International Union for the Protection of New Varieties of Plants

E

Technical Working Party on Testing Methods and Techniques

TWM/2/5 Add.

Second Session
Virtual meeting, April 8 to 11, 2024

Original: English
Date: April 8, 2024

ADDENDUM TO: UNIFORMITY ASSESSMENT USING MOLECULAR MARKERS

Document prepared by experts from the United Kingdom

Disclaimer: this document does not represent UPOV policies or guidance

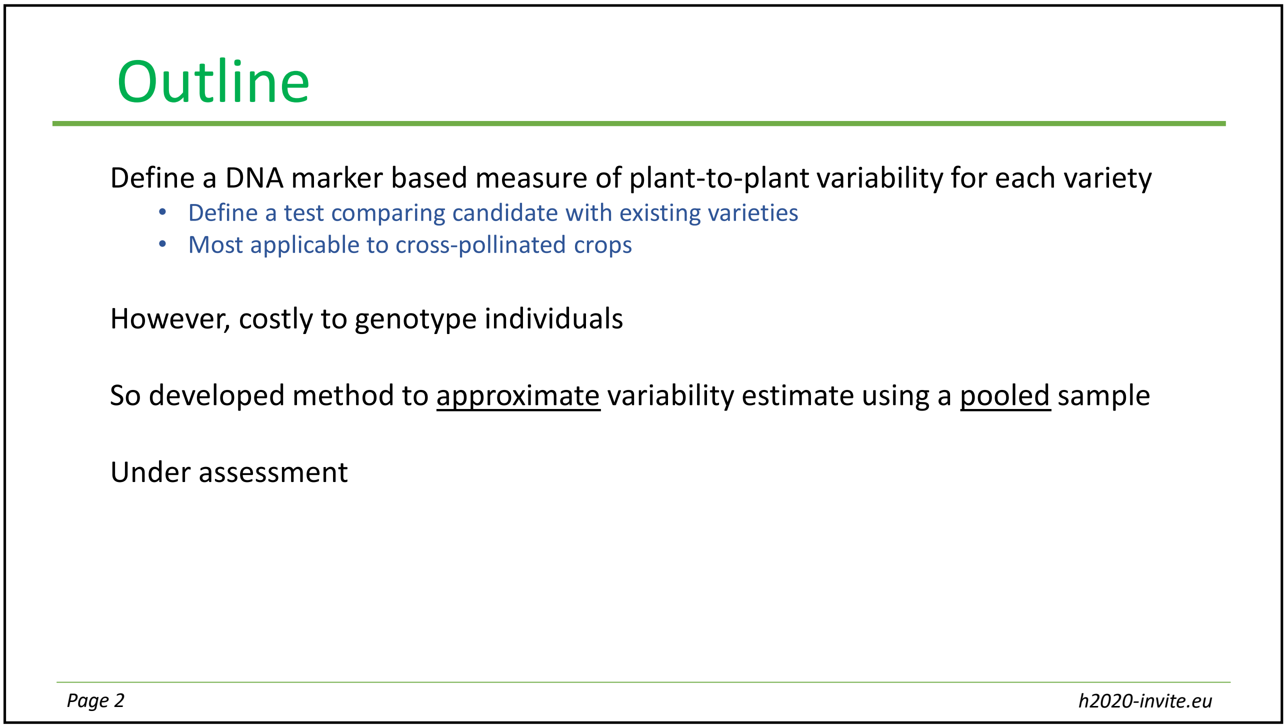
The annex to this document contains a copy of a presentation “Uniformity assessment using molecular markers”, made by an expert from the United Kingdom, at the second session of the Technical Working Party on Testing Methods and Techniques (TWM).

[Annex follows]

Uniformity Assessment using Molecular Markers (TWM/2/5)

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No 817970.

