UPOV

International Union for the Protection of New Varieties of Plants

E

Technical Working Party on Testing Methods and Techniques

TWM/3/24

Third Session
Beijing, China, April 28 to May 1, 2025

Original: English
Date: April 22, 2025

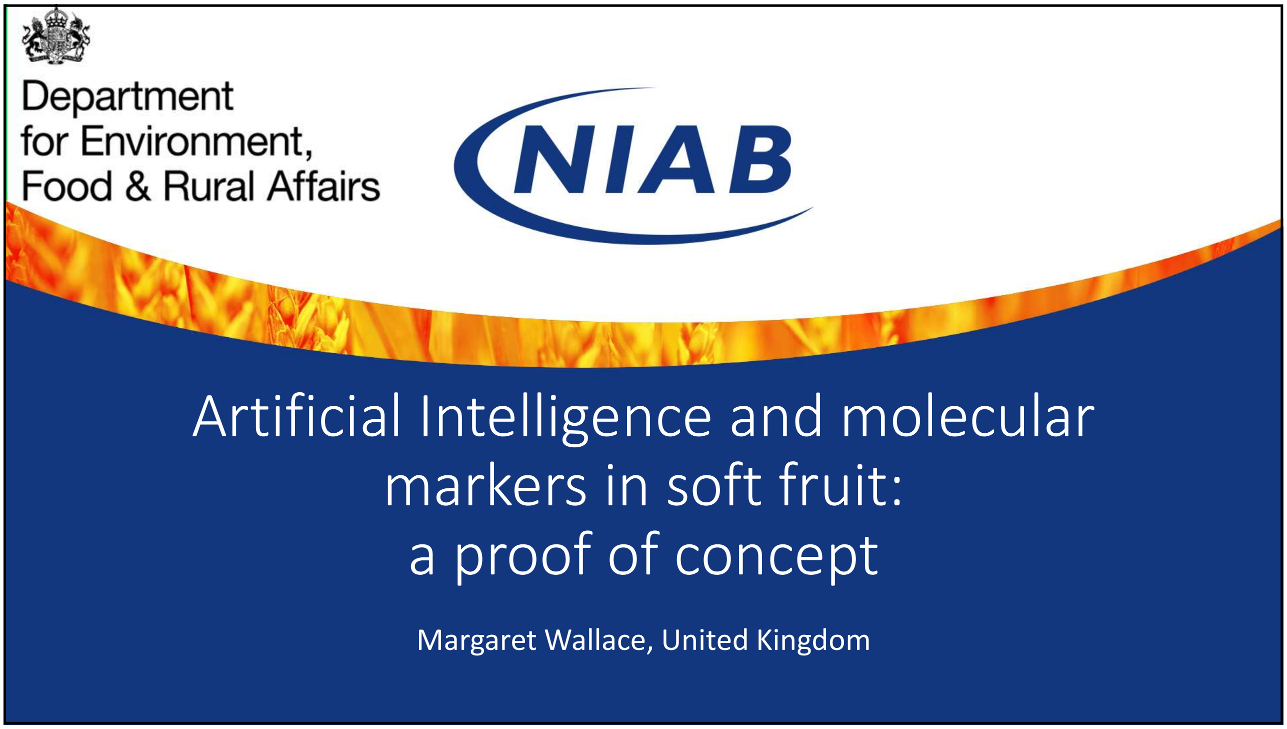
ARTIFICIAL INTELLIGENCE AND MOLECULAR MARKERS IN SOFT FRUIT: A PROOF OF CONCEPT

Document prepared by an expert from the United Kingdom

Disclaimer: this document does not represent UPOV policies or guidance

The annex to this document contains a copy of a presentation “Artificial Intelligence and molecular markers in soft fruit: a proof of concept”, to be made by an expert from the United Kingdom, at the third session of the TWM.

