How to assess the safety of new plant varieties in the years to come?

MEACB, 23 November 2017

Esther Kok

RIKILT
WAGENINGEN UR

Acknowledgements

RIKILT Wageningen UR

Jeroen van Dijk

Martijn Staats

Marleen Voorhuijzen

Rico Hagelaar

Viola Ghio

Gijs Kleter

WUR - Biometris

Hilko van der Voet

WUR – Plant breeding

Ronald Hutten

Richard Visser

University of Nijmegen

University of Eindhoven

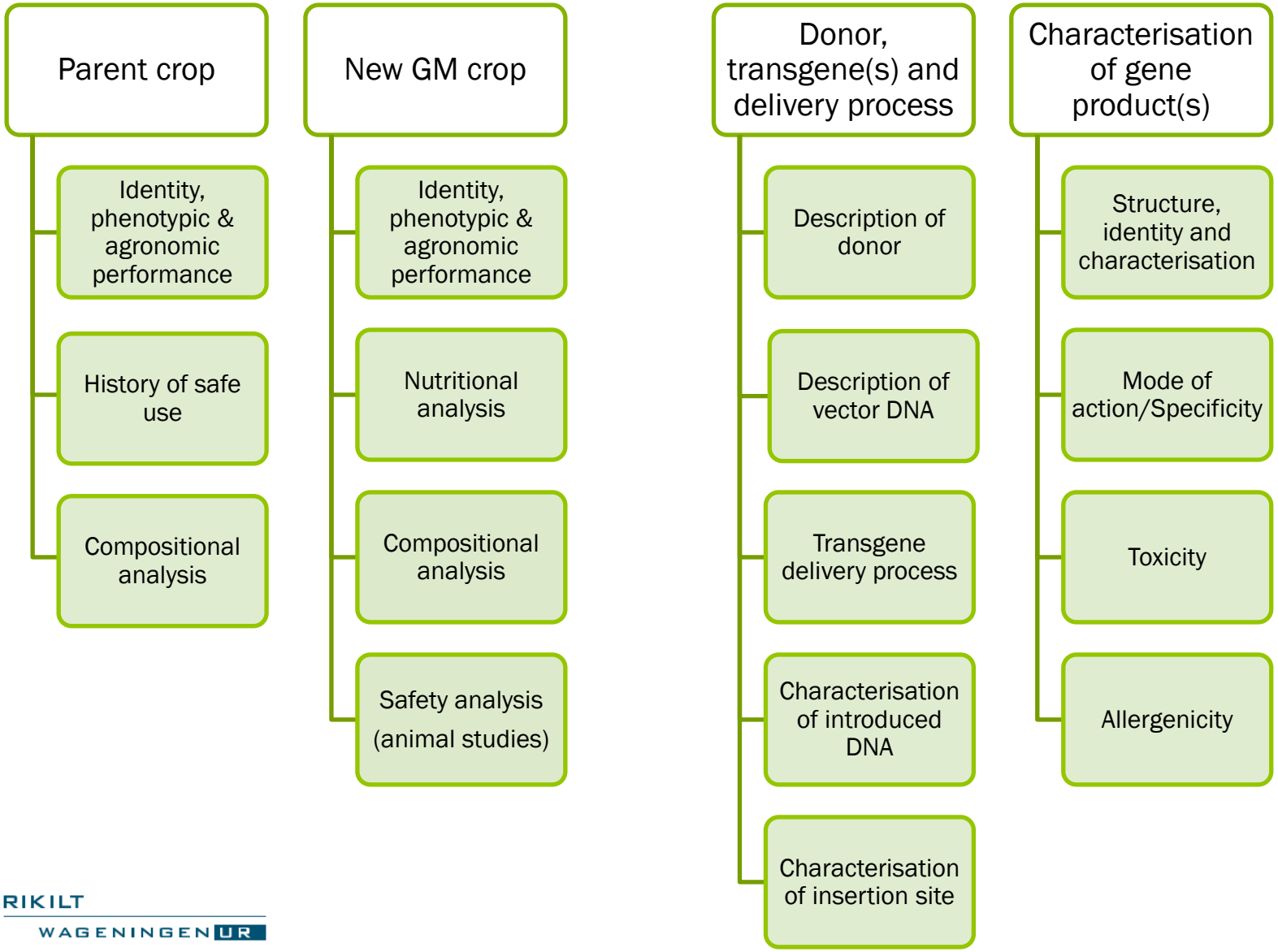
Jeroen Jansen

Edwin van den Heuvel

Alberto Brini

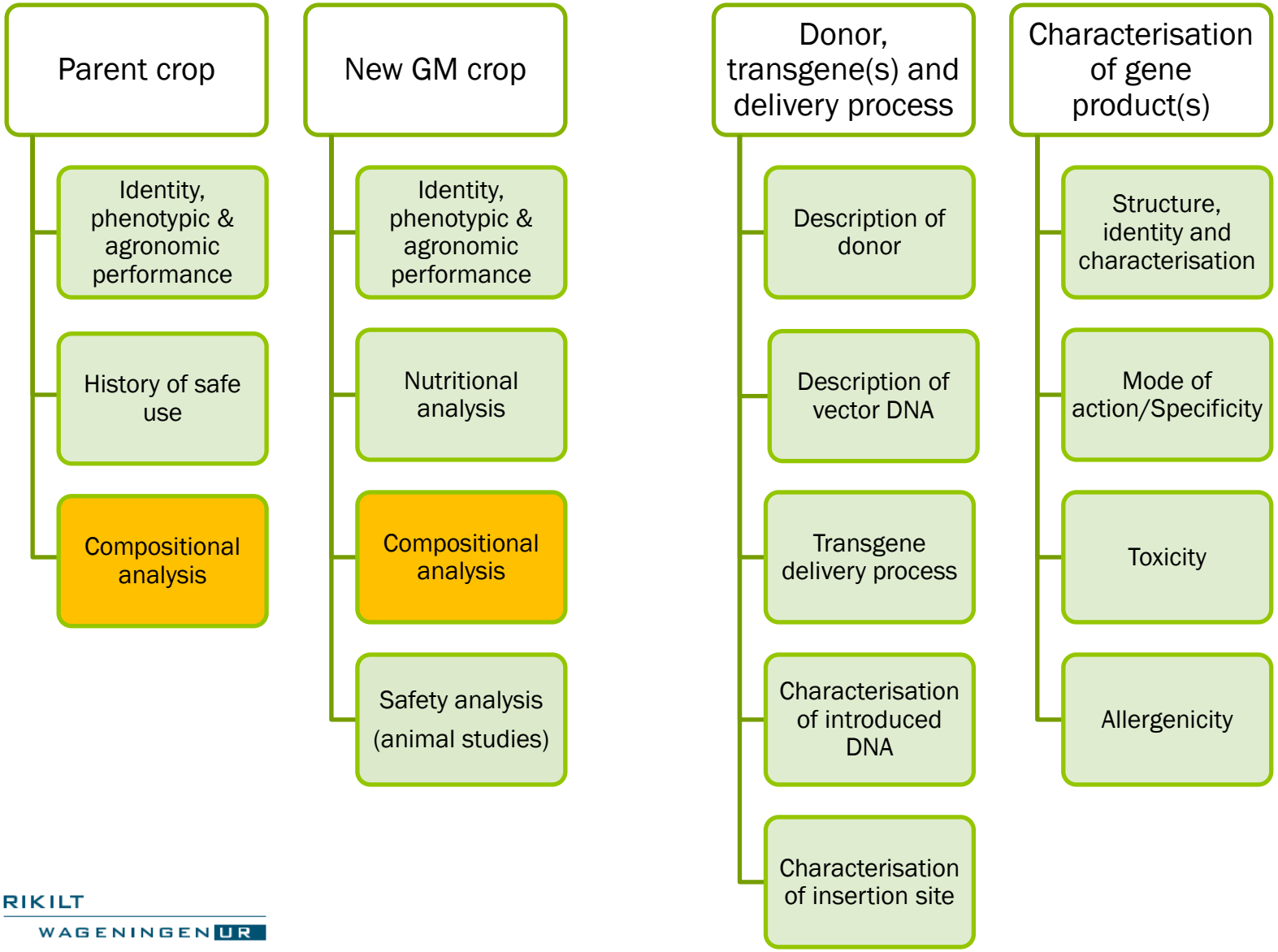
RIKILT WAGENINGEN UR

SAFETY ASSESSMENT OF A NEW GM VARIETY



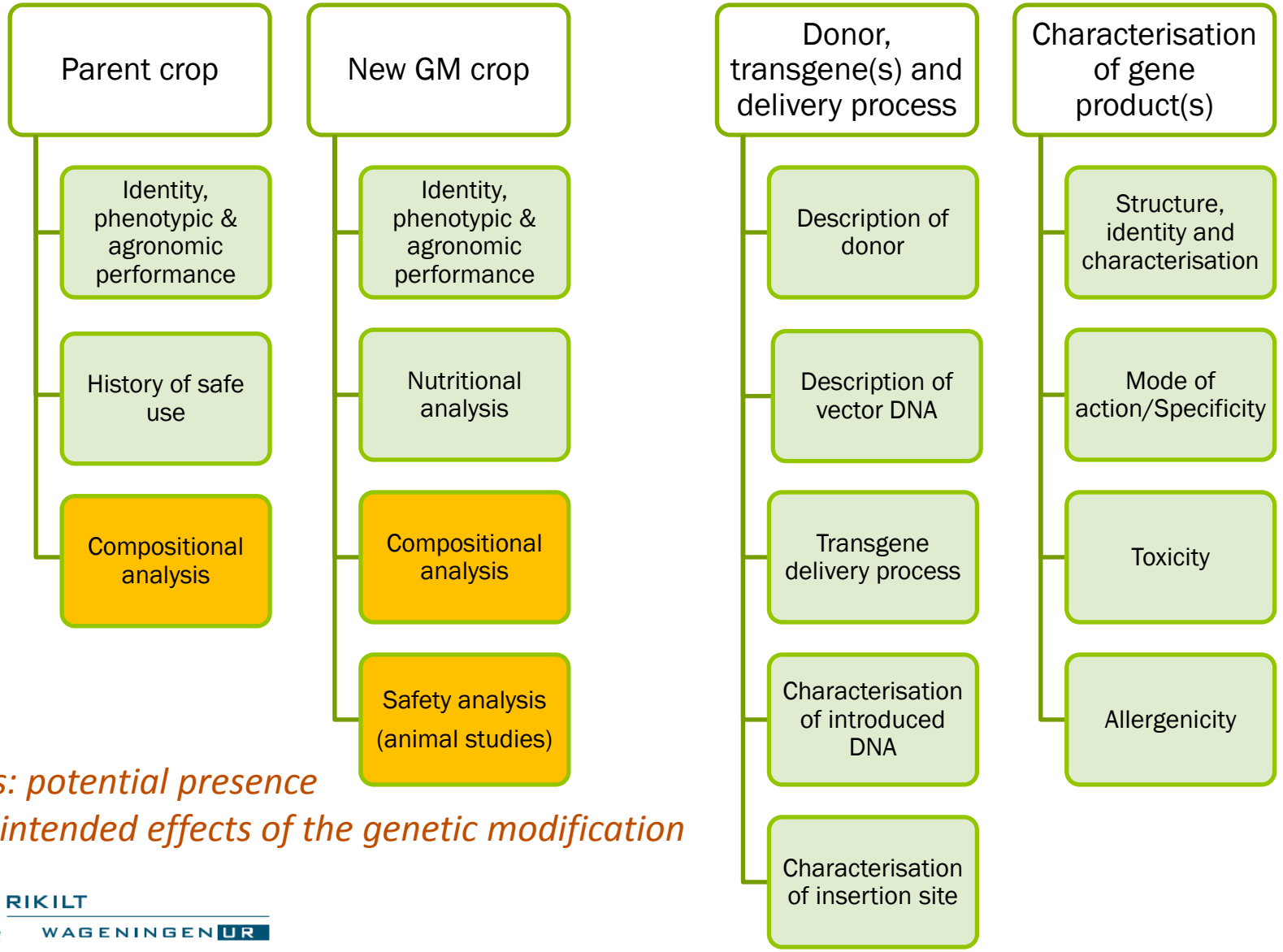
RIKILT
WAGENINGEN UR

SAFETY ASSESSMENT OF A NEW GM VARIETY



RIKILT
WAGENINGEN UR

SAFETY ASSESSMENT OF A NEW GM VARIETY



Focus: potential presence of unintended effects of the genetic modification

RIKILT WAGENINGEN UR

Unintended effects

Why may potential unintended effects not be relevant?

- We have a long history of innovative plant breeding with very few examples of adverse effects
- Plant breeders take their responsibility to develop new crop varieties that are safe and nutritious
- It is unlikely that a safe variety is transformed into an unsafe variety as the result of unintended effects

Unintended effects

Why may potential unintended effects be relevant?

- A range of new and powerful techniques (Crispr-Cas, synthetic biology) allow the rapid introduction of new RNAs, proteins and secondary metabolites, unknown to our food supply chain, possibly even unknown to nature.
- Plant breeding programmes are becoming shorter with less time (years/harvests) to assess new varieties for altered characteristics

Unintended effects

- Two types of potential unintended effects:

❑ Insertional effects

