

Developing a strategy to apply SNP molecular markers in the framework of winter oilseed rape DUS testing

Agreement Nr. **7515058**

Co-funded by the Community Plant Variety Office (CPVO) Research and Development Section and the Groupe d'Étude et de contrôle des Variétés Et des Semences (GEVES), June 2019 – September 2021

Project coordinator: GEVES (FR)

Project partner: Bundessortenamt (DE)

Associated partner: NIAB (UK)

Duration of the project: 2 years*

Total estimated cost: 253 492 €

***Note.** The project started officially on 25/06/2019 and was originally planned to finish on the 24/06/2021. However, due to the Covid-19 pandemics, CPVO agreed to a no cost extension until 30 September 2021.

Summary of the project

GEVES (France) and Bundessortenamt (Germany) carried out a joint project with the objective to develop a novel approach that integrates genotypic information based on single nucleotide polymorphism (SNP) markers within the framework of DUS testing in winter oilseed rape to optimize the management of the reference collection and avoid having to redescribe the entire reference collection every year.

The objectives of the project were:

- to genotype part of the reference collection and produce a large and consistent molecular dataset for ~2000 WOSR varieties;
- to use the molecular data to filter and optimise the set of SNP markers;
- to evaluate different approaches and design a new model that combine genetic and phenotypic data to optimize DUS testing in WOSR.

Drawing upon the results of the project, we propose a new approach derived from UPOV BMT Model “Combining phenotypic and molecular distances in the management of variety collections” that uses network analysis for describing genetic relationships between candidate and reference varieties and defining an “optimal reference collection” whose composition is deduced from the candidates’ genotypes. In this model, graph clustering tools are used to identify clusters of genetically related varieties and only references from clusters that comprise candidate varieties are included in growing trials. Depending on the method used in the process of the distinctness criteria (COYD or GAIA), the model can reduce by 20 to 35% the size of growing trials in the first year with the preliminary thresholds of pairwise genetic distance that we defined.

So far, the method has only been tested *in silico*. A second follow-up project is required to test the model in the field and calibrate the genetic thresholds.

List of partners

GEVES (project coordinator)

Groupe d'Étude et de contrôle des Variétés et des Semences
25 rue Georges Morel, CS 90024, 49071 Beaucouzé cedex, France

Contact person: Arnaud Remay

Scientific coordinators: Arnaud Remay and Muriel Thomasset

Data analysis was performed by Marc Delêtre and Muriel Thomasset. The conceptual approach was developed by Marc Delêtre who wrote the final report with contributions from Muriel Thomasset, Arnaud Remay and Anne Bernole.

Bundessortenamt (project partner)

Bundessortenamt
Osterfelddamm 80, 30627 Hannover, Germany

Main contacts: Beate Rücker, Elisabeth Thiemt

NIAB (associated partner)

National Institute of Agricultural Botany
93 Lawrence Weaver Road, Cambridge CB3 0LE, United Kingdom

Main contact: Margaret Wallace

Table of contents

SUMMARY OF THE PROJECT	1
LIST OF PARTNERS	2
TABLE OF CONTENTS	3
INTRODUCTION	5
MILESTONE 1: BUILDING A LIBRARY OF MOLECULAR PROFILES FOR THE WOSR REFERENCE COLLECTION	7
1) TASK 1 – SELECTION OF VARIETIES: DEFINING A CONSENSUAL REFERENCE COLLECTION	7
2) TASK 2 – GENOTYPING	7
i. Preparation of seed bulks	7
ii. DNA extraction protocol	7
iii. PCR amplification protocol	8
iv. Allele calling protocol	8
3) TASK 3 – DATA ANALYSIS AND OPTIMIZATION OF THE SNP SET	8
i. Selection of additional SNP markers	8
ii. Filtering criteria	9
iii. Controls, replicates and pseudo-replicates	11
iv. Performance of the final SNP set	12
MILESTONE 2: DEVELOPING A NEW MODEL USING SNP GENETIC DATA	13
MAIN DIFFERENCES BETWEEN THE FRENCH AND GERMAN DUS SYSTEMS IN WOSR	13
MALE STERILITY SYSTEMS AND THEIR INFLUENCE ON DUS TESTING	14
1) TASK 1 – DEVELOPING A NEW MODEL	17
i. Using unsupervised graph-based community detection methods for identifying clusters of genetically related varieties	19
ii. Description of the method	19
iii. Defining genetic thresholds	22
iv. Expected results	24
2) TASK 2 – ASSESSMENT OF THE NEW MODEL	26
i. Using coefficients of relatedness (kinship) to calibrate the thresholds	29
ii. Evaluating the consistency of clustering results	32
iii. Applicability of the network approach to the GAIA system (FR)	36
iv. Applicability of the network approach to the COYD system (DE)	37
v. Implementation	39
vi. Varietal turn-over	39
MILESTONE 3: FEASIBILITY STUDY	42
1) CHOICE OF EXTERNAL PROVIDERS: KASPAR VS. NGS-BASED TECHNOLOGIES	42
2) EVALUATION OF SERVICE	42
i. Transferability	42
ii. Flexibility, delays/reactivity, availability	44

iii. Deliverables	44
iv. Comparability of results: reproducibility, repeatability	44
v. Value for money	46
3) CONCLUSIONS	47
MILESTONE 4: FINAL MEETING	48
1) CONCLUSIONS	48
2) FOLLOW-UP	50
REFERENCES	51
APPENDIXES	54
i. Complete list of the 548 SNP markers tested	54
ii. Distribution of SNPs on the genome	64

Introduction

Winter oilseed rape (*Brassica napus* L., WOSR) is the third most important oilseed crop in the world and the main oilseed crop in Europe, where it is mainly grown for edible oil and biodiesel production. Encouraged by EU Renewable Energy Directive (RED 2009/28/EC) that set as a target that 10% of energy in transport should be from renewable sources by 2020, the production of oilseeds in Europe has rapidly grown over the last decade to produce biofuel and the EU is now the world's largest producer of biodiesel¹ from rapeseed oil (approx. 60% of the global output).

With 1.4-1.6 million ha [Mha] each (~20% of total EU production), France and Germany are the two largest producers of WOSR in Europe, followed by Poland (0.8-0.9 Mha), UK (0.5-0.7 Mha) and Czech Republic (0.4 Mha) (EUROSTAT 2021). In the last 10 years, breeding efforts have focused on producing hybrid varieties to harness hybrid vigour (heterosis). Compared to inbred lines, F₁ hybrids exhibit better adaptability and higher stress resistance, and hybrid varieties now represent the majority of applications submitted in France and Germany for plant breeders' rights (PBRs) and/or for listing in national catalogues (NLI). When parental lines are included, this represents almost 200 new varieties per country that must undergo DUS testing each year.

In oilseed rape, the expression of phenotypic characters is very sensitive to changes in environmental conditions. To perform valid comparisons, the entire reference collection must be considered during the test to account for this high inter-annual variability. To date, the theoretical reference collection for oilseed rape represents approximately 8 500 varieties of common knowledge (worldwide). The technical collection, which totals ~2 500 varieties in Europe for winter OSR, includes all varieties bred for similar agroclimatic zones and/or tested and described by entrusted examination offices (EOs). Varieties bred in the USA and Canada, where genetically modified crops are authorized, are not included in field trials in Europe, whereas they are banned under European legislation, nor are varieties with a very distinct genetic makeup (varieties from China, Japan, Mexico, Russia, Uruguay) or adapted to different agri-environmental conditions. Old varieties that deviate significantly from the genetic pool of varieties under study given the evolution of ongoing breeding efforts are also excluded from the technical collection. The effective collection (~2 350 varieties) corresponds to the part of the technical collection for which EOs have seed samples at their disposal (Fig. 1).

Examination offices have adopted different strategies to reduce the number of varieties to be included in growing trials. However, all varieties protected by PBRs and varieties in national catalogues must be systematically included. The effective collection usually also comprises a set of varieties that show relatively stable characters and are used for calibrating DUS evaluations. The composition of the collection is also tailored to the characteristics of candidate varieties of that particular year; for example, if candidates include high erucic acid content lines then all erucic lines in the collection are included in the growing trial (whereas they would not otherwise, to reduce the size of field trials). In all, DUS technicians must observe between 1 000 and 1 600 varieties every year. This represents significant costs and huge technical and logistical challenges in terms of planning and infrastructure (e.g. limited availability of agricultural lands). GEVES (France) and Bundessortenamt (Germany) carried out a joint project with the objective to develop novel approaches that integrate genetic and phenotypic information to avoid having to redescribe the entire reference collection every year without compromising the validity of DUS decisions. The project investigated how single nucleotide polymorphism markers (SNPs) could be used to reduce the number of varieties that need to be included in growing trials in the 1st year of DUS testing.

¹ Rapeseed oil methyl ester (RME)

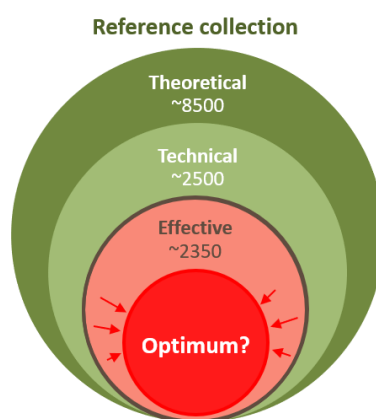


Fig. 1 | A schematic view of the reference collection showing how the effective collection is nested within the technical and theoretical collections. The project’s aim was to develop a method to define an “optimal reference collection” to reduce the size of growing trials.

The project is a follow-up of a pilot project led by GEVES and NIAB (UK) in 2016-2017 that tested an initial set of 500 SNPs (CPVO 2018). This built upon the conclusions of a project led by NIAB in 2005-2007 (CPVO 2008) that explored the potential of simple sequence repeat markers (SSRs) for managing the WOSR reference collection which concluded that multi-allelic markers are not suitable for DUS testing in a partially out-crossing tetraploid species such as oilseed rape and advocated instead in favour of functional SSRs (i.e. gene specific markers linked to characteristics of interest) or bi-allelic markers such as SNPs.

Project aims

The objectives of the project were:

- to genotype a large part of the reference collection and produce a consistent molecular dataset for ~2000 WOSR varieties;
- to use the molecular data to filter and optimise the set of SNP markers;
- to design new models and evaluate different approaches that combine genetic and phenotypic information based on SNP markers and historical field data to optimize DUS testing in WOSR.

We designed and tested a new approach derived from UPOV Model “Combining phenotypic and molecular distances in the management of variety collections” (UPOV TGP/15/3 [2020]) that uses network analysis to identify sets of related varieties based on their proximity (genetic similarity). In this model, graph clustering tools are used to identify an “optimal reference collection” whose composition is deduced from the candidates’ genotypes. We show that graph clustering tools can help infer genetic structure within large reference collections and select more efficiently reference varieties to reduce the workload in winter oilseed rape DUS testing without impacting the frameworks (COYD or GAIA) underlying decisions on variety distinctness.

Milestone 1: Building a library of molecular profiles for the WOSR reference collection

1) Task 1 – Selection of varieties: defining a consensual reference collection

The Bundessortenamt (BSA) and GEVES defined together a consensual list of 1989 WOSR varieties to be included in the molecular dataset. Varieties protected by plant variety rights were included in priority (737 in total), followed by varieties in national collections from other countries entrusted by CPVO for conducting DUS testing in oilseed rape (CZ, DE, DK, ES, FR, PL, SK, and formerly UK). Both EOs agreed to start with varieties in the French and German national reference collections for which morphological data were available.

Selection criteria

Varieties were selected to include all varietal types (restored hybrids, three-way hybrids, sterile hybrids, CMS restorer lines, sterile lines, synthetic lines) and to represent a range of phenotypic variation. Additionally, example varieties listed in CPVO-TP/36/03 and UPOV-TG/36/6 and varieties that are regularly used by experts as references in field trials and known to have relatively stable characters were also included. In total, 1 968 WOSR varieties, including 1531 varieties that were common between the French and German collections, were selected for genotyping². The collection comprised 646 hybrids, including 23 semi-dwarf varieties, 32 imidazolinone (IMI) herbicide tolerant varieties, 20 high oleic/low linoleic acid (HOLL) content, three high erucic acid (HE) content, 22 three-way hybrids, and 18 sterile hybrids. Parental lines of the hybrid varieties selected for genotyping were also included in the collection, including three dwarf lines, 20 HOLL lines, 13 HE lines, and eight IMI lines (1290 conventional lines in total). Six candidate varieties that failed DUS tests for default of distinction and 15 varieties that failed DUS tests for lack of homogeneity were also included. In addition, 41 spring OSR varieties and one forage (winter) variety were included as outgroup.

2) Task 2 – Genotyping

All molecular work was performed at GEVES. Following the conclusions of the pilot project, DNA extractions were performed on bulks of 32 seeds. While bulking partially masks intravarietal genetic diversity (a phenomenon called allele dilution), it produces unique and consistent molecular profiles for WOSR varieties and permits the detection of alleles with a minor allele frequency (MAF) > 20%.

i. Preparation of seed bulks

For each variety, bulks of 32 seeds were prepared by hand or using a Contador electronic seed counter (Pfeuffer, Germany). Seeds were ground with a Retsch MM301 mixer mill (Retsch GmbH) for 20 seconds at 30 Hz using ø7 mm stainless steel beads.

ii. DNA extraction protocol

DNA was extracted using a partially automated paramagnetic bead-based method, using NucleoMag 96 Plant DNA extraction kits (Macherey-Nagel GmbH & Co., Düren, Germany). Briefly, 1000 µl of MC1 lysis buffer and 10 µl RNase A (at 12 mg/ml concentration) were added to 150 mg plant material in a 2 ml tube. DNA was then incubated for 30 min at 65°C, mixing samples 2-3 times by vortexing during incubation. After incubation, samples were centrifugated for 10 min at 14 000 × g. The lysate was then transferred onto 96-well plates for DNA extraction using an automated Biomek FX^P Beckman workstation (Beckman Coulter Life Sciences, Indianapolis, USA). To bind DNA to the beads, 400 µl MC2 buffer was then added to 400 µl lysate and 20 µl NucleoMag[®] V-beads. Plates were left for 2

² To comply with IPR, the complete list of varieties cannot be reproduced in the report.

min on a magnetized base, the supernatant was then discarded, and beads were washed with 80 μ L MC3 buffer. Tubes were agitated for 5 min and left for 2 minutes on a magnetic separator. After discarding the supernatant, samples were washed with 80 μ L MC4 buffer following a similar protocol. After a final wash with 80 μ L 80% EtOH, DNA was eluted in 150 μ L of MC6 buffer, after 5 min agitation at room temperature and the supernatant transferred into 96-well elution plates. Final DNA concentrations were checked by spectrometry using a FLUOstar® Omega microplate reader (BMG Labtech GmbH, Germany) and normalized to 20 ng/ μ L. DNA (170 μ L per sample) was then transferred into 384-well plates and diluted to 2 ng/ μ L using a Hamilton Microlab® STAR automated liquid handling platform (Hamilton Bonaduz AG, Bonaduz, Switzerland).

iii. PCR amplification protocol

Polymerase Chain Reaction (PCR) amplifications were carried out in 4 μ L total volume in a 384-well plate in an Hydrocycler™ (LGC Genomics). Each well contained 1.95 μ L of 2 $\times$ KASP-TF master mix, 0.053 μ L KASP assay mix, and 2 μ L (~4ng) DNA template. KASP assay mixes are specific to each SNP and contain two allele-specific forward primers and a common reverse primer. The KASP-TF master mix contains the two universal FRET (fluorescent resonance energy transfer) cassettes (FAM and HEX) and PCR reagents (ROX™ passive reference dye, Taq polymerase, free nucleotides and MgCl₂). The PCR cycling conditions were: 15 min at 94°C, 10 annealing cycles starting by 20 s at 94°C then 1 min at 61°C (dropping to 55°C at a rate of 0.6°C per cycle), followed by 26 amplification cycles starting with 10 s at 94°C and followed by 1 min at 55°C.

iv. Allele calling protocol

PCR plates were read using PHERAstar FSX microplate reader (BMG Labtech GmbH, Germany). Data analysis and allele calling were performed using the Kraken software (LGC Genomics). If the signature genotyping groups had not formed after the initial amplification, three additional amplification cycles ('recycling') were performed, and the samples were read again. Recycling was repeated as many times as necessary.

3) Task 3 – Data analysis and optimization of the SNP set

One of the aims of the project was to optimize the set of SNP markers by replacing SNPs that failed during the pilot project and filtering markers according to their ease of scoring (unequivocal allele calling), discriminating power, reproducibility, and genome specificity.

i. Selection of additional SNP markers

For the pilot project, a set of 500 SNP markers with a uniform distribution across *B. napus* genome was selected from the 60K Illumina Infinium™ array developed by Clarke *et al.* (2016). Using genotyping data from INRAE (French National Research Institute for Agriculture & Environment) from 406 oilseed rape varieties characterized with that SNP array, 48 additional SNPs were chosen within regions of the chromosomes where some of the previously selected SNPs failed to amplify or showed little or no polymorphism, leaving gaps in the coverage of the genome. Because *B. napus* is allotetraploid³, SNPs were first screened for their specificity to one of the two genomic components (either A [*B. rapa*] or C [*B. oleracea*]). The new SNPs were then selected based on their position on both physical and genetic maps (Appendix i and ii) according to the same criteria as applied during the pilot project (flanking sequences suitable for designing primers for KASP™ assays, polymorphism information content (*PIC*) > 0.3, heterozygosity < 0.05, missing data < 0.03). When no SNP was available that matched these criteria, the best SNP on the next segment was chosen instead.

³ *Brassica napus* is amphidiploid (AACC genome, 2n = 38). It is the result from interspecific hybridization between *B. rapa* L. (AA genome, 2n = 20) and *B. oleracea* L. (CC genome, 2n = 18).

ii. Filtering criteria

All 508 SNPs⁴ were first tested on 384 samples (one plate). Based on these initial results, we chose to discard 65 markers that appeared to be monomorphic or failed to converge into distinct clusters even after several additional recycling steps.

A common problem we encountered was the lack of clear separation between the heterozygous group and one of the homozygous groups. This can be due to the preferential amplification of one of the allele-specific primers (a phenomenon called allele competition), or a dosage-effect due to working with seed bulks and to the heterogeneous nature of the plant material used (i.e. some varieties may not be entirely fixed at that locus and consist of a mixture of homozygous plants with a minority of heterozygous individuals). As a rule, when this occurred, we preferred marking the data as missing rather than making the wrong call.

Genotyping resumed with the remaining 443 SNPs. To confirm the validity of allele calls, a first round of preliminary analyses was performed by comparing the observed genotypes of hybrid varieties with the genotypes expected given the genotypes of the parents. This validation step permitted establishing some rules for reading plates in a consistent manner to reduce interpersonal variability in calling genotypes. Based on these rules, plates were read a second time and results reviewed by committee made of research technicians and data analysts to validate genotype calls. Following this step, twelve additional SNPs that proved difficult to score were discarded.

The final marker set was determined by filtering SNPs using the following criteria:

- *Missing data rate: < 15%*

Missing data can result from difficulties to score alleles because data points did not resolve into clearly separate clusters on the KASP genotyping plot, or because the SNP failed to amplify in several samples. As a rule, all SNPs with more than 15% missing data were discarded (5 SNPs in total).

- *Heterozygosity in parental lines: < 30%*

An excess of heterozygosity in parental lines, which are generally assumed to be homozygous on all or most loci, can indicate genotyping errors due to difficulty to call alleles because the heterozygous group migrated towards one of the homozygous groups on the KASP genotyping plots, making genotype scoring difficult. SNP loci yielding ambiguous readings are difficult to score and not desirable for routine applications (UPOV INF/17/1 [2010]). All SNPs scored as heterozygous for more than 30% of the 1301 lines included in the collection were flagged as potentially problematic and discarded (3 SNPs in total).

- *Conformity with the hybrid formula: < 30% mismatch*

For all hybrid varieties included in the collection, results were compared with the genotypes expected given the genotypes of the parents. Altogether, 651 comparisons (630 restored hybrids and 21 three-way hybrids) were performed to assess the conformity of hybrid formulas. SNP loci with over 30% mismatch between the expected and observed genotypes across all hybrid varieties were flagged as problematic (42 SNPs in total).

⁴ Including the 452 SNPs validated during the pilot project, eight SNPs from the pilot project that amplified but were monomorphic (for the 23 varieties initially tested), and the 48 new SNPs selected to replace SNPs that failed.

- *Concordance between controls and technical/biological replicates: < 30% mismatch*

To maximize agreement between allele calling across the different genotyping runs, three samples from the pilot project whose genotype was known for most of the loci were repeated on each of the 24 DNA extraction plates (equivalent to 4 repeats of each control per 384-well genotyping plate). To further assess the concordance of results across runs, DNA extraction and genotyping were also repeated for 60 samples (lines and hybrids) using the same material (technical replicate) or a different bulk of 32 seeds (biological replicate). As a rule, loci with over 30% non-conformity between controls and across replicates were discarded (12 SNPs in total).

In total, 369 markers passed the first four filters. The final decision whether to keep or discard problematic SNP markers was based on all criteria together, taking care that it does not result in significant gaps in the genome coverage (whether based on the physical or genetic map).

- *Linkage disequilibrium (LD)-based SNP pruning: $r^2=0.95$*

In addition to screening markers for quality, SNPs were also tested for independence. Linkage disequilibrium (LD) is the non-random association of alleles at two or more loci within a population (Slatkin 2008). Adjacent loci are said to be in linkage disequilibrium when there is a strong correlation between polymorphisms (alleles) observed between loci. Strong linkage disequilibrium results in redundant information, therefore a cut-off value was applied to prune loci in excess pairwise LD.

For biallelic markers, one of the most common measures for LD is the r^2 statistic, which measures correlation between alleles at two loci based on the presence/absence of a particular allele at the first locus and the presence/absence of a particular allele at the second locus. Genotyping data was pruned in R 4.0.5 (R Project for Statistical Computing, 2021) using the package *SNPRelate* (Zheng *et al.* 2012), with a 5% rate of missing data and minor allelic frequency of 0.05. No pair of SNPs were found in complete LD ($r^2 = 1$) and only one pair (C03-287 / C03-288) for $r^2 = 0.99$. As a cut-off value, we retained only one SNP for analysis from pairs or groups of SNPs with a $r^2 \geq 0.95$. The LD-pruning step removed nine additional SNPs, resulting in a final set of 360 markers. The number of SNPs discarded vs. the number of SNPs kept on each chromosome is shown in Fig. 2. Although care was taken that discarding these markers would not result in large gaps in the genome coverage, minor gaps remain, particularly on chromosomes A04 and A08 (middle section) for the A-subgenomic component, and chromosomes C07 and C09 (distal end) for the C-subgenomic component. The distribution of the final panel of SNPs across the genome is shown in Appendix ii.

