nature genetics

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41588-019-0480-1

In the format provided by the authors and unedited.

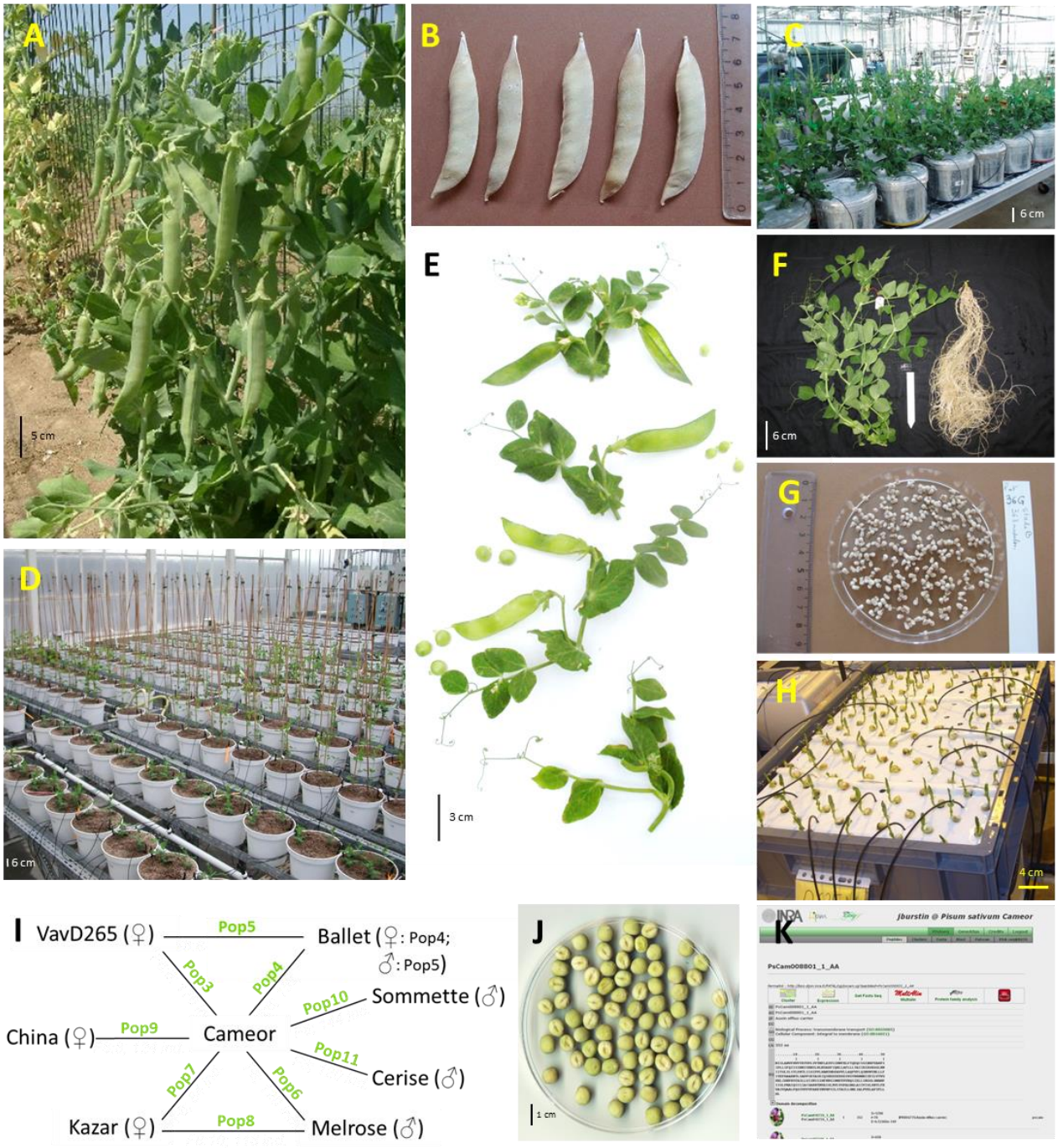
A reference genome for pea provides insight into legume genome evolution

Jonathan Kreplak[1,20], Mohammed-Amin Madoui[2,20], Petr Cápal[3], Petr Novák[4], Karine Labadie[5], Grégoire Aubert[1], Philipp E. Bayer[6], Krishna K. Gali[7], Robert A. Syme[8], Dorrie Main[9], Anthony Klein[1], Aurélie Bérard[10], Iva Vrbová[4], Cyril Fournier[1], Leo d'Agata[5], Caroline Belser[5], Wahiba Berrabah[5], Helena Toegelová[3], Zbyněk Milec[3], Jan Vrána[3], HueyTyng Lee[6,19], Ayité Kougbeadjo[1], Morgane Térézol[1], Cécile Huneau[11], Chala J. Turo[12], Nacer Mohellibi[13], Pavel Neumann[4], Matthieu Falque[14], Karine Gallardo[1], Rebecca McGee[15], Bunyamin Tar'an[7], Abdelhafid Bendahmane[16], Jean-Marc Aury[5], Jacqueline Batley[6], Marie-Christine Le Paslier[10], Noel Ellis[17], Thomas D. Warkentin[7], Clarice J. Coyne[15], Jérome Salse[11], David Edwards[6], Judith Lichtenzveig[18], Jiří Macas[4], Jaroslav Doležel[3], Patrick Wincker[2] and Judith Burstin[1]*

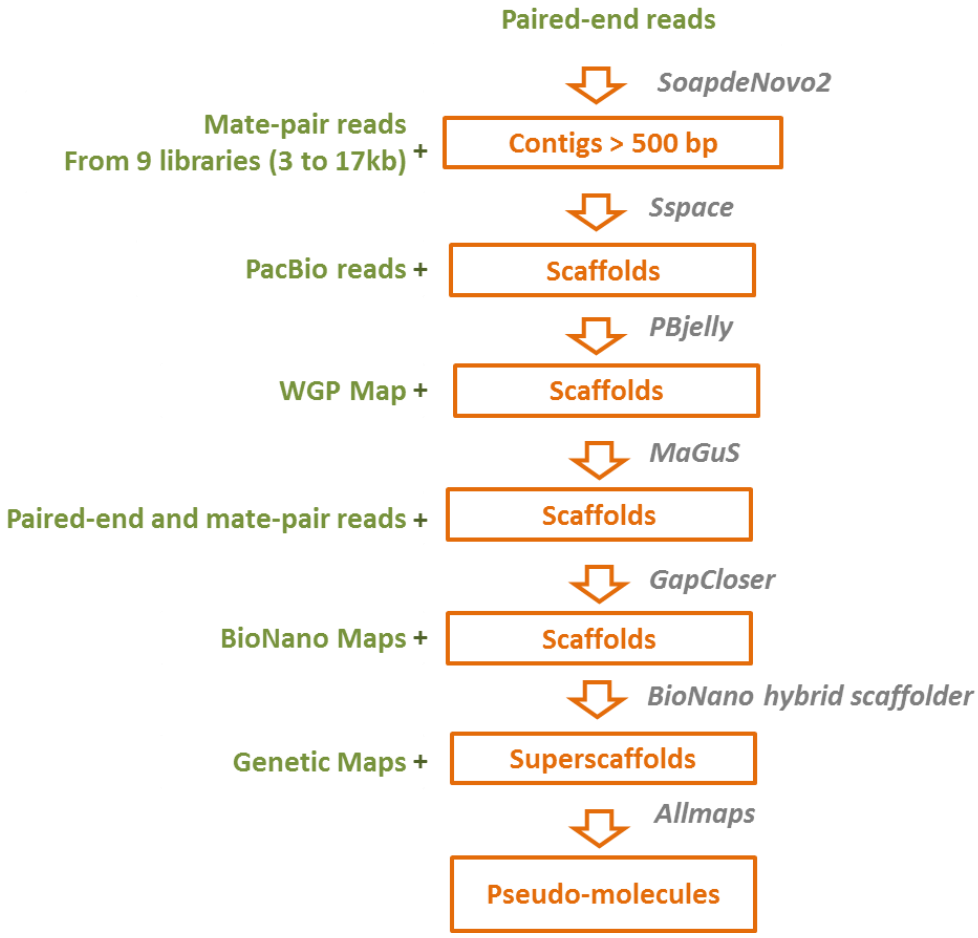
[1]Agroécologie, AgroSup Dijon, INRA, Université Bourgogne Franche-Comté Bourgogne, Université Bourgogne Franche-Comté, Dijon, France. [2]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Université Evry, Université Paris-Saclay, Evry, France. [3]Institute of Experimental Botany, Centre of the Region Haná for Biotechnological and Agricultural Research, Olomouc, Czech Republic. [4]Biology Centre, Czech Academy of Sciences, České Budějovice, Czech Republic. [5]Genoscope, Institut François Jacob, CEA, Université Paris-Saclay, Evry, France. [6]School of Biological Sciences and Institute of Agriculture, University of Western Australia, Perth, Western Australia, Australia. [7]Crop Development Centre/Department of Plant Sciences, University of Saskatchewan, Saskatoon, Saskatchewan, Canada. [8]Centre for Crop and Disease Management, Curtin University, Bentley, Western Australia, Australia. [9]Department of Horticulture, Washington State University, Pullman, WA, USA. [10]Etude du Polymorphisme des Génomes Végétaux, INRA, Université Paris-Saclay, Evry, France. [11]UMR 1095 Génétique, Diversité, Ecophysiologie des Céréales, INRA, Université Clermont Auvergne, Clermont-Ferrand, France. [12]Centre for Crop and Disease Management, School of Molecular and Life Science, Curtin University, Bentley, Western Australia, Australia. [13]URGI, INRA, Université Paris-Saclay, Versailles, France. [14]GQE–Le Moulon, INRA, University of Paris-Sud, CNRS, AgroParisTech, Université Paris-Saclay, Gif-sur-Yvette, France. [15]USDA Agricultural Research Service, Pullman, WA, USA. [16]Institute of Plant Sciences Paris-Saclay, INRA, CNRS, University of Paris-Sud, University of Evry, University Paris-Diderot, Sorbonne Paris-Cite, University of Paris-Saclay, Orsay, France. [17]School of Biological Sciences, University of Auckland, Auckland, New Zealand. [18]School of Agriculture and Environment, University of Western Australia, Perth, Western Australia, Australia. [19]Present address: Department of Plant Breeding, IFZ Research Centre for Biosystems, Land Use and Nutrition, Justus Liebig University, Giessen, Germany. [20]These authors contributed equally: Jonathan Kreplak, Mohammed-Amin Madoui. *e-mail: judith.burstin@inra.fr

A reference genome for pea provides insight into legume genome evolution

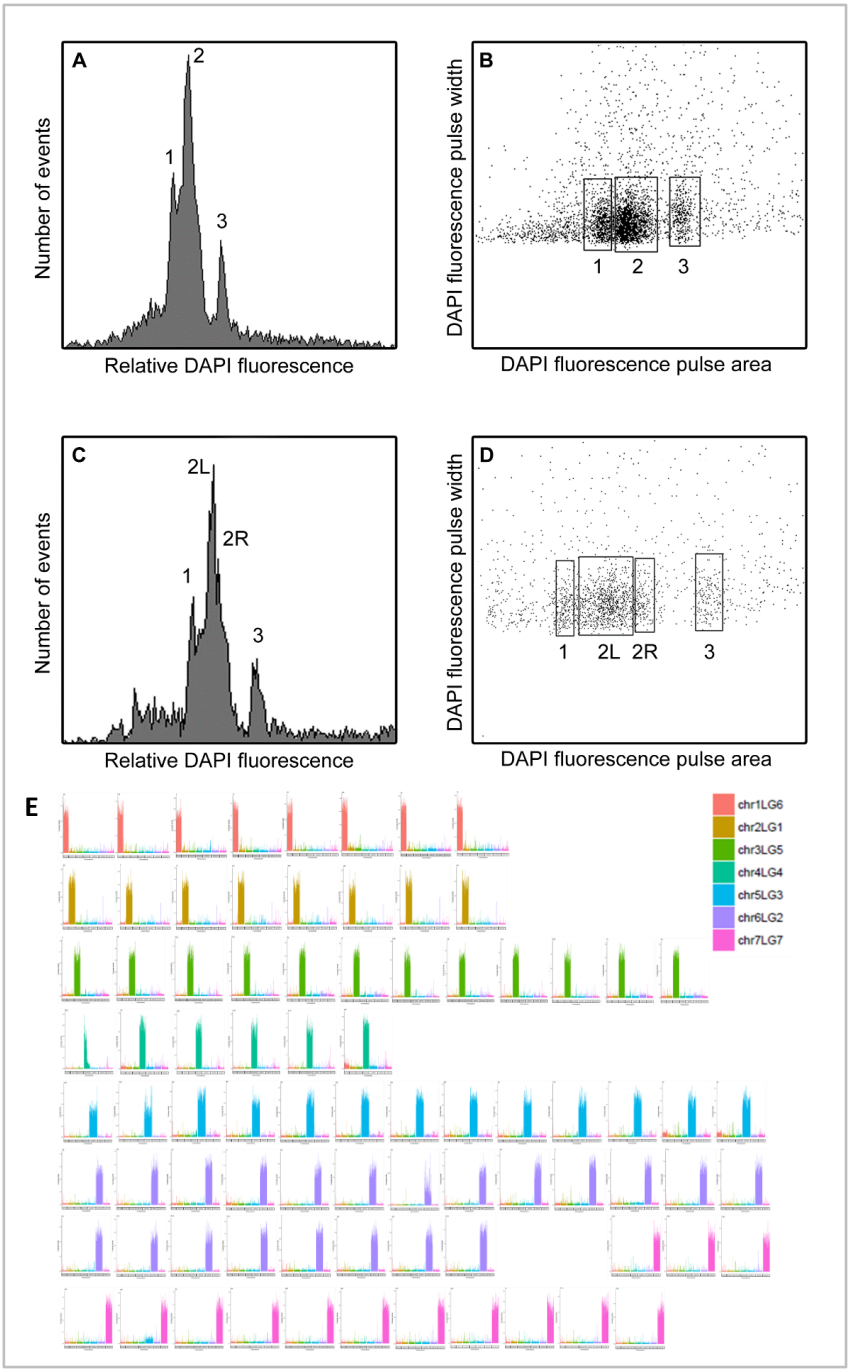
Supplementary Figures



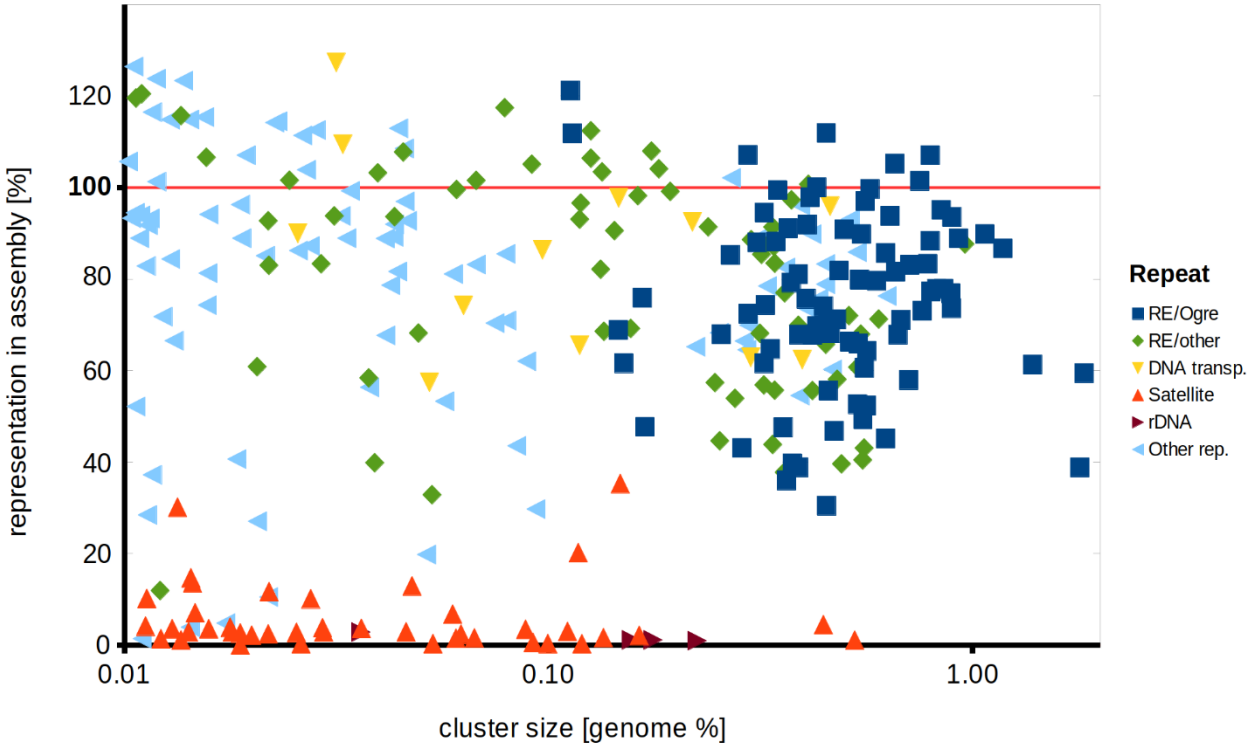
Supplementary Figure 1. Caméor, the pea reference genotype for genome sequencing. A. Trellised plants grown at the INRA experimental farm of Bretenière; B. Pods; C. Plants grown in hydroponic conditions at the INRA glasshouses; D. Caméor TILLING population grown in glasshouses; E. Flowering and podding nodes; F. Shoot and root plant parts; G. Nodules; H. Germinating seeds; I. Recombinant inbred lines populations derived from Caméor[2]; J. Seeds; K. Gene expression atlas.