❑ Secundary trait effects

Unintended effects

- Two types of potential unintended effects:

❑ Insertional effects

❑ ***Secundary trait effects***

RIKILT
WAGENINGEN UR

Unintended effects

- Starting-point: it is unlikely that a safe variety is transformed into an unsafe variety as the result of unintended effects

Unintended effects

- Starting-point: it is unlikely that a safe variety is transformed into an unsafe variety as the result of unintended effects

- So we need a basic and pragmatic approach to screen for potential adverse effects related to, primarily, the new trait

Unintended effects

- Starting-point: it is unlikely that a safe variety is transformed into an unsafe variety as the result of unintended effects
- So we need a basic and pragmatic approach to screen for potential adverse effects related to, primarily, the new trait
- Link up as much as possible to data the plant breeder will have already!

Unintended effects

- Hazard identification on the basis of:
  - Molecular characterisation
  - Phenotypic analysis
  - Agronomic performance
  - Compositional analysis (targeted analyses)
  - Animal feeding trials with whole foods

Unintended effects

- Hazard identification on the basis of:
  - Molecular characterisation
  - Phenotypic analysis
  - Agronomic performance
  - **Compositional analysis (targeted analyses)**
  - Animal feeding trials with whole foods

Unintended effects

- Hazard identification on the basis of:
  - Molecular characterisation
  - Phenotypic analysis
  - Agronomic performance
  - Compositional analysis (targeted analyses)
  - Animal feeding trials with whole foods

*In the **GRACE** project:*

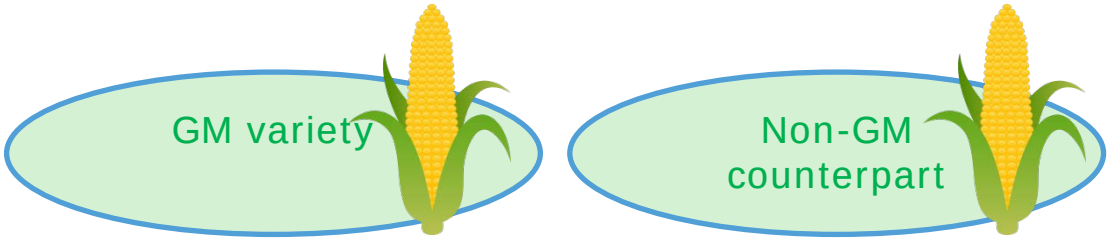
- animal feeding trials with whole foods

