The theory behind EuroForMix

and its functionalities

updated for version 3.0

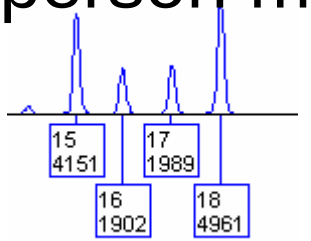
by Øyvind Bleka

EUROFORGEN
Network of Excellence

Oslo
University Hospital

Motivation

- Assume this is a 2-person mixture evidence profile:

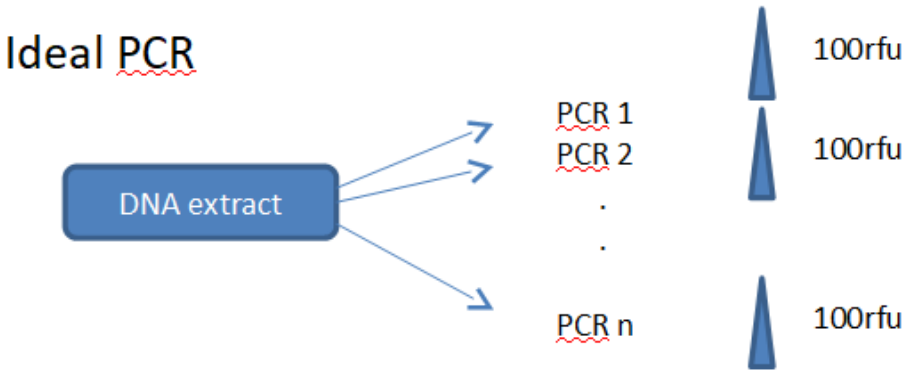


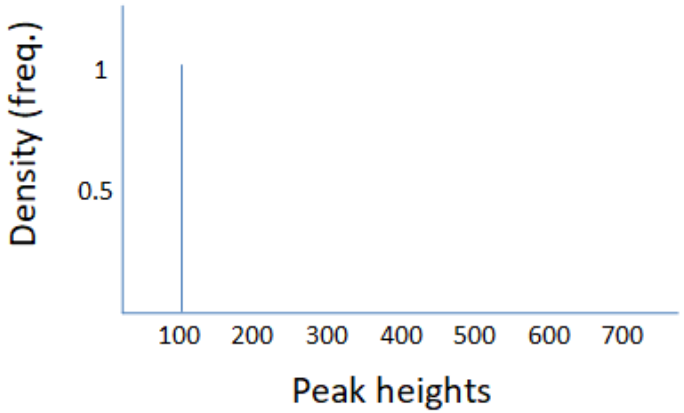
- The qualitative information is given as: 15 16 17 18
  - LRmix Studio uses this information
  - Contributors can have several genotype combinations

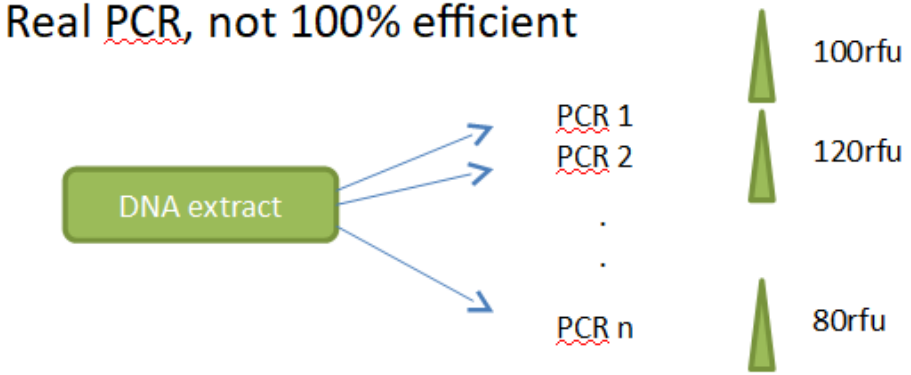
- The peak heights give extra quantitative information:
  - They reflect the amount of DNA coming from each contributors
  - The contributors are much more likely to have genotypes 15/18 and 16/17, than other combination (remember prev. lecture)

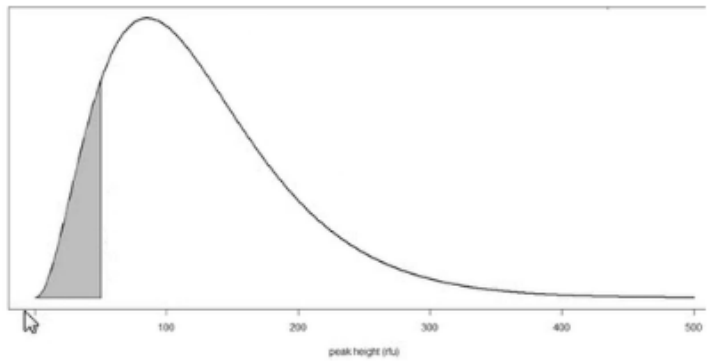
Peak heights are stochastic

- It's difficult to reveal what is really going on by just looking at the peak heights directly, since they are stochastic.
- We assume that the distribution of peak heights arising in the PCR amplification of STRs follows a Gamma distribution

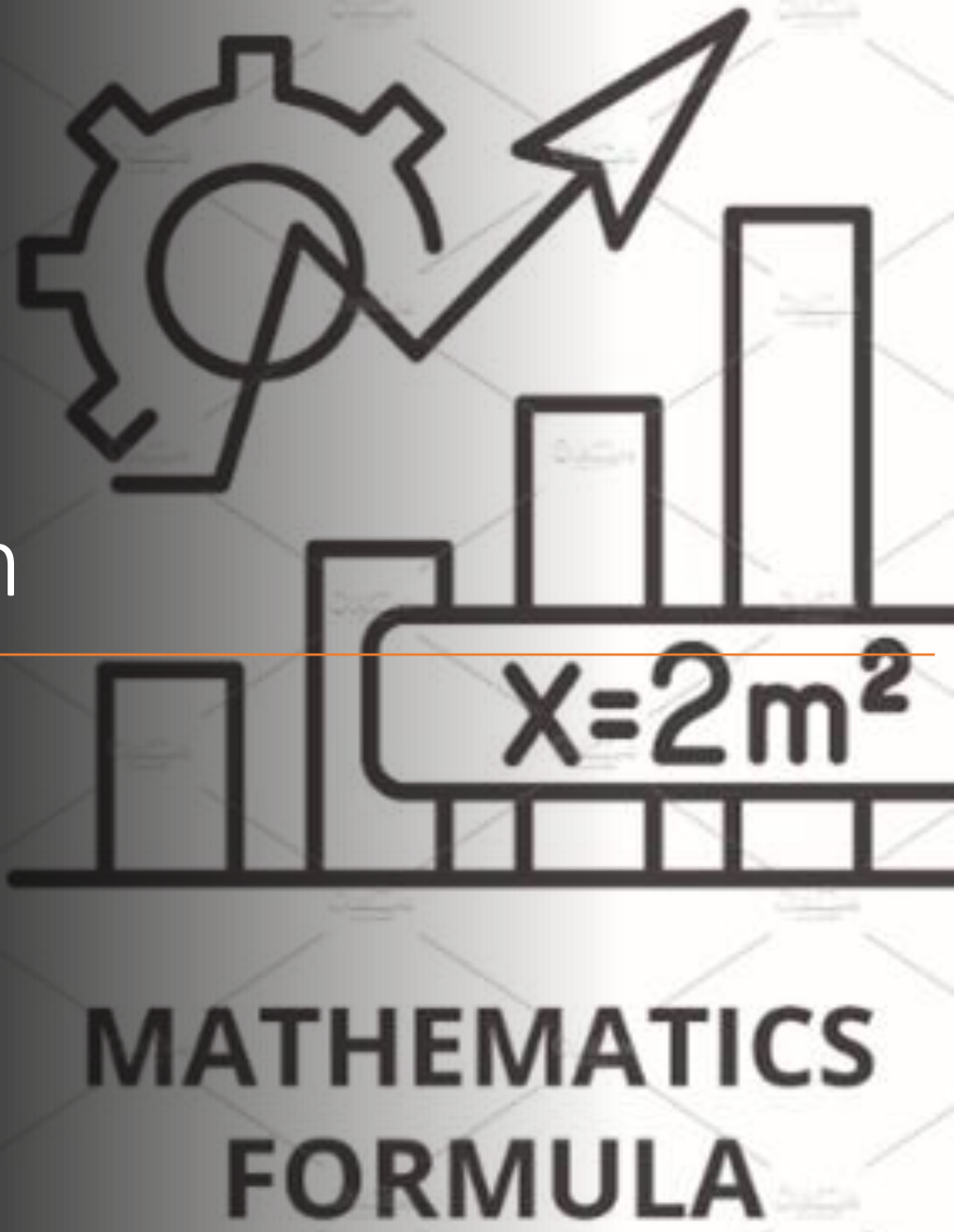








Model theory of the PH distribution



Gamma model for PH-distribution

Probability density function (pdf):

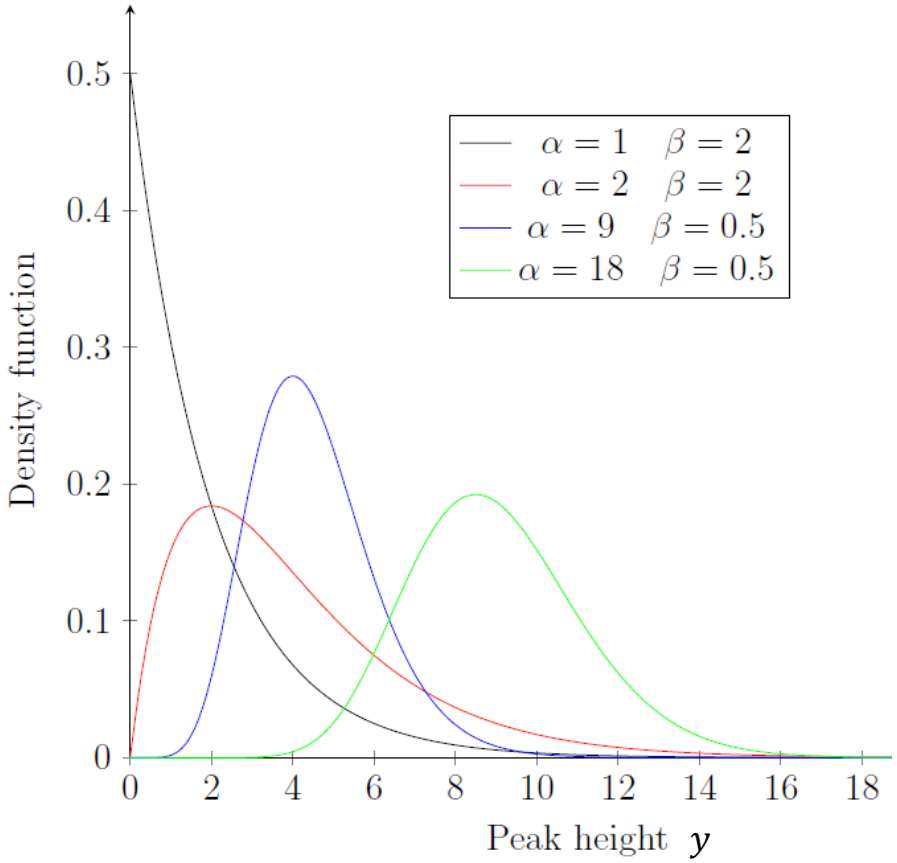
$$p(y|\alpha,\beta) = gamma(y|\alpha,\beta)$$

α is the **shape** parameter

β is the **scale** parameter

We say that P.H. Y is a **stochastic variable** taking value y

$$Y \sim gamma(\alpha,\beta)$$



EuroForMix works with a re-parameterized model:

$\alpha = \frac{1}{\omega^2}$ is the shape parameter

$\beta = \mu\omega^2$ is the scale parameter

$$p(y|\mu,\omega) = gamma(y|\omega^{-2},\mu\omega^2) \quad \Leftrightarrow \quad Y \sim gamma(\omega^{-2},\mu\omega^2)$$

Interpretation of the PH distribution

$$p(y|\mu,\omega) = gamma(y|\omega^{-2}, \mu\omega^{2})$$

$\approx Normal(\mu, \sigma = \mu * \omega)$
For small ω (around 0.1)

Interpretation:

