The AFLP package: generating objective and repeatable genetic AFLP data

useR! 2011

ir. Thierry Onkelinx[1], ing. Pieter Verschelde[1], Sabrina Neyrinck[2], ir. Gerrit Genouw[2], Dr. ir. Paul Quataert[1]

Research Institute of Nature and Forest, Belgium
[1]Biometry and Quality Assurance
[2]Laboratories

16/08/11

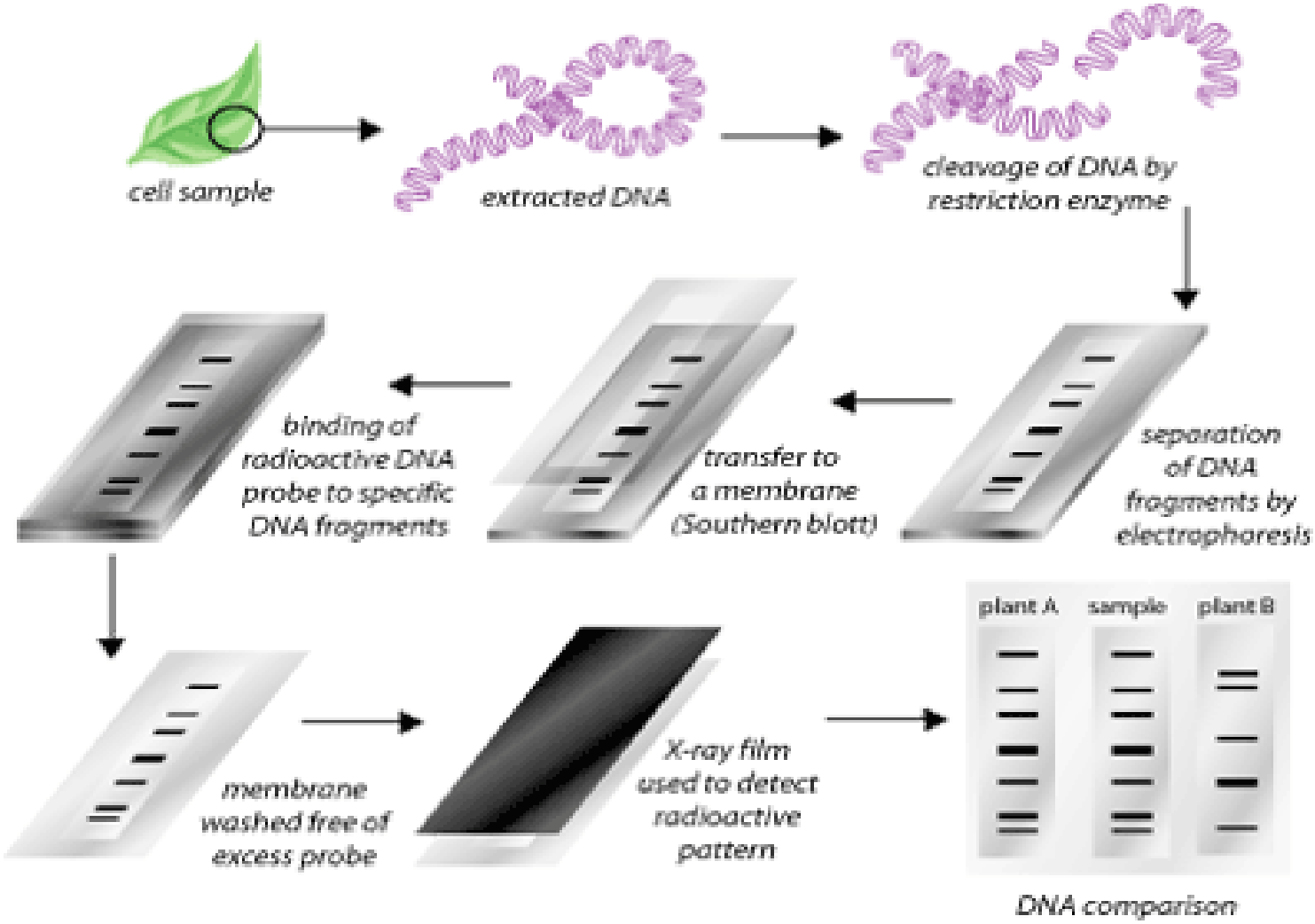
inbo
Instituut voor Natuur- en Bosonderzoek
Research Institute for Nature and Forest