*- **detailed compositional analyses** - **same maize materials***

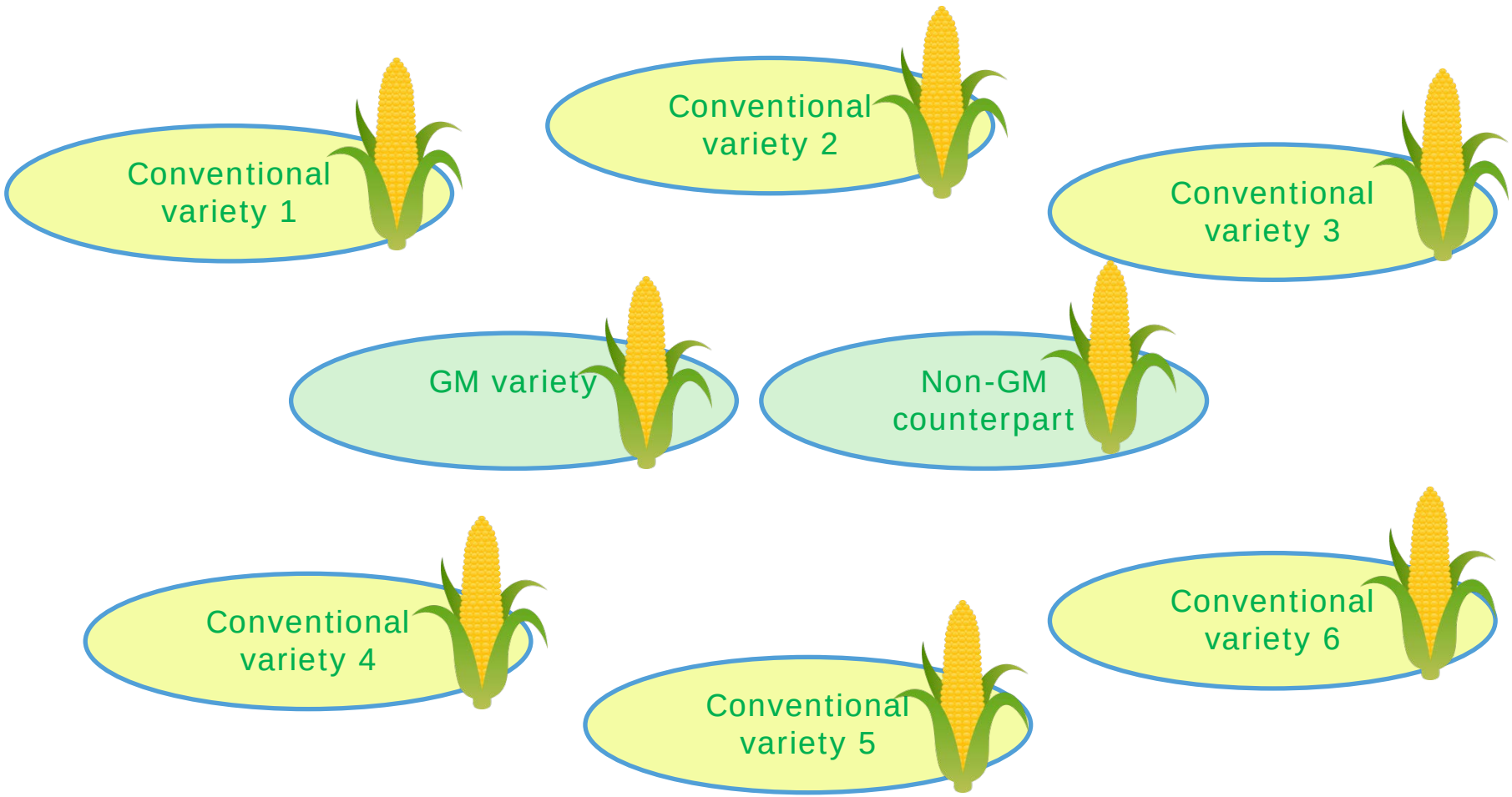
RIKILT WAGENINGEN UR

GRACE
GMO Risk Assessment and Communication of Evidence

Compositional analysis (targeted)



Compositional analysis (targeted)



Compositional analysis, targeted vs omics analysis

Targeted analyses:

- key nutrients (macronutrients/micronutrients),
- key anti-nutrients, including natural toxins

Omics analyses:

- Transcriptome: all transcribed DNA products (RNA)
- Proteome: all proteins
- Metabolome: all secondary metabolites

Targeted versus omics analyses

Targeted analyses

- Few hundreds of end-points
- Limited coverage of individual metabolic routes
- Advanced data analysis is required (comparison with conventional varieties)
- Natural variation needs to be included!

Unintended effects

Targeted analyses

- Few hundreds of end-points
- Limited coverage of individual metabolic routes
- Advanced data analysis is required (comparison with conventional varieties)
- Natural variation needs to be included!

Omics analyses

- Many thousands of end-points
- Broad coverage of individual metabolic routes
- Advanced data analysis is required (comparison with conventional varieties)
- Natural variation needs to be included!

Omics analyses

Omics analyses lead to very large datasets.

The question is: how to analyse for meaningful differences in the omics profiles, given the fact that there is much natural variation between plants due to e.g.

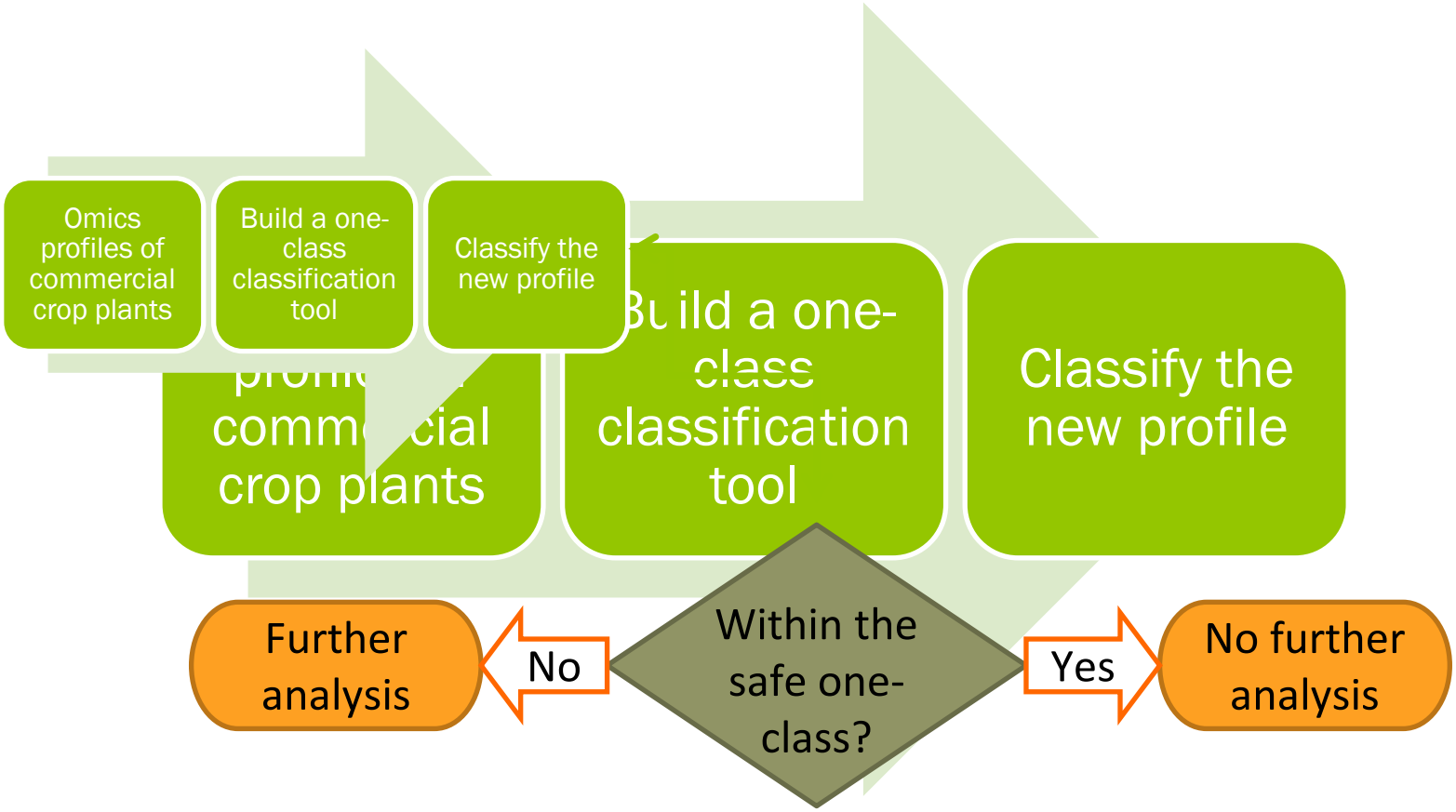
- Genotype
- Environmental conditions of growth
  (soil and climatological conditions)

Model developed with Wageningen UR Biometris (statisticians)
and University of Nijmegen, dept of Chemometrics

Basic criterium: profiles of varieties that can not be considered as safe should fall outside of the one class

RIKILT WAGENINGEN UR

Compare transcriptomics profiles

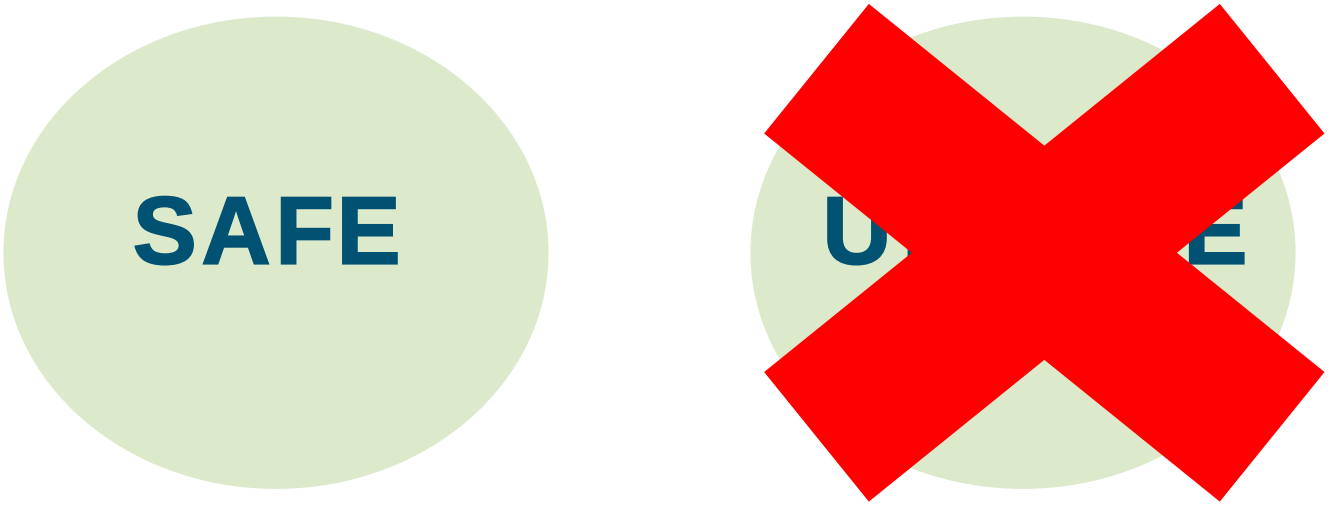


Construction of the one class model (SIMCA)

SIMCA is in fact a PCA model with additional functionality, so a SIMCA class inherits most of the functionality of a PCA class.

