

# FINAL REPORT:

## TEST OF THE POTENTIAL USE OF SNP MARKERS ON OILSEED RAPE VARIETIES

---

Agreement number 7510544, co-funded by the Community Plant Variety Office (CPVO) Research and Development Section, APHA, and GEVES  
November 2016 to December 2017.

**Institution responsible for the project:** APHA (UK), GEVES (FR) and NIAB (UK)

**Other institutes involved:** BSA (DE), Department of Variety Testing (DK), COROBU (PL), UKSUP (SK), INIA (ES), UKZUZ (CZ), ESA (UE)

**Duration of the project:** 1 year

**Total estimated cost:** 76 372 €



## **SUMMARY OF THE PROJECT**

Following the recommendations and conclusions of a first project co funded by the CPVO on the use of molecular markers (SSR) on oilseed rape varieties, a new project is proposed, using another type of molecular markers (SNP). Before assessing this new type of molecular marker on a large number of varieties, more knowledge and background are needed. It is necessary to know whether these markers can be used without difficulty for DUS purpose on partially out-crossing allotetraploid species such as OSR, and whether a bulking strategy could be considered for future applications. Two laboratories from two different countries (France and United Kingdom) will participate in this first step. A set of 500 SNPs will be tested on different matrices. Only the most efficient markers will be kept, and bulking strategy will be assessed.

The main objective of the project is to select a reliable marker set, as well as an applicable procedure for routine genotyping. By reviewing the results, a protocol to genotype different varieties with marker combinations could be proposed, and a project would follow with the aim of combining genotypic and phenotypic data to optimise OSR reference collection management.

### **Objectives:**

The project aimed to examine the potential use of SNP markers as a tool for the management of OSR reference collection.

The detailed objectives were:

- Selection and validation of a suitable set of SNP markers
- Method optimisation: assess the use of bulks of plants or seeds instead of individual plants

## **LIST OF PARTNERS**

### **Institutions responsible for the project:**

**GEVES (coordinator)**, 25 rue Georges Morel, CS 90024, 49071 Beaucouzé cedex, FR

**APHA**, The Animal and Plant Health Agency, Plant Variety Rights and Seeds Office, Eastbrook, Shaftsbury Road, Cambridge CB2 8DR, UK

**NIAB**, Huntingdon Road, Cambridge CB3 0LE, UK

### **Other institute involved:**

**BSA**, Bundessortenamt, Osterfelddamm 80, 30627 Hannover, DE

**COBORU**, Research Centre for Cultivar Testing, 63-022 Słupia Wielka, PL

**UKSUP**, Central Controlling and Testing Institute in Agriculture, Akademická 4, 949 01 Nitra, SK

**Department of Variety Testing**, Teglværksvej 10, Tystofte, DK-4230 Skælskør, DK

**INIA**, Dirección Técnica de Evaluación de Variedades y Productos Fitosanitarios, Crtra. de La Coruña km 7,200, 28040 Madrid, ES

**UKZUZ**, Central Institute for Supervising and Testing in Agriculture, Hroznová 2, 656 06 Brno, CZ

**ESA**, 23, Rue du Luxembourg, 1000 Brussels, BE

**Name of the contact person:** Corbel Anne-Lise

**Name of the scientific coordinator:** Remay Arnaud

**Other major contributors:** Helen Appleyard, Anne Bernole, Elizabeth Scott, Muriel Thomasset, Cheryl Turnbull, Margaret Wallace

## LIST OF CONTENT

|                                                                                                                                                      |     |
|------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| SUMMARY OF THE PROJECT .....                                                                                                                         | I   |
| LIST OF PARTNERS .....                                                                                                                               | II  |
| LIST OF CONTENT .....                                                                                                                                | III |
| LIST OF FIGURES .....                                                                                                                                | IV  |
| LIST OF TABLES .....                                                                                                                                 | IV  |
| 1. INTRODUCTION .....                                                                                                                                | 1   |
| 2. OBJECTIVES ADRESSED .....                                                                                                                         | 2   |
| 3. CURRENT SITUATION.....                                                                                                                            | 2   |
| 3.1. Molecular markers in OSR .....                                                                                                                  | 2   |
| 3.2. Molecular markers and OSR DUS testing .....                                                                                                     | 3   |
| 3.3. Choice of SNP platform and technology .....                                                                                                     | 3   |
| 3.4. Pooling strategy.....                                                                                                                           | 4   |
| 3.5. Experimental approaches.....                                                                                                                    | 4   |
| 4. PHASE 1- Marker selection, primer design / varieties selection .....                                                                              | 1   |
| 4.1. Milestone: selection of a set of SNP markers, synthesis of molecular primers and distribution of the primers to the two laboratories (FR) ..... | 1   |
| 4.2. Milestone: selection of varieties to be tested and material requirements (FR and UK).....                                                       | 5   |
| 4.3. Milestone: Growing plant material and DNA extraction (FR) .....                                                                                 | 7   |
| 5. PHASE 2 - Amplification test/markers screening .....                                                                                              | 9   |
| 5.1. Experimental approaches.....                                                                                                                    | 9   |
| 5.2. Results .....                                                                                                                                   | 9   |
| 6. PHASE 3 - Working in Bulk and size of the Bulks .....                                                                                             | 13  |
| 6.1. Milestone: SNP selection marker for phase 3 .....                                                                                               | 13  |
| 6.2. Milestone: scoring individuals and synthetic DNA (FR and UK) .....                                                                              | 14  |
| 6.3. Milestone: scoring seed and plant bulks with validation from synthetic bulks (FR and UK) .....                                                  | 23  |
| 7. PHASE 4 - Variety screening & data collation by partners for method validation (FR and UK) ....                                                   | 26  |
| 7.1. Experimental approach .....                                                                                                                     | 26  |
| 7.2. Milestone: Data analysis and review of the results .....                                                                                        | 26  |
| 8. PHASE 5 - Meeting with all entrusted EOs, ESA and the CPVO.....                                                                                   | 31  |
| 9. LITERATURE .....                                                                                                                                  | 33  |

## **LIST OF FIGURES**

|                                                                                                                                                                                                                                                         |    |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| <b>Figure 1.</b> A schematic representation of the initial selection process used to identify 500 SNPs for conversion into KASP assay. ....                                                                                                             | 3  |
| <b>Figure 2.</b> Distribution of the 500 SNPs selected during the initial filtering process, according to their PIC value. ....                                                                                                                         | 4  |
| <b>Figure 3.</b> Snapshot displaying SNP genotyping. Snapshot (a) represents a failed pattern (in this case SNP BN_A02_020) and snapshot (b) shows a polymorphic pattern with 3 clusters (in this case SNP BN_A06_112).....                             | 10 |
| <b>Figure 4.</b> Physical distribution on the 19 chromosomes of the 500 SNPs whether they failed KASP development .....                                                                                                                                 | 11 |
| <b>Figure 5.</b> Bitmap of polymorphism for pairwise comparisons of the 23 varieties (rows) and 50 SNPs in common (columns).....                                                                                                                        | 12 |
| <b>Figure 6.</b> The SNP selection process for Phase 3.....                                                                                                                                                                                             | 13 |
| <b>Figure 7.</b> UPGMA dendrogram based on genetic distances as estimated by 50 SNPs (FR) among the four varieties. ....                                                                                                                                | 15 |
| <b>Figure 8.</b> Histogram showing the results for allele 2 of Var 23 for the marker BN_A09_157 .....                                                                                                                                                   | 17 |
| <b>Figure 9.</b> Snapshot displaying SNP genotyping. Snapshot represents pattern for the SNP BN_C03_268 for individuals genotyping and DNA bulk (8, 16, 24, 32) for the four varieties of phase 3.....                                                  | 18 |
| <b>Figure 10.</b> UPGMA dendrogram based on genetic distances as estimated by 50 SNPs (UK) among the four varieties. ....                                                                                                                               | 19 |
| <b>Figure 11.</b> Histogram showing the results for allele 1 of variety 22 for the marker BN_C03_259.....                                                                                                                                               | 20 |
| <b>Figure 12.</b> FR Allele call from different size of DNA bulk and allele frequency estimated from individual experimental designs. ....                                                                                                              | 21 |
| <b>Figure 13.</b> UK Allele call from different size of DNA bulk and allele frequency estimated from individual experimental designs. ....                                                                                                              | 22 |
| <b>Figure 14.</b> Detection of allele with lowest allele frequency (FR data). ....                                                                                                                                                                      | 24 |
| <b>Figure 15.</b> Detection of allele with lowest allele frequency (UK data). ....                                                                                                                                                                      | 25 |
| <b>Figure 16.</b> Histogram of the pairwise genetic distances between the 2 sampling of each variety (FR)26                                                                                                                                             |    |
| <b>Figure 17.</b> Distribution of the genetic distances among the 23 varieties .....                                                                                                                                                                    | 27 |
| <b>Figure 18.</b> Histogram of the pairwise genetic distances between the 2 sampling of each variety (UK) .....                                                                                                                                         | 28 |
| <b>Figure 19.</b> Snapshot displaying SNP genotyping. Snapshot (a) represents a pattern for the SNP BN_A03_058 268 for individuals genotyping (phase 2) and snapshot (b) shows for the same SNP for the same varieties but genotyped in seed bulks..... | 29 |
| <b>Figure 20.</b> Bitmap of polymorphism for pairwise comparisons between repetitions carried out by each laboratory on the 23 varieties (rows) and 38 SNPs in common (columns).....                                                                    | 30 |

## **LIST OF TABLES**

|                                                                                                                                                                                                               |    |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| <b>Table 1.</b> Characteristics of the 23 selected varieties. ....                                                                                                                                            | 6  |
| <b>Table 2.</b> Summary of diversity indices number of alleles (Na), observed (Ho) and expected (He) heterozygosities per variety calculated for the individual plants and the difference size of bulk). .... | 16 |

## 1. INTRODUCTION

Oilseed rape (*Brassica napus*) is a prominent oil and fodder crop worldwide. It is grown for the production of animal feed, vegetable oil for human consumption and biodiesel; leading producers include the European Union, Canada, the United States, Australia, China and India.

With the development in the breeding of this species, number of varieties of common knowledge is increasing significantly. This together with the fact that more and more hybrids are being bred has led to a substantial increase in the size of the reference collection (all components used to produce the hybrid must also be included in the reference collection).

All the Examination Offices (EOs) in Europe who conduct DUS tests on oilseed rape (OSR) have experienced difficulties due to the increased workload and the need for more space when conducting the trials. A new variety has to be studied in comparison with all the varieties of the reference collection. Each candidate that passes the DUS test is added to the reference collection. Considering the reference collection has more than doubled in size since 2005, number of varieties (and therefore numbers of comparisons) is the main limiting factor.

The management of this collection became a great concern for every office and they are all interested in reducing the workload and costs of testing by eliminating unnecessary reference varieties in the trials. Reducing number of reference varieties to be grown prior to DUS growing trial is a key feature of the process. UPOV (TGP 9) states that selection of varieties of common knowledge to be compared can be narrowed down by using well documented variety descriptions. One possible way is to use DNA profiling of varieties as a management tool. The model *“Combining Phenotypic and Molecular Distances in the Management of Variety Collection”* has long-since been accepted by UPOV members (see UPOV/INF/18/1 *“Possible use of Molecular Markers in the Examination of Distinctness, Uniformity and Stability (DUS)”*) and is included in TGP/15/1 *“Guidance on the Use of Biochemical and Molecular Markers in the Examination of Distinctness, Uniformity and Stability (DUS)”*. This model is already in use at GEVES on different species (spring barley and maize) based on results of studies using different molecular markers types (SSR and SNP).

A first research project “Management of winter oilseed rape reference collections” CPV5766 (2005-2007) which used SSRs (Simple Sequence Repeats), highlighted problems of using SSR markers for the management of OSR reference collection. Using SSRs has proven to be particularly challenging in an out-crossing tetraploid species such as OSR and even more when bulk of seedlings are being used to generate variety profiles in different laboratories. Moreover, SSR markers were multi allelic showing DNA profile difficult to analyse. The project concluded, among other points, that the investigation of other types of markers such as SNPs (Single Nucleotide Polymorphisms) when available would be beneficial. The reasoning behind the conclusion was that because SNP markers are bi-allelic, results are supposedly easier to interpret than with SSR markers, and should be more conclusive. After the publication of a set of 1536 free-access SNP markers in February 2013 (Delourme *et al.* 2013), GEVES initiated a project to improve the management of OSR reference collections using this set of SNP markers and solicited the participation of the following partners: all the examination offices entrusted for OSR (DE, DK, UK, CZ, SK, ES, PL) and the support of ESA.

However, due to the lack of technical knowledge on the use of SNPs on oilseed rape, all the partners agreed that a preliminary project to study the feasibility of using SNPs for routine work should be considered. As a consequence, a project was developed to test the potential future use of SNP markers on OSR. This project aims to develop techniques that could improve the quality of DUS tests, increase efficiency and facilitate exchanges between EOs.

Effort made on SNP utilisation in *B. napus* is circumscribed by one main objective: the potential use of SNP markers for the WOSR reference collection management.

## **2. OBJECTIVES ADRESSED**

The project aimed to examine the potential use of SNP markers as a tool for the management of OSR reference collection.

The detailed objectives were:

- Selection and validation of a suitable set of SNP markers;
- Method optimization: assess the use of a bulk of plants or seeds instead of individual plants.

## **3. CURRENT SITUATION**

### **3.1. Molecular markers in OSR**

The crucifer *B. napus* is an allotetraploid species (AC genome,  $n=19$ ) that originated from the spontaneous hybridisation between the two closely related diploid species *B. rapa* (genome A,  $n=10$ ) and *B. oleracea* (genome C,  $n=9$ ).

During the last two decades, development of molecular markers offers several advantages for plant variety characterisation. Restriction Fragment Length Polymorphisms (RFLPs), followed by PCR based markers such as Randomly Amplified Polymorphic DNA (RAPD) and Amplified Fragment Length Polymorphisms (AFLP) have at one time or another been used in oilseed rape to study genetic diversity and to produce the first molecular maps (Landry *et al.* 1991; Lombard and Delourme 2001; Udall, Quijada, and Osborn 2005). With the improvement of PCR technologies, new molecular markers have arisen: Simple Sequence Repeat (SSR), also known as microsatellite markers (Batley *et al.* 2007; Hopkins *et al.* 2007; Iniguez-Luy, Voort, and Osborn 2008; Tommasini *et al.* 2003). SSR markers known to be robust and highly polymorphic have been used to study genetic diversity in Brassica (Bus *et al.* 2011; Charters *et al.* 1996; Hasan *et al.* 2005; Havlíčková *et al.* 2014). The fact that SSR markers are abundant in the genome has led to the generation of higher density molecular maps (reviewed by Iniguez-Luy and Federico 2011).

The recent availability of sequences for Brassicas and the rapid rise of novel bioinformatics' approaches have given impulse to the development of more efficient and high-throughput molecular markers. Single nucleotide polymorphism markers (SNP) (Chagné *et al.* 2007) have become one of the most popular markers. A SNP is defined as a single base change in a DNA sequence or small insertions / deletions that occur in a significant proportion (more than 1%) of a large population. They occur throughout the genome and represent the most frequent type of polymorphism and thus are available in large numbers for almost any organism (Edwards *et al.* 2007). Moreover, they can be considered as relatively stable, with no significant change from generation to another. SNPs are usually bi-allelic, although, in theory, four different variants (4 nucleotides) could be observed. The different features of SNPs (low mutation rate, stability and their abundance) make them a potential tool to analyse genetic diversity, to study evolutionary process, to deal with genomic inheritance and to draw high density genetic map in crop species including species with ancient genome duplications, such as *B. napus*. Several studies have focused on the successful discovery and development of SNPs (Clarke *et al.* 2016, 2013; Dalton-Morgan *et al.* 2014; Delourme *et al.* 2013; Durstewitz *et al.* 2010; Trick *et al.* 2009). In a recent study, Hayward *et al.* (2012) detailed the latest technologies for SNP discovery as well as their potential use in Brassicas.

