UPOV
International Union for the Protection of New Varieties of Plants

E

Technical Working Party for Vegetables

TWV/56/21

Fifty-Sixth Session
Virtual meeting, April 18 to 22, 2022

Original: English
Date: April 6, 2022

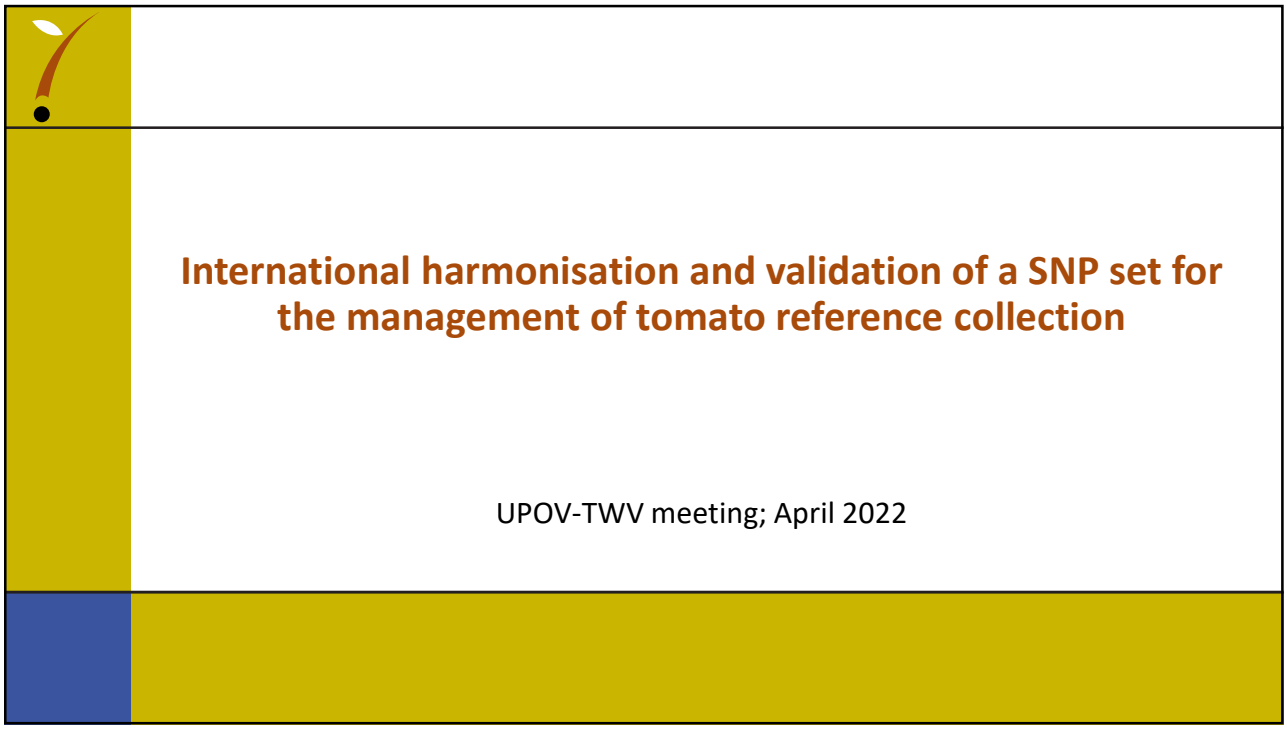
PRESENTATION ON THE USE OF MOLECULAR TECHNIQUES IN DUS EXAMINATION

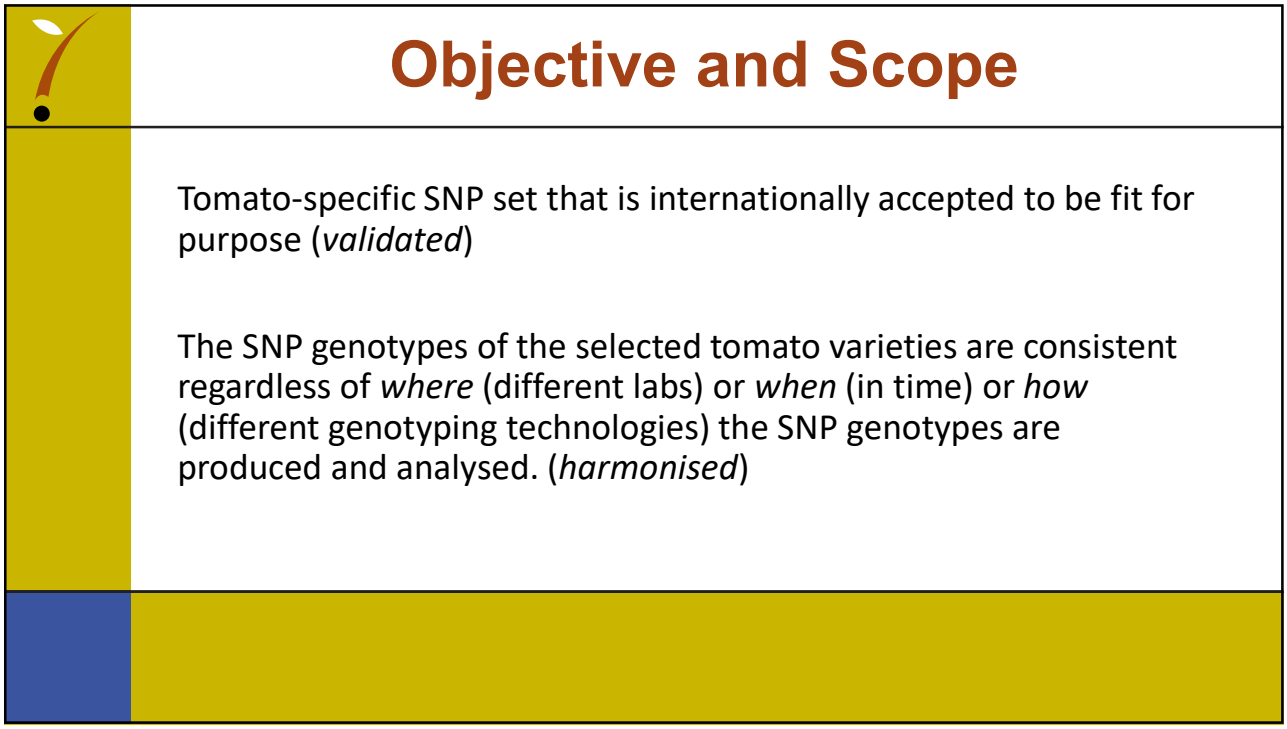
Document prepared by an expert from the Netherlands

Disclaimer: this document does not represent UPOV policies or guidance

The annex to this document contains a copy of a presentation "International harmonisation and validation of a SNP set for the management of tomato reference collection", to be made by an expert from the Netherlands, at the fifty-sixth session of the TWV.

[Annex follows]


International harmonisation and validation of a SNP set for the management of tomato reference collection
UPOV-TWV meeting; April 2022


Objective and Scope
Tomato-specific SNP set that is internationally accepted to be fit for purpose (*validated*)
