

ANÁLISE COMPUTACIONAL INTEGRATIVA DE DADOS
TRANSCRIPTÔMICOS DE SOJA (*GLYCINE MAX*)

FABRICIO BRUM MACHADO

CAMPOS DOS GOYTACAZES – RJ
DEZEMBRO 2020

ANÁLISE COMPUTACIONAL INTEGRATIVA DE DADOS
TRANSCRIPTÔMICOS DE SOJA (*GLYCINE MAX*)

FABRICIO BRUM MACHADO

Tese apresentada ao Centro de Biociências e Biotecnologia, da Universidade Estadual do Norte Fluminense Darcy Ribeiro, como parte das exigências para obtenção do título de Doutor em Biotecnologia Vegetal

CAMPOS DOS GOYTACAZES – RJ
DEZEMBRO 2020

FICHA CATALOGRÁFICA

UENF - Bibliotecas

Elaborada com os dados fornecidos pelo autor.

M149

Machado, Fabricio Brum.

Análise computacional integrativa de dados transcriptômicos de soja(*Glycine Max*) /
Fabricio Brum Machado. - Campos dos Goytacazes, RJ, 2020.

79 f. : il.

Inclui bibliografia.

Tese (Doutorado em Biotecnologia Vegetal) - Universidade Estadual do Norte Fluminense
Darcy Ribeiro, Centro de Biociências e Biotecnologia, 2020.

Orientador: Thiago Motta Venancio.

1. Atlas transcricional. 2. expressão gênica. 3. RNA-seq. 4. *splicing*. 5. gene de referência..
I. Universidade Estadual do Norte Fluminense Darcy Ribeiro. II. Título.

CDD - 660.6

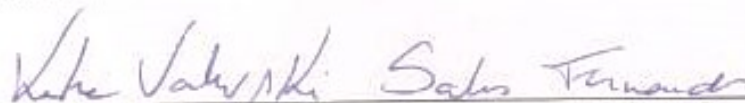
ANÁLISE COMPUTACIONAL INTEGRATIVA DE DADOS
TRANSCRIPTÔMICOS DE SOJA (GLYCINE MAX)

FABRICIO BRUM MACHADO

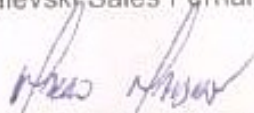
Tese apresentada ao Centro de Biociências e
Biotecnologia, da Universidade Estadual do
Norte Fluminense Darcy Ribeiro, como parte
das exigências para obtenção do título de
Doutor em Biotecnologia Vegetal

Aprovada em 17 de dezembro de 2020.

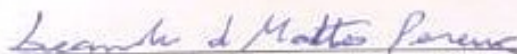
Comissão examinadora:



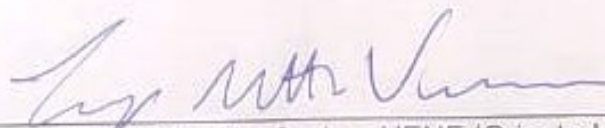
Dr^a. Kátia Valevski Sales Fernandes- UENF



Dr. Marcelo Maraschin – UFSC



Dr. Leandro de Mattos Pereira – Databiomics



Dr. Thiago Motta Venâncio – UENF (Orientador)

DEDICATÓRIA

Aos meus queridos pais Antonio Celso e Toni, também aos meus irmãos Filipe e Carolina pelo amor incondicional e por estarem sempre ao meu lado me dando força".

DEDICO.

AGRADECIMENTOS

Gostaria de agradecer primeiramente a Deus, pela vida e por me proporcionar estar apresentando esta tese, além de todas as pessoas e toda a equipe envolvida em minha recuperação.

Aos meus pais Celso e Ioni, meus irmãos Filipe e Carolina por todo o amor.

À minha companheira, Tatiane Sanches, pelo apoio nesses mais de 15 anos de caminhada.

Ao meu orientador, Thiago Venâncio, por me receber em seu laboratório, acreditar em mim, me ajudar e apoiar durante todo esse caminho. Meu muito obrigado, por todo apoio nos momentos que mais precisei.

Aos meus amigos do Laboratório LQFPP, que fizeram essa caminhada mais leve e principalmente pela amizade, em especial aos envolvidos na publicação desse trabalho: Kanhu, Fabrício, Raj e Francisnei.

Gostaria de agradecer a todos do PPGBV, à Universidade Estadual do Norte Fluminense Darcy Ribeiro (UENF), por me proporcionar a oportunidade de realizar o Doutorado em Biotecnologia Vegetal, fornecendo todo o suporte estrutural e acadêmico.

Por fim, agradeço à Capes, pela concessão da bolsa de estudos.

