expVIP: a Customizable RNA-seq Data Analysis and Visualization Platform[1][OPEN]

Philippa Borrill[2], Ricardo Ramirez-Gonzalez[2], and Cristobal Uauy*

John Innes Centre, Norwich NR4 7UH, United Kingdom (P.B., C.U.); and Genome Analysis Centre, Norwich NR4 7UH, United Kingdom (R.R.-G.)

ORCID IDs: 0000-0002-7623-8256 (P.B.); 0000-0001-5745-7085 (R.R.-G.); 0000-0002-9814-1770 (C.U.).

The majority of transcriptome sequencing (RNA-seq) expression studies in plants remain underutilized and inaccessible due to the use of disparate transcriptome references and the lack of skills and resources to analyze and visualize these data. We have developed expVIP, an expression visualization and integration platform, which allows easy analysis of RNA-seq data combined with an intuitive and interactive interface. Users can analyze public and user-specified data sets with minimal bioinformatics knowledge using the expVIP virtual machine. This generates a custom Web browser to visualize, sort, and filter the RNA-seq data and provides outputs for differential gene expression analysis. We demonstrate expVIP's suitability for polyploid crops and evaluate its performance across a range of biologically relevant scenarios. To exemplify its use in crop research, we developed a flexible wheat (*Triticum aestivum*) expression browser (www.wheat-expression.com) that can be expanded with user-generated data in a local virtual machine environment. The open-access expVIP platform will facilitate the analysis of gene expression data from a wide variety of species by enabling the easy integration, visualization, and comparison of RNA-seq data across experiments.

The global demand for staple crops is predicted to double by 2050 (FAO, 2009; Tilman et al., 2011), which will require an annual increase in yield of approximately 2.4% (Ray et al., 2013). However, currently, yields of the major crops maize (*Zea mays*), rice (*Oryza sativa*), wheat (*Triticum aestivum*), and soybean (*Glycine max*) are increasing only at 1.6%, 1%, 0.9%, and 1.3% per year, respectively (Ray et al., 2013). The advent of the genomics era represents a great opportunity to accelerate the pace of yield increase in staple crops, for example, by facilitating novel breeding strategies (Heffner et al., 2009) and providing unprecedented numbers of genetic markers (Bevan and Uauy, 2013). In particular, transcriptome sequencing (RNA-seq) is a widely adopted genomics approach in crops due to its relatively low cost (Wang et al., 2009), its suitability for nonmodel organisms (Ekblom and Galindo, 2011), and the multiple downstream applications of the data generated. These features have driven the generation of a wealth of expression data with over 9,000 RNA-seq samples currently available at public repositories, such as the National Center for Biotechnology Information (NCBI)/ENA for the major agricultural crops (Table I).

Although several public databases containing gene expression data for plant species exist (Lawrence et al., 2007; Ouyang et al., 2007; Dash et al., 2012), these resources do not make full use of the expression data available in SRAs, frequently relying on a subset of experiments or microarray data. Similarly, pipelines have been proposed to allow the reanalysis of expression data that provide useful functionality but limit the number of samples that can be analyzed (D'Antonio et al., 2015), have limited visualization outputs (Fonseca et al., 2014), or require the user to process their own data before uploading to a visualization tool (Nussbaumer et al., 2014). In most cases, visualization tools are static and do not allow meaningful comparison of data. In addition, many studies used disparate transcriptome assemblies or annotations that hinder the possibility to compare results across different biological samples (Gillies et al., 2012; Pfeifer et al., 2014). Thus, despite the significant investment in RNA-seq studies across the major agricultural crops, these data remain largely underutilized and inaccessible to the majority of breeders and biologists due to the lack of common platforms and resources to analyze the data.

[1] This work was supported by the Biotechnology and Biological Sciences Research Council (grant nos. BB/J004588/1 and BB/J003557/1 to C.U. and Anniversary Future Leader fellowship no. BB/M014045/1 to P.B.) and by a Norwich Research Park Ph.D. studentship and a Genome Analysis Centre funding and maintenance grant to R.R.-G.

[2] These authors contributed equally to the article.

* Address correspondence to cristobal.uauy@jic.ac.uk.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Cristobal Uauy (cristobal.uauy@jic.ac.uk).

P.B. catalogued data for inclusion, performed wet-lab experiments, and carried out the RNA-seq alignments and differential expression analysis; R.R.-G. developed the expVIP source code, the virtual machine, and the wheat expression browser Web site; all authors contributed to the conception of the project and writing the article.

[OPEN] Articles can be viewed without a subscription.

www.plantphysiol.org/cgi/doi/10.1104/pp.15.01667

We have developed expVIP (expression Visualization and Integration Platform), an adaptable platform to create a gene expression interface for any species with a transcriptome assembly available. We provide a user-friendly virtual machine implementation allowing breeders and biologists to access this resource on a desktop personal computer. expVIP takes an input of RNA-seq reads (from single or multiple studies), quantifies expression per gene using the fast pseudoaligner kallisto (Bray et al., 2015), and creates a database containing expression and sample information. This platform allows comparisons across studies, and the output is viewable as a Web browser interface with intuitive and interactive filtering, sorting, and export options.

We have implemented expVIP on wheat to demonstrate its potential to be applied to crop species. In particular, our analysis of wheat data demonstrates the pipeline's ability to handle data from polyploid species, a key aspect for agricultural research, since many of the world's major crops are polyploid or have undergone recent whole-genome duplication events (Bevan and Uauy, 2013; Table I). In the case of wheat, we reanalyzed 418 RNA-seq samples from 16 studies including diverse developmental time courses, tissues, pathogen infections, and abiotic stresses. We conducted a series of analyses to demonstrate its utility for candidate gene characterization and its potential to compare across independent studies and generate novel hypotheses. Using expVIP, we developed a wheat expression browser (www.wheat-expression.com) as a community resource to access publicly available wheat RNA-seq data.

Table I. *Publicly available RNA sequencing samples in the NCBI short read archive (SRA) for the top 10 crops based on production (FAO, 2015) and additional agricultural species (as of August 5, 2015)*

Ploidy levels and evidence of recent whole-genome duplication (WGD) events are shown.

Species (Common Name)	Samples in the SRA Database	Ploidy (Recent WGD)
Saccharum officinarum (sugarcane)	46	8×/10×
Zea mays (maize)	3,514	2× (WGD)
Oryza sativa (rice)	1,264	2×
Triticum aestivum (wheat)	799	6×
Solanum tuberosum (potato)	337	4×
Manihot esculenta (cassava)	61	2×
Glycine max (soybean)	972	2× (WGD)
Beta vulgaris (sugar beet)	32	2×
Solanum lycopersicum (tomato)	830	2×
Hordeum vulgare (barley)	269	2×
Musa acuminata (banana)	73	2×/3× (WGD)
Sorghum bicolor (sorghum)	128	2×
Brassica spp. (field mustard and oilseed rape)	835	2×/4×
Phaseolus vulgaris (common bean)	106	2×
Gossypium hirsutum (cotton)	468	4×
Vitis vinifera (grape)	448	2×

RESULTS

Pipeline for Expression Analysis and Browser Interface

We developed expVIP (Fig. 1), which pseudoaligns and quantifies short reads from RNA-seq experiments to detect and visualize gene expression data through a user-friendly interface. expVIP requires three input files: the RNA-seq reads, a reference transcriptome, and the metadata from the RNA-seq studies. Since the reference transcriptome is user specified, expVIP can facilitate the analysis of RNA-seq data from any species and can use custom reference sequences. expVIP is available in two formats from Github: (1) the source code and (2) a virtual machine implementation that allows easy use of the pipeline and data display from a desktop machine without requiring bioinformatics expertise (see "Materials and Methods").

To illustrate the uses and flexibility of expVIP, we have implemented it to create a wheat gene expression browser (www.wheat-expression.com; Supplemental Text S1), which until now has been lacking in this important crop species. This browser can be used directly with the available wheat expression data, or users can add their own wheat RNA-seq reads to place their data within a wider context of previously published studies. Similar gene expression browsers can be easily developed for any species using the virtual machine or source code.

Global Analysis in Wheat: Validation of Methods

We used expVIP to analyze 16 wheat gene expression studies from the SRA across a range of tissues, developmental stages, and stress conditions (Table II). In total, these included 418 individual samples containing over 11 billion reads, of which 7.4 billion mapped to the reference International Wheat Genome Sequencing Consortium (IWGSC) gene models from EnsemblPlants containing 103,274 genes (Supplemental Table S1). The median number of reads per study was 213 million, with 137 million reads mapped per study.

We found that 99% of genes (102,259) had at least one read mapping to them, and 85% of genes (88,528) were expressed in at least one sample at over 2 transcripts per million (tpm), which has been advocated as the cutoff for real expression over noise (Wagner et al., 2013). Using this cutoff, on average, 34% of genes (35,549) were expressed per sample, with a minimum expression of 11% of genes (10,899) at 20 DPA in the starchy endosperm and a maximum of 48% of genes (50,224) expressed in the spike at anthesis.

We found that, across all samples, there was a weak (adjusted $r^2 = 0.07$), albeit significant ($P = 1.48 \times 10^{-8}$), relationship between the number of mapped reads and the number of genes expressed. This indicates that,

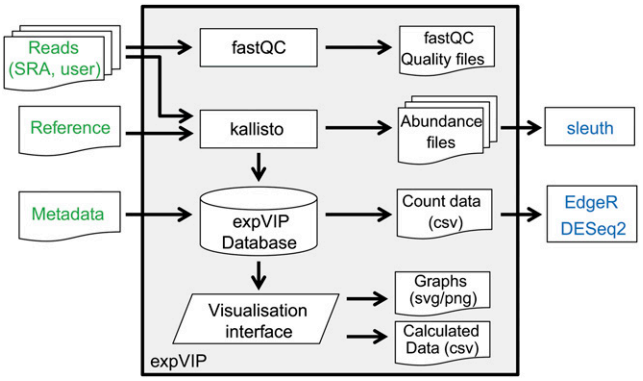

Figure 1. Implementation of expVIP. User inputs are highlighted in green. Downstream differential gene expression analysis (blue) can be performed on expVIP outputs, which are preformatted for this use. External programs are in rectangles, document symbols represent inputs and outputs, the trapezoid represents the visualization interface, and the cylinder represents the expVIP relational database.

although our samples varied widely in their number of mapped reads (1.1–63.6 million), this did not limit comparisons between studies (Supplemental Fig. S1).

We investigated whether, despite coming from diverse studies, tissue-specific expression patterns could be detected. We found that, in general, expression profiles were similar between samples from the same tissue (Fig. 2). For example, grain samples (Fig. 2, red) originating from seven independent studies were found in one main group and leaf and stem samples (Fig. 2, green) from nine studies largely belonged to two groups. However, in some cases, samples from different tissues clustered together, including root samples, which grouped with leaf/stem and spike samples. To further examine the expression patterns of genes in different tissues, we identified the 10 most highly expressed genes in grain and leaves (Supplemental Table S2). We found that, in the grain, six out of the 10 most highly expressed genes encode components of gluten, which is the principal storage protein in wheat grain (Shewry, 2009). In the leaves, several of the most highly expressed genes are related to photosynthesis (Andersson and Backlund, 2008). These results indicate that our data analysis reflects the expected gene expression profiles and supports combining of data from diverse studies.

Accurate Read Mapping Enables Homeologue Specificity

Many crop species are polyploids that contain closely related homeologous genomes, which share highly similar nucleotide sequences within coding regions. This poses a challenge for assigning short reads to the correct gene copy (homeologue). To assess whether kallisto could correctly assign reads to the relevant homeologue, we used a unique genetic resource available in wheat: nullitetrasomic lines (Sears, 1954). Normal bread wheat contains three copies of most genes, one on each of the A, B, and D homeologous chromosomes, and these genes share over 95% identity in coding sequences (Krasileva et al., 2013). In nullitetrasomic lines, one chromosome is specifically deleted (nulli) and compensated by an additional copy of a homeologous chromosome (tetra). Nullitetrasomic lines for chromosome 1 had been sequenced previously (SRP028357), and we used the data in our analysis.

For this analysis, we selected only genes present as three homeologous copies on group 1 chromosomes, with at least one homeologue expressed at over 2 tpm in the wild type (2,645 genes in shoots and 3,445 genes in roots). We compared the expression of genes located on

Table II. *SRA studies analyzed with expVIP*

Study Identifier	Summary	Total Reads	Mapped Reads and Percentage	Reference
DRP000768	Phosphate starvation in roots and shoots	118,053,746	84,529,715 (72%)	Oono et al. (2013)
ERP003465	*Fusarium* head blight-infected spikelets	1,827,362,091	1,357,197,955 (74%)	Kugler et al. (2013)
ERP004505	Grain tissue-specific developmental time course	873,709,556	475,184,621 (54%)	Pfeifer et al. (2014)
SRP004884	Flag leaf down-regulation of *GPC*	209,427,573	121,855,143 (58%)	Cantu et al. (2011)
SRP013449	Grain tissue-specific developmental time course	132,702,451	82,417,257 (62%)	Gillies et al. (2012)
SRP017303	Stripe rust-infected seedlings	33,361,836	13,732,210 (41%)	Cantu et al. (2013)
SRP022869	*Septoria tritici*-infected seedlings	100,582,632	63,155,877 (63%)	Yang et al. (2013)
SRP028357	Shoots and leaves of nullitetra group 1 and group 5	3,304,500,117	2,258,692,000 (68%)	Leach et al. (2014)
SRP029372	Grain tissue-specific developmental time course	101,477,759	17,525,439 (17%)	Li et al. (2013)
SRP038912	Comparison of stamen, pistil, and pistilloidy expression	217,315,378	153,009,134 (70%)	Yang et al. (2015)
SRP041017	Stripe rust and powdery mildew infection time course	395,463,786	272,228,560 (69%)	Zhang et al. (2014)
SRP041022	Developmental time course of synthetic hexaploid	134,641,113	84,583,556 (63%)	Li et al. (2014)
ERP008767	Grain tissue-specific expression at 12 DPA	45,213,827	26,420,708 (58%)	Pearce et al. (2015)
SRP045409	Drought and heat stress time course in seedlings	921,578,806	533,928,182 (58%)	Liu et al. (2015)
ERP004714	Developmental time course of cv Chinese Spring	1,536,051,415	1,066,712,760 (69%)	Choulet et al. (2014)
SRP056412	Grain developmental time course with the 4A dormancy quantitative trait locus	1,875,916,011	808,809,053 (43%)	Barrero et al. (2015)

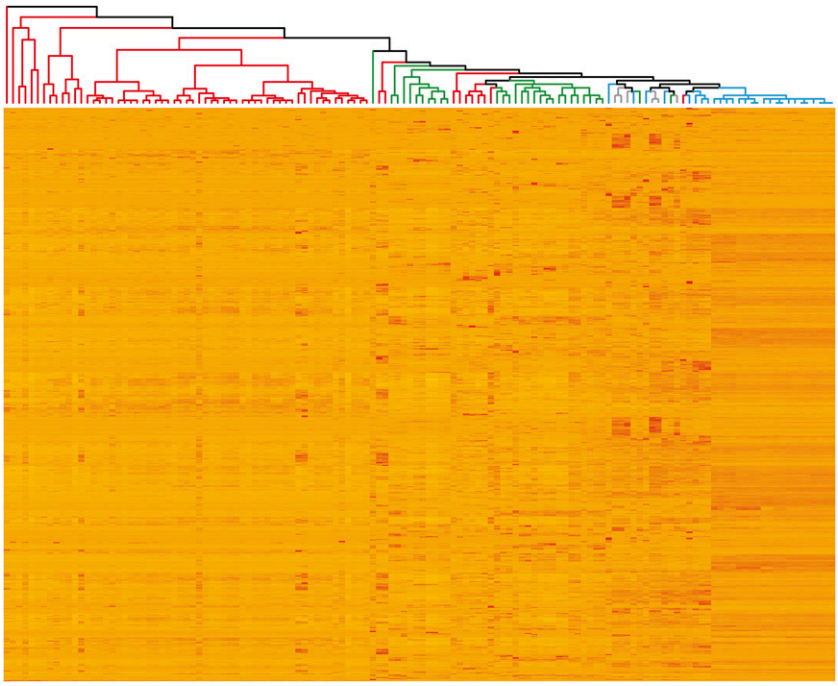

Figure 2. Similarity of expression profiles between samples (columns), with replicate samples averaged and excluding samples from nullitetrasomic lines. One thousand randomly selected genes are represented, one gene per row. Only genes expressed in at least one sample over 2 tpm were used. Colors on the dendrogram indicate the tissues from which samples originate: grain (red), spike excluding grain (blue), leaves/stem (green), and roots (gray).

chromosomes 1A, 1B, and 1D between wild-type and nullitetrasomic lines (Fig. 3). In wild-type plants, average gene expression was quite even between the three homeologous genomes (36.6%, 30.9%, and 32.5% for A, B, and D in shoots and 33.4%, 32.3%, and 34.4% for A, B, and D in roots). Similarly, in nullitetrasomic lines for the homeologue, which was present with two copies (as in the wild type), expression was 34% of total in shoots and 33.8% in roots. In contrast, expression of the homeologue that was deleted in the nullitetrasomic

