Orphan genes are involved in drought adaptations and ecoclimatic-oriented selections in domesticated cowpea

Guojing Li[1, 2], Xinyi Wu[1], Yaowen Hu[1], Maria Muñoz-Amatriaín[3, #], Jie Luo[4], Wen Zhou[1], Baogen Wang[1], Ying Wang[1], Xiaohua Wu[1], Lijuan Huang[1,5], Zhongfu Lu[1], Pei Xu[1, 2,*]

[1]Institute of Vegetables, Zhejiang Academy of Agricultural Sciences, Hangzhou, China

[2]State Key Lab Breeding Base for Sustainable Control of Plant Pest and Disease, Zhejiang Academy of Agricultural Sciences, Hangzhou, China

[3]Department of Botany and Plant Sciences, University of California Riverside, Riverside, California, USA

[4]Central Laboratory of Zhejiang Academy of Agricultural Sciences, Zhejiang Academy of Agricultural Sciences, Hangzhou, China

[5]College of Horticulture, Northwest Agriculture and Forestry University, Yangling, China

*Correspondence: peixu@mail.zaas.ac.cn

[#]Current address: Department of Soil and Crop Sciences, Colorado State University, Fort Collins, Colorado, USA

Highlights: Orphan genes (species-specific genes) in cowpea are found to involve in drought adaptations and proposed to have contributed to balancing the adaptive and agronomic traits during domestication and improvement.

Abstract

Orphan genes (OGs) are genes that are restricted to a single species or a particular taxonomic group. To date, little is known about the functions of OGs in domesticated crops. Here, we report our findings on the relationships between OGs and environmental adaptation in cowpea (*Vigna unguiculata*). We identified 578 expressed OGs, of which 73.2% were predicted to be noncoding. Transcriptomic analyses revealed a high rate of OGs that were drought-inducible in roots when compared with conserved genes. Coexpression analysis further revealed the possible involvement of OGs in stress response pathways. Overexpression of *UP12_8740*, a drought-inducible OG, conferred enhanced tolerance to osmotic stresses and soil drought. By combining Capture-Seq and fluorescence-based Kompetitive allele-specific PCR (KASP), we efficiently genotyped SNPs on OGs across a 223-accession cowpea germplasm collection. Population genomic parameters, including PIC, *He*, π, and Tajima's D statistics, that were calculated based on these SNPs showed distinct signatures between the grain- and vegetable-type subpopulations of cowpea. This work reinforces the idea that OGs are a valuable resource for identifying new genes related to species-specific environmental adaptations and fosters new insights that artificial selection on OGs might have contributed to balancing the adaptive and agronomic traits in domesticated crops in various eco-climatic conditions.

Introduction

Orphan genes (OGs), or lineage-specific genes, are genes that are restricted to a single species or a particular taxonomic group. Depending on the mining method and phylogenetic resolution, OGs constitute from 1% to over one-third of the total gene number in a genome (Arendsee *et al*., 2014; Prabh and Rödelsperger, 2016). Nevertheless, from a molecular ecological perspective, the total number of OGs across all living lineages has been shown to exceed that of conserved genes (NOGs hereafter) (Tautz and Domazet-Lošo, 2011; Arendsee *et al*., 2014). Widespread computational work has documented the sequence and expression signatures of OGs in eukaryotes, such as shorter gene length and intron size, higher level of repeat sequences, and generally low and tissue-specific expression levels (Campbell *et al*., 2007; Guo *et al*., 2007; Yang *et al*., 2013; Xu *et al*., 2015). However, until recently, our understanding of OGs has been derived mainly from comparative genomic studies, and little is known about the population genomic signatures of OGs in a given species. For domesticated crops in particular, how natural and artificial selection have impacted OGs compared with NOGs remains unexplored.

There is accumulating evidence showing that OGs are functionally essential to species-/taxa-specific morphological speciation or environmental adaptation. For example, the human-specific gene *FLJ33706* is associated with brain function (Li *et al*., 2010). The *EED1* gene, found only in the pathogen *Candida albicans*, is crucial for hyphal extension and maintenance of filamentous growth (Martin *et al*., 2011). In plants, only a few OGs have been functionally characterized, and some were identified by positional cloning (Li *et al*., 2015; Fan *et al*, 2016). In general, despite the emergence of the theory that OGs arise because they confer selective advantages to an organism (Khalturin *et al*., 2009; Tautz and Domazet-Lošo, 2011), there has been a limited number of studies supporting this idea from a genomic perspective. In the seminal review of Arendsee *et al*. (2014), how the analysis of transcriptomics and genomics data at the subspecies level, combined with targeted experimental studies, can illuminate the emergence and functionality of new orphan genes was listed as an "outstanding question".

A shortage of fresh water is a severe agricultural problem facing plant scientists. Traditional model plants, including Arabidopsis and rice, are not inherently drought-

tolerant species; thus, they might be of limited use in efforts toward improving crop drought tolerance. With a relatively small diploid genome and an excellent adaptability to dry conditions, cowpea (*Vigna unguiculata*), a traditional legume crop feeding millions of people in the arid and semiarid environments of the world, is emerging as a new model for drought tolerance research (Muchero *et al*., 2013; Muñoz-Amatriaín *et al*., 2017). There are currently two major subspecies of cultivated cowpea: the short-podded grain cowpea that is prevalent in dry Africa, and the long- and succulent-podded vegetable cowpea that was domesticated in humid southeastern Asia (Timko *et al*., 2007; Xu *et al*., 2017). The different drought tolerances of the two cultivated subpopulations in contrasting eco-climates make this crop ideal for studying the relationships of OGs and environmental adaptions. In this report, we present a comprehensive study on orphan genes in this legume crop. In particular, we report, for the first time to our knowledge, a subspecies population genomic analysis on OGs.

Materials and methods

Plant materials

Plant materials used in this study include: a 223-accession cowpea germplasm collection (Table S1), which is a subset of the 299-line germplasm population used for domestication history study of cowpea (Xu *et al*., 2017); the cowpea cultivar "B128" and the adzuki bean cultivar "Suhong No. 1" used for comparative genomic hybridization; and the cowpea cultivar "II7E0826" used for hairy-root transformation.