SUMÁRIO

RESUMO.....	ix
ABSTRACT	x
ÍNDICE DE ABREVIACÕES.....	xi
ÍNDICE DE TABELAS	xii
ÍNDICE DE FIGURAS	xiii
1 CAPÍTULO 1 - INTRODUÇÃO GERAL.....	1
1.1 INTRODUÇÃO	1
1.1.1 A Soja	1
1.1.2 Origem, domesticação e introdução da soja no Brasil	2
1.1.3 Importância da soja na economia mundial e no Brasil.....	3
1.1.4 Fatores limitantes na produção agrícola e da soja.....	5
1.1.5 Estudos de transcriptoma por RNA-Seq	6
1.1.6 Abordagens genômicas no estudo da soja	8
1.2 REFERÊNCIAS	10
CAPÍTULO 2 - Systematic analysis of 1,298 RNA-Seq samples and construction of a comprehensive soybean (<i>Glycine max</i>) expression atlas.....	17
2.1. SUMMARY.....	17
2.2 INTRODUCTION.....	18
2.3 RESULTS AND DISCUSSION.....	19
2.3.1 Unsupervised sample clustering reveals three major clades comprising underground, aerial, and seed tissues	24
2.3.2 Systematic analysis of hundreds of RNA-Seq libraries support the expression of the vast majority of the soybean genes.....	27
2.3.3 Housekeeping genes	30
2.3.4 Tissue-specific gene expression	33

2.3.5	Nodule-specific genes.....	37
2.3.6	Endosperm-specific genes.....	39
2.3.7	Flower-specific genes	39
2.3.8	Identification of novel transcripts.....	40
2.3.9	Data availability through a user-friendly web interface.....	47
2.4	CONCLUSIONS.....	49
2.5	METHODS	49
2.5.1	Soybean genome and annotation data	49
2.5.2	Soybean RNA-Seq data.....	49
2.5.3	Preprocessing and quality control	50
2.5.4	Transcript assembly and gene expression estimation	50
2.5.5	Sample clustering	51
2.5.6	Identification of novel genes and splicing isoforms	52
2.5.7	Analysis of the top 1000 highest expressed gene lists.....	52
2.5.8	Identification of housekeeping genes.....	52
2.5.9	Assessment of tissue-specific expression.....	53
2.5.10	Gene orthologs and enrichment tests	54
2.5.11	Web server	54
2.5.12	Data statement	55
2.6	ACKNOWLEDGEMENTS	55
2.7	REFERENCES.....	55

RESUMO

A soja (*Glycine max*) é uma das mais importantes culturas na agricultura mundial, constituindo uma fonte crucial de proteína e óleo na alimentação humana e animal. Desde sua domesticação, a cultura da soja apresentou enorme progresso, especialmente devido ao desenvolvimento de cultivares melhoradas. Contudo, ainda há perdas relevantes em sua produção em decorrência de estresses bióticos e abióticos. A publicação do genoma da soja, em 2010, abriu diversas frentes de estudo molecular e genômico desta espécie, incluindo análises transcriptômicas em diversas condições e tecidos. No entanto, análises integrativas desses dados, visando identificar padrões transcricionais globais em tecidos específicos, ainda são escassos. Dessa forma, o objetivo deste trabalho foi integrar os dados transcriptômicos disponíveis publicamente, buscando identificar grandes padrões e especificidades neste conjunto de dados. Foram coletados dados de 1.298 amostras de transcriptoma de soja disponíveis publicamente no banco de dados SRA (Short Read Archive, do NCBI), que foram filtrados, processados e mapeados sistematicamente no genoma de referência de soja (Wm82.a2.v1). Dos genes anotados, 94% (52.737 / 56.044) mostraram expressão detectável em pelo menos uma das amostras. Foram revelados três grupos principais através do agrupamento não-supervisionado, incluindo amostras de partes subterrâneas, aéreas e relacionadas a sementes. Foram encontrados 452 genes com níveis de expressão uniformes e constantes ao longo das amostras, o que indica que estes sejam os genes *housekeeping* de soja. Além destes, 1.349 genes apresentaram transcrição tecido-específica. Identificamos ainda que 95% (70.963 / 74.490) dos transcritos conhecidos se sobrepõem aos identificados em nosso trabalho, enquanto 3.256 dos transcritos representam potenciais novas isoformas de *splicing*. Além das descobertas supracitadas, todo o conjunto de dados integrados neste trabalho pode ser obtido ou acessado por meio de uma interface amigável disponível em <http://venanciogroup.uenf.br/resources/>. Este atlas transcriptômico poderá acelerar as pesquisas em diferentes áreas, desempenhando, portanto, um importante papel em programas de biotecnologia e melhoramento da soja.

Palavras-chave: Atlas transcricional, expressão gênica, RNA-seq, *splicing*, gene de referência.