1



Outline

Define a DNA marker based measure of plant-to-plant variability for each variety

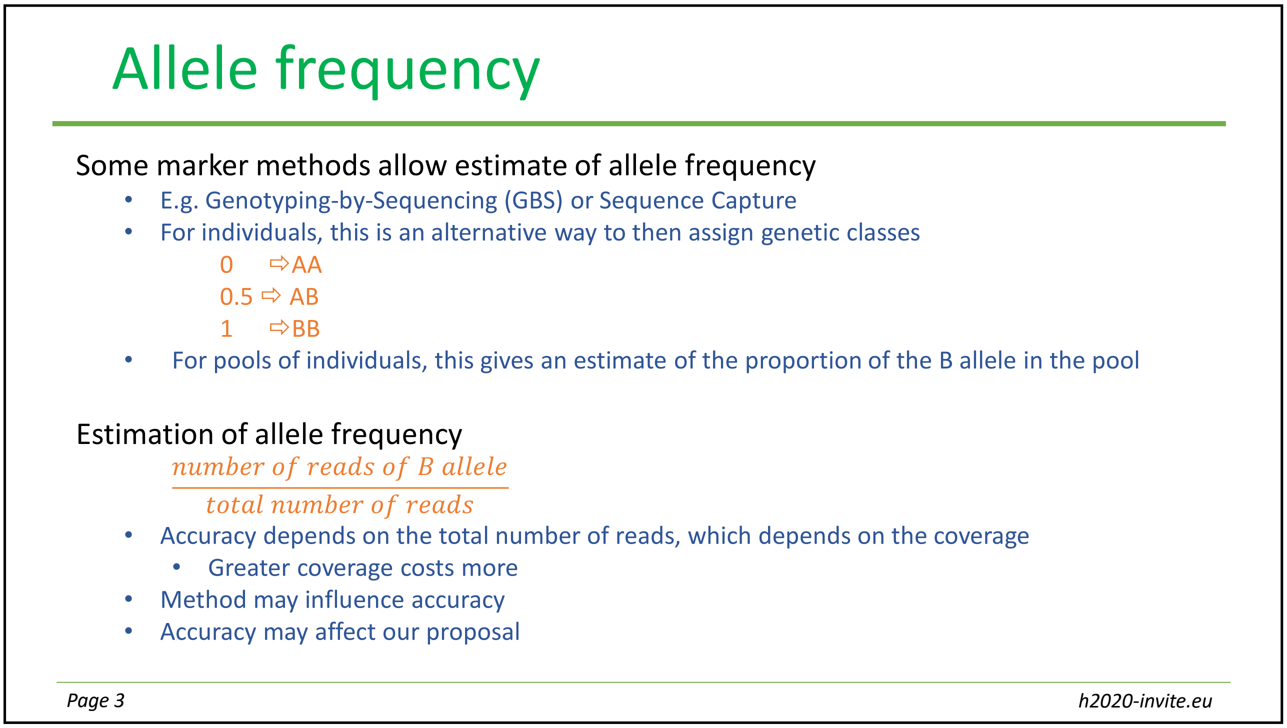
- Define a test comparing candidate with existing varieties
- Most applicable to cross-pollinated crops

However, costly to genotype individuals

So developed method to approximate variability estimate using a pooled sample

Under assessment

2



Allele frequency

Some marker methods allow estimate of allele frequency

- E.g. Genotyping-by-Sequencing (GBS) or Sequence Capture
- For individuals, this is an alternative way to then assign genetic classes

  0 ⇨AA
  0.5 ⇨ AB
  1 ⇨BB
- For pools of individuals, this gives an estimate of the proportion of the B allele in the pool

