Plant Molecular Biology (2005) 59:515–532
DOI 10.1007/s11103-005-0536-8

A 6374 unigene set corresponding to low abundance transcripts expressed following fertilization in *Solanum chacoense* Bitt, and characterization of 30 receptor-like kinases★

Hugo Germain[1], Stephen Rudd[2], Corine Zotti[1,3], Sébastien Caron[1], Martin O'Brien[1], Sier-Ching Chantha[1], Marie Lagacé[1], François Major[4] and Daniel P. Matton[1,*]
[1]*Institut de Recherche en Biologie Végétale (IRBV). Département de sciences biologiques, Université de Montréal, 4101 Sherbrooke Est, H1X 2B2, Montréal, QC, Canada (*author for correspondence; e-mail dp.matton@umontreal.ca);* [2]*Munich Information Center for Protein Sequences, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany and Centre for Biotechnology, FIN-20521, Tykistökatu 6, Turku, Finland;* [3]*Centre d'innovation Génome Québec et Université McGill, 740 avenue du Docteur-Penfield, H3A 1A4, Montréal, QC, Canada;* [4]*Département d'informatique et de recherche opérationelle, Université de Montréal, C. P. 6128, H3C 3J7, Montréal, QC, Canada*

Received 16 May 2005; accepted in revised form 4 July 2005

Key words: embryo development, expressed sequence tag (EST), fertilization, negative selection, real-time PCR, receptor-like kinase (RLK), *Solanum chacoense*, virtual subtraction

Abstract

In order to characterize regulatory genes that are expressed in ovule tissues after fertilization we have undertaken an EST sequencing project in *Solanum chacoense*, a self-incompatible wild potato species. Two cDNA libraries made from ovule tissues covering embryo development from zygote to late torpedo-stage were constructed and plated at high density on nylon membranes. To decrease EST redundancy and enrich for transcripts corresponding to weakly expressed genes a self-probe subtraction method was used to select the colonies harboring the genes to be sequenced. 7741 good sequences were obtained and, from these, 6374 unigenes were isolated. Thus, the self-probe subtraction resulted in a strong enrichment in singletons, a decrease in the number of clones per contigs, and concomitantly, an enrichment in the total number of unigenes obtained (82%). To gain insights into signal transduction events occurring during embryo development all the receptor-like kinases (or protein receptor kinases) were analyzed by quantitative real-time RT-PCR. Interestingly, 28 out of the 30 RLK isolated were predominantly expressed in ovary tissues or young developing fruits, and 23 were transcriptionaly induced following fertilization. Thus, the self-probe subtraction did not select for genes weakly expressed in the target tissue while being highly expressed elsewhere in the plant. Of the receptor-like kinases (RLK) genes isolated, the leucine-rich repeat (LRR) family of RLK was by far the most represented with 25 members covering 11 LRR classes.

Abbreviations: DAP, days after pollination; EST, expressed sequence tag; ORK, ovule receptor kinase; RLK, receptor-like kinase

Introduction

In order to gain information about the signal transduction elements and other genes that are

★ Genbank Accession numbers: DN976421 to DN984171 and DR398320.

weakly expressed during plant embryogenesis, we have undertaken a medium scale EST sequencing project using ovule tissues enriched (virtually subtracted) for rare mRNAs (Li and Thomas, 1998; Nelson *et al*., 1999). In classical EST sequencing projects ESTs are sequenced randomly (Van der Hoeven *et al*., 2002; Shimada *et al*., 2003; Fei *et al*., 2004), and the occurrence of a single EST species in the dataset is based on its expression level in the cell or tissue analyzed and, to a lesser extent, on its mRNA stability within the cell. While this approach is rapid and provides valuable information on differential gene expression levels between tissues, it requires the sequencing of large numbers of clones in order to obtain information on the weakly expressed genes. Therefore, using this method, weakly expressed genes can only be obtained at a great cost. Furthermore, genes that are only expressed in a narrow temporal window and genes expressed in only a few cells would be largely underrepresented in non-normalized libraries. Several methods can be used in an attempt to normalize the relative abundance of cDNAs within the library. For example, a simple method is to hybridize the target library with known commonly occurring abundant cDNAs and to select for only the clones that do not hybridize (Adams, 1995; Gong *et al*., 1997). This approach has been used with success, but requires some prior knowledge regarding cDNA frequencies. Other methods, such as oligonucleotide fingerprinting (OFP) (Radelof *et al*., 1998), hybridization kinetics approach (Soares *et al*., 1994), and methods involving the selection of mRNA based on a cap-trapping technique in order to increase the number of cDNAs containing the 5′ end of the mRNA (Carninci *et al*., 2000) have also been used. Although they have proven to be useful and efficient, these techniques all have important limitations that have stood in the way of their widespread use. OFP is a very expensive method, due to the requirement of a large number of synthetically synthesized oligonucleotides, while hybridization kinetics can easily lead to artifacts in the subtracted library composition if all parameters are not perfectly controlled (Bonaldo, 1996). The high number of contigs that are found when one uses the method being developed by Carninci *et al*., suggest that this process is still in need of some improvements (Clark *et al*., 2001).

We used *Solanum chacoense*, a self-incompatible close relative of the potato and tomato which produces a large number of easily isolated ovules inside a fleshy tomato-like fruit, to dissect the molecular events occurring at precise time points following pollination and fertilization as well as during embryogenesis and seed development. The genomic data resulting from the EST project (obtained from two cDNA libraries covering the period from fertilization to late torpedo staged embryos), combined with the data generated from the completion of the sequencing of the *Arabidopsis* genome (Initiative, 2000), and the availability of more than 120 000 EST sequences analyzed in tomato (Van der Hoeven *et al*., 2002), were used to identify receptor-like kinase (RLK) involved in seed and embryo development. The presumed function of receptor-like kinases is to perceive extracellular stimuli and to communicate this information to an intracellular signaling pathway in order to ensure appropriate cellular responses. Information such as positional cues needed by the cell to ensure proper cell fate, or the signal given by the incoming pollen tube could be mediated by RLKs. RLKs are the most abundant kinase family in plants and, in *Arabidopsis*, they comprise more than 2% of the proteins characterized in its genome (Shiu and Bleecker, 2001). Despite the extensive array of developmental processes affected or regulated by RLKs, such as plant defense, meristem development, pollen/pistil interactions, hormone signaling, pollen development, cell morphogenesis and differentiation, organ shape and abscission, and somatic embryogenesis, only a limited number of RLKs have been characterized as being involved in reproductive development and none have been identified as having a specific function in zygotic embryo development (reviewed in Becraft, 2002).

In this paper we have used a modified subtraction screen based on the virtual subtraction procedure of Li and Thomas (1998), and the negative screen of Nelson *et al*. (1999) as a method to enrich EST pools for rare messenger RNAs expressed during embryo development with a focus on receptor kinase signaling. Using a target tissue self-probe hybridization approach, only the clones that yielded a weak radioactive signal following hybridization were selected for sequencing. This method significantly increased the frequency of unigene sequences. Furthermore, no bias towards non-target tissue expression was observed in the study of all the receptor-like kinases isolated as

determined by an extensive quantitative real-time PCR analysis.

Material and methods

cDNA library construction and mass excision of the cDNA clones

All plant material was collected from *S. chacoense* genotype G4 (self-incompatibility alleles $S_{12}S_{14}$). For fertilization-related events, *S. chacoense* genotype V22 (self-incompatibility alleles $S_{11}S_{13}$) was used as the pollen donor. The target tissues used for the cDNA libraries synthesis were obtained by hand-dissection of ovaries from 2 to 6 days after pollination (DAP) (depericarped ovaries) or from 7 to 17 DAP ovules (fertilization occurs between 36 and 42 h post-pollination). The total RNA from these tissues was isolated as described previously (Jones *et al.*, 1985) and polyadenylated RNA was prepared by Oligo(dT) spin column chromatography using oligo dT agarose beads type VII (GE Healthcare, Baie d'Urfée, QC, Canada). Complementary DNA (cDNA) synthesis was done according to the manufacturer's instruction in the Uni-Zap vector, and library packaging was done with the GigaPack Gold phage extracts (Stratagene, La Jolla, CA, USA). The titer of the primary libraries were approximately 2.8×10^7 pfu/ml (2–6 DAP library) and 1.2×10^7 pfu/ml (7–17 DAP library), with a mean clone length of 1.5 kb. For both libraries, an aliquot representing 10 times the primary library titer was mass excised according to the manufacturer's instruction. Mass excised plasmids were plated on an LB agar supplement with 50 μg/ml kanamycin and 100 μg/ml ampicilin (Sigma-Aldrich Canada Ltd., Oakville, ONT, Canada) to obtain approximately 150 non-overlapping colonies per 132 mm Petri dish. 20 736 individual clones were randomly hand selected from both libraries and transferred to 384-well plates (deep well plates, 200 μl per well).

Virtual subtraction

Sterile Hybond N^+ nylon membranes (GE Healthcare) were placed on Petri dishes containing kanamycin and ampicilin, and clones were transferred to the membranes with a 384 prong microwell plate copier. After overnight growth at 37 °C, a picture of every membrane was taken in order to eliminate colonies that had grown poorly, and that would have also been selected as corresponding to weakly expressed mRNAs. Colonies were denatured on the membranes using the following protocol. Membranes were removed from Petri dishes using forceps and placed in a 15 cm × 15 cm container with 15 ml of denaturation solution (0.5 N NaOH and 1.5 M NaCl) and left for 5 min. The membranes were then transferred to the neutralizing solution (1.5 M NaCl and 0.5 M Tris–Hcl pH 7.4) for 5 min. Membranes were then soaked in 2 × SSC for 30 s, air-dried and UV cross-linked (120 mJ/cm^2). Probes used for the subtraction process were synthesized from double stranded cDNA obtained during the library construction procedure. These corresponded to the unused fraction of the long double stranded cDNAs obtained after size fractionation (fractions 5–8 spanning from 500 bp to > 4 kb) mixed with the smaller rejected cDNAs (fractions 9–11 spanning from 200 bp to ~2.5 kb). The amount of the probe was empirically determined as being optimal when using 100 ng per 10 hybridized membranes. Using too little of the probe rendered the subtraction less efficient as more colonies hybridized weakly, and would thus have been selected. The membranes were exposed at −85 °C with one intensifying screen on Kodak Biomax MR film (Interscience, Markham, ONT, Canada). Prior to the selection of weakly hybridizing colonies, the autoradiogram obtained was compared with the corresponding picture of the bacterial colony growth. Colonies that had not grown were not selected for transfer. From densitometric scan analyses of the autoradiograms, as well as from eye inspection, only the colonies corresponding to the lowest fifth of the hybridizing signal were transferred to 96-well plates containing Terrific Broth (Difco, Sparks, MD, USA) for plasmid purification and sequencing.

