**MATERIALS AND METHODS**

**Plant material**

The pea core collection used in this study consisted of 325 accessions that have been carefully selected from a large *Pisum* spp. collection of >3000 accession that were initially provided by USDA (Department of Agriculture, USA), JIC (John Innes Center, UK), CRF INIA (Centro nacional de Recursos Fitogenéticos, Spain), CGN (CPRO-DLO, The Netherland), IPK (Leibniz Institute of Plant Genetics and Crop Plant Research, Germany) and ICARDA (International Center for Agricultural Research in the Dry Areas, Syria). The collection is representative of the different *Pisum* species and subspecies including accessions from *P. sativum* L.*, P. fulvum* Sm.*, P. abyssinicum* A. Braun and the subspecies *sativum, arvense, elatius, pumilio, syriacum, cinereum, jomardii, thebaicum* and *transcaucasicum* of *P. sativum.* In addition, it comprises cultivated, wild and landraces of worldwide origin with large genetic and morphologic diversity (Fig 1. and Supl. Table 1).

**DNA extraction, library construction, and sequencing**

Pea core collection was genotyped with the DArTSeq approach by DiversityArray Ltd (Australia). For this, third composed leaves from 20 two weeks-old seedlings of each accessions grown under controlled condition were harvested, pooled together, flash frozen in liquid nitrogen and lyophilized. Then, DNA was extracted following to the method stipulated by Diversity Arrays P/L, Canberra, Australia as described previously (Barilli et al., 2018). DNA quality was assessed by electrophoresis on 0.8% on agarose gel and quantified by fluorescence at XXX nm on a microtiter plate reader (XXX) with Quantifluor® DsDNA system following manufacturer recommendations (Promega Corporation, Madison, USA). The DNA samples were then adjusted at 20 ng/µl prior to DArT marker analysis using the high density Pea DArTseq 1.0 array (50,000 markers) adapted for wild *Pisum spp.* accessions as described previously (Barilli et al., 2018). Complexity reduction with the *Pst*I,-*Mse*I restriction enzymes, library construction, amplification and Illumina sequencing were performed by Diversity Arrays Technology Pty Ltd (Canberra, Australia) as described in Barilli et al (2015). DArTSeq sequence analysis retrieve two set of markers, SNPs and presence–absence sequence variants (Silico-DArT), collectively referred to as DArT-Seq markers. Data cleaning was then performed for both type of DArT markers to remove low quality and non-polymorphic markers as described before (Rispail et al., 2018, Montilla-Bascón et al., 2015). Accordingly, DArT markers with > 20% missing data, minor allele frequency (MAF) < 5% and heterozygosity > 0.1% were removed from the analysis. In parallel, DArT markers were mapped onto the *Pisum* reference genome sequences (Keplack et al 2019) by Blast (threshold parameter: E-value= 5e-04 y min % identify=80%). Distribution of mapped markers to each chromosome was performed with the LinkageMapView package in R (Ouellette et al., 2017)

**Population structure of the pea core collection**

Population structure of the Pisum core collection was inferred with the Silico-DArT dataset after filtering markers in linkage disequilibrium (LD) with PLINK v1.9 (Chang et at. 2015). LD filtering was performed with the pruning method with a window size of 200 markers and a r2 threshold of 0.1 leading to a total of 4,000 Silico-DArT markers of which 2880 were homogeneously distributed onto the 7 pea chromosomes (Suppl. Table 2) and 583 mapped to unanchored contigs. Upon LD filtering, population structure was established with STRUCTURE 2.3.4 (Prichard XXX) using the admixture model with correlated allele frequencies between populations, which was shown as the optimum model for subtle population structure (Falush et al. 2003). Ten independent simulations were performed for each k from k= 1 to k=15. Each simulation consisted of 10,000 burn-in and 20,000 iterations. Longer burn-in or MCMC did not change significantly the results. The optimal number of k and the percentages of admixture of each accession (Q-matrix) were then given by STRUCTURE HARVESTER (Earl and vonHoldt, 2012) according to the Δk method (Evanno et al., 2005). For subsequent analysis an accession was assigned to a subpopulation when it had more than 60% membership to this subpopulation. Visualization of STUCTURE Q matrix was done with the online software STUCTUREPLOT (Ramasamy et al., 2014). Principal component analysis (PCAs) was also performed with the full Pisum Silico-DArT dataset to infer the structure of the Pisum collection. PCA were estimated and drawn in R (R Development Core Team, 2020) with the ggfortify (Tang et al., 2016) and ggplot2 packages (Wickham, 2016) under the RStudio (RStudio Team, 2020).

**Phylogenetic relationship of the pea core collection.**

The phylogenetic relationship of the 325 pea accessions was inferred using the MEGA X.2.4 software (Kumar et al., 2018) with the full Silico-DArT dataset. For this, a p-distance matrix (Nei and Kumar, 2000) was estimated from the Silico-DArT matrix after pairwise deletion of gap using the HKY substitution model (Hasegawa et al., 1985) with gamma distribution which was the optimal substitution model according to the BIC criterion as estimated with MEGA X software. Pairwise gap deletion remove all site with more than 5% alignment gaps, missing data, and ambiguous bases from the marker matrix leading to a total of 11,635 polymorphic site. Then, a phylogenetic tree was reconstructed with the NJ Neighbor-Joining method (Satou and Nei, 1987) with 1000 bootstrapping replicates based on the p-distance matrix. Upon phylogenetic tree reconstruction, the tree was edited with the MEGA tree editor to colour each branch according to the STRUCTURE subpopulation each accession belonged to.

**Linkage disequilibrium**

The disequilibrium matrix summarising pair wise measures of LD was estimated with the SilicoDArT dataset with TASSEL 5 software (REF) with a sliding window of 100 markers. LD test was performed for all intrachromosomal marker pairs. To investigate the average LD decay (LD50) in the whole genome and per chromosome, significant intra-chromosomal r2 values were plotted against the physical distance (Mbp) between markers with R using the function LDit developed by Ross-Ibarra group (Davis University, USA; https://github.com/rossibarra/r\_buffet/blob/master/LDit.r). Average LD50 was then estimated in R as described in Marioni et al. (2011).