Allegro Genotyping solution : Application in crops

Damien D. Hinsinger

Damien.hinsinger@inrae.fr

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

INRAE

EPGV
Etude du Polymorphisme des Génomes Végétaux

université PARIS-SACLAY

Etude du Polymorphisme des Génomes Végétaux (EPGV)

- INRAE unit (US 1279), BAP department
- Located in Genoscope (CEA-Evry, France)

- Use all sequencing technologies available at Genoscope / CNRGH

- Focus on variations in plant genomes:
  - > Single Nucleotide Polymorphism (SNPs)
  - > Structural Variants (SVs)

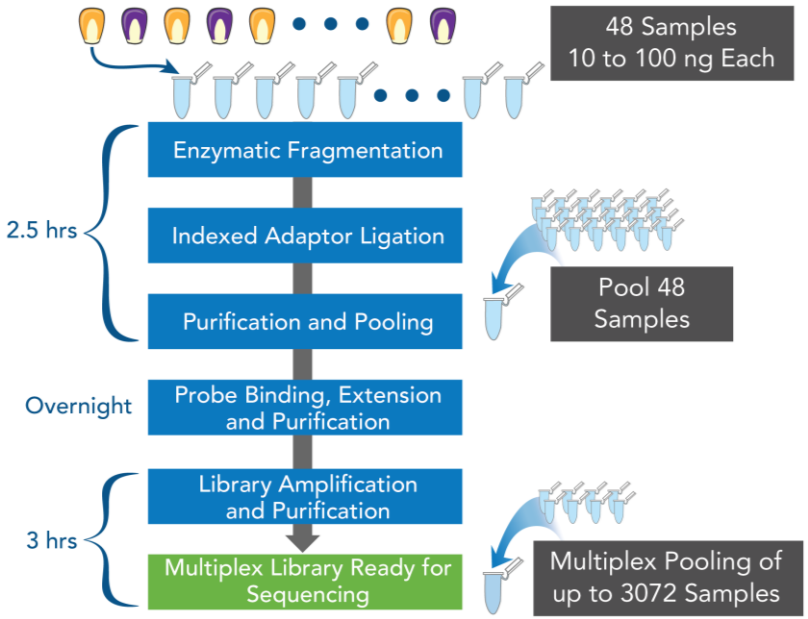
- Different levels of polymorphism (intra-individual, intra- & inter- species)

- Genome sequencing and genotyping

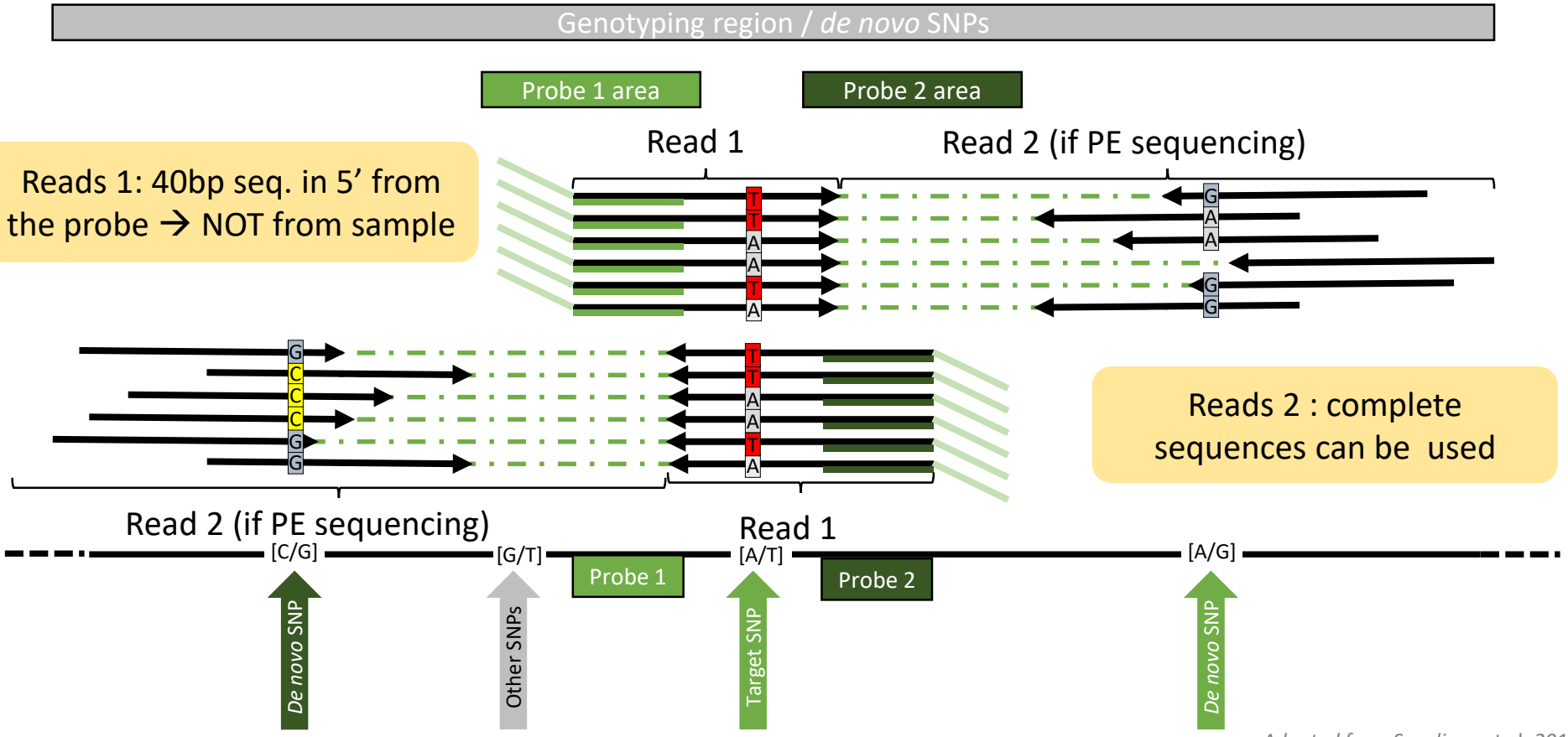
Support and collaboration for research projects in plant genomics

