

1     **Systematic analysis of 1,298 RNA-Seq samples and construction**  
2     **of a comprehensive soybean (*Glycine max*) expression atlas**

3     Fabricio Brum Machado<sup>1,#</sup>, Kanhu C. Moharana<sup>1,#,\*</sup>, Fabricio Almeida-Silva<sup>1</sup>, Rajesh K.  
4     Gazara<sup>1</sup>, Francisnei Pedrosa-Silva<sup>1</sup>, Fernanda S. Coelho<sup>1</sup>, Clícia Grativol<sup>1</sup>, Thiago M.  
5     Venancio<sup>1,\*</sup>

6     <sup>1</sup>Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biotecnologia e  
7     Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro; Campos dos  
8     Goytacazes, Brazil. # These authors contributed equally to this work.

9

10

11     \* Corresponding authors

12     Av. Alberto Lamego 2000 / P5 / 217; Parque Califórnia

13     Campos dos Goytacazes, RJ

14     Brazil

15     CEP: 28013-602

16     TMV: [thiago.venancio@gmail.com](mailto:thiago.venancio@gmail.com); KCM: [kcm.eid@gmail.com](mailto:kcm.eid@gmail.com)

17

18

19

## 20 **Abstract**

21 Soybean (*Glycine max* [L.] Merr.) is a major crop in animal feed and human nutrition,  
22 mainly for its rich protein and oil contents. The remarkable rise in soybean transcriptome  
23 studies over the past five years generated an enormous amount of RNA-seq data,  
24 encompassing various tissues, developmental conditions, and genotypes. In this study,  
25 we have collected data from 1,298 publicly available soybean transcriptome samples,  
26 processed the raw sequencing reads, and mapped them to the soybean reference  
27 genome in a systematic fashion. We found that 94% of the annotated genes  
28 (52,737/56,044) had detectable expression in at least one sample. Unsupervised  
29 clustering revealed three major groups, comprising samples from aerial, underground,  
30 and seed/seed-related parts. We found 452 genes with uniform and constant expression  
31 levels, supporting their roles as housekeeping genes. On the other hand, 1,349 genes  
32 showed heavily biased expression patterns towards particular tissues. A transcript-level  
33 analysis revealed that 95% (70,963/74,490) of the known transcripts overlap with those  
34 reported here, whereas 3,256 assembled transcripts represent potentially novel splicing  
35 isoforms. The dataset compiled here constitute a new resource for the community, which  
36 can be downloaded or accessed through a user-friendly web interface at  
37 <http://venanciogroup.uenf.br/resources/>. This comprehensive transcriptome atlas will  
38 likely accelerate research on soybean genetics and genomics.

39

## 40 Introduction

41 Soybean (*Glycine max* [L.] Merr.) is one of the most important legume crops worldwide.  
42 It is critically important in human nutrition, animal feed, and biotechnological  
43 applications. Global climate change and increased food demand resulting from a growing  
44 human population have been fueling the development and application of  
45 biotechnological methods to generate better cultivars (Iizumi et al., 2014). In recent years,  
46 various omics approaches have been deployed to improve productivity of several crops,  
47 including soybean. An important achievement in soybean omics-based research was the  
48 availability of whole-genome sequencing data, which helped identify molecular markers  
49 (e.g. single nucleotide polymorphisms, SNPs) (Schmutz et al., 2010; Deshmukh et al., 2014)  
50 that are instrumental in the identification of genes associated with various phenotypes of  
51 interest. Further, the soybean whole-genome sequencing project has also contributed to  
52 the substantial rise in soybean transcriptome studies (Libault et al., 2010; Severin et al.,  
53 2010; Garg and Jain, 2013; O'Rourke et al., 2017), initially dominated by microarray  
54 platforms and later by RNA-Seq technologies.

55 To date, several studies reported spatiotemporal changes occurring in various  
56 soybean tissues using RNA-seq. The two first soybean RNA-Seq studies were published by  
57 Libault *et al.* (Libault et al., 2010) and Severin *et al.* (Severin et al., 2010). The former  
58 reported the sequencing of 14 (mainly root and nodule) tissues, whereas the latter  
59 evaluated several tissues and seed developmental stages. Dozens of other studies  
60 followed, such as those addressing different life cycle stages (Jones and Vodkin,  
61 2013; Bellieny-Rabelo et al., 2016; Gazara et al., 2019), conditions (Belamkar et al., 2014),  
62 and cultivars/lines (Goettel et al., 2014). The accumulation of plant transcriptomic data in  
63 public repositories [e.g. Sequence Read Archive (SRA) at the National Center for  
64 Biotechnology Information (NCBI)] inspired the development of unified collections or  
65 atlases, such as those found for *Arabidopsis thaliana* (Fucile et al., 2011), *Medicago*  
66 *truncatula* (He et al., 2009), *G. max* (Supplementary Table S1), as well as multi-species  
67 atlases (Dash et al., 2012), which are often reused by the scientific community. Specifically  
68 in soybean, Kim *et al.* constructed the SoyNet ([www.inetbio.org/soynet](http://www.inetbio.org/soynet)) database using  
69 734 microarrays and 290 RNA-seq samples (Kim et al., 2017), while Wu *et al.* uncovered  
70 a nodulation-related co-expression module by analyzing 1,270 microarray samples  
71 generated with Affymetrix gene chips (Wu et al., 2019).

72 Despite the previous efforts to integrate soybean transcriptomes, there is a  
73 massive amount of soybean RNA-Seq data that remain largely unexplored. Here, we have  
74 collected data from 1,298 publicly available soybean RNA-seq samples from the NCBI SRA  
75 database. We systematically processed and mapped sequencing reads to the soybean  
76 reference genome. Transcriptional levels were estimated to allow a systematic global  
77 gene expression analysis, aiming to elucidate the dynamics of transcriptional regulation

78 across this broad range of samples, tissues, and cultivars. Further, the collected and  
79 processed data are readily available to allow both, automatic analysis and single-gene  
80 investigations using an easy-to-use interface at our lab website  
81 (<http://venanciogroup.uenf.br/resources/>).

82

## 83 **RESULTS AND DISCUSSION**

### 84 **Data gathering, processing, and mapping to the reference genome reveal an overall high** 85 **quality of the publicly available soybean RNA-Seq data**

86 We performed an extensive literature mining process to gather as many as possible  
87 soybean RNA-seq datasets. A total of 1,742 raw read sequencing files were downloaded  
88 from the NCBI SRA database (Supplementary Table S2). Reads obtained from the same  
89 biological sample were combined in a single FASTQ file (or in two files, for paired-end  
90 data; \*\_1.fq and \*\_2.fq). This resulted in 1,298 samples (65% single-end and 35% paired-  
91 end) from 84 BioProjects comprising sixteen different broad tissue categories in various  
92 developmental stages (Supplementary Table S3). Approximately 35% (458/1298) of the  
93 samples lacked cultivar/genotype information in SRA. Among the other 840 samples, we  
94 found 157 different soybean cultivar names, although this is likely an overestimation  
95 because of authors calling the same cultivars with slightly different names during data  
96 submission. The cultivar Williams 82, which had the genome sequenced, represented 23%  
97 (302/1,298) of the total samples. Leaves were the most abundant tissue, representing  
98 46% (603/1,298) of the samples (Figure 1). Three libraries from unknown tissue sources  
99 were excluded. We have also found that 76% (986/1,295) of the libraries were unstranded  
100 (Supplementary Table S3).

101 Reads from each RNA-seq library were mapped to the reference genome,  
102 assembled, and used for estimating gene expression (Figure 2). Whenever present,  
103 adapter sequences were trimmed. Reads with average quality lower than 20 were  
104 excluded. An average of 32,210,805 million reads pairs per sample with paired-end data  
105 and 29,579,316 million reads per sample with single-end data were used for read  
106 mapping. Mapped and uniquely mapped reads correspond to an average of 87.9% and  
107 81%, respectively (Supplementary Table S4 and Supplementary Figure 1). Further, we  
108 excluded 47 samples for which: i) 50% or more of the reads failed to map or; ii) 40% or  
109 more of the reads failed to uniquely map. After these exclusions, 1,248 samples were kept  
110 for further downstream analysis.

111 Several methods used to analyze RNA-seq data (e.g. differential gene expression)  
112 rely on read count normalization strategies (Robinson and Oshlack, 2010;Po-Yen et al.,  
113 2011), such as Reads Per Kilobase Million (RPKM) (Mortazavi et al., 2008), Fragments Per  
114 Kilobase Million (FPKM), and Transcripts Per Million (TPM) (Wagner et al., 2012), out of  
115 which the latter has been proposed to be more consistent across technical replicates

116 (Wagner et al., 2012;Conesa et al., 2016;Li and Li, 2018). Here, we normalized data using  
117 TPM for most of the downstream analysis. Nevertheless,  $\log_2$  transformed raw read  
118 counts are more commonly used for quality control steps such as unsupervised sample  
119 clustering (Jordan et al., 2015). In addition, many popular tools used for differential gene  
120 expression analysis (e.g. DESeq2, edgeR) require raw read counts instead of normalized  
121 read counts. Therefore, after read mapping, we estimated transcript abundances in the  
122 form of raw read counts per transcript and TPM. Transcript-level expression values were  
123 also aggregated to estimate expression at gene level. Gene expression values across 1,248  
124 samples were then used in further downstream analysis.

125

### 126 **Unsupervised sample clustering reveals three major clades comprising underground, 127 aerial, and seed tissues**

128 In transcriptomics studies, gene and samples are often clustered to identify sub-groups  
129 with similar transcriptional profiles (Liu and Si, 2014;Marini and Binder, 2019). While gene  
130 clustering helps identify co-expressed genes, sample clustering is instrumental to detect  
131 broad transcriptional similarities between samples, as well as to identify potential  
132 technical artifacts and mislabeled samples. Among several methods, distance-based  
133 hierarchical clustering, *K-means* clustering, and dimensional-reduction-based  
134 visualization methods (e.g. principal component analysis, PCA) are commonly used.  
135 Recently, t-Distributed Stochastic Neighbor Embedding (t-SNE) has been shown to  
136 provide a better global structure of sample sub-groups than several other methods (Dey  
137 et al., 2017). Here, we employed three sample clustering methods to identify outliers and  
138 overall pairwise sample similarity. We used a gene expression matrix as input to perform  
139 hierarchical clustering, *K-means* clustering, and t-SNE analysis. These analyses uncovered  
140 three major groups comprising samples from aerial, underground, and developmental or  
141 seed tissues (Figure 3) (Severin et al., 2010). Interestingly, however, we found an  
142 additional cluster comprising samples from leaves and shoots from drought-stress-  
143 related and leaf senescence samples. Although not entirely novel, these results are part  
144 of an important step to check for technical issues or biases that could, for example, result  
145 in the clustering of samples from the same sequencing batch or research group. Four  
146 shoot samples and one root sample clustered with seed-embryo samples. After  
147 confirming this result with the t-SNE and *K-means* clustering, we excluded these samples.  
148 Overall, sample clustering supports a high quality level of the publicly available RNA-Seq  
149 samples analyzed here, as only 0.4% (5/1248) of the samples were excluded after the  
150 clustering analysis.