Plasmid DNA preparation and sequencing

Colonies were grown overnight at 37 °C in Terrific Broth containing 50 μg/ml kanamycin and 100 μg/ml ampicilin and agitated at 200 rpm on a rotary shaker. Minipreps were performed using Qiagen 96-well miniprep (Qiagen, Mississauga, ONT, Canada) according to the manufacturer's instructions. Plasmid DNA was quantified using absorbance at

260 nm. Sequencing was done at the Centre d'Innovation Génome Québec (Montréal, Canada). Plasmid DNA was purified using 96-well plates Milipore MultiScreen Plasmid (Milipore, Nepean, ONT, Canada). Approximately 200 ng (2 μl) of plasmid DNA and 8 μl of a reaction mixture containing 5 μl of water, 1.5 μl 5$\times$ sequencing buffer, 0.5 μl T3 primer at 20 μM, and 1 μl Big Dye Terminator (Applied Biosystems) was used for the sequencing reaction. The sequencing reactions were performed on a GeneAmp PCR System 9700 (Applied Biosystems), and the cycling conditions were: 96 °C/3 s, 50 °C/15 s, 60 °C/4 min for 30 cycles. Amplification products were analyzed using an automated ABI 3700 DNA Analyzer (Applied Biosystems).

Sequence analysis

EST sequences were assembled into the Sputnik relational database (Rudd *et al.*, 2003) and were cleaned of vector sequence or polylinker remnants using the crossmatch application (http://www.phrap.org). The cleaned sequences were then clustered and assembled using the HarvESTer software (Biomax informatics) that utilizes a suffix tree based clustering algorithm (HPT) and the CAP3 assembly algorithm. The result of the analysis is a collection of singleton and cluster consensus sequences. These were recovered from the Harvester application and were imported back into Sputnik. A control EST dataset, named VIRSC (virtual subtraction control), was generated by randomly selecting (using the Python random function) 50 pools of 4001 (the average number of sequences of SV5 and SV6) TAMU *Lycopersicon esculentum* Ovary ESTs (total of 12 000 EST). These were independently clustered and assembled as described above. For unique match analysis, all plant EST sequences from the EMBL database were downloaded from the EBI (ftp.ebi.ac.uk/pub/databases/embl/release). The sequences were loaded into the open Sputnik database and were taxonomically filtered into groups that represented the eurosids, the euasterids, the caryophyllids and the monocots. Separate datasets representing the available gene sequences from the complete plant genomes were also included. Blast tables were created for each taxonomic group and the *S. chacoense* sequences were compared to each dataset using BLASTN; results were filtered arbitrarily at 1^{E-10}. All *S. chacoense* sequences were classified for overlap with the sequence collections. The number of sequences that overlapped with each sequence collection were counted as were the number of sequences that appeared unique to the single collection. Classification of the 30 RLKs was performed based on the Blastp homology of the available kinase and extracellular domains to their closest relatives in *Arabidopsis* according to Shiu and Bleecker (2001). Both 5′ and 3′ sequences obtained from each RLK ESTs were blasted and retrieved the same *Arabidopsis* ortholog, strengthening the classification.

Real-time RT-PCR

For total RNA isolation, tissues were ground in liquid nitrogen, and the powder was kept at −80 °C until RNA extractions were performed. RNA was extracted using the RNeasy Plant Mini Kit from Qiagen. RNA concentration was determined by measuring its absorbance at 260 nm and verified (adjusted if necessary) by agarose gel electrophoresis and ethidium bromide staining. Two micrograms of total RNA in a final volume of 11.7 μl was denatured at 65 °C for 15 min and quickly transferred to ice. 14.3 μl of an RT mix containing 1X First Strand Buffer, 0.01 M dithiothreitol (both from Invitrogen, Burlington, ONT, Canada), dNTP 1mM/each (Roche), 0.1 μg of oligo $d(T)_{20}$ (Sigma), 10 units of M-MLV RT (Invitrogen), and 2.15 units of RNase Inhibitor (Roche, Laval, QC, Canada) were added to the denatured RNA for a total volume of 26 μl. The following RT program was used: 25 °C for 10 min, 37 °C for 60 min, 99 °C for 5 min in a GeneAmp PCR System 9700 (Aplied Biosystem). Real-time PCR was performed in triplicate and, as a control, ubiquitin primers derived from a ubiquitin cDNA sequence obtained from the library were used. Primers for specific receptor kinase were made from the 3′-UTR of each cDNA and were designed to generate a product between 150 and 250 bp in length. All primer pairs used are listed in Table S1 (supplementary material). About 4 μl of RT was added to a PCR mix of 46 μl containing 20.45 μl sterile water, 1$\times$ PCR buffer (Qiagen), 2.5 mM $MgCl_2$, 0.2 mM of each dNTP, 0.25 μM of each primers, 8% glycerol, 3% DMSO, 3.3 μl of a 10 000 $\times$ dilution of SYBR-Green I (Sigma) and 0.025 unit of HotStart Taq DNA polymerase (Qiagen). Each primer was first tested for amplicon size and the presence of primer

dimers through vizualization of PCR product on agarose gel, following ethidium bromide staining. For real-time PCR amplification, the following PCR program was used: 95 °C for 10 min, 94 °C for 30 s, 6 °C for 30 s, 72 °C for 30 s; steps 2–4 were repeated 40 times in a Mx4000® Multiplex QPCR System (Stratagene), and fluorescence readings were done at 60 °C. All primers were obtained from Sigma-Aldrich.

Results

Library screening for low abundance mRNAs

Functional genomics studies depend on the identification of all expressed genes, the transcriptome, within a given organ, tissue, cell type, or under specific physiological conditions. One major hurdle in EST sequencing projects is the high redundancy of sequenced clones and the poor representation of genes that are weakly expressed. We have used a modified subtraction screen based on the virtual subtraction procedure of Li and Thomas (1998), and the negative screen of Nelson *et al*. (1999), as a method to enrich EST pools for rare messenger RNAs. The method is based on the hybridization of arrayed cDNA clones with a radiolabeled probe made from total cDNAs prepared from the target tissues that have been used to construct the libraries to be screened. Since the probe used to screen the library corresponds exactly to the RNA profile of the tissue used for the library, without an amplification step, this results in a non-biased, sensitive probe, that represents the actual transcriptome in a given tissue. The rationale behind the use of a ''self-probe'' is that any weakly abundant transcripts will also be underrepresented in the probe, and thus, a low hybridization signal will be observed for the corresponding cDNAs following membrane hybridization (Figure 1). Two cDNA libraries were made from 2 to 6 DAP depericarped ovaries (named SV6) and from isolated ovules harvested 7–17 DAP (named SV5), which covered seed development from early fertilization to the period when the ovules were bearing late torpedo stage embryos. After mass excision of the libraries and initial plating, bacterial colonies were transferred to 384-well plates (54 plates for each library, for roughly 20 000 colonies per library). The colonies were plated on nylon membranes and grown overnight. Each membrane was photographed and compared with the hybridization profile obtained, in order to eliminate

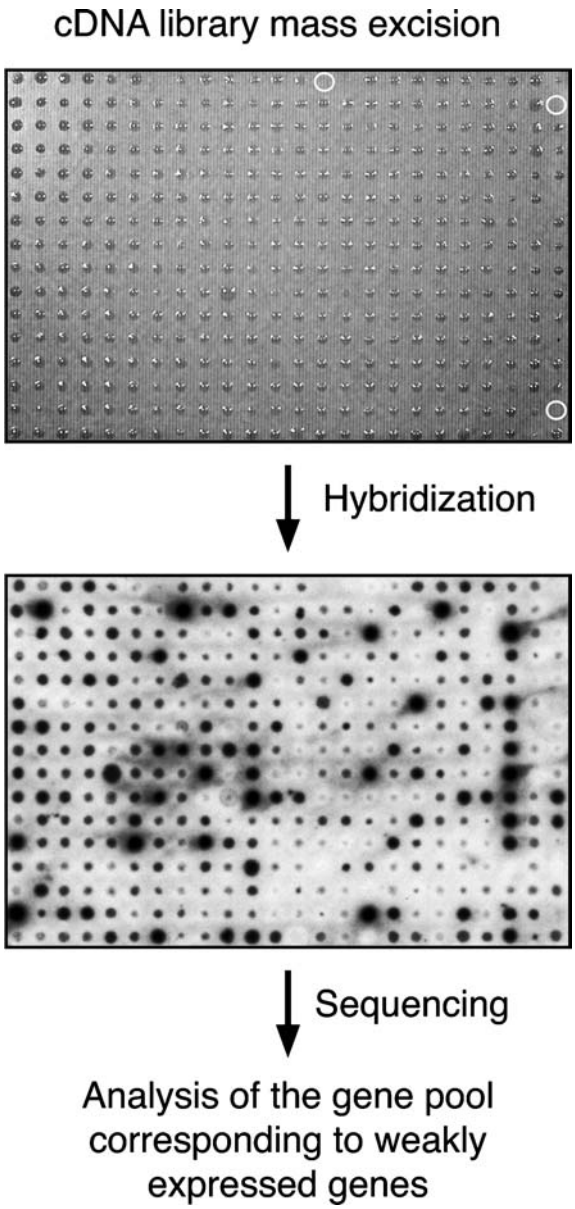

Figure 1. Schematic representation of the virtual subtraction approach. In the first step, mass-excised clones were arrayed onto nylon membranes placed on Petri dishes containing LB media and the appropriate antibiotic and incubated overnight at 37 °C. Spots where no growth was observed were noted and not selected for sequencing. In the second step, plasmid DNA was cross-linked to the nylon membrane and the probe was hybridized to the membranes. After appropriate washing, the film was exposed to the membrane. All clones yielding a very weak signal or none at all (except for those that had not grown in the first step) were selected for sequencing.

colonies that did not grow (Figure 1, white circles). After densitometric scanning of the autoradiograms and eye inspection, colonies that displayed a hybridization signal corresponding to the lowest twenty percent were selected for further analysis. The size of the colony was also taken into account for correction purposes. The selected colonies were transferred into 96-well plates for sequencing.

