An exact, efficient, and flexible representation of statistical models for DNA profiles

Therese Graversen

Department of Mathematical Sciences, University of Copenhagen

Wednesday 9th November 2016

Overview

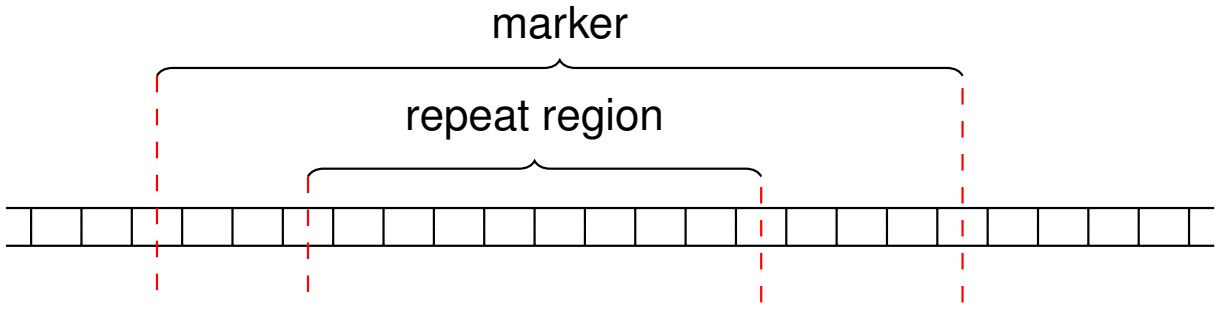
I want to give users the possibility to work with DNA mixture models just like any other statistical model.

- Statistical framework: joint model for EPG and DNA profiles.
- Statistical model checking
- Development of computational methodology

Forensic identification using DNA

DNA for forensic identification

STR marker: An identifiable area (locus) on a chromosome



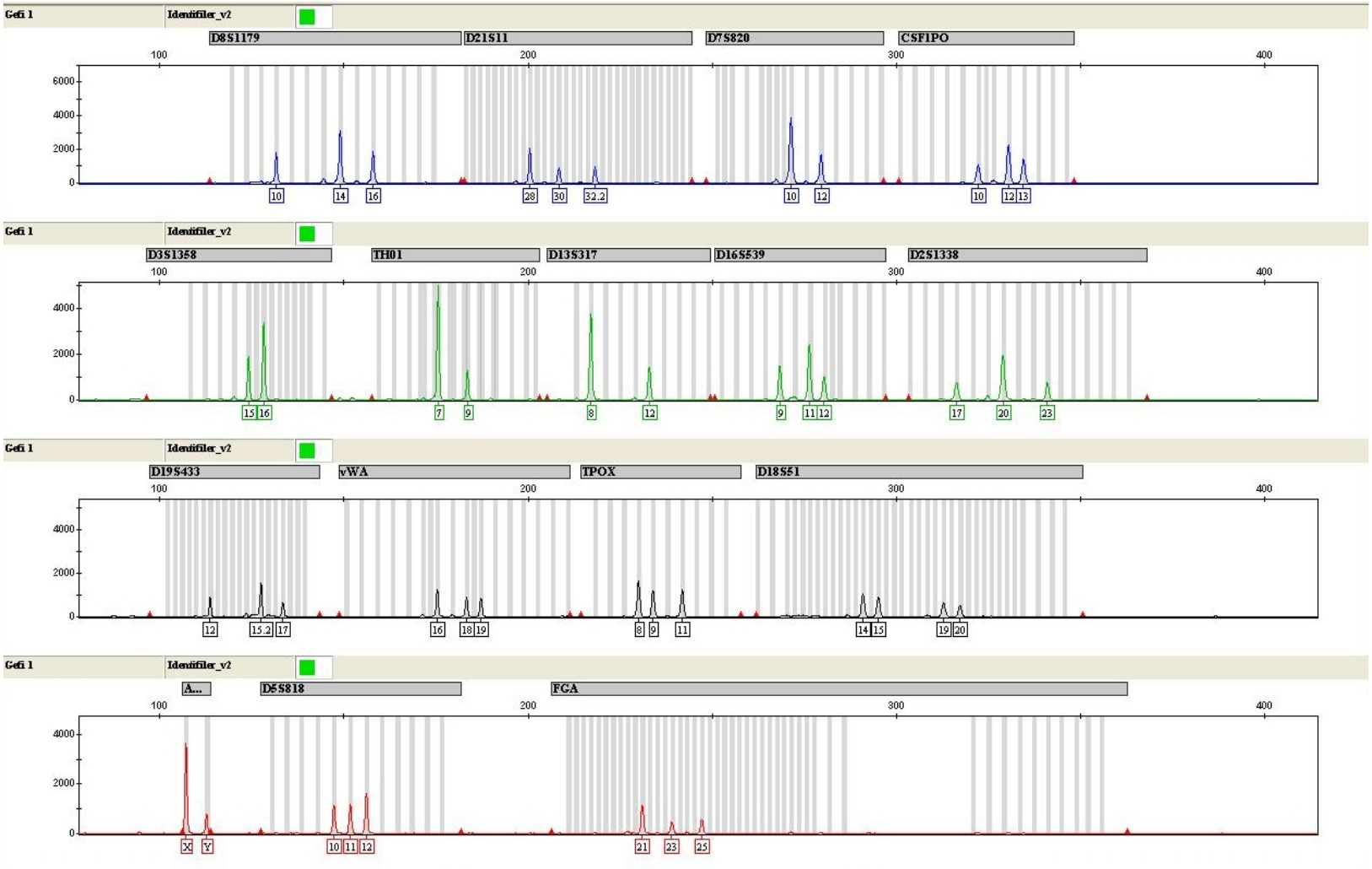
Allele: The DNA sequence at a marker.
Named by the number of repetitions.
TH01, allele 8: $[AATG]_8$

Genotype: Unordered pair of alleles, e.g. (8, 9)

DNA profile: Genotypes across a set of markers

The electropherogram (EPG)

DNA fragments are separated according to length and dye.



Peak heights reflect the amount of each allele.

Artefacts in the EPG

The EPG is not a perfect representation of the DNA sample.

Missing peaks:

Dropout: No peak is detected.

Extra peaks:

Stutter: PCR process gives copies one repeat too short.

Dropin: Small amounts of DNA irrelevant to the case.

DNA mixtures

DNA mixture

A sample of DNA from multiple (≥ 1) *contributors*.

DNA mixtures

DNA mixture

A sample of DNA from multiple (≥ 1) *contributors*.

Hypothesis

An explanation of the mixture in terms of a set of potential contributors.

We distinguish between

Known contributors: The DNA profiles are known to us.

Unknown contributors: The DNA profiles are unknown to us. Typically given a distribution according to a reference population.

A statistical framework

Establishing a good framework

Questions of interest revolve around using peak heights to say something about the underlying DNA profiles.

