ is the population frequency of allele a, and $r(a)$ is the repeat-length value used for allele a. The fragment length for the grouped unobserved allele was then calculated as

$$f_{\ell,\varnothing} = o_\ell + s_\ell \bar{a}_{\ell,\varnothing}. \tag{A.4}$$

After obtaining $f_{\ell,a}$, the degradation factor was computed as

$$D_{\ell,a}(\beta) = \beta^{(f_{\ell,a}-125)/100}. \tag{A.5}$$

The value 125 bp is used as the reference fragment length. Therefore, $D_{\ell,a}(\beta) = 1$ for an allele of length 125 bp. If $\beta = 1$, there is no degradation. If $0 < \beta < 1$, alleles longer than 125 bp are down-weighted relative to alleles around the reference length.

Table A.1 lists the slope and offset values used for each locus.

Appendix A.2 : Allele population frequencies

Tables A.2 and A.3 list the allele population frequencies used in the LR calculations. These frequencies are used to assign prior probabilities to unknown contributor genotypes.

Table A.1: PPF6C locus-specific constants used to approximate allele fragment lengths. The period is included for reference; the implementation uses the slope and offset according to Equations (A.1) and (A.2).

Locus	Period	s_ℓ	o_ℓ	Locus	Period	s_ℓ	o_ℓ
D3S1358	4	4.27	59.52	vWA	4	4.12	84.52
D1S1656	4	4.08	121.80	D21S11	4	4.08	105.83
D2S441	4	4.11	179.87	D7S820	4	4.02	249.62
D10S1248	4	3.98	224.88	D5S818	4	4.04	298.77
D13S317	4	4.08	285.67	TPOX	4	4.02	378.66
Penta E	5	4.87	344.56	D8S1179	4	4.33	41.97
D16S539	4	4.34	59.02	D12S391	4	4.10	77.99
D18S51	4	4.11	103.99	D19S433	4	4.02	172.24
D2S1338	4	4.06	182.51	SE33	4	4.06	256.52
CSF1PO	4	4.07	297.45	D22S1045	3	3.05	413.97
Penta D	5	5.08	366.87	FGA	4	4.10	86.02
TH01	4	4.28	54.43				

Table A.2: NFI allele frequencies, loci D1S1656–TH01.