The SNP genotypes of the selected tomato varieties are consistent regardless of *where* (different labs) or *when* (in time) or *how* (different genotyping technologies) the SNP genotypes are produced and analysed. (*harmonised*)

Objective and Scope

Validated and harmonised SNP set and SNP genotypes are prerequisites for…

- International DNA database for tomato
- Use the database in DUS procedure for the management of reference collection (UPOV-models)

First step: this project

General Project information

Project started July 2019 (grant agreement between CPVO and Naktuinbouw)

Budget €295.000; co-financed by CPVO for 90%

Duration 30 months (December 2021) – extended with 20 months (August 2023)

Delay:

- Legal arrangements like Project Partner Agreement and Agreement on ownership and use of plantmaterial and DNA samples
- Requesting consent of the titleholders
- Covid 19

Project Partners

CPVO entrusted EOs for tomato DUS testing

Observers

Partners - Within Europe

GEVES (France)
COBORU (Poland)
NÉBIH (Hungary)
INIA (Spain)
DGAV (Portugal)
CREA (Italy)
Naktuinbouw (Netherlands)

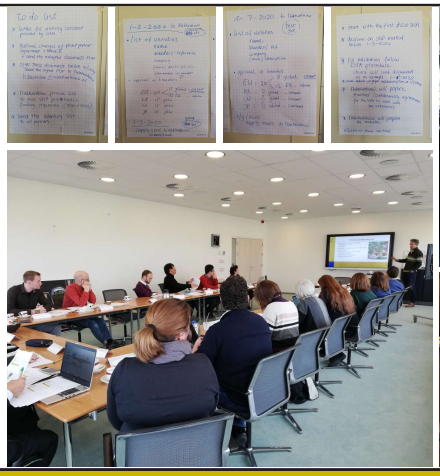
€

Euroseeds (Europe - breeders)
CPVO (Europe – financial support)

