



Assessing Varietal Heterogeneity in Pooled Samples of Cross-Pollinated Crops Using Genetic Markers

Adrian M I Roberts, Philippe Barre, Sabrina Delaunay, Bernadette Julier, Stephen Byrne, Dan Milbourne

► To cite this version:

Adrian M I Roberts, Philippe Barre, Sabrina Delaunay, Bernadette Julier, Stephen Byrne, et al.. Assessing Varietal Heterogeneity in Pooled Samples of Cross-Pollinated Crops Using Genetic Markers. Grass and Forage Science, 2025, 80, pp.e70009. <10.1111/gfs.70009>. <hal-05261604>

HAL Id: hal-05261604

<https://hal.inrae.fr/hal-05261604v1>

Submitted on 15 Sep 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY 4.0 - Attribution - International License

ORIGINAL ARTICLE OPEN ACCESS

Assessing Varietal Heterogeneity in Pooled Samples of Cross-Pollinated Crops Using Genetic Markers

Adrian M. I. Roberts¹  | Philippe Barre²  | Sabrina Delaunay²  | Bernadette Julier²  | Stephen Byrne³  | Dan Milbourn³ 

¹Biomathematics and Statistics Scotland (BioSS), Edinburgh, UK | ²INRAE, URP3F, Lusignan, France | ³Teagasc, Carlow, Republic of Ireland

Correspondence: Adrian M. I. Roberts (adrian.roberts@bioss.ac.uk)

Received: 19 May 2025 | **Revised:** 3 July 2025 | **Accepted:** 14 July 2025

Funding: This work was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement no 817970. Adrian M.I. Roberts was also supported by funding from the Scottish Government's Rural and Environment Science and Analytical Services Division. HORIZON EUROPE Food, Bioeconomy, Natural Resources, Agriculture and Environment.

Keywords: cross-pollinated | genetic markers | pooled | purity | uniformity | variation

ABSTRACT

Under the regulatory framework of the International Union for the Protection of New Varieties of Plants (UPOV) for plant variety protection, a new variety must be assessed for uniformity, as well as distinctness and stability (DUS). For cross-pollinated crops such as perennial ryegrass (*Lolium perenne*) and lucerne (*Medicago sativa*), where cultivars are heterogeneous populations, assessment of uniformity poses particular challenges. These challenges stem from the fact that such evaluations can be costly and require substantial labour and time, as they typically involve the measurement of numerous comparator characteristics across multiple individual plants representative of a candidate cultivar, often over a period exceeding 1 year. Although DUS testing is based on phenotypic characteristics, genetic markers are currently permitted for use in planning of trials, and for distinctness assessment in particular situations. However, genetic markers have not yet been employed for distinctness assessment. Similarly, under OECD Seed Schemes, marketed seed is assessed for varietal purity through the observation of plants in field plots. This study proposes a new measure of genetic heterogeneity based on genetic markers, along with a method to estimate this in pooled samples to reduce costs. The method relies on a system for estimating allele frequencies at SNPs, such as genotyping-by-sequencing or sequence capture. We demonstrate the potential of this approach with example data for perennial ryegrass. The approximation based on pooled samples provided satisfactory estimates of genetic heterogeneity compared with direct assessment based on individual plants, and proved to be sensitive to contamination of samples by other varieties.

1 | Introduction

Commercial varieties of cross-pollinated agricultural crops such as perennial ryegrass (*Lolium perenne*) and lucerne (*Medicago sativa*) are synthetic cultivars, sometimes referred to as “population varieties” (Posselt 2010). This process exists because of the necessity of producing large amounts of true botanical seed in plant species with low multiplication rates combined with an obligate outcrossing reproductive mode, frequently driven by genetically encoded systems for self-incompatibility in these

species. Seed for these varieties is produced in open-pollinated nurseries, where measures such as containment, physical barriers, and distance are used to ensure that cross pollination only occurs exclusively between members of the same varietal source. The original parental individuals of the variety can vary widely in number depending on the crop and the objectives of the breeder. However, these individuals are seldom retained, as their maintenance typically requires repeated rounds of clonal propagation; instead, only the seed source material is retained as a heterogeneous bulk from the first generation of multiplication.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2025 The Author(s). *Grass and Forage Science* published by John Wiley & Sons Ltd.

Considering the definition of a variety as “a more precisely defined group of plants, selected from within a species, with a common set of characteristics” (UPOV FAQs (A)), synthetic cultivars present a challenge for both variety registration and subsequent regulatory and commercial quality control. This is due to the fact that they exhibit a level of natural heterogeneity but have a requirement to exhibit this variation within a defined set of limits, and over multiple cycles of production. Such varieties are subject to regulatory and commercial controls to check that levels of heterogeneity are acceptable. In particular, many countries accede to the UPOV convention on plant variety protection. Under this system, new plant varieties are assessed for Distinctness, Uniformity and Stability (DUS). Here, we focus on the uniformity component of DUS assessment. In cross-pollinated species, uniformity is assessed by comparing the plant-to-plant variability in set phenotypic characteristics with the variability seen in existing varieties (UPOV 2002). These assessments can be expensive and both labour and time intensive, as they may be based on measurements of numerous comparator characteristics on multiple individual plants, generally over more than 1 year. For instance, for perennial ryegrass, in the European Union (CPVO 2019) and the United Kingdom (APHA 2022), candidate varieties are assessed for DUS in at least two, but usually three, years of trials. Trials have at least three replicates, with 60 plants in total per trial for each variety. At least 22 characteristics are evaluated, many by measurements on individual plants. For such quantitative characteristics, uniformity is assessed using the Combined-Over-Years Uniformity Criterion (COYU), which is based on a comparison of the within-plot standard deviations, adjusted for any relationship with expression level, between the candidate and existing varieties (UPOV 2022).

The potential for using genetic markers for plant variety identification is well established. Whilst the phenotypic characteristics are key to DUS assessments under UPOV principles, genetic markers may be used to aid the process (UPOV FAQs (B)). To date, genetic markers have mainly been utilised for trial management, by selecting either similar varieties to plant next to the candidate, or to identify clearly distinct pairs of varieties to reduce the trial size. Two application models have been identified and are in current use in some crops: “Characteristic-specific molecular markers” (UPOV Application Model 1), requiring a tight linkage between trait and markers; and “Combining Phenotypic and Molecular Distances in the Management of Variety Collections” (UPOV Application Model 2), currently based on the relationship between genetic and phenotypic distances (UPOV 2020). Less attention has been given in the UPOV framework to the use of molecular markers to support decisions on uniformity. Despite this, there have been proposals to utilise characteristic specific markers (coherent with distinctness testing in UPOV model1) for uniformity analysis. For example, in barley, Saccomano et al. (2020) demonstrated the technical feasibility of using diagnostic KASP markers designed to DUS traits to identify “off-types” in seed batches, meaning that genetic markers targeting distinctness could also contribute to supporting decisions on either uniformity (in DUS) or seed purity. Yang et al. (2021) proposed that markers could be used to assess uniformity, based on distances between individuals or pools of individuals, by comparing distances. Wang et al. (2016) suggested that research is needed to establish a method for uniformity assessment for ryegrass species based on genetic markers.