ABSTRACT

Soybean (*Glycine max*) is one of the most important crops in the world, constituting an important source of protein and oil for human and animal nutrition. Since its domestication, the soybean crop has made enormous progresses, especially as a result of the development of improved cultivars. However, there are still significant losses in its production because of biotic and abiotic stresses. The publication of the soybean genome, in 2010, boosted several molecular and genomic studies of this species, including transcriptomic analyses in different conditions and tissues. However, integrative analyses of these data, aiming to identify global transcriptional patterns in specific tissues, are still scarce. Thus, the objective of this work was to integrate the publicly available transcriptomic data, seeking to identify large patterns and specificities in this data set. Data were collected from 1,298 publicly available soybean transcriptome samples in the SRA database (NCBI Short Read Archive), which were systematically filtered, processed and mapped to the soybean reference genome (Wm82.a2.v1). Of the annotated genes, 94% (52,737 / 56,044) showed detectable expression in at least one of the samples. Three main groups were revealed through unsupervised clustering, including samples of underground, aerial and seed-related parts. There were 452 genes with uniform and constant expression levels throughout the samples, which indicates that these are the soybean housekeeping genes. In addition, 1,349 genes showed tissue-specific transcription. We also identified that 95% (70,963 / 74,490) of known transcripts overlap with those identified in our work, while 3,256 of the transcripts represent potential new splicing isoforms. In addition to the aforementioned findings, the entire set of data integrated in this work can be obtained or accessed through a user-friendly interface available at <http://venanciogroup.uenf.br/resources/>. This transcriptomic atlas will likely accelerate research in different areas, thus playing an important role in biotechnology and soybean breeding programs.

Keywords: Transcriptional atlas, gene expression, RNA-seq, splicing, reference gene.

ÍNDICE DE ABREVIações

AG = AGAMOUS

AP1 = APETALA1

AP2 = APETALA2

AP3 = APETALA3

AS = Splicing alternativo ou *alternative splicing*

cDNA = DNA complementar

CDS = Sequências codificadoras ou *coding sequences*

CoV = Coeficiente de variação ou *coefficient of variation*

EMB = *Arabidopsis* EMBRYO-DEFECTIVE

FPKM = *Fragments Per Kilobase Million*

HK = *Housekeeping genes*

MFC = *Maximum to Minimum Coefficient*

NCBI = *National Center for Biotechnology Information*

NIN = *Nodule Inception*

PI = PISTILLATA

RNA-Seq = *RNA sequencing*

RPG = *Rhizobium-Directed Polar Growth*

RPKM = *Reads Per Kilobase Million*

SNPs = Polimorfismo de nucleotídeo único ou *single nucleotide polymorphism*

SRA = *Sequence Read Archive*

TF = Fator de transcrição ou *transcription factor*

TPM = *Transcripts Per Million*

TSS = *Transcription start site*

τ = Tau

ÍNDICE DE TABELAS

Table 1: Tissue-specific transcription factors.	37
Table 2: Number of transcripts in each transcript-classification code defined by GffCompare.....	41
Table 3: Number of alternative splicing events (AS). The first column illustrates the possible AS isoforms. The boxes represent exons and lines connect adjacent exons in the mature transcript.	43

ÍNDICE DE FIGURAS

Figure 1: Number of samples analyzed in this study and a graphical representation of each tissue.	20
Figure 2: Pipeline used to create the soybean RNA-Seq atlas.	21
Figure 3: Hierarchical clustering of samples using their transcriptional profiles. Per gene raw read counts were used to perform hierarchical clustering using the R function <code>hclust()</code> with default parameters. Samples were grouped into three major clades: aerial, underground, and seed-embryo related. A minor group of samples containing drought-stress-related leaves and shoots was also identified. The upper-left panel shows the sample clustering using t-SNE. Five samples (four from shoot: SAMN04932642, SAMN04932648, SAMN04932639, SAMN04932645 and one from root: SAMN02197701), labeled in the inside plot, showed very unexpected clustering patterns and were excluded from further analysis. An interactive 3D version of the t-SNE sample clustering is available at http://venanciogroup.uenf.br/resources/	26
Figure 4: Global gene expression patterns of the housekeeping (HK) genes. A. Scatter plot of mean vs standard deviation showing uniform and stable expression of 452 HK genes. The gray dots represent all the non-HK expressed genes ($TPM \geq 1$ in at least one sample). The word cloud represents KEGG pathways enriched in HK genes (p -value < 0.05). B. Global expression patterns of HK genes. Three main clusters were found with K-means clustering, which were then hierarchically clustered.	32
Figure 5: Violin plot showing the distribution of Tau indexes of housekeeping, tissue-specific, and the remaining genes. Tau values range between 0 and 1, with low values indicating a stable and constitutive expression and higher values supporting tissue-specificity.	33
Figure 6: Heatmap showing the number of up-regulated genes in the tissues from the rows when compared with those from the columns. Gene up-regulation was determined by using a \log_2 (foldchange) ≥ 2 and adjusted p -value ≤ 0.05 using the moderated t-statistic in the limma package.	34

Figure 7: Global transcriptional patterns of tissue-specific genes. Expression values are represented as $\log_2(\text{TPM})$ values in 1243 samples. 35

Figure 8: Web interface to browse and download the expression data analyzed in this study. A. Users can search, visualize and download average expression levels in each tissue or; B retrieve expression values in batch in particular samples, tissues, or BioProjec. This resource is available at: <http://venanciogroup.uenf.br/resources/>..... 48

Figure S1: Stacked histograms of read mapping statistics across tissues. 22

Figure S2: Distribution of Spearman's rank correlation coefficients between normalized expression values (in TPM) estimated with kallisto and stringtie across samples. 24

Figure S3: Tissue- and sample-wise distribution of expressed genes ($\text{TPM} \geq 1$). A. Number of expressed genes in each tissue. B. Number of samples in which genes are expressed..... 28