Other Partners - Outside Europe

IVF of CAAS (China)
KSVS (Republic of Korea)
NARO (NCSS) (Japan)

Kick-off meeting (dec 2019)



Project structure - different phases

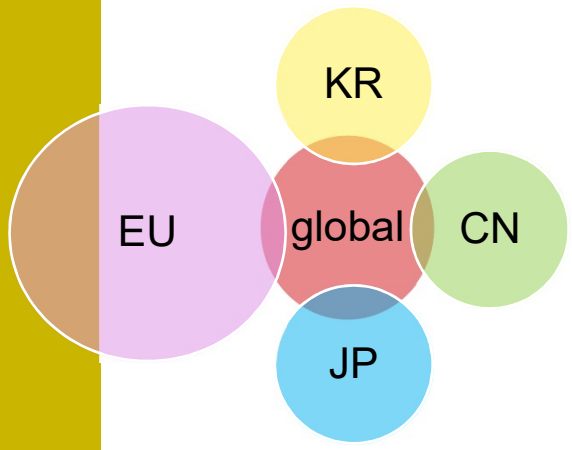
Preparation
- Variety selection
- Request for consent from breeders
- Collection of plant material
- SNP selection and share information
- Exchange information on genotyping methods
- MTA

Lab Work
- CT1: Optimization starting with 500 SNPs and select final SNP set
- CT2: Validation of SNP performance within each lab
- CT2: Investigate SNP performance between labs; different methods in different labs
- PT3: Test blind samples

Analysis & Report
- Individual labs
- Comparing labs
- Proficiency test
- Statistical analysis
- Writing, exploitation and dissemination

Variety Selection

Different sets of varieties



European sets

Korean sets

Chinese sets

Japanese sets

Training set = Developmental set
Used for development and optimization

Test set = Validation set
Used for validation within and between labs

Global Developmental / Training set and Global Validation / Test set

All varieties that are in common for all project partners

Criteria for training / developmental sets

- Representation of a broad genetic diversity (all types and all characteristics)
- varieties that are morphologically close but distinct, (variety pairs that might have caused some discussion in the DUS test and/or an extra year of testing was required to consider them distinct)
- different companies (different germplasms)
- No wild species

Criteria for test / validation sets

Samples that should be distinct

- ✓ Genetically very similar varieties or lines, NILs, RILs
- ✓ Parents and offspring
- ✓ Varieties with similar morphological descriptions, different in just one/few characteristics (e.g. resistance)
- ✓ Varieties with similar morphological descriptions with different pedigree or from different companies

Samples that should not be distinct

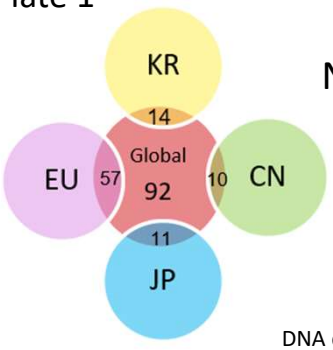
- ✓ Duplicated DNA templates (technical replicates)
- ✓ Different DNA samples from the same variety / seed lot (biological replicates)
- ✓ Different individual plants from the same variety / seed lot (expected to have identical or nearly identical genotypes)
- ✓ Plant material from the same variety but different origins, different seed lots (expected to have identical or nearly identical genotypes)

Varieties - Overview

	paper selection total number of varieties	consent of the breeder	no consent needed	seeds received by Naktuinbouw	DNA extracted	Used in GLB set	Used in EU set
España	42	42	0	42	42	0	17
Portugal	40	5	0	5	5	4	1
France	54	41	13	54	54	10	21
Italy	40	40	0	40	40	15	8
Poland	40	36	0	36	36	0	25
Hungary	40	31	0	31	31	15	5
Netherlands	157	128	0	128	128	13	14
Republic of Korea	15	0	15	15	14	14	0
China	10	10	0	10	10	10	0
Japan	15	11	0	11	11	11	0
total	***453***	***344***	***28***	***372***	***371***	***92***	***91***