151

152 **Systematic analysis of hundreds of RNA-Seq libraries support the expression of the**  
153 **vast majority of the soybean genes**

154 After comparing the reference transcript annotations (for 56,044 genes) with the  
155 merged consensus transcript assembly, we excluded 1.3% (759/56,044) of the genes  
156 because of overlapping gene predictions. Next, we applied a minimum TPM threshold of  
157 1 to define a gene as expressed and found that 92.1% (51,644/56,044) of the known  
158 soybean protein-coding genes were expressed in at least one sample. The remaining  
159 genes had their TPM values set to zero and classified as not expressed. An average of  
160 31,063 genes were expressed per sample. The tissues with the greatest numbers of  
161 expressed genes were inflorescence (37,108 genes) and flower (average of 36,051 genes)  
162 (Supplementary Figure 2A), whereas nodules had the lowest number of expressed genes  
163 (average of 25,718 genes). We also found 16,916 genes expressed in at least 1,150  
164 samples (Supplementary Figure 2B), including 1,758 genes that are expressed in all 1,243  
165 samples. On the other hand, 6% (3,233/56,044) of the genes were not expressed (TPM <  
166 1) in any sample, out of which 82% had coding regions comprising less than 500 codons  
167 (Supplementary Figure 3). As a final data quality check, we analyzed the top 1,000  
168 expressed genes from each tissue category using MapMan pathway bins (see Methods).  
169 For example, contrasting gene expression profiles of roots and leaves uncovered several  
170 expected transcriptional patterns of photosynthesis genes in the latter (Supplementary  
171 Figure 4).

172

173 **Housekeeping genes**

174 Given the wide coverage of tissues and conditions, we also sought to identify  
175 housekeeping (HK) genes based on the assumption that these genes are constitutively  
176 and robustly expressed across broad conditions (Czechowski et al., 2005; Hu et al., 2009).  
177 Further, several of these genes have also been used as references in real-time  
178 quantitative polymerase chain reaction (RT-qPCR) assays (Supplementary Table S5).  
179 Hence, by using a large collection of RNA-Seq datasets as the one presented here, one  
180 can not only evaluate commonly used reference genes, but also propose new ones. By  
181 employing a previously developed method (Hoang et al., 2017), we inferred 452 HK genes  
182 (Supplementary Table S6). We evaluated expression levels of each gene in tissues with at  
183 least 10 samples and found that HK genes had very low expression variation (Figure 4A).  
184 To identify HK genes, we used a score that consists of the product of the Coefficient of  
185 Variation and ratio of the maximum to the minimum expression level (see methods for  
186 details). Genes with scores within the 1st quartile were classified as HK genes. Further,

187 we used a tissue-specificity index  $Tau$  ( $\tau$ ) (Yanai et al., 2004; Kryuchkova-Mostacci and  
188 Robinson-Rechavi, 2017) to estimate tissue specificity and verify whether our predicted  
189 HK genes were broadly expressed or not. The  $\tau$  values scale from 0 to 1, where low and  
190 high values indicate widely expressed and more tissue-specific genes, respectively. The  $\tau$   
191 scores of the HK genes ranged from 0.053 to 0.379, supporting their stable expression  
192 level (Figure 6).

193 According to their expression levels, HK genes were grouped in three broad  
194 clusters (Figure 4B). Importantly, 7 previously proposed HK genes (Yim et al., 2015) were  
195 present in our list (Figure 4), out of which four (*ACT11.C*, *B-actin*, *CYP.B* and, *ELF1 $\alpha$* ) belong  
196 to cluster 1 (highly expressed, Figure 4A), confirming that high expression is typically an  
197 important factor in choosing reference genes. Conversely, given its expression  
198 fluctuations (Figure 4), we do not recommend using *UBQ10*, which has also been  
199 proposed as a reference gene.

200 Pathway enrichment analysis of the 452 putative HK genes revealed that these  
201 genes are involved in various biological processes such as RNA degradation, mRNA  
202 surveillance, and TCA cycle (Figure 4B). We found an enrichment of orthologs of  
203 *Arabidopsis* essential genes (Meinke, 2019) among the HK genes (Fisher's Exact test; p-  
204 value = 1.76e-2). Given their roles in basic biological processes, we also verified the  
205 conservation of the HK genes in other 14 species on Phytomine and found that 85%  
206 (385/452) of them have orthologs in at least 10 other species (Supplementary Table S6),  
207 as opposed to an average of 181.6 ( $\pm$  11.6) in 5 random lists of 452 non-HK genes.

208

### 209 **Tissue-specific gene expression**

210 We compared the global expression patterns between tissues to identify tissue-specific  
211 genes (Figure 5). We selected 359 samples that belong to the same tissues and clustered  
212 together (Supplementary Table S7), which resulted in the exclusion of four tissue  
213 categories. The 12 tissues were compared with each other (a total of 144 comparisons),  
214 resulting in a total of 1,349 genes up-regulated in a single tissue as compared to all the  
215 others (Figure 7; Supplementary Table S8). Importantly, 96% of these genes (1,300/1,349)  
216 had  $\tau$  indexes greater than 0.8 and median  $\tau$  of 0.9704 (Figure 6). Given their strong  
217 preferential expression in particular tissues, we called these genes as tissue-specific.

218 The number of tissue-specific genes ranged from 4 in pods to 358 in nodules.  
219 Collectively, nodule (26.5%) and endosperm (301; 22%) account for nearly half of the  
220 tissue-specific genes. The lower number of tissue-specific genes in leaf, shoot, cotyledon,  
221 and pod can be explained by the physiological or developmental relatedness of some  
222 samples (e.g. cotyledon and seed). Notably, 39% (520/1,349) of the tissue-specific genes  
223 identified here were also identified by Severin et. al (Severin et al., 2010) using a much

224 smaller set of samples, supporting the general high quality and reproducibility of the  
225 publicly available soybean transcriptomes. Strikingly, nearly 12% (168/1,349) of the  
226 tissue-specific genes were transcription factors (TFs) (Table 1), which is a remarkable  
227 enrichment (Fisher's Exact Test, p-value = 2.94e-11) considering the overall abundance of  
228 TFs in the soybean genome (Moharana and Venancio, 2019). Among the tissue-specific  
229 TFs, 27, 21, and 20 genes belong to the MYB, C2H2, and ERF families, respectively. Of the  
230 27 MYB TFs, 20 were specific to flower (n=8), hypocotyl (n=7), and endosperm (n=5). Of  
231 the 21 C2H2 genes, 12 were specific to nodule (n=6) and endosperm (n=6). Ten out of 20  
232 ERF genes and six out of 10 WRKY genes were specific to hypocotyl. Finally, 8 of 9 MIKC  
233 type MADS TFs were flower-specific. Several interesting tissue-specific genes are  
234 discussed in the sections below.  
235

### 236 ***Nodule-specific genes***

237 Symbiotic N<sub>2</sub> fixation takes place in root nodules of several Fabaceae species. Nodulation  
238 had a single origin in the common ancestor of the N<sub>2</sub>-fixing clade, followed by multiple  
239 independent losses (Griesmann et al., 2018). Among the genes lost in non-nodulating  
240 species, *Nodule Inception* (NIN) and *Rhizobium-Directed Polar Growth* (RPG) were  
241 reported to be of paramount importance for the origin of root nodules (Griesmann et al.,  
242 2018). As mentioned above, nodule is the tissue with the greatest number of tissue-  
243 specific genes in soybean, a trend that has also been reported in other legumes (Benedito  
244 et al., 2008). Soybean nodules have been shown to correlate poorly with other tissues at  
245 the transcriptional level (Severin et al., 2010), a finding that we corroborated here.

246 We found several nitrogen fixation genes as nodule-specific, including two  
247 leghemoglobin (*Glyma.10G199000*, *Glyma.20G191200*) and ten nodulin genes. The TF  
248 families mostly represented among the 29 nodule-specific TFs were NIN-like (n=6) and  
249 C2H2 (n=6). A higher percentage of NIN-like and C2H2 nodule-specific TFs have been also  
250 described previously (Libault et al., 2010; Severin et al., 2010). Importantly, NIN-like and  
251 C2H2 TFs are important in nitrate signaling (Konishi and Yanagisawa, 2013) and  
252 symbiosome differentiation during nodule development (Sinharoy et al., 2013). We also  
253 found three nodule-specific ERF TFs that are conserved in *Phaseolus vulgaris* and  
254 *Medicago truncatula* and are essential for nodule differentiation and development  
255 (Vernié et al., 2008).

256 We found 12 soybean nodule-specific genes within the experimentally validated  
257 list of over 200 nodulins described previously (Roy et al., 2019). These 12 genes include  
258 the above mentioned ERF TFs, NIN (*Glyma.04G000600*), C2H2 (*Glyma.07G135800*), and  
259 GRAS (*Glyma.16G008200*). Next, we analyzed the 28 genes from a nodule-related module  
260 identified in a co-expression network derived from soybean microarray data (Wu et al.,



261 2019). Notably, 9 of these 28 genes were identified as nodule-specific in our analysis: one  
262 leghemoglobin (*Glyma.10G199000*), two NIN-like TFs (*Glyma.02G311000*,  
263 *Glyma.14G001600*), two purine biosynthesis genes (*Glyma.08G001000*,  
264 *Glyma.11G221100*), one iron transporter (*Glyma.05G121600*), one zinc finger protein-  
265 related (*Glyma.08G044700*), one sulfate transporter (*Glyma.18G018900*), and a formyl  
266 transferase (*Glyma.19G115900*).  
267

### 268 ***Endosperm-specific genes***

269 The endosperm plays important roles during seed development. *Ar. thaliana* endosperm-  
270 specific genes are associated with cell cycle, DNA processing, chromatin assembly, protein  
271 synthesis, cytoskeleton- and microtubule-related processes, and cell/organelle  
272 biogenesis and organization (Day et al., 2008). Out of the 301 endosperm-specific genes  
273 reported here, 9 (*Glyma.19G040600*, *Glyma.09G194500*, *Glyma.01G147300*,  
274 *Glyma.19G058100*, *Glyma.19G044000*, *Glyma.04G187100*, *Glyma.03G219800*,  
275 *Glyma.02G255900*, and *Glyma.08G129200*) encode chromatin modifiers such as histone  
276 acetyltransferases, histone-lysine n-methyltransferases, histone deacetylases, and  
277 histone demethylases. Further, 17 endosperm-specific genes encode F-box proteins and  
278 8 genes encode BTB-POZ and MATH domain proteins, which likely operate in the  
279 ubiquitin-proteasome pathway (Smalle and Vierstra, 2004; Figueroa, 2005). We also found  
280 36 endosperm-specific TFs, including 6 and 5 C2H2 and MYB TFs, respectively. Together,  
281 these results clearly show a number of endosperm-specific genes as involved in  
282 transcriptional and post-transcriptional regulatory processes.  
283

### 284 ***Flower-specific genes***

285 The genetic basis of floral development has been widely studied in several plants,  
286 including *Ar. thaliana* and *Antirrhinum majus* (Soltis et al., 2007; Bowman et al., 2012).  
287 According to the ABCDE model, most of the genes involved in the regulation of flower  
288 development encode MADS and AP2/ERF TFs (Chi et al., 2017). The combinatory action  
289 of these genes regulates the development of various distinct floral parts. For example, *Ar.*  
290 *thaliana* sepal development is regulated by the MADS-box gene *APETALA1* (AP1) together  
291 with the ERF TF *APETALA2* (AP2). Similarly, two MADS-box genes, *APETALA3* (AP3) and  
292 *PISTILLATA* (PI), regulate petal/stamen development, whereas the MADS-box gene  
293 *AGAMOUS* (AG) regulates carpel development. These basic regulators of flower  
294 development are also conserved in other angiosperms (Becker, 2003; Zhao et al., 2017).  
295 Further, 491 genes have been suggested to be involved in soybean flower development  
296 (Jung et al., 2012).

