ddRAD-seq para el análisis de asociación genómico de los principales caracteres agronómicos y calidad organoléptica en melocotonero

N. Ksouri[1], G. Sanchez[2], B. Contreras-Moreira[3*] and Y. Gogorcena[1*]

[1]Group of Genomics of Fruit Trees and Grapevine, Department of Pomology, EEAD-CSIC, Zaragoza, Spain.
[2]Biotechnology Lab, Estación Experimental Agropecuaria (EEA) San Pedro, INTA, San Pedro (2930), Argentina.
[3]Computational and Structural Biology Group, Department of Genetics and Plant Production, EEAD-CSIC, Zaragoza, Spain.
*Corresponding authors: aoiz@eead.csic.es and bcontreras@eead.csic.es

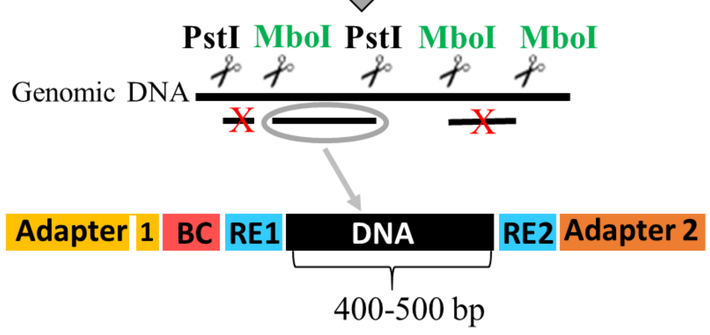
1. Introduction

- Double-digest RAD sequencing (ddRAD-Seq) is a powerful method for SNP discovery at a genome-wide scale. It relies on breaking the genome into a certain size of DNA fragments using two restriction enzymes: i) one common cutter with a short recognition site and ii) a low frequency cutter having a large recognition motif (See workflow (a) in MM section). Amplification step in library preparation can introduce PCR artefacts. Those are expected to skew allele frequency by increasing homozygosity leading to false genotype calls. In this context, we carried out a comparison study between DNA-variants generated with duplicates and those generated after removing with either "SAMtools:rmdup" or "Stacks clone_filter" (See workflow (b) in MM section).
- On the other hand, the accuracy of genetic variants identification is a crucial step towards understanding phenotypical traits and monitoring breeding programs. Thereby, a good combination of computational tools for alignment and variant calling is crucial to tackle the possible artifacts. In response to this challenge, three variant callers (BCFtools, Freebayes and GATK-HaplotypeCaller) were combined on top of the BWA-mem read mapper. SNPs derived from the intersection of these callers were used for a genome wide association study to identify genetic variants associated with agronomic and organoleptic traits (In this poster, we illustrate only one study case trait). A diverse set of 90 peach accessions were sampled from the Experimental Station of Aula Dei (CSIC) located at Zaragoza (northern of Spain). This germplasm collection includes landraces and modern breeding lines from different origins.