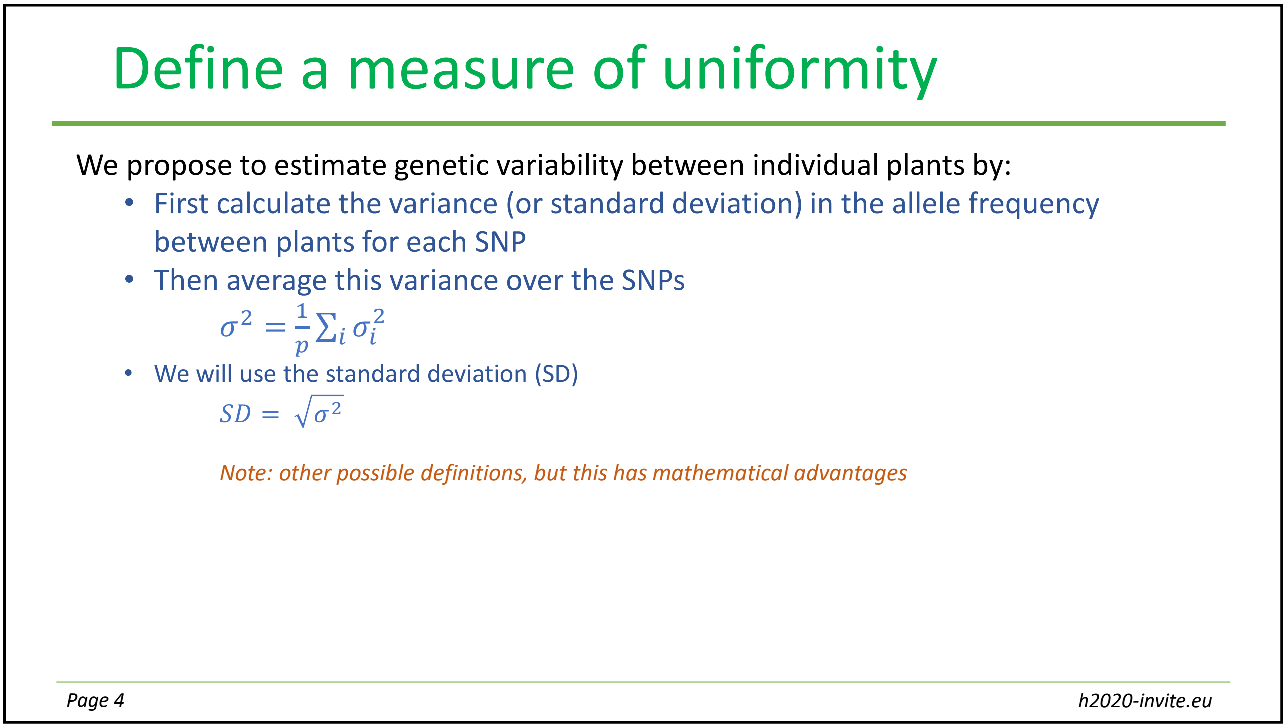
Estimation of allele frequency

$$\frac{number\ of\ reads\ of\ B\ allele}{total\ number\ of\ reads}$$

- Accuracy depends on the total number of reads, which depends on the coverage
  - Greater coverage costs more
- Method may influence accuracy
- Accuracy may affect our proposal

Page 3 *h2020-invite.eu*

3



Define a measure of uniformity

We propose to estimate genetic variability between individual plants by:

- First calculate the variance (or standard deviation) in the allele frequency between plants for each SNP
- Then average this variance over the SNPs

$$\sigma^2 = \frac{1}{p}\sum_i \sigma_i^2$$

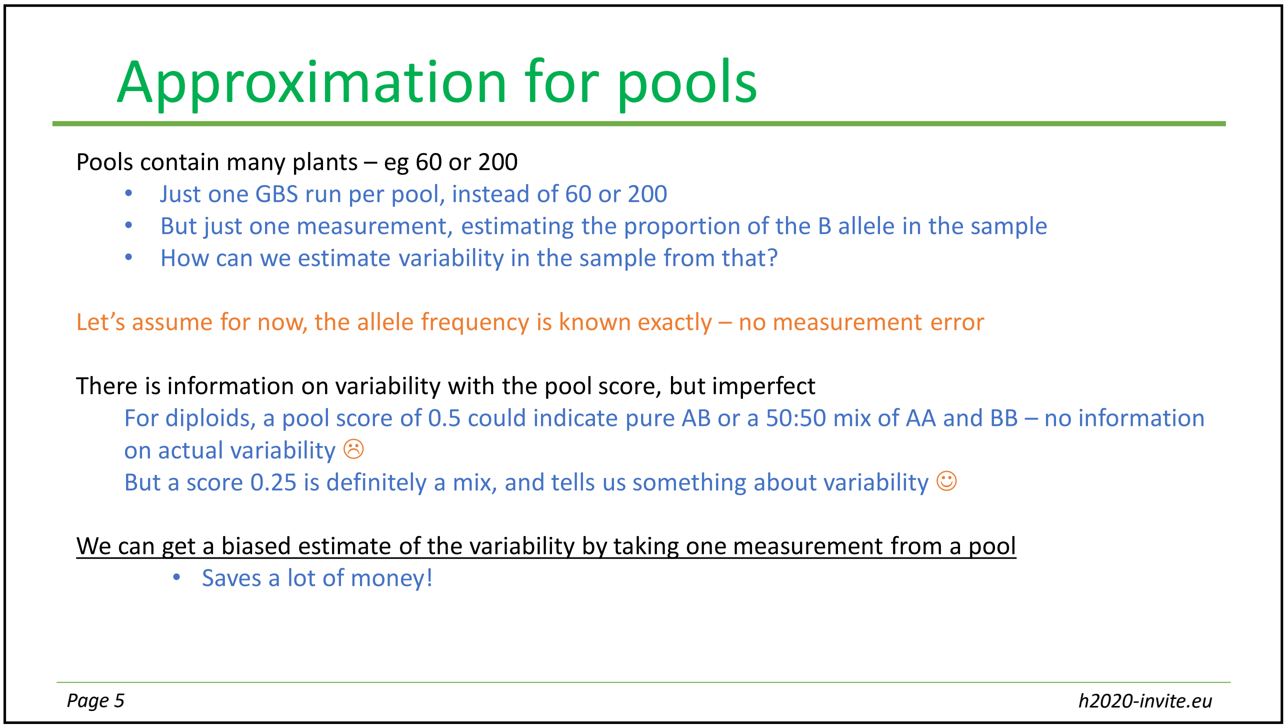
- We will use the standard deviation (SD)