After registration of a variety, many countries apply processes to ensure the quality of seed sold, following procedures set out in the Organisation for Economic Co-operation and Development (OECD) Seed Schemes. Varietal purity constitutes a critical element of the certification process. Through the use of control plots, the number of plants that do not conform to the specified variety is assessed and compared with fixed standards. Whereas for most species, the standard is based on the proportion of off-types, in some species such as perennial ryegrass the assessment is based on the number of off-types in a set area. This approach is adopted “because determining the number of plants per unit area is very difficult, if not impossible, for these species” (OECD 2023). Similar to the DUS criteria within UPOV, under the OECD Seed Schemes, genetic markers are seen as supplementary tool, and the certification status of a seed lot cannot be established solely on the basis of genetic markers (UPOV 2024). At the current time, genetic markers can only be applied for variety identification, but their use for variety purity is being considered.

Similar to the issues around phenotypic uniformity, using molecular markers in synthetic varieties is not straightforward due to their heterogeneity. Taking the aforementioned example of DUS testing in perennial ryegrass, one approach might be to genotype a similar number of individual representatives of the variety as applied to the morphological testing. However, such an approach is likely to be prohibitively expensive. An alternative approach is to pool a similar number of individual representatives of the variety and genotype the pool of individuals to determine allele frequencies. Previous work has demonstrated that genotyping by sequencing can be deployed to calculate genome-wide allele frequency profiles from such pools of individuals representing a cultivar (Byrne et al. 2013; Julier et al. 2018; Verwimp et al. 2018). Furthermore, studies have shown the potential of using these allele frequencies in genetic association analysis and genomic prediction (Fè et al. 2015, 2016; Cericola et al. 2018; Guo et al. 2018; Pembleton et al. 2018; Blanco-Pastor et al. 2021; Keep et al. 2020, 2021; Konkolewska et al. 2023; Pégard et al. 2023). Therefore, it is feasible that such allele frequency data could also be exploited to assess uniformity as part of DUS testing, for applications in assessment of variety purity, both as part of official tests and for internal quality controls for seed producers.

In this paper, a method is proposed to exploit allele frequency estimates from genotyping-by-sequencing data for assessing genetic uniformity and varietal purity. It is demonstrated with example data sets for perennial ryegrass.

2 | Materials and Methods

2.1 | Approach

First a definition of genetic variance (or heterogeneity) from plant to plant is proposed. There are a number of alternatives, but this one has mathematical advantages for later calculations. It is designed for marker systems which estimate the allele frequency.

Consider first diploids. For a single nucleotide polymorphism (SNP) marker on an individual plant, then an allele frequency

of 0 corresponds to the homozygote with the major allele (AA), 0.5 to the heterozygote (AB), and 1 to the homozygote with the minor allele (BB). For a sample of plants from a variety, genotyped individually, there would be a mixture of allele frequencies 0, 0.5 and 1 seen at the SNP. The variability of the genotypes for a particular marker can be expressed through the variance of these allele frequencies, and the overall genetic variability can be defined as the mean of these variances over all the markers. So, the genetic variance, σ^2 , is given by

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

where p is the number of markers, and σ_i^2 is the sampling variance of the allele frequency over individual plants for marker i , given by $\frac{1}{n-1} \sum_j (f_{ij} - \bar{f}_i)^2$, where f_{ij} is the allele frequency for marker i in plant j , $\bar{f}_i$ is the mean allele frequency for the marker, and there are n plants.

Genotyping individual plants is more expensive than genotyping a pooled sample from the variety. With methods such as Genotyping-by-Sequencing (GBS), or Sequence Capture, it is possible to estimate the mean allelic frequency in a pooled sample. Of course, with a pooled sample, information about the variability of individuals is partly lost. However, substantial information is retained, and we claim that this is sufficient for assessment of genetic variability in pools. For example, a mean allele frequency of 0 or 1 for a marker implies that the sample consists of pure AA or BB genotypes respectively, representing a variance of 0. However, a mean allele frequency of 0.5 is completely ambiguous; it could arise from pure AB (variance 0), a 50:50 balance of AA and BB (variance approximately 0.25), or a mix of pure AB with a balance of AA and BB (variance between 0 and 0.25). A mean allele frequency of 0.25 can only arise from a mixture, for example, AA and AB (variance ~ 0.06), and so the variance must be positive. As some information on variability is retained, it is possible to find an approximation, $\hat{\sigma}^2$, for the genetic variance, σ^2 , based on the mean allele frequencies for each marker, μ_i :

$$\hat{\sigma}^2 = \frac{1}{p} \sum_i f(\mu_i) \quad (2)$$

where for diploids, the function $f(\mu_i)$ is given by:

$$\text{for } \mu_i \leq \frac{1}{2}: f(\mu_i) = \mu_i \left(\frac{1}{2} - \mu_i \right)$$

$$\text{for } \mu_i > \frac{1}{2}: f(\mu_i) = \left(\mu_i - \frac{1}{2} \right) (1 - \mu_i)$$

It can be shown that $\hat{\sigma}^2$ is an underestimate of σ^2 , with the bias depending on the frequency of the less common homozygote in the sample:

$$\hat{\sigma}^2 = \sigma^2 - \frac{1}{2p} \sum_i \min(p_{AA,i}, p_{BB,i}) \quad (3)$$

where $p_{AA,i}$ and $p_{BB,i}$ are the (unknown) frequencies of the AA and BB homozygotes in the sample for marker i . A mathematical justification for this approximation and of the bias is given in the Appendix A.

The quantity $\min(p_{AA,i}, p_{BB,i})$ represents the proportion of the least common homozygote in the sample for marker i . It is expected that this value should be low in practice, and thus the estimate $\hat{\sigma}^2$ should be close to σ^2 . For example, if the allele frequency was 0.25 at a marker, it might be expected that most of the individuals in the pool would be AA and AB, with very few BB.

The quality of the approximation will depend on the frequency of AB genotypes; the approximation is less effective when this is high since $\min(p_{AA,i}, p_{BB,i})$ could be relatively large. The overall effectiveness is evaluated in this report using real data sets, as simulation would require assumptions about typical genotypic make-up within populations.