2. Materials and Methods

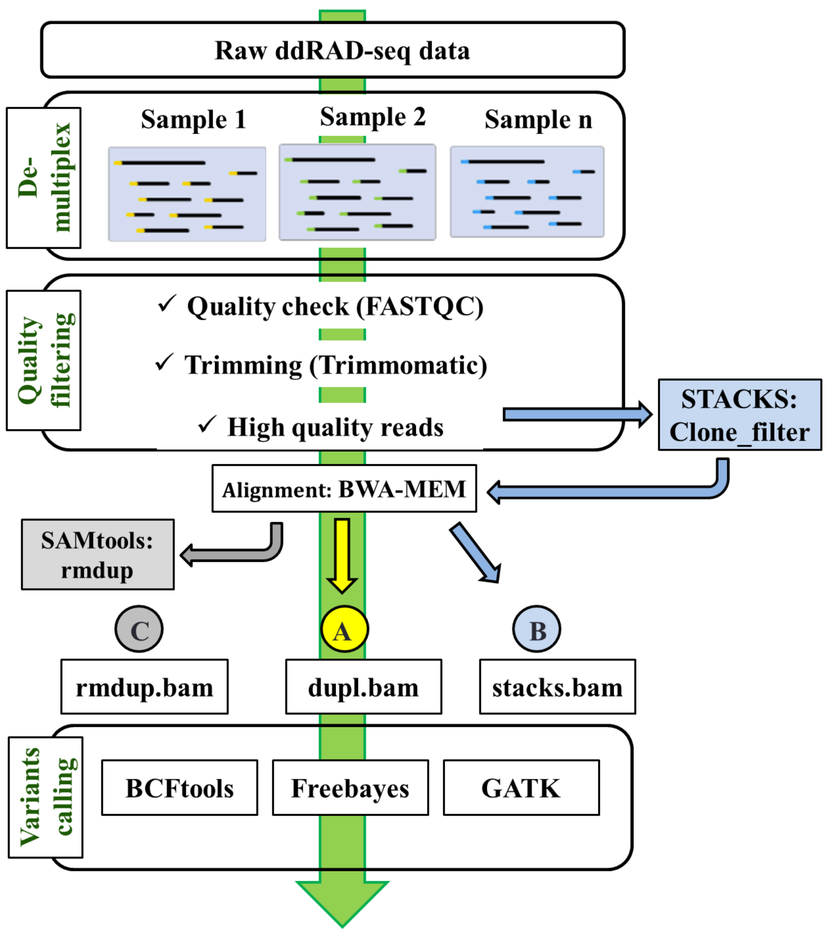
a. In vivo

- ddRAD libraries preparation at IABiMo INTA-CONICET, Argentina. (peach samples from different origin).
- Digestion: PstI (rare), MboI (frequent)
- 400-500 bp size selection
- Individual barcodes (BC)
- Pools formation
- Indexed adapters + 2 PCRs

