Co-Expression Network-based Analysis associated with potato initial resistance

Lang Yan[1,2], Xianjun Lai[1], Yan Wu[1], Xuemei Tan[1], Haiyan Wang[1], and Yizheng Zhang[1*]

[1]College of Life Sciences, Sichuan University, Key Laboratory of Bio-resources and Eco-environment, Ministry of Education, Sichuan Key Laboratory of Molecular Biology and Biotechnology, Center for Functional Genomics and Bioinformatics, Chengdu 610065, Sichuan, China

[2]Laboratory of Potato Functional Genome and Application, Xichang College, Liangshan, 615000, Sichuan, China

[*]To whom correspondence should be addressed. E-mail: yizzhang@scu.edu.cn

ABSTRACT

RNA sequencing (RNA-seq) providing genome-wide expression datasets has been successfully used to study gene expression patterns and regulation mechanism among multiple samples. Gene co-expression networks (GCNs) studies within or across species showed that coordinated genes in expression patterns are often functionally related. For potatoes, a large amount of publicly available transcriptome datasets have been generated but an optimal GCN detecting expression patterns in different genotypes, tissues and environmental conditions, is lacking. We constructed a potato GCN using 16 published RNA-Seq datasets covering 11 cultivars from native habitat worldwide. The correlations of gene expression were assessed pair-wisely and biologically meaningful gene modules which are highly connected in GCN were identified. One of the primitively native-farmer-selected cultivars in the Andes, ssp.*Andigena*, had relative far distance in gene expression patterns with other modern varieties. GCN in further enriched 134 highly and specifically co-expressed genes in ssp.*Andigena* associated with potato disease and stress resistance, which underlying the dramatic shift in evolutionary pressures during potato artificial domestication. In total, the network was consisted of into 14 gene models that involves in a variety of plant processes, which sheds light on how gene modules organized intra- and inter-varieties in the context of evolutionary divergence and provides a basis of information resource for potato gene functional studies.

Keywords: potato, RNA-seq, gene expression patterns, co-expressing genes

Introduction

As the fourth highest food crop with global production over 350 million tons, tropical-originated potato has been cultivated worldwide nowadays with a wide variety of habitats ranging from cool highland zones to tropical lowlands Pino et al. (2007); Kikuchi et al. (2015); Hawkes et al. (1990, 1994). Although applied over relatively short timescales, artificial selection has dramatically altered the tuber sizes and shapes, color in skins and fleshes, providing a wide range of germplasm resources accessible for research and breeding purposes Machida-Hirano (2015). Along with the publishing of potato reference genome, represented by the doubled monoploid *Solanum tuberosum* group *Phureja* clone Consortium et al. (2011), researchers have insight into the genetic performances in potatoes, from basic gene structure and biological function to their expression profiles. While advances over the past years have generated a growing number of gene expression datasets which provide unbiased snapshots of gene expression dynamics among worldwide potato cultivars, it remains a key challenge to deduce the underlying gene regulatory circuits in the process of the potato genetic diversification.

To address this, ever more gene expression datasets can be merged and integrated to develop a system-wide research Hughes et al. (2000); Wu et al. (2002); Childs et al. (2011). Gene co-expression networks (GCNs) has emerged as a tool that incorporated large-scale gene expression analyses, which describing the relatedness between gene expression patterns in a pairwise fashion Schmid et al. (2005); Lee et al. (2009); Wilkins et al. (2009). Utilizing Pearson correlation coefficient (PCC) to measure expression correlation between gene pairs from individual experiment, GCNs assembled via the Weighted Gene Coexpression Network Analysis (WGCNA) method could cluster gene pairs with PCC larger than a chosen threshold value into modules Langfelder and Horvath (2008). According to a ’guilt-by-association’ paradigm, genes with associated biological functions often have similar expression patterns, the genes (nodes) and connections (edges) in the module could represent co-expression dynamics in different gene expression datasets. Although several GCN studies have been done, such as COB Schaefer et al. (2014), CORNET De Bodt et al. (2012) in maize and SCoPNet Bassel et al. (2011) in Arabidopsis, most of them integrated the microarray expression datasets of cultivars worldwide. Given that RNA-Seq has become the favored technique for detecting genome-wide expression patterns, the increasing number of RNA-Seq in potatoes implies an RNA-Seq based potato GCN protocol would be valuable to the scientific community Crookshanks et al. (2001); Ronning et al. (2003); Flinn et al. (2005); Rensink et al. (2005); Li et al. (2006); Wang et al. (2009).

Here, we integrated 16 published RNA-Seq datasets from 11 distinct potato cultivars worldwide to identify divergences of general gene expression patterns intra- and inter-varieties. Based on the correlations of gene expression patterns, we constructed a potato WGCNA-based gene co-expression network, containing 12,346 genes and assembling into 14 distinct gene modules with an average module size of 881 genes. This potato GCN enriched genes highly expressed and correlated to certain cultivars or tissues into modules, which highlighting functions in potato disease and stress responses, photomorphogenesis and tuber dormant release development .etc. Importantly, these identified modules could further used to analyze the transcription factors which acting as regulators controlling gene expression patterns. The potato GCN study could provide us the valuable resources with potato gene expression rules, help design proper crossing schemes in potato breeding across different geographical areas and be useful for potato systems biology analysis.

Results

Overview of RNA-Seq datasets

Publicly available potato gene expression data were downloaded from NCBI bio-project database, a collection of diverse biological data across multiple potato varieties and developmental tissues Gong et al. (2015); Campbell et al. (2014); Petek et al. (2014); Liu et al. (2015b,a); Lin et al. (2015). Only data including single- or pair-end RNA-Seq runs in the last five years were considered and there were totally 16 datasets with above 3 replicates remained, which monitored potato transcriptome from various cultivars and tissues under normal growth conditions (Table 1). In particular, four datasets in leaves represented distinct varieties as 'Deisree', 'Sarpo Mira', 'Igor' and 'SW92-1015', and another four were tubers in dormant and sprouting status of var.'Longshu3' and var.'Russet Burbank'. The classic American commercial variety 'Atlantic', local Chinese varieties var.'Xindaping' and var.'Heimeiren' and primitive native-farmer-selected cultivars ssp.*Andigena* were also included. It is noteworthy that Andigena potatoes represents primitive cultivar group with a distinct gene pool compared to other modern potatoes, which used to identify widespread alteration of gene expression patterns during artificial domestication.

Differential gene expression patterns between primitive and modern cultivars

The collected raw data files (.sra) were downloaded, mapped against to the *S. tuberosum* group *Phureja* DM v4.03 reference genome Trapnell et al. (2012); Consortium et al. (2011), and processed into gene expression matrix represented by Fragments Per Kilobase of transcript per Million mapped reads (FPKM)-normalized gene expression values. The genes with average FPKM values < 3.0 were filter out, resulting in a gene expression matrix involving in 12,346 genes (rows) and 16 runs (columns). After classifying gene expression values into negligible (FPKM < 0.5), extremely low ($0.5 \leqslant$ FPKM < 3), low ($3 \leqslant$ FPKM < 50), moderate ($50 \leqslant$ FPKM < 100) and high (FPKM $\geqslant$100), we found similar expression patterns in most modern cultivars ,in which relatively less genes had negligible expression (191 genes in average) while a large number of genes had expression ranges from 3 to 50 FPKM (5,287 genes in average) (Figure 1). However, the primitive cultivar ssp.*Andigena* had distinct gene expression patterns in comparison with modern cultivars, even the comparisons were only involved in tubers. Nearly 3,000 genes in Andigena potatoes were tend to be silent with expression value < 0.5 FPKM and other 4,584 genes expressed extremely lowly which were in the range between 0.5 and 3 FPKM. The general expression pattern in primitive ssp.*Andigena* were leveled off without the expression peak in a certain region.