### 3.2. Molecular markers and OSR DUS testing

UPOV/INF/17/1 named “*Guidelines for DNA-profiling: molecular marker selection and database construction (“BMT Guidelines”)*” was adopted and published. This document highlights several recommendations based on criteria that should be assessed when choosing molecular marker in order to manage reference collections:

- Reproducibility of results between laboratories / detection platform and repeatability over time;
- Useful polymorphism level and discriminatory power;
- Known position on the genome (genetic map);
- Single locus markers with no null alleles are desirable;
- Publicly available;
- Strong amplification pattern, easy to score.

In the previous project conducted with SSR markers, multi allelic banding patterns were often observed and divergence in the reading between the different laboratories occurred. With SNP markers, due to their bi-allelic nature, chances to obtain more repeatable and reliable results are higher, as their scoring will in general be straightforward. However due to the same bi-allelic nature, a large number of SNP markers have to be used to maintain a sufficient discriminating power.

SNP calling within polyploidy plants might be more challenging due a dosage effect. To avoid problems often encountered when working with polyploids (false polymorphism, presence of homeologous genes, allele dosage; Mason 2015), primers design will focus on targeting markers present in only one of the parental genome (A or C). This selection is possible, as most of the SNPs are now mapped.

Keeping in mind that the long-term objective is to combine phenotypic and molecular distances in the management of OSR, increasing the number of markers and homogenizing their distribution through the whole genome may improve link between molecular and morphological distances.

Since UPOV/INF/17/1 recommends that more than one laboratory should be involved in the evaluation phase to select suitable markers for a given species, two laboratories from two different countries (France and United Kingdom) participated in this project.

### 3.3. Choice of SNP platform and technology

A large variety of SNP detection platforms and technologies are currently available. The choice of assay and technology will depend on the number of SNPs, sample size, flexibility (genotyping few samples with many SNPs or many samples with few SNPs), the level of multiplexing possible, length of the template sequence, assay development requirements, the type of equipment needed and cost design. Even if platforms for SNP calling will likely continue to change and evolve, SNP data that has already been obtained using older platforms will remain valid (Semagn *et al.* 2014). Both laboratories involved in the present project (France and United Kingdom) have previous successful experiences and access to the KASP technology (currently called Kompetitive Allele Specific PCR, or KASP™) from LGC Genomics (<http://www.lgcgenomics.com>). KASP is a very flexible assay, regarding the number of samples as well as the number of SNPs. Moreover, one of the most attractive features of KASP is the cost effectiveness and the duration of the synthesis of the assay. The KASP assay is a PCR based fluorescent SNP genotyping system. This technology uses a unique form of allele-specific PCR using a Fluorescence Resonance Energy Transfer (FRET) based homogeneous format. KASP is routinely used on sorghum and maize at GEVES and on barley and wheat at NIAB. Due to its flexibility, and regarding the number of samples, KASP technology is well adapted for this project.

### 3.4. Pooling strategy

When using molecular markers, working with bulks is more cost effective than individual genotyping. To describe a variety, bulks provide one unique profile for a number of plants per variety i.e. one profile per variety instead of individual genotyping resulting in a number of profiles per variety. Therefore, bulk strategy consists in the conversion of alleles' frequencies into a binary code of present-absent alleles scoring.

Moreover, only alleles above a given threshold will be expected to be scored. The phenomena called "allele dilution" will be more important when increasing the size of bulks. While bulk of DNA may mask partially a part of the genetic diversity of a variety, it has to remain effective to study genetic relationship among different varieties. Also, reducing the bulk size will increase sampling variance, rendering the DNA profile inconsistent through replicated sampling. To our knowledge there is no consensus method to define bulking strategy.

The study aimed to optimise bulk size based on preliminary estimations of allele frequency from each variety. We analysed the reproducibility of the results through replicated samplings and studied the impact of bulking different type of tissues (seeds or leaves).

### 3.5. Experimental approaches

The study was carried out in 5 sequential phases:

- Phase 1: Marker selection/ varieties choice / DNA preparation
- Phase 2: Amplification test/ marker screening:
  - test of 500 selected SNPs on 23 individual plants: 275 SNPs per lab (50 SNPs in common)
- Phase 3: Working in bulk and size of the bulk
  - Selection of 90 SNPs => 50 SNPs per lab (10 in common)
  - Work on 4 varieties => individual genotyping of 32 plants per varieties + synthetic DNA bulk
  - According to previous results => genotyping of leaves and/or seed bulk? Size?
- Phase 4: Variety screening
  - Method chosen in phase 3 applied to the 19 remaining varieties
- Phase 5: Data analysis and review

## 4. PHASE 1- Marker selection, primer design / varieties selection

### 4.1. Milestone: selection of a set of SNP markers, synthesis of molecular primers and distribution of the primers to the two laboratories (FR)

- Primer selection

The first phase was used to select SNP set and to distribute it to the partner.

The recent development and use of different arrays in oilseed rape (Clarke *et al.* 2013; Dalton-Morgan *et al.* 2014; Delourme *et al.* 2013) has led to the development and commercial release of a *Brassica* 60K Illumina Infinium™ array, containing a total of 52 157 SNPs (Clarke *et al.* 2016). For our purpose (essentially germplasm characterisation) we did not need to use the whole SNP set from the array. From our knowledge of previous work on other species, we decided to select a subset of 500 SNPs with a uniform distribution through the genome.

INRA (French National Institute of Agricultural Research) granted access to genotyping data obtained on 406 oilseed rape varieties with the 60K Brassica SNP array as well as the list of the 60K SNP array and their position on the *B. napus* genome to allow the selection of a test marker set. The genotyping datasets were essential to optimise the selection. First, SNPs on the 60K SNP array were identified on different genepools (winter-type, spring-type, etc), as they may behave differently in terms of polymorphism for different populations. Out of the 406 genotypes, only the 170 unique “varieties” of winter oilseed rape known to belong to the French reference collection were considered. The calculation of PIC (Polymorphism Information Content), percentage of missing data, percentage of observed and expected heterozygous and number of alleles was carried out on the 170 genotypes by GEVES with internal R script.

The first criteria for the SNP set was an even distribution of the markers throughout the *B. napus* genome based on the physical position. To have an even genome coverage, the genome was divided in 500 segments of equal size, each chromosome having a different number of segments.

For each segment we applied different filtering steps:

1. A filter was applied to select SNPs with a known position and associated with a single linkage group using the available annotation of A (*B. rapa*) and C (*B. oleracea*) diploid Brassica genomes.
2. We selected SNPs with a high percentage of call rate over 97% (equivalent to less than 3% of missing data).
3. We eliminated SNPs with a high percentage heterozygote call rate (over 5%).
4. We discarded SNPs that are monomorphic on our dataset.
5. The last criteria were based on PIC values:
  - a. On each segment we first have selected SNPs with a PIC value greater than 0.3;
  - b. For segments not represented following selection with criteria (a), SNPs with a PIC value between 0.2 and 0.3, taking the highest value when possible, were selected;
  - c. Finally, for the remaining segments with no suitable SNP after criteria (a) and (b) had been applied, SNPs with PIC value between 0.1 and 0.2 were selected.

At the end of this filtering procedure, 431 SNPs (as shown in green, Figure 1) were selected. However, this did not give full coverage of all segments of the genome (69 segments were not represented). SNPs on the nearby segments were selected, applying the same filtering steps resulting in the selection of 61 more SNPs (as shown in blue, Figure 1).

After this second filter, eight segments still did not display any suitable SNP. These segments are therefore not covered by the marker set. To reach the target marker number (500) in this initial phase,

the eight “best SNPs” were selected. “Best SNPs” were defined by the highest PIC value (with no consideration of the genome segment) with no missing data, and lowest rate of heterozygous.

The distribution of the 500 markers on the 19 chromosomes of *B. napus* according to their PIC value is shown in Figure 2. For a large part of chromosome C07 it was difficult to find SNPs with PIC value greater than 0.1.

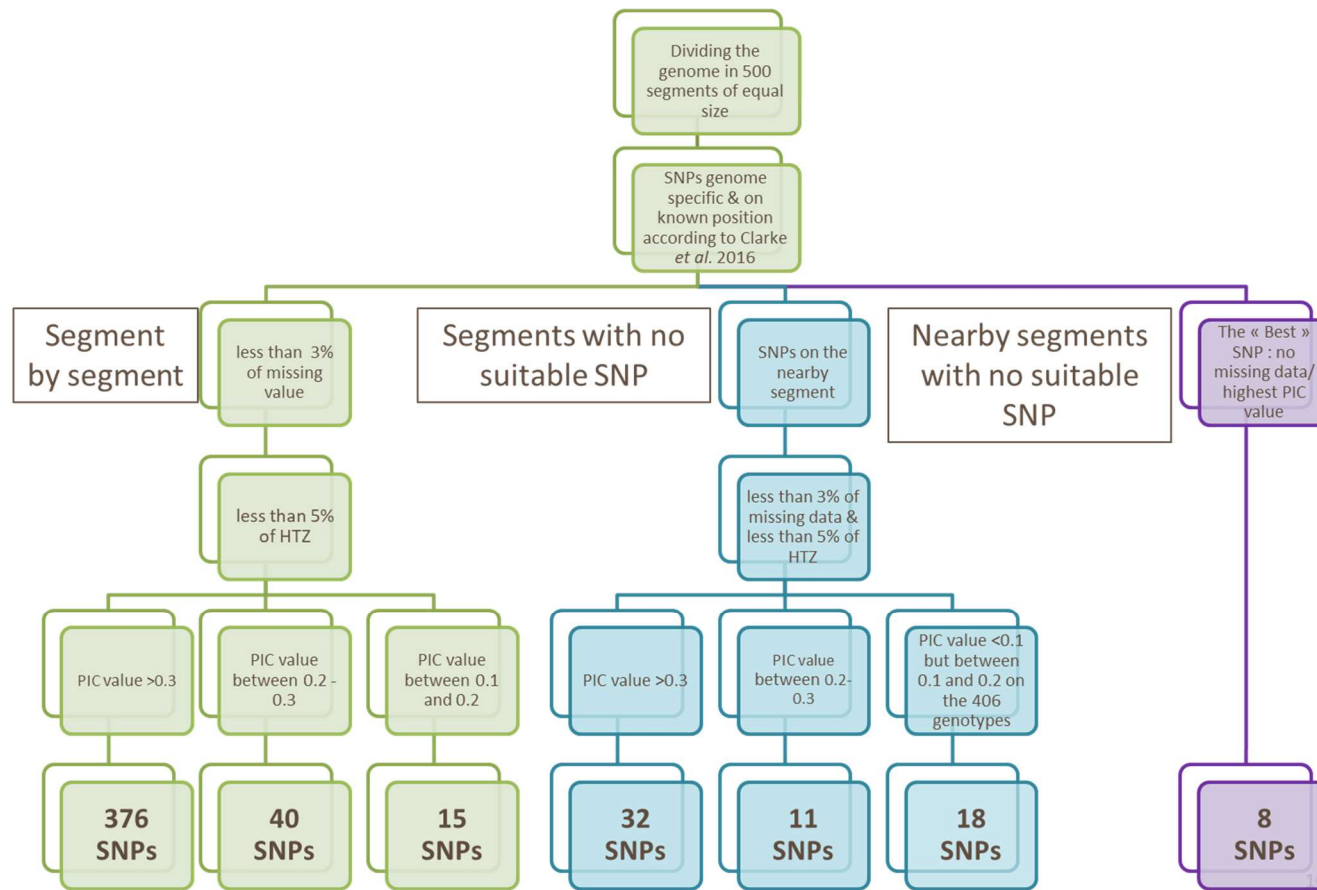
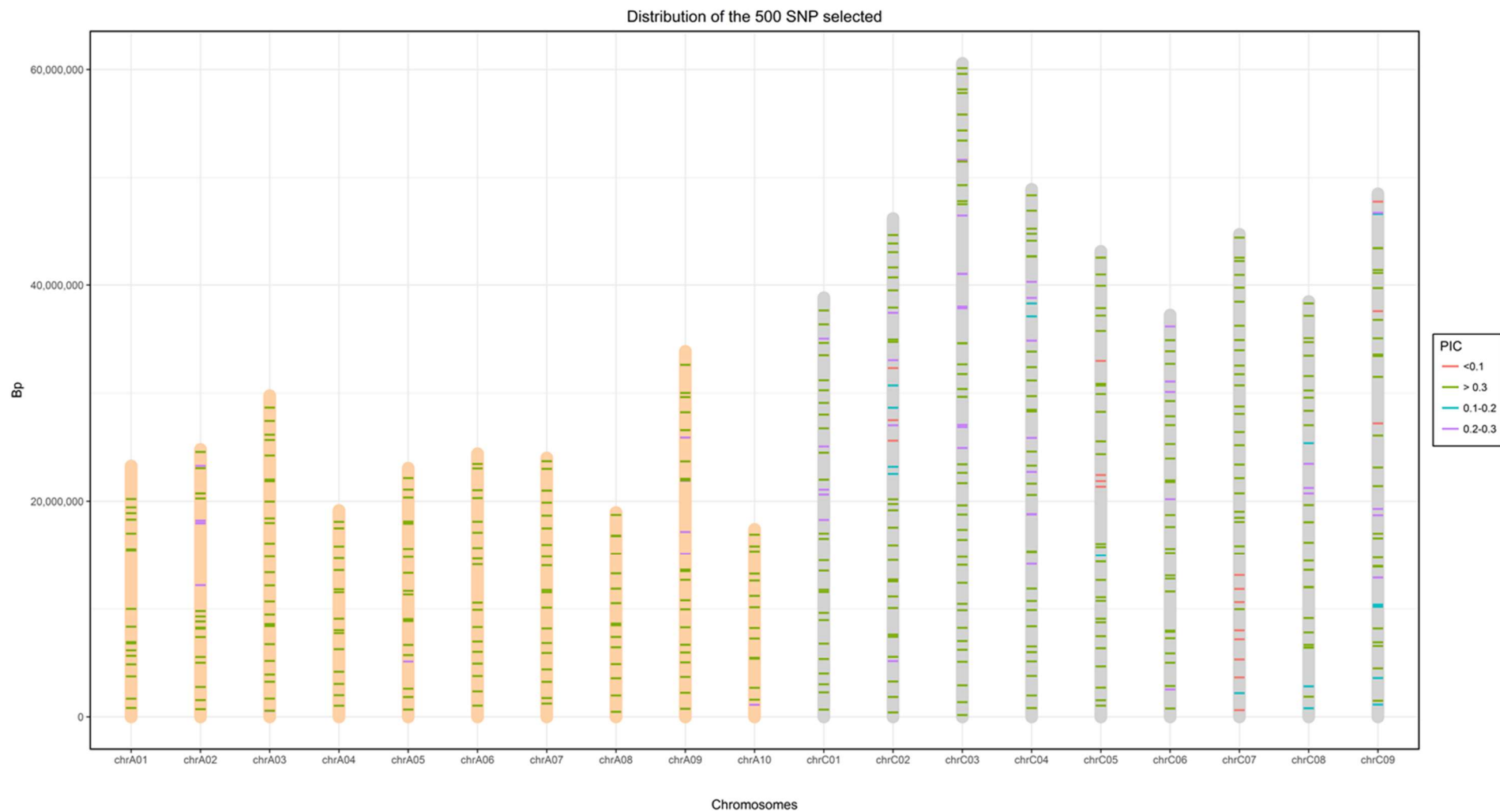


Figure 1. A schematic representation of the initial selection process used to identify 500 SNPs for conversion into KASP assay.