$$SD = \sqrt{\sigma^2}$$

Note: other possible definitions, but this has mathematical advantages

Page 4 *h2020-invite.eu*

4



Approximation for pools

Pools contain many plants – eg 60 or 200

- Just one GBS run per pool, instead of 60 or 200
- But just one measurement, estimating the proportion of the B allele in the sample
- How can we estimate variability in the sample from that?

Let's assume for now, the allele frequency is known exactly – no measurement error

There is information on variability with the pool score, but imperfect

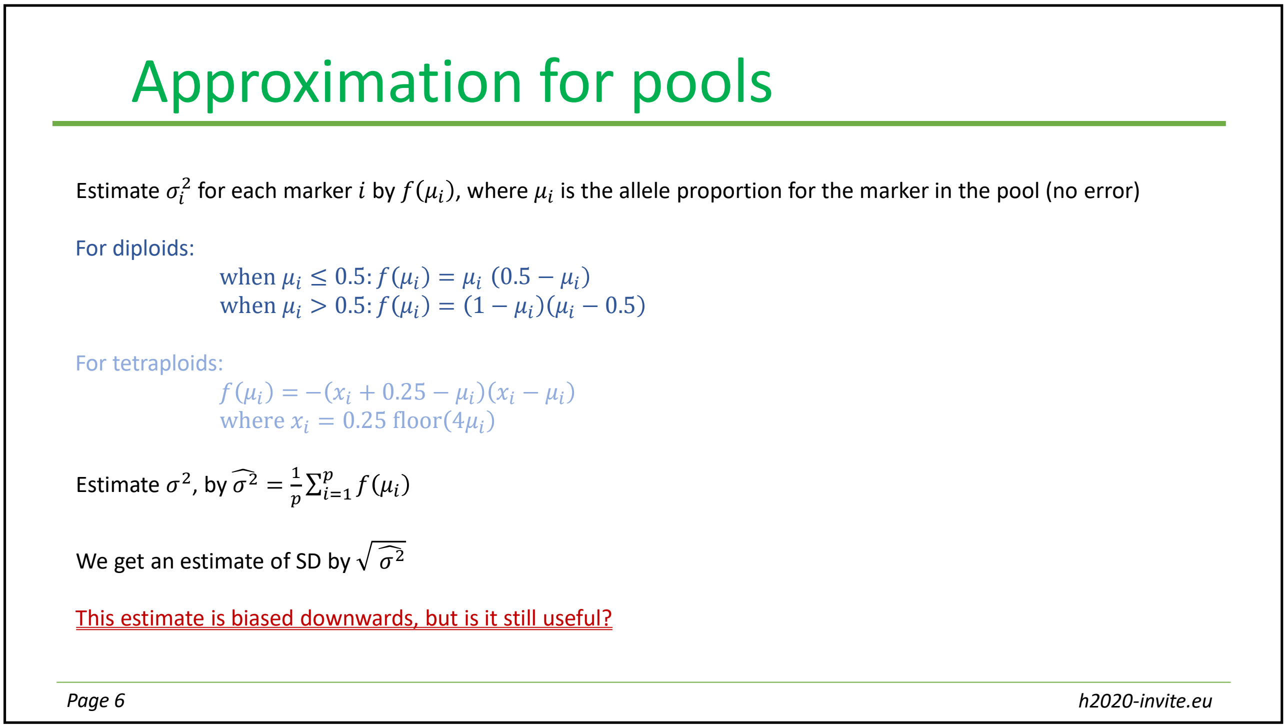
For diploids, a pool score of 0.5 could indicate pure AB or a 50:50 mix of AA and BB – no information on actual variability ☹