Allegro Targeted Genotyping V2

- **Technological approach:** targeted SNPs sequencing using **SPET** (Single Primer Enrichment Technology): extension of specific hybridized probes, followed by amplification
- **Target detection:** 2 specific probes/SNP
  1 for each strand
- **Molecular tags:** 6 bases (to remove PCR duplicates)
- **SNP number:** 5,000 to 100,000 probes panels
  (2,500 to 50,000 SNPs),
  more on demand
- **Samples number:** 48 ind. / library
  up to 3,072 ind. / seq. multiplex
- **Input DNA quantity:** 10 - 100 ng / sample
- **Sequencing:** SE 100 (TECAN specifications)
  PE 150 possible



Allegro Targeted Genotyping V2: Principle



Adapted from Scaglione et al. *2019*

Project A: Design

- Genotyping objectives

  - 25,000 target SNPs in coding régions

    (Kreplak *et al.*, INRAE Dijon from Ogulten *et al.*, 2018)

  - 336 individuals (1 sample = 1 individual ; no pool of ind.)

Project A: Design

Identification of needs & project design & order

Submit SNP coordinates to expert design team

Receive design report & approve or improve

Custom system delivery

25,292 SNPs with SNP context region

201 bases fasta seq.

SNP in central position (pos. 101)

Probe stat : *for each design probe :*

- *genome coordinates*
- *target name*
- *Sequence*
- *Tm*
- *number of genome alignments*

BED file for target regions and probes

Target level success : target SNP coverage

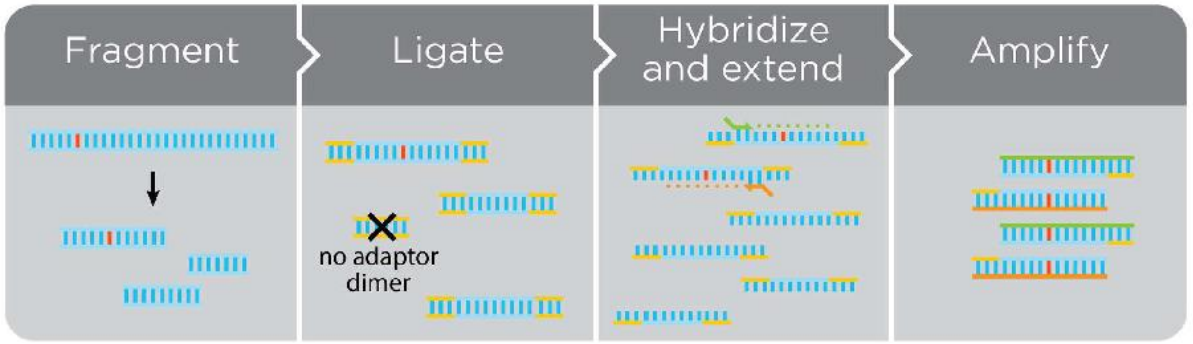
Manufactured & shipped in 4-6 weeks

Subset of 25,000 SNPs (24,958 covered by 2 probes) : 49,958 probes

Libraries construction

- 336 individuals from the collection
- 60 ng DNA / sample
- Interpool biological replicates