**Figure 2.** Distribution of the 500 SNPs selected during the initial filtering process, according to their PIC value.

- **Primer design**

The reference genome of the winter type 'Darmor- Bzh' was available on the *Brassica napus* Genome Browser (<http://www.genoscope.cns.fr/brassicanapus/cgi-bin/gbrowse/colza/>). As the position of each SNP was known (Clarke *et al.* 2016), the flanking regions on either side of the SNP position (70 base pairs before and after) were identified. All selected SNPs, with their nature are available in Annex 1.

We submitted all the sequences into the Kraken™ software (LGC Genomics). For three of the SNPs, marker design was not possible (mainly due to percentage of GC content). Selection of a further three SNPs in the corresponding segments with the highest PIC value was carried out to maintain the target number of markers.

A private provider, IDT-DNA, synthesised the KASP assays.

- **Primer distribution**

The partners tested half of the SNPs and, to prevent any technical problems leading to mis-scoring, 50 SNPs were tested in common between the two partners. SNPs were sorted according to their physical position in the genome of *B. napus* (Annex 1); every other SNP was selected to be used by the second lab (NIAB). To choose the 50 SNPs in common, the first two SNPs on each chromosome were selected as well as two last on six of the chromosomes (Annex 1).

## **4.2. Milestone: selection of varieties to be tested and material requirements (FR and UK)**

The two laboratories (FR and UK) agreed with a selection of 23 varieties which included inbred lines and hybrids, known to be homogeneous or heterogeneous either on field and molecular biology observations. Summary of the different hybrids and lines can be found in Table 1. Five hybrids (shown in bold in Table 1) were included in this study as well as their parental lines. Breeders have been informed of the use of their varieties in this pre-project, but to maintain confidentiality, varieties names/references have been removed.

For Phase 3, to model our bulking strategy, work was carried out on a sub selection of four varieties from the 23 (varieties in green in Table 1).

**Table 1. Characteristics of the 23 selected varieties.**

| Type          | Code          | Characteristics                                                                |
|---------------|---------------|--------------------------------------------------------------------------------|
| <b>HYBRID</b> | <b>Var_1</b>  | Cytoplasmic male sterility hybrid                                              |
| LINE          | Var_22        | Maintainer line                                                                |
| STERILE LINE  | Var_9         | Parental line                                                                  |
| RESTORER LINE | Var_10        | Parental line                                                                  |
| <b>HYBRID</b> | <b>Var_6</b>  | Cytoplasmic male sterility hybrid from a different genetic origin              |
| STERILE LINE  | Var_4         | Parental line                                                                  |
| RESTORER LINE | Var_3         | Parental line                                                                  |
| <b>HYBRID</b> | <b>Var_19</b> | Cytoplasmic male sterility hybrid from a different genetic origin              |
| STERILE LINE  | Var_16        | Parental line                                                                  |
| RESTORER LINE | Var_17        | Parental line                                                                  |
| <b>HYBRID</b> | <b>Var_5</b>  | Genetic male sterility hybrid                                                  |
| STERILE LINE  | Var_11        | Parental line                                                                  |
| LINE          | Var_12        | Maintainer line                                                                |
| STERILE LINE  | Var_13        | Parental line                                                                  |
| LINE          | Var_21        | Parental line                                                                  |
| <b>HYBRID</b> | <b>Var_14</b> | Genetic male sterility hybrid from a different genetic origin                  |
| STERILE LINE  | Var_20        | Parental line                                                                  |
| LINE          | Var_15        | Parental line                                                                  |
| LINE          | Var_7         | Statistically similar / different in field in close comparison                 |
| LINE          | Var_2         |                                                                                |
| LINE          | Var_18        | Standard variety for electrophoresis                                           |
| LINE          | Var_8         | Variety which is heterogenous on electrophoresis but uniform in the field      |
| LINE          | Var_23        | Variety heterogenous in the field (on leaves characteristics and on fertility) |

### 4.3. Milestone: Growing plant material and DNA extraction (FR)

One partner (GEVES) carried out all DNA extractions, with a portion being sent to the other partner (NIAB). This allowed bulk extraction of large amounts of DNA needed for all phases of the project and that both laboratories were using the same material. All DNA extractions were carried out using Macherey Nagel Kit in accordance with manufacturer's recommendations. Accurate DNA quantifications were made using measured with a SPECTROstar Omega® UV-vis Spectrophotometer (BMG Labtech). Before use, all DNA solutions were diluted to 20 ng/μl<sup>-1</sup>.

- **Growing material**

Seedlings were grown in a germination room until the fourth true-leaf stage (at least 3 weeks after germination) to give enough green plant material for each of the extractions for all phases of the project. Each tray contained at least 50 seeds of one variety. The maximum number of plants needed per variety was 32, but extra seeds were sown to ensure that adequate plant material was produced.

- **DNA for Phase 2**

For phase 2 (first screening on 500 SNPs) one plant per variety was harvested. For each plant over 3000 ng of DNA was needed, therefore two fully expanded leaves were used, with final DNA concentration ranging from 25 to 350 ng/μl<sup>-1</sup>.

- **DNA for Phase 3**

The requirement for DNA for phase 3 was different than in phase 2. Four varieties were selected to represent different criteria (inbred, hybrid, homogenous, non-homogenous) had been selected. All DNA sample types (as described below) were produced using the same individual plants with the obvious exception of the seed bulks.

- *Individual plants:*

32 plants from each variety were analysed to determine the degree of diversity within each variety. Since only 50 SNPs per lab were tested in phase 3, less material per individual was needed, therefore only two leaf discs from each plant were sampled. After DNA extraction final concentration ranged between 20 and 320 ng.μl<sup>-1</sup>.

- *DNA Bulks:*

Synthetic bulks were made by mixing aliquots of the samples used for the individual plants. Bulks containing the DNA of 8, 16, 24 or 32 individuals were produced. To check the consistency of the results, bulks were repeated three times (three independent random samplings) except for the bulk of 32 which can only be made up of all individuals. To determine which plants were to be put in each bulk, each plant/DNA sample was numbered and a random selection for each bulk was made. DNA samples were pooled together after individual extractions had been diluted to the same concentration (20 ng.μl<sup>-1</sup>).

- *Leaf bulks:*

For each bulk, two leaf discs from the individual plants were freeze dried separately and then pooled together before grounding. The final concentration ranged from 15 ng/μl<sup>-1</sup> for a bulk of 8 leaves, up to 125 ng/μl<sup>-1</sup> for the bulk of 32 leaves.

- *Seed bulks:*

Four different seeds bulks (containing 8, 16, 24 or 32 individuals) with 3 replicates of each bulk were prepared. The DNA samples of these bulks were extracted from different individuals

compared to all other samples in this phase of the project (seed is not viable after DNA extraction using this method).

- **DNA for Phase 4**

Extractions for Phase 4 were carried out following the analysis of results from Phase 3 as experimental approach was dependent on the conclusions of that section.

## **5. PHASE 2 - Amplification test/markers screening**

In order to screen the suitability of the chosen SNPs, amplification tests were carried out on the 500 SNPs was carried out by both labs.

### **5.1. Experimental approaches**

Genotyping with KASP was performed in two different platforms with different equipment.

#### **a. FR**

The total reaction volume for PCR amplification was 4 µl in a 384-well plate, composed of 1.95 µl of 2× KASP master mix, 0.053 µl primer mix, and 2 µl template (4ng of genomic DNA). KASP master mix contained the FRET reporting system (FAM, HEX) and PCR reagents. Primer mixes, unique for each SNP, contained two alleles - specific primers and one common primer. PCR was performed as follows: Hot start at 94°C for 15 min, followed by 10 cycles of 94°C for 20 s and 61°C for 1 min (dropping -0.6°C per cycle to achieve a 55°C annealing temperature), then followed by 26 cycles of amplification (94°C 10 s; 55°C 1min).

Assays were performed in a hydrocycler (LGC Genomics) and fluorescence was detected using a Microplate Reader FLUOstar Omega (BMG Labtech). Data was then analysed using Kraken software (LGC Genomics) to identify SNP genotypes. If the signature genotyping groups had not formed after the initial amplification, additional amplification cycles (3) were applied, and the samples were read again. The last step (additional cycles) was redone as many times as necessary depending of the SNP.

#### **b. UK**

The total reaction volume for PCR amplification was 5 µl in a 384-well plate, composed of 2.5 µl of 2× KASP master mix, 0.07 µl primer mix, and 2.5 µl template (2.5ng of genomic DNA). KASP mix was the same between the two laboratories, as well as PCR conditions except for the total number of amplification's cycles which was only 25 for UK.

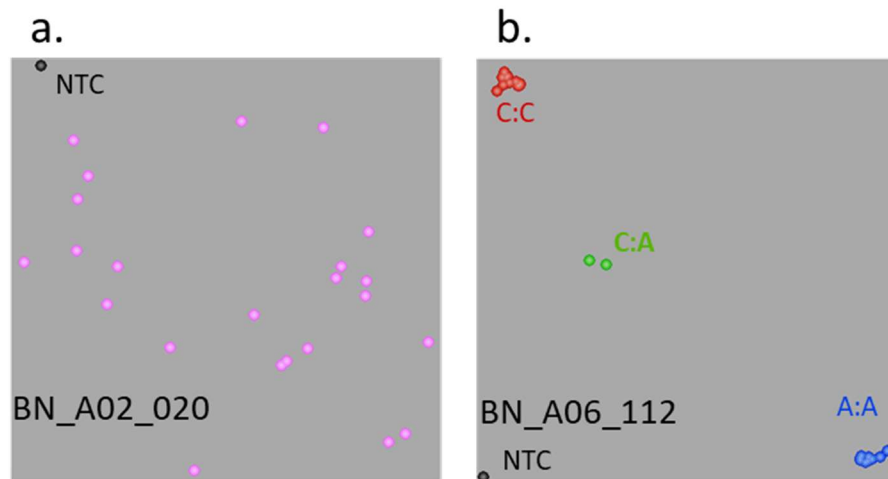
Different equipment was used for amplification: ABI biosystems Proflex PCR thermocycler and ABI Biosystems Quantstudio 7 Flex. Fluorescence was detected using ABI Biosystems Quantstudio 7 Flex. Identification of SNP genotypes was performed using ABI biosystems Quantstudio Real Time PCR software. Like France, additional amplification cycles (5) were applied if the signature genotyping groups had not formed after the initial amplification and the samples were read again. Additional cycles were done as many times as necessary.

### **5.2. Results**

#### **a. Development and validation of KASP assay**

Of the 500 SNPs tested, 36 SNPs failed (6.8%) so genotype call was not possible (Figure 3.a). Of the remaining markers, 12 SNPs were found to be monomorphic on the 23 varieties tested and therefore not kept for further analysis in this project. The majority of the remaining 454 polymorphic SNPs showed a pattern with three distinct clusters: AA, AB, and BB (Figure 3.b). This classic diploid cluster pattern is expected for segregation of a biallelic marker at a single locus.

Physical positions of the SNPs which failed or which were monomorphic are shown in Figure 4 (red SNP), SNPs which were polymorphic are indicated in blue on the same figure. On the chromosome C07, SNPs on one arm showed very low PIC level on the 170 genotypes (Figure 2), the same SNPs failed amplification (Figure 4).



**Figure 3. Snapshot displaying SNP genotyping.** Snapshot (a) represents a failed pattern (in this case SNP BN\_A02\_020) and snapshot (b) shows a polymorphic pattern with 3 clusters (in this case SNP BN\_A06\_112). NTC for Negative Template Control.

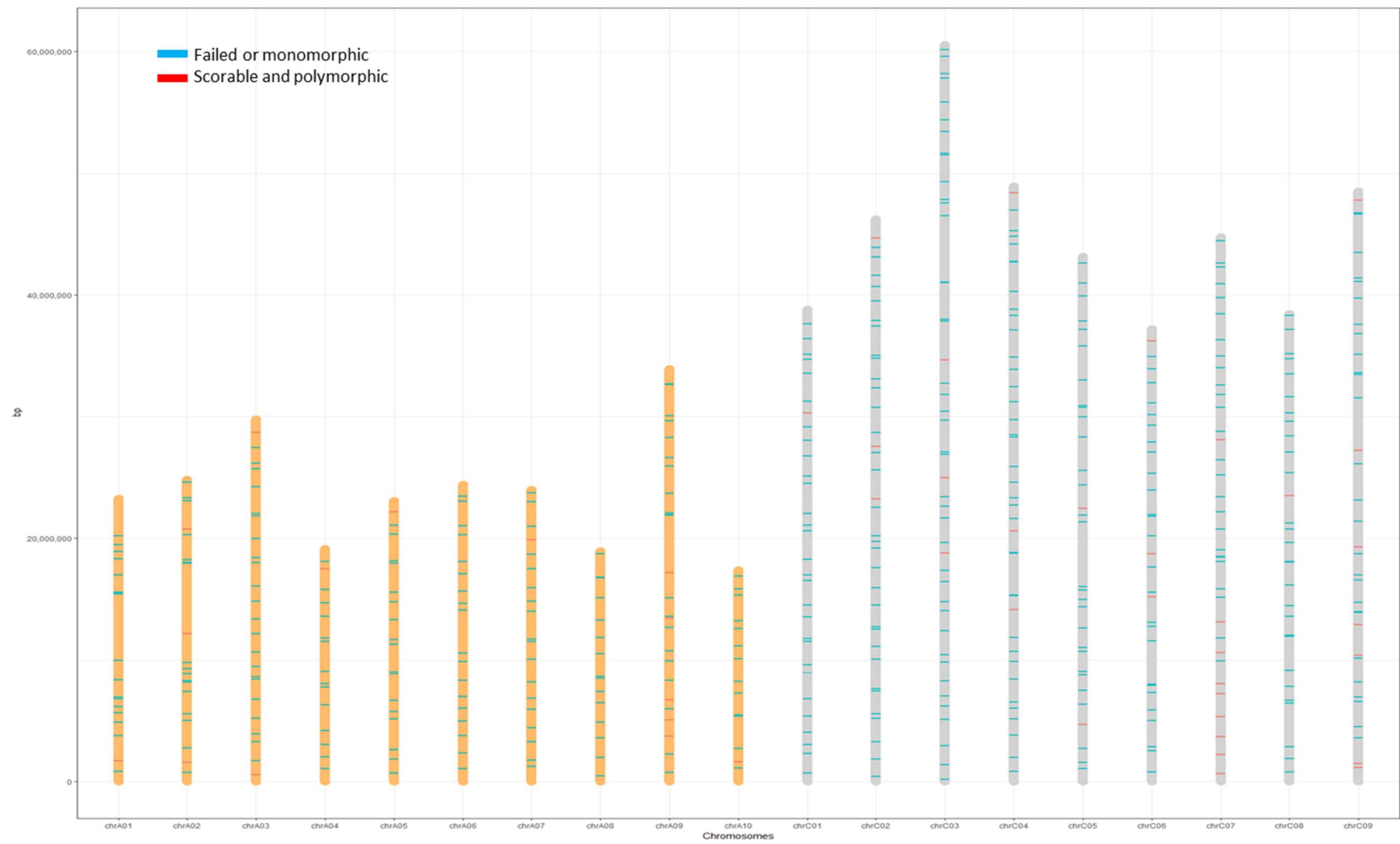
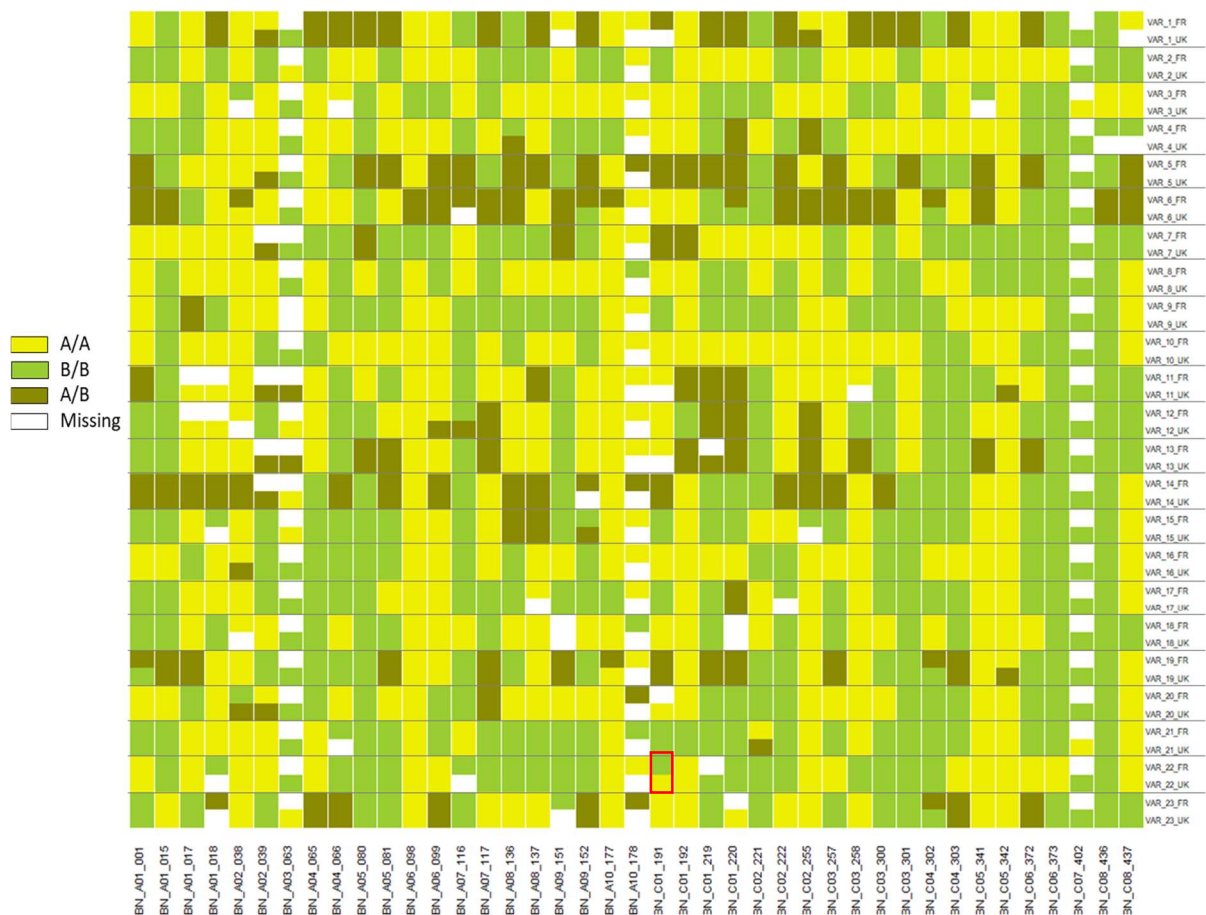