But a score 0.25 is definitely a mix, and tells us something about variability ☺

We can get a biased estimate of the variability by taking one measurement from a pool

- Saves a lot of money!

5



Approximation for pools

Estimate σ_i^2 for each marker i by $f(\mu_i)$, where μ_i is the allele proportion for the marker in the pool (no error)

For diploids:

$$\text{when } \mu_i \leq 0.5\text{: } f(\mu_i) = \mu_i\ (0.5 - \mu_i)$$
$$\text{when } \mu_i > 0.5\text{: } f(\mu_i) = (1 - \mu_i)(\mu_i - 0.5)$$

For tetraploids:

$$f(\mu_i) = -(x_i + 0.25 - \mu_i)(x_i - \mu_i)$$
$$\text{where } x_i = 0.25\ \text{floor}(4\mu_i)$$

Estimate σ^2, by $\widehat{\sigma^2} = \frac{1}{p}\sum_{i=1}^{p} f(\mu_i)$

We get an estimate of SD by $\sqrt{\widehat{\sigma^2}}$

This estimate is biased downwards, but is it still useful?

6

Assessment with example data

A. Simulated pools for 30 populations, with no measurement or sampling error

Supplied by Teagasc:

Arojju et al. BMC Genetics (2018)
https://doi.org/10.1186/s12863-018-0613-z
Byrne et al. Scientific Reports (2017) https://www.nature.com/articles/s41598-017-03232-8

30 diploid families of perennial ryegrass (including 10 synthetic cultivars)

~60 plants from each genotyped individually

Use this to simulate pools (without error) then compare actual variance vs estimate

B. Simulation of contamination events from example A data

“Add” 1 or 5 plants from another variety

Observe effect on variability

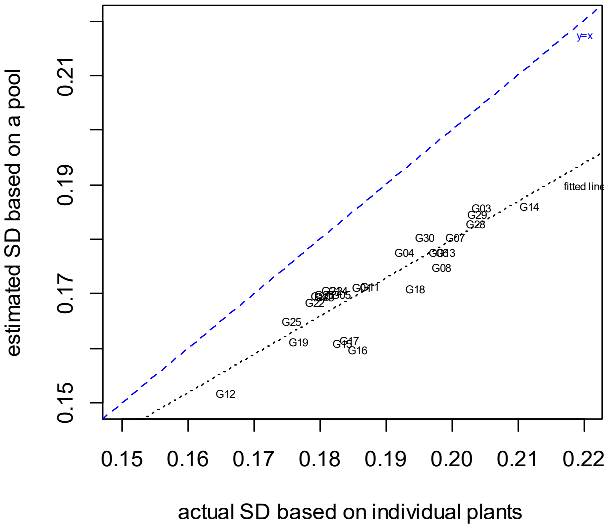
C. 4 varieties with actual pools, includes measurement and sampling error