I want to be able to address at least:

1. Likelihood ratios comparing two proposed hypotheses.
2. Inferring genotypes of contributors to the sample (deconvolution).

Establishing a good framework

Questions of interest revolve around using peak heights to say something about the underlying DNA profiles.

I want to be able to address at least:

1. Likelihood ratios comparing two proposed hypotheses.
2. Inferring genotypes of contributors to the sample (deconvolution).

Focus on modelling data rather than targeting specific questions of interest.

We get a unified, self-consistent, and flexible framework.

DNA mixture models

Our framework is a joint statistical model for
the EPG (peak heights $\boldsymbol{Z}$) and the DNA profiles (genotypes $\boldsymbol{g}$):

$$\Pr(\boldsymbol{Z}, \boldsymbol{g} \mid \mathcal{H}, \psi)$$

where

- $\boldsymbol{Z}$: set of peak heights across the EPG(s).
- $\boldsymbol{g}$: set of genotypes for the contributors under $\mathcal{H}$.
- $\mathcal{H}$: hypothesis under consideration.
- ψ: fixed, but unknown, model parameters; e.g. the proportion of DNA from each contributor.

R package `DNAmixtures` offers an implementation of this statistical framework.

DNA mixture models

The model is naturally specified in two parts

$$\Pr(\boldsymbol{Z}, \boldsymbol{g} \mid \mathcal{H}, \psi) = \Pr(\boldsymbol{Z} \mid \boldsymbol{g}, \psi) \Pr(\boldsymbol{g} \mid \mathcal{H})$$

A model for peak heights for fixed genotypes:

$$\Pr(\boldsymbol{Z} \mid \boldsymbol{g}, \mathcal{H}, \psi) = \Pr(\boldsymbol{Z} \mid \boldsymbol{g}, \psi),$$

and a model for contributors' genotypes:

$$\Pr(\boldsymbol{g} \mid \mathcal{H}, \psi) = \Pr(\boldsymbol{g} \mid \mathcal{H}).$$

Likelihood function

The likelihood function is based on the observed peak heights

$$\Pr(\boldsymbol{Z} \mid \mathcal{H}, \psi) = \sum_{\boldsymbol{g}} \Pr(\boldsymbol{Z} \mid \boldsymbol{g}, \psi) \Pr(\boldsymbol{g} \mid \mathcal{H}).$$

The sum is generally huge.

The (exact) likelihood function can be computed easily, using the method of Graversen and Lauritzen (2014).

Is this a useful approach?

Likelihood ratios

Compare the likelihood of the observed peak heights under $\mathcal{H}_p$ and $\mathcal{H}_d$:

$$\text{LR} = \frac{\Pr(\boldsymbol{Z} \mid \mathcal{H}_p, \psi_p)}{\Pr(\boldsymbol{Z} \mid \mathcal{H}_d, \psi_d)}.$$

I can compute the likelihood $\Pr(\boldsymbol{Z} \mid \mathcal{H}, \psi)$
for any hypothesis $\mathcal{H}$ and any parameter ψ.

Mixture deconvolution

First, formulate a hypothesis $\mathcal{H}$ characterising the contributors, e.g. $\mathcal{H} : K_1 \,\&\, K_2 \,\&\, U$.

Identify various sets of posterior most probable profiles for the unknown contributors under $\mathcal{H}$.

The joint model

$$\Pr(\boldsymbol{Z}, \boldsymbol{g} \mid \mathcal{H}, \psi)$$

implies a posterior distribution for the genotypes of contributors under $\mathcal{H}$,

$$\Pr(\boldsymbol{g} \mid \boldsymbol{Z}, \mathcal{H}, \psi)$$

Establishing a good framework

Ideally, we distinguish between

- A statistical model
- How we use the model in targeting questions of interest
- How we compute necessary quantities
- Software implementing any of the above

Challenging an interpretation

Statistical interpretation of a DNA mixture

Think of an interpretation as consisting of

- Hypothesis $\mathcal{H}$ specifying contributors.
- Model assumptions on genotypes, $\Pr(\boldsymbol{g} \mid \mathcal{H})$.

  Example: Basic model for unknown contributors, US Caucasian population.
- The peak height model $\Pr(\boldsymbol{Z} \mid \boldsymbol{g}, \psi)$.

  Example: Model of Cowell et al. (2015)
- Parameters ψ.

  Example: MLE $\hat{\psi}$ under $\mathcal{H}$

Challenging the interpretation

We should justify that an interpretation adequately describes the case at hand.

The model framework allows us to do this systematically.

Challenging the interpretation

We should justify that an interpretation adequately describes the case at hand.

The model framework allows us to do this systematically.

Why challenge?

- When calculating the LR, what if neither of the compared hypotheses provide a good explanation of the mixture?

Challenging the interpretation

We should justify that an interpretation adequately describes the case at hand.

The model framework allows us to do this systematically.

Why challenge?

- When calculating the LR, what if neither of the compared hypotheses provide a good explanation of the mixture?
- The peak height model could be unsuitable

Challenging the interpretation

We should justify that an interpretation adequately describes the case at hand.

The model framework allows us to do this systematically.

Why challenge?

- When calculating the LR, what if neither of the compared hypotheses provide a good explanation of the mixture?
- The peak height model could be unsuitable
- Model for the unknown contributors could be unsuitable

Challenging the interpretation

We should justify that an interpretation adequately describes the case at hand.

The model framework allows us to do this systematically.

Why challenge?

- When calculating the LR, what if neither of the compared hypotheses provide a good explanation of the mixture?
- The peak height model could be unsuitable
- Model for the unknown contributors could be unsuitable
- Errors in data, or preprocessing.

Challenging the interpretation

We should justify that an interpretation adequately describes the case at hand.

The model framework allows us to do this systematically.

Why challenge?

- When calculating the LR, what if neither of the compared hypotheses provide a good explanation of the mixture?
- The peak height model could be unsuitable
- Model for the unknown contributors could be unsuitable
- Errors in data, or preprocessing.

Note that validation studies are also challenging interpretations.