www.inbo.be

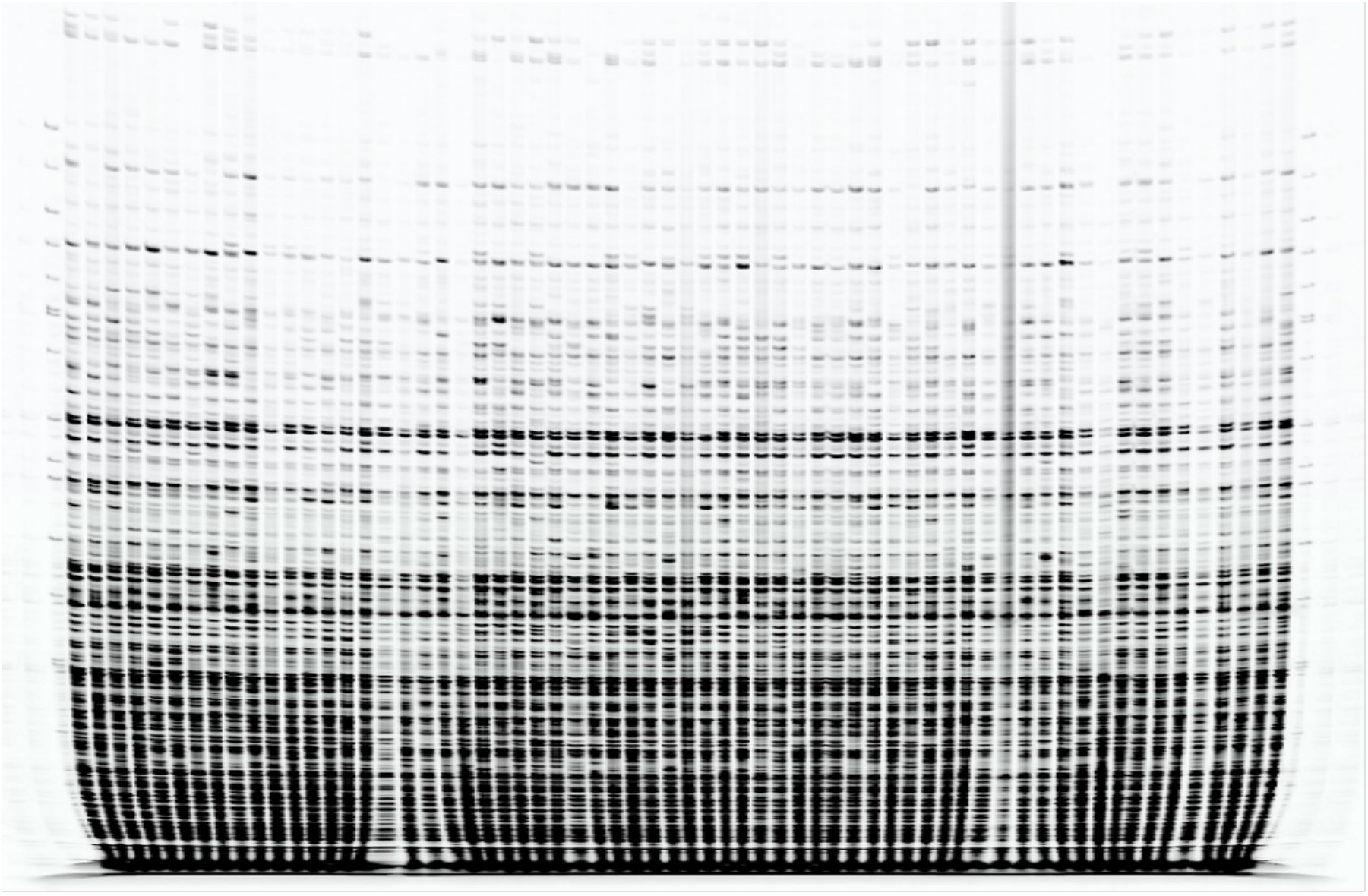
Outline

AFLP?

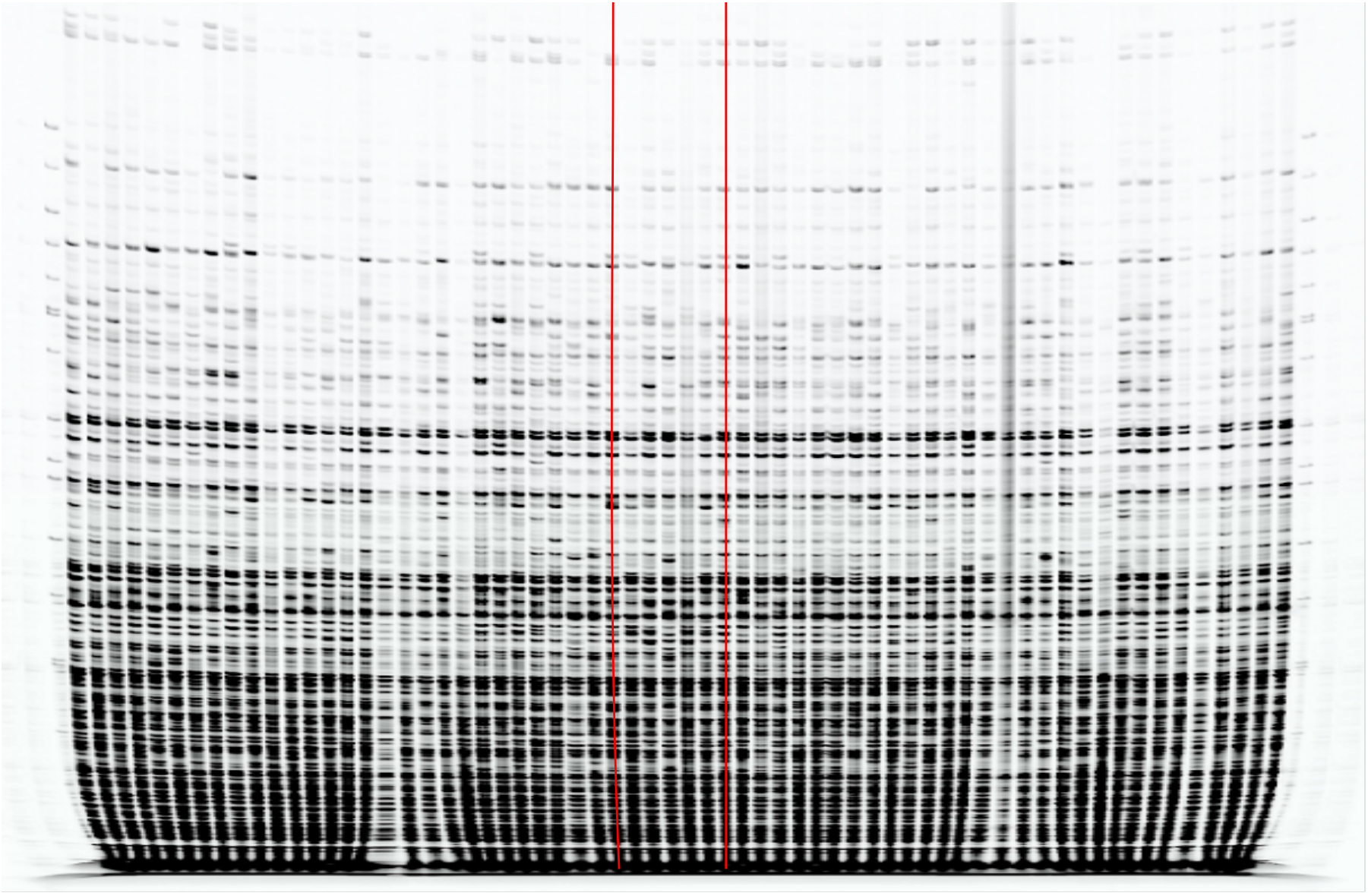
Amplification **F**ragment **L**ength **P**olymorphism



