the plant journal

SEB Society for Experimental Biology

The Plant Journal (2015) 84, 1–19 doi: 10.1111/tpj.12967

RESOURCE

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

Susete Alves-Carvalho[1,†], Grégoire Aubert[1,†], Sébastien Carrère[2], Corinne Cruaud[3], Anne-Lise Brochot[1], Françoise Jacquin[1], Anthony Klein[1], Chantal Martin[1], Karen Boucherot[1], Jonathan Kreplak[1], Corinne da Silva[3], Sandra Moreau[2], Pascal Gamas[2], Patrick Wincker[3], Jérôme Gouzy[2] and Judith Burstin[1,*]

[1]*Institut National de la Recherche Agronomique, UMR1347, 17 rue Sully, BP 86510, 21065 Dijon Cedex, France,*
[2]*Laboratoire des Interactions Plantes Micro-Organismes, Institut National de la Recherche Agronomique/Centre National de la Recherche Scientifique, 24 chemin de Borde Rouge, 31326 Castanet Tolosan, France, and*
[3]*GENOSCOPE, 2 Rue Gaston Crémieux, 91000 Evry, France*

Received 21 April 2015; revised 9 July 2015; accepted 16 July 2015; published online 22 August 2015.
*For correspondence (e-mail judith.burstin@dijon.inra.fr).
†These authors contributed equally to this work.

SUMMARY

Next-generation sequencing technologies allow an almost exhaustive survey of the transcriptome, even in species with no available genome sequence. To produce a Unigene set representing most of the expressed genes of pea, 20 cDNA libraries produced from various plant tissues harvested at various developmental stages from plants grown under contrasting nitrogen conditions were sequenced. Around one billion reads and 100 Gb of sequence were *de novo* assembled. Following several steps of redundancy reduction, 46 099 contigs with N50 length of 1667 nt were identified. These constitute the 'Caméor' Unigene set. The high depth of sequencing allowed identification of rare transcripts and detected expression for approximately 80% of contigs in each library. The Unigene set is now available online (http://bios.dijon.inra.fr/FATAL/cgi/pscam.cgi), allowing (i) searches for pea orthologs of candidate genes based on gene sequences from other species, or based on annotation, (ii) determination of transcript expression patterns using various metrics, (iii) identification of uncharacterized genes with interesting patterns of expression, and (iv) comparison of gene ontology pathways between tissues. This resource has allowed identification of the pea orthologs of major nodulation genes characterized in recent years in model species, as a major step towards deciphering unresolved pea nodulation phenotypes. In addition to a remarkable conservation of the early transcriptome nodulation apparatus between pea and *Medicago truncatula*, some specific features were highlighted. The resource provides a reference for the pea exome, and will facilitate transcriptome and proteome approaches as well as SNP discovery in pea.

Keywords: next-generation sequencing, *de novo* assembly, *Pisum sativum* L., gene expression atlas, nitrogen symbiotic fixation, nodule development.

INTRODUCTION

In recent years, much effort and resources in the field of plant functional biology have been devoted to the development of generic genomic tools for a selected number of model species chosen for their biological amenability to sequencing (small genomes), transformation (infection by *Agrobacterium*) and phenotyping (small plant size and short life cycle). The function of many genes has been identified through knockout mutant analysis. High-throughput sequencing technologies now provide an opportunity to fill the gap between model and cultivated species. Pea (*Pisum sativum* L.) is one of the most cultivated grain legumes worldwide, and is particularly adapted to temperate areas. Pea seeds are a valuable source of protein for the human diet as well as for animal feed. Being a legume, pea is also an environmentally friendly crop, due to its ability to establish a symbiosis with soil nitrogen-fixing bacteria from the *Rhizobiaceae* family. By fixing atmospheric nitrogen through plant–rhizobium symbiosis, pea crops allow a reduction in the use of nitrogen fertilizers in cropping systems. This reduces agricultural energy costs and minimizes production of greenhouse gases, as the equivalent of 1.5 tonnes of oil is needed to apply 1 tonne of nitrogen to a fertilized crop (McLaughlin *et al*., 2000). In addition, introduction of a grain legume crop to crop rotations benefits biodiversity of cropping systems and reduces disease and pest transmission. Therefore, pea is an important species for enhancing ecological services of agro-systems and food security. However, with a genome of approximately 4.5 Gb, pea remains a genomic orphan. Genomic data, and, in particular, knowledge of the genes expressed in this species, are essential for deciphering the genetic control of important traits, such as adaptation to climate change, tolerance to pests and diseases, seed composition and quality, and nitrogen-fixing efficiency.

Understanding the establishment and functioning of the symbiosis between pea and its bacterial nitrogen-fixing partner is a prerequisite for optimizing its efficiency. The molecular determinants of the beneficial interaction between leguminous plants and nitrogen-fixing bacteria have been widely studied in the model plants *Medicago truncatula* and *Lotus japonicus* (Schultze and Kondorosi, 1998; Den Herder and Parniske, 2009; Madsen *et al*., 2010; Oldroyd *et al*., 2011; Popp and Ott, 2011; Oldroyd, 2013; Udvardi and Poole, 2013; Roux *et al*., 2014). This symbiosis is established through specific molecular communication between the plant and the bacteria: the plant roots exude flavonoid signals that are perceived by rhizobia, which in turn secrete Nod factors. Perception of Nod factors by specific receptors in the roots triggers a signaling cascade involving a calcium spiking signal and expression of a range of transcription factors, eventually leading to formation on the root of a novel organ called a nodule and its subsequent bacterial colonization. Rhizobia released in the plant cells are surrounded by membranes of plant origin, forming the so-called symbiosomes. After a differentiation process (Kondorosi *et al*., 2013), hosted rhizobia fix nitrogen and supply reduced nitrogen to the plant, which in return provides carbohydrates, proteins and minerals to the bacteria, as well as the low-oxygen environment required for functioning of the bacterial nitrogenase. Unlike *Lotus japonicus* and tropical legumes, but similarly to *Medicago truncatula*, pea develops indeterminate nodules hosting *Rhizobium leguminosarum* symbionts that provide the plant with amides (Den Herder and Parniske, 2009). The nitrogen-fixing symbiosis process is complex, and assessing the generic nature of its determinants between model legumes and pea is of major importance for applied genetics in the context of agroecology.

Gene expression atlases have been developed for a number of plant species, and have allowed identification of the expression networks underlying important plant biological processes. Most of these gene atlases were developed using microarrays or high-throughput RNA sequencing (RNA seq) data mapped onto a reference genome (Benedito *et al*., 2008; Li *et al*., 2010; Severin *et al*., 2010; Zenoni *et al*., 2010; Garg *et al*., 2011; Lan *et al*., 2013; Verdier *et al*., 2013; O'Rourke *et al*., 2014). A few recent studies have used *de novo* assembly of RNA-seq data (Sierro *et al*., 2013; Zhang *et al*., 2013). Indeed, next-generation sequencing technologies, with massively parallel sequencing producing tens of gigabases of sequences in a single run, provide unprecedented perspectives for species lacking a sequenced genome. *De novo* assembly of next-generation sequencing reads allows the discovery of almost all expressed genes in a plant tissue. RNA-seq datasets have recently been produced for several plants. In pea, normalized libraries for seeds, meristems, flowers, leaves, cotyledons, seedlings, epi- and hypocotyls (Franssen *et al*., 2011), young and mature leaves, stems, immature pods and seeds (Kaur *et al*., 2012), and young plantlets (Duarte *et al*., 2014) have been sequenced. However, these experiments were designed for SNP discovery and did not produce a reference Unigene set that is easily available to researchers, nor quantitative gene expression profiles, because of library normalization.

Here, we describe the development of a new resource for pea, and more generally for legume research. After *de novo* assembly of high-throughput sequences obtained for 20 cDNA libraries encompassing various below- and above-ground plant tissues, and for various developmental stages and nutritive conditions, we produced a full-length Unigene set of expressed sequences. This Unigene set was then used to develop the pea RNA-seq gene atlas, and to identify major putative regulators of pea nitrogen-fixing symbiosis.

RESULTS

More than a billion short read sequences of high quality (fewer than 1% of reads contained N), corresponding to approximately 100 Gb sequence, were produced by high-throughput Illumina sequencing of 20 pea cDNA libraries (Table 1).

Construction of a pea Unigene set and deployment of the pea gene atlas portal

Three strategies for *de novo* assembly of RNA-seq data were evaluated in terms of recovery of the most complete full-length cDNA sequence set. The first assembly included 402 871 contigs with a N50 length of 2724 nt, the second included 144 194 contigs with a N50 length of 2323 nt, and the third included 98 688 contigs with a N50 length of 2407 nt (Table 2). Only contigs longer than 200 nt and

Table 1 Description of the 20 libraries that were submitted to RNA-seq

Description	Organ	Stage	Nutrition	CNS identifier	Sequencer	Paired-end read number	Read length (bp)
RootSys_A_HN	Root system	A	High-nitrate, hydroponics	BOSU	GAIIx	34 196 953	108
RootSys_A_LN	Root system	A	Low-nitrate, hydroponics	DOSU	GAIIx	34 335 068	109
Root_B_LN	Roots	B	Low-nitrate, hydroponics	EOSU	GAIIx	34 163 532	109
Root_F_LN	Roots	F	Low-nitrate, aeroponics	NOSU	HiSeq2000	94 570 863	104
Nodule_A_LN	Nodules	A	Low-nitrate, hydroponics	QOSU	GAIIx	37 744 938	109
Nodule_B_LN	Nodules	B	Low-nitrate, hydroponics	ROSU	GAIIx	34 754 968	109
Nodule_G_LN	Nodules	G	Low-nitrate, aeroponics	POSU	HiSeq2000	94 304 059	104
Shoot_A_HN	Shoot	A	High-nitrate, hydroponics	AOSU	GAIIx	25 090 649	108
Shoot_A_LN	Shoot	A	Low-nitrate, hydroponics	COSU	GAIIx	35 506 890	109
Leaf_B_LN	Leaves	B	Low-nitrate, hydroponics	FOSU	GAIIx	37 847 422	109
LowerLeaf_C_LN	Lower leaves[a]	C	Low-nitrate, hydroponics	IOSU	GAIIx	32 113 899	109
UpperLeaf_C_LN	Upper leaves[b]	C	Low-nitrate, hydroponics	KOSU	GAIIx	38 067 818	109
Tendril_BC_LN	Tendrils	B + C	Low-nitrate, hydroponics	ACOSU	HiSeq2000	79 866 150	104
Stem_BC_LN	Stems	B + C	Low-nitrate, hydroponics	ABOSU	HiSeq2000	86 854 716	104
Peduncle_C_LN	Peduncles	C	Low-nitrate, hydroponics	AAOSU	HiSeq2000	83 334 538	104
ApicNode_B_LN	Apical node	B	Low-nitrate, hydroponics	SOSU	GAIIx	36 562 361	109
Flower_B_LN	Flowers	B	Low-nitrate, hydroponics	TOSU	GAIIx	33 452 129	109
Pods_C_LN	Pods[c]	C	Low-nitrate, hydroponics	VOSU	HiSeq2000	90 949 002	104
Seeds_12dap	Seeds	E	High-nitrate, pots	LOSU	GAIIx	36 355 226	109
Seed_5dai	Seeds	D	Water	GOSU	GAIIx	38 680 145	109

Stage A represents 7–8 nodes, 5–6 opened leaves; stage B represents the start of flowering; stage C represents 20 days after the start of flowering; stage D represents germination, 5 days after imbibition; stage E represents 12 days after pollination; stage F represents 8 days after sowing; stage G represents 18 days after sowing, i.e. 10 days after inoculation.

[a] Nodes below the flowering node (N−1, N−3, N−5).

[b] Nodes above the flowering node (N + 1, N + 3, N + 5).