Figure 4. Physical distribution on the 19 chromosomes of the 500 SNPs whether they failed KASP development

## b. Results comparison between FR and UK

50 SNPs were genotyped by both laboratories on the same DNA from the 23 varieties. The mismatch rate among pairwise comparisons of the 23 varieties was 6.5% excluding mismatch due to missing data. Out of the 50 SNPs, nine SNPs were not able to be scored in either laboratory, due to weak or no amplification, or ambiguity in clearly differentiating the different clusters. For three other SNPs, calling was done in one lab (Figure 5), but not in the other laboratory. These SNPs were kept for Phase 4 but not for Phase 3. Two types of genotyping mismatches were observed among pairwise comparisons: a SNP scored as different homozygotes in the two laboratories, or a SNP scored as a homozygote in one but as a heterozygote in another laboratory. The second of the two types of mismatch being the most common (Figure 5).

There was a strong need for harmonisation of reading and genotype calling between the laboratories. Such mismatches genotype calls were minimised when genotype plots of each marker were checked and re-scored manually. After recalling and corrections, the mismatch rate was less than 2%. There was only one instance of the two laboratories calling different homozygotes for the same variety using the same marker (Figure 5, Var 22, SNP BN\_C01\_191). Given that the two laboratories were working on the same DNA, the remaining differences may be due to a genotyping error or still some error in genotype calling or problem during experiment (particularly DNA distribution which is very crucial for correct cluster calling).



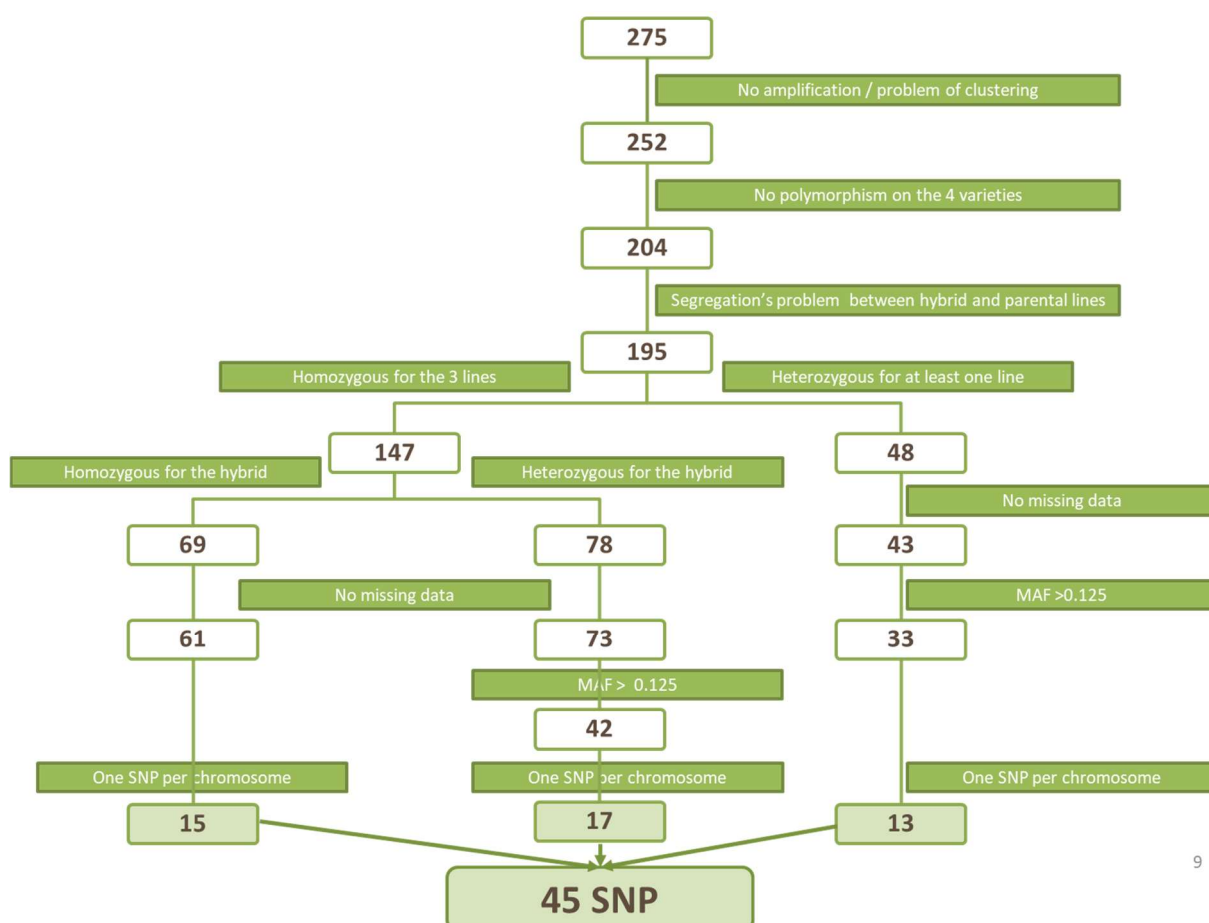
**Figure 5. Bitmap of polymorphism for pairwise comparisons of the 23 varieties (rows) and 50 SNPs in common (columns). Only SNPs that were scored are represented on the plot and results are following harmonisation.**

## 6. PHASE 3 - Working in Bulk and size of the Bulks

### 6.1. Milestone: SNP selection marker for phase 3

Four varieties, representing inbred (three lines) and hybrid (one line) breeding and differing degrees of uniformity (Table 1) were selected in Phase 3. This subset from the initial variety list was used to test the bulking strategies. The marker set for this stage of the project was also selected using a list of criteria, namely the Polymorphic Information Content (PIC), minor allele frequency (MAF), and observed heterozygosity, which were calculated using R software (R Core Team, 2013) and internal script.

The SNP selection process was a cascade involving several steps with the number of SNPs reduced after each step to reach a target of 90 markers. Figure 6 shows the details of the FR selection process; however, the UK process was similar in steps. Firstly, SNPs were filtered by the following steps: (1) polymorphic and no calling failure on the 23 varieties, (2) polymorphic on the 4 varieties, (3) no segregation problem between hybrid and parental lines. Afterwards the subset was split in 3 categories, whether SNPs were: (1) homozygous for the 4 varieties, (2) homozygous for the 3 lines but heterozygous for the hybrid, (3) heterozygous for at least one of the line. Only markers with no missing data and a MAF greater than 0.125 were kept. To narrow down the selection further (45 from each laboratory), one SNP per chromosome was chosen, as far as possible. A total of 90 SNPs was selected, 50 SNPs were run in each laboratory with 10 markers in common.



**Figure 6. The SNP selection process for Phase 3.** The specific filtering criteria have been reported for each filtering stage, as well as the number of resulting SNPs for FR

## 6.2. Milestone: scoring individuals and synthetic DNA (FR and UK)

### a. Experimental approach and statistical analyses

Work was carried out on the subset of four varieties out of the 23 varieties selected in Phase 1 (Table 1). One maintainer line was known to be homogenous (Var 22), one hybrid (Var 6), one line heterogenous in electrophoresis but homogenous in the field (Var 8) and one line heterogenous in the field for different morphological characters (Var 23) were selected. For each variety after DNA extraction of the 32 individual plants, SNP profiles of individual seeds were used to compute observed allele frequencies for each marker which would be considered as the expected frequency for each variety. Minor allele frequencies (MAF) were also calculated from the results of single seed assays. Genetic diversity was estimated through number of alleles ( $N_a$ ), observed ( $H_o$ ) and expected ( $H_e$ ) heterozygosity, for each locus and variety/bulk. Distance-based analysis was used to assess the relationship among the individuals within each variety. Rogers pairwise genetic distances (Rogers 1972) were computed using R (R Core Team, 2013) and the package *poppr* (Kamvar *et al.* 2014). The resulting similarity matrix was used to construct a dendrogram by means of the UPGMA (unweighted pair-group method with arithmetical averages). The reliability of branches was evaluated by the bootstrapping procedure using 100 random sampling of molecular markers.

For DNA bulk, since KASP technology is not a quantitative assay, only three classes of genotypes were recovered, the two homozygote classes (AA or BB) and one heterozygous class (AB). Bulk DNA results were converted into an arbitrary code resulting in an estimation of allele frequency:

Homozygous individual A/A: freq Allele1 = 1 and freq Allele2 = 0

Heterozygous individual A/B: freq Allele1 = 0.5 and freq Allele2 = 0.5

Expected allele frequency and bulk calling were compared.

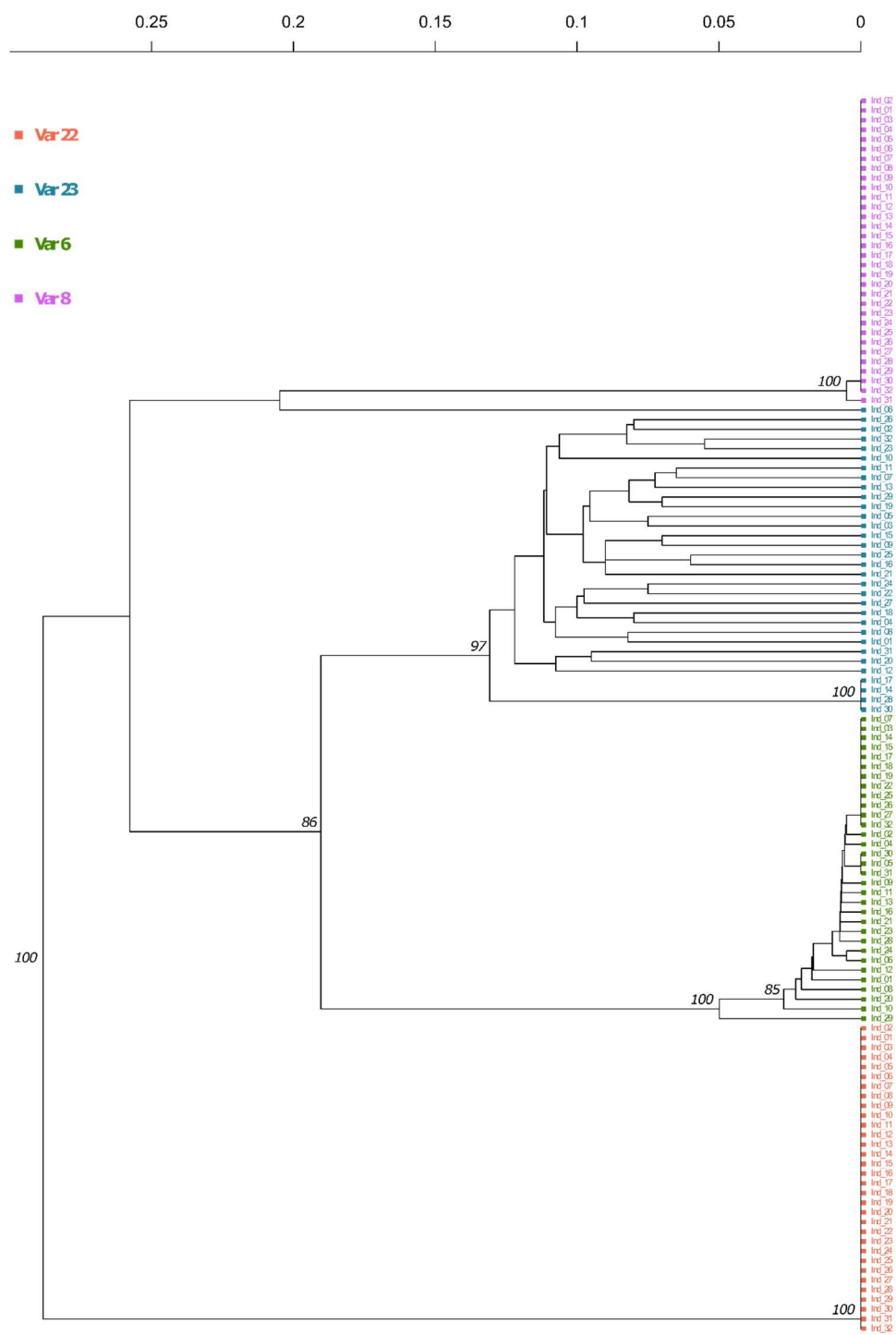
### b. Results FR

#### • Intra-varietal variability and individual genotyping

The dendrogram based on SNP markers distinguished the four varieties, which clustered into four groups (Figure 7). For two varieties (Var 22 and Var 8) no or very low level of intra-variety variability was observed (Table 2 and Figure 7). In Var 22 all individuals were identical and in Var 8 only one SNP for one individual was found heterozygous while all the others were homozygous. Except for the SNP previously mentioned no heterozygous was found within each of the two varieties. This could indicate that the varieties are fixed, however sub selection of SNP for phase 3 may have influence on the results as we selected SNPs for the fact that they were homozygous.