297           Recently, several studies reported transcriptional changes during flowering time  
298 in legumes (Weller and Ortega, 2015). We found 182 flower-specific genes, including at  
299 least 20 members of the *plant invertase/pectin methylesterase inhibitor* (PMEI)  
300 superfamily, which is involved in cell wall modification in *Ar. thaliana* (Zhao et al., 2015).  
301 Specific PMEIs are highly expressed in specific wheat floral parts, such as anthers and  
302 pollen tubes (Rocchi et al., 2012), playing a significant role in flower development (Wormit  
303 and Usadel, 2018). In addition, we found 20 flower-specific TFs, mostly from the MYB  
304 (40%, 8/20) and MIKC-type MADS (40%, 8/20) families. Finally, out of 8 these MIKC genes,  
305 two AGAMOUS-like (*Glyma.03G019400*, *Glyma.07G081300*) and three PISTILLATA  
306 (*Glyma.06G117600*, *Glyma.13G034100*, *Glyma.14G155100*) were among the 36 flower-  
307 specific genes reported by Jung *et al.* (Jung et al., 2012).  
308

### 309 **Identification of novel transcripts**

310           We compared the genomic coordinates of the transcripts assembled in our atlas  
311 with those available in Phytozome and categorized them in nine classes (Table 2). We  
312 found that 95% (70,963/74,490) of the transcripts precisely matched known transcripts  
313 (class =). We also investigated class-J and class-U categories, which account for 3,256 and  
314 23 transcripts, respectively. Class-J comprises multi-exon transcripts with at least one  
315 known exon junction, while class-U encompasses transcripts located in intergenic regions.  
316 While class-J transcripts include new isoforms of known genes, those from class-U are  
317 useful to identify potentially new genes. We found that 30% (983/3256) of the class-J  
318 transcripts and 17% (4/23) of the class-U transcripts had TPM  $\geq$  1 in 907 and 1,207  
319 samples, respectively. Only one of the four class-U expressed transcripts (TU4871,  
320 Chr02:12125821-12127123) encode a protein longer than 50 aa, which contains a reverse  
321 transcriptase-like RNase\_H (PF13456) domain, supporting that it is likely a mobile  
322 element. In two of these expressed class-U transcripts (TU28093, TU56508), only one  
323 exon showed high read coverage (Supplementary Figure 5).

324           All the 3,256 class-J transcripts were further analyzed for alternate splicing (AS)  
325 events using ASprofile (Florea et al., 2013). AS events were categorized in one of six  
326 categories: (i) exon-skipping; (ii) multiple exon-skipping; (iii) alternative transcription start  
327 site (TSS); (iv) alternative transcription termination sites (TTS); (v) intron retention and;  
328 (vi) alternate 5' and/or 3' exon ends. We detected 6,582 AS events, mostly TSS and TTS  
329 (Table 3). Several novel AS events were supported by hundreds of split reads  
330 (Supplementary Figure 6-8). For example, TU62356 from *Glyma.17G195900* (CASEIN  
331 KINASE 1-LIKE PROTEIN 4) is a novel isoform with a skipped exon (Supplementary Figure  
332 6). Interestingly, we found no support for this alternative isoform in other tissues.  
333

### 334 **Data availability through a user-friendly web interface**

335 We developed a simple user-friendly web interface to allow researchers to easily explore  
336 1,243 soybean transcriptome samples. Through this interface (Figure 8), one can explore  
337 the expression of a particular gene in multiple tissues, with the aid of an image illustrating  
338 all the available tissues. Alternatively, users can also retrieve expression profiles of  
339 multiple genes in batch, with multiple filtering options (e.g. by tissue, BioProject, study).  
340 The outputs can be exported as plain text files. We strongly believe that this website will  
341 optimize data reuse and help research groups in their own projects. This service can be  
342 freely accessed at <http://venanciogroup.uenf.br/resources/>.

343

### 344 **Conclusions**

345 We have culled a large collection of publicly available RNA-seq datasets to construct a  
346 transcriptome atlas in soybean. We implemented a pipeline with state-of-art methods to  
347 map and quantify gene expression levels in 16 different broad tissue categories. This atlas  
348 allowed us to identify constitutive and tissue-specific genes. The constitutively expressed  
349 genes might, for example, be used as reference genes in RT-qPCR experiments, whereas  
350 tissue-specific genes might help scientists test hypotheses in downstream experiments  
351 and functional genomics studies. To optimize data reuse, we elaborated a simple web  
352 interface to allow the community to quickly access and browse the collected data. We  
353 believe this atlas will be an invaluable resource not only for basic research projects, but  
354 also in the development of novel strategies to improve soybean productivity to meet  
355 increasing global food demands.

356

### 357 **Methods**

#### 358 **Soybean genome and annotation data**

359 Soybean genomic sequences and gene annotation data (assembly version:  
360 Gmax\_275\_Wm82.a2.v1) were obtained from Phytozome (Schmutz et al.,  
361 2010;Goodstein et al., 2012). The gene annotation file contained 56,044 and 88,647 genes  
362 and transcripts, respectively. The gene annotation file containing exon-intron boundaries  
363 (GFF3 format) was used as a reference guide in read mapping. We excluded 759  
364 overlapping genes from the analysis. The gene description file was used to obtain various  
365 annotations such as GO, KEGG, KOG, and *Arabidopsis* ortholog descriptions.

366

#### 367 **Soybean RNA-Seq data**

368 To identify soybean transcriptome sequencing projects, we searched the NCBI SRA  
369 database (<https://www.ncbi.nlm.nih.gov/sra>) and the metadata were exported by using

370 *Run selector* (<https://trace.ncbi.nlm.nih.gov/Traces/study/>). We also searched Soybean  
371 RNA-seq studies in the literature (up to May 2018) to find additional datasets. We  
372 enriched this list of studies with various other details, such as PubMed ID and experiment  
373 details obtained by using NCBI *e-fetch*. Using these metadata, we excluded miRNA/siRNA  
374 samples and a few other samples showing technical issues such as: i) empty FASTQ files;  
375 ii) paired-end samples with single-end reads and; iii) paired-end reads of unequal lengths.  
376 Collectively, we downloaded a total of 1,742 *.sra* files (Supplementary table S2), which  
377 were decompressed using *sra-toolkit* (v.2.5.7) (Leinonen et al., 2010).

378

### 379 **Preprocessing and quality control**

380 Quality assessment of FASTQ files was performed using FASTQC  
381 (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). Datasets were  
382 processed using Trimmomatic (v0.36) (Bolger et al., 2014) to remove reads with average  
383 base quality lower than 20 or containing adapter sequences. Library strandedness was  
384 determined with the *infer\_experiment.py* script from RSeQC (Wang et al., 2012) using a  
385 mapping of 20% of the reads of each sample to the soybean genome in a fast-forward  
386 manner using Bowtie2 (Langmead and Salzberg, 2012).

387

### 388 **Transcript assembly and gene expression estimation**

389 We aligned the reads to the *Gl. max* reference genome (Gmax\_275\_Wm82.a2.v1) by  
390 using STAR (v.2.5.3a) (Dobin et al., 2013) with default parameters, along with the soybean  
391 gene annotation file containing exon-intron boundaries (in GFF3). When required, STAR  
392 also splits reads to find novel exon-intron boundaries or splice sites. The log files were  
393 processed to obtain read mapping statistics. Next, StringTie (v. 1.3.4) (Pertea et al., 2015)  
394 was used to assemble transcripts and estimate normalized gene expression. We  
395 performed transcriptome assemblies for each of the 16 tissues separately. In StringTie,  
396 we set the following parameters: i) at least 5 reads with at least 25% of the total read  
397 length covering both sides of an exon junction boundary (`-j 5 -a 0.25*read_length`); ii)  
398 average read depth for a transcript of at least 10 (`-c 10`) and; iii) library strandedness,  
399 when applicable. The resulting 16 assembled transcript annotations from each tissue  
400 were combined with TACO v0.7.3 (Niknafs et al., 2017). GffCompare (v0.10.5)  
401 (<https://ccb.jhu.edu/software/stringtie/gffcompare.shtml>) was used to compare  
402 assembled and reference transcripts. Further, *featureCount* (subread-v1.6.2) (Liao et al.,  
403 2014) was used to count the number of reads per feature at transcript and gene levels,  
404 while normalized expression was estimated in TPM using StringTie (`-e` option).

405

## 406 **Sample clustering**

407 We assessed the sample clustering patterns by submitting 41,011 genes with mean  $\log_2$   
408 (read count+1)  $\geq 1$  to: i) hierarchical clustering; ii) t-SNE clustering and; iii) *K-means*  
409 clustering. These analyses were performed using R functions ([www.r-project.org](http://www.r-project.org)) *cor()*,  
410 *hclust()*, and *kmeans()*. For t-SNE clustering, we used the *t-SNE* R package (Krijthe, 2015)  
411 with clustering parameters *max\_iter*= 5000 and *perplexity*= 50. For hierarchical clustering,  
412 sample dissimilarity ( $1 - \text{Pearson Correlation Coefficients}$ ) values were used to infer  
413 pairwise sample distances. The resulting tree was inspected for unexpected sample  
414 clustering patterns. t-SNE separated samples in 35 sub-clusters. Thus, we ran the *K-means*  
415 clustering analysis to find 35 centroids (*k*= 35).

416

## 417 **Identification of novel genes and splicing isoforms**

418 To identify novel genes and isoforms, we analyzed the GffCompare output files.  
419 Transcripts not overlapping with any known reference transcript were assigned to class  
420 U. The nucleotide sequences of the class U transcripts were extracted and translated  
421 using TransDecoder (v. 3.0.1). Protein domains were predicted using HMMER3 v. 3.1b2  
422 (all default parameters except domain e-value < 0.01) ([hmmer.org](http://hmmer.org)) and the Pfam  
423 database (release 32.0) (El-Gebali et al., 2019). Read coverage of these novel genes were  
424 visualized with Gbrowse, available on Soybase  
425 (<https://soybase.org/gb2/gbrowse/gmax2.0>). Class-J transcripts were classified as  
426 putative novel isoforms. Splice junctions of these transcripts in GTF format were  
427 compared against all known splice junctions using ASprofile v.b-1.0.4 (Florea et al., 2013).  
428 The number of reads supporting a splice junction was visualized as sashimi plots using  
429 Integrated Genome Viewer (v2.4.10)(Robinson et al., 2011).

430

## 431 **Analysis of the top 1000 highest expressed gene lists**

432 The top 1000 genes with the greatest average TPM in each tissue category were analyzed  
433 using MapMan (v3.5.1R2) (Thimm et al., 2004). To assign pathway bins, amino acid  
434 sequences of these gene lists were compared against *Arabidopsis* peptide database using  
435 Mercator4 (v. 2.0) (Schwacke et al., 2019).

436

## 437 **Identification of housekeeping genes**

438 We selected 11 tissues with at least 10 samples, which resulted in a total of 1,225 samples.  
439 The variability in gene expression was evaluated as previously described (Hoang et al.,  
440 2017). The following criteria were applied to identify HK genes:

- 441 i. A gene with TPM < 1 in a given sample was considered as not expressed (these  
442 TPM values were set to 0);
- 443 ii. Genes must be expressed in all 1,225 samples. This step resulted in 1,809  
444 genes;
- 445 iii. The mean TPM of each gene was calculated by taking the average of the gene  
446 expression across all samples;
- 447 iv. The Coefficient of Variation (CoV) was computed by taking the standard  
448 deviation divided by the mean expression of a gene;
- 449 v. The ratio of the maximum to minimum (MFC) was calculated by dividing the  
450 largest by the smallest TPM value. A product score (MFC-CoV) was calculated  
451 based on the product of CoV and MFC for each gene;
- 452 vi. Genes with MFC-CoV scores within the 1<sup>st</sup> quartile were classified as HK genes.