[c] Young, shiny, green 2 cm pods.

exhibiting no N were retained. We evaluated the level of redundancy and/or potential alternative transcription by searching homologous sequences of 100 randomly chosen sequences within each assembly (Table 3). We further evaluated the quality of the three assemblies by comparison with 30 known pea gene sequences obtained from Genbank. This allowed us to detect the presence of the corresponding transcripts, assess the size and structure of open reading frames, and detect transcript variants and errors such as chimeras for contiguous genes with overlapping 3′ UTR, trans-self chimeras as described by Yang and Smith (2013), and some artifact tandem repeats added during transcriptome assembly, as reported on the OASES users forum (http://listserver.ebi.ac.uk/pipermail/oases-users/2012-December/000294.html) (Table 3 and Figure S1). The results of these quality controls are summarized in Table 3. The first assembly, obtained using increasing *k*-mers and by running all library datasets separately, was the most efficient in recovering the expected full-length spliced form, and also spliced and unspliced variants (Table S1). It also contained more errors and chimeras. In the second assembly, obtained using a unique *k*-mer of 75 and by running all library datasets separately, transcripts were usually found in their expected spliced form, but splice variants

Table 2 Statistics for the pea RNA-seq contig assemblies produced in this study or publicly available

	Contig number	N50 value	Median contig length	Mean contig length	Maximum contig length	Minimum contig length	Number of sequences with N
Test assembly multiple *k*-mer	402 871	2724	1790	2090	75 916	200	0
Test assembly *k*-mer 75, per library	144 194	2323	1621	1852	86 503	200	0
Test assembly *k*-mer 75, all libraries	98 688	2407	1470	1742	21 929	200	0
Merged ORF dataset	735 782	1377	720	964	16 197	201	0
ORF 1st cleaning step	203 717	1380	645	924	16 197	201	0
ORF 2d cleaning step	79 144	1407	708	1017	16 596	202	0
ORF 3d cleaning step	52 477	1590	780	1127	16 601	202	0
PsCameor_Unigene	46 099	1666	867	1199	16 601	142	0
PsCameor_Uni_Low copy	40 204	1725	985	1277	16 601	203	0
PsCameor_Uni_High copy	5704	686	525	661	7212	203	0
PsCameor_Organelles	191	980	493	712	2726	142	0
PsCamTri1E_LC	49 017	1483	628	988	14 685	201	0
PsCamTri2E_LC	31 823	1627	948	1206	15 531	201	0
Franssen assembly (before)	81 449	418	245	336	6258	41	22 979
Kaur assembly (before)	100 078	375	266	277	6587	3	0
CR-EST pea sequences (before)	9376	612	612	572	1562	11	2928
Duarte assembly (before)	68 581	956	705	844	5250	200	133
NCBI *Pisum* sequences (before)	18 511	554	509	497	1171	30	1466
Franssen assembly (after)	43 194	460	271	418	5204	200	0
Kaur assembly (after)	60 608	391	346	413	6587	200	0
CREST pea sequences (after)	8065	612	612	583	1562	200	0
Duarte assembly (after)	68 448	956	705	844	5250	200	0
NCBI *Pisum* sequences (after)	17 638	553	512	507	1171	200	0

The first three assemblies are the test assemblies obtained by running the multiple k-mer strategy, the *k*-mer of length 75 separately on the 20 libraries, and the *k*-mer of length 75 on all libraries; the ORF assemblies are the total ORF assemblies obtained after the multiple *k*-mer strategy before redundancy reduction, and after each step of redundancy reduction; the PsCam_Unigene assembly was obtained after all the cleaning steps; the PsCam_HighCopy and PsCam_LowCopy assemblies correspond to genes present in high copy number and in low copy number in the genome; PsCam_Organelles includes contigs corresponding to mitochondria and chloroplast sequences; PsCam_-Tri1E_LC and PsCamTri2E_LC are the assemblies obtained using the Trinity program; the 'Franssen', 'Kaur' and 'Duarte' assemblies and the CR-EST and NCBI sequences are published pea sequence data before and after sequences smaller than 200 nt and/or containing N were discarded.

were missing. The third assembly, obtained using a unique *k*-mer of 75 and by running all library datasets together, contained the fewest errors and noise but was also the least exhaustive: mostly unspliced or partially spliced variants were found.

In order to build the pea Unigene set, we designed a pipeline taking into account the results of the previous tests (Figure 1). The Velvet program (https://www.ebi.ac.uk/~zerbino/velvet/) was run using the multiple *k*-mer strategy on the various libraries separately, in order to obtain the most exhaustive contig assembly. Because tests had shown that Velvet/Oases assemblies contain contigs representing multiple splicing forms of the same gene and also chimeras, the second step extracted the longest ORF from all contig families through several successive clustering steps, so that only one representative of each gene is included in the Unigene set. After each step, the redundancy was evaluated by searching homologs of 30 randomly chosen sequences in the assembly using FASTA (http://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml). The goal was to obtain the least redundant dataset possible. At the end of the redundancy reduction process, we obtained 52 477 extended ORF sequences, including ORFs with 200 nt at the 3′ and 5′ ends, that defined the pea

Table 3 Quality control of the assemblies obtained using four *de novo* assembly strategies

	1st strategy	2nd strategy	3rd strategy	Unigene strategy
Completeness	+++	+	+	+++
Full length	+++	+	+	+++
Expected spliced forms	+++	+	–	+
Alternative forms	++	–	–	–
Unspliced forms	+	+	+	+
Errors	+	–	–	–
Redundancy	+++	++	+	–
Noise	+++	++	+	–

Completeness, full length and errors were estimated by comparing assemblies to 30 known gene sequences. Potential alternative transcription and noise, i.e. sequences with no open reading frame, were estimated from 100 randomly chosen sequences within each assembly. +++, high; ++, medium; +, low; –, weak.

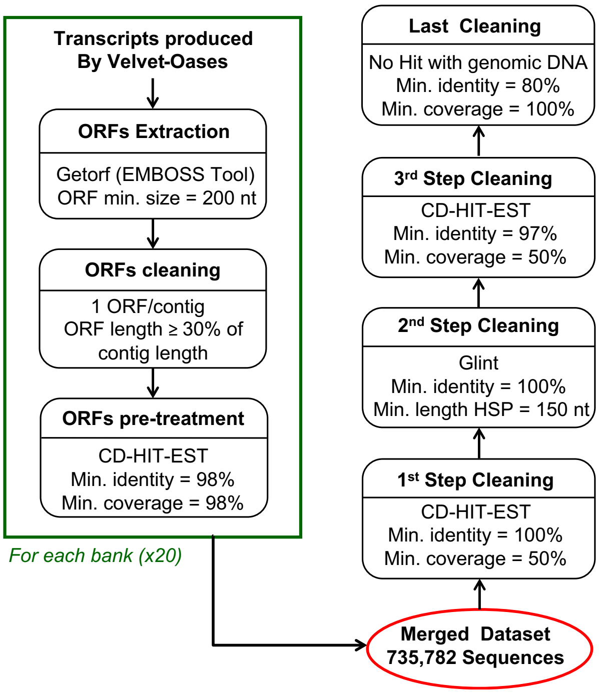


Figure 1. Pipeline of production of the pea Unigene set.

Unigene set. The statistics for the pea RNA-seq contig assemblies and the distribution of sequence lengths through the various steps are shown in Table 2. Annotation by homology of the 52 477 ORF sequences revealed that some contigs were highly homologous to non-plant genes. Using pea genomic sequence data for the 'Caméor' genotype (approximately 35 Gb were available when the Unigene set was produced; C. Aluome, M.C. Le Paslier and D. Brunel, INRA UR1279, personal communication), we removed from the ORF set all contigs that did not match any genomic read, i.e. 6378 sequences (Table 2), including 2742 that were annotated by homology with the NCBI nucleotide collection database (http://www.ncbi.nlm.nih.gov/). Many of these annotated contigs corresponded either to rhizobium-expressed or freshwater eukaryote-expressed genes, which were found especially in below-ground organ libraries, or virus-expressed genes, which were found especially in leaf libraries. However, because some pea genes may have been absent from genomic reads and discarded in the process due to the moderate depth of genomic sequences (approximately sevenfold the genome), this set of sequences is made available for searching as 'PsCam_Discarded' in the pea gene atlas portal (http://bios.dijon.inra.fr/FATAL/cgi/pscam.cgi). Comparison of the Unigene set with pea genomic sequence data further allowed us to identify genes present in high copy number in the pea genome (mean coverage higher than 50 reads per nucleotide). Many of these contigs displayed diverse transposase- or retroelement-encoding motifs.

Altogether, the *de novo* assembly pipeline provided two Unigene sets: a low-copy-number Unigene set containing 40 204 contigs (PsCam_LowCopy, Table 2), and a high-copy-number Unigene set consisting of 5704 contigs (PsCam_HighCopy, Table 2). Another 191 organelle sequences were added to the Unigene set. Following deployment of the pea Unigene data in the Functional Analyses porTAL (http://lipm-bioinfo.toulouse.inra.fr/download/FATAL/), the pea Unigene set is now publicly available at http://bios.dijon.inra.fr/FATAL/cgi/pscam.cgi. For each contig, a peptide and nucleotide annotation sheet are available and different queries are possible (FASTA/BLAST/PATSCAN). The contig annotation is provided on the peptide sheet. The annotation was performed both by searching for protein-encoding domains and by using the annotation of the nearest homologs in public databases. The expression profiles of the contigs may also be visualized in the gene atlas part of the website. The reliability of the gene atlas expression profiles was checked by comparing the expression level of ten genes with contrasting patterns, as determined (i) by RNA-seq (RPKMnorm) and (ii) by quantitative RT-PCR, using three biological replicates. The two methods were found to give very similar profiles (mean $r^2 = 0.91$, Figure 2). In order to obtain an overview of the biological functions and metabolic pathways of the contigs that are differentially expressed between pea tissues, the annotation tool Mercator (http://mapman.gabipd.org/web/guest/app/mercator) was used. This tool allowed assignment of MapMan bin codes for each of the 40 395 PsCam_LowCopy and organelle contigs. All contigs (16 812 sequences) that were annotated by Mercator but lacked an original InterProScan annotation (Quevillon *et al.*, 2005) were assigned the bin code '35.3_not assigned.disagreeing hits'.

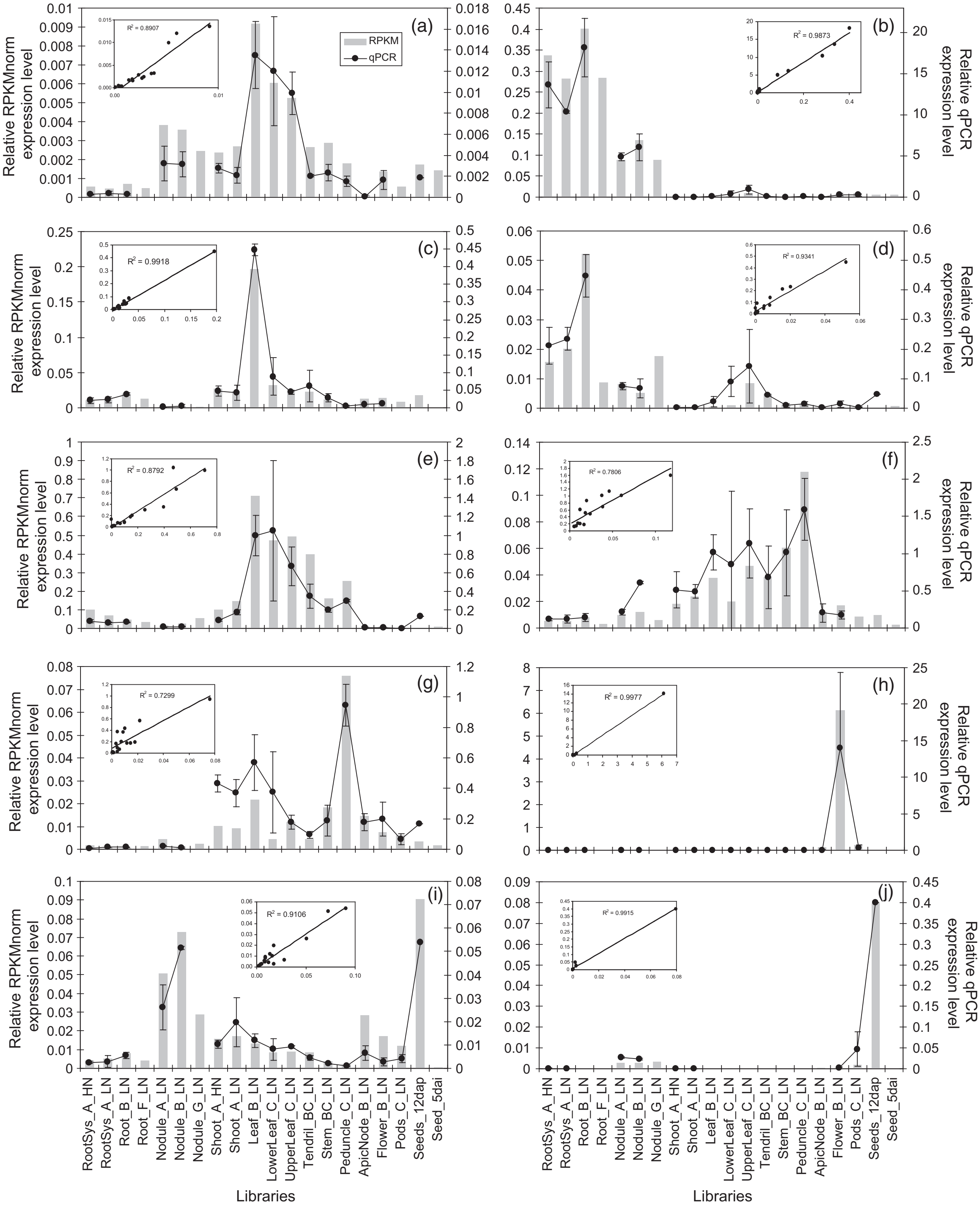


Figure 2. Validation of gene atlas RNA-seq expression profiles by quantitative PCR for (a) PsCam036305, (b) PsCam042603, (c) PsCam030773, (d) PsCam007587, (e) PsCam050022, (f) PsCam042625, (g) PsCam033628, (h) PsCam044575, (i) PsCam042632 and (j) PsCam036471.
Inserts show the correlation between RPKMnorm and the quantitative PCR expression levels.

Among the 23 583 remaining annotated contigs, including approximately 5% of the bin codes corrected manually, 18 476 contigs were assigned a putative function.