Expectation $E[y] = \mu$ ← «P.H.expectation»

Coefficient-of-variation $\frac{Sd[y]}{E[y]} = \omega$ ← «P.H.variation»
(a standardized measure for variation)

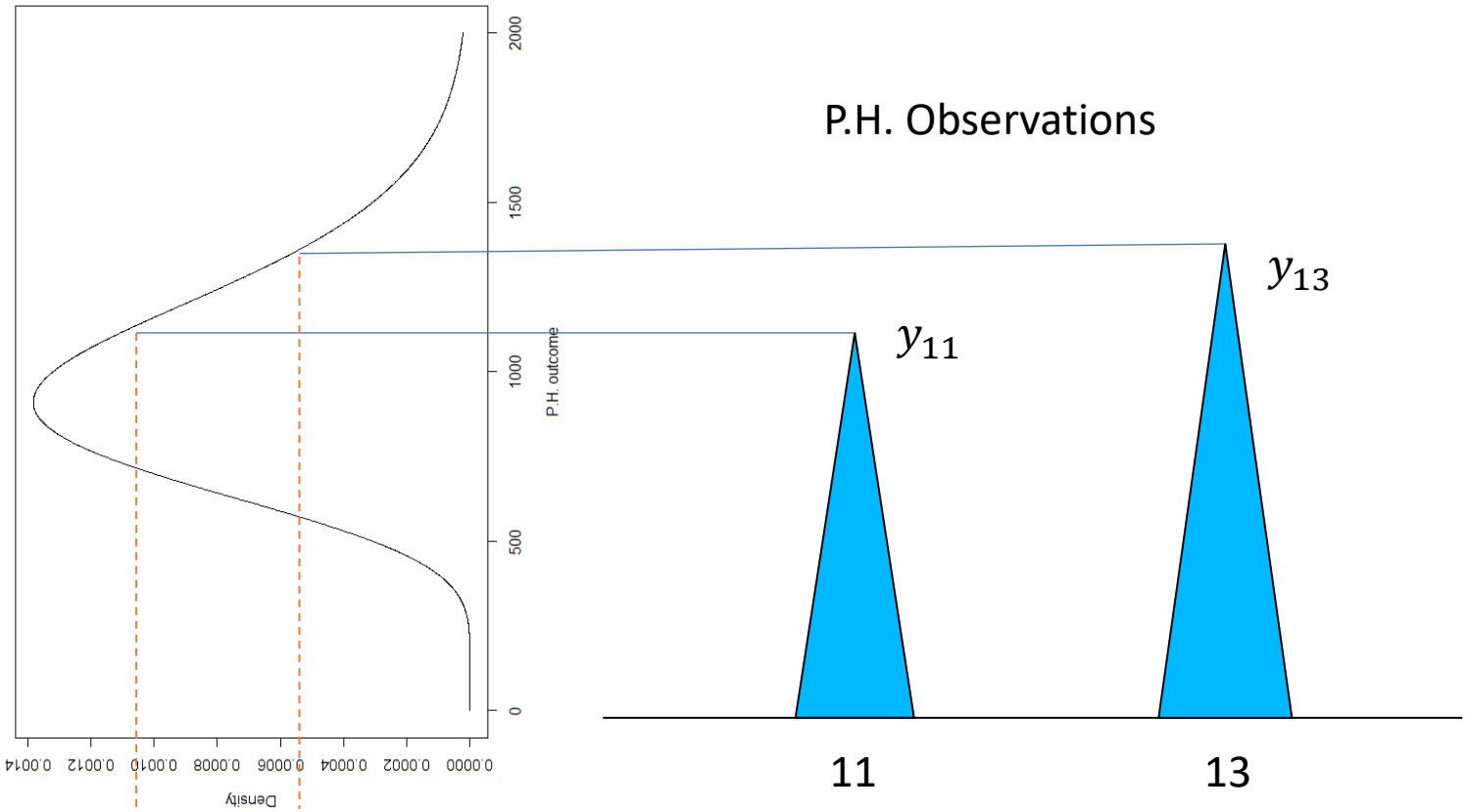
μ, ω are part of the ***unknown model parameter set*** $\boldsymbol{\theta}$ (same as $\boldsymbol{\beta}$ from prev.lecture) defining the peak height model $p(y|\boldsymbol{\theta})$

The parameters in $\boldsymbol{\theta}$ are unknown and must be estimated based on data (EPG)

Example: PH-distribution for one contributor

For heterozygot variant: $Y_{11}, Y_{13} \sim gamma(\omega^{-2}, \mu\omega^{2})$

$\mu = 1000$
$\omega = 0.2$



Allele weights (likelihoods): $p(y_{11}|\mu, \omega)$ and $p(y_{13}|\mu, \omega)$

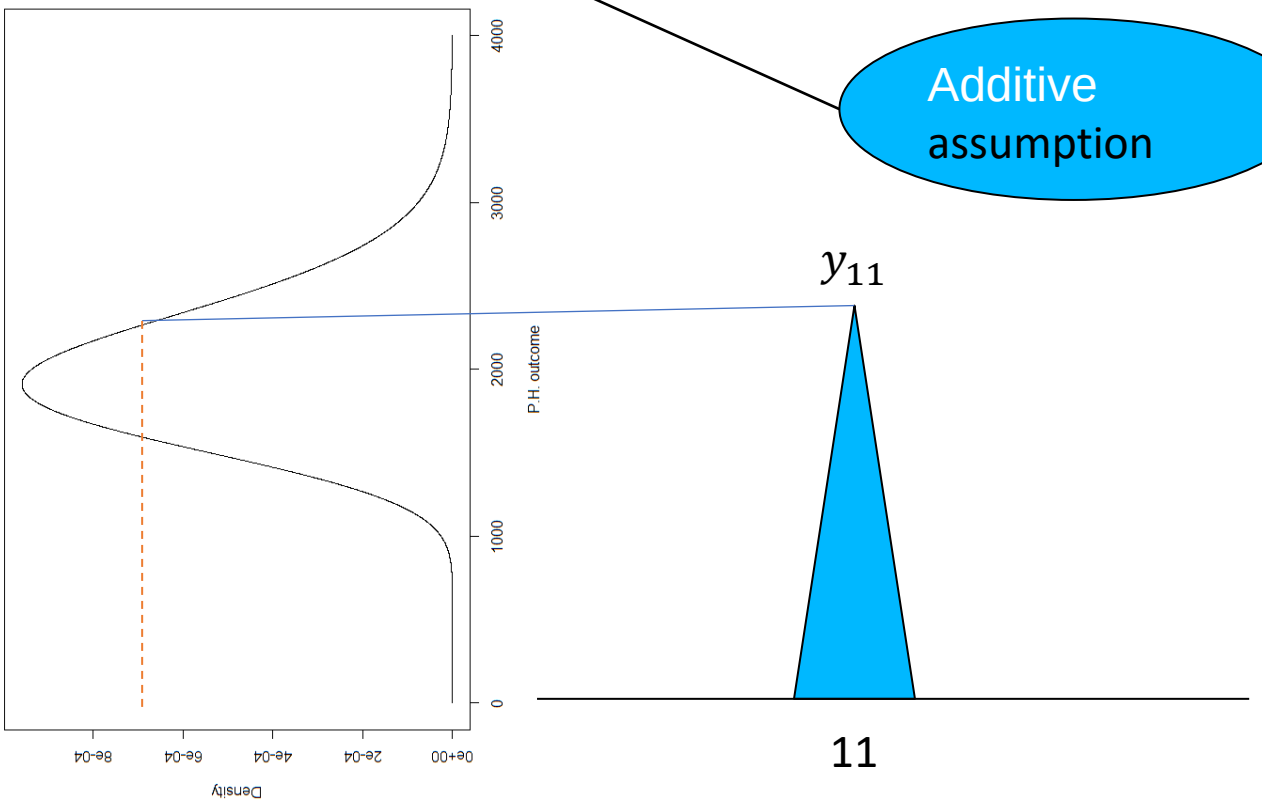
Example: PH-distribution for one contributor

For homozygot variant:

$$Y_{11} \sim gamma(2 * \omega^{-2}, \mu\omega^2)$$

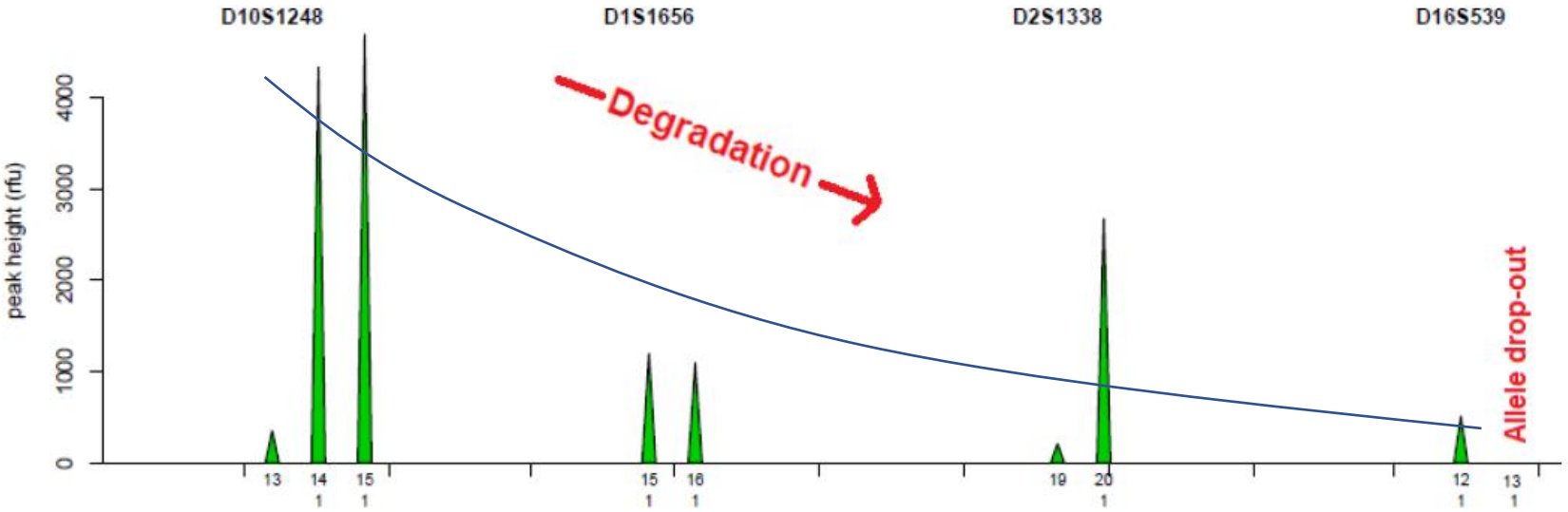
Additive assumption

$\mu = 1000$
$\omega = 0.2$

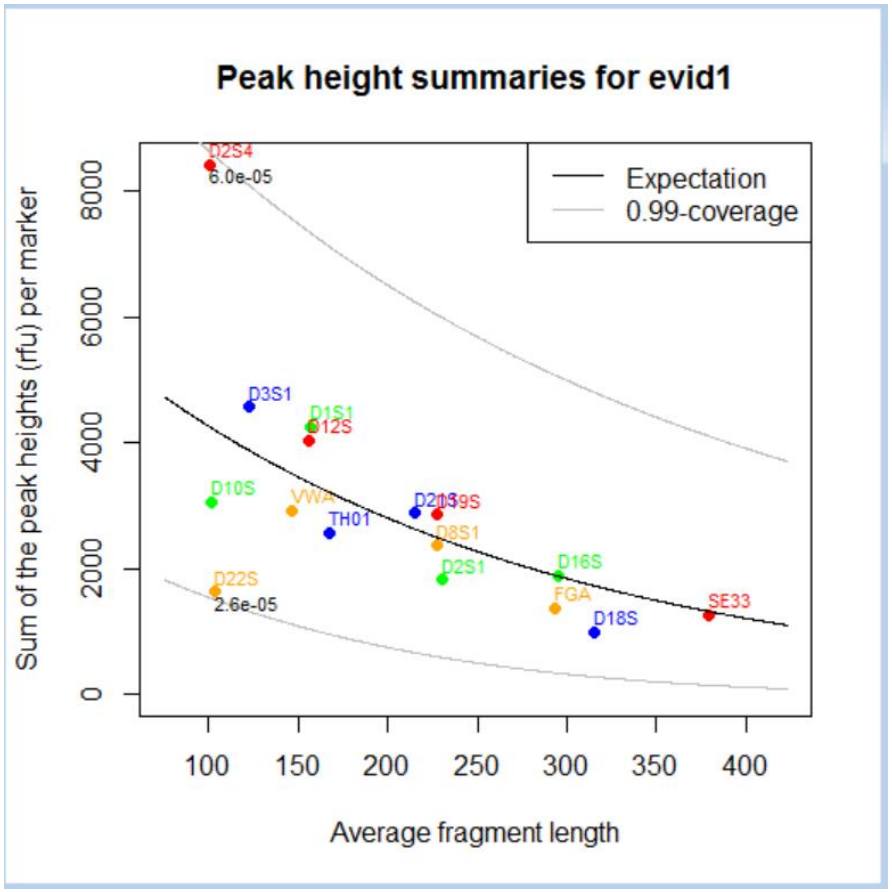


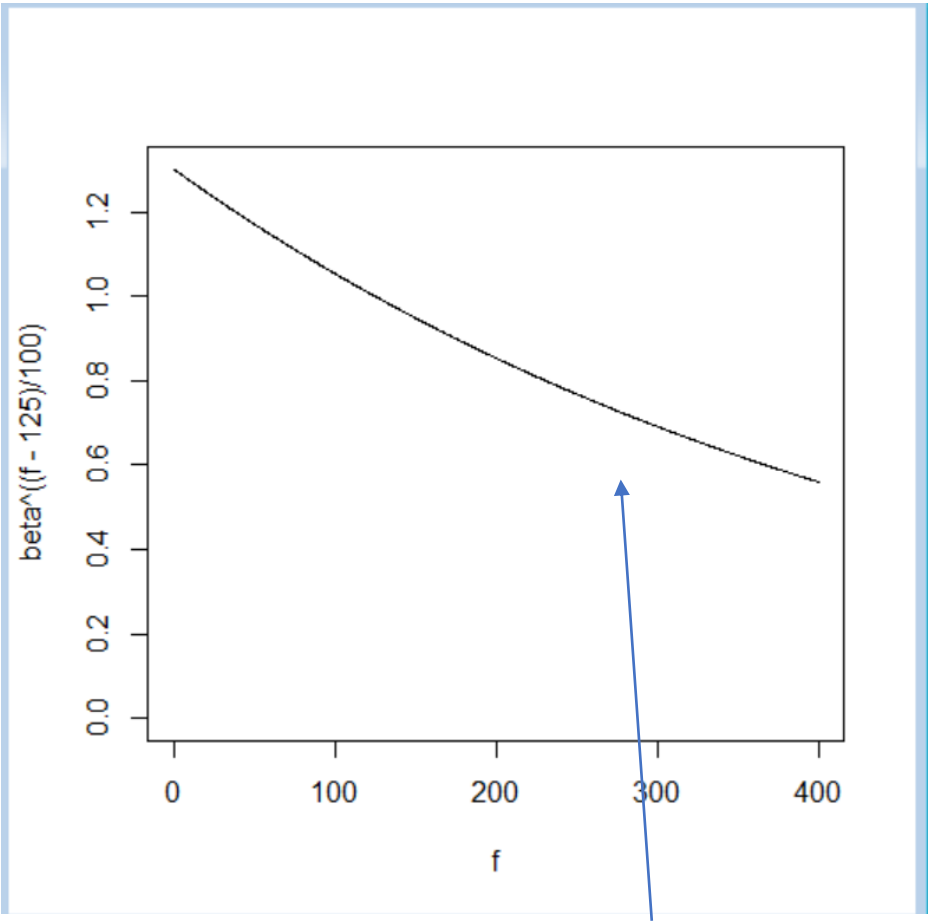
Allele weight (likelihood): $p(y_{11}|\mu, \omega)$

Samples with “lesser quality”



Probabilistic model for degradation





Expected peak heights proportional to $\beta^{\frac{f_{m,a}-125}{100}}$

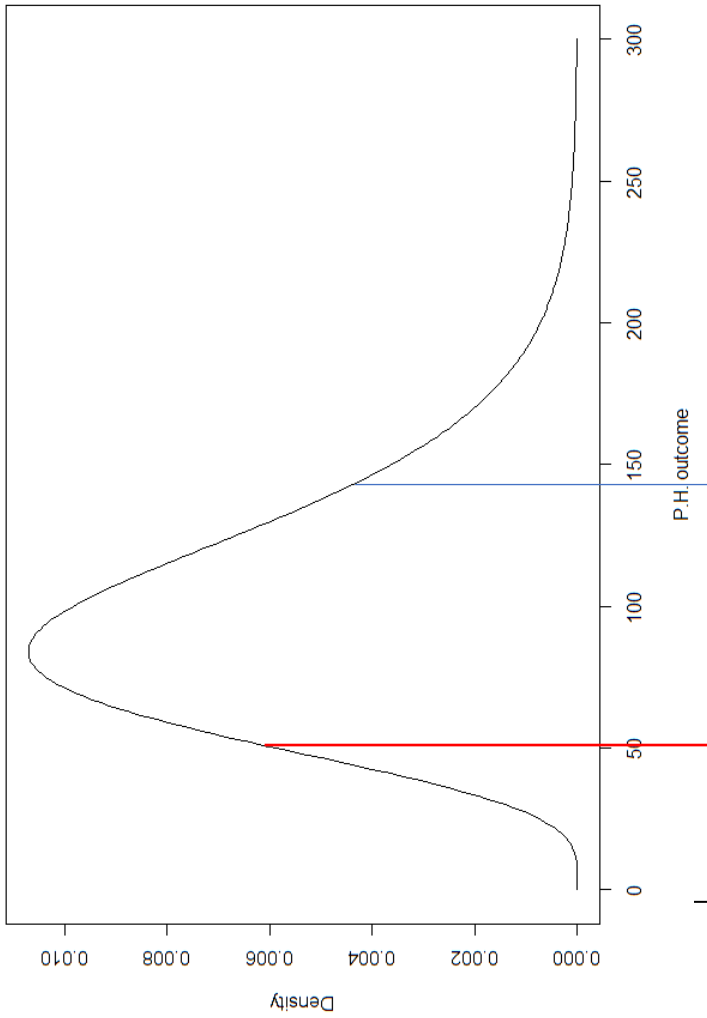
$$p(y_a|\mu,\omega,\beta) = gamma\left(y_a\middle|\omega^{-2}\beta^{(f_a-125)/100},\mu\omega^2\right)$$

β *added to model parameters* $\boldsymbol{\theta}$

Probabilistic model for dropout

Assume the model for heterozygot variant:

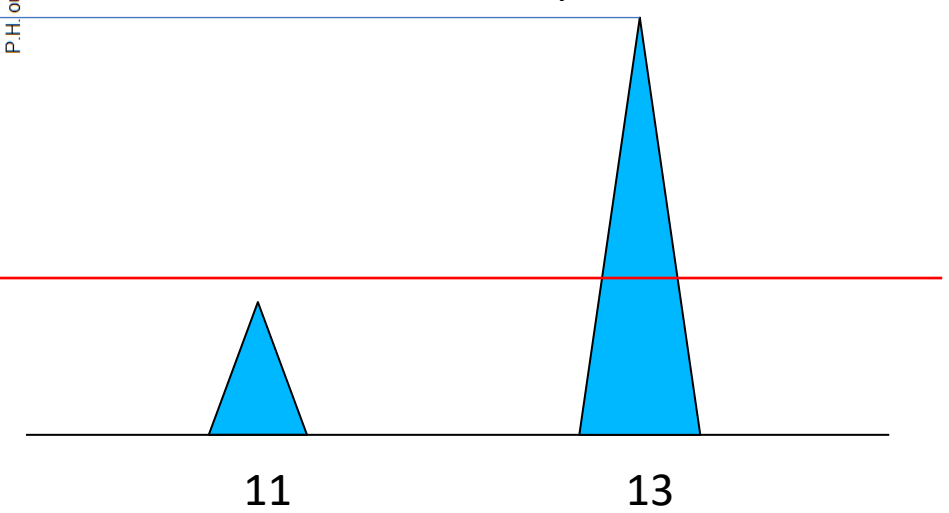
$\mu = 100$
$\omega = 0.4$



$$Y_{11}, Y_{13} \sim gamma(\omega^{-2}, \mu\omega^{2})$$

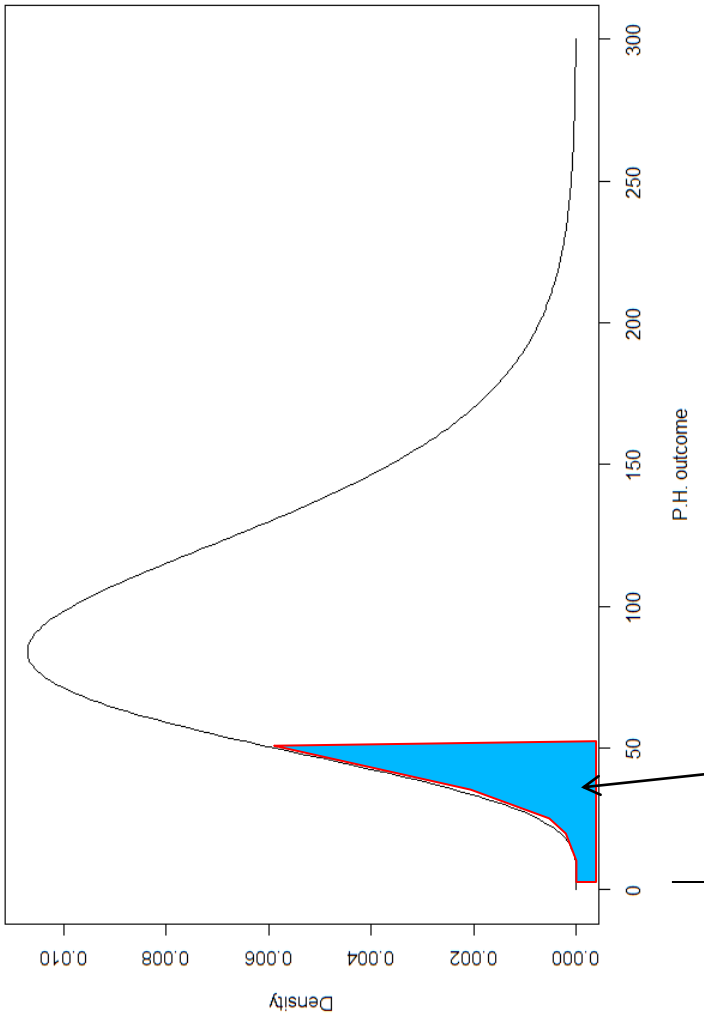
Observed PHs are above 50 RFU (AT)

What if it «accidently» was below AT? A drop-out



Probabilistic model for dropout

$\mu = 100$
$\omega = 0.4$



The Dropout probability is defined as

«The probability that a peak height is below AT for the given model»

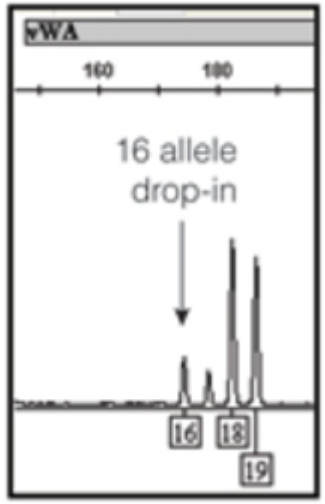
$$P(Y < AT|\mu, \omega) = \int_0^{AT} p(x|\mu, \omega)dx$$