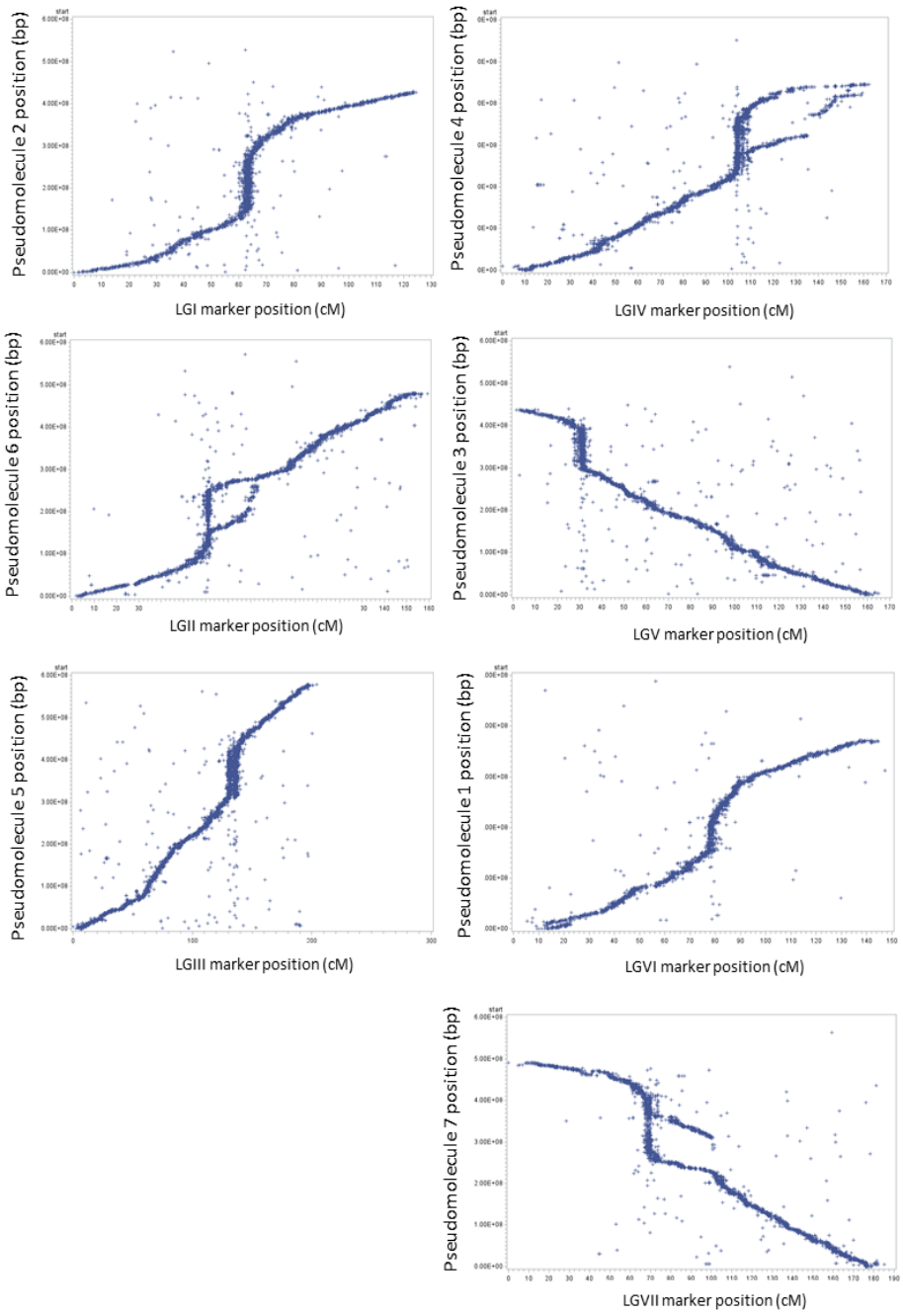
Supplementary Figure 2. Pea genome assembly pipeline



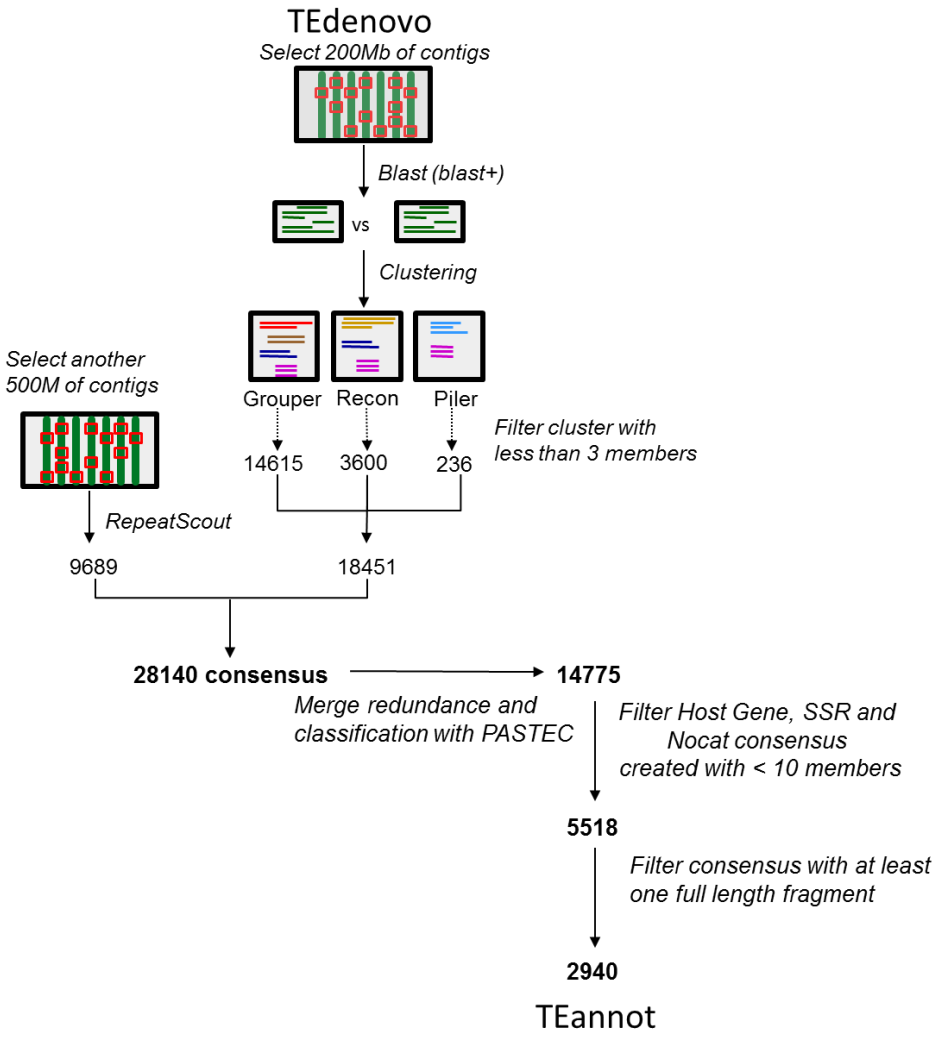
Supplementary Figure 3. Chromosome sorting by flow-cytometry. Monovariate (A, C) and bivariate (B, D) flow karyotypes obtained after the analysis of DAPI-stained suspensions of mitotic metaphase chromosomes of *P. sativum* cv. Caméor. The monovariate flow karyotype (A) comprises a partially resolved peak (marked 1), which represents chromosome 1, a second large composite peak (marked 2) representing chromosomes 2, 3, 4, 6 and 7, and a third well resolved peak (marked 3) representing chromosome 5. Higher resolution of chromosome peaks in the second experiment (C) lead to discrimination of a shoulder on composite peak 2 (marked 2R). The position of sort windows is indicated on bivariate flow karyotypes DAPI fluorescence pulse area vs. fluorescence pulse width (B, D). Mapping of flow-sorted 'Caméor' single chromosome amplified DNA resequencing reads onto pseudomolecules of genome assembly v1a (E).



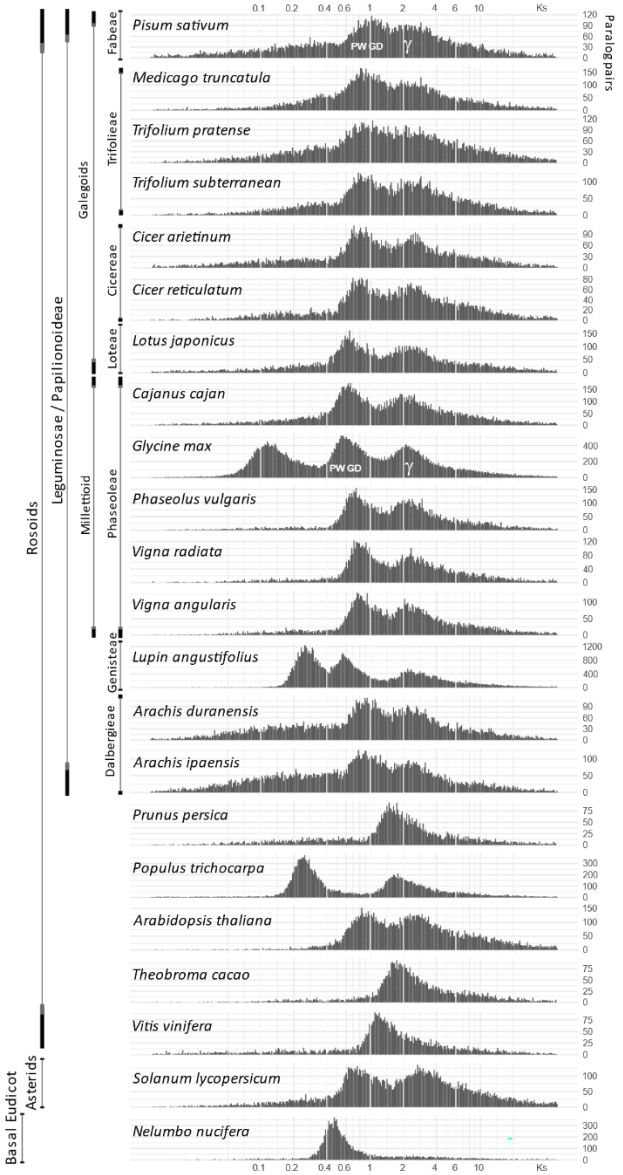
Supplementary Figure 4. Repetitive element representation in the assembly



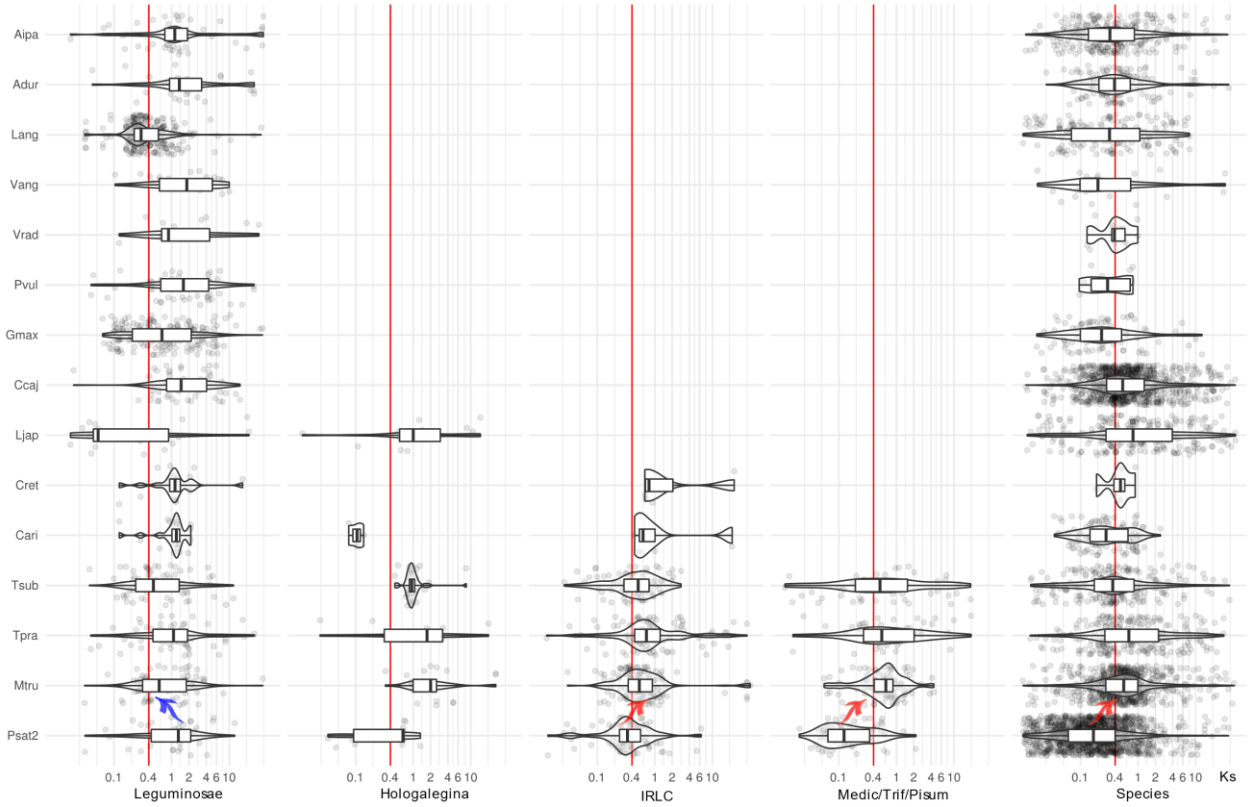
Supplementary Figure 5. Plots of genetic marker position on the genetic map of recombinant inbred line Pop6 ('Caméor'x'Melrose') versus on pseudomolecules showed low recombination regions in the pea genome.



Supplementary Figure 6. Repetitive DNA annotation pipeline

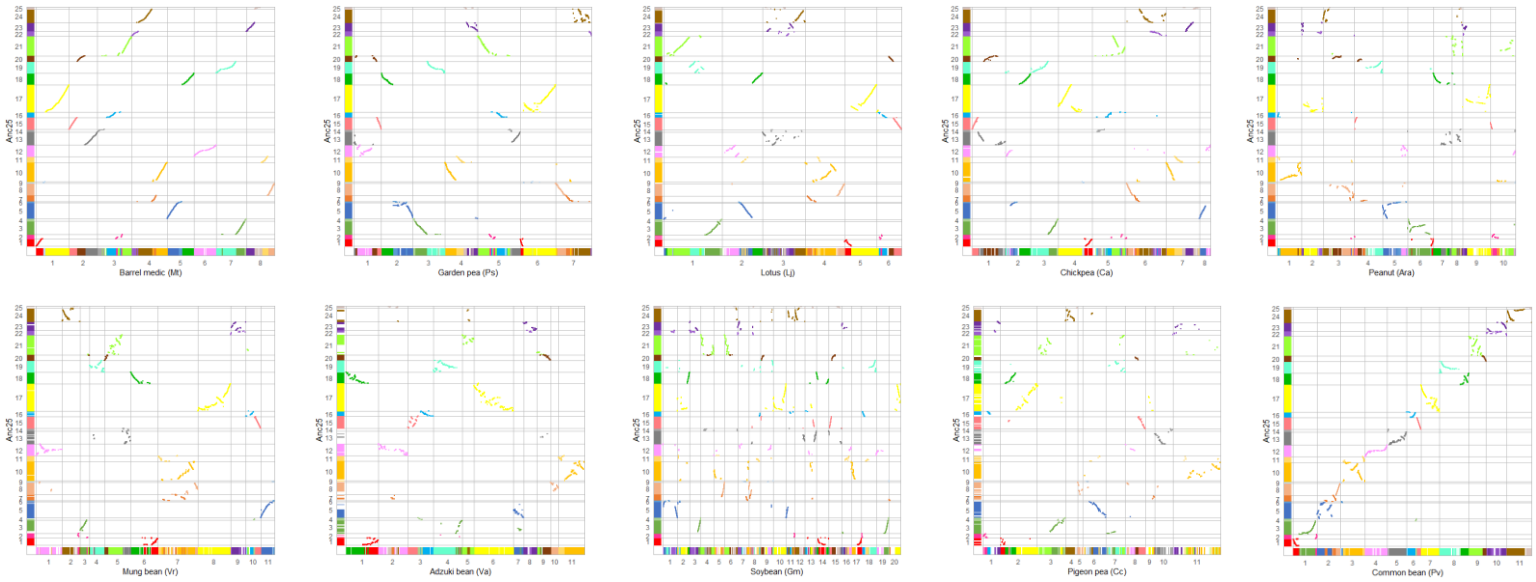


Supplementary Figure 7. Paleopolyploidy events and shift in mutation rates in the *Pisum* lineage. Distribution of paralog pairwise synonymous substitution per synonymous site (Ks). Ks distribution peaks resulting from whole genome triplication common to all core Eudicots is marked with γ (Bowen et al 2003), that resulting from the whole genome duplication common to the Papilionoideae is marked with PWGD for the histograms of pea and *G. max*, a genome often used as reference. Note x-axis is presented in a log-scale; Ks values are non-transformed values; Ks = 0.1, 0.4, 1, 2 and 6 are demarked with white lines for clarity.

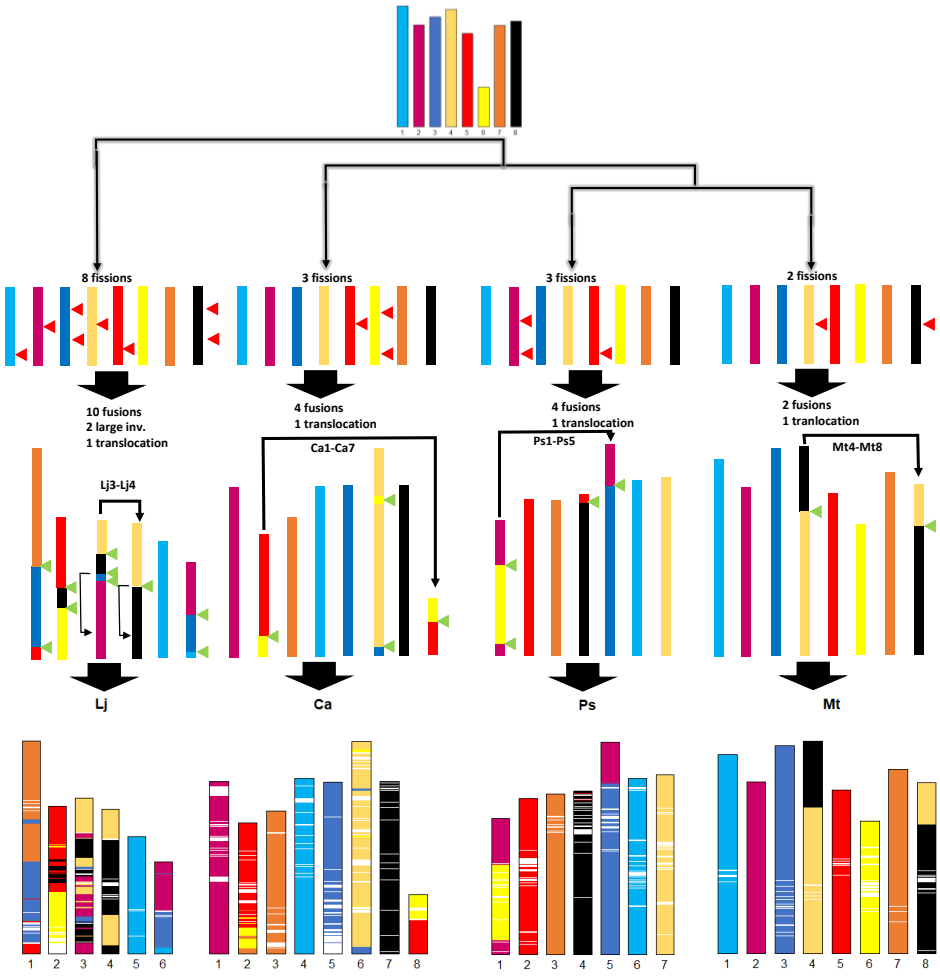


Supplementary Figure 8. Gene gain and gene duplication**.** Distribution of paralog pairwise synonymous substitutions per synonymous site (Ks) in paralog pairs classified as specific to (from left to right) the Papilionoideae subfamily, the Hologalegina cluster, the IRLC cluster, the Trifolieae/Fabeae tribes (MTP for *Medicago*, *Trifolium* and *Pisum*), or species-specific. Data density is denoted by violin and box plots. Data points are represented by a transparent grey circle. Note x-axis is presented in a log-scale; Ks values are non-transformed values; Ks = 0.4 is demarked with red for clarity.