Finally, we addressed the question of isoforms in the PsCam_LowCopy Unigene set using the OrthoMCL method (Li *et al.*, 2003), which defines transcript families, searches for orthologous families between two species, and provides information about paralogous families in the species. Using the *M. truncatula* genome (version 4.0, http://jcvi.org/medicago) as a reference, we found that similar numbers of transcripts were singletons or part of multi-gene families (Table S2) in the two species.

Evaluation of the pea Unigene set

The PsCam_LowCopy contig length and number were compared to two low-copy assemblies obtained using the Trinity program (Grabherr *et al.*, 2011): PsCamTri1E_LC and PsCamTri2E_LC (Table 2 and Figure S2). PsCam_LowCopy exhibited a higher mean and median contig length than PsCamTri1E_LC and a higher contig number than PsCamTri2E_LC. Only PsCam_LowCopy was retained for further evaluation. However, a number of PsCam_LowCopy contig sequences contain artifact micro-repeats, probably due to selection of the longest ORF, that are not present in the Trinity assemblies.

In order to assess the representativeness of the Unigene set, we estimated to what extent previous pea RNA-seq assemblies or gene datasets were represented by its contigs, and, conversely, to what extent the Unigene set contigs were represented by previous pea sequence datasets. First, we applied the same trimming as that used in our pipeline to the publicly available sequence datasets: all sequences exhibiting one N or more were excluded, and contigs shorter than 200 nt were discarded (Table 2). We then performed blast analysis (http://blast.ncbi.nlm.nih.gov/Blast.cgi) of pea sequences from Franssen *et al.* (2011), Kaur *et al.* (2012), Duarte *et al.* (2014), the CR-EST database (http://pgrc.ipk-gatersleben.de/cr-est/) and the NCBI (http://www.ncbi.nlm.nih.gov/) database against the pea Unigene set (Table 4), and found that 98, 93, 98, 98 and 97% of the 'Franssen', 'Kaur', 'Duarte', CR-EST and NCBI pea sequences, respectively, showed a hit (*e*-value = 10^{-4}) with at least one pea Unigene sequence. This suggests that the pea Unigene set developed herein includes the vast majority of previously described sequences. This is reinforced by the fact that, among the sequences that showed no hit with the pea Unigene set, only 30% had a functional annotation in 'Franssen' sequences, 43% in CR-EST sequences, 1.2% in 'Kaur' sequences and 3% in 'Duarte' sequences. Conversely, 'Franssen', 'Kaur', 'Duarte', CR-EST and NCBI sequences showed a hit with 54, 77, 79, 17 and 40%, respectively, of the Unigene sequences. Our pea Unigene set thus identified many pea sequences that were not found in previous pea transcript assemblies. We further assessed the quality and representativeness of the Unigene set by estimating the proportion of predicted peptides from six published genomes that were represented by the pea Unigene set (Table 5): between 25 636 and 29 233 predicted peptide homologs (*e*-value = 10^{-10}) from the most distant species *Arabidopsis thaliana* to the legume species *C. cajan* were found in the pea Unigene set. Among them, between 19 697 and 25 299 were best reciprocal homologs (threshold *e*-value = 10^{-10} both ways), suggesting that they are orthologs.

Different pea transcriptomes are revealed by the low-copy-number and high-copy-number Unigene sets

The distribution of contigs' Gene Ontologies (GO) from the low-copy-number and high-copy-number Unigene sets revealed a different constitution of the two sets (Figure S3). Whereas all cellular processes and molecular function classes were represented in the low-copy-number Unigene set, only a few were highlighted in the high-copy-number Unigene set, notably nucleic acid metabolic processes. Annotations by homology available for 1136 contigs from the high-copy-number Unigene set revealed a large proportion of contigs encoding reverse transcriptases, gag capsid proteins and polynucleotidyl transferases, and a small proportion of transposases. The level of expression of contigs also varied greatly between the two sets. A mean 93–63% of PsCam_LowCopy contigs

Table 4 Comparison of the PsCam_Unigene library with published pea sequences

	Databases versus Unigene set				Unigene set versus databases	
	Hits	Hits (%)	No hit	No hit annotated	Hits	Hit (%)
Franssen assembly	42 285	98	769	233	24 858	54
Kaur assembly	56 286	93	4105	50	35 396	77
CR-EST pea sequences	7936	98	130	56	7865	17
Duarte assembly	66 782	98	1666	56	36 145	78
NCBI *Pisum* sequences	17 161	97	478	478	18 378	40

Published pea sequence data from Franssen *et al.* (2011), Kaur *et al.* (2012), Duarte *et al.* (2014), CR-EST and NCBI sequences were compared by blast ($e = 10^{-4}$) to the PsCam_Unigene contigs, and, conversely, PsCam_Unigene contigs were compared by blast to these sequences. The table indicates the number and proportion of public sequences showing homology with PsCam_Unigene contigs, the number of public sequences with no hit on the PsCam_Unigene contigs, and, among these, the number of sequence that were not annotated, and, conversely, the number and proportion of PsCam_Unigene contigs showing homology with public sequences.

Table 5 Comparison of the PsCam_Unigene library with to published plant genome sequences

	Peptide number	Median peptide length	Mean peptide length	Maximum peptide length	Minimum peptide length	No of hits with PsCam_LowCopy*	No of reciprocal hits with PsCam_LowCopy**
Cajca	48 680	242	318	5241	49	29 233	25 299
Lotja	37 971	186	257	4294	12	26 708	23 823
Medtr	62 319	275	353	5386	17	29 112	23 188
Glyma	75 778	265	336	24 182	12	29 128	20 618
Chickpea	28 268	321	387	5341	49	28 014	25 274
Arath	35 386	349	409	5393	16	25 636	19 697

The predicted cDNA sequences from the genomes of *Cajanus cajan* (Cajca), *Lotus japonicus* (Lotja), *Medicago truncatula* (Medtr), *Glycine max* (Glyma), chickpea (*Cicer arietinum*) and *Arabidopsis thaliana* (Arath) retrieved from public databases were compared with the PsCam_Unigene library by uni-directional and bi-directional BLAST ($e = 10^{-10}$). The table indicates the statistics for the cDNA sequences of the various species and the number of uni-directional and bi-directional hits.

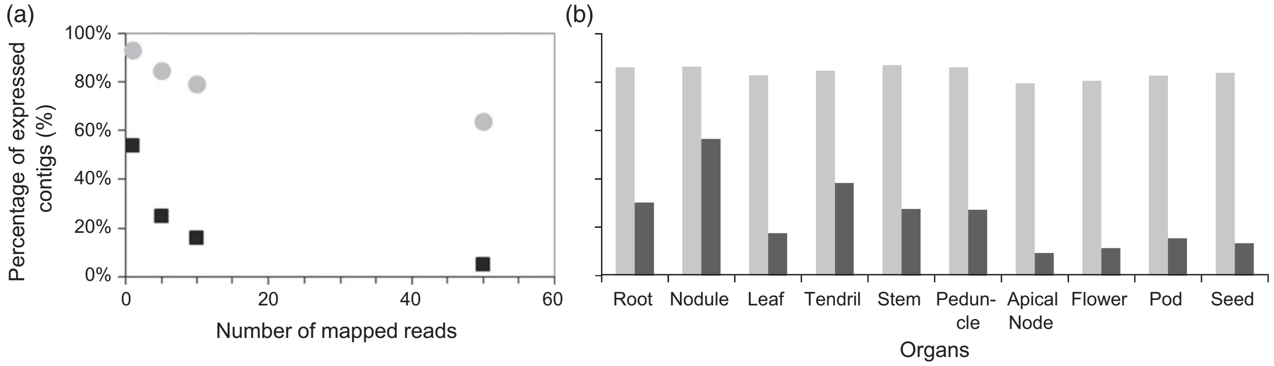


Figure 3. Mean percentage of expressed contigs in each library (%).
(a) Expressed contigs according to the number of mapped reads.
(b) Expressed contigs according to the organ, with a threshold of five mapped reads per expressed contig.
Gray, percentage of expressed PsCam_LowCopy contigs; black, percentage of expressed PsCam_HighCopy contigs.

were expressed over the 20 libraries, with a threshold of 1–50 mapped reads, respectively, while only 54–5% of PsCam_HighCopy contigs appeared to be expressed with the same thresholds, suggesting a lower representation of PsCam_HighCopy contigs (Figure 3). Between 86 and 92% of PsCam_LowCopy contigs and between 16 and 66% PsCam_HighCopy contigs were expressed with at least five reads mapped per expressed contig (Figure 3b). The clustering of the various organs upon principal component analysis of PsCam_LowCopy RPKMnorm revealed a clear specialization of gene expression among nodules, roots, seeds, and above-ground vegetative tissues (Figure 4a), whereas principal component analysis of PsCam_High Copy RPKMnorm mainly distinguished nodule and peduncle libraries from the others (Figure 4b). A heatmap of Pearson correlations of PsCam_LowCopy expression levels (Figure 5) clearly separated nodules from other plant tissues. Among these, roots and seeds appeared closer to each other than to above-ground vegetative tissues, which were further sub-divided into a group comprising young and/or meristematic tissues (flowers, pods, apical nodes and shoots) and a group comprising mature tissues (leaves, tendrils, peduncles and stems).

All PsCam_LowCopy transcripts were classified into five groups according to their expression level and tissue specificity: not differentially expressed (fold change < 3 between the tissues showing the highest and the second highest expression levels), preferentially expressed (fold change ≥ 3 and <10), very preferentially expressed (fold change ≥ 10 and <100), specifically expressed (fold change ≥ 100 and <1000) and very specifically expressed in one tissue (fold change ≥ 1000). Transcripts exclusively expressed in one tissue but with low expression (RPKMnorm < 0.2) were excluded from this classification and referred to as ‘low expressed’. The specificity classes included 4036, 1974, 661 and 86 genes, respectively, that were preferentially, very preferentially, specifically or very specifically expressed in one tissue (Figure S4 and Table S3). Seeds, flowers and nodules exhibited the highest number of specifically and very specifically expressed contigs, followed by the apical node and roots (Table S3). The distribution of specifically and very specifically expressed transcripts according to their functional annotation using the BLAST2GO program (https://www.blast2go.com/) is shown in Figure S5. A higher representation of tricarboxylic acid/organic acid transformation, redox and

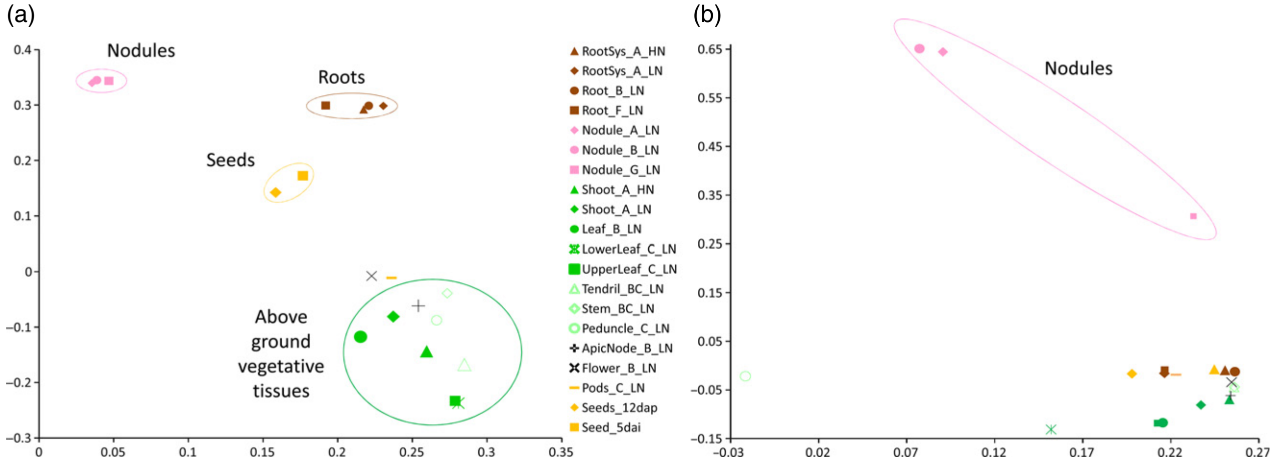


Figure 4. Principal component analysis of RPKMnorm for the PsCam_LowCopy Unigene set (a) and the PsCam_HighCopy Unigene set (b) in the 20 libraries.