[Annex follows]



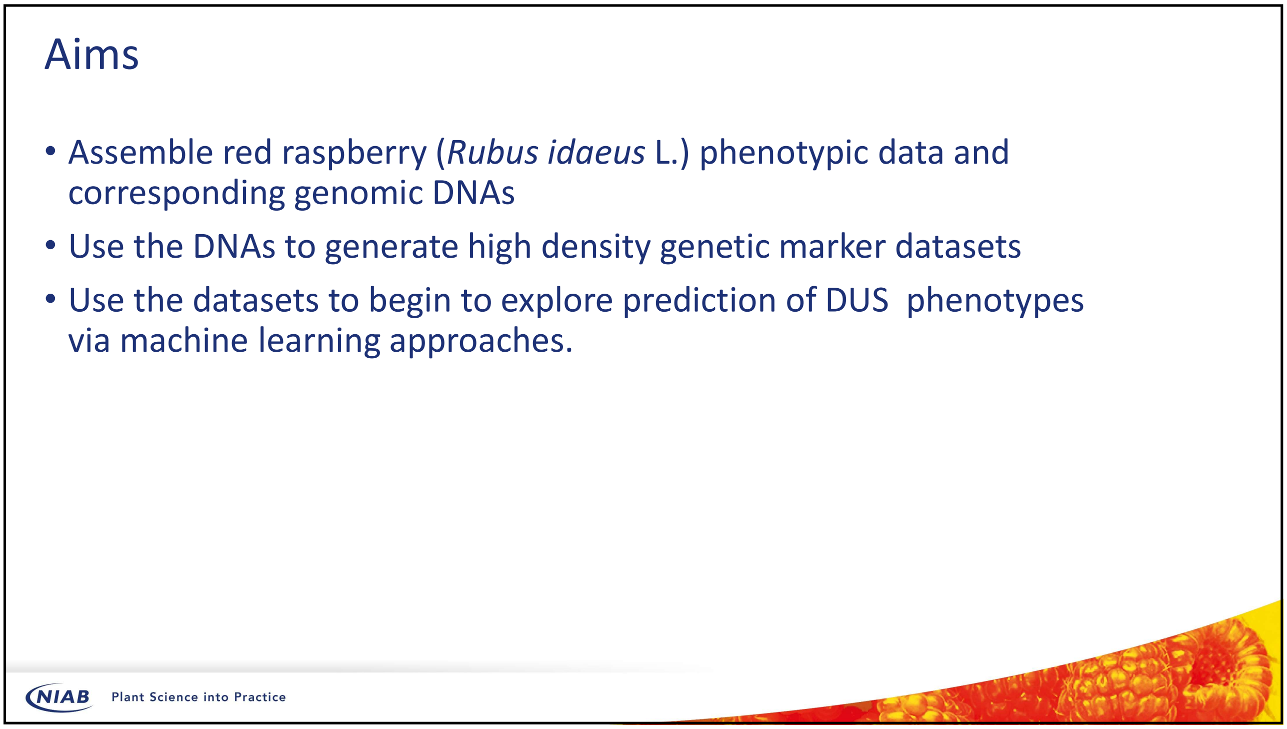
Department for Environment, Food & Rural Affairs