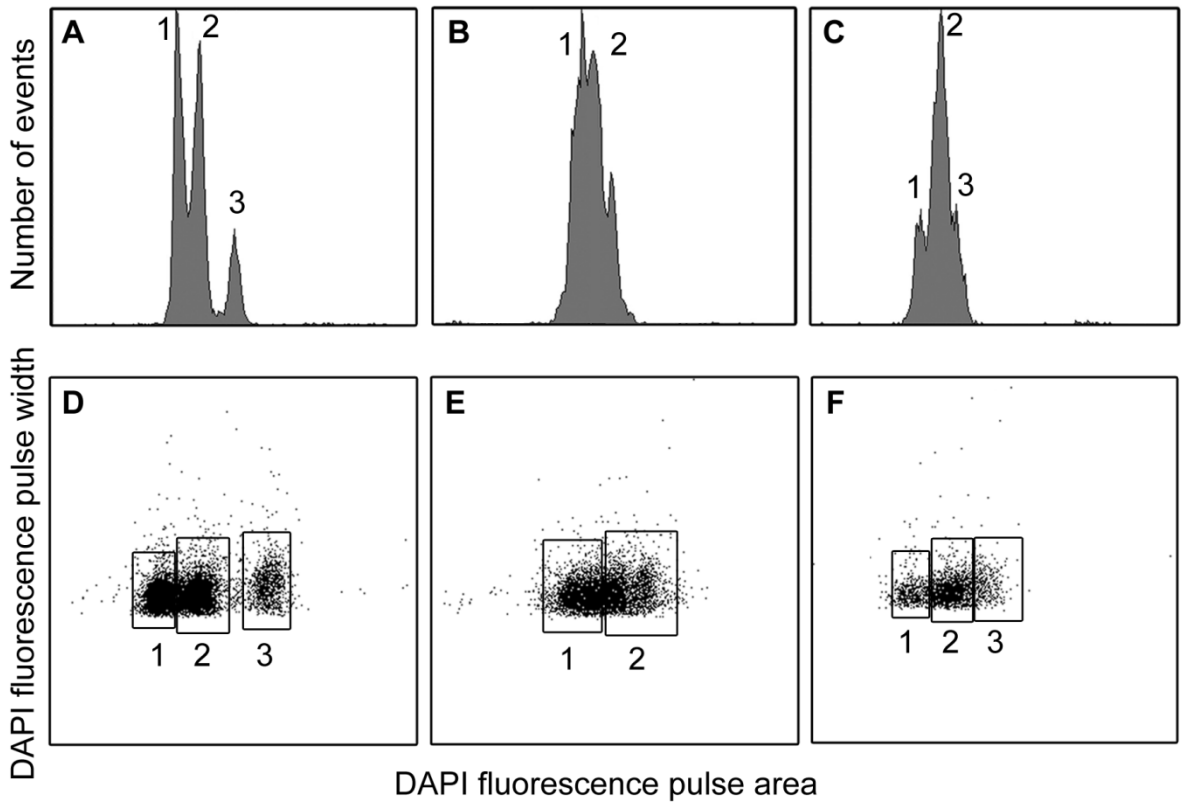
a



b

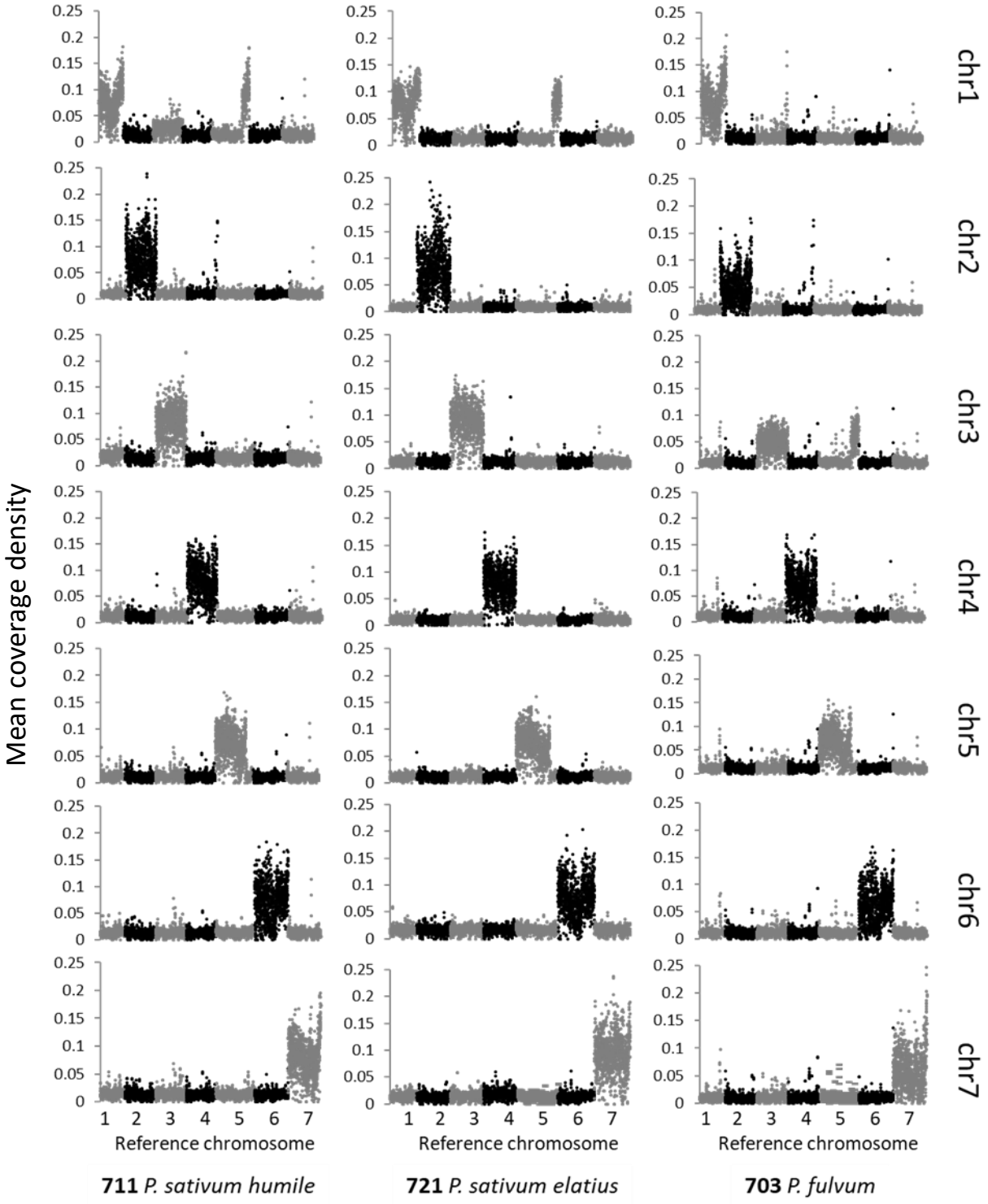


Supplementary Figure 9. Legume genome synteny. **a**. Dotplot-based deconvolution of the synteny relationships between ALK (y-axis) and *Phaseolus vulgaris (Pv), Cajanus cajan (Cc), Glycine max (Gm), Vigna angularis (Va), Medicago truncatula (Mt), Lotus japonicus (Lj), Cicer arietinum (Ca), Pisum sativum (Ps), Vigna radiata (Vr), Arachis, duranensis(Ara) (x-axis).* The chromosomes are depicted as a mosaic of a 25 color-code reflecting the 25 inferred CARs. The synteny relationships identified between the ancestral genomes and the modern species are illustrated with colored diagonals in the dotplot. **b.** Evolutionary scenario of the modern Galegoid genome (bottom) from the reconstructed AGK of 8 proto-chromosomes (top). Fusions, fissions, large inversions and translocation are illustrated on the scenario to reach the modern karyotype of (from left to right) *Lotus japonicus (Lj), Cicer arietinum (Ca), Pisum sativum (Ps), Medicago truncatula (Mt).*

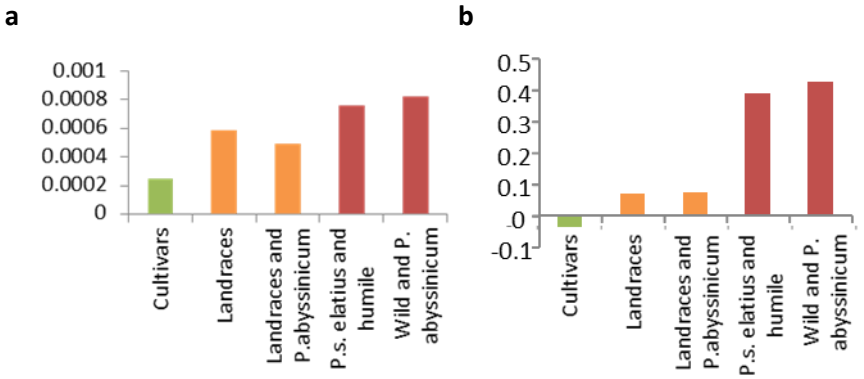


Supplementary Figure 10. Flow karyotypes of mitotic metaphase chromosomes of *P. fulvum*, *P. sativum southern humile* and *P. sativum elatius.* Monovariate (A - C) and bivariate (D - F) flow karyotypes obtained after the analysis of DAPI-stained suspensions of mitotic metaphase chromosomes of accession 703 (A, D), accession 711 (B, E), and accession 721 (C, F). The positions of sort windows are indicated on bivariate flow karyotypes of DAPI fluorescence pulse area vs. fluorescence pulse width (D - F).

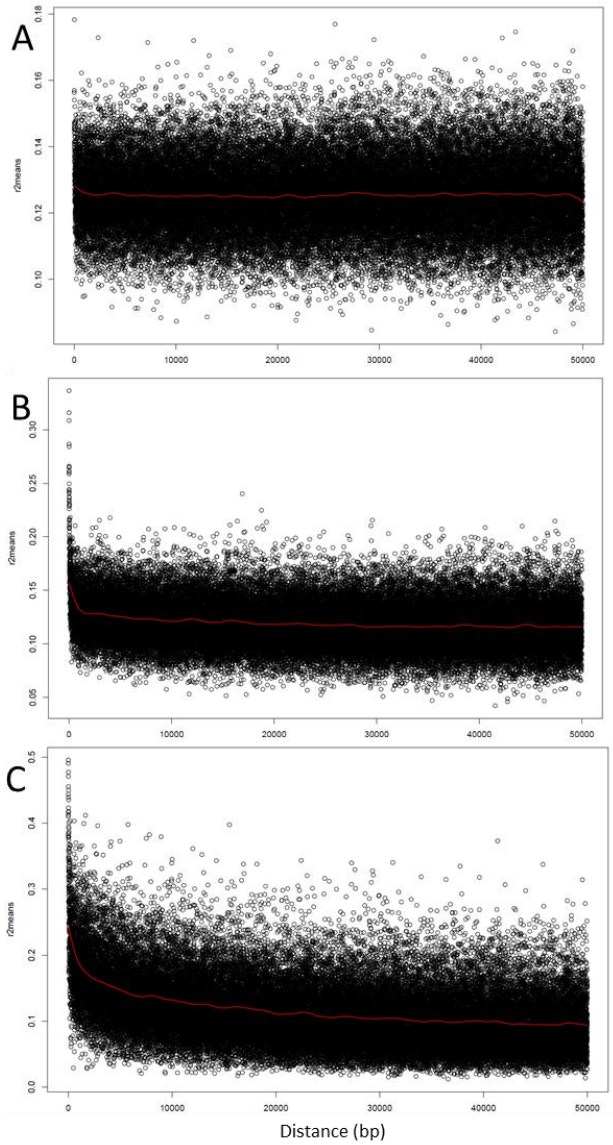
.



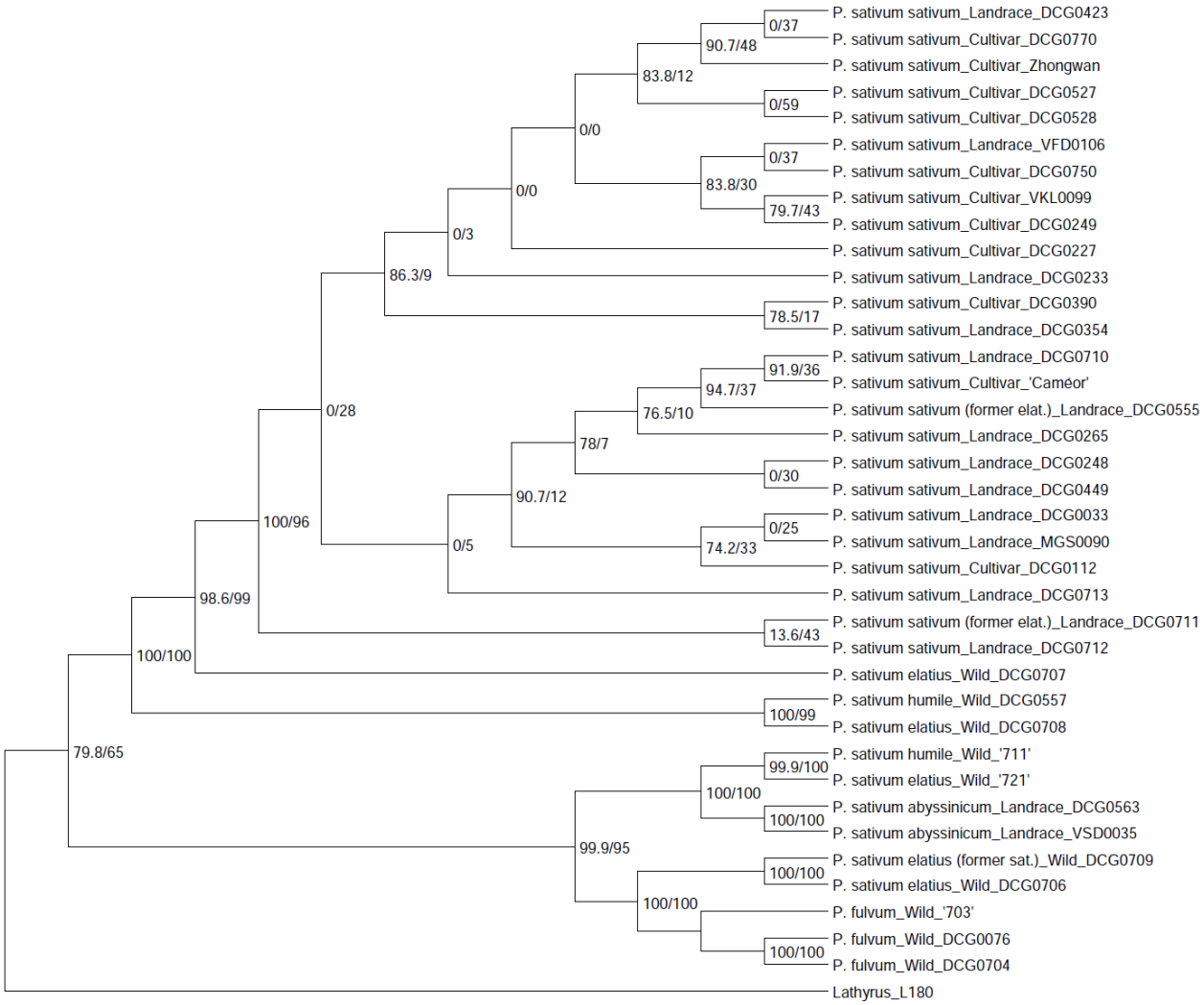
Supplementary Figure 11. Mean mapping coverage on the pea genome reference of resequencing reads of single chromosome amplified DNA samples, in three wild pea accessions.



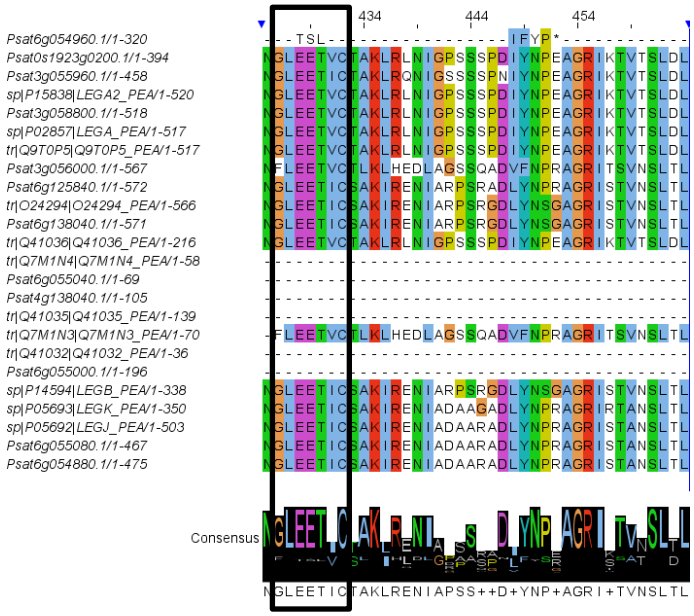
Supplementary Figure 12. *Pisum* diversity indices: **a**. Mean nucleotide diversity π[127] and **b**. Tajima's D[128] mean values in the different groups of accessions.



Supplementary Figure 13. Linkage disequilibrium decreases rapidly with distance between SNPs. Linkage disequilibrium was computed between pairs of single nucleotide polymorphisms (r2) in the 'wild' (A), 'landrace' (B), and cultivar' (C) groups of accessions.



Supplementary Figure 14. Phylogenetic tree of 38 re-sequenced accessions based on chloroplastic SNP diversity.



Supplementary Figure 15. Alignment of Legumin protein sequences from the pea genome, UNIPROT, and the pea gene atlas. Focus on the cleavage site between the basic and acidic polypeptides of the pre-protein (black box).

Supplementary Tables

Supplementary Table 1: 'Caméor' sequencing libraries

Sample preparation	Library type	Library insert sizes	Sequencing technologies	Read length (bp)	Genome coverage	NCBI or EMBL Accession number
Caméor nuclear DNA (Floraclean Plant DNA isolation kit)	Illumina overlapping PE	300 bp	HiSeq2000	2x100	71	PRJEB30482
	Illumina PE (tightly sized)	500 bp	HiSeq2500	2x150	18	PRJEB30482
		600 bp			27	PRJEB30482
		800 bp			30	PRJEB30482
				Total paired-end	***146***	
Caméor total DNA (MTAB extraction)	Illumina MP libraries (gel sized)	3-5 kb	HiSeq2000	2x100	37	PRJEB30482
		5-8 kb			39	PRJEB30482
		8-11 kb			33	PRJEB30482
Caméor nuclear DNA (MTAB extraction)	Illumina MP libraries (SageELF sized)	4.5 kb	HiSeq2500	2x150	3	PRJEB30482
		5.5 kb			3	PRJEB30482
		6.7 kb			2	PRJEB30482
		8.1 kb			3	PRJEB30482
		9.3 kb			3	PRJEB30482
		12 kb			6	PRJEB30482
		17 kb			6	PRJEB30482
				Total mate-pair	***135***	
Caméor (MTAB extraction)	SMRTbell	15-50 kb	PacBio RSII	N/A	13	PRJNA509681
Flow-sorted chromosomes amplified DNA	Illumina PE	450-800	HiSeq2500	2x250	~20	PRJEB30482
Flow-sorted chromosomes amplified DNA	Illumina PE	500 bp	HiSeq2000	2x100	~9	PRJNA507688

Supplementary Table 2: Statistics of intermediary assemblies

Assembly step	Soapde Novo2	Sspace	PBjelly	MaGuS	GapCloser	Chimera correction	Bionano	V1 Allmaps
Assembly unit	Contig	Scaffold	Scaffold	Scaffold	Scaffold	Scaffold	Super-scaffold	Pseudo-molecule
Total assembly size (Gb)	2.4	3.5	3.6	3.6	3.5	3.6	3.9	3.9
Assembly size (Gb) without N	2.4	2.5	2.9	2.8	3.1	3.2	3.2	3.2
N50	3,959	328,763	334,846	559,080	553,390	2,935	2,013	4
L50 (kb)	14.64	3.24	3.27	1.99	1.99	369.78	415.94	446,350.68
N90	695	83,940	85,393	143,658	142,513	10,171	9,775	1,084
L90	836	1,099	1,111	671	670	9,221	9,492	158,103
%N	0	30	20	22	12	12	19	19
Unit's max (kb)	79.6	1,896.1	1,922.8	2,889.5	2,846.6	2,165.2	7,603.4	579,269.1

Supplementary Table 3: Statistics of whole-genome Bionano maps.

1: Filtered Molecules statistics (>150Kb)	
Number of molecules	1,519,711
Total length (Mb)	405,574.52
Average length (kb)	266.88
Molecule N50 (kb)	271.75
Label density (per 100kb)	9.85
Theoretical Reference Coverage (x)	109.93
2: Molecules aligned to the NGS assembly	
Number of molecules aligned	608 43
Molecule fraction align	0.40
Total align length (Mb)	158,047.50
Effective Coverage (x)	43.63
Average align length (kb)	259.80
3: Optical assembly	
N Genome Maps	3 28
Total Genome Map Length (Mb)	3,689.38
Average Genome Map Length (Mb)	1.12
Median Genome Map Length (Mb)	0.90
Genome Map n50 (Mb)	1.44
Total Reference Length (Mb)	3,582.45
Total Genome Map Length / Reference Length	1.03
N Genome Maps aligned	1,344 (0.41)
Total Aligned Length (Mb)	626.28
Total Aligned Length / Reference Length	0.17
Total Unique Aligned Length (Mb)	616.68
Total Unique Aligned Length / Reference Length	0.17

Supplementary Table 4: Statistics of Bionano optical map generated from DNA of flow-sorted chromosomes.