  (2 replicated individuals)



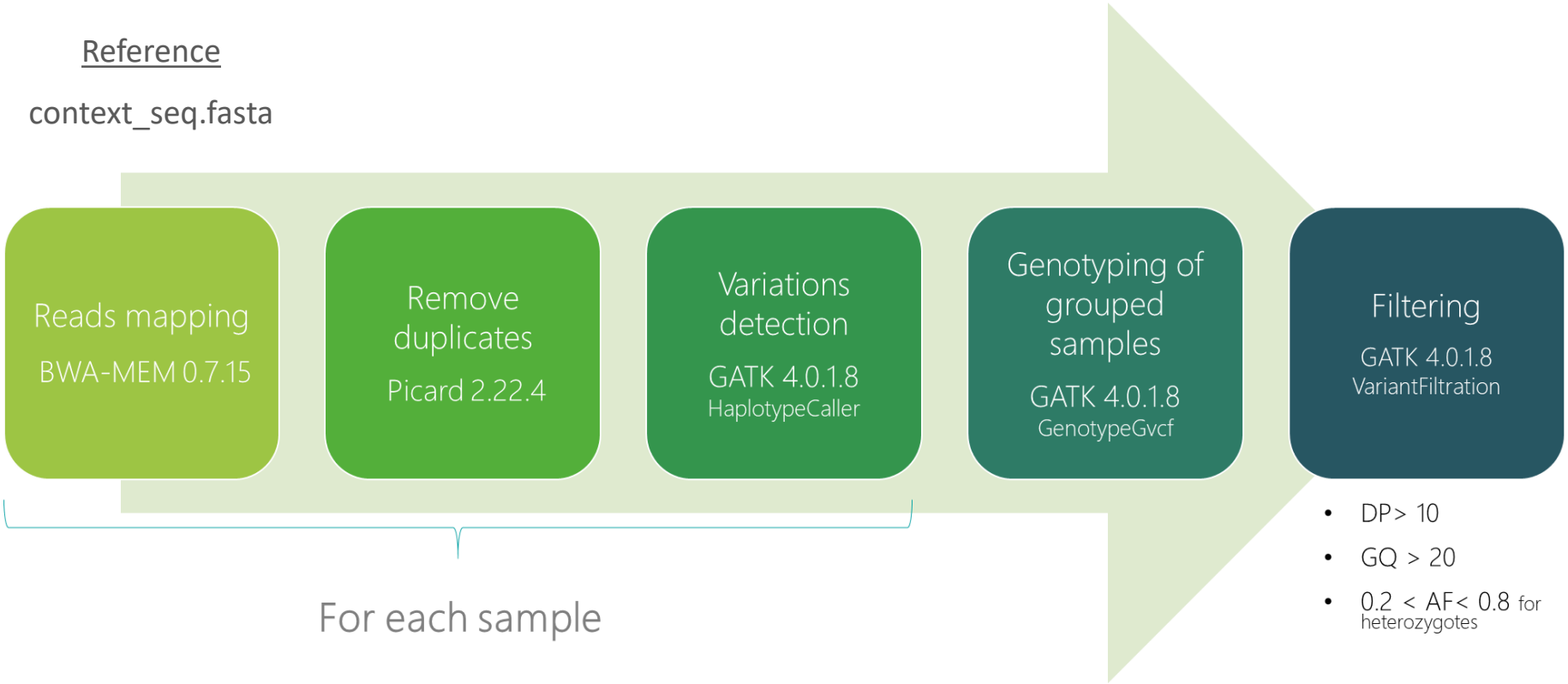
- Individual indexing of samples, then pooling
- Next steps of libraries construction in pools (1 pool = 1 tube)

Sequencing

- 7 pools of 48 individuals
- NovaSeq SP PE 2x150 cycles

Average sequencing depth >=Q30 :96x/target/individual
(min: 37x; max: 207x)