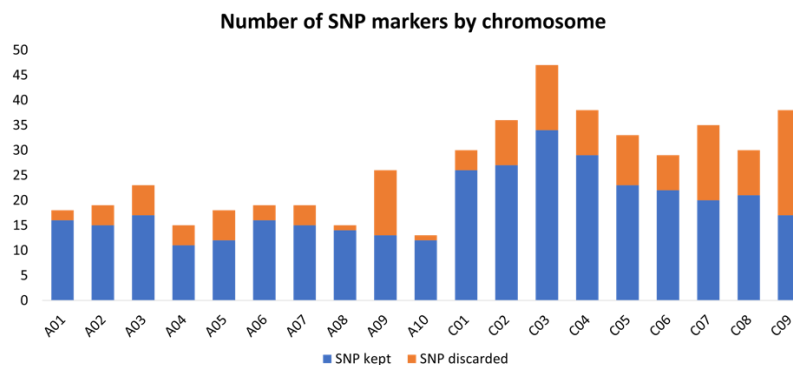


Fig. 2 | Final number of SNPs by chromosome. The number of SNPs discarded after quality filtering and LD-pruning is shown against the theoretical number of SNPs (proportional to the size of each chromosome) needed to achieve even coverage of the genome.

iii. Controls, replicates and pseudo-replicates

To assess repeatability of the results and ensure consistency across genotyping runs, three varieties from the pilot project were included as controls in each of the 24 plates prepared for DNA extraction (Fig. 3-A). In addition, an additional bulk of 32 seeds was prepared for 60 varieties (30 hybrids and 30 parental lines) to provide biological replicates (Fig. 3-B). Seeds were ground and the material split into two aliquots to provide technical replicates (60 in total) (Fig. 3-C). DNA from each aliquot was used for independent DNA extraction, PCR and genotyping. Bulks from distinct seed lots were also prepared for 60 varieties to provide ‘pseudo’-replicates (Fig. 3-D). One additional seed lot (i.e. two in total) was analysed for 44 varieties, and two additional seed lots (three in total) were analysed for 8 varieties.

Based on the set of 360 SNP markers we retained, we obtained 99% match ($\pm 2\%$) between controls (all 24 replicates), 97% match ($\pm 3\%$) between biological replicates, and 98% match ($\pm 3\%$) between technical replicates. For pseudo-replicates, genotypes were 95% congruent ($\pm 4\%$) between seed lots. These high rates of concordance between controls and replicates demonstrate the reliability of the SNP panel and further validate our choice to work with bulks of 32 seeds, which ensures that data derived from bulk DNA are representative of the (dominant) genotype of the variety.

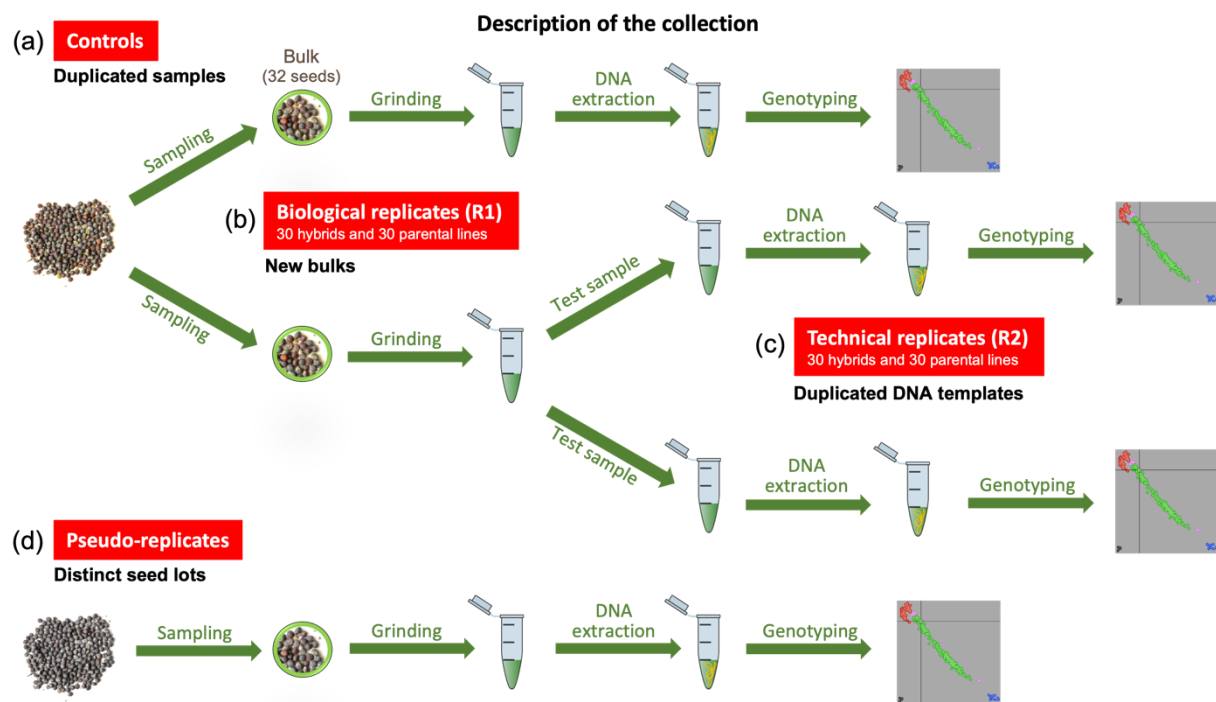


Fig. 3 | Sampling strategy for controls (a), biological replicates (b), technical replicates (c) and pseudo-replicates (d). Biological replicates (distinct bulks) were included to check data replicability, i.e. whether the bulking strategy yields consistent genotypic profiles (i.e. to estimate the amount of noise and/or allele dilution due to working with seed bulks). Technical replicates (duplicated DNA samples with a new DNA extraction) were used to check data repeatability and measure the rate of genotyping errors. Finally, pseudo-replicates that involved distinct DUS seed lots were also included to test data reproducibility and the marker set's ability to detect changes in allele frequencies (which will translate into heterogeneous genotypic profiles) between distinct seed lots of the same variety.

iv. Performance of the final SNP set

Plotting the number of distinct multilocus genotypes (MLG) resolved by SNPs markers against an increasing number of markers showed that with 360 SNPs selected a plateau is quickly reached, and that ~175 loci would suffice to capture the genetic diversity of the entire collection (Fig. 4). The saturation of the genotype discrimination curve illustrates the resolving power of the final set of markers. Out of a total of 1 936 512 comparisons, only 11 pairs of varieties showed a $GD = 0$. Nine of these pairs corresponded to sterile A-line and their corresponding maintainer (B-line) in the CMS system, in which A and B are expected to be isogenic. Of the remaining two pairs, one involved double haploids obtained through the same original crossing, and the final pair was comprised of two varieties from the same breeder that differed only due to missing data on two loci.

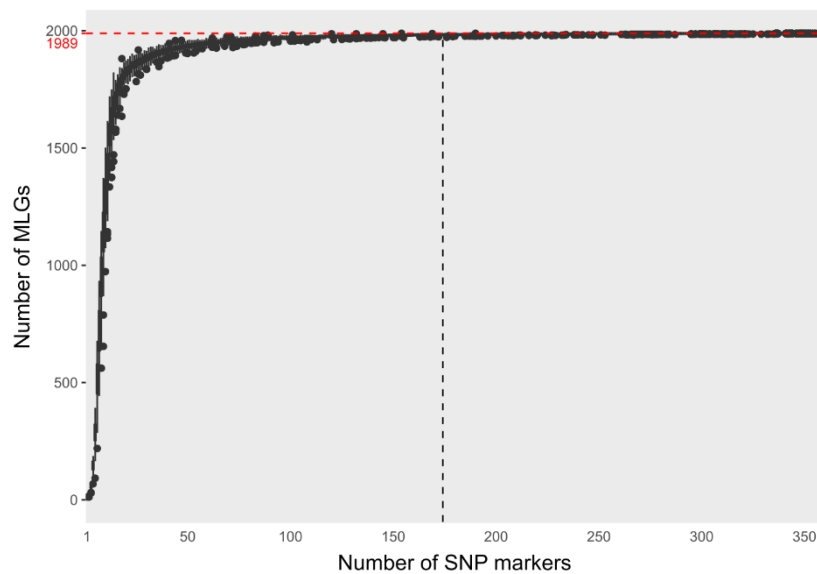


Fig. 4 | Resolving power of the final SNP set. The number of multilocus genotypes (MLG) resolved (y-axis) is shown as a function of the number of SNP markers used (x-axis). In total, the molecular library includes 1989 distinct genotypes, all of which could be resolved using only about half of the 360 markers.

Milestone 2: Developing a new model using SNP genetic data

The aim of the project was to design a new approach that uses SNP genetic data to reduce the size of growing trials from the first year of study onwards without affecting the validity of DUS decisions, irrespective of the method used in the process of the distinctness criteria (GAIA or COYD).

Main differences between the French and German DUS systems in WOSR

In France, morphological descriptions from the first growing cycle are used to identify the potential similar varieties that need to be compared side by side during the second growing cycle. Visual observations and measurements converted into notes are analysed using GAIA, a software developed by GEVES that calculates a phenotypic distance between pairs of varieties (UPOV 2003, 2007). Because characters do not all contribute the same to the overall measure of morphological similarity, weighing matrices are used to refine the calculation of morphological distances between varieties⁵. Matrices are defined for each character, in which the weight varies depending on the amplitude of the difference observed or measured between varieties in their levels of expression of a character (i.e. higher weights are assigned to larger differences). These matrices are set and regularly reviewed by DUS experts. GAIA returns a list with pairs of varieties for which the sum of weightings across all characteristics is lower than a chosen threshold (usually 6 for WOSR); these varieties are grown side-by-side in the second year of DUS testing, whereas those who exceed the threshold are considered “distinct +” and excluded from the next DUS cycle (Fig. 5-A). To further reduce the number of pairwise comparisons, isoenzymatic profiles are also compared and contribute to the global distance between two varieties. Under CPVO revised protocol TP/036/3 (CPVO 2020), examination offices using the GAIA method are exempt of assessing some of the quantitative characteristics that are otherwise compulsory. In addition to GAIA, DUS experts in France use the combined-over-years method (COYD) for assessing distinctness on characteristics related to siliques.

In Germany, the evaluation of distinctness is based on two growing cycles for candidates and reference varieties. For characteristics measured on individual plants, data are analysed using COYD to assess distinctness (Fig. 5-B). COYD involves three steps: first, (1) calculating for each character the variety means from 2-3 consecutive years of field trial and calculating the over-year means for candidate and reference varieties; (2) analysis of variance (ANOVA) to determine for each character the least significant difference (*LSD*) at the chosen statistical threshold (1% for WOSR); and (3) comparing differences between varieties in their over-years means with *LSD*_{1%} values. A pair of varieties is considered distinct if the difference between their over-years means is greater than the *LSD*_{1%} for at least one character. Varieties that are not declared distinct after two years are evaluated for another year. If a variety shows no significant difference after three years, it is rejected for lack of distinctness.

In WOSR, hybrid varieties now represent most of the applications for PBRs and/or listing in national catalogues. In the French and German systems, both hybrids and their (new) parental lines undergo DUS testing.

⁵ Weighting matrices are defined by DUS experts to account for correlations between characteristics and attribute more weights to larger differences, according to the reliability of the characteristic (i.e. its sensitivity to variations in environmental conditions).

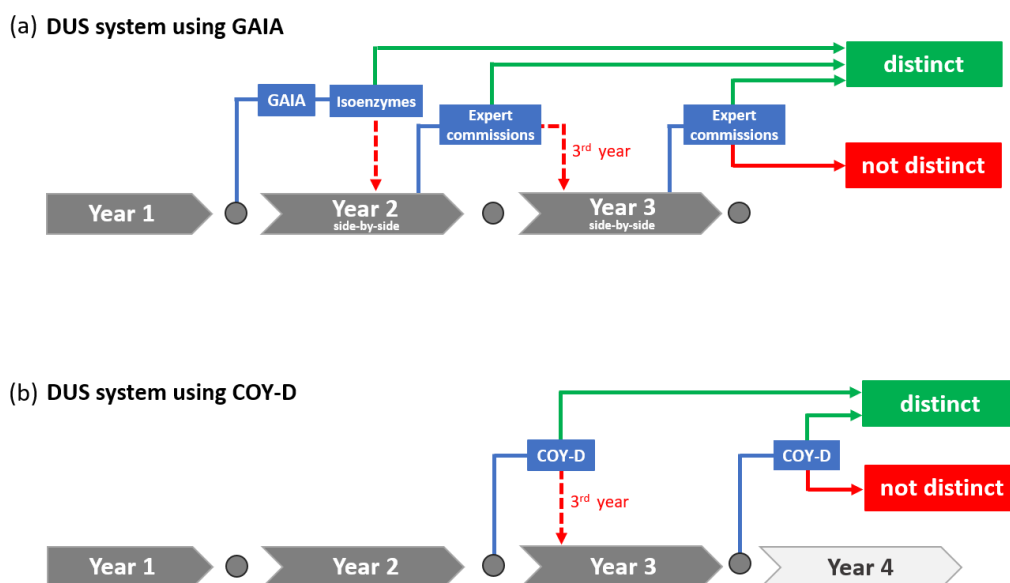


Fig. 5 | General description of the GAIA and COYD systems. (a) In the GAIA system, selection of most similar varieties is assessed at the end of the 1st year by summing up phenotypic differences between varieties. Varieties that exceed a given threshold are considered distinct, those that do not are evaluated during the second year of DUS testing. To reduce the number of pairwise comparisons, isoenzymatic profiles are also compared. If the profiles are different, the candidates are declared distinct; if not, the varieties undergo a trial for distinctness assessment during the second year of testing where they are grown side-by-side with the varieties they were found to be morphologically close. If after the 2nd cycle a candidate is still not found distinct, pairwise comparisons are reconducted for another year. **(b)** In the COYD system, distinctness is assessed at the end of the 2nd year. Varieties that are not declared distinct after two years are evaluated for another year. If a candidate shows no significant difference after three years of data collection, it is rejected for lack of distinctness. Because sowing of new trials starts at the end of August but results of data analyses are not available until November, all candidates are planted in the 3rd year.

Male sterility systems and their influence on DUS testing

Plant breeding methods used to produce hybrid varieties greatly influence the organization of DUS trials in France and Germany. Several systems have been developed that exploit male sterility as a pollination control system to produce commercial hybrid seeds with high hybridization rate (for a review see Chen and Liu [2014]). Male sterile plants do not produce viable pollen but can produce viable seeds when pollinated with pollen from a male fertile plant. In Europe, breeding uses either the *Ogu*-INRA cytoplasmic genetic male sterility system ("CMS") or the genic male sterility Lembke (MSL) system ("GMS").

- DUS protocol for hybrids produced with the CMS system

Several sources of CMS have been identified in *B. napus*: *pol* (which is widely used in China), *nap*, *nig*, *hau*, and *ogu*. In Europe, the *Ogu*-INRA system (Ogura 1968) is predominant. CMS is a three-line system which requires a male sterile line, a maintainer line, and a restorer line (Fig. 6-A). In this system, sterility is maternally inherited and leads to complete sterility under normal environmental conditions. Male sterility is determined by interaction between mitochondrial genes contained in the cytoplasm and a nuclear gene (*Rf*) which overrides cytoplasmic sterility ('restorer gene'). The sterile line (A-line) is developed through repeated backcrossing of a CMS donor with a sterile cytoplasm and lacking the *Rf* gene as the female parent with a maintainer line (B-line) with normal cytoplasm as the

male recurrent parent. After 5-7 backcrosses, cytoplasmic sterility is fixed in the A-line and its nuclear genome is near-identical to that of the recurrent parent⁶. Hybrid fertility is restored using a genetically distinct restorer line (R-line) that possesses the *Rf* gene. In the Ogu-INRA system, restorer lines are introgressed with a restorer gene derived from radish (*Raphanus sativus* L.). These R-lines are therefore genetically distinct from other lines and treated as a distinct category for DUS purpose⁷.

Because the sterile line (A-line) and its maintainer line (B-line) are in theory isogenic, they are expected to be morphologically similar (except for characteristics linked to fertility). A-lines are declared distinct *de facto* if the corresponding B-lines are distinct. Analysis of pairwise genetic distances between A-lines and their corresponding B-lines showed that *GD* was < 0.05 for $\approx$ 95% of CMS A:B pairs included in the collection (Fig. 6-B).

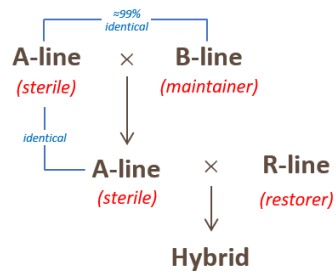
- *DUS protocol for hybrids produced with the GMS system*

The GMS Lembke system (MSL) involves four lines: a male sterile line (MSL-A) which is crossed with a maintainer (MSL-B) to produce the maternal line (MSL-C), which in turn is crossed with a restorer to produce the hybrid (Fig. 6-C). Any inbred line can be used to restore hybrid fertility. In this system, the sterile line and its maintainer can be genetically distinct (Fig. 6-D), and all lines (MSL-A, MSL-B, MSL-C and the restorer) undergo DUS testing.

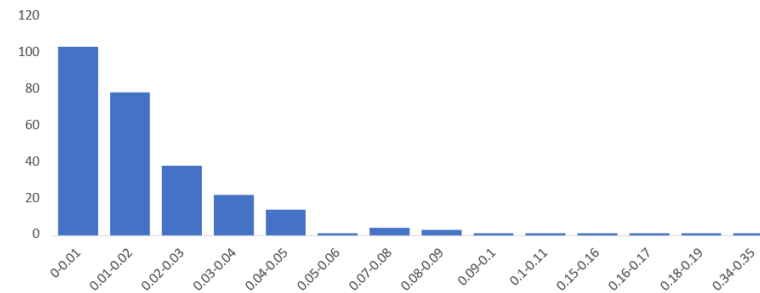
⁶ After five backcrosses, the two lines are $\approx$ 98.4% similar. Usually, the A-line is produced after 6-7 backcrosses.

⁷ Genetic analyses based on SNP data confirmed that R-lines form a genetically distinct group from other commercial lines (data not shown).

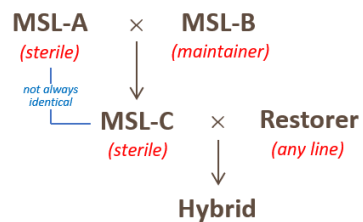
(a) Cytoplasmic genetic male sterility (CMS) system



(b) Frequency distribution of pairwise *GD* between A-lines and B-lines (CMS system)



(c) Genic male sterility (GMS) system



(d) Frequency distribution of pairwise *GD* between A-lines and B-lines (GMS system)

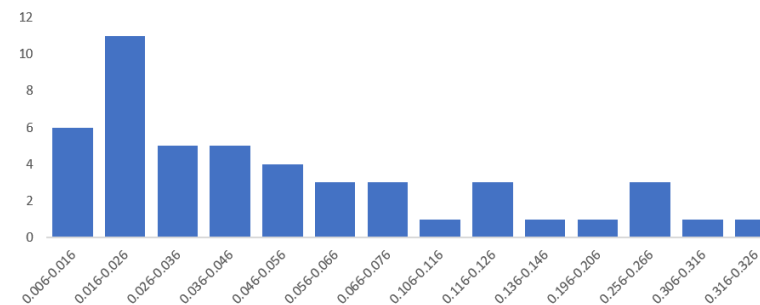


Fig. 6 | CMS-based and GMS-based hybrid breeding systems. In the CMS system (a), sterile lines (A-lines) and their maintainer (B-lines) are expected to be genetically identical (>99% similar). Analysis of pairwise genetic distances showed that *GD* was < 0.05 for ≈ 95% of CMS A:B pairs ($n = 268$) (b). In the GMS system (c), A-lines and B-lines can be very similar in the case of a backcross but they can also be completely unrelated. Pairwise *GD* exceeded 0.05 in 35% of the GMS A:B pairs included in the collection ($n = 48$) (d). All four lines (MSL-A, MSL-B, MSL-C and the restorer) undergo DUS testing.

1) Task 1 – Developing a new model

Selecting the most appropriate reference varieties to be grown in the first DUS cycle is the only approach that can reduce the size of the growing trial in both GAIA- and COY-based systems. One approach is to select a subset of varieties that ‘captures’ the morphological diversity of the reference collection. Such a core collection should include example varieties displaying extreme phenotypes for all diagnostic characteristics used to assess distinctness in order to establish the lower and upper bounds for the observed range of expression of phenotypic characters. The total number of varieties composing this core collection should also permit the definition of statistically meaningful ordinal classes based on the frequency distribution of values for each character to achieve optimal classification of varieties into nine classes (or fewer, depending on characteristics). With a robust methodology to characterize morphological diversity within the collection and position varieties within a multivariate space representing this diversity, it might be possible to identify such a core collection and select varieties using a stratified random sampling design. However, we could not identify a morphological distance that showed enough discriminating power to identify clusters of varieties based on their phenotypes.

Gower distance g (Gower 1971) is particularly adapted to analyse morphological description data because it can handle mixtures of nominal, ordinal and (a)symmetric binary variables, but its power is limited by the number of possible unique combinations of morphological traits used in DUS testing, and by the responsiveness of these traits to environmental fluctuations that can result in scoring inconsistencies (Yang *et al.* 2021). The comparison of pairwise Gower distance calculated using historical DUS data with Rogers’ genetic distance (Rogers 1972) revealed a positive but nonlinear relationship between genetic and morphological distances, with data points showing a characteristic triangular distribution where high morphological distance is always associated with high pairwise genetic distance, but both high and low genetic distances can be observed for lower values of g (Fig. 7). These results show the limits of Gower metrics as a measure of morphological dissimilarity and confirm previous findings that morphological distance does not correlate well with genetic distance unless pedigree relationships are taken into account (Jones *et al.* 2013).