The approximation can be extended to other ploidies, though it can be more complicated. For autotetraploids, $f(\mu_i)$ is given by:

$$f(\mu_i) = -\left(x_i + \frac{1}{4} - \mu_i\right)(x_i - \mu_i) \quad (4)$$

where $x_i = \frac{1}{4} \text{floor}(4\mu_i)$, and floor indicates rounding down to the nearest integer.

The formulation of the bias for tetraploids is more complex than for diploids, but again depends on the frequencies of less likely genotypes given the mean allele frequency. For example, when μ_i is between 0 and $\frac{1}{8}$, then

$$\sigma_i^2 - f(\mu_i) = \frac{1}{8}p_{AABB,i} + \frac{3}{8}p_{ABBB,i} + \frac{6}{8}p_{BBBB,i} \quad (5)$$

The estimate of genetic variance will be affected by various sources of error, as well as the bias when using the approximation. These include:

- Sampling variation due to sampling a finite number of plants.
- The number of markers.
- The read depth of the genotyping approach.
- Measurement error of the genotyping approach, beyond read depth. This can include bias around the homozygotes, and overall bias, as well as other forms of measurement error. Understanding of measurement error and bias for these methods is an emerging area.

The influence of these errors on the estimate might be expected to be greater for tetraploids than for diploids, as there are more genetic classes represented over the same scale range (zero to one).

2.2 | Example Data Sets for Evaluation of the Approach

To evaluate the approach, two data sets are used based on diploid ryegrass populations. Data set A allows evaluation of the bias without being concerned about the other sources of error. Data set B includes all sources of error in the comparison between the estimate of genetic variance based on individual plants and

the pooled estimate. In addition, data set A data is used to investigate the sensitivity of the method to contamination of one variety by another.

2.2.1 | Data Set A

This perennial ryegrass data set is described in Byrne et al. (2017). In outline, the data set covers up to 60 plants from each of 10 synthetic cultivars (G01-G10), eight full-sib families (G11-G18), eight half-sib families (G19-G26), and four ecotypes (G27-G30), all being diploid. The plants were genotyped individually using genotyping-by-sequencing after genome complexity reduction with the restriction enzyme ApeKI (Elshire et al. 2011). All genotyping-by-sequencing can be found in the NCBI sequence read archive (SRA) under the BioProject ID: PRJNA352789. The sequence data was aligned to a reference genome available at the time (Byrne et al. 2015), and genotypes were called with a minimum genotype quality score of 20. There were 63,159 SNPs after filtering. Each selected SNP had at least 50% non-missing observations over the samples. For this work, the data set was reduced further to samples having at least 10,000 SNPs, and populations having at least 40 qualifying samples (Table S1).

This data set did not include samples genotyped as pools. Instead, pooled samples were simulated by calculating mean allele frequencies. The means and variances for each SNP were weighted according to the number of samples with non-missing data in the population. The genetic variance was calculated by averaging these weighted variances over the SNPs. This was compared with the approximation for the genetic variance, calculated with formula (2), based on the mean of the weighted means over the SNPs.

2.2.2 | Simulating Variety Contamination Based on Data Set A

Contamination events were simulated by “adding” individuals from one population to another, based on the populations in Data set A. Firstly, one individual from the contaminating population was added at random to the complete set of individuals from the contaminated population, and then the variance calculations were repeated for the simulated pool. This was repeated, but adding five contaminant individuals (from the same contaminating population) to the set of individuals for the contaminated population, again at random. In each case, 10 separate random sets of contaminated samples were produced to evaluate variability in the impact of contamination. Two cases are shown: for the pair of populations with the highest correlation between mean SNP frequencies, G05 and G06, and for the pair with the lowest correlation, G05 and G01. In both cases, G05 was treated as the contaminated population.

2.2.3 | Data Set B

This perennial ryegrass data set was based on four commercial diploid varieties: Aberavon, Clerpin, Mara and Medea. The varieties were genotyped by sequence capture both in pools and for individuals.

For pools, 300 seeds per variety (Blanco-Pastor et al. 2019) were sown in a pot (9 × 9 cm) and leaves were collected as described in Byrne et al. (2013). Leaves were air dried in boxes with silica gel. DNA was extracted as described in Julier et al. (2018). Five replicates were extracted for each variety. For individuals, leaves from 60 plantlets per variety were sampled separately for DNA extraction as described above. Two replicates were extracted for each individual.

Based on previous information (Blanco-Pastor et al. 2019; Byrne personal communication), a list of 4881 SNPs was set up in order to cover the genome published by Byrne et al. 2015 and to target some QTLs (Table S2). MyBaits (Arbor Biosciences) was able to design and synthesise two probes of 80 bases for 4668 SNP. The two probes overlapped on 40 bases, with the SNP located in the middle of the overlap.

The DNA libraries were obtained with an adapted protocol of Mascher et al. (2013) used in Bataillon et al. (2022) on wild thyme. Briefly, the protocol included: (i) a preparation of DNA fragments (fragmentation + repair with NEB Next Ultra II FS DNA Module), ligation of P5/P7 adapters including barcodes with Invitrogen T4 DNA Ligase and amplification with NEB Q5 High-Fidelity 2× Master Mix, and (ii) a double capture with MyBaits probes defined above used with the sequence capture protocol of Roche Diagnostic SeqCap EZ Hybridisation Kit, Sequence Capture Developer Reagent (Roche diagnostic) and Invitrogen Dynabeads MyOne Streptavidin C1. The preparation of DNA fragments was performed on 480 samples (5 plates of 96 wells including the 4 varieties with for each 5 pools and about 60 individuals). The 96 samples of each plate were pooled before the double capture leading to 5 libraries. The double capture was performed with a first hybridisation at 47°C for 64 h and a second hybridisation at 47°C for 66 h. The 5 libraries were amplified with 2X KAPA HiFi Hot Start Ready Mix (Roche diagnostic) with primers including indexes i5 and i7. Eventually, the 5 libraries were pooled for sequencing.

Sequencing was performed by the GeT-PlaGe INRAE platform (<https://get.genotoul.fr/la-plateforme/get-plage/>) on one lane of Novaseq SP with a read length of 300 bases and in paired-end. A total of 189 million read pairs were obtained for the 480 samples, with an average per sample of 382,907 and ranging from 9453 to 942,876 (<https://doi.org/10.57745/ETSZVF>).

Sequence analysis was performed as follows: demultiplexing with Stacks: process radtags v. 2.66, assembling paired-ends with PEAR (Zhang et al. 2014), trimming with Cutadapt v4.2 on Python 3.11.2, aligning on the perennial ryegrass reference sequence from Byrne et al. (2015) with bwa mem v. 0.7.17-r1188 and allele counting for each position of each alignment with Bam-readcount v. 1.0.1-unstable-8-c7c76e6 (commit c7c76e6). Then a homemade pipeline was used to extract biallelic SNP with a global (on the 480 samples) allele frequency between 5% and 95% on positions with read depth higher than 10. For each of these SNP, the minor allele frequency was extracted for each sample. A total of 452 249 SNPs were identified on the 480 samples. We selected the 428 samples with less than 10 000 missing data. We obtained 54 142 SNPs with less than 100 missing data (<https://doi.org/10.57745/ETSZVF>) in which 8 543 presented no missing data.