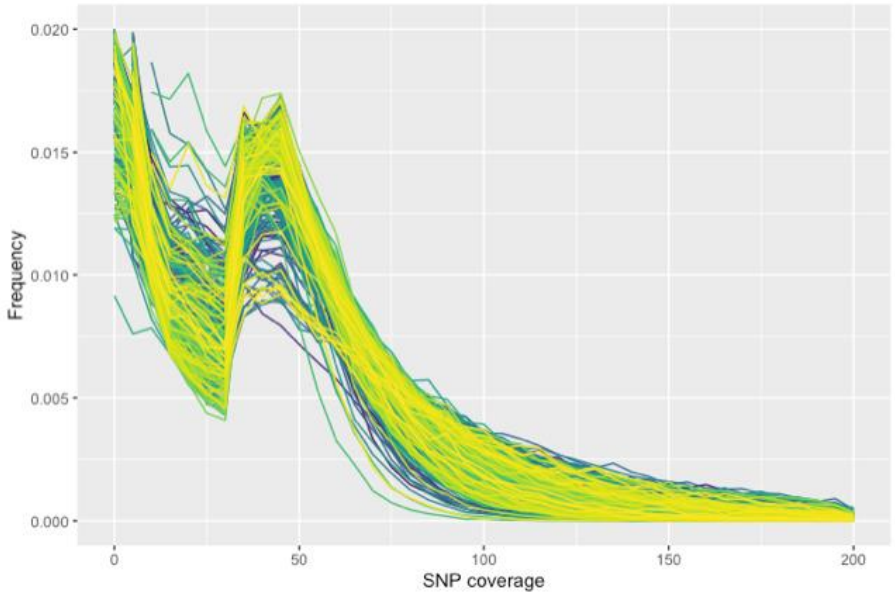
Data analysis

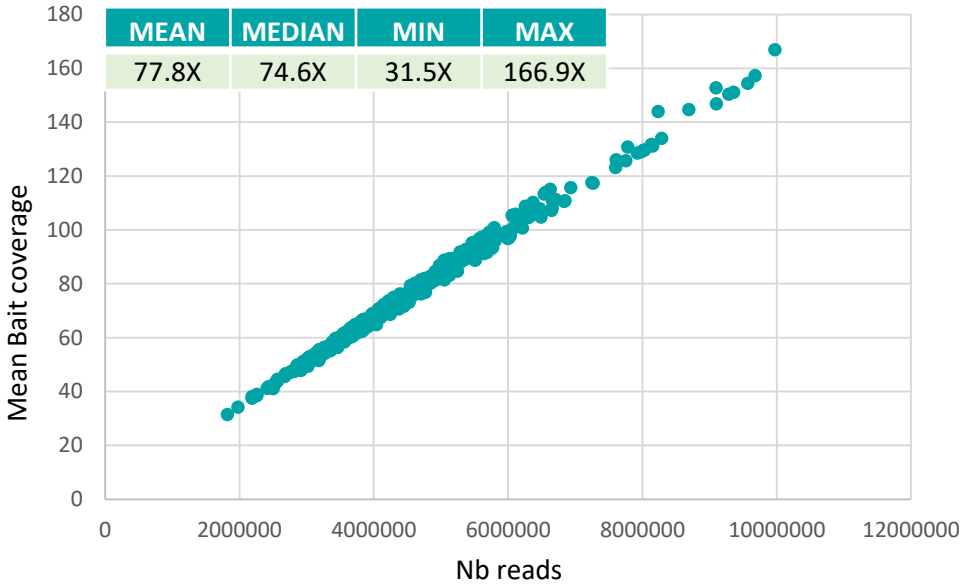


25,000 target SNPs coverage

- 25,000 target SNPs coverage : 50 - 100 X

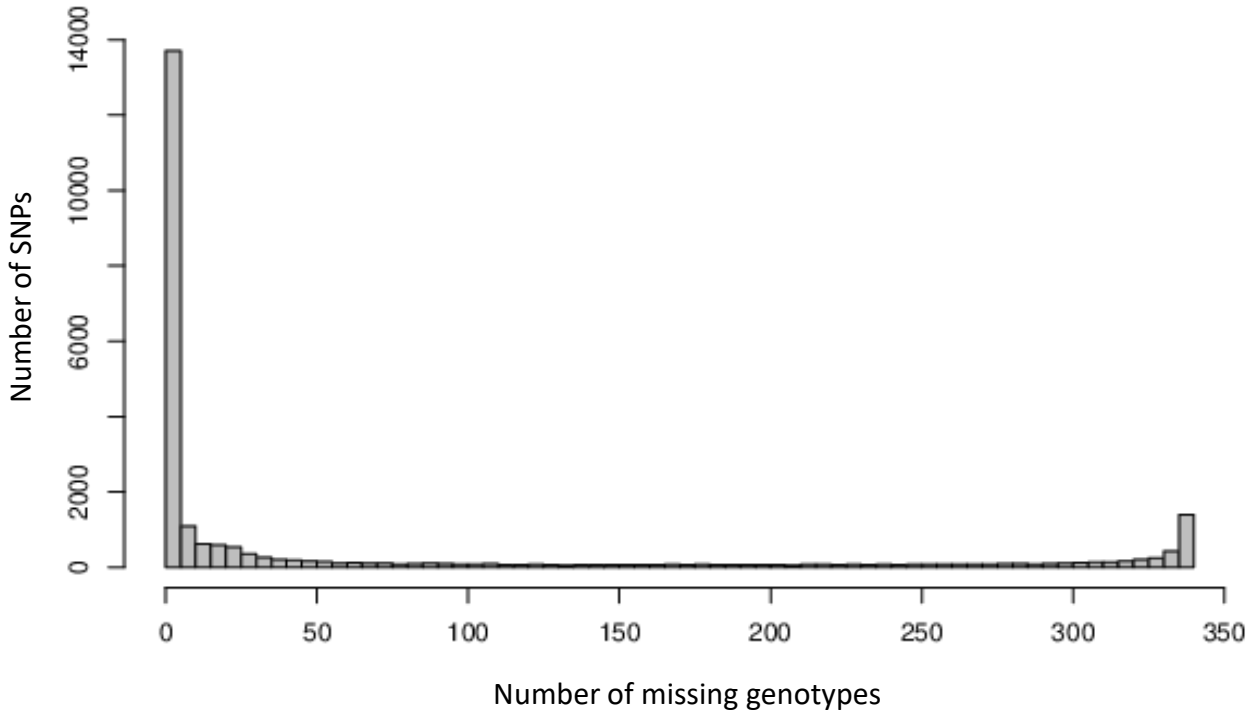
- >96% reads aligned, and 77% of those reads mapped uniquely on a target region