To identify the divergence of gene expression patterns in further between primitive and modern cultivars, the pair-wisely correlations of gene expression were calculated between datasets according to the spearman correlation coefficient (Figure 2, Supplementary Figure S1). The leaf tissues of 4 varieties ('Deisree', 'Igor', 'SarpoMira' and 'SW92') had the highest correlation coefficient between each pair ((r $>$0.85, p-values $<$0.05) compared to the rest tuber groups, indicating the higher correlations of gene expression patterns between tissues than that between cultivars (Table 2). Among tubers in different developing stage, gene expression had significant higher correlation coefficients within cultivars than that across cultivars, such as the dormant and non-dormant tubers of var.'RussetBurbank' (r=0.88, p-value $<$ 0.05) and that of var.'Longshu3' (r=0.85, p-value $<$0.05). Some cultivars were close-related in breeding process or geographical area, resulting in the significantly correlated gene expression across cultivars, such as tuber skins between var.'Heimeiren' and var.'Xindaping' (r=0.89, p-value

$<$0.05), tubers between var.'Atlantic' and var.'RussetBurbank' (r=0.83, p-value $<$0.05), tubers fleshes in var.'Xindaping' and tubers in var.'Longshu3' (r=0.86, p-value $<$0.05). Nevertheless, since the ssp.*Andigina* had no any significant correlations with other modern cultivars in expression levels ($r < 0.65$, p-value < 0.05), relationships between primitive and domesticated potatoes should be drifting apart.

Construction of gene co-expression modules

Based on correlations of gene expressions pair-wisely, GCNs were constructed using the weighted gene co-expression network analysis method (WGCNA), which is a systems biology approach aimed at understanding the gene expression patterns in networks instead of individual genes.Langfelder and Horvath (2008). The GCNs were consisted of 12,346 genes, involving in 14 distinct gene modules that contained 31-3283 genes with an average module size of 881 genes (Figure 3, Supplementary Figure S2). Each module represented genes with highly correlated expression profiles, either in a single or a few related datasets. The majority (80.5%) of genes assigned to co-expression modules have functional annotation. Thus, for genes with unknown functions, GCNs supplied potato gene annotation informations which placed unknown genes in a functional context and inferred their role according to their interaction with known meaningful genes. Functional enrichment analysis was performed on each module using total genes as population. A total of 213 GO terms were significantly enriched (Table 2). Since GO terms indicated the gene functional types in a certain module, it also contributed to understand the metabolic pathways which unknown genes involved in.

Module eigengenes (MEs) for each module are the first principal component of a given module and can be considered as a representative of the module's gene expression profileHollender et al. (2014). The MEs of 14 modules were calculated to check genes with similar expression patterns between different datasets, which showed in the heat map (Figure 4, Supplementary Figure S3). The dark red at the row-column intersection indicates the tissue specifically and highly expressed genes in the certain module. The result showed that gene expression patterns across different modules varied a lot. Six out of 14 modules are comprised of genes highly expressed in a single dataset ($r > 0.8$, $P < 10^{-3}$), which could be regarded as genes specifically correlated with a certain tissue/cultivar. Figure 5 displayed these tissue-specific gene expression modules which were mainly distributed on apical and lateral tubers of ssp.*Andigena* (Module #12 and #11), stolon of var.'Ningshu4' (Module #4), non-dormant tuber of var.'RussetBurbank' (Module #3) and tuber skin of var.'Heimeiren' and 'Xindaping' (Module #9 and #10). In addition, some modules contained genes that were co-expressed highly in multiple datasets, implying that these genes could participate in particular regulations in multi-tissues/varieites. For example, Module #1 and #2 were both co-expressed highly in leaves and enriched in GO terms of cytochrome b6f complex (GO:0009512) and chloroplast part (GO:0044434), indicating the genes in these two modules were related to leaf development and morphogenesis. Module #7 and #8 were co-expressed highly in all of the datasets except in ssp.*Andigina*, corresponding with the hypothesis that some genes under selections during the domestication process had changed there general gene expression patterns.

Gene modules revealing certain developmental processes

The standardized MEs above were represented by the principal gene expression profile in modules, some of which showed the relatively high and specific expression level in certain tissues/cultivars. We then concerned about the biological processes of these tissue-specifically expressed genes participated in. Our GCNs identified four particularly relevant modules. Module #1 and #2 were enriched with genes involved in leaf-specific expression in multiple cultivars, and contributes to the process of leaf morphogenesis and photosynthesis, including chlorophyll biosynthetic and electron transport chain in cytochrome b6/f for transferring electrons within the cyclic electron transport pathway of photosynthesis activity. A number of genes for the development of photosystem were only included in these two modules: magnesium protoporphyrin IX methyltransferase (PGSC0003DMG400014243), light harvesting (PGSC0003DMG400020492) and G-protein coupled photoreceptor activity (PGSC0003DMG400010034), etc.

We firstly identified the expression pattern of genes enriched in Module 1 and 9, which had the tissue-specific expression in leaves of multiple varieties. Pair-wisely comparison of the gene expression difference and specificity were conducted in leaves among var.'Deisree', var.'Igor' and var.'Mira' (var.'SW92' was excluded due to lack of background information). The three varieties had 32, 33 and 77 specifically expressed genes (SEGs) respectively, which might be related to their varietal specificity like cultivated lineages and local habits etc. (Figure 6A). Among all the expressed genes, there were 63 differentially expressed genes (DEGs) between 'Deisree' and 'Mira', 83 between 'Deisree' and 'Igor', and 26 between 'Igor' and 'Mira'. Moreover, there were some co-DEGs identified between two groups (Figure 6C,D), such as PGSC0003DMG400021142, the DGE in group 'Deisree' and 'Mira' (460.32 vs 2.22 FPKM) as well as the DGE between 'Igor' and 'Mira' (258.79 vs 2.22 FPKM). Importantly, two genes differentially expressed among all three cultivars, like PGSC0003DMG400011751, the expression value were 1881.08, 111.74 and 6.65 in 'Deisree', 'Igor' and 'Mira' respectively, indicating that the significant expressing volatility among different varieties. As expected, majority of the enriched genes in Module 1 and 9 of the co-expression network were DEGs and concentrated on photosynthesis process, including chlorophyll biosynthetic and electron transport chain in cytochrome b6/f for transferring electrons within the cyclic electron transport pathway of photosynthesis activity. The main functions of these genes encoded were magnesium protoporphyrin IX methyltransferase, light harvesting and G-protein coupled photoreceptor activity, etc.

Compared to leaf morphogenesis, gene expression and regulation of tuber sprouting, which is a major yield-determining trait, attracted more attention of researchers. Two modules involved in tuber dormant releasing process were identified, providing us the principal regulatory candidates involved in the tuber sprouting development. Genes in Module #3 specifically and highly expressed in sprouting tubers of var.'Russet Burbank', which predominantly encoded proteins such as auxin-induced protein, ethylene response factor, MADS-box protein, SAUR family protein, wound induced protein, etc. Seventy-three genes in Module #6 was specifically highly expressed in sprouting tubers of the Chinese local cultivar 'Longshu3', and showed functional specialization for tuber-specific and sucrose-responsive element binding factor, proteinase inhibitor, MYB domain class transcription factor, heat shock protein, etc.

Among the RNA-seqs in tubers, we compared the gene expression patterns of dormant and non-dormant tubers in both var.'Longshu3' ('LD' and 'LN' for short) and var.'Russet Burbank' ('RD' and 'RN'), in which the genes related to dormant releasing process had enriched in Module 3 and 6 of the co-expression networks. Apart from the 11,594 genes expressed in all four tubers, there were more genes specifically expressed in sprouting than that in dormant tissues, demonstrating the increasing of the gene expression activities in the dormancy release process. Then, the comparison of gene differential expression pair-wisely showed 121 (LDvsLN), 68 (RDvsRN), 31 (LDvsRD), and 69 (LNvsRN) genes had significantly differential expression, respectively. It is noteworthy that there were totally 19 genes had strong variety-differences in expression levels since they had co-differential expression. Moreover, genes in module 3 and 6, which were specifically and highly expressed in sprouting tubers of var.'Russet Burbank' and var.'Longshu3' respectively, provided us the principal regulatory candidates involved in the tuber dormant release development. Although these two modules were all consisted of tuber-specific genes, their gene functions and metabolic process enriched had differences between varieties. Module 3 was predominantly enriched in proteins such as auxin-induced protein, ethylene response factor, MADS-box protein, SAUR family protein, wound induced protein, etc. while Module 6 was enriched in tuber-specific and sucrose-responsive element binding factor, proteinase inhibitor, MYB domain class transcription factor, heat shock protein, demonstrating the differential expression patterns in two varieties with distinct backgrounds.