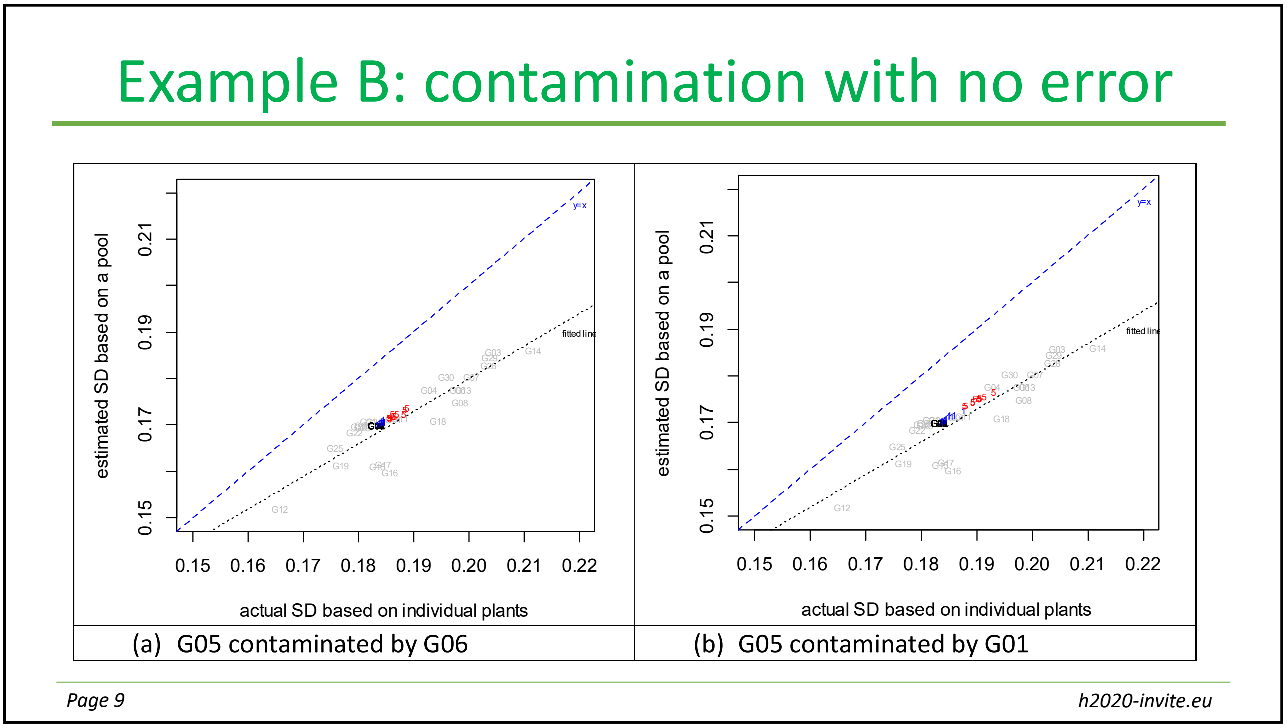
Supplied by INRAE as part of INVITE

4 perennial ryegrass varieties, measured individually and in pools, using sequence capture