Assuming
$Y \sim gamma(\omega^{-2}, \mu\omega^2)$

$P(Y \leq AT|\mu = 100, \omega = 0.4) = 0.08$

Pr(dropout)=0.08

Probabilistic model for drop-in



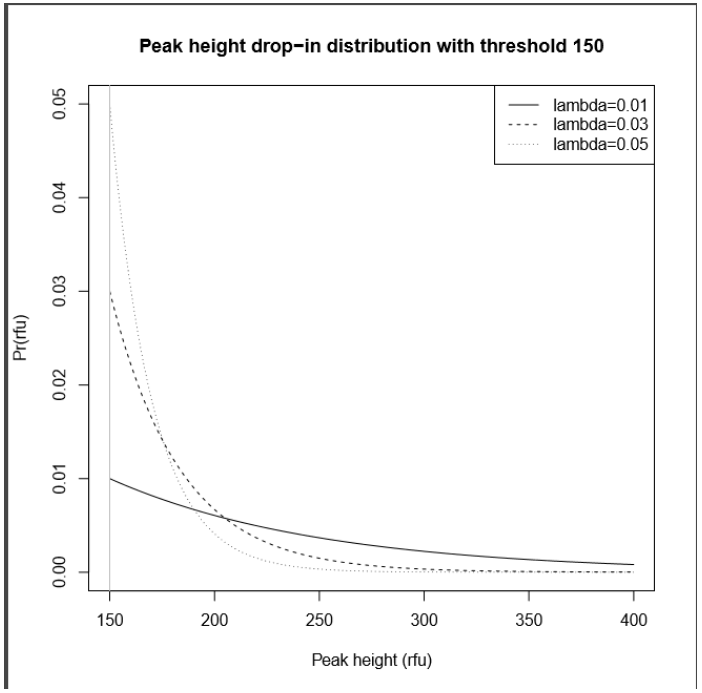
Estimating drop-in frequency:

$$C = \frac{n}{N \times L}$$

n=Number of drop-ins
N=Number of samples
L=Number of loci

Distribution of the drop-in peak heights:

$$p(y|\lambda) = \lambda e^{-\lambda(y-T)}$$



NB: Curve also depends on the analytical threshold (T)

Estimating the drop-in model

Estimate λ from your own lab data:

$$\lambda = \frac{n}{\sum_i (x_i - T)}$$

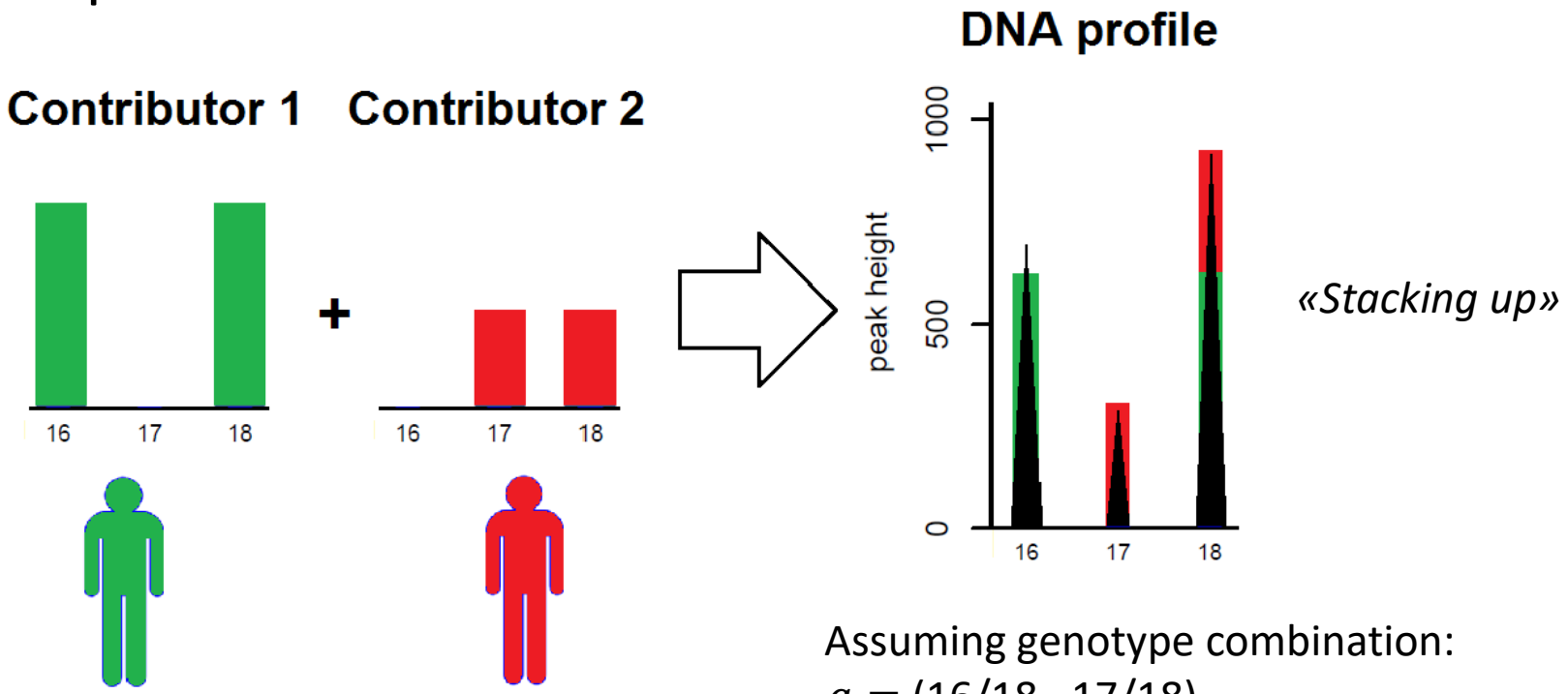
NB: Estimates depends on the analytical threshold (T) to be used in analysis!

Model for drop-in:

$$\mathrm{P}(y \text{ is dropin}) = C * p(y|\lambda)$$

C, λ are assumed as a **known** part of the model (pre-calibrated)
They are not part of the unknown model parameter set $\boldsymbol{\theta}$

Multiple contributors



Assuming genotype combination:
$g = (16/18\ ,\ 17/18)$

Contribution per allele *a*/contr *k*: $n_{a,k}$
decided by genotypes *g*

Allele	Contr 1	Contr 2
16	1	0
17	0	1
18	1	1

$*$

Mix-prop param $\boldsymbol{\pi}$

	Mix-prop
Contr 1	π_1=0.67
Contr 2	π_2=0.33

$=$

Total contribution per allele *a*

Allele	Total
16	π_1
17	π_2
18	π_1+π_2

The model for «total/stacked» PH

Contribution per allele *a*/contr *k*: $n_{a,k}$

Allele	Contr 1	Contr 2
16	1	0
17	0	1
18	1	1

$*$

Mix-prop param $\boldsymbol{\pi}$

	Mix-prop
Contr 1	π_1=0.67
Contr 2	π_2=0.33

$=$

Total contribution per allele *a*

Allele	Total
16	π_1
17	π_2
18	$\pi_1+\pi_2$

Each P.H component (per allele a/contr k) distributed as

$$Y_{a,k} \sim gamma(n_{a,k} * \pi_k * \omega^{-2}, \mu\omega^2)$$

Additivity assumption $Y_a = \sum_k Y_{a,k}$
gives distribution for «total» PH:

$$Y_a = \sum_k Y_{a,k} \sim gamma(\omega^{-2} \sum_k n_{a,k} * \pi_k, \mu\omega^2)$$

The relative contribution is part of the shape argument

Proportional to P.H. expectation

$\boldsymbol{\pi}$ *added to model parameters* $\boldsymbol{\theta}$

Example for allele 18

965 rfu

Assumption for $\boldsymbol{\theta}$*:*

$\mu = 1000$

$\omega = 0.2$

$\pi_1 = 0.67, \pi_2 = 0.33$

Genotype assumption:

$g =$ (16/18 , 17/18)

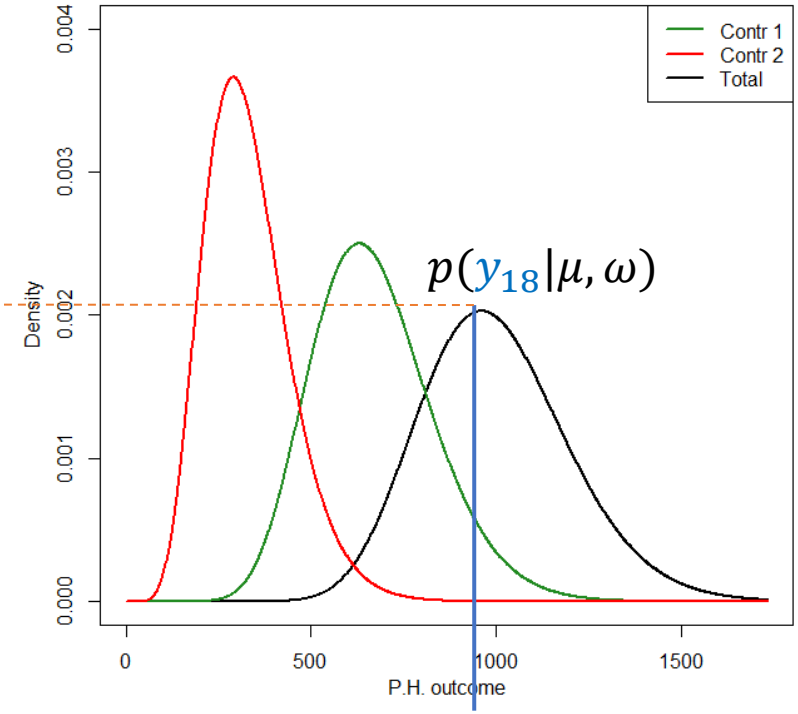
$n_{18,1} = 1, n_{18,2} = 1$

Each components distributed as

$Y_{a,k} \sim gamma(n_{a,k} * \pi_k * \omega^{-2}, \mu\omega^2)$

Distribution of peak height (total)

$Y_a \sim gamma(\omega^{-2} \sum_k \pi_k * n_{a,k}, \mu\omega^2)$



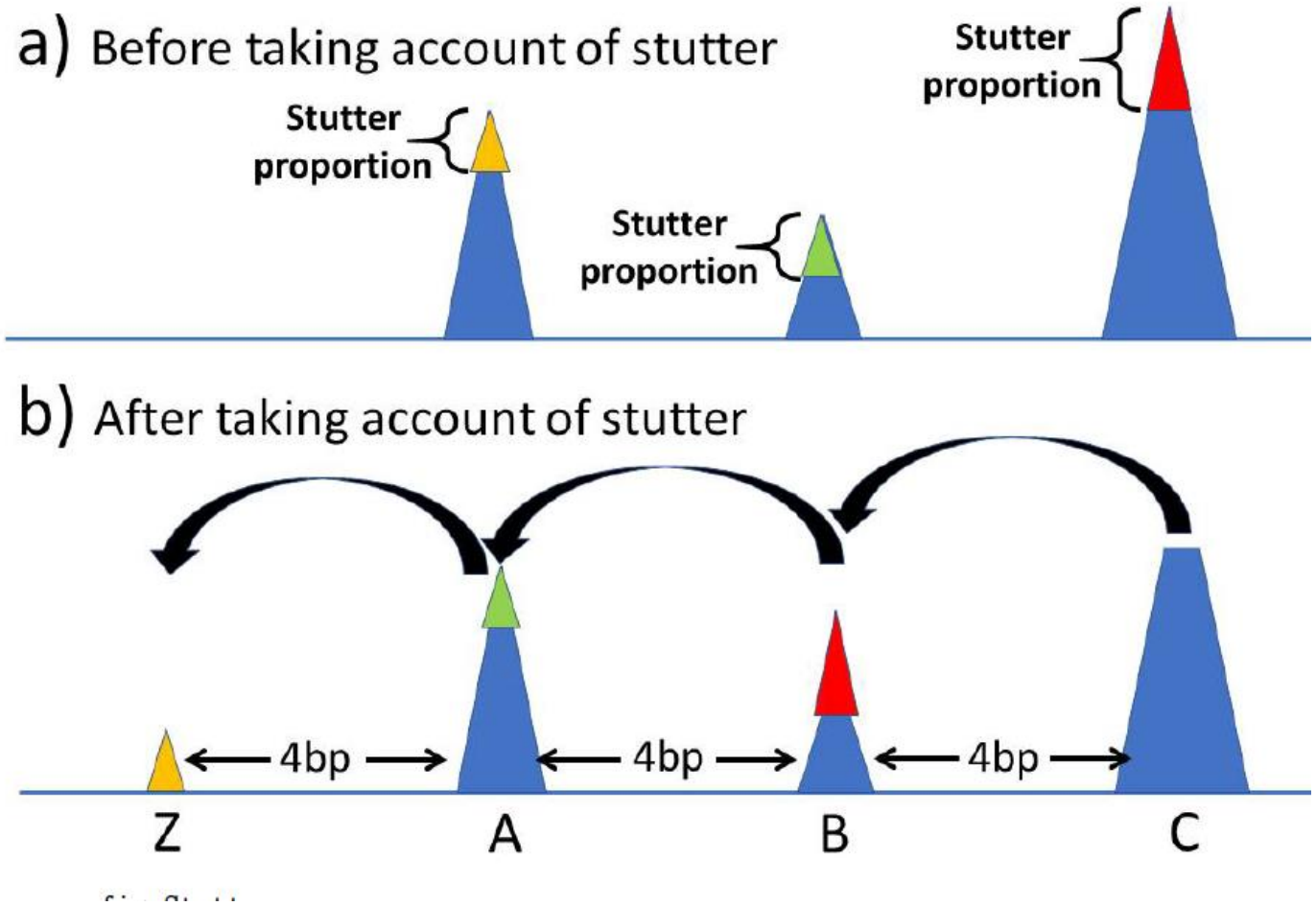
$Y_{18,1} \sim gamma(1 * \pi_1 * \omega^{-2}, \mu\omega^2)$