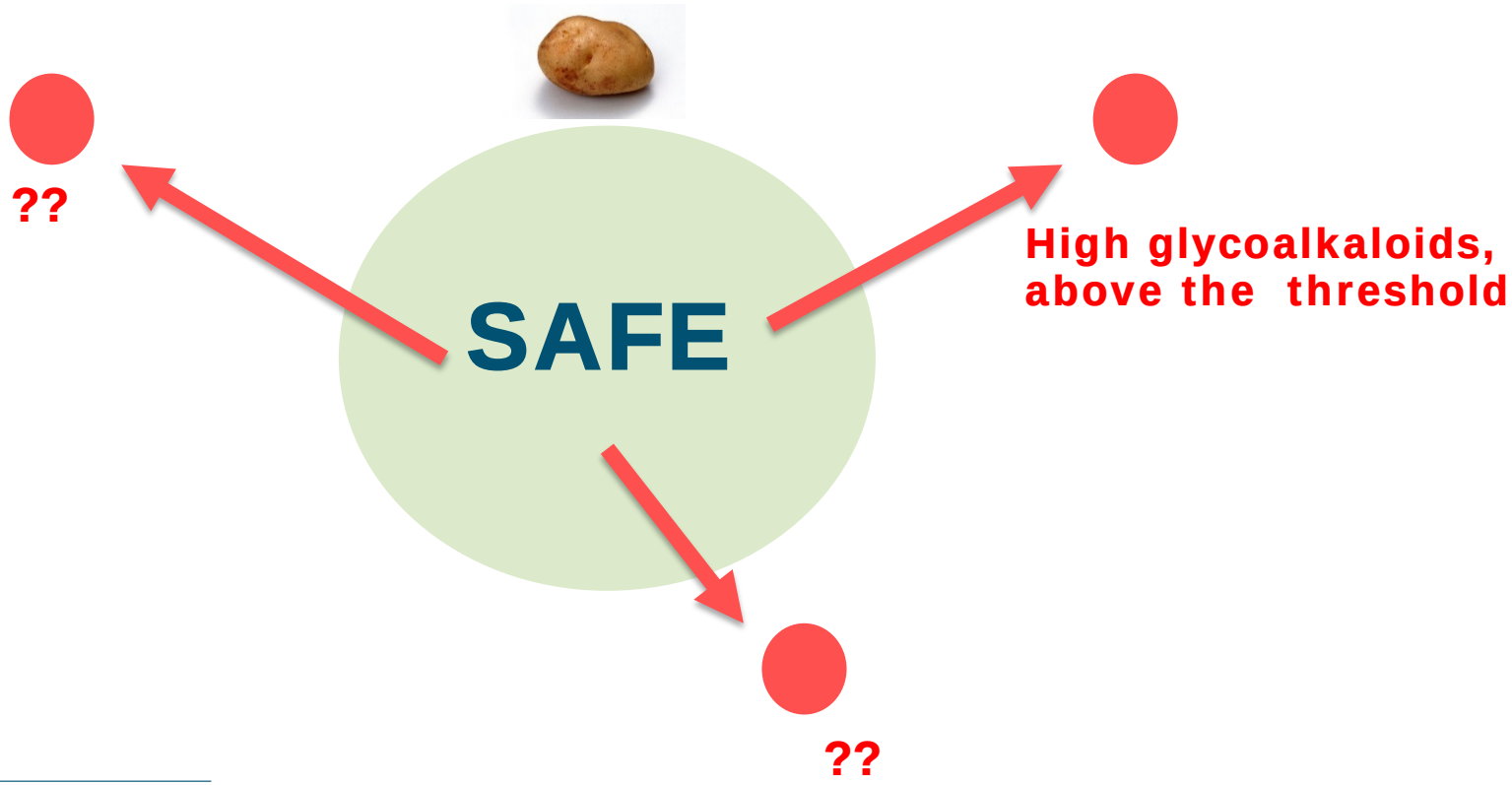
Construction of the one class model (SIMCA)

SIMCA is in fact a PCA model with additional functionality, so SIMCA class inherits most of the functionality of PCA class.



Construction of the one class model (SIMCA)

SIMCA is in fact a PCA model with additional functionality, so SIMCA class inherits most of the functionality of PCA class.



Construction of the one class model (SIMCA)

SIMCA is in fact a PCA model with additional functionality, so SIMCA class inherits most of the functionality of PCA class.

- *Assess whether new varieties are similar to (commercial) varieties that we consider as safe*

Construction of the one class model (SIMCA)

SIMCA is in fact a PCA model with additional functionality, so SIMCA class inherits most of the functionality of PCA class.

- *Assess whether new varieties are similar to (commercial) varieties that we consider as safe*

- *If aberrant profiles are observed: assess the differences for their toxicological relevance*

Construction of the one class model (SIMCA)

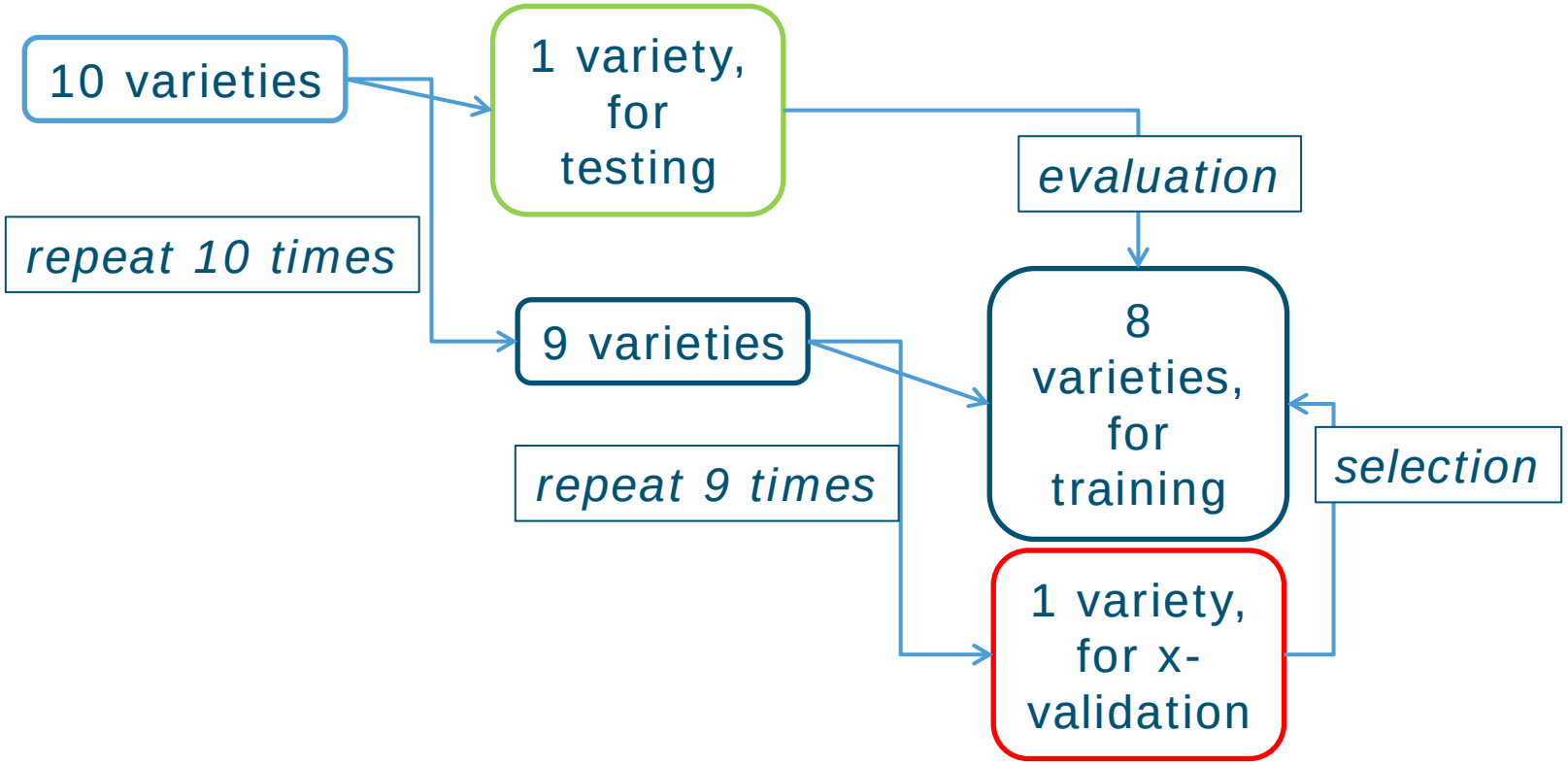
SIMCA is in fact a PCA model with additional functionality, so SIMCA class inherits most of the functionality of PCA class.