As both individual plants and pools were genotyped, it was possible to make a direct comparison between the genetic variance based on individuals and the approximation based on the pools. In this case, the comparison includes measurement error from the sources noted above. For calculations of mean and variance, samples with more than 50% missing values over the SNPs were removed. Based on results for individual plants, it was observed that there was bias in the allele frequency estimates for some SNPs. This was monitored by trimming the scores to the interval 0.15 to 0.85. In principle, these cases should correspond to heterozygotes and thus average 0.5. However, the trimmed mean was somewhat below 0.5 for a number of SNPs. 39% of SNPs had trimmed means outside the interval 0.45 to 0.55 and were labelled as showing bias. Results were produced both based on all SNPs and excluding those SNPs labelled as being biased.

3 | Results

3.1 | Data Set A

In Figure 1, the approximated estimated genetic standard deviations based on the simulated pools are compared with the actual calculated standard deviations based on individual plants from Data set A. The bias was clearly evident, with the pool-based estimates consistently lying below the actual values. However, the approximate values sat fairly closely to a straight line, and the rank ordering of the populations was reasonably conserved. The 10 populations with the highest genetic variance included all three of the Ecotypes, five of the seven synthetic cultivars,

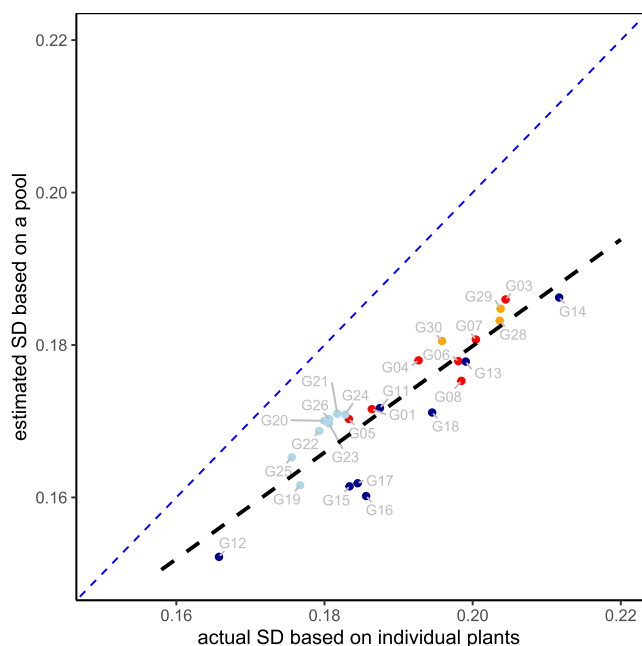


FIGURE 1 | Data set A: Estimated standard deviations for each population based on simulated pools compared with the actual standard deviations based on individual plants. The points are coloured by the type of population: Red for synthetic cultivar; dark blue for full-sib families; light blue for half-sib families; and orange for ecotypes. The blue dashed line marks the line of equality between the approximation and the actual values. The black dashed line is a fitted linear regression.

and two Full-sib families (of eight). The population with the lowest variance was G12, a Full-sib family.

3.2 | Contaminated Simulation Based on Data Set A

Figure 2 shows the simulated contamination events based on Data set A. The contaminated samples were more variable than the pure populations, and the larger degree of contamination was reflected in higher variability. The variability depended on the individual from the contaminating population—with some scatter in standard deviations over the different random selections. The contamination of G05 by G01 (Figure 2B) had a greater impact than by G06 (Figure 2A), with the contaminated pools having higher genetic variability, reflecting the closer genetic profiles between G05 and G06. The approximate approach based on pools was well correlated with the exact calculation.

3.3 | Data Set B

The comparison of allele frequencies evaluated on one pool and on 42–46 individuals on the 8543 SNP without missing data showed good correlations with a mean R^2 of 0.89 (0.85–0.93) over the four varieties, a mean slope of 0.91 (0.87–0.95), and a mean 95% prediction interval width of 0.20 (0.16–0.25) (Figure 3). The correlations between individuals and pools were improved when using the average of the five replicates of pools with R^2 of 0.96, slope of 0.98, and mean prediction interval width of 0.12.

Figure 4 shows the comparison of the pool approximation versus direct calculation based on individual plants for Data set B. The level of variability was higher when markers showing bias were included, and these results arguably appear more in line with those of Data set A. Although the inclusion of markers with bias did impact the results, the relative positions of the varieties were maintained. So it may be that this form of bias has little effect on decision-making as long as the same marker set is used. The four varieties had lower genetic variance estimates than most of the populations in Data set A. It was not clear whether this was due to the selection of varieties, the difference in SNPs between the two data sets, or the difference in marker technology. The variety Aberavon had the lowest genetic variance; this variety is known to have a narrow genetic basis (Auzanneau et al. 2007).

4 | Discussion

An approach for evaluating genetic heterogeneity has been proposed, along with a method for utilising pooled samples to minimise costs. Based on the results from the perennial ryegrass example data sets, the variance approximation for pooled samples demonstrates potential for use in practice.

This method was developed for use in uniformity assessment as part of the DUS testing process. It should be noted that under UPOV rules, uniformity assessment must be based on phenotypic characteristics, which are also utilised for DUS. Instead, it