Expressed sequence tag analysis

Single-pass sequences from the 5′-termini of 7741 cDNAs were obtained, with an average of more than 500 bp per EST sequence (data not shown). Table 1 shows the results from the sequence clustering analysis of the SV5 and SV6 sequence collections. In both cases, the redundancy of sequenced clones was low. This is reflected by the relatively high number of singletons and concomitantly low number of contigs obtained. A mean of only two sequences per contig was found in both cases. A direct comparison of statistical data from our EST set and an EST set generated from the large tomato EST sequencing project (Van der Hoeven *et al*., 2002) cannot be made due to the differences in the number of sequences to be compared. As the number of sequences increases, the likelihood of finding new unique genes decreases (diminishing returns). In order to compare the efficiency of the subtraction method used, a control EST dataset named VIRSC (virtual subtraction control) was generated by randomly selecting 50 pools of 4001 ESTs (the average of the SV5 and SV6 sets) from the non-normalized TAMU *Lycopersicon esculentum* ovary ESTs (Van der Hoeven *et al*., 2002). These 4001 randomly chosen ESTs represent an equivalent control, introducing minimal bias related to sample size (saturation) and tissue sampling. Since 50 independent tomato ovary EST pools were generated and averaged, this also reduced the sampling bias within the pool of 12 000 ovary ESTs. Analysis of our EST data set (Table 1) reveals that the number of unigenes obtained was 91% for SV5 library and 82% for SV6 library. By comparison, the non-subtracted and averaged VIRSC datasets contained only 54% of unigenes. This high percentage of unigenes obtained is reflected in the low numbers of sequences found in contigs, the number of sequences per contigs when compared to the VIRSC data set (2.23 in the pooled SV5/SV6 dataset versus 3.42 in the tomato ovary VIRSC set), and in the percentage of the gene set covered with a limited number of ESTs sequenced (numbers based on the predicted gene number (35 000) in the tomato genome, Van der Hoeven *et al*., 2002). Thus, the subtraction method used provides both a way to obtain weakly represented transcripts, as well as an efficient and cost effective method to obtain substantial gene set coverage per sequence run.

In order to assess the number of totally new genes in our dataset (genes not found in any of the publicly available databases), we performed a computational comparison between our unigene set and other unigene databases, including *Arabidopsis*, rice, poplar, and *Medicago* genomes, the database of some of the core Eudicot, all Monocots, the animal and fungi kingdom and a non-redundant unigene database (see Table 2). Since fully sequenced genomes represent a very large proportion of unigenes of the Eurosid and Asterid classes, all completed genome unigenes were remove from higher classes and analyzed independently. Short

Table 1. Comparison of non-subtracted and subtracted EST sequence sets from tomato and *S. chacoense*.

	S. chacoense			Tomato	
	Subtracted			Non-subtracted	
	SV5	SV6	SV5/SV6	Ovary	Total
Total number of EST	3406	4335	7741	4001	120 892
ESTs in contigs	520	1327	2470	2573	106 833
Total number of contigs	227	557	1103	752	13 215
Total number of unigenes	3113	3565	6374	2180	27 274
Percentage of unigenes	91	82	82	54	23
Average number of repeat of a sequence in a contig	2.29	2.38	2.23	3.42	8.08
Percentage of the gene set covered	8.9	10.2	18	6.2	78
Percentage of the gene set covered per sequence run	0.26	0.23	0.23	0.15	0.064

Table 2. Occurrence of matches (known genes) and unique matches (novel genes) against the main unigene databases.

Database	Matches	Unique matches
Arabidopsis genome	4304	29
Rice genome	1612	2
Poplar genome	2951	4
Medicago genome	1475	4
Euasterid unigenes	2792	6
Eurosid unigenes	3860	26
Caryophyllid unigenes	2481	6
Monocot unigenes	2633	6
Vertebrates	1531	0
Invertebrates	2111	2
Fungi	1560	0
Non-redundant protein database	4565	70
Short sequences	28	10
Low coding potential	1323	1100
All *S. chacoense* sequences	6272	387

sequences (less than 150 bp), and sequences with low coding potential (possibly corresponding to 3′UTR) were rejected from the analysis to avoid an artificial inflation of the number of unique matches. It is to be noted that *S. chacoense* belongs to the Asterid clade. A total of 387 *S. chacoense* sequences (5% of the total EST set, 6% of the unigene set) corresponded to unique matches. These correspond to sequences that by definition have no match elsewhere in the DNA sequence databases tested, are long (>150 nt) and have significant protein coding potential. A similar figure (5%) as been observed for other EST collections (Van der Hoeven *et al.*, 2002).

Functional annotation

The subtracted *S. chacoense* unigenes and randomly selected EST pools from tomato ovaries (VIRSC) were compared with the *Arabidopsis thaliana* proteome using the BLASTX algorithm (Altschul *et al.*, 1990). Homology based annotation was performed using the closest match obtained against the *Arabidopsis* proteome with an arbitrary threshold for expected values smaller than or equal to 1^{E-10}. All ESTs with an *E*-value greater than 1^{E-10} were classified as unknown proteins. Each clone was classified in a functional category according to the criteria used by the Munich Information Center for Protein Sequences (MIPS – http:// mips.gsf.de). Using these criteria, 72% of all *S. chacoense* ESTs could be fitted into defined functional categories (Figure 2). Percentages from EST belonging to each functional categories were compared with the tomato ovary VIRSC set (Table 3). Although overall similar, a statistical analysis (χ^2-test) was performed to determine if these differences were significant and to determine if the subtraction strategy had introduced a bias into the gene pool isolated. A χ^2 value of 78.3322 ($P < 1^{E-5}$) confirmed that there were significant differences between the subtracted set versus the non-subtracted set for seven of the functional categories (Table 3). Four functional categories were slightly more represented in the virtual subtracted pool (Cell Growth, Cell Division and DNA Synthesis; Protein Synthesis; Protein Destination; Cellular Organization) while three functional categories were slightly underrepresented (Metabolism; Cellular Communication; Development). If we take into account the percentage of unigenes obtained both in the subtracted and non-subtracted gene set (the redundancy is quite higher in the non-subtracted EST set) and apply it per functional categories, then only the Development category would be slightly less represented in the *S. chacoense* subtracted pool.

Receptor-like kinases expression analysis

Three obvious drawbacks of the subtraction method used can be considered. Firstly, if a bacterial colony did not grow well, it could have been chosen as corresponding to a weakly expressed gene. This problem was easily resolved by comparing the profile of bacterial growth to the

hybridization results (Figure 1). Secondly, colonies that grew well but which harbored a plasmid devoid of an insert would have also been selected, since they would have given a background hybridization signal. From the initial library construction, we knew that very few empty vectors were present, since restriction analysis of 50 randomly selected colonies revealed no vectors without an insert. A more thorough analysis was performed on the EST sequenced and less than 0.1% of the colonies (six in total) were found having empty vectors. Furthermore, only 28 sequenced plasmids

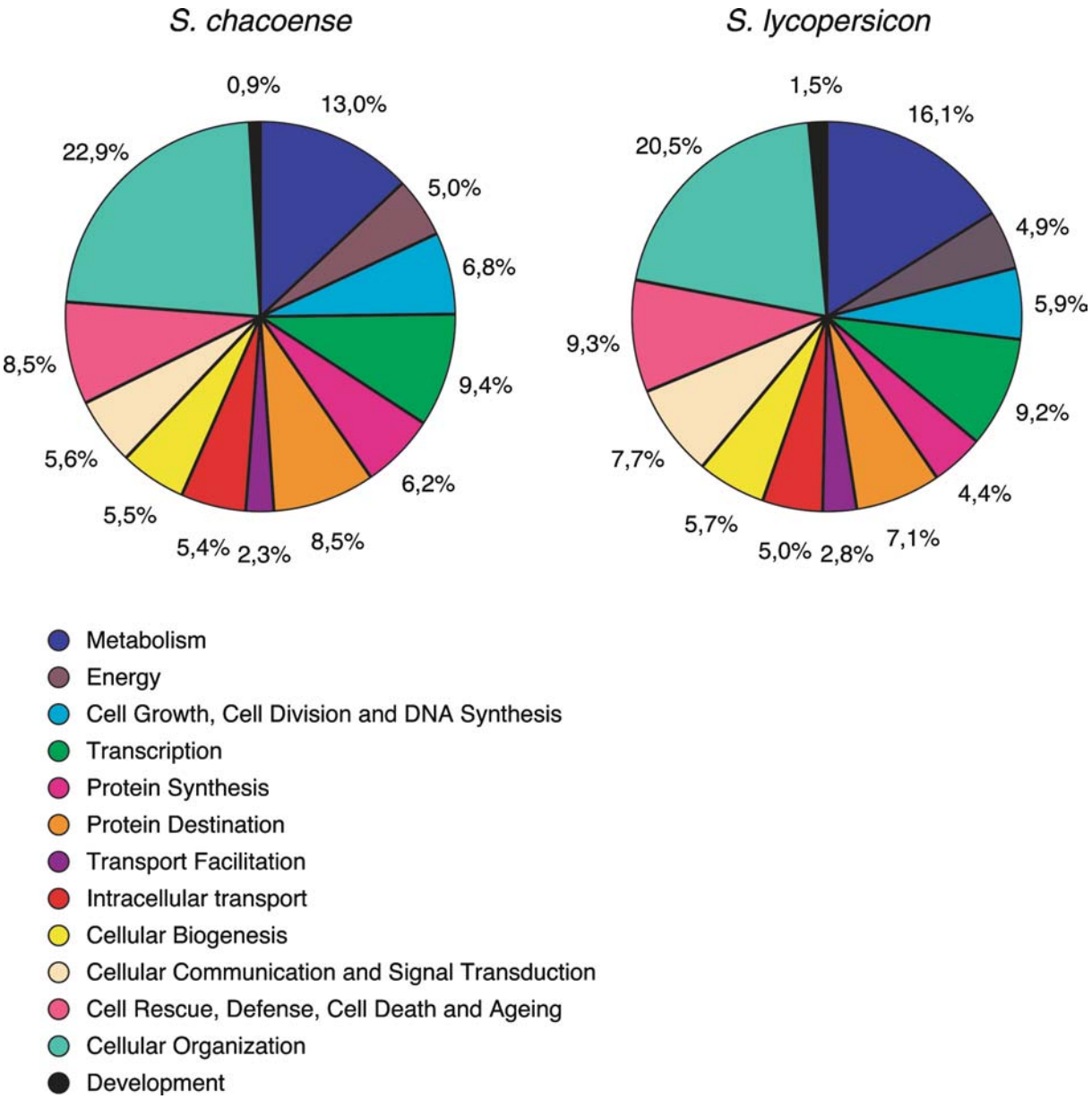

Figure 2. The subtracted *S. chacoense* unigenes and the tomato ovaries VIRSC set were Blasted (BlastX) against the *Arabidopsis thaliana* proteome. Annotation was done based on the closest match obtained against the *Arabidopsis* proteome using a threshold of an expected value smaller or equal to 1^{E-10}. All EST having an *E*-value larger than 1^{E-10} were classified as unknown proteins or as not yet clearly classified. Role categorization was attributed according to the criteria used by the Munich Information Center for Protein Sequences (MIPS) (http:// mips.gsf.de).

contained inserts of less than 150 bp. Thirdly, an important bias may exist within our sampling and classification of ovule sequences. Since a 'self-probe' strategy was used, there is a possibility that the genes selected are a mixture of ovule-specific and ovule-predominant genes and genes that are lowly expressed in the target tissue (ovule) but strongly expressed in other tissues. To test this hypothesis, we have selected all the receptor-like kinases identified from our gene pool in female reproductive tissues and determined their expression profile using real-time quantitative RT-PCR analyses (Figure 3). Thirty individual RLKs were retrieved from our gene pool and sequenced from both ends. No redundancy was found among theses RLKs as none was found twice. Classification of the 30 RLKs was performed based on the homology of the kinase and extracellular domains to their closest relatives in *Arabidopsis* according to Shiu and Bleecker (2001) (Table S3, supplementary material). All the receptor kinases were named from ORK1 to ORK30 for Ovule Receptor Kinase, after the tissues in which they were found. Twenty-five corresponded to the LRR-RLK group (Table 3), the largest RLK group in *Arabidopsis*, and which represents more than 50% of all RLKs. Surprisingly, 28 of the analyzed RLK showed their strongest expression level in ovary or young fruit tissues (Figure 3). The only two RLK analyzed that did not show their strongest expression level in the female reproductive tissues were members of the L-lectin family. These RLK, ORK23 and ORK30, showed their peak expression levels in pollen and leaves respectively. With the exception of ORK21, which was undetectable in pollen and of ORK23, whose expression was highest in pollen, pollen mRNA was used as the calibrator for all quantitative RT-PCR analyses.