Based on:

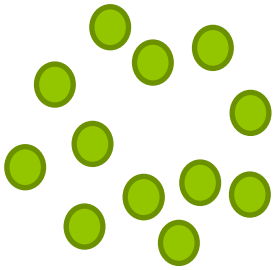
- A training set (commercial varieties considered as safe)
- A data set for cross-validation (commercial varieties considered as safe)
- A test set (for evaluation: well-characterised samples)

Construction of the one class model (SIMCA)

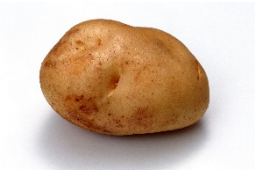
A double loop of cross-validation, for model selection, and testing, for model evaluation - 10 conventional varieties to build the model



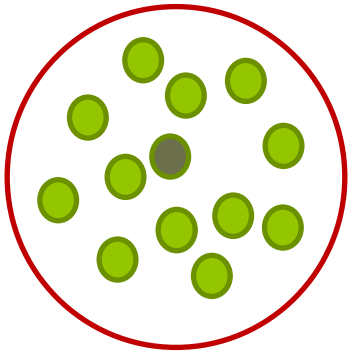
Omics analysis: one class model (SIMCA)



Safe

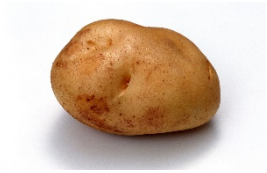


Omics analysis: one class model (SIMCA)



Safe

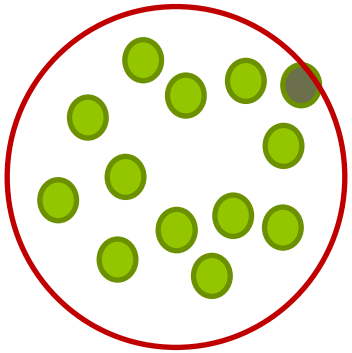
Parent



Omics analysis: one class model (SIMCA)

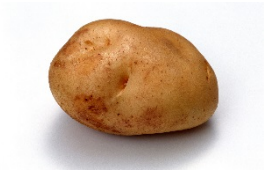
Quality check for the parent line

(or genetically close comparator):



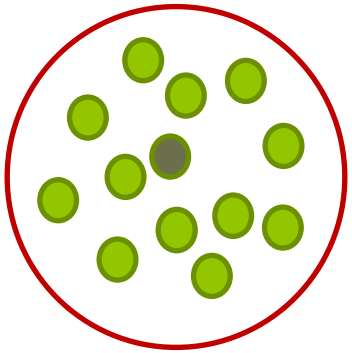
Safe

Parent



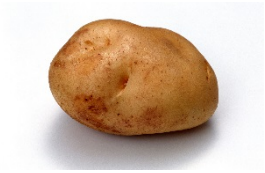
Model of insufficient quality!

Omics analysis: one class model (SIMCA)

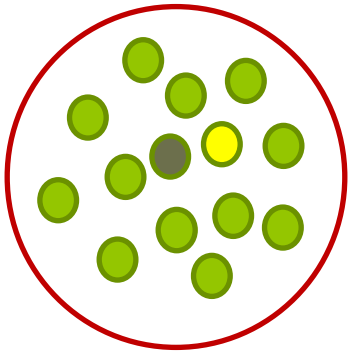


Safe

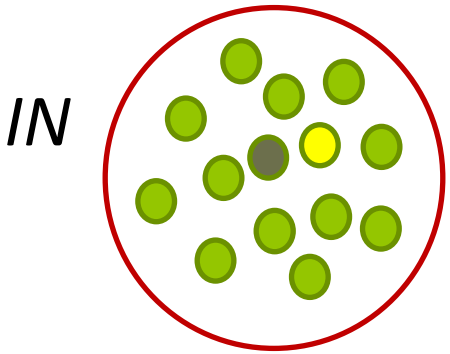
Parent



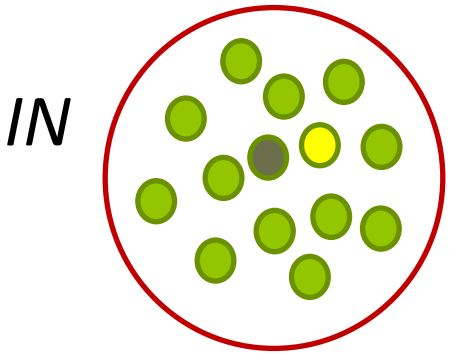
Omics analysis: one class model (SIMCA)



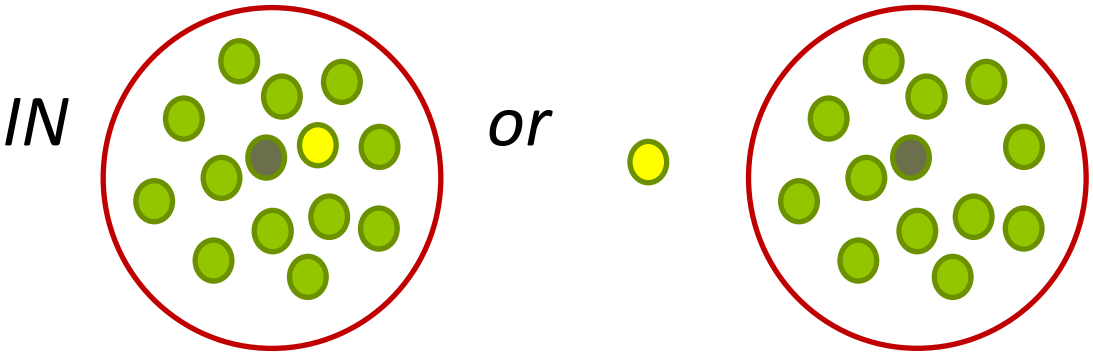
Omics analysis: one class model (SIMCA)



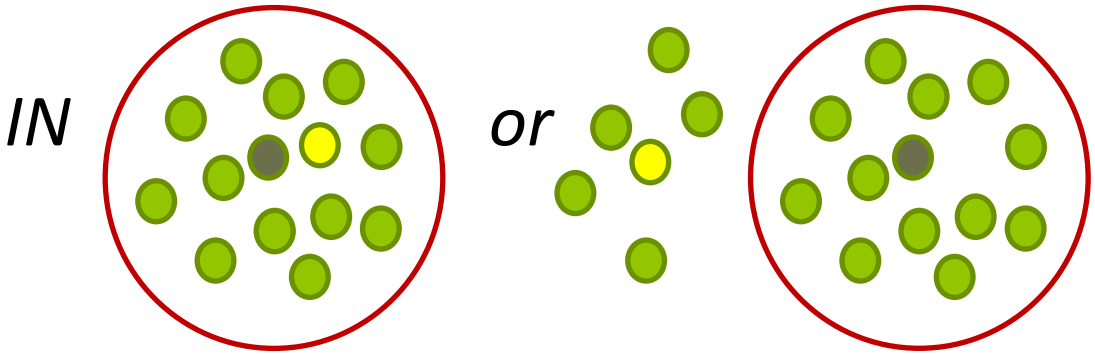
Omics analysis: one class model (SIMCA)



Omics analysis: one class model (SIMCA)



Omics analysis: one class model (SIMCA)



RIKILT
WAGENINGEN UR

Omics analysis: one class model (SIMCA)

Safety assessment of plant varieties using transcriptomics profiling and a one-class classifier

Jeroen P. van Dijk [a,*,1], Carla Souza de Mello [a,b,1], Marleen M. Voorhuijzen [a], Ronald C.B. Hutten [c], Ana Carolina Maisonnave Arisi [b], Jeroen J. Jansen [d], Lutgarde M.C. Buydens [d], Hilko van der Voet [e], Esther J. Kok [b]

[a] RIKILT, Wageningen UR, Wageningen, The Netherlands
[b] Federal University of Santa Catarina, Brazil
[c] Plant Breeding, Wageningen UR, Wageningen, The Netherlands
[d] Institute for Molecules and Materials, Radboud University Nijmegen, The Netherlands
[e] Biometris, Wageningen UR, Wageningen, The Netherlands

ELSEVIER

Regulatory Toxicology and Pharmacology

Volume 70, Issue 1, October 2014, Pages 297–303

RIKILT
WAGENINGEN UR

GRACE omics analyses

	Variables	Profiles
Potato metabolomics RIKILT	100213	44
Potato transcriptomics RIKILT	47582	104
Maize transcriptomics RIKILT	39787	16
Maize transcriptomics CRAG	39621	8
Maize metabolomics RIKILT	128873	46