Target SNPs coverage in line with expectations
Strong correlation between coverage and sequencing depth

Genotypes description from 24,443 biallelic target SNPs after filtering



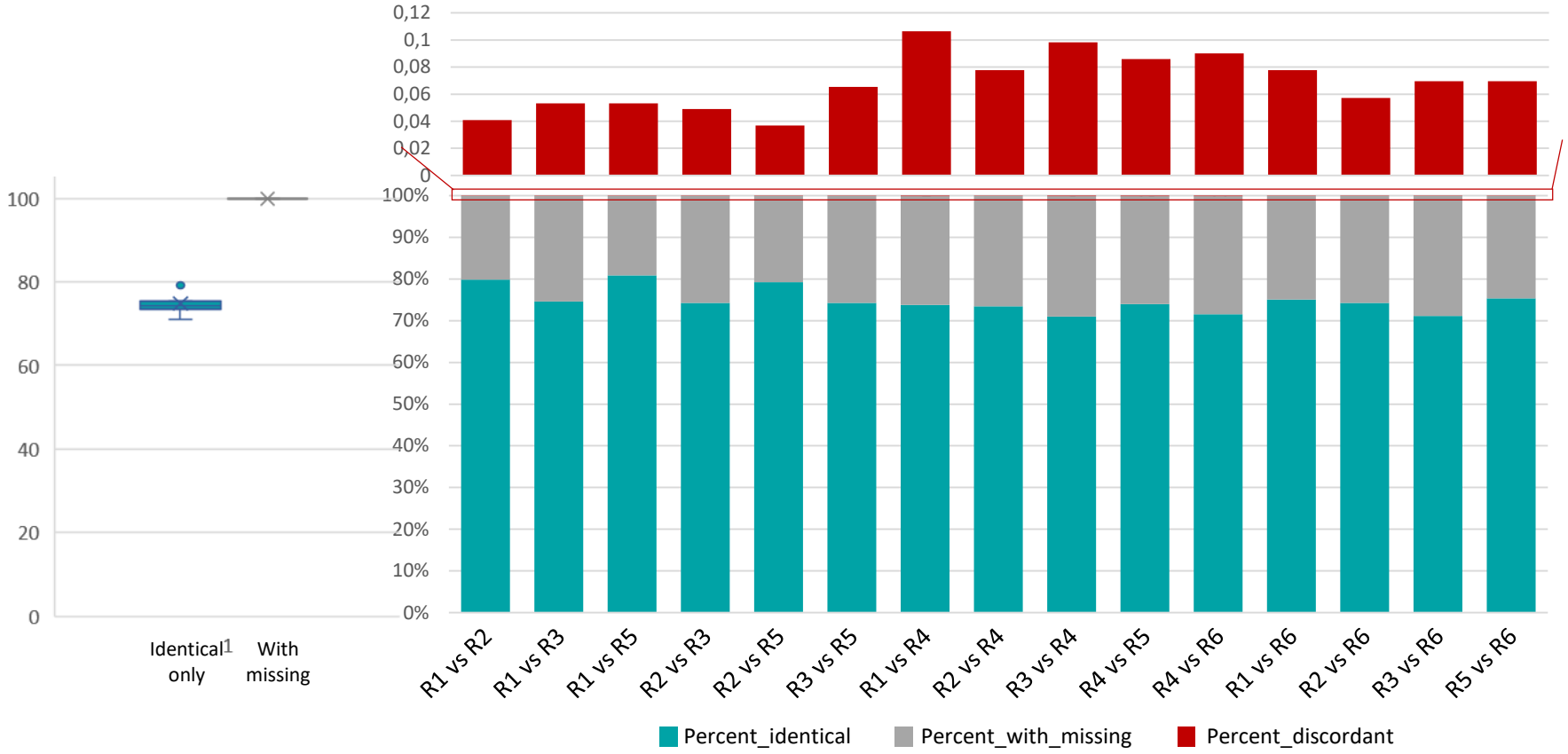
23,236 genotyped SNPs (**95 %**) :