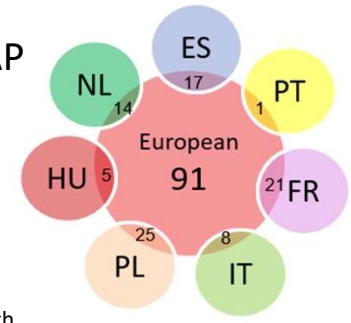
GLB set vs EU set

Plate 1



NO OVERLAP

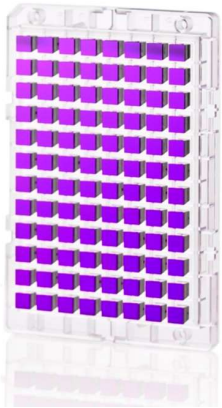
Plate 2



DNA quality was very high
200 ul of 20 ug/ul

SNP selection as input for this project

- Axiom Array with 51.214 SNPs
- 192 tomato varieties
  (representative for the diversity)



Providing SNP information to partners

- We start with the best 500 SNPs
- Positions of these 500 SNPs on the reference genome were shared
- Flanking sequences 100 bp upstream and 100 bp downstream of the SNP position suitable for primer design were shared

Coordinator provide 500 – 1000 SNPs and flanking regions

Genotyping methods

	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

Lab work

3 Comparative tests with different aims

1st Labmeeting
April 2021

1. CT for optimization and final selection of SNPs

2nd Labmeeting
March 2022

2. CT for method validation (in each lab & and comparison between labs)

3rd Labmeeting
Jan 2023

3. PT for identification of blind samples

Two online labmeetings

First lab meeting: April 22, 2021

Second lab meeting: March 31, 2022

Data analysis on SNP 'Performance'

1. How many SNP assays are successful in producing genotypes for the varieties?

2. Are the genotypes for the SNPs produced consistent between the partners?

3. Can the varieties of the developmental sets be distinguished?

Input file for all analyses

	SNP1-Partner A	SNP2-Partner A	SNP3-Partner A	SNP4-Partner A	SNP5-Partner A		SNP500-Partner A
Var1-Partner A	RR	RA	RA	RA	-		AA
Var2-Partner A	RA	RR	AA	-	-		RR
Var3-Partner A	AA	RR	RA	-	-		RA
Var92-Partner A	RA	RR	AA	AA	-		RA

500x7 SNPs

Two different alleles: R = Reference and A = Alternative

92x7 varieties in GLB set
91x4 varieties in EU set

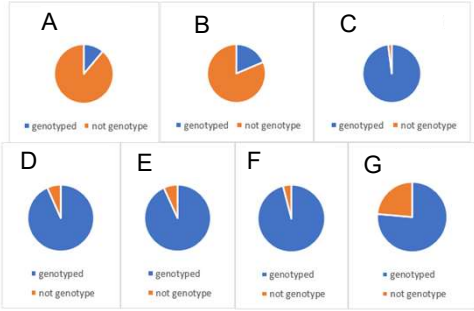
Successful SNPs in GLB set

a SNP is successful when a genotype was obtained for at least one sample (variety).

Plate 1 (GLB set) for 7 partners

	GLB	GLB	GLB	GLB	GLB	GLB	GLB
	A	B	C*	D	E	F	G
# SNPs genotyped	56	93	490*	467	466	480	382
% SNPs genotyped	11%	19%	98%*	93%	93%	96%	76%
# SNPs not genotyped	444	407	10	33	34	20	118
% SNPs not genotyped	89%	81%	2%	7%	7%	4%	24%

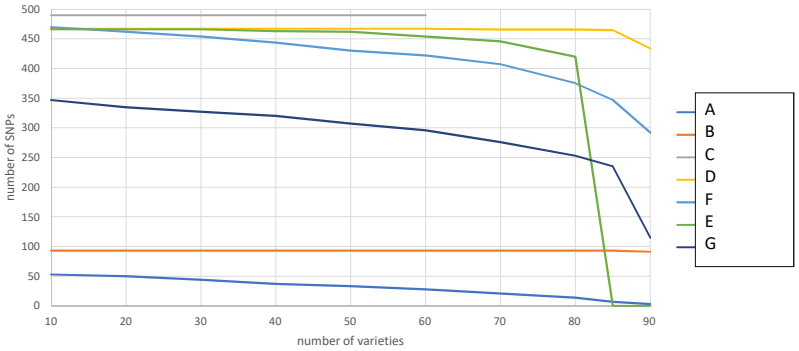
Rate of successful SNPs from the initial 500