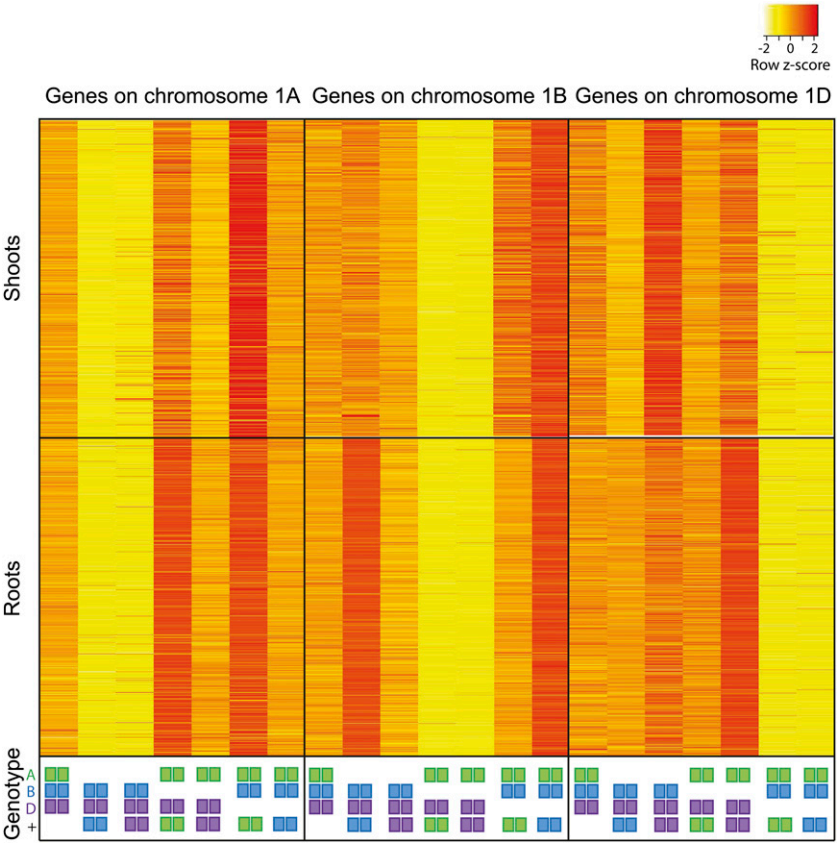

Figure 3. Expression of genes with three homeologous copies on chromosome 1 in nullitetrasomic wheat lines in shoots and roots. Genotypes for chromosome 1 are indicated by colored squares: A genome in green, B genome in blue, and D genome in purple. Squares listed at bottom (+) indicate extra copies (tetra); the absence of squares indicates deletion (nulli) of the entire chromosome.

lines was strongly decreased to 5.9% and 5.3% of total in shoots and roots, respectively. Expression of the homeologue that was present with four copies (2× the wild type) rose to 60.1% and 60.9% of total for shoots and roots, respectively. These results demonstrate that, even in the extreme case where expression from one homeologue has been abolished completely by chromosomal deletion, our pipeline can accurately distinguish from which homeologue gene expression originated. Analysis of a manually curated set of 52 tetraploid wheat homeologues showed that they share 97.3% ± 1.2% DNA sequence identity and that the distance between adjacent variants decreases exponentially, with an average separation of approximately 38 bp. This determines that 8% of single-nucleotide polymorphisms (SNPs) between A and B genome homeologous are over 100 bp apart (Krasileva et al., 2013). This would prevent reads containing these widely spaced SNPs from being unambiguously mapped to one homeologue, explaining why we observe a residual level of expression from the deleted chromosome in the nullitetrasomic lines.

Comparison of kallisto with bowtie2 Combined with eXpress

Since kallisto is a newly released pseudoalignment tool for the quantification of RNA-seq data, we compared its performance with a more conventional RNA-seq quantification pipeline using bowtie2 and eXpress. We found that kallisto and bowtie2 had very similar overall alignment rates (62.7% and 63.4%, respectively; Supplemental Table S1). kallisto identified slightly more genes as expressed in at least one sample at over 2 tpm: 88,528 compared with the 87,842 genes identified by bowtie2 + eXpress. As an assessment of accuracy, we compared their performance using the nullitetrasomic wheat lines described previously. We found that kallisto was slightly more accurate than bowtie2 + eXpress: on average, kallisto assigned 5.6% of total gene expression to have originated from the deleted chromosome, whereas bowtie2 + eXpress assigned 7% of total gene expression (Supplemental Table S3). These results support the use of kallisto, given its fast running times and high accuracy (Bray et al., 2015).

Powerful Visualization and Data Integration Platform

expVIP is highly flexible, as it allows the user to supply metadata to classify samples according to different categories (based on their biological question), which are then uploaded into the database. The visualization interface allows users to group, filter, sort, and download their data according to the categories specified in the metadata. This design provides control over precise categories to be used in the database, and the visualization interface will adjust accordingly. For example, we classified the expression data at www.wheat-expression.com by broad and specific categories for age, tissue, disease/abiotic stress, and variety (Supplemental Tables S1 and S4). This hierarchical structure allows users to group data for an initial high-level assessment and then open up data into specific samples, analogous to main effects and simple effects in statistical analyses (Supplemental Table S4). This structure can be modified as required by users by simply modifying the metadata input file or by providing a different nomenclature for classification, such as Plant Ontology temporal and anatomy accession identifiers (Avraham et al., 2008). We describe below how this visualization interface can be used to facilitate research.

Candidate Gene Function Prediction

Fine-mapping frequently results in a candidate gene list within a defined genetic interval. Understanding gene expression patterns can help narrow down this list but typically requires the development of homeologue-specific quantitative PCR (qPCR) primers, which is challenging and time consuming in polyploids. Using the wheat expression browser, we are now able to rapidly investigate in silico candidate gene expression patterns.

For example, a physical contig containing seven candidate genes for grain preharvest sprouting resistance was published recently (Barrero et al., 2015). Therefore, we organized and sorted the data based on the tissue origin of the RNA-seq sample. We displayed the expression data for the six candidate genes in this region with genome annotation either as a heat map (Fig. 4A) or individual bar graphs (Fig. 4B). We find that one gene is expressed at very low levels below 2 tpm in all tissues (*Traes_4AL_DD1B27086.2*) and that three genes are most highly expressed in roots (*Traes_4AL_9A01E952D.1*, *Traes_4AL_1C557F688.1*, and *Traes_4AL_65DF744B71.3*), with very little expression in the grain, where genes involved in precocious germination would be expected. Two closely related genes show expression solely in the grain: *Traes_4AL_BFAB568BF.1* and *Traes_4AL_F99FCB25F.1*, with the latter having much higher expression.

To further define the expression patterns, we displayed the age and specific tissue of the samples. This filtering and dynamic sorting is available in both heat map and bar graph modes. Focusing on *Traes_4AL_F99FCB25F.1* displayed as a bar graph (Fig. 5A), we see that this gene is most highly expressed during the latter stages of grain development, consistent with a role in grain dormancy imposition, and that expression is strongest in whole grain and mostly absent in seed coat and endosperm tissues (Fig. 5B), suggesting that expression might originate from the embryo. The color code of the graph dynamically alters to reflect the most recent category selected by the user. The two candidate genes highlighted by this analysis (*Traes_4AL_BFAB568BF.1* and *Traes_4AL_F99FCB25F.1*) were recently shown to act as positive regulators of dormancy (Barrero et al., 2015).

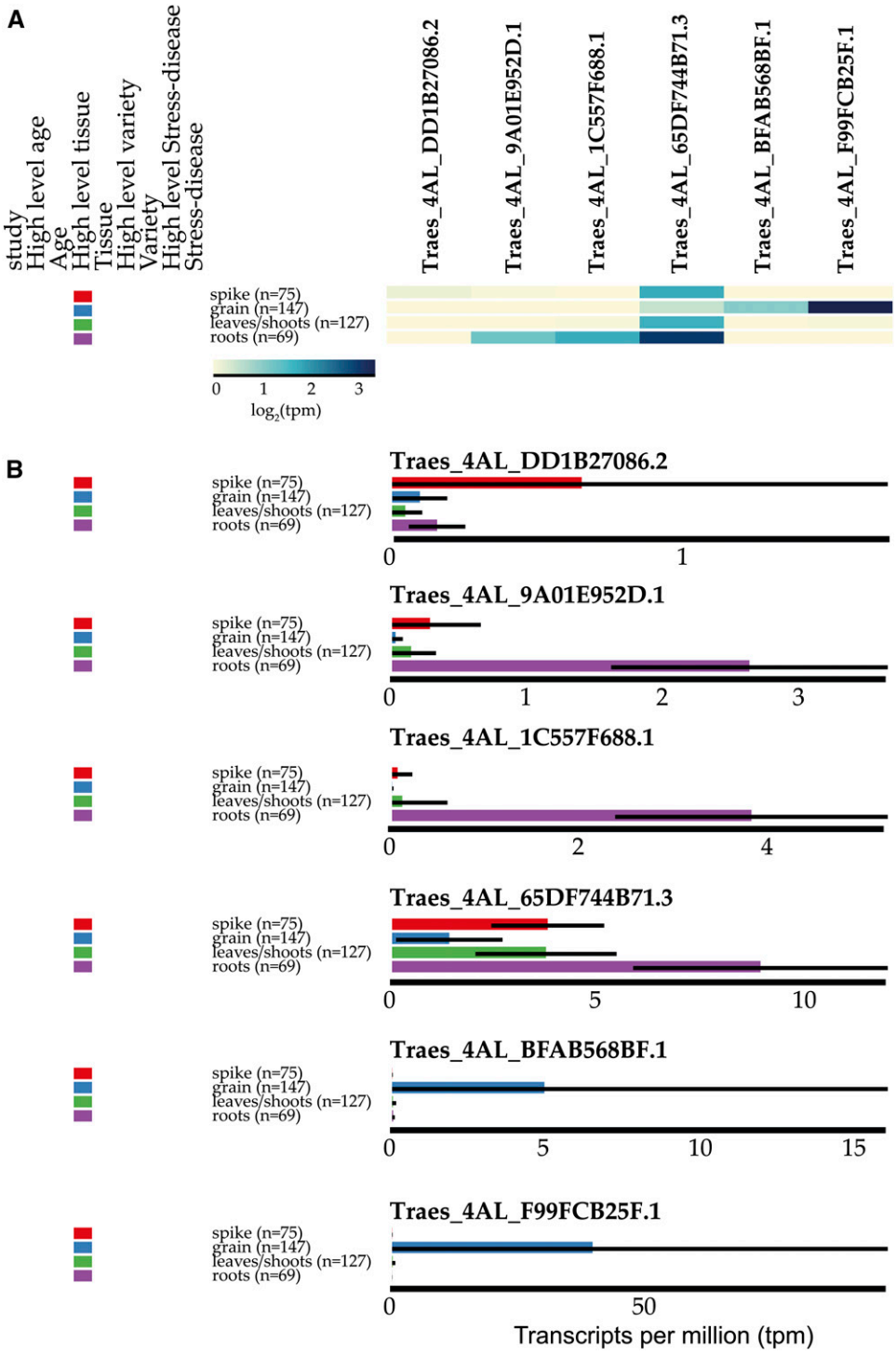

Figure 4. A simple search on www.wheat-expression.com reveals gene expression patterns of six candidate genes within a quantitative trait locus region for preharvest sprouting. The data may be displayed as a heat map for all six genes simultaneously (A), with the intensity of the blue color indicating the expression level [$\log_2$(tpm)]. Alternatively, each gene may be displayed individually as a bar graph (B) in tpm. The display was configured to average data according to the high-level tissue; hence, all samples coming from spike (red), grain (blue), leaves/shoots (green), and roots (purple) are averaged according to their respective categories. Genes are ordered from lowest expressed (left [A] and top [B]) to highest expressed (right [A] and bottom [B]). Note that axes in B are not equal because expVIP recalculates the axis for each gene individually.

Identification of Stable Reference Genes

To compare gene expression levels, a widely used method is qPCR, which requires stably expressed reference genes across all samples being compared. The integrated data available from expVIP allow quick analysis to identify potential novel reference genes. To identify reference genes suitable for wheat across diverse tissues, developmental stages, and stress and disease conditions, we included 321 out of the total 418 wheat samples included at www.wheat-expression.com (we excluded 97 samples that were from nullitetrasomic samples to avoid bias against the missing chromosomes in those samples). We found that 3,170 genes were expressed at over 2 tpm in all 321 samples. We calculated the coefficient of variation as a measure of the stability of expression across all samples. These varied from

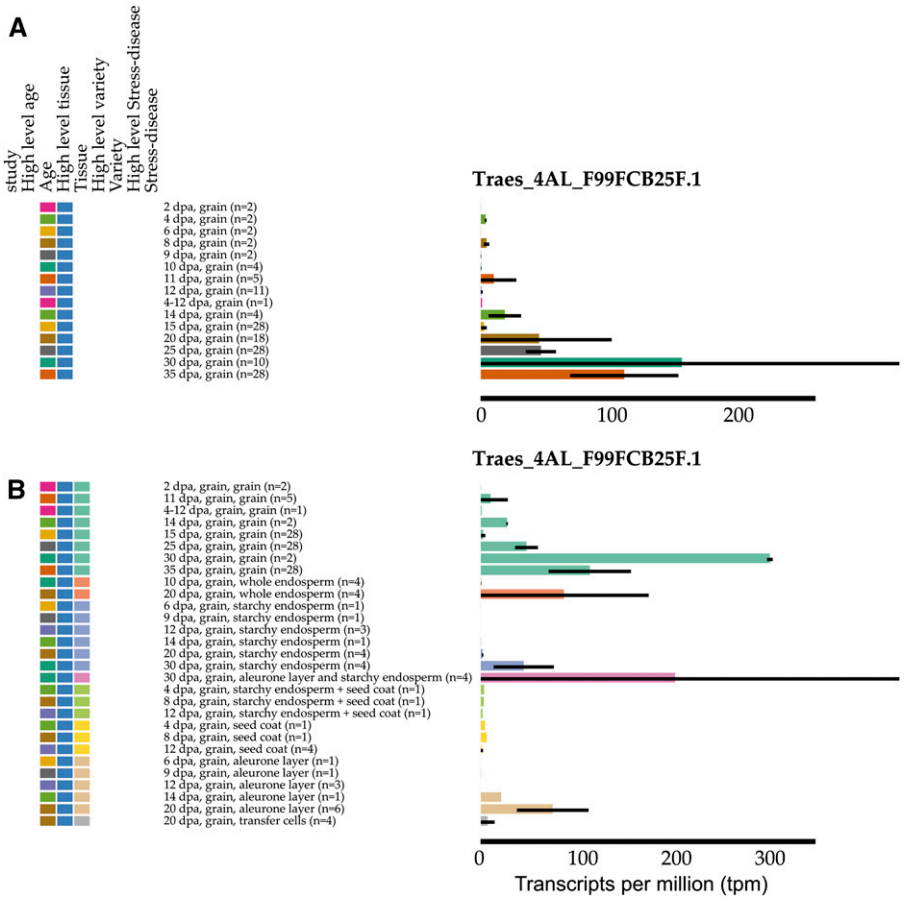

Figure 5. Expression of *Traes_4AL_F99FCB25F.1* in grains categorized by age (A) and age and tissue (B). The colors represent age (A) or tissue (B): the color coding of the graph is determined by the most recent category clicked by the user.

32.7% for the most stable gene to 318% for the least stable gene, with the median coefficient of variation being 61.6% (Fig. 6A; Supplemental Table S5). We investigated whether genes commonly used as reference genes in qPCR were stably expressed in our samples. We found that 1,736 genes were more stably expressed than 13 commonly used reference genes (Yan et al., 2003; Tenea et al., 2011; Qi et al., 2012; Fig. 6A; Supplemental Table S6), seven of which were not expressed in all samples at over 2 tpm (Fig. 6A). We selected the 20 most stable genes (Fig. 6B) and found a much narrower range of variation in expression levels compared with the commonly used reference genes (Fig. 6C). These stably expressed genes had a range of different functions, including ubiquitin-mediated protein degradation, DNA binding, and signal transduction (Table III).

To test whether these newly identified stable genes could be used in qPCR as reference genes, we designed homeologue-specific primers for five genes. The efficiencies ranged from 93.3% to 97.1% (Table IV). To test the stability of these primers, we extracted RNA and synthesized complementary DNA (cDNA) from a diverse range of 30 conditions (Supplemental Table S7), including various tissues, developmental stages, varieties, and disease/stress conditions. We found that all five genes had low coefficients of variation using qPCR (4.4%–8.4%), suggesting that they are suitable for use as reference genes (Table IV). We found that the coefficients of variation measured by qPCR were lower than those found by RNA-seq analysis. This may be due to the qPCR using a smaller panel of samples (30 conditions) compared with the 321 samples included in the RNA-seq analysis. Furthermore, the qPCR analysis used more homogeneous sample extraction methods than the RNA-seq samples, which were from a diverse range of studies carried out in different laboratories, which might have introduced extra variability.