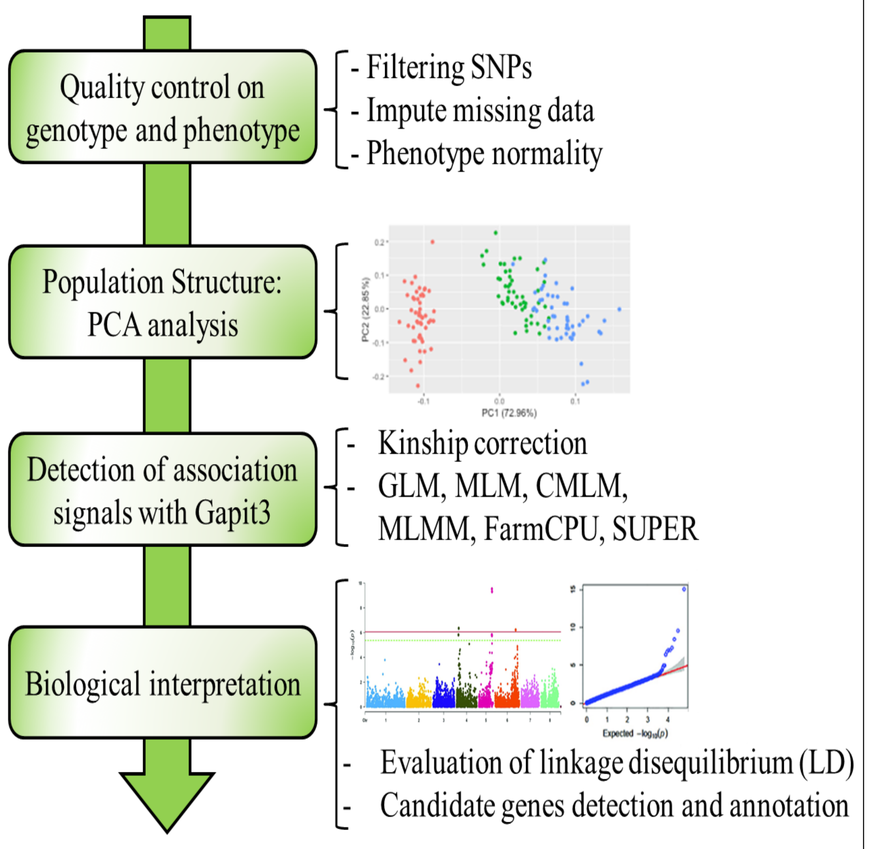


- Paired-end sequencing NovaSeq 6000

b. In silico

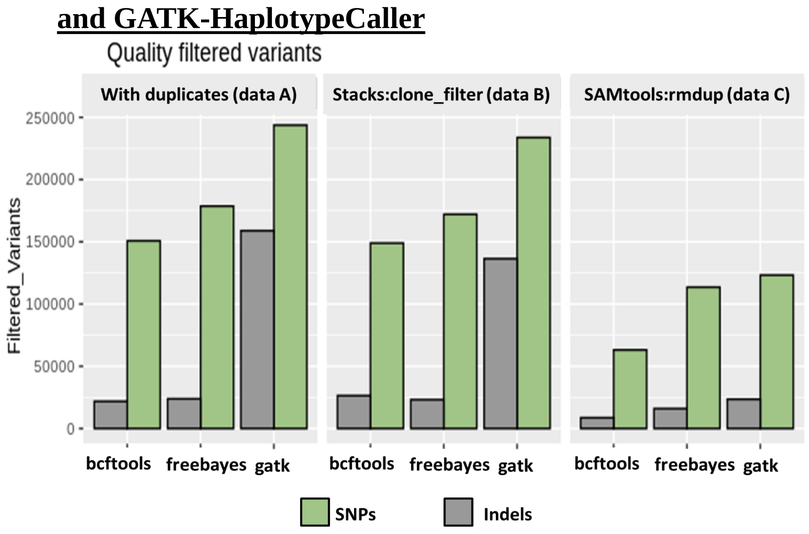


c. GWAS essential modules



3. Results and discussion

I. Robustness of three variant callers: BCFtools, Freebayes and GATK-HaplotypeCaller



✓After discarding false positive calls, we observed that GATK revealed the highest number of both SNPs and indels across all data types.

II. SNPs intersection from three variant callers

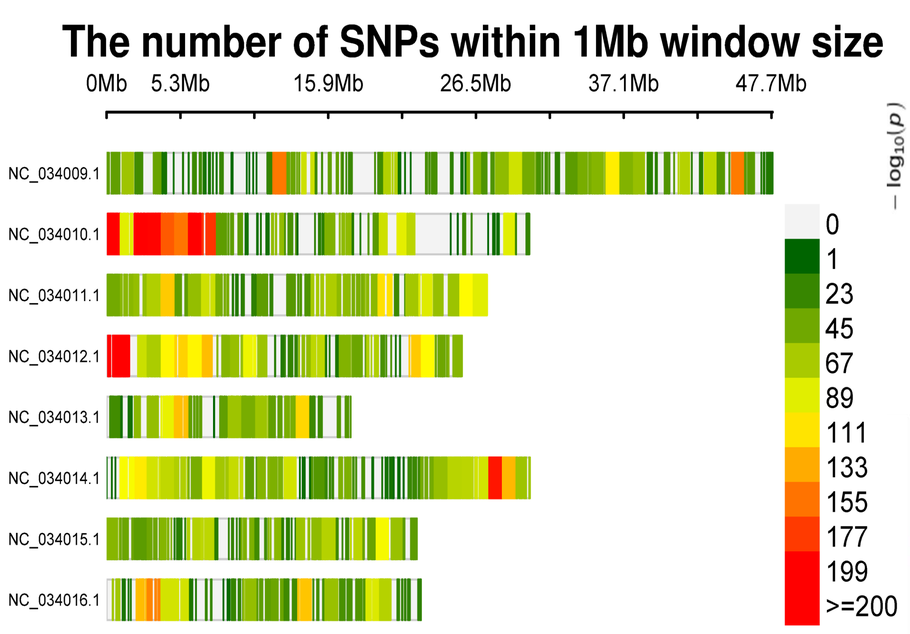
Data_type	Intersected SNPs	Homo	Hetero
A. With duplicates	90,660	62,815	65,716
B. Stacks:clone_filter	88,576	63,395	75,422
C. SAMtools:rmdup	42,519	28,949	31,940

✓SNPs number was reduced to the half when duplicates removal was conducted with SAMtools:rmdup.

✓93% of SNPs were shared between data A and B with 7% uniquely identified when duplicated reads were kept. Inspection of these calls revealed that they were initially detected as raw SNPs in Stacks de-duplicated data set. However, they were filtered out as they did not pass the filter of minimum depth of reads (DP ≥ 5).

✓→ **PCR duplicates artificially inflate the read depth**.