Primitive cultivar-specific module involves in initial resistance in ssp.*Andigena*

GCNs can be visualized as network maps using Cytoscape, in which the nodes represent genes and the connecting lines (edges) between genes represent gene correlationsShannon et al. (2003). Here we present an particular interest case of genes in a primitive cultivar-specific module (Module 11) that co-expressed specifically in lateral of ssp.*Andigena* (Figure 4). This network consisted of 134 correlative genes in which 3 hub genes had connections >80 and another 5 hub genes had connections >70 (Supplementary Table S1). Strikingly, 39 out of all hub genes (29%) were annotated as unknown functions and nearly half (10 out of 24) were functional unknown when only considering of the complicated regulatory genes with connections >60. These highly and specifically expressed genes with unknown functions in ssp.*Andigena*, which also existed in the primitive ssp.*Phureja* clone (the reference genome), might have correlations with the primitive cultivated potatoes under purifying selection during domestication.

The hub gene with the highest edge number (83 edges) is PGSC0003DMG401006052, a secA-type chloroplast protein transport factor, which is one of the important components in the protein translocation sec pathway in chloroplast. Other highly connected hub genes encoded guanine nucleotide regulatory protein, elicitor inducible LRR receptor, protein phosphatase 2c, senescence-associated protein, etc. Of particular importance is that lots of functional genes highly expressed in ssp.*Andigena* tended to be associated with potato disease and stress resistance. Previous studies have identified such genes in some model plants. For example, genes encoding the elicitor inducible LRR protein (EILP) was activated by treatment of salicylic acid or inoculation of *Pseudomonas syringae*, and the product of EILP amplified the sensitivity to disease stress and involved in non-host disease resistance in tobaccoTakemoto et al. (2000); the enhanced expression of senescence-associated genes (SAGs) could improve the ability in response to diseases caused by fungi, bacteria, and viruses through triggering the hypersensitive response (HR) or eliciting necrotic symptoms induced by virulent fungi and bacteria during infections, which had been observed in both the model plant *Arabidopsis thaliana* and a commercially important grapevine cultivarEspinoza et al. (2007); SAG29 in *Arabidopsis* could help the organism survive in the high salinity and other osmotic stress conditions through regulating the cell viability which may serve as a molecular link that integrates environmental stress responses into senescing processSeo et al. (2011). In addition, some functional genes with less connections in this network also showed important functions contributing to the potato growth and signal transduction processes such as gene encoding Auxin:hydrogen symporter, Auxin-induced protein X10A and F-box protein as an auxin receptor.

Discussion

Gene expression data have expanded the availability of genetic resources, which not only had been used as the supplement for gene functional annotation in a genomic scale, but also help to detect the individual gene expression patterns among tissues, varieties, or environmental conditions. In this study, we combined the gene expression correlation analyses and gene co-expression network to identify the correlations and divergences of gene expression, which may imply common functions or regulatory pathways among multiple potato cultivars. Making use of the integrated public RNA-Seq datasets could contribute to the identification of highly correlated genes in multiple parallel datasets. It also provides us an opportunity to study the gene regulatory mechanisms related to some important biological processes in systematic level, which is unattainable in single or few transcriptomes. The GCN established in this study enriched the highly correlated and expressed genes in the processes of photomorphogenesis and tuber dormant release development as well as stress resistance related genes specifically and highly expressed in primitive ssp.*Andigena*. The gene expression divergence inter-varieties breeding from different native habitats worldwide had never been reported before.

Gene expression patterns analyses between different cultivars in this study provide us some clues in the future breeding. Firstly, gene expression correlation in parallel tissues across cultivars could be associate with their relationships during breeding process. As shown in the Table 2, the Chinese local cultivars, 'Ningshu4' and the American traditional commercial variety 'Atlantic', which were bred from different geographical area, had high correlation coefficient in their gene expression patterns, implying their close relationship in breeding. Var.'Atlantic' should be the exotic breeding accessions in the breeding of 'Ningshu4'. Secondly, this study supported the previous hypothesis that long-term storage of potatoes is largely associated with the genetic control in potato itself instead of external environmental factors like soil and weather during potato growthSuttle (2007). The GCNs identified two gene modules, both of which had enriched genes specifically and highly expressed in sprouting tubers of var.'Longshu3' and var.'Russet Burbank', respectively. The var.'Longshu3' and var.'Russet Burbank' were both self-stable with relatively long dormant period, belonging to mid-late maturity varieties. However, some dormancy-controlled genes and their regulated pathways during the dormancy release process had different expression patterns in two varieties. For example, the genes encoding pectinesterase and xyloglucan endotransglucosylase, which should be the indicators of metabolic shifting into cell wall loosening, were found to be up-regulated expression only in non-dormant tubers of var.'Russet Burbank'. As reported before, the high expression of xyloglucan endotransglucosylase linked with protein increase that alters turgor pressure or cell wall pH valueresulting in the cell expansion Van Sandt et al. (2007); Senning et al. (2010); De Vos et al. (2012); Wolf et al. (2012). The pectinesterase has also been shown to be associated with increase in cell wall H^+ Wolf et al. (2012). Oppositely, non-specific lipid-transfer proteins and lipid binding proteins encoded genes were significantly highly-expressed and enriched in the dormant tubers of var.'Longshu3' rather than var.'Russet Burbank'. As most of the genes required for lipid degradation were expressed before bud emergence, which lead to energy release and supply of carbon resources for the production of sucrose to bud growth, the related genes in var.'Longshu3' might positively participate in lipid metabolism during the process of dormancy releasePalta et al. (1993); Liu et al. (2015b).

Moreover, our results supported the conclusion that long-term storage of potatoes is largely associated with the genetic control in the variety itself except for the external environmental factors like soil and weather during potato growthSuttle (2007). The var.'Longshu3' and var.'Russet Burbank' were both shelf-stable with relatively long dormant period, belonging to mid-late maturity varieties. The pair-wisely comparison of LD-LN and RD-RN demonstrated that although DEGs appeared during the dormancy release process, the dormancy-controlled genes and their regulated pathways between two varieties were not completely identical.

Lastly but most significantly, the specifically- and highly-expressed gene modules and their regulatory network in ssp.*Andigena* are valuable resources to identify the genetic locis being targets of artificial selection during potato domestication. Potatoes are native to the Andes of South America, where represent at best genetic diversity among potato germplasmSpooner et al. (2005a). Andigena (*Solanum tuberosum L.* subsp.*andigena Hawkes*) is the most primitive cultivated potato in the Andean highlands and is a likely ancestor of the worldwide grown modern potatoes *S.tuberosum* Hawkes et al. (1990); Hawkes (1956); Sukhotu and Hosaka (2006). Andigena potatoes hold genotypes of wild species and gradually form the abundant modern cultivated populations in ever-changing environment selected by nature and human beings, thus it is important in germplasm resources as a primary gene pool for improving worldwide grown potatoes in breedingSukhotu and Hosaka (2006). Along with the domestication and long-distance dispersal of seed tubers spreading from South America to Europe, Africa, Asia, and diffusing world widely, ssp.*Andigena* had shown multiple differences with the modern cultivars morphologically and physiologically, which are also reflected on the gene expression patternsGrun (1990); Spooner et al. (2005b); Ames and Spooner (2008). In our study, the expression of multiple gene in ssp.*Andigena* had decreased which is silenced or expressed extremely lowly. Thus, Andigena had no significant correlations with modern varieties in gene expression level, demonstrating the massive variation in transcriptional levels between primitive and modern cultivars. Moreover, Andigena remained valuable traits in resistance and nutritional contents which had lost in modern potato cultivars, not only conferring resistance to late blight, potato virus X, potato virus Y, nematodes, tuber moth, etc., but also acting as as a source of antioxidant phytochemicals and mineral micronutrients like carbohydrates, vitamin C, phenolic, carotenoid etc. Andre et al. (2007). For this reason, Andigena had its wide variability in tuber shape, flavor and cooking quality, which differ from the modern domesticated cultivarsAndre et al. (2007). As genetic improvement in modern potato breeding is facing increasingly narrow genetic basis and the decline of the genetic diversityChimote et al. (2004); Collares et al. (2004); Fernie et al. (2006), the introduction of primitive cultivars like Andigena, a very important gene pool, could provide the luxuriant and distinct gene resources for breeding and agricultural production. Although the multiple especially-correlated genes identified and enriched in Andigena showed a considerable variability in resistance and other important functions, nearly half of the genes in the network correlated specifically with Andigena had unknown gene functions, implying more attention should be drawn to the molecular studies in the primitive potato species in the future.