Source sequences and BLAST method

To identify OGs in cowpea, we started with 3,480 EST-derived unigenes (assembly P12, http://harvest.ucr.edu/, 29,728 unigenes in total) with no annotated orthologs in UniProt (UniRef-90, Aug 16, 2008). These ESTs came from Sanger sequencing of 17 cDNA libraries representing diverse spatial, temporal, and stress-responsive RNA profiles (Muchero *et al*., 2009). While some OGs not represented in the 17 cDNA libraries may be missed in our analysis, the primary advantage of this strategy is that all input sequences were from real biological entities, thus avoiding the risk of pseudogene contamination or genome annotation artifacts. A BLASTn search was performed against eight available legume genomes as of December 2015,

including *Medicago*, *Lotus japonicus*, chickpea, soybean, pigeon pea, mung bean, common bean, adzuki bean, and the genomes of the model plants Arabidopsis and rice. The web addresses for downloading the aforementioned sequences or the sequence accession numbers are as follows: *Lotus japonicus*: ftp://ftp.plantgdb.org/download/Genomes/LjGDB/; Pigeon pea: http://www.icrisat.org/gt-bt/iipg/Genome_Manuscript.html (V1.0); Chickpea: http://www.icrisat.org/gt-bt/ICGGC/GenomeManuscript.htm; *Medicago*: https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Mtruncatula (*Mtruncatula*_198); Soybean: https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Gmax (*Gmax*_189); Common bean: https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Pvulgaris (Pvulgaris_218); Mung bean: Vradiate_ver6, GenBank assembly accession GCA_000741045.2; Adzuki bean: JZJH00000000 at GenBank; Arabidopsis: ATgenomeTAIR9.171; Rice: *Osativa*_193. Prior to BLAST, the 3480 unigene sequences were trimmed for 3'-poly(A) or 5'-poly(T) using Trimest (http://www.bioinformatics.nl/cgi-bin/emboss/trimest). Unigenes that were not matched to any of the compared genomes were retained for further comparative genome hybridization analysis. The criteria for declaring a sequence match were a matched sequence length $\geqslant$100 bp and an *e*-value $\leqslant 1E^{-10}$.

Microarray-based genome hybridization and statistic methods

An Agilent SurePrint G3 Custom Microarray (GenBank accession number GSE113644) representing 29,728 unigenes from the cowpea unigene assembly P12 and including both orphan and conserved genes was used for hybridization. The microarray-based genomic hybridization was used here as a high-throughput analog to traditional southern hybridization, the classical method for detecting specific DNA fragments in a genome. The principle of this method is similar to the detection of single genes that are species-specific in that the less similar the sequences between the probe and DNA are, the lower the signal intensity from the hybridization; however, high-throughput is utilized here by the use of a cDNA microarray. We performed this experiment to partly overcome the drawback (empirical criteria for determining a "no hit") of using computational analysis alone. For comparative genomic hybridization, genomic DNA from cowpea and adzuki bean were labeled with fluorochromes Cy3

and Cy5, respectively, purified using the genomic DNA purification module (Agilent Technologies), combined, and mixed with SeqCap EZ Developer Reagent (final concentration 100 ng/µl, Roche). The SeqCap EZ Developer Reagent is an alternative of human Cot-1 DNA specifically designed for plants to minimize nonspecific binding of probes and DNA. After being denatured at 95°C, the mixture was applied to the microarray, and the hybridization was performed at 57°C for 40 hours (Oligo aCGH hybridization kit, Agilent Technologies). After hybridization, the microarray was washed with Oligo aCGH wash buffer at 37°C, and slides were then dried. By using a low washing temperature, hybridization signal intensity data could be obtained from both cowpea and adzuki bean, allowing us to develop a statistical approach for validation of OGs (see the next paragraph). The arrays were then scanned using an Agilent Scanner (Agilent Technologies), and $\log_2$-intensities were extracted from raw microarray images files using the Agilent Feature Extraction software version 9.0.

We used a simple but robust statistical method to determine genes that were specific to cowpea. First, we extracted hybridization signal intensities from the 28,333 genes that were found to be nonspecific to cowpeas (meaning, genes with orthologs in related species) in the aforementioned BLAST analysis procedure. The vast majority of the signal intensities were lower in adzuki beans than in cowpeas due to sequence mismatch. By taking a 0.95 percentile of the cowpea-to-adzuki bean hybridization signal intensity ratio, a ratio = 2.1 was adopted as the lower limit to determine cowpea-specific genes. By taking a 0.05 percentile of the cowpea gene hybridization signal intensity, the absolute signal intensity value of 345 was taken as the lower limit of positive signal to exclude hybridizations backgrounds. In doing so, the genomic hybridization also efficiently ruled out possible microbial sequence contaminants in the starting cowpea unigenes set.

Sequence characterization of the OGs

Basic sequence statics such as length and GC-content were calculated using custom Perl scripts. The coding potential of each sequence was analyzed using Coding-Non-Coding Index (CNCI) trained with Arabidopsis for plant transcripts (Sun *et al*. 2013). CNCI reports the likelihood of protein-coding for a given short sequence based on sequence intrinsic composition independent of sequence completeness, making it a good fit for this study. Transposable elements in the orphan gene

sequences were predicted using RepeatMasker (version open-4.0.6, http://www.repeatmasker.org/) with the "DNA source" model set as "Arabidopsis".

Differentially expressed gene (DEG) analysis and drought-associated weighted gene coexpression network module inference

Differential expressions of the OGs and NOGs were analyzed using our previous Nimblegene microarray gene expression data (NCBI accession number GSE63636) on drought-stressed and non-stressed cowpea plants (cultivar "B128") in root and leaf tissues (Xu *et al*., 2015). Briefly, the expression data from the probes were normalized using quantile normalization, and expression data for genes were generated using the Robust Multichip Average (RMA) algorithm (Irizarry *et al*., 2003). Comparisons were made between the stressed and non-stressed conditions. Multiple test corrections were performed based on the FDR (Benjamini and Hochberg, 1995). The cutoff for declaring a gene differentially expressed was: fold change $\geqslant$ 2, FDR $\leqslant$ 0.05. Gene ontology (GO) functional enrichment analysis was performed using Goatools (https://github.com/tanghaibao/Goatools).

The coexpression gene network modules were inferred from DEGs using a weighted gene coexpression network analysis (WGCNA) implemented in R. The automatic one-step network construction and module detection method with default settings were used, which include an unsigned type of topological overlap matrix (TOM), a power β of 10, a minimal module size of 20, and a branch merge cut height of 0.25. The module eigengene value was calculated and used to test the association of the modules with drought treatment. The gene coexpression network topology was displayed using the system preferred algorithm in Cytoscape (Shannon *et al*. 2003). The transcriptomic data from roots and leaves of "B128" under well-watered and drought-stressed conditions were used in coexpression analysis (GSE63636).

SureSelect capture assay design, in-solution hybridization and sequencing

SureSelect baits (Agilent design ID 0666761) were designed based on the cDNA sequences of the 578 identified OGs. A total of 15,723 120-mer RNA baits covering approximately every 25 bp in OG sequences were designed that tilled across 342.432 kb. The bait sequences are listed in Notes S1.