For the hybrid (Var 6) a high level of heterozygous was observed (Table 2) as expected from this type of breeding. Moreover, heterogeneity was observed across loci in this variety; both homozygous and heterozygous SNPs were observed at some loci within the set of 32 individuals indicating intra-varietal variability (Figure 7).

Var 23, known to be heterogenous, in the field showed not only a high rate of heterozygous (Table 2) but also the highest level of intra-varietal variability (Figure 7). For this variety, except for four individuals which were identical, the 28 remaining showed differences among them. Presence of individuals scored as homozygous for alternate SNP alleles was observed.



**Figure 7. UPGMA dendrogram based on genetic distances as estimated by 50 SNPs (FR) among the four varieties. Bootstrap values (Numbers above branches) were reported when above 80%.**

**Table 2. Summary of diversity indices number of alleles (Na), observed (Ho) and expected (He) heterozygosities per variety calculated for the individual plants and the difference size of bulk). Only mean between the 3 different samplings for each size of bulk is shown.**

| Variety | Pooled DNA | Number of individual | Na     | Ho     | He     |
|---------|------------|----------------------|--------|--------|--------|
| Var 22  | No         | 32                   | 1      | 0      | 0      |
|         | Yes        | 8                    | 1      | 0      | 0      |
|         | Yes        | 16                   | 1      | 0      | 0      |
|         | Yes        | 24                   | 1      | 0      | 0      |
|         | Yes        | 32                   | 1      | 0      | 0      |
| Var 8   | No         | 32                   | 1.0006 | 0.0006 | 0.0006 |
|         | Yes        | 8                    | 1      | 0      | 0      |
|         | Yes        | 16                   | 1      | 0      | 0      |
|         | Yes        | 24                   | 1      | 0      | 0      |
|         | Yes        | 32                   | 1      | 0      | 0      |
| Var 23  | No         | 32                   | 1.5161 | 0.2406 | 0.2797 |
|         | Yes        | 8                    | 1.532  | 0.5267 | 0.2678 |
|         | Yes        | 16                   | 1.54   | 0.54   | 0.27   |
|         | Yes        | 24                   | 1.54   | 0.551  | 0.29   |
|         | Yes        | 32                   | 1.54   | 0.5417 | 0.3    |
| Var 6   | No         | 32                   | 1.5743 | 0.5619 | 0.2939 |
|         | Yes        | 8                    | 1.58   | 0.58   | 0.29   |
|         | Yes        | 16                   | 1.58   | 0.58   | 0.29   |
|         | Yes        | 24                   | 1.58   | 0.58   | 0.29   |
|         | Yes        | 32                   | 1.58   | 0.58   | 0.29   |

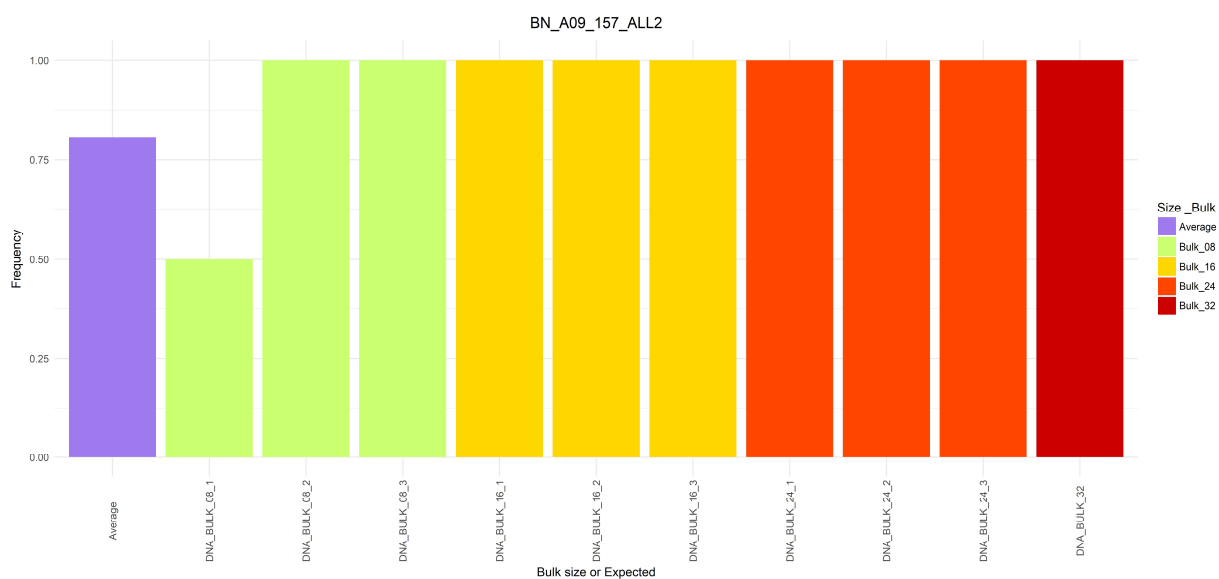
- Variation within bulks of same size and between bulks of different size**

It should be reiterated that each of DNA pools was constructed with different individuals (even if some are in common). Thus, differences between allele calling for DNA bulk incorporated sampling variance.

Allele frequencies were computed for each SNP and each variety (Purple bar on Figure 8). On the same plot, the results for the allele call for each bulk size and for the repetitions of each bulk were added. (Green for bulk of size 8, yellow for bulk of size 16, orange for bulk of size 24 and red for bulk of size 32). As consequence two plots per SNP and per variety were produced (meaning 400 plots). Each plot was observed individually.

For each size of bulk with equal amounts of DNA from 8, 16, 24 or 32 individual samples, very little differences were observed between repetitions of bulk sizes except for three bulks of 8 synthetic DNA (See example Figure 8) and one bulk of 16, and only for the two varieties showing intra-varietal variability. Differences were observed for four different SNPs only. Heterogeneity between bulks of the same size can result from presence of individuals comprising in the bulk that were homozygous for alternate SNP alleles or heterozygous. Reducing size of bulks increased sampling variance which may render DNA profile inconsistent.

As expected for the four bulks mentioned in the paragraph above, results were similar between bulks of the different size.



**Figure 8. Histogram showing the results for allele 2 of Var 23 for the marker BN\_A09\_157**

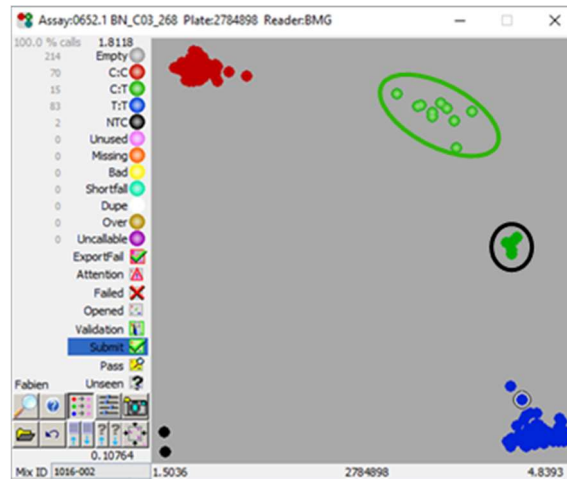
- **Comparison with expected allele frequency and threshold of detection**

The expectation was that the association between allele calling from bulk DNA and from individual genotyping would be greatest when both sets of estimates were based on the same individuals. Therefore, allele call from bulks including the 32 individuals would have been closer to the estimates of allele frequency estimated from the 32 independent individuals. Since only few differences were observed between bulks of difference sizes (bulk 8 and 16), correlation was equivalent between bulks of different size (Figure 12).

Bulk DNA genotyping was analysed as the conversion of allele calling into a binary code resulting from allele present-allele absent scoring. As mentioned previously, if the two alleles were present they were scored 0.5/0.5. It was expected that only those allele frequencies above a certain threshold were likely to be scored. Comparisons between allele calling from different size of DNA bulk and allele frequency estimated from individual experimental designs were carried out variety per variety (Figure 12).

For varieties which showed no intra-varietal variability and no heterozygous individuals (Var 22 and Var 8), allele callings from the bulk designs were similar to the estimates alleles frequency. For the hybrid (Var 6), allele remained undetected if MAF was less than 10% except for one SNP (BN\_C03-268). For this SNP, MAF was around 7% but all the bulks (size 8, 16, 24 and 32) were scored as heterozygous (Green circle on Figure 9). However, results from heterozygous bulk did not cluster with heterozygous individuals (Dark circle in Figure 9). Major allele frequency estimated for this SNP was 93%. On this SNP, effect of allele dosage was clearly observed (Figure 9), however since KASP is not a quantitative assay, it was decided that every dot which did not cluster with homozygous would be scored as heterozygous with no regards to allele dosage. SNPs with high rate of heterozygous individuals (MAF superior to 40%) were always scored as heterozygous.

For the variety 23, threshold for allele detection was more variable. For bulks of 8 and 16 which had shown different results, allele with a MAF over 35% remained undetected, this threshold goes down to 20% for bulks of 24 and 32 individuals.



**Figure 9. Snapshot displaying SNP genotyping.** Snapshot represents pattern for the SNP BN\_C03\_268 for individuals genotyping and DNA bulk (8, 16, 24, 32) for the four varieties of phase 3. Results of the bulks for Var 6 are circled in green. Dot circled in black are the heterozygous individuals for Var 6.

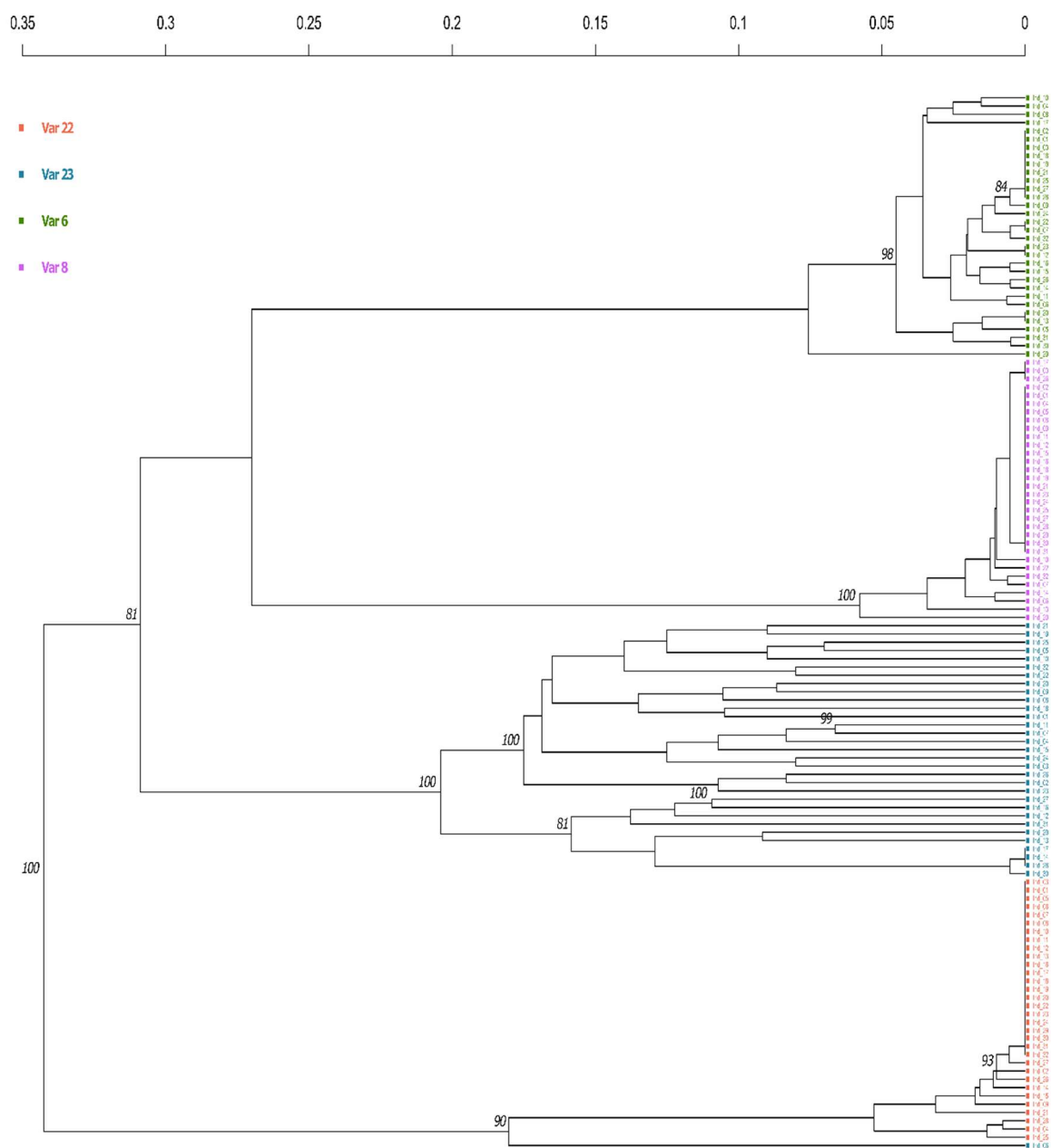
The study also showed that, although bulk DNA sampling analysis partially masked genetic diversity within varieties, it was effective to evaluate their genetic relationships as all bulks when analysed on the dendrogram clustered with their own variety (Data not shown).

### c. UK

#### • Intra-varietal variability and individual genotyping

The dendrogram (Figure 10) based on genetic distances among the four varieties as estimated by the 50 SNPs used in the UK (10 in common with FR) identified similar relationships as with the FR SNP set. Overall, there was more intra-varietal variation. Var 22 and 8 had the lowest amount of intra-varietal variation; however, both varieties showed more within-variety variation than in the FR dataset. As expected from a hybrid breeding system, there was a greater degree of variation, with only nine individuals being identical.

As with the FR SNPs, the greatest level of intra-varietal variation was observed in Var 23, with only three identical individuals. Of all varieties, only Individual 6 (ind\_06) of Var 23 did not cluster with other individuals from the same variety. According to the UK data, this specific individual was more closely related to Var 22 (FR SNP set linked it more closely with Var 8). Given that this variety showed the most within variety diversity (and appears non-homogenous in the field), it is understandable that within that set there is an unrelated individual.



**Figure 10. UPGMA dendrogram based on genetic distances as estimated by 50 SNPs (UK) among the four varieties. Bootstrap values (Numbers above branches) were reported when above 80%.**

- **Variation within bulks of same size and between bulks of different size**

The results from the UK were analysed using the same binary code as the FR data to describe the presence/absence of an allele. In some cases, the size of the bulk had little effect on the observed allele frequency (in line with FR results). However, the UK noted that there were more instances where the bulk size did affect the observed allele frequencies (an example is shown in Figure 11). The bulk made

up of all 32 individuals was the most successful bulk size at replicating the allele frequency calculated from the individuals.