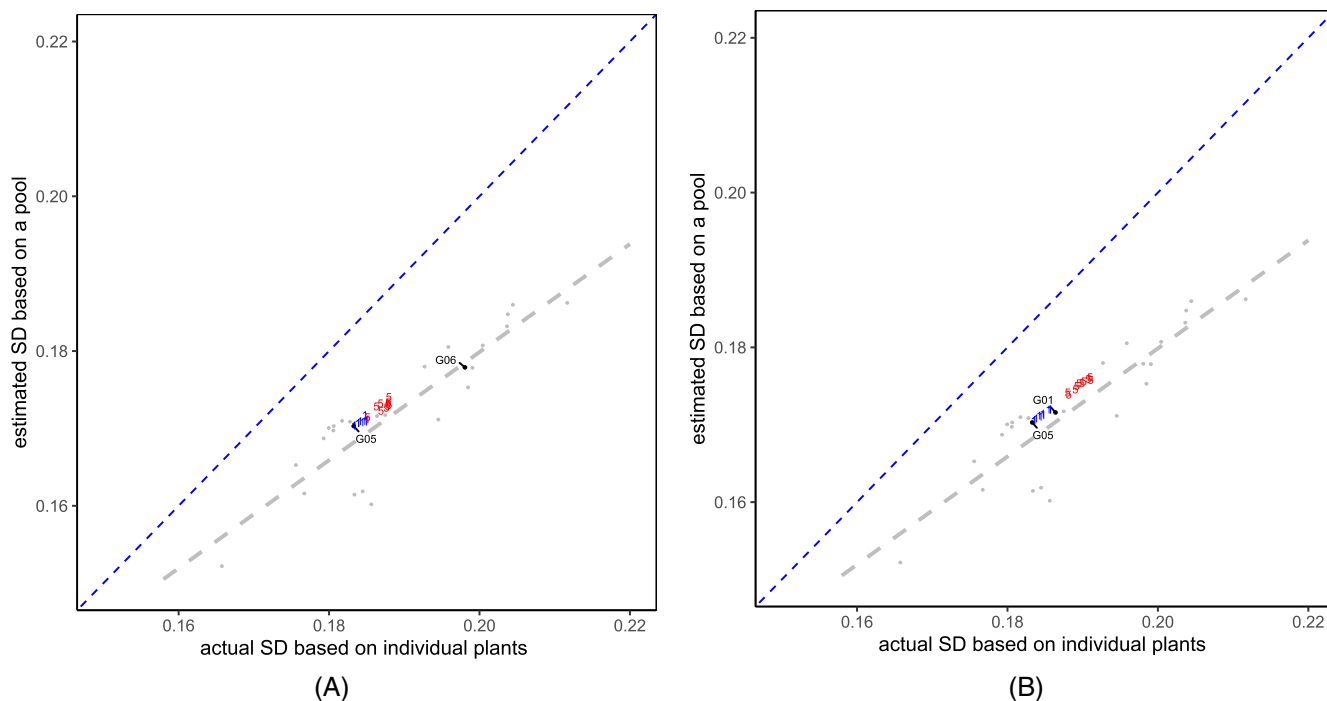


FIGURE 2 | Contamination simulation: Estimated standard deviations based on simulated pools compared with the actual standard deviations based on individual plants. The original uncontaminated cultivar and the contaminating cultivar are highlighted in black. The other populations for Data set A are in grey for reference. Contamination events by 1 added contaminant plant are indicated by a blue “1”, and by 5 contaminant plants by a red “5”. (a) is where G05 is contaminated by G06, a relatively similar cultivar, (b) is where G05 is contaminated by G01, a less similar cultivar.

is proposed that this method be used to complement uniformity decisions in two ways:

- By identifying candidates with clear uniformity issues very early in the DUS process;
- As supplementary information for borderline cases.

Comparison with existing varieties could be carried out using a similar test as the Combined-Over-Years Uniformity Criterion (UPOV 2002, part II section 9), but without a need to adjust for the level of expression. An upper tolerance for variability could be calculated using a one-sided Student *t*-test, based on the mean variability (M_R) for the reference varieties and the variance of the variabilities of the reference varieties (V_R):

$$\text{upper tolerance} = M_R + t_{p,R-1} \sqrt{V_R \left(\frac{1}{R} + 1 \right)}$$

where R is the number of reference varieties and $t_{p,R-1}$ is the one-tailed *t*-value for probability p with $R-1$ degrees of freedom. In COYU, the logarithm of the standard deviations between plants is used here, but alternative transformations could be considered. The variance approximation for pooled samples should allow clear and correct comparisons in many cases. In borderline cases, the error induced by the approximation requires consideration. Depending on the context, it may be that further testing is required based on individual samples or smaller pools, or alternatively, the candidate sample is given the benefit of doubt. A benefit of using genetic markers rather than phenotypic data is that comparisons can easily be made with all screened varieties. Indeed, there is the possibility of

fixing a threshold for a given species and type, combined with the marker system, since this form of assessment is independent of the effect of environment.

As well as uniformity assessment as part of DUS, genetic heterogeneity could be used for assessing varietal purity post-registration, both for statutory and industry checks of varietal purity for seed production. The contamination example shows that the method has potential in estimating the degree of variety purity. It is also clear that it is sensitive to the closeness of the relationship between the variety and contaminant. Although this approach was developed specifically for cross-pollinated crops where varieties are populations, it may have potential in other crops as well. If it is to be used on pools, it does depend on marker technology for estimating allele frequency. In practice, a test statistic might be based on the ratio of the genetic variance for the sample compared with a pure sample of the variety in question. A process of calibration would be needed to establish a relevant threshold that relates to the OECD standard (OECD 2023).

The assessment of the impact of the sources of error associated with the allele frequency estimate requires further investigation. Data set B did indicate that these errors may not be a substantial problem in practice, but the data set only included four varieties. The impact of one form of bias was investigated. Although the bias in allele frequencies was rather common over the SNP set studied, the initial finding is that this may not matter in practice. The number and distribution of markers is undoubtedly important, and this should be investigated in the contexts of the method's different uses. It seems likely, though, that an acceptable level of precision can be obtained with a much lower number of SNPs (a few thousand) than