The five novel genes tested had equivalent stability to five of the most stable commonly used reference genes across the 30 conditions tested (6.8% ± 1.7% and 6.4% ± 1.4%, respectively; Supplemental Table S8). The commonly used reference genes were originally identified in flag leaves (Tenea et al., 2011) and had lower coefficients of variation (3% ± 1%) than the novel genes (5.5% ± 2.3%) in this tissue. However, in the grain, the novel reference genes had much lower coefficients of variation (2.7% ± 0.5%) than the commonly used reference genes (6.6% ± 2.5%), indicating that, under specific sets of conditions, these novel reference genes outperform current reference genes. The strong stability in grain samples may reflect the origin of samples used to identify the novel reference genes: 147 out of the 321 samples used originated from grains. These results indicate that the expVIP platform can help to identify

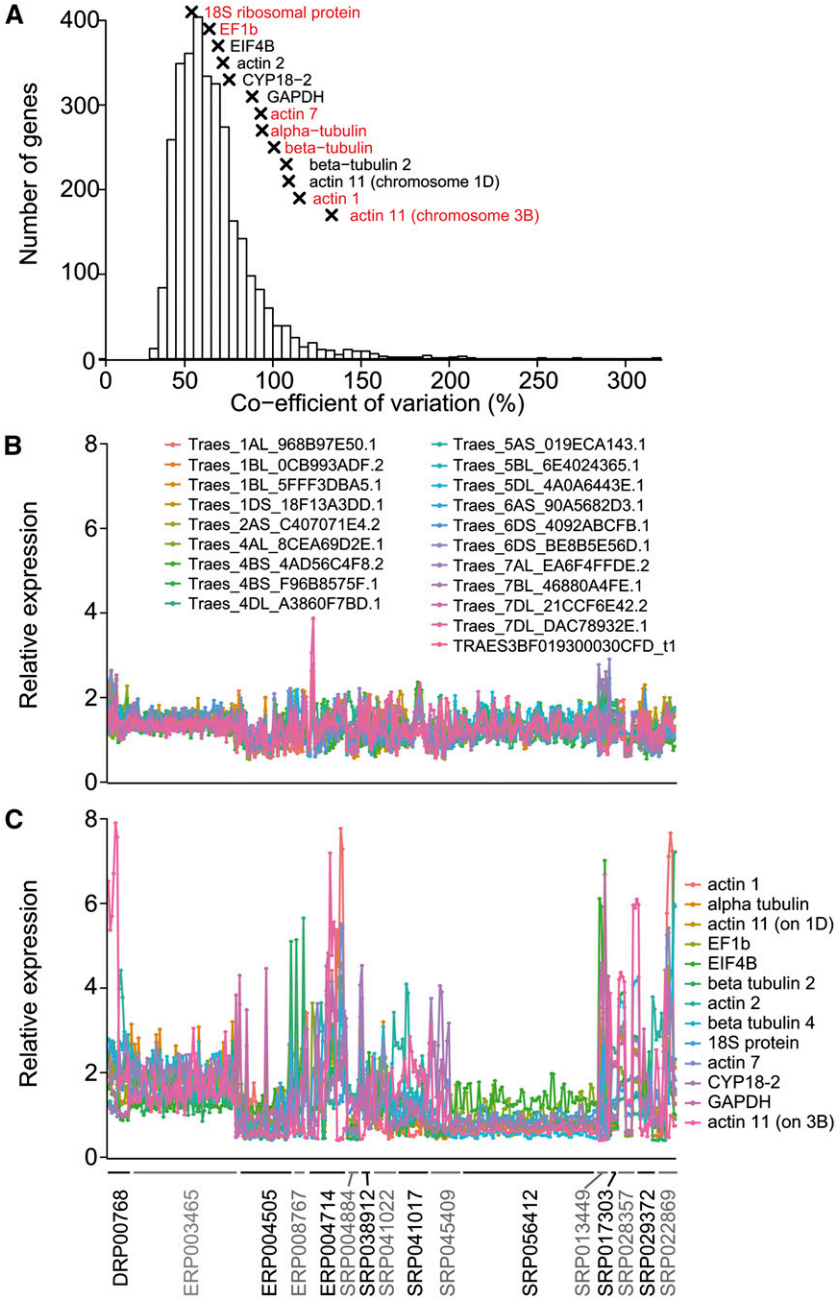

Figure 6. Stability of gene expression between samples. A, Coefficient of variation for genes that are expressed at over 2 tpm in all samples. Commonly used reference genes are indicated by crosses (x), and reference genes in red are not expressed at over 2 tpm in all samples. B and C, Expression of the 20 most stably expressed genes (B) and 13 commonly used reference genes (C) across 321 wheat samples belonging to 16 studies indicated on the *x* axis. The expression level of each gene in a sample is relative to the average expression level of this gene across all samples. Abbreviations are as follows: elongation factor 1-β (EF1b), eukaryotic translation initiation factor 4B (EIF4B), cylophilin A (CYP18-2), and glyceraldehyde 3-phosphate dehydrogenase (GAPDH).

stably expressed genes for use in qPCR, which can be tailored to individual needs either across different tissues or focusing on a particular tissue of interest.

Comparative Analyses to Generate Novel Biological Insights

expVIP allows easy integration of data for differential gene expression analysis. Using the output from kallisto, we used its companion tool sleuth (Pimentel et al., 2015) to identify genes that were differentially expressed in disease and stress conditions compared with control conditions. For this analysis, we included all samples from seedling stage wheat leaves that had replicates. These included two different SRA studies, which comprised samples from 12 different conditions (Table V; for details, see Supplemental Table S1).

In order to find genes that are differentially expressed in multiple conditions, we used a relaxed threshold to identify differentially expressed genes ($q < 0.05$). In total, 53% of genes (54,207 genes) were differentially expressed in at least one stress condition compared with the control. The number of differentially expressed genes varied from 2,018 genes after 48 h of stripe rust infection to 34,221 genes after 6 h of combined drought and heat stress (Fig. 7A). In general, the abiotic stresses caused more genes to be differentially expressed than

Table III. *Twenty most stably expressed genes across all 321 wheat samples*

Ensembl Transcript Identifier	Mean Expression Level (tpm)	Coefficient of Variation (%)	Putative Function[a]
Traes_1DS_18F13A3DD.1	13	33	RING zinc finger domain superfamily protein
Traes_5AS_019ECA143.1[b]	13	33	Ion channel
Traes_7BL_46880A4FE.1	8	33	Ser/Thr protein kinase
Traes_6DS_4092ABCFB.1	7	34	Uncharacterized protein
Traes_6DS_BE8B5E56D.1[b]	24	34	Ser/Thr protein kinase
Traes_6AS_90A5682D3.1	21	34	Ser/Thr protein kinase
Traes_1AL_968B97E50.1[b]	15	34	ATP-dependent zinc metalloprotease FTSH8
Traes_2AS_C407071E4.2	9	34	WRKY family transcription factor family protein
Traes_4BS_F96B8575F.1	6	34	Uncharacterized protein
Traes_4DL_A3860F7BD.1	9	35	DEAD box ATP-dependent RNA helicase38
Traes_1BL_0CB993ADF.2	10	35	VHS and GAT domain-containing protein
Traes_7DL_DAC78932E.1	9	35	DGCR14-related
Traes_7DL_21CCF6E42.2	9	35	GCIP-interacting family protein
TRAES3BF019300030CFD_t1	14	35	Uncharacterized protein
Traes_1BL_5FFF3DBA5.1	15	35	Ubiquitin family protein
Traes_5DL_4A0A6443E.1	12	35	Uncharacterized protein
Traes_4AL_8CEA69D2E.1[b]	31	35	Ubiquitin-conjugating enzyme
Traes_7AL_EA6F4FFDE.2	13	35	Zinc finger protein
Traes_4BS_4AD56C4F8.2[b]	13	36	Uncharacterized protein
Traes_5BL_6E4024365.1	9	36	Gal oxidase/Kelch repeat superfamily protein

[a]For genes that were not annotated in wheat, putative functions were assigned by orthology to rice, maize, and Arabidopsis genes according to EnsemblPlants.

[b]Gene stability tested by qPCR.

under disease conditions (on average, 27,212 compared with 6,429 genes), and in abiotic stress, more genes were up-regulated than down-regulated, whereas the reverse pattern was observed in disease conditions.

We found that the majority of genes were differentially expressed in multiple conditions (Fig. 7B), indicating that transcriptional responses to different stresses are shared. Comparing between abiotic and disease stress, we found that 38% (20,553 genes) of differentially expressed genes were found in both cases. We detected enrichment for 32 Gene Ontology (GO) terms among the genes differentially expressed in 10 or more abiotic and disease conditions (false discovery rate [FDR] < 0.05; Supplemental Tables S9 and S10). Nineteen of these related to biological processes rather than molecular function or cellular compartment (Table VI). The two most strongly enriched GO terms (GO:0018298 and GO:0009765) were related to chlorophyll *a*/*b*-binding proteins, whereas the third most strongly enriched GO term (GO:0006457; protein folding) included three HSP90 family heat shock proteins, three calreticulin/calnexin proteins, and three cyclophilin-type peptidyl-prolyl cis-trans-isomerase domain-containing proteins. Evidence was also found for the regulation of gene expression, and 14 transcription factors were differentially expressed across 10 or more conditions, including members of the NAC, MYB, basic-Leu zipper, zinc finger, and AP2/ERF families. Many of these large gene families have been shown in plants to be involved in abiotic and biotic stress responses (Singh et al., 2002;

Table IV. *Homeologue-specific primers designed for five of the most stably expressed genes identified from 321 wheat samples*

The stability of the expression of these five genes was tested across 30 independent conditions, including different tissues, developmental stages, varieties, and disease infection (for details, see Supplemental Table S7).

Ensembl Transcript Identifier	Primer Sequences (5′–3′)	Primer Efficiency (%)	Coefficient of Variation (%)
Traes_4AL_8CEA69D2E.1	CGGGCCCGAAGAGAGTCT ATTAACGAAACCAATCGACGGA	97.1	7.1
Traes_4BS_4AD56C4F8.2	TCGTTGCTTGAGGAAAATG CATGACCGTCTTATTTATGGCA	93.7	8.2
Traes_1AL_968B97E50.1	TTTGCACAGTATGTACCAAATGAG TCTTCCAATCAAAACCTCCTCT	95.0	5.8
Traes_5AS_019ECA143.1	TCTAAATGTCCAGGAAGCTGTTA CCTGTGGTGCCCAACTATT	96.0	4.4
Traes_6DS_BE8B5E56D.1	CATGCTCTGGGATTTATCCAT CTGGATCATTTCCGGTGC	93.3	8.4

Table V. *Samples used to compare gene expression responses to abiotic and biotic stresses*

Study	Age	Conditions	Replicates
SRP041017	7 d	Stripe rust, 24 h	3
		Stripe rust, 48 h	3
		Stripe rust, 72 h	3
		Powdery mildew, 24 h	3
		Powdery mildew, 48 h	3
		Powdery mildew, 72 h	3
SRP045409	7 d	Drought stress, 1 h	2
		Drought stress, 6 h	2
		Heat stress, 1 h	2
		Heat stress, 6 h	2
		Drought and heat stress, 1 h	2
		Drought and heat stress, 6 h	2

Feller et al., 2011; Nakashima et al., 2012), but this joint analysis identified precise candidates in wheat based on available experimental data that can be further characterized.

We identified nine genes that were differentially expressed in all 12 conditions. Examining the expression of these genes in the wheat expression browser gives further insight into their expression patterns across all 16 studies. For example, the ortholog of the endosomal targeting *BRO1* gene *Traes_2AL_2DFED03C9.2* is strongly up-regulated in abiotic stress conditions (Fig. 8A, purple bar), and opening up the data to look into individual stresses, we find that it is not up-regulated in phosphorous starvation (Fig. 8B, purple bars labeled P-10d). *Traes_2AL_2DFED03C9.2* is down-regulated in the majority of disease conditions (Fig. 8B, yellow bars), except in the spike infected with *Fusarium graminearum* (Fig. 8B, yellow bars labeled fu30h–fu50h) and after 6 d of stripe rust infection (Fig. 8B, yellow bars labeled sr6+d). This visualization also shows that *Traes_2AL_2DFED03C9.2* is expressed in all tissues (roots, leaves/stems, spikes, and grains) and is not restricted to seedling leaves, the tissue from which it was identified by our analysis. Selecting the homeologue option allows the expression of homeologous genes to be examined side by side (Fig. 8C). In this case, all three homeologues show a similar pattern of expression in the various samples, and all three homeologues are differentially expressed in 11 or 12 abiotic stress and disease conditions. The expVIP visual interface also allows individual studies to be selected; in this case, the two original studies also can be displayed on their own to visualize the differences identified by sleuth (Supplemental Fig. S2).

DISCUSSION

Highly Accurate Pipeline

A major challenge in the analysis of RNA-seq data, particularly in polyploid crop species, is the assignment of short reads to the correct copy of a gene. Using nullitetrasomic wheat lines, we have shown that kallisto as implemented through expVIP accurately assigns reads to the correct homeologue. The visualization interface makes expression data across a wide range of conditions easily available, enabling researchers and breeders to rapidly check the expression patterns of individual homeologues. This will allow a more precise understanding of gene regulation beyond the broad general trends usually reported in wheat with non-homeologue-specific qPCR primers. The ability to query homeologue-specific expression data will also complement growing knowledge about sequence diversity between homeologues. A recent genome-wide analysis between landraces and elite varieties suggested that, during domestication, positive selection was usually restricted to an advantageous mutation within a single homeologue (Jordan et al., 2015). This highlights that understanding of homeologue-specific variation in both sequence and expression will be fundamental for future advances in wheat improvement.

Utility for Functional Genomic Research in Wheat

Until recently, marker availability had been a major constraint in wheat research; however, developments

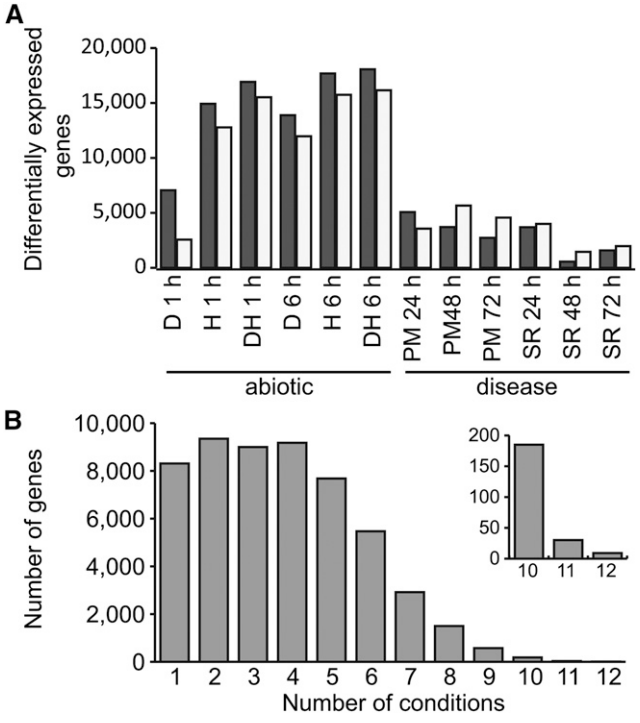

Figure 7. Differentially expressed genes ($q < 0.05$) in abiotic stress and disease conditions. A, Numbers of up-regulated genes (black bars) and down-regulated genes (gray bars) in individual stress conditions. D, Drought; H, heat; DH, drought and heat combined; PM, powdery mildew; SR, stripe rust. B, Number of genes that are differentially expressed in multiple abiotic stress and disease conditions.