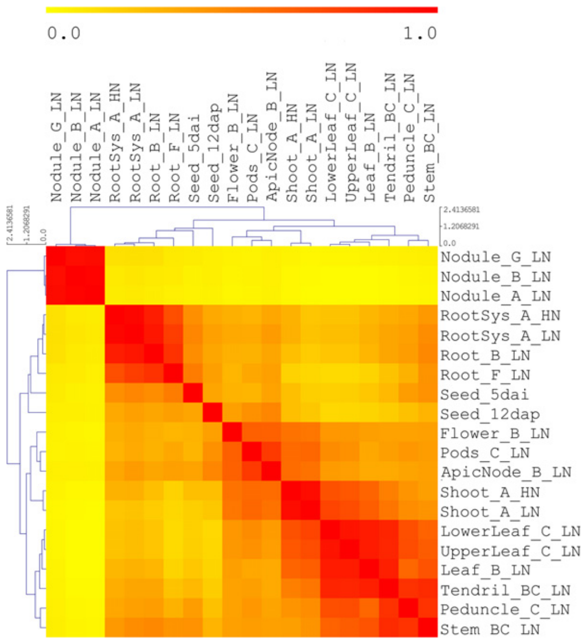


Figure 5. Comparison of the transcriptomes of the various pea tissues. RPKMnorm pairwise Pearson correlation coefficients were used to generate the heatmap. The color scale indicates the degree of correlation. Samples were clustered on the basis of Euclidean distance using MultiExperiment Viewer software (http://www.tm4.org/mev.html).

nucleotide metabolism molecular functions was found for nodule-enhanced genes. Flowers showed a higher representation of genes involved in the biodegradation of xenobiotics, major carbohydrate metabolism, mitochondrial electron transport, and cell wall metabolism than other organs. Among the 2207 pea transcripts identified as encoding transcription factors, 237, 93 and 21 transcription factors, respectively, were preferentially, very preferentially and specifically expressed in one pea tissue compared to the others (Figure S4 and Table S4). Figure S6 shows the distribution of the families of the 114 very preferentially and specifically expressed transcription factors, and suggests that different transcription factor families may be preferentially expressed in different organs.

The pea gene atlas reveals the transcriptional apparatus of the nodulated pea root

K-means hierarchical clustering of the PsCam_LowCopy contigs identified 20 gene expression profiles among the various organs and conditions (Figure S7 and Table S5). Most clusters showed up-regulated expression in specific organs or group of organs according to the expected relatedness of these organs as revealed by the principal component analysis and hierarchical clustering (Figures 4 and 5). Transcripts up-regulated in the three nodule libraries were grouped in cluster 1, and transcripts that were more specifically up-regulated in Nodule_G were grouped in cluster 6, possibly related to differences of expression in aeroponic and hydroponic conditions. Transcripts up-regulated in roots were included in clusters 13 and 16; those up-regulated in seeds 12 days after pollination, germinating seeds 5 days after imbibition, flowers and apical nodes were grouped in clusters 3, 4, 18 and 20, respectively. Transcripts expressed in aerial vegetative organs were often grouped together: shoots and leaves in cluster 7, leaves at stages B and C in clusters 8 and 9, and stem, tendrils, peduncles in clusters 10 and 11. Interestingly, cluster 17 contained 1803 contigs that are preferentially expressed in young meristematic organs (apical nodes, flowers, pods and seeds at 12 days after pollination). Contigs corresponding to the cell cycle, cell organization, DNA repair and synthesis, and protein folding were well represented in this cluster. Cluster 5 included contigs that are up-regulated in the Root_B_LN and Leaf_B_LN samples, and showed a high representation of transcripts involved in signaling; cluster 14 contained contigs up-regulated in both nodules and seed tissues at 12 days after pollination, and showed a high representation of transcripts corresponding to mitochondrial electron transport/ATP synthesis.

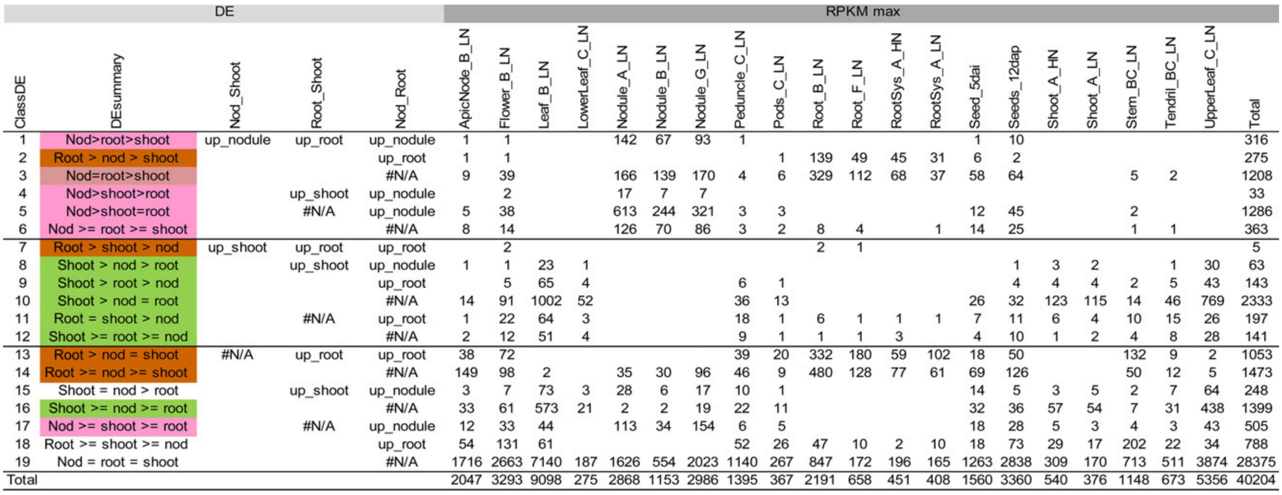
A statistical analysis of differential expression of PsCam_LowCopy transcripts among shoot (including shoot and leaf libraries), root and nodule tissues using DEseq (http://www-huber.embl.de/users/anders/DESeq/) revealed differential regulation of numerous contigs (Figure 6). A combination of 19 differential expression patterns was derived from the three comparisons between nodule and shoot, root and shoot, and nodule and shoot (Figure 6). Differential expression classes 1, 4, 5, 6 and 17 included 2503 transcripts that were significantly up-regulated in nodules, classes 2, 7, 13 and 14 included 2806 transcripts that were up-regulated in roots, classes 8, 9, 10, 11, 12 and 16 included 4276 transcripts that were up-regulated in shoots, while class 19 comprised the vast majority that were not differentially expressed among the three types of organs (28 375). TOPGO graphs (http://www.bioconductor.org/packages/release/bioc/html/topGO.html) of contigs differentially expressed between nodule and shoot tissues (Figure S8) highlighted differential expression patterns of GO families involved in nodule morphogenesis, and to a lesser extent in oxygen transport, ion and oxygen binding, and those whose expression is localized inside membranes. TOPGO graphs of contigs differentially expressed between roots and nodules (Figure S9) indicated differential expression of genes involved in nodule morphogenesis, ion binding, and those whose expression is localized at the periphery of the cell in the apoplast, cell wall and membrane compartments. The differential expression classes were consistent with *K*-means clustering groups: the vast majority of significantly differentially expressed genes were found in different *K*-means groups (Table S6).

By combining information on differential expression and specificity (Table S3), we identified transcripts that were both significantly and highly up-regulated in nodule, root or shoot libraries. Table S7 lists the transcripts that were significantly and very preferentially, specifically or very specifically expressed in nodules. These 842 transcripts displayed a diversity of putative functions that are consistent with the role of this organ in the plant. A number of nodulin genes were found in this list. Other annotated contigs included: contigs related to the response to or biosynthesis of auxin (five contigs), jasmonate (three contigs) or ethylene (three contigs), which may be involved in the hormonal control of nodulation, 16 kinases, 28 transcription factors, six calcium- or calmodulin-binding proteins, 16 transmembrane proteins possibly required for the signaling cascade, 10 ubiquitin-related contigs and three subtilisin proteases with possible regulatory roles through protein modification, 11 contigs encoding caffeic acid *O*-methyltransferase, expansin, glycosyl hydrolase, pectate lyase, pectinacetylesterase and polygalacturonase, possibly involved in comprehensive reorganization of the cell wall during establishment of symbiosis, 20 defense-related contigs, including nine γ-thionin-like proteins, possibly involved in the control of plant immunity, and two contigs homologous to ankyrin repeat proteins that may be required for symbiosome persistence, as reported in *L. japonicus* by Kumagai *et al.* (2007). During symbiosis, the bacteroids fully rely on the plant for energy, metabolites and constituents: 66 contigs encoding transporters from various protein families were up-regulated in nodules, including seven ABC transporter-like transporters, ten major intrinsic protein-like transporters, five MATE efflux transporters, three nodulin MtN21/EamA-like transporters, 11 for amino acid and ureide, four for sulfate, one for sulfate/molybdate, two for iron, two for potassium, one for magnesium, one for malate, six for sugar, and six peptide or peptide nitrate transporters. Six cysteine desulfurylase and rhodanese-related sulfur transferases were also up-regulated. During nitrogen fixation, the transport of oxygen and the maintenance of hypoxia required for nitrogenase activity may be mediated by seven contigs encoding leghemoglobin, and by a carbonic anhydrase (PsCam046442). Interestingly, one contig encoding a 2-on-2 (truncated) hemoglobin was also highly up-regulated in nodules. The energy for bacteroid nitrogen fixation may be provided though PepC (PsCam054230) and sucrose synthase (PsCam032992). The two asparagine synthase and four glutamine synthase contigs may be implicated in metabolism of symbiotically fixed nitrogen before its export to the plant. Several cysteine or serine proteases may be involved in nodule senescence. Fourteen contigs potentially involved in reaction to abiotic stresses, including oxidative stress, were also up-regulated.

DISCUSSION

De novo assembly and redundancy reduction of the pea Unigene set

The bioinformatics procedures for *de novo* short sequence assembly after next-generation sequencing are under continuous and rapid development (Grabherr *et al.*, 2011; Chang *et al.*, 2015). Because no widely used method had been defined for *de novo* assembly at the time of our analysis, we tested various strategies in order to optimize the final assembly. The Velvet tool (Zerbino and Birney, 2008) used for producing the assemblies searches for a given sequence length of *k* nucleotides (*k*-mer), by overlapping all possible $k-1$ sequences and building de Bruijn graphs (Zerbino and Birney, 2008) that connect the perfectly overlapping sequences. This permits a reduction in the complexity of data treatment, and



DE					RPKM max																					
ClassDE	DEsummary	Nod_Shoot	Root_Shoot	Nod_Root	ApicNode_B_LN	Flower_B_LN	Leaf_B_LN	LowerLeaf_C_LN	Nodule_A_LN	Nodule_B_LN	Nodule_G_LN	Peduncle_C_LN	Pods_C_LN	Root_B_LN	Root_F_LN	RootSys_A_HN	RootSys_A_LN	Seed_5dai	Seeds_12dap	Shoot_A_HN	Shoot_A_LN	Stem_BC_LN	Tendril_BC_LN	UpperLeaf_C_LN	Total
1	Nod>root>shoot	up_nodule	up_root	up_nodule	1	1			142	67	93	1						1	10						316
2	Root > nod > shoot			up_root	1	1							1	139	49	45	31	6	2						275
3	Nod=root>shoot			#N/A	9	39			166	139	170	4	6	329	112	68	37	58	64			5	2		1208
4	Nod>shoot>root		up_shoot	up_nodule		2			17	7	7														33
5	Nod>shoot=root		#N/A	up_nodule	5	38			613	244	321	3	3					12	45			2			1286
6	Nod >= root >= shoot			#N/A	8	14			126	70	86	3	2	8	4		1	14	25			1	1		363
7	Root > shoot > nod	up_shoot	up_root	up_root		2								2	1										5
8	Shoot > nod > root		up_shoot	up_nodule	1	1	23	1											1	3	2		1	30	63
9	Shoot > root > nod			up_root		5	65	4				6	1						4	4	4	2	5	43	143
10	Shoot > nod = root			#N/A	14	91	1002	52				36	13					26	32	123	115	14	46	769	2333
11	Root = shoot > nod		#N/A	up_root	1	22	64	3				18	1	6	1	1	1	7	11	6	4	10	15	26	197
12	Shoot >= root >= nod			#N/A	2	12	51	4				9	1	1	1	3		4	10	1	2	4	8	28	141
13	Root > nod = shoot	#N/A	up_root	up_root	38	72						39	20	332	180	59	102	18	50			132	9	2	1053
14	Root >= nod >= shoot			#N/A	149	98	2		35	30	96	46	9	480	128	77	61	69	126			50	12	5	1473
15	Shoot = nod > root		up_shoot	up_nodule	3	7	73	3	28	6	17	10	1					14	5	3	5	2	7	64	248
16	Shoot >= nod >= root			#N/A	33	61	573	21	2	2	19	22	11					32	36	57	54	7	31	438	1399
17	Nod >= shoot >= root		#N/A	up_nodule	12	33	44		113	34	154	6	5					18	28	5	3	4	3	43	505
18	Root >= shoot >= nod			up_root	54	131	61					52	26	47	10	2	10	18	73	29	17	202	22	34	788
19	Nod = root = shoot			#N/A	1716	2663	7140	187	1626	554	2023	1140	267	847	172	196	165	1263	2838	309	170	713	511	3874	28375
Total					2047	3293	9098	275	2868	1153	2986	1395	367	2191	658	451	408	1560	3360	540	376	1148	673	5356	40204

Figure 6. Number of differentially expressed contigs among nodule, root and shoot tissues.
Contig expression levels were compared pairwise using DEseq (false discovery rate < 0.05), and various classes of differential expression were defined. Pink indicates preferential expression in nodules, brown indicates preferential expression in roots, and green indicates preferential expression in shoots. The contigs from these classes were distributed according to the tissue showing maximum RPKMnorm expression.

thus reduces the computational power requirements. The choice of *k*-mer is crucial: the lower the *k*-mer, the larger the de Bruijn graph and the more comprehensive the contig assembly, but more computational power is needed. It has been reported that shorter *k*-mers are needed for genes expressed at a low level, whereas longer *k*-mers are better for highly expressed genes (Gruenheit *et al*., 2012). A commonly used solution is to use a *k*-mer that performs well on a subset of sequences. However, because it is difficult to find an ideal *k*-mer for *de novo* assembly, another solution is to use various *k*-mers to obtain a better coverage of all transcripts (Surget-Groba and Montoya-Burgos, 2010). We tested these two options, and the increasing *k*-mer method proved the most useful in our case.