Key idea

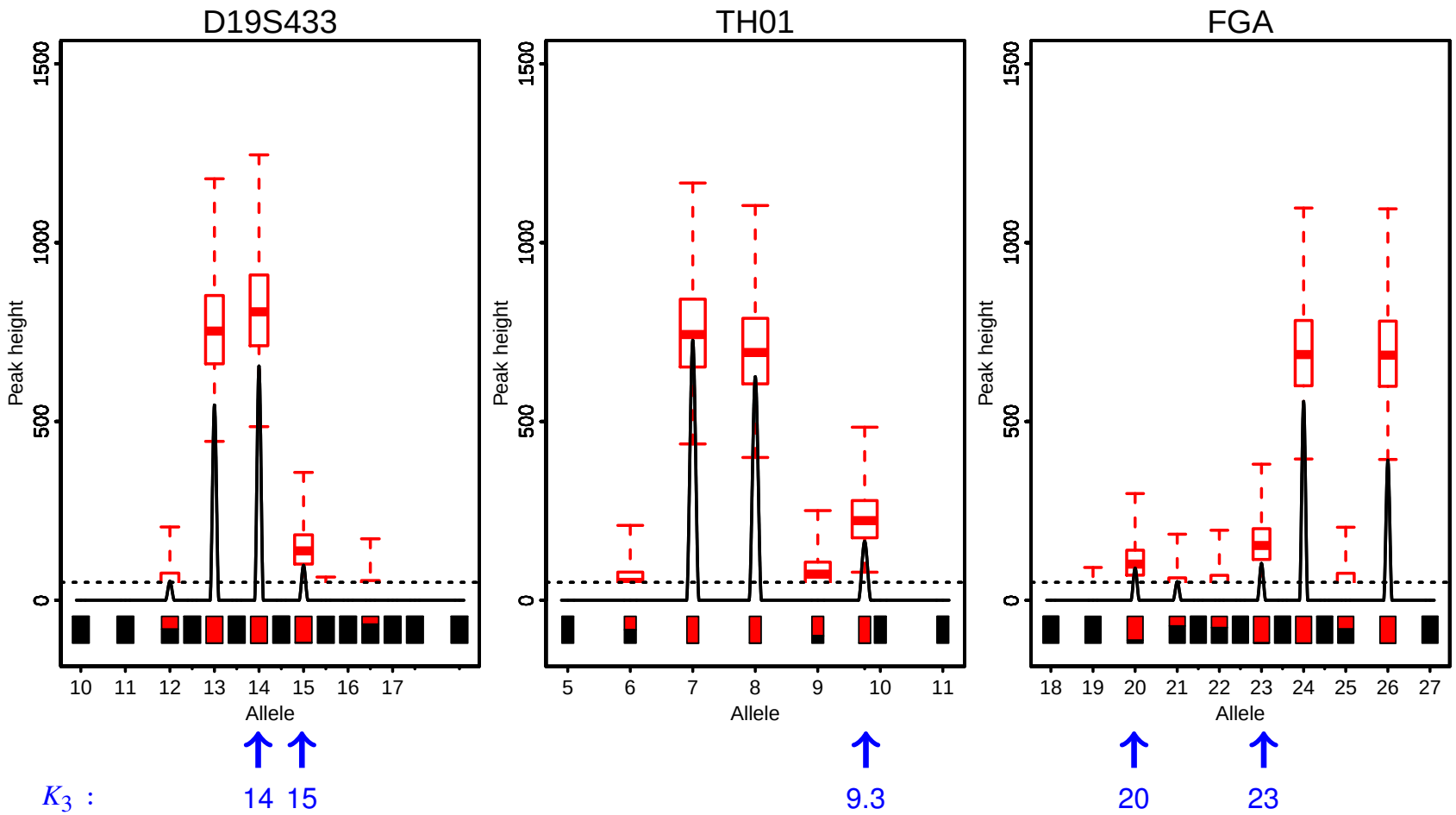
Compare observed quantities to their theoretical distribution under a given interpretation of the DNA mixture.

Investigate and assess assumptions by

- visual diagnostic methods
- statistical tests

Variability of peak heights

$$\mathcal{H}_p : K_1 \,\&\, K_2 \,\&\, K_3 \,\&\, U$$

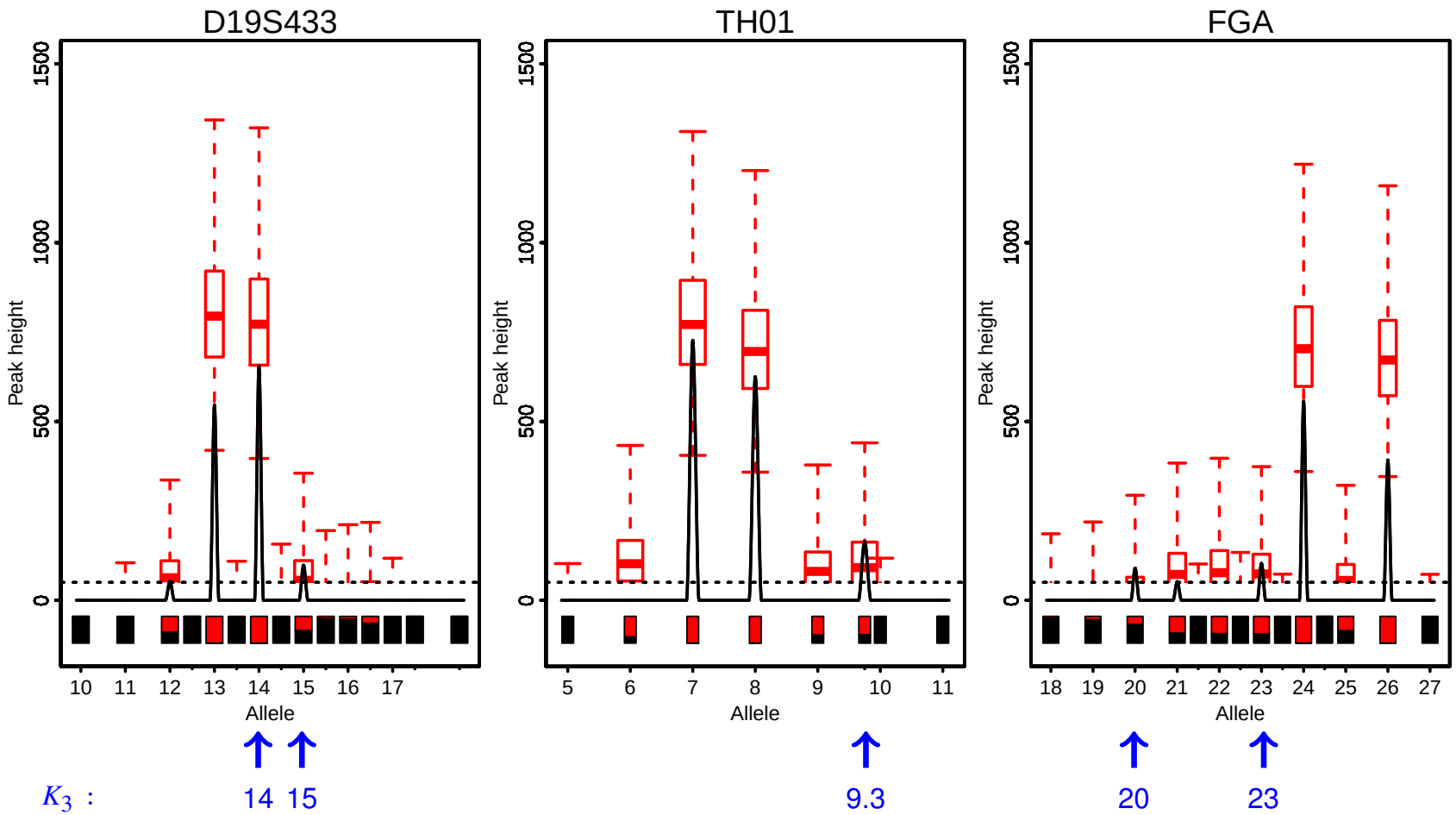


Black bars: probability of no peak.