Figure S4: Length of coding regions with undetectable expression levels ($\text{TPM} < 1$).28

Figure S5: Pathway analysis of the top 1,000 highest expressed genes in leaves and roots. As expected, we found that photosynthesis genes are enriched in the former. The small groups of boxes in each pathway represent genes involved in that process. The color of these boxes ranges from dark red to dark blue representing extremely high expression and low expression, respectively..... 29

Figure S6: Wiggle plots showing read coverage of potentially novel genes at four unannotated loci in the soybean genome. We selected five samples in which these genes had the highest expression levels. A: TU4871, B: TU28093, C: TU72199, D: TU56508..... 42

Figure S7: Sashimi plot of Glyma.17G195900 (CASEIN KINASE 1-LIKE PROTEIN 4) showing the number of reads supporting splice junctions in two nodule and two root samples. The tracks below the plot represent splicing isoforms. Exons within the highlighted region indicate variation in splicing patterns. The top track (TU62356) is the novel isoform, comprising an exon skipping event, which is supported by 10 reads from nodule samples. 44

Figure S8: Sashimi plot of Glyma.14G052400 (Glycine rich protein family) showing the number of reads supporting the splice junctions in two nodule samples. The tracks

below the plot represent splicing isoforms. The exon within the highlighted region indicates variation in splicing patterns due to the skipping of exon 2 in TU50862. This new isoform also has some small variations in exon junctions in exons 3 and 4. The skipping of exon 2 in the new isoform is supported by 2,412 reads from two nodule samples. 45

Figure S9: Sashimi plot of Glyma.06G276400 (Cysteamine dioxygenase/Persulfurase) showing the number of reads supporting the splice junctions in two nodule samples. The tracks below the plot represent splicing isoforms. Exons within the highlighted region indicate variation in splicing patterns. The top track (TU21492) is a novel isoform comprising a different length in exon 2, along with two terminal 3' exons instead of one, giving rise to a new combination of exons. 46

1 CAPÍTULO 1 - INTRODUÇÃO GERAL

1.1 INTRODUÇÃO

1.1.1 A Soja

A soja pertence à ordem Fabales, família Fabaceae (leguminosa) e gênero *Glycine*. Nesse gênero se encontram as espécies *Glycine max* (*G. max*) e *Glycine soja* (*G. soja*). *G. max* é a espécie mais cultivada no mundo, enquanto a *G. soja* é uma espécie selvagem aparentada (Mishra e Verma, 2010). As variedades comerciais de soja possuem algumas características peculiares. São plantas de caule híspido (contém pelos) e pouco ramificado. Além disso, dependendo da cultivar, a estatura da planta pode variar entre 60 e 110 cm. Suas folhas podem ser de quatro tipos diferentes: o primeiro par de folhas é representado por cotilédones, seguido por um segundo par de folhas primárias, folhas de folhagem trifolioladas e os profilos (Lersten e Carlson, 2004).

A soja apresenta sistema radicular predominantemente axial fasciculado, com nódulos induzidos e ocupados por simbioses. A soja normalmente estabelece simbiose com diferentes tipos de rizóbio fixadores de nitrogênio, como: *Bradyrhizobium japonicum*, *Bradyrhizobium elkanii*, *Bradyrhizobium liaoningense*, *Bradyrhizobium yuanmingense*, *Rhizobium tropici*, *Rhizopus oryzae* e *Mesorhizobium tianshanense*. A eficiência da fixação simbiótica de nitrogênio em soja por aplicação de inoculantes depende muito da especificidade do hospedeiro simbiótico (Hayashi et al., 2012; Yang et al., 2010). Tais microrganismos auxiliam na manutenção da fertilidade do solo, reduzindo a demanda de fertilizante nitrogenado nas safras seguintes e facilitando as práticas de cultivo em rodízio (Gazzoni e Dall'agnol, 2018; Salvagiotti et al., 2008).

As flores da soja são completas e ocorrem em racemos terminais ou axilares. Seu número varia de 2 a 35 flores por racemo que, quando abertas, medem de 3 a 8 mm. O fruto da soja é do tipo vagem e contém de uma a cinco sementes. Pode ser achatado, arredondado, reto ou curvado, apresentando pubescência de coloração cinza clara, cinza escura, marrom clara, marrom média e marrom escura (Nogueira et

al., 2015). Na polinização, os estames diáfanos são elevados, de modo que as anteras fiquem próximas ao estigma, permitindo que o pólen se espalhe diretamente sobre o mesmo, resultando em até 99% de autofecundação (Palmer et al., 2011). O genoma da soja é diploide (2n) e está contido em 40 cromossomos (Joshi et al., 2014). As sementes de soja acumulam proteínas, carboidratos e lipídeos, que abastecem a germinação e as etapas iniciais do desenvolvimento pós-germinativo (Singh, 2010). O embrião de soja é composto por dois cotilédones, radícula, hipocótilo e epicótilo.