Methods

Raw RNA-Seq Dataset and Normalization of the Gene Expression

The RNA-Seq datasets used for the analysis including differential gene expression and the construction of the network were downloaded from NCBI SRA database Gong et al. (2015); Campbell et al. (2014); Petek et al. (2014); Liu et al. (2015b,a). These RNA-Seq datasets contained 16 tissues of 11 potato varieties and the SRA accessions are shown in Table 1. To simplify the further comparison, only the tissues under the normal growth conditions were selected. The reads of these datasets were filtered using Cutadapt v1.9Martin (2011) and mapped to the potato reference genomeConsortium et al. (2011) (PGSC_DM_v4.03_pseudomolecules.fasta & PGSC_DM_v4.03_genes.gff) from the potato genome database (http://solanaceae.plantbiology.msu.edu/pgsc_download.shtml) using GMAP/GSNAP (http://research-pub.gene.com/gmap/)Wu and Watanabe (2005); Martin (2011). Samtools kit was used to convert the data format and Cufflinks was used to normalize the gene expression levels as fragments per kilobase of exon model per million mapped reads (FPKM)Li (2011); Trapnell et al. (2012). The genes with extremely low expression (FPKM < 3) in average of all tissues were also filtered out in the construction of the co-expression gene network. Differential expression of transcripts was determined using CuffDiff v2.21 Trapnell et al. (2012) and Spearman correlation coefficient analyses between each two datasets were conducted using R packages.

Gene Co-Expression Network Construction

The gene expression matrix of potato was used to generate the gene co-expression network using weighted gene co-expression network analysis method (WGCNA)Langfelder and Horvath (2008). WGCNA is a systems biology method for describing the correlation patterns among genes across multiple samples and is based on the WGCNA R software package, a comprehensive collection of R functions for performing various aspects of weighted correlation network analysis. The thresholding power β was used to calculate adjacency between each two genes and 10 were chosen as optimal thresholding power among a set of candidate powers from 1 to 20 based on the function pickSoftThreshold returned results. The adjacency was calculated and then transformed into Topological Overlap Matrix (TOM) using function TOM similarity and the corresponding dissimilarity was calculated. The gene hierarchical clustering tree was produced using hierarchical clustering setting the minimum module size as 30 and the tree height cut as 0.25. The Module Eigengenes (MEs) were calculated and used to the further analyses. To visualize the network, the edge file and node file of the module were generated for dissimilarity and displayed on Cytoscape softwareShannon et al. (2003).

Functional Enrichment and Clustering

The Goatools v1.9 based on Python was introduced to analyze the enrichment of genes belonging to the different expression module in order to learn the gene GO terms and functions for each moduleHaibao et al. (2015). The genes participated in network construction were selected to get Gene ontology categories from GO database through Blast2Go (http://www.blast2go.com/b2glaunch/start-blast2go), with parameters of 20 hits and an e-value of 10e^{-6}. GO terms conforming to p-value through Bonferroni Correction 0.05 were defined as significantly enriched GO terms. The genome annotation file described above was used as the reference. Only GO terms for Biological Process are shown.

References

Ames, M., and Spooner, D.M. (2008). Dna from herbarium specimens settles a controversy about origins of the european potato. American Journal of Botany **95**:252–257.

Andre, C.M., Ghislain, M., Bertin, P., Oufir, M., del Rosario Herrera, M., Hoffmann, L., Hausman, J.F., Larondelle, Y., and Evers, D. (2007). Andean potato cultivars (solanum tuberosum l.) as a source of antioxidant and mineral micronutrients. Journal of agricultural and food chemistry **55**:366–378.

Bassel, G.W., Glaab, E., Marquez, J., Holdsworth, M.J., and Bacardit, J. (2011). Functional network construction in arabidopsis using rule-based machine learning on large-scale data sets. The Plant Cell **23**:3101–3116.

Campbell, M., Suttle, J., Douches, D.S., and Buell, C.R. (2014). Treatment of potato tubers with the synthetic cytokinin 1-(α-ethylbenzyl)-3-nitroguanidine results in rapid termination of endodormancy and induction of transcripts associated with cell proliferation and growth. Functional & integrative genomics **14**:789–799.

Childs, K.L., Davidson, R.M., and Buell, C.R. (2011). Gene coexpression network analysis as a source of functional annotation for rice genes. PloS one **6**:e22196.

Chimote, V., Chakrabarti, S., Pattanayak, D., and Naik, P. (2004). Semi-automated simple sequence repeat analysis reveals narrow genetic base in indian potato cultivars. Biologia plantarum **48**:517–522.

Collares, E.A.S., Choer, E., and Pereira, A.d.S. (2004). Characterization of potato genotypes using molecular markers. Pesquisa Agropecuária Brasileira **39**:871–878.

Consortium, P.G.S., et al. (2011). Genome sequence and analysis of the tuber crop potato. Nature **475**:189–195.

Crookshanks, M., Emmersen, J., Welinder, K.G., and Lehmann Nielsen, K. (2001). The potato tuber transcriptome: analysis of 6077 expressed sequence tags. FEBS letters **506**:123–126.

De Bodt, S., Hollunder, J., Nelissen, H., Meulemeester, N., and Inzé, D. (2012). Cornet 2.0: integrating plant coexpression, protein–protein interactions, regulatory interactions, gene associations and functional annotations. New Phytologist **195**:707–720.

De Vos, D., Dzhurakhalov, A., Draelants, D., Bogaerts, I., Kalve, S., Prinsen, E., Vissenberg, K., Vanroose, W., Broeckhove, J., and Beemster, G.T. (2012). Towards mechanistic models of plant organ growth. Journal of experimental Botany **63**:3325–3337.

Espinoza, C., Medina, C., Somerville, S., and Arce-Johnson, P. (2007). Senescence-associated genes induced during compatible viral interactions with grapevine and arabidopsis. Journal of Experimental Botany **58**:3197–3212.

Fernie, A.R., Tadmor, Y., and Zamir, D. (2006). Natural genetic variation for improving crop quality. Current opinion in plant biology **9**:196–202.

Flinn, B., Rothwell, C., Griffiths, R., Lägue, M., DeKoeyer, D., Sardana, R., Audy, P., Goyer, C., Li, X.Q., Wang-Pruski, G., et al. (2005). Potato expressed sequence tag generation and analysis using standard and unique cdna libraries. Plant molecular biology **59**:407–433.

Gong, L., Zhang, H., Gan, X., Zhang, L., Chen, Y., Nie, F., Shi, L., Li, M., Guo, Z., Zhang, G., et al. (2015). Transcriptome profiling of the potato (solanum tuberosum l.) plant under drought stress and water-stimulus conditions. PloS one **10**:e0128041.

Grun, P. (1990). The evolution of cultivated potatoes. Economic Botany **44**:39–55.

Haibao, T., Debra, K., Brent, P., Patrick, F., Kenta, S., Fidel, R., Jeff, Y., and Chris, M. (2015). Goatools: Tools for gene ontology. Zenodo 10.5281/zenodo.31628.

Hawkes, J., Bradshaw, J., Mackay, G., et al. (1994). Origins of cultivated potatoes and species relationships. Potato genetics. 3–42.

Hawkes, J.G. (1956). Taxonomic studies on the tuber-bearing solanums. 1: Solanum tuberosum and the tetraploid species complex. In Proceedings of the Linnean Society of London, vol. 166, 97–144, Wiley Online Library.

Hawkes, J.G., et al. (1990). The potato: evolution, biodiversity and genetic resources. Belhaven Press.

Hollender, C.A., Kang, C., Darwish, O., Geretz, A., Matthews, B.F., Slovin, J., Alkharouf, N., and Liu, Z. (2014). Floral transcriptomes in woodland strawberry uncover developing receptacle and anther gene networks. Plant Physiology **165**:1062–1075.

Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., et al. (2000). Functional discovery via a compendium of expression profiles. Cell **102**:109–126.

Kikuchi, A., Huynh, H.D., Endo, T., and Watanabe, K. (2015). Review of recent transgenic studies on abiotic stress tolerance and future molecular breeding in potato. Breeding science **65**:85–102.

Langfelder, P., and Horvath, S. (2008). Wgcna: an r package for weighted correlation network analysis. BMC bioinformatics **9**:1.