Sequence capture was performed using a SureSelect solution phase hybridization assay (SureSelect XT Reagent Kit, Agilent). DNA from eight distantly

related cowpea accessions, including G15, G47, G93, G118, G128, H237, G314, and G334, which dispersed in the principal component analysis (PCA, Fig. S1) plot indicating broad genetic diversity, were used for paired-end library preparation according to the standard protocols. The capture assay was hybridized with barcoded DNA libraries according to Agilent's in-solution hybridization instructions. The SeqCap EZ Developer Reagent was added to the hybridization system to reduce nonspecific capturing. The captured-libraries were sequenced using a HiSeq 2500 with a 100-nt paired-end run.

SNP calling

Raw sequencing data were filtered using SeqPrep software (https://github.com/jstjohn/SeqPrep). Key parameters used were: "-q 20 -L 25 -B AGATCGGAAGAGCGTCGTGT -A AGATCGGAAGAGCACACGTC. Sickle software (https://github.com/najoshi/sickle) was applied with default parameters for 3'-end sequence trimming to acquire clean sequence data. The clean reads were aligned to reference sequences (OG sequences in the unigene assembly P12) using Bowtie-0.12.5 (Langmead *et al*., 2009). The parameters -m1 and -n2 were applied so that reads with more than one alignment would be suppressed, with two mismatches allowed between the reference sequence and the first 28 nucleotides of a read. Alignment files generated by Bowtie were processed using SAMtools (v 0.1.6) (Li, 2011) to produce output files containing information about the depth of coverage and variant counts. Varscan2 (http://massgenomics.org/varscan) was used to call SNPs with the following parameters: min-avg-qual = 0, min-coverage = 1, min-reads = 20, min-var-freq = 0.01; min-freq-for-hom = 0.75. Only variants that had a minimum read depth of 4 at the variant position were considered useful for genotype assignment. For assigning genotypes, a referential or alternative genotype was called if 100% or 0% of the variants were the same as the reference, respectively; otherwise, a missing call was assigned.

Validation of SNPs by Sanger sequencing

PCR primers were designed for 20 randomly selected OGs that harbor 79 putative SNPs called through targeted sequencing (Table S2). After amplification, unambiguous PCR products were purified and sequenced by the Sanger method. Sequence alignment was performed using MacVector (www.macvector.com).

Conversion of SNPs into KASP markers and genotyping

We targeted 50- to 100-bp sequences flanking each side of the intended SNPs for designing primers suitable for KASP assay. Wherever more than one SNP was present in the sequence of interest, the additional SNPs were marked using IUPAC codes. By the above standard, 36 of the 390 SNP-containing OGs were not eligible for KASP assay design. For the remaining OGs, one SNP from each OG was selected for KASP assay design. Two allele-specific forward primers differing at their 3' ends, where the target SNP is located, and one common reverse primer were designed for each SNP locus. Genotyping of a germplasm collection consisting of 223 cowpea lines with KASP assays was performed according to LGC Genomics' instructions. The genotyping data obtained based on fluorescence detected from the KASP assay was viewed graphically through SNPviewer2 (version 3.2.0.9, LGC Genomics, 2013). Due to the unusual sequence characteristics of some OGs, such as the high level of repeats and low complexity, 25 KASP assays failed. Among the successful KASP assays, 217 revealed polymorphisms in the germplasm collection, and the rest detected polymorphisms only between the reference OG sequences (from far-related African cowpea) and the germplasm lines (but not within the latter).

Population genomic parameter calculation and display

Population genomic parameters, including polymorphism information content (PIC), expected heterozygosity (*He*), nucleotide diversity (π) and Tajima's D statistic, were calculated based on SNPs in the OGs and the NOGs. For the OGs, the SNP data from the KASP assay were used, and for the NOGs, published SNP data (48,216 SNPs in inferred cowpea gene homologs to *Phaseolus vulgaris* genes) were extracted for the 223 accessions (Muñoz-Amatriaín *et al*. 2017; Xu *et al*., 2017). The SNP densities in this analysis were 1 SNP per OG and 2 SNPs per NOG. All accessions showing a rate of heterozygosity calls $\leqslant$ 0.25 and a rate of missing calls $\leqslant$ 0.25 were retained for further analysis. PIC was calculated using the method of Botstein *et al*. (1980); *He* (for two alleles) and π were evaluated as in Muñoz-Amatriaín *et al*. (2017) and Xu *et al*. (2017), respectively. Tajima's D was calculated using BioPerl based on the formula of Tajima (1983). All parameters were calculated for each of the SNP sites. For display purposes, the results of the NOGs were averaged and plotted via a kernel-smoothing moving method on 500-kb sliding windows to generate

genome-wide distributions, and the results of OGs are shown as individual dots. Significance cutoffs for Tajima's D statistics were determined by taking the 5% percentile across the genome.

Vector construction and hairy-root transformation

For overexpression of the drought-induced cowpea orphan gene *UP12_8740*, the intronless gene sequence was amplified from genomic DNA using *pfu* DNA polymerase (Dingguo Biotech, Beijing) with the primers 5'-CTTGGTGGGAAAGGTTTTGA-3' and 5'-GGACCAAGGCGATAAGATGA-3', and the 1003-bp amplicons were mobilized into the pCAMBIA1301 vector, replacing the original GUS fragment in the vector to generate cauliflower mosaic virus 35S promoter-driven OE constructs. Note that because noncoding genes typically lack sequence characteristics at their 5'-end to indicate the full functional length, we amplified the *UP12_8740* region from the DNA template (*UP12_8740* is an intron-free gene) several hundreds of base pairs upstream the 5'-end of the unigene sequence, resulting in a length of 1003 bp for the amplicon. The recombinant plasmids were transformed into *Agrobacterium rhizogenes* R1000 (courtesy of Michael P Timko, University of Virginia) competent cells by electroporation. Transgenic composite plants were generated using the hairy root transformation method with the cowpea accession "II7E0826" as the recipient (Mello *et al*., 2012). II7E0826 is a landrace showing high efficiency of hairy-root transformation (Zhou *et al.* 2018). After transformation, the seedlings were transferred to 1/2 MS solutions for recovery for one day, before being subjected to osmotic stress or transferred to pots with soil (see next paragraph). Overexpression of the transgene in roots was verified by RT-PCR and qRT-PCR. For qRT-PCR reactions, gene-specific primers were designed against the *UP12_8740* sequence near the 3'-end and against an *EF-1α* reference gene also near the 3'-end. The primers used are as follows: UP12_8740-qRT-F 5'-GAATGGATGATCAAGCAAATAGG-3' and UP12_8740-qRT-R 5'-ACCAAGGCGATAAGATGAATAAAA-3'; EF-1α-F: 5'-TGACAGGCGTTCTGGTAAGG-3' and EF-1α-R: 5'--GATGATGGTGGAAACTTCAAAC-3'.