1.1.2 Origem, domesticação e introdução da soja no Brasil

É amplamente aceito que a soja cultivada moderna teve sua origem e domesticação a partir de ancestrais da soja selvagem provenientes do Leste Asiático, mais especificamente da China, ao longo do Vale do Rio Huang He (conhecido também como Rio Amarelo), entre 5.000 a 9.000 anos atrás (Dong et al., 2004; Lee et al., 2011). Esta hipótese foi corroborada por estudos genômicos recentes envolvendo variedades de soja presentes atualmente nessa região (Han et al., 2016).

O processo de domesticação e melhoramento da soja envolveu uma ampla gama de mudanças evolutivas que passaram por vários estágios contínuos. Estudos genéticos recentes revelaram alguns genes que estariam envolvidos na domesticação da soja, levando à adaptação ou melhoramento da cultura ao longo do tempo (Meyer e Purugganan, 2013). Assim, características relacionadas à perda e quebra de vagens e dispersão de sementes (Dong et al., 2004); à dureza da semente, que inclui a permeabilidade à água em sementes secas e dureza de sementes cozidas (Sun et al., 2015); ao crescimento do caule (Tian et al., 2010) e; ao tempo de floração, estão entre as prioridades no processo de domesticação e melhoramento da soja para a agricultura (Cober e Morrison, 2010; Sedivy et al., 2017; Wu et al., 2017). Análises históricas de cultivares lançadas ao longo dos últimos 90 anos indicam que um dos principais fatores que impactam o rendimento com o cultivo de soja é o aumento do número de sementes por planta (Jin et al., 2010; Morrison et al., 2000).

A soja foi levada da China para Europa em 1740 e, em 1804, introduzida na América do Norte. A soja chegou ao Brasil pela Bahia em 1882, sendo disseminada para outras regiões do país a partir de 1914. Contudo, foi na Região Sul que a soja melhor se adaptou, tornando-se a principal cultura da região, que passaria a ser então o principal fornecedor de soja para o mercado interno brasileiro (Dall'agnol, 2016). Naquele momento, o Brasil também iniciava um esforço para alavancar a produção de suínos e aves, pressionando a demanda por farelo de soja para nutrição animal. Em 1966, a produção comercial de soja no Brasil já era estratégica, correspondendo a cerca de 500 mil toneladas anualmente. Os investimentos em pesquisa no Brasil permitiram a adaptação da soja às regiões de baixas latitudes, resultando num marcante aumento da produtividade com o avanço para a Região Centro Oeste a partir das décadas de 80 e 90. Por fim, o progresso no manejo da cultura e o desenvolvimento de novas cultivares elevaram o Brasil ao topo do ranking dos maiores produtores mundiais de soja (Silva et al., 2019).

1.1.3 Importância da soja na economia mundial e no Brasil.

A soja tornou-se a colheita milagrosa do século XXI (Thakare et al., 2006), particularmente pela grande variedade de usos comerciais e alto valor de mercado, que fizeram dela a leguminosa mais valiosa do mundo (Hershman et al., 2011). Ressalte-se ainda que a maior parte da dieta humana, avícola e pecuária é derivada de cereais e leguminosas (Mandal e Mandal, 2000), resultando num cenário em que a soja corresponde a mais de um quarto da alimentação humana e animal no mundo (Ash, 2017; Tian et al., 2010).

O óleo e o farelo são os dois principais derivados da soja (Warrington et al., 2015). A maior parte dos valores gerados com a sua produção se relacionam majoritariamente com a farinha, seguida do óleo e da venda de grãos (Pettersson e Pontoppidan, 2013). O farelo é importante na alimentação humana, animal e na fabricação de produtos. O grão de soja tem também grande importância econômica pelas inúmeras aplicações industriais, resultando em crescente demanda global

atualmente. A demanda no mercado internacional e a estabilização da soja como importante fonte de proteínas, fizeram com que o investimento na sua produção apresentasse crescimento expressivo (Hirakuri e Lazzarotto, 2011). No ano de 2020, a produção mundial de soja ultrapassou a marca de 337 milhões de toneladas (USDA, 2020).

O teor de óleo em grãos de soja é de 20%, que é tipicamente utilizado para fornecer energia ao embrião em desenvolvimento durante a germinação (Copeland e McDonald, 2001). Na indústria, o óleo de soja é usado, por exemplo, para extrair lecitina, na fabricação de molhos para salada, margarinas, caldo, sabão, gorduras e outros produtos. Os grãos de soja também são ricos em proteínas, que respondem por 40% do seu peso total (Singh, 2010) e tornam a soja um fator chave na alimentação humana e na agropecuária. Além dos altos teores de proteína e óleo, a matéria orgânica residual da soja pode ser utilizada como forragem na alimentação animal e no adubo verde (Qiu e Chang, 2010).