Transcriptional regulation of the ORKs following pollination and fertilization

Since most of the isolated RLKs were predominantly expressed in female reproductive tissues, we determined if these RLKs were regulated at the transcriptional level following pollination, fertilization, and later in ovules during early seed development. An arbitrary threshold value of 1.5 fold variation was set and genes were scored as induced after pollination or reduced after pollination by comparing expression levels at 0 h (unpollinated) and 24 h after pollination. Similarly, genes were scored as induced after fertilization or reduced after fertilization by comparing expression profiles between 24 and 48 h after pollination (fertilization takes place from 36 to 42 h post-pollination in *S. chacoense*), and later during early embryogenesis by comparing the average expression level between 8 DAP and 12 DAP with control ovules (0 h). These later stages (8 DAP and 12 DAP) correspond to ovules bearing embryos from the 8 to 16 cells stages to globular embryos,

Table 3. Comparison of observed versus expected frequency of each functional categories using a table frequency analysis with linear-log between the *S. chacoense* ESTs and the tomato VIRSC EST set.

Class	Observed numbers	%	Expected numbers	%	Significant differences
Metabolism	829	13.0	644	16.1	−
Energy	319	5.0	196	4.9	
Cell Growth, Cell Division and DNA Synthesis	433	6.8	236	5.9	+
Transcription	599	9.4	368	9.2	
Protein Synthesis	395	6.2	172	4.3	+
Protein Destination	542	8.5	284	7.1	+
Transport Facilitation	147	2.3	112	2.8	
Intracellular Transport	344	5.4	200	5	
Cellular Biogenesis	351	5.5	228	5.7	
Cellular Communication/Signal Transduction	357	5.6	308	7.7	−
Cell Rescue, Defense, Cell Death and Ageing	542	8.5	372	9.3	
Cellular Organization	1460	22.9	820	20.5	+
Development	57	0.9	60	1.5	−
	6375		4000		

Significant differences: a statistical analysis (χ^2-test) was performed to determine if the differences were significant and to determine if the subtraction strategy had introduced a bias into the gene pool isolated. A χ^2 value of 78.3322 ($P < 1^{E-5}$) confirmed that there were significant differences between the subtracted set versus the non-subtracted set for seven of the functional categories.

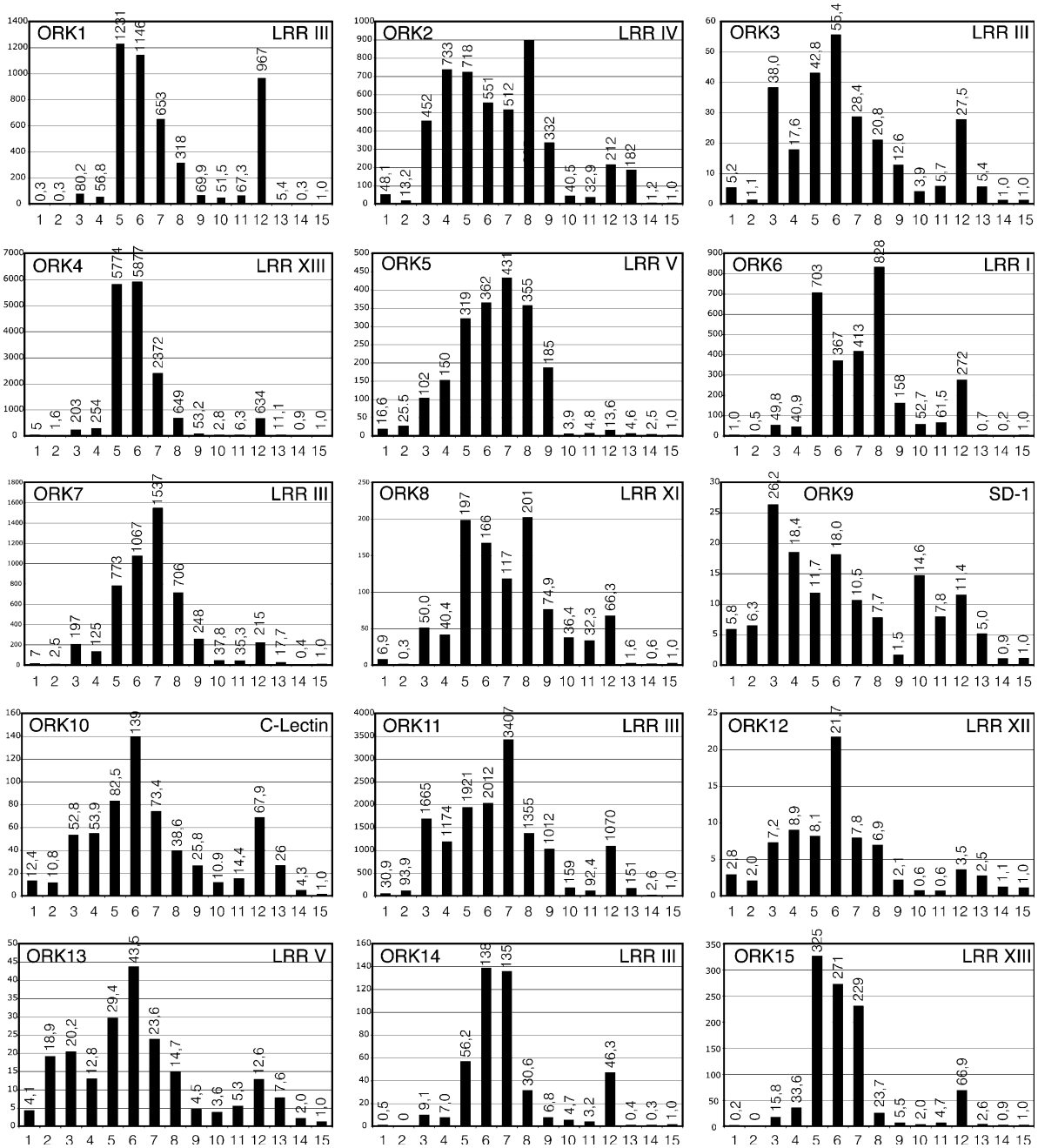

Figure 3. Expression profile of the 30 *S. chacoense* Ovule Receptor Kinases isolated from our subtracted pool. Two (2) μg of total RNA from various tissues was used in a real-time quantitative PCR experiment. Ubiquitin was used as the normalizer, and each amplification was performed in triplicate. In all cases (except ORK23 and ORK30) pollen was used as the calibrator. ORK23 and 30 use 'Style 0 h' as their calibrator. Lanes: 1. Style 0 h, 2. Style 48 h, 3. Ovary 0 h, 4. Ovary 24 h, 5. Ovary 48 h, 6. Ovary 72 h, 7. Ovary 96 h, 8. Ovary 8 days, 9. Ovary 12 days, 10. Roots, 11. Stem, 12. Leaves, 13. Petals, 14. Anthers, 15. Pollen. RLK specific primers were designed from the 3'UTR of each RLKs.