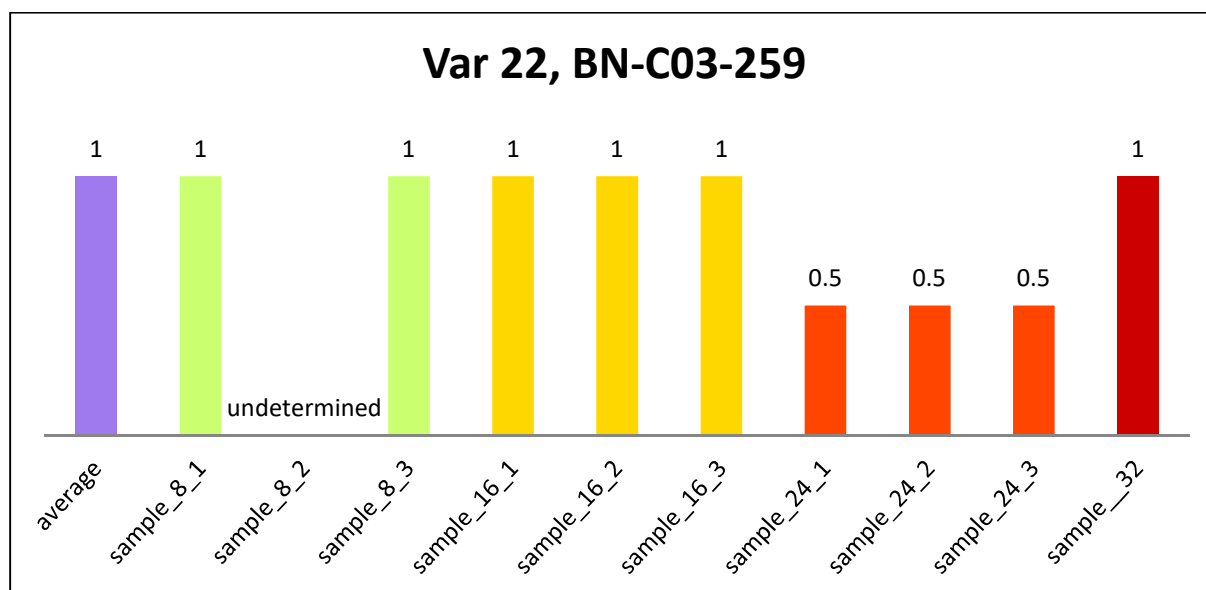


Figure 11. Histogram showing the results for allele 1 of variety 22 for the marker BN\_C03\_259.

- **Comparison with expected allele frequency and threshold of detection**

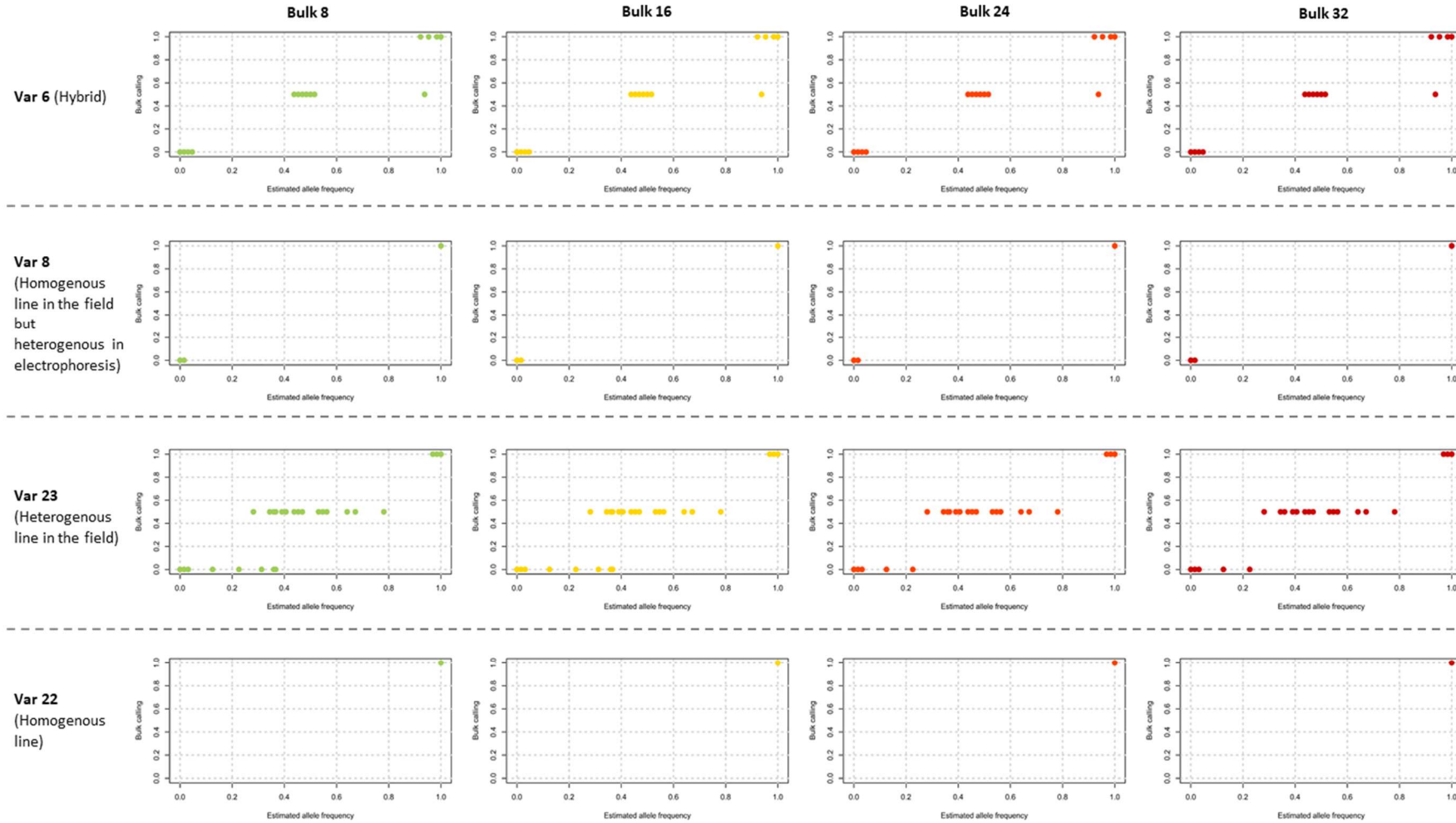
As described previously, the UK data set contained more variation than the FR dataset. This is apparent when comparing Figure 12 and Figure 13. However, the UK dataset still shows that there were few observable differences between bulks of different sizes.

#### d. Conclusions

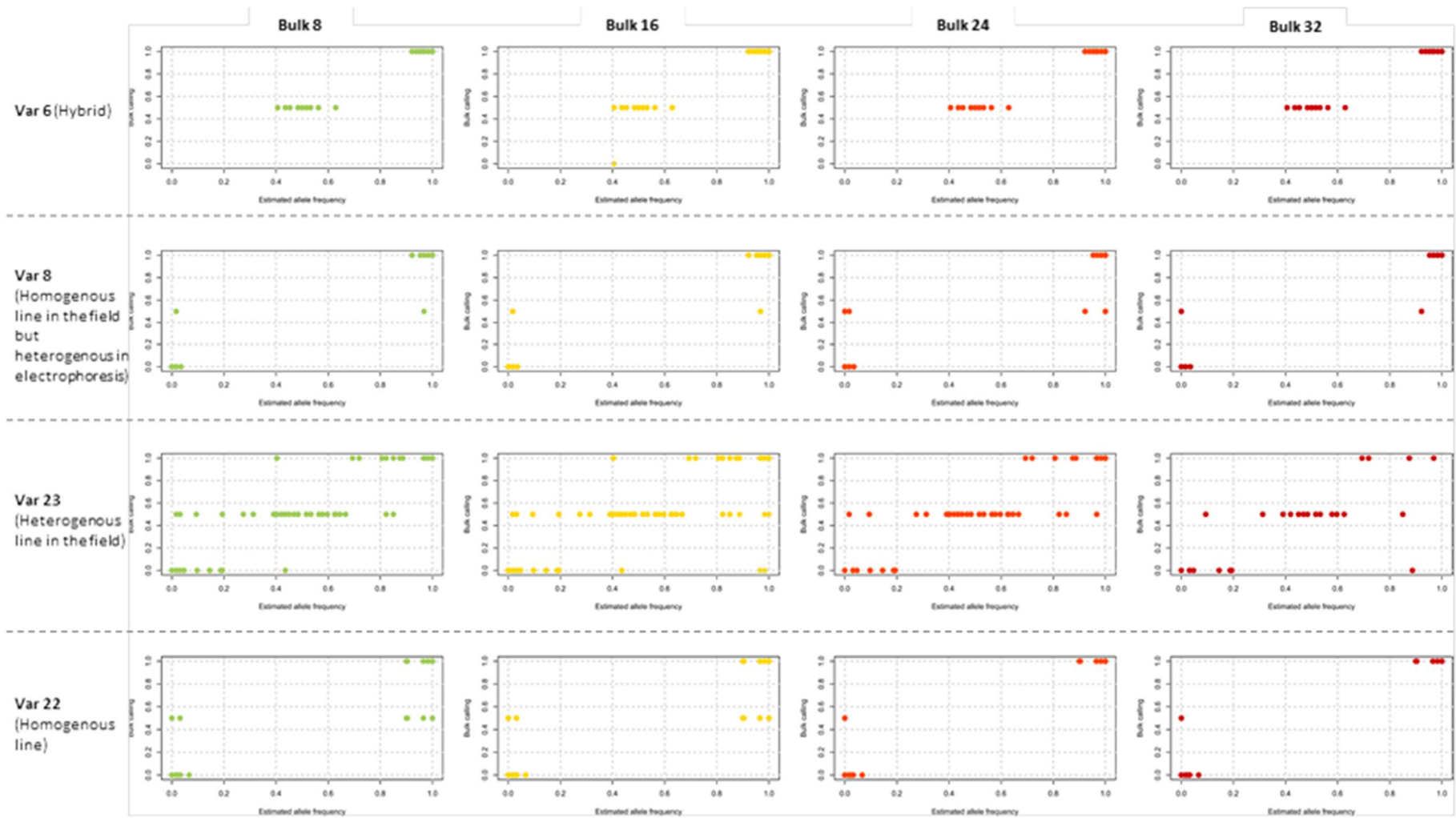
Evaluated allele frequency from bulk DNA worked well with very uniform varieties, even with hybrids with a high rate of heterozygous SNPs.

The challenge remained for heterogenous variety, which showed more variable allele frequencies. In this case bulk masked partially this variability since allele may only be called 0, 0.5 or 1. However when working in bulks of minimum size of 24 individuals, alleles with a minor frequency above 20% were always detected.

The UK results were not as clear cut as the results from FR. The systems in use in the two laboratories are based on the same principles, but the differences in practice could explain the difference in results. For example, the software used to analyse the allele calls are not the same, with the UK system being the more ambiguous/interpretive. That said the overall conclusions reached by the two laboratories were the same. To evaluate the use of leaf or seed bulks, since bulks of 8 synthetic DNA were the less consistent; it was decided to keep only bulks of size 16, 24 and 32.



**Figure 12.** FR Allele call from different size of DNA bulk and allele frequency estimated from individual experimental designs. The x-axis and y-axis categorizations correspond to individual and pooling values of SNPs, respectively. Allele 1 for the 50 SNPs are represented at the same time.



**Figure 13. UK Allele call from different size of DNA bulk and allele frequency estimated from individual experimental designs.** The x-axis and y-axis categorizations correspond to individual and pooling values of SNPs, respectively. Allele 1 for the 50 SNPs are represented at the same time.

### 6.3. Milestone: scoring seed and plant bulks with validation from synthetic bulks (FR and UK)

#### a. Experimental approaches and analysis

Using information gained from the synthetic bulk results, sampling strategies for the plant and seed bulks were assessed and only seed and plant bulks of size 16, 24, 32 were used. Work was carried out on the same subset of 4 varieties than in the previous phase (Table 1).

Bulk DNA results were converted as before into a code resulting in a rough estimation of allele frequency:

Homozygous individual A/A: freq Allele1 = 1 and freq Allele2 = 0

Heterozygous individual A/B: freq Allele1 = 0.5 and freq Allele2 = 0.5

Expected allele frequency calculated previously from the genotype of the 32 individual plants and bulk calling were compared.

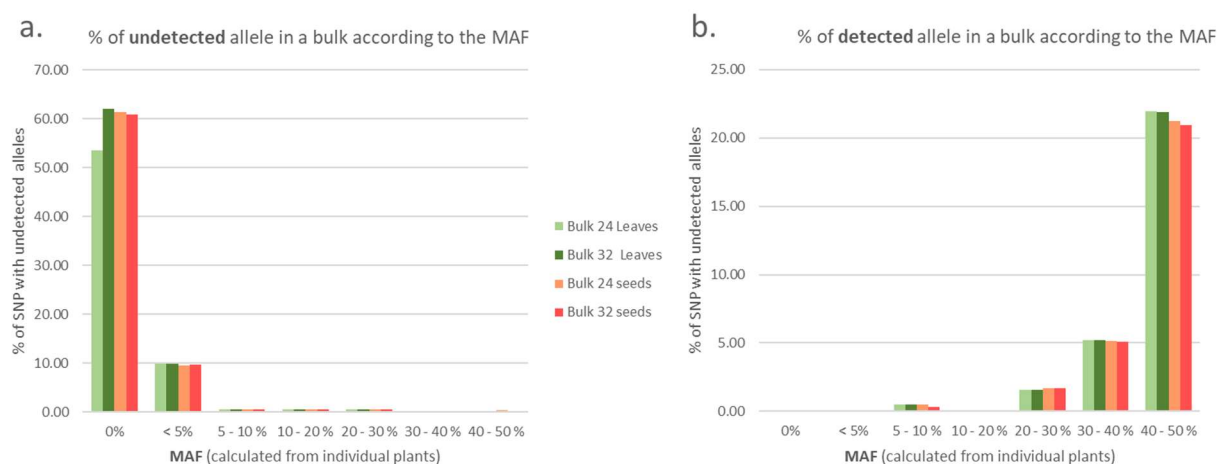
#### b. Results FR

- **Optimum size and type of bulk**

Each SNP and variety were analysed separately, comparisons were carried out between size and type (Leaf/seeds or DNA). Bulks of different and different type size (16, 24, 32 DNA/ leaves /seeds) gave similar results. Except for hybrid variety and only one SNP C03\_268, one bulk of 16 seeds scoring was different and the second allele was not detected. Constituent for the DNA bulks and leaf bulks were the same individuals however for seed bulks it was not the case, different sample was carried out for each bulk. The difference observed for this bulk may be due to the size of the bulk which would have increased the sampling variance or to a genotyping error.

- **Allele dilution/threshold of detection**

Only alleles above a given threshold will be expected to be scored. The phenomena called allele dilution will be more important when increasing the size of bulks. Allele with a low MAF should remain undetected in the bulk. According to the MAF calculated from the results of single seed assays and for each SNP, we reported for each SNP if allele with the lowest frequency was scored or not (Figure 12). As expected with no regards to bulk type or size, alleles not detected in the individual plants (MAF=0) were not scored in the bulk (Figure 14 a). For several SNPs (5 out of 50) with MAF inferior to 5%, second alleles were never detected in the bulks (Figure 14 a and b). For one SNP with a MAF superior to 40%, allele remained undetected in seed bulk. The SNP is difficult to read, and we could have faced errors in genotype call or genotyping errors. Increasing the size of the bulk did not seem to affect the threshold of detection since different size of bulk gave similar results (Figure 14).



