CPVO · OCVV
Community Plant Variety Office
Office Communautaire des Variétés Végétales

Submission date 10/10/2018

CPVO R&D

AGREEMENT NUMBER – 7515057

1. Title

International harmonisation and validation of a SNP set for the management of tomato reference collection.

2. Details of Coordinator(s) (Institution, address and contact data)

Naktuinbouw
Sotaweg 22
2371 GD Roelofarendsveen
The Netherlands
+31 71 3326262
Contact person(s):
DUS, Variety and Trails; Raoul Haegens (r.haegens@naktuinbouw.nl)
Laboratory R&D; Sebastiaan Flanderhijn (s.flanderhijn@naktuinbouw.nl)

3. General Information

Species	Duration (in month)	Granted extension (in month)
Solanum lycopersicum L.	30	20

Total costs	Percentage of (co)financing
€ 295.000	100%

International harmonisation and validation of a SNP set for the management of tomato reference collection – Final Report

Objective and scope
The scope of this project is to develop an internationally harmonized and validated marker set (SNPs) suitable to genetically differentiate tomato varieties.

In more detail, this project aims to test SNPs for performance by different participating laboratories using different genotyping methods. The different executed phases are: A) Selection of genotyping method, definition of initial set of SNPs and variety selection; B) optimizing methodology per lab, genotyping selected samples; C) select markers based on performance, fit for purpose validation of the selected marker set, method validation of the applied genotyping methods based on the selected marker set.

Deliverables

- The final SNP set shall be suitable to support the management of the reference collection for DUS testing of tomato.
- The aim is to have a final set of 200-500 SNP markers, harmonized among lab partners.
- Fit for purpose validation of the SNP marker set performed by each participating laboratory.
- Validation of the genotyping method used by each of the participating labs.
- Validation of the SNP marker set, tested by different methods and in different laboratories.
- A final report describing the work that has been done within the framework of the project.
- Publication of the final SNP set and the validation results (annex 1). The final SNP set will be publicly available, free of access with the freedom to operate. The SNP set can be used by any authority responsible and or entrusted for tomato DUS testing. In addition, by any breeder to assess their candidate variety before its deposit for official registration for listing or to obtain Plant Breeders' Rights.

Varieties and sample selection

In order to test the pre-selection of 500 SNPs to their full extent, a well-founded representation of the common knowledge was deemed necessary. At the start of the project, each project partner provided a selection of Common Knowledge varieties, to allow the best possible selection to be used within the project. To aid and ensure the selections met the project's requirements, specific criteria were established by the partners.

These criteria, which guided the creation of the Developmental set, outlined several essential aspects:
- The inclusion of varieties that collectively represented a wide spectrum of genetic diversity, incorporating all types and characteristics.
- The incorporation of varieties that, while morphologically highly similar, possessed characteristics significant enough to be deemed distinct. These varieties potentially sparked debates during DUS tests, or even required an additional year of testing to definitively establish their distinctiveness.
- The inclusion of varieties derived from different germplasms.
- Exclusion of wild types to maintain the focus on cultivated varieties.

In addition, the Validation set required additional criteria to meet its specific objectives:
- Selection of samples that, when compared to each other, should not be distinguishable. This might include A) Different lots of the same variety B) DNA isolation replicates obtained from the same plant (biological replicates) D) DNA replications obtained from a single DNA isolation (technical replicates).

An overview of the varieties / samples selected per partner is provided in **Table 1.**

Table 1: Overview variety selection per partner.

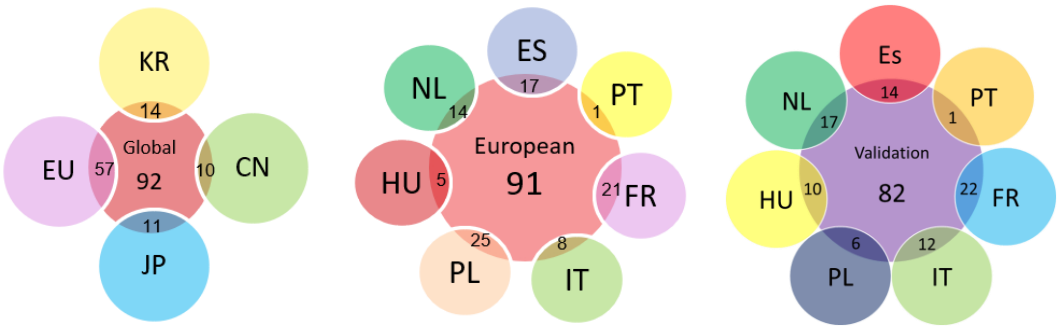
Project partner	Total number of varieties proposed as paper selection	Consent of the breeder needed and obtained	No consent needed	Seed samples received by Naktuinbouw	Usable DNA sample obtained from varieties	Used in Global Developmental set	Used in European Developmental set	Used in Validation set
Spain	42	42	0	42	42	0	17	14
Portugal	40	5	0	5	5	4	1	1
France	54	41	13	54	54	10	21	22
Italy	40	40	0	40	40	15	8	12
Poland	40	36	0	36	36	0	25	6
Hungary	40	31	0	31	31	15	5	10
Netherlands	157	128	0	128	128	13	14	17
Korea	15	0	15	15	14	14	0	0
China	10	10	0	10	10	10	0	0
Japan	15	11	0	11	11	11	0	0
total	***453***	***344***	***28***	***372***	***371***	***92***	***91***	***82***

After the initial variety selection process, 3 sub selections have been defined by the coordinator to be used within the genotyping experiments. The 3 defined selections were; Global Developmental set, European Developmental set and the Validation set. Throughout the project each partner performed two rounds of genotyping experiments. The first round consisted only of the Developmental sets. The second round consisted only of the Validation set. (The Validation set was not defined at the start of the project, this was done after the first round of genotyping.) The main purpose of the Developmental sets was to determine which SNPs would be suitable for genotyping in an international harmonized format. The Validation set was used to test the selected SNPs on their performance at each lab as well as testing each lab on their capability of producing consistent and repeatable genotype results.

During the first round of genotyping all European lab partners (CREA, INIA, GEVES and Naktuinbouw) analyzed two plates of samples (Global Developmental and European Developmental set), while the non-European lab partners (KSVS, CAAS and NARO) analyzed one plate of samples (Global Developmental set). There was no overlap between the 57 EU varieties included in the Global developmental set and the varieties in the European developmental set. This was a strategic choice to maximize the number of different varieties to be tested. In total, 183 different varieties have been genotyped during the first round of genotyping experiments.

During the second round of genotyping all the lab-partners performed the genotyping experiments on the Validation set. A visual overview of the 3 plates of DNA sent out by the coordinator is given in **Figure 1.**

Figure 1: Visual overview of the 3 plates of DNA sent out by the coordinator.
From left to right, an overview of the Global Developmental set, the European Developmental set and the Validation set. The larger circle in the middle represents the entire plate with the total number of varieties/samples included. The smaller circles surrounding the plates represent the number of varieties included per partner.



Lab partners and technique of choice:

As stated, the primary goal of the project was to establish an internationally harmonized SNP-set that could be used by all lab partners, regardless of their geographical location, and most importantly, regardless of the genotyping platform used.

Each partner was able to choose their own preferred genotyping workflow, from which 2 approaches stand out. Roughly half of the partners used either a single amplicon based platform (e.g. KASP) or a multiplexed amplicon approach and were either planning on performing the experiments within their own lab or outsource it to a trusted service provider. An overview per partner is provided in **Table 2**. In scope of the project, the diversity of workflows was deemed beneficial, so that the harmonization would not be restricted to a single technique and would find a broader harmonization and acceptance.

Table 2: Overview of techniques and genotyping platforms used per partner.

	Partner	Genotyping method	Reference	Service provider or own lab
1	Partner A	KASP	LGC	Own lab
2	Partner B	KASP	LGC	Own lab, with fluidigm juno system
3	Partner C	SeqSNP - Allegro Targeted Genotyping kit	Biosearch technologies	Biosearch technologies
4	Partner D	KASP	LGC	Own lab
5	Partner E	SeqSNP - Allegro Targeted Genotyping kit	NuGEN	NuGEN
6	Partner F	Agri-Seq	Thermofisher	Thermofisher
7	Partner G	GT-Seq	(Campbell et al. 2015)	Own lab

DNA extraction and shipment

As the main goal was to work towards a harmonized SNP marker set, any influence derived from the DNA isolation or biological influence was unfavorable. To minimize this influence, each partner had access to the exact same DNA samples. As there is a small chance of genetic variation in Tomato, all samples collected originated from a single plant per sample.

Seeds from all selected and received varieties were sown in the Naktuinbouw greenhouse under code. This included the varieties and samples present within the three defined plates as well as the remaining varieties which had not been selected. Single plants of approx. 15 cm in height were sampled (I.e. When the 3rd pair of leaves where in a developing state). Fresh leaf material was harvested in 50 ml Greiner tubes, frozen at -80°C overnight and subsequently freeze-dried. DNA was extracted using the QIAGEN DNeasy plant maxi kit (Cat. No. / ID: 68163). DNA concentration was determined on a Qubit fluorometer, and integrity Quality (e.g. degradation) was visually checked by agarose gel electrophoresis.

As by partner request, a suitably volume of DNA was provided by the coordinator. Each partner received roughly between 50 - 200µl of DNA, with a concentration of 20ng/µl.

Plates with DNA belonging to the EU and Global Developmental sets were shipped to each individual lab project partner in the period May-June 2021. The plates belonging to the Validation set were sent out in the period of May-June 2022, after the 2nd lab meeting.

Genotyping & data analysis

The first round of genotyping experiments was performed by all partners in the period from June 2021 till January 2022. The coordinator received the last dataset in January 2022. Until March 2022 all genotyping data from all partners was analyzed and compared. After the data analysis, the 2nd lab meeting was held. In this meeting the results were shared, and the decision was made to continue with a smaller SNP marker set based on SNP performance.

The genotyping experiments for the Validation set were performed mostly during the period of June 2022 until January 2023. The last experiments of the method validation were performed in May 2023. During the period of January 2023 until March 2023, all received data was analyzed. Afterwards the data was presented to the lab partners during the 3rd lab meeting in March 2023.

Data analysis on performance of all SNPs: Comparing genotype data for either Global Developmental set and European Developmental set of tomato varieties for all partners (Phase 2.1)

After all lab-partners finalized the genotyping experiments, their results were sent to the coordinator for data analysis, with 3 main objectives and research questions to be answered:

1. Which SNPs are successful in producing genotypes for the varieties and which SNPs are not?
2. Are the genotypes produced by successful SNPs consistent between the partners?
3. Can the varieties of the developmental sets be distinguished from each other?

As an input for the data analysis the coordinator received excel files from each of the 7 individual partners. All the excel files shared a similar layout, for each of the varieties a profile was reported with the SNP call per loci.

Due to the different genotyping platforms and software used, each excel file with data was distinguishable based on small differences in annotations. For example, an uncalled SNP was reported as “-2/-2”, “N/A” or simply as a blank value. Also, heterozygous scores were reported as e.g. “A/T” or “T/A”, which is in principle the same result, with a difference in DNA strand annotation. To account for the different annotations, all the SNP calls were transformed into a 5 categorial system. In this system, homozygous A (AA) equals to 1, TT equals to 2, GG equals to 3, CC equals to 4, any form of heterozygosity was translated to 5. Missing data was transformed into blank values, which get ignored by the software in downstream analysis. Due to the nature of the 5 categories, some data might get lost. As any form of heterozygosity was deemed “5”, differences in allele calls could remain unnoticed (e.g. “A/T” vs. “C/T”).
After uniformization of the SNP data all results were combined together, resulting into 2 datasets; the Global Developmental set with results of 7 partners, genotypes of 92 samples, based on 500 SNPs and; the European Developmental set with results of 4 partners, genotypes of 91 samples, based on 500 SNPs.

1. SNPs that are successful in revealing a genotype per sample for each of the lab partners

As a start of the data analysis, per lab partner, the number of successful genotyped SNPs was calculated. Herein a successful SNP was defined as "*a SNP being able to produce at least 1 genotype call within the tested varieties*". Performing this calculation for each partner on their respective tested plates (EU and Global Developmental) resulted in the numbers presented in **Table 3.** Per partner, per plate a visual aid is given in **Figure 2** and **Figure 3**.

As easily observable by **Figure 2** and **3**, some partners were more successful SNPs than others. Partners A and B were able to generate data for 11 & 19 percent of the SNPs, whilst partners C, D, E and F were able to generate data for 93 percent. Partner G subsided more in the middle with data for 76 percent of the SNPs. No effect has been observed between the European and Global Developmental set for the European partners.