Sample	**Pea (Mol intensity 0.6)**	**Pea (Mol intensity 0.4)**
Genome size	4.45 Gb	
No of chromosomes sorted	2,800,000 + 2,800,000	
Amount of DNA obtained	3.5 µg + 3.5 µg	
Raw data	**Pea_merged.bnx**	**Pea_merged_filtered.bnx**
Raw data (> 150 Kb; SNR 2.75)	1,087 Gb	1,002 Gb
No of molecules (> 150 Kb; SNR 2.75)	4,974,212	4,727,218
Mols N50 (> 150 Kb; SNR 2.75)	210 kb	205 kb
Labelling density (> 150 Kb; SNR 2.75)	9.7 labels/100 kb	9.3 labels/100 kb
Genome coverage	244 x	225 x

1: Filtered Molecules statistics (>150Kb)	
Number of molecules	4,586,875
Total length (Mb)	975,280.9
Average length (kb)	212.624
Molecule N50 (kb)	206.120
Label density (per 100kb)	9.436
Theoretical Reference Coverage (x)	283.3
2: Molecules aligned to the NGS assembly	
Number of molecules aligned	289,427
Molecule fraction align	0.063
Total align length (Mb)	36,498.3
Effective Coverage (x)	10.602
Average align length (kb)	126.1
3: Optical assembly	
N Genome Maps	6,879
Total Genome Map Length (Mb)	3,755.206
Average Genome Map Length (Mb)	0.546
Median Genome Map Length (Mb)	0.426
Genome Map n50 (Mb)	0.678
Total Reference Length (Mb)	3,582.45
Total Genome Map Length / Reference Length	1.048
N Genome Maps aligned	1,745 (0.25)
Total Aligned Length (Mb)	559.667
Total Aligned Length / Reference Length	0.156
Total Unique Aligned Length (Mb)	556.212
Total Unique Aligned Length / Reference Length	0.155

Supplementary Table 5: Correspondence between pseudomolecule labels in the pea genome assembly v1a, and those from earlier publications for linkage groups and chromosomes.

Pseudomolecules (present study)	Linkage groups[2]	Chromosomes[12]	Chromosomes[27]	Chromosomes[26]	Chromosomes[25]
chrom1LG6	VI	1	5	1	5
chrom2LG1	I	2	6	2	2
chrom3LG5	V	3	1	3	1
chrom4LG4	IV	4	4	4	4
chrom5LG3	III	5	3	5	6
chrom6LG2	II	6	2	6	7
chrom7LG7	VII	7	7	7	3

Supplementary Table 6: RepeatExplorer characterization of repeat content of the paired-end reads.

Repeats	Genome %
LTR/gypsy	***55.61***
TatIV_Ogre	46.68
Athila	5.42
Tekay	3.44
TatV	0.05
CRM	0.02
LTR/copia	***11.72***
SIRE	7.43
Angela	2.45
Ivana	1.03
TAR	0.22
Tork	0.15
Ale	0.11
Ikeros	0.04
LTR/TRIM	0.29
DNA transposon	***1.94***
MuDR_Mutator	1.1
EnSpm_CACTA	0.73
hAT	0.06
Helitron	0.05
rDNA	***0.6***
45S	0.56
5S	0.04
Satellites	2.64
Unclassified	4.28
ALL	***76.8***

Supplementary Table 7: Statistics of repetitive elements in the pea genome v1a.

Class	Total length (bp)	% contigs total length	Order	**Copy number**	Total length (bp)	% contigs total length	Repeat type	Lineage	Total length (bp)	% contigs total length
Class I (RXX)	**2.457E+09**	**77.78%**		**1945520**						
		including	DIRS (RYX)	**31802**	6.690E+07	2.118%				
			LINE (RIX)	**80076**	3.264E+07	1.033%				
			SINE (RSX)	**35413**	7.385E+06	0.234%				
			LTR (RLX)	**1707747**	2.298E+09	72.726%				
						including	Ty1/copia (RLC)		3.797E+08	12.018%
								AleI/Retrofit	7.114E+06	0.225%
								AleII	1.616E+07	0.512%
								Angela	6.417E+07	2.031%
								Bianca	3.637E+05	0.012%
								Ivana/Oryco	5.239E+07	1.658%
								Maximus/SIRE	2.186E+08	6.920%
								TAR	8.481E+06	0.268%
								Tork	1.235E+07	0.391%
								unclass(Ale)	3.377E+04	0.001%
							Ty3/gypsy (RLG)		1.228E+09	38.883%
								Athila	1.742E+08	5.513%
								chromovirus	8.122E+07	2.571%
								Ogre/Tat	9.731E+08	30.800%
Class II (DXX)	**1.720E+08**	**5.44%**		**246432**						
		including	Helitron (DHX)	**9182**	3.383E+06	0.107%				
			Maverick (DMX)	**85**	1.921E+04	0.001%				
			TIR (DTX)	**205905**	1.600E+08	5.065%				
						including	hAT		3.021E+06	0.096%
							CACTA		1.494E+07	0.473%
							Mutator		3.917E+06	0.124%
rDNA	1.402E+04	0.00%								
Unclassified	9.428E+06	0.30%								
HostGene	6.027E+06	0.19%								
							Pararetrovirus		1.649E+06	0.052%

Supplementary Table 8: Statistics of structural and functional gene annotation of the pea genome assembly v1a. Lengths are indicated in base pairs.

Features		Genes	Truncated Genes	Total
# genes		44,756	29	44,785
# mRNAs		57,835	47	57,882
# exons		283,368	287	283,655
# CDSs		265,652	245	265,897
#mono-exonic genes		10,781	3	10,784
Mean gene length		2,784	6,025	2,786
Mean coding sequence length		1,016	1,455	1,016
Mean exon length		325	406	325
Min exon length		2	2	2
Max exon length		16,939	4,954	16,931
Mean intron length		423	562	423
Min intron length		20	38	20
Max intron length		69,895	8,374	69,855
Mean number of exons per gene		6.33	9.90	6.00
Ratio of CDS/gene lengths		0.36	0.24	0.36
# annotated mRNAs				
Functional Annotation by	**Tools**	InterProscan5		46,561
		TrapID		42,996
		KASS		10,625
	Terms	InterPro		37,484
		GO		35,393
		KEGG		3,822
		Reactome		7,345
		MetaCyc		2,566
No functional annotation				16,454

Supplementary Table 9: Ks values for the Papilionoideae-specific mode. The mode range was calculated as one sigma around the mode.

Species	Mode	Range	
Pisum sativum	1.00	0.75	1.33
Medicago truncatula	0.83	0.66	1.03
Trifolium pratense	0.95	0.67	1.34
Trifolium subterraneum	0.84	0.68	1.05
Cicer arietinum	0.80	0.64	1.02
Cicer reticulatum	0.81	0.64	1.04
Lotus japonicus	0.65	0.53	0.80
Cajanus cajan	0.64	0.51	0.81
Glycine max[a]	0.61	0.49	0.76
Phaseolus vulgaris	0.72	0.59	0.87
Vigna radiata	0.86	0.68	1.08
Vigna angularis	0.83	0.67	1.03
Lupinus angustifolius[b]	0.63	0.49	0.80
Arachis duranensis	0.94	0.72	1.23
Arachis ipaensis	0.87	0.71	1.08

a) *G. max* lineage specific mode and range: 0.12 (0.08-0.16).
b) *L. angustifolius* lineage specific mode and range: 0.27 (0.21-0.34).

Supplementary Table 10: Relative amounts of the major TE families in *M. truncatula, P. vulgaris,* and *G. max*, according to De Vega et al.[65] and *L. japonicus* according to Sato et al.[66], as compared with pea.

	Present study Pisum sativum		de Vega et al. 2015 Trifolium pratense		Medicago truncatula		Phaseolus vulgaris		Glycine max		Sato et al. 2008 Lotus japonicus	
	Superfamily Coverage (bp)	Fraction of contig sequence	Fold-change / pea coverage	Superfamily Coverage (bp)	Fold-change / pea coverage	Superfamily Coverage (bp)	Fold-change / pea coverage	Superfamily Coverage (bp)	Fold-change / pea coverage	Superfamily Coverage (bp)	Fold-change / pea coverage	Coverage
Total TE	2 638 701 042	**0.8352**	20.68	127 576 578	14.16	186 328 751	7.77	339 606 927	4.24	622 604 981	24.67	106 974 100
Class 1 TEs												
LTR Gypsy	1 228 464 134	0.3888	162.35	7 566 938	31.54	38 954 999	9.32	131 786 780	4.56	269 298 878	42.94	28 606 300
LTR Copia	379 700 833	0.1202	15.57	24 391 421	12.50	30 381 006	4.79	79 334 021	2.45	154 704 057	16.35	23 225 200
SINEs	7 384 573	0.0023	4.99	1 480 120	2.39	3 085 624	44.95	164 280	5.51	1 340 657	235.18	31 400
LINEs	32 640 756	0.0001	1.66	19 637 581	1.27	25 763 983	0.65	50 090 542	1.28	25 568 379	84.32	387 100
Total Class 1	2 457 319 695	0.7778	38.67	63 552 730	21.75	112 990 132	8.56	286 983 344	5.13	479 305 937	39.38	62 394 400
Class 2 TEs												-
Total Class 2	171 953 356	0.0544	2.92	58 907 531	2.35	73 116 505	3.43	50 158 313	1.23	139 622 182	16.03	10 728 400
Unclassif TE	9 427 991	0.0030	1.84	5 116 317	42.45	222 114	3.82	2 465 270	2.56	3 676 862	0.28	33 851 300

Supplementary Table 11. Reconstructed Legume ancestral karyotypes. The table delivers for each synteny block (in lines) the orthologous chromosomes among the Millettoid (*Phaseolus vulgaris* (Pv), *Cajanus cajan* (Cc), *Vigna angularis* (Va), *Vigna radiata* (Vr), *Glycine max* (Gm)) and among the Galegoid (*Medicago truncatula* (Mt), *Cicer arietinum* (Ca), *Pisum sativum* (Ps), *Lotus japonicus* (Lj)) as well as *Arachis, duranensis (Ara)*. Reconstructed CARs for ALK (25 and 17), AGK (8) and AMK (16) are delivered respectively in the second, third, fourth and fifth columns. The color code (first column) is as in the main Figure 4.

Color_Code	ALK_19	ALK_25	AGK_8	AMK-16	Vr	Cc	Gm	Va	Pv	Mt	Lj	Ca	Ps	Ara
1	1	1	chr1	chr1	chr6	chr2	chr13,chr14,chr17	chr1	chr1	chr1	chr5	chr4	chr6	chr6
2	2	2	chr3	chr3	chr2,chr3	chr1	chr11,chr18	chr4	chr1	chr3	chr6	chr5,chr6	chr5	chr3,chr4,chr10
3	3	3	chr7	chr2,chr12	chr3	chr2,chr3,chr7	chr3,chr16,chr19	chr4,chr7	chr1	chr7	chr1	chr3	chr3	chr6,chr10
4	3	4	chr7	chr10	chr10	chr3	chr3,chr19	chr3	chr6	chr7	chr1	chr3	chr3	chr6
5	4	5	chr5	chr4	chr11	chr6,chr8	chr1,chr2,chr9,chr11	chr7	chr2	chr5	chr1,chr2	chr2,chr8	chr2,chr4	chr5
6	4	6	chr5	chr6	chr11		chr1,chr2		chr3	chr5	chr2	chr2	chr2	chr5
7	5	7	chr4	chr5	chr7	chr9	chr5,chr7,chr8,chr13,chr20	chr2	chr2	chr4	chr4	chr6	chr7	chr3,chr4
8	5	8	chr4	chr5	chr1,chr7	chr5,chr6,chr8,chr10	chr5,chr8	chr10	chr2	chr8	chr2,chr4	chr6	chr7	chr3,chr5
9	6	9	chr1	chr6	chr11	chr6	chr7,chr9,chr20		chr3	chr1		chr4	chr6	chr10
10	7	10	chr4,chr8	chr6	chr7	chr6,chr11	chr2,chr5,chr7,chr13,chr17	chr11	chr3	chr4	chr4	chr6,chr7	chr4,chr7	chr1,chr10
11	7	11	chr8	chr6	chr7	chr6,chr11	chr2,chr5,chr8,chr9,chr16,chr18	chr11	chr3	chr8	chr2,chr4	chr7	chr4	chr1,chr3
12	8	12	chr6	chr7	chr1	chr8	chr1,chr2,chr5,chr7,chr9,chr13,chr16,chr19	chr2	chr4	chr6	chr2	chr2,chr6,chr	chr1	chr5,chr10
13	9	13	chr2	chr8	chr4,chr5	chr10	chr8,chr12,chr13,chr15	chr9	chr5	chr2	chr3	chr1	chr1,chr5	chr7,chr8
14	9	14	chr3	chr8	chr5	chr10	chr4,chr6,chr11,chr15		chr5	chr3	chr3	chr5	chr5	chr5,chr8,chr10
15	10	15	chr2	chr9	chr10	chr2,chr8	chr8,chr9,chr12,chr13,chr15,chr16	chr3	chr6	chr2	chr6	chr1	chr1	chr4,chr10
16	11	16	chr3	chr9	chr10	chr1	chr8,chr11,chr13,chr18,chr20	chr3	chr6	chr3	chr6	chr5,chr6	chr5	chr1,chr3
17	12	17	chr1	chr11	chr8	chr2,chr3,chr9	chr2,chr9,chr10,chr11,chr13,chr20	chr6	chr7	chr1	chr5,chr6	chr4	chr6	chr2,chr9
18	13	18	chr5	chr1	chr6	chr2	chr2,chr13,chr14	chr1	chr8	chr5	chr1,chr2	chr2	chr2	chr7,chr8
19	14	19	chr7	chr12	chr4	chr2,chr7	chr1,chr2,chr7,chr8,chr9,chr16,chr18	chr4	chr8	chr7	chr1	chr2,chr3	chr3	chr4,chr6
20	15	20	chr2	chr13,chr14	chr3,chr5	chr8	chr4,chr6,chr9,chr15	chr9	chr9	chr2	chr6	chr1	chr1	chr1,chr3,chr4
21	16	21	chr3	chr13	chr3,chr5,chr10	chr3,chr7,chr11	chr4,chr6	chr5,chr9	chr9	chr3	chr1	chr5	chr5	chr3,chr7,chr8,chr10
22	17	22	chr4	chr15	chr9	chr1,chr11	chr1,chr3,chr7,chr8	chr8	chr10	chr4	chr3	chr6	chr7	chr2,chr4,chr9
23	17	23	chr8	chr15	chr9	chr1,chr11	chr3,chr7,chr8,chr16	chr8	chr10	chr8	chr3	chr7	chr4	chr2,chr8,chr9
24	18	24	chr4	chr16	chr2	chr4,chr6	chr6,chr9,chr11,chr12	chr2,chr5	chr11	chr4	chr3	chr6	chr7	chr3,chr8
25	19	25	chr8	chr16	chr2	chr4	chr6,chr8,chr12,chr13,chr15	chr2	chr11	chr8	chr3	chr7	chr4	chr3,chr8

Supplementary Table 12. Number of single-chromosome amplified DNA samples selected for sequencing for accessions '703', '721', and '711'.

Sort window	P. fulvum '703'	P. sativum elatius '711'	P. sativum humile '721'
1	13	23	8
2	19	28	28
3	8	---	11
All	40	51	47

Supplementary Notes

Table of Contents

1. Genome assembly

1.1 Plant material and DNA preparation

The French pea cultivar 'Caméor' was used for genome sequencing (Supplementary Figure 1). Pea is a primarily self-pollinating species. Nevertheless, to obtain full homozygosity, plants used for tissue production for genome sequencing were produced from seeds increased in insect-proof glasshouses after three generations of single-seed descent. Homogeneity was evaluated using SSR markers. Total DNA was extracted from fresh leaves following two protocols, either using the Floraclean Plant DNA isolation kit or using Myristyl-Trimethyl-Ammonium-Bromide (MTAB) according to Baurens et al.[1] without the column purification step to limit DNA degradation.

1.2 Whole genome sequencing

Sequence data was obtained for DNA libraries using various sample preparation and sequencing methods. The library features are given in Supplementary Table 1 and methods are described below.

1.2.1 Illumina sequencing

Illumina sequencing was conducted at Genoscope (Evry, France). Four Illumina paired-end (PE) libraries were prepared starting from Caméor total DNA. Two independent DNA fragmentations were performed using the E210 Covaris instrument (Covaris, Woburn, USA) in order to generate fragments of around 300 bp (for the overlapping library), or 600 bp (for the library with three insert sizes of 500 bp, 600 bp, and 800 bp). Sequencing libraries were constructed using the NEBNext DNA Sample Prep Master Mix Set (New England Biolabs, Ipswich, USA). DNA fragments were PCR-amplified using Platinum Pfx DNA polymerase (Invitrogen, Carlsbad, USA) and P5 and P7 primers. Amplified library fragments were size selected on 3% agarose gels around 300 bp, or on 2% agarose gels around 500 bp, 600 bp and 800 bp. Library traces were validated on an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, USA) and quantified by qPCR using the KAPA Library Quantification Kit (Kapa Biosystems, Wilmington, MA, USA) on a MxPro instrument (Agilent Technologies). The PE libraries were sequenced using Illumina HiSeq 2000 or HiSeq 2500 platforms (Illumina, San Diego, USA) generating a total of 2.9 billion paired-end reads as described in Supplementary Table 1.

The mate-pair (MP) libraries were prepared using the Nextera Mate Pair Sample Preparation Kit (Illumina). Briefly, genomic DNA (4 µg) was simultaneously enzymatically fragmented and tagged with a biotinylated adaptor. Fragments were size selected (3-5; 5-8 and 8-11 Kb) using gel electrophoresis and then circularized overnight with a ligase. Linear, non-circularized fragments were digested, and circularized DNA was fragmented to 300-1000 bp size range using Covaris E210. Biotinylated DNA was immobilized on streptavidin beads, end-repaired, then 3'-adenylated. Illumina adapters were added. DNA fragments were amplified using Illumina adapter-specific primers and then purified. Finally, libraries were quantified by qPCR and library profiles were evaluated using Agilent 2100 Bioanalyzer (Agilent Technologies). Each library was sequenced using 100 bp read chemistry on a paired-end flow cell on Illumina HiSeq 2000 (Illumina). Later, a new series of seven MP libraries were prepared (still starting from 4 µg DNA) using the same Nextera technology as described above but performing the size selection on a Sage Science Electrophoretic Lateral Fractionator (Sage Science, Beverly, USA). This system allowed obtaining narrow-sized MP libraries, isolating 12 different discrete size fractions from a single sample loading. We selected

seven fractions (4.5, 5.5, 6.7, 8.1, 9.3, 12 and 17 Kb) to continue the MP library preparation and sequencing on Illumina HiSeq 2500.

1.2.2 PacBio sequencing

PacBio sequencing was conducted at the Laboratory for Biotechnology and Bioanalysis, Washington State University. Genomic DNA samples were submitted to PacBio library construction and sequencing as follows. Ten to fifteen µg gDNA was sheared using Covaris G-Tubes for 10min at 1350xg using Beckman Coulter Minifuge 16 centrifuge. The sheared DNA was concentrated and cleaned using 0.45x Ampure XP beads (Beckman Coulter, Brea, USA). Pacific Biosciences Single Molecule, Real Time (SMRT) bell library was prepared following the protocol (P/N 100-286-000-5) provided by Pacific Biosciences (www.pacb.com) using the SMRTbell Template Prep kit 1.0 (P/N 100-259-100). The resulting SMRTbell libraries were size selected using BluePippin (Sage Science) according the manufacturer's instructions. The cut off limit was set to 15kb-50kb to select SMRTbell library molecules with an average size of 20kb or larger. The Pacific Biosciences Binding and Annealing calculator was used to determine the appropriate concentrations for the annealing and binding of the SMRTbell libraries. The libraries were annealed and bound to the P5 or P6 DNA polymerase for sequencing using the DNA/Polymerase Binding Kit P5 (P/N100-256-000) or P6 v2.0 (P/N100-372-700). The only deviation from standard protocol was the extension of binding times from 30 min to 1-3 hours. The bound SMRTbell libraries were loaded onto the SMRT cells using the standard MagBead protocol, and the MagBead Buffer Kit v2.0 (P/N 100—642-800). The standard MagBead sequencing protocol was followed using the DNA Sequencing Kit 4.0 v2 (P/N 100-612-400) either P5/C3 or P6/C4 chemistry. Sequencing data was collected for 6-hour movie times and Stage Start was enabled to capture the longest single reads possible on the PacBio RS II instrument at Washington State University. This resulted, using the P5 chemistry in the production of 28 Gbp in total, with N50 of 9,500 kb, and using the P6 chemistry, in 41 Gbp in total with N50 of 15,917 kb.