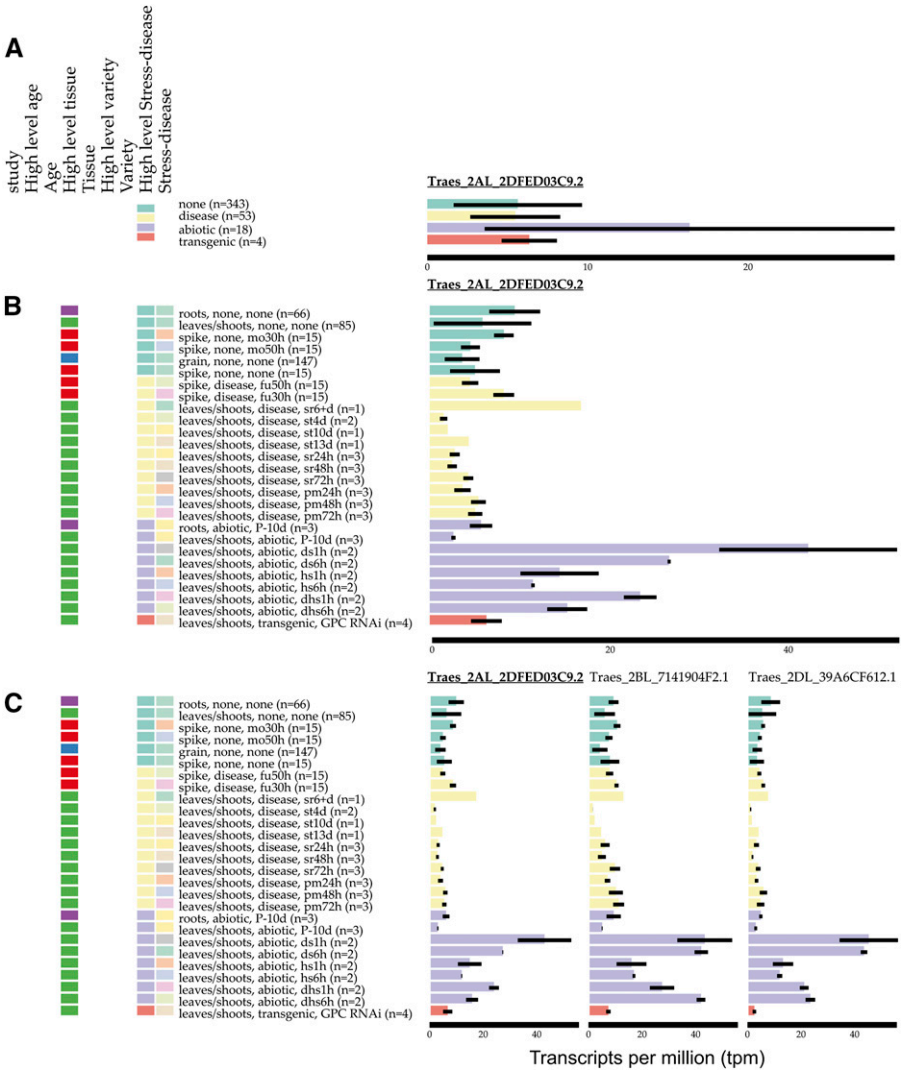

Figure 8. Example of gene expression visualization using expVIP for the gene *Traes_2AL_2DFED03C9.2*, with samples grouped according to their High level stress-disease (A), *Traes_2AL_2DFED03C9.2*, with additional categorization of samples including lower level Stress-disease and High level tissue (B), and *Traes_2AL_2DFED03C9.2* and its B and D homeologues, which are differentially expressed in 11 and 12 abiotic and disease conditions, respectively (C). The data shown here include expression data from all studies, not just the studies examined for differential expression. Samples are ordered by their High level stress-disease status: none (green), disease (yellow), abiotic (purple), and transgenic (orange).

in SNP- and sequence-based genotyping have removed these limitations (Borrill et al., 2015). The focus has now shifted toward the understanding of gene function, which is being accelerated by the availability of a draft reference genome (International Wheat Genome Sequencing Consortium, 2014) and next-generation sequencing-enabled mapping approaches (Ramirez-Gonzalez et al., 2015). The availability of a comprehensive gene expression visualization platform in wheat will facilitate the functional characterization of genes by providing researchers with information regarding where they might be acting. We have demonstrated that the expression browser rapidly delivers information about tissue-specific expression patterns and can help narrow down candidate genes within mapping intervals through both heat-map and single-gene analyses. Furthermore, we have used these data to propose genes with high stability across a wide range of conditions that might represent better reference genes for qPCR than those traditionally used, particularly in grains.

Opportunities for Meta-Analysis

Using the data generated by expVIP for wheat, we compared between samples from a diverse range of abiotic stress and disease conditions, leveraging the unified analysis platform. We found that slightly more genes were up-regulated than down-regulated in abiotic stresses, whereas in disease conditions, the opposite pattern was observed: this contrasts with a previous meta-analysis of rice abiotic and biotic stress microarray experiments, where 60% of differentially express genes were down-regulated under abiotic stress and 60% of differentially expressed genes were up-regulated under biotic stress (Shaik and Ramakrishna, 2014). These results may be different because the rice analysis included additional stress conditions that might have influenced overall trends, microarrays having an incomplete gene complement, or biological differences between species. expVIP will facilitate the meta-analysis of RNA-seq experiments, which has been difficult so far due to non-unified methods of analysis, in contrast to microarray

Table VI. *Enriched biological processes in genes differentially expressed in 10 or more abiotic and disease conditions*

GO Accession No.	Term	Percentage of Differentially Expressed Genes	Percentage of Transcriptome	FDR
GO:0018298	Protein-chromophore linkage	4.0	0.1	3.00E-09
GO:0009765	Photosynthesis, light harvesting	4.0	0.2	5.20E-05
GO:0006457	Protein folding	5.2	0.7	0.0011
GO:0009651	Response to salt stress	2.9	0.2	0.0041
GO:0006970	Response to osmotic stress	2.9	0.2	0.0066
GO:0065007	Biological regulation	17.8	8.2	0.014
GO:0065008	Regulation of biological quality	5.7	1.5	0.034
GO:0045449	Regulation of transcription	10.3	4.1	0.041
GO:0009889	Regulation of biosynthetic process	10.3	4.3	0.044
GO:0010556	Regulation of macromolecule biosynthetic process	10.3	4.3	0.044
GO:0031326	Regulation of cellular biosynthetic process	10.3	4.3	0.044
GO:0019219	Regulation of nucleobase, nucleoside, nucleotide, and nucleic acid metabolic process	10.3	4.3	0.044
GO:0051171	Regulation of nitrogen compound metabolic process	10.3	4.3	0.044
GO:0080090	Regulation of primary metabolic process	10.3	4.5	0.048
GO:0006355	Regulation of transcription, DNA dependent	9.8	4.1	0.048
GO:0030001	Metal ion transport	4.0	0.9	0.048
GO:0051252	Regulation of RNA metabolic process	9.8	4.1	0.048
GO:0009628	Response to abiotic stimulus	4.6	1.2	0.048
GO:0010468	Regulation of gene expression	10.3	4.5	0.048

experiments, which have been better catalogued and compared (Zimmermann et al., 2004; Parkinson et al., 2007; Wagner et al., 2013). Although differences were seen between abiotic and disease transcriptional responses, 38% of differentially expressed genes were identified in both abiotic and disease conditions, which is similar to the proportion identified in a comparison of gene expression in rice of drought and bacterial responses (39% shared genes; Shaik and Ramakrishna, 2013).

The majority of genes differentially expressed in 10 or more stress conditions did not show the same direction of expression change in all stresses. For example, three homeologues of an endosome-targeting *BRO1* gene (*Traes_2AL_2DFED03C9.2*, *Traes_2BL_7141904F2.1*, and *Traes_2DL_39A6CF612.1*) were up-regulated in abiotic stresses and down-regulated in disease conditions. Manipulating endosomal trafficking by overexpressing a RAB5 GTPase in Arabidopsis (*Arabidopsis thaliana*) enhanced salt-stress tolerance (Ebine et al., 2012), and endocytic trafficking is also known to be important for disease resistance (Teh and Hofius, 2014), indicating that *BRO1* represents a candidate gene to manipulate abiotic stress and disease responses. Several transcription factors from diverse families are also up- and down-regulated in stress conditions; for example, the NAC transcription factor *Traes_5BL_4497A137C.1* is up-regulated in response to abiotic stress and during early stripe rust infection but down-regulated later during stripe rust and powdery mildew infection. Analogously, the basic helix-loop-helix (bHLH) transcription factor *Traes_5DL_2A286B481.1* is up-regulated during the first 1 h of drought, heat, and drought combined with heat stress, but after 6 h in all three conditions it is down-regulated, suggesting a specific temporal role. In Arabidopsis, *bHLH92*, the ortholog of *Traes_5DL_2A286B481.1*, is also induced by abiotic stresses, but its up-regulation is maintained at both 6 and 24 h (Jiang et al., 2009). The ability to combine studies from multiple environmental conditions will allow novel hypothesis generation to deepen our understanding of conserved and divergent responses to abiotic and biotic stresses.

Application to a Range of Species

We demonstrate that expVIP can be used to reanalyze studies using a common reference, allowing accurate and easy comparison between data from different sources. We applied our pipeline to polyploid wheat and generated an open-access expression browser (www.wheat-expression.com). However, the expVIP pipeline and browser interface can be implemented readily into other species to facilitate functional gene characterization. This is especially relevant given the speed with which genomics is progressing: the best reference genomes and transcriptomes change constantly, making it difficult to compare between RNA-seq studies that have used different references. This problem is also exemplified in more mature systems such as rice, where two different genome annotations are widely used: Rice Annotation Project gene models and Michigan State University gene models (Ohyanagi et al., 2006; Ouyang et al., 2007). Although these annotations share many similar genes, they cannot be compared directly. expVIP facilitates the rapid reanalysis of data sets that were originally evaluated with different reference sequences to enable such comparisons on a common set of gene models (Supplemental Text S2).

The flexible expVIP metadata structure can accommodate formal ontologies such as Plant Ontology accession identifiers, which can be linked through established parent-child relationships. This is immediately possible for the temporal and anatomical components of ontologies that are well described and documented (Avraham et al., 2008). However, although ontologies for stress treatments (abiotic and biotic) have been proposed (Walls et al., 2012), they are not commonly implemented. Looking forward, the use of a common platform such as expVIP to analyze RNA-seq data from multiple species will facilitate cross-species comparisons of gene expression between orthologs. Orthologous relationships between genes for multiple plant species are well established (Rouard et al., 2011; Goodstein et al., 2012; Bolser et al., 2015), and they will become increasingly precise as additional genomes are sequenced. This would allow the inclusion of an additional species category within the visualization interface to compare the expression of orthologs across multiple species. However, this will require the research community to improve and engage more actively with the use of ontologies to describe the origin of diverse RNA-seq samples.

The availability of expVIP as a virtual machine will facilitate its application to any species with a transcriptome reference. expVIP is based on the lightweight pseudoaligner kallisto (which we have shown to perform as well if not more accurately that bowtie2 + eXpress), which will allow rapid analysis on a desktop machine without the need for bioinformatics infrastructure. This opens up intuitive and interactive data visualization of gene expression data to researchers using both unpublished and publicly available data.

CONCLUSION

The pipeline and visualization interface we have developed will open up the analysis of gene expression data from a wide variety of species to researchers and breeders. Our application to wheat gene expression data provides a community resource that will aid the functional analysis of wheat genes for their use in research and breeding programs. Moving into the future, the volume of RNA-seq expression data will only increase, and the value from reanalysis and integration of data cannot be underestimated. This is especially relevant given the frequent release of improved reference genomes, which, while welcomed, poses a challenge when comparing RNA-seq data that have been aligned to previous releases. This open-access platform makes a first step toward enabling the easy integration, visualization, and comparison of RNA-seq data across experiments.

MATERIALS AND METHODS

Data Preparation

Reads

We downloaded the wheat (*Triticum aestivum*) gene expression data from the SRA database at NCBI available on August 12, 2015. Study ERP004714 was incomplete and missing the required metadata in the SRA, so the data were downloaded directly from https://urgi.versailles.inra.fr/files/RNASeqWheat/. For consistency of analysis, we only included data sets generated using RNA-seq on the Illumina platform, both paired and single-end reads. We excluded small RNA studies and studies with fewer than 50 million total reads. The SRA studies included in this analysis are listed in Table II with a short description (full details are given in Supplemental Table S1).

Reference

The wheat transcriptome reference was downloaded from EnsemblPlants release 26 (Choulet et al., 2014; International Wheat Genome Sequencing Consortium, 2014).

Metadata

Experiment metadata were downloaded from the SRA and supplemented by manual curation from the associated publications. This manual curation was used to define the factors that were used for the classification of studies in the visualization interface. For the wheat expression browser, we defined factors as study, age, tissue, variety, and stress-disease treatment. These factors were grouped at a high level and also at the individual level to allow more meaningful comparisons (Supplemental Table S4). The homeologues of each gene were extracted from EnsemblCompara release 26 (Vilella et al., 2009) and added as metadata to the genes. Detailed documentation on how to load metadata into expVIP is available online (https://github.com/homonecloco/expvip-web/wiki).

Expression Analysis

We implemented an initial sample quality control using fastQC (version 0.10.1; Andrews, 2010), which reports the fastQC quality files for the user to assess. Wheat gene expression quantification was carried out using kallisto version 0.42.3 (Bray et al., 2015) and the wheat transcriptome described previously. For paired-end reads, kallisto was run using default parameters with 100 bootstraps (-b 100). For single-end reads, kallisto was run using 100 bootstraps (-b 100) in the single-end read mode (–single), and the average fragment length used was 150 bp (-l 150) with an SD of 50 (-s 50); these values were taken as an average of reported fragment lengths for the studies included. For comparison, a more traditional analysis (not included in expVIP) was carried out where reads were aligned to the IWGSC transcriptome version 2.26 using bowtie2 (version 2.2.4) using the parameters recommended by eXpress (Roberts and Pachter, 2013): output in sam format (-S), maximum insert size of 800 bp (-X 800), and unlimited multimappings (-a). Counts per gene and tpm were calculated using eXpress version 1.5.1 using the default parameters except that sequence-specific biases were ignored (-no-bias-correct) due to some samples having too few fragments to accurately learn bias parameters, so the bias correction was turned off for all samples to maintain a uniform treatment across samples.

Differential gene expression analysis was carried out on the kallisto output abundance files using sleuth (Pimentel et al., 2015). Default settings were used, except that the maximum bootstraps considered was 30 (max_bootstrap = 30). For the integrated disease and stress analysis, each sample was compared with the control sample from the study from which it originated. Genes with an FDR-adjusted P (q) < 0.05 were considered differentially expressed.

Visualization Interface

The outputs from kallisto were merged into two separate files: the raw estimated counts and tpm for all samples. Those files were loaded into an MySQL 5.5 relational database along with a Web server using the framework Ruby on Rails 4.2. expVIP is released as a Biogem (Bonnal et al., 2012). The visualization of the expression is implemented as a BioJS (Corpas et al., 2014) component, using the Web development frameworks D3v3, jQuery 2.1, and jQuery-UI 1.11.

Availability of expVIP

The source code to prepare and set up the expVIP database and graphical interface are available in Github: https://github.com/homonecloco/expvip-web. The BioJS component to visualize the expression data are available at the BioJS registry: http://biojs.io/d/bio-vis-expression-bar. The expVIP virtual machine, the data displayed in the Web interface, and the detailed documentation are available on the wiki page https://github.com/homonecloco/expvip-web/wiki.

qPCR Analysis of Reference Gene Stability

Tissue samples were collected in liquid nitrogen for a range of tissues, developmental stages, varieties, and disease conditions (Supplemental Table S7). All plants were grown in greenhouses in soil under 16-h-light/8-h-dark, 20°C day/12°C night conditions, except cv Maris Huntsman seedlings, which were grown on moist filter paper in petri dishes in the dark at 20°C. Frozen samples were ground to a fine powder, and RNA was extracted using TRI Reagent (Sigma) according to the manufacturer's instructions, except for grain samples, which were extracted according to a phenol-based method (Box et al., 2011) with the addition of 20% (v/v) Plant RNA Isolation Aid (Ambion) to the RNA extraction buffer. RNA samples were diluted to 250 ng μL^{-1}, treated with RQ1 DNase (Promega), and reverse transcribed using Moloney murine leukemia virus (Invitrogen) according to the manufacturer's instructions. qPCR was carried out using LightCycler 480 SYBR Green I Master Mix (Roche) with each primer at a final concentration of 0.25 μM and 0.05 μL of cDNA in a 10-μL reaction using 384-well plates. The qPCR program run on the LightCycler 480 (Roche) was as follows: preincubation at 95°C for 5 min; 45 amplification cycles of 95°C for 10 s, 58°C for 10 s, and 72°C for 20 s with the final melt-curve step cooling to 60°C and then heating to 97°C with five reads per 1°C as the temperature increased. For all sample/primer combinations, melt curves were inspected to have only a single product. Crossing thresholds were calculated using the second derivative method provided in the LightCycler 480 SW 1.5 software (Roche). Primer efficiencies were calculated using a serial dilution of cDNA.

Analysis of GO Term Enrichment

GO term enrichment was calculated using Singular Enrichment Analysis provided by agriGO (Du et al., 2010) using default settings. The genes differentially expressed in 10, 11, and 12 abiotic and disease conditions were supplied as the query list, along with GO terms downloaded from EnsemblPlants biomart (release 26). The entire IWGSC version 2.26 transcriptome was used as the reference using GO terms downloaded from EnsemblPlants biomart.

Supplemental Data

The following supplemental materials are available.

Supplemental Figure S1. Comparison of the number of genes expressed per sample and the number of mapped reads.

Supplemental Figure S2. Demonstration of filtering within the expVIP interface.

Supplemental Table S1. Detailed wheat metadata per sample.

Supplemental Table S2. Ten most highly expressed genes in wheat grain and leaf tissues.

Supplemental Table S3. Comparison of the accuracy of kallisto and bowtie2 using nullitetrasomic wheat lines.

Supplemental Table S4. Structure of wheat RNA-seq metadata for www.wheat-expression.com.

Supplemental Table S5. Means, SD, and covariance of transcript expression across 321 wheat samples.

Supplemental Table S6. Stability of reference gene expression across 321 wheat samples.

Supplemental Table S7. Samples used to test the stability of expression of qPCR primers.

Supplemental Table S8. Comparison of coefficients of variation between five novel reference genes and five commonly used reference genes across 30 conditions.

Supplemental Table S9. GO term enrichment among genes expressed under stress and disease conditions.

Supplemental Table S10. Genes differentially expressed in 10 stress conditions, fold change, and function.

Supplemental Text S1. Tutorial for expVIP graphic interface (Wheat Expression Browser example).

Supplemental Text S2. Application of expVIP to rice allows the integration of previous studies.