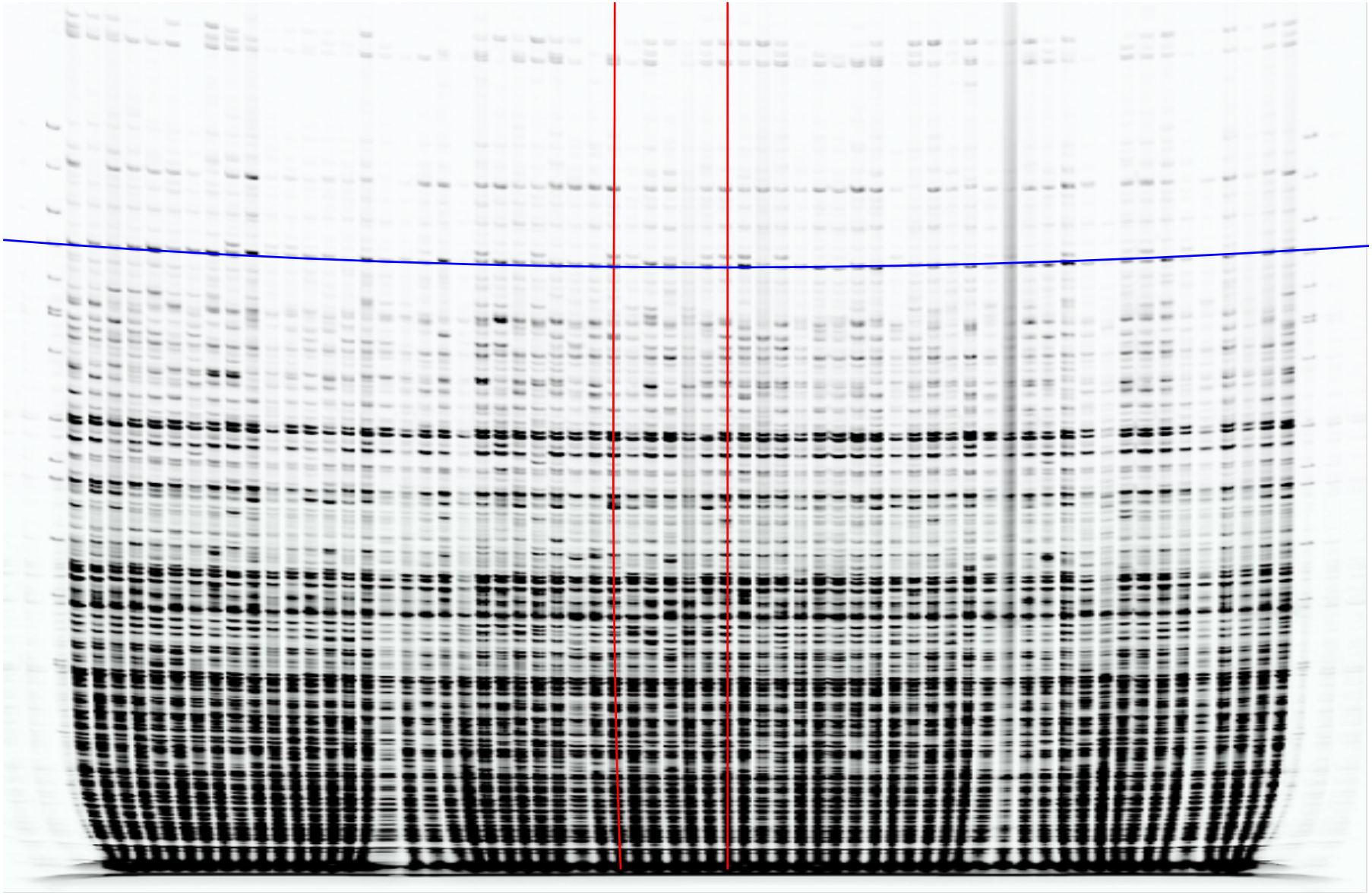
One plate of an AFLP slab gel



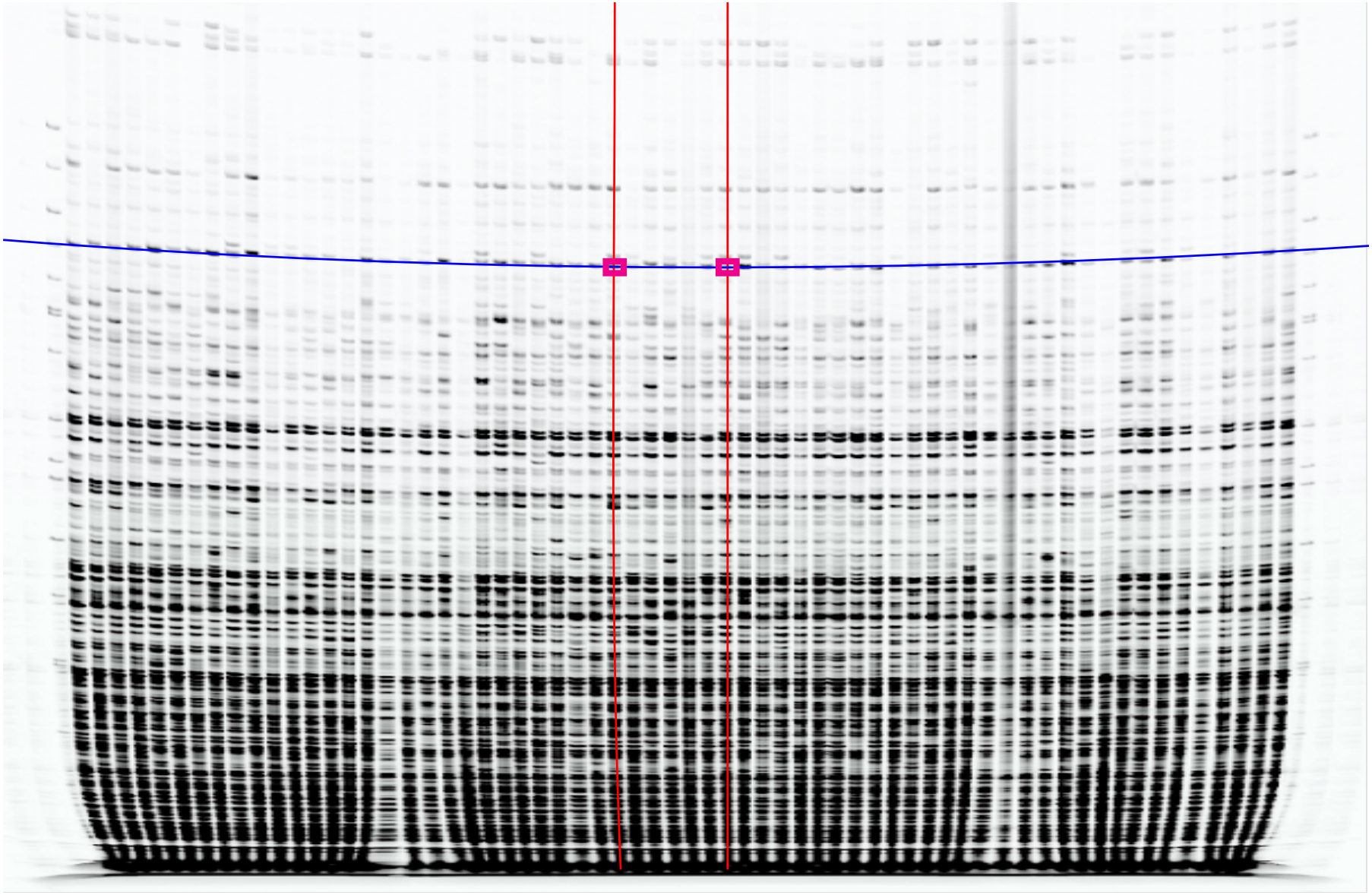
Red line: replicates



Blue line: marker = fragments with same length



Magenta rectangles: bands with fluorescence



Problems

- Binary data required
  - Fluorescence $\Rightarrow$ presence/absence (scoring)
- Manual scoring
  - Tedious
  - Subjective
  - If scientist uses knowledge about population $\Rightarrow$ possible bias
- Lots of steps in the lab
  - Some affect entire subset (e.g. replicate, marker, plate,. . .)
  - Hard for humans to take into account
  - Replication is required to asses lab effects
- Single threshold is not a good option
  - Adjustments required for lab effects
  - Gradient in signal strength along fragment lengths

Goals of the AFLP package

1. Randomise specimens and add replicates to asses lab effect and avoid bias

Goals of the AFLP package

1. Randomise specimens and add replicates to asses lab effect and avoid bias
2. Convert fluorescence into binary data
   - Correct for lab effects
   - Find optimal thresholds

Goals of the AFLP package

1. Randomise specimens and add replicates to asses lab effect and avoid bias
2. Convert fluorescence into binary data
   - Correct for lab effects
   - Find optimal thresholds
3. Evaluate repeatability
   - All replicates per specimen should be identical
   - Overall repeatability = lab quality
   - Repeatability per marker = useful for further analysis

Goals of the AFLP package

1. Randomise specimens and add replicates to asses lab effect and avoid bias
2. Convert fluorescence into binary data
   - Correct for lab effects
   - Find optimal thresholds
3. Evaluate repeatability
   - All replicates per specimen should be identical
   - Overall repeatability = lab quality
   - Repeatability per marker = useful for further analysis
4. Do it fast, objective and reproducible

Normalisation

- Import fluoresence from other software

Normalisation

- Import fluoresence from other software
- Model average fluorescence with mixed model: `lme4` (Bates et al., 2011)
- Model takes lab effects into account

```
log(Fluorescence) ~ Marker + (1|Marker) +
(1|Replicate) + (1|Plate)
```