Lee, T.H., Kim, Y.K., Pham, T.T.M., Song, S.I., Kim, J.K., Kang, K.Y., An, G., Jung, K.H., Galbraith, D.W., Kim, M., et al. (2009). Ricearraynet: a database for correlating gene expression from transcriptome profiling, and its application to the analysis of coexpressed genes in rice. Plant Physiology **151**:16–33.

Li, H. (2011). A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics **27**:2987–2993.

Li, X.Q., Griffiths, R., Lagüe, M., DeKoeyer, D., Rothwell, C., Haroon, M., Stevens, B., Xu, C., Gustafson, V., Bonierbale, M., et al. (2006). Est sequencing and analysis from cold-stored and reconditioned potato tubers. In VI International Solanaceae Conference: Genomics Meets Biodiversity 745, 491–495.

Lin, T., Lashbrook, C.C., Cho, S.K., Butler, N.M., Sharma, P., Muppirala, U., Severin, A.J., and Hannapel, D.J. (2015). Transcriptional analysis of phloem-associated cells of potato. BMC genomics **16**:1.

Liu, B., Zhang, N., Wen, Y., Jin, X., Yang, J., Si, H., and Wang, D. (2015a). Transcriptomic changes during tuber dormancy release process revealed by rna sequencing in potato. Journal of biotechnology **198**:17–30.

Liu, Y., Lin-Wang, K., Deng, C., Warran, B., Wang, L., Yu, B., Yang, H., Wang, J., Espley, R.V., Zhang, J., et al. (2015b). Comparative transcriptome analysis of white and purple potato to identify genes involved in anthocyanin biosynthesis. PloS one **10**:e0129148.

Machida-Hirano, R. (2015). Diversity of potato genetic resources. Breeding science **65**:26–40.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal **17**:pp–10.

Palta, J.P., Whitaker, B.D., and Weiss, L.S. (1993). Plasma membrane lipids associated with genetic variability in freezing tolerance and cold acclimation of solanum species. Plant Physiology **103**:793–803.

Petek, M., Rotter, A., Kogovšek, P., Baebler, Š., Mithöfer, A., and Gruden, K. (2014). Potato virus y infection hinders potato defence response and renders plants more vulnerable to colorado potato beetle attack. Molecular ecology **23**:5378–5391.

Pino, M.T., Skinner, J.S., Park, E.J., Jeknić, Z., Hayes, P.M., Thomashow, M.F., and Chen, T.H. (2007). Use of a stress inducible promoter to drive ectopic atcbf expression improves potato freezing tolerance while minimizing negative effects on tuber yield. Plant Biotechnology Journal **5**:591–604.

Rensink, W., Hart, A., Liu, J., Ouyang, S., Zismann, V., and Buell, C.R. (2005). Analyzing the potato abiotic stress transcriptome using expressed sequence tags. Genome **48**:598–605.

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., et al. (2003). Comparative analyses of potato expressed sequence tag libraries. Plant Physiology **131**:419–429.

Schaefer, R.J., Briskine, R., Springer, N.M., and Myers, C.L. (2014). Discovering functional modules across diverse maize transcriptomes using cob, the co-expression browser. PloS one **9**:e99193.

Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Schölkopf, B., Weigel, D., and Lohmann, J.U. (2005). A gene expression map of arabidopsis thaliana development. Nature genetics **37**:501–506.

Senning, M., Sonnewald, U., and Sonnewald, S. (2010). Deoxyuridine triphosphatase expression defines the transition from dormant to sprouting potato tuber buds. Molecular breeding **26**:525–531.

Seo, P.J., Park, J.M., Kang, S.K., Kim, S.G., and Park, C.M. (2011). An arabidopsis senescence-associated protein sag29 regulates cell viability under high salinity. Planta **233**:189–200.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research **13**:2498–2504.

Spooner, D., Nunez, J., Rodriguez, F., Naik, P., and Ghislain, M. (2005a). Nuclear and chloroplast dna reassessment of the origin of indian potato varieties and its implications for the origin of the early european potato. Theoretical and Applied Genetics **110**:1020–1026.

Spooner, D.M., McLean, K., Ramsay, G., Waugh, R., and Bryan, G.J. (2005b). A single domestication for potato based on multilocus amplified fragment length polymorphism genotyping. Proceedings of the National Academy of Sciences of the United States of America **102**:14694–14699.

Sukhotu, T., and Hosaka, K. (2006). Origin and evolution of andigena potatoes revealed by chloroplast and nuclear dna markers. Genome **49**:636–647.

Suttle, J.C. (2007). Dormancy and sprouting. Potato Physiology and Biotechnology. Advances and perspectives, edition 287–305.

Takemoto, D., Hayashi, M., Doke, N., Nishimura, M., and Kawakita, K. (2000). Isolation of the gene for eilp, an elicitor-inducible lrr receptor-like protein, from tobacco by differential display. Plant and Cell Physiology **41**:458–464.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. Nature protocols **7**:562–578.

Van Sandt, V.S., Suslov, D., Verbelen, J.P., and Vissenberg, K. (2007). Xyloglucan endotransglucosylase activity loosens a plant cell wall. Annals of Botany **100**:1467–1473.

Wang, Z., Gerstein, M., and Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. Nature reviews genetics **10**:57–63.

Wilkins, O., Nahal, H., Foong, J., Provart, N.J., and Campbell, M.M. (2009). Expansion and diversification of the populus r2r3-myb family of transcription factors. Plant Physiology **149**:981–993.

Wolf, S., Hématy, K., and Höfte, H. (2012). Growth control and cell wall signaling in plants. Annual review of plant biology **63**:381–407.

Wu, L.F., Hughes, T.R., Davierwala, A.P., Robinson, M.D., Stoughton, R., and Altschuler, S.J. (2002). Large-scale prediction of saccharomyces cerevisiae gene function using overlapping transcriptional clusters. Nature genetics **31**:255–265.

Wu, T.D., and Watanabe, C.K. (2005). Gmap: a genomic mapping and alignment program for mrna and est sequences. Bioinformatics **21**:1859–1875.

Acknowledgements

The authors would like to thank Dr.Xianjun Lai (from SCU) for his assistance with some scripts and codes in bioinformatic work. Thanks also to Xia Zhan for scientific English language editorial assistance.

Author contributions statement

L.Y. and YZ.Z. conceived the analyses, L.Y collected the data, L.Y., Y.Q. and ZR.F. performed the calculations and analysis, L.Y. completed the manuscript. All authors discussed the results and reviewed the manuscript.

Competing financial interests

The authors declare no competing financial interests.

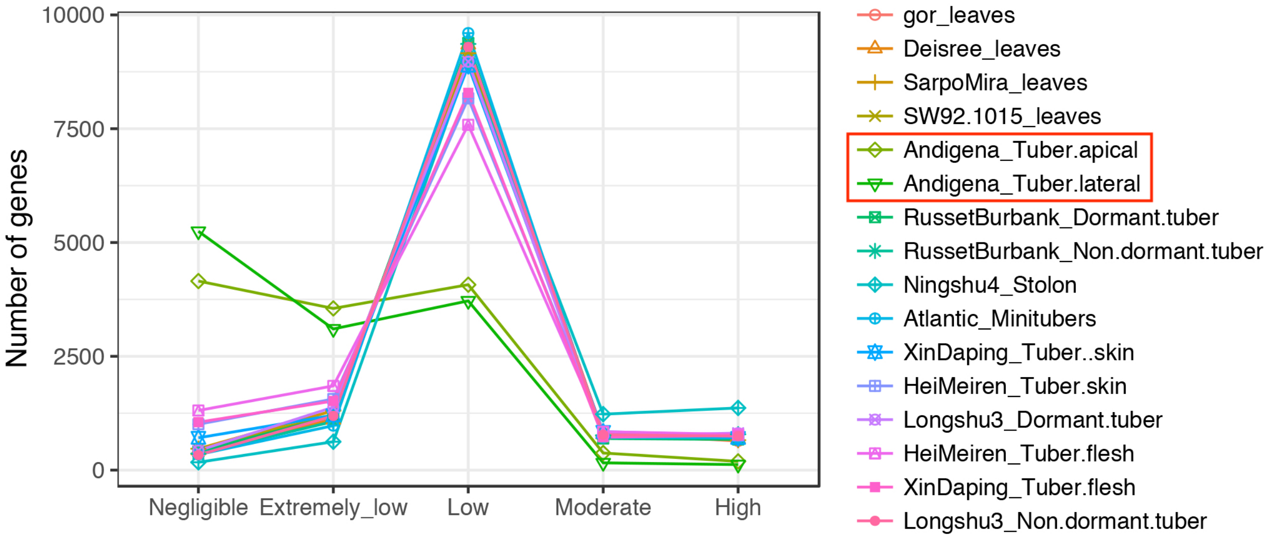


Figure 1. Gene expression patterns in different datasets. The genes with FPKM values were classified into negligible (FPKM < 0.5), extremely low ($0.5 \leqslant$ FPKM < 3), low ($3 \leqslant$ FPKM < 50), moderate ($50 \leqslant$ FPKM < 100) and high expressed (FPKM $\geqslant 100$).