NIAB

Artificial Intelligence and molecular markers in soft fruit: a proof of concept

Margaret Wallace, United Kingdom

1

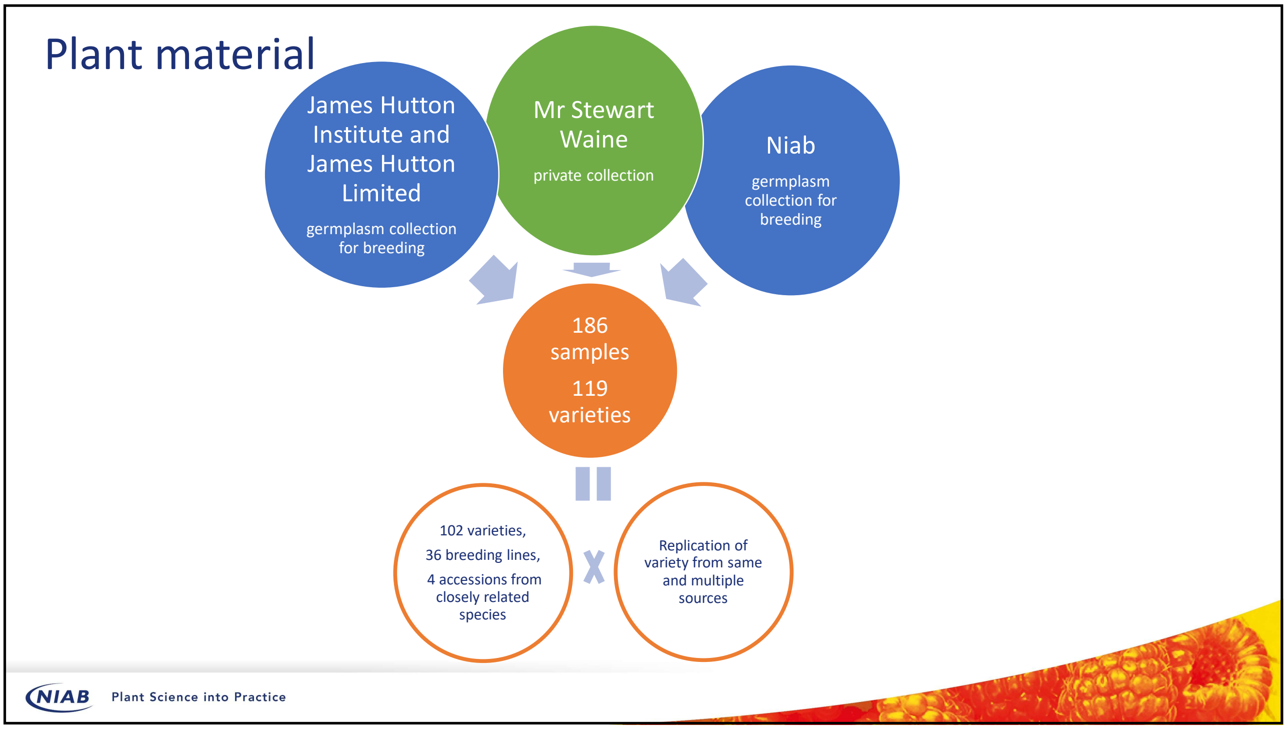


Aims

- Assemble red raspberry (*Rubus idaeus* L.) phenotypic data and corresponding genomic DNAs
- Use the DNAs to generate high density genetic marker datasets
- Use the datasets to begin to explore prediction of DUS phenotypes via machine learning approaches.