O comércio mundial da soja foi mantido unicamente pela China até o início da Segunda Guerra Mundial. A partir do final do conflito, os EUA dominaram o comércio mundial da commodity durante mais de 20 anos, quando, a partir do final da década de 1960, Brasil e Argentina tornaram-se importantes produtores e exportadores (Silva et al., 2019). Atualmente, juntamente com os EUA, estes países respondem por 83% da produção mundial na safra 2019/2020 (CONAB, 2020). O rendimento da soja nesses três países têm aumentado constantemente ao longo das últimas duas décadas, frequentemente acompanhado do aumento de área plantada (Ainsworth et al., 2012). Até 2019, o Brasil era o segundo maior produtor de soja do mundo, atrás apenas dos EUA, com rendimento superior a três toneladas por hectare (<http://www.fao.org>) (Black, 2000; CONAB, 2020; Hirakuri e Lazzarotto, 2011). Em 2020, o Brasil passou a ser o maior produtor mundial de soja (CONAB, 2020).

Dentro do sistema intensivo de produção, as características desejáveis que contribuem positivamente para aumentar a estabilidade e potencial de rendimento de cultivares de soja são: maior resistência à doenças, resistência aos insetos-praga e aos nematoides associados à cultura; boa resistência ao acamamento e a deiscência

precoce; boa qualidade fisiológica da semente; adaptação a condições locais de ambiente e a escolha do tipo de planta adequada ao sistema agrícola utilizado na região produtora, visando uma maior produtividade (Prior et al., 2006).

1.1.4 Fatores limitantes na produção agrícola e da soja

Mesmo com o sistemático aumento de produtividade, há fatores que afetam negativamente a produção de soja, como os estresses bióticos (ataque de patógenos, pragas, microrganismos) e abióticos (condições ambientais adversas).

Os estresses abióticos são os principais responsáveis pela diminuição na produção e no rendimento da soja, dentre os quais podemos citar: seca, excesso de salinidade, metais pesados e as temperaturas extremas, dentre os quais a desidratação e os extremos de temperatura são os que afetam mais significativamente o desenvolvimento vegetal (Shinozaki e Yamaguchi-Shinozaki, 2000), resultando em fortes perdas de produtividade. No âmbito dos estresses abióticos, os desafios se intensificaram em decorrência das mudanças climáticas (Ahmad et al., 2016a; Ahmad et al., 2016b; Lake et al., 2012; Ray et al., 2015). Dentre os eventos climáticos que podem afetar a produção estão o El Niño Oscilação Sul (ENSO), o Dipolo do Oceano Índico (IOD), a Variabilidade do Atlântico Tropical (TAV) e a Oscilação do Atlântico Norte (NAO), que juntos respondem por 7% das perdas na produção global de soja (Anderson et al., 2019). No contexto nacional, apenas nas safras de 2003/2004 a 2014/2015, as perdas devido a eventos de seca foram estimadas em cerca de US \$ 46,6 bilhões (Fuganti-Pagliarini et al., 2017).

Em menor proporção, o estresse biótico causa prejuízos em escala global para todo o agronegócio da soja, como no caso do fungo *Phakopsora pachyrhizi*, causador da ferrugem asiática, e do inseto praga Percevejo-marrom-da-soja, *Euschistus heros* (Bueno et al., 2011; Da Graca et al., 2016; De Oliveira et al., 2018). No caso dos microrganismos causadores de doenças, a indução de mecanismos de defesa na planta durante o contato com o patógeno aumenta a demanda por fotoassimilados e provoca alterações no metabolismo primário vegetal. Ademais, em alguns casos, o

desenvolvimento de áreas cloróticas e necróticas nas folhas reduz a área ativa para realização da fotossíntese (Berger et al., 2007).

Além dos estresses acima listados, o cultivo de soja em áreas não tradicionais frequentemente requer estratégias de melhoramento genético para o desenvolvimento de cultivares adaptadas aos novos locais (Grainger e Rajcan, 2014; Granier et al., 2006). A seleção direta para estabilidade de produção com base em testes em vários locais tem sido tradicionalmente usada para o desenvolvimento de variedades adaptadas a condições ambientais adversas (Manavalan et al., 2009). Desta forma, há um esforço na comunidade científica para elucidar os mecanismos moleculares envolvidos na resposta à condições adversas (Fait et al., 2006; Prior et al., 2006; Weber et al., 2005), sob a premissa de que estes seriam críticos no desenvolvimento de novas cultivares.

1.1.5 Estudos de transcriptoma por RNA-Seq

O uso de tecnologias ômicas como aquelas relacionadas à transcriptômica, proteômica e metabolômica, aumentaram a compreensão das complexas redes regulatórias associadas à adaptação e tolerância em plantas (Urano et al., 2010; Yin e Struik, 2010).

Especificamente nos estudos de transcriptômica, tecnologias baseadas em hibridização e em sequenciamento já vem sendo empregadas há mais de duas décadas. Dentre as tecnologias baseadas em hibridização, podemos citar os macro e microarranjos de DNA (Alkharouf et al., 2006). Por sua vez, as tecnologias baseadas em sequenciamento compreendem uma maior variedade de métodos, como as bibliotecas de EST (*Expressed Sequence Tag*) (Adams et al., 1991); a técnica de SAGE (*Serial Analysis of Gene Expression*) (Velculescu et al., 1995) e suas versões derivadas LongSAGE (Sara et al., 1995) e SuperSAGE (Matsumura et al., 2003). Embora tais tecnologias tenham sido extremamente importantes por vários anos, apenas na segunda metade da década de 2000 que foi desenvolvida a mais impactante delas, a tecnologia de RNA sequencing (RNA-Seq) (Mortazavi et al., 2008).