$Y_{18,2} \sim gamma(1 * \pi_2 * \omega^{-2}, \mu\omega^2)$

$Y_{18} \sim gamma((1 * \pi_1 + 1 * \pi_2) * \omega^{-2}, \mu\omega^2)$

$= gamma(\omega^{-2}, \mu\omega^2)$

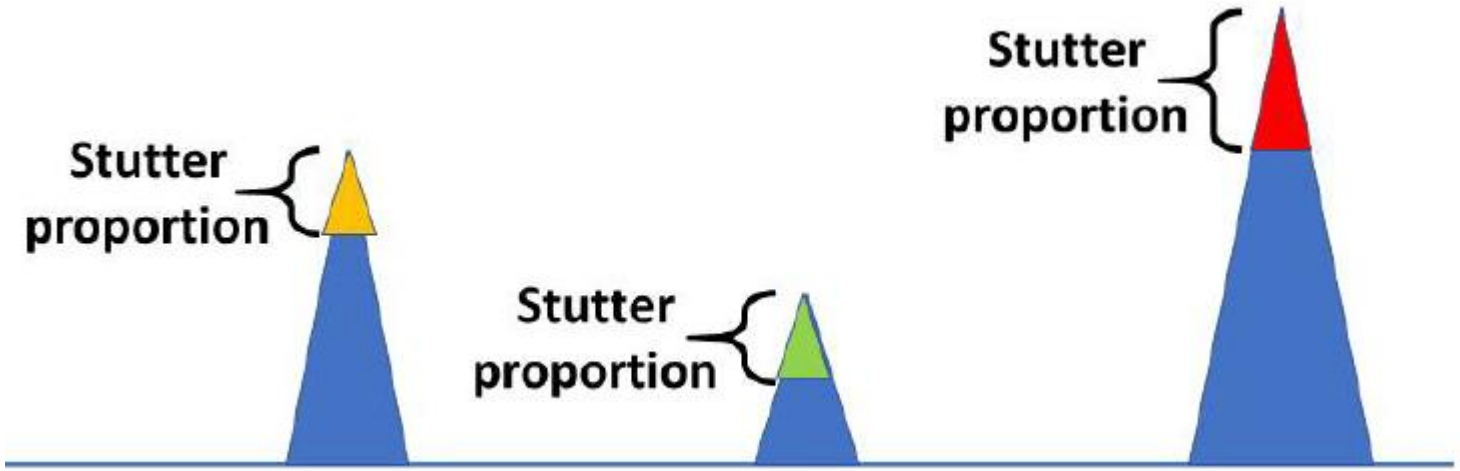
Backward stutters



Assuming that some **proportion** of the peak heights at allele *a* is moved to allele *a-1*

EuroForMix also support **Forward stutters**: prop. of peak heights at allele *a* is moved to allele *a+1*

Distribution for the stutter proportions



The stutter proportion sizes follows the beta-distribution with expectation ξ

$$\sim beta(\xi * \text{something}, (1-\xi) * \text{something})$$

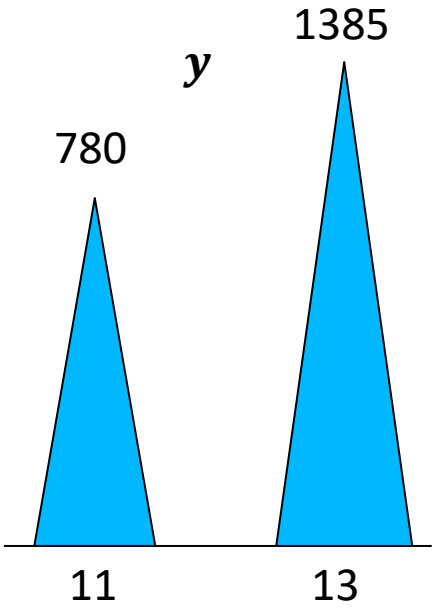
ξ *added to model parameters* $\boldsymbol{\theta}$

Ok so now we have defined the PH-model $\mathrm{p}(\mathrm{y}|\boldsymbol{\theta}, \mathrm{g})$ for given contribution *g* and model parameter $\boldsymbol{\theta}$

But how can we get to the likelihood ratio (LR) from here?

«The likelihood function»

Observed profile E (data): $\boldsymbol{y} = (y_{11}, y_{13}) = (780,1385)$



<u>Hypothesis Hp (prosecution):</u>
Suspect with genotype 11/13 is a contributor to E

The likelihood function is obtained by **inserting** the data into the probability density function (pdf) of the assumed model, $p(\mathbf{Y} = \mathbf{y}|\boldsymbol{\theta})$

Multiplying together the allele weights

$$Lik(\boldsymbol{\theta}|H_p) = p(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{\theta}, H_p) = gamma(780, \omega^{-2}, \mu\omega^2) * gamma(1385, \omega^{-2}, \mu\omega^2)$$

Hence the likelihood function is a function over the unknown parameters

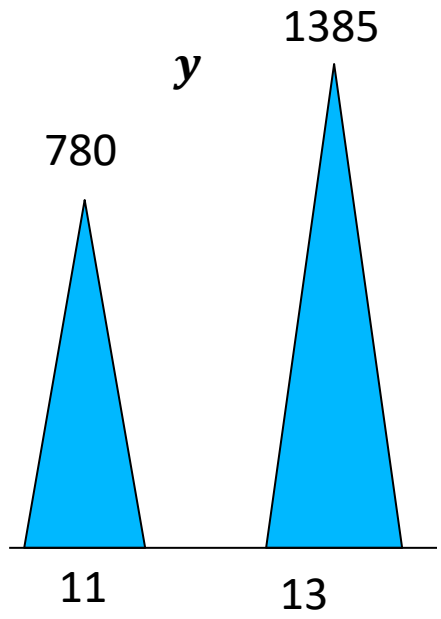
$$\boldsymbol{\theta} = (\mu, \omega, \boldsymbol{\pi}, \beta, \xi)$$

P.H. properties *Mix-prop* *Degrad-slope* *Stutter-prop*

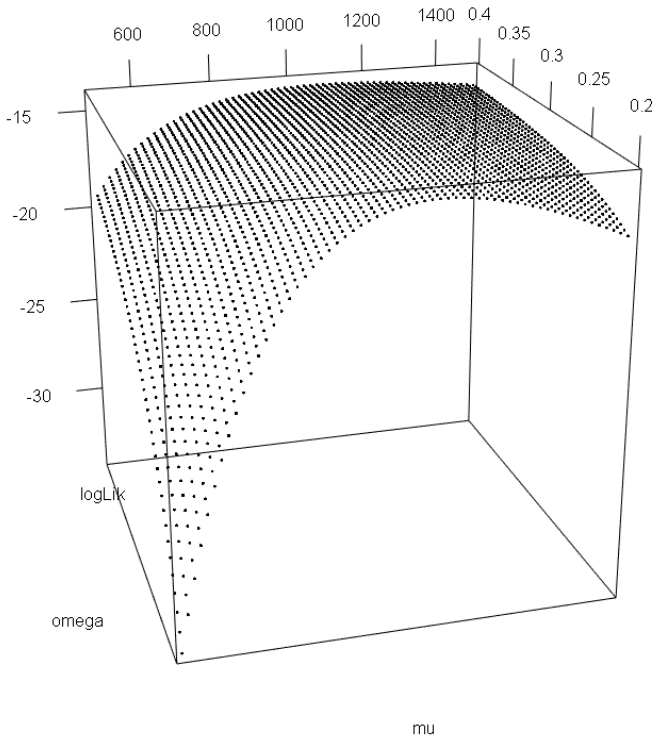
Visualizing the likelihood function

P.H. observations

$$Lik(\boldsymbol{\theta}|H_p) = gamma(780, \omega^{-2}, \mu\omega^2) * gamma(1385, \omega^{-2}, \mu\omega^2)$$



$\log Lik(\theta)$



Maximum Likelihood Estimation aims to maximize $Lik(\boldsymbol{\theta}|H_p)$

MLE gives the parameter values $\boldsymbol{\theta}$ best describing the data E

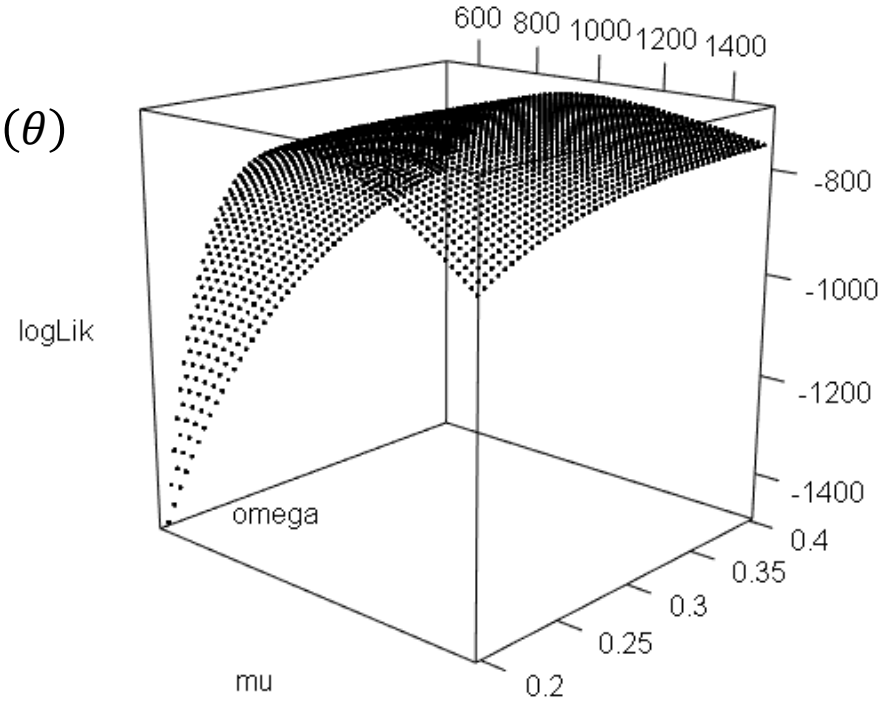
Example with a full profile

Consider n=100 (independent) peak heights in a full profile. The likelihood function is given as

$$L(\theta) = \prod_{i=1}^{n} p(y_i|\theta)$$

The more observed data the more «steep» the likelihood function becomes.