Redundancy reduction was another issue in production of the pea Unigene set. It is necessary to discard different versions of the same contig obtained using different *k*-mers or from different graphs with the same *k*-mer. The situation is even more complex for transcript isoform classes for which splice variants must be reduced but multi-gene family members must be preserved. Previous RNA-seq experiments in pea mainly targeted development of a reference for SNP discovery (Franssen *et al*., 2011; Kaur *et al*., 2012; Duarte *et al*., 2014). These authors pointed out the difficulty of *de novo* assembly and especially the issue of redundancy and paralogy. They described production of numerous short contigs, indicating that the bioinformatics treatment of the data is a limiting step. The number of contigs in the three preliminary assemblies from this study illustrated this difficulty (Table 2): they were much larger than the number of expected genes in a plant species, and included different splicing forms of the same gene. Among the first three assemblies obtained, the first provided the widest range of transcript forms and is well suited to explore the splicing variations of a gene.

After the various cleaning and redundancy reduction steps described in Figure 1, the pea Unigene set was defined. It includes 46 099 sequences and provides a comprehensive full-length gene catalog of pea. The number of genes is similar to figures obtained from most other legume plant genomes: the *A. thaliana* genome includes 25 498 protein-coding genes (Arabidopsis Genome Initiative 2000) and the *Cicer arietinum* genome includes 28 269 gene models (Varshney *et al*., 2013), while the *M. truncatula* genome includes 62 388 genes (Young *et al*., 2011), the *Glycine max* genome includes 46 430 genes (Schmutz *et al*., 2010), and the *C. cajan* genome includes 48 690 genes (Varshney *et al*., 2011). Furthermore, the N50 lengths of the pea Unigene set are consistent with gene lengths in plants (Yandell and Ence, 2012) and the size of the predicted peptides was very similar to the size of predicted peptides for orthologous *M. truncatula* and *G. max* genes (Figure S10), suggesting that contigs probably correspond to full-length sequences. Interestingly, the longest contigs of the first assembly corresponded to chloroplast gene contigs. Because organelle-encoded genes are contiguous on the chromosome, their assembly as one contig is logical.

We tested the representativeness of the PsCam_Unigene set compared to other available pea transcript sequences (Table 4), and found that the Unigene set was the most complete assembly. The vast majority of previously identified pea sequences are present in the Unigene set, whereas other assemblies lack at least 22% of its sequences. Additionally, most sequences that were found in other studies but not in our Unigene set were not annotated, suggesting that they may be assembly artifacts. Furthermore, the comparison of our Unigene set to sequenced genomes identified between 25 636 and 29 233 best homologs and between 19 697 and 25 299 best reciprocal homologs in *A. thaliana*, *C. arietinum*, *G. max*, *C. cajan*, *L. japonicus* and *M. truncatula* genomes (Table 5). More specifically the number of transcript families detected using the OrthoMCL method (Li *et al.*, 2003) was similar to that for *M. truncatula* (Table S2). Finally, we characterized the pea Unigene set by mapping genomic sequences that are available for cv. 'Caméor' onto it. This enabled us to define a PsCam_LowCopy Unigene set that comprises transcripts corresponding to genes present in low-copy-number in the genome, and a PsCam_HighCopy Unigene set that comprises transcripts whose coding sequences are present in high-copy-number in the genome.

A unique functional tool for analysis of gene expression in pea

RNA-seq has been described as a very robust and sensitive tool for transcriptomics ('t Hoen *et al.*, 2008; Wang *et al.*, 2009; Garg and Jain, 2013). The RNA-seq data obtained for the various tissues sampled in our study yielded a mean of 50.7 million reads per sample (Table 1). The concordance between differential expression and *K*-means clustering analyses (Table S6) suggests a wide dynamic range of RNA-seq data, allowing clear differentiation among organs and stages. Probably due to this very high depth of sequencing, expression of most PsCam_LowCopy contigs was detected in almost all plant tissues (Figure 3), as previously reported in other experiments (Wang *et al.*, 2009), and more than 80% of contigs appeared to be expressed in each library (Figure 3). Surprisingly, the PsCam_HighCopy contig expression was low, but not zero. While most GO classes are represented in the PsCam_LowCopy Unigene set (Figure S3), the PsCam_HighCopy Unigene set mainly included transcripts associated with retrotransposon and transposon metabolism. Our results showed that these transposable elements integrated in the pea genome are expressed at low levels in most plant tissues. The significance of this finding may merit more attention in the future. Even more intriguing was the pattern of expression of PsCam_HighCopy contigs. Low levels of expression were found in embryogenic tissues (pods, flowers, apical nodes and seeds), with higher levels in nodules and peduncles (Figure 3). Furthermore, the profile of the transcriptome in the various tissues given by the PsCam_HighCopy Unigene set clearly differed from that for the PsCam_LowCopy Unigene set (Figure 4).

PsCam_LowCopy transcriptomes provide a significant insight into the functioning of the pea plant. Principal component analysis of the transcriptomes (Figure 4) revealed functional relationships between tissues: the first principal component analysis axis separated nodules from other plant tissues, and the second axis differentiated below- and above-ground tissues. Seeds were intermediate on both axes. The two seed stages diverged on the second principal component analysis plan. Indeed, the 12 days after pollination stage of seed development corresponds to the transition phase between embryogenesis and seed filling, while seeds at 5 days after imbibition represent an early stage of germination. Similarly, using microarray expression data for *M. truncatula*, Benedito *et al.* (2008) differentiated three groups of organs: below-ground (including roots and nodules), above-ground (including vegetative tissues, flowers and pods), and seeds at various developmental stages. Verdier *et al.* (2013), using microarray data for *L. japonicus*, and Severin *et al.* (2010), using RNA-seq expression data for *G. max*, obtained a similar clustering of organs. This is globally consistent with the classifications we found, even though the boundaries may change according to the samples included in the experiment. The distribution of gene ontologies for contigs specifically expressed in a tissue (>100-fold change between the highest and the second highest expression levels) (Figure S5) further highlighted the specific nature of nodules, seeds, roots and flowers. More than half of the specifically expressed APETALA2/Ethylene Responsive Factors were found in nodules, while more than half of the specifically expressed B3-type transcription factors were found in seeds. Specifically expressed basic helix-loop-helix proteins were equally distributed in seeds and roots. The highest numbers of specifically expressed transcription factor genes were found in nodules, seeds and flowers (343, 180 and 160 respectively, Table S4). While nodules, flowers and seeds gave rise to numerous specific gene expression profiles (Figure S5 and Table S3), roots display more subtle differences in gene expression compared to other plant vegetative organs than nodules, flowers and seeds do. Altogether, all results (Figures 4, 5 and S6) indicate the specific molecular apparatus of nodules, a unique organ that develops on roots to host the nitrogen-fixing symbiosis with rhizobacteria.

The pea nodulated root 'way of life'

Root nodules play an essential role in symbiosis by hosting symbiotic bacteria and providing an appropriate cellular environment for nitrogen fixation. This symbiosis has been extensively studied, and following genome sequencing of model legumes, numerous molecular determinants of legume nodulation have been identified, especially in *M. truncatula* and *L. japonicus*. Sets of genes playing important roles in the perception of specific *Rhizobium* Nod factors, the initiation of symbiotic infections and nodule organogenesis, and later stages of nodule development have been identified (Den Herder and Parniske, 2009; Oldroyd *et al.*, 2011; Popp and Ott, 2011; Limpens *et al.*, 2013;

Udvardi and Poole, 2013; Roux *et al.*, 2014). In pea, a number of mutations affecting nodule development have been identified through synteny (*PsSym7*: Kaló *et al.*, 2005; *PsSym33*: Ovchinnikova *et al.*, 2011; *PsSym35*: Borisov *et al.*, 2003; *PsSym8*: Edwards *et al.*, 2008; *PsSym9*: Lévy *et al.*, 2004; *PsSym10*: Madsen *et al.*, 2003; *PsSym19*: Endre *et al.*, 2002; *PsSym29*: Krusell *et al.*, 2002; *PsNod3*: Schnabel *et al.*, 2011; *PsSym28*: Krusell *et al.*, 2011), suggesting largely conserved nodulation determinism amongst legume species. Here, we used reciprocal FASTA searches to identify probable pea orthologs for most significant regulators of nodulation (Figure 7 and Table S8), and most of them were expressed very similarly to the corresponding *M. truncatula* genes in terms of timing and location. These include genes controlling Nod factor perception (*MtLYK3* and *MtNFP*), signal transduction upstream of Ca^{2+} spiking (*MtDMI2*, *NUP85*,*NUP133*, *NENA* and *MtDMI1*), and signal transduction downstream of Ca^{2+} spiking, including transcription factors and genes leading to bacterial infection and nodule organogenesis (*MtDMI3*, *MtIPD3*, *NSP1*, *NSP2*, *ERN1* and *NIN*). Among the 4912 transcripts that were found to be differentially regulated in pea nodules in comparison with roots (adjusted *P*-value < 0.05, fold change >2), 2451 were up-regulated and 2461 were down-regulated. Likely orthologs were found in *M. truncatula* for most of these pea contigs (82.9%), and the majority of orthologous pairs (2375, i.e. 58.3% of the 4075 predicted orthologous genes, Roux *et al.*, 2014) were found to be regulated in a similar way in pea and *M. truncatula*, while only 311 (7.6%) were regulated in opposite ways, and the difference in expression for the rest was below the thresholds chosen for defining differential regulation. Opposite regulation was notably observed for contigs belonging to multi-gene families (e.g. 11 cytochrome P450 genes), which may reflect differential evolution within a family or possibly incorrect prediction of orthologous genes among closely related genes. An intriguing difference between the pea and *M. truncatula* nodule transcriptomes was the poor conservation of the large Nodule-specific Cysteine-Rich (NCR) family, which encodes cysteine-rich peptides that are strongly up-regulated in indeterminate nodules and play a key role in bacteroid differentiation (Fedorova *et al.*, 2002; Mergaert *et al.*, 2003; Kondorosi *et al.*, 2013). One hundred putative NCR-encoding nodule-specific transcripts were found in the pea transcriptome (86 in the PsCam_LowCopy Unigene set, five in the PsCam_HighCopy Unigene set and nine in the PsCam_Discarded set) using HMM profiles CRP1130–CRP1530 as described by Zhou *et al.* (2013). Even if the number of identified NCRs is under-estimated because of the redundancy elimination step, this is much less than the number of NCRs identified in the *M. truncatula* nodule transcriptome (over 600, Roux *et al.*, 2014). In addition, none of them had high sequence homology with *M. truncatula* genes. Consistently, a recent study of NCR expression and sequence diversity within *M. truncatula* suggested rapid and recent evolution of this gene family (Nallu *et al.*, 2014). It will be interesting to investigate whether this difference between pea and *M. truncatula* results from differential amplification and evolution of the NCR gene family in the two legume species, and whether it influences the nitrogen-fixing symbiotic efficiency.

Based on studies in pea, nodules have long been known to express so-called nodulin genes that are poorly expressed in other organs such as roots. Figure 7 shows data obtained for a series of classical nodulin genes, which, as expected, were found to exhibit dramatic levels of up-regulation in nodules. Surprisingly, no orthologous contig was found for the well-known early nodulin gene *MtENOD11* (Journet *et al.*, 2001). By contrast, RNA-seq data revealed two new genes encoding proline-rich putative cell-wall proteins paralogous to *PsENOD12A* and *PsENOD12B* (Govers *et al.*, 1991), respectively, while only one *ENOD12* gene was found in *M. truncatula*. All four *PsENOD12* genes were up-regulated in nodules, but the two *PsENOD12A* genes were also expressed in some aerial organs (Figure 7). The pea Unigene set also included the probable pea ortholog of *VsENBP1* (Table S8), which regulates the expression of *PsENOD12B* (Hansen *et al.*, 1999). Analyzing more deeply the 842 transcripts (Table S7) that were both significantly and highly up-regulated in nodules compared to other organs (adjusted *P*-value < 0.05, fold change > 10) revealed candidates for components of nodule development (including signaling, regulatory and organogenesis components) as well as nodule function (including bacteroid energy source, oxygen and mineral nutrition, stress management, nitrogen compound metabolism and export). Signaling and transport were the most highly represented classes of functions, as reported in other species (González-Guerrero *et al.*, 2013; Udvardi and Poole, 2013; O'Rourke *et al.*, 2014).