Os estudos de RNA-seq baseiam-se no sequenciamento em altíssima escala de fragmentos de RNA, denominados *reads*, que são analisados utilizando métodos baseados em mapeamento, como STAR (Dobin e Gingeras, 2015), TopHat2 (Kim et al., 2013) e Bowtie 2 (Langmead e Salzberg, 2012) ou de pseudo-contagens para estimar a expressão gênica, como por exemplo os programas Salmon (Patro et al., 2017) e Kallisto (Bray et al., 2016).

O método de RNA-Seq é potencialmente capaz de caracterizar todo o transcriptoma e quantificar os diferentes níveis transcricionais em uma determinada amostra, sendo atualmente o método mais utilizado para análises de expressão gênica (Levin et al., 2010; Zhang et al., 2018). Estudos comparativos entre os conjuntos de dados gerados por RNA-Seq e microarranjos, a partir do mesmo conjunto de amostras, mostraram uma alta correlação entre os perfis de expressão gênica gerados pelas duas tecnologias. Contudo, o RNA-Seq foi superior na detecção de transcritos de baixa abundância, na discriminação de isoformas de splicing na identificação de variantes genéticas. O RNA-Seq também demonstrou uma faixa dinâmica mais ampla do que os microarranjos, o que permitiu a detecção de genes mais diferencialmente expressos. A análise dos dois conjuntos de dados também mostrou vantagens na prevenção de problemas inerentes aos microarranjos, como hibridização cruzada, hibridização inespecífica e intervalo de detecção limitado de sondas individuais. Como o RNA-Seq não depende de hibridização, esta tecnologia não apresenta tais problemas, o que simplifica e potencializa a análise de dados (Zhao et al., 2014).

Diversas aplicações das tecnologias de RNA-Seq vêm sendo realizadas para estudos dos padrões de expressão em diferentes tecidos e em diferentes condições fisiológicas de plantas, inclusive nas respostas aos diversos estresses bióticos e abióticos (Barbazuk et al., 2008; Reddy, 2007; Reddy et al., 2013; Reddy et al., 2012; Reddy e Shad Ali, 2011). Uma das respostas a esse estresse é a expressão de um grande número de genes, cujos produtos podem estar envolvidos em diversas funções adaptativas, como reguladores da expressão gênica e da transdução de sinal, bem como na proteção e na detoxificação das células (Shinozaki e Yamaguchi-Shinozaki, 2007; Shinozaki et al., 2003; Wally e Punja, 2010).

1.1.6 Abordagens genômicas no estudo da soja

Grande parte das angiospermas descendem de ancestrais que sofreram duplicações integrais de genoma (*whole-genome duplication*, WGD), seja através de auto- quanto de aloploidia. Evidências recentes indicam que todas as angiospermas existentes tenham sofrido pelo menos um evento de WGD em sua linhagem evolutiva, sugerindo que tais eventos sejam importantes na geração de diversidade (Akoz e Nordborg, 2019; Jaillon et al., 2007; Jiao et al., 2011; Moharana e Venancio, 2020) Adicionalmente, diversos estudos têm demonstrado a alta prevalência deste tipo de fenômeno em plantas (Landis et al., 2018; One Thousand Plant Transcriptomes, 2019; Semon e Wolfe, 2007).

O primeiro sequenciamento do genoma da soja (da cultivar Williams 82), foi realizado em 2010 (Schmutz et al., 2010), abrindo enormes perspectivas na exploração de aspectos bioquímicos e fisiológicos da espécie. Esta iniciativa mostrou que aproximadamente 75% dos genes estão presentes em múltiplas cópias no genoma da soja, correspondendo a 12.253 famílias gênicas. Boa parte destas famílias multigênicas foram geradas por eventos de WGD, seguidos de diversificação, perda de genes e rearranjos cromossômicos. Além das WGDs, as expansões em tandem também contribuíram para o surgimento de famílias multigênicas em soja, como observado previamente em genes codificadores de NBS-LRR, proteínas F-box, proteínas de resposta a auxina e proteínas contendo outros domínios comumente encontrados em famílias multigênicas em genomas vegetais (Bellieny-Rabelo et al., 2013; Schmutz et al., 2010).

Estudos de re-sequenciamento têm sido realizados, tanto na soja cultivada, quanto na selvagem, utilizando polimorfismos de nucleotídeo único (*Single Nucleotide Polymorphisms*, SNPs), a fim de determinar a origem, domesticação e seleção da soja cultivada atualmente. Dentre estes, podemos destacar estudos realizados com as cultivares encontradas na Coreia (Chung et al., 2014), que abriram caminho para uma investigação mais ampla de 302 acessos de soja, incluindo 62 da soja silvestre (*G.*

soja) e 240 de *G. max* (130 variedades locais e 110 cultivares melhoradas) (Zhou et al., 2015). No cenário nacional, o re-sequenciamento genômico de 28 cultivares concluiu que há espaço para a introdução de novos alelos potencialmente vantajosos para a produção da soja no país (Dos Santos et al., 2016).