ACKNOWLEDGMENTS

We thank Nikolai Adamski and Oluwaseyi Shorinola (John Innes Centre [JIC]) for discussions; Martha Clarke, Clare Lewis, Paul Nicholson, and Marianna Pasquariello (JIC) for RNA samples; members of the JIC Crop Genetics Department for beta testing of www.wheat-expression.com; Robert Davey (TGAC) for downloading RNA-seq data from NCBI; and Michael Burrell (NCBI Computing Infrastructure for Science group) for assistance in installing kallisto.

Received October 29, 2015; accepted February 10, 2016; published February 11, 2016.

LITERATURE CITED

Andersson I, Backlund A (2008) Structure and function of Rubisco. Plant Physiol Biochem **46:** 275–291

Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc (September 9, 2015)

Avraham S, Tung CW, Ilic K, Jaiswal P, Kellogg EA, McCouch S, Pujar A, Reiser L, Rhee SY, Sachs MM, et al (2008) The Plant Ontology Database: a community resource for plant structure and developmental stages controlled vocabulary and annotations. Nucleic Acids Res **36:** D449–D454

Barrero JM, Cavanagh C, Verbyla KL, Tibbits JF, Verbyla AP, Huang BE, Rosewarne GM, Stephen S, Wang P, Whan A, et al (2015) Transcriptomic analysis of wheat near-isogenic lines identifies *PM19-A1* and *A2* as candidates for a major dormancy QTL. Genome Biol **16:** 93

Bevan MW, Uauy C (2013) Genomics reveals new landscapes for crop improvement. Genome Biol **14:** 206

Bolser DM, Kerhornou A, Walts B, Kersey P (2015) Triticeae resources in Ensembl Plants. Plant Cell Physiol **56:** e3

Bonnal RJP, Aerts J, Githinji G, Goto N, MacLean D, Miller CA, Mishima H, Pagani M, Ramirez-Gonzalez R, Smant G, et al (2012) Biogem: an effective tool-based approach for scaling up open source software development in bioinformatics. Bioinformatics **28:** 1035–1037

Borrill P, Adamski N, Uauy C (2015) Genomics as the key to unlocking the polyploid potential of wheat. New Phytol **208:** 1008–1022

Box MS, Coustham V, Dean C, Mylne JS (2011) Protocol: a simple phenol-based method for 96-well extraction of high quality RNA from Arabidopsis. Plant Methods **7:** 7

Bray N, Pimentel H, Meslted P, Pachter L (2015) Near-optimal RNA-Seq quantification. arXiv 1505.02710

Cantu D, Pearce SP, Distelfeld A, Christiansen MW, Uauy C, Akhunov E, Fahima T, Dubcovsky J (2011) Effect of the down-regulation of the high *Grain Protein Content* (*GPC*) genes on the wheat transcriptome during monocarpic senescence. BMC Genomics **12:** 492

Cantu D, Segovia V, MacLean D, Bayles R, Chen X, Kamoun S, Dubcovsky J, Saunders DGO, Uauy C (2013) Genome analyses of the wheat yellow (stripe) rust pathogen *Puccinia striiformis* f. sp. *tritici* reveal polymorphic and haustorial expressed secreted proteins as candidate effectors. BMC Genomics **14:** 270

Choulet F, Alberti A, Theil S, Glover N, Barbe V, Daron J, Pingault L, Sourdille P, Couloux A, Paux E, et al (2014) Structural and functional partitioning of bread wheat chromosome 3B. Science **345:** 1249721

Corpas M, Jimenez R, Carbon SJ, García A, Garcia L, Goldberg T, Gomez J, Kalderimis A, Lewis SE, Mulvany I, et al (2014) BioJS: an open source standard for biological visualisation. Its status in 2014. F1000 Res **3:** 55

D'Antonio M, D'Onorio De Meo P, Pallocca M, Picardi E, D'Erchia AM, Calogero RA, Castrignanò T, Pesole G (2015) RAP: RNA-Seq Analysis Pipeline, a new cloud-based NGS web application. BMC Genomics **16:** S3

Dash S, Van Hemert J, Hong L, Wise RP, Dickerson JA (2012) PLEXdb: gene expression resources for plants and plant pathogens. Nucleic Acids Res **40:** D1194–D1201

Du Z, Zhou X, Ling Y, Zhang Z, Su Z (2010) agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Res **38:** W64–W70

Ebine K, Miyakawa N, Fujimoto M, Uemura T, Nakano A, Ueda T (2012) Endosomal trafficking pathway regulated by *ARA6*, a RAB5 GTPase unique to plants. Small GTPases **3:** 23–27

Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity (Edinb) **107:** 1–15

FAO (2009) Global agriculture towards 2050. http://www.fao.org/fileadmin/templates/wsfs/docs/Issues_papers/HLEF2050_Global_Agriculture.pdf (September 9, 2015)

FAO (2015) FAOSTAT. http://faostat3.fao.org (September 9, 2015)
Feller A, Machemer K, Braun EL, Grotewold E (2011) Evolutionary and comparative analysis of MYB and bHLH plant transcription factors. Plant J **66:** 94–116
Fonseca NA, Petryszak R, Marioni J, Brazma A (2014) iRAP: an integrated RNA-seq analysis pipeline. BioRxiv. http://biorxiv.org/content/early/2014/06/06/005991 (September 9, 2015)
Gillies SA, Futardo A, Henry RJ (2012) Gene expression in the developing aleurone and starchy endosperm of wheat. Plant Biotechnol J **10:** 668–679
Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, et al (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res **40:** D1178–D1186
Heffner EL, Sorrells ME, Jannink JL (2009) Genomic selection for crop improvement. Crop Sci **49:** 1–12
International Wheat Genome Sequencing Consortium (2014) A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science **345:** doi/10.1126/science.1251788
Jiang Y, Yang B, Deyholos MK (2009) Functional characterization of the Arabidopsis bHLH92 transcription factor in abiotic stress. Mol Genet Genomics **282:** 503–516
Jordan KW, Wang S, Lun Y, Gardiner LJ, MacLachlan R, Hucl P, Wiebe K, Wong D, Forrest KL, Sharpe AG, et al (2015) A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. Genome Biol **16:** 48
Krasileva KV, Buffalo V, Bailey P, Pearce S, Ayling S, Tabbita F, Soria M, Wang S, Akhunov E, Uauy C, et al (2013) Separating homeologs by phasing in the tetraploid wheat transcriptome. Genome Biol **14:** R66
Kugler KG, Siegwart G, Nussbaumer T, Ametz C, Spannagl M, Steiner B, Lemmens M, Mayer KF, Buerstmayr H, Schweiger W (2013) Quantitative trait loci-dependent analysis of a gene co-expression network associated with Fusarium head blight resistance in bread wheat (*Triticum aestivum* L.). BMC Genomics **14:** 728
Lawrence CJ, Schaeffer ML, Seigfried TE, Campbell DA, Harper LC (2007) MaizeGDB's new data types, resources and activities. Nucleic Acids Res **35:** D895–D900
Leach LJ, Belfield EJ, Jiang C, Brown C, Mithani A, Harberd NP (2014) Patterns of homoeologous gene expression shown by RNA sequencing in hexaploid bread wheat. BMC Genomics **15:** 276
Li A, Liu D, Wu J, Zhao X, Hao M, Geng S, Yan J, Jiang X, Zhang L, Wu J, et al (2014) mRNA and small RNA transcriptomes reveal insights into dynamic homoeolog regulation of allopolyploid heterosis in nascent hexaploid wheat. Plant Cell **26:** 1878–1900
Li HZ, Gao X, Li XY, Chen QJ, Dong J, Zhao WC (2013) Evaluation of assembly strategies using RNA-seq data associated with grain development of wheat (*Triticum aestivum* L.). PLoS ONE **8:** e83530
Liu Z, Xin M, Qin J, Peng H, Ni Z, Yao Y, Sun Q (2015) Temporal transcriptome profiling reveals expression partitioning of homeologous genes contributing to heat and drought acclimation in wheat (*Triticum aestivum* L.). BMC Plant Biol **15:** 152
Nakashima K, Takasaki H, Mizoi J, Shinozaki K, Yamaguchi-Shinozaki K (2012) NAC transcription factors in plant abiotic stress responses. Biochim Biophys Acta **1819:** 97–103
Nussbaumer T, Kugler KG, Bader KC, Sharma S, Seidel M, Mayer KFX (2014) RNASeqExpressionBrowser: a web interface to browse and visualize high-throughput expression data. Bioinformatics **30:** 2519–2520
Ohyanagi H, Tanaka T, Sakai H, Shigemoto Y, Yamaguchi K, Habara T, Fujii Y, Antonio BA, Nagamura Y, Imanishi T, et al (2006) The Rice Annotation Project Database (RAP-DB): hub for *Oryza sativa* ssp. *japonica* genome information. Nucleic Acids Res **34:** D741–D744
Oono Y, Kobayashi F, Kawahara Y, Yazawa T, Handa H, Itoh T, Matsumoto T (2013) Characterisation of the wheat (*Triticum aestivum* L.) transcriptome by de novo assembly for the discovery of phosphate starvation-responsive genes: gene expression in Pi-stressed wheat. BMC Genomics **14:** 77
Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L, et al (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Res **35:** D883–D887
Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, Holloway E, Kolesnykov N, Lilja P, Lukk M, et al (2007) ArrayExpress: a public database of microarray experiments and gene expression profiles. Nucleic Acids Res **35:** D747–D750
Pearce S, Huttly AK, Prosser IM, Li YD, Vaughan SP, Gallova B, Patil A, Coghill JA, Dubcovsky J, Hedden P, et al (2015) Heterologous expression and transcript analysis of gibberellin biosynthetic genes of grasses reveals novel functionality in the GA3ox family. BMC Plant Biol **15:** 130
Pfeifer M, Kugler KG, Sandve SR, Zhan B, Rudi H, Hvidsten TR, Mayer KF, Olsen OA (2014) Genome interplay in the grain transcriptome of hexaploid bread wheat. Science **345:** 1250091
Pimentel H, Bray N, Meslted P, Pachter L (2015) sleuth: RNA-Seq analysis. http://pachterlab.github.io/sleuth/ (September 9, 2015)
Qi B, Huang W, Zhu B, Zhong X, Guo J, Zhao N, Xu C, Zhang H, Pang J, Han F, et al (2012) Global transgenerational gene expression dynamics in two newly synthesized allohexaploid wheat (*Triticum aestivum*) lines. BMC Biol **10:** 3
Ramirez-Gonzalez RH, Segovia V, Bird N, Fenwick P, Holdgate S, Berry S, Jack P, Caccamo M, Uauy C (2015) RNA-Seq bulked segregant analysis enables the identification of high-resolution genetic markers for breeding in hexaploid wheat. Plant Biotechnol J **13:** 613–624
Ray DK, Mueller ND, West PC, Foley JA (2013) Yield trends are insufficient to double global crop production by 2050. PLoS ONE **8:** e66428
Roberts A, Pachter L (2013) Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Methods **10:** 71–73
Rouard M, Guignon V, Aluome C, Laporte MA, Droc G, Walde C, Zmasek CM, Périn C, Conte MG (2011) GreenPhylDB v2.0: comparative and functional genomics in plants. Nucleic Acids Res **39:** D1095–D1102
Sears E (1954) The aneuploids of common wheat. University of Missouri Agricultural Experiment Station Research Bulletin 572
Shaik R, Ramakrishna W (2013) Genes and co-expression modules common to drought and bacterial stress responses in *Arabidopsis* and rice. PLoS ONE **8:** e77261
Shaik R, Ramakrishna W (2014) Machine learning approaches distinguish multiple stress conditions using stress-responsive genes and identify candidate genes for broad resistance in rice. Plant Physiol **164:** 481–495
Shewry PR (2009) Wheat. J Exp Bot **60:** 1537–1553
Singh K, Foley RC, Oñate-Sánchez L (2002) Transcription factors in plant defense and stress responses. Curr Opin Plant Biol **5:** 430–436
Teh OK, Hofius D (2014) Membrane trafficking and autophagy in pathogen-triggered cell death and immunity. J Exp Bot **65:** 1297–1312
Tenea GN, Peres Bota A, Cordeiro Raposo F, Maquet A (2011) Reference genes for gene expression studies in wheat flag leaves grown under different farming conditions. BMC Res Notes **4:** 373
Tilman D, Balzer C, Hill J, Befort BL (2011) Global food demand and the sustainable intensification of agriculture. Proc Natl Acad Sci USA **108:** 20260–20264
Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. Genome Res **19:** 327–335
Wagner GP, Kin K, Lynch VJ (2013) A model based criterion for gene expression calls using RNA-seq data. Theory Biosci **132:** 159–164
Walls R, Smith B, Elser J, Goldfain A, Stevenson D, Jaiswal P (2012) A plant disease extension of the infectious disease ontology. *In* R Cornet, R Stevens, eds, 3rd International Conference on Biomedical Ontology. http://ceur-ws.org/ (September 9, 2015)
Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet **10:** 57–63
Yan L, Loukoianov A, Tranquilli G, Helguera M, Fahima T, Dubcovsky J (2003) Positional cloning of the wheat vernalization gene *VRN1*. Proc Natl Acad Sci USA **100:** 6263–6268
Yang F, Li W, Jørgensen HJL (2013) Transcriptional reprogramming of wheat and the hemibiotrophic pathogen *Septoria tritici* during two phases of the compatible interaction. PLoS ONE **8:** e81606
Yang Z, Peng Z, Wei S, Liao M, Yu Y, Jang Z (2015) Pistillody mutant reveals key insights into stamen and pistil development in wheat (*Triticum aestivum* L.). BMC Genomics **16:** 211
Zhang H, Yang Y, Wang C, Liu M, Li H, Fu Y, Wang Y, Nie Y, Liu X, Ji W (2014) Large-scale transcriptome comparison reveals distinct gene activations in wheat responding to stripe rust and powdery mildew. BMC Genomics **15:** 898
Zimmermann P, Hirsch-Hoffmann M, Hennig L, Gruissem W (2004) Genevestigator: Arabidopsis microarray database and analysis toolbox. Plant Physiol **136:** 2621–2632

Supplemental figures

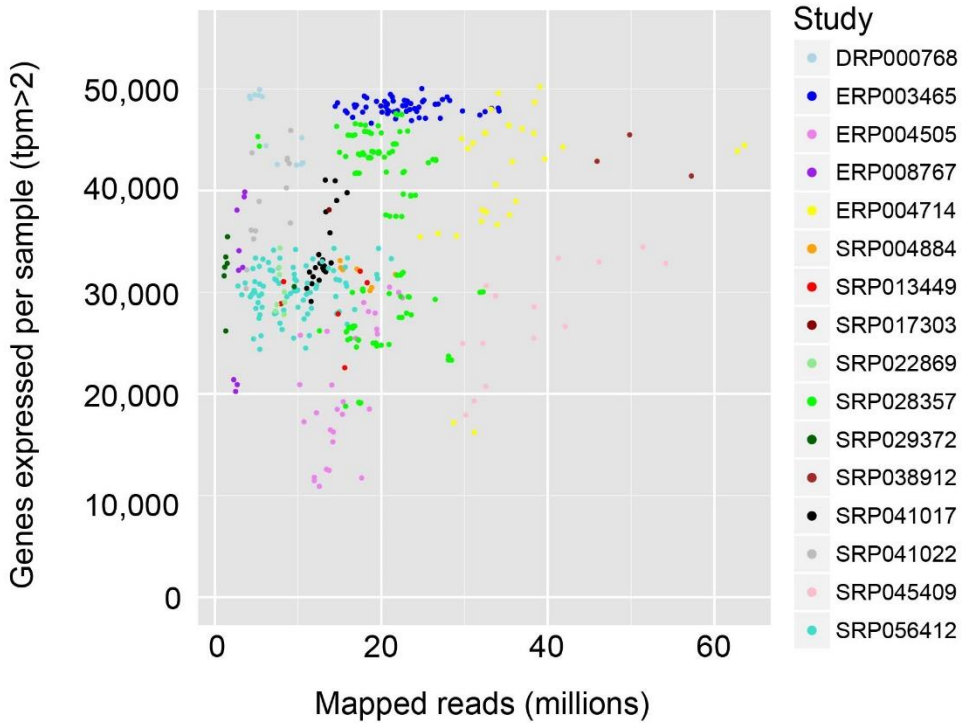

Figure S1. Number of genes expressed per sample compared to the number of mapped reads per sample. Samples are colour coded by study.