1.3 Raw data processing and estimation of genome size

The k-mer spectrum was built with 30x of 150 bp reads using the GenomeScope program (http://qb.cshl.edu/genomescope/). The estimated genome size of 'Caméor' through this method (4.426 Gb) was consistent with previous estimates obtained by flow-cytometry[3].

1.4 De novo assembly of the pea genome

The seven chromosomal molecules representing the pea nuclear genome were assembled in a step-wise manner. The assembly pipeline is summarized in Supplementary Figure 2. Shot-gun Illumina and PacBio sequence reads were combined to obtain scaffolds. The first assembly was improved with layers of data from physical maps (Whole Genome Profiling, WGP), additional reads, optical maps (Bionano maps), various high-density linkage maps (Genetic maps) and synteny to the *M. truncatula* genome.

The statistics of the different intermediary assemblies obtained in the step-wise assembly pipeline summarized in Supplementary Figure 2. are presented in Supplementary Table 2.

1.4.1 De novo assembly and scaffolding of Illumina reads

Overlapping paired-end reads were corrected using Musket[4] and merged with an in-house script. The merged reads were assembled into contigs using SOAPdenovo2 2.04[5] with 127 nt k-mer and the –R option in the 'pregraph' step. Contigs shorter than 500 nt were removed. The contigs were scaffolded with

SSPACE 2.0[6] using the Nextera reads obtained for mate-pair libraries. Only the links validated by five read pairs were considered and scaffolds shorter than 2 kb were removed.

1.4.2 Scaffolding using Whole Genome Profiling map

A physical map was produced using 295,680 bacterial artificial chromosome (BAC) clones of cv. 'Caméor' pooled in a multi-dimensional manner. The BAC library was provided by INRA IPS2, Paris-Saclay, France and is available at INRA CNRGV (https://cnrgv.toulouse.inra.fr/library/genomic_resource/Psa-B-Cam). The estimated average insert size of the BAC library is 100 kb and its genome coverage is 9.3X. The BAC DNA was digested with restriction enzymes *Hin*dIII/*Mse*I; restriction fragments were ligated, PCR amplified, and sequenced using Illumina HiSeq 2000 platform (100 nt read length). The reads were assigned to individual BAC clones based on their occurrence in pooled samples of BAC clones in each dimension. The BAC clones were finally assembled using an improved version of FPC software (Keygene N.V.) and the physical map was built according to Gali et al.[7]. The SSPACE scaffolds were then re-scaffolded with MaGuS 1.0[8] using the mate-pair reads and the sequence-based physical map generated using the Whole Genome Profiling technology[9] (Keygene N.V., Wageningen, The Netherlands).

1.4.3 Gap-closing using Illumina and PacBio reads

The gaps in scaffolds were closed with GapCloser[10,11] using paired-end, mate-pair and PacBio reads. Gap-closed super-scaffolds constituted a first assembly, which was visually inspected for inter-chromosomal and intra-chromosomal chimeras and edited in accordance.

1.4.4 Correction of inter-chromosomal chimeric scaffolds

To confirm chromosome allocation of sequence contigs and to detect inter-chromosomal chimeric scaffolds, we used sequence data obtained from single chromosomes sorted by flow-cytometry following the methodology detailed below.

1.4.4.1 Chromosome sorting by flow-cytometry

Suspensions of intact mitotic chromosomes were prepared as described by Neumann et al.[12], with modifications. To prepare one sample, 30 seeds were germinated in a glass petri dish on moistened filter paper. Seedlings with approximately 3 cm primary roots were transferred onto a plastic tray filled with Hoagland's solution containing 1.25 mM hydroxyurea (HU) for 18 hours. Then the roots were incubated in HU-free Hoagland's solution for 4.5 h and immediately after in 10 µM amiprophos-methyl (APM) in Hoagland's solution for 2 h. All incubations were performed in the dark at 25 ± 1°C and all solutions were aerated. Finally, the seedlings were transferred to a tray filled with ice water and incubated overnight in a refrigerator. The synchronized roots were cut 1 cm from the tip and fixed in 2% formaldehyde in Tris buffer for 30 min at 5°C. Then the roots were washed three times for 5 min in Tris buffer and meristem tips of 25 roots were cut and transferred to a polystyrene tube containing 1 ml LB01 lysis buffer[13], and chromosomes were released mechanically by a Polytron PT 1200 homogenizer (Kinematica AG, Luzern, Switzerland) at 13,000 rpm for 18 s. The homogenate was passed through a 20 µm pore size nylon mesh, stained by DAPI (4´,6-diamidino 2-phenylindole) at final concentration of 2 µg/ml and analyzed on FACSAria II SORP flow cytometer and sorter (BD Immunocytometry Systems, San José, USA) at rates of 1500–2000 particles per second. Relative DAPI fluorescence intensities of the particles were acquired on histograms of DAPI

fluorescence pulse area (Supplementary Figure 3 A, C). The results were also displayed as dot-plots of DAPI fluorescence pulse area vs. fluorescence pulse width (Supplementary Figure 3 B, D).

1.4.4.2 Single chromosome sorting and DNA amplification

To sort single copies of chromosomes, the flow-cytometry instrument was set for "single cell one drop" mode and sort windows were set on a dot-plot of DAPI fluorescence pulse area vs. fluorescence pulse width. In order to increase the probability of collecting a similar number of copies of each of the seven pea chromosomes, several sort windows were set corresponding to subpopulations differing in DAPI fluorescence. Two series of 49 chromosomes of *P. sativum* cv. Caméor were prepared, each series theoretically comprising seven copies of each pea chromosome. In the first experiment, three windows corresponding to the tree peaks on monovariate flow karyotype were selected (Supplementary Figure 3 B). In total, 7 chromosomes were sorted from window 1, 35 chromosomes from window 2 and 7 chromosomes from window 3. Slightly higher resolution was achieved in the second experiment, leading to the resolution of a shoulder on the major composite peak 2 (Supplementary Figure 3 C). This provided an opportunity to sort chromosomes from four populations (Supplementary Figure 3 D). In total, 7 chromosomes were sorted from window 1, 28 chromosomes from window 2L, 7 chromosomes from window 2R and 7 chromosomes from window 3.

Multiple displacement amplification (MDA) of chromosomal DNA was done according to Cápal et al.[14]. Briefly, GenomiPhi V2 sample buffer (GE Healthcare, Little Chalfont, UK) was mixed with 10mg/ml proteinase K (Sigma-Aldrich, St. Louis, USA) at ratio 10:1 just prior to use and chromosomes (one chromosome per tube) were flow-sorted directly into 0.2 ml PCR tubes containing 3 µl of the mix. Each sorted chromosome was spun down and incubated in the GenomiPhi V2 buffer supplemented with proteinase at 50°C overnight. The proteinase was inactivated by heating to 85°C for 15 min and the samples were stored at -20°C until use. Then, 1.5 µl of lysis buffer (600mM KOH, 100mM DTT, 10mM EDTA pH 8) was added to each sample, followed by incubation at 30°C for 15 min. The reactions were then neutralized with 1.5 µl neutralization buffer (Tris-HCl pH 8, 300mM HCl), spun down and kept on ice until use. DNA amplification master mix consisting of 4 µl sample buffer, 9 µl reaction buffer and 1 µl enzyme (all reagents from GenomiPhi V2 kit) was added to each sample. The amplification was performed at 30°C for 4 hours and the enzyme was then inactivated at 65°C for 10 min. The negative control was processed in exactly the same way, except that no chromosome was sorted into the tube. All pipetting steps were performed in a UV-irradiated biohazard cabinet using Axygen sterile filter pipette tips (Corning, Tewksbury, USA). MDA products were checked on 1.5% agarose gel, purified by Agencourt Ampure XP magnetic beads (Beckman Coulter, Brea, USA), dissolved in 35 µl ddH_2O and DNA concentration was estimated by Nanodrop 1000D spectrophotometer (Thermo Fischer Scientific, Wilmington, USA), yielding on average 1.96 ug DNA/sample in the first batch and 2.89 ug DNA in the second batch.

Thirty-five single-chromosome DNA samples with the highest amount of amplified DNA were selected from the first series and sequenced. MDA products obtained from the second series were checked by PCR for the presence of PisTR-B tandem repeat[15] using the following primer sequences: Forward atttgggtactttaaactaac; Reverse gaatgatgaaaatgttgatgt with 5 ng DNA amplified DNA as a template. All 49 MDA products were sent for sequencing.

1.4.4.3. Sequencing of Flow-sorted chromosomes' amplified DNA

Thirty-five flow-sorted single chromosome amplified DNA samples were sequenced at UWA. Paired-end Illumina libraries were sequenced to a total coverage of 9x and deposited in the SRA under BioProject

PRJNA507688. Forty-nine samples were sequenced at GENOSCOPE. Flow-sorted chromosomes' amplified DNA (250 ng) was sheared in a Covaris microTube using the Covaris E210 System (Covaris, Woburn, MA). The peak size of DNA was around 600 bp. Libraries were prepared using the NEBNext DNA Sample Prep Master Mix kit with a 'on beads' protocol as described[16]. Libraries were then subjected to quality control as described in §1.2.1 and sequenced on an Illumina HiSeq 2500 in rapid mode, with 2x250 base paired end reads, reaching at least about 2 million reads per sample.

1.4.4.4 Chimeric scaffolds correction

Sequencing reads obtained for eighty-one chromosome libraries were mapped and samples were assigned to chromosomes (Supplementary Figure 3 E): 8 libraries were assigned to chr1LG6, 8 to chr2LG1, 9 to chr3LG5, 6 to chr4LG4, 11 to chr5LG3, 18 to chr6LG2, 12 to chr7LG7. Nine libraries corresponded to a mix of several chromosomes and three failed. Mapping chromosome-specific reads identified scaffolds that contained contigs from different chromosomes and allowed splitting them into smaller scaffolds.

1.4.4.5. Construction of a high-density linkage map and correction of intra-chromosomal chimeric scaffolds

Skim-based genotyping by sequencing (skim GBS) uses low-coverage (1-10x) whole genome sequencing and is a two-stage method that requires a reference genome sequence, genomic reads from parental individuals and individuals of the population[17]. TruSeq Nano DNA Libraries (Illumina) were prepared for the parental lines 'Caméor' and 'Melrose' and the 162 recombinant inbred lines derived from their cross (Pop6[2]) according the manufacturer's instructions. The library insert size was ~500 bp and paired-end skim-GBS data (2x125 bp) was produced using the HiSeq 2000 at the Australian Genome Research Facility (AGRF, Melbourne, Australia; NCBI Bioproject PRJNA507685). Approximately 10X and 6.5X coverage was obtained for the parental lines 'Caméor' and Melrose', respectively, with an average of 1x for the progeny ranging from 0.3X to 1.7X. SNPs and genotypes were called using SGSautoSNP and the skimGBS pipeline[17,18]. Reads were mapped to the reference using SOAPaligner/soap2 v2.21[19] and the following options: insert size 0 to 1000, report reads aligning non-repetitively. Subsequent mapping of the progeny reads to the same reference and comparison with the parental SNP file enables the calling of the parental genotype. According to the SGSautoSNP protocol, read data were not trimmed or filtered[17,18].

The map for Pop6 was built at the contig level in order to minimize missing data and improve the mapping resolution. Contigs' genotyping data for the 162 recombinant inbred lines were imputed from skimGBS SNPs as follows. SNPs were assigned to contigs; contigs were genotyped according to calling of the SNP present on this contig; when all SNPs in a contig were not of the same genotype, the contig was discarded from the analysis. Map construction was done based on contig-level genotypes, as described in Tayeh et al.[2]. This map included 64,038 contig positions (Supplementary Data 1), imputed from genotyping data for 473,583 SNPs, and represents the highest density genetic map published so far for pea. From the position of contigs on the genetic map, we identified potential intra-chromosomal chimeras: when contigs belonging to a given scaffold were placed more than 10 centi-Morgans apart on the genetic map, we manually split the scaffold.

1.4.5. Super-scaffolding using optical BioNano maps

Corrected scaffolds were assembled into super-scaffolds with the help of optical maps as follows.

1.4.5.1. Generation of whole genome optical BioNano map

Nuclei were isolated from young leaves of 'Caméor' plants grown in the dark, following IrysPrep DNA isolation protocol (BioNano Genomics, San Diego, USA). Briefly, three grams of pea young leaves were

fixed with formaldehyde and blended with a tissue homogenizer in Plant Homogenization Buffer plus (BioNano Genomics). After filtration steps and washing treatments, the nuclei were purified on density gradient (BioNano Genomics) and embedded in agarose gel plugs (CHEF Genomic DNA Plug Kit, BioRad). Embedded nuclei were incubated two times (two hours and O/N) with lysis buffer (BioNano Genomics) and proteinase K (Qiagen) at 50°C following by a RNAse treatment of 1h at 37°C. The plugs were washed, and agarose was solubilized with 2 units of GELase (Epicentre). Extracted high molecular weight (HMW) DNA was drop dialyzed for 2.5 hours. DNA concentrations were measured using the Quant-iT dsDNA Assay Kit (Life Technologies).

The NLRS DNA (Nicked, Labelled, Repaired and Stained DNA) was performed following the IrysPrep Reagent Kit protocol (BioNano Genomics). Briefly, 900 ng of DNA was digested with 10 units of nicking endonuclease (New England BioLabs) Nt.BspQI (GCTCTTC) for 2 h at 37 °C. Nicked DNA was then incubated for 1 h at 72 °C with fluorescently labelled dUTP and Taq Polymerase (New England BioLabs). The ligation of nicks was performed with Taq ligase (New England BioLabs) in the presence of dNTPs. DNA was counterstained with YOYO-1 (Life Technologies). NLRS DNA samples were loaded into IrysChips® (BioNano Genomics) and run on the Irys® instrument (BioNano Genomics). Data were collected until ≥ 110 fold coverage of long molecules (≥ 150 kb) was achieved.

The raw molecules were filtered using BioNano IrysView software (version 2.5.1) and molecule longer than 150 kb and with at least 8 label sites were kept. We obtained 1,519,711 molecules with a N50 of 271.7 kb (Supplementary Table 3.1.). The NGS assembly was *in silico* digested with the *Bsp*QI enzyme using the BioNano tool fa2cmap.pl. The molecules were aligned to the NGS assembly using RefAligner (Supplementary Table 3.2). We performed the assembly of the filtered molecules using the BioNano assembly pipeline (Pipeline version 4618, RefAligner and Assembler version 4704) with the parameters used for human samples as recommended for large genomes. We launched five iterations and used the "auto-noise" option which calculates the optimal noise parameters. It produced 3,281 genome maps with a N50 of 1.4 Mb and a cumulative length of 3.4 Gb (Supplementary Table 3.3). We finally performed hybrid scaffolding based on the NGS assembly and the genome maps to improve the contiguity of the initial scaffolds. We used the BioNano hybridScaffold pipeline (version 4741) with the default parameters and the automatic resolution of conflict. This option was used to split NGS scaffolds which were detected as being chimeras. We obtained 1,730 hybrid scaffolds with a total length of 1.7 Gb and a N50 of 1.4 Mb. The final assembly was composed of 24,623 scaffolds with a cumulative length of 3.9 Gb and a N50 of 416 kb. The rate of undetermined bases (N) has increased to 19.4% due to the hybrid scaffolding.

1.4.5.2. Generation of Bionano optical map from DNA of flow-sorted chromosomes

Suspensions of intact mitotic chromosomes were prepared from *P. sativum* cv. Caméor, stained and analyzed by flow cytometry as described above (§1.4.4.1-2). In order to sort large quantities of chromosomes needed for preparation of high molecular weight DNA for optical map construction, the "4-way purity" sort mode was selected, and the sort window was set on a dot-plot of DAPI fluorescence pulse area vs. fluorescence pulse width so that the window included the populations representing all seven pea chromosomes. The monovariate flow karyotype comprised the partially resolved peak representing chromosome 1, a large composite peak representing chromosomes 2-4 and 6-7, and a well resolved peak representing chromosome 5. In order to purify all chromosomes for the preparation of high molecular weight DNA to construct optical map, the sorting window was set to include all populations representing the seven *P. sativum* chromosomes.

A total of 5.6 million pea chromosomes, corresponding to 7 µg DNA, were flow-sorted and embedded in eight agarose miniplugs of a total volume of 160 µl, each comprising 700,000 chromosomes. DNA of chromosomes in the plugs was purified by proteinase K (Roche) as described by Šimková *et al.*[20]. Purified DNA was used to construct Bionano optical map following the protocol of Staňková *et al.*[21]. Based on the frequency of recognition sites in the whole genome sequence assembly, *Nt.BspQI* nickase (GCTCTTC recognition site) with the estimated frequency of 9.6 sites/100 kb was selected for DNA labeling. A total amount of 2.7 µg of chromosomal DNA was nicked using 90 U of *Nt.BspQI* (New England BioLabs, Beverly, USA) at 37°C for two hours in NEBuffer 3. DNA samples were then labelled at nicking sites with Alexa546-dUTP fluorochrome and the backbone of fluorescently labelled DNA was stained with YOYO-1 following the IrysPrep Reagent Kit protocol (Bionano Genomics, San Diego, USA). Labelled and stained DNA was loaded on the Irys chips and analyzed on the Irys platform (Bionano Genomics). A total of 1260 and 1087 Gb raw data were generated of which 1002 Gb comprising molecules >150 kb, corresponding to 225 genome equivalents was selected and used to assemble Bionano optical map (Supplementary Table 4). *De novo* map assembly was performed by a pairwise comparison of all single molecules and graph building[22] applying parameters recommended for large genomes. A p-value threshold of $1e^{-10}$ was used during the pairwise assembly, $1e^{-11}$ for extension and refinement steps, and $1e^{-15}$ for the final refinement.

1.4.6 Pseudomolecule construction based on linkage maps and inter-specific synteny

To produce pseudo-molecules, hybrid scaffolds derived from the Bionano optical maps were anchored using Allmaps[23] and six high-density genetic maps derived from populations with Caméor as a common parent. The maps included those described in Tayeh et al[2] for populations 4 (6,642 markers), 5 (6,031 markers), 7 (7,012 markers), and 9 (7,639 markers) and that built for population 6 genotyped using skimGBS as part of this study (64,038 positions). Synteny between the assembled pea genome and that of the model legume *Medicago truncatula* (v4)[24] was used as an additional anchoring criterion although with a lower relative weight. After a first run, Allmaps was relaunch for three pseudomolecules (chr2LG1, chr3LG5, chr6LG2) scoring a Spearman correlation between physical and genetic map lower than 0.9. Parameters were tuned specifically by lowering the weight of Pop 6 map. This resulted in the pea genome assembly v1a that comprises seven pseudomolecules.

In total 10,357 scaffolds were anchored to the seven linkage groups based on the LG group assignment (Supplementary Data 2), and 1,155 were anchored by at least three markers, allowing for a confident determination of their orientation. Adjacent super-scaffolds in each chromosome were separated by 100 “N”s. The total length of anchored sequences was 3.23 Gb, which accounts for 82.5% of the pea genome assembly.