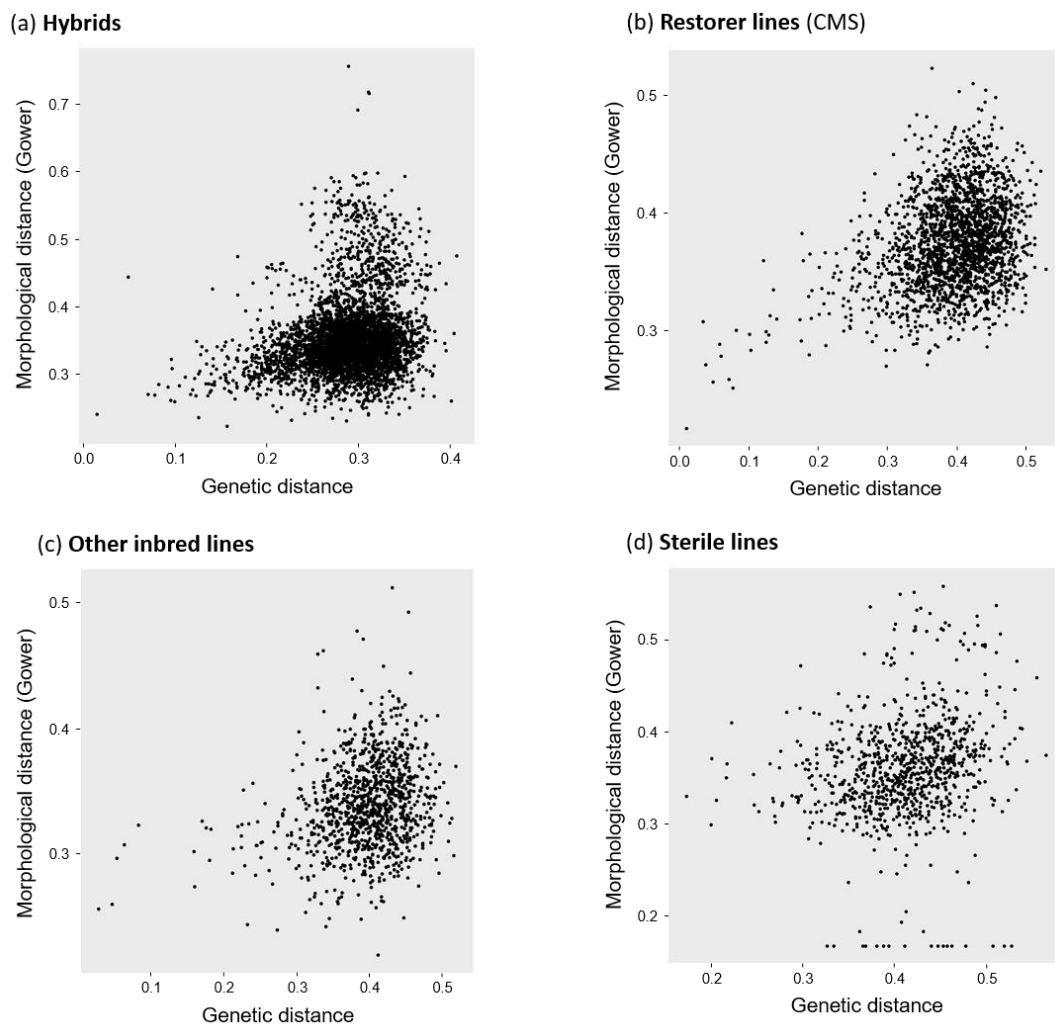


Fig. 7 | Correlation between morphological and genetic distance for hybrids **(a)**, restorer lines **(b)**, inbred lines **(c)** and sterile lines **(d)**. Genetic distance is based on Rogers GD . Morphological distance was estimated by calculating Gower distance g using varietal descriptions converted into notes.

Using molecular markers to structure the collection and select a ‘core-collection’ of exemplar varieties representative of the collection’s overall genetic diversity (i.e. selecting only a subset of varieties from each genetic cluster) is another approach to reduce the size of field trials. However, following intensive breeding to develop “00” varieties with low erucic acid and glucosinolate content, the genetic basis of winter oilseed rape has become very narrow (Gourrion *et al.* 2020). Visualizing pedigree relationships between varieties using information provided in technical questionnaires revealed most varieties to be interconnected (data not shown), and it proved difficult to analyse the population structure of the collection using classical population genetics methods (e.g. Bayesian clustering approaches such as STRUCTURE [Pritchard *et al.* 2000] or multivariate methods such as DAPC [Jombart *et al.* 2010]) (data not shown).

i. Using unsupervised graph-based community detection methods for identifying clusters of genetically related varieties

In recent years, there has been a growing interest in network-based methods to infer genetic population structure in large scale genotyping datasets (Neuditschko *et al.* 2012; Greenbaum *et al.* 2016; Steinig *et al.* (2016); Kuismin *et al.* 2017; Greenbaum *et al.* 2019). Networks based on genetic similarity between varieties are an intuitive approach to explore the genetic structure of a collection and identify groups of varieties that are genetically related. By transforming the genetic distance matrix into a network where each node represents a variety and edges' weight represents the genetic distance between varieties, it is possible to approach the problem of identifying clusters from a graph partitioning point of view.

Graph clustering methods aim at detecting “communities”, that is, densely connected regions of a network comprising sets of nodes (here, varieties) that interact more often or more strongly with each other than with the rest of the network. Unlike model-based approaches such as Structure (Pritchard *et al.* 2000), Admixture (Alexander *et al.* 2009) or BAPS (Corander *et al.* 2003), network-based methods make no prior assumptions about the population genetic model (e.g. populations at Hardy-Weinberg equilibrium) or about the number of subpopulations. They are also computationally more efficient.

ii. Description of the method

We propose a model in which the composition of the reference collection is deduced from the genotype of candidate varieties. In this method, a matrix of pairwise genetic distances is calculated between all varieties. The distance matrix is then transformed into an undirected weighted network, in which nodes (varieties) are connected by edges whose weight corresponds to the genetic distance between nodes (Fig. 8-A). The network's properties can be then explored using graph-theory. First, the network is simplified by pruning all edges corresponding to a genetic distance superior to a predefined genetic threshold to produce a sparse network⁸ (Fig. 8-B). The goal is to reveal clusters of densely connected varieties that share some degree of genetic relatedness. A community detection algorithm is then run to identify clusters of varieties based on their proximity in the network (Fig. 8-C), and only clusters that comprise at least one candidate variety are included in growing trials (Fig. 8-D).

- Detecting communities

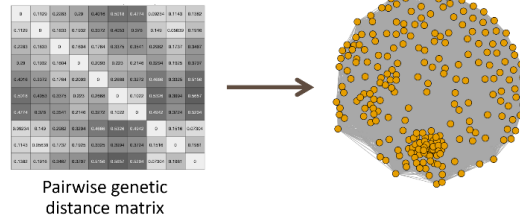
Several algorithms have been proposed to detect community structure within networks (e.g. Louvain [Blondel *et al.* 2008; Rotta and Noack 2011], Leiden [Traag *et al.* 2019], Vienna [Biedermann *et al.* 2020])⁹. The Leiden heuristic is an improvement of the Louvain multi-level algorithm, one of the best performing and fastest algorithms for community detection (Lancichinetti and Fortunato 2009; Yang *et al.* 2016). Unlike some other algorithms, the Leiden algorithm takes edge weights into account, which are treated as distances¹⁰.

⁸ In contrast to a dense network, in which all nodes are connected to all other nodes in the network, a sparse network contains only a subset of the maximum number of links (edges) possible between all nodes. Within a sparse network, the corresponding adjacency matrix, which describes interactions between nodes, contains mostly '0' (no interaction).

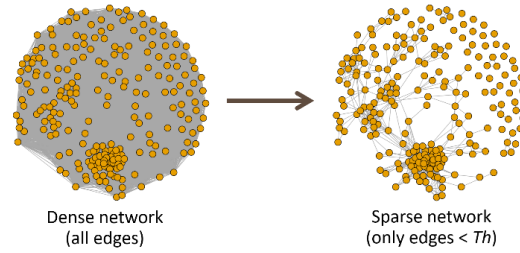
⁹ For a review of methods for community detection, see Fortunato (2010), Lancichinetti and Fortunato (2009), Harenberg *et al.* (2014), and Yang *et al.* (2016). In recent years, research has been focussing on harnessing deep learning techniques for analysing structures within large networks (Su *et al.* 2021).

¹⁰ Community detection algorithms can treat edge weights either as a measure of the strength of the association between two nodes, i.e. a similarity (1- D) or as a distance (D), in which case high values indicate weakly connected dyads.

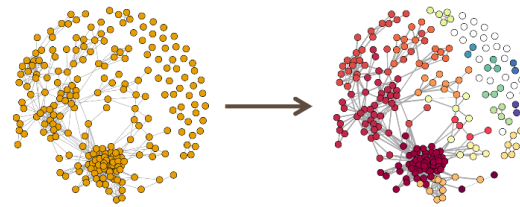
(a) Genetic-similarity network building



(b) Edge pruning



(c) Community detection



(d) Cluster selection

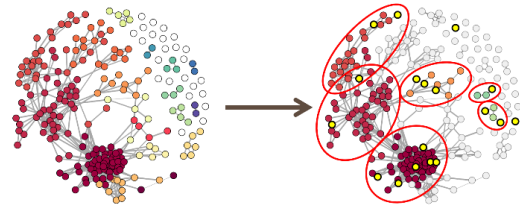


Fig. 8 | Description of the network-based optimisation method. A matrix of pairwise genetic distances is calculated between all varieties then transformed into an undirected weighted network, in which nodes (varieties) are connected by edges whose weight corresponds to the genetic distance between nodes **(a)**. All edges corresponding to a genetic distance superior to a predefined threshold (GD_{Th}) are removed to simplify the network **(b)** and a community detection algorithm is run to identify clusters of varieties **(c)**. Clusters that comprise at least one candidate variety (yellow circles) are included in growing trials; the rest of the reference collection is excluded **(d)**.

Community detection proceeds in three phases: (1) local moving of nodes, (2) refinement of the partition, and (3) aggregation of the network based on the refined partition, using the non-refined partition to create an initial partition for the aggregate network. In the local move procedure, only nodes whose neighbourhood has changed are visited. The refinement is done by restarting from a singleton partition within each cluster and gradually merging the subclusters. Contrary to the Louvain algorithm, which can only merge clusters, the Leiden algorithm can also split clusters. By running the algorithm iteratively, it is possible to improve clustering quality by moving nodes between clusters towards an optimal structure where all clusters are well-connected. Whereas the Louvain algorithm can lead to arbitrarily connected communities, the approach implemented in the Leiden algorithm guarantees that the assignment of nodes to clusters is stable (Traag *et al.* 2019).

- Implementation

The method was developed in R 4.0.5 (R Core Team 2021). Pairwise genetic distances were calculated using custom scripts implemented in the *BioGeves* R package developed by GEVES, and the package *igraph* (Csardi and Nepusz 2006) was used to transform the resulting matrix of genetic distances into a network and remove edges above a predefined threshold. The package *leiden* (Kelly 2020), which implements the Leiden community detection algorithm (Traag *et al.* 2019), was used to identify clusters of varieties within the network. The default Reichardt and Bornholdt Potts optimisation model (Reichardt and Bornholdt 2006) was used with 100 iterations to ensure that clustering converges towards a stable solution where all nodes are locally optimally assigned (Traag *et al.* 2019). The *leiden* package requires Python and the package *reticulate* (Ushey *et al.* 2021) must also be installed to create an interface between R and Python.

We tested the method using Rogers distance (Rogers 1972) but other measures could be used. Rogers distance is an “identity-by-state” (IBS) measure of similarity between individuals that counts allele mismatches between individuals at each locus. Only loci with no missing data (i.e. for which data is present in both individuals) are considered. The distance thus reflects the percentage of loci that differ between both varieties based on their number of common markers. As an alternative, Greenbaum *et al.* (2016) proposed a frequency-weighted similarity metrics that considers genetic distances between individuals and differences in allelic frequencies between subpopulations. The metrics measures diversity relative to the entire population, taking allele frequencies into account. In this case, the metrics measures similarity between varieties (not a distance). Using a frequency-weighted allele-sharing similarity measure that gives higher weight to rare alleles¹¹ may be particularly interesting for multiallelic markers such as SSRs, but the benefits may be more limited for biallelic SNPs.

Other tools (not tested as part of the project) have been developed to explore population genetic structure from molecular datasets without depending on setting a predefined genetic threshold. In CONE (Community Oriented Network Estimation), Kuismin *et al.* (2017) implemented a data-dependent approach that uses generalized linear models (GLM) and a least absolute shrinkage and selection operator (LASSO) regularization technique to introduce sparseness in the network. CONE is implemented in R. However, while CONE proved capable of identifying subspecies in crops, phylotypes in pathogens, and geographic clusters of diversity in human populations, it might not be sensitive enough to reveal the finer genetic structure of reference collections of plant varieties where genetic variation may be limited, with many closely related individuals and no clear population structure.

Another method, NetView P (Steinig *et al.* 2016), uses mutual k -nearest-neighbours (mk-NN) thresholds to build population networks from a genetic distance matrix based on genome-wide SNPs. NetView P can infer fine-scale population structure and even recover structure at family-level. NetView was developed for Python and requires the number of mutual nearest neighbours, k , to be defined. This parameter can be tuned to investigate different levels of genetic similarity (Neuditschko *et al.* 2012), where small values of k focus on connecting fewer, genetically more similar individuals.

NetView and CONE are both focussed on revealing population substructures but offer no real means to control the level of genetic differentiation within clusters, making it harder to interpret results in the context of DUS testing. In the approach we propose, the focus is on identifying groups of closely related individuals based on historical DUS observations, using morphological similarity between varieties to calibrate the maximum pairwise genetic distance to consider for building the network.

¹¹ Sharing rare alleles makes two individuals more similar than sharing common ones.

iii. Defining genetic thresholds

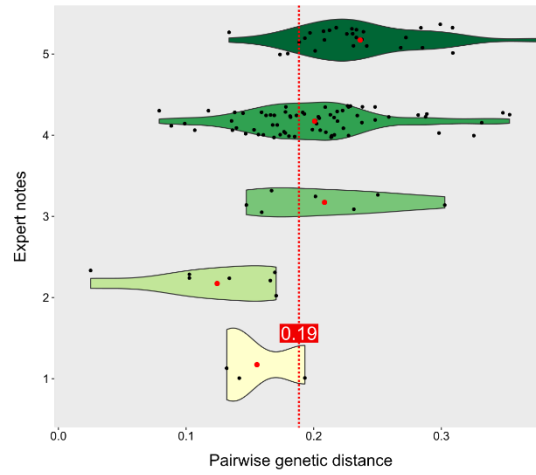
A system of expert notes was used to calibrate the genetic thresholds against morphological similarity between varieties. Using historical data from GEVES, pairs of varieties were ranked from 1 to 5 in increasing order of degree of distinctness, according to how early the pairs were declared distinct in the DUS testing process. The following scale was used:

Similarity scale	
5	Varieties declared distinct at leaf stage before the 1 st commission are considered <u>very distinct</u>
4	Varieties declared distinct at leaf stage (1 st commission) are considered <u>distinct</u>
3	Varieties declared distinct at leaf stage (2 nd commission) are considered <u>somewhat similar</u>
2	Varieties declared distinct at flowering stage are considered <u>similar</u>
1	Varieties not declared distinct after evaluation at the flowering stage are considered <u>very similar</u>

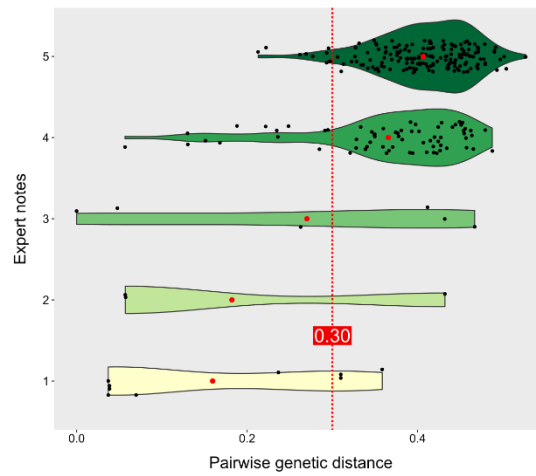
This notation system differs from the system used in previous projects, in which notes were attributed by experts who evaluated distinctness by comparing varieties growing side-by-side in the field (see BMT/16/10 [UPOV 2017]). Here, the scale is based on a global appreciation of the degree of morphological similarity/difference between varieties *a posteriori*, using recordings of their individual characteristics during the second year of evaluation (GAIA system). This notation system is not equivalent to an actual expert note and somewhat introduces a bias as some varieties may be declared “very distinct” at a juvenile state (when evaluated solely on foliar characteristics) but appear more similar at a later stage of development. As a result, it is not possible to make a direct comparison between these *a posteriori* expert notes and the number of characteristics found to be significantly distinct between a pair of varieties using the COYD system: a pair of varieties may be declared distinct based on a single vegetative character (e.g. leaf dentation), and therefore receive a note of ‘5’, while another pair of varieties that also differs by a single character – but one that is observed at a later developmental stage (e.g. silique length)—would receive a note of ‘1’. Despite these shortcomings, these “pseudo-expert notes” are very useful as a first approximation of distinctness for identifying a target range of genetic distances to serve as a threshold above which pairs of varieties are genetically distant enough to be most likely to show significant differences for one or several characters (in the COYD system) and be most unlikely to warrant pairwise comparisons in the field (in the GAIA system).

Based on the analysis of the correlation between these expert notes and pairwise genetic distances based on Rogers’ distance (*GD*), three preliminary thresholds were defined, one for each category identified by DUS experts: *GD* = 0.20 for hybrid varieties, *GD* = 0.30 for restorer lines (CMS system), and *GD* = 0.25 for sterile lines and other parental lines (which include maintainers and other conventional lines). These initial thresholds correspond to the maximum genetic distances observed for the lowest expert notes, corrected for pairs of varieties that DUS examination showed to be clearly distinct (Fig. 9).

(a) Hybrids



(b) Restorer lines (CMS)



(c) Inbred lines (excl. sterile lines)

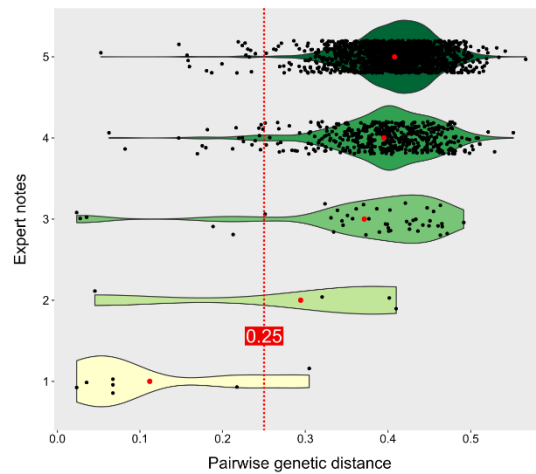


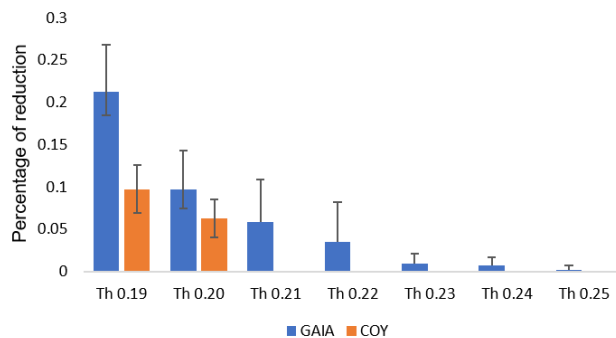
Fig. 9 | Relationship between expert notes and Rogers' genetic distance for restored hybrids (a), restorer lines (CMS system) (b), and other lines (c). Jitter was added along the y-axis to show the distribution of data points along the x-axis (pairwise genetic distance) for each expert note. The position of the mean is indicated with a red dot. Dashed lines indicate the maximum *GD* observed for pairs of varieties with the lowest expert note. The target genetic threshold was defined after correcting for pairs of varieties that received a low expert note despite being clearly distinct (based on characteristics observed at a later stage of development). No expert note was available for the sterile lines to define a genetic threshold.

iv. Expected results

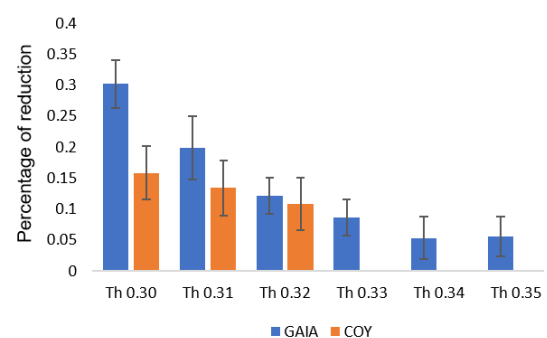
Using historical data provided by GEVES and BSA, the network-based optimization method was run *in silico* to evaluate the impact of the method on the reduction of the number of varieties to include in 1st year trials under the COYD and GAIA frameworks. Separate simulations were run for each category (hybrids, restorer lines¹², sterile lines, and all other lines), using lists of varieties included in field trials between 2008 and 2017, with information on their status ('reference', '1st year', '2nd year', or '3rd year'). Several values were tested below and above the preliminary thresholds identified for each category. For each threshold and category, we estimated the fraction of the reference collection that could be excluded from growing trials when selecting only clusters that included candidate varieties.

Based on available molecular data, the reduction that could be achieved ranged between 10-20% for hybrids, 15-30% for restorer lines, and 30-55% for other parental lines (Fig. 10). Lowering the genetic threshold for hybrid varieties from 0.20 to 0.19 resulted in a significant gain, reducing on average the number of reference varieties to include in fields trials by 20% in the GAIA system and 10 % in the COYD system.

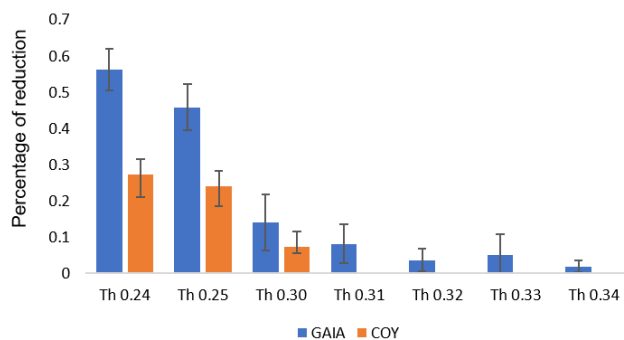
(a) Hybrids



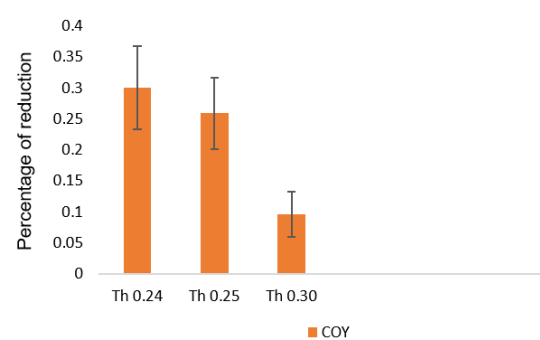
(b) Restorer lines (CMS)



(c) Other inbred lines



(d) Sterile lines



¹² In the CMS (cytoplasmic male sterility) system.

Fig. 10 | Average impact of the network-based optimization method on the size of growing trials at different genetic thresholds for hybrid varieties **(a)**, restorer lines **(b)**, and inbred lines **(c)** under a GAIA-based (blue) and COY-based (orange) DUS framework. Sterile lines do not undergo DUS testing under the GAIA framework **(d)**. Averages are based on a 10-year period (2008-2017) for GAIA and 7-year period (2011-2017) for COYD. For hybrid varieties, the reduction was estimated by comparing the number of reference varieties selected by the method to the number of reference varieties that would be included in growing trials based on information about the candidates' pedigree (i.e. all hybrid varieties with a parental line in common). For parental lines, the reduction was estimated compared to a scenario where all references in collection are included in field trials.