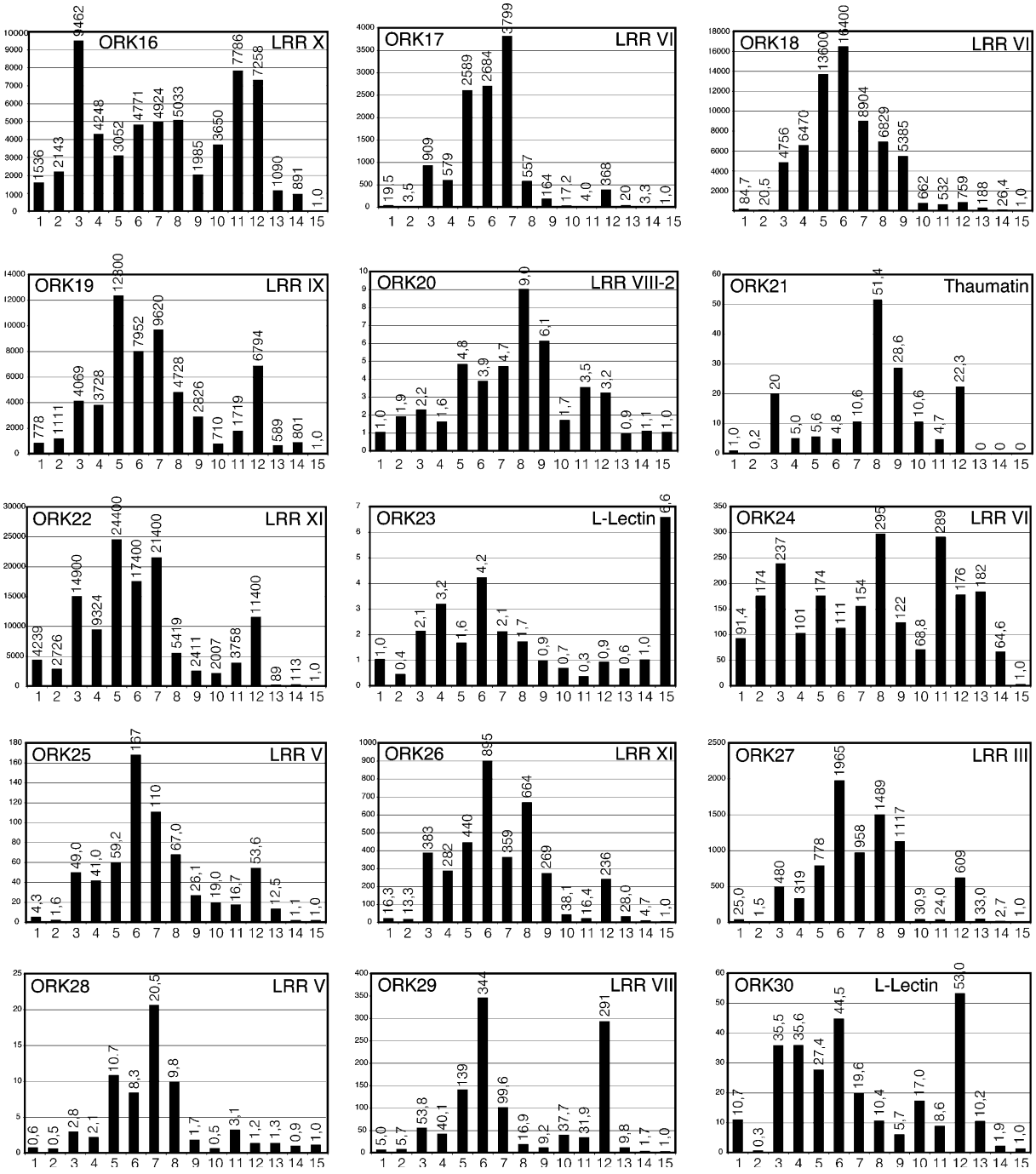

Figure 3. (Continued).

Table 4. Induction and repression level after pollination (24 h/0 h), fertilization (48 h/24 h) and later during embryogenesis (average of 8 days and 12 days/0 h).

ORK	Class	Pollination				Fertilization		Later during embryogenesis	
		Induction		Repression					
		Style	Ovary	Style	Ovary	Induction	Repression	Induction	Repression
ORK1	III					21.7		2.4	
ORK2	VI		1.6	3.6					
ORK3	III			4.6	2.2	2.4			2.3
ORK4	XIII			3.1		22.7		1.7	
ORK5	V	1.5	1.5			2.1		2.6	
ORK6	I			2.0		17.2		9.9	
ORK7	III			2.8	1.6	6.2		2.4	
ORK8	XI			19.8		4.9		2.8	
ORK9	SD-1						1.6		5.7
ORK10	C-Lectin					1.5			1.6
ORK11	III	3.0				1.6			
ORK12	XII								1.6
ORK13	V	4.5			1.6	2.3			2.1
ORK14	III			12.0		8.1		2.0	
ORK15	XIII		2.1	3.5		9.7			
ORK16	X				2.2				2.7
ORK17	VI			5.5	1.6	4.5			2.5
ORK18	VI			4.1		2.1			
ORK19	IX					3.3			
ORK20	VIII-2	1.9				3.0		3.4	
ORK21	Thaumatin			4.2	4.0			2.0	
ORK22	XI			1.6	1.6	2.6			3.8
ORK23	L-lectin		1.5	2.4			1.9		1.6
ORK24	VI	1.9			2.3	1.7			
ORK25	V			2.8					
ORK26	XI					1.6			
ORK27	III			16.2	1.5	2.4		2.7	
ORK28	V					5.2		2.0	
ORK29	VII					3.5			4.1
ORK30	L-lectin			31.5					4.4

respectively. Furthermore, pollination effects can be detected both locally (in the style) or at a distance in the ovary before the pollen tubes reach the ovary (Lantin *et al.*, 1999). These two possibilities were thus treated separately (Table 4). Figure 4 shows a Venn diagram representation showing the relationship between the 30 *S. chacoense* ORKs and their expression profile during pollination and fertilization. The most striking results were observed in the induced after fertilization class. Twenty-three out of thirty (77%) ORK genes were induced following fertilization and, of these, 22 belonged to the large LRR-RLK class. Only ORK10 belonged to the C-lectin class and its induction level was at the selected threshold value only. Of these, seven ORKs showed a greater than 5-fold increase in mRNA levels: ORK 28 (5.2-fold), ORK7 (6.2-fold), ORK14 (8.1-fold), ORK15 (9.7-fold), ORK6 (17.2-fold), ORK1 (21.7-fold), and ORK4 (22.7-fold). Conversely, 16 ORKs (53%) showed a significant decrease in expression following pollination in the style. Five of these showed a greater than 5-fold decrease in mRNA levels: ORK17 (5.5-fold), ORK-14 (12.0-fold), ORK27 (16.2-fold), ORK8 (19.8-fold), and ORK30 (31.5-fold). Out of those that were induced after fertilization in the ovary, 48% (11/23) had their expression concomitantly reduced in the style. Also noteworthy, while there are only 5 out the 30 genes that are not part of the LRR family, four of these five genes were among the non-induced or repressed genes.

In the 'later during embryogenesis' class eleven ORKs were induced while another eleven ORKs had their expression level reduced above the threshold level. Again, in the genes had their expression

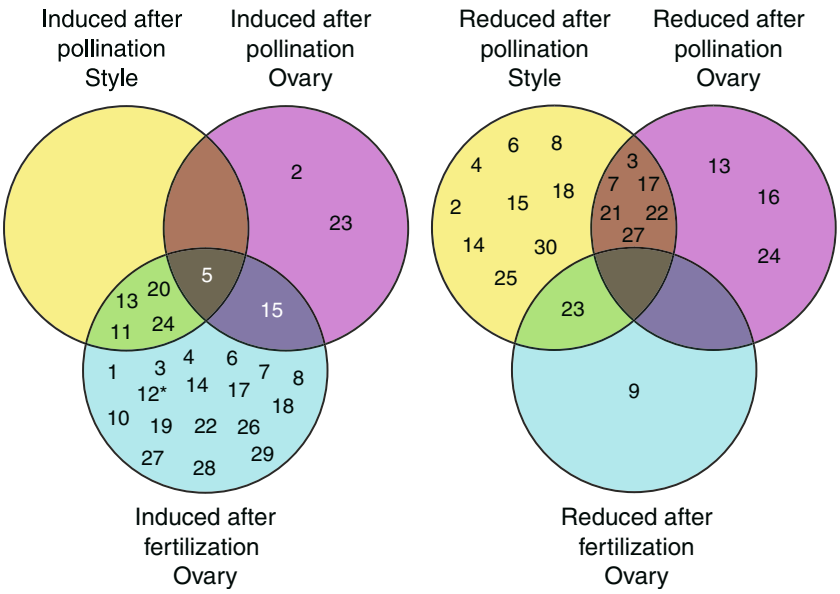

Figure 4. Venn diagram showing the relationship between the 30 *S. chacoense* ORKs and their expression profile during pollination and fertilization. Individual ORKs were placed in their respective expression categories based on the fold-induction or fold-repression threshold value of 1.5 as determined in Table 4. Numbers correspond to their respective ORK. *Although ORK12 did not show any immediate mRNA increase 48 h after pollination, a 3-fold transient increase was observed 72 h after pollination.

level reduced, of the five ORKs that did not belong to the LRR-RLK class, four were repressed. Another interesting aspect is revealed by the presence of a group of thirteen ORKs that are modulated by pollination, but at a distance in the ovary, since their expression level can be increased (ORK2, ORK5, ORK15, and ORK23) or decreased (ORK3, ORK7, ORK13, ORK16, ORK17, ORK21, ORK22, ORK24, and ORK27) before the pollen tubes reach the ovary. Thus, expression of these RLKs is either modulated developmentally, or by long-distance signaling cues as determined before for other genes (Lantin *et al.*, 1999).

Discussion

During plant sexual reproduction, pollination and fusion of the gametes are expected to trigger a vast array of signal transduction events that will lead to a reorganization of the expressed gene set, which will in turn direct growth and development of the embryo and give rise to a mature fruit. In the present work, which aimed at gaining a better knowledge of the transcriptome induced following fertilization and early embryogenesis, we report the generation of more than 7700 EST sequences from *Solanum chacoense* ovules taken at diverse stages of embryo development, with a particular focus on genes involved in signal transduction events. The presence of a self-incompatibility system in *S. chacoense* not only enables a total control over pollination and fertilization timing, but also makes it possible to address issues related to genes differentially regulated by compatible and incompatible pollinations as well as genes regulated at a distance in the ovary following pollination (Lantin *et al.*, 1999).

Despite the fact that the fruits of several Solanaceous species have been extensively domesticated (e.g. tomato, eggplant, tomatillo and tamarindo) and are very important from an economic point of view, gene expression studies in the female gametophyte have not been extensively characterized until recently (Drews and Yadegari, 2002; Van der Hoeven *et al.*, 2002; Hu *et al.*, 2003; de Folter *et al.*, 2004; Hennig *et al.*, 2004; Sprunck *et al.*, 2005). Several key aspects of plant development are known and are suspected to be controlled by protein kinases. Because of their capacity to amplify very subtle stimulus, components of signaling modules, such as receptor kinases, mitogen-activated protein kinases (MAPKs), and small