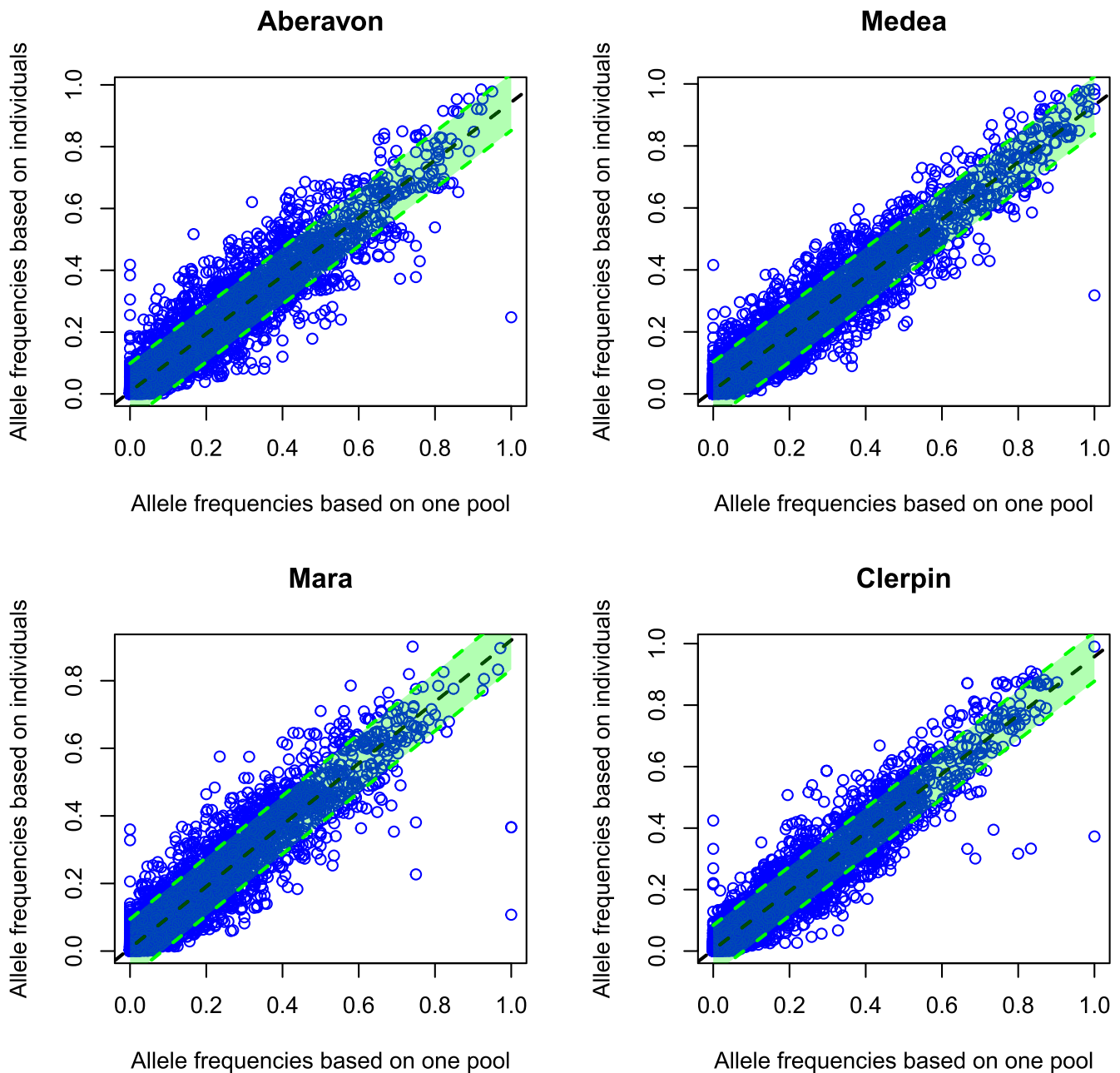


FIGURE 3 | The relationship between allele frequencies estimated on individuals and on one pool for each variety. The dashed line marks equality between the two ($y=x$), with the green area representing the 95% prediction intervals for a fitted linear regression line.

typically seen in GBS data sets (several hundreds of thousands) (Julier et al. 2018).

In practice, the GBS data can be utilised to establish a new genotyping platform that may address some issues discussed above. Targeted genotyping by sequencing approaches, such as Flex-Seq, are robust to allelic variation due to longer probes. Furthermore, as they target specific loci in a more uniform manner than GBS based on genome complexity with restriction enzymes, higher read depths per locus can be achieved. This will enable more accurate estimates of allele frequencies from complex pools such as diploid and tetraploid synthetic cultivars. As it is a targeted approach, the markers can be evenly distributed across chromosomes and even supplemented with some target loci (e.g., markers in linkage with QTL). This more targeted

sequencing approach is likely to represent a more robust, reproducible, and cost-effective tool for the purposes outlined here. Genotyping may be potentially used for other aspects of the variety registration process, thus improving efficiency. These include reference collection management and distinctness assessment, as well as for enhancing estimates of performance.

Further work would be beneficial to fully establish the robustness and utility of the variance approximation for pools. This would include consideration of marker technology, and the influence of coverage and number of markers, as well as application to other examples of cross-pollinated crops. A deeper exploration of measurement error associated with allele frequency estimation and its impact on the genetic variance estimate is needed. The utility and robustness of the method should be evaluated more fully

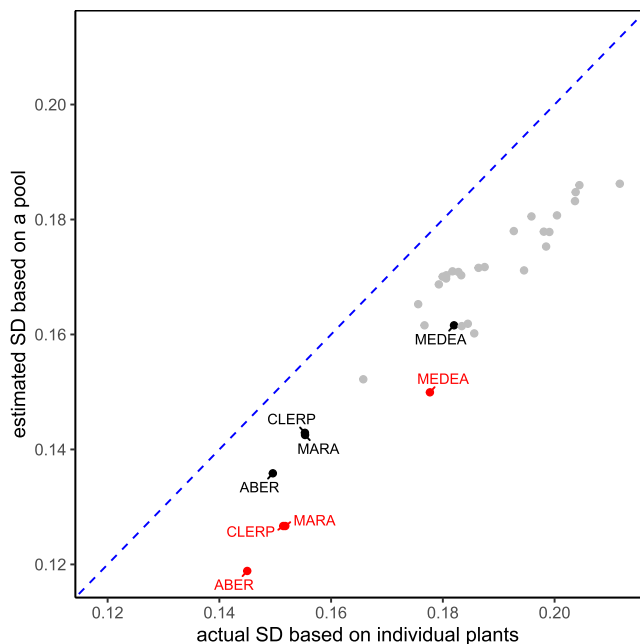


FIGURE 4 | Data set B: Estimated standard deviations for each population based on actual pools compared with the actual standard deviations based on individual plants. The four varieties are shown in black when all markers were included, and in red when only markers with little overt bias were included. The results for Data set A are in grey for reference. The blue dashed line marks the line of equality between the approximation and the actual values.

in the DUS and varietal purity contexts, including comparisons with existing non-genetic processes. For example, can it identify samples with poor uniformity or varietal purity?

5 | Conclusions

The proposed method for evaluating genetic heterogeneity in pooled samples demonstrates potential for use in practice. Applications include uniformity assessment, as part of new plant variety registration, and for varietal purity checks, as part of production controls. Although this study concentrated on perennial ryegrass, the described approaches are applicable to other crops whose varieties are synthetic populations, such as lucerne and clovers.

Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 817970. Adrian Roberts was also supported by funding from the Scottish Government's Rural and Environment Science and Analytical Services Division. Thanks to many project partners for invaluable discussions, in particular Cecile Collonnier of CPVO. The authors thank the technicians of UR P3F from INRAE involved in this project, including F. Gelin for growing the plants in the greenhouses and S. Blugeon for the bioinformatic analysis. They also thank the INRAE GeT-PlaGe team for their sequencing service.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All data analysed as data set A has been deposited into the NCBI sequence read archive (SRA) under the BioProject ID: PRJNA352789. Data set B is available at <https://doi.org/10.57745/ETSZVF>.