- 10,262 with no missing data (**42 %**)
- 17,068 SNPs with < 10% missing genotypes ./. (69.8 %)

1,207 SNPs without genotype (4.8 % = FAILURES)

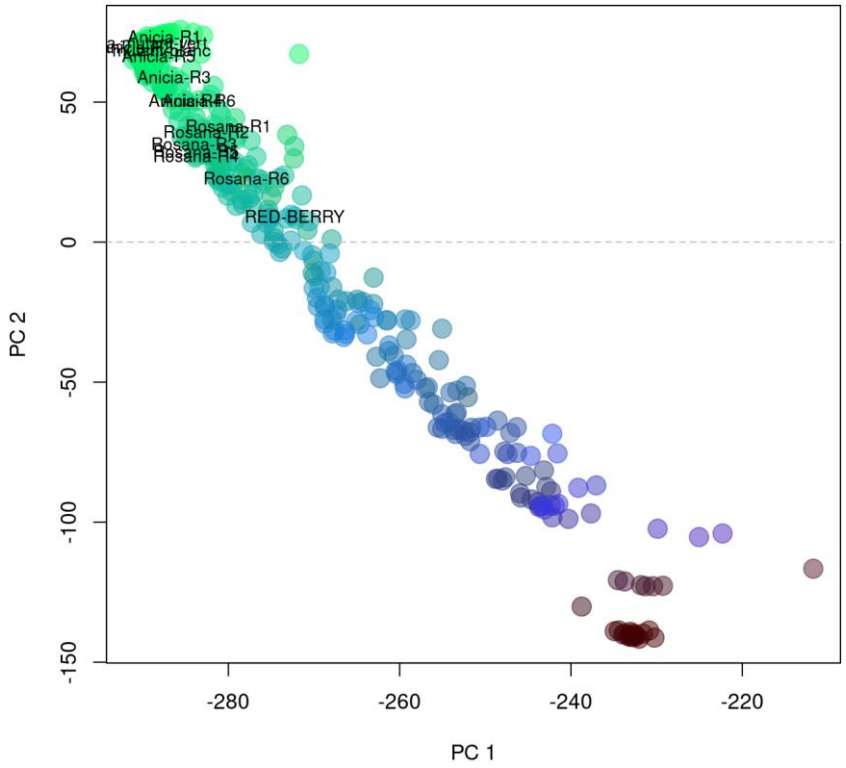
Overall, few missing data

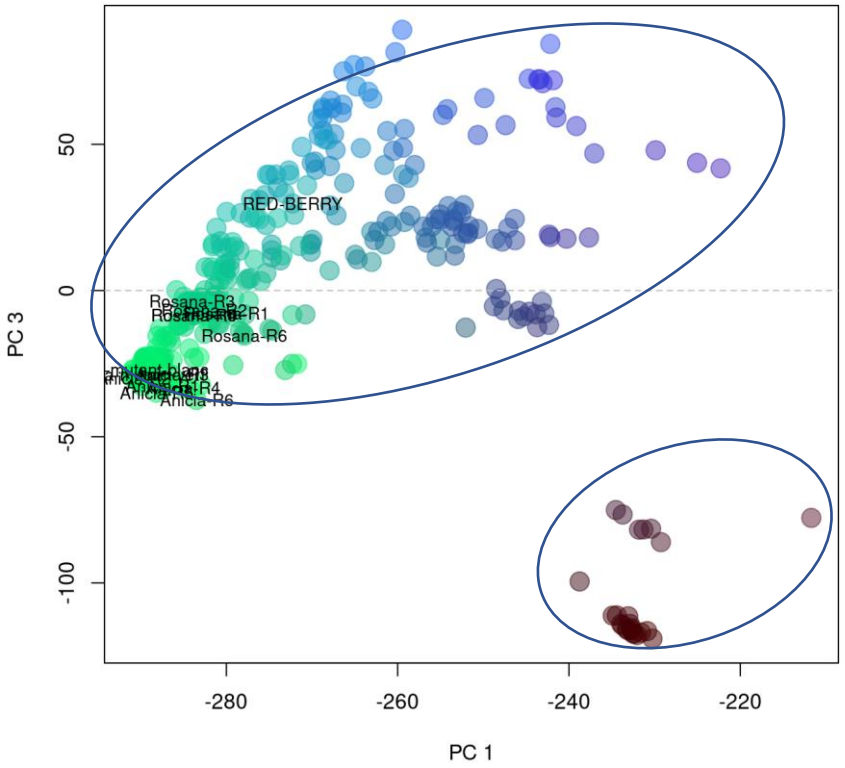
Biological reproducibility :
same variety, different individuals from the same seeds lot



High reproducibility (>99,9% R1, R2, R3 and R5 ; >99% across the 6 biological replicates)