**Figure 14. Detection of allele with lowest allele frequency (FR data).** (a) Percentage of SNP with undetected alleles across all loci tested according to the frequency estimated from individual genotyping. (b) Percentage of SNP with detected alleles across all loci tested according to the frequency estimated from individual genotyping. Each bulk type and size represented with a different colour.

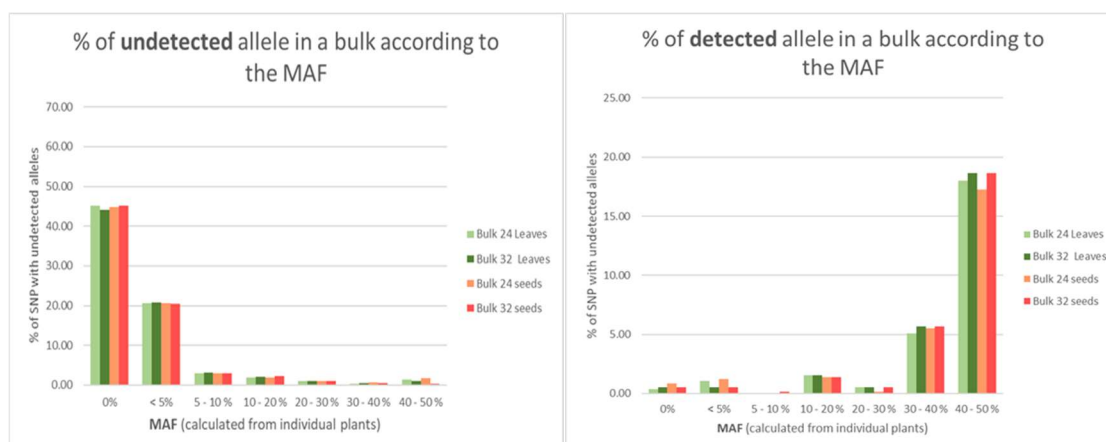
### c. Results UK

#### • Allele dilution/threshold of detection

It is worth noting that in 6 instances (0.25%) within the UK dataset, the results from a bulk sample did not match any of the individuals that had been included in the making of the bulk. For example, the result from Var 8, marker BN\_A04\_066 showed the genotype of C/T whereas all individuals included in the bulk were C/C. This is not easily explained and is most likely due to laboratory error. There were 11 instances of the observed genotype of the bulk being dictated by a few individuals within the bulk. In some cases, this was to be expected due to the level of known heterogeneity. For example, the results of three bulk samples (one of the 16's and one of the 24's) in Var 23, marker BN\_A01\_011 showed the same genotype as 11 of the 32 individuals. This genotype was also observed in one of the bulks from 24 seeds. On the other hand, in some cases less than three individuals dictated the genotype of the bulk, which is less likely to occur.

The bulks of seeds were not made using the same individuals and as such a direct comparison cannot be made. However, it is worth noting that 28 instances occurred where the genotype observed was not as expected based on the experience with the individuals observed in Phase 2 (eleven of which were previously unobserved genotypes).

The resulting dataset showed a greater degree of undetected alleles in the bulk according to the calculated MAF (Figure 15) compared to the FR dataset. The variation within a variety was higher, therefore there were more allele types that go unseen within a bulk sample. Due to time constraints and the scope of this project, the UK laboratory was unable to investigate these irregularities fully. This enquiry will be conducted independently of this project.



**Figure 15. Detection of allele with lowest allele frequency (UK data).** (a) Percentage of SNPs with undetected alleles across all loci tested according to the frequency estimated from individual genotyping. (b) Percentage of SNPs with detected alleles across all loci tested according to the frequency estimated from individual genotyping. Each bulk type and size represented with a different colour.

#### d. Conclusions

Again, as for Section 7.2, despite the incongruencies between the two datasets, the overall conclusions of the two laboratories were in agreement. Comparisons between the frequencies estimated from the bulks and the frequencies estimated from individual genotyping showed that for range tested, bulk size or bulk type had no significant effect on frequency estimations. It is important at this stage to consider the longer-term aspirations of this work. It is envisaged that this project is the first step in developing a system that will be adopted into the statutory DUS testing process. This means that any system will have to be robust as well as efficient (both in time and in resources). With that in mind, both laboratories agreed that extracting DNA from seeds was the preferred option as it avoids the growing stage. The quantity and quality of DNA was not affected by extracting from seeds, coupled with the efficacy of the allele detection, therefore making the seed bulks the most obvious choice. As there was little difference between the sizes of the bulk in either frequency estimations or repeatability (except bulks of 8 individuals), therefore based on the opinion of technical specialists, the next phase of the project would be carried out using bulks of 32 seeds.

## 7. PHASE 4 - Variety screening & data collation by partners for method validation (FR and UK)

### 7.1. Experimental approach

To secure our results and to optimise the accuracy of allele frequency estimation, two bulks of 32 seeds per variety were genotyped with the whole set of SNPs. DNA extractions were carried out according to the methodology previously described in phase 2. After DNA quantification and dilution of 200 µl of DNA at 20ng.µl<sup>-1</sup>, DNA solution was divided between the two laboratories. PCR and revelations were performed independently by each lab according to the two different protocols from phase 2. Only SNP which worked and were polymorphic in phase 2 were analysed (i.e. 236 for Fr and 256 for UK with 38 in common).

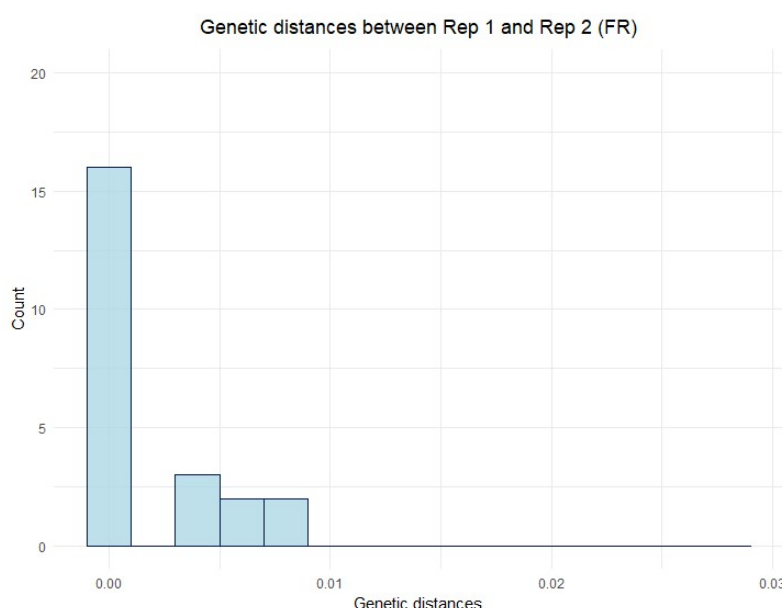
Genetic distance (Rogers 1972) was used to compare the results for the two-independent sampling for each variety. Results for the 38 SNPs in common between the two laboratories were also compared.

### 7.2. Milestone: Data analysis and review of the results

#### a. FR

- **Genetic distances between two different samplings of 32 seeds**

228 SNPs in average were used to calculate Rogers genetic distances between repetitions. Minimum distance of 0 and maximum distance of 0.009 were observed (Figure 16). Seventeen pairwise comparisons were equal to 0 indicating that sampling did not introduce sampling error. Differences between repetitions relied on a maximum of three SNPs. Heterogeneity between repetition was due to two different sources: heterozygous SNP reported for one repetition while the other was reported homozygous or repetition reported as homozygous for alternate SNP alleles. Differences between repetitions could be due to the sampling which may have introduced variance or to the error rate of genotyping. Indeed, LGC Genomics has reported KASP genotyping error rate inferior to 0.3%. One SNP BN\_C05\_364 was reported three times with alternate SNP alleles. This could indicate an error due to the sampling or a problem with primer design (multi locus, allele competition, allele specificity).



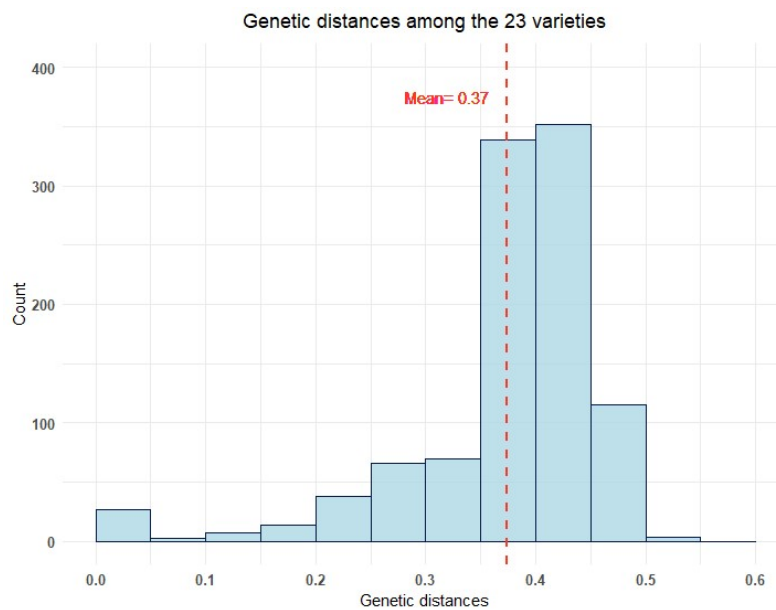
**Figure 16. Histogram of the pairwise genetic distances between the 2 sampling of each variety (FR)**

- **Segregation between parental lines and registered hybrid**

Five registered hybrids were included in this project. Segregation between parental lines and hybrids were studied. According to the genotypes of the parental line, genotypes of the expected hybrids was determined and compared to the genotype observed. For one SNP BN\_A08\_146, hybrids were all expected heterozygous C/T but always observed T/T, suggesting a problem of allele competitiveness during the PCR. To a slightly lower extent (for four hybrids out of five), this problem was encountered for six more SNPs (BN\_A07\_130; BN\_A08\_142; BN\_C02\_221; BN\_C02\_228; BN\_C03\_296; BN\_C04\_338). For a further project, these SNPs should be analysed carefully.

- **Genetic distances among variety**

Only the two different samplings from the same variety showed a genetic distance equal to 0.0 (Figure 17). Genetic distances (Rogers, 1972) between a sterile variety and its maintainer were always inferior to 0.1. Mean of the genetic distances on the 23 varieties was 0.37 with a maximum of 0.513. The SNP set selected was efficient to discriminate the 23 varieties.

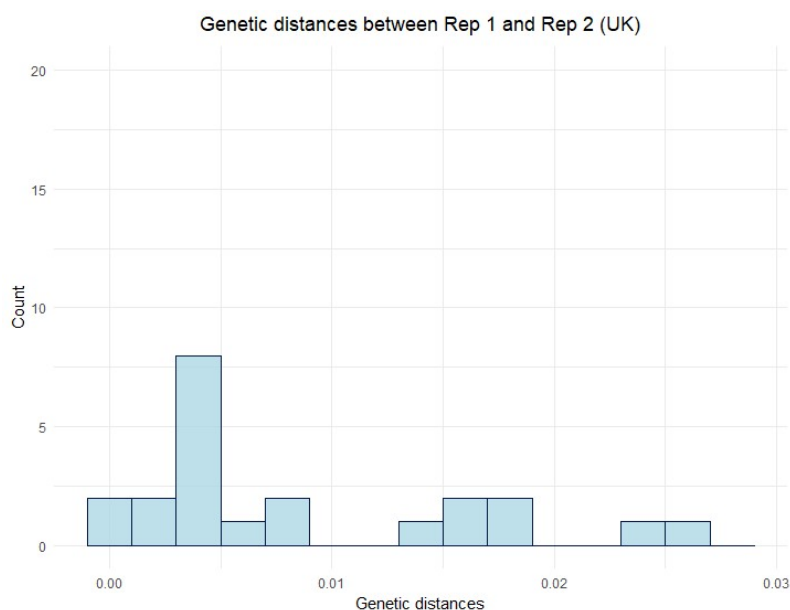


**Figure 17. Distribution of the genetic distances among the 23 varieties**

## **b. UK**

- **Genetic distances between two different samplings of 32 seeds**

Rogers's genetic distance was calculated between the two samples (32 seeds in each) of each variety (Figure 18). The genetic distance was more variable in the UK data set. Heterogeneity between the two samples could be attributed to the two reasons reported for the FR data, e.g. sampling or primer design. However, it is worth noting that this dataset was produced using the same techniques as per Phase 3, so it is probable that errors occurring then, were also a contributing factor to the variation in genetic distances here.



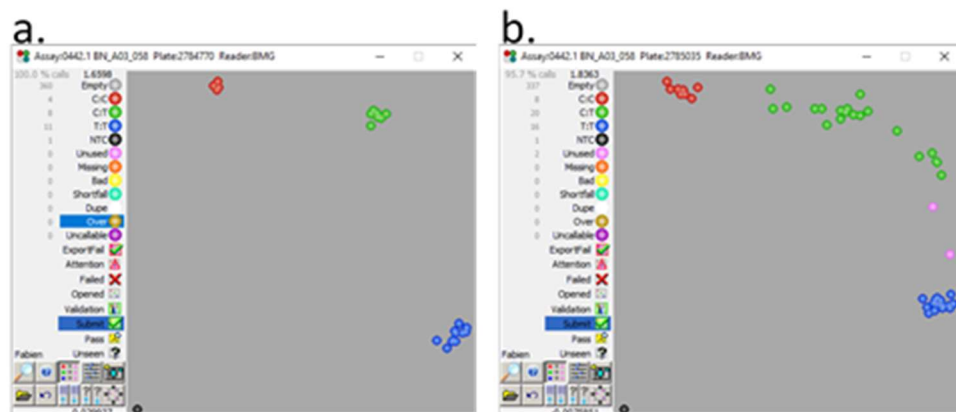
**Figure 18. Histogram of the pairwise genetic distances between the 2 sampling of each variety (UK)**

### c. Comparison between FR and UK

38 SNPs were common to the two laboratories on the same DNA. Repetitions from UK and from FR were compared two by two (Figure 20). Out of the 3496 data points, 136 missing data were observed (Figure 20) due to amplification problem or problem during the clustering process. Each variety was sampled two times and genotyped by each laboratory meaning a total of 1748 pairwise comparisons. Removing comparisons which showed a missing data in one or the other laboratory, a total of 1621 pairwise comparisons were analysed. 41 differences were observed between the two laboratories (i. e. 2.5% of differences). Out of the 41 differences 40 were due to different callings between homozygous and heterozygous. Two SNPs both showed six of the observed differences. For these two SNPs, different ways of calling heterozygous genotypes were observed. Only one SNP for one repetition showed the presence of two different homozygous probably due to the genotyping error. Assignment of heterozygous is challenging specially working in bulk due to different allele dose present in each bulk. These results highlight the fact that there is a real need for harmonisation of genotypes calling. Since primer design and first amplification test were carried out during the present project, no previous ring test could have been performed, however for a further project it is crucial to take into account the importance of such test between different laboratories.

### d. Pooling DNA vs. individual genotyping

Calling procedure when working on SNP profiles from bulked sample is more complex than analysis of single individuals. Clusters and especially heterozygous clusters are less sharp due to allele dosage when working in bulks (Figure 19). If points are intermediate between two clusters, they should be called as missing data instead of being wrongly assigned.