Normalisation

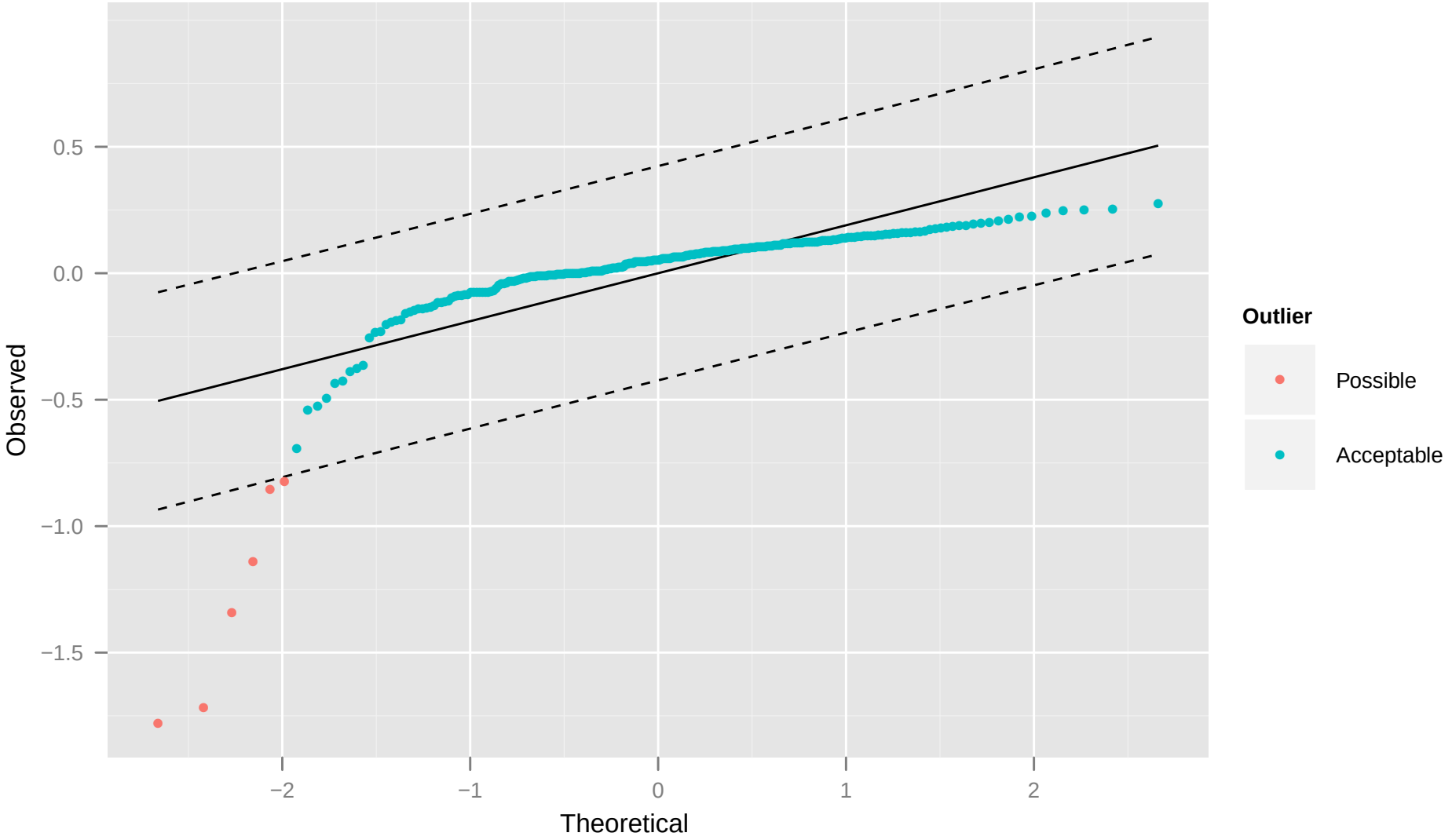
- Import fluoresence from other software
- Model average fluorescence with mixed model: `lme4` (Bates et al., 2011)
- Model takes lab effects into account
  `log(Fluorescence) ~ Marker + (1|Marker) + (1|Replicate) + (1|Plate)`
- Fragment present $\Rightarrow$ stronger than average signal
- Fragment absent $\Rightarrow$ weaker than average signal
- Residuals = normalised fluorescence

Normalisation

- Import fluoresence from other software
- Model average fluorescence with mixed model: `lme4` (Bates et al., 2011)
- Model takes lab effects into account
  `log(Fluorescence) ~ Marker + (1|Marker) + (1|Replicate) + (1|Plate)`
- Fragment present $\Rightarrow$ stronger than average signal
- Fragment absent $\Rightarrow$ weaker than average signal
- Residuals = normalised fluorescence

- QQ-plots random effects and residuals $\Rightarrow$ outliers
- Outliers marked, not deleted

QQ-plot random effect replicates



Classification

- Two types of markers
  - Monomorphic marker: fragment present in all/none specimens
  - Polymorphic marker: fragment present in some specimens