signaling ligands, are generally expressed at low levels and are weakly represented in most EST sequencing projects (Hu *et al*., 2003). In order to increase the number of unigenes found, recent large scale EST sequencing projects in Solanaceous species (Van der Hoeven *et al*., 2002; Ronning *et al*., 2003) have relied on the use of a large number of diversified libraries, while other methods used to decrease gene redundancy (based on library normalization) have also been developed. Despite all the disadvantages linked to sequence redundancy, EST sequencing remains a method of choice for gene discovery as well as a cost-effective method for tagging thousands of genes from a given organism. This is reflected by the fact that nearly 6 millions (5 817 901) *viridiplantae* ESTs had been deposited at NCBI (Jan 01, 2005). Yet only 53 *S. chacoense* cDNA sequences have been deposited at NCBI and no single ESTs have been deposited so far. Thus, the *S. chacoense* data presented here is therefore the first large scale sequencing effort addressing this organism that has been widely used for introgression of valuable traits in the cultivated potato (Hawkes, 1990), and for gametophytic self-incompatibility studies (Matton *et al*., 1999; O'Brien *et al*., 2002). In order to analyze the transcriptional program induced during embryo development, libraries from mixed-stages fertilized ovules were selected for a medium scale EST sequencing project. To increase the isolation of weakly expressed mRNAs; reduce the redundancy of the EST dataset; and decrease the cost related to DNA sequencing, the selection of the clones to be sequenced was preceded by a virtual subtraction (negative selection) of the target tissue cDNAs. The virtual subtraction screen was first mentioned by Li and Thomas (1998) and was developed to isolate tissue-specific genes, and genes represented by low-abundance mRNAs in plant embryos. It uses random-primed PCR cDNA probes derived from non-target tissue mRNAs converted to double-stranded cDNAs. A similar strategy termed negative subtraction hybridization was used by Nelson *et al*. (1999) to isolate weakly expressed transcripts from human prostate tissue. It used high-density cDNA clone arrays that were hybridized to a first-strand cDNA probe derived from mRNAs isolated from the same tissue used to build the library to be screened. The authors selected the cDNA clones representing the lowest quartile of the hybridization intensities for sequencing. The methodology resulted in a significant decrease in redundancy from the negative selection libraries compared to non-normalized library, but only limited EST numbers were used, thus introducing a bias due to the limited sampling depth of the libraries (average of less than a 1000 clones sampled for each library). Nonetheless, from this limited data set, a 3-fold decrease in redundancy in ESTs was observed, from 33% in a non-normalized library to 11% in the negative selection library. In two recent studies, two research groups also used a negative selection screen based on the hybridization of the target tissue library with a cDNA probe derived from first-strand reverse transcribed mRNAs isolated either from the same tissues (Vodkin *et al*., 2004) or from fungal cells subjected to different physiological conditions (Ray *et al*., 2004). In the study involving plant tissue, Vodkin *et al*. (2004) used the negative selection screen to isolate weakly expressed genes from an immature soybean cotyledon library. This strategy increased the number of new genes identified from the ESTs sequenced from the negatively subtracted pool when compared to the non-normalized library. Again, the EST numbers used were too low (only 931 sequences came from the non-subtracted library and 1528 were sequenced from the negatively subtracted pool), thus introducing a bias due to the limited sampling depth of the libraries. For the current study, we used a self-probe strategy that used doubled-stranded cDNAs labeled to high specific activity. Since the probe used to screen the library corresponds exactly to the RNA profile of the tissue used for the library, this results in a non-biased sensitive probe, that represents the actual transcriptome in a given tissue. In our hands, the use of a radioactively-labeled first-strand cDNA probe derived from reverse-transcribed mRNAs showed a lower sensitivity than the use of a double-stranded cDNA pool as the starting material (Chantha, S.-C. and Matton, D. P., unpublished results). With a first-strand cDNA probe, almost half the colonies would have been scored as negatives (close to background level), compared to approximatively 20% with the double-stranded cDNA probe procedure used in this study. Furthermore, when we used a first-strand cDNA probe strategy, the same profiles were obtained, irrespective of the mRNA source used. For

example, no differences were observed between a probe made from whole flowers compared to ovaries, or from a mixture of root, stem, leaf and anther tissues, or from unpollinated pistils, suggesting a lesser degree of discrimination with the use of such probes (Chantha, S.-C. and Matton, D. P., unpublished results), and indicating that the same highly expressed genes are found in all tissues examined. Without library normalization, recent EST sequencing efforts of the same magnitude as the one presented here, produced a much lower percentage of unigenes, from 43.1% for 7106 ESTs sequenced from sunflower embryonic libraries (Ben *et al*., 2005), to 47.7% for 11 954 ESTs from various tissues in cassava (Lopez *et al*., 2004), and to 50.5% for 15 781 ESTs from common bean libraries also covering various tissues (Ramirez *et al*., 2005). These numbers are quite similar with the one obtained from our VIRSC non-normalized control set (54%, Table 1). Thus the virtual subtraction used in the present study, considerably increased the number of unigenes isolated from each library screened (91% from the SV5 and 82% from the SV6 library, respectively), for a total of 82% considering the two libraries together (6374 unigenes out of 7741 ESTs). Furthermore, comparison of the two libraries that correspond to different developmental stages indicates only limited redundancy at the gene expression level since only 304 unigenes were shared between the two libraries. A total of 6678 unigenes (3113 and 3565 unigenes for SV5 and SV6, respectively) was obtained from the sum of the two libraries, but this was only reduced to 6374 unigenes when the two libraries were merged for analysis (Table 1). Thus the contribution of each library is quite distinct and complementary, and suggests a complex transcriptional shift between the various stages of the ovules and the embryo it bears during early stages of embryo development. This was also observed using stage specific libraries in the sunflower embryo EST project (Ben *et al*., 2005).

Another interesting observation is that a total of 387 (5%) *S. chacoense* sequences corresponded to unique matches (Table 2). A similar figure had been observed for the tomato EST collection when compared to available plant sequences (Van der Hoeven *et al*., 2002). Thus the subtraction not only led to the identification of ovule-expressed genes but to the finding of new genes, corresponding to roughly 1% of the predicted *Solanum lycopersicon* genome with only a limited sequencing effort, and although already more than 400 000 sequences from *Solanum tuberosum* and *Solanum lycopersicon* ESTs have been deposited in Genbank (see the taxonomy browser at http://www.ncbi.nlm.nih.gov/Taxonomy/Browser for details about the number of deposited sequences).

The functional annotation of our 7741 ESTs set revealed that, although the genes sequenced were selected on the basis of their weak expression level, no major difference was observed between the subtracted set and a non-normalized EST set (the VIRSC set of tomato ovary ESTs), as far as percentage of genes found per category (Figure 2). This implies that lowly expressed genes are expressed throughout all functional categories and are not limited to specific categories like Development or Cellular communication and signal transduction. Interestingly, in a study on the transcriptional program during early reproductive stages in *Arabidopsis thaliana*, Hennig *et al*. (2004), came to the conclusion that the transcriptional reprogramming during reproductive development involved genes from all major functional classes, although the authors found a significant enrichment for genes encoding proteins involved in metabolism, transcription, and cellular organization. Similarly, these three functional categories were the most represented in our gene pool (Figure 2).

Our specific interest in signal transduction events involved in ovule and embryo development led us to analyze all the receptor-like kinases found in our gene pool by quantitative RT-PCR analyses over a wide range of tissues and developmental stages (Figure 3). Of the 30 RLKs isolated 25 corresponded to the LRR-RLK group (Table 4), the largest RLK group in *Arabidopsis*. Surprisingly, 28 of the analyzed RLK showed their strongest expression level in ovary or young fruit tissues (Figure 3). The only two RLK analyzed that did not show their strongest expression level in the female reproductive tissues were members of the L-lectin family. This strongly suggests that the subtraction procedure did not introduce any bias toward genes weakly expressed in the target tissue while being strongly expressed elsewhere in the plant. Despite the fact that several of the 25 LRR-RLK found in this EST project fall within subfamilies containing well characterized members, such as *CLAVATA1* (Clark *et al*., 1997) and *HAESA* (Jinn

et al., 2000) from subfamily LRR XI, *BRI1* (Li and Chory, 1997) in subfamily LRR X, *ERECTA* (Torii *et al.*, 1996) in subfamily LRR XIII, and *FLS2* (Asai *et al.*, 2002) in subfamily LRR XII, none of the 30 RLK found was an ortholog of a RLK with an already known function, suggesting that many more RLKs are to be found having important roles during reproductive development. To our knowledge only one of our ORK ortholog has been previously characterized in *Arabidopsis* (Tarutani *et al.*, 2004). The authors have analyzed T-DNA lines for RLK902 (At3g17840 corresponding to ORK11) and generated a double-mutant containing T-DNA insertions in RLK902 and RLK1 (a very similar RLK) and could not observed any phenotype. When the RLK902 promoter was fused to the GUS reporter gene, expression was observed in the root tips, the lateral root primordia, the stipules, and the floral abscission zones. RLK902 and RLK1 are the two closest relative of a large family but display very different GUS staining, suggesting a different function.

Twenty-three out of thirty (77%) ORK genes were induced following fertilization and, of these, 22 belonged to the large LRR-RLK class. Interestingly, a coordinated regulation was found between half the ORKs that were being up-regulated in the ovary after fertilization and down-regulated in the style after pollination (Figure 4). This could suggest that, among the RLKs characterized, some might be involved in pollen-pistil interactions, their expression following the progression of the pollen tube in the style and ovary, while others are clearly fertilization-related only, showing a very low level of expression in style and ovary before pollination and fertilization, and a steep increase in mRNA expression levels following fertilization, as determined for ORK28, ORK7, ORK14, ORK15, ORK6, ORK1, and ORK4 (Figures 3 and 4). On the other hand, genes belonging to the SD-1, L/C-lectin and thaumatin family may not play an active role during fertilization and fruit growth since none are induced at fertilization time and/or are repressed later during embryogenesis above the selected threshold level. The fact that no single ORK was found to be solely pollination-induced might be the consequence of the tissue sampled for the EST project since from the gynoecium, only depericarped ovaries and isolated ovules were used to make the libraries. Apart from responses resulting from a direct interaction between pollen and pistil, another interesting class of ORKs is exemplified by ORKs that are transcriptionally induced or repressed at a distance, before the pollen tube reach the ovary (Figure 4). Pollen tube growth in the style is known to cause major cellular deterioration and ultimately death of the transmitting tissue (Cheung, 1996). Among the cues that have been shown to modulate gene expression at a distance in the ovary, wounding and jasmonates have been shown to be potent inducers (Lantin *et al.*, 1999). Ethylene production, has also been shown to be highly regulated during pollination and would also be a good candidate in modulating responses at a distance in pistil tissues (Singh *et al.*, 1992).

It is noteworthy that, among the 30 ORKs analyzed, 28 showed their peak expression in ovule or ovary tissues, and for 23 of these, leaf was their second highest expressing tissue. One L-lectin RLK (ORK30) had its highest expression level in leaf and then in gynoecium tissue. In a transcriptional profiling experiment using the *Arabidopsis* ATH1 microarray, Hennig *et al.* (2004), also noted that flowers expressed mainly the same genes as leaves, but that there were many more flower-specific than leaf-specific genes. These two results are in agreement with Goethe's hypothesis proposed more than 200 years ago that flower organs represent modified leaves (Goethe, 1831). We also observed that genes belonging to the LRR family nearly all responded positively (induced) following fertilization in the ovary while the opposite pattern was observed for all other families analyzed. In line with these results we observed that some families, namely LRR III, LRR V, LRR VI, and LRR XIII appear to be over-represented in the gynoecium, from 2 to 10 fold, when compared to their respective representation among the *Arabidopsis* RLKs (Table S2, supplementary material). These findings suggest that the female reproductive tissue uses a rather unique set of receptor kinases, which is not expressed at a high level in other tissue and that families of RLK might have evolved to co-regulate precise developmental processes such as embryogenesis.