Drought treatment and transgenic phenotyping

Two types of drought stress were included: one imposed by 6% (roughly equals water potential of -0.9 bars) or 10% (roughly equals water potential of -1.3 bars) PEG solutions for osmotic stresses (Michel, 1983), and the other by withholding water in pots for a progressive soil water deficit. Phenotyping was performed with a series of independent *UP12_8740*-OE lines and pCAMBIA1301 empty vector-transformed CK plants. All the plants were placed in a growth chamber with the ambient temperature cycle of 28°C (day)/22°C (night) and a 16 h/8 h photoperiod. The morphology of the above-ground parts of the plants was visually inspected. Measurements of relative electrical conductance followed the protocol of Li *et al*. (2014). Seedling biomass gain was calculated by subtraction between the seedling weight measured right before and 7 days after PEG treatment.

Results and Discussion

Identification and characterization of expressed OGs in cowpea

Through a BLASTn search against the eight legume genomes and the genomes of Arabidopsis and rice, we found that the number of cowpea unigenes unmappable to the subject genomes decreased with increase in phylogenetic relatedness (Fig. 1A). Combined, the searches led to the detection of 598 unigenes with no match in any of the compared genomes. The number of cowpea-specific genes was further reduced to 578 genes by microarray-based genomic hybridization with DNA of cowpea and adzuki bean (*Vigna angularis*), a close relative of cowpea, where only 20 fewer unigenes were recognized as cowpea-specific (Fig. 1B). Combining the *in silico* and wet-bench results, those 578 unigenes were determined to be the cowpea OGs (Table S3), accounting for 2% of the total cowpea unigenes in the database. By BLASTn searching against the cowpea genome assembly of Muñoz-Amatriaín *et al*. (2017), the chromosome locations of the OGs were determined. OGs were distributed in all eleven chromosomes, and clustered distributions were clearly seen in many chromosomes, particularly Vu03, Vu04 and Vu08 (Fig. 2A, Table S3). The clustered distribution of OGs in certain chromosome regions is not unique to cowpea (Gross *et al*. 2007; Xu *et al*. 2015). In *Pseudomonas fluorescens*， a cluster of orphan genes was found to be involved in the biosynthesis of a lipopeptide natural product (Gross *et al*.

2007). It is therefore possible that OGs within clusters may be regulated in a coordinated way.

Cowpea OGs showed unique sequence characteristics similar to those from other species (Tautz and Domazet-Lošo, 2011; Arendsee *et al*., 2014): generally shorter sequence length, lower GC content, and enriched repetitive sequences compared with NOGs (Fig. 2B, C). Using Coding-Non-Coding Index (CNCI), a tool tailored for predicting the coding potential of a sequence based on the sequence intrinsic composition, 423 of the OGs (73.2%) were predicted to be noncoding (Table S3). Notably, we only found 7 transposable element (TE)-like sequences in the OGs, which contrasts with the high frequency (53%) found in primate OGs (Toll-Riera *et al*. 2009). *De novo* evolution out of noncoding genomic regions and transposon-associated activities are considered to be forces contributing to the formation of new orphan genes (Tautz and Domazet-Lošo, 2011). Our results appear to support that the former may be predominant mechanism in cowpea and raise an interesting question regarding the origins of OGs in various lineages.

Cowpea OGs exhibited unique regulatory modes in response to drought stress

To investigate the expression patterns of cowpea OGs in different tissues, we reanalyzed our previous transcriptomic data on four different types of tissues under normal growth conditions (Xu *et al*., 2015, 2017). We found that, in general, OGs had a lower expression than NOGs in all four tissues (Fig. 3A), similar to the expression patterns in many other species (Yang *et al*., 2013; Xu *et al*., 2015). More interestingly, the numbers of actively expressed OGs (GeneSpring normalized expression value > 0) were similar among the four tissues, which were 148 in leaf and 145 in each of the remaining three tissues. Compared to flowers and roots, developing seeds and leaves had more tissue-specific expressed OGs (Fig. 3B). To further interrogate the relationship between OGs and drought adaptation, we also reanalyzed transcriptomic data available from drought-stressed and non-stressed cowpea plants (Xu *et al*., 2015). We found a high proportion of OGs (120/578, 20.7%) that were drought-responsive in roots, with the vast majority (106/120, 88.3%) being upregulated, in comparison to only 0.8% (228/29150) in the NOGs (Fig. 3C, Table S4). Only 3.8% (22/578) of the OGs were differentially expressed under drought stress in leaves, as opposed to 17.3% of the NOGs. It is known that plant drought response first involves root sensing of soil water deficit and then root-sourced signals being transported to leaves (Schachtman

and Goodger, 2008). Thus, the observed root-biased transcriptional regulation of the OGs suggests their possible role in drought sensing and signal triggering, which agrees with the noncoding (and thus regulatory) nature of the majority of the OGs. A weighted gene coexpression network analysis (WGCNA) for all DEGs (including OGs and NOGs) in relation to drought further identified three significant coexpression modules ($P < 0.05$, Fig. S1), where 85 OGs were involved and enriched in gene ontology (GO) terms, including "response to stimulus" and "response to oxidoreductase activity" ($P < 0.05$) (Table S5). We found coexpression of many OGs with genes known to be involved in stress response/oxidoreductase activities, such as *UP12_18699* (VuOG323) and *UP12_16084* (encoding properoxidase precursor) (Fig. 3D, E, Table S6). These correlations provide a means for inferring the OG functions and suggest the incorporation of certain OGs into conserved stress networks as a mechanism underlying the drought adaptation of cowpea.

Overexpression of UP12_8740 in roots conferred enhanced drought tolerance of composite transgenic plants

As a more direct way to demonstrate the functionality of specific OGs in drought tolerance, we selected *UP12_8740*, a putative noncoding OG showing the largest expression difference (7.4-fold) in roots upon drought stress, to generate a series of overexpression (OE) plants using the hairy-root transformation approach. Hairy-root transformation is a method commonly used to generate composite transgenic plants for recalcitrant species (Mellor *et al*., 2012). By comparing the *UP12_8740*-OE and empty vector-transformed CK lines, we found more vigorous seedling phenotypes in five *UP12_8740*-OE lines than in CK lines after 24 hours of treatment with 6% or 10% PEG solutions to induce osmotic stress (Fig. 4A, Fig. S2). qRT-PCR showed that the *UP12_8740* gene expression levels were higher in *UP12_8740*-OE lines than in CK lines; this indicates that enhanced osmotic stress tolerance was related to *UP12_8740* overexpression. In addition, the individual *UP12_8740*-OE lines showed greater (albeit slightly in some cases) biomass gain than the CK lines under osmotic stress (Fig. 4B). In the assay of leaf relative electrical conductance (REC), a physiological parameter indicative of cellular damage under stress conditions, we found that the REC values were significantly smaller in *UP12_8740*-OE lines than in control plants (Fig. 4B), suggesting more effective membrane structure maintenance in response to osmotic stress in the *UP12_8740*-OE

lines. By comparing the performance of transgenic seedlings and CK lines under progressive soil drought conditions, we could observe attenuated leaf damage phenotypes and more vigorous growth in the *UP12_8740*-OE lines than the wild type and empty vector-transformed CK lines, though the leaf and root morphologies of these lines were similar before transplanting into soil (Fig. 4C, D). Collectively, these data support that *UP12_8740* is a novel gene conferring drought tolerance in cowpea and reinforce that OGs are a valuable resource for identifying new genes involved in specific environmental adaptations. Since drought responses were evaluated only at the seedling stage due to limitations of the hairy-root transformation approach, our results may not necessarily indicate adult-plant drought resistance. Evaluating yields of the stable homozygous OG transgenic lines under normal and drought stressed conditions will be necessary to assess the agronomic potential of harnessing the OGs for improvement of crop yields.