2) Task 2 – Assessment of the new model

The assumption behind the proposed method is that morphological similarity is correlated with genetic relatedness, an assumption corroborated by observations by DUS experts that varieties that are difficult to distinguish in the field are usually genetically related¹³. Current DUS testing rules for hybrid varieties follow a similar reasoning; reference varieties are chosen by analysing the pedigree of the candidate varieties, and only varieties that have a parent in common with one or several candidates are included in growing trials. Here, proximity between varieties in the network is used as a proxy of their likelihood to have similar levels of expression for relevant morphological characteristics under similar environmental conditions (i.e. their phenotype may be influenced by similar genotype × environment interactions). Although one cannot exclude that two varieties may have converged in their morphological characteristics despite coming from distinct gene pools (the “convergence scenario” described by Gilliland *et al.* 2020), such a scenario is not borne by our analysis of 10 years of historical DUS field data¹⁴ and candidates that failed DUS testing in the past always involved varieties that were genetically related to some degree.

By removing all ‘superfluous’ links corresponding to large genetic distances, the method reveals subgraphs consisting of densely connected varieties that share a common ancestry. The ability of the method to detect sets of genetically related varieties is particularly well illustrated in the case of hybrid varieties, for which the structure of the network closely reflects pedigree relationships between varieties (Fig. 11). Similarly, we found clear correlations between cluster membership and genetic relatedness among inbred lines (Fig. 12). Lines that were obtained from the same original cross followed by dihaploidisation and selfing (e.g. W0386 and W0388¹⁵ [W1057 × W0387]), or whose formula involved a common parent (e.g. W0225, W0309, W0853 and W0863, all of which have W0094 as a common parent), always clustered together, regardless of the genetic threshold chosen. The approach was found to be performing equally well for pairs of varieties that failed DUS test for lack of distinctness (e.g. W0133-W0285) or that were found to be morphologically similar (e.g. W0423-W0946 and W0870-W0965).

¹³ This corresponds to the “non-divergence scenario” described in Gilliland *et al.* (2020), in which a candidate has not sufficiently diverged from its parental line.

¹⁴ 15 years of data from GEVES and 10 years of data from BSA, including morphological descriptions for all varieties selected for genotyping (measurements and data transformed into notes), the lists of varieties (references and candidates) included in 1st year field trials (GEVES and BSA) and the lists of varieties selected by GAIA for side-by-side comparison in the 2nd year their GAIA notes (GEVES), as well as results of original COYD analysis with the corresponding morphological data (BSA).

¹⁵ To respect confidentiality, varietal names have been replaced using a project-specific coding system.

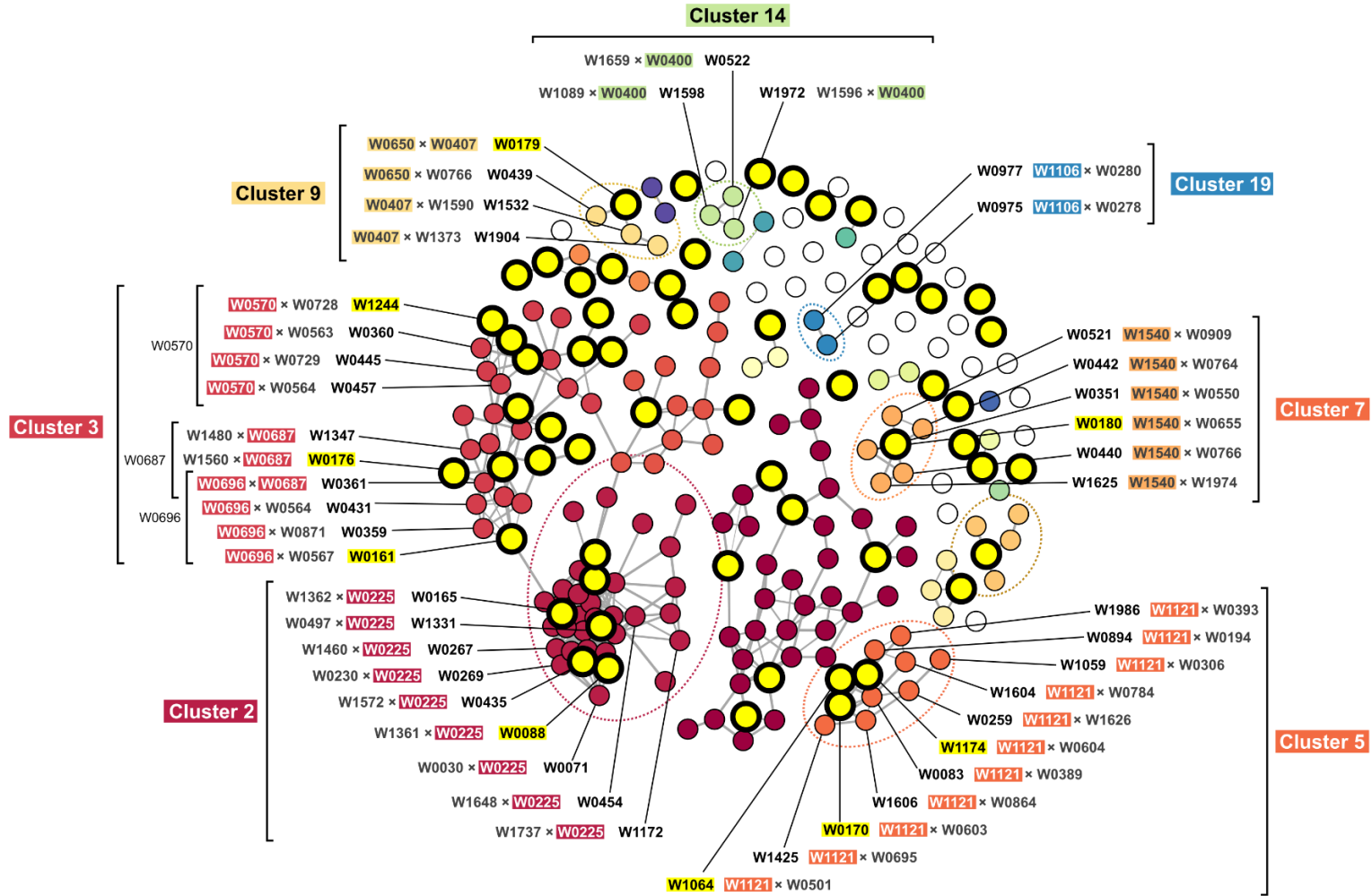


Fig. 11 | Clustering results for hybrid varieties in trial in France in 2015 ($GD_{Th} = 0.19$, resolution = 0.5). The method identified 21 clusters and 35 singletons. Each cluster is represented by a distinct colour. Candidate varieties are highlighted in yellow. Singletons, which correspond to varieties that became disconnected from the network after pruning edges above the chosen genetic threshold, are shown as blank circles. Details about pedigree relationships between varieties are given for a number of varieties. The clusters identified correspond to groups of varieties sharing a common parent, but some clusters can also encompass several groups of varieties sharing a common ancestry (e.g. cluster 3). These clusters split into subclusters when increasing the resolution. In the chosen example, clusters 14 and 19, which do not comprise candidate varieties, would be excluded from growing trials.

i. Using coefficients of relatedness (kinship) to calibrate the thresholds

While they did not find any significant correlation between genetic and morphological data, Tommasini *et al.* (2003) showed that in WOSR pairwise genetic distances (as measured by SSRs) were correlated with pedigree relationships between varieties. Analysing the relationship between pairwise genetic distance (*GD*) and coancestry, as measured by kinship coefficients, is useful for calibrating genetic thresholds. In barley, mean phenotypic distance was found to be highly correlated with genotypic distances when varieties were grouped according to their pedigree relationships (CPVO 2011; Jones *et al.* 2013).

- Pedigree based kinship coefficients

The pedigree-based kinship coefficient θ predicts the expected proportion of an individual's genome that is identical-by-descent (IBD) to another's based on their known pedigree relationships. θ represents the probability that two homologous alleles drawn from each of two individuals are IBD and ranges from 0 (varieties are unrelated) to 0.50 (varieties are genetically identical). The kinship coefficient between a parent and its offspring is 0.25, while half-siblings (such as hybrids sharing a common parent) have a kinship coefficient of 0.125.

Among restored hybrids, the average *GD* for half-siblings (hybrids with a parental line in common; $\theta = 0.125$) was 0.192 ($\pm .036$), compared to 0.302 ($\pm .035$) for unrelated varieties (Fig. 13-A). By setting the threshold between 0.19 and 0.20, the method reveals clusters of varieties that share a common parent, in accordance with current DUS rules. Among restorer lines, the average *GD* for full siblings (varieties obtained from the same cross followed by double haploidization and selfing; $\theta = 0.25$) was 0.191 ($\pm .084$) and 0.296 ($\pm .053$) for half-siblings (varieties obtained via crosses involving a common parental line), compared to 0.412 ($\pm .057$) for unrelated varieties and/or varieties whose pedigree information is unknown¹⁶ ($\theta = 0$) (Fig. 13-B). By setting GD_{Th} to 0.30 for restorer lines, the method targets varieties with a common ancestry, and more likely to show distinction problems¹⁷.

- Marker-based kinship coefficients

Marker-based estimates provide another measure of kinship that does not depend on the availability of pedigree information. However, low rates of recombination among markers located on the same chromosome and high levels of inbreeding can introduce "Mendelian noise" that causes realized IBD to differ from expected IBD. Hence, pedigree-based kinship provides only a theoretical and not a real value of genetic relatedness (Goudet *et al.* 2018), while kinship coefficients based on genetic markers provide an estimate of the realized IBD.

¹⁶ Pedigree information was only available for 58 pairs of varieties.

¹⁷ The relationship between genetic distance and pedigree-based kinship coefficients could not be tested for other inbred lines (including sterile lines) due to a lack of information about their ancestry.

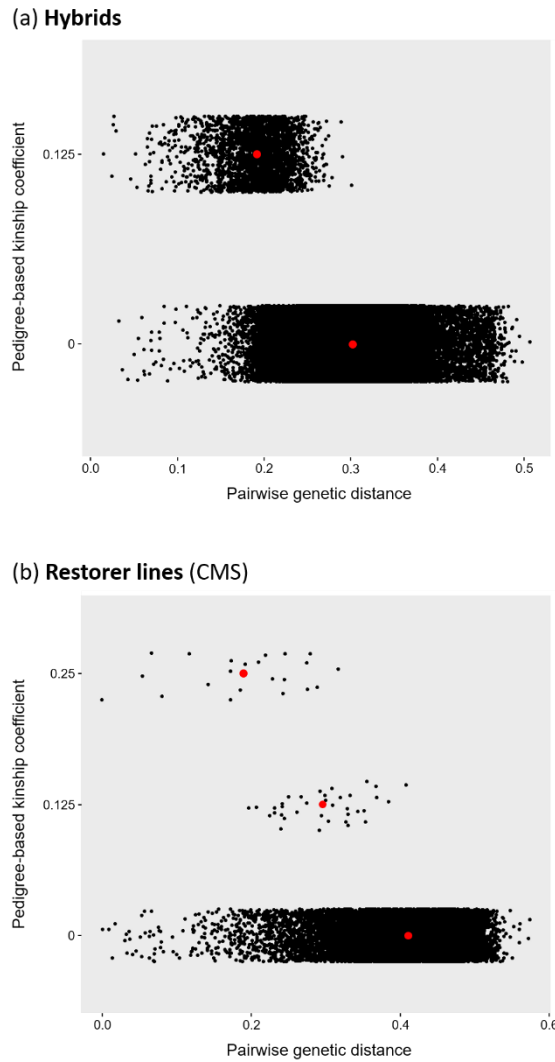


Fig. 13 | Relationship between Rogers' genetic distance and pedigree-based kinship coefficient among 646 restored hybrids **(a)** and 234¹⁷ restorer lines **(b)** based on 360 SNP markers. Jitter was added on the y-axis to better show the distribution of data points along the x-axis (pairwise genetic distance) at different values of θ . The red dot indicates the position of the mean.

Calculating marker-based kinship coefficients between varieties requires first phasing alleles, i.e. determining which allele was inherited from the maternal line and which one from the paternal line. Phasing alleles is a probabilistic estimation of haplotypes using genotypic data. Haplotype phasing was performed on the entire collection using Beagle 4.0 (Browning and Browning 2007). Unlike most recent versions of the software, Beagle 4.0 uses information about pedigree relationships between varieties to improve accuracy of reconstructing haplotype sequences from genotype data. The hap-IBD program (Zhou *et al.* 2020a) was then used to detect IBD haplotype segments and IBDkin (Zhou *et al.* 2020b) to calculate kinship coefficients from IBD segments inferred from the phased genotype data.

Plotting marker-based kinship coefficients against pairwise genetic and morphological distances showed a negative correlation between genetic distance and genetic relatedness between varieties, while morphological distance increased with genetic distance (Fig. 14), suggesting that the more genetically related varieties are, the more similar they are morphologically.

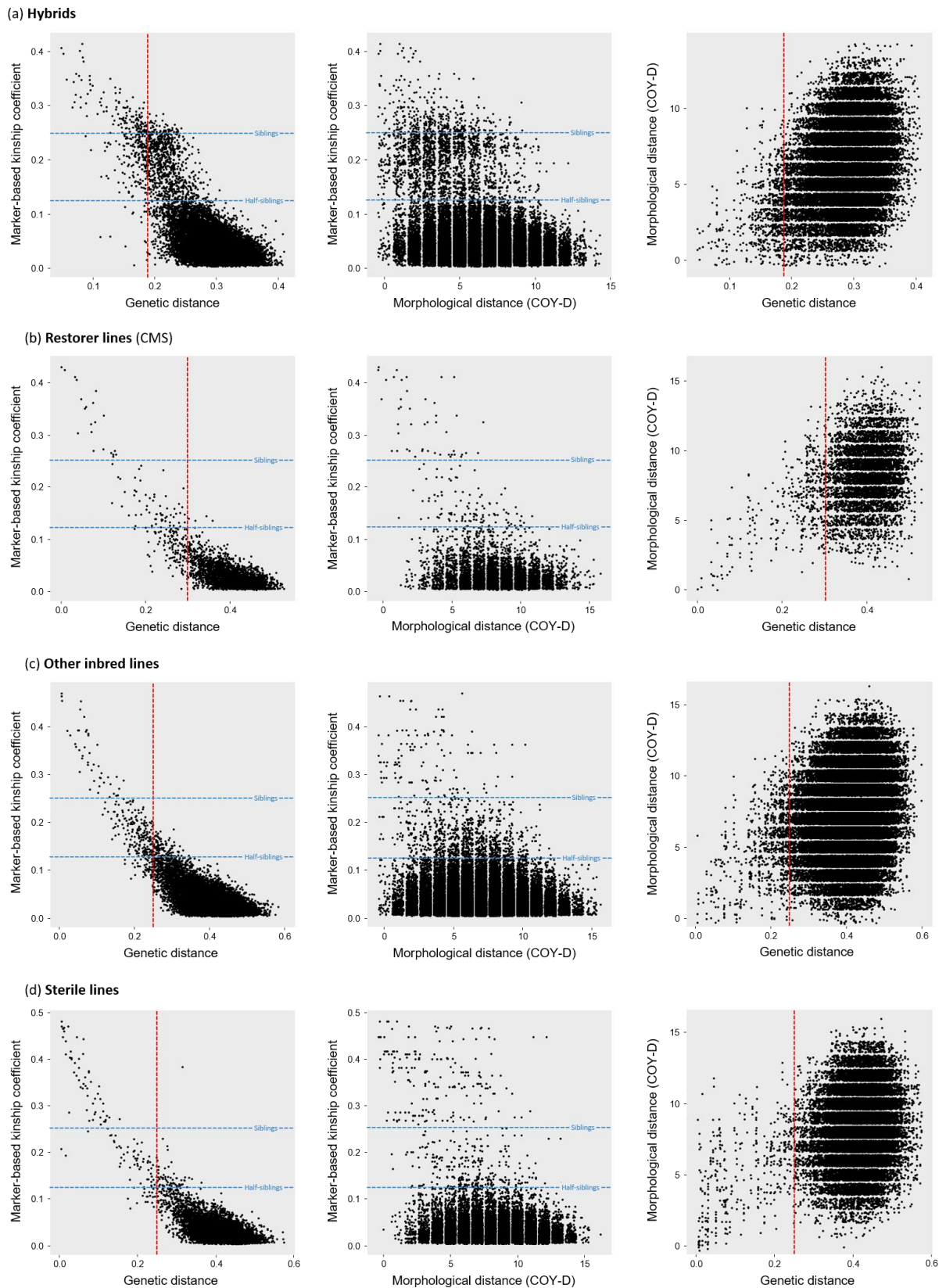


Fig. 14 | Relationship between genetic relatedness and pairwise genetic and morphological distance for hybrids (a), restorer lines (b), inbred lines (c) and sterile lines (d). Morphological distance is expressed as the number of characters for which differences between varieties were significantly different based on COYD analysis. Genetic distance is based on Rogers GD . Dashed red lines indicate the genetic thresholds defined for each category. Dashed blue lines indicate the theoretical kinship coefficients for siblings ($\theta = 0.25$) and half-siblings ($\theta = 0.125$).

ii. Evaluating the consistency of clustering results

Different optimisation methods exist for identifying the best solution to partitioning networks into communities (Fortunato 2010; Harenberg *et al.* 2014). By default, the *leiden* R package implements Reichardt and Bornholdt Potts model (Reichardt and Bornholdt 2006), which tries to maximize the number of internal edges in a community while simultaneously keeping small the size of communities.

Modularity (Newman and Girvan 2004) is an important property of biological networks (Fletcher *et al.* 2013). Modularity measures the tendency of a network to divide into clusters (or communities) in which nodes are more densely connected to each other (strong within-community links) than they are to the rest of the network (weak between-community links). Modularity is an objective measure of how good the community structure found by the algorithm is, but it suffers from a resolution limit that is dependent on the size of the network and the degree of interconnectedness of the clusters (Fortunato and Barthelemy 2007). This effect can be mitigated by setting a resolution parameter (γ).

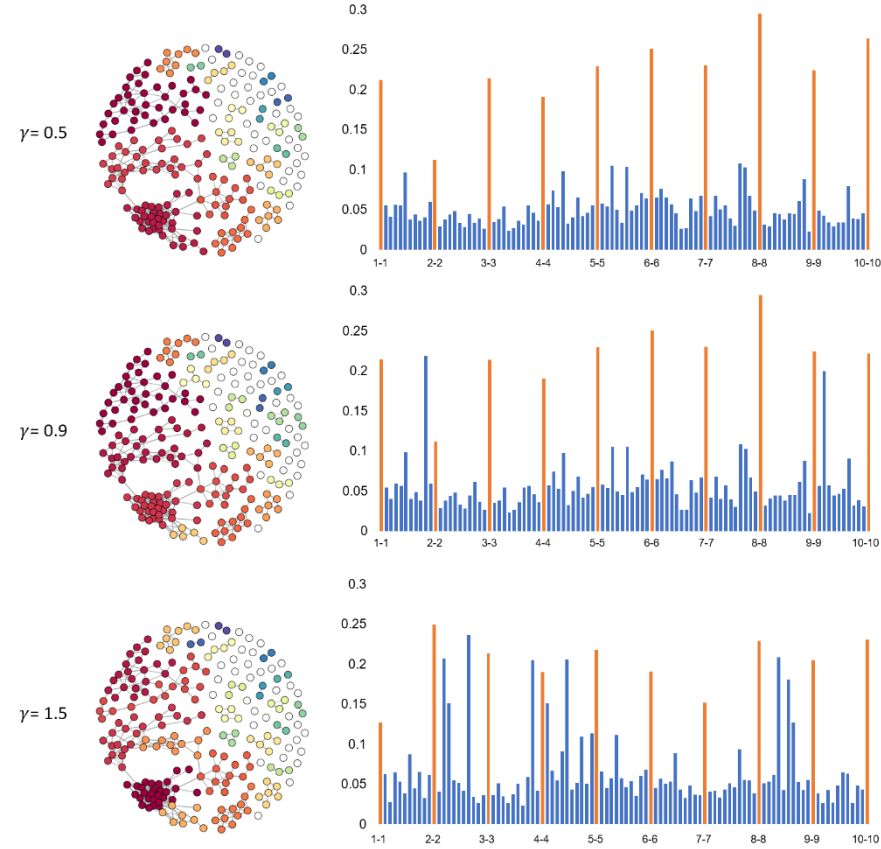
- *Setting the resolution parameter*

The resolution parameter γ determines the size of the clusters detected by the algorithm. γ controls the relative importance between the density of internal edges (within clusters) and external density (between clusters). Changing resolution does not change the nature of the clustering, only the level of definition, and γ can be used to tune the number of clusters to detect. High resolutions ($\gamma > 1$) lead to more communities (splitting clusters into smaller groups), whereas low resolutions ($\gamma < 1$) lead to fewer communities (merging sub-clusters). Increasing γ could help achieve higher discrimination when lowering further the genetic threshold is not possible; however, it can also yield spurious clustering results.

The impact of increasing/lowering γ on the genetic quality of cluster membership was tested by calculating the average genetic relatedness among varieties within and between clusters at different resolutions. Within clusters, varieties were generally more related to each other (higher θ) than when comparing varieties between clusters. This held true while the resolution parameter was kept below 1. At higher resolution ($\gamma > 1$), clusters split into subclusters no longer reflecting varieties' pedigree relationships (Fig. 15).

Although there is no rule-of-thumb for defining the value of the resolution parameter, we recommend keeping $\gamma < 1$. Resolution should be adjusted based on a visual assessment of the network and a comparison of clustering results obtained at different resolutions. Based on our observations, we recommend setting γ to 0.5 for hybrid varieties and 0.9 for inbred lines. These values were defined empirically, aiming to maximize genetic relatedness among varieties within clusters. This provides an objective method for choosing a value for the resolution parameter that yields the most biologically meaningful clusters.