Missing data ~ 13% per biological replicate (and sample)

Characterization of species A ressources





Biological replicates grouped →Accurate and reproducible approach

2 main groups: mainly wild individuals vs cultivated individuals

Continous diversity in cultivated individuals

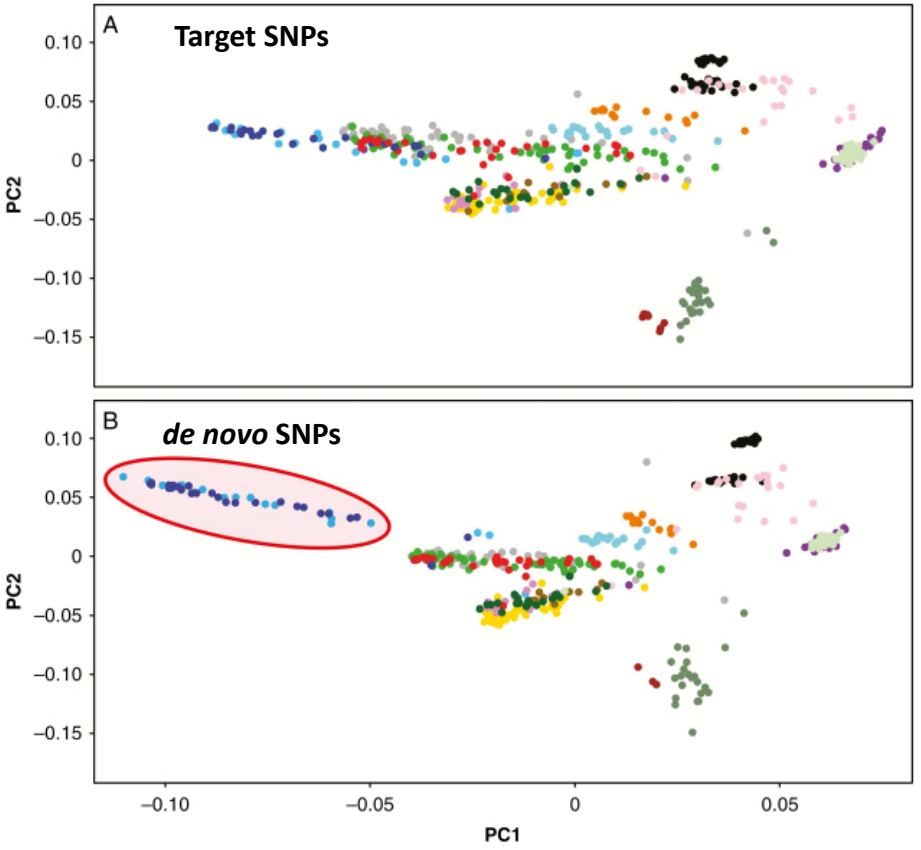
de novo SNPs discovery by read R2
more distant SNPs compared to R1