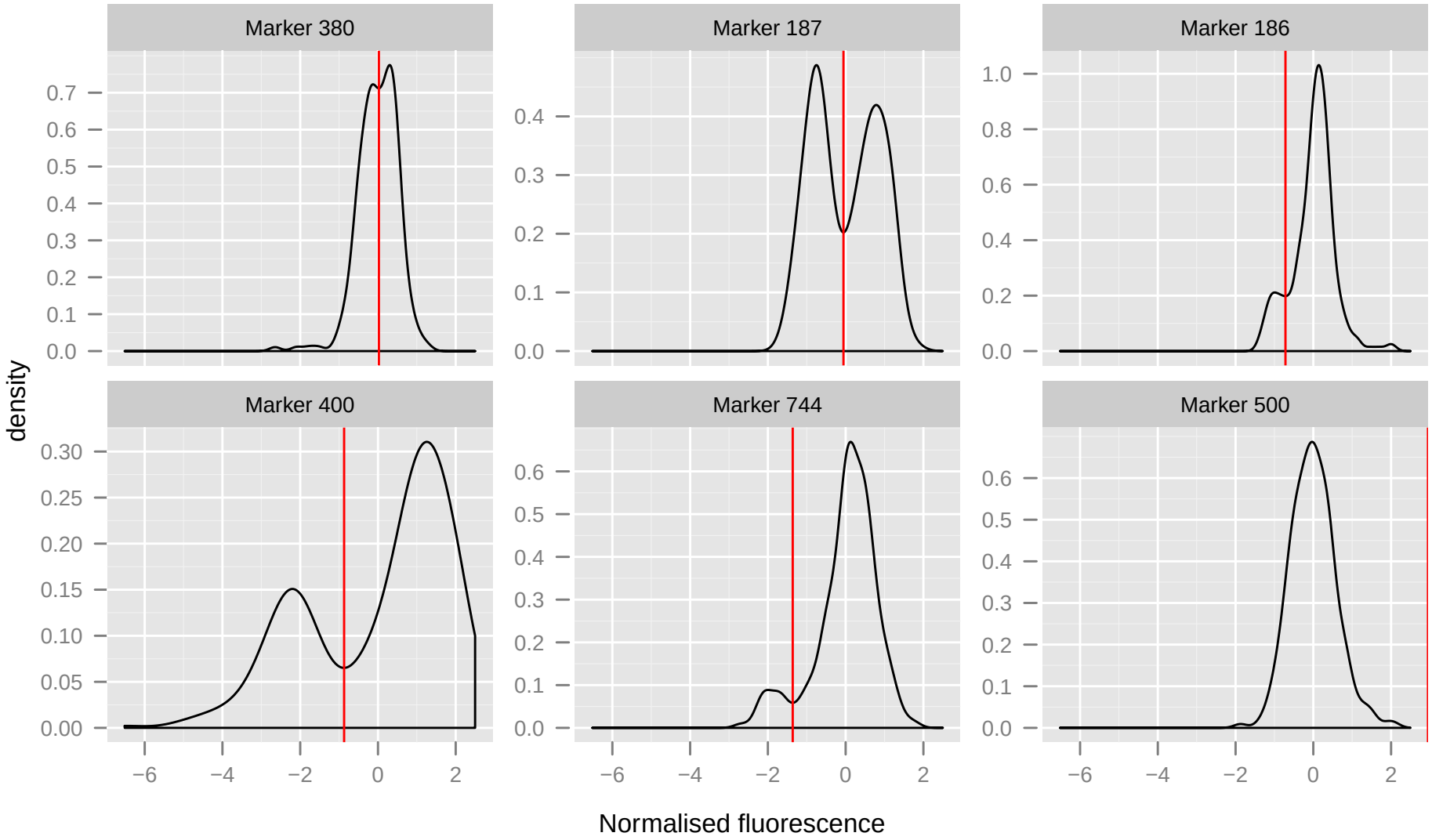
Classification

- Two types of markers
  - Monomorphic marker: fragment present in all/none specimens
  - Polymorphic marker: fragment present in some specimens
- Probability density distribution of (normalised) fluorescence per marker
  - Monomorphic markers $\Rightarrow$ unimodal distribution
  - Polymorphic markers $\Rightarrow$ bimodal distribution

Classification

- Two types of markers
  - Monomorphic marker: fragment present in all/none specimens
  - Polymorphic marker: fragment present in some specimens
- Probability density distribution of (normalised) fluorescence per marker
  - Monomorphic markers $\Rightarrow$ unimodal distribution
  - Polymorphic markers $\Rightarrow$ bimodal distribution
- Threshold $\Rightarrow$ minimum density between two maxima $\Rightarrow$ variable among markers
- (Normalised) fluorescence above threshold $\Rightarrow$ present