(a) Hybrids



(b) Restorer lines (CMS)

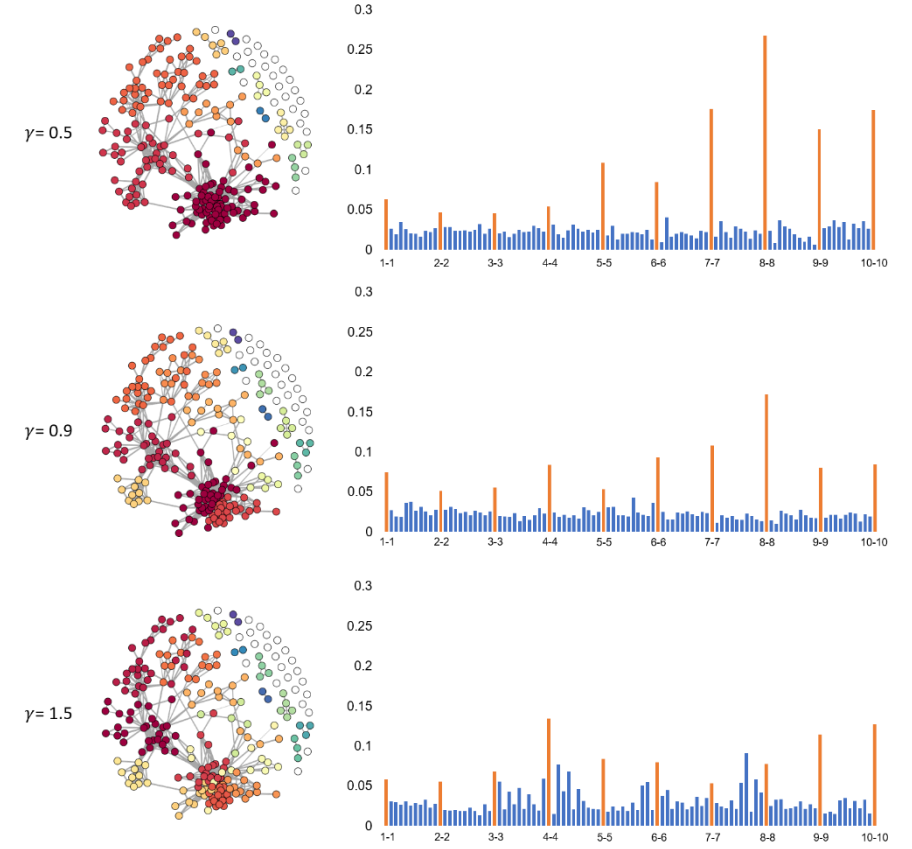


Fig. 15 | Impact of varying the resolution parameter γ on clustering results and the average genetic relatedness between varieties within and between clusters for restored hybrids (a) ($GD_{Th} = 0.19$) and restorer lines (b) ($GD_{Th} = 0.30$). Results for three different values of γ are shown (0.5, 0.9, 1.5). Average values of kinship (marker-based kinship coefficient) among varieties between clusters are shown in blue. Average values among varieties within clusters are shown in orange.

- *Testing the stability of communities*

The capacity to detect communities within the network, as well as the stability of the solution attained by the community detection algorithm, are dependent on the genetic threshold chosen to simplify the network by removing superfluous edges that describe ‘weak’ connections between varieties (i.e. large genetic distances) and mask information about the finer population structure of the collection.

To evaluate the consistency of the method for the chosen genetic thresholds, clustering results from ten independent runs were compared using data from 2016 and 2017. Results proved very robust for hybrids and restorer lines but more unstable for inbred lines. For restorer lines, the marginal fraction in which assignments could vary across runs concerned ~1 % of the total number varieties in test ($n = 234$). For other parental lines, the marginal fraction represented ~5 % of the collection ($n = 685$) (Fig. 16). Increasing resolution slightly increased stability for inbred lines but increased variability in the case of restorer lines. No impact was observed for hybrids.

It is possible to increase the stability of clustering results by adjusting the genetic thresholds. In the case of inbred lines, expert notes suggest it is possible to decrease the genetic threshold below 0.25 as no pair of varieties showing distinction problems was found above $GD = 0.20$ (Fig. 9). Lowering GD_{Th} to 0.24 significantly increased the stability of clustering results, whereas setting the threshold higher to 0.30 increased the marginal-to-stable fraction ratio to 40/60 (Fig. 17). This illustrates the importance of setting genetic thresholds carefully to optimize clustering results. If the thresholds are too high, the algorithm may not be able to find a stable solution. Several clustering solutions may be possible, and some varieties might end up being variably assigned to one cluster or another.

Based on the results of these tests, we propose the following thresholds:

- **Hybrids:** 0.19
- **Restorer lines (CMS):** 0.30
- **Inbred lines:** 0.24
- **Sterile lines:** 0.25

These values are calibrated to the French and German reference collections and can still be refined. However, a comparison of the clusters returned by the method using these thresholds and varietal listings from 2008 until 2017 showed clustering results to be very stable and coherent between years.

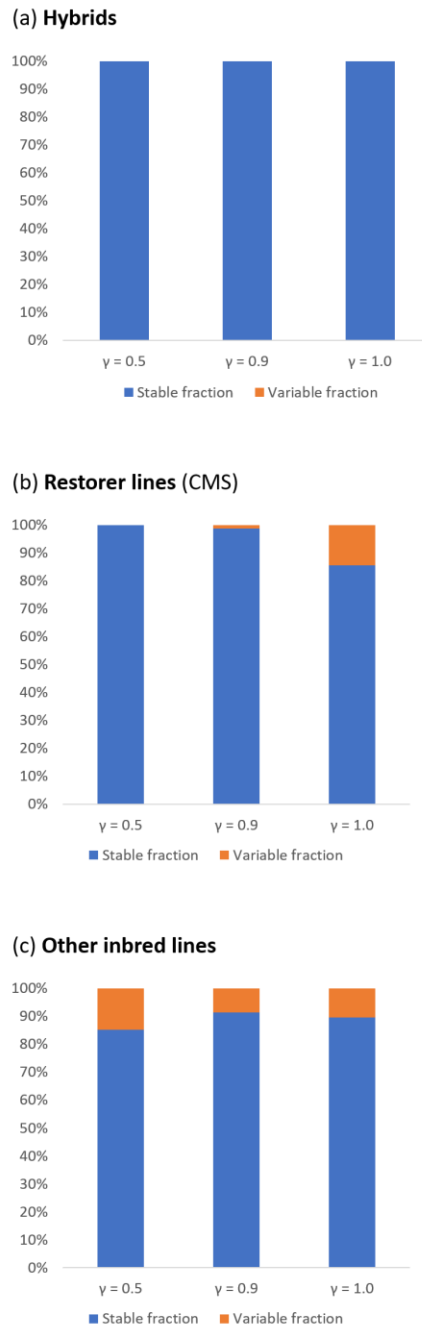


Fig. 16 | Impact of varying the resolution parameter on the stability of clustering results. Three values of γ were tested for hybrids **(a)** ($GD_{Th} = 0.19$), restorer lines **(b)** ($GD_{Th} = 0.30$) and other inbred lines **(c)** ($GD_{Th} = 0.25$). The stable-to-variable fraction ratio was assessed by comparing clustering results between ten independent runs with three different values of γ (0.5, 0.9, 1.5). The stable fraction (blue) represents varieties whose cluster membership did not vary across the runs. The variable fraction (orange) represents varieties whose cluster membership varied across runs.

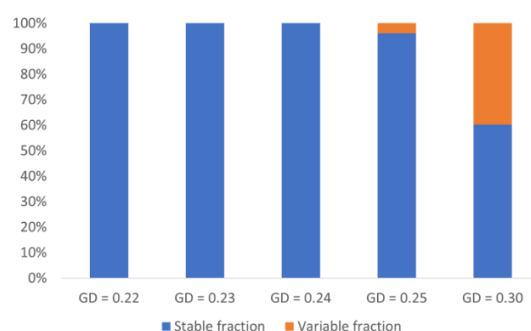


Fig. 17 | Impact of varying the genetic threshold on the stability of clustering results for inbred lines. The stable-to-variable fraction ratio was assessed by comparing clustering results between ten independent runs for different values of GD_{Th} . The stable fraction (blue) represents varieties whose cluster membership did not vary across the runs. The variable fraction (orange) represents varieties whose cluster membership varied across runs.

- Singletons

As we lower the genetic threshold and remove additional edges, some varieties may become disconnected from the network. These varieties are termed ‘singletons’. Although most singletons are reference varieties, some candidates might also end up disconnected from the network. The network approach suggests that these candidates are genetically distinct from all other varieties in the collection (at the chosen threshold). However, this does not mean that singleton candidates should be considered as ‘distinct +’. Under the new model we propose, molecular data is not used to declare varieties as distinct but to guide the selection of reference varieties and select the most pertinent subsets of the reference collection to include in growing trials, based on the molecular profile of candidate varieties. Singleton varieties will therefore be systematically included in field trials alongside other candidates and reference varieties selected by the method. However, it is also possible to run the network analysis iteratively for these singleton candidates, increasing each time the genetic threshold by small increments (e.g. 0.01) to identify the next closest reference variety(s).

iii. Applicability of the network approach to the GAIA system (FR)

a. Methodology

We tested our method using lists of varieties from previous DUS campaigns (2008-2017) provided by GEVES and compared clustering results with pairs of varieties flagged by GAIA as possibly requiring direct comparisons in the second year of DUS testing. The aim was to check whether excluding part of the reference collection from growing trials would lead to overlooking decisive comparisons for establishing distinctness and failing to detect distinction problems with candidate varieties.

b. Impact

Comparisons with historical data did not reveal any instance where our network-based clustering approach excluded a variety that experts would have wanted to compare side-by-side. With no exception, pairs of varieties that were judged ‘very similar’ (expert note = 1) clustered together, meaning also the corresponding varieties would have been included in growing trials. Checking retrospectively experts’ decisions at the end of the second DUS cycle on pairs of varieties flagged by GAIA, after the first cycle also revealed many instances were pairwise comparisons turned out to be unnecessary. In these cases, varieties had been often declared ‘distinct’ at an early stage of the DUS procedure and received high expert notes (4 or 5). Varieties within these pairs also often (but not always) fell into distinct clusters. None of the candidate varieties that ended up as singletons showed any problem of distinctness, and all were declared distinct after one or two years of DUS testing.

iv. Applicability of the network approach to the COYD system (DE)

Similarly, we evaluated the incidence of the new model on the calculation of COY statistics to evaluate whether reducing the size of the effective reference collection would impact DUS decisions. In the combined-over-years method for distinction (COYD) methodology, a pair of varieties is considered distinct if the difference between their over-years means is greater than the Least Significant Difference (*LSD*) observed for at least one character.

LSD is calculated as follows:

$$LSD_p = t_p \times \sqrt{2} \times SE(\bar{x})$$

where t_p is the corresponding value on the Student's t distribution table for $n-1$ degrees of freedom at the significance level p , and $SE(\bar{x})$ is the standard error of the mean $\bar{x}$.

As n (the size of the effective reference collection) decreases, t_p , $SE(\bar{x})$ and LSD_p should all increase, so that for a given character the difference between varietal over-years means should be even greater (compared to the full experimental setting) for a pair to be declared distinct for that character. In theory, therefore, reducing the size of the reference collection in growing trials should result in more conservative DUS decisions. However, because the choice of reference varieties is guided by the genotype of candidate varieties and leads to targeting in priority reference varieties that are genetically related the most to the candidates, assuming a strong influence of the genotype in the expression of morphological characters, the non-random selection of reference varieties could also lead to a contraction around the mean (i.e. less variation) and a lower LSD_p .

a. Methodology

We ran COYD analysis before and after applying the network-based optimization method, using varietal listings and morphological data from past DUS campaigns provided by BSA. Data from seven years (2011-2017) were used to simulate five full 3-year DUS campaigns (2011-2013, 2012-2014, 2013-2015, 2014-2016, and 2015-2017), and results obtained with the complete reference collection were compared to those obtained with the reduced collection, in which only references from clusters that comprised one or more candidate varieties were included.

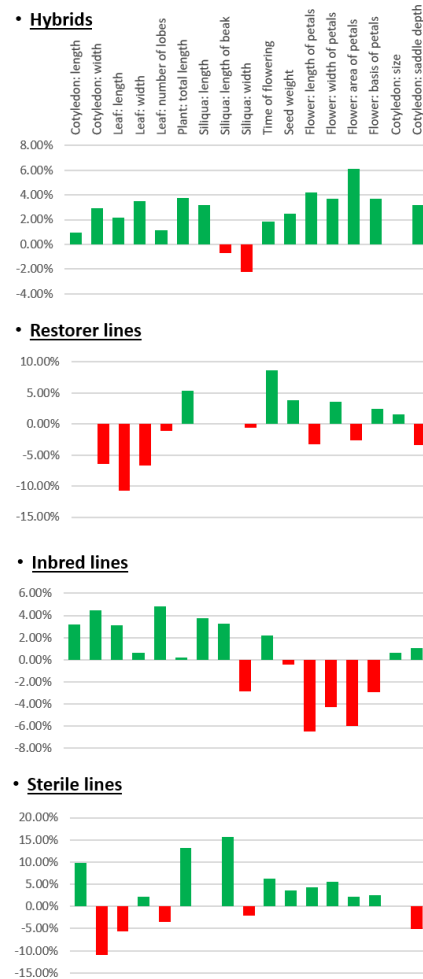
b. Impact

In practice, results showed that the impact of reducing the size of the reference collection varied greatly in direction (positive or negative) and magnitude, depending on morphological descriptors and varietal types (Fig. 18). In most cases, changes to LSD_p values had no impact on DUS decisions. Only in a few exceptions did reducing the size of the reference collection turn a difference between two varieties for a given character from 'significant' to 'non-significant', and even more rarely the other way round. In all cases, however, changes in the individual characteristics had no effect on the overall results and confirmed the original decisions made by DUS experts.

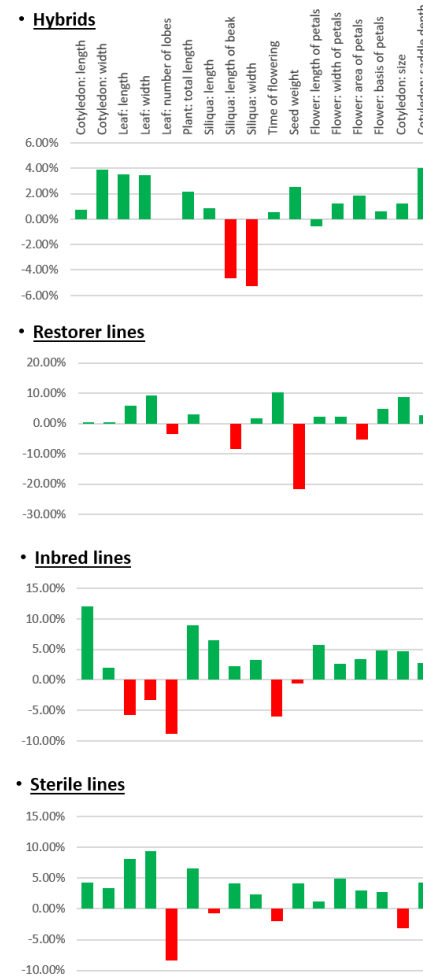
c. Limits

Because molecular data were available for only about 80% of the reference collection, the datasets used for comparing results of COYD analyses before/after running the model are only partially representative of the original data. As we move back in time, the representativity of the molecular dataset decreases, from 75% in 2017 to 53% in 2011. Therefore, the "full" dataset used in our simulations already represents a reduction of the reference collection compared to the original datasets. Furthermore, tests were carried out with quantitative characters only. Despite these limitations, results from those *in silico* COYD analyses with a reduced reference collection led to the same conclusions as the real trial.

(a) 2015-2017



(b) 2014-2016



(c) 2013-2015

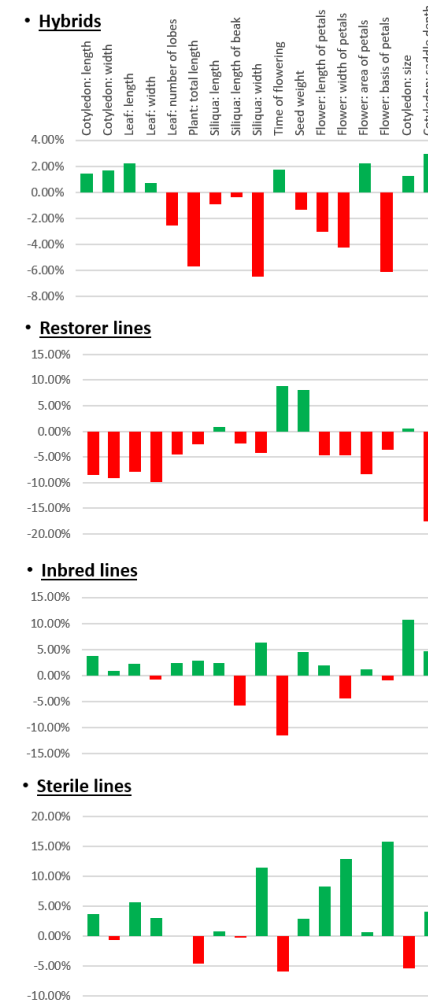


Fig. 18 | Impact of excluding part of the reference collection on the calculation of COYD 1% statistics for three 3-year DUS campaigns. (a) 2015-2017, (b) 2014-2016 and (c) 2013-2015. Depending on characteristics, reducing the number of varieties either increased (in green) or decreased (in red) the LSD_p .

v. Implementation

For the proposed model to be compatible with the COY method, morphological data must be available for all varieties included in the system for two to three years in a row. To ensure that the set of reference varieties (SRV) deduced from candidate varieties remain in the system for the whole duration of DUS tests, the dataset used to run the network-based method should include candidate varieties from the corresponding year but also candidates from the two previous years (Fig. 18). The collection might also include a number of example varieties established by each EO and that may be revised periodically to reflect changes in the conditions of trials.

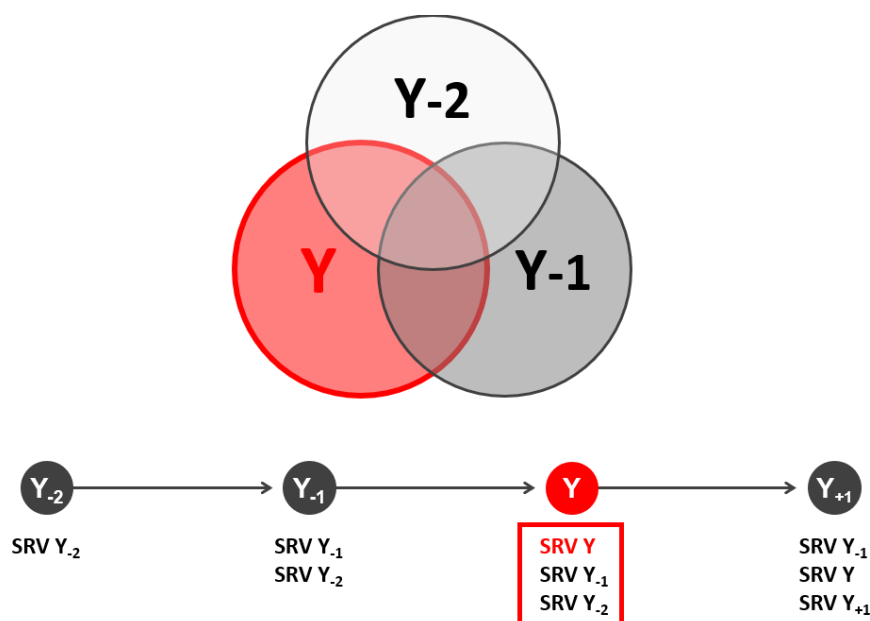


Fig. 19 | Rotation of varieties in the optimum reference collection over a 3-year rolling period. In any given year Y, the reference collection will include the set of reference varieties (SRV) deduced from the candidates' genotype (SRV_Y) and reference varieties for the two previous years (SRV_{Y-1} and SRV_{Y-2}).

The molecular reference library will need to be updated every year with genotypic information from candidate varieties and will need to also include any new variety that was listed abroad and added to the reference collection.

vi. Varietal turn-over

As the proposed method involves excluding a different part of the reference collection every year, we compared the composition of the set of reference varieties (SRV) identified by the method over a 10-year period (2008-2017) to analyse the varietal turn-over, using lists of candidate varieties provided by GEVES.

In theory, there should be enough variability among candidates from one year to another to ensure that most varieties within the collection get described periodically. In practice, however, a small fraction of the collection was left out. Turnover was highest for hybrids, with 95% of reference varieties¹⁸ being selected at least once and 29% at least twice (Fig. 20-A). Only 0.3% were systematically included in SRVs, which is coherent with the current DUS protocol, in which only hybrids that share a common parent with one or several candidate varieties are included in growing

¹⁸ Varieties for which genetic data are available.

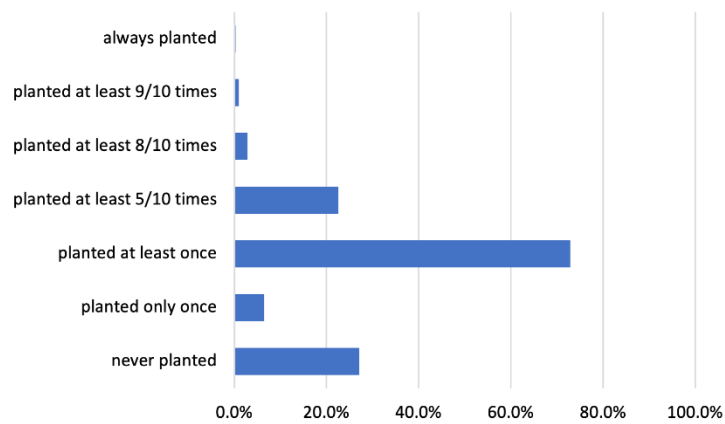
trials. The probability that a variety would be selected every year is therefore very low. In comparison, 90% of restorer lines were selected at least once and nearly a third (28%) were systematically included in field trials (Fig. 20-B). This is also not surprising, as some varieties are very popular and used regularly in breeding schemes to restore fertility in CMS hybrids¹⁹. Similarly, many lines are very popular in the GMS system. Nearly half of inbred lines (49%) were selected 8/10 times. 19% of the collection were systematically included in SRVs and nearly as much were never selected over the 10-year period tested (Fig. 20-C).

These results indicate that even with a model that excludes part of the reference collection every year, there would be a good varietal turnover in all categories²⁰. However, as the composition of the optimal reference collection depends on the genetic background of candidate varieties, in the long run we can expect part of the reference collection to be gradually excluded from field trials as breeding efforts shift in focus and the genetic makeup of candidate varieties evolves.

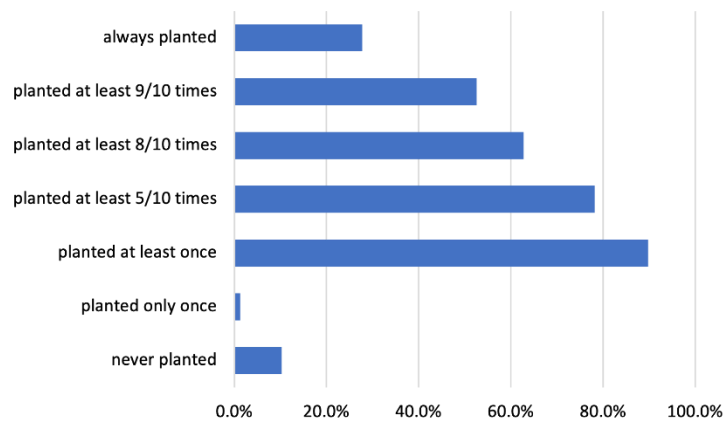
¹⁹ Two varieties have been used to restore fertility in close to 50 different CMS hybrids. Similarly, more than 20 lines have been used as restorers in GMS based hybrids.

²⁰ As this exercise was done using data from GEVES only (GAIA system), the varietal turnover in the case of sterile lines was not evaluated but should be comparable to other lines. Under a COYD framework, where varieties in trials must remain in the system for the whole duration of tests (2-3 years), the varietal turnover is expected to be generally higher compared to GAIA.