NIAB Plant Science into Practice

2



Plant material

James Hutton Institute and James Hutton Limited

germplasm collection for breeding

Mr Stewart Waine

private collection

Niab

germplasm collection for breeding

186 samples

119 varieties

102 varieties,

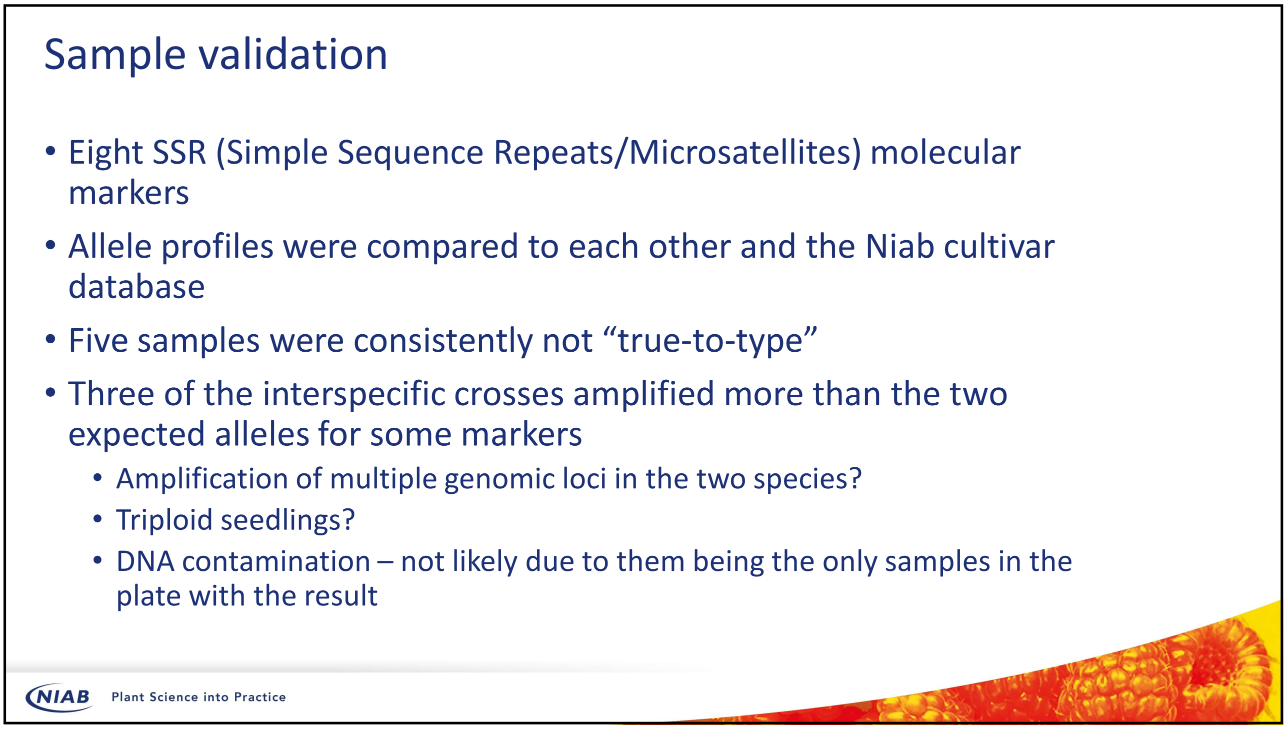
36 breeding lines,

4 accessions from closely related species

Replication of variety from same and multiple sources

NIAB Plant Science into Practice

3

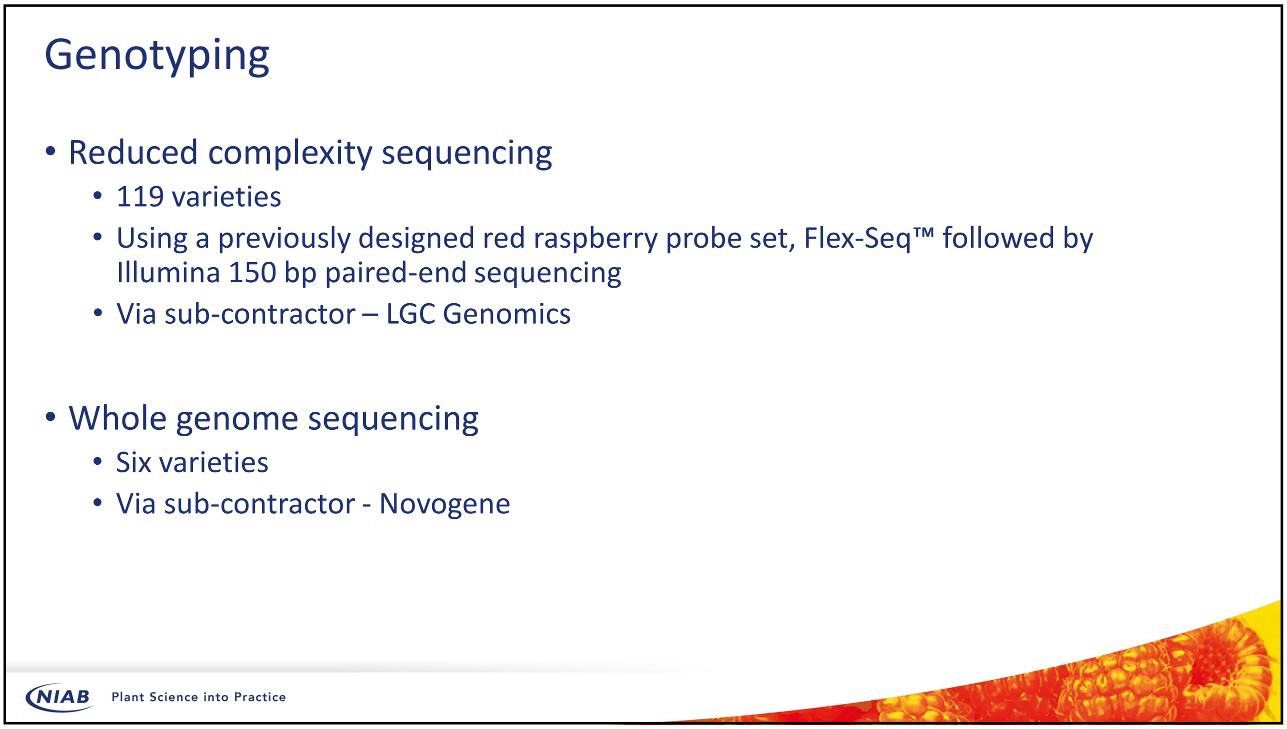


Sample validation

- Eight SSR (Simple Sequence Repeats/Microsatellites) molecular markers
- Allele profiles were compared to each other and the Niab cultivar database
- Five samples were consistently not “true-to-type”
- Three of the interspecific crosses amplified more than the two expected alleles for some markers
    - Amplification of multiple genomic loci in the two species?
    - Triploid seedlings?
    - DNA contamination – not likely due to them being the only samples in the plate with the result