Densities of the normalised fluorescence per marker with threshold



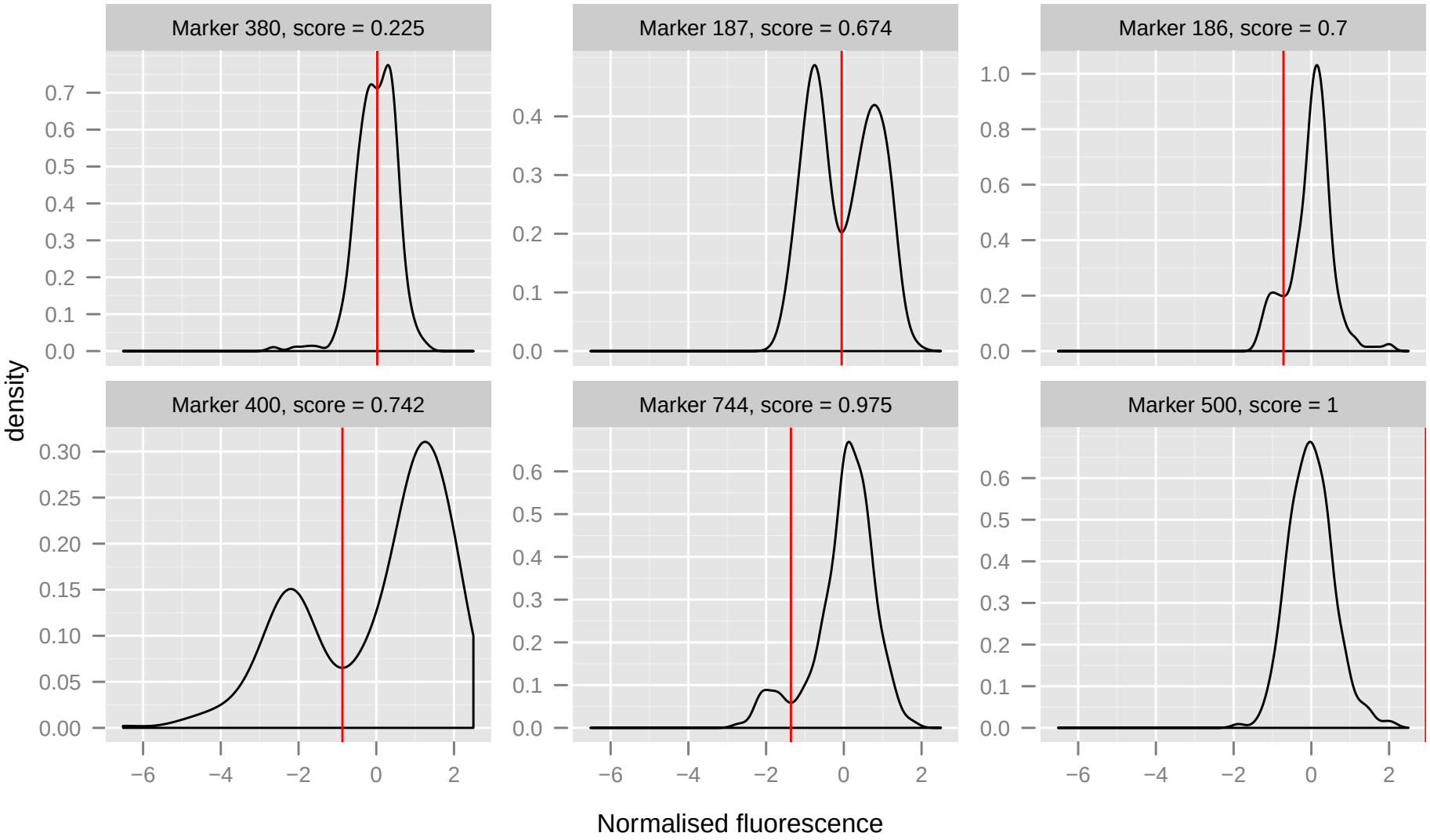
Repeatability

- Compare classification of different replicates from same specimen
- $R = 1 - \frac{\sum E_{ij}}{\sum M_{ij}}$
- Varies between 0 ($\forall i,j : E_{ij} = M_{ij}$) and 1 ($\forall i,j : E_{ij} = 0$)
- Our repeatability score = complement of 'technical difference rate' TDR (Bonin et al., 2004)
  - TDR only defined for exactly 2 replicates per specimen
  - Our formula = generalisation of TDR

Repeatability

- Compare classification of different replicates from same specimen
- $R = 1 - \frac{\sum E_{ij}}{\sum M_{ij}}$
- Varies between 0 ($\forall i,j : E_{ij} = M_{ij}$) and 1 ($\forall i,j : E_{ij} = 0$)
- Our repeatability score = complement of 'technical difference rate' TDR (Bonin et al., 2004)
  - TDR only defined for exactly 2 replicates per specimen
  - Our formula = generalisation of TDR

Densities of the normalised fluorescence per marker with threshold and repeatability score



Further analysis

- Classification = **start** genetic analysis
- Results available as data.frame $\Rightarrow$ easy to use with another package
- Direct methods for `hclust()` and `princomp()`
- Suggestions welcome