III. SNP density plot across the 8 chromosomes of peach



V. Conclusions

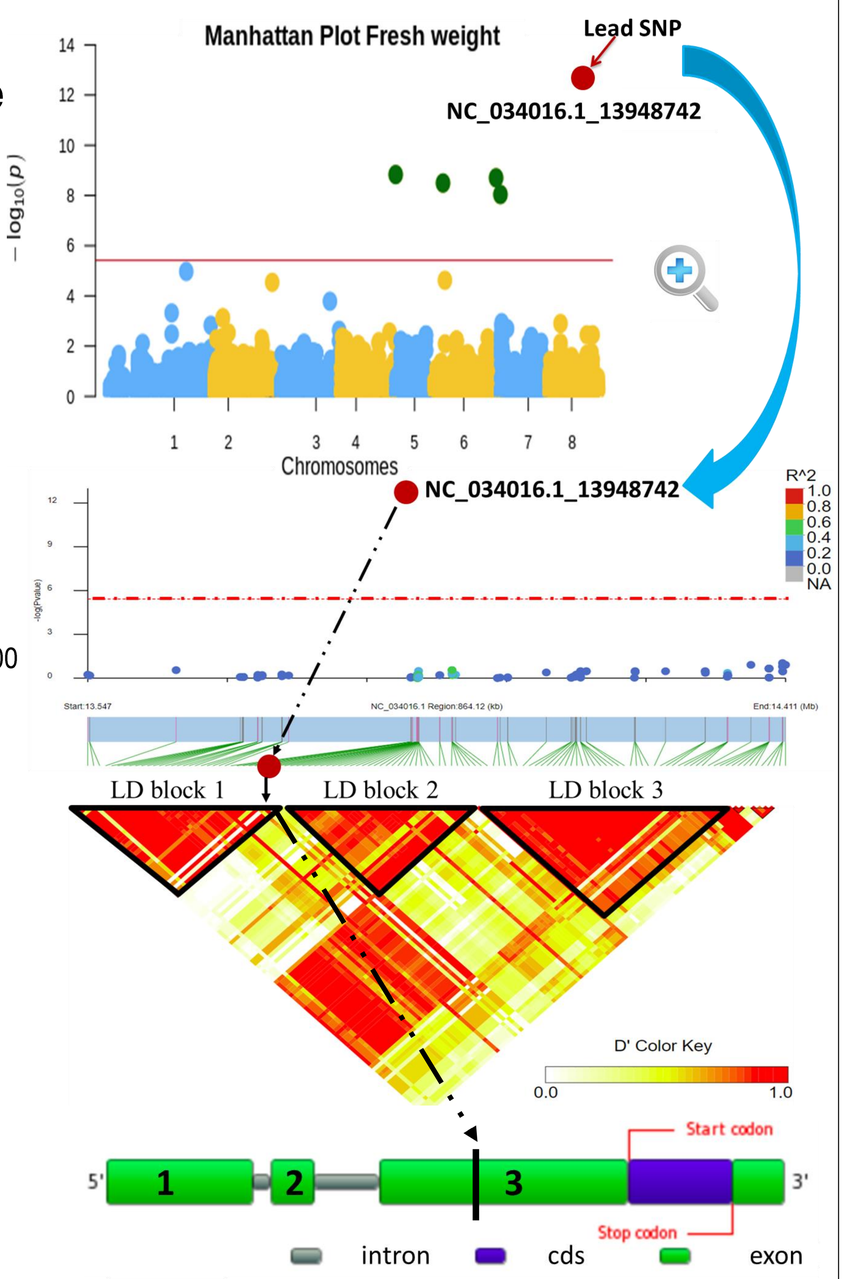
✓A Manhattan plot, illustrates the association statistical significance as –log10(*p*-value) in the y-axis against chromosomes in the x-axis.

✓**NC_034016.1_13948742** is considered as the **lead SNP** as it has the smallest *p*-value. It falls in the LD block 1 and **Prupe.8G109900** is the target gene of this SNP.

✓**NC_034016.1_13948742** is found to be located in the gene exon, more precisely in exon 3.

✓Newly identified SNPs may provide new opportunities for *Prunus persica* genome research, including informed genomics assisted breeding with high density SNP markers

IV. Manhattan plot of genome wide association of a case study trait



Prupe.8G109900: Leucine-rich repeat protein

Acknowledgements: This work was funded by the Spanish Ministry of Economy and the Government of Aragón with grants AGL2014-52063-R, AGL2017-83358-R, 2020AEP119, A09_20R and A08_20R which were co-financed with FEDER funds. N.K. was funded with a pre-doctoral contract awarded by the Government of Aragón.