NIAB Plant Science into Practice

4

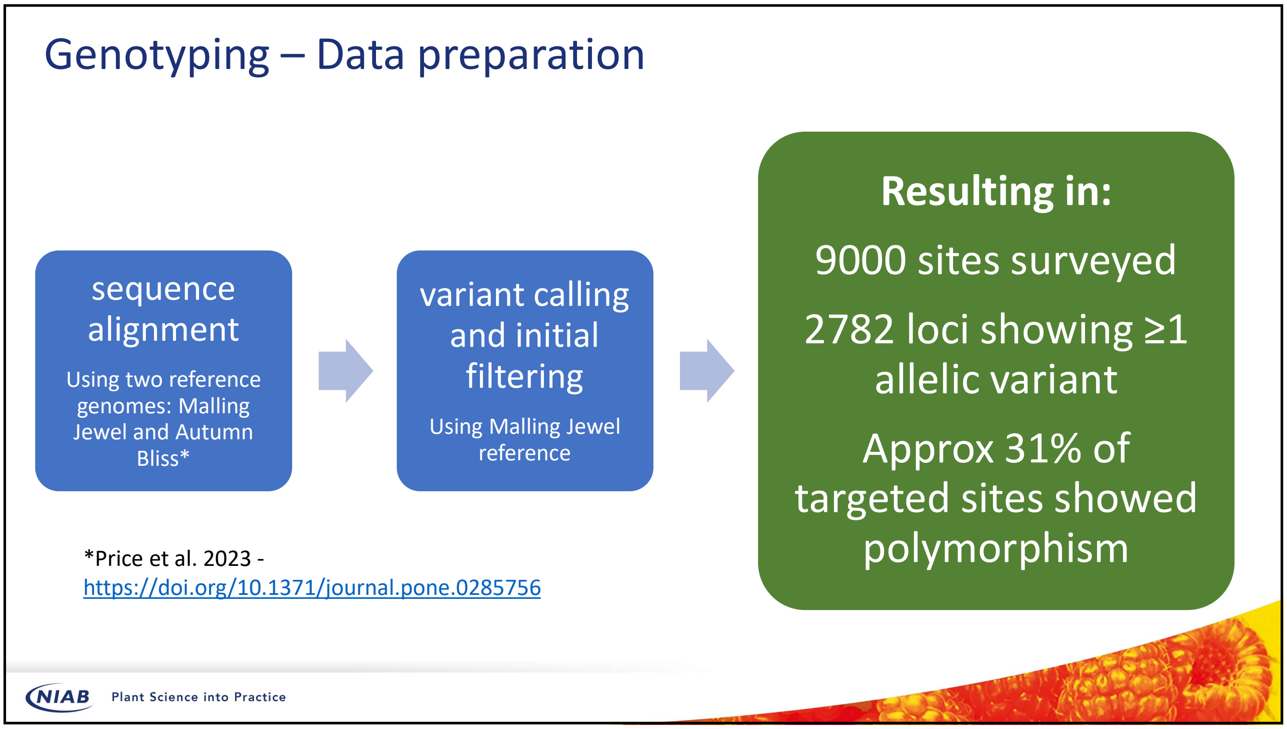


Genotyping

- Reduced complexity sequencing
  - 119 varieties
  - Using a previously designed red raspberry probe set, Flex-Seq™ followed by Illumina 150 bp paired-end sequencing
  - Via sub-contractor – LGC Genomics
- Whole genome sequencing
  - Six varieties
  - Via sub-contractor - Novogene

NIAB Plant Science into Practice

5



Genotyping – Data preparation

sequence alignment

Using two reference genomes: Malling Jewel and Autumn Bliss*

variant calling and initial filtering

Using Malling Jewel reference

Resulting in:

9000 sites surveyed

2782 loci showing ≥1 allelic variant

Approx 31% of targeted sites showed polymorphism

*Price et al. 2023 - https://doi.org/10.1371/journal.pone.0285756

NIAB Plant Science into Practice

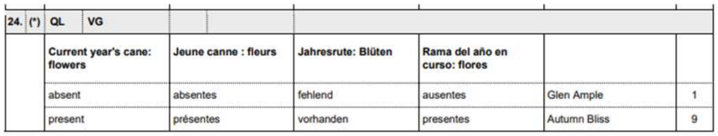
6

Machine learning/deep learning augmented datasets

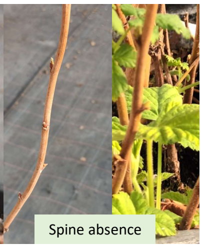
- Converted the raw genotypes to 1 (type 1 homozygote), 0 (heterozygote), -1 (type 2 homozygote), NA (missing value)
- Limited number of samples with genotype and phenotype information.
- Used minor class oversampling to ensure that the minority “score” was represented.

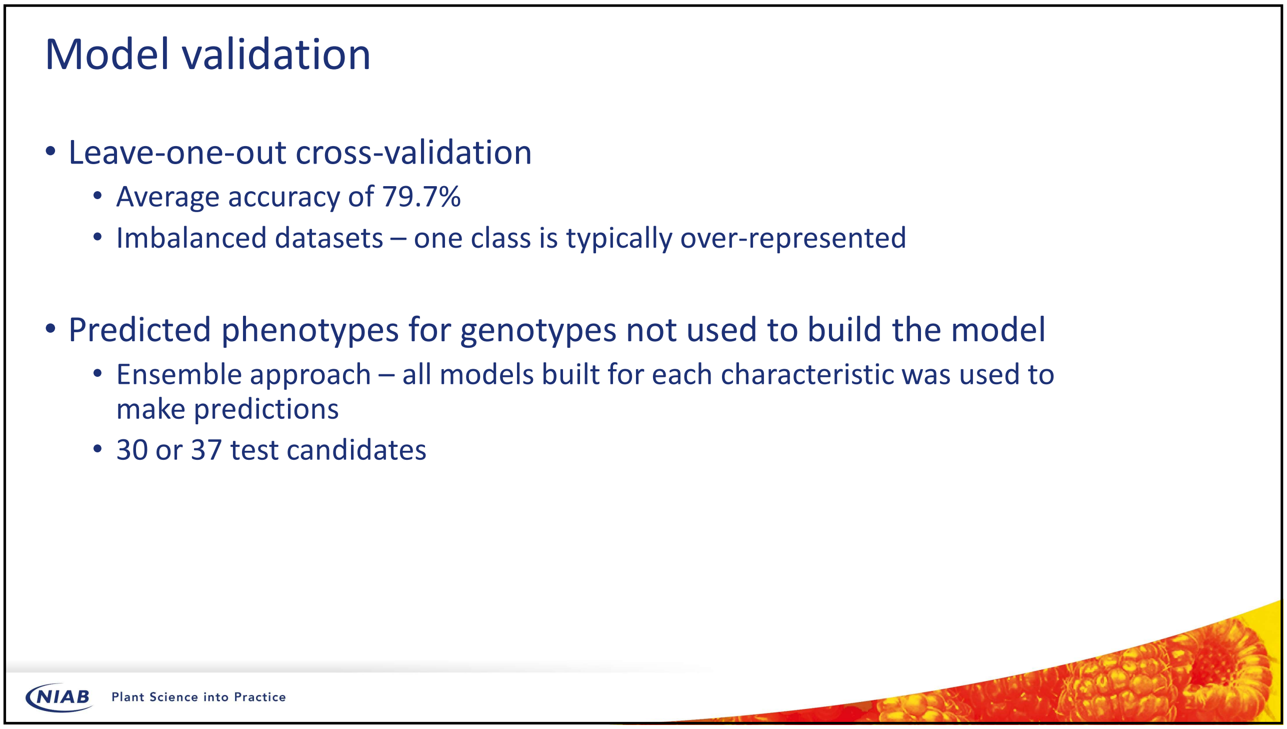
Analysis exploration using supervised machine learning