Single-nucleotide polymorphisms (SNPs) on OGs in the cowpea germplasm collection

To provide insights into OGs from a population genomic perspective, we developed a method combining Capture-Seq and fluorescence-based Kompetitive Allele-Specific PCR (KASP) for an efficient and cost-effective genotyping of SNPs. A SureSelect Target Enrichment Capture Assay (Agilent design ID 0666761) harboring 15,723 biotinylated cRNA oligonucleotides covering 572 of the 578 OGs was hybridized to DNA libraries constructed from eight distantly related genotypes (the SNP discovery panel, Fig. S3) followed by sequencing, which produced a rate of on-target reads of ~ 45% for each library and an average on-target sequencing coverage of 7.5 ×. A total of 1560 SNPs were detected from transcribed regions in 390 of the 572 OGs. The average exonic SNP density of OGs was therefore estimated to be 4 SNPs per gene. We current have no information about SNP density of NOGs in the same diversity panel, but according to resequencing of a 37-accession cowpea diversity panel, there are on average 6 exonic SNPs per gene (Munoz-Amatriain, personal communication). Through PCR amplification and Sanger sequencing of 20 randomly selected OGs (Table S2), 77 of the 78 intended SNPs were verified, demonstrating the high accuracy of SNP discovery provided by this strategy. We then selected one SNP from each OG for conversion to KASP markers, of which 379 ultimately succeeded. In genotyping the 223-line cowpea germplasm collection (Table

S1), we obtained unambiguous genotypic data for 354 of the KASP markers, which were used to assess population genomics parameters of the OGs.

Subspecies-level population genomic signatures of OGs

We found that the overall polymorphism information content (PIC), expected heterozygosity (*He*) and nucleotide diversity (π) were all higher in OGs than in NOGs (Fig. 5A), suggesting more relaxed selective constraints in the former. Tajima's D value, a statistic that detects deviations from the neutral mutation model, was found to be over three times greater in OGs than in NOGs, implying higher levels of non-neutral selection events for the OGs in cowpea. When subdividing the population into the grain-type (subpopulation 1) and vegetable-type (subpopulation 2) subgroups according to previous population structure analysis (Xu *et al*., 2017), we found that genome-wide Tajima's D values calculated based on OGs were positive in subpopulation 1 and negative in subpopulation 2 (Fig. 5A). The overall Tajima's D patterns in the two subpopulations calculated based on NOGs, though similar to that of OGs, were much less contrasting. In subpopulation 1, more OGs exhibited significant positive Tajima's D values ($\geqslant$1.72, Fig. 5B), suggesting more widespread balancing selection, whereas in subpopulation 2, more OGs were suggested to be under neutral (Tajima's D value ~ 0) or purifying (Tajima's D value $\leqslant$ -1.07) selection.

As mentioned before, the grain- and vegetable-type cowpeas were well adapted to the dry African climate and the humid Southeast Asian climate, respectively. The grain cowpea has obviously confronted more challenging and versatile soil water constraints. The more dominant balancing selection that was detected, which is known to be beneficial for plant adaptation by maintaining genetic diversity (Wu *et al*., 2017), is consistent with the importance of OGs in drought adaptation in grain cowpea. In vegetable cowpea, which has undergone selections primarily for pod quality traits such as length and tenderness (Xu *et al*., 2017), the higher level of purifying selection appears to be a reflection of artificial selection toward balancing the adaptive and agronomic traits in the less drought-prone regions. We postulate that some mutations in OGs that are beneficial for drought tolerance may be disadvantageous for pod quality or other favorable agronomic traits in vegetable cowpea and have hence been selected against during domestication. We noted that

UP12_8740 had a considerable Tajima's D value (1.503) in subpopulation 1, while it was monomorphic in subpopulation 2, indicating that it may be one of the aforementioned selection target genes.

The discovery of OGs has challenged current molecular biological paradigms and theories of evolution to some degree. Along with the recent advances in the study of the origin and turnover of OGs (Tautz and Domazet-Lošo, 2011; Palmieri *et al*., 2014), our functional and population genomic work in domesticated cowpea would be helpful for better understanding the ecological and agricultural importance of OGs. We support that OGs confer a selective advantage to an organism and further propose that, in domesticated crops, selections on OGs may have played an important role in balancing the adaptive and agronomic traits.

Supplementary material

Supplementary data are available at JXB online.

Fig. S1 Module-trait relationship in WGCNA analysis. Each row corresponds to a colored module eigengene (ME) detected by WGCNA. The correlation between each ME and the analyzed drought response trait and the corresponding *P*-value (in parenthesis) are displayed in each cell. **P* < 0.05; ***P* < 0.01.

Fig. S2 Osmotic stress tolerance phenotypes of independent *UP12_8740*-OE lines, empty vector-transformed CK lines and WT plants under 6% or 10% PEG treatment.

Fig. S3 Principal component analysis (PCA) of the cowpea germplasm used in this study, showing the distribution of the eight accessions chosen for the SNP discovery panel.

Table S1 Cowpea germplasm lines used in this study.

Table S2 Primer sequences for the amplification of DNA fragments flanking 20 randomly selected SNPs.

Table S3 The 578 cowpea OGs, their sequence characteristics and their genome locations.

Table S4 List of differentially expressed OGs in roots and leaves of the cowpea cultivar “B128” in response to drought stresses.

Table S5 GO enrichments in the three significant gene co-expression modules.

Table S6 Co-expression relationships related to cowpea OGs.

Notes S1 Bait library sequence used in capture-Seq.

Acknowledgments

This study is supported by the National Key Research & Development Program of China (2016YFD0100204-32); the National Natural Science Foundation of China (31272183, 31861143044); and the Major Science and Technology Project of Plant Breeding in Zhejiang Province (2016C02051-7-3); Pei Xu is also supported by the National Program for the Support of Top-notch Young Professionals. M.M.A. is supported by the NSF BREAD project "Advancing the Cowpea Genome for Food Security" (Award #1543963). We would like to thank Ye Tao and Liang Zeng (Biozeron Biotech, Shanghai) for their technical assistance in the bioinformatics analysis.