Earlier estimates from reassociation kinetics had proposed that the number of genes expressed in the male-gametophyte was as high as 20 000 genes (Goldberg *et al.*, 1989), and since the female gametophyte is an actively growing tissue, there is no reason to believe that this figure would be lower

in the later. More recent analyses based on microarray hybridization suggests that around 15 500 genes are expressed in flowers (Hennig *et al.*, 2004), although the detection limit of the microarrays compared to quantitative PCR might underestimate the actual number of genes expressed in reproductive tissues. Nonetheless, out of 7741 ESTs sequenced from a normalized (virtually subtracted) library, 6374 unigenes were isolated, thus representing a large fraction of the total number of genes expressed in female reproductive tissues. This new EST dataset combined with the production of cDNA arrays derived from it should be an invaluable tool in addressing these issues as well as the transcriptional profiling of genes during fertilization and early embryogenesis in a solanaceous species.

Acknowledgements

We thank Gabriel Téodorescu for plant care and maintenance. We also thank Jean-Nicolas Sylvestro for helping with the colony plating and hybridization; Kateryna Vyetrogon, Daphné Dupéré-Richer, Thomas Chevalier-Laliberté, and Harmanjit Singh for helping out with EST annotations; Homère Lamarre-Ménard for his work on our database web interface, and Stéphane Daigle for statistical analyses. Special thanks also to Professor Terry L. Thomas (Texas A&M University) for his guidance and support at the beginning of this project. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and from the Canada Research Chair program. H. Germain, M. O'Brien, and S.-C. Chantha are the recipient of Ph.D. fellowships from NSERC and from Le Fonds Québécois de la Recherche sur la Nature et les Technologies (FQRNT, Québec). D. P. Matton holds a Canada Research Chair in Functional Genomics and Plant Signal Transduction.

References

Adams, M.D., Kerlavage, A.R., Fleischmann, R.D. Fuldner, R.A., Bult, C.J., Lee, N.H., Kirkness, E.F., Weinstock, K.G., Gocayne, J.D. and White, O. 1995. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nature 377: 3–174.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. Lipman, D.J. 1990. Basic local alignment search tool. J. Mol. Biol. 215: 403–410.

Asai, T., Tena, G., Plotnikova, J., Willmann, M.R. Chiu, W.L., Gomez-Gomez, L., Boller, T., Ausubel, F.M. and Sheen, J. 2002. MAP kinase signalling cascade in Arabidopsis innate immunity. Nature 415: 977–983.

Becraft, P.W. 2002. Receptor kinase signaling in plant development. Ann. Rev. Cell Dev. Biol. 18: 163–192.

Ben, C., Hewezi, T., Jardinaud, M.F., Bena, F., Ladouce, N., Moretti, S., Tamborindeguy, C., Liboz, T., Petitprez, M. and Gentzbittel, L. 2005. Comparative analysis of early embryonic sunflower cDNA libraries. Plant Mol. Biol. 57: 255–270.

Bonaldo, M.F., Lennon, G. and Soares, M.B. 1996. Normalization and subtraction: two approaches to facilitate gene discovery. Genome Res. 6: 791–806.

Carninci, P., Shibata, Y., Hayatsu, N., Sugahara, Y. Shibata, K., Itoh, M., Konno, H., Okazaki, Y. Muramatsu, M. and Hayashizaki, Y. 2000. Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. Genome Res. 10: 1617–1630.

Cheung, A.Y. 1996. The pollen tupe growth pathway: its molecular and biochemical contributions and responses to pollination. Sex. Plant Reprod. 9: 330–336.

Clark, M., Hennig, S, Herwig, R, Clifton, S.W., Marra, M.A., Lehrach, H. and Johnson, S.L.Group tW, WU-GSCnEST Group 2001. An Oligonucleotide Fingerprint Normalized and Expressed Sequence Tag Characterized Zebrafish cDNA Library. Genome Res. 11: 1594–1602.

Clark, S.E., Williams, R.W. and Meyerowitz, E.M. 1997. The CLAVATA1 gene encodes a putative receptor kinase that controls shoot and floral meristem size in Arabidopsis. Cell 89: 575–585.

de Folter, S., Busscher, J., Colombo, L., Losa, A. and Angenent, G.C. 2004. Transcript profiling of transcription factor genes during silique development in Arabidopsis. Plant Mol. Biol. 56: 351–366.

Drews, G.N. and Yadegari, R. 2002. Development and function of the angiosperm female gametophyte. Ann. Rev. Genet. 36: 99–124.

Fei, Z., Tang, X., Alba, R.M., White, J.A., Ronning, C.M., Martin, G.B., Tanksley, S.D. and Giovannoni, J.J. 2004. Comprehensive EST analysis of tomato and comparative genomics of fruit ripening. Plant J. 40: 47–59.

Goethe, and J.W.v., 1831. Essai sur la métamorphose des plantes. Translation of Versuch die Metamorphose der Pflanzen zu erklären. J. G. Cotta, Stuttgart.

Goldberg, R.B., Barker, S.J. and Perez-Grau, L. 1989. Regulation of gene expression during plant embryogenesis. Cell 56: 149–160.

Gong, Z., Yan, T., Liao, J., Lee, S.E., He, J. and Hew, C.L. 1997. Rapid identification and isolation of zebrafish cDNA clones. Gene 201: 87–98.

Hawkes, J.G. 1990. The Potato: Evolution, Biodiversity, and Genetic Resources. Smithsonian Institution Press, Washington, DC.

Hennig, L., Gruissem, W., Grossniklaus, U. and Kohler, C. 2004. Transcriptional programs of early reproductive stages in Arabidopsis. Plant Physiol. 135: 1765–1775.

Hu, W., Wang, Y., Bowers, C. and Ma, H. 2003. Isolation, sequence analysis, and expression studies of florally expressed cDNAs in Arabidopsis. Plant Mol. Biol. 53: 545–563.

Initiative, A.G. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408: 796–815.

Jinn, T.L., Stone, J.M. and Walker, J.C. 2000. HAESA, an Arabidopsis leucine-rich repeat receptor kinase, controls floral organ abscission. Genes Dev. 14: 108–117.

Jones, J.D.G., Dunsmuir, P. and Bedbrook, J. 1985. High level expression of introduced chimeric genes in regenerated transformed plants. Embo J. 4: 2411–2418.

Lantin, S., O'Brien, M. and Matton, D.P. 1999. Pollination, wounding and jasmonate treatments induce the expression of a developmentally regulated pistil dioxygenase at a distance, in the ovary, in the wild potato Solanum chacoense Bitt. Plant Mol. Biol. 41: 371–386.

Li, J. and Chory, J. 1997. A putative leucine-rich repeat receptor kinase involved in brassinosteroid signal transduction. Cell 90: 929–938.

Li, Z. and Thomas, T.L. 1998. PEI1, an embryo-specific zinc finger protein gene required for heart-stage embryo formation in Arabidopsis. Plant Cell 10: 383–398.

Lopez, C., Jorge, V., Piegu, B., Mba, C., Cortes, D. Restrepo, S., Soto, M., Laudie, M., Berger, C., Cooke, R., Delseny, M., Tohme, J. and Verdier, V. 2004. A unigene catalogue of 5700 expressed genes in cassava. Plant Mol. Biol. 56: 541–554.

Matton, D.P., Luu, D.T., Xike, Q., Laublin, G., O'Brien, M., Maes, O., Morse, D. and Cappadocia, M. 1999. Production of an S RNase with dual specificity suggests a novel hypothesis for the generation of new S alleles. Plant Cell 11: 2087–2097.

Nelson, P.S., Hawkins, V., Schummer, M., Bumgarner, R., Ng, W.L., Ideker, T., Ferguson, C. and Hood, L. 1999. Negative selection: a method for obtaining low-abundance cDNAs using high-density cDNA clone arrays. Genet. Anal. 15: 209–215.

O'Brien, M., Kapfer, C., Major, G., Laurin, M., Bertrand, C., Kondo, K., Kowyama, Y. and Matton, D.P. 2002. Molecular analysis of the stylar-expressed Solanum chacoense small asparagine-rich protein family related to the HT modifier of gametophytic self-incompatibility in Nicotiana. Plant J. 32: 985–996.

Radelof, U., Hennig, S., Seranski, P., Steinfath, M., Ramser, J., Reinhardt, R., Poustka, A., Francis, F. and Lehrach, H. 1998. Preselection of shotgun clones by oligonucleotide fingerprinting: an efficient and high throughput strategy to reduce redundancy in large-scale sequencing projects. Nucleic Acids Res. 26: 5358–5364.

Ramirez, M., Graham, M.A., Blanco-Lopez, L., Silvente, S., Medrano-Soto, A., Blair, M.W., Hernandez, G., Vance, C.P. and Lara, M. 2005. Sequencing and Analysis of Common Bean ESTs. Building a Foundation for Functional Genomics. Plant Physiol. 137: 1211–1227.

Ray, A., Macwana, S., Ayoubi, P., Hall, L.T., Prade, R. and Mort, A.J. 2004. Negative subtraction hybridization: an efficient method to isolate large numbers of condition-specific cDNAs. BMC Genomics 5: 22.

Ronning, C.M., Stegalkina, S.S., Ascenzi, R.A., Bougri, O., Hart, A.L., Utterbach, T.R., Vanaken, S.E., Riedmuller, S.B., White, J.A., Cho, J., Pertea, G.M., Lee, Y., Karamycheva, S., Sultana, R., Tsai, J., Quackenbush, J., Griffiths, H.M., Restrepo, S., Smart, C.D., Fry, W.E., Van Der Hoeven, R., Tanksley, S., Zhang, P., Jin, H., Yamamoto, M.L., Baker, B.J. and Buell, C.R. 2003. Comparative analyses of potato expressed sequence tag libraries. Plant Physiol. 131: 419–429.

Rudd, S., Mewes, H.W. and Mayer, K.F. 2003. Sputnik: a database platform for comparative plant genomics. Nucleic Acids Res. 31: 128–132.

Shimada, T., Kita, M., Endo, T., Fujii, H., Ueda, T., Moriguchi, T. and Omura, M. 2003. Expressed sequence tags of ovary tissue cDNA library in Citrus unshiu Marc. Plant Sci. 165: 167–168.

Shiu, S.H. and Bleecker, A.B. 2001. Receptor-like kinases from Arabidopsis form a monophyletic gene family related to animal receptor kinases. Proc. Natl. Acad. Sci. USA 98: 10763–10768.

Singh, A., Evensen, K.B. and Kao, T.-H. 1992. Ethylene synthesis and floral senescence following compatible and incompatible pollinations in *Petunia inflata*. Plant Physiol. 99: 38–45.

Soares, M.B., Bonaldo, M.F., Jelene, P., Su, L., Lawton, L. and Efstratiadis, A. 1994. Construction and characterization of a normalized cDNA library. Proc. Natl. Acad. Sci. USA 91: 9228–9232.