1.5 Naming pseudomolecules in the pea genome assembly v1a

How the various chromosome designations corresponded to each other and to the linkage group designations was not easy to decipher due to the long history of pea genetic studies. The pea karyotype comprises seven chromosomes: two sub-metacentric chromosomes (1 and 2) and five acrocentric chromosomes (3, 4, 5, 6 and 7). Chromosomes 4 and 7 have a secondary constriction corresponding to the 45S rRNA gene cluster. Unlike most other species, pea chromosomes have not been classified according to their length and morphology, but rather relative to the successive assignations in linkage maps developed throughout history. At least four different chromosome numbering systems have been

used[25,26,27,12]. To be able to relate the pea genome assembly v1a to earlier genetic mapping results, we named pseudomolecules according to the latest reference genetic map[2] and to the chromosome numbering used by Neumann et al.[12] (Supplementary Table 5)

1.6 Evaluation of pea genome v1a quality

1.6.1 Analysis of proportion of sequenced reads represented in assembly

A sample of paired-end reads (2x150 bp) equivalent to 10X the genome length (287,923,574 reads in total) was mapped against the pseudomolecules in the final assembly using BWA[19]. Only 0.08% of the reads did not map to the reference.

1.6.2 Evaluation of potential structural mis-assemblies

The genome was evaluated for potential mis-assemblies using DELLY[28]. Paired-end reads were mapped to the genome and simulations of various potential structural variants were evaluated. Taking 5 paired-end reads as the lowest possible support for mis-assembly, the structural variant (SV) rate is one mis-assembly every 88.4kb.

Calculation of potential mis-assemblies based on pair-end reads mapping

Number of supported paired end per split-read mapping (PE/SR)

	5	10	15	20	25	30	40	50
Deletions	9,218	3,428	2,670	2,471	2,372	2,313	2,249	2,220
Duplications	5,313	1,115	474	283	199	153	97	74
Inversions	2,419	679	317	185	122	94	58	43
Translocations	27,364	8,739	4,133	2,490	1,588	1,091	602	381
Total	44,319	13,971	7,609	5,449	4,306	3,681	3,046	2,768
Variant rate (bp)	88,453	280,593	515,201	719,428	910,395	1,064,972	1,286,987	1,416,243

1.6.3 Representation of repeated elements in genomic reads and in the assembly

Clustering of 3,972,596 paired-end reads (100 bp) was performed using the computational pipeline RepeatExplorer v2[29]. The automated repeat classification of repeat clusters representing more than 0.01% of the genome was manually curated. A TAREAN analysis[30] of 1,954,369 paired-end reads (100 bp) further identified 30 high-confidence and 7 low-confidence putative satellites, most of them matching previously characterized satellites from *Pisum*.

The representation of these various repeats in the assembly was then assessed as follows. Assembled scaffolds were split into 120 bp fragments and blasted (-e 1e-20 -W 11 -r 2 -q -3 -G 5 -E 2) against sequences of repeat clusters obtained previously using RepeatExplorer (Supplementary Table 6). Each fragment was assigned to one cluster based on its best hit. The proportion of Illumina reads in each cluster relative to the total number of analyzed reads provides an estimate of abundance of corresponding repeats. Accordingly, the proportion of assembled sequences mapped to each cluster relative to the total size of the assembly provides the estimate of the representation of the same repeat in the assembly. The analysis revealed that most of the abundant repeats (cluster sizes >0.1 % of the genome) are under-represented in the assembly (Supplementary Figure 4). Most satellite repeats and rDNA are largely under-represented or missing as shown by the red triangles close to the zero line on Supplementary Figure 4. We calculated

that there are 957 Mbp of missing repeated sequences which account for most of the difference between the assembled contig length (3.16 Gb) and expected genome size (4.42 Gb).

1.7. Placing centromeres

Since centromeres lack meiotic recombination, their positions on genetic maps are demonstrated as regions made of tightly linked markers which are in reality physically distant. Plotting genetic distances of the markers from the pea high-density genetic map against their locations in the assembly revealed such non-recombining regions for all seven pseudomolecules (Supplementary Figure 5). Marker positions on the high-density skim-GBS map built for the RIL population Pop6[2] were plotted against their position on pea pseudomolecules.

The markers' coordinates were used for labelling centromeres in the assembly. To verify centromere placements in the assembly, their estimated locations were compared with pea chromosomes' morphology and localization of selected unique and repetitive sequences performed using fluorescence in situ hybridization (FISH, Figure 1). *Pisum* centromeres were previously cytogenetically characterized[31,32], revealing their peculiar morphology consisting of extended primary constrictions containing multiple domains of centromeric chromatin. Thirteen families of satellite repeats were found to be associated with centromeres, some of which were specific for a single chromosome while others were amplified on multiple chromosomes[12]. Two examples of these centromeric satellites localized on Caméor chromosomes are provided on Figure 1d; TR11/19 is the centromere-specific satellite, while TR10 is distributed in centromeres as well as in terminal regions of several chromosomes. The most abundant pea satellite PisTR-B is shown on Figure 1e. This repeat provides FISH signals allowing discrimination of all chromosomes within the pea karyotype and construction of their cytogenetic maps. Comparison of centromere positions on these cytogenetic maps to their estimated positions in the assembly revealed their consistent agreement (Figure 1c). Moreover, arrays of several centromeric satellites were found to be correctly placed in the predicted centromeric regions of the pseudomolecules, in line with their centromeric locations revealed by FISH (Figure 1d). We have also investigated positions in the assembly of seven single-copy EST-based FISH markers that were localized inside or close to the primary constrictions of chromosomes 2, 4 and 6 (example of the marker c1722 is provided in Figure 1F). All of them were placed in the assembly to the correct centromere, although four of them were misplaced compared to the FISH data (Figure 1c). These discrepancies probably reflect incorrect placement of some scaffolds due to the repeat-rich nature of centromeric regions.

Further, we tested the level of collinearity by calculating correlation coefficients between the pseudomolecules sequence and high-marker-density genetic maps. For that purpose, we used data derived from RIL populations 4, 5, 6, 7 and 9[2] (Supplementary Data 2).

2. Annotation and characterization of repetitive DNA, genes and miRNA

2.1 Repetitive sequences

2.1.1. Annotation of repetitive sequences

The annotation of repetitive DNA followed both homology-based prediction and *de novo* identification of repeats. We used the REPET package version 2.6[33,34] to identify and annotate repetitive elements in the contigs of the pea genome V1a sequence as described in Supplementary Figure 6. TEdenovo was run on

700 Mb taken from the longest scaffolds of each chromosome to detect repeats present in at least three copies: 200 Mb were aligned onto themselves to identify repeats and RepeatScout[35] was applied to screen the remnant 500Mb for repetitive low complexity DNA. Repeat sequences were then clustered by multiple alignments to produce a library of consensus sequences. Finally, these repeat consensus sequences were classified according to their characteristics and redundancy using PASTEC with Repbase (v20.05). TEannot then mapped the repeat consensus sequence library produced by TEdenovo against the genome using in a two-step approach[36].

The first step identified consensus sequences with at least one full-copy fragment in the genome. The second step identified the copies of these elements in the genome. The annotation of transposon protein domains was further refined using DANTE-Protein Domain Finder, a new tool available at the RepeatExplorer server, which employs LAST searches[37] against custom database of transposon protein domains[29,38]. The hits were filtered to cover at least 80% of the reference sequence, with minimum identity of 35% and minimum similarity of 45%, allowing for max three interruptions (frameshifts or stop codons). The relative amounts of the different repetitive element class, order and family is given in Supplementary Table 7. TE annotation was done on contigs of the pea genome v1a using REPET[33]. TE classes were defined according to Wicker et al.[39], and TE lineages were defined according to Novak et al. [29] (Supplementary Table 7).

2.2 Gene prediction and annotation

2.2.1 Gene prediction and functional database annotation

For gene model prediction, repeats were masked using maskfasta from bedtools 2.26.0[40]. *De novo* prediction was carried out using AUGUSTUS v3.0.3[41] and Fgenesh v7.1.1[42] trained on the *Medicago truncatula* matrix. Protein homology searches were done for different sources of sequences. Protein sequences from *Cicer arietinum* (GA_v1.0), *Glycine max* (275_Wm82.a2.v1), and *Medicago truncatula* (Mt4.0v1) were mapped onto the genome using TBLASTN[43]. Hits with an E-value < $1e^{-50}$ and more than 50 % of the protein length mapped were retained. UniProt and Swissprot sequences were mapped onto the genome using TBLASTN and hits with E-value< 1e-20 were retained. Pea DNA and RNA sequences from IPK (http://pgrc.ipk-gatersleben.de/cr-est/files/pea/) and NCBI (http://www.ncbi.nlm.nih.gov/nucest/?term=(Pisum+sativum)+AND+%22Pisum+sativum%22%5Bporgn%3Atxid3888%5D) were aligned to the genome using BLASTN with an E-value < 1e-50 and identity criteria ≥ 0.98. Retained sequences were analyzed using Exonerate v2.2.0[44] to generate protein-based gene models. Furthermore, to refine the annotation and identify splice junctions, RNA-Seq reads from a series of libraries were aligned to the genome assembly using the ultrafast universal RNA-seq aligner STAR (version STAR_2.4.0j[45]: 20 RNA-Seq libraries from various plant tissues of Caméor at different plant growth stages (188,446,568 reads)[46] and 12 highly dense libraries generated from leaf tissue of cultivar Kaspa inoculated with isolates of the fungal complex causing Ascochyta blight or mock-inoculated (160,332,071 reads)[47] available in NCBI Bioproject PRJNA510273. A set of assembled transcripts were obtained from the alignments using StringTie (v1.2.2)[48] and Trinity-GG (v2.0.6)[49]. Integration of all above gene models and identification of alternative gene splice sites were conducted using the annotation pipeline PASA v2.0.2 which includes Evidence Modeler v1.1.1[50]. In total, the annotation procedure yielded 57,835 transcripts and 44,756 gene models.

Putative gene functions were assigned using the best match to Swiss Prot and TrEMBL databases[51]. Motifs and domains were searched using InterProScan v5[52] against all default protein databases including ProDom, PRINTS, PfamA, SMART, TIGRFAM, PrositeProfiles, HAMAP, PrositePatterns, SITE, SignalP, TMHMM, Panther, Gene3d, Phobius, Coils and CDD. In addition, we used TrapID[53], and the PLAZA 2.5 reference database[54] to assign each transcript to a reference gene family and transfer functional annotation including GO for each transcript. Additionally, an embedded pipeline of EuGene 4.2a[55,56] was launched using the same proteins and RNA-seq databases. This annotation procedure yielded 34,137 gene models and was used to curate gene models manually (Supplementary Table 8).

2.2.2 Non-coding RNA prediction and annotation

Two methods were used to detect putative lncRNA. First, FEELnc[57] was used on StringTie assembled transcripts produced for the gene annotation. NcRNA genes were also predicted by tRNAScan-SE[58], rfamscan (Rfam release 12)[59] and RNAmmer (RDNAs)[60] integrated in the EuGene pipeline. For lncRNA, only elements predicted by the two methods were kept in the annotation.

For the identification of miRNA, developing seeds of 'Caméor' were harvested at two stages (12 days and 22 days after pollination). RNA was purified and small RNA libraries were produced and sequenced according to Lelandais-Briere et al.[61]. Reads were pooled and trimmed using fastx clipper and a minimum length of 15nt. ShortStacks (v3.8.5)[62] was employed to map and identify miRNA. Putative miRNA responding to all ShortStacks miRNA analysis criteria (Y) or without miRNA-star (N15) were mapped against miRbase[63] version 22 mature miRNA sequences using ssearch36 and only alignment with at least 95% of identity were conserved. Only N15 with at least one annotation with a known plant miRNA were kept. Y without annotation were considered as putative new miRNA. This analysis resulted into 54 miRNAs with an annotation featuring 25 different families and 14 putative new miRNAs. Finally, targets were predicted using TargetFinder[64] and kept only if their score was superior at 3. Fifty-nine miRNAs showed at least one putative target (Supplementary Data 3).

3. Genome evolution

In studying the pea genome evolution, we followed three lines of research, one on gene orthology among plant genomes (§3.1), the second on transposable element evolution (§3.2), the other into the reconstruction of the pea paleo-genome (§3.3).

3.1 Comparative gene divergence in the context of the Eudicots and focus on the Leguminosae

3.1.1 Gene orthology

Genome expansion in plants is driven by two major phenomena leading to sharp increases in GS: polyploidization (whole genome replication) and proliferation of transposable elements. Regular gains and deletions of DNA loci contribute to changes in genome size in a milder and continuous manner. As a starting point in the exploration of genome expansion in pea, a whole-genome comparative analysis was conducted based on genome size and homologous relationships between gene-coding loci (CDS) using a dataset composed of all species in the Leguminosae family with sequenced genomes (all in the Papilionoideae clade), and the reference species of the core and basal Eudicot clades (Supplementary Data 4[24,65-96]; Figure 2b).

To identify putative paralogous and orthologous gene clusters, protein-coding gene sets from pea and 21 other Eudicot species, including all sequenced species in the Leguminosae family (all in the Papilionoideae clade), reference species of core Eudicots clades and the basal Eudicot *Nelumbo nucifera* (Supplementary Data 5) were analyzed using Orthofinder v2.1.2 and its defaults parameters[97] with the Diamond v0.9.14 option instead of BLAST[98]. Here homology relationships were inferred based on gene sequence similarity and phylogenetics but not synteny. Orthogroups are defined as clusters with at least two homologous genes in the same or different species and are presumed to derive from a single gene in the common ancestor of the taxa. Prior to the analysis, genome assemblies and annotations were subjected to minor amendments to exclude plastid sequence data, inconsistencies in the headings format between fasta and gff3 files, spurious stop codons or sequences with premature stop codons, and alternative transcripts. In cases where there were two or more transcript variants, the longest transcript was selected to represent the coding region (input data is summarized in Supplementary Data 5, including total number of CDS).

The pea genome ranks fifth in total CDS number (44,791; Supplementary Data 5), after pigeon-pea *Cajanus cajan*, *M. truncatula*, *L. angustifolius*, and *G. max*, which contain 9, 13, 18 and 25% more genes, respectively. The latter two genomes have undergone recent paleo-polyplodization events (Figure 2b). A total of 29,549 candidate gene-families, or orthogroups, were identified and represented 86% of all CDS in the 22-Eudicot dataset. For most species, the great majority of the genes were clustered into orthogroups (Supplementary Data 6). In contrast, the pea genome contained the lowest percentage of genes assigned to gene families, 67% genus-specific genes, followed by *M. truncatula* with 77% (Supplementary Data 6). The number of pea genes clustered into the orthogroups common to all Eudicots and those specific to the Papilionoideae were second lowest, 32% and 0.15%, respectively, following *C. cajan* with 31.5% and 0.14%, respectively.

In addition to the large percentage of genus-specific genes (33% of lineage specific CDS unassigned to clusters, Supplementary Data 5), the pea genome contains the largest proportion of genus-specific orthogroups and genes (656 clusters, 1639 CDS in total; Supplementary Data 5; 3.7% of all 44,791 CDS and 10% of all genus-specific genes; Supplementary Data 5), both indicative of a prolific gene gain process after the Fabeae- Trifolieae divergence estimated to have occurred between 24.7 and 17.5 MYA[99].

3.1.2 Paralogs' sequence divergence

The sequence divergence for all possible pairs of paralogs within each orthogroup (see 3.1.1) was estimated based on Ks. Protein sequences were aligned using MUSCLE v3.8.31 (Edgar, 2004) and converted into codon aligned nucleotides using the bioruby-alignment package[100]. Ks values were calculated through maximum likelihood estimation (MLE) using the 'codeml'[101] and 'yn00'[102] programs in the PAML package[103] and using the following parameters: runmode = -2, set-type = 1 (codon sequences), alpha fixed to 0, codonFreq = 2 (F2X4). To do so, we created an in-memory sqlite database including the whole genome assemblies and annotations to identify pairs of paralogs based on the Orthogroups.csv file (Supplementary Data Table 3.2). For all Ks distribution histograms, the x-axes were drawn in a log-scale with non-transformed Ks values to represent the decreasing relative importance of differences as the Ks value increases resulting from the stochastic nature and saturation of Ks calculations[104]. The range of values, 0.01-50, were binned into 400 interval-bins. To reduce the exponential effect of spurious homologs on background noise, we filtered the data based on orthogroup size. The orthogroup size (*i.e.* number of genes per orthogroup) affects the histograms' shape[105]. The larger the orthogroup size the more likely the

orthogroup includes spurious homologs with Ks values spreading throughout the Ks range (*eg*. groups with 20 or more genes). The histograms in Supplementary Figure 7 represent paralogs pairs in orthogroups of 8 to 20 genes or less. For each species, the orthogroup size was determined based on the genome multiples for events leading to the Eudicot divergence onwards (Supplementary Data 5).

Whole genome paleo-polyploidy events have been described in plants: γ common to all core Eudicots[105], PWGD common to all Papilionoideae within the Leguminosae family[106]; and others that are lineage specific (LS): N-LS[96], S-LS[107], β and α[105,108], SWGD[88], L-LS[107], G-LS[77]. To survey the whole-genome duplication events in the pea genome, the distribution of paralogs pairwise synonymous substitutions (Ks) were plotted for the 22 species in the Eudicot set (Supplementary Figure 7). As expected from numerous earlier studies, no evidence for recent whole-genome replications were observed in pea genome Ks histograms. The bimodal nature of the pea Ks histogram is in line with the paleo-polyploidy events reported for all other sequenced Galegoids genomes: the whole-genome duplication event common to all Papilionoideae (PWGD) estimated to have occurred 55 MYA and, the whole-genome triplication event common to all core Eudicots, Superrosoids and Superasterids (syn. Arabidopsis "gamma"). However, two aspects of the pea Ks histograms are notable. First, the right shift in the pea PWGD-peak compared to other Papilionoideae species (pea mode Ks = 1 and peak range 0.75-1.33, Medicago Ks = 0.83 0.66-1.03; Supplementary Figure 7 and Supplementary Table 9). The pea higher Ks mode is indicative of a high whole-genome mutation rate. The mutation rate is substantially higher than that observed in G. max, often used as a reference (mode Ks = 0.61, and peak range 0.49-0.76).

Another notable aspect of the Ks histogram of pea is the high paralog-pair density of the PWGD-peak left-tail (Ks < 0.4). Similar histograms were evident for low-Ks paralogs in the Trifolieae and the Dalbergieae species (Supplementary Figure 7). However, the species appear to differ in the evolutionary pathways underlying paralogs with low Ks. When the paralog pairs are classified according to taxonomy lineages (Figure 2b, 2d; Supplementary Figure 8), about 75% of the Pisum-specific paralog pairs show Ks ≤ 0.4; similar results were observed for the pea paralogs clustered in orthogroups specific to the *Pisum, Medicago* and *Trifolium* genera, and to a lesser extent, pea paralogs clustered in orthogroups specific to the ILRC clade. Lineage-specific paralog pairs in all other species show higher density close by but at lower Ks than their respective PWGD-peak (Supplementary Figure 7 and 8).

3.2 Diversity of transposable element inter and intra-species

3.2.1 Transposable elements' representation in legume genomes

The representation of TE in the pea genome was compared with published data obtained for other legume species and showed that the larger genome size of pea is largely accounted for by LTR gypsy retroelement expansion as compared to other legumes, as well as SINE expansion as compared with *P. vulgaris, G. max,* and *L. japonicus* (Supplementary Table 10).