453

454 HK genes were also analyzed using the tissue-specificity index  $\tau$  (Yanai et al.,  
455 2004;Kryuchkova-Mostacci and Robinson-Rechavi, 2017). The  $\tau$  values ranged from 0  
456 (broad expression) to 1 (exclusive expression).  $\tau$  for each gene was calculated by using the  
457 formula:

458

459 
$$\tau = \frac{\sum_{i=1}^n (1 - \hat{x}_i)}{n - 1}; \hat{x}_i = \frac{x_i}{\max_{1 \leq i \leq n} (x_i)}$$

460 where

461  $x_i$  = expression of the gene in tissue  $i$ .

462  $n$  = number of tissues.

463

#### 464 **Assessment of tissue-specific expression**

465 We used the  $\log_2$  transformed TPM values for this analysis. Each of the 12 tissues was  
466 compared against each other (a total of 144 comparisons) to find significantly over-  
467 expressed genes using *limma* (Ritchie et al., 2015). We used  $\log_2$  (fold-change)  $\geq 2$  and  
468 adjusted p-value  $\leq 0.05$  (moderated t-statistic) to identify significantly over-expressed  
469 genes. If a gene  $G$  is over-expressed in a tissue  $T$  in comparison to the other 11 tissues,  $G$   
470 was considered as specifically expressed in  $T$ . We also used  $\tau$  to assess tissue-specific  
471 expression by applying a minimum threshold of 0.8, as previously recommended  
472 (Kryuchkova-Mostacci and Robinson-Rechavi, 2017).

473

#### 474 **Gene orthologs and enrichment tests**

475 We obtained the gene descriptions from Phytomine  
476 (<https://phytozome.jgi.doe.gov/phytomine/begin.do>), which is an InterMine (Lyne et al.,

477 2015) interface to genomic data from Phytozome (Goodstein et al., 2012). We used  
478 Phytomine to assess the conservation of HK genes in 14 different species (*Ph. vulgaris*,  
479 *Me. truncatula*, *Vigna unguiculata*, *Ar. thaliana*, *Oryza sativa*, *Gossypium raimondii*,  
480 *Carica papaya*, *Vitis vinifera*, *Sorghum bicolor*, *Zea mays*, *Amborella trichopoda*,  
481 *Selaginella moellendorffii*, *Physcomitrella. Patens*, and *Volvox carteri*). To estimate the  
482 conservation of non-HK genes, we created 5 sets of 452 randomly selected genes from  
483 the 55,592 non-HK genes. Each of these sets were searched for orthologs in the above  
484 mentioned 14 species. GO enrichment was performed on Phytomine (corrected p-value  
485 < 0.05). We performed Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway  
486 enrichment using KOBAS 3.0 (Ai and Kong, 2018). We used the Fisher's Exact test to assess  
487 the enrichment of essential genes and TFs in particular gene sets. The list of 510  
488 *Arabidopsis* EMBRYO-DEFECTIVE (EMB) genes (Meinke, 2019) were searched on  
489 Phytomine and the corresponding 1,010 soybean orthologs were retrieved. The list of  
490 soybean TFs was obtained from a recently published work (Moharana and Venancio,  
491 2019).

492

#### 493 **Web server**

494 The TPM and read count values for 54,877 genes across 1243 samples were stored in a  
495 relational database implemented in MySQL and hosted on an Apache HTTP web server.  
496 The front-end to this database was developed using Python/html/CSS. Interactive  
497 visualizations were implemented using *D3.js* (<https://d3js.org/>) and *Plotly.js*  
498 (<https://plot.ly/>) javascript libraries. The online server is publicly available at  
499 <http://venanciogroup.uenf.br/resources/>.

500

#### 501 **Acknowledgements**

502 This work was supported by Fundação Carlos Chagas Filho de Amparo à Pesquisa do  
503 Estado do Rio de Janeiro (FAPERJ; grants E-26/010.002019/2014, E-26/102.259/2013, and  
504 E-26/203.014/2018), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior -  
505 Brasil (CAPES; Finance Code 001), and Conselho Nacional de Desenvolvimento Científico  
506 e Tecnológico (CNPq). The funding agencies had no role in the design of the study and  
507 collection, analysis, and interpretation of data and in writing.

508

#### 509 **References**

510 Ai, C., and Kong, L. (2018). CGPS: A machine learning-based approach integrating multiple gene  
511 set analysis tools for better prioritization of biologically relevant pathways. *Journal of*  
512 *Genetics and Genomics* 45, 489-504.10.1016/j.jgg.2018.08.002

- 513 Becker, A. (2003). The major clades of MADS-box genes and their role in the development and  
514 evolution of flowering plants. *Molecular Phylogenetics and Evolution* 29, 464-  
515 489.10.1016/s1055-7903(03)00207-0
- 516 Belamkar, V., Weeks, N.T., Bharti, A.K., Farmer, A.D., Graham, M.A., and Cannon, S.B. (2014).  
517 Comprehensive characterization and RNA-Seq profiling of the HD-Zip transcription factor  
518 family in soybean (*Glycine max*) during dehydration and salt stress. *BMC Genomics* 15,  
519 950.10.1186/1471-2164-15-950
- 520 Bellieny-Rabelo, D., De Oliveira, E.A., Ribeiro, E.S., Costa, E.P., Oliveira, A.E., and Venancio, T.M.  
521 (2016). Transcriptome analysis uncovers key regulatory and metabolic aspects of soybean  
522 embryonic axes during germination. *Sci Rep* 6, 36009.10.1038/srep36009
- 523 Benedito, V.A., Torres-Jerez, I., Murray, J.D., Andriankaja, A., Allen, S., Kakar, K., Wandrey, M.,  
524 Verdier, J., Zuber, H., Ott, T., Moreau, S., Niebel, A., Frickey, T., Weiller, G., He, J., Dai, X.,  
525 Zhao, P.X., Tang, Y., and Udvardi, M.K. (2008). A gene expression atlas of the model  
526 legume *Medicago truncatula*. *The Plant Journal* 55, 504-513.10.1111/j.1365-  
527 313X.2008.03519.x
- 528 Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina  
529 sequence data. *Bioinformatics* 30, 2114-2120.10.1093/bioinformatics/btu170
- 530 Bowman, J.L., Smyth, D.R., and Meyerowitz, E.M. (2012). The ABC model of flower development:  
531 then and now. *Development* 139, 4095-4098.10.1242/dev.083972
- 532 Chi, Y., Wang, T., Xu, G., Yang, H., Zeng, X., Shen, Y., Yu, D., and Huang, F. (2017). GmAGL1, a  
533 MADS-Box Gene from Soybean, Is Involved in Floral Organ Identity and Fruit Dehiscence.  
534 *Frontiers in Plant Science* 8.10.3389/fpls.2017.00175
- 535 Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., Mcpherson, A., Szczesniak,  
536 M.W., Gaffney, D.J., Elo, L.L., Zhang, X., and Mortazavi, A. (2016). A survey of best practices  
537 for RNA-seq data analysis. *Genome Biology* 17.10.1186/s13059-016-0881-8
- 538 Czechowski, T., Stitt, M., Altmann, T., Udvardi, M.K., and Scheible, W.R. (2005). Genome-wide  
539 identification and testing of superior reference genes for transcript normalization in  
540 *Arabidopsis*. *Plant Physiol* 139, 5-17.10.1104/pp.105.063743
- 541 Dash, S., Van Hemert, J., Hong, L., Wise, R.P., and Dickerson, J.A. (2012). PLEXdb: gene expression  
542 resources for plants and plant pathogens. *Nucleic Acids Res* 40, D1194-  
543 1201.10.1093/nar/gkr938
- 544 Day, R.C., Herridge, R.P., Ambrose, B.A., and Macknight, R.C. (2008). Transcriptome Analysis of  
545 Proliferating *Arabidopsis* Endosperm Reveals Biological Implications for the Control of  
546 Syncytial Division, Cytokinin Signaling, and Gene Expression Regulation. *Plant Physiology* 148,  
547 1964-1984.10.1104/pp.108.128108
- 548 Deshmukh, R., Sonah, H., Patil, G., Chen, W., Prince, S., Mutava, R., Vuong, T., Valliyodan, B., and  
549 Nguyen, H.T. (2014). Integrating omic approaches for abiotic stress tolerance in soybean.  
550 *Frontiers in Plant Science* 5.10.3389/fpls.2014.00244
- 551 Dey, K.K., Hsiao, C.J., and Stephens, M. (2017). Visualizing the structure of RNA-seq expression  
552 data using grade of membership models. *PLoS Genet* 13,  
553 e1006599.10.1371/journal.pgen.1006599



- 554 Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and  
555 Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15-  
556 21.10.1093/bioinformatics/bts635
- 557 El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson,  
558 L.J., Salazar, G.A., Smart, A., Sonnhammer, E.L I., Hirsh, L., Paladin, L., Piovesan, D., Tosatto,  
559 S.C e., and Finn, R.D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Res*  
560 47, D427-D432.10.1093/nar/gky995
- 561 Figueroa, P. (2005). Arabidopsis Has Two Redundant Cullin3 Proteins That Are Essential for  
562 Embryo Development and That Interact with RBX1 and BTB Proteins to Form Multisubunit E3  
563 Ubiquitin Ligase Complexes in Vivo. *The Plant Cell Online* 17, 1180-  
564 1195.10.1105/tpc.105.031989
- 565 Florea, L., Song, L., and Salzberg, S.L. (2013). Thousands of exon skipping events differentiate  
566 among splicing patterns in sixteen human tissues. *F1000Research* 2,  
567 188.10.12688/f1000research.2-188.v1
- 568 Fucile, G., Di Biase, D., Nahal, H., La, G., Khodabandeh, S., Chen, Y., Easley, K., Christendat, D.,  
569 Kelley, L., and Provart, N.J. (2011). ePlant and the 3D data display initiative: integrative  
570 systems biology on the world wide web. *PLoS One* 6, e15237.10.1371/journal.pone.0015237
- 571 Garg, R., and Jain, M. (2013). Transcriptome Analyses in Legumes: A Resource for Functional  
572 Genomics. *The Plant Genome* 6, 0.10.3835/plantgenome2013.04.0011
- 573 Gazara, R.K., De Oliveira, E.a.G., Rodrigues, B.C., Nunes Da Fonseca, R., Oliveira, A.E.A., and  
574 Venancio, T.M. (2019). Transcriptional landscape of soybean (*Glycine max*) embryonic axes  
575 during germination in the presence of paclobutrazol, a gibberellin biosynthesis inhibitor. *Sci*  
576 *Rep* 9, 9601.10.1038/s41598-019-45898-2
- 577 Goettel, W., Xia, E., Upchurch, R., Wang, M.L., Chen, P., and An, Y.Q. (2014). Identification and  
578 characterization of transcript polymorphisms in soybean lines varying in oil composition and  
579 content. *BMC Genomics* 15, 299.10.1186/1471-2164-15-299
- 580 Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W.,  
581 Hellsten, U., Putnam, N., and Rokhsar, D.S. (2012). Phytozome: a comparative platform for  
582 green plant genomics. *Nucleic Acids Res* 40, D1178-1186.10.1093/nar/gkr944
- 583 Griesmann, M., Chang, Y., Liu, X., Song, Y., Haberer, G., Crook, M.B., Billault-Penneteau, B.,  
584 Lauressergues, D., Keller, J., Imanishi, L., Roswanjaya, Y.P., Kohlen, W., Pujic, P., Battenberg,  
585 K., Alloisio, N., Liang, Y., Hilhorst, H., Salgado, M.G., Hoher, V., Gherbi, H., Svistoonoff, S.,  
586 Doyle, J.J., He, S., Xu, Y., Xu, S., Qu, J., Gao, Q., Fang, X., Fu, Y., Normand, P., Berry, A.M., Wall,  
587 L.G., Ane, J.M., Pawlowski, K., Xu, X., Yang, H., Spannagl, M., Mayer, K.F.X., Wong, G.K.,  
588 Parniske, M., Delaux, P.M., and Cheng, S. (2018). Phylogenomics reveals multiple losses of  
589 nitrogen-fixing root nodule symbiosis. *Science* 361.10.1126/science.aat1743
- 590 He, J., Benedito, V.A., Wang, M., Murray, J.D., Zhao, P.X., Tang, Y., and Udvardi, M.K. (2009). The  
591 Medicago truncatula gene expression atlas web server. *BMC Bioinformatics* 10,  
592 441.10.1186/1471-2105-10-441
- 593 Hoang, V.L.T., Tom, L.N., Quek, X.C., Tan, J.M., Payne, E.J., Lin, L.L., Sinnya, S., Raphael, A.P.,  
594 Lambie, D., Frazer, I.H., Dinger, M.E., Soyer, H.P., and Prow, T.W. (2017). RNA-seq reveals