Authors contributions

P.X. and G.L. devised the research. G.L., P.X., Y.H. J.L. and M.M.A. performed the bioinformatics analysis. X.Y.W., Y.H., W.Z., Y.W., L.H., B.W., X.H.W. and Z.L. performed the experiments. G. L, P.X. and M.M.A. drafted the manuscript.

References

Arendsee ZW, Li L, Wurtele, ES. 2014. Coming of age: orphan genes in plants. Trends in Plant Science **19**:698-708.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate, a practical and powerful approach to multiple testing. Journal of The Royal Statistical Society Series B-Statistical Methodology **57**: 289-300.

Botstein D, White RL, Skolnick M, Davis RW. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. American Journal of Human Genetics **32**: 314-331.

Campbell MA, Zhu W, Jiang N, Lin H, Ouyang S, Childs KL, Haas BJ,

Hamilton JP, Buell CR. 2007. Identification and characterization of lineage-specific genes within the *Poaceae*. Plant Physiology. 145:1311-22.

Fan Y, Yang J, Mathioni SM, Yu J, Shen J, Yang X, Wang L, Zhang Q, Cai Z, Xu C, Li X, Xiao J, Meyers BC, Zhang Q. 2016. PMS1T, producing phased small-interfering RNAs, regulates photoperiod-sensitive male sterilityin rice. Proceedings of the National Academy of Sciences of the United States of America. **113**:15144-15149.

Gross H, Stockwell VO, Henkels MD, Nowak-Thompson B, Loper JE, Gerwick WH. 2007. The genomisotopic approach: a systematic method to isolate products of orphan biosynthetic gene clusters. Chemistry & Biology. **14**:53-63.

Guo WJ, Li P, Ling J, Ye SP. 2007. Significant comparative characteristics between orphan and nonorphan genes in the rice (*Oryza sativa* L.) genome. Computational and Functional Genomics. **1**:21676.

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics **4**: 249-264.

Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TC. 2009. More than just orphans: are taxonomically-restricted genes important in evolution? Trends in Genetics. **25:**404-13.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology **10**: R25.

Li AX, Han YY, Wang X, Chen YH, Zhao MR, Zhou SM, et al. 2014. Root-specific expression of wheat expansin gene *TaEXPB23* enhances root growth and water stress tolerance in tobacco. Environmental and Experimental Botany. **110**: 73-84.

Li CY, Zhang Y, Wang Z, Zhang Y, Cao C, Zhang PW, Lu SJ, Li XM, Yu Q, Zheng X, Du Q *et al.* 2010. A human-specific *de novo* protein-coding gene associated with human brain functions. PLoS Computational Biology **6**: e1000734.

Li L, Wurtele ES. 2015 The QQS orphan gene of Arabidopsis modulates carbon and nitrogen allocation in soybean. Plant Biotechnology Journal **13**:177-187.

Li, H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics **27**:2987-2893.

Martin R, Moran GP, Jacobsen ID, Heyken A, Domey J, Sullivan DJ, Kurzai O, Hube B. 2011. The Candida albicans-specific gene EED1 encodes a key regulator of hyphal extension. PLoS One**. 6:**e18394.

Mellor KE, Hoffman AM, Timko MP. 2012. Use of *ex vitro* composite plants to study the interaction of cowpea (*Vigna unguiculata* L.) with the root parasitic angiosperm *Striga gesnerio*ides. Plant Methods **8**:22.

Michel BE. 1983. Evaluation of the water potentials of solutions of polyethylene glycol 8000 both in the absence and presence of other solutes. Plant Physiology **72**:66-70.

Muchero W, Diop NN, Bhat PR, Fenton RD, Wanamaker S, Pottorff M, Hearne S, Cisse N, Fatokun C, Ehlers JD, Roberts PA *et al.* 2009. A consensus genetic map of cowpea (*Vigna unguiculata* L. Walp) and synteny based on EST-derived SNPs. Proceedings of the National Academy of Sciences of the United States of America **106**:18159-18164.

Muchero W, Roberts PA, Diop NN, Drabo I, Cisse N, Close TJ, Muranaka S, Boukar O, Ehlers JD. 2013. Genetic architecture of delayed senescence, biomass, and grain yield under drought stress in cowpea. PLoS One **8**: e70041.

Muñoz-Amatriaín M, Mirebrahim H, Xu P, Wanamaker SI, Luo M, Alhakami H, Alpert M, Atokple I, Batieno BJ, Boukar O, Bozdag S *et al.* 2017. Genome resources for climate-resilient cowpea, an essential crop for food security. Plant Journal **89**:1042-1054.

Palmieri N, Kosiol C, Schlötterer C. 2014. The life cycle of Drosophila orphan genes. eLife. **3**:e01311.

Prabh N, Rödelsperger C. 2016. Are orphan genes protein-coding, prediction artifacts, or non-coding RNAs? BMC Bioinformatics **17**:226.

Schachtman DP, Goodger JQ. 2008. Chemical root to shoot signaling under drought. Trends in Plant Science **13**:281-287.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research **13**: 2498-2504.

Sun L, Luo H, Bu D, Zhao G, Yu K, Zhang C, Liu Y, Chen R, Zhao Y. 2013. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. Nucleic Acids Research **41**: e166.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics **105**: 437-460.

Tautz D, Domazet-Lošo T. 2011. The evolutionary origin of orphan genes. Nature Reviews Genetics **12**:692-702.

Timko MP, Ehlers JD, Roberts PA. 2007. Cowpea. In: Kole C (ed) Pulses, sugar and tuber crops. Theoretical and applied genetics, genome mapping and molecular breeding in plants. vol 3. Springer, Berlin, pp 49-67.

Wu Q, Han TS, Chen X, Chen JF, Zou YP, Li ZW, Xu YC, Guo YL. 2017. Long-term balancing selection contributes to adaptation in Arabidopsis and its relatives. Genome Biology **18**: 217.

Xu P, Moshelion M, Wu X, Halperin O, Wang B, Luo J, Wallach R, Wu X, Lu Z, Li G. 2015. Natural variation and gene regulatory basis for the responses of asparagus beans to soil drought. Frontiers in Plant Science **6**: 891.

Xu P, Wu X, Muñoz-Amatriaín M, Wang B, Wu X, Hu Y, Huynh BL, Close TJ, Roberts PA, Zhou W, Lu Z, Li G. 2017. Genomic regions, cellular components and gene regulatory basis underlying pod length variations in cowpea (*V. unguiculata* L. Walp). Plant Biotechnology Journal **15**:547-557.