Table 3: Number of SNPs (from the initial 500) that were (un)successful in genotyping.
Per partner, per sample set, the number and percentage of (un)unsuccessful SNPs are shown. * Only 65 samples of the 92 were tested

Set	Global Developmental set							European Developmental set			
Partner	A	B	C*	D	E	F	G	D	E	F	G
# SNPs genotyped	56	93	490*	467	466	480	382	467	466	481	386
% SNPs genotyped	11%	19%	98%*	93%	93%	96%	76%	93%	93%	96%	77%
# SNPs not genotyped	444	407	10	33	34	20	118	33	34	19	114
% SNPs not genotyped	89%	81%	2%	7%	7%	4%	24%	7%	7%	4%	23%

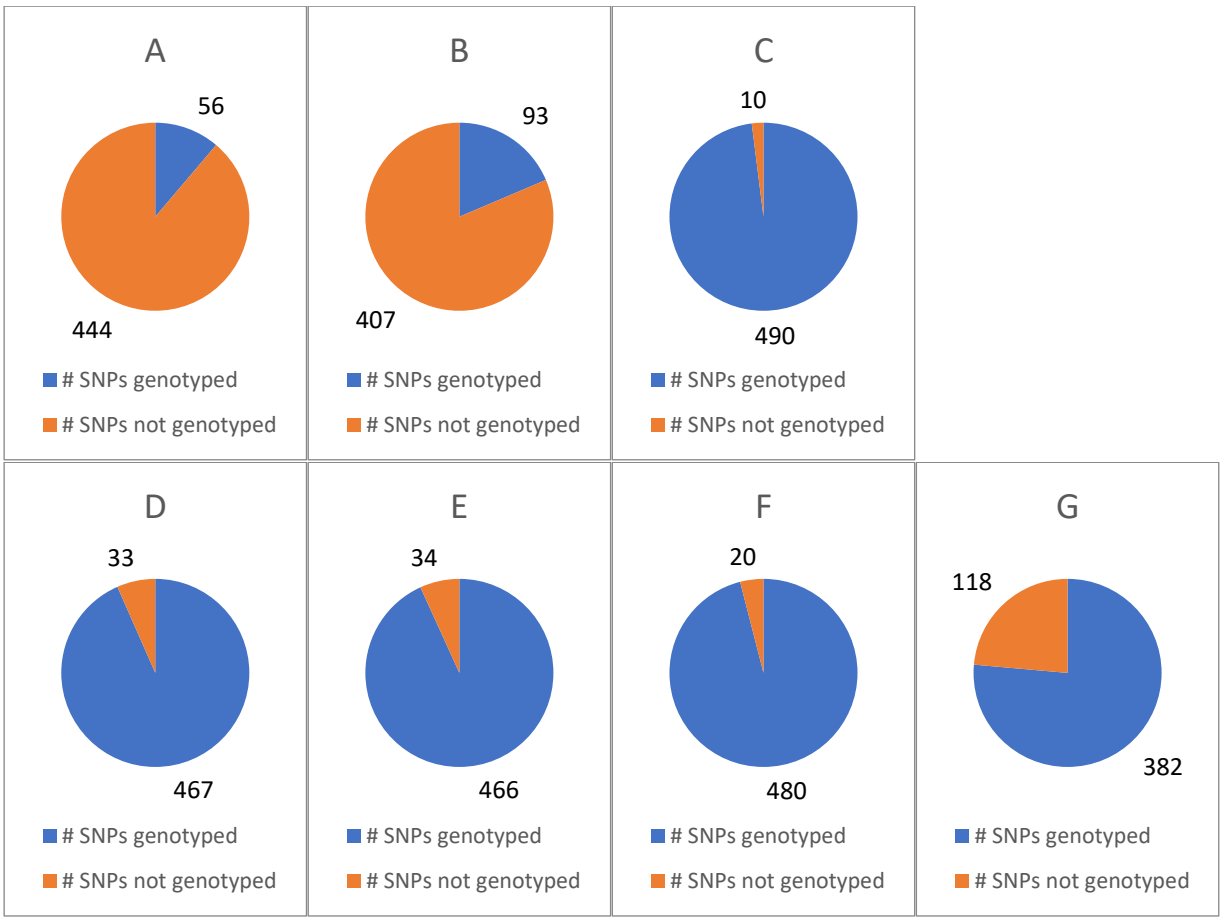


Figure 2: Number of successful SNP assays (blue) vs number of SNPs that did not reveal a successful genotype (orange) for the Global Developmental set for all lab partners.

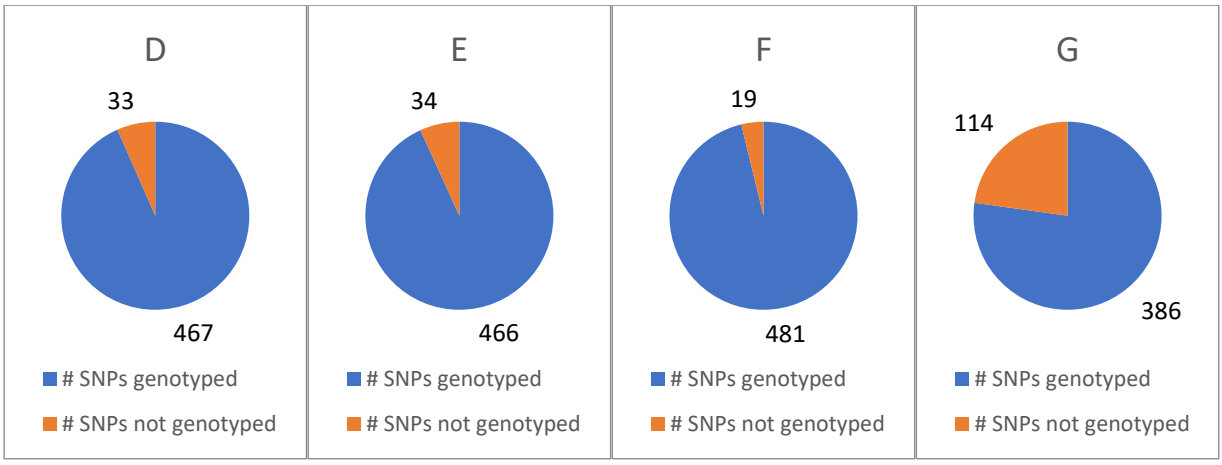


Figure 3: Number of successful SNP assays (blue) vs number of SNPs that did not reveal a successful genotype (orange) for the European Developmental set for all European lab partners.

The predefined definition of a successful SNP was not adequate to answer the question how many SNPs are successful at each partner. To give a better insight into the amount of data generated per partner, the number of SNP calls were counted per SNP, over all genotyped varieties. The results per partner, per plate have been visualized in **Figures 4** and **5**. For example, in **Figure 4**, partner D was able to steadily generate 466 SNP calls in 85 varieties, with a small decline to 434 SNP calls in 90 varieties. Opposed to partner G, where a steady decline can be observed from 350 SNPs in 10 varieties down to 235 SNP calls in 85 varieties.

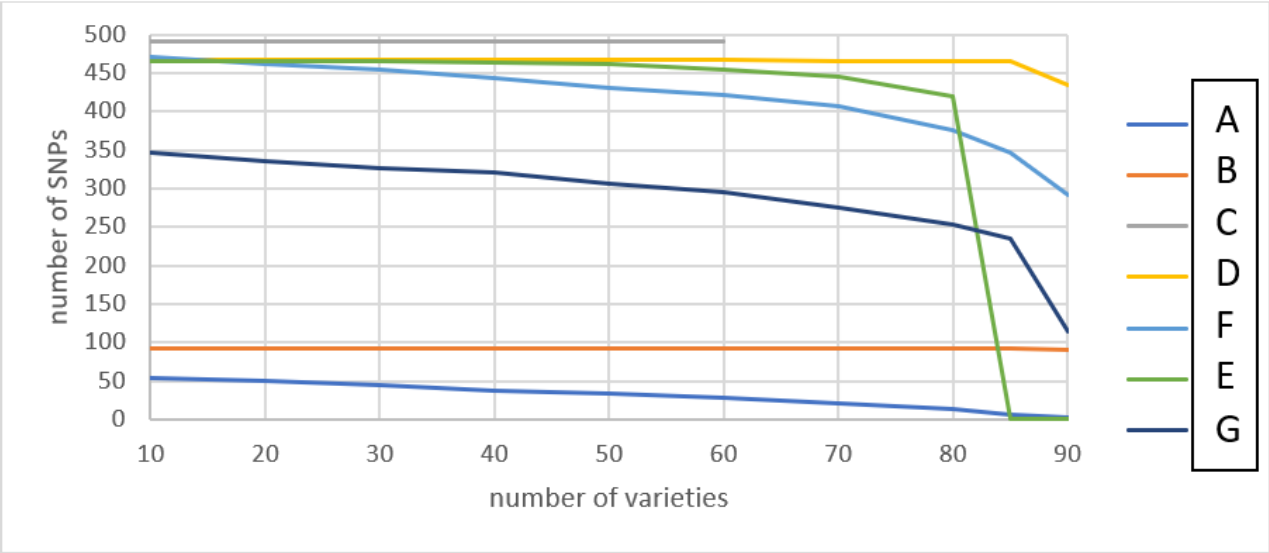


Figure 4: Number of successful SNPs per number of varieties from the Global Developmental set for each partner. On the vertical axis the number of SNPs is plotted against the number of varieties on the horizontal axis. Each line represents a partner.

Comparing the results of the Global and European Developmental set, the same trends can be observed per partner. As visualized in **Figure 5**, partners D, E and F were able to consistently produce more SNP calls than partner G.

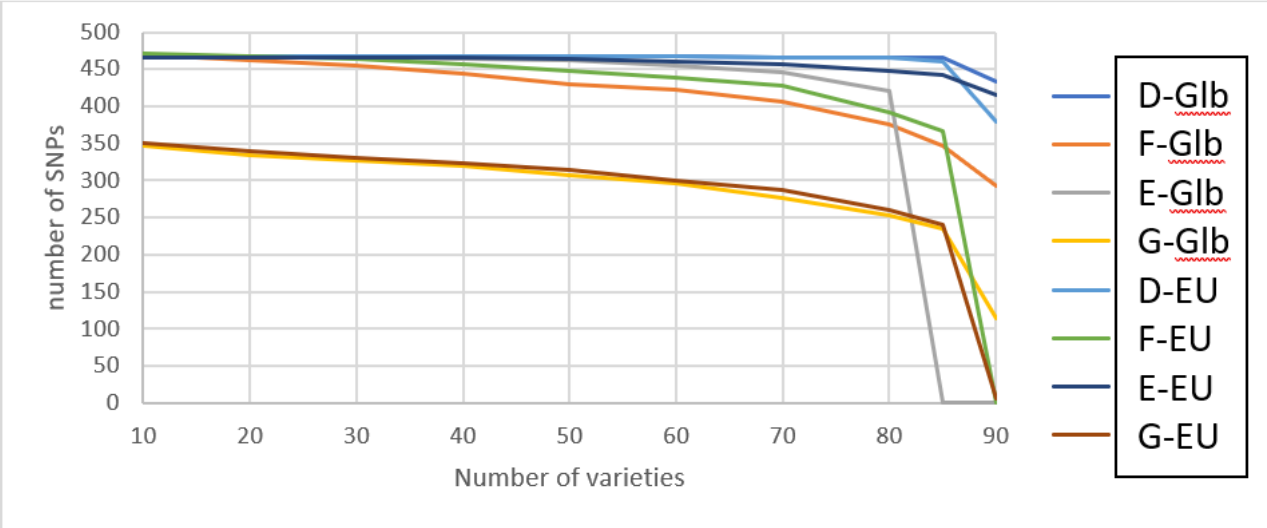


Figure 5: Comparison of successful SNPs per number of varieties between the Global and European Developmental sets for the EU lab partners. On the vertical axis the number of SNPs is plotted against the number of varieties on the horizontal axis. Each line represents a partner.

Based on the data presented in Figures 2 to 5 and Table 3, we were not able to draw a conclusion regarding the preferable technology or genotyping method.

2. Comparing genotype data for all partners to analyze consistency of the SNPs between the partners

In analyzing the consistency of the (successful) SNPs between the partners, we started again with a big input file containing the genotypes from all partners for all 500 SNPs for all 92 varieties of the Global set. See Table 4.

Table 4: Header of the input file with genotypes from all partners for all 500 SNPs for all 92 varieties.