Since Mendel's discovery of the rules of genetics, pea has long been a model species for plant geneticists and physiologists, and a wealth of characterized mutants are available (for example at http://data.jic.ac.uk/cgi-bin/pgene). The development of a complete set of expressed genes for pea and the deployment of the first pea gene atlas represents a significant step forward in pea genomics. This is a unique resource to allow searches for candidate gene sequences, or for designing microarrays or undertaking large-scale proteomics studies. It is a gateway towards genomic-enabled improvement of pea, taking advantage of the extensive synteny observed between pea and its closely related model species *M. truncatula*, for which numerous significant traits have been deciphered. Our results will enhance the capacity to transfer knowledge from *M. truncatula* to pea, using induced mutants (Dalmais *et al.*, 2008) or natural variants of candidate genes for important traits.

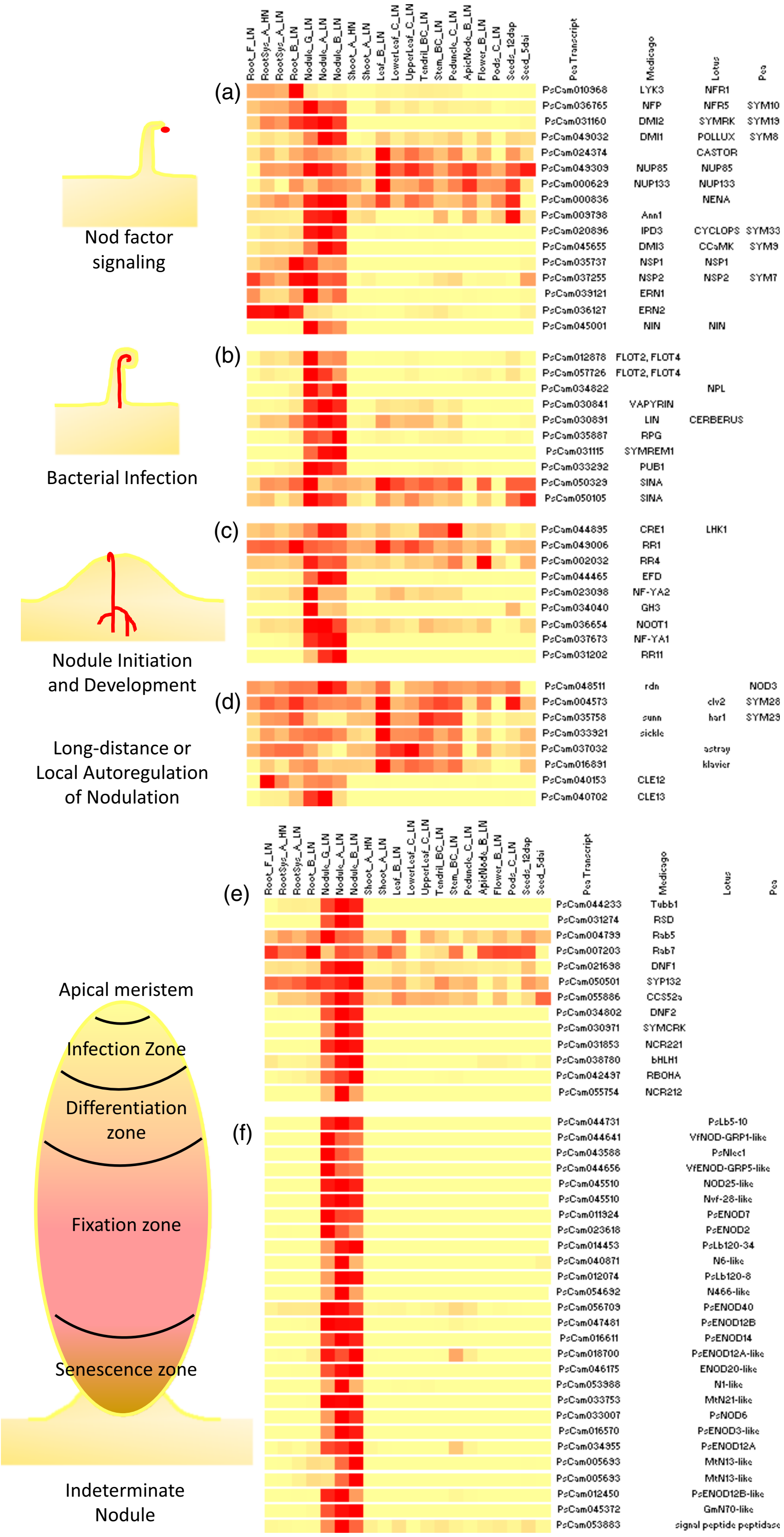


Figure 7. Expression profiles of pea putative orthologs of major determinants of nodule organogenesis and bacterial infection.

Correspondence between PsCam_LowCopy contigs and *M. truncatula* and *L. japonicus* major nodulation genes.

(a) Genes involved in Nod factor signal perception and transduction, such as the receptor-like kinases encoded by *MtLYK3* and *MtNFP*, genes whose products are involved in calcium oscillations (*MtDMI2*, *MtDMI1*, *LjCASTOR*, *NUP85*, *NUP133* and *LjNENA*) and perception (*MtDMI3*, *MtIPD3*), as well as transcription factors initiating the nodulation response (*NIN*, *ERN1* and *ERN2*).

(b) Genes involved in the initiation of bacterial infection at the epidermis, and possibly in the persistence of the nodule.

(c) Genes involved in nodule initiation in the cortex.

(d) Genes involved in long-distance or local regulation of nodulation.

(e) Genes involved in nodule differentiation and persistence.

(f) Genes previously identified as markers of nodulation in several species.

The putative location of expression and biological processes were inferred based on knowledge of orthologous genes in *M. truncatula* and *L. japonicus* (Oka-Kira and Kawaguchi, 2006, Den Herder and Parniske, 2009; Oldroyd *et al.*, 2011; Popp and Ott, 2011; Udvardi and Poole, 2013; Roux *et al.*, 2014). Note that several genes (such as *NF-YA1*, *NIN*, *ERN1* and *CRE1*) actually take part in multiple processes (Nod factor signaling, infection and nodule organogenesis).

EXPERIMENTAL PROCEDURES

Plant material, cDNA library preparation and sequencing

Plant tissues were obtained from three experiments performed in spring 2010 (Table S9). In a first experiment, seeds from the pea cultivar 'Caméor' were sterilized for 10 min in chloride solution (6 g L^{-1}), and germinated in Petri dishes at 25°C for 3 days. Then, seedlings were grown under hydroponic conditions in glasshouses at the Institut National de la Recherche Agronomique, Dijon, France. The nutrient solutions were aerated using a bubbler, inoculated with the P221 *R. leguminosarum* strain at a concentration of 107 colony forming units/ml, and changed regularly. A quarter of the plants were grown at 14 mM nitrogen (high nitrogen) and the rest of plants were grown at 0.625 mM nitrogen (low nitrogen). Plant tissues at three developmental stages were swiftly harvested into refrigerated plates and plunged into liquid nitrogen (Tables 1 and S9). At each stage, measurements were made on four plants to characterize the development stages corresponding to the various libraries: plant height, number of nodes, number of first flowering node (stage B and C), number of flowers, aboveground, root and nodule biomass, length of the tap root, number of nodules, and chlorophyll content as estimated by SPAD chlorophyll meter (Table S10). In a second experiment performed in the glasshouses of the Institut National de la Recherche Agronomique, Dijon, France, seeds were germinated in Petri dishes at 25°C, and 12 germinating seeds were harvested 5 days after imbibition. Twelve other seedlings were transferred to 7 L pots filled with a mix of attapulgite and clay beads, and irrigated with a nutrient solution at 14 mM nitrogen. Seeds were harvested from the second flowering node of these plants 12 days after pollination, i.e. at the transition stage between seed embryogenesis and filling. In a third experiment performed in the growth chamber at the Institut National de la Recherche Agronomique, Toulouse, France, 50 seeds of cv. 'Caméor' were germinated after sterilization in chloride solution for 10 min. Then, plants were grown under aeroponic conditions, as described in the Medicago handbook (http://www.noble.org/MedicagoHandbook/), at 22°C, 75% humidity, 200 μE m^{-2} sec^{-1} light intensity, with a 16 h light/8 h dark photoperiod. The aeroponic nutrient medium was supplemented with 0.5 mM ammonium nitrate. After 8 days, whole root systems were harvested from 18 plants just before inoculation as controls, and the other plants were inoculated with *R. leguminosarum* strain P221. Nodules were harvested from 18 plants at 10 days post-inoculation, and immediately frozen in liquid nitrogen. All tissues were stored at −80°C prior to RNA extraction. Total RNA was extracted using an RNeasy plant mini kit (Qiagen, http:/www.qiagen.com) after grinding, first with a mortar and pestle, and then with a MM301 mill (Retsch, www.retsch.fr) for 2 × 30 sec. RNAs were quantified with using Nanodrop spectrophotometer (http://www.nanodrop.com), and their quality was checked using a bioanalyser (Agilent, http://www.agilent.com). Then, poly(A) mRNA was isolated by two successive purification steps on oligo(dT)$_{25}$ magnetic beads using a Dynabeads® mRNA purification kit (Thermo Fisher Scientific, http://www.thermofisher.com). Synthesis of double-stranded cDNA was performed using a Superscript double-strand cDNA synthesis kit (Thermo Fisher Scientific), and 0.5–1 μg cDNA per sample was used to prepare cDNA libraries. Fragmentation was performed using a E210 Covaris instrument (Covaris, http://covarisinc.com) (duty cycles 10%, intensity 5, 200 cycles per burst, 90 sec). Libraries were prepared using an NEBNext® DNA Library Prep Master Mix Set for Illumina® (New England Biolabs, http://www.neb.com). Libraries were paired-end sequenced as recommended by Illumina (https://www.illumina.com, one lane per library, 14 libraries using the Genome Analyzer II platform and six libraries using the HiSeq2000 platform). Two files containing one-end sequences were generated for each library (Table S11).

De novo assembly methods

After trimming and cleaning reads, the mean insert length and its standard deviation were estimated by mapping paired reads on expressed sequence tags of pea available in the NCBI database (http://www.ncbi.nlm.nih.gov/) (Table S11). Then, we tested three methods for assembly of the RNA-seq datasets obtained from the 18 vegetative tissues. The first method assembled the sequences of each library, with increasing k-mer values, using the Velvet/Oases tools (Zerbino and Birney 2008; Schulz *et al.*, 2012) . At each iteration (n), contigs of the preceding iteration ($n-1$) were used as long reads, and all non-redundant reads of iterations ($n-1$) and (n) were used for iteration ($n+1$). For computing capacity reasons, the range of k-mers was adjusted to the size of the sequence datasets (Table S12). Then, the contigs obtained for the various libraries were merged and subjected to TGICL++ (http://lipm-bioinfo.toulouse.inra.fr/download/compendium/), a program suite based on TGICL (Pertea *et al.*, 2003) that reduces redundancy. The second method of assembly consisted of running Velvet/Oases tools separately on all libraries using a unique *k*-mer. A *k*-mer of 75 was chosen from preliminary tests using *k*-mers of length 27–75 to minimize chimeras of neighboring genes with overlapping 3′ UTRs. Then transcripts from all libraries were merged before running TGICL++. The third method consisted of running Velvet/Oases on all sequences from the various libraries merged together, and then running TGICL++. This third method was the most computationally intensive: it was run on a 1 TB RAM server at GENOTOUL (Toulouse, France). The first method was the longest: it was run in parallel on several 256 Gb RAM servers at GENOUEST (Rennes, France). Finally, the contig assembly obtained after TGICL++ was cleaned using an in-house script removing all sequences of low complexity, lengths below 200 nt and/or containing one or more N. These tests were performed without including the seed libraries.

After analyzing the three assemblies produced by these methods, we set up a fourth *de novo* assembly pipeline that we used to produce the pea Unigene set (Figure 1). The pipeline extracted the assembly of contigs corresponding to the non-redundant set of longest ORFs from the Velvet/Oases transcripts assembly of each library. This assembly was then cleaned using an in-house script removing all sequences below 200 nt and/or whose ORF represented <30% of the contig length. In the case of chimeras, we only kept the longest ORF per contig, because the shorter ORFs were usually incomplete in hetero-chimera and useless in self-chimera. After these cleaning steps, we extracted ORF sequences with 200 nt upstream and downstream sequences in order to include flanking UTRs in subsequent steps. Then, extended ORF sequences from all libraries were merged, and we obtained a dataset of 735 782 contigs. We reduced contig redundancy by several consecutive clustering steps, using CD-HIT-EST (Li and Godzik, 2006; parameters were at least 50% of the smallest sequence aligned, with 100% identity), GLINT (http://lipm-bioinfo.toulouse.inra.fr/download/; parameters were at least 150 nt aligned with 100% identity), and then CD-HIT-EST again with less-stringent parameters (at least 50% of the smallest sequence aligned, with 97% identity). Then, we removed non-plant sequences. Using the Uniprot TREMBL virus sequence database at ftp://ftp.uniprot.org/pub/databases, we discarded sequences showing homologies to virus sequences. Using a set of Illumina

HiSeq2000 reads obtained from 'Caméor' genomic DNA sequencing sevenfold (C. Aluome, M.C. Le Paslier and D. Brunel, INRA UR1279 personal communication), we further discarded all contigs that did not match any genomic read. Matching parameters were 80% identity on 100% read length. Finally, we labeled all contigs whose genomic coverage was above 50 as 'high copy genes'. On the other hand, we integrated sequences homologous to organelle sequences, some of which were smaller than 200 nt.