- 595 more consistent reference genes for gene expression studies in human non-melanoma skin  
596 cancers. *PeerJ* 5, e3631.10.7717/peerj.3631
- 597 Hu, R., Fan, C., Li, H., Zhang, Q., and Fu, Y.F. (2009). Evaluation of putative reference genes for  
598 gene expression normalization in soybean by quantitative real-time RT-PCR. *BMC Mol Biol*  
599 10, 93.10.1186/1471-2199-10-93
- 600 Iizumi, T., Luo, J.-J., Challinor, A.J., Sakurai, G., Yokozawa, M., Sakuma, H., Brown, M.E., and  
601 Yamagata, T. (2014). Impacts of El Niño Southern Oscillation on the global yields of major  
602 crops. *Nat Commun* 5.10.1038/ncomms4712
- 603 Jones, S.I., and Vodkin, L.O. (2013). Using RNA-Seq to profile soybean seed development from  
604 fertilization to maturity. *PLoS One* 8, e59270.10.1371/journal.pone.0059270
- 605 Jordan, I.K., Reeb, P.D., Bramardi, S.J., and Steibel, J.P. (2015). Assessing Dissimilarity Measures  
606 for Sample-Based Hierarchical Clustering of RNA Sequencing Data Using Plasmode Datasets.  
607 *PLoS One* 10, e0132310.10.1371/journal.pone.0132310
- 608 Jung, C.H., Wong, C.E., Singh, M.B., and Bhalla, P.L. (2012). Comparative genomic analysis of  
609 soybean flowering genes. *PLoS One* 7, e38250.10.1371/journal.pone.0038250
- 610 Kim, E., Hwang, S., and Lee, I. (2017). SoyNet: a database of co-functional networks for  
611 soybean *Glycine max*. *Nucleic Acids Res* 45, D1082-D1089.10.1093/nar/gkw704
- 612 Konishi, M., and Yanagisawa, S. (2013). Arabidopsis NIN-like transcription factors have a central  
613 role in nitrate signalling. *Nat Commun* 4, 1617.10.1038/ncomms2621
- 614 Krijthe, J.H. (2015). Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut  
615 Implementation.
- 616 Kryuchkova-Mostacci, N., and Robinson-Rechavi, M. (2017). A benchmark of gene expression  
617 tissue-specificity metrics. *Brief Bioinform* 18, 205-214.10.1093/bib/bbw008
- 618 Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods*  
619 9, 357-359.10.1038/nmeth.1923
- 620 Leinonen, R., Akhtar, R., Birney, E., Bonfield, J., Bower, L., Corbett, M., Cheng, Y., Demiralp, F.,  
621 Faruque, N., Goodgame, N., Gibson, R., Hoad, G., Hunter, C., Jang, M., Leonard, S., Lin, Q.,  
622 Lopez, R., Maguire, M., McWilliam, H., Plaister, S., Radhakrishnan, R., Sobhany, S., Slater, G.,  
623 Ten Hoopen, P., Valentin, F., Vaughan, R., Zalunin, V., Zerbino, D., and Cochrane, G. (2010).  
624 Improvements to services at the European Nucleotide Archive. *Nucleic Acids Res* 38, D39-  
625 45.10.1093/nar/gkp998
- 626 Li, W.V., and Li, J.J. (2018). Modeling and analysis of RNA-seq data: a review from a statistical  
627 perspective. *Quantitative Biology* 6, 195-209.10.1007/s40484-018-0144-7
- 628 Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for  
629 assigning sequence reads to genomic features. *Bioinformatics* 30, 923-  
630 930.10.1093/bioinformatics/btt656
- 631 Libault, M., Farmer, A., Joshi, T., Takahashi, K., Langley, R.J., Franklin, L.D., He, J., Xu, D., May, G.,  
632 and Stacey, G. (2010). An integrated transcriptome atlas of the crop model *Glycine max*, and  
633 its use in comparative analyses in plants. *Plant J* 63, 86-99.10.1111/j.1365-  
634 313X.2010.04222.x
- 635 Liu, P., and Si, Y. (2014). Cluster Analysis of RNA-Sequencing Data. 191-217.10.1007/978-3-319-  
636 07212-8\_10

- 637 Lyne, R., Sullivan, J., Butano, D., Contrino, S., Heimbach, J., Hu, F., Kalderimis, A., Lyne, M., N.  
638 Smith, R., Štěpán, R., Balakrishnan, R., Binkley, G., Harris, T., Karra, K., A. T. Moxon, S.,  
639 Motenko, H., Neuhauser, S., Ruzicka, L., Cherry, M., Richardson, J., Stein, L., Westerfield, M.,  
640 Worthey, E., and Micklem, G. (2015). Cross-organism analysis using InterMine. *genesis* 53,  
641 547-560.10.1002/dvg.22869
- 642 Marini, F., and Binder, H. (2019). pcaExplorer: an R/Bioconductor package for interacting with  
643 RNA-seq principal components. *BMC Bioinformatics* 20.10.1186/s12859-019-2879-1
- 644 Meinke, D.W. (2019). Genome-wide identification of EMBRYO-DEFECTIVE (EMB) genes required  
645 for growth and development in Arabidopsis. *New Phytol.*10.1111/nph.16071
- 646 Moharana, K.C., and Venancio, T.M. (2019). Polyploidization events shaped the transcription  
647 factor repertoires in legumes (Fabaceae). *bioRxiv*, 849778.10.1101/849778
- 648 Mortazavi, A., Williams, B.A., Mccue, K., Schaeffer, L., and Wold, B. (2008). Mapping and  
649 quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5, 621-  
650 628.10.1038/nmeth.1226
- 651 Niknafs, Y.S., Pandian, B., Iyer, H.K., Chinnaiyan, A.M., and Iyer, M.K. (2017). TACO produces  
652 robust multisample transcriptome assemblies from RNA-seq. *Nat Methods* 14, 68-  
653 70.10.1038/nmeth.4078
- 654 O’rourke, J.A., Graham, M.A., and Whitham, S.A. (2017). Soybean Functional Genomics: Bridging  
655 the Genotype-to-Phenotype Gap. 151-170.10.1007/978-3-319-64198-0\_10
- 656 Perteza, M., Perteza, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., and Salzberg, S.L. (2015).  
657 StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat*  
658 *Biotechnol* 33, 290-295.10.1038/nbt.3122
- 659 Po-Yen, W., Phan, J.H., Fengfeng, Z., and Wang, M.D. (2011). Evaluation of normalization methods  
660 for RNA-Seq gene expression estimation. 50-57.10.1109/bibmw.2011.6112354
- 661 Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers  
662 differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids*  
663 *Res* 43, e47-e47.10.1093/nar/gkv007
- 664 Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov,  
665 J.P. (2011). Integrative genomics viewer. *Nat Biotechnol* 29, 24-26.10.1038/nbt.1754
- 666 Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression  
667 analysis of RNA-seq data. *Genome Biology* 11, R25.10.1186/gb-2010-11-3-r25
- 668 Rocchi, V., Janni, M., Bellincampi, D., Giardina, T., and D'ovidio, R. (2012). Intron retention  
669 regulates the expression of pectin methyl esterase inhibitor (Pmei) genes during wheat  
670 growth and development. *Plant Biol (Stuttg)* 14, 365-373.10.1111/j.1438-8677.2011.00508.x
- 671 Roy, S., Liu, W., Nandety, R.S., Crook, A.D., Mysore, K.S., Pislariu, C.I., Frugoli, J.A., Dickstein, R.,  
672 and Udvardi, M.K. (2019). Celebrating 20 years of genetic discoveries in legume nodulation  
673 and symbiotic nitrogen fixation. *Plant Cell*, tpc.00279.02019.10.1105/tpc.19.00279
- 674 Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen,  
675 J.J., Cheng, J., Xu, D., Hellsten, U., May, G.D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya,  
676 M.K., Sandhu, D., Valliyodan, B., Lindquist, E., Peto, M., Grant, D., Shu, S., Goodstein, D.,  
677 Barry, K., Futrell-Griggs, M., Abernathy, B., Du, J., Tian, Z., Zhu, L., Gill, N., Joshi, T., Libault,  
678 M., Sethuraman, A., Zhang, X.C., Shinozaki, K., Nguyen, H.T., Wing, R.A., Cregan, P., Specht,