Antennas: 99% prediction intervals.

Variability of peak heights

$$\mathcal{H}_d : K_1 \,\&\, K_2 \,\&\, U_1 \,\&\, U_2$$



Black bars: probability of no peak.

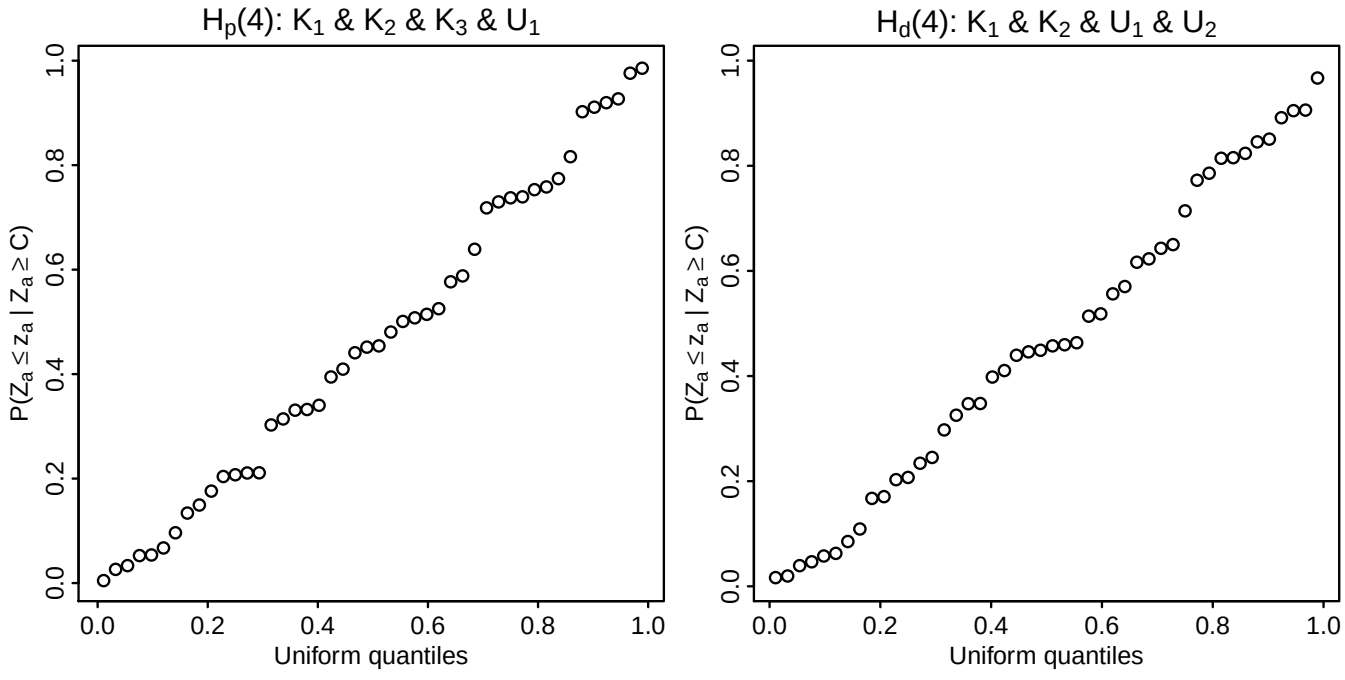
Antennas: 99% prediction intervals.

Probability integral transforms

Transform peak heights to be uniformly distributed

If the model is correct, the probability integral transforms of peak heights above threshold are uniformly distributed, i.e.

$$\Pr(Z_a \leq z_a \,|\, Z_a \geq C) \sim \text{unif}[0, 1]$$



Presence/absence of peaks

Given an interpretation,
can we predict which peaks are present in the EPG?

Presence/absence of peaks

Consider all allelic positions in the EPG in some order.

At position a the peak is either *present* or *absent*.

Presence/absence of peaks

Consider all allelic positions in the EPG in some order.

At position a the peak is either *present* or *absent*.

A penalty

$$Y_a = -\log \Pr(E_a \mid Z_b, b < a)$$

assigns a large positive number to a poor prediction of the observed event E_a.

Presence/absence of peaks

Consider all allelic positions in the EPG in some order.

At position a the peak is either *present* or *absent*.

A penalty

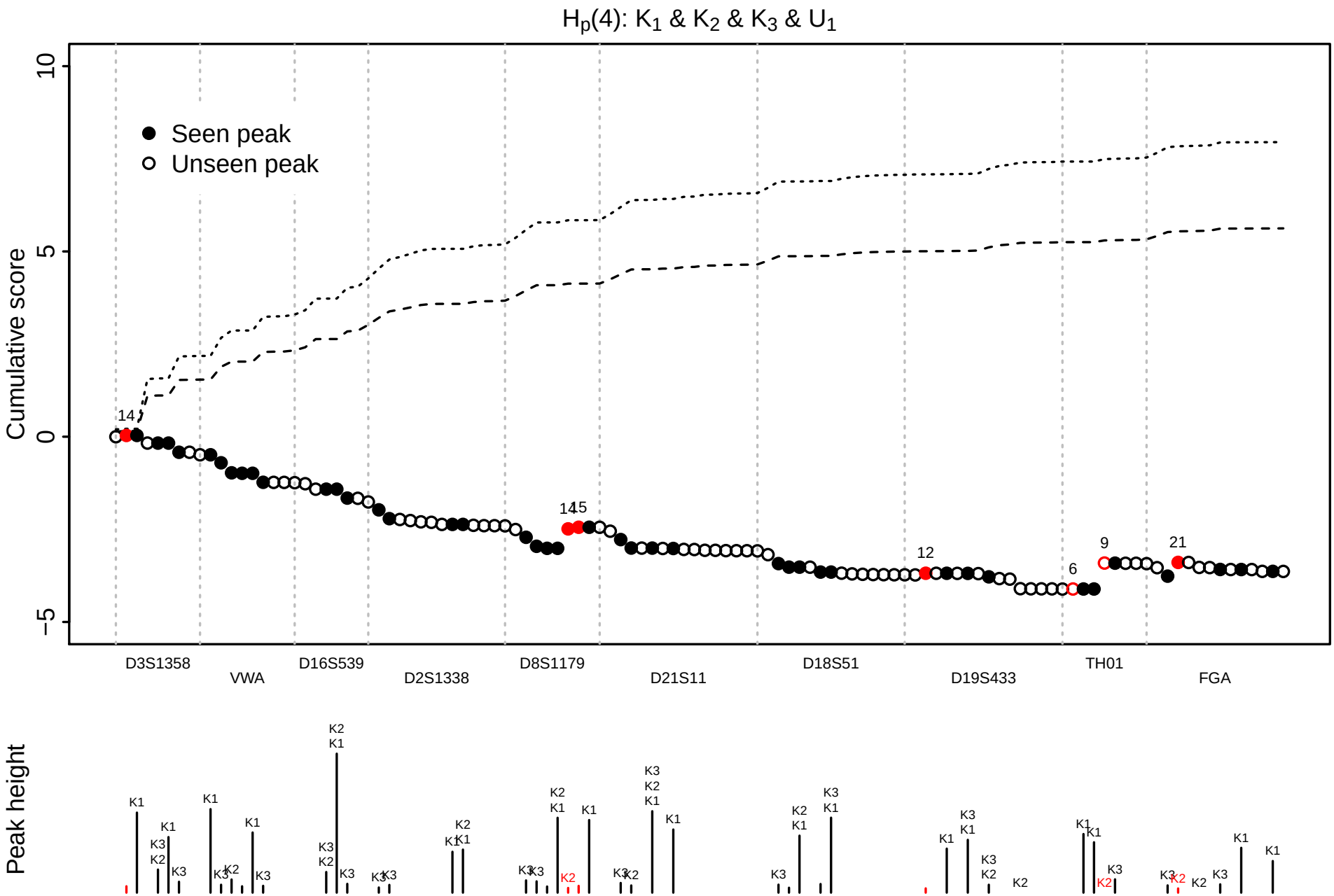
$$Y_a = -\log \Pr(E_a \mid Z_b, b < a)$$