- Modelling using Random Forests
  - Built with 500 trees
- Made predictions for three characteristics
  - Anthocyanin presence/absence
  - Spine presence/absence
  - Fruiting habit – floricane/primocane



24. (*)	QL VG					
	Current year's cane: flowers	**Jeune canne : fleurs**	**Jahresrute: Blüten**	**Rama del año en curso: flores**		
	absent	absentes	fehlend	ausentes	Glen Ample	1
	present	présentes	vorhanden	presentes	Autumn Bliss	9



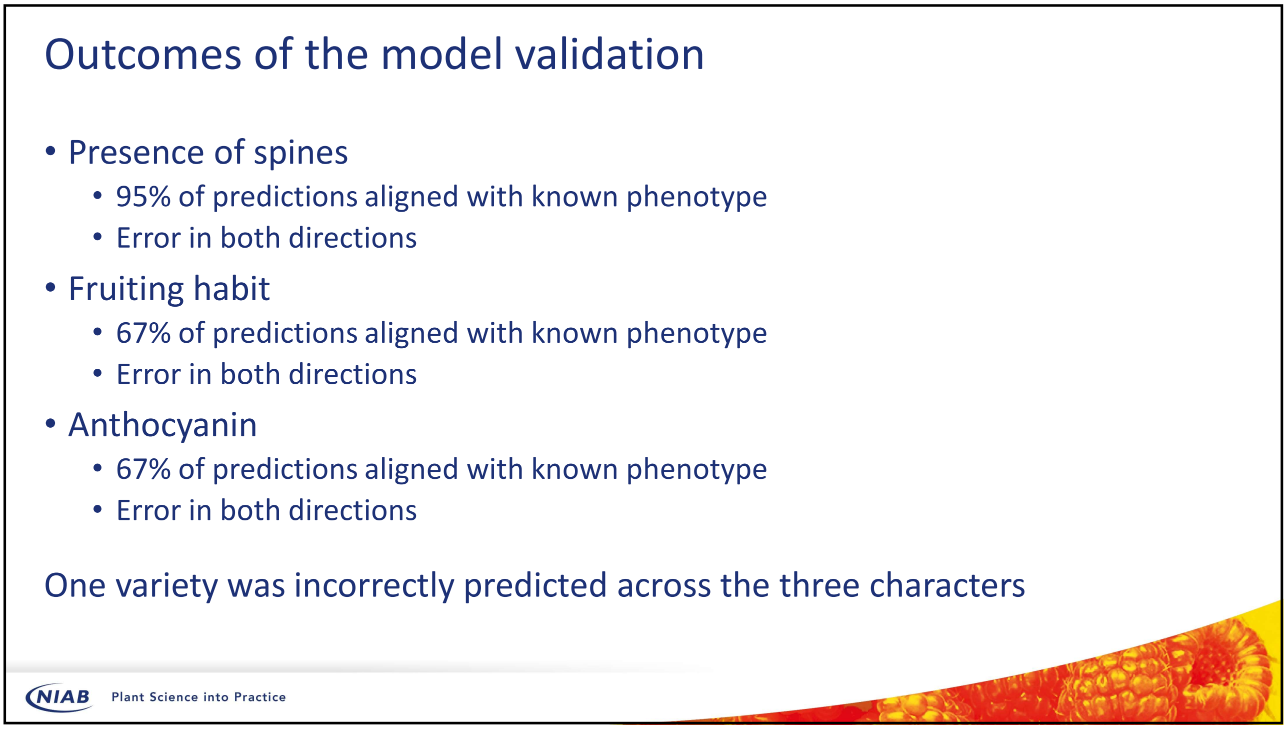


Model validation

- Leave-one-out cross-validation
  - Average accuracy of 79.7%
  - Imbalanced datasets – one class is typically over-represented
- Predicted phenotypes for genotypes not used to build the model
  - Ensemble approach – all models built for each characteristic was used to make predictions
  - 30 or 37 test candidates