- 679 J., Grimwood, J., Rokhsar, D., Stacey, G., Shoemaker, R.C., and Jackson, S.A. (2010). Genome  
680 sequence of the palaeopolyploid soybean. *Nature* 463, 178-183.10.1038/nature08670
- 681 Schwacke, R., Ponce-Soto, G.Y., Krause, K., Bolger, A.M., Arsova, B., Hallab, A., Gruden, K., Stitt,  
682 M., Bolger, M.E., and Usadel, B. (2019). MapMan4: A Refined Protein Classification and  
683 Annotation Framework Applicable to Multi-Omics Data Analysis. *Molecular Plant* 12, 879-  
684 892.10.1016/j.molp.2019.01.003
- 685 Severin, A.J., Woody, J.L., Bolon, Y.T., Joseph, B., Diers, B.W., Farmer, A.D., Muehlbauer, G.J.,  
686 Nelson, R.T., Grant, D., Specht, J.E., Graham, M.A., Cannon, S.B., May, G.D., Vance, C.P., and  
687 Shoemaker, R.C. (2010). RNA-Seq Atlas of Glycine max: a guide to the soybean transcriptome.  
688 *BMC Plant Biol* 10, 160.10.1186/1471-2229-10-160
- 689 Sinharoy, S., Torres-Jerez, I., Bandyopadhyay, K., Kereszt, A., Pislariu, C.I., Nakashima, J., Benedito,  
690 V.A., Kondorosi, E., and Udvardi, M.K. (2013). The C2H2 transcription factor regulator of  
691 symbiosome differentiation represses transcription of the secretory pathway gene  
692 VAMP721a and promotes symbiosome development in *Medicago truncatula*. *Plant Cell* 25,  
693 3584-3601.10.1105/tpc.113.114017
- 694 Smalle, J., and Vierstra, R.D. (2004). The Ubiquitin 26s Proteasome Proteolytic Pathway. *Annual*  
695 *Review of Plant Biology* 55, 555-590.10.1146/annurev.arplant.55.031903.141801
- 696 Soltis, D.E., Chanderbali, A.S., Kim, S., Buzgo, M., and Soltis, P.S. (2007). The ABC Model and its  
697 Applicability to Basal Angiosperms. *Annals of Botany* 100, 155-163.10.1093/aob/mcm117
- 698 Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., Selbig, J., Müller, L.A., Rhee, S.Y.,  
699 and Stitt, M. (2004). mapman: a user-driven tool to display genomics data sets onto diagrams  
700 of metabolic pathways and other biological processes. *The Plant Journal* 37, 914-  
701 939.10.1111/j.1365-313X.2004.02016.x
- 702 Vernié, T., Moreau, S., De Billy, F., Plet, J., Combiér, J.-P., Rogers, C., Oldroyd, G., Frugier, F., Niebel,  
703 A., and Gamas, P. (2008). EFD Is an ERF Transcription Factor Involved in the Control of Nodule  
704 Number and Differentiation in *Medicago truncatula*. *Plant Cell* 20, 2696-  
705 2713.10.1105/tpc.108.059857
- 706 Wagner, G.P., Kin, K., and Lynch, V.J. (2012). Measurement of mRNA abundance using RNA-seq  
707 data: RPKM measure is inconsistent among samples. *Theory in Biosciences* 131, 281-  
708 285.10.1007/s12064-012-0162-3
- 709 Wang, L., Wang, S., and Li, W. (2012). RSeQC: quality control of RNA-seq experiments.  
710 *Bioinformatics* 28, 2184-2185.10.1093/bioinformatics/bts356
- 711 Weller, J.L., and Ortega, R.L. (2015). Genetic control of flowering time in legumes. *Frontiers in*  
712 *Plant Science* 6.10.3389/fpls.2015.00207
- 713 Wormit, A., and Usadel, B. (2018). The Multifaceted Role of Pectin Methyltransferase Inhibitors  
714 (PMEIs). *International Journal of Molecular Sciences* 19, 2878.10.3390/ijms19102878
- 715 Wu, Z., Wang, M., Yang, S., Chen, S., Chen, X., Liu, C., Wang, S., Wang, H., Zhang, B., Liu, H., Qin,  
716 R., and Wang, X. (2019). A global coexpression network of soybean genes gives insights into  
717 the evolution of nodulation in nonlegumes and legumes. *New*  
718 *Phytologist*.10.1111/nph.15845
- 719 Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-  
720 Saban, S., Safran, M., Domany, E., Lancet, D., and Shmueli, O. (2004). Genome-wide midrange

721 transcription profiles reveal expression level relationships in human tissue specification.  
722 *Bioinformatics* 21, 650-659.10.1093/bioinformatics/bti042  
723 Yim, A.K., Wong, J.W., Ku, Y.S., Qin, H., Chan, T.F., and Lam, H.M. (2015). Using RNA-Seq Data to  
724 Evaluate Reference Genes Suitable for Gene Expression Studies in Soybean. *PLoS One* 10,  
725 e0136343.10.1371/journal.pone.0136343  
726 Zhao, S., Zhang, Y., Gordon, W., Quan, J., Xi, H., Du, S., Von Schack, D., and Zhang, B. (2015).  
727 Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation  
728 of gene overlap. *BMC Genomics* 16.10.1186/s12864-015-1876-7  
729 Zhao, T., Holmer, R., De Bruijn, S., Angenent, G.C., Van Den Burg, H.A., and Schranz, M.E. (2017).  
730 Phylogenomic Synteny Network Analysis of MADS-Box Transcription Factor Genes Reveals  
731 Lineage-Specific Transpositions, Ancient Tandem Duplications, and Deep Positional  
732 Conservation. *Plant Cell* 29, 1278-1292.10.1105/tpc.17.00312  
733  
734

735 **Tables**

736

737 **Table 1:** Tissue-specific transcription factors.

Transcription factor family	Cotyledon	Endosperm	Flower	Hypocotyl	Leaves	Nodule	Pod	Root	Seed	Shoot	Suspensor	Total
MYB		5	8	7	2		1	2	1		1	27
ERF		1	1	10		3		3			2	20
C2H2		6		1		6	2	2			4	21
NAC				2		1			1		4	8
bHLH	2	1		2				4				9
WRKY				6				2			2	10
MYB_related		2	1	1								4
LBD			1					1			1	3
G2-like	1	1						1				3
NF-YB		1				2						3
M-type		2				1						3
MIKC			8					1				9
HD-ZIP		2									2	4
GRAS				1		2						3
bZIP		2				4						6
B3		2									2	4
AP2						2					1	3
ZF-HD		2										2
YABBY			1									1
WOX											3	3
SRS						1						1
SBP										1		1
NZZ/SPL		2										2
Nin-like						6						6
NF-YC		3										3
NF-YA						1						1
HSF				1								1
GRF										1		1
GATA	1											1
Dof				1								1
CPP		1										1
C3H		3										3
<b>Total</b>	<b>4</b>	<b>36</b>	<b>20</b>	<b>32</b>	<b>2</b>	<b>29</b>	<b>3</b>	<b>16</b>	<b>2</b>	<b>2</b>	<b>22</b>	<b>168</b>

738

739






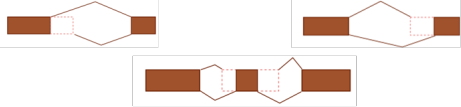
740 **Table 2:** Number of transcripts in each transcript-classification code defined by  
 741 GffCompare.

Class code	Description	# of transfrags
=	Complete, exact match of intron chain	70,963
j	Multi-exon with at least one exon junction match	3256
c	Contained in reference (intron compactable)	78
e	Single exon transfrag partially covering intron, possible pre-mRNA	70
k	Containment of reference (reverse containment)	69
u	Unknown, intergenic	23
o	Other same strand overlap with reference exon	23
x	Exonic overlap on opposite strand	4
p	Possible polymerase run-on (no actual overlap)	4

742

743

744 **Table 3:** Number of alternative splicing events (AS). The first column illustrates the  
 745 possible AS isoforms. The boxes represent exons and lines connect adjacent exons in the  
 746 mature transcript.

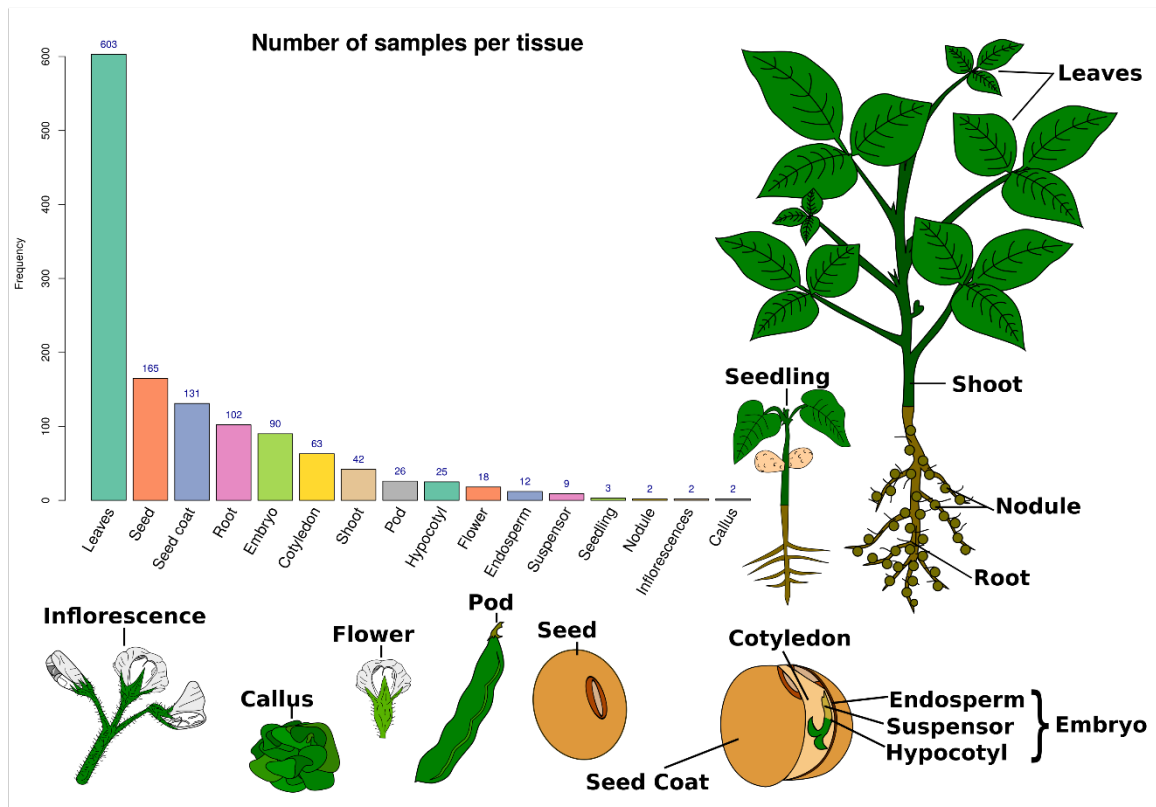
Exon junctions	Event type	Number of events
	Exon skipping (SKIP)	218
	Multiple exon skipping (MSKIP)	40
	Retention of single or multiple introns (IR/MIR)	190
	Alternative transcript start (TSS)	2831
	Alternative transcript termination (TTS)	2761
	Alternative exon ends (AE)	542
	Total	6582