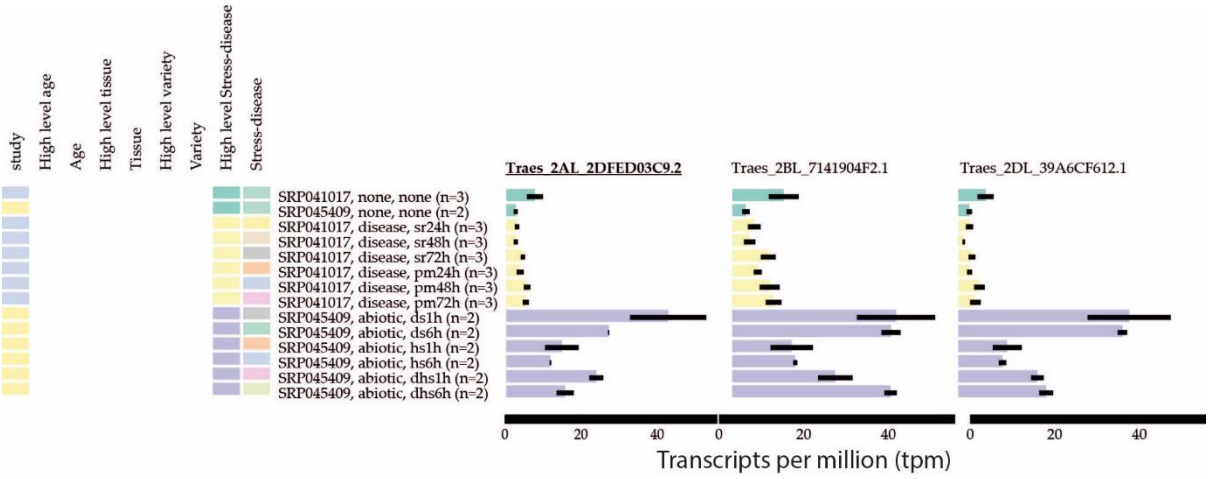

Figure S2. Demonstration of data filtering within expVIP interface. The expression of Traes_2AL_2DFED03C9.2 is shown only from studies which were used for differential gene expression analysis (SRP041017 for disease and SRP045409 for abiotic stress).

Supplemental tables

Tissue	Ensembl transcript ID	Average expression (tpm)	Function
grain	Traes_5BS_FDE980DE1.1	5,808	Grain softness protein
grain	TRAES3BF027700070CFD_t1	6,898	Serine protease inhibitor (annotation of orthologue)
grain	Traes_3DS_D718FF51C1.2	7,042	Alpha-amylase inhibitor 0.19
grain	Traes_1AS_8FAD6A69F.1	9,155	unknown
grain	Traes_1DS_67B7153A8.1	10,350	Glutenin (annotation of orthologue)
grain	Traes_1DS_66B67E9B41.2	20,423	Glutenin (annotation of orthologue)
grain	Traes_4AL_661613B77.1	63,949	Alpha/beta-gliadin MM1
grain	Traes_4AL_4FF5B8837.1	70,002	Alpha/beta-gliadin A-I
grain	Traes_5BL_E68C461B3.1	77,505	Alpha/beta-gliadin (annotation of orthologue)
grain	Traes_1DL_D861501F5.1	96,186	Glutenin, high molecular weight subunit 12
leaf	Traes_4DL_7CF374FEE.1	4,518	Ribulose bisphosphate carboxylase large chain (annotation of orthologue)
leaf	Traes_3DS_4EF0DAA39.1	4,794	Photosystem I reaction center subunit XI (annotation of orthologue)
leaf	Traes_3DL_B80EC2366.1	6,041	putative nontranslating CDS
leaf	TRAES3BF032400040CFD_t1	6,402	unknown
leaf	Traes_4DL_85C90C56C.2	7,058	Ribulose bisphosphate carboxylase large chain
leaf	TRAES3BF011000020CFD_t1	7,785	Histidine kinase 3 (annotation of orthologue)
leaf	Traes_6BL_5979B5341.1	10,819	unknown
leaf	TRAES3BF007300450CFD_t1	12,421	unknown
leaf	TRAES3BF007300320CFD_t1	21,923	unknown
leaf	Traes_7DS_037CD9FCA.1	23,120	unknown

Table S2. Ten most highly expressed genes in wheat grain and leaf tissues from seven and nine independent studies, respectively.

			Shoots							Roots						
			euploid	N1AT1B	N1AT1D	N1BT1A	N1BT1D	N1DT1A	N1DT1B	euploid	N1AT1B	N1AT1D	N1BT1A	N1BT1D	N1DT1A	N1DT1B
kallisto	Average gene expression level in shoots for genes located on chromsome (tpm):	1A	16.5	3.0	3.6	42.0	19.5	32.6	22.6	20.7	3.6	4.7	42.9	24.0	38.8	22.0
		1B	14.0	31.3	16.8	3.0	3.2	16.6	32.0	20.0	40.2	20.9	2.4	3.2	20.4	39.4
		1D	14.7	17.2	29.7	18.7	29.1	3.1	3.4	21.3	23.9	39.3	25.5	45.7	3.4	4.1
		1A+ 1B+ 1D	45.2	51.5	50.0	63.7	51.9	52.3	58.0	62.1	67.7	64.8	70.8	73.0	62.5	65.5
	Percentage of gene expression originating from:	1A	36.6%	5.8%	7.3%	66.0%	37.6%	62.4%	39.0%	33.4%	5.3%	7.2%	60.5%	32.9%	62.0%	33.6%
		1B	30.9%	60.7%	33.5%	4.7%	6.3%	31.7%	55.2%	32.3%	59.3%	32.2%	3.4%	4.4%	32.6%	60.2%
		1D	32.5%	33.5%	59.3%	29.3%	56.1%	5.9%	5.8%	34.4%	35.4%	60.6%	36.0%	62.7%	5.4%	6.2%
bowtie2 + eXpress	Average gene expression level in shoots for genes located on chromsome (tpm):	1A	9.1	2.2	2.4	20.5	11.0	16.8	11.5	13.5	2.7	3.4	26.3	14.9	24.0	13.9
		1B	7.4	15.3	8.2	1.8	2.0	8.6	15.8	13.0	24.7	13.2	2.4	2.6	12.9	24.6
		1D	7.6	8.6	14.8	8.9	14.7	1.9	1.9	13.8	15.2	24.3	16.0	28.1	2.7	3.2
		1A+ 1B+ 1D	24.1	26.1	25.4	31.1	27.7	27.4	29.2	40.2	42.7	40.8	44.7	45.6	39.6	41.8
	Percentage of gene expression originating from:	1A	37.8%	8.5%	9.4%	65.7%	39.6%	61.5%	39.4%	33.6%	6.4%	8.3%	58.8%	32.6%	60.6%	33.4%
		1B	30.7%	58.5%	32.3%	5.8%	7.2%	31.5%	54.1%	32.2%	58.0%	32.3%	5.3%	5.7%	32.7%	58.9%
		1D	31.5%	33.0%	58.3%	28.5%	53.2%	7.0%	6.5%	34.2%	35.7%	59.5%	35.8%	61.7%	6.7%	7.7%

Table S3. Comparison of accuracy for two methods of read alignment and quantification. Reads from nullitetrasomic wheat lines were aligned and quantified using the pseudo aligner and quantifier kallisto (upper part of table) or the conventional aligner bowtie2 combined with eXpress for quantification (lower part of table). Gene expression on chromosome group 1 (the chromosome group for which whole chromosomes were added or deleted in these lines) was compared to assess accuracy of alignment and quantification.

This tutorial is based on the `Wheat Expression Browser`. However, the principles are the same for any transcriptome study which is powered by the expVIP graphical interface.

Home Page

The home page allows the user to insert a gene name to search and to define which studies are to be included in the visualisation interface. By default all studies are selected, but users can select/deselect a study by simply clicking on the specific button.

You can also compare expression between two genes by introducing both gene names in the boxes and pressing the `Compare` button.

Alternatively you can compare expression across multiple genes (up to 50) to generate a heatmap. You can add a list of genes separate by commas or one gene per line in the `Multiple genes` box.

All gene names are based on the transcriptome reference used for expVIP: for the case of the Wheat Expression Browser we used the IWGSC transcriptome available through Ensembl Plants release 26.

Visualisation interface

Single gene or two-gene comparison

Once the gene expression loads the page includes several features. These are shown below and explained point by point:

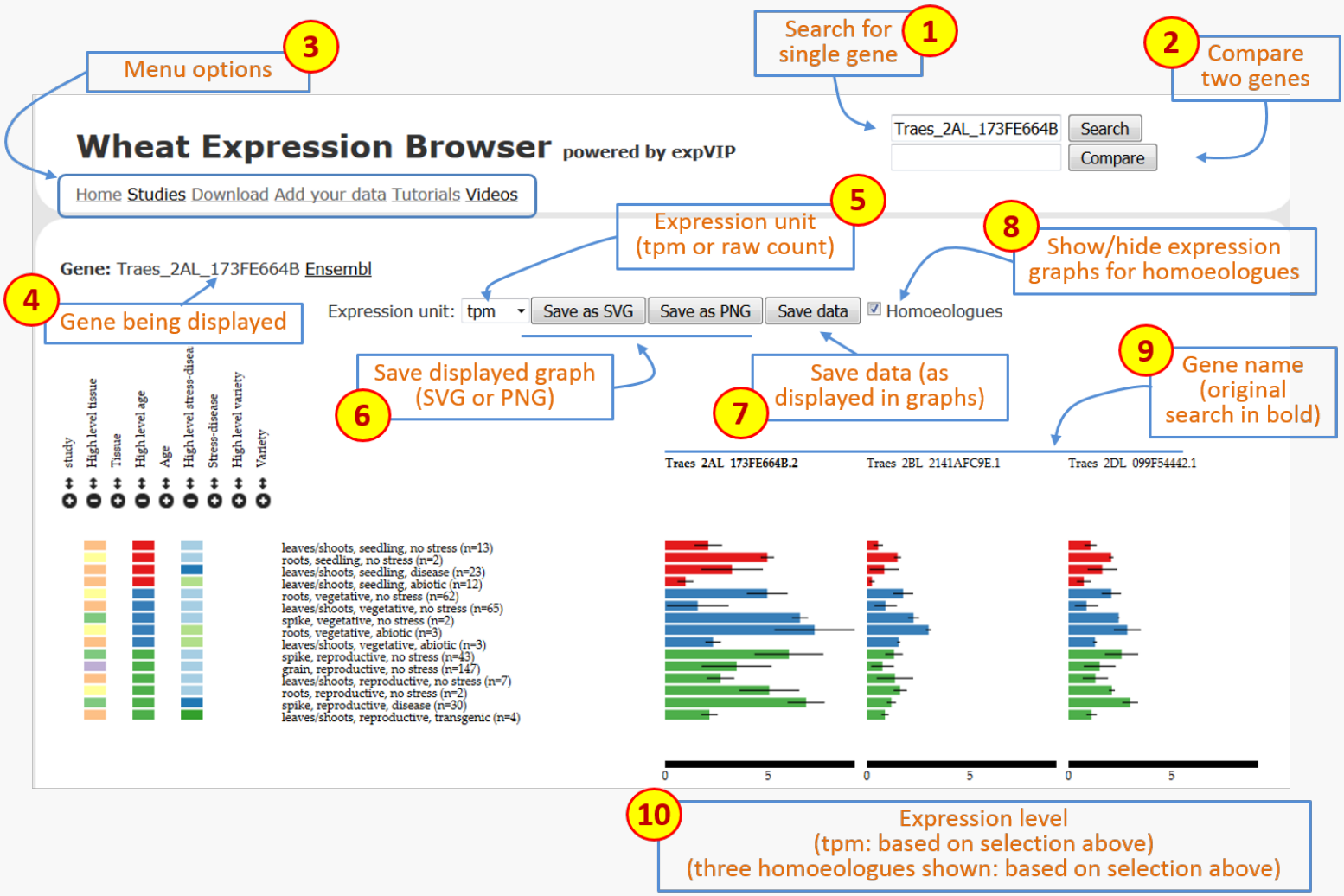

Figure 1: Overall description of features on Wheat Expression Browser

1. `Search box`: at any point you can type or copy a new gene name (based on Ensembl Plants nomenclature) and generate a new set of expression data.
2. `Compare box`: you can type a second gene name and press the `Compare` button to generate two expression graphs drawn at the same scale.
3. `Menu options`: this includes a series of links to different options:
    - `Home`: return to `home` screen.
    - `Studies`: opens up a popup screen with a summary and short description of each study and a link to manuscript.
    - `Download`: link to download all the wheat expression database including `tpm` and `counts` and associated metadata.
    - `Add your data`: link to GitHub to download virtual machine.
    - `Tutorials`: link to Wheat Expression Browser Tutorial.
    - `Videos`: link to Wheat Expression Browser Video Tutorial.
4. `Gene`: shows the gene which is currently being displayed with link to Ensembl Plants gene page.
5. `Expression unit`: allows user to select the expression unit used to visualise the expression data. This can be either “transcript per million (`tpm`)” or “estimated counts (`counts`)”. We have not provided RPKM given the inconsistencies generated across samples when using this measure. A detailed discussion can be found in Wagner *et al* (2012). It is

important to mention that `tpm` is preferred over RPKM since it allows an easier comparison for abundances between samples. However it is important to stress that while `tpm` serves as a relative measure to compare genes across experiments, a proper normalisation and statistical analysis with differential gene expression programs must be performed. `expVIP` generates outputs which allow easy implementation of `sleuth`, `DESeq` and `EdgeR`.

6. `Save graph`: these two buttons allow users to save the current graphs in either `SVG` (to work on Adobe Illustrator) or as `PNG` files. The graphical file will render based on the current selection and order of factors as displayed on the screen.

7. `Save data`: this allows the user to download a `csv` file with the data based on the current selection and order of factors as displayed on the screen. The data will include the standard errors and the number of samples that make up each value.

8. `Homoeologues`: by clicking on this button, the `Wheat Expression Browser` will display the expression graphs of known homoeologues of the original primary gene. This gene name will remain in bold and the homoeologous graphs will be displayed according to A, B, D genome ordering. When homoeologues are displayed the same expression scale is used across graphs and the sorting and filtering of factors is simultaneous to allow easier comparison.

9. `Gene names`: gene name for corresponding graph. When homoeologues are shown the original gene used for the search is shown in bold.

10. `Expression level`: the expression level adjusts according to the expression of each set of gene homoeologues. The scale remains consistent across homoeologues to allow easier comparison. The values are based on the unit selected in the `expression unit` box (see point 5 above).

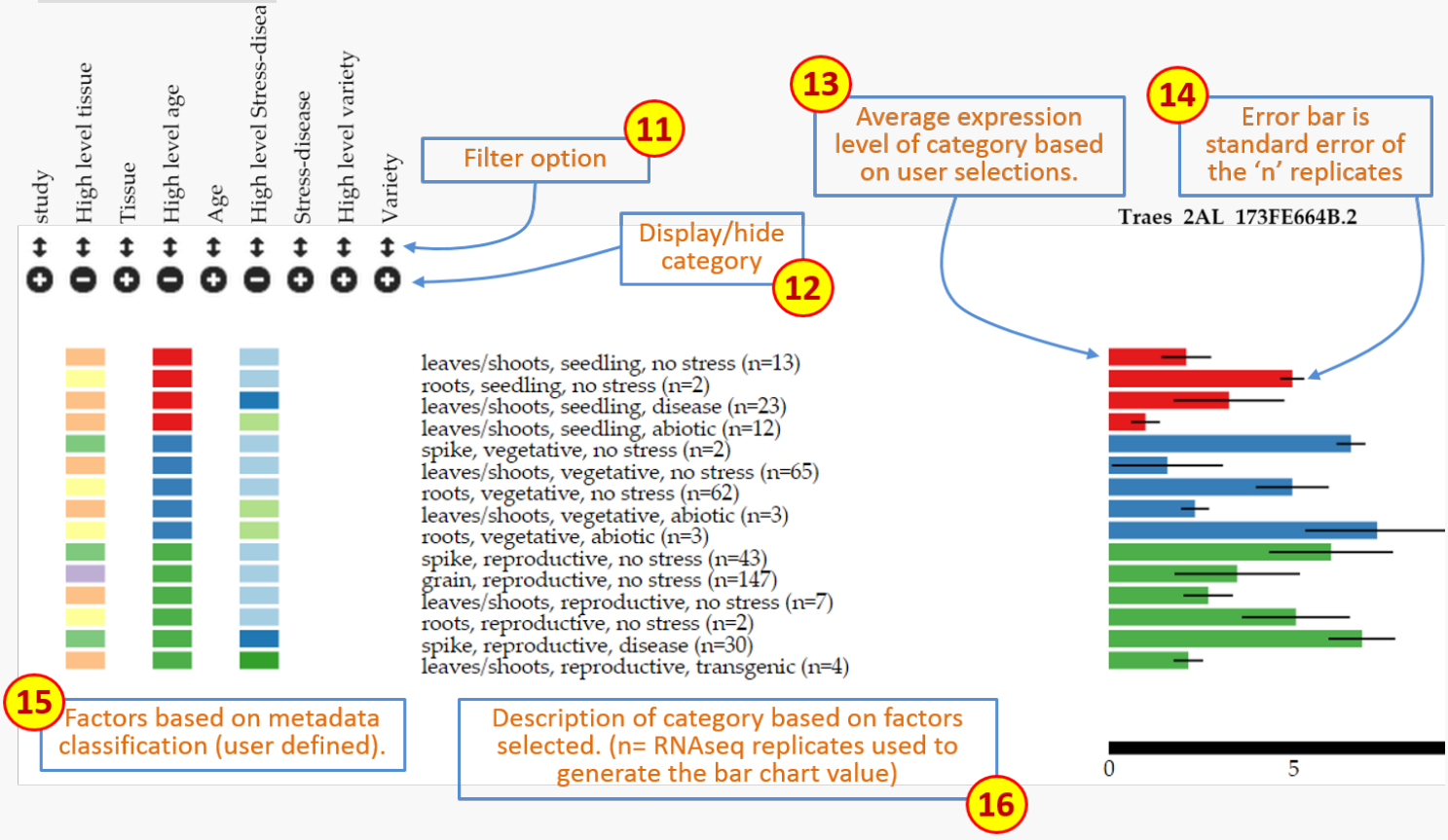

Figure 2: Overall description of features on Wheat Expression Browser (continued)

11. `Filter`: This feature open a pop-up window which reveals all the levels within the particular category. All levels are pre-selected, but users can choose to display specific levels by selecting or deselecting them accordingly. If a level is

deselected, then the data associated with this factor is removed from the graph. Within the pop-up window levels can also be re-arranged according to the user's preference by dragging the level to the specific position within the pop-up window (see Features section below).