GRACE
GMO Risk Assessment and Communication of Evidence

RIKILT
WAGENINGEN UR

GRACE omics analyses

GRACE
GMO Risk Assessment and Communication of Evidence

Omics models:

Potato metabolomics: model built based on **10** conventional potato varieties:
- GM variety (*phytophtera* – resistant): inside the one class
- 6 experimental varieties (genetically more distant, fit for human consumption): outside the one class

Potato transcriptomics: model built based on **10** conventional potato varieties:
- GM variety (*phytophtera* – resistant): inside the one class
- 9 experimental varieties (genetically more distant, fit for human consumption): outside the one class
- 2 experimental varieties (genetically more distant, fit for human consumption): inside the one class

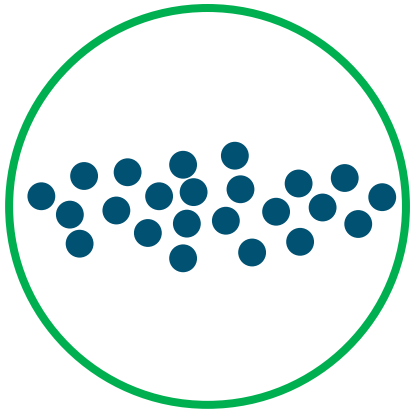
Maize metabolomics: model built based on **7** conventional maize varieties:
- 2 fungus-infected samples: outside the one class

Maize transcriptomics (kernels): model built based on **14** conventional maize varieties:
- 1 GM variety (MON810): inside the one class

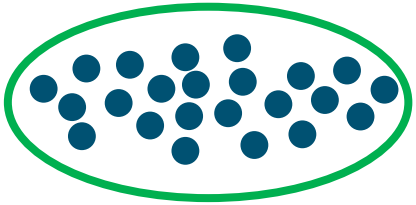
Maize transcriptomics (embryos): model built based on **6** conventional maize varieties:
- 2 GM varieties (MON810): inside the one class

Construction of the one class model (SIMCA)

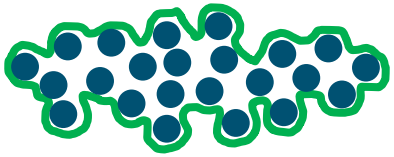
how many principal components for the model?



Few components,
Loose description



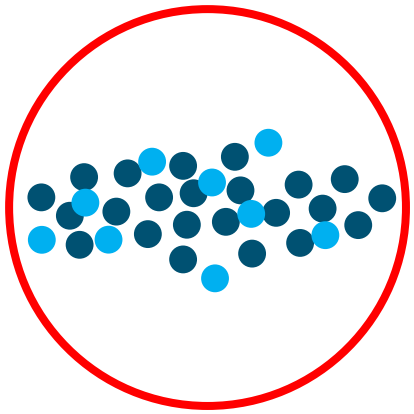
More components,
Fairly accurate description



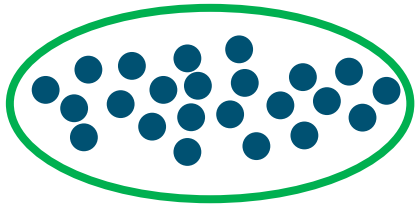
Many components
Very accurate description

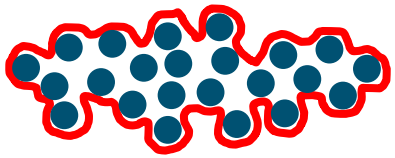
Construction of the one class model (SIMCA)

Use an independent dataset of the same category
Each extra component tightens the model.
Criterion: add as much components until one of the independent datasets falls out.



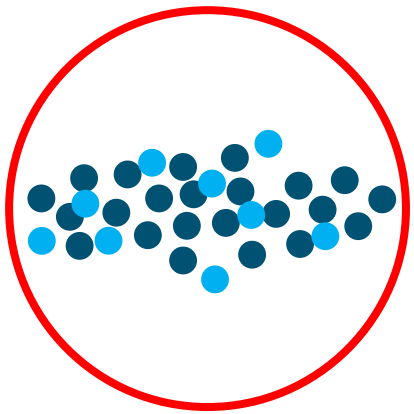
Too wide

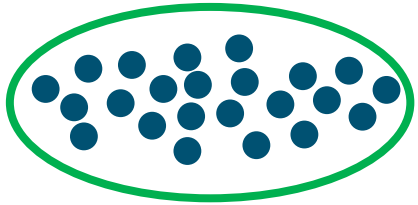


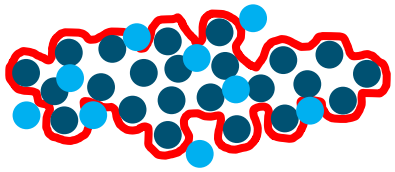


Construction of the one class model (SIMCA)

Use an independent dataset of the same category
Each extra component tightens the model.
Criterion: add as much components until one of the independent datasets falls out.





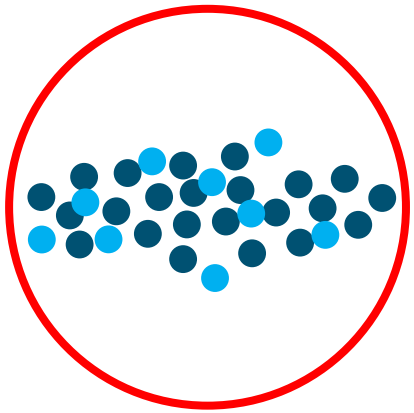


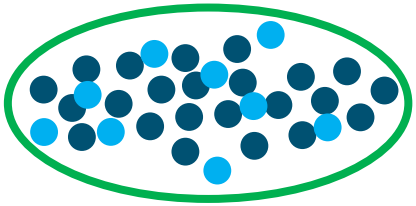
Too tight

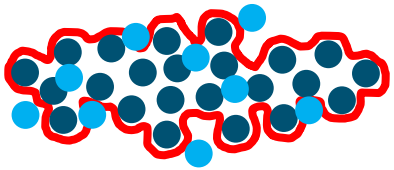
RIKILT
WAGENINGEN UR

Construction of the one class model (SIMCA)

Use an independent dataset of the same category
Each extra component tightens the model.
Criterion: add as much components until one of the independent datasets falls out.



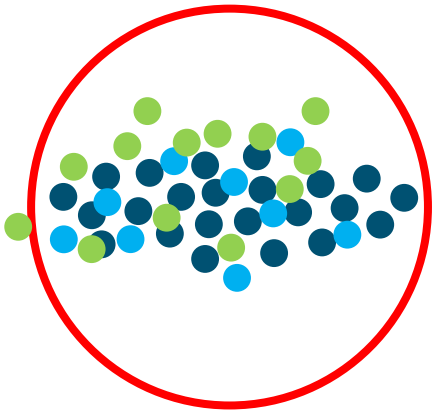




Just right

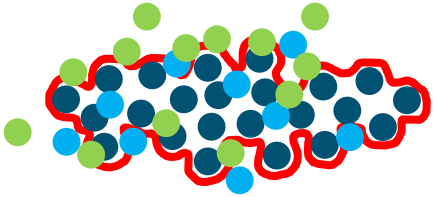
Construction of the one class model (SIMCA)

Use a third independent dataset of the same category to evaluate the model



Very few false positives

A few false positives

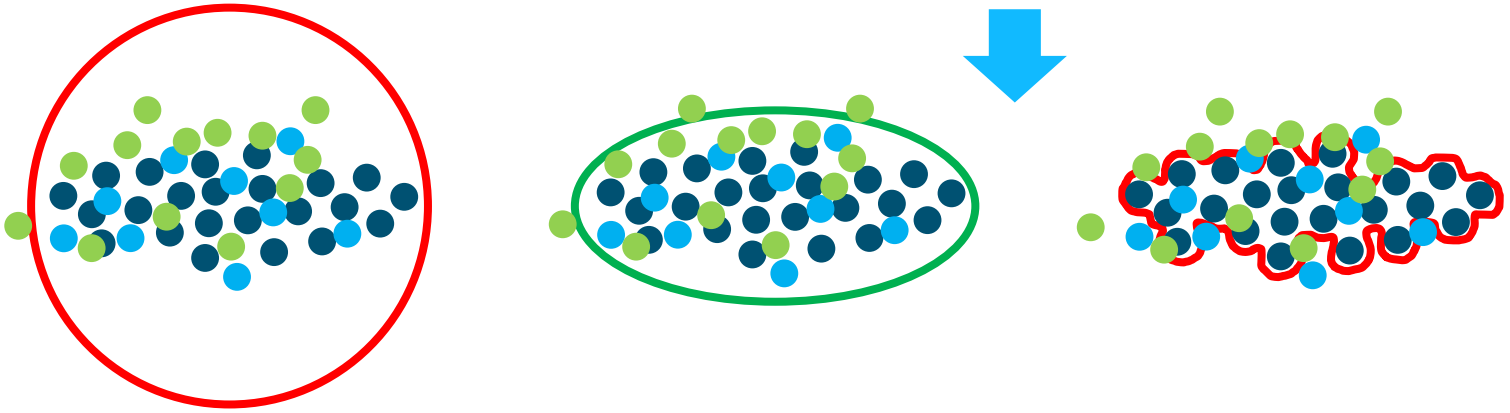


Too many false positives

RIKILT
WAGENINGEN UR

Construction of the one class model (SIMCA)