~60 plants per pool, with replicates

Example A: no error

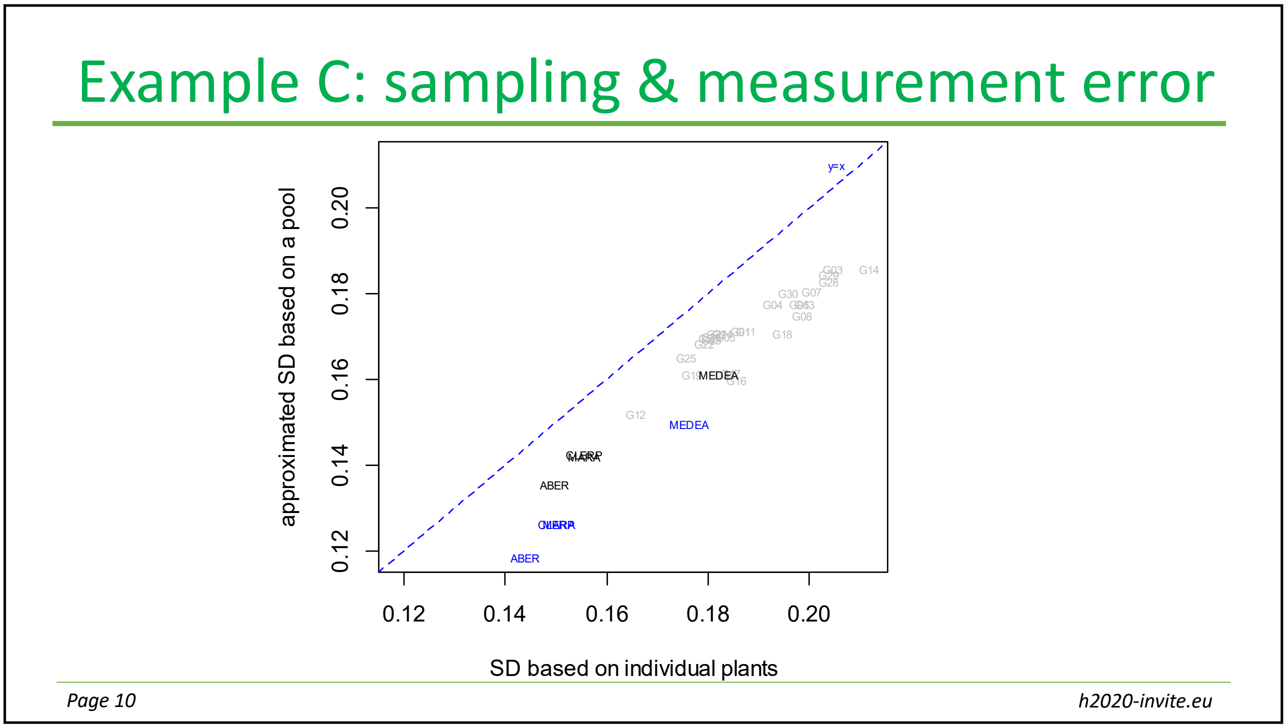




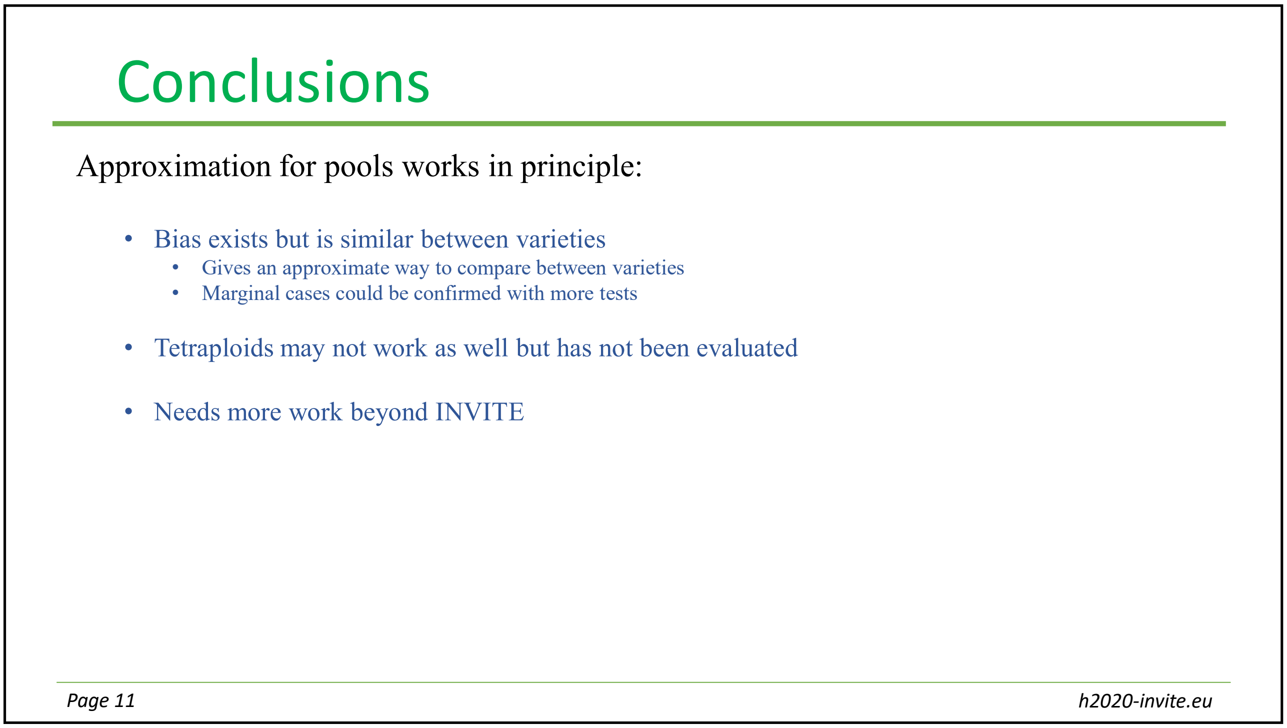
Example B: contamination with no error

(a) G05 contaminated by G06

(b) G05 contaminated by G01



Example C: sampling & measurement error



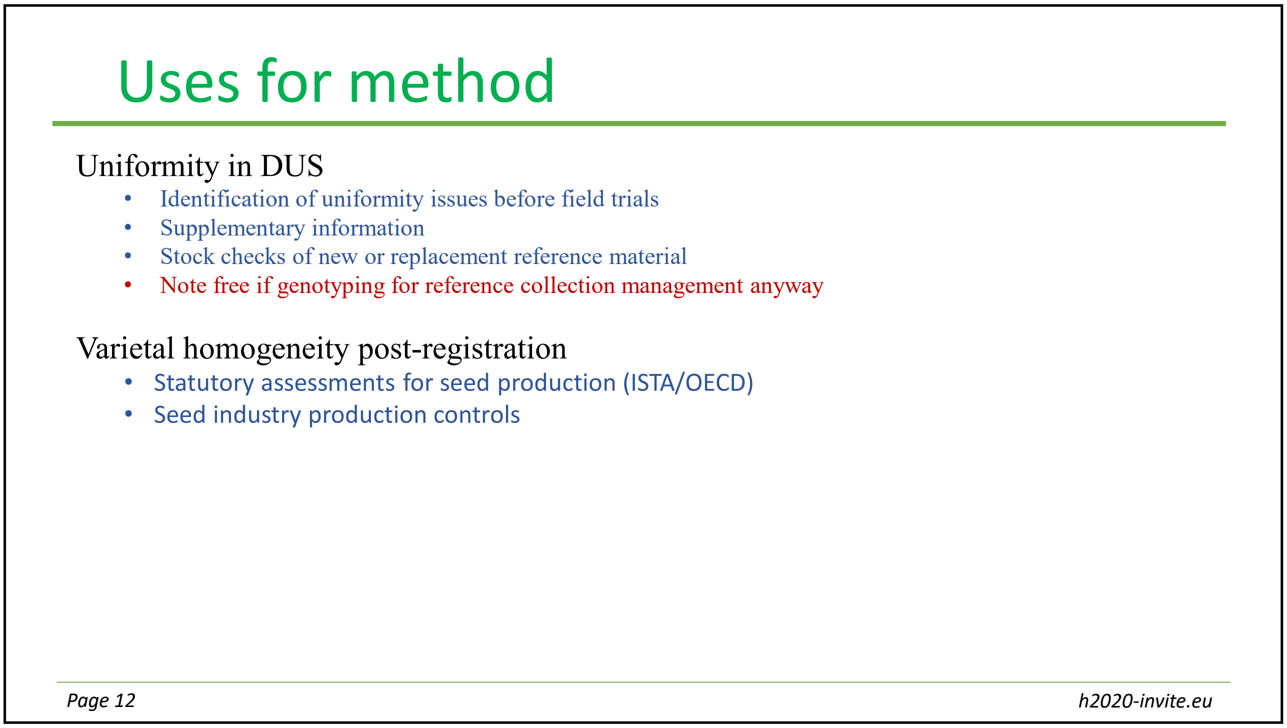
Conclusions

Approximation for pools works in principle:

- Bias exists but is similar between varieties
  - Gives an approximate way to compare between varieties
  - Marginal cases could be confirmed with more tests
- Tetraploids may not work as well but has not been evaluated
- Needs more work beyond INVITE

Page 11 *h2020-invite.eu*

11



Uses for method

Uniformity in DUS

- Identification of uniformity issues before field trials
- Supplementary information
- Stock checks of new or replacement reference material
- Note free if genotyping for reference collection management anyway

Varietal homogeneity post-registration

- Statutory assessments for seed production (ISTA/OECD)
- Seed industry production controls

Page 12 *h2020-invite.eu*

12

invite

Stay informed:

Website: www.h2020-invite.eu
Email: a.roberts@bioss.ac.uk

13

[End of Annex and of document]