$\log L(\theta)$



This means only some parameter choices are likely to the data

The likelihood function for a given hypothesis

<u>The alternative hypothesis Hd (defence):</u>
An unknown with genotype 11/13 is a contributor to E

$$Lik(\boldsymbol{\theta}|H_d) = gamma(780, \omega^{-2}, \mu\omega^2) * gamma(1385, \omega^{-2}, \mu\omega^2) * (2 * p_{11} * p_{12})$$

$p(Y = y|\theta, g)$

$P(g|H_d)$

In general the likelihood function is a sum over all possible genotypes of the unknowns defined under hypothesis *H*:

$$Lik(\theta|H) = \sum_{\boldsymbol{g}} p(Y = y|\theta, g)\, P(g|H)$$

Hypothesis H decides the outcome of possible genotypes for the contributors

Calculating the LR based on Maximum Likelihood (ML)

The likelihood ratio for the two hypotheses *Hp* and *Hd:* LR = $\frac{P(E|H_p)}{P(E|H_d)}$

Our model for the data E depends on the unknown model parameter set $\boldsymbol{\theta}$ (one set for each hypothesis) giving

$$\text{LR}(\boldsymbol{\theta}_p, \boldsymbol{\theta}_d) = \frac{P(E|H_p,\theta_p)}{P(E|H_d,\theta_d)} = \frac{Lik(\theta_p|H_p)}{Lik(\theta_d|H_d)}$$

With the ML approach we maximize the likelihood function under each of the hypotheses seperately giving

$$LR_{ML} = \frac{\max\limits_{\theta_p} Lik(\theta_p|H_p)}{\max\limits_{\theta_d} Lik(\theta_d|H_d)}$$

The parameter set $\boldsymbol{\theta}$ maximizing the Likelihood function L($\boldsymbol{\theta}$) is called the maximum likelihood estimates $\boldsymbol{\theta}_{MLE}$

The likelihood function for multiple markers

The generalization to *M* markers is simple:

With **the product rule** the likelihood function for the full profile is given as

$$Lik(\boldsymbol{\theta}|\boldsymbol{y}, H) = \prod_{m=1}^{M} Lik_m(\boldsymbol{\theta}|\boldsymbol{y_m}, H)$$

with the evaluation for marker *m* given as before:

$$Lik_m(\boldsymbol{\theta}|\boldsymbol{y_m}, H) = \sum_{\boldsymbol{g}} p(Y_m = y_m|\theta, g)\, P(g|H)$$

The likelihood function for replicates

The generalization to *R independent identical distributed (iid)* replicates $(\boldsymbol{y_1}, \ldots, \boldsymbol{y_R})$ is simple:

Consider one marker with the replicated set of peak heights $\boldsymbol{y} = (\boldsymbol{y_1}, \ldots, \boldsymbol{y_R})$

The likelihood function for this marker is then given as

$$Lik(\boldsymbol{\theta}|\boldsymbol{y}, H) = \sum_{\boldsymbol{g}} P(g|H) \prod_{\boldsymbol{r=1}}^{\boldsymbol{R}} p(\boldsymbol{Y_r} = \boldsymbol{y_r}|\boldsymbol{\theta}, g)$$

The model in EuroForMix assumes that $\boldsymbol{\theta}$ is common across all markers and replicates

Deconvolution based on probabilistic model

Based on the defined model we can derive the posterior probablity of any genotypes *g (Bayes' Theorem)*

$$P(g|\theta, H, \boldsymbol{y}) = c * p(Y = y|\theta, g)\ P(g|H)$$

c is a normalization constant such that $\sum_g P(g|\theta, H)$

Hence $c^{-1} = \sum_{\boldsymbol{g}} p(Y = y|\theta, g)\ P(g|H)$ = $Lik(\boldsymbol{\theta}|\boldsymbol{y}, H)$

In EuroForMix we use $\theta = \theta_{MLE}$

and the probabilites $P(g|\theta_{MLE}, H, \boldsymbol{y})$

are ranked for all genotype combinations *g*

Computational expenses for the genotype sum

The required computation of the likelihood for a given parameter set $\boldsymbol{\theta}$ at a specific marker:

$$Lik(\boldsymbol{\theta}|H) = \sum_{\boldsymbol{g}} p(Y = y|\theta, g)\, P(g|H)$$

The size of the sum depends on the problem:

Evidence *E*: Observing *n* peak heights above AT (*n* alleles)

Hypothesis *H*: K contributors to E, where *U* are unknown

Genotype outcome for 1 contributor: $\frac{(n+2)(n+1)}{2}$

Combined genotype outcome for K contributors where U are unknown:

$$n_g = \left[\frac{(n+2)(n+1)}{2}\right]^U$$

Allele	1	2	.	n	Q
1	1/1				
2	1/2	2/2			
.	.	.	.		
n	1/n	2/n	.	n/n	
Q	1/Q	2/Q	.	n/Q	Q/Q

Allele Q is a «drop-out» allele

Examples

$$n_g = [(n+2)(n+1)/2]^U =$$

#alleles n	#unknowns u	Total combs n_g
3	2	100
4	2	225
3	3	1000
6	3	21 952
3	4	10 000
6	4	614 656
8	4	4 100 625

This is again summed up for each marker!

The computation required to calculate the likelihood function depends on the number of alleles, markers and unknowns

Computation expenses for optimising the likelihood function

- To calculate the LR we must optimize $Lik(\boldsymbol{\theta}|H)$
- Under both *Hp* and *Hd*

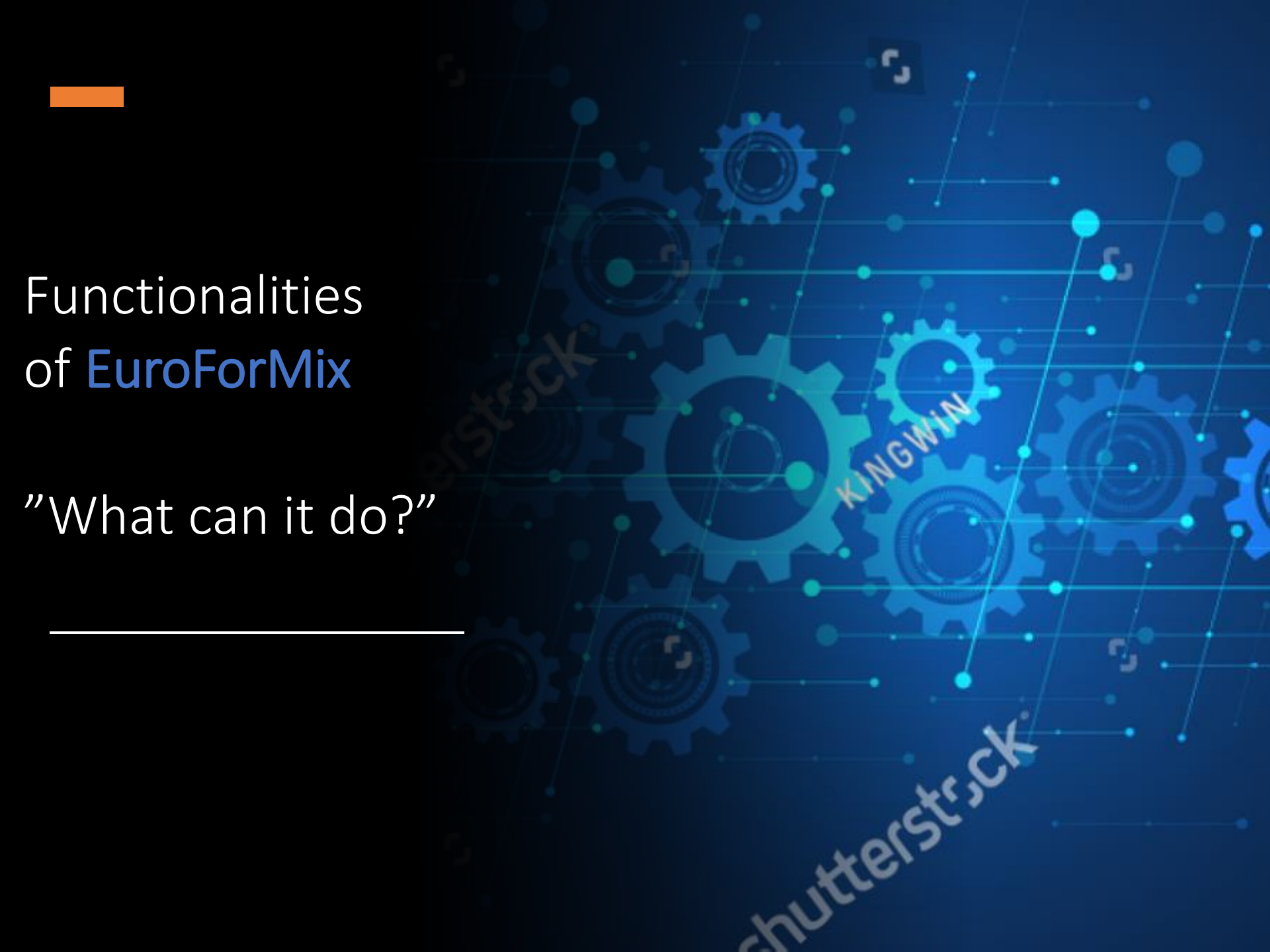
- With full model variant $\boldsymbol{\theta} = (\mu, \omega, \boldsymbol{\pi}, \beta, \xi)$
- Where the mix-proportion is $\boldsymbol{\pi} = (\pi_1, .., \pi_{K-1})$
- The dimension of $\boldsymbol{\theta}$ is $|\boldsymbol{\theta}| = K + 3$

The number of evaluations required in optimizations to obtain MLE increases with the dimension of $\boldsymbol{\theta}$.

EuroForMix uses a Newton-type algorithm to optimize $Lik(\boldsymbol{\theta}|H)$ possibly requiring up to 100 evaluations.

Functionalities of EuroForMix

”What can it do?”



Data visualization and simple allele comparison

Printing to R-console

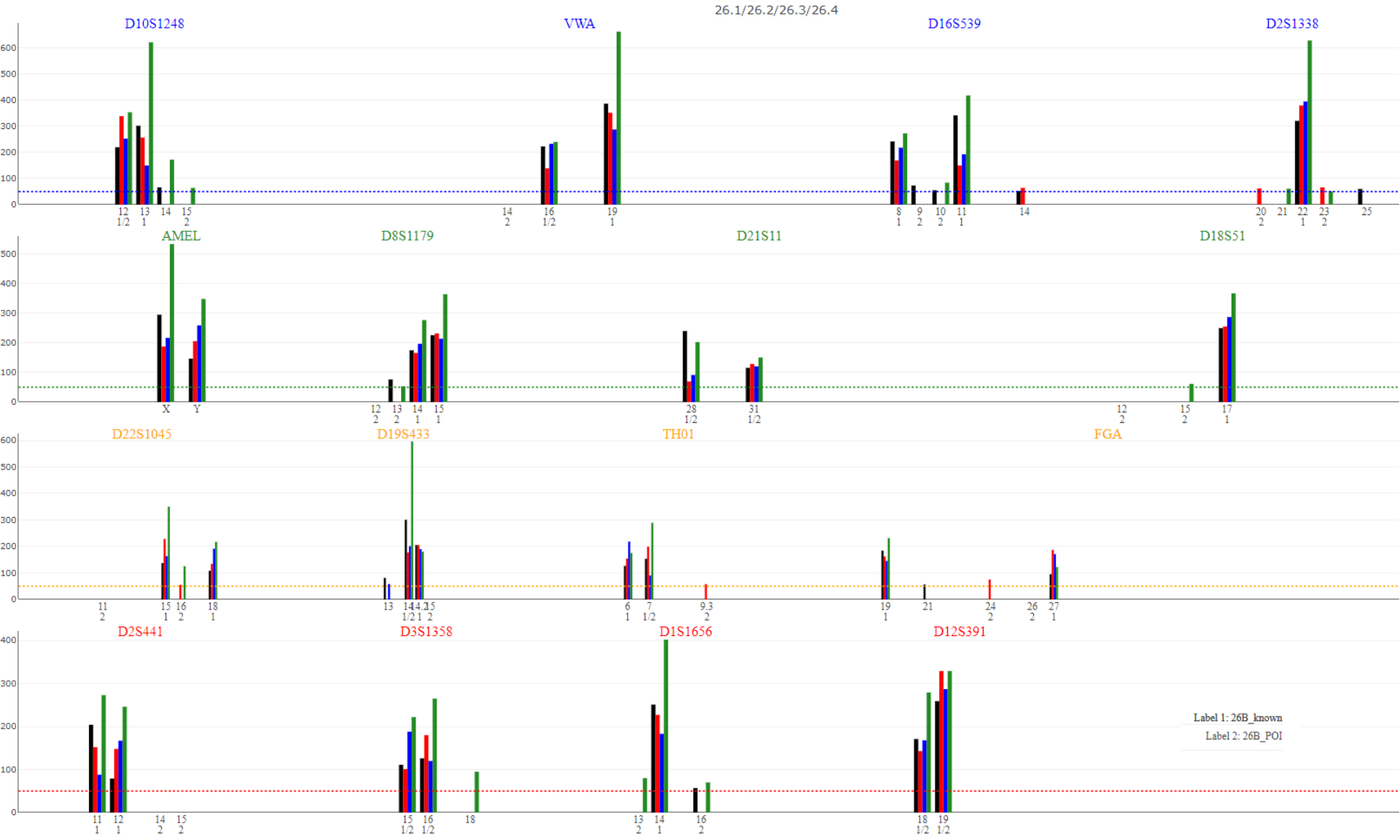
```
[1] "Samplename: evid1"
         Allele                Height
AMEL     "X/Y"                 "2136/1015"
D3S1358  "14/15/16"            "178/2405/1982"
TH01     "6/7/9.3"             "419/282/1871"
D21S11   "27/29"               "1128/1750"
D18S51   "15/17"               "467/524"
D10S1248 "13/14/15"            "1856/155/1045"
D1S1656  "12/15/16/16.3/17.3"  "1140/601/488/155/1877"
D2S1338  "17/19/20/23"         "290/619/259/649"
D16S539  "9/10/11/12"          "217/312/743/619"
D22S1045 "15/16"               "1017/610"
VWA      "14/15/17"            "1250/440/1232"
D8S1179  "10/13/14/15"         "206/352/978/827"
FGA      "21/22"               "664/714"
D2S441   "9/10/11/14"          "200/3362/1168/3693"
D12S391  "18/18.3/19/21/22"    "297/1446/751/171/1370"
D19S433  "13/14/15.2"          "1157/781/922"
SE33     "29.2/30.2/33.2"      "221/473/570"
```

```
          P1          P2
D3S1358   "16/15"     "16/15"
TH01      "9.3/9.3"   "6/7"
D21S11    "29/27"     "29/35"
D18S51    "17/15"     "11/14"
D10S1248  "15/13"     "13/13"
D1S1656   "12/17.3"   "15/16"
D2S1338   "23/19"     "17/20"
D16S539   "11/12"     "9/10"
D22S1045  "15/16"     "15/15"
VWA       "14/17"     "15/17"
D8S1179   "14/15"     "10/13"
FGA       "22/21"     "22/25"
D2S441    "10/14"     "11/11"
D12S391   "18.3/22"   "18/19"
D19S433   "13/15.2"   "14/14"
SE33      "30.2/33.2" "27.2/29.2"
[1] "Number of matching alleles with samplename evid1:"
          P1 P2
AMEL      NA NA
D3S1358    2  2
TH01       2  2
D21S11     2  1
D18S51     2  0
D10S1248   2  2
D1S1656    2  2
D2S1338    2  2
D16S539    2  2
D22S1045   2  2
VWA        2  2
D8S1179    2  2
FGA        2  1
D2S441     2  2
D12S391    2  2
D19S433    2  2
SE33       2  1
MAC       32 27
nLocs     16 16
```