Use a third independent dataset of the same category to evaluate the model



Very few false positives

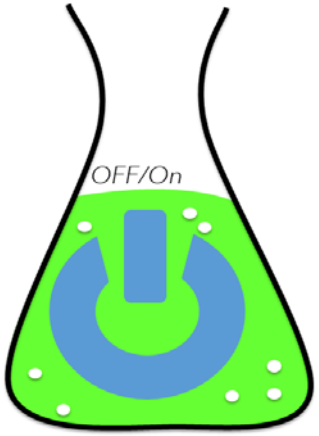
A few false positives

Too many false positives

Technische Universiteit
Eindhoven
University of Technology

RIKILT: safety evaluation of new plant varieties

>COAST>



RIKILT
WAGENINGEN UR

Where innovation starts

Objective

Safety evaluation of new experimental varieties of potato samples:

- Development of a ROBUST statistical methodology
- Classification of the experimental samples according to commercial potatoes with an history of safe use

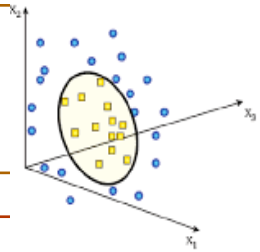
LIST OF METHODS

FEATURE SELECTION

- Selects a subset of relevant features for model construction;
- Avoid the course of dimensionality (#variables > # samples);
- Simplification of the models to make them easier to interpret by users.

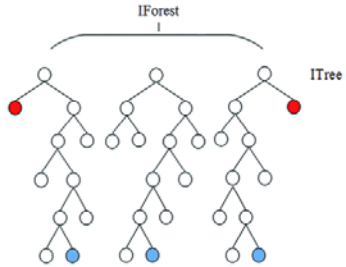
CLASSIC and ROBUST SIMCA

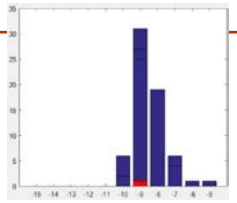
- (Special) PCA treatment of the data;
- ROBUST PCA: improve sensitivity to outliers and to skewed data;
- Different choices of critical values for the one class classifier.

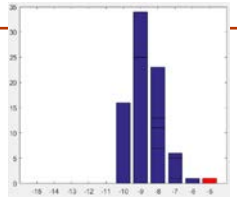


ISOLATION FOREST

- Random Forest of decision trees to detect data anomalies;
- Decision Tree = Isolation of samples in the data (potatoes) by repeatedly selecting at random a feature (gene) from a subsample of the data, then randomly selecting a split in the feature;
- Recursive partitioning -> tree structure: the tree's path length is the measure of abnormality;
- Decision criteria: all the samples ordered according to the measure of abnormality. Ex: outlier samples have the shortest measure.





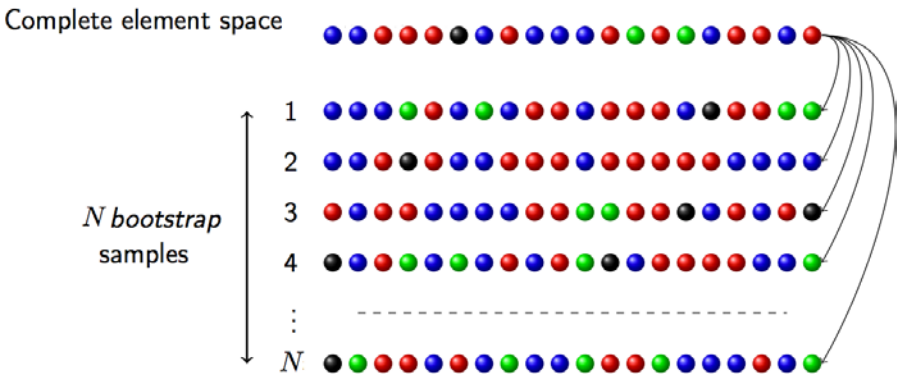


>COAST>

Technical Details

MODEL VALIDATION

- Leave one sample (commercial potato) out
  - SIMCA model on the remaining samples
    - o **Bootstrap** N times the remaining data
    - o Choose the optimal #PC based on the classification rate (expected 95%IN=54/57)
  - Fit SIMCA with the chosen #PC
- Sensitivity (expected 95%IN=55/58)



SIMCA model for class "Commercial" summary

Info:
Significance level (alpha): 0.05
Selected number of components: 1

	Expvar	Cumexpvar	Sens (cal)
Comp 1	28.67	28.67	0.91
Comp 2	12.86	41.53	0.90
Comp 3	7.17	48.70	0.88
Comp 4	5.59	54.29	0.88
Comp 5	4.08	58.37	0.83
Comp 6	3.08	61.45	0.83
Comp 7	2.59	64.04	0.79
Comp 8	2.37	66.41	0.81
Comp 9	2.04	68.44	0.81
Comp 10	1.94	70.39	0.83
Comp 11	1.70	72.09	0.81
Comp 12	1.65	73.74	0.81
Comp 13	1.55	75.29	0.81
Comp 14	1.45	76.74	0.79
Comp 15	1.37	78.11	0.79
Comp 16	1.23	79.34	0.76
Comp 17	1.18	80.52	0.79
Comp 18	1.11	81.63	0.79
Comp 19	0.98	82.61	0.76
Comp 20	0.93	83.54	0.83
Comp 21	0.91	84.45	0.84
Comp 22	0.86	85.31	0.88
Comp 23	0.81	86.11	0.86
Comp 24	0.77	86.89	0.81

CLASSIFICATION OF THE EXPERIMENTAL POTATOES

- SIMCA model on the commercial potatoes
  - Bootstrap N times the commercial data;
  - Choose the optimal #PC.
- Fit SIMCA with the chosen #PC;
- Project the Experimental potatoes in the final model.

Preliminary Results – CLASSICAL SIMCA: MODEL VALIDATION

(SIMCA trained on the log normalized data)

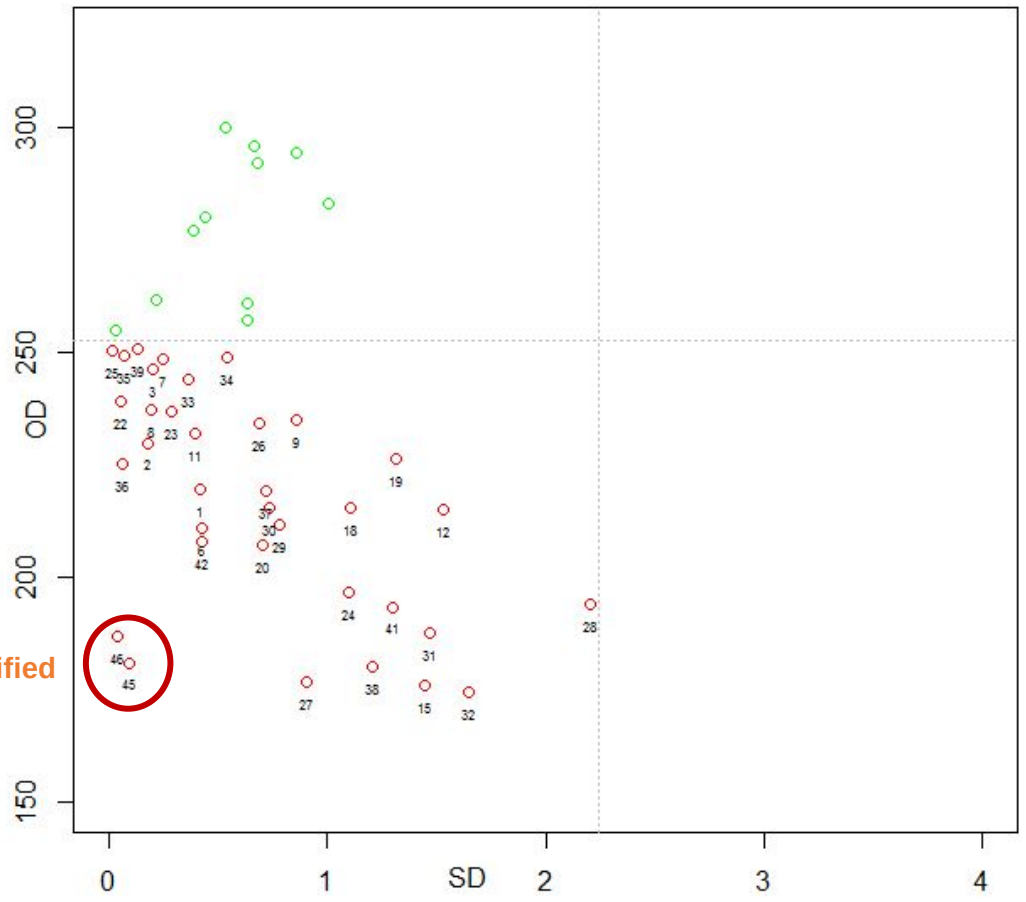
Sensitivity (TRUE POSITIVES): $91.38\% \in [91.2; 98.1]\%$