- Allow detection of finer-scale differenciation

- Allow stronger clustering (e.g. poplar, Scaglione *et al*. 2019 - 1 probe / SNP)

	Expected	Observed
Average target region cov.	100 X	98,85 X
Standard filters * :		
- Target SNPs	98,134	66,922 (68.2%)
- *de novo* SNPs	N.A.	453,170
Stringent filters ** :		
- Target SNPs	98,134	51,943 (52.9%)
- De novo SNPs	N.A.	203,478
Failed SNPs	N.A.	Approx. 10 %

* QD < 2.0 || MQ < 40.0 || MQRankSum < –12.5
** QD < 2.0 || MQ < 40.0 || MQRankSum < –12.5 & <8 reads ; <75% ind. called



Work in progress for Species A…

From Scaglione et al. *2019*

What have been done in plants ?

Species (sequencing)	Nb ind.	Nb Target SNPs (2 probes)	Origin of SNPs	Obs. target SNPs (2 probes)	Failure rate %	Accuracy %	Raw *de novo* SNPs	filtered *de novo* SNPs	Reference
Tests and benchmarking									
Tomato (SE150bp)	400	5,000 (0)	Reseq.	4,577 (0)	1.2-10,4*	99.8	N.A.	7,425	Barchi et al. (2019)
Eggplant (SE 150bp)	422	5,082 (0)	Reseq.	4,628 (0)	0.02-4,5*	99.9	N.A.	26,103	Barchi et al. (2019)
Maize (SE 150bp)	10	46,358 (24,843)	chip	27,236 (14,474)	24,3 **	97.2 (50X) 95.9 (30X)	N.A.	N.A.	Scaglione et al. (2019)
Poplar (PE 130bp)	540	98,134 (0)	chip	51,943 (0)	~ 10	N.A.	N.A.	203,478	Scaglione et al. (2019)
Application cases									
Endemic Solanaceae (SE 150bp)	74	5,093 (N.A.)	Reseq.	284 (raw) / 34 (filtered) (N.A.)	N.A.	100	6,220	1,387	Gramazio et al. (2020)
Oil palm (SE 150bp)	96	5,000 (0)	Reseq.	4,308	N.A.	N.A.	N.A.	8,765	Herrero et al. (2020)
Lentil (PE 150bp)	276	46,520 (19,103)	Transcipt.	N.A. ***	N.A.	N.A.	2,043,680	56,349	Dissanayake et al. (2021)

* missing data for cultivated individuals – missing data for wild individuals (failure rate not indicated)
** loci with <3 reads
*** authors chose to use only *de novo* SNPs

What have been done in plants ?

Species (sequencing)	Nb ind.	Nb Target SNPs (2 probes)	Origin of SNPs	Obs. target SNPs (2 probes)	Failure rate %	Accuracy %	Raw *de novo* SNPs	filtered *de novo* SNPs	Reference
Tests and benchmarking									
Tomato (SE150bp)	400	5,000 (0)	Reseq.	4,577 (0)	1.2-10,4*	99.8	N.A.	7,425	Barchi et al. (2019)
Eggplant (SE 150bp)	422	5,082 (0)	Reseq.	4,628 (0)	0.02-4,5*	99.9	N.A.	26,103	Barchi et al. (2019)
Maize (SE 150bp)	10	46,358 (24,843)	chip	27,236 (14,474)	24,3 **	97.2 (50X) 95.9 (30X)	N.A.	N.A.	Scaglione et al. (2019)
Poplar (PE 130bp)	540	98,134 (0)	chip	51,943 (0)	~ 10	N.A.	N.A.	203,478	Scaglione et al. (2019)
Application cases									
Endemic Solanaceae (SE 150bp)	74	5,093 (N.A.)	Reseq.	284 (raw) / 34 (filtered) (N.A.)	N.A.	100	6,220	1,387	Gramazio et al. (2020)
Oil palm (SE 150bp)	96	5,000 (0)	Reseq.	4,308	N.A.	N.A.	N.A.	8,765	Herrero et al. (2020)
Lentil (PE 150bp)	276	46,520 (19,103)	Transcript.	N.A. ***	N.A.	N.A.	2,043,680	56,349	Dissanayake et al. (2021)

91.5%
x1.6
91.1%
x5.6
58.7%
52.9%
x3.9
0.6%
x40.8
86.0%
x2.0

* missing data for cultivated individuals – missing data for wild individuals (failure rate not indicated)
** loci with <3 reads
*** authors chose to use only *de novo* SNPs