Visualization of peak heights

Possible to highlight allele info when hoving with mouse (fragment length, rfu)

Visualization of replicates

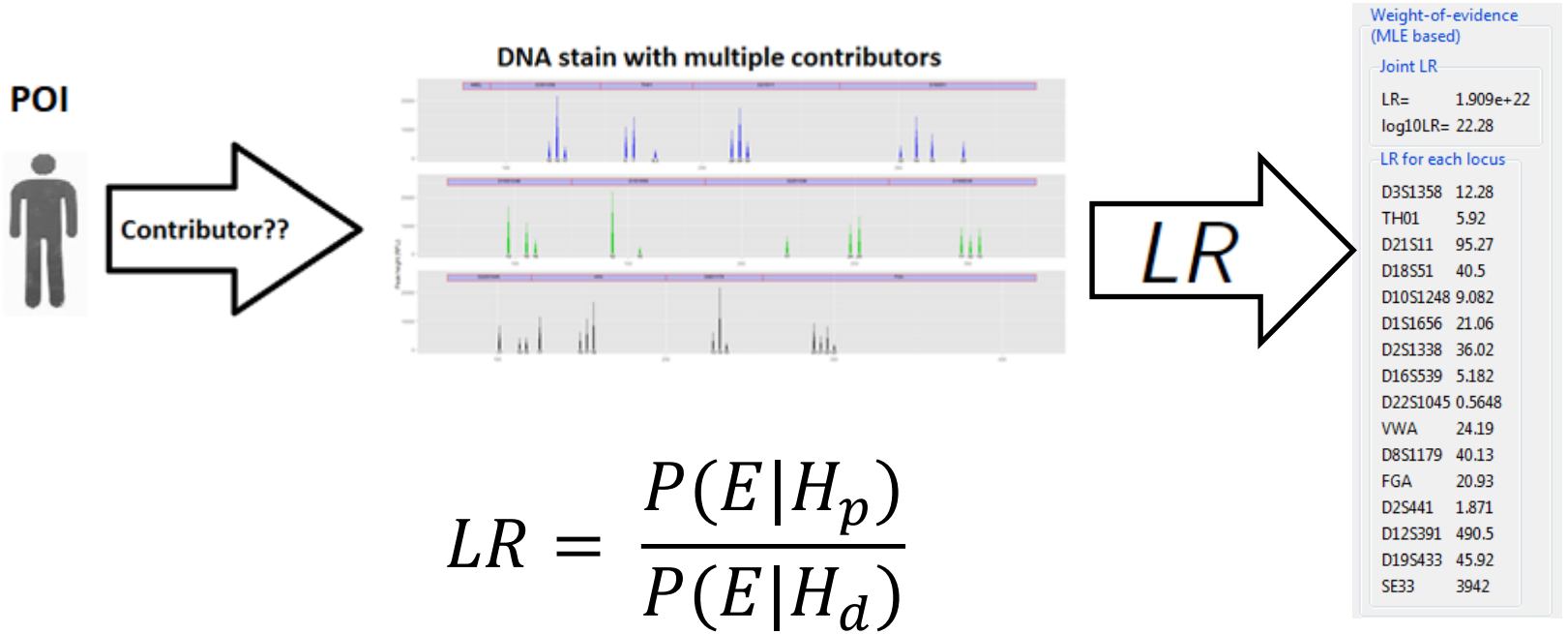


Possible to highlight allele info when hoving with mouse (fragment length, rfu, sample)

Weight-of-evidence calculations using LR

Establishing two hypotheses: Hp (prosecution) and Hd (defence)

- H_p: The person of interest (POI) is a contributor to the evidence *E*
- H_d: An unknown individual (unrelated*) is a contributor to *E*



$$LR = \frac{P(E|H_p)}{P(E|H_d)}$$

*EuroForMix supports that the unknown individual is related to a typed reference (for instance the POI)

Weight-of-evidence **modules**

$$LR = \frac{P(E|H_p, \theta_p)}{P(E|H_d, \theta_d)}$$

Maximum Likelihood approach

$$LR_{ML} = \frac{\max_{\theta_p} P(E|H_p, \theta_p)}{\max_{\theta_d} P(E|H_d, \theta_d)}$$

Model penalty

Model selection with AIC:

$$\widehat{M} = \max_{m} \left\{ \max_{\theta_m} \log P(E|H_d, \theta_m, m) \; - |\theta_m| \right\}$$

Calculations

- Quantitative LR (Maximum Likelihood based)
- Optimal quantitative LR (automatic model search)
- Quantitative LR (Bayesian based)
- Qualitative LR (semi-continuous)

Bayes Factor: Bayesian approach

$$LR_{BF} = \frac{\int P(E|H_p, \theta_p) p(\theta_p) d\theta_p}{\int P(E|H_d, \theta_d) p(\theta_d) d\theta_d}$$

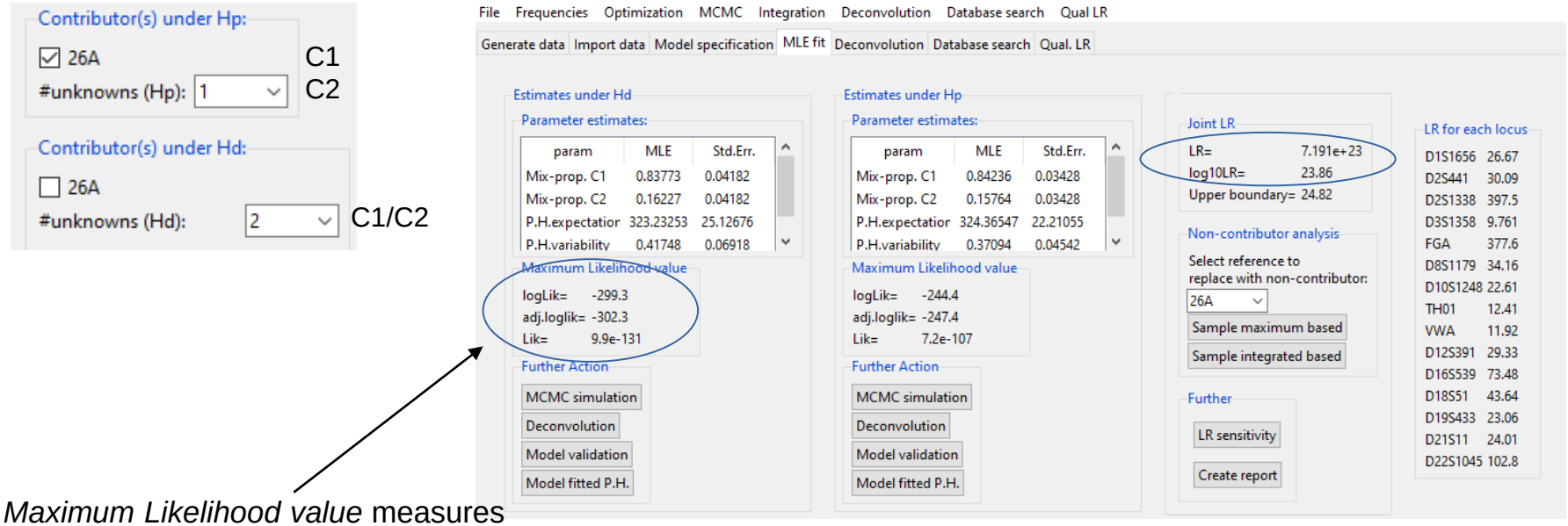
Quantitative

Qualitative

LRmix/LRmix studio approach

ML approach: LR_{ML}

Example of Likelihood Ratio result (ML approach)



The *Maximum Likelihood value* measures model performance:
"How well does the model fit data?"
Want the model with largest adj.loglik (AIC criterion)

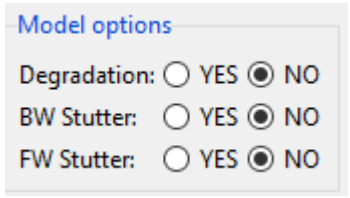
<u>Further action:</u>

- MCMC simulation (to infer distr. of param)
- LRsensitivity (conservative LR)

- Model validation (Hp/Hd)
- Create report (results)

- Show Model fitted P.H. (Hp/Hd)
- Deconvolution (Hp/Hd)

Automatic Model selection approach



Model options (TRUE/FALSE)
DEG = degradation
BW=Backward stutter
FW=Forward stutter

Score for model selection

NOC = number of contributors

adjLogLik = logLik - #param

logLik= natural logarithm of maximum likelihood

Log10LR for Person of Interest (POI)

MxPOI = mixture proportion of PO

Model comparison results

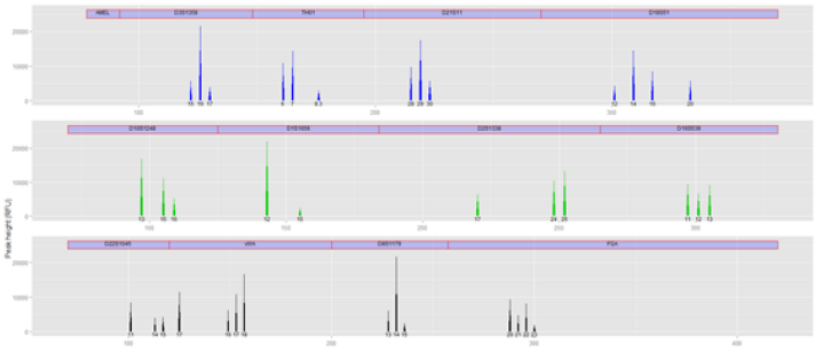
NOC	DEG	BWstutt	FWstutt	logLik	adjLogLik	log10LR	MxPOI	SignifHp	SignifHd
1	FALSE	FALSE	FALSE	-314.69	-316.69	20.64	1.00	0	0
1	TRUE	FALSE	FALSE	-313.40	-316.40	20.57	1.00	0	0
1	FALSE	TRUE	FALSE	-302.47	-305.47	19.24	1.00	0	0
1	TRUE	TRUE	FALSE	-301.72	-305.72	19.36	1.00	0	0
2	FALSE	FALSE	FALSE	-299.34	-302.34	23.86	0.84	0	0
2	TRUE	FALSE	FALSE	-297.12	-301.12	23.86	0.83	0	0
2	FALSE	TRUE	FALSE	-299.00	-303.00	24.09	0.86	0	0
2	TRUE	TRUE	FALSE	-296.91	-301.91	24.08	0.85	0	0
3	FALSE	FALSE	FALSE	-299.52	-303.52	23.85	0.84	0	0
3	TRUE	FALSE	FALSE	-297.24	-302.24	23.85	0.83	0	0
3	FALSE	TRUE	FALSE	-299.08	-304.08	24.09	0.86	0	0
3	TRUE	TRUE	FALSE	-296.98	-302.98	24.07	0.85	0	0