(a) Hybrids



(b) Restorer lines (CMS)



(c) Other inbred lines (including sterile lines)

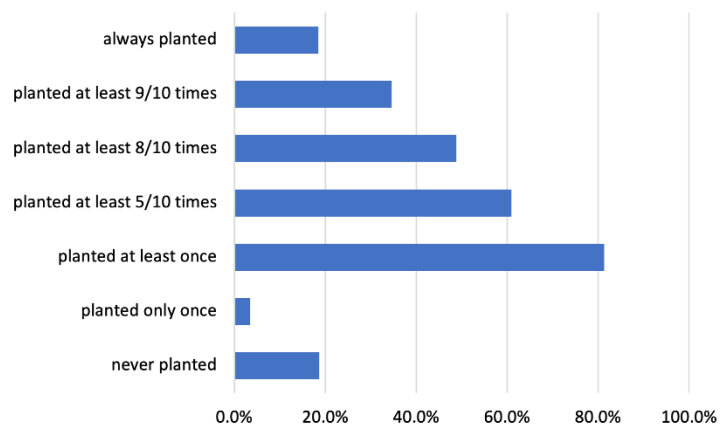


Fig. 20 | Varietal turn-over for hybrids (a), restorer lines (b) and other inbred lines (c) over a 10-year period (2008-2017).

Milestone 3: Feasibility study

DUS testing in winter oilseed rape is highly constrained by time, with a very short window between application, reception of seed lots for candidate varieties and sowing deadline. Usually, EOs receive applications and seed samples at the start of August, while sowing takes place in France in September. To have sowing lists ready on time, genotypic data should be ready by August, 21st at the latest, leaving a window of approximately two weeks to genotype all candidate varieties and their parental lines (about 200 accessions each year), analyse the data and communicate results to DUS experts so that they can draw the lists of reference varieties to be included in growing trials. In Germany, the sowing date is end of August / beginning of September, the whole process has to be adopted to an even smaller window.

The project evaluated the possibility to outsource genotyping to external providers for EOs that do not have in-house facilities to carry out molecular analyses. The aims were to evaluate the transferability of the markers and to evaluate the concordance of results obtained using different genotyping technologies.

1) Choice of external providers: KASPar vs. NGS-based technologies

We carried out a comparative analysis of cost-efficiency of the services provided by two sequencing platforms, each using a different technology. Whereas Provider A uses the same technology as GEVES (KASPTM genotyping assay), Provider B uses a targeted genotyping-by-sequencing approach.

Building upon the development of high-throughput sequencing technologies, genotyping-by-sequencing has become very cost-effective and allows genotyping hundreds or thousands of individuals at hundreds or thousands of custom-designed loci simultaneously. Targeted sequencing is a next generation sequencing (NGS)-based multiplexed PCR amplicon sequencing approach in which only target SNPs are amplified using specific PCR primers. Unlike SNP arrays that have fixed marker density and SNP targets, the number of SNPs that can be included is flexible (up to 100 000 markers) and scalable with the number of samples to genotype.

DNA extractions were performed by the providers. Each provider received bulks of 32 seeds from 72 varieties, including the 23 varieties from the pilot project, each of which was repeated once (46 samples in total), and 49 varieties chosen to represent different hybrid types (simple restored hybrids, three-way hybrids, sterile hybrids), with their parental lines. Many varieties from this list were selected from those for which we had technical replicates and pseudo-replicates (distinct seed lots). The list also included ten “problematic” varieties - five with a high percentage of missing data (>20%) and five (inbred lines) with a high percentage of heterozygosity (>50%). In total, the dataset comprised 95 samples.

2) Evaluation of service

i. Transferability

Because Provider A uses the same technology as GEVES, there was no experimental set-up required and the platform was asked to process the 95 samples with the set of 360 markers that we retained for analysis. In contrast, transferring the markers to the NGS-based technology (Provider B) required designing *de novo* primers using a reference genome and precise information about the position of the SNPs on each chromosome. The availability of a complete genome of reference is a prerequisite

for designing high-quality oligo probes and ensure high accuracy of variant calling. It is necessary to avoid repetitive regions and for assessing the number of potential off-target hits²¹ for each probe.

²¹ Off-target hits mean that the probe may bind to several positions in the genome. Each binding site that is not near the targeted SNP is considered “off-target”. Probes with more than one match in the genome usually fail.

Provider B was asked to genotype all 95 samples using the 443 SNP set we initially retained after the first quality screening step. Because Provider B uses a sequencing-based approach, we chose to include the 83 SNPs that we originally discarded to test whether better results could be achieved for these markers using a different technology. Designing the probes took about a week. Provider B returned different sets of primers, either with high specificity levels (no off-target hit) or with normal specificity levels (up to four off-target hits). The design failed for one SNP (C09-484) that mapped more than 10x to the reference genome despite the marker working very well in the KASP assay. High specificity probes (forward and reverse) were ordered for 301 SNP loci and a mixture of high and normal specificity probes were ordered for the remaining 58 SNPs.

ii. Flexibility, delays/reactivity, availability

Both providers received the samples mid-January and sent back results in March. Producing the primer kit for the NGS-based technology took five weeks. Once primers were ready, sequencing and data analysis took only a week. Altogether, both platforms took two months to return results, but these delays can be considerably shortened in a routine situation where the genotyping project can be scheduled and given priority. For the targeted sequencing approach, there will also be no need to redesign the primers. Producing the primer kit can take up to six weeks, but once the set of markers is fixed this step can be anticipated, and Provider B typically sends back the results within a week. Both providers requested samples to be sent as powders for DNA extractions. Preparing seed bulks is labour-intensive, and time must be factored in for the preparation of samples when planning analyses. Finally, preparing samples as powder require specific equipment for grinding seeds (e.g., mixer mill), although it can also be done with a food processor (providing that blades are thoroughly cleaned between samples to avoid contamination).

iii. Deliverables

Provider A returned raw data and provided the software (Fluidigm SNP Genotyping Analysis software) for analysing the results. This has the advantage of ensuring results consistency as the end-user is calling the genotypes, but analysing data takes time. Provider B sent the raw sequencing data in FastQC format and genotype data in Variant Calling format (VCF). Data were sent on an encrypted USB key containing a report and a breakdown of the bioinformatic workflow, with distinct folders containing data after each step (quality trimming, adapter clipping, and alignments to the reference genome). Some information on the tools used to analyse the data is also provided (Bowtie2²² for sequence alignment, FreeBayes²³ for variant calling) so that it is possible to repeat the analysis if necessary.

iv. Comparability of results: reproducibility, repeatability

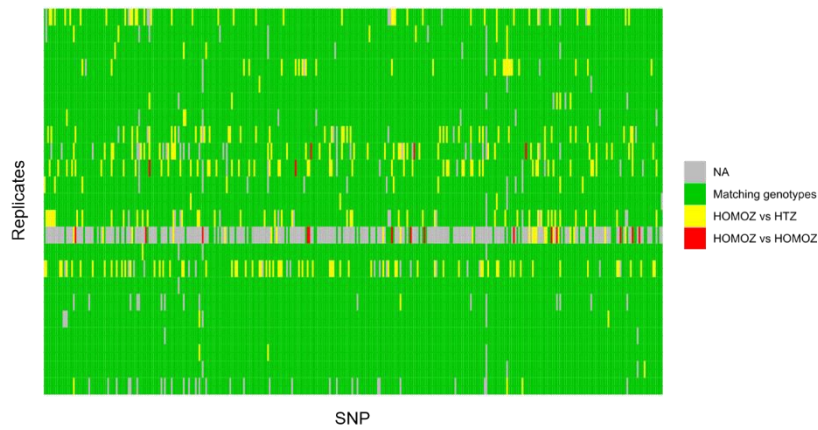
- Concordance between replicates

As a mean to evaluate repeatability and the reliability of the marker set, each provider was asked to aliquot the DNA extracted from the set of 23 varieties that were used for the pilot project, and to genotype each aliquot as a technical replicate. Genotypes were 95% concordant between replicates for Provider A (Fig. 21-A) and 98.2% for Provider B (Fig. 21-B), when excluding missing data.

²² <http://bowtie-bio.sourceforge.net/bowtie2/index.shtml>

²³ <https://github.com/freebayes/freebayes>

(a) **Provider A** (KASP assay)



(b) **Provider B** (targeted sequencing)

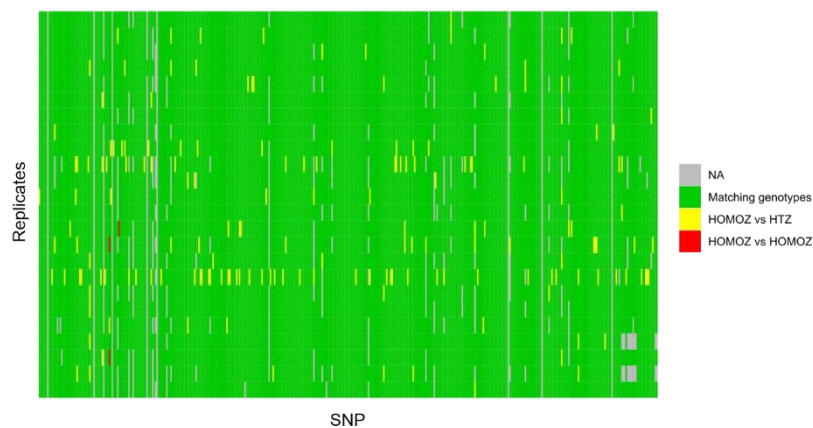


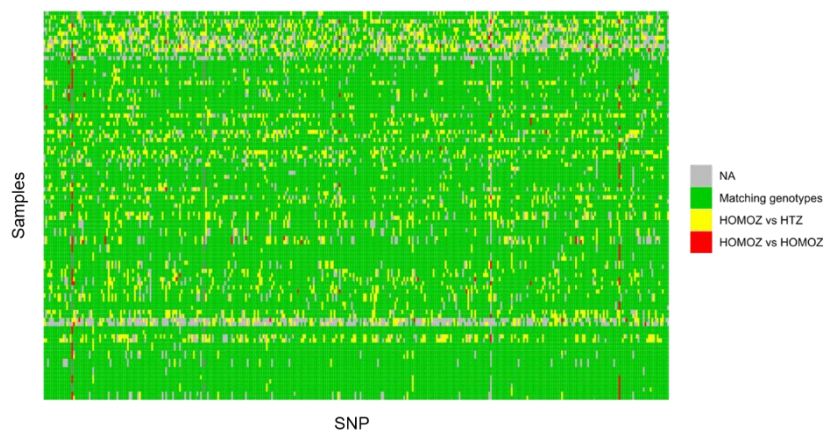
Fig. 21 | Concordance between replicates for (a) Provider A (KASP assay, 360 SNPs) and (b) Provider B (targeted sequencing approach, 442 SNPs). Each provider was asked to repeat genotyping for 23 varieties (46 samples in total).

The lower concordance rate for Provider A is partly the result of a problem during DNA extraction for one variety (W0527), for which data was missing for 70% of the markers (247/360) and discordant for 32% of the SNPs that could be scored in both samples. There were also many instances where a variety was called homozygous at one locus and heterozygous at the same locus in the technical replicate. These discrepancies were probably caused by insufficient recycling and lack of clear separation between clusters, resulting in difficulties to call the genotypes. Concordance between replicates was very high for Provider B but seven SNPs failed to amplify across all 23 samples tested.

- *Concordance between platforms*

Across platforms, results were 91.4% concordant between GEVES and Provider A, with three loci (A02-019, C05-341, and C08-447) showing high levels of disagreement (35.6%, 35.7% and 49.5%, respectively), two of which (A02-019 and C05-341) had also a high percentage of missing data (55.2% and 37.5%, respectively) (Fig. 22-A). The percentage of concordance was higher between GEVES and Provider B (95%) for the 359 SNPs for which primers could be designed. Three SNPs did not work (A03-057, C04-303, and C06-394), and two SNPs (A02-021 and C05-356) had very low levels of concordance (14.7% and 22.6%, respectively) (Fig. 22-B). Rates do not include the ten problematic varieties for which, as expected, there was a high rate of disagreement.

(a) **Provider A** (KASP assay)



(b) **Provider B** (targeted sequencing)

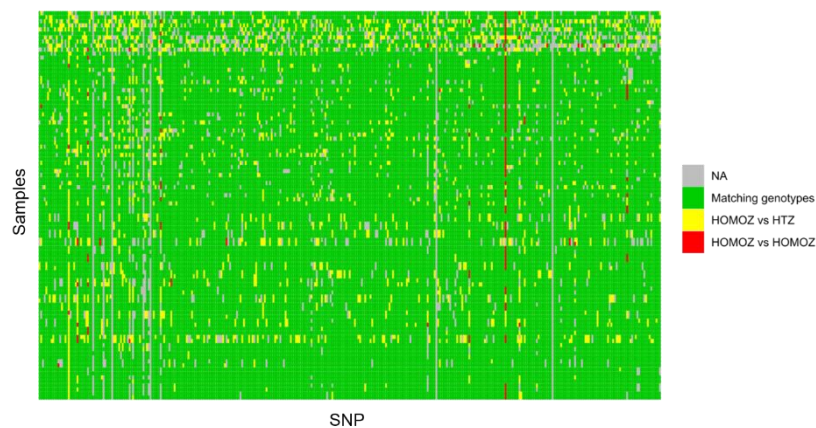


Fig. 22 | Concordance between platforms. Comparison of genotype calls for 96 samples between (a) GEVES and Provider A and (b) between GEVES and Provider B. Comparisons were performed for the 359 SNPs common to all platforms (Provider B could not design primers for C09-484).

These results suggest that a final screening step may be necessary to optimize the SNP set. We identified four markers that showed high discrepancies between the three datasets (A04-069, A05-520, C04-328, C05-356). This might indicate that the corresponding SNPs are either multiallelic or multilocus and that it might be better not to include them in the final set of markers.

v. Value for money

Genotyping 96 samples at 360 SNPs with Provider A cost € 10 000 (about € 100 per sample), with the synthesis of KASP oligos representing alone $\approx$ 50% of the cost. To calculate the true cost per sample, it is necessary to consider the number of PCR reactions that can be performed from the minimum amount of assay that can be ordered. A single KASP assay aliquot can be used to perform 2 500 genotyping reactions, while only 96 reactions were necessary for the ring test.

In comparison, genotyping with Provider B cost € 4 000. There was no difference in price for analysing 360 or 443 SNPs (i.e. a fixed cost of € 40 per sample). Here too, the actual cost depends on the quantity of primers ordered and would decrease in the context of routine analysis, with the larger the genotyping project the better the value in terms of cost per datapoint.

3) Conclusions

The goal of the ring test was not to compare providers in terms of quality or reliability but to test the feasibility of outsourcing molecular work to a third-party to assess the transferability of the set of markers and the reproducibility of genotyping results.

As part of the pilot project, GEVES and NIAB conducted an initial ring test to assess the reproducibility of results between laboratories. Both GEVES and NIAB used the KASP™ array technology but while GEVES used the Kraken software (LGC Genomics) to call alleles, NIAB used the QuantStudio™ Real Time PCR software (Applied Biosystems). Results from this initial ring test stressed the need for harmonizing protocols for reading and calling genotypes. The comparison of results between GEVES and Provider A highlights similarly the need for setting clear technical specifications for the genotyping protocol, in particular the number of recycling steps to be applied for each SNP, and setting guidelines for interpreting results, which have been shown to be highly influenced by interindividual variability in calling alleles.

Targeted sequencing yields comparatively more objective results as it is not subject to visual inspection of clustering results, and thus does not depend on the experience of the technician in charge of analysing the data. However, data quality is highly dependent on the initial design of the probes. The experiment showed that not all SNPs could be transferred onto the NGS platform and that several loci that worked well for KASP did not amplify after conversion. Finally, expertise in bioinformatics, while not essential, would be advantageous to rerun analyses, if required, to confirm genotypes.

Milestone 4: Final meeting

The final meeting took place online on the 8th of September 2021 with members of CPVO and representatives of UKZUZ (Czech Republic), TystofteFoundation (Denmark), GEVES (France), BSA (Germany), COBORU (Poland), UKSUP (Slovakia), and NIAB (United-Kingdom).

1) Conclusions

The method performance is constrained by marker density and the genetic heritability of DUS traits, which can be quite complex. Ensuring complete and even coverage of the genome is a prerequisite for the method to avoid any bias. Through this project, we identified an optimized set of 360 SNPs markers with good genome coverage and more than double the resolving power required to discriminate among all varieties in a molecular library that represents over 80% of the French and German reference collections for WOSR (Fig. 4).

Using this data, we proposed a new approach that uses network analysis for defining an “optimal reference collection” whose composition is deduced from the candidates’ genotypes. In this model, graph clustering tools are used to identify clusters of related varieties based on their genetic proximity within a network representing genetic relationships between candidate and reference varieties, and only references from clusters that comprise candidate varieties are included in growing trials. Using the thresholds that we defined, it is, in theory, possible to reduce the size of DUS growing trials by 20 to 35% on average in the first cycle, depending on whether COYD or GAIA is used to evaluate distinctness. The model is compatible with both statistical frameworks and accommodates for differences in DUS protocols among EOs without impacting DUS decisions or leading to exclude reference varieties that are important for DUS testing.

A schematic view of the implementation of the model is described in Fig. 23. For the proposed model to work, the reference library of molecular profiles needs to be updated every year to include varieties that have been added recently to the reference collection (i.e. listed in other countries). Once the closing date for applications has passed, candidate varieties are genotyped using SNP markers and pairwise genetic distances are calculated against all varieties from the reference collection and against all other candidates. The matrix is then transformed into a network and analysed as described in Fig. 8.

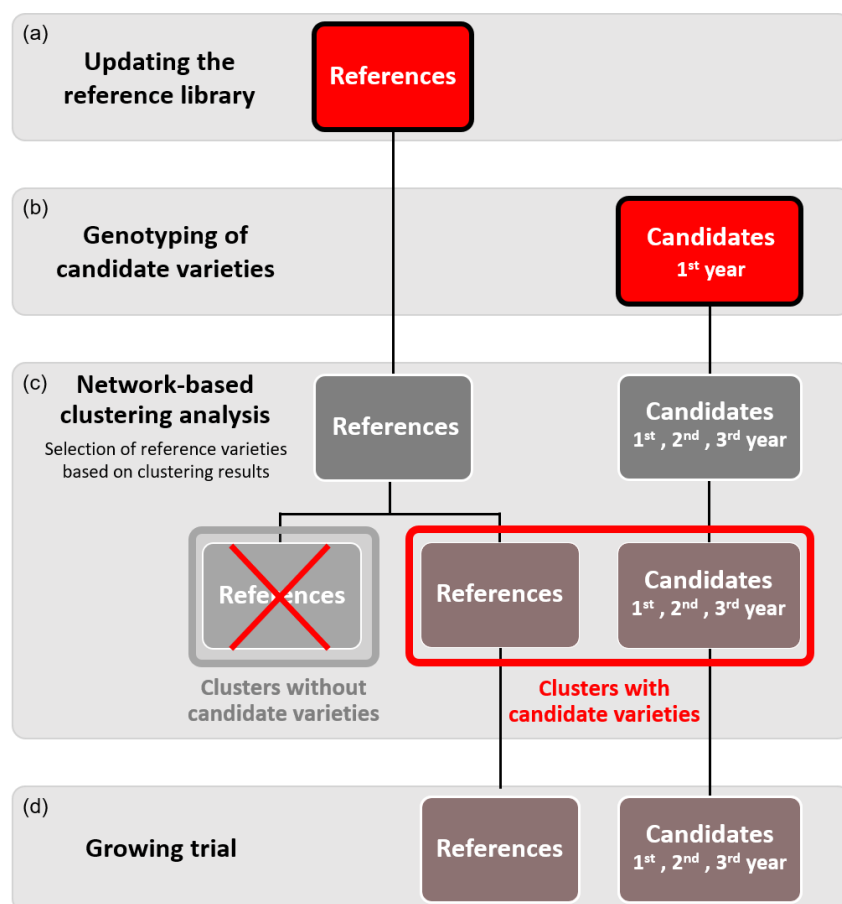


Fig. 23 | A schematic view of the proposed model for integrating DNA profiles in DUS testing of winter oilseed rape varieties. Every year, the reference library of molecular profile is updated to include varieties that have been recently added to the reference collection **(a)**. Once all applications have been received, new candidates are genotyped **(b)** and a matrix of pairwise genetic distances is calculated (one for each varietal category recognized by DUS experts²⁴) against the updated reference collection and against all other candidate varieties under DUS examination. The matrix is then analysed using the network optimization statistical workflow which returns a table with clustering results **(c)**. Only clusters that include one or several candidate varieties are selected for inclusion in growing trials; the rest of the reference collection is excluded **(d)**.

While the clusters identified by the method encompass primarily varieties for which pairwise genetic distances do not exceed the chosen threshold, they also comprise varieties that are genetically more distant but related to candidates via one or several intermediary varieties (i.e. second and third-order neighbours). In the network-based approach, distances between pairs of varieties are considered in relation to the entire collection and references are selected based on their proximity (i.e. their genetic similarity) with candidate varieties in the network.

The network-based optimization method is implemented in R, an open source and cross-platform statistical tool. It is fast and easy to implement and takes less than ten minutes to run with minimal assumptions and only two parameters to set (the genetic threshold and the resolution parameter). The conceptual approach is compatible with all types of markers (SSR, SNP, morphological...) and is independent of the distance used (genetic or otherwise). Depending on how the model performs

²⁴ E.g. restored hybrids, three-way hybrids, sterile hybrids, restorer lines, sterile lines, high erucic acid or high oleic/low linoleic acid content varieties, etc.

under the conditions of a real field trial, it could be also tested on other agricultural species with large reference collections (e.g. soybean [Achard *et al.* 2020]), for which it might be also advantageous to reduce the size of growing trials by selecting reference varieties more stringently.

2) Follow-up

So far, the method has only been tested *in silico* with genetic thresholds that were defined based on the French/German collections. The molecular dataset currently represents ~80% of the French/German technical collections. A prerequisite to test the model *in situ* will be to complete genotyping of the reference collection. Refining the molecular toolset is also an important task for a future project. Identifying a smaller, more manageable set of highly informative SNPs ('tag' SNPs) that suffices to resolve the genetic structure of the collection and provides enough information to help in selecting reference varieties would greatly help reducing the cost of analysis. This could also contribute to minimise technical constraints linked to the short window between the reception of candidate DUS seed lots and the start of DUS trials by speeding up the process of genotyping candidate varieties.

A Follow-up 2 project will be required to test the model in the field. Further interested EOs may be included. Testing the model *in situ* will not necessarily require a dedicated trial or experimental setting. However, implementing the method in the field will require adjustments to be made to ensure compatibility with current DUS protocols, e.g. ensuring that parental lines from hybrids in testing are included in field trials and that the optimal reference collection has to be based on candidates in the 1st, 2nd and 3rd cycle when COYD is used. Prior to testing the model, each EO will need to create a molecular dataset representative of their own reference collection, including all candidate varieties from the relevant test year. The genetic thresholds we defined can be useful as preliminary values for exploring data but are valid for the French and German collections only. Each EO interested in testing the method will need to calibrate these thresholds using their own reference collection to reflect regional differences in agri-environmental conditions and in the genetic makeup of candidate varieties which will reflect different focus in breeding efforts. Flagging pairs of varieties with distinction problems is also important as those are useful to refine the genetic thresholds. It is proposed to arrange a meeting with interested parties, in the first semester of 2022, to discuss how to move forward with a further project.