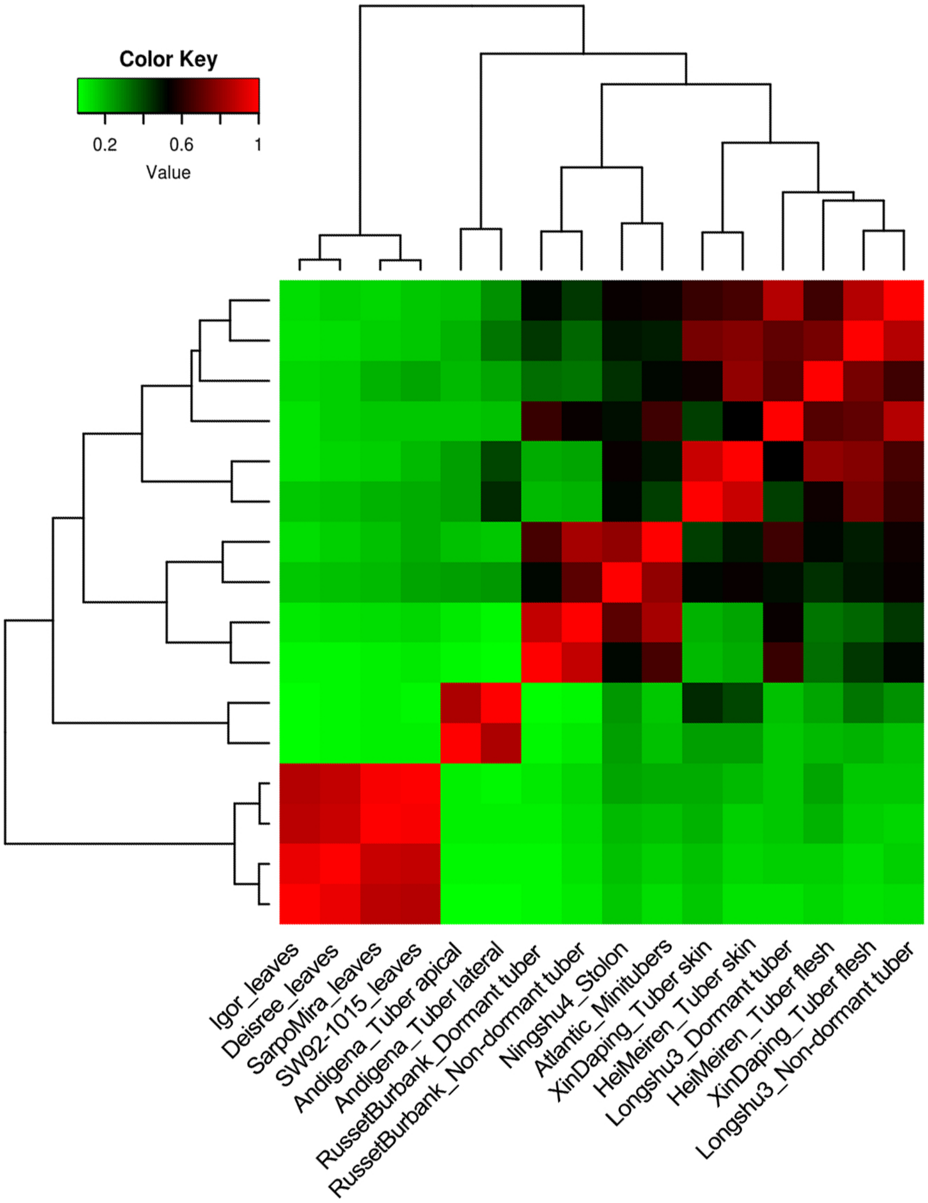


Figure 2. Cluster heat maps of log2-transformed FPKM values using the Spearman correlation coefficients among 16 datasets.

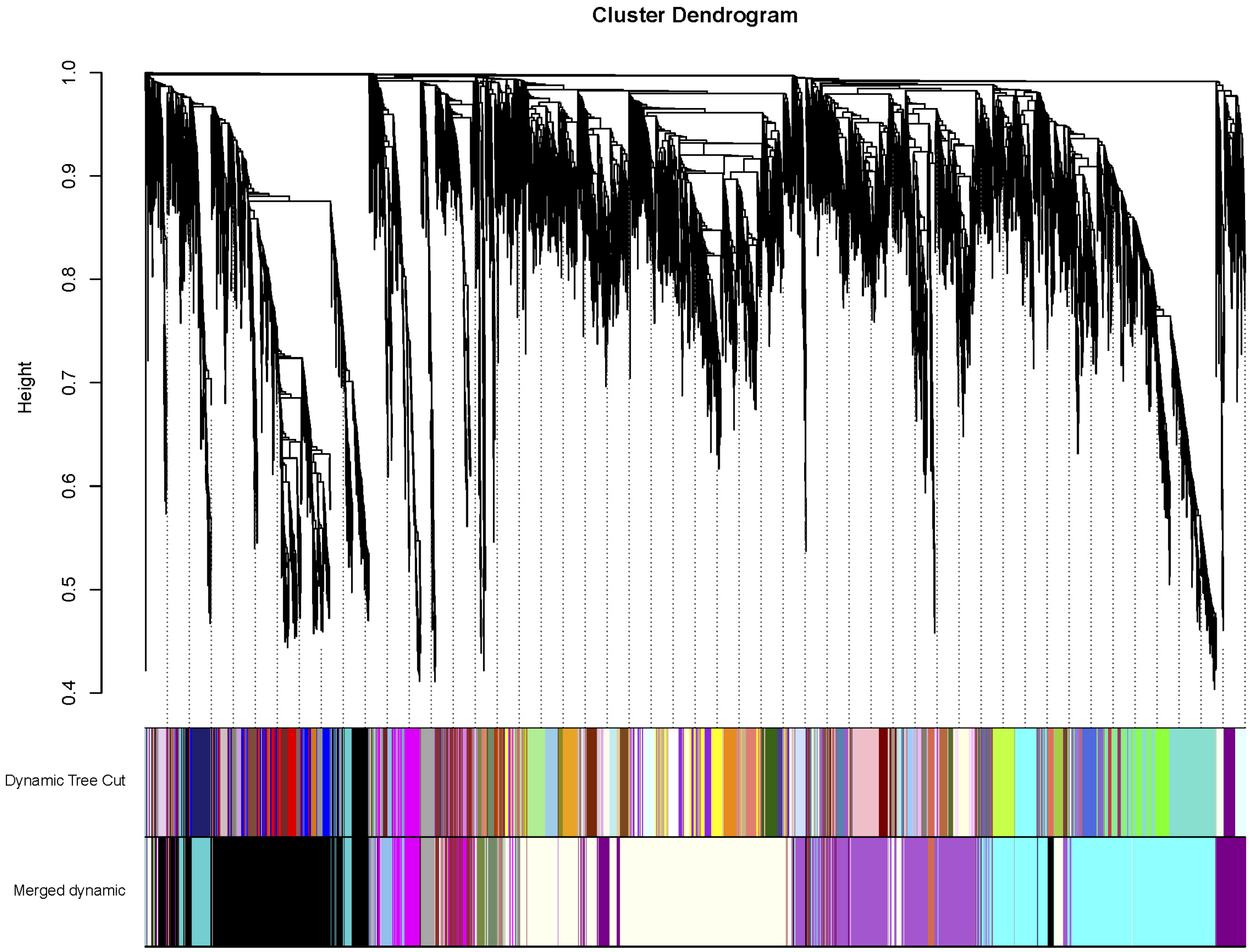
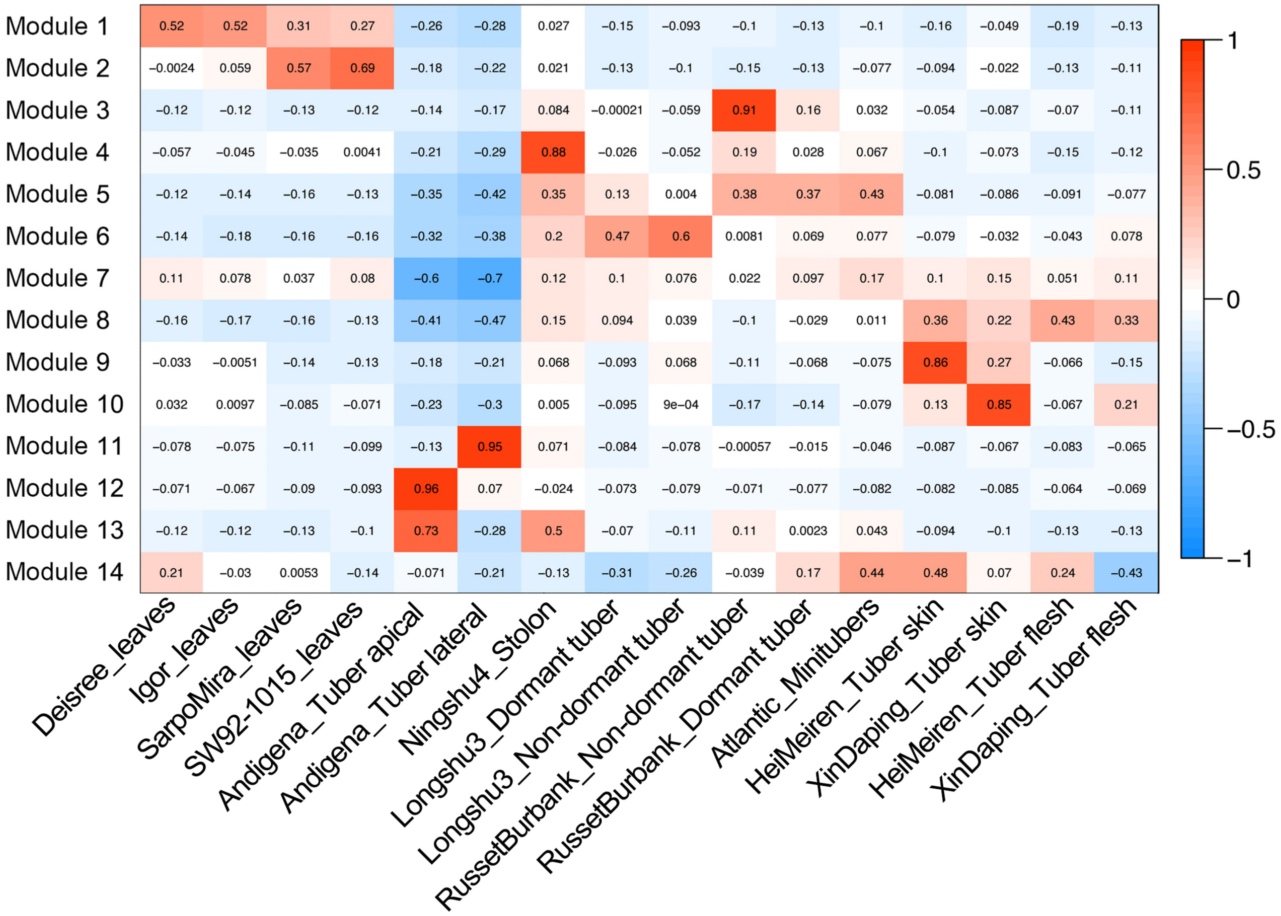