747

748

749 **Figures**

750



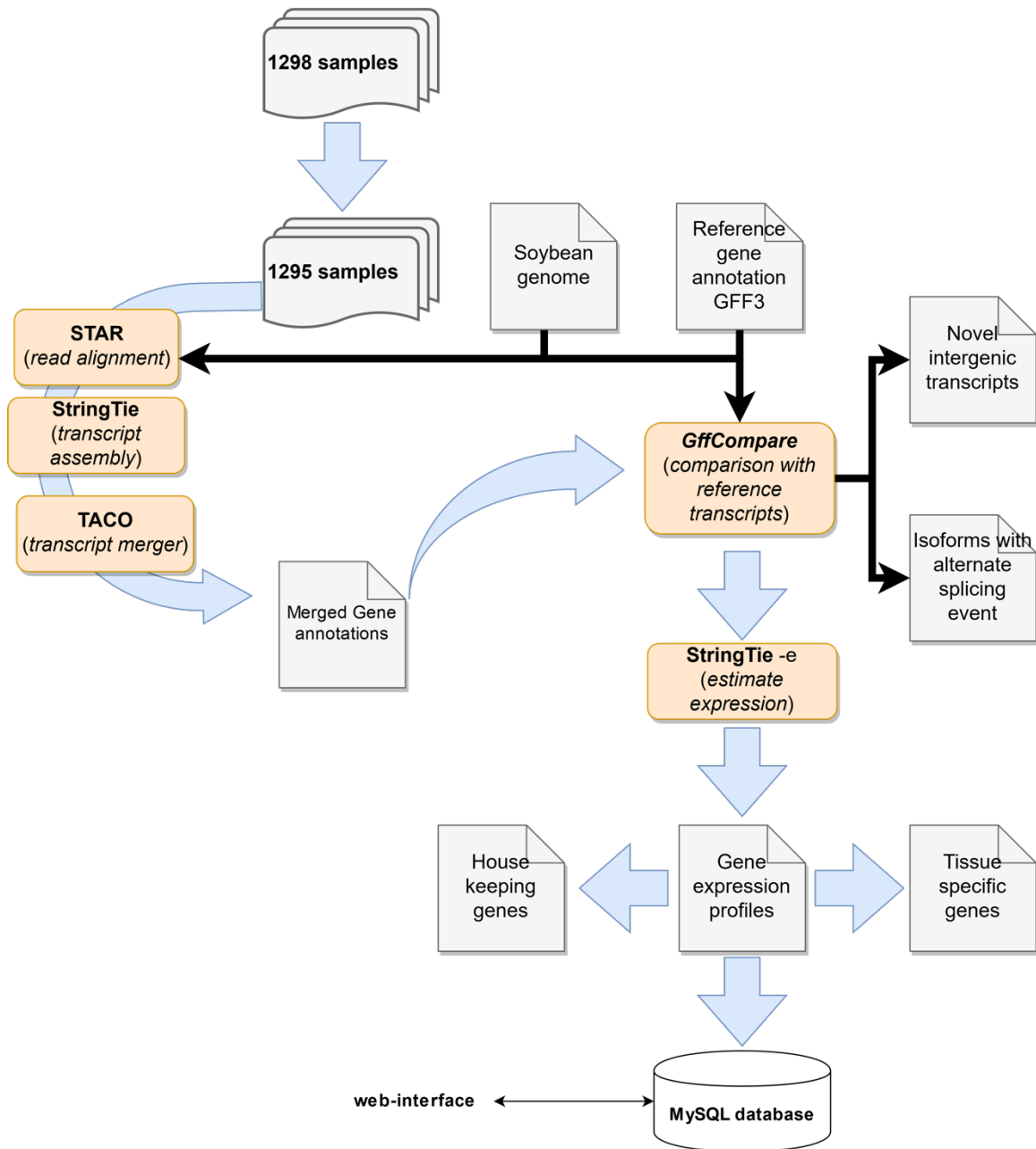
751

752

753 Figure 1: Number of samples analyzed in this study and a graphical representation of each  
754 tissue.

755

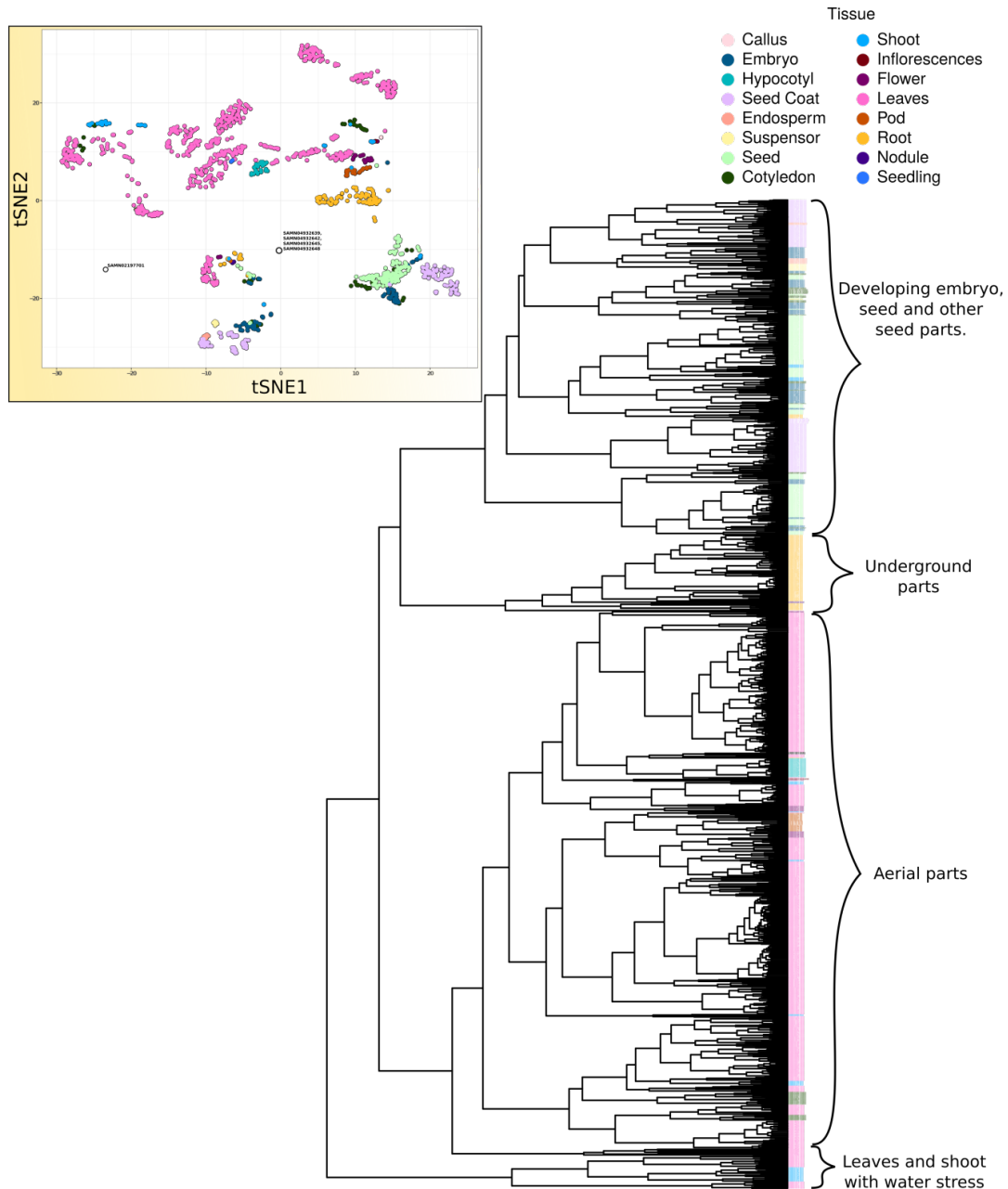




756  
757

758 Figure 2: Pipeline used to create the soybean RNA-Seq atlas.

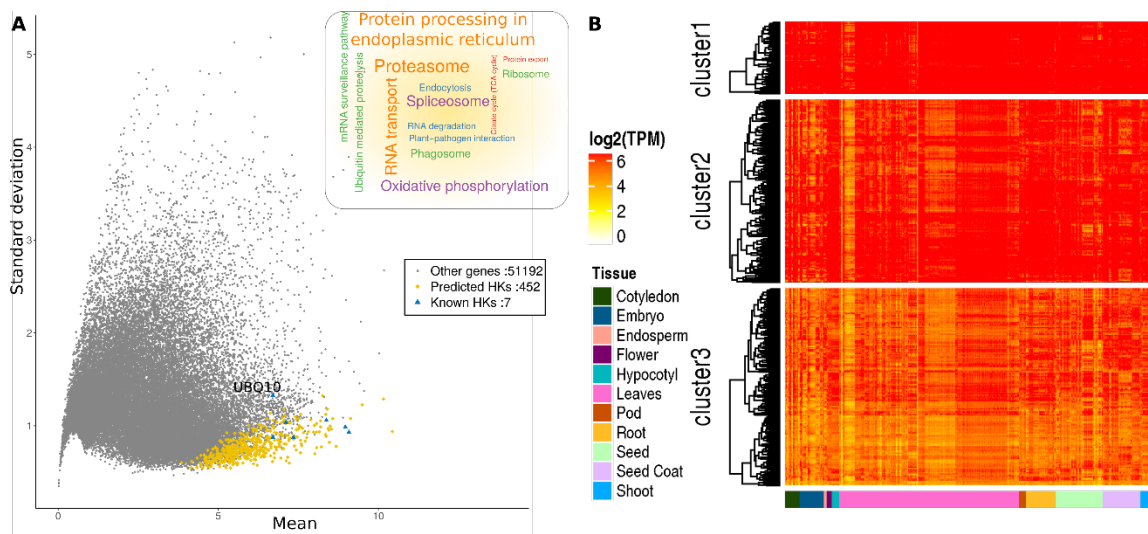
759  
760



761

762 Figure 3: Hierarchical clustering of samples using their transcriptional profiles. Per gene  
763 raw read counts were used to perform hierarchical clustering using the R function *hclust()*  
764 with default parameters. Samples were grouped into three major clades: aerial,  
765 underground, and seed-embryo related. A minor group of samples containing drought-  
766 stress-related leaves and shoots was also identified. The upper-left panel shows the  
767 sample clustering using t-SNE. Five samples (four from shoot: SAMN04932642,  
768 SAMN04932648, SAMN04932639, SAMN04932645 and one from root: SAMN02197701),  
769 labeled in the inside plot, showed a very unexpected clustering patterns and were  
770 excluded from further analysis. An interactive 3D version of the t-SNE sample clustering  
771 is available at <http://venanciogroup.uenf.br/resources/>.