References

- Achard F, Butruille M, Madjarac S, Nelson PT, Duesing J, Laffont JL, Nelson B, Xiong J, Mikel MA, Smith JSC (2020) Single nucleotide polymorphisms facilitate distinctness-uniformity-stability testing of soybean cultivars for plant variety protection. *Crop Science* **60**, 2280–2303. <https://doi.org/10.1002/csc2.20201>
- Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19**, 1655–1664. <https://doi.org/10.1101/gr.094052.109>
- Biedermann S, Henzinger M, Schulz C, Schuster B (2020) Vienna Graph Clustering. *Methods in Molecular Biology* **2074**, 215–231. https://doi.org/10.1007/978-1-4939-9873-9_16
- Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **10008**, 6. <https://doi.org/10.1088/1742-5468/2008/10/P10008>
- Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics* **81**, 1084–1097. <https://doi.org/10.1086/521987>
- Chen L, Liu Y-G (2014) Male sterility and fertility restoration in crops. *Annual Review of Plant Biology* **65**, 579–606. <https://doi.org/10.1146/annurev-arplant-050213-040119>
- Clarke WE, Higgins EE, Plieske J, Wieseke R, Sidebottom C, Khedikar Y, Batley J, Edwards D, Meng J, Li R, Taylor Lawley C, Pauquet J, Laga B, Cheung W, Iniguez-Luy F, Dyrzka E, Rae S, Stich B, Snowdon RJ, Sharpe AG, Ganai MW, Parkin IA (2016). A high-density SNP genotyping array for *Brassica napus* and its ancestral diploid species based on optimised selection of single-locus markers in the allotetraploid genome. *Theoretical and Applied Genetics* **129**, 1887–1899. <https://doi.org/10.1007/s00122-016-2746-7>
- Corander J, Waldmann P, Sillanpää MJ (2003) Bayesian analysis of genetic differentiation between populations. *Genetics* **163**, 367–374. <https://doi.org/10.1093/genetics/163.1.367>
- CPVO (2008) Management of winter oilseed rape reference collections (final report). CPVO Project Nr. CPV5766.
- CPVO (2011) A potential UPOV Option 2 approach for barley using high density SNP genotyping (final report). CPVO Project Nr. EPM 7501705.
- CPVO (2018) Test of the potential use of SNP markers on oilseed rape varieties (final report). CPVO Project Nr. 7510544.
- CPVO (2020) Protocol for tests on distinctness, uniformity and stability. *Brassica napus* L. emend. Metzg. Rape seed (adopted 21/04/2020). TP/036/3. https://cpvo.europa.eu/sites/default/files/documents/brassica_napus_emend_3.pdf
- Csardi G, Nepusz T (2006) The igraph software package for complex network research. *InterJournal, Complex Systems* **1695**, 1–9. <https://igraph.org>
- EUROSTAT (2021) Crop production in EU standard humidity (apro_cpsh1). [Data file]. Retrieved from https://ec.europa.eu/eurostat/databrowser/view/apro_cpsh1/ (last accessed 07 July 2021)
- Fletcher RJ, Revell A, Reichert BE, Kitchens WM, Dixon JD, Austin JD (2013) Network modularity reveals critical scales for connectivity in ecology and evolution. *Nature Communications* **4**, 2572. <https://doi.org/10.1038/ncomms3572>
- Fortunato S (2010) Community detection in graphs. *Physics Reports* **486**, 75–174. <https://doi.org/10.1016/j.physrep.2009.11.002>
- Fortunato S, Barthelemy M (2007) Resolution limit in community detection. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 36. <http://dx.doi.org/10.1073/pnas.0605965104>
- Gilliland TJ, Annicchiarico P, Julier B, Ghesquière M (2020). A proposal for enhanced EU herbage VCU and DUS testing procedures. *Grass and Forage Science* **75**, 227–241. <https://doi.org/10.1111/gfs.12492>
- Goudet J, Kay T, Weir BS (2018) How to estimate kinship. *Molecular Ecology* **27**, 4121–4135. <https://doi.org/10.1111/mec.14833>
- Gourrion A, Simon C, Vallée P, Delourme R, Chatre S, Dheu J-É (2020). Enlarging the genetic diversity of winter oilseed rape (WOSR) by crossing with spring oilseed rape (SOSR). *OCL (Oilseeds & fats Crops and Lipids)* **27**, 16. <https://doi.org/10.1051/ocl/2020013>
- Gower JC (1971) A general coefficient of similarity and some of its properties. *Biometrics* **27**, 857–871. <https://doi.org/10.2307/2528823>

- Greenbaum G, Rubin A, Templeton AR, Rosenberg NA (2019) Network-based hierarchical population structure analysis for large genomic data sets. *Genome Research* **29**, 2020–2033. <https://doi.org/10.1101/gr.250092.119>
- Greenbaum G, Templeton AR, Bar-David S (2016) Inference and analysis of population structure using genetic data and network theory. *Genetics* **202**, 1299–1312. <https://dx.doi.org/10.1534/genetics.115.182626>
- Harenberg S, Bello G, Gjeltrema L, Ranshous S, Harlalka J, Seay R, Padmanabhan K, Samatova N (2014) Community detection in large-scale networks: a survey and empirical evaluation. *Wiley interdisciplinary reviews. Computational Statistics* **6**, 426–439. <https://doi.org/10.1002/wics.1319>
- Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics* **11**, 94. <https://doi.org/10.1186/1471-2156-11-94>
- Jones H, Norris C, Smith D, Cockram J, Lee D, O’Sullivan DM, Mackay I (2013) Evaluation of the use of high-density SNP genotyping to implement UPOV Model 2 for DUS testing in barley. *Theoretical and Applied Genetics* **126**, 901–911. <https://doi.org/10.1007/s00122-012-2024-2>
- Kelly ST (2020) *leiden*: R implementation of the Leiden algorithm. R package version 0.3.8. <https://github.com/TomKellyGenetics/leiden>
- Kuismin MO, Ahlinder J, Sillanpää MJ (2017) CONE: community oriented network estimation is a versatile framework for inferring population structure in large-scale sequencing data. *G3: Genes, Genomes, Genetics* **7**, 3359–3377. <http://dx.doi.org/10.1534/g3.117.300131>
- Lancichinetti A and Fortunato S (2009) Community detection algorithms: A comparative analysis. *Physical Review E* **80**, 056117. <https://doi.org/10.1103/PhysRevE.80.056117>
- Neuditschko M, Khatkar MS, Raadsma HW (2012) NetView: a high-definition network-visualization approach to detect fine-scale population structures from genome-wide patterns of variation. *PLoS One* **7**, e48375. <https://doi.org/10.1371/journal.pone.0048375>
- Newman MEJ, Girvan M (2004) Finding and Evaluating Community Structure in Networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* **69**, 026113. <http://dx.doi.org/10.1103/PhysRevE.69.026113>
- Ogura H (1968) Studies on the new male sterility in Japanese radish, with special references to the utilization of this sterility towards the practical raising of hybrid seeds. *Memoirs of the Faculty of Agriculture, Kagoshima University* **6**, 39–78. <http://hdl.handle.net/10232/00003764>
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959. <https://doi.org/10.1093/genetics/155.2.945>
- R Core Team (2021) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. <https://www.R-project.org/>
- Reichardt J, Bornholdt S (2006) Statistical mechanics of community detection. *Physical Review E* **74**, 016110. <https://doi.org/10.1103/PhysRevE.74.016110>
- Rogers JS (1972) Measures of genetic similarity and genetic distance. *Studies in Genetics* **7**, 145–153.
- Rotta R, Noack A (2011) Multilevel local search algorithms for modularity clustering. *Journal of Experimental Algorithmics* **16**, 2.3.1–27. <https://doi.org/10.1145/1963190.1970376>
- Slatkin M (2008) Linkage disequilibrium – understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics* **9**, 477–485. <https://doi.org/10.1038/nrg2361>
- Steinig EJ, Neuditschko M, Khatkar MS, Raadsma HW, Zenger KR (2016) NetView p: a network visualization tool to unravel complex population structure using genome-wide SNPs. *Molecular Ecology Resources* **16**, 216–227. <https://doi.org/10.1111/1755-0998.12442>
- Su X, Xue S, Liu F, Wu J, Yang J, Zhou C, Hu W, Paris C, Nepal S, Jin D, Sheng QZ, Yu PS (2021) A comprehensive survey on community detection with deep learning, in *IEEE Transactions on Knowledge and Data Engineering*. <https://doi.org/10.1109/TKDE.2021.3104155>
- Tommasini L, Batley J, Arnold G, Cooke R, Donini P, Lee D, Law J, Lowe C, Moule C, Trick M, Edwards K (2003) The development of multiplex simple sequence repeat (SSR) markers to complement distinctness, uniformity and stability testing of rape (*Brassica napus* L.) varieties. *Theoretical and Applied Genetics* **106**, 1091–1101. <https://doi.org/10.1007/s00122-002-1125-8>
- Traag VA, Waltman L, Van Eck NJ (2019) From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports* **9**, 1–12. <https://doi.org/10.1038/s41598-019-41695-z>
- UPOV (2002) Guidelines for the conduct of tests for distinctness, uniformity and stability in Rape seed (*Brassica napus* L. *oleifera*). TG/36/6 Corr. Geneva, Switzerland. <https://www.upov.int/edocs/tgdocs/en/tg036.pdf>

- UPOV (2003) GAIA software: Crop expert phenotypic distances between varieties. TWC/21/4. Technical working party on automation and computer programs, 21st session Tjele, Denmark, June 10–13, 2003. Geneva, Switzerland. https://www.upov.int/edocs/mdocs/upov/en/twc/21/twc_21_4.pdf
- UPOV (2007) GAIA software. TWC/25/13. Technical working party on automation and computer programs, 25th session Sibiu Romania, September 3–6, 2007. Geneva, Switzerland. https://www.upov.int/edocs/mdocs/upov/en/twc/25/twc_25_13.pdf
- UPOV (2010) Guidelines for DNA-Profiling: Molecular marker selection and database construction (BMT Guidelines). INF/17/1. Geneva, Switzerland. https://www.upov.int/edocs/infdocs/en/upov_inf_17.pdf
- UPOV (2017) The use of molecular markers (SNP) for maize DUS testing: Development and official applications to assess distinctness of hybrid varieties (France). International Union for the Protection of New Varieties of Plants (UPOV). 16th session of the Working Group on Biochemical and Molecular Techniques and DNA-profiling in particular, 7-10 November 2017, La Rochelle, France. BMT/16/10. https://www.upov.int/edocs/mdocs/upov/en/bmt_16/bmt_16_10.pdf
- UPOV (2019) Trial design and techniques used in the examination of distinctness, uniformity and stability. TGP/8. Geneva, Switzerland. https://www.upov.int/edocs/tgpdocs/en/tgp_8.pdf
- UPOV (2020) Examining distinctness. TGP/15/3. Geneva, Switzerland. https://www.upov.int/edocs/tgpdocs/en/tgp_15.pdf
- Ushey K, Allaire JJ, Tang Y (2021) *reticulate*: Interface to 'Python'. R package version 1.19. <https://CRAN.R-project.org/package=reticulate>
- Yang CJ, Russell J, Ramsay L, Thomas W, Powell W, Mackay I (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>
- Yang Z, Algesheimer R, Tessone CJ (2016) A comparative analysis of community detection algorithms on artificial networks. *Scientific Reports* **6**, 30750 <https://doi.org/10.1038/srep30750>
- Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328. <https://dx.doi.org/10.1093/bioinformatics/bts606>
- Zhou F, Mahler S, Toivonen H (2012) Simplification of networks by edge pruning. In: Berthold MR (eds) *Bisociative knowledge discovery. Lecture Notes in Computer Science*, vol 7250, pp. 179–198. Springer, Berlin, Heidelberg. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-31830-6_13
- Zhou Y, Browning SR, Browning BL (2020a) A fast and simple method for detecting identity-by-descent segments in large-scale data. *American Journal of Human Genetics* **106**, 426–437. <https://doi.org/10.1016/j.ajhg.2020.02.010>
- Zhou Y, Browning SR, Browning BL (2020b) IBDkin: fast estimation of kinship coefficients from identity by descent segments. *Bioinformatics* **36**, 4519–4520. <https://doi.org/10.1093/bioinformatics/btaa569>

Appendixes

i. Complete list of the 548 SNP markers tested

Characteristics of the 548 SNPs tested (including the 500 SNPs tested as part of the pilot project and the 48 new SNPs). SNPs are mapped on the Darmor-bzh v4.1 reference genome (Clarke *et al.* 2016).

SNP	Chromosome	Physical position	Genetic position	Alleles
BN_A01_001	A01	824790	5.2	A/C
BN_A01_002	A01	1680730	18.7	T/C
BN_A01_003	A01	3747996	33.8	T/C
BN_A01_004	A01	4856648	42.3	T/C
BN_A01_005	A01	5651413	48.4	T/C
BN_A01_006	A01	6154320	51.2	A/G
BN_A01_007	A01	6794703	54.4	A/G
BN_A01_008	A01	6916644	56.2	T/C
BN_A01_009	A01	8334397	65.2	A/G
BN_A01_010	A01	9984798	74.3	A/C
BN_A01_011	A01	15453898	79.8	T/C
BN_A01_012	A01	15487141	79.8	A/G
BN_A01_013	A01	15556380	79.8	A/G
BN_A01_014	A01	17004077	81.8	A/G
BN_A01_015	A01	18303706	84.8	A/G
BN_A01_016	A01	18901219	87.2	T/C
BN_A01_017	A01	19439099	87.2	T/C
BN_A01_018	A01	20206623	92.5	T/C
BN_A01_501	A01	1400752	14.9	T/G
BN_A01_502	A01	14240261	78.3	T/C
BN_A01_503	A01	20344118	93.8	T/G
BN_A01_504	A01	20465173	95.1	A/C
BN_A01_505	A01	21484318	106.1	A/G
BN_A01_506	A01	1939611	20.5	A/G

SNP	Chromosome	Physical position	Genetic position	Alleles
BN_A02_019	A02	716411	35.3	A/G
BN_A02_020	A02	1559029	57.8	T/G
BN_A02_021	A02	2768243	41.9	A/G
BN_A02_022	A02	5008644	96.3	A/G
BN_A02_023	A02	5532985	97.5	T/C
BN_A02_024	A02	7382074	115	T/G
BN_A02_025	A02	8162298	111.5	T/G
BN_A02_026	A02	8255337	115	A/G
BN_A02_027	A02	8825512	115.6	T/C
BN_A02_028	A02	9281526	115.2	T/C
BN_A02_029	A02	9785309	118.3	A/C
BN_A02_030	A02	12180723	124.7	T/C
BN_A02_031	A02	17954962	124.7	T/C
BN_A02_032	A02	17997856	124.7	A/G
BN_A02_033	A02	18208657	124.7	A/C
BN_A02_034	A02	20262202	132.7	A/G
BN_A02_035	A02	20723010	135.4	A/G
BN_A02_036	A02	20736861	135.4	A/G
BN_A02_037	A02	23067298	170.3	T/C
BN_A02_038	A02	23281113	172.1	A/G
BN_A02_039	A02	24567493	186.9	A/G
BN_A02_507	A02	3894723	56.2	T/C
BN_A02_508	A02	12741048	101.6	T/C

SNP	Chromosome	Physical position	Genetic position	Alleles
BN_A03_040	A03	553192	5.3	T/C
BN_A03_041	A03	573865	5.3	T/G
BN_A03_042	A03	1689959	18.5	A/G
BN_A03_043	A03	3249677	39.4	A/C
BN_A03_044	A03	3912230	47.3	T/C
BN_A03_045	A03	5178903	60.7	A/G
BN_A03_046	A03	6727279	72.4	A/G
BN_A03_047	A03	8405439	79.6	T/C
BN_A03_048	A03	8583785	82.5	A/C
BN_A03_049	A03	9459111	84.5	T/C
BN_A03_050	A03	10676837	85.2	T/C
BN_A03_051	A03	12160059	97.8	T/G
BN_A03_052	A03	13382300	101.6	T/C
BN_A03_053	A03	14849424	113.3	T/C
BN_A03_054	A03	16074819	121.5	A/C
BN_A03_055	A03	17988350	131.9	A/C
BN_A03_056	A03	18417565	132.2	A/G
BN_A03_057	A03	19974521	146.7	T/G
BN_A03_058	A03	21851205	159.7	T/C
BN_A03_059	A03	22003574	159.7	T/C
BN_A03_060	A03	24234727	164.6	T/C
BN_A03_061	A03	25671206	171.7	T/G
BN_A03_062	A03	26157992	178.6	T/C
BN_A03_063	A03	27429561	179.2	A/C
BN_A03_064	A03	28671764	181.7	A/G
BN_A03_509	A03	1353372	16.7	T/C
BN_A03_510	A03	28699941	181.7	T/C
BN_A03_511	A03	29519077	181.7	T/G
BN_A03_512	A03	29734315	181.7	T/C

SNP	Chromosome	Physical position	Genetic position	Alleles
BN_A04_065	A04	1037957	6.2	T/G
BN_A04_066	A04	2002973	9.9	T/C
BN_A04_067	A04	3049836	11.5	A/G
BN_A04_068	A04	4171118	13.1	A/C
BN_A04_069	A04	6261591	15.6	A/G
BN_A04_070	A04	7754487	13.9	A/C
BN_A04_071	A04	8018477	(unknown)	A/G
BN_A04_072	A04	9078543	13.9	A/G
BN_A04_073	A04	11529656	27.3	T/G
BN_A04_074	A04	11794984	27.3	A/G
BN_A04_075	A04	13579345	43.6	T/G
BN_A04_076	A04	14687908	54.1	A/G
BN_A04_077	A04	15809998	66.4	T/C
BN_A04_078	A04	17499319	78.1	T/C
BN_A04_079	A04	18094049	83.1	A/C
BN_A04_513	A04	10644573	22.7	T/C
BN_A04_514	A04	12387665	31.8	A/C
BN_A04_515	A04	15445316	63	A/G
BN_A04_516	A04	17398099	76.8	T/C
BN_A04_517	A04	9947700	20.3	T/G

SNP	Chromosome	Physical position	Genetic position	Alleles
BN_A05_080	A05	678757	2.1	T/G
BN_A05_081	A05	1839608	11.7	A/G
BN_A05_082	A05	2605937	21.8	A/G
BN_A05_083	A05	5120232	45.5	A/C
BN_A05_084	A05	5711517	50.7	T/G
BN_A05_085	A05	6651271	(unknown)	T/G
BN_A05_086	A05	8857758	59.1	T/C
BN_A05_087	A05	9045193	59.1	T/C
BN_A05_088	A05	11320722	59.1	T/C
BN_A05_089	A05	11656698	62.3	A/C
BN_A05_090	A05	13321112	59.1	A/G
BN_A05_091	A05	14807439	(unknown)	A/G
BN_A05_092	A05	15586113	61.7	A/G
BN_A05_093	A05	17930951	73.5	A/G
BN_A05_094	A05	18124607	74.8	T/G
BN_A05_095	A05	20351180	110	T/G
BN_A05_096	A05	21079757	117	A/C
BN_A05_097	A05	22148785	122.2	A/G
BN_A05_518	A05	3277719	29.8	A/G
BN_A05_519	A05	18805551	81.8	T/C
BN_A05_520	A05	20349488	108.7	A/G

SNP	Chromosome	Physical position	Genetic position	Alleles
BN_A06_098	A06	1045812	48.7	T/C
BN_A06_099	A06	2360914	57.8	T/G
BN_A06_100	A06	3775795	87	T/C
BN_A06_101	A06	4927798	96.9	T/C
BN_A06_102	A06	6019698	105.4	A/G
BN_A06_103	A06	6967397	111.2	A/G
BN_A06_104	A06	8305642	113.6	A/G
BN_A06_105	A06	9902591	116.9	A/G
BN_A06_106	A06	10558764	117.5	T/C
BN_A06_107	A06	14121755	117.5	A/G
BN_A06_108	A06	14657393	117.5	T/C
BN_A06_109	A06	15660439	117.8	T/C
BN_A06_110	A06	17086470	121.4	A/C
BN_A06_111	A06	18106719	124.6	A/G
BN_A06_112	A06	20298154	144.6	A/C
BN_A06_113	A06	21026646	151.4	T/C
BN_A06_114	A06	23040319	169.1	A/C
BN_A06_115	A06	23452633	175.9	A/G
BN_A06_521	A06	299826	0.4	A/C
BN_A06_522	A06	19228278	133.8	T/G
BN_A06_523	A06	2594640	62.595	A/G
BN_A06_524	A06	787710	4.4	A/G

SNP	Chromosome	Physical position	Genetic position	Alleles
BN_A07_116	A07	1237631	10.4	T/G
BN_A07_117	A07	1737124	10.7	A/G
BN_A07_118	A07	3237817	11.9	A/G
BN_A07_119	A07	4388133	12.5	T/G
BN_A07_120	A07	5909982	12.5	A/G
BN_A07_121	A07	6827762	12.5	T/C
BN_A07_122	A07	8173848	13.5	A/G
BN_A07_123	A07	10098407	22.8	T/C
BN_A07_124	A07	11533793	38.9	T/G
BN_A07_125	A07	11534201	38.9	A/G
BN_A07_126	A07	11738168	41.3	A/C
BN_A07_127	A07	14028422	58.5	A/G
BN_A07_128	A07	14839056	61.1	T/G
BN_A07_129	A07	15954103	79	A/C
BN_A07_130	A07	17484633	95.1	A/G
BN_A07_131	A07	18677722	103.1	A/G
BN_A07_132	A07	19874430	113.9	T/C
BN_A07_133	A07	20983286	119.1	T/C
BN_A07_134	A07	23003258	132.2	A/C
BN_A07_135	A07	23707093	136.2	T/G
BN_A07_525	A07	12680897	44.4	A/G
BN_A07_526	A07	11526727	38.9	A/G
BN_A07_527	A07	19881264	113.9	T/C

SNP	Chromosome	Physical position	Genetic position	Alleles
BN_A08_136	A08	469043	1.4	T/C
BN_A08_137	A08	1979050	3.5	A/G
BN_A08_138	A08	3571164	4.3	T/C
BN_A08_139	A08	4879516	4.3	A/G
BN_A08_140	A08	6437976	4.3	A/C
BN_A08_141	A08	7388369	4.7	A/G
BN_A08_142	A08	8478419	7	A/C
BN_A08_143	A08	8625305	7	T/C
BN_A08_144	A08	10515602	12.1	T/C
BN_A08_145	A08	11851803	21.8	A/C
BN_A08_146	A08	13277331	29.8	T/C
BN_A08_147	A08	15126794	(unknown)	T/G
BN_A08_148	A08	16776098	57.6	T/C
BN_A08_149	A08	16825096	59.1	T/G
BN_A08_150	A08	18739906	74.5	A/C
BN_A08_528	A08	17422209	66	A/G
BN_A08_529	A08	10338218	10.45	A/C
BN_A08_530	A08	11051828	16	A/G
BN_A08_531	A08	12509902	29	T/C
BN_A08_532	A08	14118474	36.1	A/C
BN_A08_533	A08	14773171	41.2	A/G