We evaluated the Unigene contig set by various means. We compared simple statistics of contig length and distribution with previously obtained pea datasets. We searched all published pea sequences (Künne *et al.*, 2005; Franssen *et al.*, 2011; Kaur *et al.*, 2012; Duarte *et al.*, 2014) and expressed sequence tags at http://www.ncbi.nlm.nih.gov/nucest (TaxID 3888, *Pisum sativum*) for the best homologs of our pea Unigene set using BLASTN (http://blast.ncbi.nlm.nih.gov/Blast.cgi). We also searched all predicted peptide sequences available in November 2014 for *A. thaliana* version 10 (Arath, http://www.arabidopsis.org/), *C. cajan* (Cajca, http://www.icrisat.org/gt-bt/iipg/Home.html and http://gigadb.org/dataset/100028), *G. max* (Glyma, http://soybase.org/dlpages/index.php), *M. truncatula* version 4 (Medtr, http://www.jcvi.org/medicago/), *C. arietinum* (http://gigadb.org/dataset/100076) and *L. japonicus* (Lotja, ftp://ftp.kazusa.or.jp/pub/lotus/lotus_r2.5/) for the best homologs of our pea Unigene set using BLASTP and for the best reciprocal homolog using an in-house script (Bordat *et al.*, 2011).

We also assessed the *de novo* sequencing package TRINITY (Grabherr *et al.*, 2011). Two assemblies were obtained by testing different methods: the first using the default parameters of the package (called Trinity 1) and the second using a minimal k-mer value of 5 (called Trinity 2). The two assemblies were then cleaned as described above for the Unigene set, by removing all sequences with low complexity, lengths below 200 nt, and by reducing redundancy using CD-HIT-EST. ORFs were then predicted for each assembly using Getorf (EMBOSS, http://emboss.sourceforge.net/) (called Trinity1E or Trinity2E) before merging. The merged files were subjected to (i) the four consecutive cleaning steps using CD-HIT-EST described above, and (ii) to the mapping of 'Cameor' genomic DNA reads onto the Unigene contigs to differentiate low copy number and high copy number genes. We finally obtained two sets of sequences (a low-copy-number set and a high-copy-number set) for each method of assembly with TRINITY (Trinity 1E and Trinity 2E); these were compared with the Unigene set.

Annotation of the Unigene sequences and pea gene atlas portal deployment

An analysis of Unigene predicted peptides using INTERPROSCAN (Quevillon *et al.*, 2005) was performed on the GENOTOUL platform. An HMMsearch using the HMMER3 package (Eddy, 2011) was performed using default values to detect NCR and complete the annotation. This analysis of peptide domains provided a functional annotation. Transcript family clusters from *Pisum sativum* PsCam_LowCopy contigs and *Medicago truncatula* genes (version 4.0) were determined using ORTHOMCL 2.0.9 (Li *et al.*, 2003). Similarities between sequences were calculated by all-by-all BLASTP (http://blast.ncbi.nlm.nih.gov/Blast.cgi) with an *e*-value of 1e-05. Default parameters were used. Alternative transcripts in the *Medicago truncatula* set were not removed.

A quantification of the level of expression of all contigs was performed using various procedures: (i) COUNT, the number of paired-end reads mapped per contig in each library relative to the total number of mapped paired-end reads in each library, (ii) RPKM, the number of reads per kilobase per million reads (Mortazavi *et al.*, 2008), and (iii) RPKMnorm, the RPKM divided by the geometric mean of the RPKM for three control genes: histone H1 (PsCam009820), actin (PsCam042218) and EF1α (PsCam042119). All this information (contig nucleic and peptidic sequences, INTERPROSCAN annotation, and expression level metrics) for all Unigene contigs was introduced into the BIOS database (http://bios.toulouse.inra.fr/). Then, the Functional Analyses porTAL (http://lipm-bioinfo.toulouse.inra.fr/) was deployed: the structure of the portal was created and the NCBI-blast (http://blast.ncbi.nlm.nih.gov/Blast.cgi) and WU-blast (http://blast.wustl.edu) annotation was launched. All the results are stored in the BIOS database, and are available online at http://bios.dijon.inra.fr/FATAL/cgi/pscam.cgi, which allows searching of the database for sequences and visualization of expression data. The reliability of the expression data visualized in the gene atlas was checked by comparing it to the RPKM and RPKMnorm expression profiles obtained by quantitative RT-PCR expression for the ten genes listed in Table S13. Experiments were performed using a Lightcycler 480 thermocycler (Roche, http://www.roche.com) as described by Gallardo *et al.* (2007) with three technical replicates and three biological replicates. Relative expression of the genes was calculated using the $2^{-\Delta\Delta C_T}$ method (Livak and Schmittgen, 2001), and using the three control genes histone H1 (PsCam009820), actin (PsCam042218) and EF1α (PsCam042119) for normalization (as used for RPKMnorm calculation). The primers used are listed in Table S13.

A MAPMAN bin file (http://mapman.gabipd.org/web/guest/mapman; Thimm *et al.*, 2004) was prepared using MERCATOR (http://mapman.gabipd.org/web/guest/app/mercator; Lohse *et al.*, 2014) by comparing contigs against already classified proteins. BLASTP was used to search all predicted peptides against the Arabidopsis Information Resource (TAIR10) proteins, SwissProt & Uniprot plant proteins, and the Institute for Genomic Research (TIGR5) rice proteins. RPSBLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) was used for comparison with the conserved domain database (CDD, http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml), clusters of orthologous groups (COG, http://www.ncbi.nlm.nih.gov/COG/) and INTERPROSCAN. Database hits with bit scores <50 were ignored as not significantly similar. The results of all searches were compiled into one table with preliminary MAPMAN bin codes assigned to each of the Unigene predicted peptides. Some minor manual corrections were added to the table by comparing these bin codes with those assigned to pea homologs in existing MAPMAN mapping for *A. thaliana* (TAIR9_Jan2010), *M. truncatula* (Mt3.5_v3_0411) and *G. max* (Gmax_109_peptides).

Differential gene expression analyses

The transcriptomes of the 20 libraries as shown by the PsCam_LowCopy and PsCam_HighCopy transcriptomes were compared using principal component analysis of the RPKMnorm expression values (proc princomp, SAS Institute, http://www.sas.com). Then, a more specific analysis of the PsCam_LowCopy gene set was performed: a heatmap was generated after mean linkage hierarchical clustering of pairwise Pearson correlation coefficients of RPKMnorm in the 20 libraries, using MULTIEXPERIMENT VIEWER software (http://www.tm4.org/mev.html). Differential gene expression among libraries was analyzed by *K*-means hierarchical clustering using Genesis ($K = 20$, Sturn *et al.*, 2002, http://genome.tugraz.at/) and DEseq (DESeq Bioconductor package in R, Anders and Huber, 2010). Count data normalization and identification of differentially expressed contigs were performed by pairwise comparisons using a negative binomial distribution. Pairwise comparisons of root libraries (RootSys_A_HN, RootSys_A_LN, Root_B_LN and Root_F_LN), nodule libraries (Nodule_G_LN, Nodule_B_LN and Nodule_A_LN) and shoot libraries (Shoot_A_HN, Shoot_A_LN, Leaf_B_LN, LowerLeaf_C_LN and UpperLeaf_C_LN) were performed with replications in the DeSeq analyses. A false discovery rate threshold of 0.05 (Benjamini and Hochberg, 1995) was applied to identify significantly differentially expressed contigs between pea tissues. Differentially expressed sequences were visualized using TOPGO (http://topgo.bioinf.mpi-inf.mpg.de/). Differential expression between root and nodule tissues was compared with published *M. truncatula* data (Roux *et al*., 2014).

Data deposition

Raw next-generation sequencing read sequences have been deposited at the National Center for Biotechnology Information under Bioproject ID PRJNA267198 (http://www.ncbi.nlm.nih.gov/bioproject/267198).

ACKNOWLEDGEMENTS

We are grateful to the GENOUEST platform (http://www.genouest.org) for allowing access to their computing capacity, to Christelle Aluome, Marie-Christine Le Paslier and Dominique Brunel (Etude du Polymorphisme des Génomes Végétaux, Institut National de la Recherche Agronomique, UR1279, Evry, France) for pea genomic reads, to Céline Rond and Hervé Houtin for skilful plant follow-up and harvesting, to Vincent Durey and Anne-Sophie Voisin for hydroponic systems, to the INRA Dijon glasshouse facility team for running the facility, to Carlos Molina for seed tissue production, and to R. Thompson, N. Tayeh, K. Gallardo and G. Duc for proof-reading of the manuscript. This research was funded by the Agence Nationale de la Recherche (project ANR-09-GENM-026 'GENOPEA').

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Analysis of the six reading frames of sequences of the multiple k-mer contig assembly using RESEARCH software.

Figure S2. Distribution of contig lengths in three assemblies of pea RNA-seq reads.

Figure S3. Distribution of gene ontologies of the pea Unigene contigs.

Figure S4. Gene expression specificity across pea tissues.

Figure S5. Distribution of the putative molecular functions of specifically expressed pea contigs.

Figure S6. Distribution of pea transcription factors expressed at least specifically in one tissue, according to family membership.

Figure S7. K-means clustering of RPKMnorm expression profiles of PsCam_LowCopy Unigene contigs.

Figure S8. TOPGO analysis of genes differentially expressed between nodules and shoots.

Figure S9. TOPGO analysis of genes differentially expressed between roots and nodules.

Figure S10. Distribution of differences in predicted peptide lengths between pea and *M. truncatula* or *G. max* orthologous genes.

Table S1. Analysis of splicing variants retrieved from the various RNA-seq assemblies for eight known pea genes.

Table S2. Number of transcript clusters generated for the PsCam_-LowCopy and PsCam_HighCopy Unigene sets using OrthoMCL (Li *et al*., 2003).

Table S3. Distribution of contigs showing low, not preferential, preferential, very preferential, specific and very specific tissue expression according to differential expression and tissue localization of maximum RPKMnorm expression.

Table S4. List of contigs putatively encoding transcription factors, and showing very preferential or specific expression.

Table S5. Contig distribution among *K*-means clustering groups according to tissue localization of maximum RPKMnorm expression.

Table S6. Contig distribution among *K*-means clustering groups according to differential expression classes as defined in Figure 6.

Table S7. List of contigs significantly and very preferentially, specifically or very specifically expressed in nodules.

Table S8. List of best homologs of major determinants of nodulation in *M. truncatula* and *L. japonicus*.

Table S9. List of tissues sampled for RNA-seq, according to the experiment, stage of harvest, and growing conditions.

Table S10. Development and growth characteristics of four plants harvested at stages A (6 April 2010), B (19 April 2010) and C (7 May 2010).

Table S11. RNA-seq read specifications.

Table S12. Contig assembly characteristics after Velvet/Oases steps, using the multiple k-mer strategy.

Table S13. Primer sequences used for validation of *in silico* expression levels by quantitative PCR.

REFERENCES

Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106.

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

Benedito, V.A., Torres-Jerez, I., Murray, J.D. *et al.* (2008) A gene expression atlas of the model legume *Medicago truncatula*. *Plant J.* **55**, 504–513.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B Methodol.* **57**, 289–300.

Bordat, A., Savois, V., Nicolas, M. *et al.* (2011) Translational genomics in legumes allowed placing *in silico* 5460 unigenes on the pea functional map and identified candidate genes in *Pisum sativum* L. *G3*, **1**, 93–103.

Borisov, A.Y., Madsen, L.H., Tsyganov, V.E. *et al.* (2003) The Sym35 gene required for root nodule development in pea is an ortholog of Nin from *Lotus japonicus*. *Plant Physiol.* **131**, 1009–1017.

Chang, Z., Li, G., Liu, J., Zhang, Y., Ashby, C., Liu, D. and Huang, X. (2015) Bridger: a new framework for *de novo* transcriptome assembly using RNA-seq data. *Genome Biol.* **16**, 29.

Dalmais, M., Schmidt, J., Le Signor, C. *et al.* (2008) UTILLdb, a *Pisum sativum in silico* forward and reverse genetics tool. *Genome Biol.* **9**, R43.

Den Herder, G. and Parniske, M. (2009) The unbearable naivety of legumes in symbiosis. *Curr. Opin. Plant Biol.* **12**, 491–499.

Duarte, J., Rivière, N., Baranger, A. *et al.* (2014) Transcriptome sequencing for high throughput SNP development and genetic mapping in Pea. *BMC Genomics*, **15**, 126.

Eddy, S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195.

Edwards, A., Heckmann, A.B., Yousafzai, F., Duc, G. and Downie, J.A. (2008) Structural implications of mutations in the pea SYM8 symbiosis gene, the DMI1 ortholog, encoding a predicted ion channel. *Mol. Plant Microbe Interact.* **20**, 1183–1191.

Endre, G., Kereszt, A., Kevei, Z., Mihacea, S., Kaló, P. and Kiss, G.B. (2002) A receptor kinase gene regulating symbiotic nodule development. *Nature*, **417**, 962–966.