References

- APHA. 2022. "United Kingdom Variety Testing/Plant Breeders' Rights Technical Protocol for Official Examination of Distinctness, Uniformity and Stability (DUS): Perennial Ryegrass." Published by the Animal and Plant Health Agency, Cambridge, United Kingdom. <https://www.gov.uk/guidance/dus-protocols-for-testing-plant-varieties#dus-protocols>.
- Auzanneau, J., C. Huyghe, B. Julier, and P. Barre. 2007. "Linkage Disequilibrium in Synthetic Varieties of Perennial Ryegrass." *Theoretical and Applied Genetics* 115: 837–847. <https://doi.org/10.1007/s00122-007-0612-3>.
- Bataillon, T., P. Gauthier, P. Villesen, S. Santoni, J. D. Thompson, and B. K. Ehlers. 2022. "From Genotype to Phenotype: Genetic Redundancy and the Maintenance of an Adaptive Polymorphism in the Context of High Gene Flow." *Evolution Letters* 6: 189–202. <https://doi.org/10.1002/evl3.277>.
- Blanco-Pastor, J. L., P. Barre, T. Keep, et al. 2021. "Canonical Correlations Reveal Adaptive Loci and Phenotypic Responses to Climate in Perennial Ryegrass." *Molecular Ecology Resources* 21, no. 3: 849–870. <https://doi.org/10.1111/1755-0998.13289>.
- Blanco-Pastor, J. L., S. Manel, P. Barre, et al. 2019. "Pleistocene Climate Changes, and Not Agricultural Spread, Accounts for Range Expansion and Admixture in the Dominant Grassland Species *Lolium Perenne* L." *Journal of Biogeography* 46, no. 7: 1451–1465. <https://doi.org/10.1111/jbi.13587>.
- Byrne, S., A. Czaban, B. Studer, F. Panitz, C. Bendixen, and T. Asp. 2013. "Genome Wide Allele Frequency Fingerprints (GWAFFs) of Populations via Genotyping by Sequencing." *PLoS One* 8, no. 3: e57438. <https://doi.org/10.1371/journal.pone.0057438>.
- Byrne, S. L., P. Conaghan, S. Barth, et al. 2017. "Using Variable Importance Measures to Identify a Small Set of SNPs to Predict Heading Date in Perennial Ryegrass." *Scientific Reports* 7: 3566. <https://doi.org/10.1038/s41598-017-03232-8>.
- Byrne, S. L., I. Nagy, M. Pfeifer, et al. 2015. "A Synteny-Based Draft Genome Sequence of the Forage Grass *Lolium Perenne*." *Plant Journal* 84, no. 4: 816–826. <https://doi.org/10.1111/tjp.13037>.
- Cericola, F., I. Lenk, D. Fè, et al. 2018. "Optimized Use of Low-Depth Genotyping-By-Sequencing for Genomic Prediction Among Multi-Parental Family Pools and Single Plants in Perennial Ryegrass (*Lolium Perenne* L.)." *Frontiers in Plant Science* 9: 369. <https://doi.org/10.3389/fpls.2018.00369>.
- CPVO. 2019. "Protocol for Tests on Distinctness, Uniformity and Stability: *Lolium*." Published by Community Plant Variety Office (CPVO), Angers, France. <https://cpvo.europa.eu/en/applications-and-examinations/technical-examinations/technical-protocols/cpvo-technical-protocols>.
- Elshire, R. J., J. C. Glaubitz, Q. Sun, et al. 2011. "A Robust, Simple Genotyping-By-Sequencing (GBS) Approach for High Diversity Species." *PLoS One* 6: e19379. <https://doi.org/10.1371/journal.pone.0019379>.
- Fè, D., B. H. Ashraf, M. G. Pedersen, et al. 2016. "Accuracy of Genomic Prediction in a Commercial Perennial Ryegrass Breeding Program." *Plant Genome* 9, no. 3: 11.0110. <https://doi.org/10.3835/plantgenom2015.11.0110>.
- Fè, D., F. Cericola, S. Byrne, et al. 2015. "Genomic Dissection and Prediction of Heading Date in Perennial Ryegrass." *BMC Genomics* 16, no. 1: 921. <https://doi.org/10.1186/s12864-015-2163-3>.