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

Successful SNPs in GLB set

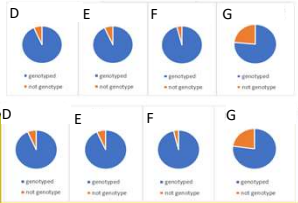
Number of successful SNPs per number of varieties from the Global developmental set for each partner

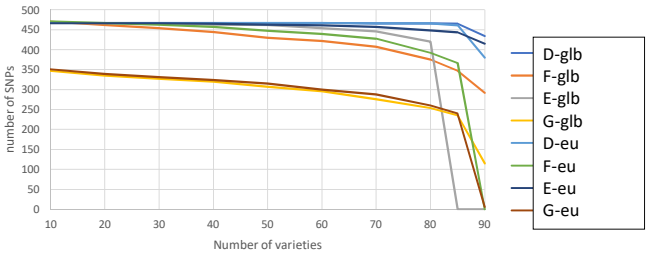


Successful SNPs: GLB vs EU set

a SNP is successful when a genotype was obtained for at least one sample.

	Plate 1 (GLB set) for 4 EU partners				Plate 2 (EU set) for 4 EU partners			
	GLB	GLB	GLB	GLB	EU	EU	EU	EU
	D	E	F	G	D	E	F	G
# SNPs genotyped	467	466	480	382	467	466	481	386
% SNPs genotyped	93%	93%	96%	76%	93%	93%	96%	77%
# SNPs not genotyped	33	34	20	118	33	34	19	114
% SNPs not genotyped	7%	7%	4%	24%	7%	7%	4%	23%





comparison of successful SNPs per number of varieties between the GLB set and the EU set for the 4 EU lab partners

Conclusions

- The number of successful SNP assays is very variable between the 7 partners
- For most partners the number of successful SNPs drops for >80 varieties. So, in most of the SNP datasets we can observe missing data for 0-20 varieties
- Not the same varieties are missing in the several datasets. The missing varieties are randomly divided and different for each partner. From this observation we can conclude that DNA quality is not the reason for genotype failure of a particular variety.
- From these results we cannot draw a conclusion on which technology or genotyping method is preferable
- The number of successful SNPs for each of the EU partners is very consistent: the results on the GLB and EU sets of varieties are very consistent for each partner
- Whether a SNP is successful or not, is independent on the set of varieties

Input file for all analyses

	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

Two different alleles
R = Reference (refgenome)
A = Alternative

sample 1	sample 2	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

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

Are the genotypes produced for each SNP consistent between the partners?

Consistency of genotypes

We want to select SNPs that produce consistent genotypes by all partners for each variety

	SL3.0ch01_346524_E	SL3.0ch01_346524_D	SL3.0ch01_346524_F	SL3.0ch01_346524_C	SL3.0ch01_507890_E	SL3.0ch01_507890_D	SL3.0ch01_507890_F	SL3.0ch01_507890_C
SL3.0ch01_346524_E	100	40,97	100	100	18,91	24,1	21,52	17,86
SL3.0ch01_346524_D	40,97	100	42,94	36,93	28,66	31,53	30,11	23,08
SL3.0ch01_346524_F	100	42,94	100	100	18,91	23,37	21,03	16,93
SL3.0ch01_346524_C	100	36,93	100	100	20	24,62	22,58	16,93
SL3.0ch01_507890_E	18,91	28,66	18,91	20	100	89,64	91,03	94,55
SL3.0ch01_507890_D	24,1	31,53	23,37	24,62	89,64	100	100	90,77
SL3.0ch01_507890_F	21,52	30,11	21,03	22,58	91,03	100	100	91,94
SL3.0ch01_507890_C	17,86	23,08	16,93	16,93	94,55	90,77	91,94	100

a snapshot of the total similarity matrix for the pair-wise comparison of successful SNPs per partner for the Global Developmental set

Matrix comparing
(all SNPs for all partners) x (all SNPs for all partners)

dataset is not compete: missing SNP assays for partners – input is successful SNPs for each partner

Per SNP we compare the genotypes obtained by partner X to the genotypes obtained by partner Y

When the genotypes are consistent, the similarity is 100

We also calulated the average similarity per SNP over the genotyes of all partners as an expression of consistency

Green SNP=70,14 Blue SNP=92,99

Consistency of genotypes

For 494 SNPs we obtained successful genotypes for at least 2 partners
The number of pair-wise combinations that is used to calculate the average similarity for all 494 SNPs is 4614.