12. `Display/hide category`: Each individual category can be displayed or hidden by pressing the `+/-` button. When a category is displayed, the expression graphs will re-arrange according to the new category which has been introduced. If a category is hidden, then the graphs will also adjust accordingly. Data is not removed when doing this, rather it is grouped within the categories selected such that the total samples displayed remains the same. The colours within the category correspond to unique values or levels (up to 24 different colours) and are also used in the bar graphs corresponding to the expression data.

13. `Expression bars`: These bars represent the expression level of the "n" samples which are grouped according to the factors chosen based on the selection criteria (11 and 12 above). When hovering over the bar with the mouse a small tooltip will indicate the expression level (`tpm` or `counts`) and the standard error (sem) used for the error bars (see 14)

14. `Error bars`: Standard error of the means for the "n" expression values on which the bar graph is based.

15. `Factors`: Coloured rectangles represent the categories which are displayed according to the factors chosen based on the selection criteria (11 and 12 above). When hovering above the rectangles a tooltip will appear to show the long name of the level being examined.

16. `Description`: Text description of the factors chosen based on the selection criteria (11 and 12 above) and the number of RNAseq samples (n) which meet this specific criterion.

Multiple gene comparisons

17. `Expression unit`: For heatmaps, log2(tpm) is suggested as the expression unit as this provides better resolution to compare multiple genes across several categories.

18. `Heatmap`: Expression data is represented as a heatmap. As for single genes, categories can be sorted and filtered using the same tools. Gene names appear on the top of each column. Currently, up to 50 genes can be visualised in one heatmap. In Figure 3, for example, the two right-most genes are expressed solely in grains, with one being expressed to higher levels as suggested by the dark blue colour.

19. `Scale`: Colour scale for the expression values in the heatmap. The values adjust according to the highest tpm value being displayed within the current heatmap visualisation. Since tpm values below 2 are considered as very low expressed genes and log2 values of tpm<1 result in negative expression values, we forced tpm values below 1 to have a log2 value of cero (i.e. log2(<1)=0).

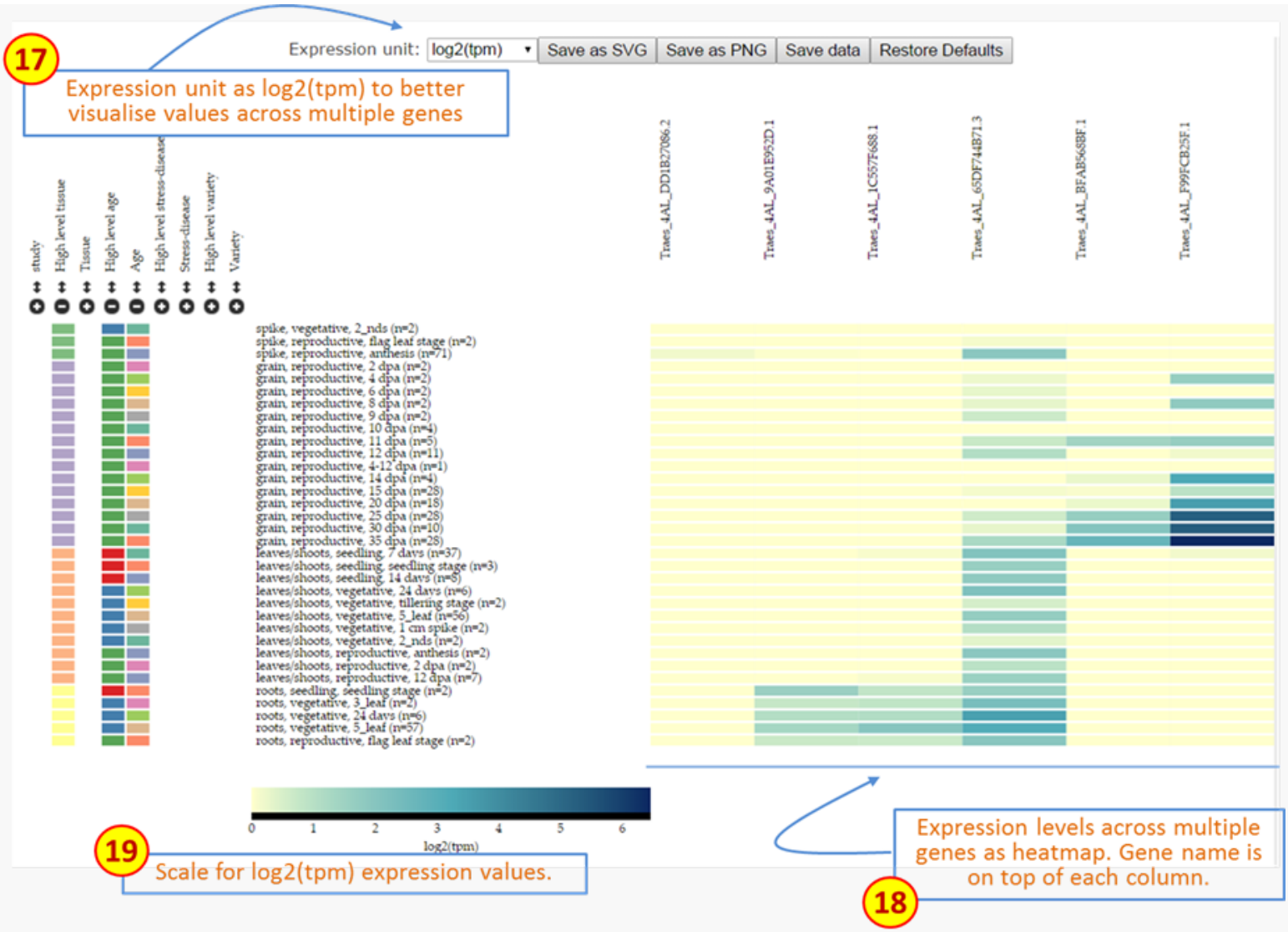

Figure 3: Description of features on Wheat Expression Browser using Multiple gene comparisons.

Features

Sorting

Factors can be sorted within each category in two ways.

1. The first is by simply clicking the mouse on top of the coloured rectangles underneath the heading. For example in Figure 2 samples are sorted on `High level age` from `seedling` (red), `vegetative` (blue) to `reproductive` (green). If the user clicks on any of the coloured rectangles in the `High level tissue` category, then the graph is automatically reorganised based on this factor. In this case it includes four categories as defined by the user in the metadata and the bar graphs on the right hand side change colour according to the latest factor used for sorting. The previous factor used (in this case `high level age`) remains as a secondary sorting factor (Figure 3).

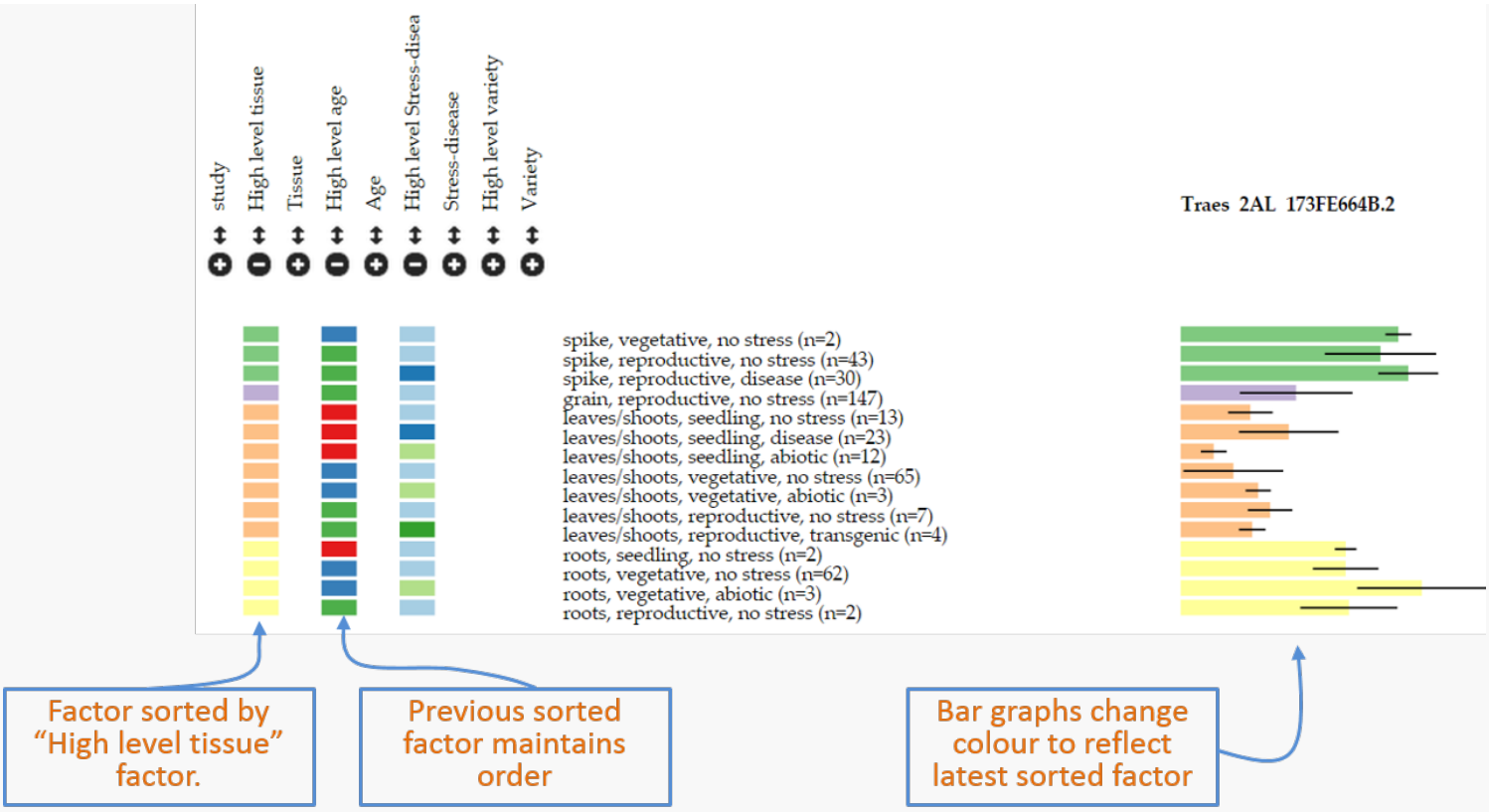

Figure 4: Example of new sorting of data based on clicking of rectangles within "high level tissue".

2. Alternatively, the user can define the exact order of factors within the browser interface. To do so the `filter` option (point 11 above) can be used. By clicking on the double arrow button the user opens a pop-up window which shows the levels within the factor. In this example by pressing the double-arrow underneath `high level tissue` a pop-up with four levels appears based on the order as determined in the user defined metadata (`spike`, `grain`, `leaves/shoots`, `roots`). To rearrange this, the user can simply click, hold and drag the level to the desired position. This will automatically re-arrange the data based on the new order and the corresponding graph and legends will follow suit. The bottom panel of Figure 5 shows a new order of `roots`, `leaves/shoots`, `spike` and `grain`.

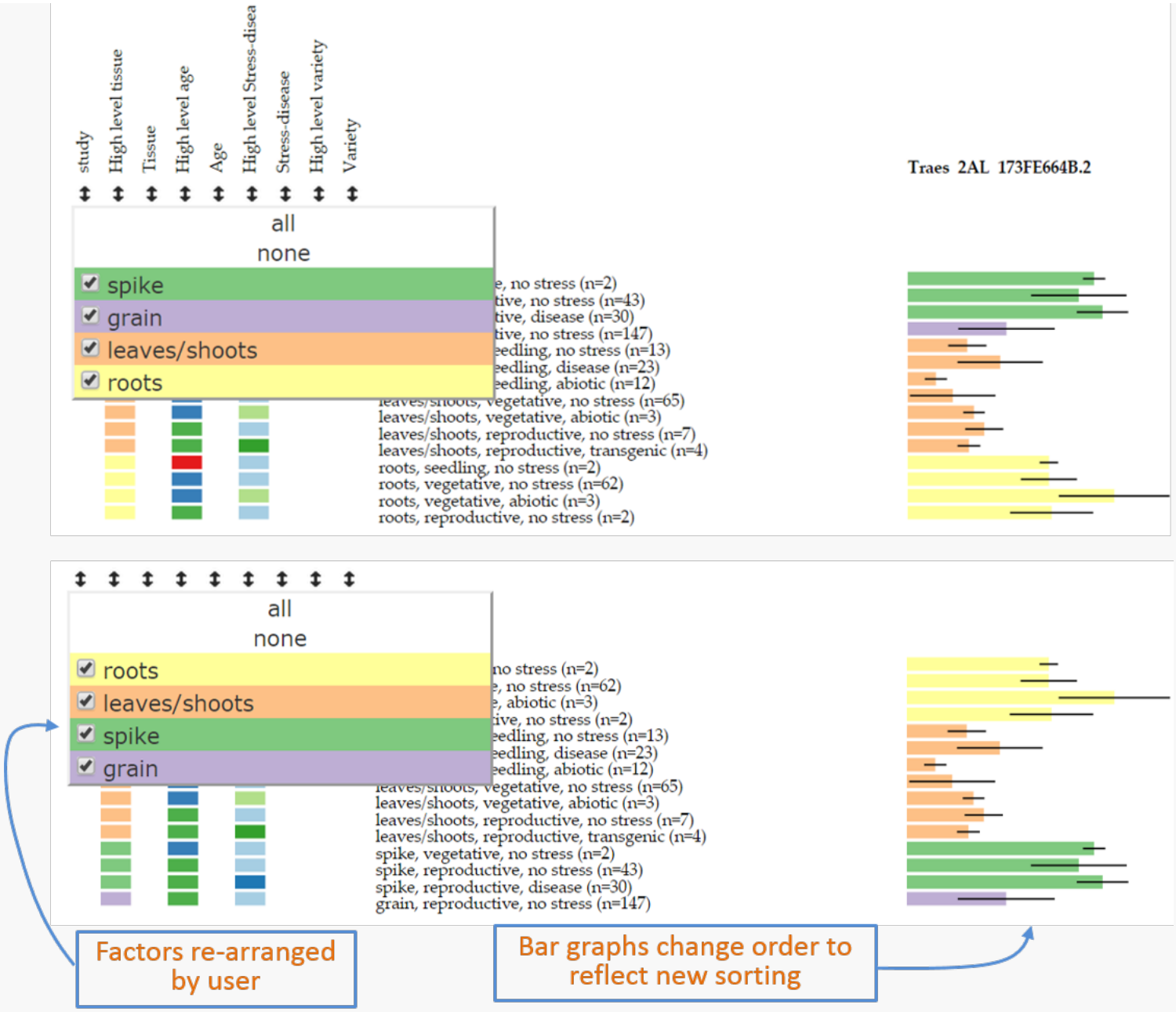

Figure 5: Example of sorting of data based on new user defined order within the filter pop-up window.

Filtering

In cases it may be required to remove certain samples from the visualisation. Note that displaying or hiding a category (point 12 above) does not remove the underlying data from the visualisation: this just simply groups the data within the selected category. Therefore to remove samples from the visualisation the user can open the filter pop-up as described for the `Sorting` option. Individual levels within the category can then be removed by using the “check-box” on the left hand side of the level name. By de-selecting a given level (in the example for Figure 6 we have deselected `leaves/shoots` and `spike`), samples defined as such will be removed from the analysis and will not be shown in the bar graphs. In Figure 6 now only two levels remain (`roots` and `grains`) and hence the bar graphs only show these two levels. Notice that the numbers of samples which comprise each bar graph are the same as those on Figure 5. The pop-up window also includes an `all` and `none` option to rapidly select/deselect individual samples. The filtering option can be used on any factor: for example to remove a complete study from the analysis the easiest way is to select the `study` filtering pop-up on the far left and deselect the study in question.