NIAB Plant Science into Practice

9



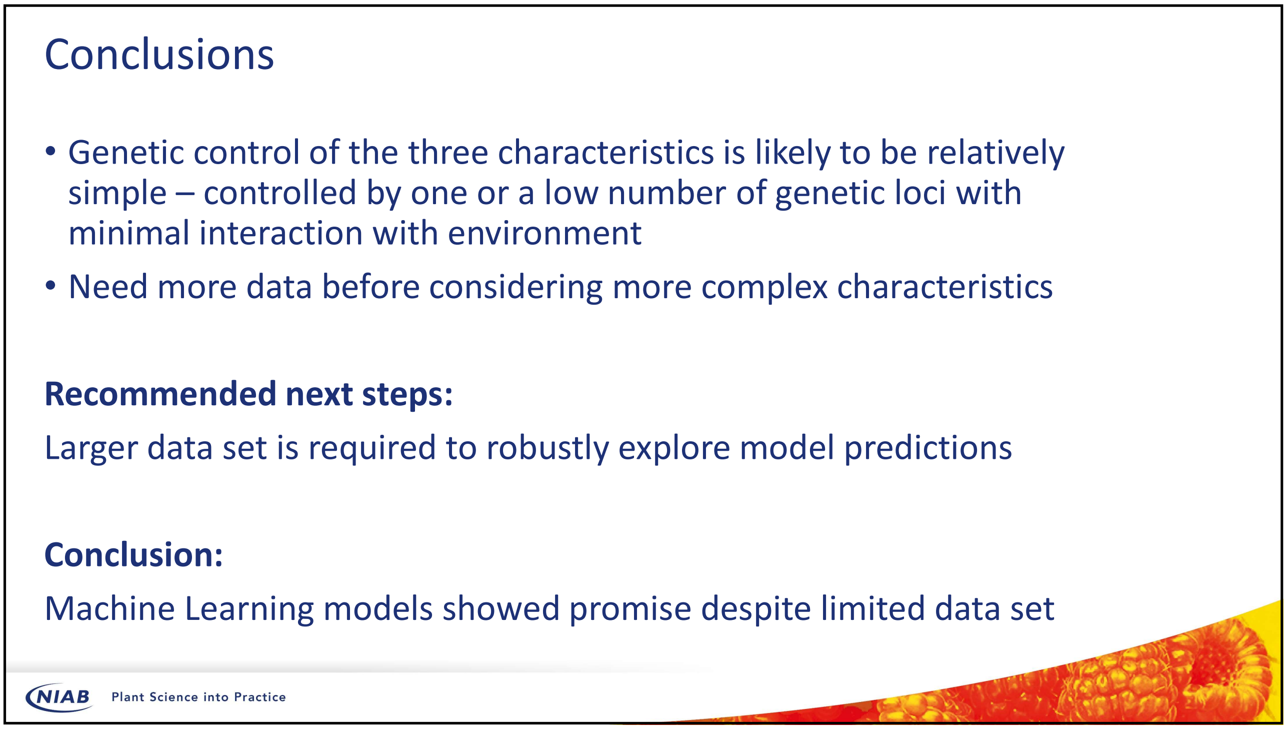
Outcomes of the model validation

- Presence of spines
  - 95% of predictions aligned with known phenotype
  - Error in both directions
- Fruiting habit
  - 67% of predictions aligned with known phenotype
  - Error in both directions
- Anthocyanin
  - 67% of predictions aligned with known phenotype
  - Error in both directions

One variety was incorrectly predicted across the three characters

NIAB Plant Science into Practice

10



Conclusions

- Genetic control of the three characteristics is likely to be relatively simple – controlled by one or a low number of genetic loci with minimal interaction with environment
- Need more data before considering more complex characteristics

Recommended next steps:

Larger data set is required to robustly explore model predictions

Conclusion:

Machine Learning models showed promise despite limited data set

NIAB Plant Science into Practice

11

Acknowledgements

Niab Researchers

James Cockram
Oghenejokpeme Orhobor
Lawrence Percival-Alwyn
Felicidad Fernández
Andrea Gutiérrez

Funded by the Defra Research, Development and Evidence (RDE) Framework

Department for Environment, Food & Rural Affairs

Provision of plant material

- Nikki Jennings (James Hutton Limited)
- Louise Gamble (James Hutton Institute)
- Mr Stewart Waine

NIAB Plant Science into Practice

12

NIAB

niab.com

niab @niabgroup niab.uk niab_UK

13

[End of Annex and of document