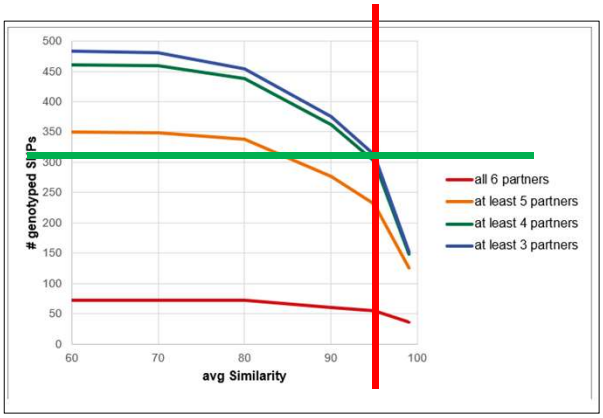
	#SNPs genotyped by N partners				
average similarity range	N = 6	N = 5	N = 4	N = 3	N = 2
>99	36	90	22	5	1
>95	55	176	71	11	3
>90	60	217	85 (Blue SNP)	13	3
>80	72	266	100	16	5
>70	72	277	110 (green SNP)	22	6
>60	72	278	111	23	6
<60	*0*	*1*	*0*	*0*	*3*

Consistency of genotypes

varieties with a successful genotype was not taken into consideration in the table.

However, to determine SNP 'performance' also includes the successful genotyping on as many as possible varieties.

Selection of high performance SNPs
- Average sim is most important! Be strict! >95
- Not much difference between 3 and 4 partners. Be strict! At least 4 partners



Consistency of genotypes

For 494 SNPs we obtained successful genotypes for at least 2 partners
The number of pair-wise combinations that is used to calculate the average similarity for all 494 SNPs is 4614.

	#SNPs genotyped by N partners				
average similarity range	N = 6	N = 5	N = 4	N = 3	N = 2
>99	[illegible]	[illegible]	[illegible]	[illegible]	1
>95	55	176	71	[illegible]	3
>90	[illegible]	[illegible]	[illegible] (blue SNP)	[illegible]	3
>80	72	266	100	16	5
>70	72	277	110 (green SNP)	22	6
>60	72	278	111	23	6
<60	*0*	*1*	*0*	*0*	*3*

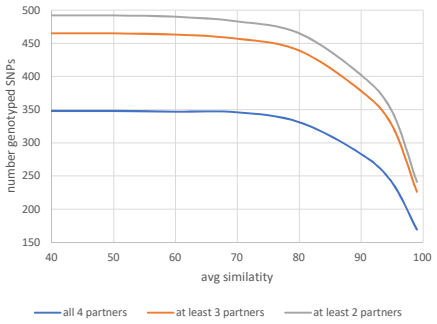
Best performing SNPs:
Very high average sim: >95
At least 4 partners: N≥4

#SNPs: 55+176+71=302

As the Global Set

Consistency of genotypes

	#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*



Best performing SNPs:
Very high average sim: >95
At least 3 partners: N≥3

#SNPs: 240+86=326

As the European set

Varieties

Criteria for developmental set:

- Representation of a broad genetic diversity (all types and all characteristics)
- varieties that are morphologically close but distinct, (variety pairs that might have caused some discussion in the DUS test and/or an extra year of testing was required to consider them distinct)
- different companies (different germplasms)
- No wild species

Input file for all analyses

	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

Two different alleles
R = Reference
A = Alternative

sample 1	sample 2	IBS
RR	RR	2 (both alleles in common)
RR	RA	1 (one allele in common)
RR	AA	0 (no allele in common)

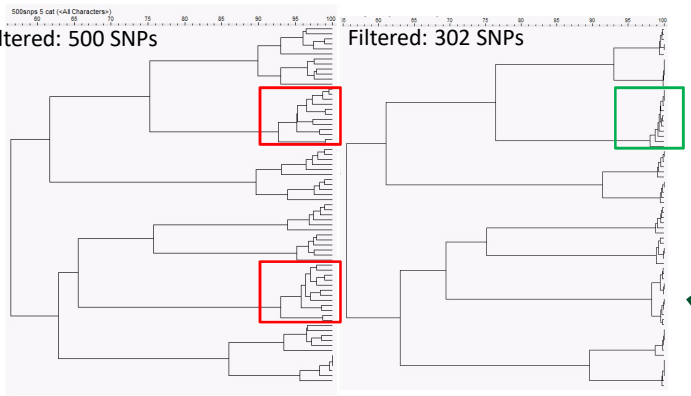
Calculation of the Identity-by-State value

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

Can the varieties of the developmental sets be distinguished?
But, are the variety-samples of all partners clustering together?