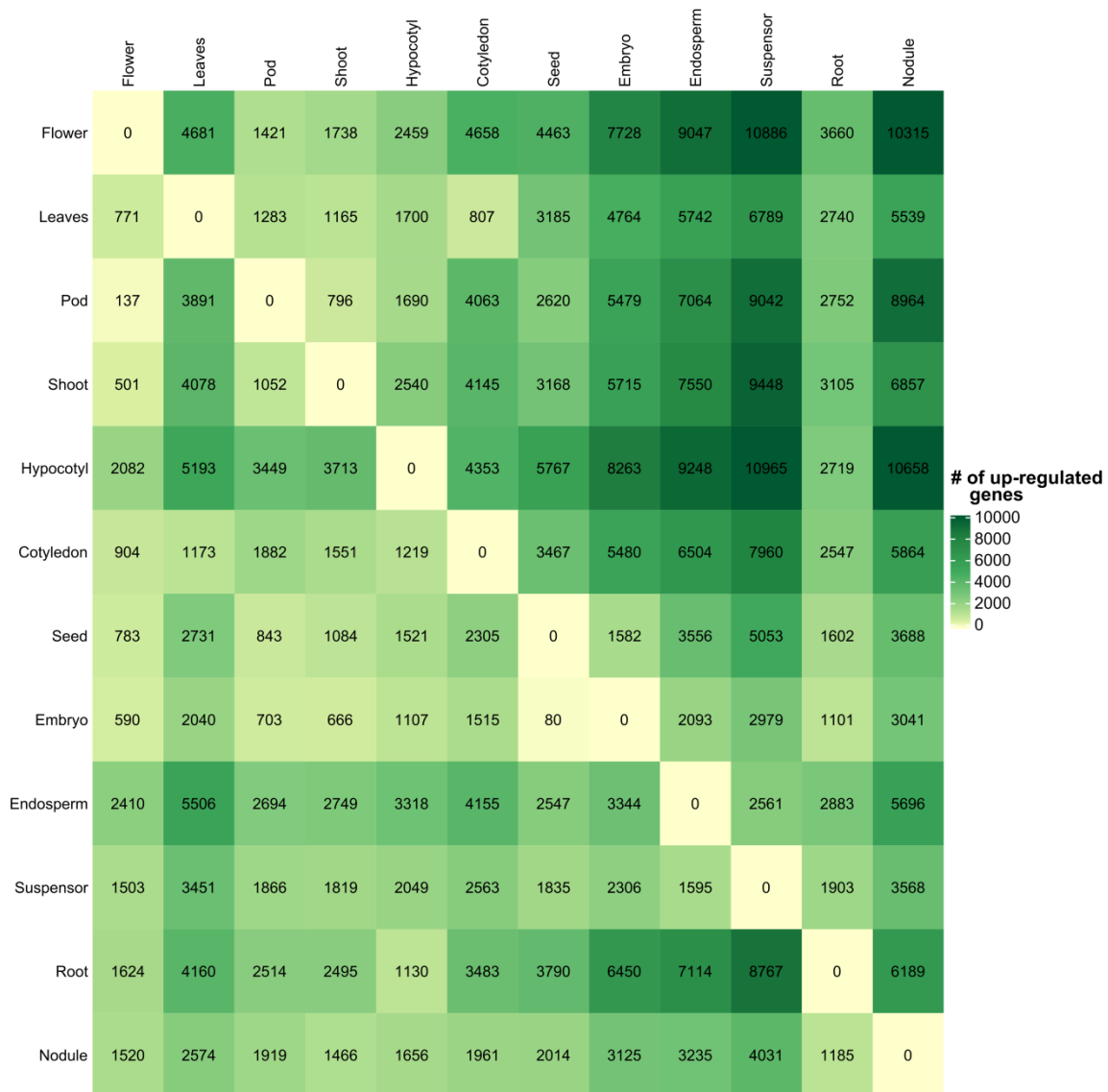
772



773

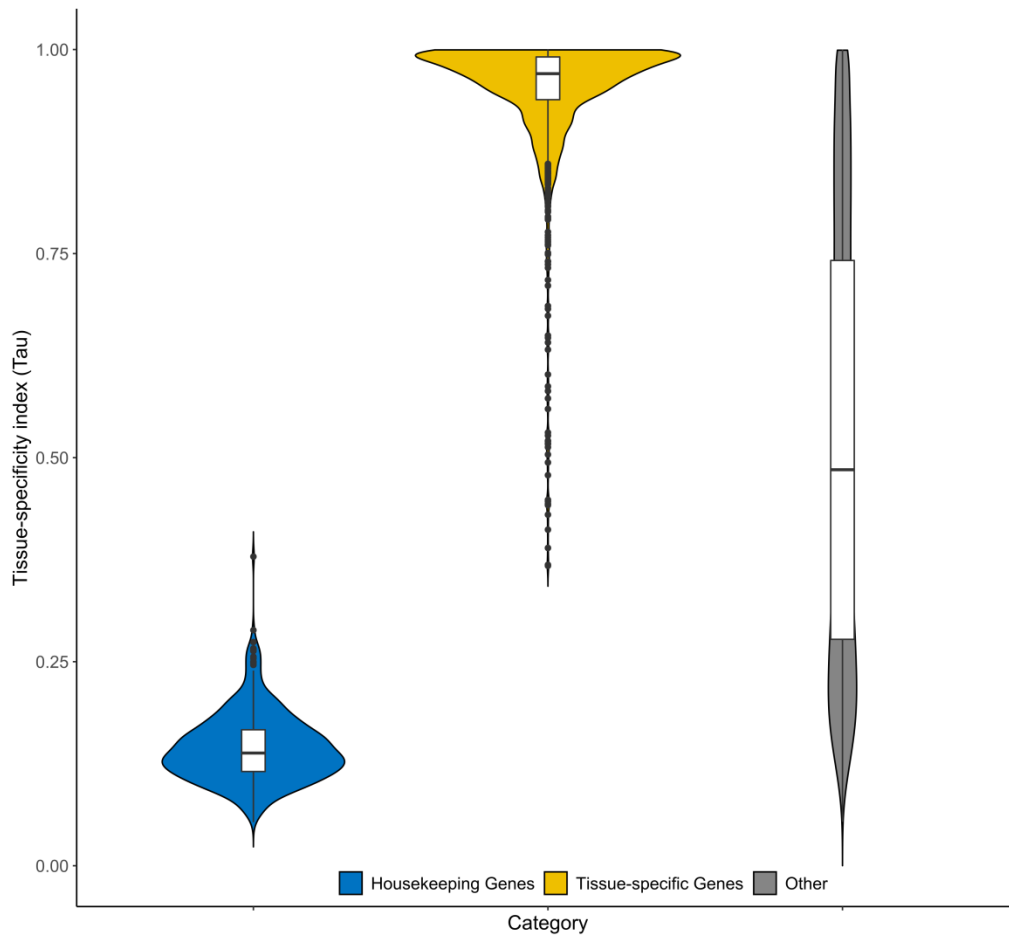
774 Figure 4: Global gene expression patterns of the housekeeping genes. A. Scatter plot of  
775 mean vs standard deviation showing uniform and stable expression of 452 housekeeping  
776 (HK) genes. The gray dots represent all the non-HK expressed genes (TPM  $\geq$  1 in at least  
777 one sample). The word cloud represents KEGG pathways enriched in HK genes (p-value <  
778 0.05). B. Global expression patterns of HK genes. Three main clusters were found with K-  
779 means clustering, which were then hierarchically clustered.

780



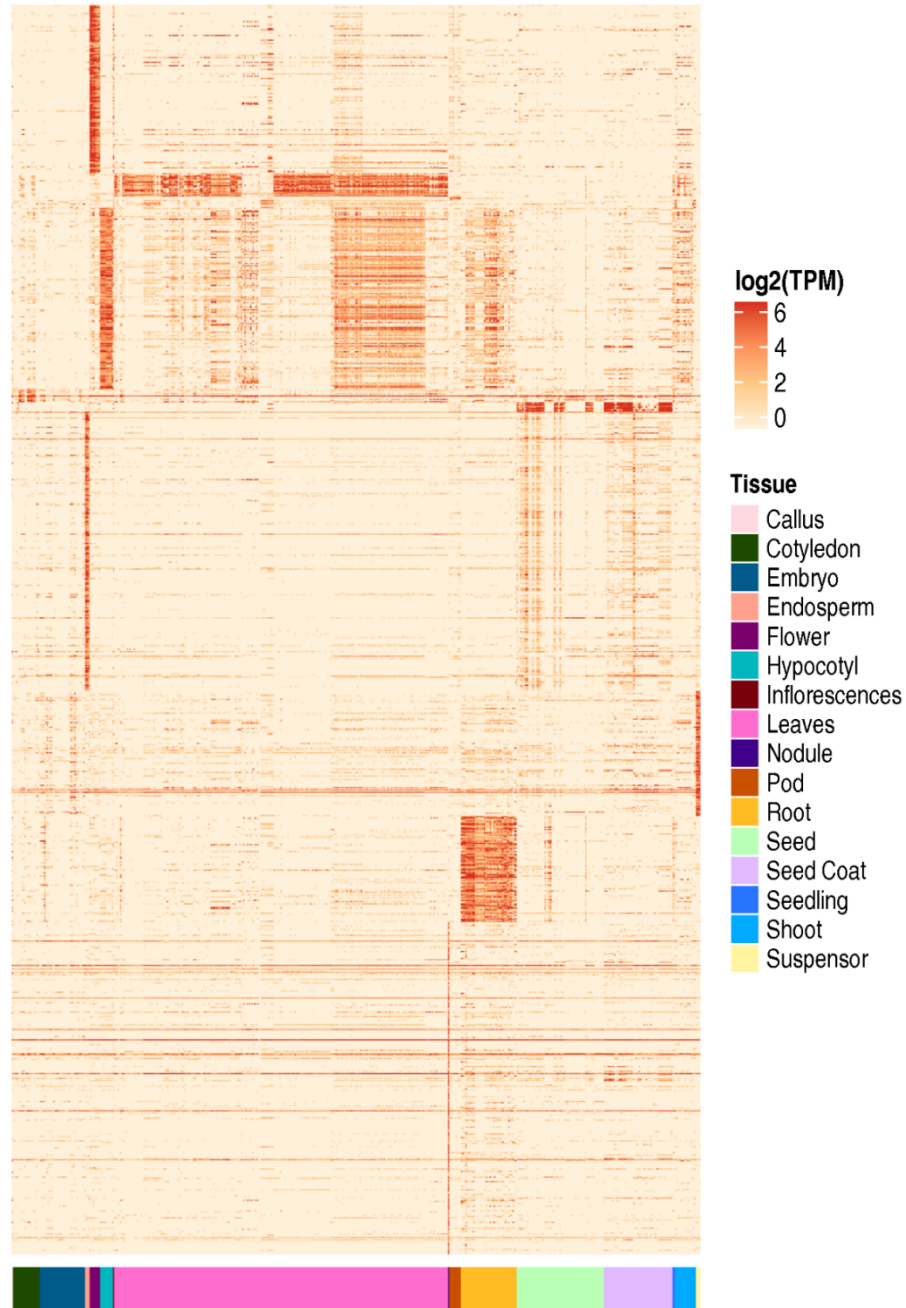
781

782 Figure 5: Heatmap showing the number of up-regulated genes in the tissues from the  
 783 rows when compared with those from the columns. Gene up-regulation was determined  
 784 by using a  $\log_2$  (fold-change)  $\geq 2$  and adjusted p-value  $\leq 0.05$  using the moderated t-  
 785 statistic in the *limma* package.



786

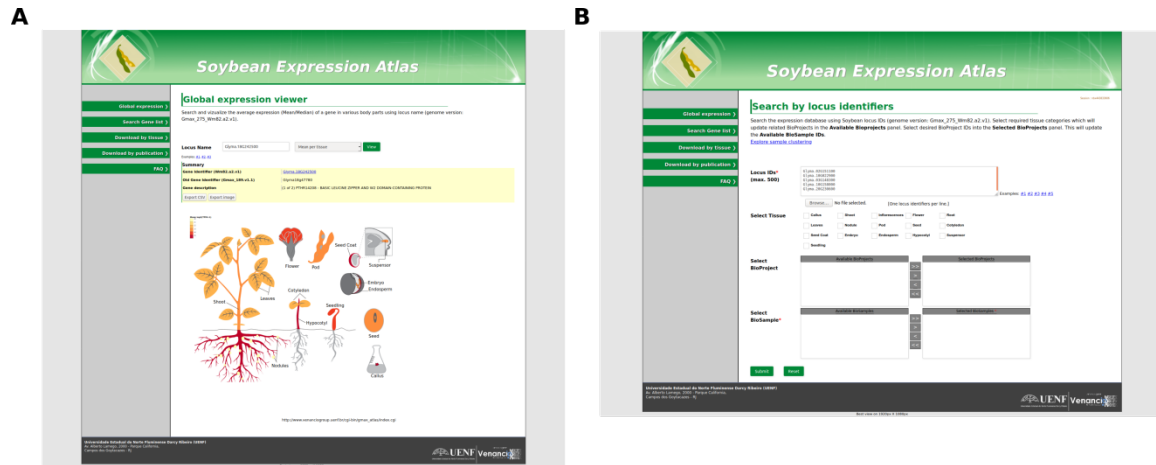
787 Figure 6: Violin plot showing the distribution of Tau indexes of housekeeping, tissue-specific, and  
788 the remaining genes. Tau values range between 0 and 1, with low values indicating a stable and  
789 constitutive expression and higher values supporting tissue-specificity.



790

791 Figure 7: Global transcriptional patterns of tissue-specific genes. Expression values are

792 represented as  $\log_2(\text{TPM})$  values in 1243 samples.



793

794 Figure 8: Web interface to browse and download the expression data analyzed in this  
795 study. A. Users can search, visualize and download average expression levels in each  
796 tissue or; B retrieve expression values in batch in particular samples, tissues, or  
797 BioProjects. This resource is available at: <http://venanciogroup.uenf.br/resources/>.