	SNP1-A	SNP1-B	SNP1-C	SNP2-A	SNP2-B	SNP2-C	SNP3-A	SNP3-B	SNP3-C		SNP500
Var1	RR	RR	RA	RA	RA	RA	-	RA	RA		AA
Var2	RA	RA	RA	RA	RR	-	-	-	AA		RR
Var3	AA	AA	RA	-	RR	RR	-	-	RA		RA
Var92	RA	RA	RA	RR	RR	-	AA	AA	-		RA

To analyze the consistency of the genotypes produced by the 7 partners for SNP1, similarity values were calculated of each pair-wise combination. To determine the genetic similarity between the genotypes for a particular SNP produced by the partners, a similarity value was calculated by comparing the categorical values (= scored alleles for each SNP) of each SNP-partner combination compared to each other SNP-partner combination, and in the end all SNP-partner combinations to all other SNP-partners combinations. A similarity value is a number between 0 and 1 that indicates for the rate of genetic similarity between the two genotypes for SNP-partners combinations being compared. The similarities of all SNP-partner combinations were calculated based on Identity-by-State (IBS) method. How the similarity value is calculated is explained below.

Table 5: Different IBS value correspond to the number of alleles in common in the pair-wise comparison between samples.

SNP1-Partner A	SNP1-PartnerB	IBS
RR	RR	2 (both alleles in common)
RR	RA	1 (one allele in common)
RR	AA	0 (no allele in common)

Calculation of the Identity-by-State value

$$\frac{(\#\ \text{markers with IBS state 2}) + (0.5 * \#\ \text{markers with IBS state 1})}{\text{Number of non-missing markers.}}$$

To compare the SNP genotypes from all partners with each other, a matrix of (all SNPs x 7 partners) x (all SNPs x 7 partners) was prepared. Since every partner started with 500 SNPs, the theoretical number of values (pairwise comparisons) in the matrix was 3.500 x 3.500 = 12.250.000. In practice it was a little less, due to SNPs failing to genotype and exclusion of data.

The genotypes obtained by Partner A were mainly heterozygous for all varieties. These results deviated significantly from all other partners. For that reason, the genotypes of Partner A were not included in the comparison. Not for all varieties complete genotypes were obtained by each individual partner. The missing data caused significant gaps in the matrix. Leaving out the SNPs that did not produce a genotype, we still have data for 2.387 SNPs-partner combinations. So, the matrix of similarities for all pairwise comparisons contained 2.387 x 2.387 = 5.697.769 datapoints. Only a small part of this information was used to compare the SNP genotypes from all partners with each other. A snapshot of this matrix is shown in **Table 6**.

Table 6: A snapshot of the total similarity matrix for the pair-wise comparison of successful SNPs per partner for the Global Developmental set.

	SL3.0ch01_346524_Partner E	SL3.0ch01_346524_Partner D	SL3.0ch01_346524_Partner F	SL3.0ch01_346524_Partner C	SL3.0ch01_507890_Partner E	SL3.0ch01_507890_Partner D	SL3.0ch01_507890_Partner F	SL3.0ch01_507890_Partner C
SL3.0ch01_346524_Partner E	100	40,97	100	100	18,91	24,1	21,52	17,86
SL3.0ch01_346524_Partner D	40,97	100	42,94	36,93	28,66	31,53	30,11	23,08
SL3.0ch01_346524_Partner F	100	42,94	100	100	18,91	23,37	21,03	16,93
SL3.0ch01_346524_Partner C	100	36,93	100	100	20	24,62	22,58	16,93
SL3.0ch01_507890_Partner E	18,91	28,66	18,91	20	100	89,64	91,03	94,55
SL3.0ch01_507890_Partner D	24,1	31,53	23,37	24,62	89,64	100	100	90,77
SL3.0ch01_507890_Partner F	21,52	30,11	21,03	22,58	91,03	100	100	91,94
SL3.0ch01_507890_Partner C	17,86	23,08	16,93	16,93	94,55	90,77	91,94	100

Here we see the pairwise combinations for two SNPs (SL3.0ch01_346524 and SL3.0ch01_507890) which were successful for 4 partners (Partner E, D, F and C). The diagonal will always show a similarity of 100 (shown in yellow) as the SNP genotype for a particular partner is compared with itself. These pairwise comparisons are ignored when calculating the average similarity. The matrix is 'mirrored' with identical information above and below the diagonal. For the calculation of the average similarity for a SNP it is sufficient to only take the information below the diagonal into account.

In the example of **Table 6**, the average similarity for SNP SL3.0ch01_346524 is 70,14 and can be determined by calculating the average of the darker green values below the diagonal (40,97+100+100+42,94+36,93+100)/6=70,14).
The average similarity for SNP SL3.0ch01_507890 is 92,99. (89,64+91,03+94,55+100+90,77+91,94)/6=92,99.
We can observe that for SNP SL3.0ch01_346524 the average similarity is quite low because the similarity from one partner (Partner D) deviates from the other three partners in the pair-wise comparisons. When this contribution is ignored for this SNP, the average similarity would be 100. The average similarity of SNP SL3.0ch01_507890 is quite high at 92,99 on average. I addition, it is consistent between all the partners, with partner D and F producing identical data over all genotyped samples.

Given that not every partner was able to generate a complete SNP dataset, the next step was to combine both the Average Similarity with the number of partners who have contributed to that average, with results summarized in **Table 7.** The values displayed in **table 7** do not take into account the number of varieties for which a successful genotype was obtained. A SNP that was successfully genotyped by a partner for just one variety is treated the same way as a SNP that is successfully genotyped by a partner for all varieties (92 Global or 91 EU). Therefore Table 7 is solely informative on the general performance of the SNPs and less informative per partner.

Table 7: Number of SNPs successfully genotyped on Global set of varieties by N partners categorized by average similarity-ranges after pairwise comparisons.
36 SNPs (indicated green) were successfully genotyped by all 6 partners. In addition, the genotypes of all 6 partners were very consistent and the average similarity was 99 or higher. These 36 SNPs are the most robust, most consistent and best reproducible once. Subsequently these are the best candidate SNPs to be selected for the final SNP set.
176 SNPs (indicated blue) were successfully genotyped by 5 partners. In addition, the genotypes of all 6 partners were very consistent and the average similarity was 95 or higher. Whilst less consistent and robust than the 36 SNPs indicated in green, these SNPs would also be deemed as good candidate SNPs for the Final SNP Set.

	#SNPs genotyped by N partners				
average similarity range	N = 6	N = 5	N = 4	N = 3	N = 2
>99	36	90	22	5	1
>95	55	176	71	11	3
>90	60	217	85	13	3
>80	72	266	100	16	5
>70	72	277	110	22	6
>60	72	278	111	23	6
<60	*0*	*1*	*0*	*0*	*3*

All data shown above (table 7) is specific for the Global Developmental set of varieties. The same analysis was performed for the European Developmental set of varieties (table 8). This analysis only included the 4 project partners from Europe.

Table 8: Number of SNPs successfully genotyped on EU set varieties by N partners categorized by average similarity-ranges after pair-wise comparisons.

	#SNPs genotyped by N partners		
average similarity range	N = 4	N = 3	N = 2
>99	169	57	15
>95	240	86	21
>90	283	95	24
>80	331	108	26
>70	346	111	26
>60	347	116	27
<60	*1*	*1*	*0*

In preparation for the selection of the best performing SNPs, the average similarity (x-axis) was plotted against the number of successfully genotyped SNPs (y-axis). The information from **Table 7** was plotted to represent the number of partners that were successful in obtaining good genotypes (**Figure 6**). We can observe that the largest increase in number of SNPs with successful genotypes is from 6 to 5 partners. Only a small number of SNPs is gained when we would select the number of SNPs that are genotyped by at least 3 partners instead of at least 4 partners. From an average similarity of 95 and higher, a significant decrease in number of SNPs with good genotypes is observed.

The same plots have been generated for the European set (**Figure 7**). These plots looked very similar as the plots presented for the Global set displayed in **Figure 6**. This indicates that the SNP performance is not dependent on the set of varieties on which they are applied.

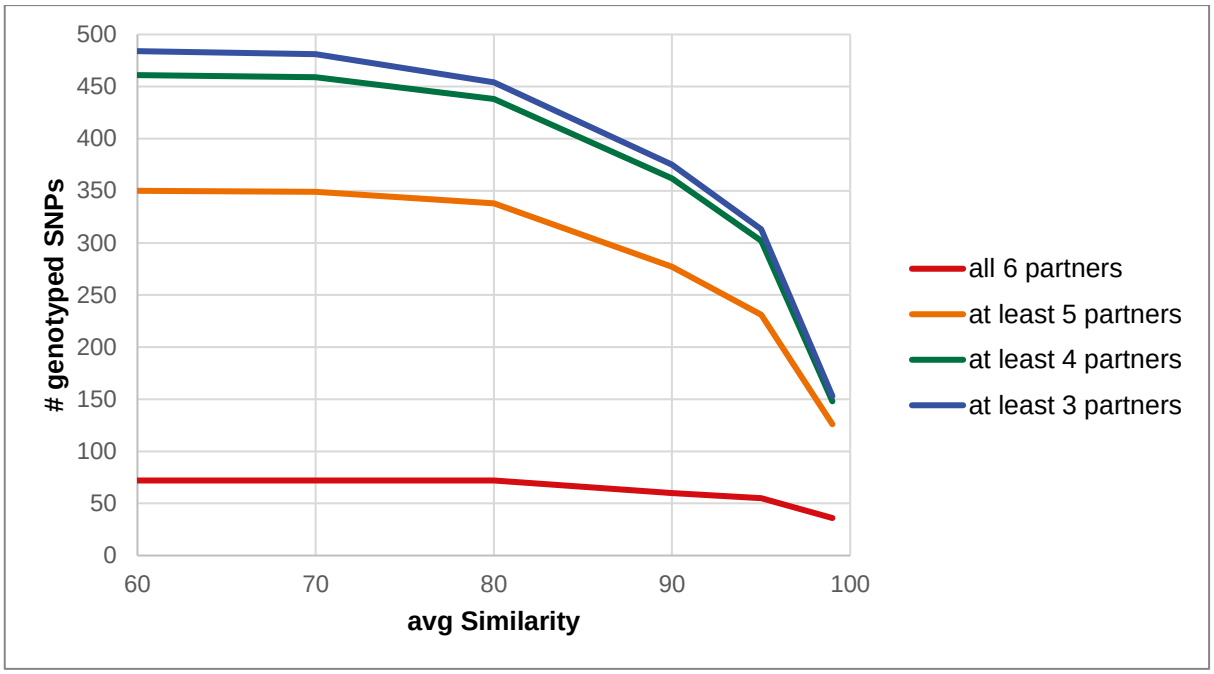


Figure 6: The relation between the average similarity (as an expression for consistency of a SNP) and the number of SNPs that were successful in genotyping in Global set given for 1) all 6 partners; 2) at least 5 partners; 3) at least 4 partners; and 4) at least 3 partners. At least 5 partners meaning 5 or 6 partners (cumulative).

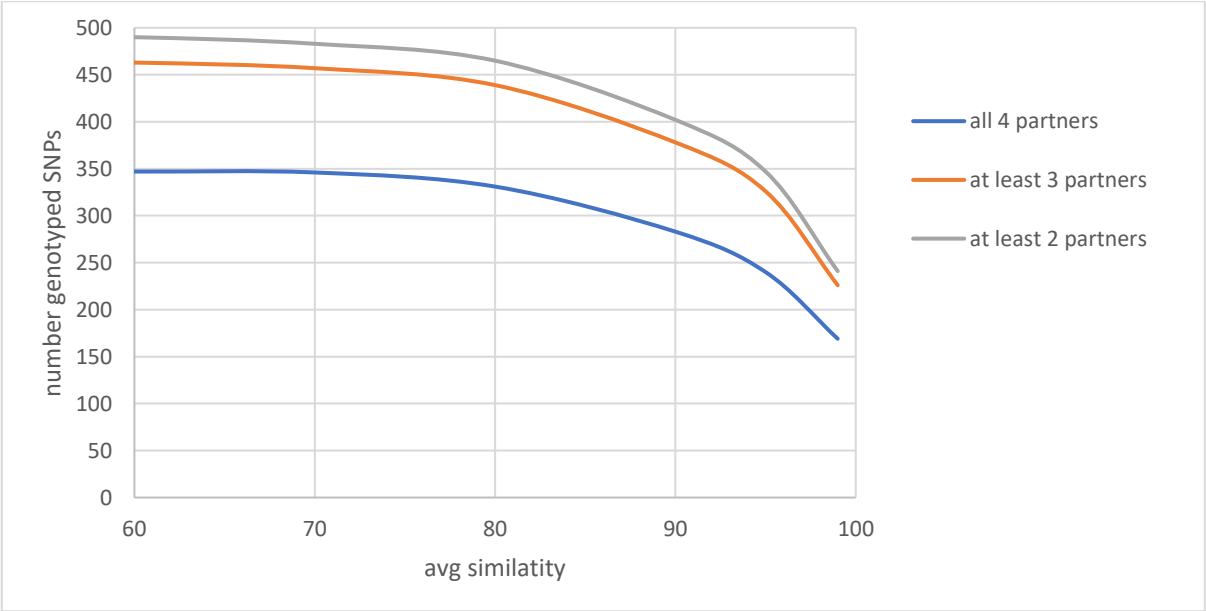


Figure 7: The relation between the average similarity (as an expression for consistency of a SNP) and the number of SNPs that were successful in genotyping in EU set given for 1) all 4 EU partners; 2) at least 3 EU partners; and 3) at least 3 EU partners. At least 3 partners means 3 or 4 partners (cumulative).