- Guo, X., F. Cericola, D. Fè, et al. 2018. "Genomic Prediction in Tetraploid Ryegrass Using Allele Frequencies Based on Genotyping by Sequencing." *Frontiers in Plant Science* 9: 1165. <https://doi.org/10.3389/fpls.2018.01165>.
- Julier, B., P. Barre, P. Lambroni, et al. 2018. "Use of GBS Markers to Distinguish Among Lucerne Varieties, With Comparison to Morphological Traits." *Molecular Breeding* 38: 133. <https://doi.org/10.1007/s11032-018-0891-1>.
- Keep, T., S. Rouet, J. L. Blanco-Pastor, et al. 2021. "Inter-Annual and Spatial Climatic Variability Have Led to a Balance Between Local Fluctuating Selection and Wide-Range Directional Selection in a Perennial Grass Species." *Annals of Botany* 128, no. 3: 357–369. <https://doi.org/10.1093/aob/mcab057>.
- Keep, T., J. P. Sampoux, J. L. Blanco-Pastor, et al. 2020. "High-Throughput Genome-Wide Genotyping to Optimize the Use of Natural Genetic Resources in the Grassland Species Perennial Ryegrass (*Lolium perenne* L.)." *G3-(Bethesda)* 10, no. 9: 3347–3364. <https://doi.org/10.1534/g3.120.401491>.
- Konkolewska, A., S. Phang, P. Conaghan, D. Milbourne, A. Lawlor, and S. Byrne. 2023. "Genomic Prediction of Seasonal Forage Yield in Perennial Ryegrass." *Grassland Research* 2, no. 3: 167–181. <https://doi.org/10.1002/glr2.12058>.
- Mascher, M., T. A. Richmond, D. J. Gerhardt, et al. 2013. "Barley Whole Exome Capture: A Tool for Genomic Research in the Genus *Hordeum* and Beyond." *Plant Journal* 76, no. 3: 494–505. <https://doi.org/10.1111/tbj.12294>.
- OECD. 2023. "Guidelines for Control Plot Tests and Field Inspection of Seed Crops." OECD, Paris, France. <https://www.oecd.org/content/dam/oecd/en/topics/policy-sub-issues/seeds/GUIDELINES%20FOR%20CONTROL%20PLOT%20TESTS%20AND%20FIELD%20INSPECTION%20OF%20SEED%20CROPS.pdf>.
- Pégard, M., P. Barre, S. Delaunay, et al. 2023. "Genome-Wide Genotyping Data Renew Knowledge on Genetic Diversity of a Worldwide Alfalfa Collection and Give Insights on Genetic Control of Phenology Traits." *Frontiers in Plant Science* 14: 1196134. <https://doi.org/10.3389/fpls.2023.1196134>.
- Pembleton, L. W., C. Inch, R. C. Baillie, et al. 2018. "Exploitation of Data From Breeding Programs Supports Rapid Implementation of Genomic Selection for Key Agronomic Traits in Perennial Ryegrass." *Theoretical and Applied Genetics* 131, no. 9: 1891–1902. <https://doi.org/10.1007/s00122-018-3121-7>.
- Saccomano, B., M. Wallace, D. M. O'Sullivan, and J. Cockram. 2020. "Use of Genetic Markers for the Detection of Off-Types for DUS Phenotypic Traits in the Inbreeding Crop, Barley." *Molecular Breeding* 40: 13. <https://doi.org/10.1007/s11032-019-1088-y>.
- UPOV. 2002. "General Introduction to the Examination of Distinctness, Uniformity and Stability." UPOV Document TG/1/3. UPOV, Geneva. https://www.upov.int/resource/en/introduction_dus.html.
- UPOV. 2020. "Guidance on the Use of Biochemical and Molecular Markers in the Examination of Distinctness, Uniformity and Stability (DUS)." UPOV Document TGP/15/3. UPOV, Geneva. https://www.upov.int/edocs/tgpdocs/en/tgp_15.pdf.
- UPOV. 2022. "Trial Design and Techniques Used in the Examination of Distinctness, Uniformity and Stability." UPOV document TGP/8/5. UPOV, Geneva. https://www.upov.int/edocs/tgpdocs/en/tgp_8.pdf.
- UPOV. 2024. "OECD Seed Schemes: Latest Developments in the Application of BMT Under the OECD Seed Schemes. UPOV Document TWM/2/19." UPOV, Geneva. https://www.upov.int/edocs/mdocs/upov/en/twm_2/twm_2_19.pdf.
- UPOV FAQs (A). "What is a Plant Variety?" Published by UPOV, Geneva. <https://www.upov.int/about/en/faq.html>.
- UPOV FAQs (B). "Does UPOV Allow Molecular Techniques (DNA Profiles) in the Examination of Distinctness, Uniformity and Stability ("DUS")?" <https://www.upov.int/about/en/faq.html#QB80>.
- Verwimp, C., T. Ruttink, H. Muylle, et al. 2018. "Temporal Changes in Genetic Diversity and Forage Yield of Perennial Ryegrass in Monoculture and in Combination With Red Clover in Swards." *PLoS One* 13, no. 11: e0206571. <https://doi.org/10.1371/journal.pone.0206571>.
- Wang, J., N. O. I. Cogan, and J. W. Forster. 2016. "Prospects for Applications of Genomic Tools in Registration Testing and Seed Certification of Ryegrass Varieties." *Plant Breeding* 135: 405–412. <https://doi.org/10.1111/pbr.12388>.
- Yang, C. J., J. Russell, L. Ramsay, W. Thomas, W. Powell, and I. McKay. 2021. "Overcoming Barriers to the Registration of New Plant Varieties Under the DUS System." *Communications Biology* 4: 302. <https://doi.org/10.1038/s42003-021-01840-9>.
- Zhang, J., K. Kobert, T. Flouri, and A. Stamatakis. 2014. "PEAR: A Fast and Accurate Illumina Paired-End Read Merger." *Bioinformatics* 30: 614–620. <https://doi.org/10.1093/bioinformatics/btt593>.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Table S1:** Number of qualifying samples by population. **Table S2:** Genomic coordinates of target SNPs utilised in the design of MyBaits to generate Data set B.

Appendix A

Proof of Genetic Variance Approximation for Diploids

With allele frequencies expressed as the proportion of the B allele at a SNP, then frequencies of 0, $\frac{1}{2}$ and 1 correspond to genotypes AA, AB, and BB respectively for a single individual.

For a pooled sample from a variety, the frequencies for AA, AB, and BB at SNP i are labelled $p_{AA,i}$, $p_{AB,i}$ and $p_{BB,i}$. Then $p_{AA,i} + p_{AB,i} + p_{BB,i} = 1$.

The allele frequency, μ_i , for the sample at SNP i is given by:

$$\mu_i = \frac{1}{2}p_{AB,i} + p_{BB,i}$$

The variance, σ_i^2 , in the allele frequencies over the individuals in the sample at SNP i is given by:

$$\sigma_i^2 = \frac{1}{4}p_{AB,i} + p_{BB,i} - \frac{1}{4}p_{AB,i}^2 - p_{BB,i}^2 - p_{AB,i}p_{BB,i}$$

We need a transform, $f(\mu_i)$, such that $f(\mu_i) \approx \sigma_i^2$. We show that the following formulation for $f(\mu_i)$ is equal to σ_i^2 when $\min(p_{AA,i}, p_{BB,i}) = 0$:

$$\text{for } \mu_i \leq \frac{1}{2}: f(\mu_i) = \mu_i \left(\frac{1}{2} - \mu_i \right)$$

$$\text{for } \mu_i > \frac{1}{2}: f(\mu_i) = \left(\mu_i - \frac{1}{2} \right) (1 - \mu_i)$$

Let's plug in the formula for μ_i . So,

For

$$\mu_i \leq \frac{1}{2}: f(\mu_i) = \frac{1}{4}p_{AB,i} + \frac{1}{2}p_{BB,i} - \frac{1}{4}p_{AB,i}^2 - p_{BB,i}^2 - p_{AB,i}p_{BB,i} = \sigma_i^2 - \frac{1}{2}p_{BB,i}$$

For

$$\mu_i > \frac{1}{2}: f(\mu_i) = \frac{3}{4}p_{AB,i} + \frac{3}{2}p_{BB,i} - \frac{1}{4}p_{AB,i}^2 - p_{BB,i}^2 - p_{AB,i}p_{BB,i} - \frac{1}{2} = \sigma_i^2 - \frac{1}{2}p_{AA,i}$$

Thus, we have:

$$f(\mu_i) = \sigma_i^2 - \frac{1}{2} \min(p_{AA,i}, p_{BB,i}), \quad \forall \mu_i$$

Therefore, the bias in $f(\mu_i)$ for estimating σ_i^2 depends on $\min(p_{AA,i}, p_{BB,i})$. And if $\min(p_{AA,i}, p_{BB,i}) = 0$, then $f(\mu_i) = \sigma_i^2$.

Note that other alternatives for $f(\mu_i)$ are possible, leading to different biases. For example, $f(\mu_i) = \mu_i \left(\frac{1}{2} - \mu_i \right)$ leads to a positive bias of $\frac{1}{4} p_{AB,i}$. However, in our example data sets, the mean of $\frac{1}{2} \min(p_{AA,i}, p_{BB,i})$ over SNPs is substantively smaller than the mean of $\frac{1}{4} p_{AB,i}$. So our preferred approximation leads to a smaller bias.