3.2.2 Transposable element diversity within Pisum species and subspecies

Resequencing reads (Bioprojects PRJNA285605, PRJNA431567, PRJEB30482) obtained for 3 *P. fulvum* accessions (DCG0494, DCG0076, DCG0704), 2 *P. abyssinicum* accessions (DCG0563,VSD0035), 7 *P. sativum elatius* accessions including 2 *humile* (DCG0705, DCG0707, DCG0708, DCG0706, DCG0771, DCG0557, DCG0709), 10 *P. sativum sativum* landraces *(*DCG0711, DCG0710, DCG0712, DCG0354, DCG0713, DCG0233, DCG0033, MGS0090, DCG0248, DCG0265) and 5 *P. sativum sativum* cultivars (VCL0042,

DCG0528, VKL0099, DCG0112, 'Caméor') were mapped on the genome using NGM by default[109]. Counts were computed using FeatureCounts[110] on specific repetitive sequence lineage domains. The reads' mapping counts onto TE domains were normalized by dividing the number of counts on a specific domain by the total number of counts on all TE domains and by the total number of occurrences of each domain in the pea genome v1a assembly per million. The variation of TE representation among the different *Pisum* species and subspecies was tested by an analysis of variance. Different models were tested by ANOVA: Model1 tested the different TE representation among *P. fulvum/P.sativum* wild*/ P.sativum sativum* groups; Model2, among *P. fulvum/P.sativum* wild*/ P.sativum* landraces*/ P.sativum* cultivars; and Model3 among *P. fulvum/P.sativum* wild*/ P.abyssinicum/P.sativum* landraces*/ P.sativum* cultivars. Results are presented in Supplementary Data 6.

3.2.3. Diversity analysis of retrotransposon protein domains in the pea genome

Regions of LTR-retrotransposon sequences coding for reverse transcriptase (RT) and integrase (INT) protein domains were identified using DANTE as described above (§2.1.1). Sequences shorter that 80% or longer than 120% of the length of the reference sequence were excluded. For construction of phylogenetic tree all-to-all pairwise comparison of corresponding DNA sequences and TN93 model[111] were used to create distance matrix. Phylogenetic tree was then constructed using neighbour-joining algorithm. The tree was created for each lineage of transposable elements separately. For rooting, we included also 10 sequences from other lineages. SIRE Ty1/Copia sequences were included as an outgroup for all other Ty1/Copia trees. For rooting of SIRE Ty1/Copia tree, Angela Ty1/Copia elements were used as outgroup. Ty3/Gypsy trees were similarly rooted using Reina, TatV, Athila or Tekay as outgroups. To estimate relative divergence times of the elements, we calculated ultrametric trees using PATHd8 program (doi: 10.1080/10635150701613783) and relative branching times (Figure 3) were extracted from the trees using R package *ape* (https://doi.org/10.1093/bioinformatics/btg412).

3.3 Reconstruction of the pea paleo-genome

An evolutionary scenario was obtained following the method described in Pont et al.[112] based on synteny relationships identified between between pea (*Pisum sativum)*, peanut diploid ancestor (*Arachis duranensis*[85]), lotus (*Lotus japonicus*[66], barrel medic (*Medicago truncatula*[106]), chickpea (*Cicer arietinum*[113]), pigeonpea (*Cajanus cajan*[76, 114]), soybean (*Glycine max*[77]), common bean (*Phaseolus vulgaris*[78]), mungbean (*Vigna radiata*[80]) and adzuki bean (*Vigna angularis*[81]). Briefly, the first step consisted in aligning the investigated genomes to define conserved/duplicated gene pairs on the basis of alignment parameters (CIP for Cumulative Identity Percentage and CALP Cumulative Alignment Length Percentage). The second step consisted in clustering or chaining groups of conserved genes into synteny blocks (excluding blocks with less than 5 genes) corresponding to independent sets of blocks sharing orthologous relationships in modern species. In the third step, conserved gene pairs or conserved groups of gene-to-gene adjacencies defining identical chromosome-to-chromosome relationships between all the extant genomes were merged into conserved ancestral regions (CARs). CARs were then merged into proto-chromosomes based on partial synteny observed between a subset (not all) of the investigated species. The ancestral karyotype can be considered as a 'median' or 'intermediate' genome consisting of proto-chromosomes defining a clean reference gene order common to the extant species investigated. From the reconstructed ancestral karyotype an evolutionary scenario was then inferred taking into account the

fewest number of genomic rearrangements (including inversions, deletions, fusions, fissions, translocations) which may have operated between the inferred ancestors and the modern genomes (Supplementary Figure 9, Supplementary Table 11).

4 Genome evolution through translocations

4.1 Plant material

Accessions of *P. sativum elatius*, *P. sativum southern humile*, and *P. fulvum* were used to identify translocations possibly involved in the evolution of the *Pisum* genus. The genotypes were chosen as "archetypes" of the following subspecies: '703' for *P. fulvum*, '721' for *P. sativum elatius*, '711' for *P. sativum southern humile*. These genotypes gave, in Ben-Ze'ev and Zohary[115], similar results as the other genotypes from the same respective groups (*P. sativum elatius*, *P. sativum southern humile*, and *P. fulvum*). However, we used 'Caméor' for comparisons and not 'Dunn' as Ben-Ze'ev and Zohary[115]. Nevertheless, our results corroborate the results obtained by these authors with 'Dunn'.

4.2 Methods

In order to identify chromosome translocations, we sequenced single chromosomes isolated by flow sorting from the three accessions and compared the sequences with the sequence assembly of *P. sativum* cv. Caméor. Preparation of suspensions of intact mitotic chromosomes, flow cytometric analysis and sorting was done as described above (§1.4.4). However, before the actual sorting, the chromosome suspensions were purified by sorting all chromosomes into sterile LB01 buffer to decrease the risk of contamination of sorted single chromosome samples. From each genotype, a total of 84 chromosomes (theoretically 12 copies of each chromosome type) were flow sorted and single chromosome DNA amplification was done as described above (§1.4.4). DNA samples were sequenced at Genoscope, Evry, France. Most sequencing libraries (2/3 DNA samples) were prepared as described above (§1.4.4) using the the NEBNext DNA Sample Prep Master Mix kit with a 'on beads' protocol. For samples with low DNA amounts (<250ng), sequencing libraries were prepared using the NEBNext Ultra II DNA library prep kit (New England BioLabs, Beverly, USA) according to manufacturer's protocol. All amplified libraries were then normalized, pooled (7 or 8 libraries per pool) and size selected (around 700-800 bp) using gel electrophoresis. Finally, libraries were subjected to a quality control as described above and sequenced on an Illumina HiSeq 2500 in rapid mode with 2 × 250 base paired end reads, reaching >= 2 million reads per sample. In order to identify which chromosome each sample corresponded to, we mapped the chromosome sequence data onto the genome assembly of *P. sativum* cv. 'Caméor'. We could thus identify samples corresponding to each pseudomolecules of the 'Caméor' assembly. Then, for each genotype and each sample, we plotted the average mapping density (y-axis), along the 'Caméor' assembly (x-axis). Single chromosome reads were also separately assembled using SPAdes[116] and markers from Tayeh et al.[2] mapped onto resulting contigs by BLAT[117] to confirm "breakpoints" in chromosomes 1, 3 and 5 in the respective lines.

4.3 Results

Classification of mitotic chromosomes by flow cytometry resulted in histograms of DAPI fluorescence intensity (flow karyotypes) with characteristic distribution of chromosome peaks (Supplementary Figure

10 A - C). The flow karyotypes of accessions '703', '711' and '721' differed from each other and from cv. 'Caméor' (§1.4.4.2). This indicated differences in relative chromosome DNA content among the four genotypes. In order to increase the probability of collecting a similar number of each chromosome type, several sort windows were set on dot-plots of DAPI fluorescence pulse area vs. fluorescence pulse width (Supplementary Figure 10 D – F).

All samples of MDA amplified DNA were checked for the presence of PisTR-B tandem repeat (see § 1.4.4.2) and a set of samples with the highest amount of DNA were selected for sequencing. The number of samples obtained from each sort window that were sent for sequencing is listed in Supplementary Table 12.

Comparisons of single-chromosome DNA sequences to the 'Caméor' assembly assigned most single chromosomes samples, one to one, to a Caméor pseudomolecule (Supplementary Figure 11). In '703', 5 libraries were assigned to chr1LG6, 3 to chr2LG1, 8 to chr3LG5/chr5LG3, 4 to chr4LG4, 10 to chr5LG3, 6 to chr6LG2, 2 to chr7LG7, and 2 libraries corresponded to a mix of several chromosomes. In '711', 6 libraries were assigned to chr1LG6/chr5LG3, 3 to chr2LG1, 6 to chr3LG5, 5 to chr4LG4, 10 to chr5LG3, 8 to chr6LG2, 7 to chr7LG7, 5 libraries corresponded to a mix of several chromosomes, and one library was off-type. In '721', 12 libraries were assigned to chr1LG6/chr5LG3, 1 to chr2LG1, 5 to chr3LG5, 9 to chr4LG4, 8 to chr5LG3, 4 to chr6LG2, 2 to chr7LG7, and 4 libraries corresponded to a mix of several chromosomes and one library failed. These comparisons revealed that in '711' and '721', part of 'Caméor' chromosome 5 was missing and was associated with chromosome 1. In 703, the same missing part was associated with chromosome 3. The break point was located at 465 Mb at the end of chromosome 5 in '703', '711' and '721'. In '711' and '703', a possible trace of translocation could be seen between chromosome 2 and chromosome 4, i.e. 2 Mb at 445 Mb at the end of chromosome 4 was mapped in samples of chromosome 2 and missing in samples of chromosome 4. In '703', there was also a short segment of 2 Mb missing on chromosome 3 (at 454 Mb) that was mapped on chromosome 3 in samples corresponding to chromosome 1. Synteny between pea and *M. truncatula* provides some clues to the probable ancestral state of these chromosomes. Chromosome 5 of pea was syntenic with chromosome 3 of *M. truncatula* from 0 to 467 Mb, and to chromosome 2 of *M. truncatula* from 467 Mb to the end. This break point roughly corresponds to the translocation break point (2 Mb difference). Moreover, *M. truncatula* chromosome 2 was mainly syntenic with pea chromosome 1. These results together with the phylogenetic tree obtained, allowed orienting the translocation events. The *Medicago-Pisum* common ancestor probably had the same karyotype, for the translocated segment (i.e., 465 Mb to the end of 'Caméor' chromosome 5), as '721' and '711'. A hypothesis would be that it evolved from chromosome 1, as in elatius/humile, to chromosome 3 in fulvum on the one hand, and to chromosome 5 in cultivated peas on the other hand. Furthermore, complex chromosome pairing during meiosis in hybrids could be explained by other short translocated chromosomal fragments.

5. Pisum diversity

5.1 Plant material and resequencing

The genomes of 44 accessions were used to study the pea genome diversity (Supplementary Data 7). Sixteen genotypes, including Caméor, were re-sequenced as described in Tayeh et al.[2], as part of the ANR

program GENOPEA (Bioproject PRJNA285605). Another 16 genotypes were re-sequenced in the Pisdom Burgundy region PARI project (FABER M. Siol, Bioproject PRJNA431567). Nuclear DNA was extracted using the Floraclean Plant DNA isolation kit as recommended by MP Biomedicals (www.mpbio.com). A Quality Control (QC) was performed for all DNA samples i.e they were checked for concentration by fluorometric measurement with Quant-iT™ PicoGreen®(Invitrogen) and for quality by measuring absorbance and checking electrophoretic profile on agarose gel. Illumina paired-end shotgun indexed libraries were prepared from one µg of DNA per genotype, using the TruSeq DNA PCR-free LT Sample Preparation Kit (Illumina Inc., https://www.illumina.com/). Briefly, library preparation was performed with low sample protocol and fragment size 350 bp. DNA fragmentation was performed by using AFA (Adaptive Focused Acoustics™) technology on focused-ultrasonicator E210 (Covaris), all enzymatic steps and clean up was realized according to manufacturer's instructions. The resulting indexed libraries, including the ligated adapter sequences, had a mean size of 564 bp (BioAnalyser® QC on Agilent 2100 High Sensitivity DNA chip). According to manufacturer's instructions, paired-end sequencing 2 × 100 sequencing by synthesis (SBS) cycles was performed on a HiSeq® 2000, TruSeq® V3 chemistry running in high output mode after cluster generation on a cBot™ system of Illumina. Additionally, three genotypes (DSP, 90-2131, Kiflica; Bioproject PRJNA509279) were sequenced by a commercial company (NovoGene, Sacramento, CA) using Illumina HiSeq, paired-end 150 bp from 350 bp insert DNA libraries and three accessions ('703', '711', '721') were resequenced at GENOSCOPE on an HiSeq2500 using the Nextera Mate Pair Sample preparation kit (Illumina) as described in §1.2 (Bioproject PRJEB30482; Note that '721' had also been sequenced in the Pisdom project). Public resequencing data for seven accessions were used.

5.2 Phenotypic evaluation

All pea re-sequenced genotypes, except Zhongwan6 for which we had no seeds, were evaluated in the glasshouse for classical growth and development traits (Supplementary Data 7). Two pots per accessions and 6 seeds per pot were sown in February 2017 in 7 L pots filled with 30% attapulgite and 70% clay balls and supplied with a 0.625 mmol N nutritive solution through drips. Temperature in the glasshouse was tempered as follows: temperatures below 18°C day/14°C night triggered heating and temperatures above 22°C day/18°C night triggered cooling. Light was supplemented to 500 W/m² during a 16h-photoperiod. As they grew, plants were tied onto 2.1 m high bamboo sticks. In total, 47 classical morphological and phenological traits were scored on the 44 genotypes: seed shape (smooth, wrinkled), flower wings color, presence of bracts (yes=1, no=0), number of flowers per node, type of leaves (afila=0, normal=1), shape of leaflets (oval / elliptical), leaflet shape according to UPOV (elliptical El, oval Ov, oval base Ovb), leaflet form (truncated / rounded / pointed), leaf shape (pointed =3, rounded =2, truncated =1), leaflet serration (serrated=1, not=0), leaflet serration (absent 0, weak 1, average 5, strong 7, very strong 9), number of leaflets at the bottom of the plant, number of leaflets at the top of the plant, shape of stipules (oval / elliptical), shape of stipules (serrated), presence of macule on leaf (no=0, few=1, many=5), presence of anthocyanin coloration in the plant (yes=1, no=0), presence of an anthocyanin ring at the base of stipule (Antho1), presence of a double anthocyanin ring at the base of stipule (Antho2), presence of anthocyanin on the stipules (Antho3), presence of anthocyanin at the base of leaflets (Antho4), presence of anthocyanin on leaflets (Antho5), presence of anthocyanin on stems (Antho6), presence of anthocyanin on pods (Antho7), presence of anthocyanin on peduncles (Antho8). We also scored the date of beginning of flowering (calendar days), the height at beginning of flowering, the date of beginning of seed filling

(calendar days), the date of flowering (calendar days), pod dehiscence at harvest (yes=1, no=0), presence of callus on pods, the number of basal ramifications, the height of first flowering node, the height of last flowering node, the number of apical ramifications, the number of 1st flowering node, the number of last flowering node, the total number of nodes, the number of pods per node, the length of peduncle, the shoot width, seed color (beige/gray=1, green=2, brown/purple/black=3), seed ornamentation (absence=0, presence=1), seed shape (round/ovoid=1, dented=2, wrinkled=3), hilum color (clear=0, dark=1), presence of dark dot next to the hilum (yes = 1), cotyledon color (yellow=1, green=2), Pod form (slightly arched=1 arched=2 straight=3, very straight=4), pod form (truncated=1, pointed=2), pod length, pod width. These classical traits were scored on two plants per accession. Productivity traits, i.e. the number of pod per plant, the number of seeds per plant, the seed weight per plant, were averaged over all harvested plants (Supplementary Data 7). Furthermore, seed protein content (%DM) was measured according the Kjehldhal and seed protein composition (PA2%, PA1%, convicilin/vicilin%, legumin%) was quantified by FPLC as described in Bourgeois et al.[118]. Just after seed harvesting, germination tests were done: five seeds per genotype were assessed on Milli-Q water for one week. The rate of germination was recorded each day after imbibition, for one week. This was repeated three times, successively. A principal component analysis (PCA) was done using these phenotypic data (Figure 5). Interestingly, the rate of germination significantly contributed to the first axes of the PCA (Supplementary Data 7) while pod dehiscence, a trait considered instrumental in domestication had a lesser contribution.

5.3 Mapping, SNP detection and filtering

Resequencing data for the 43 accessions of *Pisum* and the accession of *Lathyrus sativus* were mapped onto the pea genome v1a assembly using BWA MEM[119], keeping only unique mappings with a quality higher or equal to 30. Optical duplicates were removed with PICARD tools (http://picard.sourceforge.net/). Altogether, 95,326,251 SNPs were called using BCFtools 1.6[119] mpileup and call. All callings supported by less than three reads were re-imputed. All markers which were homozygous or heterozygous in 'Caméor' as compared to the reference were deleted using SNPSift[120]. We produced two different datasets depending on the type of analysis to be conducted. For phylogenetic analysis, 2,026,659 SNPs with less than 5 missing data and 10 heterozygotes were filtered using vcftools[121] and plink[122] (Phylogeny SNP dataset). For diversity analysis, 17,212,608 SNPs with less than 10 missing data and 10 heterozygotes were filtered (Diversity SNP dataset). In this dataset, accessions L180 and zongwhan6 were removed.

5.4 SNP diversity and phylogenetic analyses

The 'Phylogeny' SNP dataset was used to build a phylogenetic tree of the 44 accessions using IQ-Tree 1.6[123]. TVM+R10 was selected as the best model for a maximum likelihood tree using Modelfinder[124]. The tree was inferred with 1000 replicates of ultra-fast likelihood bootstrap[125] and SH-aLRT test to obtain bootstrap branch support values. The phylogeny gave a useful picture of the relationships between *Pisum* subspecies. The *Pisum fulvum* accessions clearly clustered apart from the other *Pisum* accessions, confirming the species level of this clade. *Pisum sativum* accessions were clustered consistently according to former subspecies designations, such a *P. asiaticum* or *P. transcaucasicum,* or within cultivated peas, *P. arvense* or *hortense*. Germination ability was a useful criterion to differentiate cultivated from wild accessions. The position of *P. abyssinicum* in the tree indicated that this taxon resulted from a domestication event in the *P. s. elatius* genepool independent from the domestication event that gave rise

to *P.s. sativum*. For further analyses, we divided the *Pisum* clade into three groups that exhibited differential levels of diversity. The 'wild' group included genotypes whose freshly harvested seeds did not germinate in water (*P. fulvum, P. s. elatius,* and *P.s. humile*), the 'landrace' group included traditional accessions from different regions of the world, and the 'cultivar' group included modern pea cultivars, either cultivated for dry seeds or for green seeds. Due to its specific status as being near wild peas from an evolutionary point of view, yet presenting domesticated atributes, the *P. abyssinicum* accessions were either gathered with wild or with landrace accessions according to the analysis.

We computed the number of alleles present in the different *Pisum* groups using the 'Diversity' dataset. An in-house script was used to transform SNP information into alleles coded in an allele dose 012 format. The VennCounts function of the R package limma[126] was used to calculate Venn diagrams for each group.

Nucleotide diversity (Pi[127]) was computed using vcftools with a windows of 500kb and a step of 100kb. Tajima's D[128] was computed using VCF-kit[129] with the same windows parameters as nucleotide diversity (https://github.com/AndersenLab/VCF-kit, Supplementary Figure 12)

We further investigated the level of linkage disequilibrium in the different accession groups using the 'diversity' dataset (Supplementary Figure 13). R2 was computed using PLINK within sliding windows of 10Mb. Haploblocks of LD were computed for each group of accessions using plink with an adequate maximum size for blocks "--blocks-max-kb" (10Mb for cultivars, 6Mb for landraces, 2.5Mb for wild accessions).

5.5 Chloroplast sequence diversity

A chloroplast sequence available in GenBank (KJ806203.1) was used as reference to reconstruct *Pisum sativum* cv. 'Caméor' chloroplast using 30X reads from the genome sequencing with MITObim 1,7[130]. The comparison between the two chloroplast sequences yielded only 2 SNPs.

Using KJ806203.1 as reference, re-sequencing reads obtained for 38 out of 44 accessions were aligned using bwa-mem by default. Publicly available resequencing reads for six accessions were depleted in chloroplast reads and were not included into the analysis. GATK[131] 'best practices' pipeline was run to call 4128 SNPs. Using this dataset, a phylogenic tree was computed using RAXml[132] with a GTR GAMMA plus models and 1000 bootstraps.