assigns a large positive number to a poor prediction of the observed event E_a.

Visualise this by a *prequential monitor* plot of

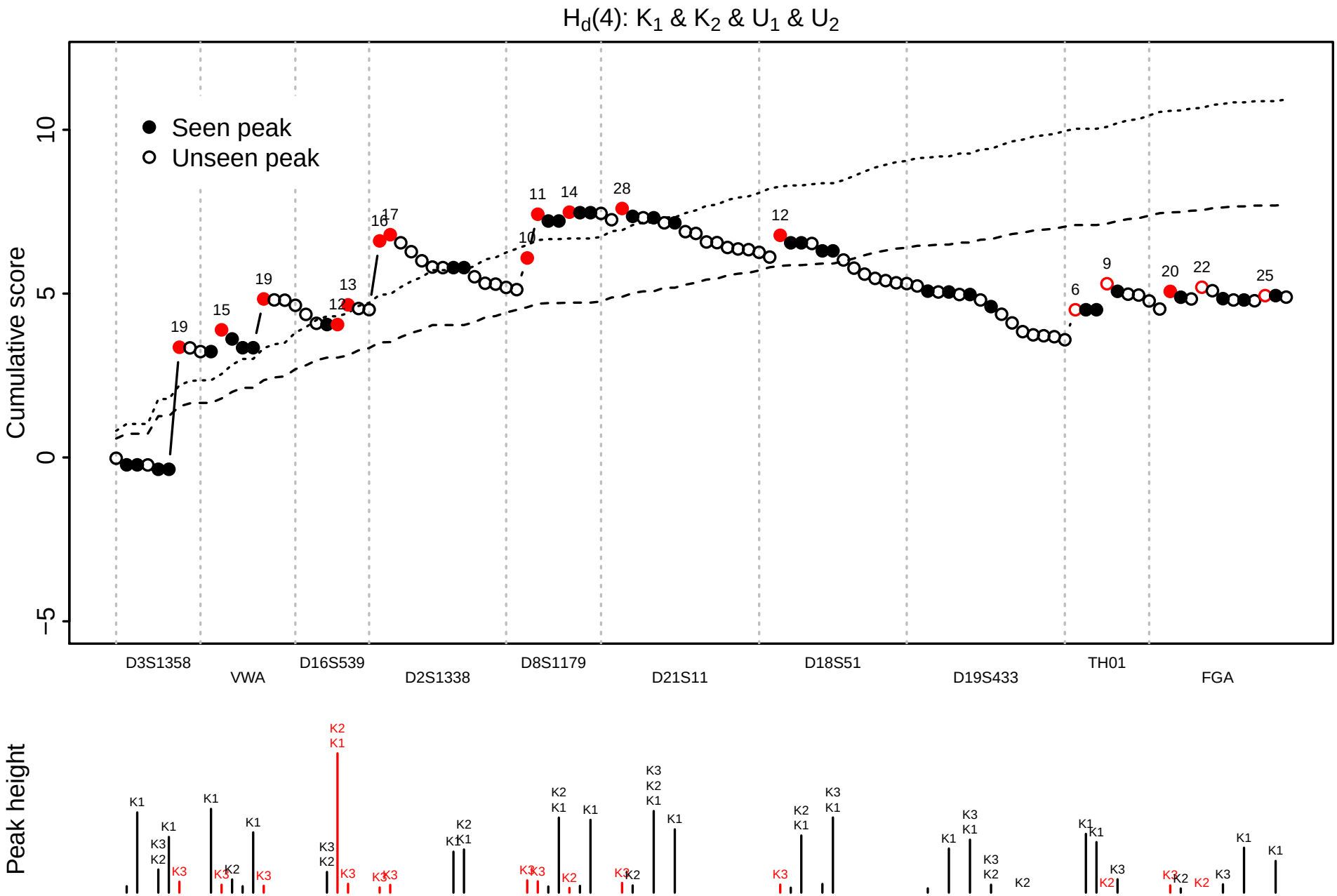
$$\sum_a (Y_a - \mathbb{E}\, Y_a)$$

against alleles a.