Sprunck, S., Baumann, U., Edwards, K., Langridge, P. and Dresselhaus, T. 2005. The transcript composition of egg cells changes significantly following fertilization in wheat (*Triticum aestivum* L.). Plant J. 41: 660–672.

Tarutani, Y., Morimoto, T., Sasaki, A., Yasuda, M.Nakashita, H., Yoshida, S., Yamaguchi, I. and Suzuki, Y. 2004. Molecular characterization of two highly homologous receptor-like kinase genes, RLK902 and RKL1, in Arabidopsis thaliana. Biosci. Biotechnol. Biochem. 68: 1935–1941.

Torii, K.U., Mitsukawa, N., Oosumi, T., Matsuura, Y., Yokoyama, R., Whittier, R.F. and Komeda, Y. 1996. The Arabidopsis ERECTA gene encodes a putative receptor protein kinase with extracellular leucine-rich repeats. Plant Cell 8: 735–746.

Van der Hoeven, R., Ronning, C., Giovannoni, J., Martin, G. and Tanksley, S. 2002. Deductions about the number, organization, and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing. Plant Cell 14: 1441–1456.

Vodkin, L.O., Khanna, A., Shealy, R., Clough, S.J. Gonzalez, D.O., Philip, R., Zabala, G., Thibaud-Nissen, F., Sidarous, M., Stromvik, M.V., Shoop, E., Schmidt, C., Retzel, E., Erpelding, J., Shoemaker, R.C., Rodriguez-Huete, A.M., Polacco, J.C., Coryell, V. Keim, P., Gong, G., Liu, L., Pardinas, J. and Schweitzer, P. 2004. Microarrays for global expression constructed with a low redundancy set of 27 500 sequenced cDNAs representing an array of developmental stages and physiological conditions of the soybean plant. BMC Genomics 5: 73.

Table S1

Click here to download Supplementary material: Table S1 supplementary material.doc

Supplementary material

Table S1. Ovule Receptor Kinases Real-Time RT-PCR primer list

RLK	PRIMER NAME	PRIMER SEQUENCE
ORK1	SV5_08G05f	AAAGGCCTACAATGAAAGATGCCCTTC
ORK1	SV5_08G05r	CAACGGTTCTTCTCATCTTTATGGAGCA
ORK2	SV5_09B07f	GTGGAGACTGCCTTAGCGGAATCAC
ORK2	SV5_09B07r	CCACTGAGGTTTCCACTGTTGGTTG
ORK3	SV5_9F11-f	ATCCCAAACCTTGCCGGACT
ORK3	SV5_9F11-r	TTCCGGCGAGAACCACAACT
ORK4	SV5_10G11f	AAAACTCCCTTTGACTTGGGTTCA
ORK4	SV5_10G11r	TCATCCACCGTAAAACATTCCATC
ORK5	SV5_40C08-f	CGCGCATACTGGGAATTCAGC
ORK5	SV5_40C08-r	TCGCGCTTCTTTGAGGTGCAT
ORK6	SV5_46H11f	GGGGTTTCCAGATTCCTTCAATTT
ORK6	SV5_46H11r	CATCCCTCTCCACTGACCTTGAG
ORK7	SV5_52D12f	TGCTGAAACTAGGAGTGGAAGTGG
ORK7	SV5_52D12r	CAACAATGCAAAGCCAAAGGTAAA
ORK8	SV5_55A01f	TTGCTGCTATTCCTTTGGCTCAT
ORK8	SV5_55A01r	CCCATTTCATCACAACATTGGAA
ORK9	SV6_10H08f	TGGATAGAAGAAAGGGCAATGCAA
ORK9	SV6_10H08r	GATGGTCTGCTAGATGCGTGTGTTC
ORK10	SV6_17D05f	CCATCAATGCGTCCGAGGAT
ORK10	SV6_17D05r	TCATGGACTGGTTCAAGCCTCA
ORK11	SV6_21C05-f	GCTGCCTCAAAGGATGGCAAGT
ORK11	SV6_21C05-r	CCTTCTTTCCCATCAACCCAAACA
ORK12	SV6_21E09f	CCAGGTTTTGATCATCTTGATGGAG
ORK12	SV6_21E09r	CCAGACGAGCAACAACTTCTTTCA
ORK13	SV6_21H05f	GCAGATTCAGGCATGGAGTAACC
ORK13	SV6_21H05r	CTGGCAGGGCACTGCATCTA
ORK14	SV6_23B07f	TCTCTTTCCCCCTCCCTTGC
ORK14	SV6_23B07r	GGATCATCAGATTTTCCAAAAGTGC
ORK15	SV6_23E05f	TCATGTCACCATGCCCGAGT
ORK15	SV6_23E05r	GGCTCCAAAAGCAAGGAGCA
ORK16	SV6_24B06f	CTGAGAAGCACGGCAAGTGG
ORK16	SV6_24B06r	TTGCAGCTCCTCCCCAAAAG
ORK17	SV6_26A08-f	GCAACCAAGAGTCCCCTGGTCA
ORK17	SV6_26A08-r	CTGAAACCTCATATAGCCAAAAGTGTTGC
ORK18	SV6_29A11f	TGGCCTCCCACATTCCATTT
ORK18	SV6_29A11r	CCAATGTGTTGAGTGCAGAGAGC
ORK19	SV6_30C11f	GGGAACCCCATCAGAGACCA

ORK19	SV6_30C11r	GCCTGCCACTTCTTAACTGCTTG
ORK20	SV6_30G09f	GGACATGGCAAGGGCAGAAG
ORK20	SV6_30G09r	TGGTGGTGTGAGTTCTTTCTCGTT
ORK21	SV6_32E08f	AAAACCGCGCCCATTTTCAT
ORK21	SV6_32E08r	CCTCACCGGAAATCAAATCCA
ORK22	SV6_34A04f	AGCAGGCAGCAAAGGTGACG
ORK22	SV6_34A04r	TTGCCTGTCACCAGCTCCAA
ORK23	SV6_36B12f	CCAGAGGCCAACCATGAAGG
ORK23	SV6_36B12r	TGCATACAAGTAAAAGCAGCTAAGAGG
ORK24	SV6_37C04f	GGGAAAGATGGTGATGGAGTGG
ORK24	SV6_37C04r	CCATAGCGCGTGAGATCAGGA
ORK25	SV6_39C12f	CCAGCCTGAGCCTGAGTTCC
ORK25	SV6_39C12r	CAACGCAATGCAACCAACAA
ORK26	SV6_39E01f	TGGGTTGATGGGAAAGAAGGAA
ORK26	SV6_39E01r	TGCAATAACTTTGGGTAGTTGGATGA
ORK27	SV6_48C03f	AGCACCATCACCAATTTACAGAACC
ORK27	SV6_48C03r	CCCCCACACAAAGGAAAATGA
ORK28	SV6_49A12-f	CTCGAGGGGCTCTATCCTGTTAAA
ORK28	SV6_49A12-r	GAGTACTCAACGTCGCATGGAAA
ORK29	SV6_49F04f	CCCGCGAAACGACCAAGTAT
ORK29	SV6_49F04r	TGCATTGTTGCCAATGCCTTT
ORK30	SV6_49G04f	GCCAACAATGCGACAAAACG
ORK30	SV6_49G04r	AAACCTCCGCGATGGTCAAA
Ubiquitin	UBQ-f	GCTGGCAAGCAGTTGGAAGAT
Ubiquitin	UBQ-r	TGGATGTTGTAGTCCGCCAGA

Supplementary material

Table S2. Numbers and percentage of *Solanum chacoense* ORK per RLK classes compared with *Arabidopsis thaliana* RLKs.

	S. chacoense Ovule ESTs		*Arabidopsis* RLKs[1]	
RLK Class	Total number	%	Total number	%
LRR I	1	3,3	45	10,8
LRRIII	6	20,0	43	10,3
LRRV	4	13,3	11	2,6
LRRVI	4	13,3	11	2,6
LRRVII	1	3,3	10	2,4
LRRVIII-2	1	3,3	12	2,9
LRRIX	1	3,3	4	1,0
LRRX	1	3,3	15	3,6
LRRXI	3	10,0	28	6,7
LRRXII	1	3,3	9	2,2
LRRXIII	2	6,7	3	0,7
L-lectin	2	6,7	44	10,6
C-Lectin	1	3,3	1	0,2
Thaumatin	1	3,3	3	0,7
SD-1	1	3,3	32	7,7
Total	30	100	271	100

[1] Data taken from Shiu and Bleecker, 2001.

Table S3
Click here to download Supplementary material: Table S3 supplementary material.doc

Supplementary material

Table S3. Ovule Receptor Kinases EST accession numbers and their closest *Arabidopsis thaliana* match

Internal database identifier	db EST	ORK #	RLK family	*Arabidopsis* closest match
SV5_08G05	DN979658	ORK1	III	At5g67280
SV5_09B07	DN979684	ORK2	VI	At1g14390
SV5_9F11	DN979716	ORK3	III	At1g60630
SV5_10G11	DN979771	ORK4	XIII	At5g62230
SV5_40C08	DN977318	ORK5	V	At4g03390
SV5_46H11	DN977960	ORK6	I	At5g48740
SV5_52D12	DN978586	ORK7	III	At3g24660
SV5_55A01	DN978827	ORK8	XI	At5g61480
SV6_10H08	DN983490	ORK9	SD-1	At4g03230
SV6_17D05	DN983985	ORK10	C-Lectin	At1g52310
SV6_21C05	DN979962	ORK11	III	At3g17840
SV6_21E09	DN979990	ORK12	XII	At3g47110
SV6_21H05	DN980024	ORK13	V	At4g03390
SV6_23B07	DN980155	ORK14	III	At2g36570
SV6_23E05	DN980191	ORK15	XIII	At2g35620
SV6_24B06	DN980260	ORK16	X	At1g34420
SV6_26A08	DN980445	ORK17	VI	At3g03770
SV6_29A11	DN980737	ORK18	VI	At3g14350
SV6_30C11	DN980857	ORK19	IX	At2g01820
SV6_30G09	DN980903	ORK20	VIII-2	At3g09010
SV6_32E08	DR398320	ORK21	Thaumatin	At5g38280
SV6_34A04	DN981113	ORK22	XI	At4g28490
SV6_36B12	DN981325	ORK23	L-Lectin	At3g53380
SV6_37C04	DN981425	ORK24	VI	At4g18640
SV6_39C12	DN981625	ORK25	V	At1g53730
SV6_39E01	DN981638	ORK26	XI	At1g28440
SV6_48C03	DN982480	ORK27	III	At5g58300
SV6_49A12	DN982561	ORK28	V	At1g53730
SV6_49F04	DN982613	ORK29	VII	At2g24230
SV6_49G04	DN982625	ORK30	L-Lectin	At2g37710