SignifHp/Hd: Checking whether the PH model fits the observed PHs. Significance level 0.01 used
Counting number of points falling outside envelope

Deconvolution

Calculating probabilities of genotypes for unknown contributors

DNA stain with multiple contributors



Unknown contributors

Marker	TopGenotype_C1	probability_C1	TopGenotype_C2	probability_C2
D3S1358	15/16	0.4759	15/16	0.9038
TH01	6/7	0.5017	9.3/9.3	0.5428
D21S11	29/99	0.3248	27/29	0.8464
D18S51	15/99	0.1886	15/17	0.7479
D10S1248	13/14	0.3008	13/15	0.9012
D1S1656	15/16	0.5664	12/17.3	0.6817
D2S1338	17/20	0.4892	19/23	0.4841
D16S539	9/11	0.2462	11/12	0.5715
D22S1045	15/16	0.3901	15/16	0.869
VWA	15/17	0.4727	14/17	0.7804
D8S1179	10/13	0.5827	14/15	0.7263
FGA	22/99	0.2268	21/22	0.7239
D2S441	11/14	0.448	10/14	0.8838
D12S391	18/19	0.4683	18.3/22	0.6742
D19S433	13/14	0.2502	13/15.2	0.4036
SE33	29.2/99	0.2773	30.2/33.2	0.5799

- 3 Modules:

All Joint

Locus	Rank	C1	C2	
D3S1358	1	15/16	15/16	0.9975
TH01	1	6/7	9.3/9.3	0.9926
D21S11	1	29/99	27/29	0.9999
D18S51	1	14/99	15/17	1
D10S1248	1	13/13	13/15	0.9995
D1S1656	1	15/16	12/17.3	1
D2S1338	1	17/20	19/23	0.9991
D16S539	1	9/10	11/12	0.9988
D22S1045	1	15/15	15/16	0.9991
VWA	1	15/17	14/17	0.9993

All Marginal (G)

Contr.	Locus	Genotype	
C2	D3S1358	15/16	0.9975
C2	TH01	9.3/9.3	0.9926
C2	D21S11	27/29	0.9999
C2	D18S51	15/17	1
C2	D10S1248	13/15	0.9995
C2	D1S1656	12/17.3	1
C2	D2S1338	19/23	0.9991
C2	D16S539	11/12	0.9988
C2	D22S1045	15/16	0.9991
C2	VWA	14/17	0.9993

All Marginal (A)

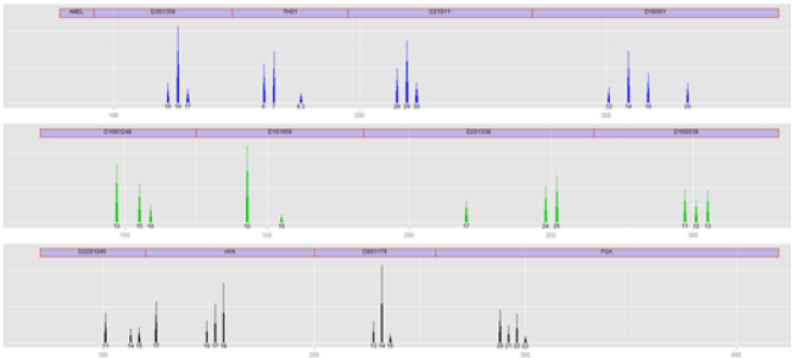
Contr.	Locus	Allele	
C2	D3S1358	15	1.0000
C2	D3S1358	16	0.9975
C2	TH01	9.3	1
C2	D21S11	27	1.0000
C2	D21S11	29	0.9999
C2	D18S51	17	1
C2	D18S51	15	1
C2	D10S1248	15	1.0000
C2	D10S1248	13	0.9995
C2	D1S1656	12	1
C2	D1S1656	17.3	1
C2	D2S1338	23	1.0000

Database-searching

- Quantitative LR (EuroForMix)
  - Peak heights utilized
- Qualitative LR (LRmix)
  - Peak heights not utilized
- Number of matching alleles (MAC)

Reference	D3S1358	TH01	D21S11	D18S51	D10S1248	D1S1656	D2S1338	D16S539	D22S1045	VWA
00-JP0001-	14/15	7/9.3	29/30	13/17	12/13	11/14	17/19	10/11	15/16	17/18
00-JP0002-	15/18	6/9	28/31.2	13/18	13/13	15/18.3	25/25	11/13	15/16	14/17
00-JP0003-	16/18	9.3/9.3	30/30	13/18	14/16	13/16	17/18	8/12	15/16	16/18
00-JP0004-	18/18	7/9.3	29/32.2	12/22	15/16	12/15	19/23	11/11	11/16	14/16
00-JP0005-	15/17	7/8	28/33.2	12/17	13/15	16/17.3	19/25	13/13	11/17	17/18
00-JP0006-	14/18	7/9.3	28/32.2	11/15	15/16	14/15.3	20/24	9/13	16/16	15/16
00-JP0007-	15/19	9.3/9.3	30/32	14/19	13/15	17.3/17.3	17/23	9/10	14/16	16/16
00-JP0008-	14/16	9/9.3	30/30.2	14/18	14/16	15.3/16.3	17/23	9/11	11/16	16/18
00-JP0009-	14/16	7/7	30/30	12/16	14/14	11/14	21/22	12/12	15/15	14/16
00-JP00010	15/16	6/6	30/32	16/17	13/16	16/18.3	21/23	9/14	14/15	18/18

DNA stain with multiple contributors



Sort table: contLR qualLR MAC nLocs

Referencename	contLR	qualLR	MAC	nLocs
00-JP00056-14_20142342311_NO-32456	242.053194610721	5.20597225119075e-06	22	16
00-JP00057-14_20142342311_NO-32457	0.207565918372396	0.00123398788172434	25	16
00-JP00075-14_20142342311_NO-32475	0.00173890105629254	3.99473400901198e-12	19	16
00-JP00044-14_20142342311_NO-32444	0.0011433526258535	2.89226989888451e-11	22	16
00-JP00067-14_20142342311_NO-32467	0.000443967859882698	1.42934120806872e-16	18	16
00-JP00041-14_20142342311_NO-32441	0.000440061290592108	7.26464344174459e-12	22	16

Summary of model features

PH model is a **gamma** distribution

- Allele drop-out probabilities based on this

Multiple contributors

- Can condition on any number of reference profiles
- Can specify any number of unknowns (practical limit is 4)

Model extensions:

- Allele drop-in
- Degradation
- Backward/Forward stutters
- Subpopulation structure («theta/Fst-correction»)
- An unknown under Hd may be related (relatedness module)

Replicated samples

- No need for making consensus samples
- Model assumes same contributors and same P.H. properties for each rep.

Other features

- ❑Includes an automatic model selector module using AIC
- ❑MCMC method to infer model parameters (Bayesian)
- ❑Visualization of Exp. Peak heights for each contributors
- ❑Supports analysis of MPS data (both SNP and STR)
  - Special support using “LUS format” for STR
- ❑Supports marker specific settings of
  - Analytical thresholds
  - Drop-in model
  - Theta/Fst-correction

Advantages of EuroForMix

- ✓ Flexible to input data
  - Any kit system (any number of markers)
  - Any number of replicates (restricted to common parameters)

- ✓ Flexible model choice
  - Models for backward or forward stutter can be included or not.
  - A model for a global degradation trend can be included or not.

- ✓ Model properties are selected as the ones **best explaining the observations**.
  - Peak height distribution fitted through observed peak heights.
  - Drop-out properties taken care of through the fitted peak height distribution.

- ✓ Uncertainty of the model parameter estimates can be taken into account
  - “Conservative LR” or ”Full Bayesian”

- ✓ Includes deconvolution and database search modules.

- ✓ The speed of Version 3 scales with number of threads and CPU speed

Approximate timeusage* to obtain LR

for EuroForMix (v3)

#unknowns	Time scale
1	~1 second to a few
2	Several seconds
3	~1 min to a few
4	~10 mins to hours

Times depends heavily on the amount of data and the following model options:

- Assuming the backward stutter model ~*doubles* the required time
- Assuming the forward stutter model ~*increases* the required time further

*The times of version 3.0 scales with CPU power (speed and number of threads)

Limitations of EuroForMix

- Effective only up to four contributors.
  - However when stutter is turned off and there are few alleles per markers, the speed is drastically increased

- Model limitations:
  - The possibility that markers have different amplification efficiency are not taken into account
  - Backward/Forward-stutters follows the same distribution for all alleles

- Deconvolution module does not take into account the uncertainty of the model parameter estimates.

- The “Full Bayesian” approach is not very robust

Publications

Forensic Science International: Genetics

Volume 21, March 2016, Pages 35–44

Research paper

EuroForMix: An open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts

Øyvind Bleka[a, b], Geir Storvik[b, 1], Peter Gill[a, c, 1]

Forensic Science International: Genetics Supplement Series

Volume 5, December 2015, Pages e405–e406

Interpretation of a complex STR DNA profile using *EuroForMix*

Øyvind Bleka[a, b], Peter Gill[a, b]

Forensic Science International: Genetics

Volume 25, November 2016, Pages 85–96

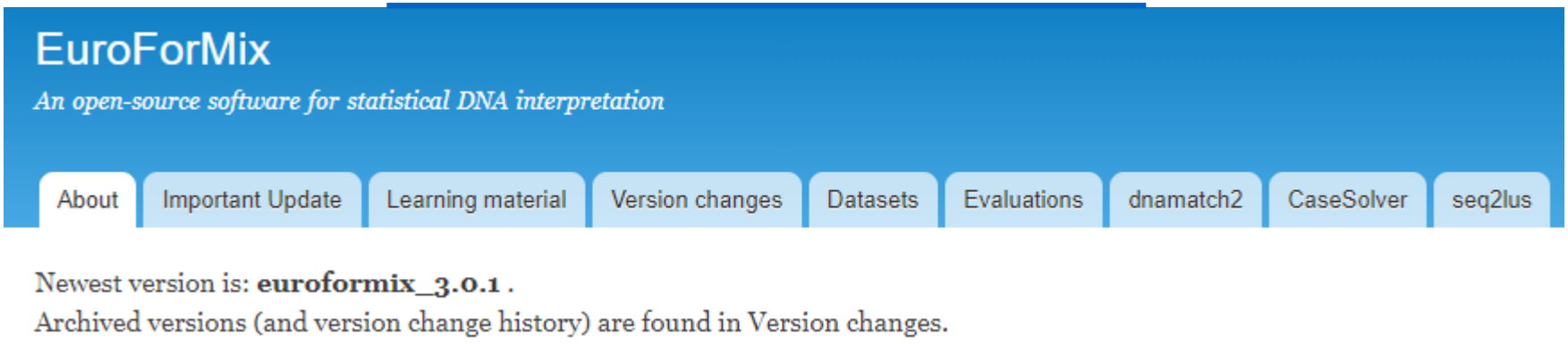
Research paper

A comparative study of qualitative and quantitative models used to interpret complex STR DNA profiles

Øyvind Bleka[a, b], Corina C.G. Benschop[c], Geir Storvik[b], Peter Gill[a, d]

Webpage

www.euroformix.com



Contains following:

- All versions (New and Archived)
- Version information
- Documentation: Manual and tutorial
- Learning materials: Tutorial and Videos
- Dataset for tutorials and publications
- Information about other software which integrates with EuroForMix:
  - *dnamatch2* – A contamination search engine
  - *CaseSolver* – A Expert system for profile comparison in case work
  - *Seq2lus* – Simple tool for converting MPS STRs to LUS format (ForenSeq)

Need help?

Don’t hesitate to email help@euroformix.com