Figure 3. Hierarchical cluster tree showing co-expression modules identified by WGCNA. Each leaf in the tree represents one gene. The major tree branches constitute 14 modules are labeled by different colors.



	Deisree_leaves	Igor_leaves	SarpoMira_leaves	SW92-1015_leaves	Andigena_Tuber apical	Andigena_Tuber lateral	Ningshu4_Stolon	Longshu3_Dormant tuber	Longshu3_Non-dormant tuber	RussetBurbank_Non-dormant tuber	RussetBurbank_Dormant tuber	Atlantic_Minitubers	HeiMeiren_Tuber skin	XinDaping_Tuber skin	HeiMeiren_Tuber flesh	XinDaping_Tuber flesh
Module 1	0.52	0.52	0.31	0.27	–0.26	–0.28	0.027	–0.15	–0.093	–0.1	–0.13	–0.1	–0.16	–0.049	–0.19	–0.13
Module 2	–0.0024	0.059	0.57	0.69	–0.18	–0.22	0.021	–0.13	–0.1	–0.15	–0.13	–0.077	–0.094	–0.022	–0.13	–0.11
Module 3	–0.12	–0.12	–0.13	–0.12	–0.14	–0.17	0.084	–0.00021	–0.059	0.91	0.16	0.032	–0.054	–0.087	–0.07	–0.11
Module 4	–0.057	–0.045	–0.035	0.0041	–0.21	–0.29	0.88	–0.026	–0.052	0.19	0.028	0.067	–0.1	–0.073	–0.15	–0.12
Module 5	–0.12	–0.14	–0.16	–0.13	–0.35	–0.42	0.35	0.13	0.004	0.38	0.37	0.43	–0.081	–0.086	–0.091	–0.077
Module 6	–0.14	–0.18	–0.16	–0.16	–0.32	–0.38	0.2	0.47	0.6	0.0081	0.069	0.077	–0.079	–0.032	–0.043	0.078
Module 7	0.11	0.078	0.037	0.08	–0.6	–0.7	0.12	0.1	0.076	0.022	0.097	0.17	0.1	0.15	0.051	0.11
Module 8	–0.16	–0.17	–0.16	–0.13	–0.41	–0.47	0.15	0.094	0.039	–0.1	–0.029	0.011	0.36	0.22	0.43	0.33
Module 9	–0.033	–0.0051	–0.14	–0.13	–0.18	–0.21	0.068	–0.093	0.068	–0.11	–0.068	–0.075	0.86	0.27	–0.066	–0.15
Module 10	0.032	0.0097	–0.085	–0.071	–0.23	–0.3	0.005	–0.095	9e–04	–0.17	–0.14	–0.079	0.13	0.85	–0.067	0.21
Module 11	–0.078	–0.075	–0.11	–0.099	–0.13	0.95	0.071	–0.084	–0.078	–0.00057	–0.015	–0.046	–0.087	–0.067	–0.083	–0.065
Module 12	–0.071	–0.067	–0.09	–0.093	0.96	0.07	–0.024	–0.073	–0.079	–0.071	–0.077	–0.082	–0.082	–0.085	–0.064	–0.069
Module 13	–0.12	–0.12	–0.13	–0.1	0.73	–0.28	0.5	–0.07	–0.11	0.11	0.0023	0.043	–0.094	–0.1	–0.13	–0.13
Module 14	0.21	–0.03	0.0053	–0.14	–0.071	–0.21	–0.13	–0.31	–0.26	–0.039	0.17	0.44	0.48	0.07	0.24	–0.43

Figure 4. Heatmap representing the strength and significance of correlations between module eigengenes and different cultivars/tissues. Each row corresponds to a module. The number of genes in each module is indicated on the left. Each column corresponds to a specific tissue. The color of each cell at the row-column intersection indicates the MEs of the module in tissues.