Fedorova, M., van de Mortel, J., Matsumoto, P.A., Cho, J., Town, C.D., Van den Bosch, K.A., Gantt, J.S. and Vance, C.P. (2002) Genome-wide identification of nodule-specific transcripts in the model legume *Medicago truncatula*. *Plant Physiol.* **130**, 519–537.

Franssen, S.U., Shresta, R.P., Bräutigam, A., Bornberg-Bauer, E. and Weber, A.P.M. (2011) Comprehensive transcriptome analysis of the highly complex *Pisum sativum* genome using next generation sequencing. *BMC Genomics*, **12**, 227.

Gallardo, K., Firnhaber, C., Zuber, H., Héricher, D., Belghazi, M., Henry, C., Küster, H. and Thompson, R. (2007) A combined proteome and transcriptome analysis of developing *Medicago truncatula* – seeds evidence for metabolic specialization of maternal and filial tissues. *Mol. Cell Proteomics*, **6**, 2165–2179.

Garg, R. and Jain, M. (2013) RNA-Seq for transcriptome analysis in non-model plants. *Methods Mol. Biol.* **1069**, 43–58.

Garg, R., Patel, R.K., Jhanwar, S., Priya, P., Bhattacharjee, A., Yadav, G., Bhatia, S., Chattopadhyay, D., Tyagi, A. and Jain, M. (2011) Gene discovery and tissue-specific transcriptome analysis in chickpea with massively parallel pyrosequencing and web resource development. *Plant Physiol.* **156**, 1661–1678.

González-Guerrero, M., Rubio-Sanz, L., Rodríguez-Haas, B., Albareda, M., Menéndez-Cerón, M., Brito, B. and Palacios, J.M. (2013) Metal transport in the rhizobium–legume symbiosis. In *Beneficial Plant–Microbial Interactions, Ecology and Applications* (González-López, J., ed.). Boca Raton, FL: CRC Press, pp. 141–163.

Govers, F., Harmsen, H., Heidstra, R., Michielsen, P., Prins, M., van Kammen, A. and Bisseling, T. (1991) Characterization of the pea *ENOD12B* gene and expression analyses of the two *ENOD12* genes in nodule, stem and flower tissue. *Mol. Gen. Genet.* **228**, 160–166.

Grabherr, M.G., Haas, B.J., Yassour, M. *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652.

Gruenheit, N., Deusch, O., Esser, C., Becker, M., Voelckel, C. and Lockhart, P. (2012) Cutoffs and k-mers: implications from a transcriptome study in allopolyploid plants. *BMC Genomics*, **13**, 92.

Hansen, A.C., Busk, H., Marcker, A., Marcker, K.A. and Jensen, E.Ø. (1999) VsENBP1 regulates the expression of the early nodulin PsENOD12B. *Plant Mol. Biol.* **40**, 495–506.

't Hoen, P.A., Ariyurek, Y., Thygesen, H.H., Vreugdenhil, E., Vossen, R.H., de Menezes, R.X., Boer, J.M., van Ommen, G.J. and den Dunnen, J.T. (2008) Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Res.* **36**, e141.

Journet, E.P., El-Gachtouli, N., Vernoud, V., de Billy, F., Pichon, M., Dedieu, A., Arnould, C., Morandi, D., Barker, D.G. and Gianinazzi-Pearson, V. (2001) *Medicago truncatula* ENOD11: a novel RPRP-encoding early nodulin gene expressed during mycorrhization in arbuscule-containing cells. *Mol. Plant Microbe Interact.* **14**, 737–748.

Kaló, P., Gleason, C., Edwards, A. *et al.* (2005) Nodulation signaling in legumes requires NSP2, a member of the GRAS family of transcriptional regulators. *Science*, **308**, 1786–1789.

Kaur, S., Pembleton, L.W., Cogan, N.O., Savin, K.W., Leonforte, T., Paull, J., Materne, M. and Forster, J.W. (2012) Transcriptome sequencing of field pea and faba bean for discovery and validation of SSR genetic markers. *BMC Genomics*, **13**, 104.

Kondorosi, E., Mergaert, P. and Kereszt, A. (2013) A paradigm for endosymbiotic life: cell differentiation of *Rhizobium* bacteria provoked by host plant factors. *Annu. Rev. Microbiol.* **67**, 611–628.

Krusell, L., Madsen, L.H., Sato, S. *et al.* (2002) Shoot control of root development and nodulation is mediated by a receptor-like kinase. *Nature*, **420**, 422–426.

Krusell, L., Sato, N., Fukuhara, I. *et al.* (2011) The *Clavata2* genes of pea and *Lotus japonicus* affect autoregulation of nodulation. *Plant J.* **65**, 861–871.

Kumagai, H., Hakoyama, T., Umehara, Y., Sato, S., Kaneko, T., Tabata, S. and Kouchi, H. (2007) A novel ankyrin-repeat membrane protein, IGN1, is required for persistence of nitrogen-fixing symbiosis in root nodules of *Lotus japonicus*. *Plant Physiol.* **143**, 1293–1305.

Künne, C., Lange, M., Funke, T., Miehe, H., Thiel, T., Grosse, I. and Scholz, U. (2005) CR-EST: a resource for crop ESTs. *Nucleic Acids Res.* **33**, D619–D621.

Lan, P., Li, W., Lin, W.D., Santi, S. and Schmidt, W. (2013) Mapping gene activity of Arabidopsis root hairs. *Genome Biol.* **14**, R67.

Lévy, J., Bres, C., Geurts, R. *et al.* (2004) A putative Ca^{2+} and calmodulin-dependent protein kinase required for bacterial and fungal symbioses. *Science*, **303**, 1361–1364.

Li, W. and Godzik, A. (2006) cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Li, L., Stoeckert, C.J. and Roos, D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189.

Li, P., Ponnala, L., Gandotra, N. *et al.* (2010) The developmental dynamics of the maize leaf transcriptome. *Nat. Genet.* **42**, 1060–1067.

Limpens, E., Moling, S., Hooiveld, G., Pereira, P.A., Bisseling, T., Becker, J.D. and Küster, H. (2013) Cell- and tissue-specific transcriptome analyses of *Medicago truncatula* root nodules. *PLoS ONE*, **8**, e64377.

Livak, K.J. and Schmittgen, T.D. (2001) Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. *Methods*, **25**, 402–408.

Lohse, M., Nagel, A., Herter, T., May, P., Schroda, M., Zrenner, R., Takayuki, T., Fernie, A.R., Stitt, M. and Usadel, B. (2014) Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. *Plant, Cell Environ.* **37**, 1250–1258.

Madsen, E.B., Madsen, L.H., Radutoiu, S. *et al.* (2003) A receptor kinase gene of the LysM type is involved in legume perception of rhizobial signals. *Nature*, **425**, 637–640.

Madsen, L.H., Tirichine, L., Jurkiewicz, A., Sullivan, J.T., Heckmann, A.B., Bek, A.S., Ronson, C.W., James, E.K. and Stougaard, J. (2010) The molecular network governing nodule organogenesis and infection in the model legume *Lotus japonicus*. *Nat. Commun.* **1**, 10.

McLaughlin, N.B., Hiba, A., Wall, G.J. and King, D.J. (2000) Comparison of energy inputs for inorganic fertilizer and manure based corn production. *Can. Agric. Eng.* **42**, 9–17.

Mergaert, P., Nikovics, K., Kelemen, Z., Maunoury, N., Vaubert, D., Kondorosi, A. and Kondorosi, E. (2003) A novel family in *Medicago truncatula* consisting of more than 300 nodule-specific genes coding for small, secreted polypeptides with conserved cysteine motifs. *Plant Physiol.* **132**, 161–173.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.

Nallu, S., Silverstein, K.A., Zhou, P., Young, N.D. and Vandenbosch, K.A. (2014) Patterns of divergence of a large family of nodule cysteine-rich peptides in accessions of *Medicago truncatula*. *Plant J.* **78**, 697–705.

Oka-Kira, E. and Kawaguchi, M. (2006) Long-distance signaling to control root nodule number. *Curr. Opin. Plant Biol.* **9**, 496–502.

Oldroyd, G.E.D. (2013) Speak, friend, and enter: signalling systems that promote beneficial symbiotic associations in plants. *Nat. Rev. Microbiol.* **11**, 252–263.

Oldroyd, G.E.D., Murray, J.D., Poole, P.S. and Downie, J.A. (2011) The rules of engagement in the legume–rhizobial symbiosis. *Annu. Rev. Genet.* **45**, 119–144.

O'Rourke, J.A., Iniguez, L.P., Fu, F. *et al.* (2014) An RNA-Seq based gene expression atlas of the common bean. *BMC Genomics*, **15**, 866.

Ovchinnikova, E., Journet, E.P., Chabaud, M. *et al.* (2011) IPD3 controls the formation of nitrogen-fixing symbiosomes in pea and *Medicago* spp. *Mol. Plant Microbe Interact.* **24**, 1333–1344.

Pertea, G., Huang, X., Liang, F. *et al.* (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, **19**, 651–652.

Popp, C. and Ott, T. (2011) Regulation of signal transduction and bacterial infection during root nodule symbiosis. *Curr. Opin. Plant Biol.* **14**, 458–467.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R. and Lopez, R. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res.* **33**, W116–W120.

Roux, B., Rodde, N., Jardinaud, M.F. *et al.* (2014) An integrated analysis of plant and bacterial gene expression in symbiotic root nodules using laser-capture microdissection coupled to RNA sequencing. *Plant J.* **77**, 817–837.

Schmutz, J., Cannon, S.B., Schlueter, J. *et al.* (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.

Schnabel, E.L., Kassaw, T.K., Smith, L.S., Marsh, J.F., Oldroyd, G.E., Long, S.R. and Frugoli, J.A. (2011) The *ROOT DETERMINED NODULATION1* gene regulates nodule number in roots of *Medicago truncatula* and defines a highly conserved, uncharacterized plant gene family. *Plant Physiol.* **157**, 328–340.

Schultze, M. and Kondorosi, A. (1998) Regulation of symbiotic root nodule development. *Annu. Rev. Genet.* **32**, 33–57.

Schulz, M.H., Zerbino, D.R., Vingron, M. and Birney, E. (2012) Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, **28**, 1086–1092.

Severin, A.J., Woody, J.L., Bolon, Y.T. *et al.* (2010) RNA-Seq atlas of *Glycine max*: a guide to the soybean transcriptome. *BMC Plant Biol.* **10**, 160.

Sierro, N., Battey, J.N., Ouadi, S., Bovet, L., Goepfert, S., Bakaher, N., Peitsch, M.C. and Ivanov, N.V. (2013) Reference genomes and transcriptomes of *Nicotiana sylvestris* and *Nicotiana tomentosiformis*. *Genome Biol.* **14**, R60.

Sturn, A., Quackenbush, J. and Trajanoski, Z. (2002) Genesis: cluster analysis of microarray data. *Bioninformatics*, **18**, 207–208.

Surget-Groba, Y. and Montoya-Burgos, J.I. (2010) Optimization of *de novo* transcriptome assembly from next-generation sequencing data. *Genome Res.* **20**, 1432–1440.

Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., Selbig, J., Müller, L.A., Rhee, S.Y. and Stitt, M. (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.* **37**, 914–939.

Udvardi, M. and Poole, P.S. (2013) Transport and metabolism in legume–rhizobia symbioses. *Annu. Rev. Plant Biol.* **64**, 781–805.

Varshney, R.K., Chen, W., Li, Y. *et al.* (2011) Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nat. Biotechnol.* **30**, 83–89.

Varshney, R.K., Song, C., Saxena, R.K. *et al.* (2013) Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nat. Biotechnol.* **31**, 240–246.

Verdier, J., Torres-Jerez, I., Wang, M., Andriankaja, A., Allen, S.N., He, J., Tang, Y., Murray, J.D. and Udvardi, M.K. (2013) Establishment of the *Lotus japonicus* gene expression atlas (LjGEA) and its use to explore legume seed maturation. *Plant J.* **74**, 351–362.

Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63.

Yandell, M. and Ence, D. (2012) A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* **13**, 329–342.

Yang, Y. and Smith, S.A. (2013) Optimizing *de novo* assembly of short-read RNA-seq data for phylogenomics. *BMC Genomics*, **14**, 328.

Young, N.D., Debellé, F., Oldroyd, G.E.D. *et al.* (2011) The *Medicago* genome provides insight into the evolution of rhizobial symbioses. *Nature*, **480**, 520–524.

Zenoni, S., Ferrarini, A., Giacomelli, E., Xumerle, L., Fasoli, M., Malerba, G., Bellin, D., Pezzotti, M. and Delledonne, M. (2010) Characterization of transcriptional complexity during berry development in *Vitis vinifera* using RNA-Seq. *Plant Physiol.* **152**, 1787–1795.

Zerbino, D.R. and Birney, E. (2008) Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829.

Zhang, J., Lee, Y., Torres-Jerez, I. *et al.* (2013) Development of an integrated transcript sequence database and a gene expression atlas for gene discovery and analysis in switchgrass (*Panicum virgatum* L.). *Plant J.* **74**, 160–173.

Zhou, P., Silverstein, K.A.T., Gao, L., Walton, J.D., Nallu, S., Guhlin, J. and Young, N.D. (2013) Detecting small plant peptides using SPADA (small peptide alignment discovery application). *BMC Bioinformatics*, **14**, 335.