3. Comparing genotype data for all partners to analyze consistency of the varieties between the partners

The varieties in the Global and European developmental sets were selected to be representative for the diversity in the species. They were also selected to be distinct varieties based on DUS characteristics. It is our expectation that these varieties are also distinct based on their SNP genotypes. In addition, we expect the 7 partners to produce consistent genotypes for the varieties.

After genotyping the provided 183 samples by all partners (i.e., 92 varieties for the Global set and 91 varieties for the European set), raw genotyping data was obtained as Excel files. The genotyping data was imported in the BioNumerics software (version 8.1) for genotyping analysis. Generated data was converted to categorical values (Homozygous reference (RR) **0**, Homozygous Alternative (AA) **2** and Heterozygous (RA) **1,** resulting in a scoring table, which was used for the statistical analysis described below.

Table 9: Snapshot as an example for the input file with genotypes from all partners for all 500 SNPs for all 92 varieties.

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7		SNP500
Var1-partnerA	RR	-	RA	AA	RA	-	RA		RR
Var1-partner B	RA	RA	RA	AA	-	-	-		RR
Var1-partner C	RR	RA	-	AA	RR	-	-		RA
Var2-partner A	-	-	AA	RR		-	AA		AA
Var2-partner B	RA	RR	AA	RA	-	-	-		RA
Var2-partner C	RA	RR	AA	-	RA	-	AA		RA
Var92	RR	AA	RA	RA	-	AA	RR		RA

To determine the genetic similarity between varieties, a similarity value was calculated by comparing the categorical values (= scored alleles for each SNP) of each genotyped sample to each genotyped sample. This calculation was based on identity-by-state (IBS) explained earlier.

To visualize the (genetic) relationship between the samples, a clustering (resulting in a dendrogram) was generated using UPGMA parameters (Unweighted Pair-Group Method, Arithmetic average). This kind of algorithms find successive clusters using previously established clusters. Two steps are performed repeatedly: 1. find and merge the two best matches (I.e., best matches being the two variety-partner combinations with the highest similarity) and 2. update the similarity matrix by averaging the scores. The resulting dendrogram is a visual representation of the genetic similarity.

To support the robustness of our strategy and quality of the analysis, data gathered from all different partners clustered together as expected. Although we can observe 'technical noise' when comparing the genotypes of all partners for the same variety, this noise is smaller than the diversity observed between different varieties. The majority of varieties could be distinguished from each other based on the 500 SNPs used. There were only a few exceptions listed in **Table 10**.
In Annex 2 the full dendrograms (based on the 500 SNPs and selection of SNPs) are provided for; the Global set, the European set and both sets combined.

Observations in the cluster analysis

- For the Global Developmental set (92 varieties) we can observe 2 sets of 2 varieties that were genetically not distinct. (see **Table 10** and **Figure 8a**)
- For the Global set, 88 varieties could be genetically distinguished. The maximum similarity between two distinct varieties is 89,8.
- 12 varieties (=6 variety-pairs) could be identified as 'close match'. The varieties were clearly distinct. The genetic diversity within the variety-cluster due to technical errors and different genotype calls between the project partners for the same variety were smaller than between the genetic diversity between the variety-clusters. The similarity was >85,0 and <89,8. (See **Figure 8a**)
- For the European Developmental set (91 varieties) we could also observe 2 sets of 2 varieties that were genetically not distinct. (see **Table 10** and **Figure 9a**)
- For the European set, 87 varieties could be genetically distinguished. The maximum similarity between two distinct varieties is 91,3.

- 9 varieties could be identified as 'close match'. That included 3 variety-pairs and 1x three closely related varieties in 1 cluster. The similarity was >85,0 and <91,3.

Table 10: Five variety-pairs were identified with as matches after genotyping with all SNPs and all partners.

Sample number of matching varieties	Contributing project partner	Information on company or description	Set
3106	Poland	Different companies, distinct varieties on morphology	European
3135			
3183	France	Indicated as close to each other	European
3170			
1703	Korea	Indicated as close to each other	Global
1705			
2473	Hungary	Same variety with different names	Global
1715	Japan		
3191	France	Same variety with different names	European
1719	Japan		Global

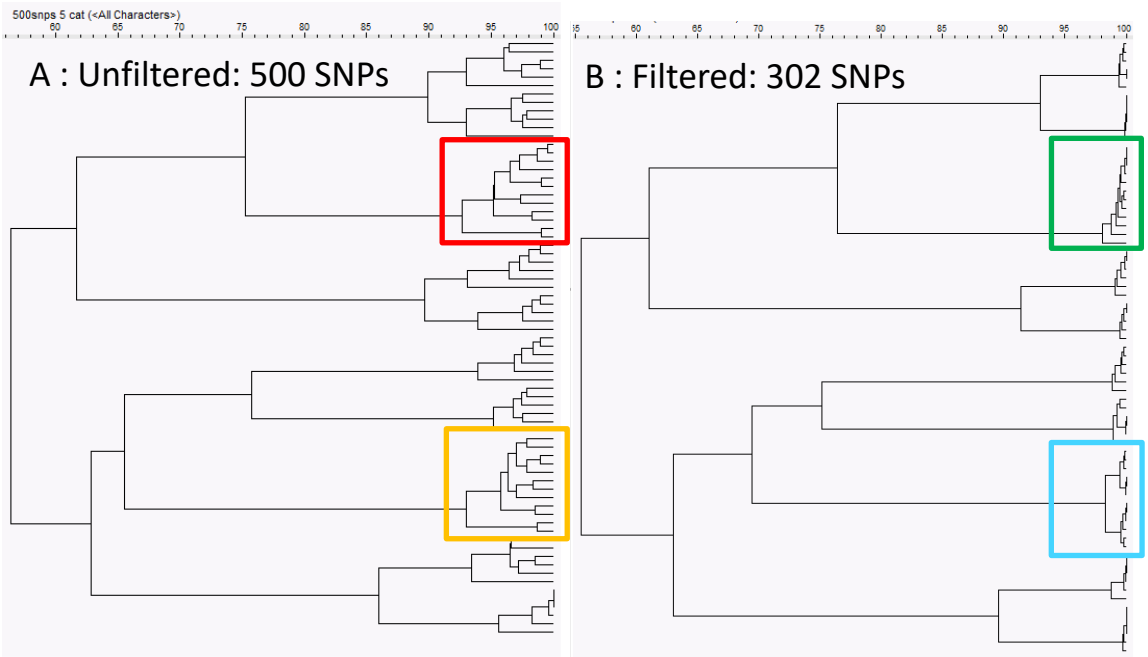


Figure 8a and b: The effect of filtering the SNPs on 'performance' on the Global set.
The effect of filtering on 'performance' is shown by comparing the clustering of some GLB similar variety-pairs based on all 500 SNPs (left=A) and the clustering of some GLB similar variety-pairs based on the proposed 302 best performing SNPs (right=B). The variety-pair indicated by the red and orange box showed a match in the unfiltered SNP set. When performing a cluster analyses based on the proposed 302 best performing SNPs, the varieties in the green box remained a match whilst the varieties in the blue box became distinguishable.

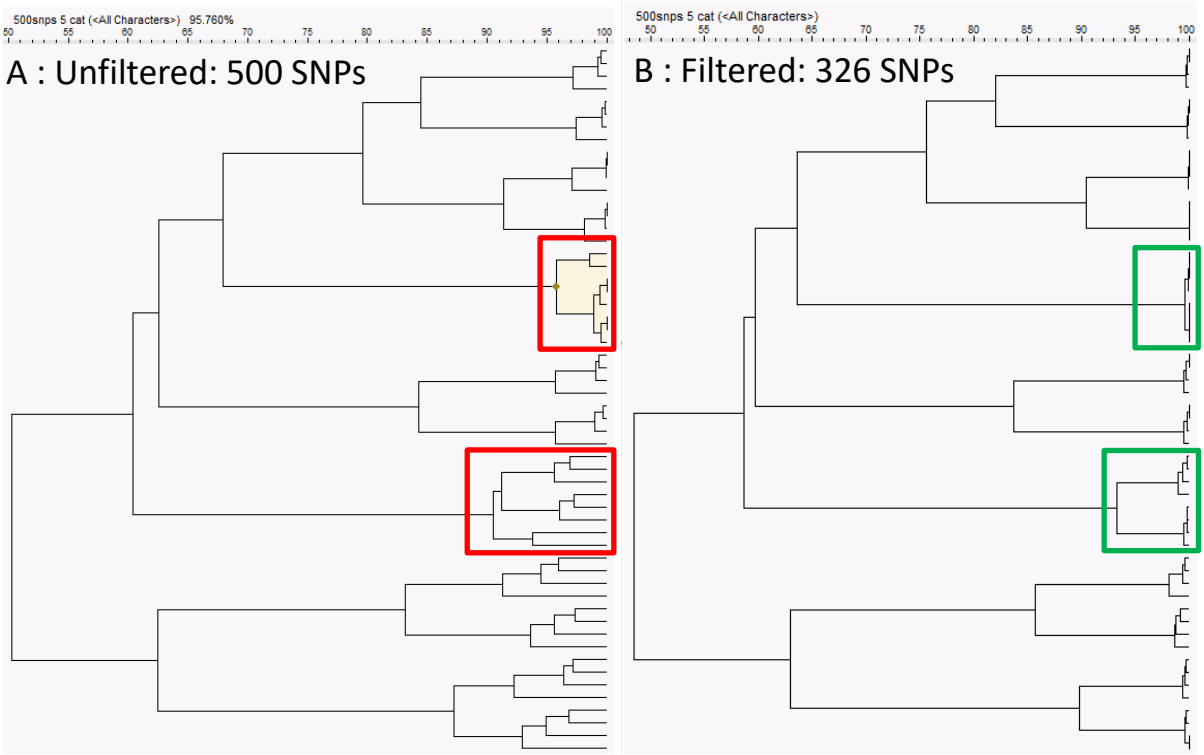


Figure 9a and b: The effect of filtering the SNPs on 'performance' on the European set.
The effect of filtering on 'performance' is shown by comparing the clustering of some EU similar variety-pairs based on all 500 SNPs (left) and the clustering of some EU similar variety-pairs based on the proposed 326 best performing SNPs (right). The variety-pairs indicated by the red boxes showed a 100% match in the unfiltered SNP set and were distinct (but very closely related) in the filtered dataset.

Table 11: Previously identified five variety-pairs with 100% match after genotyping with all SNPs and all partners (**Table 10**), re-analyzed in cluster analysis after genotyping with the filtered set of SNPs and all partners. For EU set 326 were used, for the GLB set 302 SNPs were used.

Sample number	Contributing project partner	Information on company or description	set	Different conclusion?
3106	Poland	Different companies, distinct varieties on morphology	European	No longer 100% match (99,5) 2 consistent SNPs difference between these varieties
3135				
3183	France	Indicated as close to each other	European	93,6. Clearly distinct genotypes
3170				
1703	Korea	Indicated as close to each other	Global	97,96%. But distinct clusters
1705				
2473	Hungary	Same variety with different names	Global	Still 100% match
1715	Japan			
3191	France	Same variety with different names	European	Still 100% match
1719	Japan		Global	

To decide on which best performing SNPs we will select for the final set, the lab partners reached agreement on the following questions.