Xu Y, Wu G, Hao B, Chen L, Deng X, Xu Q. 2015. Identification, characterization and expression analysis of lineage-specific genes within sweet orange (*Citrus sinensis*). BMC Genomics. **16:**995.

Yang L, Zou M, Fu B, He S. 2013. Genome-wide identification, characterization, and expression analysis of lineage-specific genes within zebrafish. BMC Genomics. **14:**65.

Zhou W. 2018. Study on factors influencing genetic transformation and establishment of the hairy-root induction system in cowpea. Master thesis. Zhejiang Normal

University. Hangzhou, China. (in Chinese with an English abstract)

Figure legends

Fig. 1 The combined computational (A) and high-throughput genomic hybridization method (B) for identifying OGs in cowpea. The numbers of unigenes that had no match in the compared genomes in each step are listed in parenthesis.

Fig. 2 Overall characteristics of OGs in cowpea. A, Genome distributions of the OGs. The location of each OG is indicated by a red dot in the inner circle. The numbers in the outer circle indicate physical distances (Mb) on the chromosomes. **B**, Comparison of the average sequence lengths between OGs and NOGs. **C**, Comparison of the GC contents between OGs and NOGs.

Fig. 3 Expression and co-expression network analyses. A, Expression patterns of OGs in leaf (LF), root (RT) (left panel) and flower (FL) and developing seeds (SD) (right panel). Note that the charts were made according to gene expression data extracted from the NimbleGen (for the left panel) and Agilent (for the right panel) microarrays, respectively. **B**, Venn diagram showing the number and relationships of expressed OGs in leaf, roots, flowers and developing seeds (10 days post anthesis). **C,** Number of differentially expressed OGs upon drought stress in leaves and roots of cowpea. **D** and **E,** Network topology between OGs and genes known to be involved in stress responses/oxidoreductase activity in the "darkred" and "purple" coexpression modules, respectively. For clarity, only part of the coexpression relationships with a correlation coefficient $\geqslant$ 0.9 are shown. For the gene names, U is an abbreviation of UP12_.

Fig. 4 Transgenic assay of *UP12_8740* using hairy-root transformation. A, Osmotic stress tolerance phenotypes of five independent *UP12_8740*-OE lines and the empty vector-transformed CK lines upon osmotic stresses. Panels 1-5: phenotypes of various *UP12_8740-OE* lines and empty vector-transformed CK lines in response to PEG-induced osmotic stresses. Note that to show the aerial parts and roots simultaneously, the photographs were shot from seedlings that had been taken out of the original glass bottles filled with PEG solutions. Please refer to Fig. S2 for the picture showing the original phenotypes of each seedling in glass bottles. **B,**

Comparison of relative electrical conductivity (REC) and seedling biomass gain between *UP12_8740*-OE lines and the CK lines. Error bar = SD. **C,** Phenotypes of *UP12_8740*-OE and CK lines without stress treatment. **D,** Phenotypes of *UP12_8740*-OE and CK lines in response to progressive soil drought. The pictures were taken 4 days (top panel), 7 days (middle panel) or 12 days (bottom panel) after water withholding. Bars = 5 cm. Note that the seedlings in the top and middle panels (No. 9 & 10) are different from those in the bottom panel (No. 11 & 12), which were generated from independent transformation experiments and subject to independent soil drought treatments. For seedlings in the bottom panel, RNA were not extracted for gene expression analysis because they were too damaged. In all panels for qRT-PCR analyses of *UP12_8740* expression levels in the transgenic and CK plants, relative expression data are from three technical replicates. Error bar = SD.

Fig. 5 Population genomic analyses of OGs at the subspecies level. A, Overall population genomics parameters estimated for OGs and NOGs in the whole germplasm collection, subpopulation 1 and subpopulation 2. **B,** Genome-wide distribution of Tajima's D values in each SNP on the OGs in the whole germplasm collection, subpopulation 1 and subpopulation 2. Dashed orange lines show the significance cutoffs for balancing and purifying selections in subpopulation 1 and subpopulation 2, respectively.