Allele	D1S1656	TPOX	D2S441	D2S1338	D3S1358	FGA	D5S818	CSF1PO	D7S820	D8S1179	D10S1248	TH01
2.2												
5												0.0071942446
6								0.0002398082				0.2098321343
7		0.0004796163					0.0004796163	0.0011990408	0.0191846523			0.1808153477
8		0.5414868106	0.0004796163				0.0035971223	0.0023980815	0.1654676259	0.0191846523		0.1110311751
8.3												0.0002398082
9		0.0947242206	0.0002398082				0.0304556355	0.0225419664	0.1798561151	0.0167865707		0.1352517986
9.3												0.3465227818
10	0.0016786571	0.0573141487	0.1887290168		0.0002398082		0.0505995204	0.2549160671	0.2592326139	0.0889688249	0.0007194245	0.0091127098
10.3								0.0002398082	0.0002398082			
11	0.0741007194	0.2582733813	0.3431654676		0.0019184652		0.3520383693	0.3069544365	0.1925659472	0.0774580336	0.0026378897	
11.3			0.0419664269						0.0002398082			
12	0.1239808153	0.0462829736	0.0479616307	0.0002398082			0.3767386091	0.3330935252	0.1278177458	0.1594724221	0.0357314149	
12.1										0.0002398082		
12.2												
12.3			0.0016786571									
13	0.0573141487	0.0014388489	0.0256594724		0.0059952038		0.1748201439	0.064028777	0.0455635492	0.3196642686	0.3160671463	
13.2												
13.3			0.0004796163									
13.4												
14	0.087529976		0.293764988	0.0002398082	0.1227817746		0.0095923261	0.012470024	0.0091127098	0.1949640288	0.3069544365	
14.1			0.0002398082									
14.2												
14.3	0.0016786571											
15	0.1314148681		0.0496402878	0.0004796163	0.2465227818		0.0014388489	0.0016786571	0.0007194245	0.0918465228	0.1964028777	
15.2												
15.3	0.0748201439											
16	0.1189448441		0.0059952038	0.0381294964	0.2417266187		0.0002398082	0.0002398082		0.0263788969	0.1175059952	
16.1	0.0002398082											
16.2												
16.3	0.0549160671											
17	0.057793765			0.1940047962	0.2163069544	0.0011996161				0.0050359712	0.020383693	
17.1	0.0007194245											
17.2												
17.3	0.1419664269											
18	0.0038369305			0.0836930456	0.1544364508	0.0179942418					0.0033573141	
18.2												
18.3	0.0537170264											
19	0.0007194245			0.1165467626	0.009352518	0.0607005758					0.0002398082	
19.1						0.0004798464						
19.2						0.0009596929						
19.3	0.0136690647											
20				0.151558753	0.0007194245	0.1341170825						
20.2						0.0002399232						
20.3	0.0009592326											
21				0.0304556355		0.1669865643						
21.2						0.0031190019						
22				0.0314148681		0.1722648752						
22.1						0.0002399232						
22.2						0.00743762						
22.3												
23				0.093764988		0.1470729367						
23.2						0.0033589251						
24				0.1227817746		0.1374760077						
24.2						0.0011996161						
25				0.112470024		0.1026871401						
25.2												
26				0.0215827338		0.037428023						
26.2												
27				0.0016786571		0.0047984645						
27.2												
28				0.0009592326		0.0002399232						
28.2												
29												
29.2												
29.3												
30												
30.2												
31												
31.2												
32												
32.2												
33												
33.2												
34												
34.1												
34.2												
35												
35.2												
38												

Table A.3: NFI allele frequencies, loci vWA–SE33.

Allele	vWA	D12S391	D13S317	PentaE	D16S539	D18S51	D19S433	PentaD	D21S11	D22S1045	SE33
2.2								0.0002398082			
5				0.0714628297				0.0007194245			
6				0.0023980815				0.0004796163			
7			0.0014388489	0.1707434053	0.0002398082			0.0067146283			
8			0.1098321343	0.0103117506	0.0167865707			0.020383693			
8.3											
9			0.0741007194	0.0105515588	0.1294964029	0.0009592326		0.2086330935			
9.3											
10	0.0002398082		0.0659472422	0.0880095923	0.0652278177	0.0079136691	0.0004796163	0.0990407674		0.0014388489	
10.3											
11			0.2964028777	0.0954436451	0.3215827338	0.0139088729	0.0050359712	0.1362110312		0.1405275779	
11.3											
12	0.0004796163		0.293764988	0.1978417266	0.2642685851	0.1570743405	0.067146283	0.2321342926		0.0131894484	0.0081
12.1							0.0007194245				
12.2							0.0002398082				0.0012
12.3											
13	0.0019184652		0.1122302158	0.093764988	0.1793764988	0.1378896882	0.2134292566	0.2122302158		0.0059952038	0.0093
13.2							0.0182254197				0.0012
13.3											
13.4								0.0011990408			
14	0.0988009592	0.0002398082	0.0455635492	0.0613908873	0.0225419664	0.1647482014	0.3661870504	0.0678657074		0.0508393285	0.0313
14.1								0.0007194245			
14.2							0.0232613909				0.0012
14.3											
15	0.0976019185	0.0417266187	0.0007194245	0.0503597122	0.0002398082	0.1429256595	0.1829736211	0.0098321343		0.3256594724	0.029
15.2							0.042206235				
15.3											
16	0.2059952038	0.0374100719		0.0522781775	0.0002398082	0.1232613909	0.0532374101	0.003117506		0.3705035971	0.0499
16.1											
16.2							0.0167865707				
16.3		0.0002398082									
17	0.2741007194	0.1004796163		0.0491606715		0.1074340528	0.0052757794	0.0002398082		0.0841726619	0.0766
17.1											
17.2							0.0033573141				
17.3		0.0227817746									
18	0.1997601918	0.1755395683		0.0242206235		0.0717026379	0.0002398082	0.0002398082		0.0067146283	0.0534
18.2							0.0009592326				0.0012
18.3		0.0220623501									
19	0.1059952038	0.1028776978		0.0105515588		0.0366906475				0.0009592326	0.08
19.1											
19.2							0.0002398082				
19.3		0.0095923261									
20	0.0143884892	0.1211031175		0.0074340528		0.0196642686					0.0545
20.2		0.0002398082									0.0058
20.3		0.0004796163									
21	0.0007194245	0.1309352518		0.0028776978		0.0091127098					0.0313
21.2											0.0116
22		0.1011990408		0.0007194245		0.0043165468					0.0081
22.1		0.0002398082									
22.2											0.0267
22.3		0.0002398082									
23		0.0748201439		0.0004796163		0.0009592326					0.0035
23.2											0.0278
24		0.0369304556				0.0014388489					0.0012
24.2									0.0002398082		0.0394
25		0.01558753									0.0023
25.2									0.0002398082		0.0406
26		0.0035971223							0.0016786571		
26.2											0.0615
27		0.0014388489							0.0429256595		
27.2											0.0777
28		0.0002398082							0.1693045564		
28.2									0.0002398082		0.0812
29									0.2011990408		
29.2									0.0007194245		0.0534
29.3									0.0004796163		
30									0.2587529976		
30.2									0.032853717		0.0568
31									0.0916067146		
31.2									0.0779376499		0.0325
32									0.0160671463		0.0012
32.2									0.073381295		0.0151
33									0.0028776978		0.0023
33.2									0.0261390887		0.007
34									0.0002398082		0.0035
34.1									0.0004796163		
34.2									0.0019184652		0.0058
35									0.0004796163		0.0035
35.2									0.0002398082		0.0023
38											0.0012