- How robust/reproducible/consistent should the genotypes be? Or, what is the minimum threshold for average similarity we accept?
- What is the minimum number of partners that must be able to produce a successful genotype for a particular SNP?
- Is the selected SNP set able to discriminate between morphologically distinct varieties?
- To a lesser extent, what is the discriminative power of the selected SNPs?

In view of an International SNP database that will be supplied by different partners, it is most important that the SNPs will reveal consistent genotypes. So, consistency of genotypes is key. As the minimum similarity between distinct varieties in the Global set was 89,8 in the and European Developmental set was 91,3 the lab partners collectively agreed to set the minimum threshold for average similarity for a SNP at >95.

After the genotyping of the Developmental sets, genotyping data from 6 partners was available. Most partners had only performed one experiment. In scope of the project each lab had the opportunity for optimization of the SNP assays. Therefore, we chose not to be too strict in the minimum number of partners that were able to produce a successful genotype for a particular SNP. Collectively the partners decided to set the minimum number of partners at 4 out of 6 for the GLB set and 3 out of 4 for the EU set.

For the Global set, when considering both proposals of minimum threshold for average similarity at >95 and minimum number of partners at 4, the number of SNPs in that fulfill these criteria is 302. This is indicated within **Figure 6.**

For the European set, when considering both proposals of minimum threshold for average similarity at >95 and minimum number of partners at 3, the number of SNPs in that fulfill these criteria is 326. This is indicated within **Figure 7.**

Comparing the SNP's selected with the previously mentioned criteria, an overlap of 297 SNPs could be observed. As all partners agreed that a single defined SNP set is preferable over 2 slightly different sets, the choice was made to continue with the 297 SNPs for the remainder of the project (visualized in **Figure 10**).

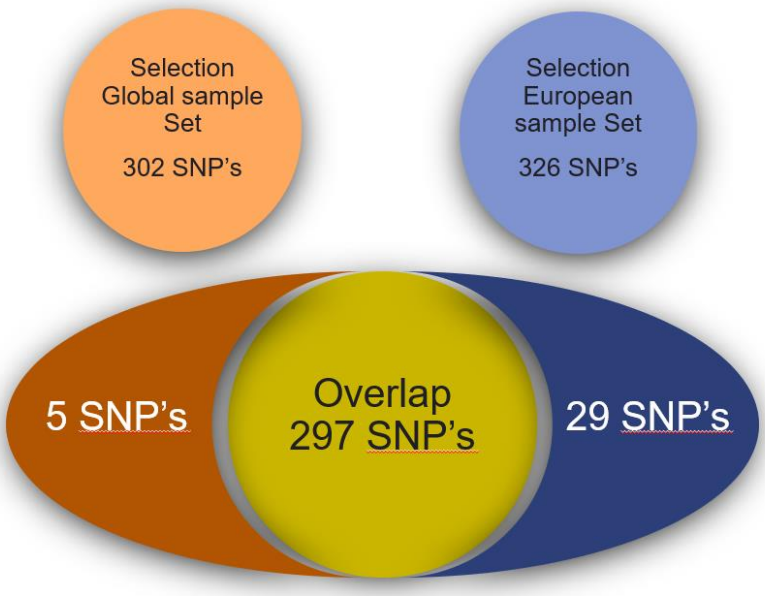


Figure 10: The overlap between the two sets of proposed selected SNPs.

Calculation of PIC-values for the SNPs as an expression of discriminative power

In the original research plan, one of the selection criteria is discriminative power of the SNPs expressed by PIC (Polymorphism Information Content) value. This calculation can only be performed on one genotype for a variety. The genotypes of the project partners for a particular variety was not 100% consistent. Therefore it was decided to calculate the PIC value based on a consensus genotype. As shown in the figure below (**Figure 11**), a high PIC value could be observed for most SNPs, which were included in de boxplot (N= 269 >0,55). Also, a small number of SNPs (N=28) yielded a lower PIC value as expected, with 2 SNPs only having a PIC value of <0,1. Based on the SNP scores of the 183 varieties included in the calculation 28 SNPs were considered as outliers. No efforts were made to remove these SNPS from the harmonized set, based on their high discriminating power during the selection process (pre-project).

Polymorphism Information Content

A widely used measure of the usefulness of a molecular marker

$$PIC_j = 1 - \sum_{i=1}^{n} p_i^2$$

i = the ith allele of the jth marker
n = the number of alleles at the jth marker
p = allele frequency

Allele frequencies	Formula	PIC
Biallelic marker		
$p_1 = 0.5, p_2 = 0.5$	$1 - (0.5^2 + 0.5^2)$	0.50
$p_1 = 0.4, p_2 = 0.6$	$1 - (0.4^2 + 0.6^2)$	0.48
$p_1 = 0.3, p_2 = 0.7$	$1 - (0.3^2 + 0.7^2)$	0.42
$p_1 = 0.2, p_2 = 0.8$	$1 - (0.2^2 + 0.8^2)$	0.32
$p_1 = 0.1, p_2 = 0.9$	$1 - (0.1^2 + 0.9^2)$	0.18
Multiallelic marker		
$p_1 = 0.33, p_2 = 0.33, p_3 = 0.33$	$1 - (0.33^2 + 0.33^2 + 0.33^2)$	0.67
$p_1 = 0.4, p_2 = 0.3, p_3 = 0.3$	$1 - (0.4^2 + 0.3^2 + 0.3^2)$	0.66
$p_1 = 0.7, p_2 = 0.2, p_3 = 0.1$	$1 - (0.7^2 + 0.2^2 + 0.1^2)$	0.46

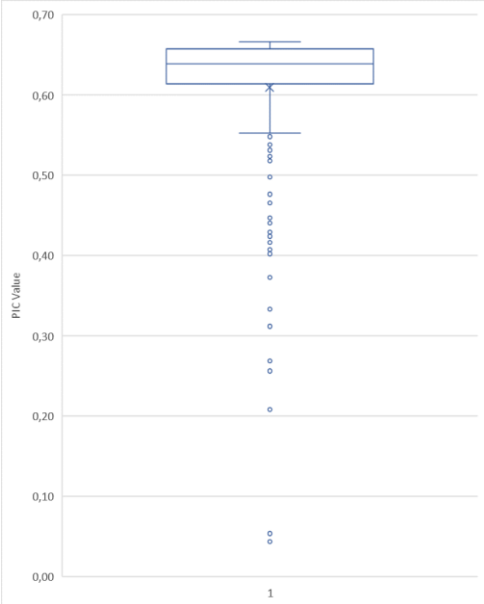


Figure 11: Box plot indication the Polymorphism Information Content for each SNP (individual dots). Based on the consensus genotype of all partners.

SNP validation

Moving on to the additional SNPs validation, a new plate of DNA samples was sent to each lab partner. Each plate contained 92 predetermined samples (i.e., technical and biological replicates, blind samples and 3 empty positions to analyze own extracted DNA), to be analyzed by each partner using the 297 agreed SNPs.

Generated SNP calls

After the genotyping was performed by each partner the results were sent back to the coordinator using the same format as in Phase 2.1.

When looking at the number of successful SNP calls generated by each partner (shown in **Figure 12**), it could be observed that Partners A, C, D, E and F were most successful. With more than 290 SNPs successfully genotyped in at least a single variety. For partners B and G only 238 and 243 SNPs successfully genotyped in at least a single variety. In comparison to the Developmental sets, Partner A showed the biggest improvement coming up from 11% successful SNPs to 98% successful genotyped SNPs. This was followed by partner B, who improved from a 19% success rate to 80% success rate. The other partners remained stable in performance.

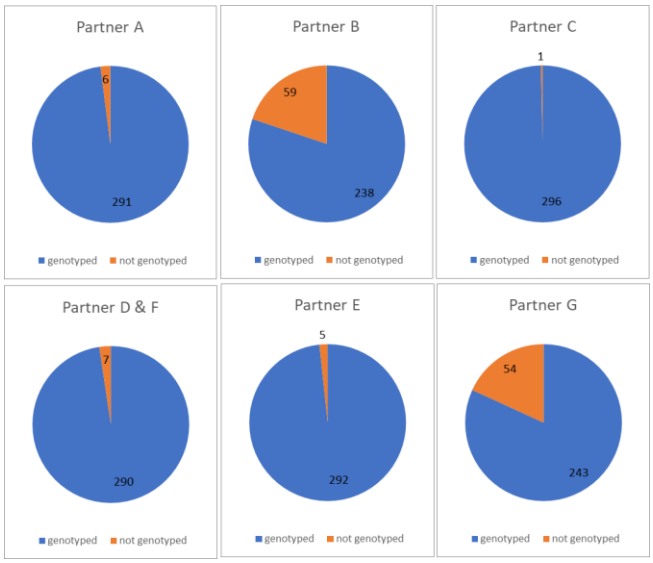


Figure 12: The number of genotyped SNPs per partner, where a SNP call was produced in at least 1 of the 92 samples.

For each partner (except partner B) it could be observed that the number of SNP calls dropped when the number of varieties was raised. Meaning that not all of the 297 SNPs were able to be called for all varieties. Noteworthy was that the number of successful SNP calls remained relatively stable for partners A, C, D and E up to the point of "number of varieties" n=85. Opposed to the steady decline in SNP calls of partners F and G, as displayed in **Figure 13**.

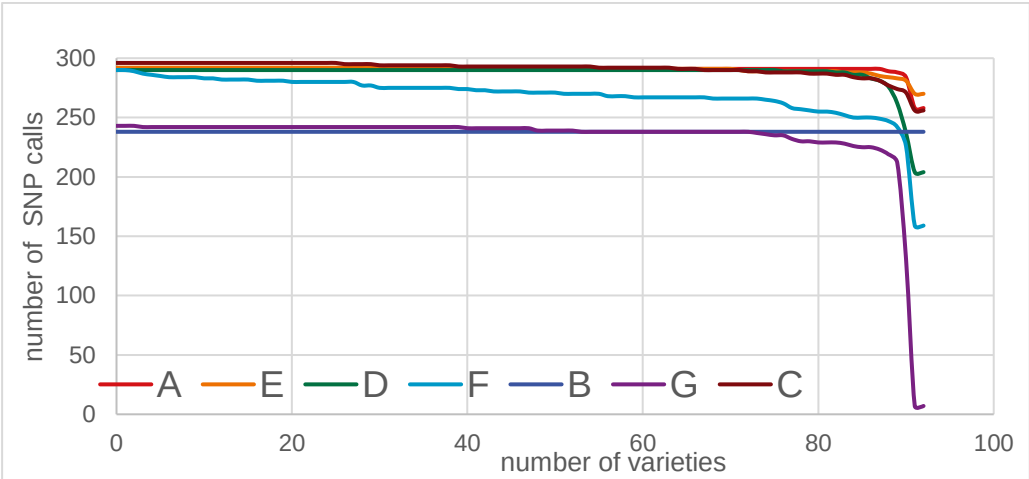


Figure 13: Number of SNPs called in relation to the number of varieties in which a SNP has been called.

Consistency of genotypes per SNP

As decided after Phase 2.1, only SNPs yielding a high consistency (avg. similarity greater than 95,0) are included in the final harmonized SNP set. To confirm whether or not the selected SNPs remained stable in their performance, the average similarity was calculated as previously described. In **Figure 14**: consistency of SNP calls, the average similarity per SNP/per partner is visualized in relation to all other partners. The blue continuous line is the average of all partners and is included as a benchmark.

In **Figure 14** it can be observed that the first 150 SNPs are very consistent with a high average similarity and very low deviation (avg similarity over all partners = 99,7). From SNP 151 until SNP 236 the average similarity is starting to decline from 99,7 down to 95,2. The decline in average similarity was expected as this was also observed within phase 2.1. Even with the average similarity declining, the performance of the first 236 SNPs upheld the original criteria of average similarity greater than 95,0. Against expectations the average similarity of SNPs 237 until 297 declined to 71,1. Their lower similarities eventually will produce more and unwanted variance within a harmonized database.