Figure 1

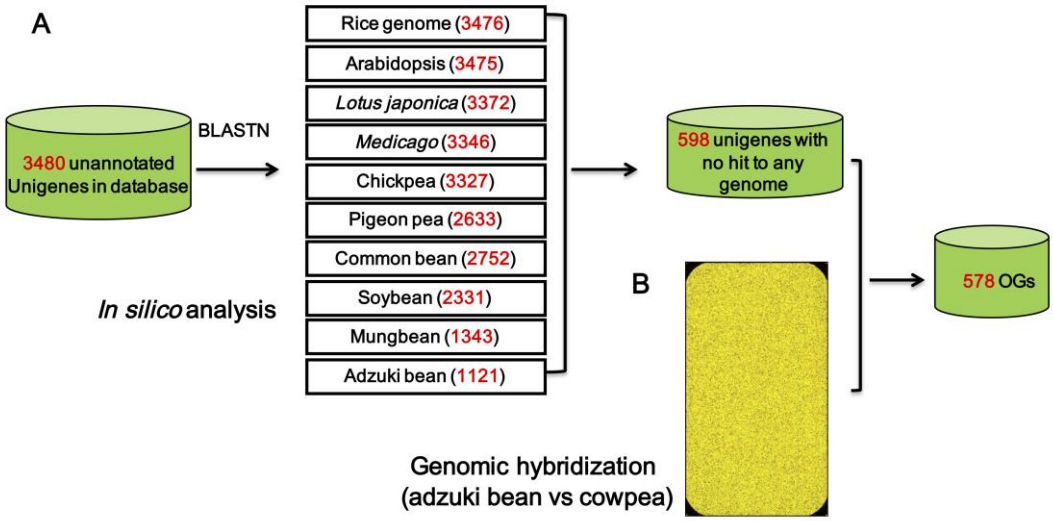

Figure 2

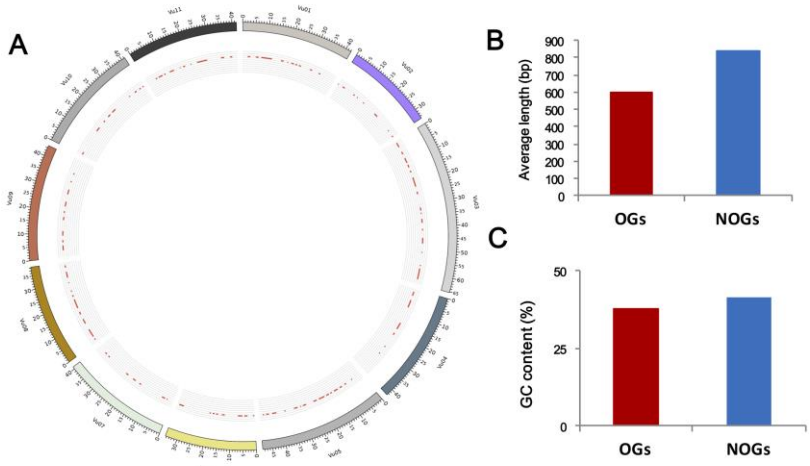

Figure 3

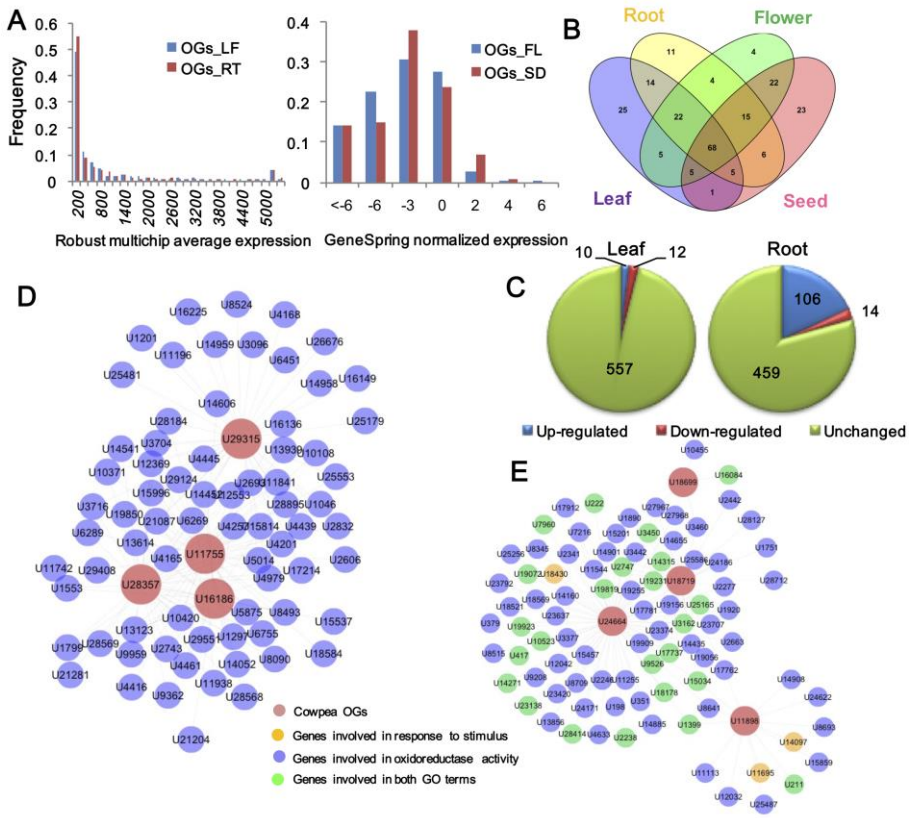

Figure 4

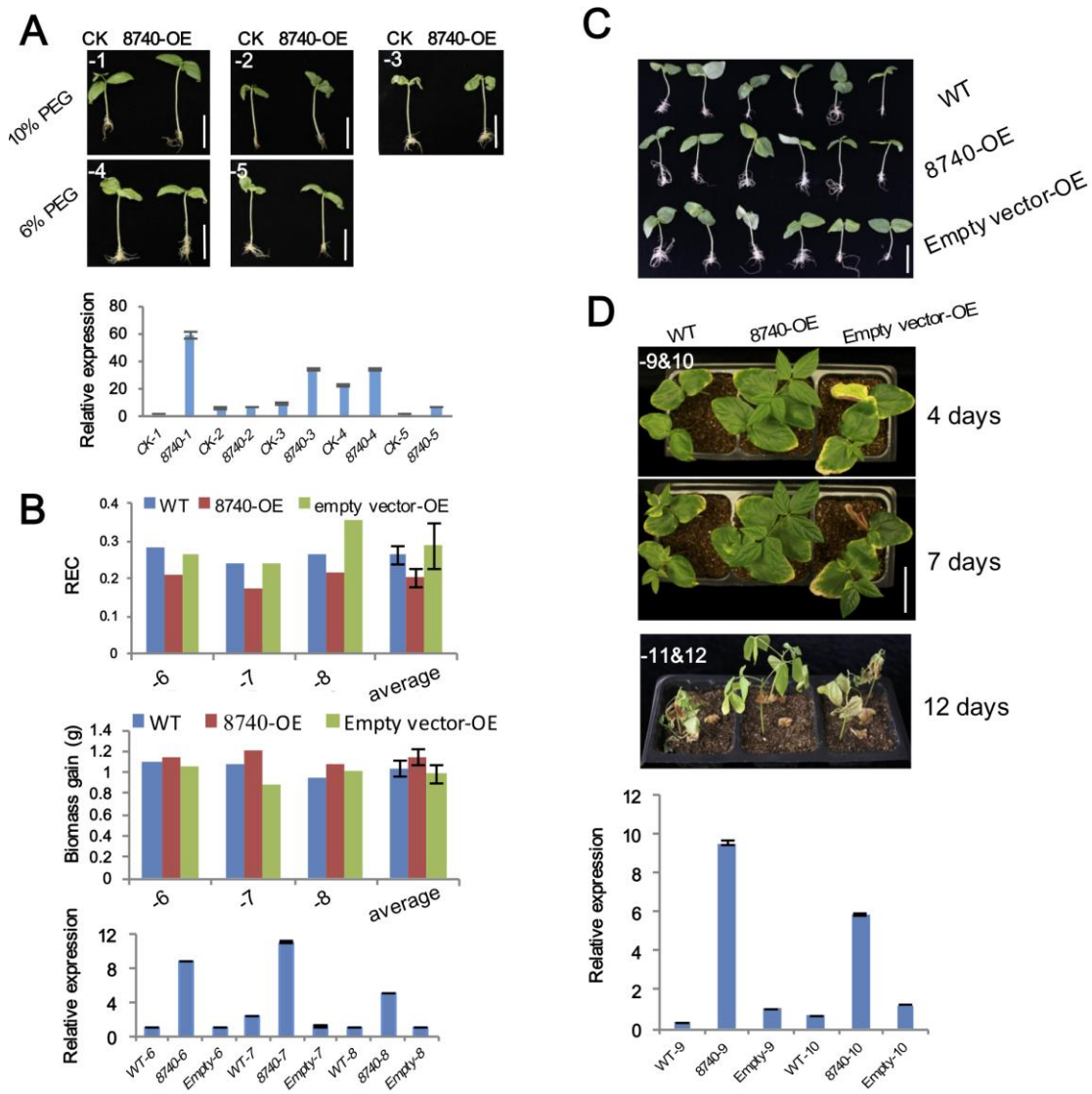

Figure 5