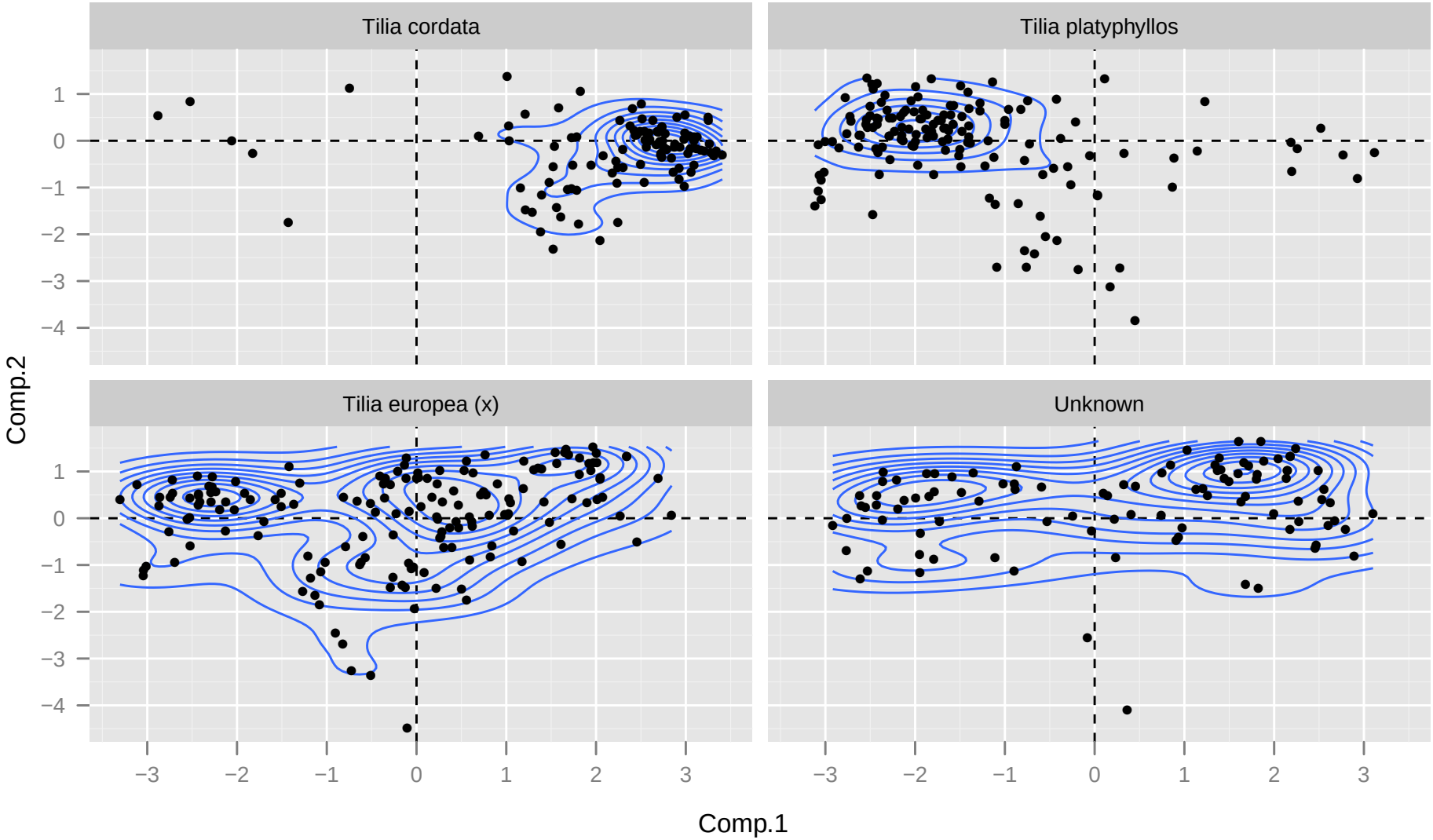
Genetic analysis of lime trees

- Two species: *Tilia cordata* and *T. platyphyllos*
- Different morphology (e.g. small vs large leaves,. . .)
- Hybrid: *Tilia europea (x)*
- Intermediate morfology
- In situ determination difficult?

Genetic analysis of lime trees

- Two species: *Tilia cordata* and *T. platyphyllos*
- Different morphology (e.g. small vs large leaves,. . .)
- Hybrid: *Tilia europea (x)*
- Intermediate morfology
- In situ determination difficult?

PCO using Jaccard distance split by field determination



The AFLP package

- Available on R-Forge
  ```
  install.packages("AFLP",
  repos="http://R-Forge.R-project.org")
  ```
- Each step $\Rightarrow$ seperate function
- Additional functionality
  - Generation of random lab design with replication
  - Imports two data formats: SAGA (Li-Cor) and ABI
  - Automatic binning of peaks available (algorithm of Arrigo et al. (2009))
- All information in one S4-object
- Output on screen, to LaTeX or no output

Strength and weakness of the package

Strength

- Reproducible and objective scoring

Weakness

Strength and weakness of the package

Strength

- Reproducible and objective scoring
- Assement of repeatability

Weakness

Strength and weakness of the package

Strength

- Reproducible and objective scoring
- Assement of repeatability
- Variable thresholds
- Corrections for lab effects

Weakness

Strength and weakness of the package

Strength

- Reproducible and objective scoring
- Assement of repeatability
- Variable thresholds
- Corrections for lab effects
- Once fluorescence is measured $\Rightarrow$ one software environment/package

Weakness

Strength and weakness of the package

Strength

- Reproducible and objective scoring
- Assement of repeatability
- Variable thresholds
- Corrections for lab effects
- Once fluorescence is measured $\Rightarrow$ one software environment/package
- Much faster than manual scoring $\Rightarrow$ months reduced to hours
- Extra markers = little extra effort $\Rightarrow$ more information
- Scoring doable in the lab $\Rightarrow$ fast feedback on quality

Weakness

Strength and weakness of the package

Strength

- Reproducible and objective scoring
- Assement of repeatability
- Variable thresholds
- Corrections for lab effects
- Once fluorescence is measured $\Rightarrow$ one software environment/package
- Much faster than manual scoring $\Rightarrow$ months reduced to hours
- Extra markers = little extra effort $\Rightarrow$ more information
- Scoring doable in the lab $\Rightarrow$ fast feedback on quality

Weakness

- Depends on quality of fluorescence measurements

Thank you for listening

Questions?

References

Arrigo, N., J. W. Tuszynski, D. Ehrich, T. Gerdes, and N. Alvarez (2009). Evaluating the impact of scoring parameters on the structure of intra-specific genetic variation using RawGeno, an R package for automating AFLP scoring. *BMC Bioinformatics*, 10:33.

Bates, D., M. Maechler, and B. Bolker (2011). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999375-39/r1282.

Bonin, A., E. Bellemain, P. Bronken Eidessen, F. Pompanon, C. Borchmann, and P. Taberlet (2004). How to track and assess genotyping errors in population genetic studies. *Molecular Ecology 13*, 3261–3273.

Onkelinx, T., P. Verschelde, S. Neyrinck, G. Genouw, and P. Quataert (2011). AFLP package. `https://r-forge.r-project.org/projects/aflp/`.