SNP	Chromosome	Physical position	Genetic position	Alleles
BN_A09_151	A09	753105	14.9	A/C
BN_A09_152	A09	2226812	26	A/C
BN_A09_153	A09	3695622	41.5	A/C
BN_A09_154	A09	5036357	59.2	A/G
BN_A09_155	A09	5957165	62.2	T/C
BN_A09_156	A09	6668848	68.5	A/C
BN_A09_157	A09	8281507	71.6	A/G
BN_A09_158	A09	9946416	76.4	T/C
BN_A09_159	A09	10773958	78.7	A/C
BN_A09_160	A09	12675514	94.3	T/C
BN_A09_161	A09	13471812	93	T/C
BN_A09_162	A09	13622907	83.4	A/G
BN_A09_163	A09	15126743	94.3	A/G
BN_A09_164	A09	17157027	95.8	T/C
BN_A09_165	A09	21894367	101.6	T/C
BN_A09_166	A09	21931781	101.6	T/C
BN_A09_167	A09	21933211	101.6	T/C
BN_A09_168	A09	22081699	102.2	T/C
BN_A09_169	A09	23692880	104.3	A/C
BN_A09_170	A09	25899202	116	A/C
BN_A09_171	A09	26580614	118	A/G
BN_A09_172	A09	28238031	132.2	T/C
BN_A09_173	A09	29637783	142.3	T/G
BN_A09_174	A09	30026113	143.1	A/G
BN_A09_175	A09	32618215	158.2	T/C
BN_A09_176	A09	32625993	159.1	A/C
BN_A09_534	A09	3791683	41.5	A/G
BN_A09_535	A09	4769933	55.1	A/G
BN_A09_536	A09	17160233	95.8	A/G

SNP	Chromosome	Physical position	Genetic position	Alleles
BN_A10_177	A10	1123045	4.8	T/C
BN_A10_178	A10	1592086	5.7	A/C
BN_A10_179	A10	2691425	10.4	A/C
BN_A10_180	A10	5380555	10.8	T/C
BN_A10_181	A10	5465235	10.8	T/C
BN_A10_182	A10	7242620	10.4	T/C
BN_A10_183	A10	8216123	11.9	A/C
BN_A10_184	A10	10133294	15.1	T/C
BN_A10_185	A10	11185458	20	A/G
BN_A10_186	A10	12610392	37.3	A/G
BN_A10_187	A10	13247713	41	A/G
BN_A10_188	A10	15342246	69	A/G
BN_A10_189	A10	15821224	76	A/G
BN_A10_190	A10	16911658	86.8	A/G
BN_A10_537	A10	11746298	26.7	T/C
BN_A10_538	A10	12372045	34.3	A/C
BN_A10_539	A10	13524058	46.7	T/C
BN_A10_540	A10	13915852	49.3	A/G
BN_A10_541	A10	14772957	60.1	T/G

SNP	Chromosome	Physical position	Genetic position	Alleles
BN_C01_191	C01	676816	3.2	A/G
BN_C01_192	C01	2273625	22.7	T/G
BN_C01_193	C01	3017329	29.4	A/G
BN_C01_194	C01	4014751	35.5	A/C
BN_C01_195	C01	5348410	(unknown)	A/G
BN_C01_196	C01	6769898	41.5	T/C
BN_C01_197	C01	8948092	50.3	T/C
BN_C01_198	C01	9611072	54.2	A/G
BN_C01_199	C01	11551257	62	A/G
BN_C01_200	C01	11758317	62	T/G
BN_C01_201	C01	13532596	66.6	T/C
BN_C01_202	C01	14495132	70.3	T/C
BN_C01_203	C01	16519503	73	T/C
BN_C01_204	C01	17005242	75.1	T/C
BN_C01_205	C01	18272820	79.9	A/G
BN_C01_206	C01	20624717	79.9	A/G
BN_C01_207	C01	21071974	79.9	A/C
BN_C01_208	C01	22001115	79.9	T/G
BN_C01_209	C01	24493672	78.5	A/C
BN_C01_210	C01	25072741	79.9	T/G
BN_C01_211	C01	26746387	79.9	T/C
BN_C01_212	C01	28021692	83	T/C
BN_C01_213	C01	29104455	83	A/G
BN_C01_214	C01	30261877	73	A/G
BN_C01_215	C01	31200490	86.3	A/G
BN_C01_216	C01	33505774	91.9	T/G
BN_C01_217	C01	34650999	95.6	T/G
BN_C01_218	C01	35044682	96.2	A/C
BN_C01_219	C01	36365984	100.9	A/G
BN_C01_220	C01	37650716	116	A/G

SNP	Chromosome	Physical position	Genetic position	Alleles
BN_C02_221	C02	413694	8.6	T/C
BN_C02_222	C02	1844992	7.9	T/C
BN_C02_223	C02	3274864	18.9	A/C
BN_C02_224	C02	5162738	40.1	A/G
BN_C02_225	C02	5549589	43.7	T/C
BN_C02_226	C02	7408306	49	A/G
BN_C02_227	C02	7605785	50.3	T/G
BN_C02_228	C02	10060782	59.4	T/C
BN_C02_229	C02	11131769	60.6	T/C
BN_C02_230	C02	12535780	67.5	T/G
BN_C02_231	C02	12712319	66.6	A/C
BN_C02_232	C02	14520401	73.3	T/C
BN_C02_233	C02	15917452	72.1	T/C
BN_C02_234	C02	17562538	73.3	T/C
BN_C02_235	C02	19166309	78.2	T/G
BN_C02_236	C02	19750793	78.5	A/C
BN_C02_237	C02	20180910	86.3	A/G
BN_C02_238	C02	22536228	83.8	A/G
BN_C02_239	C02	23190685	85.5	A/G
BN_C02_240	C02	25612313	88.1	T/G
BN_C02_241	C02	27033849	95.1	A/C
BN_C02_242	C02	27502652	89.9	A/C
BN_C02_243	C02	28654876	90.8	A/C
BN_C02_244	C02	30714314	90.8	A/G
BN_C02_245	C02	32320670	89.9	T/G
BN_C02_246	C02	33047610	89.9	T/C
BN_C02_247	C02	34741510	90.8	T/C
BN_C02_248	C02	34957893	90.8	T/C
BN_C02_249	C02	37430270	99.3	A/G
BN_C02_250	C02	37917887	101.3	A/G
BN_C02_251	C02	39511011	106.1	T/C
BN_C02_252	C02	40707599	116	A/G
BN_C02_253	C02	41630365	120.3	A/G
BN_C02_254	C02	43116765	136.5	T/G
BN_C02_255	C02	43919741	145.3	A/G
BN_C02_256	C02	44698711	150.2	T/C
BN_C02_542	C02	44843893	153.9	T/A

SNP	Chromosome	Physical position	Genetic position	Alleles
BN_C03_257	C03	174210	3.7	T/C
BN_C03_258	C03	1365915	8.6	T/C
BN_C03_259	C03	2922407	21.4	A/G
BN_C03_260	C03	5093254	37.9	T/C
BN_C03_261	C03	6203685	45.7	A/C
BN_C03_262	C03	7011358	54.1	T/C
BN_C03_263	C03	8232425	58.4	A/G
BN_C03_264	C03	9862838	64.7	T/C
BN_C03_265	C03	10445245	(unknown)	T/G
BN_C03_266	C03	12400463	72.8	T/C
BN_C03_267	C03	14071452	80.5	T/C
BN_C03_268	C03	14813052	84.8	T/C
BN_C03_269	C03	16419977	90.4	T/C
BN_C03_270	C03	17354465	91.5	T/C
BN_C03_271	C03	18771754	96.4	T/G
BN_C03_272	C03	19626762	99.6	A/G
BN_C03_273	C03	21677096	114.5	T/G
BN_C03_274	C03	22631355	121.3	T/G
BN_C03_275	C03	23418388	127.6	A/G
BN_C03_276	C03	24937645	131.6	T/C
BN_C03_277	C03	26872226	134.3	T/C
BN_C03_278	C03	27071735	134.3	A/G
BN_C03_279	C03	29665343	138.7	A/G
BN_C03_280	C03	30383682	139.9	T/G
BN_C03_281	C03	31763561	141.7	T/G
BN_C03_282	C03	32664839	141.7	A/G
BN_C03_283	C03	34605524	151.5	T/G
BN_C03_284	C03	34629718	151.5	A/G
BN_C03_285	C03	37850363	156	T/C
BN_C03_286	C03	38006755	79.9	A/G
BN_C03_287	C03	41016845	156	A/G
BN_C03_288	C03	41048584	156	A/G
BN_C03_289	C03	46498828	156	T/C
BN_C03_290	C03	47552021	163.5	T/C
BN_C03_291	C03	47825989	163.7	A/G
BN_C03_292	C03	49313549	169.4	T/C
BN_C03_293	C03	51498232	179.9	A/C
BN_C03_294	C03	51658423	180.4	T/G
BN_C03_295	C03	53438409	183.9	T/C
BN_C03_296	C03	54364129	187.2	A/G
BN_C03_297	C03	55842950	188.7	T/C
BN_C03_298	C03	57834639	199.6	T/C
BN_C03_299	C03	58165464	202	A/G
BN_C03_300	C03	59599719	213.8	T/G
BN_C03_301	C03	60131836	215.9	T/C

SNP	Chromosome	Physical position	Genetic position	Alleles
BN_C04_302	C04	818660	4.3	A/C
BN_C04_303	C04	1984637	12.5	T/C
BN_C04_304	C04	3789987	22.9	A/G
BN_C04_305	C04	5126515	33.2	T/C
BN_C04_306	C04	5984429	36.9	A/G
BN_C04_307	C04	6519105	39.5	A/G
BN_C04_308	C04	8375642	(unknown)	T/C
BN_C04_309	C04	9882343	52.7	A/C
BN_C04_310	C04	10712318	54.1	A/C
BN_C04_311	C04	11861256	53.8	T/C
BN_C04_312	C04	14163966	55.5	A/G
BN_C04_313	C04	15298788	56.2	A/C
BN_C04_314	C04	15342125	56.2	T/G
BN_C04_315	C04	18782780	59.4	A/C
BN_C04_316	C04	18818888	59.4	T/C
BN_C04_317	C04	20589816	61.9	A/C
BN_C04_318	C04	21625420	62.2	T/C
BN_C04_319	C04	22721957	59.4	A/C
BN_C04_320	C04	23299693	62.8	A/C
BN_C04_321	C04	24597218	64	A/C
BN_C04_322	C04	25869881	65.8	T/C
BN_C04_323	C04	28314481	69.5	A/G
BN_C04_324	C04	28471460	69.5	A/G
BN_C04_325	C04	29725330	71.6	T/C
BN_C04_326	C04	31176731	77.4	A/G
BN_C04_327	C04	32408106	(unknown)	T/G
BN_C04_328	C04	33844593	82.2	T/C
BN_C04_329	C04	34856080	84.9	T/C
BN_C04_330	C04	37105564	91	T/C
BN_C04_331	C04	38302582	93.3	A/C
BN_C04_332	C04	38814870	93.3	T/C
BN_C04_333	C04	40294385	96.1	T/C
BN_C04_334	C04	42703770	105.4	T/C
BN_C04_335	C04	42745572	105.4	A/G
BN_C04_336	C04	44173102	115.6	T/G
BN_C04_337	C04	44817535	128.9	T/C
BN_C04_338	C04	45281558	135.5	A/C
BN_C04_339	C04	46951935	149.5	T/C
BN_C04_340	C04	48370526	160.6	T/C

SNP	Chromosome	Physical position	Genetic position	Alleles
BN_C05_341	C05	1042687	9.9	T/G
BN_C05_342	C05	1542073	14.7	A/C
BN_C05_343	C05	2707564	25.5	A/G
BN_C05_344	C05	4668224	44.3	A/C
BN_C05_345	C05	6345629	48.5	A/C
BN_C05_346	C05	7462120	70.4	T/C
BN_C05_347	C05	8746341	72.9	A/G
BN_C05_348	C05	9065359	103	A/G
BN_C05_349	C05	10727791	76.6	A/G
BN_C05_350	C05	11057233	79.4	A/G
BN_C05_351	C05	12660974	80.1	A/C
BN_C05_352	C05	14385177	80.6	A/G
BN_C05_353	C05	14954173	80.6	T/C
BN_C05_354	C05	15752750	82.4	A/C
BN_C05_355	C05	16033775	82.4	T/G
BN_C05_356	C05	21345507	84.2	A/C
BN_C05_357	C05	21869769	84.2	A/G
BN_C05_358	C05	22428717	84.2	T/C
BN_C05_359	C05	24361417	84.7	A/G
BN_C05_360	C05	25542861	84.7	A/G
BN_C05_361	C05	28271691	84.7	T/C
BN_C05_362	C05	29921031	85	A/G
BN_C05_363	C05	30715697	85	A/G
BN_C05_364	C05	30868461	85	T/G
BN_C05_365	C05	32983091	90.7	A/G
BN_C05_366	C05	35752184	95.1	T/C
BN_C05_367	C05	37178452	100	A/C
BN_C05_368	C05	37867132	101.5	A/C
BN_C05_369	C05	39936268	116.9	A/G
BN_C05_370	C05	40981143	129.9	A/G
BN_C05_371	C05	42609171	144.2	T/C
BN_C05_543	C05	4370802	39.8	T/C

SNP	Chromosome	Physical position	Genetic position	Alleles
BN_C06_372	C06	779281	1.3	A/G
BN_C06_373	C06	2542286	2.6	A/C
BN_C06_374	C06	2855949	(unknown)	T/C
BN_C06_375	C06	5003113	11.3	T/C
BN_C06_376	C06	5871694	18.3	T/C
BN_C06_377	C06	7267212	(unknown)	T/C
BN_C06_378	C06	7864731	27.4	T/G
BN_C06_379	C06	7975388	28.2	T/C
BN_C06_380	C06	11594894	29.7	A/G
BN_C06_381	C06	12790699	31.3	T/C
BN_C06_382	C06	13083038	32.3	T/C
BN_C06_383	C06	15199533	33.2	A/G
BN_C06_384	C06	15576918	33.5	A/G
BN_C06_385	C06	17613363	39.8	T/G
BN_C06_386	C06	18723767	42.6	A/G
BN_C06_387	C06	20196280	49.5	A/C
BN_C06_388	C06	21777970	56.7	A/C
BN_C06_389	C06	21936594	56.7	T/C
BN_C06_390	C06	23956323	58.3	T/C
BN_C06_391	C06	25296765	62.8	A/C
BN_C06_392	C06	27047294	66.6	A/G
BN_C06_393	C06	27866854	66.6	T/G
BN_C06_394	C06	29269790	(unknown)	T/G
BN_C06_395	C06	30110379	69.9	A/G
BN_C06_396	C06	31075065	69.9	T/G
BN_C06_397	C06	32712271	75.3	A/G
BN_C06_398	C06	33877896	84.2	T/C
BN_C06_399	C06	34898249	91.5	T/C
BN_C06_400	C06	36174112	102.8	A/C

SNP	Chromosome	Physical position	Genetic position	Alleles
BN_C07_401	C07	626978	(unknown)	T/G
BN_C07_402	C07	2201425	1	A/G
BN_C07_403	C07	3647048	(unknown)	A/C
BN_C07_404	C07	5311940	4.6	A/C
BN_C07_405	C07	7173131	4.1	A/G
BN_C07_406	C07	8004257	(unknown)	T/G
BN_C07_407	C07	9960936	5.4	A/G
BN_C07_408	C07	10611599	2.5	T/G
BN_C07_409	C07	11828120	4.6	A/G
BN_C07_410	C07	13123473	6.5	T/C
BN_C07_411	C07	15147528	8.2	T/C
BN_C07_412	C07	15836139	10.7	T/C
BN_C07_413	C07	18085670	12.5	T/C
BN_C07_414	C07	18464106	15.3	A/G
BN_C07_415	C07	18476958	15.3	T/C
BN_C07_416	C07	19026740	15.3	T/G
BN_C07_417	C07	20741654	17.4	T/C
BN_C07_418	C07	22146112	21.4	T/C
BN_C07_419	C07	23395310	24.9	T/C
BN_C07_420	C07	25181767	29.4	A/G
BN_C07_421	C07	26404014	33.8	A/C
BN_C07_422	C07	28079925	38.8	A/G
BN_C07_423	C07	28760661	40.9	T/G
BN_C07_424	C07	30723217	45.7	T/C
BN_C07_425	C07	31751200	49	A/G
BN_C07_426	C07	32551503	53.5	A/C
BN_C07_427	C07	33974378	57.9	T/G
BN_C07_428	C07	34913620	62.5	A/G
BN_C07_429	C07	36237544	66.7	T/G
BN_C07_430	C07	38458964	79.2	A/G
BN_C07_431	C07	39760550	83.4	A/G
BN_C07_432	C07	40936422	90.9	A/G
BN_C07_433	C07	42297486	98.5	T/C
BN_C07_434	C07	42605418	(unknown)	T/C
BN_C07_435	C07	44464103	104.7	A/C
BN_C07_544	C07	27917444	38.3	A/G
BN_C07_545	C07	37596318	75.8	A/G

SNP	Chromosome	Physical position	Genetic position	Alleles
BN_C08_436	C08	801927	4.5	A/C
BN_C08_437	C08	1865293	10.2	A/G
BN_C08_438	C08	2824056	10.2	A/C
BN_C08_439	C08	6392625	13.5	A/C
BN_C08_440	C08	6435287	13.5	T/G
BN_C08_441	C08	6663763	13.5	T/C
BN_C08_442	C08	7809192	13.5	A/G
BN_C08_443	C08	9141977	13.5	T/G
BN_C08_444	C08	11972761	15.3	T/C
BN_C08_445	C08	12023411	15.3	T/G
BN_C08_446	C08	13605370	15.3	T/C
BN_C08_447	C08	14461239	15.3	T/C
BN_C08_448	C08	16165436	16.8	T/C
BN_C08_449	C08	18037769	20.8	A/C
BN_C08_450	C08	18065003	20.8	A/C
BN_C08_451	C08	19647416	29.9	T/G
BN_C08_452	C08	20727633	38.1	T/G
BN_C08_453	C08	21232166	44.2	A/G
BN_C08_454	C08	23469959	50	A/G
BN_C08_455	C08	25379127	53	A/G
BN_C08_456	C08	27039503	61	A/G
BN_C08_457	C08	28373767	70.5	T/G
BN_C08_458	C08	29584703	74.4	T/C
BN_C08_459	C08	30245503	78.6	A/G
BN_C08_460	C08	31583479	92.3	A/G
BN_C08_461	C08	33471716	100.5	T/C
BN_C08_462	C08	34723069	106.7	A/G
BN_C08_463	C08	35091887	109.4	A/G
BN_C08_464	C08	37159595	120.2	A/G
BN_C08_465	C08	38302280	121.5	T/C

SNP	Chromosome	Physical position	Genetic position	Alleles
BN_C09_466	C09	1145307	28.8	A/G
BN_C09_467	C09	1489171	30	A/C
BN_C09_468	C09	3592310	41.8	T/C
BN_C09_469	C09	4482533	46.1	T/G
BN_C09_470	C09	6552767	56.4	T/C
BN_C09_471	C09	6897184	58.3	A/C
BN_C09_472	C09	8175587	63.2	A/C
BN_C09_473	C09	10183356	66.4	T/C
BN_C09_474	C09	10374152	66.4	A/G
BN_C09_475	C09	12892018	69.9	A/G
BN_C09_476	C09	13900701	73.5	T/C
BN_C09_477	C09	13967385	69.9	T/G
BN_C09_478	C09	14757837	73.1	T/G
BN_C09_479	C09	16561656	74.9	T/G
BN_C09_480	C09	16998063	74.9	T/C
BN_C09_481	C09	18705881	80.2	A/C
BN_C09_482	C09	19297304	80.2	A/G
BN_C09_483	C09	21402578	81.6	A/G
BN_C09_484	C09	23126797	82.2	A/G
BN_C09_485	C09	26086370	82.2	T/C
BN_C09_486	C09	27205716	(unknown)	T/G
BN_C09_487	C09	31510923	82.2	A/G
BN_C09_488	C09	33434568	82.2	T/C
BN_C09_489	C09	33576273	82.2	T/C
BN_C09_490	C09	35075253	84.5	A/C
BN_C09_491	C09	36780742	85	T/C
BN_C09_492	C09	37584510	89.2	A/C
BN_C09_493	C09	39720977	91	T/C
BN_C09_494	C09	41126474	96	A/G
BN_C09_495	C09	41388753	97.8	A/G
BN_C09_496	C09	43469252	107.3	A/G
BN_C09_497	C09	43501132	107.3	T/C
BN_C09_498	C09	46646099	130.8	T/G
BN_C09_499	C09	46748516	132	A/G
BN_C09_500	C09	47784850	146.9	T/C
BN_C09_546	C09	1488283	30	T/C
BN_C09_547	C09	1679820	31.5	A/G
BN_C09_548	C09	45372117	120.2	A/G
BN_C09_482	C09	19297304	80.2	A/G
BN_C09_483	C09	21402578	81.6	A/G
BN_C09_484	C09	23126797	82.2	A/G
BN_C09_485	C09	26086370	82.2	T/C
BN_C09_486	C09	27205716	(unknown)	T/G
BN_C09_487	C09	31510923	82.2	A/G
BN_C09_488	C09	33434568	82.2	T/C
BN_C09_489	C09	33576273	82.2	T/C
BN_C09_490	C09	35075253	84.5	A/C
BN_C09_491	C09	36780742	85	T/C
BN_C09_492	C09	37584510	89.2	A/C
BN_C09_493	C09	39720977	91	T/C
BN_C09_494	C09	41126474	96	A/G
BN_C09_495	C09	41388753	97.8	A/G
BN_C09_496	C09	43469252	107.3	A/G
BN_C09_497	C09	43501132	107.3	T/C
BN_C09_498	C09	46646099	130.8	T/G
BN_C09_499	C09	46748516	132	A/G
BN_C09_500	C09	47784850	146.9	T/C
BN_C09_546	C09	1488283	30	T/C
BN_C09_547	C09	1679820	31.5	A/G
BN_C09_548	C09	45372117	120.2	A/G

Circular chromosome maps showing the physical and genetic location of the 548 SNPs tested. SNPs that were discarded after applying quality filters are shown in blue. The 16 SNPs whose position on the genetic map is unknown are indicated with an open circle. The position of the centromere is indicated with the letter “C”.