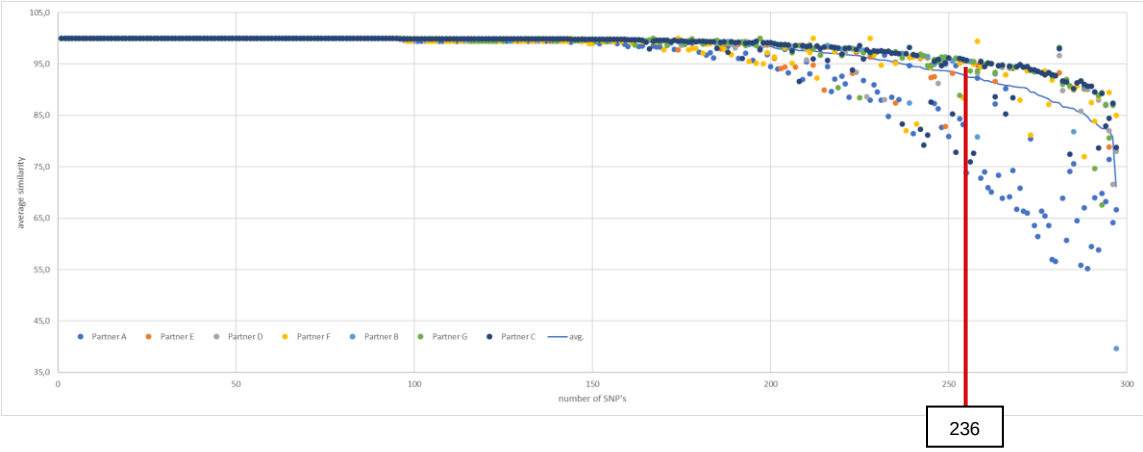


Figure 14: Consistency of SNP calls, per SNP/per partner, in relation to all other partners.
On horizontal axis, the 297 SNPs are plotted against their respective similarities (both per partner and the average over all partners). The SNPs have been sorted based on the average similarity over all partners.

Consistency of genotypes per sample

When comparing the generated genotypes per sample of each partner against all other partners the average similarity can be calculated. This is displayed in **Figure 15**. It can be observed that for the first 89 (out of 92) samples the average similarity is high (average similarity > 96,5). For the remaining 3 samples the similarity declines to 91,7. The sudden decline in average similarity might be an indication of low quality DNA samples. Furthermore it can be observed that Partner A scores consistently below the average when comparing the genotypes against the other partners.

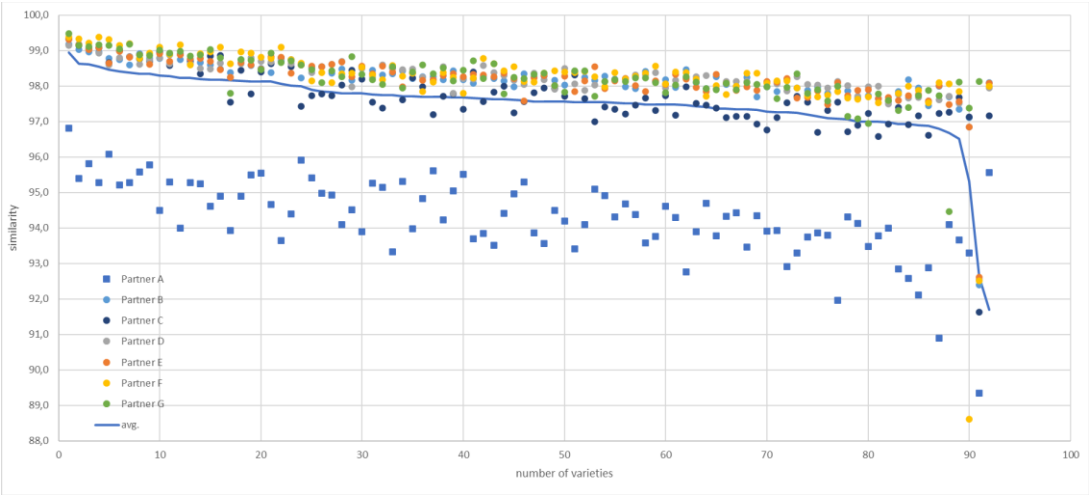


Figure 15: Consistency of genotypes, per sample/per partner, in relation to all other partners.
On the horizontal axis the number of varieties is plotted against the average similarities of the generated genotypes (both per partner and the average over all partners). The varieties have been sorted based on the average similarity over all partners.

Number of SNP calls combined with consistency

In all the previous data analyses, a successful SNP has been defined as; genotyped in at least 1 sample, with an average similarity > 95%. To give additional insights in the performance per partner the average similarity of the genotypes (**Figure 15**) has been plotted against the number of SNP calls that have contributed to that value.

Observable in **Figure 16** is the distribution of individual SNPs per partner, visualizing the relation between de average similarity and the number of SNP calls contributing to that similarity. In the bottom right corner a visual aid regarding the interpretation is given. Remarkable is the difference between the plots of partner A and B towards partner F and G. Where partner A has a high number of SNP calls (achieving very high efficiency) the similarity in relation to the other partners is relatively lower. Where partner F has SNPs who have a lower efficiency, not genotyped in all samples. The similarity of these SNPs is relatively high and comparable to the SNPs with a high number of genotype calls.

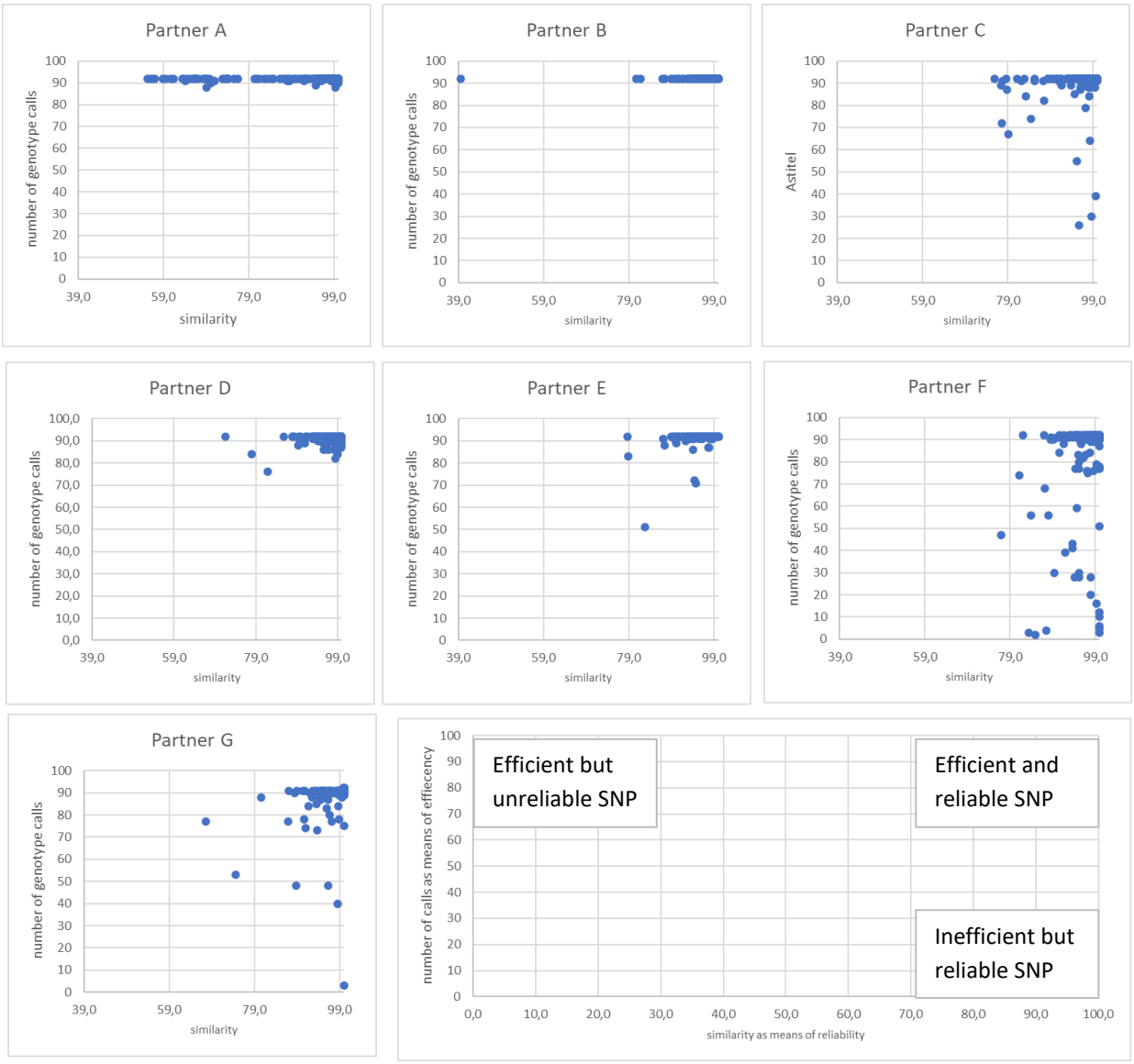


Figure 16: Distribution of individual SNPs visualizing the similarity in relation to all other partners, plotted against the number of SNP calls contributing to the similarity.

Technical replicates

Included in the set of 92 samples, five (n=10 in duplo) technical replicates were included. These should be indistinguishable from each other since the DNA is originating from the same source. The replicates included 2 samples where the isolated DNA was diluted independently and included twice. For 3 samples the DNA isolation was performed in duplo from the same plant. In **Figure 17** the clustering is visualized of all replicates of all partners (left). In addition the clustering per partner is visualized.

As visualized in **Figure 17**, all partners were able to achieve a high similarity regarding the replicates (>97,7) within their respective labs. When comparing the results of all partners the lowest similarity observed between the partners was 93,7. In the dendrogram on the left, especially the cluster of TMT 21 1145, it is observable that for all samples the similarity within a lab is higher than between labs.

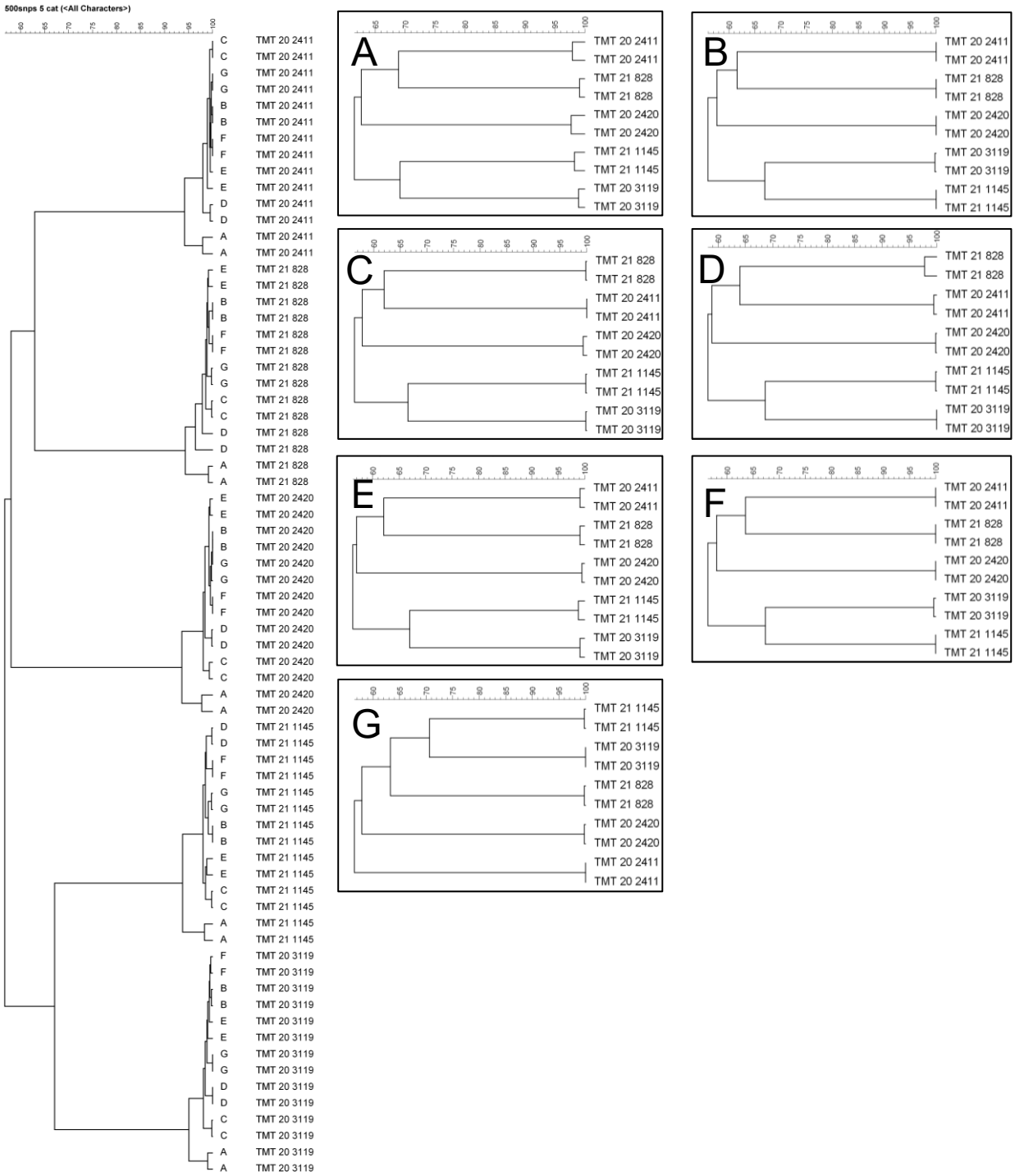


Figure 17: Clustering of technical replicates.