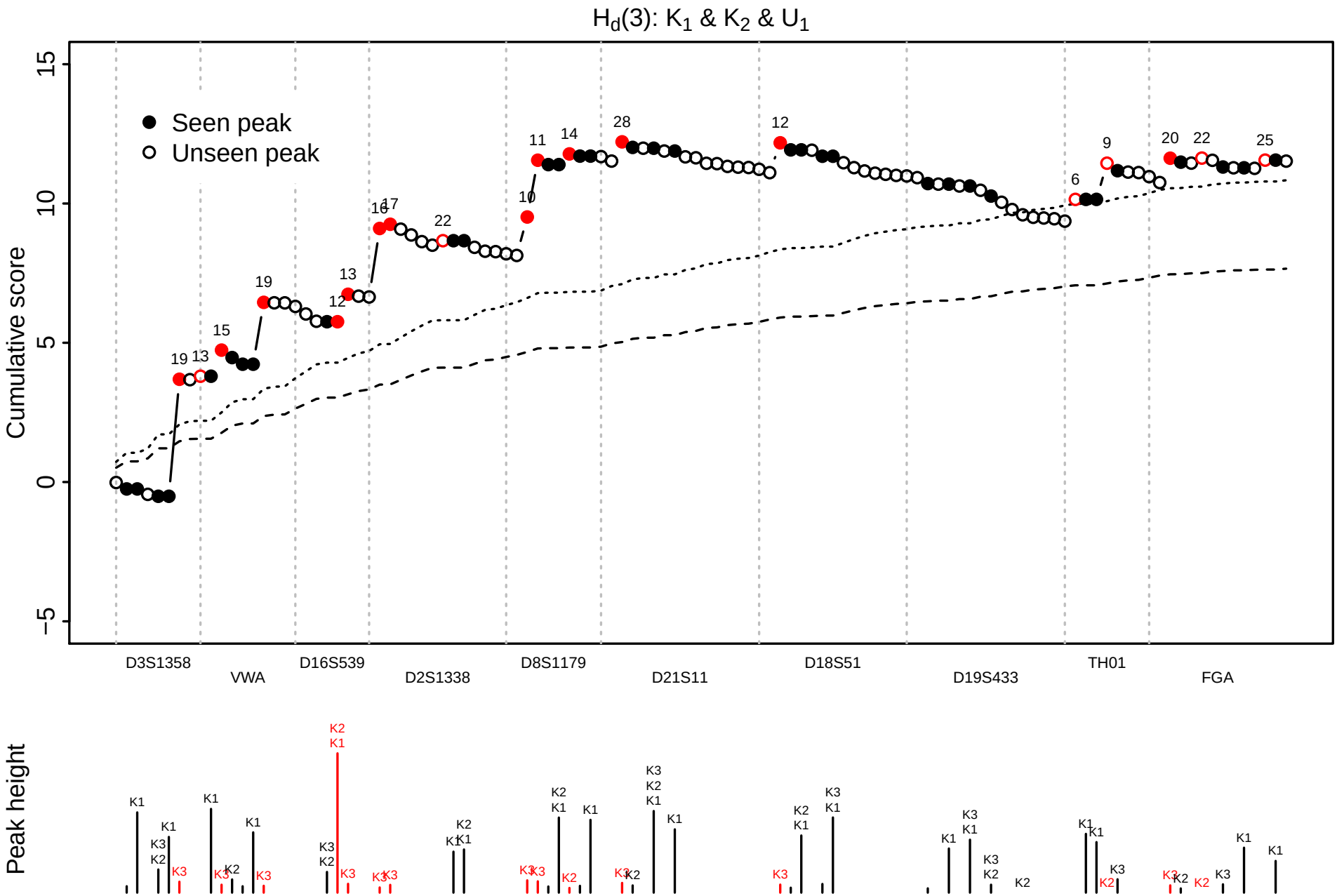
Upwards jump: worse than expected.

Downwards jump: better than expected.



Upwards jump: worse than expected.

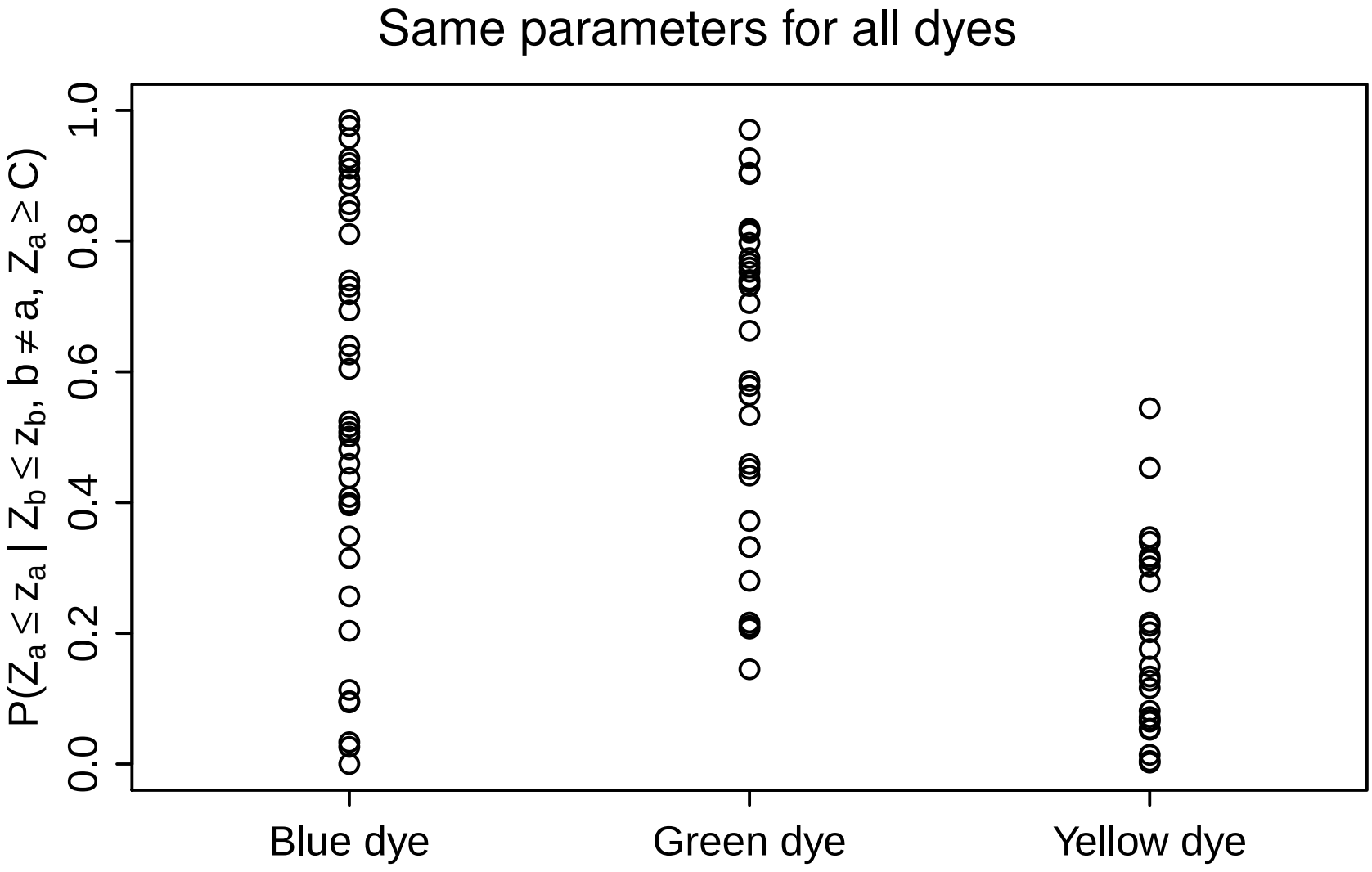
Downwards jump: better than expected.