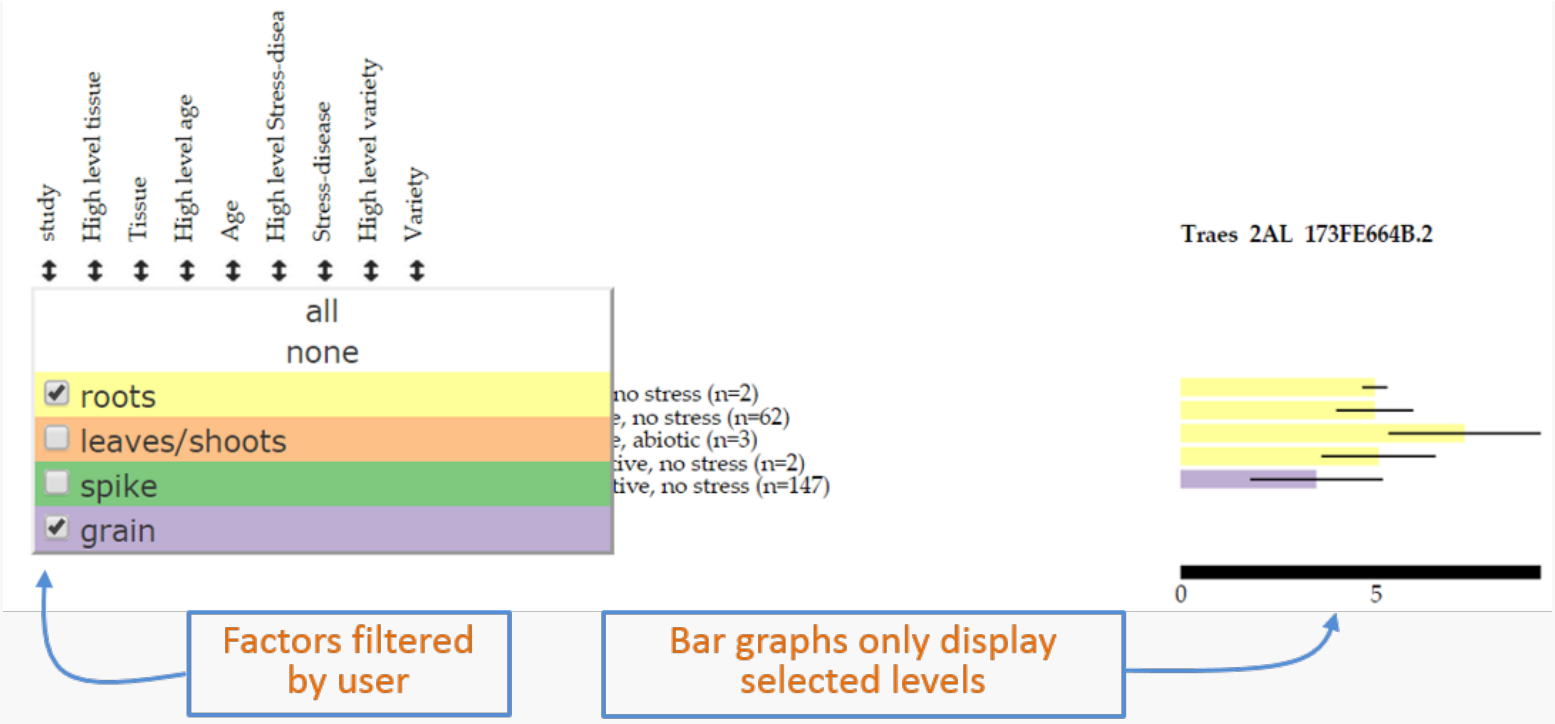

Figure 6: Example of filtering data based on user defined selection within the filter pop-up window.

Text S2. Application of expVIP to rice allows integration of previous studies

The rapid progress in sequencing technologies has meant that genomes and transcriptome references for a species are constantly improved. In some species, multiple references are available, making it difficult to compare results between studies. This problem is exemplified in rice, where two different genome annotations are widely used: the Rice Annotation Project (RAP) and Michigan State University (MSU) gene models (Ohyanagi et al., 2006; Ouyang et al., 2007).

To test the use of expVIP to integrate data we analysed two studies (SRA: DRP000716 and SRP028766) which examined gene expression changes in rice in response to phosphate starvation. DRP000716 compared the response to 22 days of phosphate (Pi) starvation in four rice varieties with varying levels of tolerance to Pi stress: the *japonica* cultivar Nipponbare with low tolerance, two *japonica* cultivars with higher tolerance IAC 25 and Vary Lava, and the *indica* cultivar Kasalath known to be highly tolerant to Pi stress (Oono et al., 2013). The second study SRP028766 investigated how the *japonica* cultivar Nipponbare responded to a time-course of Pi starvation at 1 h, 6 h, 24 h, 3 days, 7 days and 21 days (Secco et al., 2013). Both studies used 2 week old seedlings grown in hydroponic conditions which enables their comparison, however each study used a different annotation of the rice genome (DRP000716 used RAP, SRP028766 used MSU). We used expVIP to align and quantify gene expression for both studies, using the RAP gene models (release date 31.03.2015, from http://rapdb.dna.affrc.go.jp) as a common reference. We used sleuth to identify genes which were differentially expressed between varieties in DRP000716 (each variety under phosphate starvation was compared to the same variety without phosphate starvation) and between time-points in SRP028766 (phosphate starved plants were compared to plants at the same time-point which were in phosphate sufficient conditions).

Validation of previous results

DRP000716 – Inter-variety comparison**:** We found that similar numbers of genes were up and downregulated in all four varieties (Figure S3A), although Kasalath roots had fewer genes downregulated than in any other variety. We identified approximately 2-fold fewer

differentially expressed transcripts than published previously (Oono et al., 2013), which is likely due to the use of a defined reference in our study, whereas before *de novo* transcripts including multiple isoforms were also assembled. We found a total of 163 genes upregulated and 34 genes downregulated in roots and shoots of all varieties. Amongst these upregulated genes, many are known to be involved in phosphate response including *SPX1* and *SPX3* (Wang et al., 2009) and the phosphate transporter *PHT1;4* (Zhang et al., 2015). The downregulated genes included genes related to primary metabolism, e.g. ribulose bisphosphate carboxylase, and genes involved in abiotic stress responses such as *RISBZ5*, a potential negative regulator of drought and cold stress response (Liu et al., 2012).

SRP028766 – time-course**:** As previously reported we found that relatively few genes were induced within a short time period after imposition of phosphorous starvation (1h, 6h and 24 h; Figure S3B). At 3 days many genes become differentially expressed in roots, whereas there are still few genes differentially expressed in shoots. At 7 and 21 days several thousand genes are up and downregulated in roots and shoots. These results correspond well to the previously reported trends (Secco et al., 2013): we found that the early response (1 h - 3 days) to phosphate starvation involves suites of different genes at each time-point in roots and as previously reported no genes were differentially expressed in roots in common between all early time-points.

Comparison between DRP000716 and SRP028766

First we investigated whether the two different studies identified similar genes to be differentially expressed in the cultivar Nipponbare after 21 or 22 days under phosphate starvation. We found that in total 1,565 and 2,001 genes were differentially expressed in shoots and roots, respectively, across both studies (Figure S3C). We identified fewer differentially expressed genes in DRP000716 than in SRP028766, which may reflect the lower number of reads mapped in the former (15.5 million and 46.2 million, respectively). We found that amongst genes which were differentially expressed in both studies, there was a higher correlation of fold change in genes expressed in the roots (R^2 = 0.64) than in genes expressed in the shoots (R^2= 0.38) (Figure S3D). This suggests that changes in root gene

expression were more consistent between studies (more shared genes, and more highly correlated fold changes in expression), and for this reason we focused our analysis on genes differentially expressed in roots.

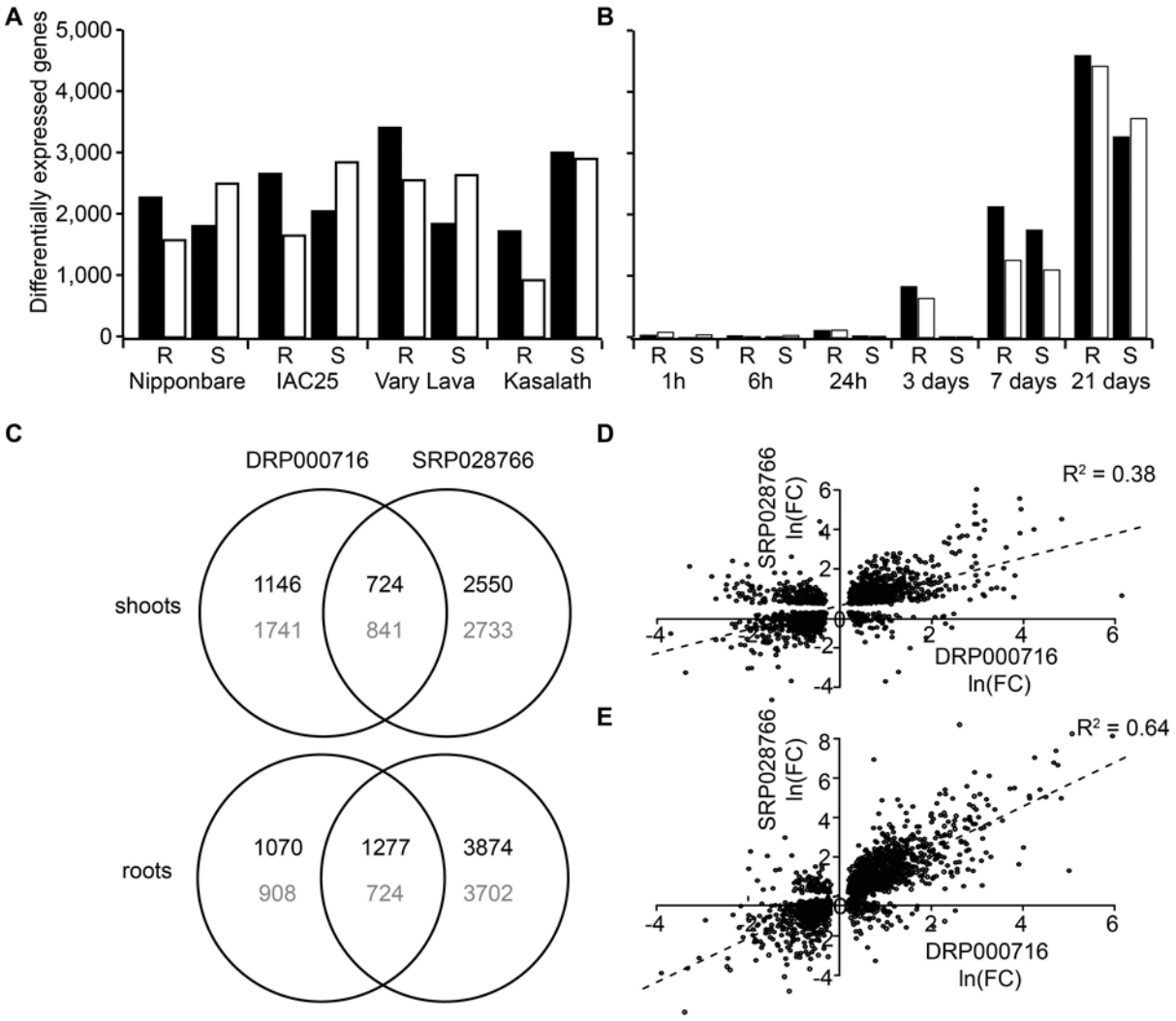

Figure S3. Comparison of differentially expressed genes identified in DRP00716 and SRP028766. (A) Genes differentially expressed after 22 days phosphate starvation in four rice cultivars. (B) Genes differentially expressed during a timecourse of phosphorous starvation in Nipponbare. In (A) and (B) filled bars represent upregulated genes, empty bars represent downregulated genes. (C) Differentially expressed genes identified in DRP000716 and SRP028766 at 22 and 21 days after phosphate starvation respectively in Nipponbare. Upregulated genes are shown in black, downregulated genes in grey. (D, E) Natural log (ln) fold change (FC) in genes differentially expressed under phosphate starvation in SRP028766 and DRP000716 in (D) shoots and (E) roots.

The integration of both studies with the same reference genes allows easy comparison between studies (above) and allows new hypotheses to be tested using existing data. We hypothesised that genes which are differentially expressed in roots of all four varieties from DRP000716 might be conserved phosphate responsive genes and should also be identified in SRP028766. To test this hypothesis we identified genes which were differentially expressed at 22 days in DRP000716 in all four varieties, and genes differentially expressed in each individual variety (Figure S4). For genes which were differentially expressed in individual varieties between 50 and 61 % were also differentially expressed in SRP028766 at 21 days (Figure S4). Amongst the genes conserved between all four varieties a higher percentage (81 %) were also detected in SRP028766: this suggests that not only are these genes conserved between varieties but they are also induced in independent experiments to a higher degree and would be strong candidates to investigate conserved phosphate responsive genes.

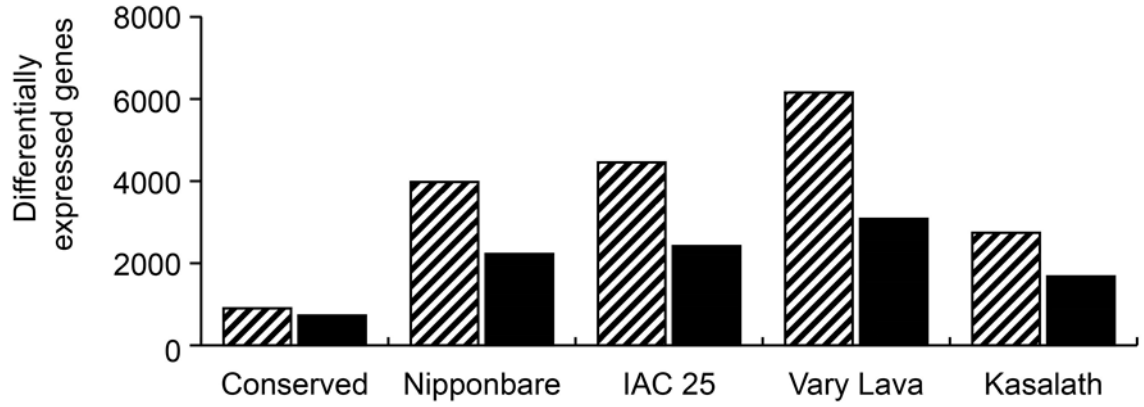

Figure S4. Intersection between genes differentially expressed in roots of all varieties and time-course expression. Differentially expressed genes identified in DRP000716 (striped bars) are also identified after 21 days phosphate starvation in SRP028766 (filled bars). Genes differentially expressed in all four varieties are called conserved.

Amongst these 726 conserved genes many have functions related to phosphate regulation including phosphate transporters (*PHT1-1*, *PHT1-4*, *PHT1-6*, *PHT1-8*, *PHT1-10*) and *SPX1*, *SPX2* and *SPX3* (Wang et al., 2009). Interestingly 91 genes have unknown functions and do not contain known interpro protein domains. An additional 32 genes contain domains of unknown function (DUF), one of which is represented in five genes: DUF581. In Arabidopsis

it has been proposed that DUF581 genes respond to specific environmental stresses and interact with SnRK1 to balance energy status (Nietzsche et al., 2014). These 5 genes (*Os03g0183500*, *Os04g0585700*, *Os06g0125200*, *Os06g0223700*, *Os09g0433800*) are upregulated 3-734 fold under phosphate starvation and may represent novel phosphate responsive genes.

Literature Cited

Liu C, Wu Y, Wang X (2012) bZIP transcription factor *OsbZIP52/RISBZ5*: a potential negative regulator of cold and drought stress response in rice. Planta **235:** 1157-1169

Nietzsche M, Schießl I, Börnke F (2014) The complex becomes more complex: protein-protein interactions of *SnRK1* with *DUF581* family proteins provide a framework for cell- and stimulus type-specific *SnRK1* signaling in plants. Frontiers in Plant Science **5**

Ohyanagi H, Tanaka T, Sakai H, Shigemoto Y, Yamaguchi K, Habara T, Fujii Y, Antonio BA, Nagamura Y, Imanishi T, Ikeo K, Itoh T, Gojobori T, Sasaki T (2006) The Rice Annotation Project Database (RAP-DB): hub for *Oryza sativa* ssp. *japonica* genome information. Nucleic Acids Research **34:** D741-D744

Oono Y, Kawahara Y, Yazawa T, Kanamori H, Kuramata M, Yamagata H, Hosokawa S, Minami H, Ishikawa S, Wu J, Antonio B, Handa H, Itoh T, Matsumoto T (2013) Diversity in the complexity of phosphate starvation transcriptomes among rice cultivars based on RNA-Seq profiles. Plant Molecular Biology **83:** 523-537

Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L, Orvis J, Haas B, Wortman J, Buell CR (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Research **35:** D883-D887

Secco D, Jabnoune M, Walker H, Shou H, Wu P, Poirier Y, Whelan J (2013) Spatio-Temporal Transcript Profiling of Rice Roots and Shoots in Response to Phosphate Starvation and Recovery. The Plant Cell **25:** 4285-4304

Wang Z, Hu H, Huang H, Duan K, Wu Z, Wu P (2009) Regulation of *OsSPX1* and *OsSPX3* on Expression of *OsSPX* domain Genes and Pi-starvation Signaling in Rice. Journal of Integrative Plant Biology **51:** 663-674

Zhang F, Sun Y, Pei W, Jain A, Sun R, Cao Y, Wu X, Jiang T, Zhang L, Fan X, Chen A, Shen Q, Xu G, Sun S (2015) Involvement of *OsPht1;4* in phosphate acquisition and mobilization facilitates embryo development in rice. The Plant Journal **82:** 556-569