Preliminary Results – CLASSICAL SIMCA: CLASSIFICATION OF THE EXPERIMENTALS

(SIMCA trained on the log normalized data)

1 PC Selected
~30% Variance Explained

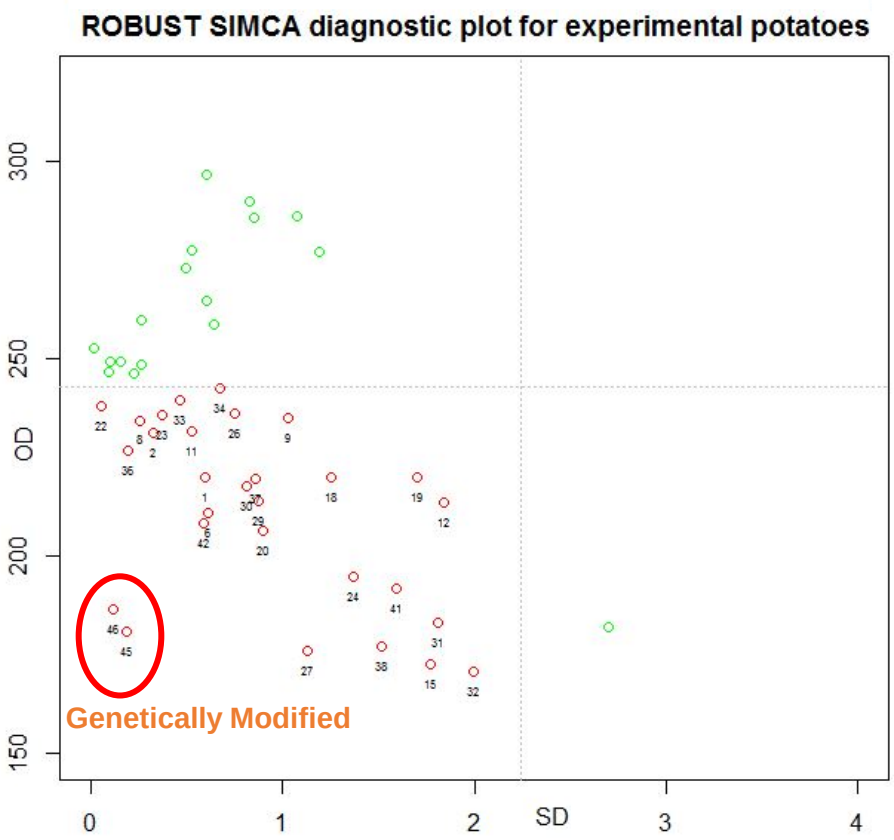
Genetically Modified



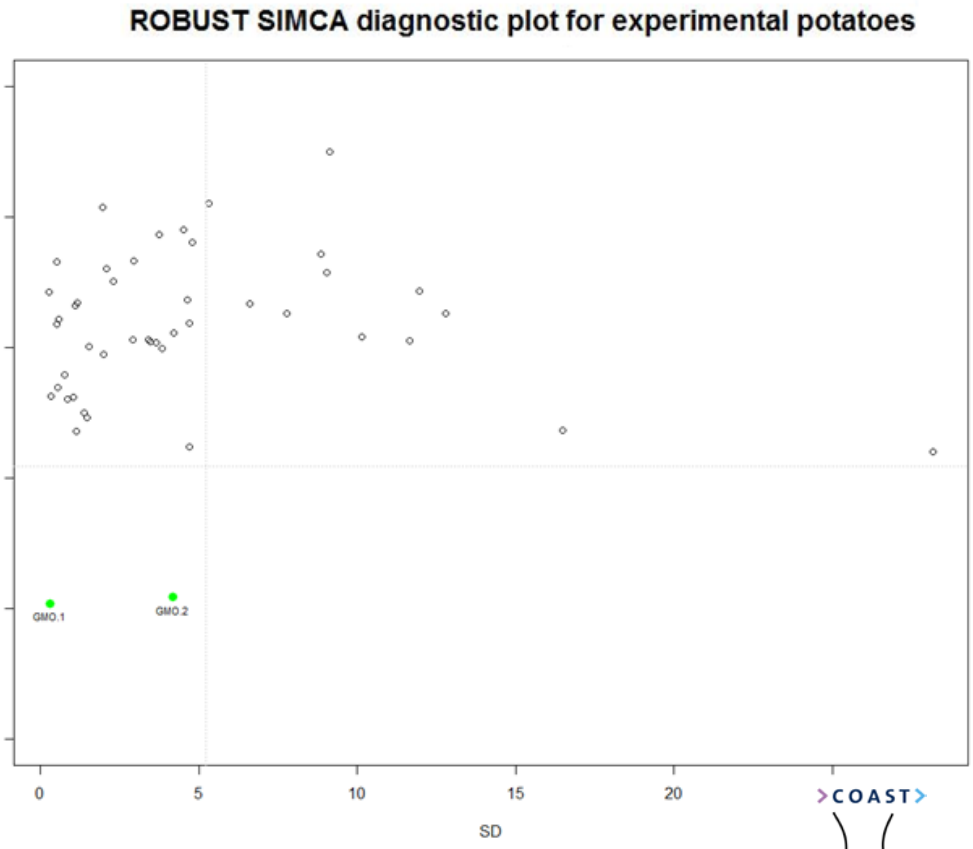
Preliminary Results – RSIMCA: CLASSIFICATION OF THE EXPERIMENTALS

(RSIMCA trained on the log normalized data)

1 PC Selected
~30% Variance Explained



5 PC Selected
~65% Variance Explained



SIMULATION DESIGN

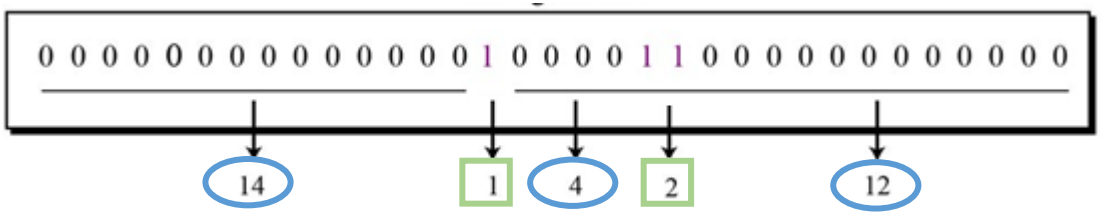
→ How do we reproduce a similar/sparse experiment?

Type	Commercial	Commercial
Unique code	C1Y1S01	C1Y1S02
Year of harvest	2011	2011
Year of Sequenc	2012	2012
SampleID	BIN2t1	BINP1t1
Variety	Bintje	Bintje
XLOC_000001	0	6
XLOC_000002	0	0
XLOC_000003	0	0
XLOC_000004	0	7
XLOC_000005	0	0
XLOC_000006	0	75
XLOC_000007	0	0
XLOC_000008	0	0
XLOC_000009	0	0
XLOC_000010	0	0
XLOC_000011	0	0
XLOC_000012	0	0
XLOC_000013	0	0
XLOC_000014	5248	21
XLOC_000015	0	0
XLOC_000016	0	0
XLOC_000017	0	0
XLOC_000018	0	0
XLOC_000019	4	75
XLOC_000020	0	0
XLOC_000021	0	0
XLOC_000022	0	0
XLOC_000023	0	0
XLOC_000024	0	0
XLOC_000025	0	0
XLOC_000026	0	0
XLOC_000027	0	0
XLOC_000028	0	0
XLOC_000029	0	0
XLOC_000030	0	0

Identification of representative **FEATURES** of the data

- **RUN LENGTH**: #times a 0 or a positive number (coded with 1) appear in the dataset.

Positive counts' Run length 0's Run length



- **NORMALIZED POSITIVE COUNTS**
- **TOTAL READ COUNT**

Study the statistical distributions for the commercial set!

Generate new data using features drawn from the corresponding statistical distributions.

Study each of the new generated data with the proposed methods.

Conclusions

- The GRACE project already showed that unintended effects can likely be more effectively traced by informative omics analyses compared to animal feeding studies with whole foods.
- Additional work is ongoing, but it seems increasingly likely that we can gain insight into complex omics datasets to the extent that we can identify relevant differences, should there be any.
- In that case all data will be available to initially assess observed differences – in specific cases additional testing may be required
- The analysis of omics data should primarily be a tool by plant breeders: they can use this to develop elite varieties that are safe. Risk assessors may evaluate their data.

Thank you very much for your attention!

esther.kok@wur.nl