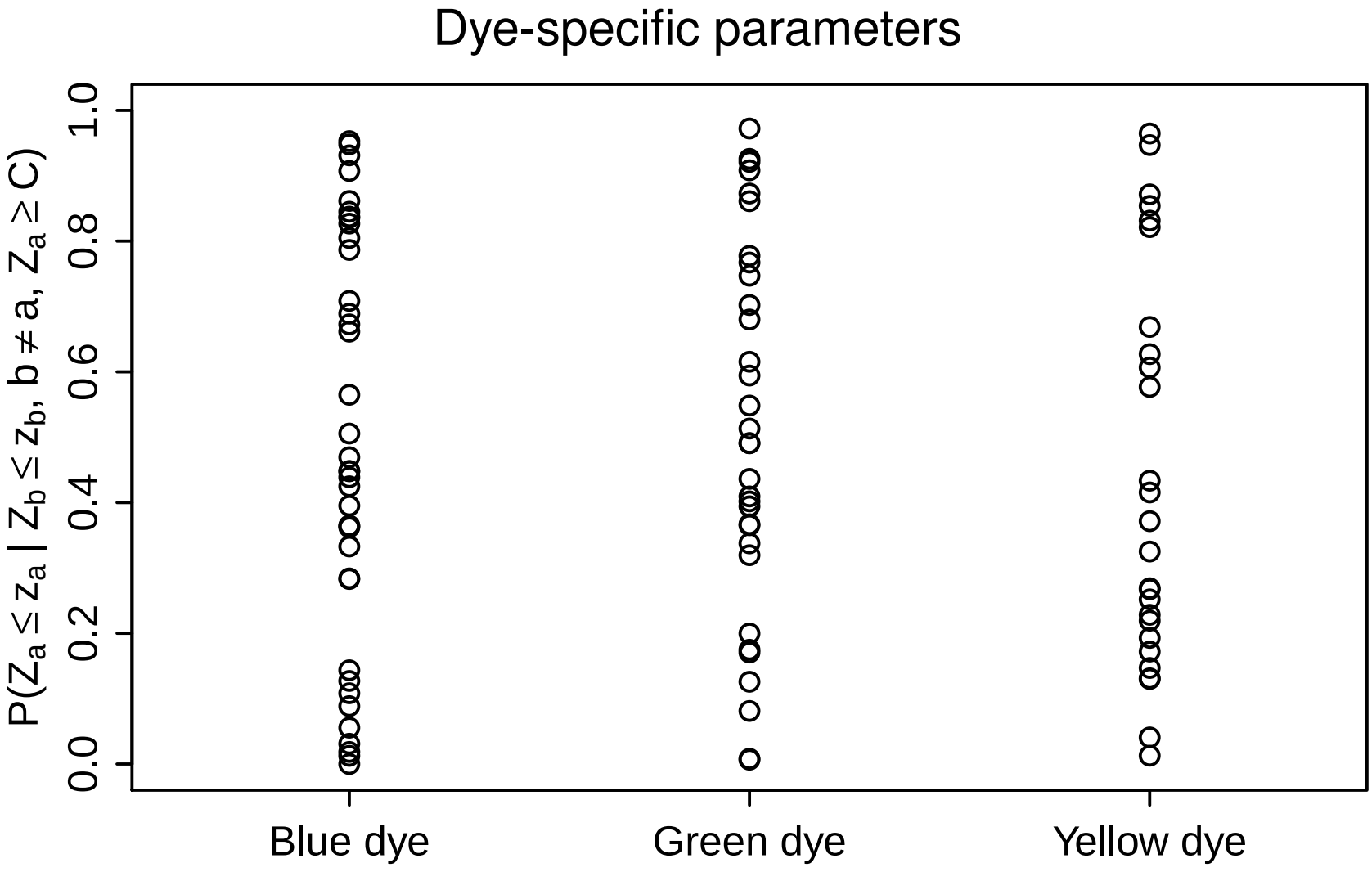
Upwards jump: worse than expected.

Downwards jump: better than expected.

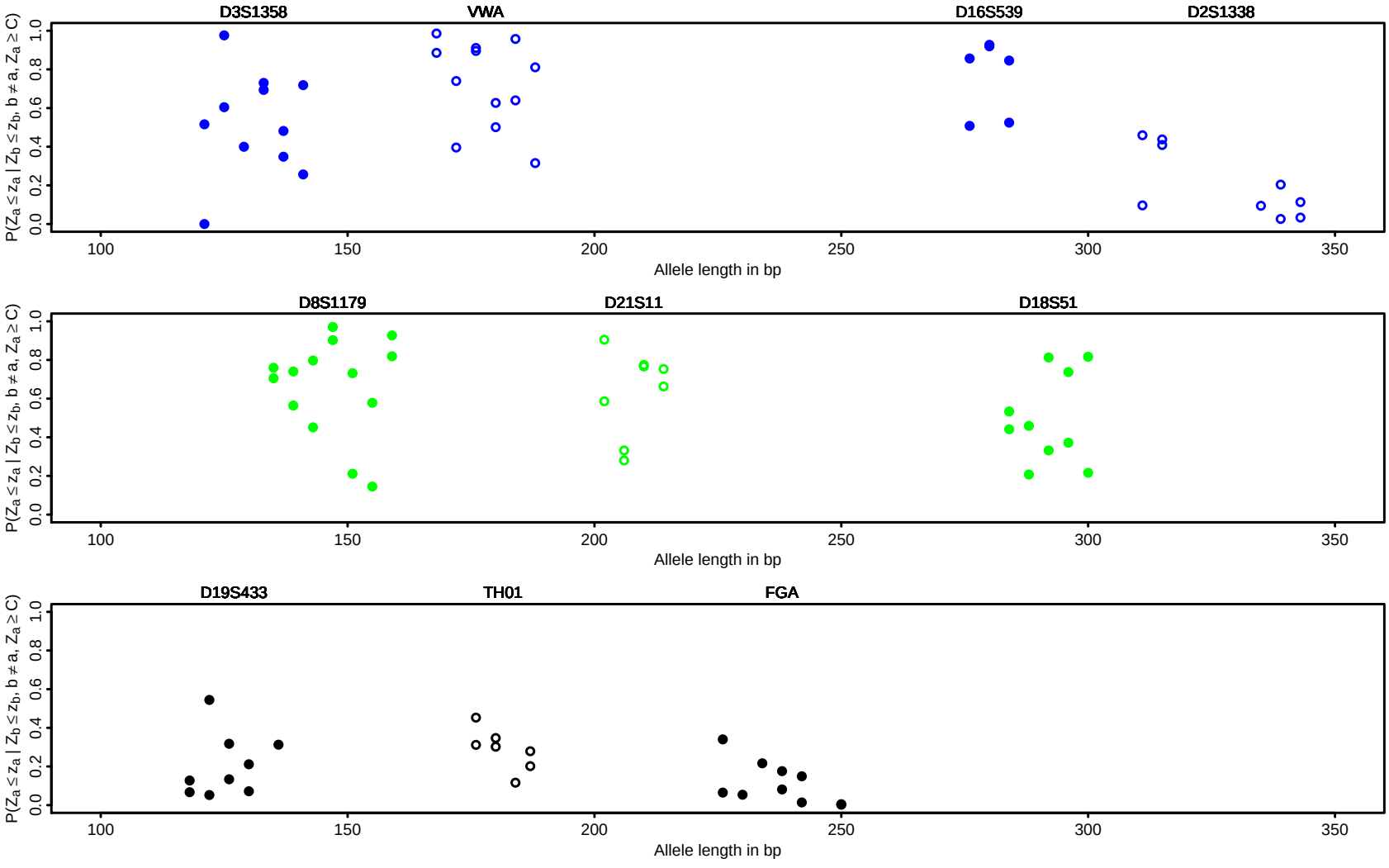
Is there a trend between dyes?