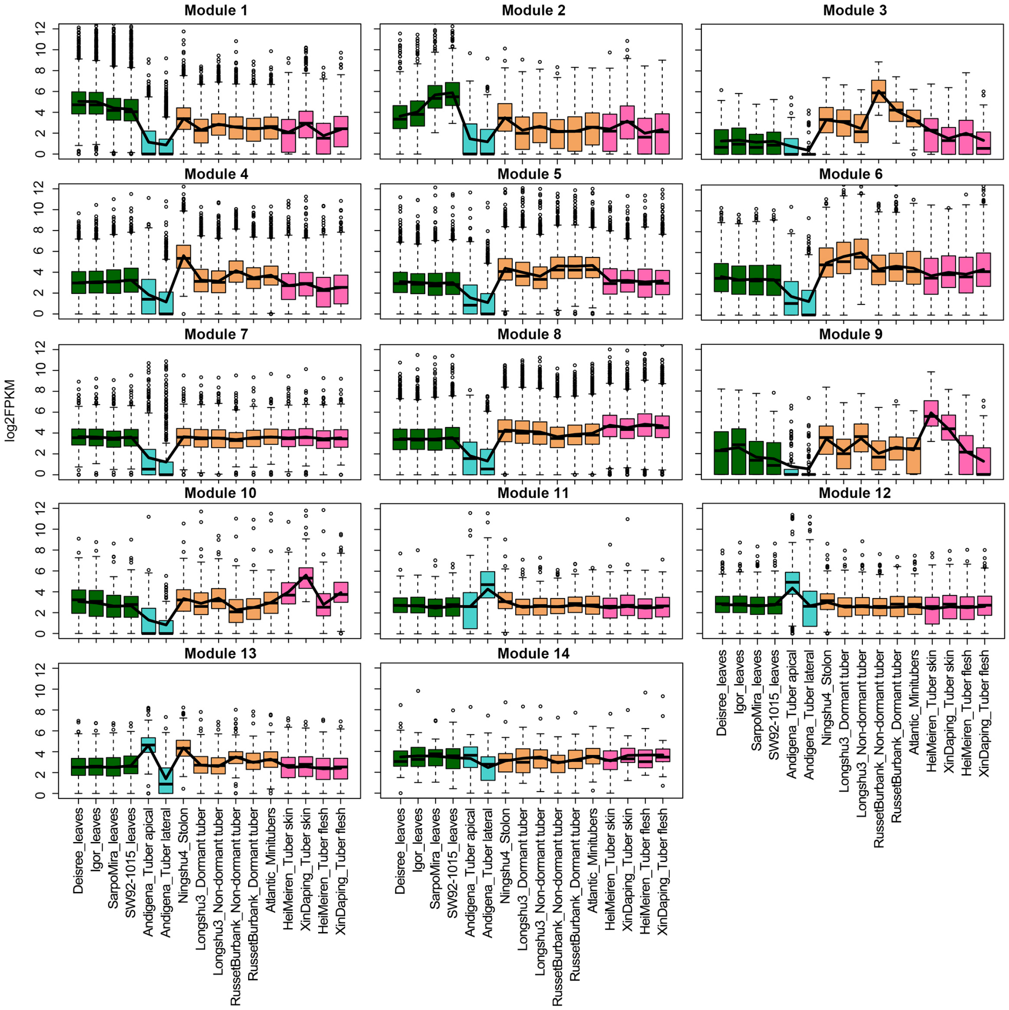


Figure 5. Boxplots of co-expression patterns of 16 datasets in the 14 modules. Gene expression values were processed using WGCNA to identify modules of highly correlated genes.

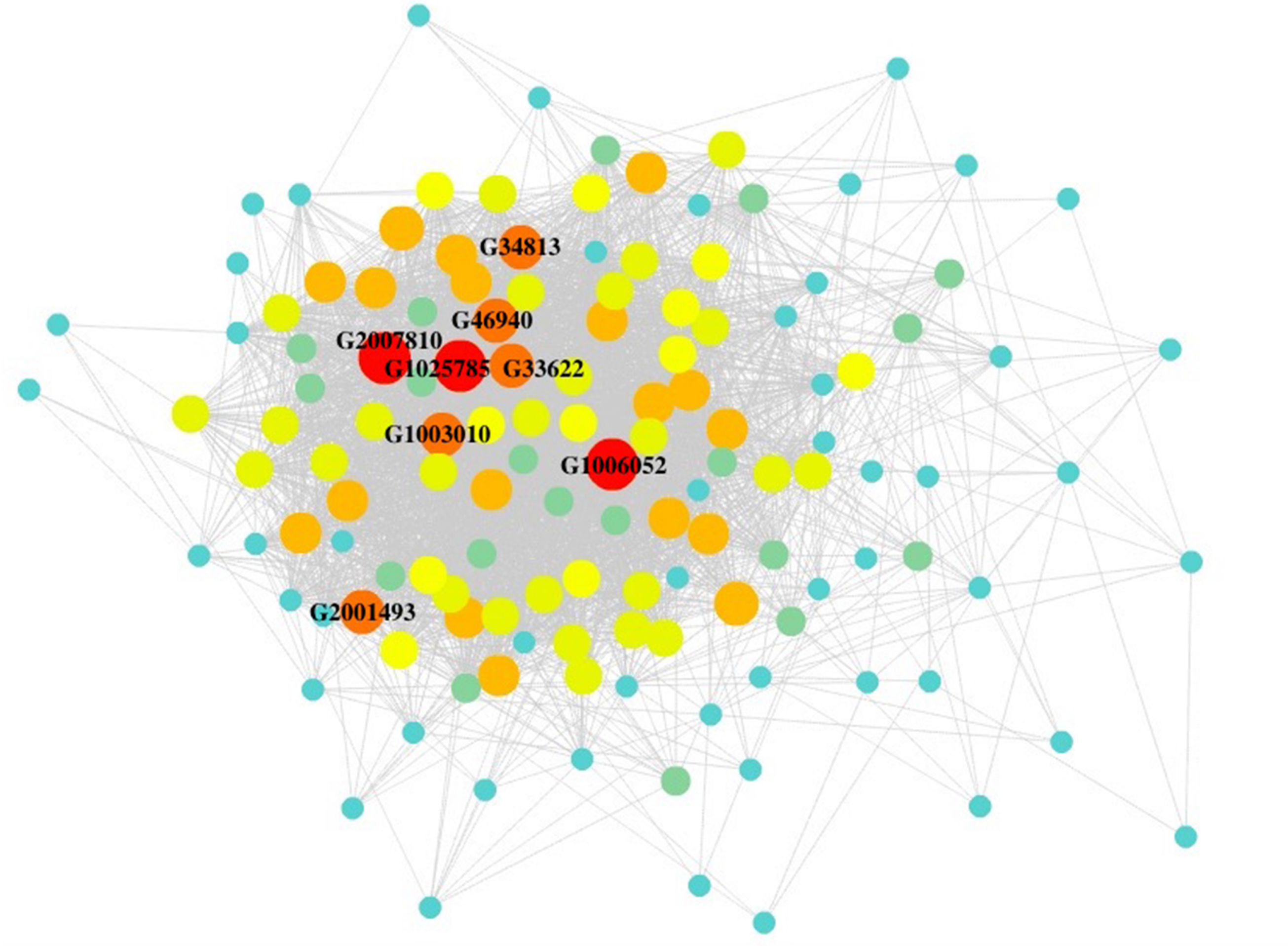


Figure 6. The correlation network 134 genes visualized by Cytoscape. Hub genes are indicated by larger circles shown in the network.

Table 1. Potato gene expression datasets used in this study.

Accession	Data Sets	Varieties	Tissues
PRJNA278091	SRR1867889	Ningshu4	Stolon tips
PRJNA262145	SRR1586393	Atlantic	Minitubers
PRJNA261838	SRR1584269	Russet Burbank	Dormant tuber
PRJNA261838	SRR1584266	Russet Burbank	Non-dormant tuber
PRJNA242936	SRR1207290	Deisree	Fully expanded leaves
PRJNA242936	SRR1207286	Igor	Fully expanded leaves
PRJEB4212	ERR305631	Sarpo Mira	Fully expanded leaves
PRJEB4212	ERR305630	SW92-1015	Fully expanded leaves
PRJNA237350	SRR1200975	Xin Daping	Tuber skin
PRJNA237350	SRR1170971	Xin Daping	Tuber flesh
PRJNA237350	SRR1200977	Hei Meiren	Tuber skin
PRJNA237350	SRR1200976	Hei Meiren	Tuber flesh
PRJNA229434	SRR1103934	Longshu3	Dormant tuber
PRJNA229434	SRR1103935	Longshu3	Non-dormant tuber
PRJNA285785	SRR2050324	Andigena	Tuber apical
PRJNA285785	SRR2050326	Andigena	Tuber lateral

Table 2. Summary of the number of entities and enriched GO terms in each validated module.

Cluster	Entities in Cluster	Functional annotation	GO terms
Module 1	2108	1708	26
Module 2	416	341	2
Module 3	78	47	2
Module 4	2421	1938	45
Module 5	1839	1519	24
Module 6	606	500	25
Module 7	439	339	14
Module 8	3283	2652	25
Module 9	102	85	8
Module 10	144	103	9
Module 11	184	140	15
Module 12	440	343	14
Module 13	255	207	4
Module 14	31	22	0