Appendix A.3 : Detection thresholds

Table A.4 lists the locus-specific detection thresholds used in the analysis. Peaks below the threshold were not treated as observed alleles. If no locus-specific threshold was specified, the default threshold of 95 RFU was used.

Table A.4: Detection thresholds used per locus.

Locus	Threshold (RFU)	Locus	Threshold (RFU)	Locus	Threshold (RFU)
AMEL	95	D10S1248	95	D19S433	135
vWA	85	D12S391	135	D1S1656	95
CSF1PO	140	D13S317	95	D21S11	85
FGA	95	D16S539	140	D22S1045	135
TPOX	85	D18S51	140	D2S1338	140
D2S441	95	D3S1358	95	D5S818	85
D7S820	85	D8S1179	135	TH01	85
SE33	135	Penta E	95	Penta D	140
DEFAULT	95				

Bibliography

[1] R. G. Cowell, T. Graversen, S. L. Lauritzen, and J. Mortera. Analysis of forensic DNA mixtures with artefacts. *Applied Statistics*, 64:1–48, 2015.

[2] Øyvind Bleka, Geir Storvik, and Peter 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, 2016.

[3] Corina C. G. Benschop, Jerry Hoogenboom, Pauline Hovers, Martin Slagter, Dennis Kruise, Raymond Parag, Kristy Steensma, Klaas Slooten, Jord H. A. Nagel, Patrick Dieltjes, Vincent van Marion, Heidi van Paassen, Jeroen de Jong, Christophe Creeten, Titia Sijen, and Alexander L. J. Kneppers. DNAxs/DNAStatistX: Development and validation of a software suite for the data management and probabilistic interpretation of DNA profiles. *Forensic Science International: Genetics*, 42:81–89, 2019.

[4] R. M. Hallema. Forensic Evidence Interpretation Using Likelihood Ratios: A Study on Prior Probabilities and LR Distributions for DNA Donors. Master's thesis, Delft University of Technology, 2025.

[5] Øyvind Bleka. Vignette to gammadnamix: An open source R-package to handle continuous model for evaluating STR DNA mixture evidences with artefacts, 2015.