Biological replicates
Included in the set of 92 samples, a multitude of biological replicates were included. These replicates consisted of different plants from the same seed lot (n=5, in duplo n=10), a different seed lot from the same variety (n=5, in duplo N=10)**,** the same variety from a different partner (origin) (n=2, in duplo n=4).

The clustering of the different plants from the same seed lot replicates is visualized in F**igure 18**.
On the left, the clustering is visualized of all replicates of all partners. On the right the clustering per partner is visualized. All partners were able to achieve a high similarity when comparing the biological replicates. The range of similarity was observed to be 95,9 to 100,00. One deviation has been observed for partner G, where the similarity between the 2 individuals of TMT 20 3128 was found to be 55,7. Explanation for this deviating profile is that only 7 of the 297 SNPs were genotyped.

When comparing the results of all partners, the observed range per replicate was 93,4 (for TMT 20 3128) up to 94,0 (for TMT 21 1181 and TMT 21 1168), with partner A being the most deviating on all samples. (The range of similarity between partners without partner A is 96,6 up to 98,0.) In contrary to the technical replicates where the similarity within a lab is higher than between labs, this has not been found the case for the biological replicates. This is mainly observable for partners B,E, F & G within samples TMT 20 3128 and TMT 21 1133.

The clustering of the different plants from the same variety originating from different seed lots is visualized in F**igure 19**. On the left, the clustering is visualized of all replicates of all partners. In addition the clustering per partner is visualized. Looking at the results per partner, all partners were able to generate similar results for the varieties included. The range of similarities varied from 96,7 to 100,0. Based on these results it can be concluded that the different seed lots represent the included varieties well.

Comparing the results of all partners, the observed range of similarities was 93,6 up to 95,4. Which is comparable with the results obtained for the other biological replicates. Again the profiles of partner A has been found the most deviant in relation to the other partners. (Excluding partner A, the range of similarities between partners is 97,4 up to 98,2).

Noteworthy is difference in clustering that can be observed in the dendrogram of **Figure 19**. For example, the cluster of TMT 20 3156 and TMT 20 3157, shows that each partner has generated profiles more comparable to each other than another partner. In contrast, partners B, C, E, F & G were able to generate almost identical profiles (similarity = 99,7) for samples TMT 20 3190 and TMT 21 1195. Another example is the cluster TMT 21 1118 and TMT 21 1119, where every partner was able to discriminate the 2 samples from one another. Given the fact that this is consistent at all partners, an acceptable explanation is that the 2 seed lots are the same variety with a slight drift in the genetic structure. Whilst samples 1167 & 1168 are a different variety than samples 1180 & 1181, none of the partners was able to clearly distinguish the different varieties. Previous research has shown that based on DNA these 2 varieties are very closely related to each other. In addition, based on morphology these 2 varieties have proven to be very similar, but distinct as well.

Figure 18: The clustering of 2 plants from the same seed lot as biological replicates is visualized.

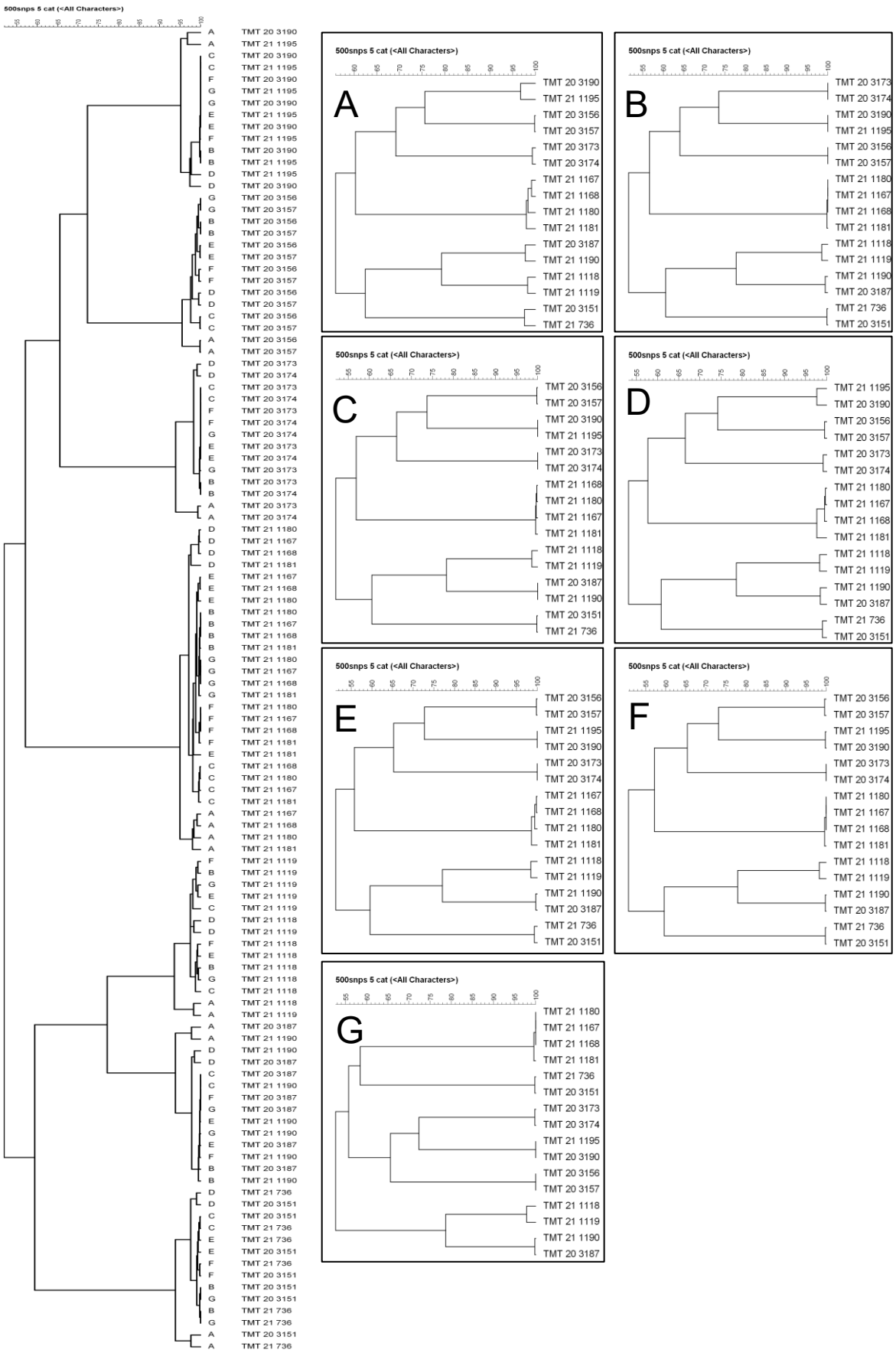


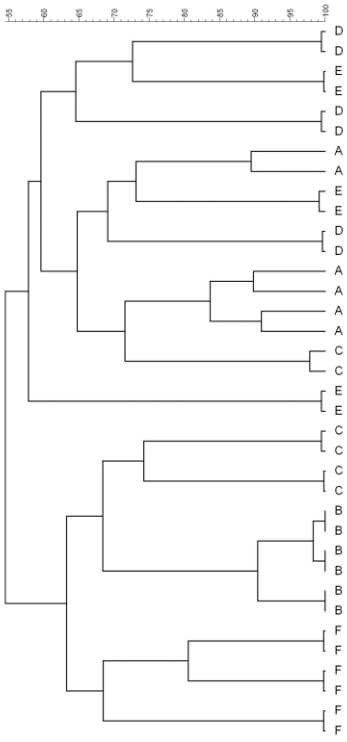
Figure 19: Clustering of different plants from the same variety originating from different seed lots.

Method validation

Besides the performance validation of the SNPs, all labs were tasked with a method validation. In the method validation, testing robustness, repeatability and reproducibility were the main objectives.

Herein the Robustness was defined as; "*measure of the capacity of an analytical procedure to remain unaffected by small variations in method parameters and provides an indication of the method's reliability during normal usage*".
As DNA quality is viewed as one of the most crucial factors in genotyping, each lab was tasked with isolating 3 samples by themselves. Each lab was able to select 3 samples which were included in the 92 samples send out by the coordinator (or previously tested in the test set). The influence of different DNA isolation techniques can be determined by comparing the results obtained by genotyping own isolated DNA to the DNA send by the coordinator.

The results obtained from this analysis are displayed in **Figure 20** in combination with the accompanied table.
As visualised by the dendrogram and similarity table, the profiles generated by partners B, C, D, E and F are highly similar with similarities greater than 97,8. The analysis of partner A was the most affected when comparing the profiles of own isolated DNA to provided DNA. Partner G failed to provide the data for this analysis.



	Sample		
partner	1	2	3
A	89,5	89,8	91,0
B	100,0	100,0	100,0
C	97,8	99,5	99,8
D	99,5	99,5	99,6
E	99,8	99,1	99,5
F	99,8	99,8	99,8
G	Na	Na	Na

Figure 20: Dendrogram and similarity table of own isolated DNA samples compared to provided DNA.
In every pair of samples, the sample on top is derived from the own isolated DNA.

Repeatability and Reproducibility

During the method validation process, repeatability and reproducibility were assessed to gauge the consistency with which each lab can generate genetic profiles. Repeatability defined as; "*Measure of consistency by genotyping a small set of samples, a multitude of times in one experiment, executed by one technician under the same conditions and machines on one day.*" Reproducibility defined as; "*Measure of consistency by genotyping a small set of samples a multitude of times, in two experiments, executed by different technicians under the same conditions but on different machines and points in time*."

As part as the method validation each partner was tasked to perform a repeatability experiment twice. The reproducibility was obtained by comparing the results of both experiments. Within the provided 92 samples, 8 samples were allocated to the repeatability and reproducibility experiments.

Due to various reasons, mainly an insufficient amount of DNA and available time, not all partners were able to perform both experiments to its full extent. Partner A and D were only able to perform the first of 2 repeatability experiments. The range of replicates varied amongst all partners from n=1 to n=12. The exact number and the average similarity across all the replicates are given in Table 12 and 13, repeatability experiments 1 and 2.

Due to the sample size of n=1 within the repeatability experiment 1 of partner A, no average similarity could be calculated. Due to circumstances, partner D only performed the repeatability experiments on 6 samples. With overall similarity scores ranging from 99,4 to 100, in both experiments, every partner has displayed their ability to achieve highly reproducible outcomes within an experiment.

Table 12: Average similarities of replicates within repeatability experiment 1.
(n/a; not available, n/p; not produced)

	Partner						
	A	B	C	D	E	F	G
Replicates	n=1	N=11	n=8	n=3	n=12	n=8	n=6
TMT 20 2406	n/a	99,8	99,8	n/p	99,6	99,7	99,9
TMT 21 1133	n/a	99,8	99,9	100	99,4	99,5	99,7
TMT 21 1145	n/a	99,9	99,9	99,8	99,4	99,9	99,9
TMT 21 1167	n/a	100	99,9	99,9	99,4	99,7	99,9
TMT 21 1168	n/a	100	99,9	n/p	99,5	99,9	100
TMT 21 1180	n/a	100	99,9	100	99,6	99,9	99,9
TMT 21 1181	n/a	100	99,9	100	99,6	99,6	99,9
TMT 21 805	n/a	99,7	99,9	100	99,5	99,7	99,6

Table 13: Average similarities of replicates within repeatability experiment 2.
(n/p; not produced)

	Partner						
	A	B	C	D	E	F	G
Replicates	n=0	N=11	n=8	n=0	n=12	n=8	n=6
TMT 20 2406	n/p	100	99,8	n/p	99,2	99,5	99,9
TMT 21 1133	n/p	100	99,7	n/p	99,8	99,6	100
TMT 21 1145	n/p	100	99,9	n/p	99,7	99,9	100
TMT 21 1167	n/p	100	99,8	n/p	99,9	99,8	99,9
TMT 21 1168	n/p	100	99,8	n/p	98,9	99,8	100
TMT 21 1180	n/p	99,9	99,9	n/p	99,9	99,8	100
TMT 21 1181	n/p	100	99,8	n/p	99,8	99,9	99,9
TMT 21 805	n/p	100	99,6	n/p	99,7	99,5	99,9

When combining the data derived from both repeatability experiments the reproducibility can be determined. The reproducibility indicates how consistent each partner is able to generate the genetic profiles of a sample from time to time. More importantly, how affected the genotyping process is by day to day variations (e.g. different apparatus, technicians, batches of chemicals).