Discriminative power - GLB

Are all varieties in GLB developmental set be distinguished with SNP set?



Before filtering:
- 88 varieties were distinct
- 2 pairs of varieties not distinct

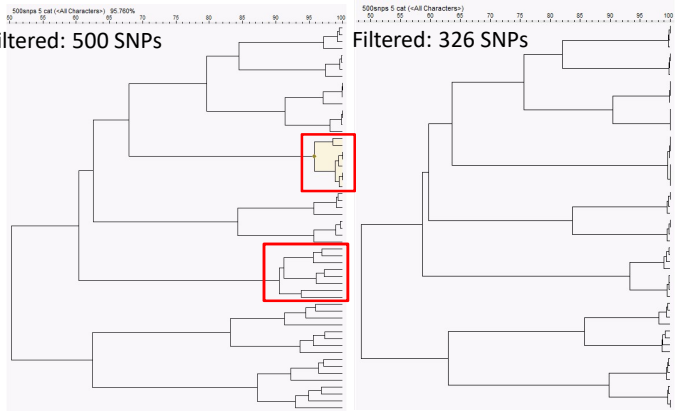
After filtering:
- 1 pair was still not distinct

After filtering:
- Reduction of 'noise'

The effect of filtering the SNPs on 'performance'

Discriminative power - EU

Are all varieties in EU developmental set be distinguished with SNP set?



Before filtering:
- 87 varieties were distinct
- 2 pairs of varieties not distinct

After filtering:
- All varieties were distinct.
- 2 SNPs difference

After filtering:
- Reduction of 'noise'

The effect of filtering the SNPs on 'performance'

Discriminative power - PIC

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

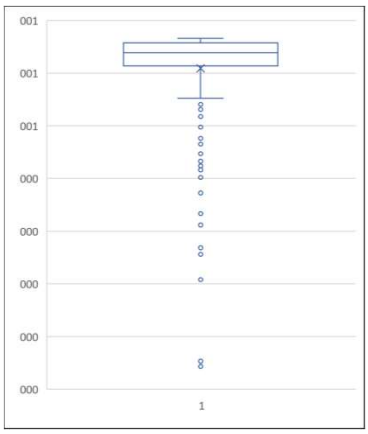
PIC-value calculation only possible on single genotype per variety

Best way = to use consensus genotype

Not possible for unfiltered set of SNPs, we only calculated PIC for filtered set

For now, we used the Partner D dataset (most complete)

This is very close to consensus as the genotypes are consistent in filtered set



Discriminative power

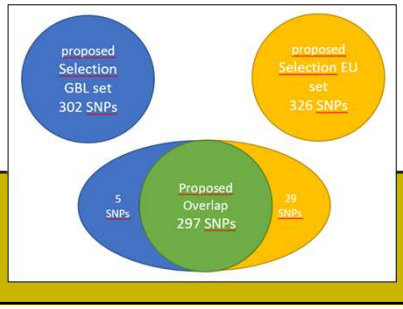
Sample number	Contributing project partner	Information on company or description	set	Different conclusion? After filtered SNP set
3106	Poland	Different companies, distinct varieties on morphology	EU	No longer 100% match (99,5) 2 consistent SNPs difference between these varieties
3135				
3183	France	Indicated as close to each other	EU	93,6. Clearly distinct genotypes
3170				
1703	Republic of Korea	?	GLB	97,96%. But distinct clusters
1705				
2473	Hungary	?	GLB	Still 100% match
1715	Japan	?		
3191	France	?	EU	Not yet checked
1719	Japan	?	GLB	

Final selection of SNPs

Agreement to proceed with 1 International SNPs set.

Agreement to use average similarity >95 and N≥4 for GLB set

Agreement to use average similarity >95 and N≥3 for EU set



overlap between GLB and EU
297 SNPs

Future work

Genotyping validation / test set varieties by all partners (1 plate; 90 samples)

Method validation by each partners individually for every method
- repeatability, reproducibility and robustness

3rd Lab meeting

Blind test

[End of Annex and of document]