The phylogenetic tree based on chloroplast polymorphism supported the scenario of *Pisum* evolution provided by nuclear SNP and translocations (Supplementary Figure 14).

6 Seed storage protein genes

A list of storage protein sequences was developed by combining sequences retrieved from the pea gene atlas, UNIPROT and NCBI and searched for homologies in the pea genome assembly (Supplementary Data 4). Candidate sequences were manually curated using protein alignments, RNA-seq data and gene models by EuGene. Known regulatory motifs were searched in the 5' region of the identified gene models (Supplementary Data 4). Best homologs were identified in Uniprot and in the M. truncatula genome v4 were also searched to check synteny relationships.

The basic and acidic polypeptides are released after the cleavage of the pre-protein polypeptide at highly conserved sites[133]. Different cleavage motifs were found in the predicted seed storage protein (SSP) genes of the pea genome: in legumins, most encoded for the GLEETIC motifs, though one encodes GLEETVC and

another FLEETVC. In Vicilin genes, the SLK and KED motifs are the most frequent (Supplementary Figure 15).

The expression profile of twenty-eight legumin, vicilin, convicilin, PA1 and PA2 encoding genes was assessed using high throughput real-time quantitative PCR using the Biomark microfluidic system from Fluidigm. All sample-gene combinations were quantified using a 96.96 Dynamic Array™ IFCs (BMK-M-96.96, Fluidigm). Pre-amplification of the samples, chip loading and real time quantitative PCR were performed according to manufacturer's protocol. Real time quantitative PCR results were analyzed using the Fluidigm real-time PCR analysis software v.4.1.3. Primers used are listed in Supplementary Data 4. Expression was normalized as in Alves-Carvalho et al.[46] and primers used for reference genes were for actin F: CTAAGGGTGAATATGATGAGTCTGG, R:GAGACACCAAAAAAGCAACCACATC; for Histone H1, F:CAGCTGTGAAGAAAGTTGCTGCG, R:CTAAACTCTCATTTCCTTCCACCTC; for Elongation Factor 1 alpha, F: GGAACAACTTGTGCAGAAGCAACC, R:GTCATCAAGAGTGTGGAGAAGAAGG. Mean expression levels were calculated over three biological replicates for all developing seed tissues at 8, 12, 16, 19, 23, 29 days after pollination, and for shoots, roots and nodules at stage A, flowers at stage B, and germinating seeds 3 days after imbibition. Two biological replicates were used for roots, nodules, leaves at Stage B, and upper leaves at Stage C. One biological replicate was used for apical nodes at Stage B, apical nodes, pods and stems at Stage C. Stages are as described in Alves Carvalho et al.[46]

7. Data Management and Visualisation

Annotation lift-over between different version of the assembly was done using CrossMap[134]. JCVI utilities (https://github.com/tanghaibao/jcvi) assembly were used to manage goldenpath, bed, fasta and gff format. Visualisation of data was done using ggplot2[135] package on R. Circos was used to build circular plot[136].

8. References

1. Baurens, F.C., Bonnot, F., Bienvenu, D., Causse, S., Legavre, T. Using SD-AFLP and MSAP to assess CCGG methylation in the banana genome. *Plant Mol. Biol. Rep*. **21**, 339-348 (2003).
2. Tayeh, N. et al. Development of two major resources for pea genomics: the GenoPea 13.2 K SNP Array and a high-density, high-resolution consensus genetic map. *Plant J.* **84**, 1257-1273 (2015).
3. Doležel, J. et al. Plant genome size estimation by flow cytometry: inter-laboratory comparison. *Ann. Bot*. **82,** (Suppl. A), 17–26 (1998)
4. Liu, Y., Schröder, J. & Schmidt, B. Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data. *Bioinformatics*, **29,** 308-315 (2012).
5. Luo, R. et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**, 18 (2012).
6. Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578-579 (2011).
7. Gali, K.K et al. Development of a sequence-based reference physical map of pea (*Pisum sativum* L.) *Front. Plant Sci.* doi: 10.3389/fpls.2019.00323 (2019)
8. Madoui, M. A. et al. MaGuS: a tool for quality assessment and scaffolding of genome assemblies with Whole Genome Profiling™ Data. *BMC Bioinformatics* **17**, 115 (2016).
9. van Oeveren, J. et al. Sequence-based physical mapping of complex genomes by whole genome profiling. *Genome Res.* **21**, 618-625 (2011).
10. Li, R. et al. The sequence and *de novo* assembly of the giant panda genome. *Nature* **463**, 311–317 (2010).

11. Li, R. et al. *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res*. **20**, 265–272 (2010).
12. Neumann, P., Pozárková, D., Vrána, J., Dolezel, J., Macas, J. Chromosome sorting and PCR-based physical mapping in pea (*Pisum sativum* L.). *Chromosome Res*. **10,** 63-71 (2002).
13. Doležel, J., Binarová, P. & Lcretti, S. Analysis of nuclear DNA content in plant cells by flow cytometry. *Biol. Plant.* **31**, 113-120 (1989).
14. Cápal, P., Blavet, N., Vrána, J., Kubaláková, M. & Doležel, J. Multiple displacement amplification of the DNA from single flow–sorted plant chromosome. *Plant J.* **84**, 838-844 (2015).
15. Neumann, P., Nouzová, M., Macas, J. Molecular and cytogenetic analysis of repetitive DNA in pea (*Pisum sativum* L.). *Genome* **44**, 716-28 (2001).
16. Alberti, A. et al. Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Sci. Data* **4,** 170093(2017).
17. Bayer, P. E. et al. High-resolution skim genotyping by sequencing reveals the distribution of crossovers and gene conversions in *Cicer arietinum* and *Brassica napus*. *Theor. Appl. Genet*. **128**, 1039-1047 (2015).
18. Lorenc, M. et al. Discovery of single nucleotide polymorphisms in complex genomes using SGSautoSNP. *Biology* **1**, 370-382 (2012).
19. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
20. Šimková, H., Číhalíková, J., Vrána, J., Lysák, M. A. & Doležel, J. Preparation of HMW DNA from plant nuclei and chromosomes isolated from root tips. *Biol. Plant.* **46**, 369-373 (2003).
21. Staňková, H. et al. BioNano genome mapping of individual chromosomes supports physical mapping and sequence assembly in complex plant genomes. *Plant Biotechnol. J.* **14**, 1523-1531 (2016).
22. Cao, H. et al. Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. *GigaScience* **3,** 34 (2014).
23. Tang, H. et al. ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biol.* **16,** 3 (2015).
24. Tang, H. et al. An improved genome release (version Mt4.0) for the model legume *Medicago truncatula*. *BMC Genomics* **15**, 312 (2014).
25. Lamprecht, H. Further studies of the linkage group Cp—Gp—Fs—Ast of *Pisum sativum. Agri Hort. Genet.* **6**, 1–9 (1948)
26. Blixt, S. Cytology of *Pisum.* III. Investigation of five interchange lines and coordination of linkage groups with chromosomes. Agri Hort. Genet. **17**, 47-75 (1959).
27. Lamm, R. & Miravalle, R. J.. A translocation tester set in *Pisum*. *Hereditas* **45**, 417-440 (1959).
28. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333-i339 (2012).
29. Novák, P., Neumann, P., Pech, J., Steinhaisl, J. & Macas, J. RepeatExplorer: A Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* **29**, 792–793 (2013).
30. Novák, P. et al. TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Res.* **45**, e111 (2017)
31. Neumann, P. Stretching the rules: monocentric chromosomes with multiple centromere domains. *PLoS Genet.* **8**, e1002777 (2012).
32. Neumann, P. et al. Centromeres off the hook: massive changes in centromere size and structure following duplication of CenH3 gene in Fabeae species. *Mol. Biol. Evol.* **32**, 1862-1879 (2015).
33. Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H. Considering transposable element diversification in *de novo* annotation approaches. *PloS One* **6**, e16526 (2011).
34. Quesneville, H. et al. Combined evidence annotation of transposable elements in genome sequences. *PLoS Comput. Biol.* **1**, e22 (2005).
35. Price, A. L., Jones, N. C., & Pevzner, P. A. *De novo* identification of repeat families in large genomes. *Bioinformatics* **21,** i351-i358 (2005)
36. Jamilloux, V., Daron, J., Choulet, F. & Quesneville, H. *De novo* annotation of transposable elements: Tackling the fat genome issue. *Proc. IEEE* **105**, 474-481 (2107).
37. Kiełbasa, S. M., Wan, R., Sato, K., Horton, P., Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome Res*. **21**, 487-493 (2011).

38. Novák, P., Neumann P. & Macas, J. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics* **11,** 378 (2010).
39. Wicker, T. et al. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973 (2007).
40. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
41. Stanke, M., Schoffmann, O., Morgenstern, B., Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics.* **7**, 62 (2006).
42. Solovyev, V., Kosarev, P., Seledsov, I. and Vorobyev, D. Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol.* **7**, S10 (2006).
43. Gertz, E. M. et al. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC Biology* **4** 41 (2006)
44. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
45. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
46. Alves-Carvalho, S. et al. . Full-length *de novo* assembly of RNA-seq data in pea (*Pisum sativum* L.) provides a gene expression atlas and gives insights into root nodulation in this species. *Plant J.* **84**, 1-19 (2015).
47. Turo C. J. Genomic analysis of fungal species causing ascochyta blight in field pea. PhD Thesis Curtin University (2016).
48. Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T. & Salzberg, S. L. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotech.* **33,** 290 (2015).
49. Grabherr M. G. et al. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotech.* **29,** 644-52 (2011).
50. Haas, B.J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol*. **9**, R7 (2008).
51. The UniProt Consortium. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res*. **39**, D214-D219 (2011).
52. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240 (2014).
53. Van Bel, M., Proost, S., Van Neste, C., Deforce, D., Van de Peer, Y. & Vandepoele, K. TRAPID: an efficient online tool for the functional and comparative analysis of *de novo* RNA-Seq transcriptomes. *Genome Biol.* **14**, R134 (2013).
54. Van Bel, M. et al. Dissecting plant genomes with the PLAZA comparative genomics platform. *Plant Physiology* **158**, 590-600 (2012).
55. Foissac, S. et al. Genome annotation in plants and fungi: EuGene as a model platform. *Current Bioinformatics* **3**, 87-97 (2008).
56. Badouin, H. et al. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature* **546**, 148-152 (2017).
57. Wucher, V. et al. FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Res.* **45**, e57-e57 (2017).
58. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955 (1997).
59. Nawrocki, E. P. et al. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.* **43**, D130-D137 (2014).
60. Lagesen, K. et al. 2007 RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100-3108 (2007).
61. Lelandais-Brière C. et al. Genome-wide *Medicago truncatula* small RNA analysis revealed novel microRNAs and isoforms differentially regulated in roots and nodules. *Plant Cell* **21,** 2780-96 (2009).
62. Johnson, N. R., Yeoh, J. M., Coruh, C. & Axtell, M. J. Improved placement of multi-mapping small RNAs. *G3: Genes, Genomes, Genet.* **6**, 2103-2111 (2016).
63. Griffiths-Jones, S., Saini, H. K., van Dongen, S. & Enright, A. J. miRBase: tools for microRNA genomics. *Nucleic Acids Res.* **36**(suppl_1), D154-D158 (2007).

64. Bo, X. & Wang, S. 2004. TargetFinder: a software for antisense oligonucleotide target site selection based on MAST and secondary structures of target mRNA. *Bioinformatics* **21**, 1401-1402 (2004).
65. De Vega, J. J. et al. Red clover (*Trifolium pratense* L.) draft genome provides a platform for trait improvement. *Sci. Rep.* **5**, 17394 (2015).
66. Sato, S. et al. Genome structure of the legume, *Lotus japonicus*. *DNA Res.* **15**, 227-239 (2008).
67. Doležel, J., Greilhuber, J., & Suda J. Estimation of nuclear DNA content in plants using flow cytometry. *Nat. Protoc.* **2,** 2233-2244 (2007).
68. Arumuganathan, K. & Earle E. D. Nuclear DNA content of some important plant species. *Plant Mol. Biol. Report* **9,** 208-218 (1991).
69. Vižintin, L., Javornik, B. & Bohanec B. Genetic characterization of selected *Trifolium* species as revealed by nuclear DNA content and ITS rDNA region analysis. *Plant Sci.* **170,** 859-866 (2006).
70. Kaur, P. et al. An advanced reference genome of *Trifolium subterraneum* L. reveals genes related to agronomic performance. *Plant Biotech. J.* **15,** 1034-1046 (2017).
71. Parween, S. et al. An advanced draft genome assembly of a desi type chickpea (Cicer arietinum L.). *Sci. Rep.* **5,** 12806 (2015).
72. Galasso, I. et al. Chromatin characterization by banding techniques, *in situ* hybridization, and nuclear DNA content in *Cicer* L.(Leguminosae). *Genome* **39,** 258-265 (1996).
73. Gupta, S. et al. Draft genome sequence of *Cicer reticulatum* L., the wild progenitor of chickpea provides a resource for agronomic trait improvement. *DNA Res.* **24,** 1-10 (2016).
74. Cheng, R. I. J. & Grant, W. F. Species relationships in the *Lotus corniculatus* group as determined by karyotype and cytophotometric analyses. *Canadian Journal of Genetics and Cytology* **15,** 101-115 (1973).
75. Greilhuber, J. & Obermayer, R. Genome size variation in *Cajanus cajan* (Fabaceae): a reconsideration. *Plant Syst. Evol.* **212,** 135-141 (1998).
76. Varshney, R. K. et al. Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nat. Biotech.* **30,** 83-89 (2012).
77. Schmutz, J. et al. Genome sequence of the paleopolyploid soybean. *Nature* **463**, 178–183 (2010).
78. Schmutz, J. et al. A reference genome for common bean and genome-wide analysis of dual domestications. *Nat. Genet.* **46,** 707-713 (2014).
79. Parida, A., Raina, S. N. & Narayan, R. K. J. Quantitative DNA variation between and within chromosome complements of *Vigna* species (Fabaceae). *Genetica* **82,** 125-133 (1990).
80. Kang, Y. J., et al. Genome sequence of mungbean and insights into evolution within *Vigna* species. *Nat. Commun.* **5,** 5443 (2014).
81. Kang, Y. J. et al. Draft genome sequence of adzuki bean, *Vigna angularis*. *Sci. Rep.* **5,** 8069 (2015).
82. Naganowska, B. et al. Nuclear DNA content variation and species relationships in the genus *Lupinus* (Fabaceae). *Ann. Bot.* **92,** 349-355 (2003).
83. Hane J. K. et al. A comprehensive draft genome sequence for lupin (*Lupinus angustifolius*), an emerging health food: insights into plant–microbe interactions and legume evolution. *Plant Biotechnol. J.* **15,** 318–330 (2017).
84. Temsch, E. M. & Greilhuber J. Genome size in *Arachis duranensis*: a critical study. *Genome* **44,** 826-830 (2001).
85. Bertioli, D. J. et al. The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nat. Genet.* **47**, 438-446 (2015).
86. Samoluk, S. S. et al. First insight into divergence, representation and chromosome distribution of reverse transcriptase fragments from L1 retrotransposons in peanut and wild relative species. *Genetica* **143,** 113-125 (2015).
87. Verde, I. et al. The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat. Genet.* **45,** 487-494 (2013).
88. Tuskan, G. A. et al. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). S*cience* **313,** 1596-1604 (2006).
89. Bennett, M. D. & Smith J. B. Nuclear DNA amounts in angiosperms. *Phil. Trans. R. Soc. Lond. B* **334,** 309-345 (1991).

90. Lamesch, P. et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40**, D1202-D1210 (2011).
91. Figueira, A., Janick, J. & Goldsbrough, P. Genome size and DNA polymorphism in *Theobroma cacao*. *J. Am. Soc. Hortic. Sci.* **117,** 673-677 (1992).
92. Motamayor, J. C. et al. The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. *Genome Biol.* **14,** r53 (2013).
93. French-Italian Public Consortium for Grapevine Genome Characterization. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**,463-467 (2007).
94. Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485,** 463-467 (2012).
95. Diao, Y. et al. Nuclear DNA C-values in 12 species in Nymphaeales. *Caryologia* **59** 25-30 (2006)
96. Ming, R. et al. Genome of the long-living sacred lotus (*Nelumbo nucifera* Gaertn.). *Genome Biol.* **14,** R41 (2013).
97. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol*. **16,** 157 (2015).
98. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59-60 (2014).
99. Lavin, M., Herendeen, P. S. & Wojciechowski, M. F. Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the tertiary. *Syst. Biol.* **54,** 575-594 (2005).
100. Bonnal, R. J. P. et al. Biogem: an effective tool-based approach for scaling up open source software development in bioinformatics. *Bioinformatics* **28,** 1035-1037 (2012).
101. Goldman, N. & Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11,** 725-736 (1994).
102. Yang, Z. & Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17,** 32-43 (2000).
103. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586-91 (2007).
104. Vanneste, K, Van de Peer, Y, Maere, S. Inference of genome duplications from age distributions revisited. *Mol. Biol. Evol.* **30,** 177–190 (2013).
105. Bowers, J. E., Chapman, B. A., Rong, J., & Paterson, A. H. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**, 433–438 (2003).
106. Young, N. D. et al. The *Medicago* genome provides insight into the evolution of rhizobial symbioses. *Nature* **480**, 520-524 (2011).
107. Cannon, S. et al. Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. *Mol. Biol. Evol.* **32**, 193–210 (2015)
108. Vision TJ, Brown DG, Tanksley SD. The origins of genomic duplication in Arabidopsis. Science 290, 2114–2117 (2000)
109. Sedlazeck, F. J., Rescheneder, P. & Von Haeseler, A. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics* **29,** 2790-2791 (2013).
110. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30,** 923-930 (2013).
111. Tamura, K. & Nei, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol*. **10** 512-526 (1993)
112. Pont, C., Wagner, S., Kremer, A., Orlando, L., Plomion, C. & Salse, J. Paleogenomics: reconstruction of plant evolutionary trajectories from modern and ancient DNA. *Genome Biol.* **20**, 29 (2019)
113. Varshney, R. K. et al. Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nat. Biotech.* **31**, 240-246 (2013).
114. Singh, N. K. et al. The first draft of the pigeonpea genome sequence. *J. Plant Biochem. Biotechnol.* **21**, 98-112 (2012).
115. Ben Ze'ev, N. & Zohary, D. Species relationships in the genus *Pisum* L. *Israel J. Bot.* **22**, 73-91 (1973).

116. Bankevich, A. et al. "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing." *J. Comput. Biol.* **19,** 455-477 (2012).
117. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656-664 (2002).
118. Bourgeois, M. et al. A PQL (protein quantity loci) analysis of mature pea seed proteins identifies loci determining seed protein composition. *Proteomics* **11**, 1581-1594 (2011).
119. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**, 2078–2079 (2009).
120. Cingolani, P. et al. Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Front. Genet.* **3**, 35 (2012).
121. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
122. Purcell, S. et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet*. **81**, 559–575 (2007).
123. Nguyen, L. T. et al. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32.** 268-274 (2014).
124. Kalyaanamoorthy, S. et al. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14,** 587–589 (2017).
125. Hoang, D. T. et al. UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35,** 518-522 (2017).
126. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43,** e47-e47 (2015).
127. Nei, M. *Molecular Evolutionary Genetics*. (Columbia University Press, New York, 1987).
128. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585-595 (1989).
129. Cook, D. E. & Andersen, E. C. 2017. VCF-kit: assorted utilities for the variant call format. *Bioinformatics*, **33**, 1581-1582 (2017).
130. Hahn, C., Bachmann, L. & Chevreux, B. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Res.* **41,** e129-e129 (2013).
131. Van der Auwera, G. A. et al. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinf.* **43,** 11-10 (2013).
132. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30,** 1312-1313 (2014).
133. Casey, R. & Domoney, C. Pea globulins. In *Seed Proteins* (eds Shewry, P.R., Casey R.) 171-208 (Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999).
134. Zhao, H. et al. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30,** 1006-1007 (2013).
135. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer, New York, 2016).
136. Krzywinski, M. I. et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* 19, 1639-1645 (2009).