**Figure 19.Snapshot displaying SNP genotyping.** Snapshot (a) represents a pattern for the SNP BN\_A03\_058 268 for individuals genotyping (phase 2) and snapshot (b) shows for the same SNP for the same varieties but genotyped in seed bulks

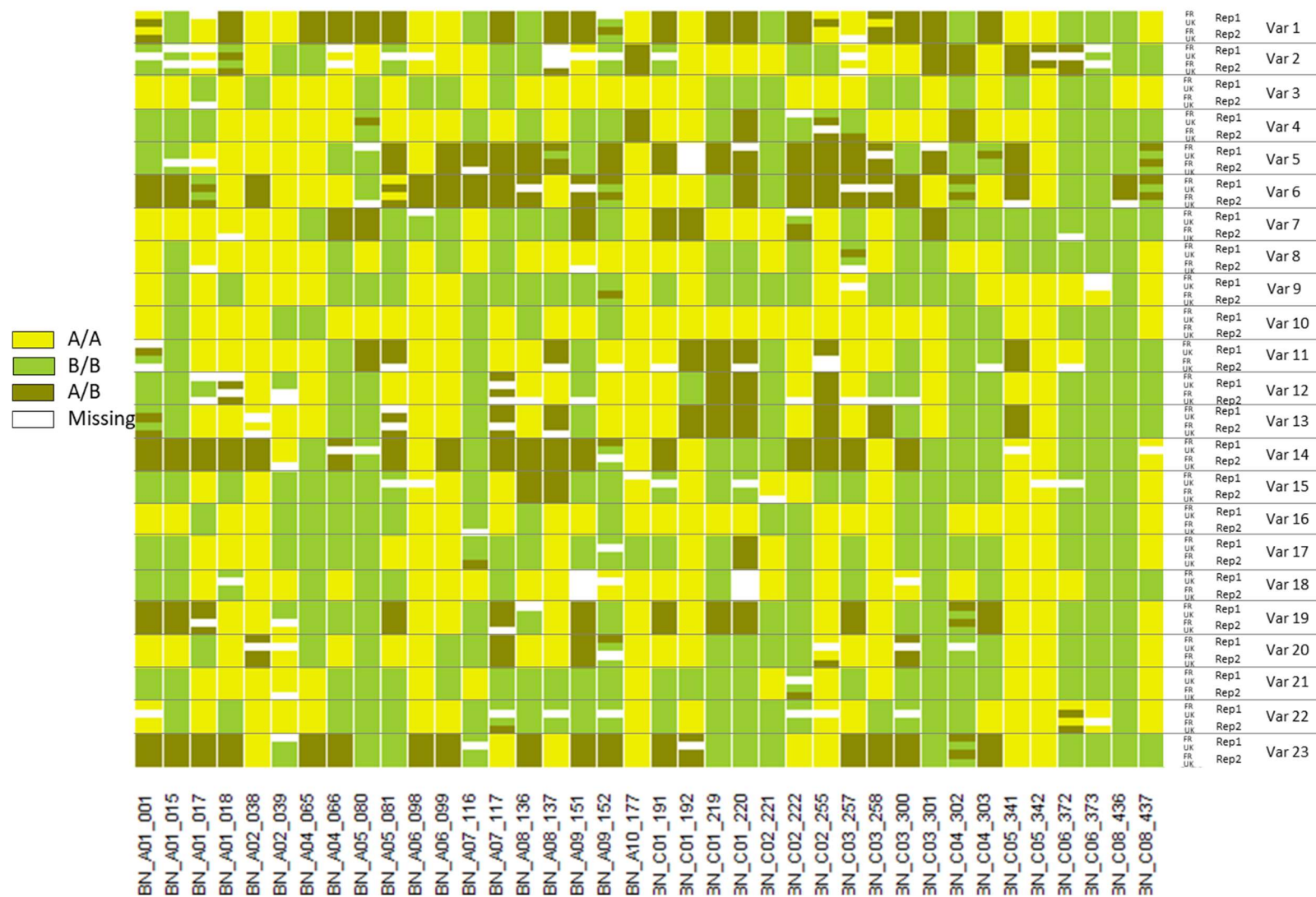


Figure 20. Bitmap of polymorphism for pairwise comparisons between repetitions carried out by each laboratory on the 23 varieties (rows) and 38 SNPs in common (columns).

## **8. PHASE 5 - Meeting with all entrusted EOs, ESA and the CPVO.**

Final meeting took place on the 7<sup>th</sup> of December at GEVES facilities, Anjouère, France. Eight EOs (Czech Republic, Denmark, France, Germany, Poland, Slovakia, Spain and United-Kingdom) were present as well as the CPVO and members of ESA.

- **Deliverables and conclusions**

SNP set and primer design for KASP assays are now available to study oilseed rape. Primer design was successful for 93% of the marker tested even if some SNPs were monomorphic on the 23 varieties studied in this project. This SNP set was efficient to discriminate among the 23 varieties studied. Keeping in mind that future project should focus on describing relationship among varieties within collections, this SNP set appeared to be a promising tool. However, improvements are still necessary. To validate a robust set of SNPs, more studies should be performed. Indeed, discriminating power, ease of cluster calling, reproducibility and genome specificity should be analysed further and on larger collection of variety, specially SNP which already showed potential problem of segregation. Moreover, for the 7% of the SNPs which failed amplification, new design could be considered.

Then, this study was designed to assess the feasibility of bulk sampling strategy and to determine optimum bulk sizes based on preliminary estimations of allele frequencies. In order to investigate the effect of bulk size and consistency through replicated sampling, DNA from different individual were pooled. Results showed that by increasing size of the bulk, the sampling variation decreased. Although number of alleles with low frequency (inferior to 0.2) remained undetected in the SNP analysis of bulk (Figures 14), the average number of allele per locus within a variety was highly correlated between analyses of individuals and bulks (Table 2), specially for varieties with very low intra-variability variation.

Even if a loss of information (specially for loci with low frequency alleles being undetected) may occur when working in bulk, our results indicated that bulks still provided reliable information to characterize winter oilseed rape varieties and to study diversity among them, based on genetic distances. As mentioned above, the 23 varieties were discriminated by the SNP set. Genotyping a greater number of varieties would allow us to better characterise diversity among oilseed rape cultivars and to assess further discriminating power of this marker set.

For bulks of minimum size of 24 individuals, very few differences between type of bulk (DNA, leaves, or seeds) were observed, thus bulking before DNA extraction offers great advantages with regard to time and cost effectiveness. According to the results working in bulk of seeds of at least 24 seeds is a good strategy to genotype different varieties of oilseed rape.

Finally working with two laboratories was challenging, even after harmonisation around 2% of differences remained between FR and UK. In this project, a time longer than initially thought was dedicated to the adjustment of the results between the two laboratories. First with KASP assay, efficiency of primer binding is dependent of the DNA sequence of the primer and hence different KASP assay reach completion at different rate (different number of PCR cycling). After initial PCR cycling; the data points within each cluster tend to be fairly spread out as consequence this could lead to a miscoring. Number of additional cycling used for each SNP, had to be determined and was a critical step which had to be discussed between the two laboratories. Moreover, there was a need to harmonise reading and genotyping calling between the two laboratories. Genotype calling for datapoint which were sparse is challenging. Attributing a data point to one or the other cluster could have created differences. After adjustment, remaining differences between the laboratories could be due to genotyping error, problem with primer specificity but also to laboratory error. Still with SNPs,

different platforms might be used however there is a real need to harmonise analysis and calling and to organise ring test when possible at first when starting a project.

- **Follow-up**

Since this preliminary project is successful, it should be considered as only the first step of a follow up project to further develop a strategy to apply SNP molecular markers in the framework of winter oilseed rape DUS testing.

For next step a go/no-go approach with the 2 following follows up will be considered. A **Follow up 1**, which should focus on 2 main issues:

1. To genotype a large number of OSR varieties and to reach an optimised SNP set for a large diversity of WOSR varieties.
2. To work on the use of this genetic data by testing existing UPOV model and by developing new ones well-adapted for this specie, both from historical field data.

According to the results, **follow up 2** would be to finalise genotyping of the whole collection and to validate and apply the chosen model in the field.

## 9. LITERATURE

- Batley, J., C. J. Hopkins, N. O. I. Cogan, M. Hand, E. Jewell, J. Kaur, S. Kaur, *et al.* 2007. 'Identification and characterization of Simple Sequence Repeat markers from *Brassica napus* expressed sequences'. *Molecular Ecology Notes* **7** (5): 886–889.
- Bus, A., N. Körber, R. J. Snowdon, and B. Stich. 2011. 'Patterns of molecular variation in a species-wide germplasm set of *Brassica napus*'. *Theoretical and Applied Genetics* **123** (8): 1413–1423.
- Chagné, D., J. Batley, D. Edwards, and J. W. Forster. 2007. 'Single Nucleotide Polymorphism genotyping in plants'. In *Association mapping in plants*, edited by Dr N. C. Oraguzie, Dr E. H. A. Rikkerink, Dr S. E. Gardiner, and Dr H. Nihal De Silva, 77–94.
- Charters, Y. M., A. Robertson, M. J. Wilkinson, and G. Ramsay. 1996. 'PCR analysis of oilseed rape cultivars (*Brassica napus* L. Ssp. *Oleifera*) Using 5' -anchored Simple Sequence Repeat (SSR) Primers'. *Theoretical and Applied Genetics* **92** (3–4): 442–447.
- Clarke, W. E., E. E. Higgins, J. Plieske, R. Wieseke, C. Sidebottom, Y. Khedikar, J. Batley, *et al.* 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** (10): 1887–1899.
- Clarke, W. E., I. A. Parkin, H. A. Gajardo, D. J. Gerhardt, E. Higgins, C. Sidebottom, A. G. Sharpe, R. J. Snowdon, M. L. Federico, and F. L. Iniguez-Luy. 2013. 'Genomic DNA enrichment using sequence capture microarrays: a novel approach to discover Sequence Nucleotide Polymorphisms (SNP) in *Brassica napus* L'. *PLoS ONE* **8** (12): e81992.
- CPV5766 (2008) Final report: Management of winter oilseed rape reference collections.
- Dalton-Morgan, J., A. Hayward, S. Alamery, R. Tollenaere, A. S. Mason, E. Campbell, D. Patel, *et al.* 2014. 'A high-throughput SNP array in the amphidiploid species *Brassica napus* shows diversity in resistance genes'. *Functional & Integrative Genomics* **14** (4): 643–655.
- Delourme, R., C. Falentin, B. F. Fomeju, M. Boillot, G. Lassalle, I. André, J. Duarte, V. Gauthier, N. Lucante, and A. Marty. 2013. 'High-density SNP-based genetic map development and linkage disequilibrium assessment in *Brassica napus* L'. *BMC Genomics* **14** (1): 120.
- Durstewitz, G., A. Polley, J. Plieske, H. Luerssen, E. M. Graner, R. Wieseke, and M. W. Ganal. 2010. 'SNP Discovery by amplicon sequencing and multiplex SNP genotyping in the allopolyploid species *Brassica napus*'. *Genome / National Research Council Canada = Génome / Conseil National De Recherches Canada* **53** (11): 948–956.
- Edwards, D., J. W. Forster, D. Chagné, and J. Batley. 2007. 'What are SNPs?' In *Association mapping in plants*, edited by Dr N. C. Oraguzie, Dr E. H. A. Rikkerink, Dr S. E. Gardiner, and Dr H. Nihal De Silva, 41–52.
- Hasan, M., F. Seyis, A. G. Badani, J. Pons-Kühnemann, W. Friedt, W. Lühs, and R. J. Snowdon. 2005. 'Analysis of genetic diversity in the *Brassica napus* L. gene pool using SSR markers'. *Genetic Resources and Crop Evolution* **53** (4): 793–802.
- Havlíčková, L., E. Jozová, A. Rychlá, M. Klíma, V. Kučera, and V. Čurn. 2014. 'Genetic diversity assessment in winter oilseed rape (*Brassica napus* L.) collection using AFLP, ISSR and SSR markers'. *Czech Journal of Genetics and Plant Breeding* **50** (3): 216–25.
- Hayward, A., A. S. Mason, J. Dalton-Morgan, M. Zander, D. Edwards, and J. Batley. 2012. 'SNP discovery and applications in *Brassica napus*'. *Journal of Plant Biotechnology* **39** (1): 49–61.
- Hopkins, C. J., N. O. I. Cogan, M. Hand, E. Jewell, J. Kaur, X. Li, G. C. Lim, *et al.* 2007. 'Sixteen new simple sequence repeat markers from *Brassica juncea* expressed sequences and their cross-species amplification'. *Molecular Ecology Notes* **7** (4): 697–700.
- Iniguez-Luy, F. L., and M. L. Federico. 2011. 'The genetics of *Brassica napus*'. In *Genetics and Genomics of the Brassicaceae*, edited by Renate Schmidt and Ian Bancroft, 291–322. Plant Genetics and Genomics: Crops and Models 9. Springer New York.

- Iniguez-Luy, F. L., A. V. Voort, and T. C. Osborn. 2008. 'Development of a set of public SSR markers derived from genomic sequence of a rapid cycling *Brassica oleracea* L. genotype'. *Theoretical and Applied Genetics* **117** (6): 977–985.
- Kamvar Z. N., J. F. Tabima, N. J. Grünwald. 2014 'Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction.' *PeerJ* **2**:e281
- Landry, B. S., N. Hubert, T. Etoh, J. J. Harada, and S. E. Lincoln. 1991. 'A genetic map for *Brassica napus* based on restriction fragment length polymorphisms detected with expressed DNA sequences'. *Genome* **34** (4): 543–552.
- Lombard, V., and R. Delourme. 2001. 'A consensus linkage map for rapeseed (*Brassica napus* L.): construction and integration of three individual maps from DH populations'. *Theoretical and Applied Genetics* **103** (4): 491–507.
- Mason, A. S. 2015. 'Challenges of genotyping polyploid species'. *Methods in Molecular Biology* (Clifton, N.J.) **1245**: 161–168.
- R Core Team (2013). R: A language and environment for statistical computing. R Foundation for statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
- Rogers, J. S. 1972. 'Measures of genetic similarity and genetic distance'. *Studies in Genetics* **7** (7213): 145–153.
- Semagn, K., R. Babu, S. Hearne, and M. Olsen. 2014. 'Single nucleotide polymorphism genotyping Using Kompetitive Allele Specific PCR (KASP): overview of the technology and its application in crop improvement'. *Molecular Breeding* **33** (1): 1–14.
- Tommasini, L., J. Batley, G. M. Arnold, R. J. Cooke, P. Donini, D. Lee, J. R. Law, *et al.* 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** (6): 1091–1101.
- Trick, M., Y. Long, J. Meng, and I. Bancroft. 2009. 'Single nucleotide polymorphism (SNP) discovery in the polyploid *Brassica napus* using solexa transcriptome sequencing'. *Plant Biotechnology Journal* **7** (4): 334–346.
- Udall, J. A., P. A. Quijada, and T. C. Osborn. 2005. 'Detection of chromosomal rearrangements derived from homeologous recombination in four mapping populations of *Brassica napus* L.' *Genetics* **169** (2): 967–979.
- UPOV. 2013. 'TGP/15/1 Guidance on the use of biochemical and molecular markers in the examination of distinctness, uniformity and stability (DUS)', Geneva.
- UPOV. 2015. 'TGP/9/2 Examining Distinctness', Geneva.
- UPOV. 2010. 'UPOV/INF/17/1 Guidelines for DNA-Profiling: Molecular marker selection and database construction ("BMT Guidelines")', Geneva.
- UPOV. 2011. 'UPOV/INF/18/1 Possible use of molecular markers in the examination of Distinctness, Uniformity and Stability (DUS)', Geneva.