Due to the absence of data regarding repeatability experiment 2 for partner A and D, the reproducibility has been calculated using the earlier genotyped samples. These samples have only been genotyped once without any replicates. The results of the reproducibility is given in **Table 14**: Reproducibility, average similarity of replicates between experiments. Some variation is observed in the profiles of partner A, with average similarities ranging from 94,8 to 97,6. The small sample size on which the average is calculated should be taken into account. Without the lowest scoring sample the overall similarities are in the range of 95,7 up to 97,8. The range of these similarities have also been observed within the technical and biological replicates, meaning partner A is evenly consistent within as between experiments. Approximately the same conclusion can be drawn for partner D, where the average similarities of the reproducibility experiment (99,6 - 100) are higher than the similarities observed within the technical and biological replicates (98,6 - 100).

For partners B, C, E, F and G the average similarities between experiments ranges from 99,4 up to 100.
Whilst per individual sample small variations can be observed between repeatability experiment 1, repeatability experiment 2 and reproducibility, the range of similarities remains stable. Hence all partners have shown that they are evenly consistent within as between experiments.

Table 14: Reproducibility, average similarity of replicates between experiments.
(n/a; not available)

	Partner						
	A	B	C	D	E	F	G
TMT 20 2406	96,5	99,9	99,8	n/a	99,4	99,6	99,9
TMT 21 1133	97,2	99,9	99,8	100	99,5	99,6	99,9
TMT 21 1145	96,9	100	99,9	99,9	99,4	99,9	100
TMT 21 1167	97,8	100	99,8	99,9	99,6	99,7	99,9
TMT 21 1168	94,8	100	99,8	n/a	99,2	99,8	100
TMT 21 1180	96,5	100	99,9	100	99,7	99,8	100
TMT 21 1181	97,6	100	99,8	99,8	99,6	99,7	99,9
TMT 21 805	95,7	99,3	99,7	99,6	99,6	99,6	99,7

Discriminative power of the SNP set

Excluding the duplicate samples used as technical and biological replicates, the plate of samples contained DNA of 74 individual varieties. The dendrogram of these 74 samples is displayed in **Figure 21,** an enlarged image is provided in Annex 2. Based on the consensus profile, 66 varieties are well distinguishable from each other with a maximum similarity of <94,1% (<90% for n=63). In addition there have been found 4 pairs of 2 varieties with a high similarity. The similarity in these pairs ranges from 98,0% to 100,0%. The high similarity based on genotyping can be explained by looking into the Distinctness based on morphology for these pairs. The pairs of similar varieties, their similarity and expert notes regarding morphology is given in **Table 15**: Observed similar varieties based on consensus genotype.

Figure 21: Dendrogram of 74 varieties included in the sample set of 92.

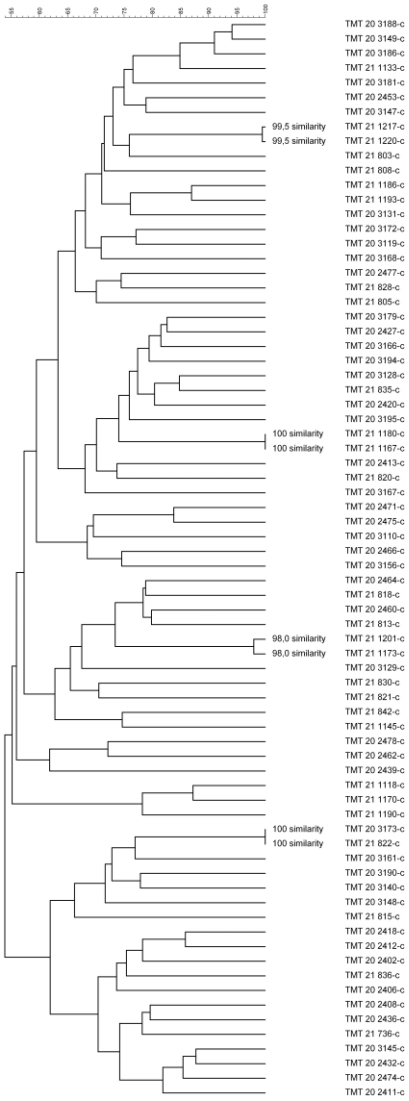
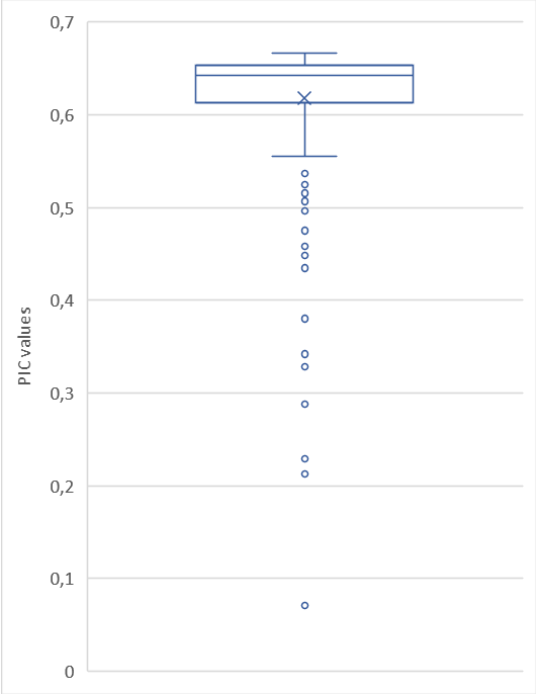


Table 15: Observed similar varieties based on consensus genotype.

Sample denomination	Observed similarity	Expert notes on morphology	Sample origin
TMT 21 1217	99,5%	Only different in resistance against TSWV	NL
TMT 21 1220			NL
TMT 21 1180	100,0%	Very similar on morphology but concluded sufficiently distinct	NL
TMT 21 1167			NL
TMT 21 1201	98,0%	Very similar on morphology	NL
TMT 21 1173			NL
TMT 20 3173	100,0%	Only different in resistance against *Verticillium*	FR
TMT 21 822			ES

Using the consensus genotype of the 74 varieties, the discriminative power otherwise known as Polymorphism Information Content (PIC) was calculated. The range of PIC values varied from the highest possible score of 0,66 down to 0,0713. When plotting the PIC values in a boxplot (shown in **Figure 22**) roughly the same pattern can be observed as previously shown in **Figure 11**. Based on the 74 varieties, 271 SNPs had a PIC value < 0,55 and 27 SNPs were marked as outliers.

Figure 22: Boxplot PIC values of 297 SNPs based on 74 genotypes.



When comparing the outliers derived from the 74 genotypes (PIC_{74}) to their earlier calculated PIC values based on 183 samples (PIC_{183}) a number of things can be observed. 9 SNPs show a PIC_{74} greater than PIC_{183}, 12 SNPs show a PIC_{74} smaller than PIC_{183} and 6 SNPs show a PIC_{74} roughly equal PIC_{183}. In addition, the PIC values have been calculated based on the 257 varieties genotyped (PIC_{257}) within the project. For most of the PIC_{74} outliers the PIC values derived from the 183 and 257 genotypes are in accordance with each other. A total of 19 SNPs have proven to be constant outliers throughout the analysis. Even though they're considered as outliers, PIC values greater than 0,4 can be observed in 14 SNPs rendering them useful for genotyping. 5 SNPs yielded a PIC value < 0,4 rendering them useful for genotyping but to a lesser extent. An overview of the 27 outlier SNPs and their respective PIC values is given in **Table 16.**

Table 16: 27 outlier SNPs based on 74 genotypes and respective PIC values.
*PIC value not considered an outlier in the analysis based on respective number of genotypes.

SNP	PIC based on 74 genotypes	PIC based on 183 genotypes	PIC based on 257 genotypes
SL3.0ch04_713521	0,54	0,41	0,58*
SL3.0ch09_2147796	0,54	0,61*	0,58*
SL3.0ch08_65435722	0,53	0,61*	0,59*
SL3.0ch08_65437054	0,53	0,61*	0,59*
SL3.0ch08_65439559	0,53	0,61*	0,59*
SL3.0ch04_709346	0,53	0,59*	0,57*
SL3.0ch10_1421931	0,53	0,61*	0,59*
SL3.0ch04_705852	0,52	0,56*	0,54
SL3.0ch02_45584521	0,51	0,47	0,50
SL3.0ch04_1177529	0,51	0,45	0,53
SL3.0ch02_39701103	0,50	0,27	0,33
SL3.0ch02_44792764	0,50	0,47	0,49
SL3.0ch12_62814879	0,48	0,47	0,47
SL3.0ch04_861510	0,48	0,43	0,44
SL3.0ch06_33298117	0,46	0,41	0,43
SL3.0ch08_63297958	0,45	0,45	0,43

SL3.0ch12_1240802	0,43	0,05	0,36
SL3.0ch02_51991732	0,38	0,50	0,47
SL3.0ch02_36154538	0,38	0,44	0,41
SL3.0ch08_65375768	0,35	0,37	0,28
SL3.0ch08_65862705	0,34	0,41	0,40
SL3.0ch08_65825308	0,33	0,42	0,40
SL3.0ch06_46016406	0,29	0,31	0,30
SL3.0ch04_1178770	0,23	0,26	0,23
SL3.0ch09_4836912	0,21	0,05	0,22
SL3.0ch09_4869306	0,21	0,21	0,22
SL3.0ch09_12404354	0,07	0,26	0,19

Discussion and Conclusions

Early on in the project, it became evident that the complexity of the legal aspects had been underestimated. This, linked with the time required to secure consent from breeders and establish the sample selection criteria, led to a substantial delay compared to the initially proposed timeline. Fortunately, after being granted a formal extension of the project by the CPVO, the work planned within the scope of the project could be finalised.

Throughout the project the lab partners have successfully genotyped 257 varieties on 297 robust SNPs.
In the developmental stage of the project these SNPs yielded an average similarity of >95% and were able to be genotyped in at least 75% of the participating labs. After the validation process of the SNPs, 236 upheld the initial criteria of being reproducible for at least 95% between the participating labs. Based on the data provided by all partners on the 297 SNPs, an average similarity of 96,5% on 89 out 92 samples was achieved. As mentioned above, only 236 SNPs upheld the quality criteria. When focusing on only the 236 high quality SNPs the average similarity of the genotypes will rise, evidently this will lower the discriminative power of the SNP-set. Whilst the discriminative power will be lower with a smaller number of SNPs in the set, it is not expected to have any major consequences on future applications. Alternatively, a workflow could be implemented focusing on the strengths of each lab. Keeping the 297 SNPs selected, but not allowing every partner to provide data for a particular SNP. By following this workflow, some profiles might be incomplete in relation to the 297 SNPs but their overall quality is assured. The optimal workflow, supported by all contributing partners, should be discussed in the period leading up to a follow-up project. This follow-up project will predominantly focus on; A) construction of a database, B) generating genetic profiles of Common Knowledge Varieties (CKV), C) implementing a tool for similar variety identification usable in DUS trials.

After the SNP set validation it can be concluded that the harmonised SNP set is able to; distinguish varieties that are morphological distinct, generate similar profiles for morphological similar varieties, generate highly similar profiles for biological and technical replicates. Furthermore, during the validation process, each lab has demonstrated their ability to generate highly similar profiles (±95%) within and between experiments. A side note on the SNP set being able to distinguish varieties that are morphological distinct; during the project only a limited number of varieties have been tested in relation to the entire catalogue of tomato. The effectiveness and abilities of the SNP set should further be assessed in a larger representation of the common catalogue of tomato.

Follow-up and future perspective

As the results of the project look promising for future application, plans will be made for a follow-up project focussing on the construction of a database using the harmonized SNPs. Ideally all of the current project partners will be involved within the framework of the follow-up. During the entirety of 2024, the coordinator together with the project partners, will define the framework of the follow-up. Herein the major aspects of the follow-up will be further defined, this includes; participating partners and their roles, user requirements and data management of the database, gauge acceptance and support amongst breeders (throughout representative organisations) for an international collaboration, legal framework and the limitations regarding the sharing of data.