Is there a trend between dyes?



Is there a trend in fragment length?



Implementation: DNAmixtures

The `DNAmixtures` R package

An implementation of the statistical framework:

- exact evaluation of likelihood function and other quantities
- easy access to various marginal and conditional distributions: probabilities, quantiles, simulation
- flexibility to explore changes in model assumptions

Statistical tools, such as

- maximum likelihood estimation with arbitrary constraints
- visual model checking methods

Computational approach

Computational approach

Due to computational complexity, it is common to reduce the genotype state space through various heuristic measures.

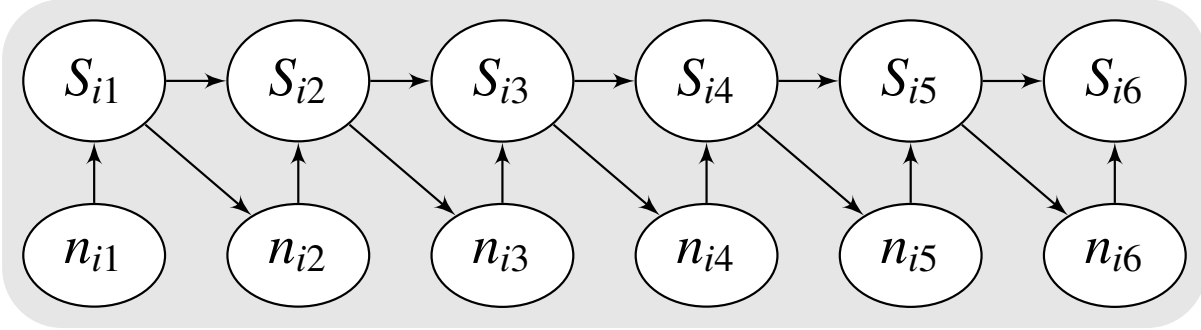
I want to do exact computations as far as possible.

Strategy: Find efficient Bayesian network representations of the model and rely on standard techniques for networks.

Probability propagation efficiently computes

- marginal distributions
- conditional distributions
- sums over the entire state space

Genotype representation

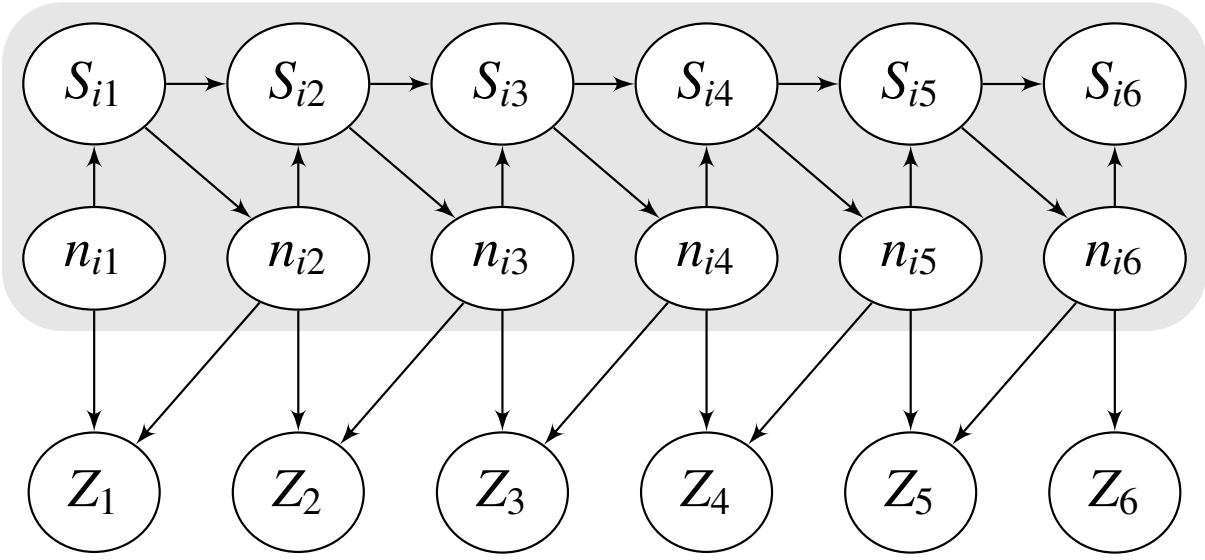


At a marker, an unknown person has 2 alleles sampled independently using (fixed and known) allele frequencies.

When allocating n_{ia} alleles of type a we only need to know how many alleles are already allocated, $S_{i,a-1}$.

More complex genotype models can be expressed similarly.

Genotype and observed peak heights



Computations can be done efficiently because

- allele counts can be “allocated” sequentially, independent of previous allocation.
- peak heights depend “locally” on the allele counts.

Bayesian network machinery

A Bayesian Network represents the joint distribution of the network variables.

Propagating information about network variables gives:

- Conditional distribution given the information
- Probability of the information

Computing sums over unknown genotypes

Expectations by propagation

Propagating information $h(x)$
in a network representing a distribution $p(x)$
computes

$$\sum_x h(x)p(x) = \mathbb{E}\, h(X)$$

Compute the likelihood function

$$f(\boldsymbol{z};\psi) = \sum_{\boldsymbol{g}} f(\boldsymbol{z} \mid \boldsymbol{g};\psi)p(\boldsymbol{g})$$

by a single propagation by choosing

$$h(\boldsymbol{g}) = f(\boldsymbol{z} \mid \boldsymbol{g};\psi).$$

Software:

R package `DNAmixtures` by Therese Graversen
`http://dnamixtures.r-forge.r-project.org/`

Model checking and computational methodology:

Computational aspects of DNA mixture analysis.
Therese Graversen and Steffen Lauritzen.
Statistics and Computing, 2014.

Statistical and Computational Methodology for the Analysis of Forensic DNA Mixtures with Artefacts.
Therese Graversen.
DPhil. University of Oxford. 2014.
`https://ora.ox.ac.uk:443/objects/uuid:4c3bfc88-25e7-4c5b-968f-10a35f5b82b0`

The statistical model:

Analysis of forensic DNA mixtures with artefacts (with discussion).
Robert G. Cowell, Therese Graversen, Steffen Lauritzen, and Julia Mortera
Journal of the Royal Statistical Society, series C. Volume 64, Issue 1, 1-48, 2015