Apesar do enorme sucesso dos programas de melhoramento genético clássico, é imperativa a necessidade de se compreender as bases genéticas e bioquímicas das características selecionadas no melhoramento vegetal, a fim de identificar potenciais alvos para futuras estratégias de melhoramento e extrapolação para outras espécies (Ainsworth et al., 2012). Neste sentido, os genomas de 26 acessos de soja, de diferentes locais do mundo, foram sequenciados e usados na construção de um pangenoma de alta qualidade, que inclui variações únicas da composição genética de algumas variedades importantes (Liu et al., 2020). Este pangenoma será, certamente, um alicerce importante tanto nas estratégias de melhoramento quanto na elucidação de funções gênicas que permeiam características agronômicas relevantes para a agricultura.

O sequenciamento do genoma abriu caminho para outros estudos complementares em larga escala, como os que vêm sendo conduzidos para estudar o transcriptoma (Ge et al., 2010; Libault et al., 2010; Severin et al., 2010), proteoma (Zhang et al., 2011) e metaboloma (Gu et al., 2017) de diferentes tecidos e condições fisiológicas da soja. Tais trabalhos, acumulados ao longo de anos, tornam os recursos de bioinformática e bancos de dados elementos essenciais para a utilização mais eficaz e democrática dos dados. Apenas em soja, os dados de RNA-Seq acumulados ao longo dos últimos 10 anos já correspondem a mais de 440 artigos publicados (<https://pubmed.ncbi.nlm.nih.gov/>), 720 BioProjects e mais de 4800 entradas do banco de dados SRA do NCBI (<https://www.ncbi.nlm.nih.gov/sra>), em consulta realizada em novembro de 2020. Esta coleção de dados é um elemento crucial tanto na pesquisa básica quanto em programas de melhoramento genético da espécie (Hakeem et al., 2012a; Hakeem et al., 2012b).

Mesmo com um acervo tão vasto de trabalhos realizados em RNA-seq de soja, há uma grande escassez de repositórios integrando estes dados após processamento

e análise sistemáticos dos mesmos, levando os pesquisadores a buscar os trabalhos individualmente, o que consome tempo e não é eficiente em decorrência das diferentes metodologias de análise empregadas, além das limitações de estrutura computacional e de expertise analítica da maioria dos grupos de pesquisa em bioquímica e biologia molecular. Estes trabalhos, naturalmente, são limitados à apenas um tecido em diferentes condições como: transcritos envolvidos na germinação da semente em uma escala de tempo (Bellieny-Rabelo et al., 2016) ou análise do transcriptoma de nódulos em diferentes estágios de desenvolvimento da soja (Yuan et al., 2017). Outros estudos são fundamentados no transcriptoma em diferentes condições de estresse, utilizando plantas suscetíveis, resistentes e controle (Miraeiz et al., 2020). Conseqüentemente, há uma clara limitação no acesso efetivo aos dados disponíveis, dificultando a realização de análises globais ou envolvendo múltiplos estudos relacionados. Neste cenário, o objetivo central do presente trabalho é a integração de todos os dados de RNA-Seq de soja disponíveis publicamente até maio de 2018, a fim de gerar um atlas transcriptômico de soja, disponível publicamente e com fácil acesso, a fim de impulsionar pesquisas moleculares de soja e maximizar a reutilização de dados, o que trará grandes benefícios para diferentes grupos de pesquisa no mundo. Ademais, realizamos diversas descobertas durante a criação e investigação deste atlas, que também estão descritas nesta tese. O desenvolvimento do atlas, nossos principais resultados e a descrição do banco de dados gerados e sua interface, estão disponíveis no capítulo 2 desta tese, em forma de um artigo científico que foi publicado em 2020 (Machado et al., 2020).

1.2 REFERÊNCIAS

- Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., et al. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, 252(5013), 1651-1656.
- Ahmad, P., Abdel, L. A. A., Rasool, S., Akram, N. A., Ashraf, M., et al. (2016a). Role of Proteomics in Crop Stress Tolerance. *Front Plant Sci*, 7, 1336.
- Ahmad, P., Rasool, S., Gul, A., Sheikh, S. A., Akram, N. A., et al. (2016b). Jasmonates: Multifunctional Roles in Stress Tolerance. *Front Plant Sci*, 7, 813.

- Ainsworth, E. A., Yendrek, C. R., Skoneczka, J. A., e Long, S. P. (2012). Accelerating yield potential in soybean: potential targets for biotechnological improvement. *Plant Cell Environ*, 35(1), 38-52.
- Akoz, G., e Nordborg, M. (2019). The Aquilegia genome reveals a hybrid origin of core eudicots. *Genome Biol*, 20(1), 256.
- Alkharouf, N. W., Klink, V. P., Chouikha, I. B., Beard, H. S., MacDonald, M. H., et al. (2006). Timecourse microarray analyses reveal global changes in gene expression of susceptible Glycine max (soybean) roots during infection by Heterodera glycines (soybean cyst nematode). *Planta*, 224(4), 838-852.
- Anderson, W. B., Seager, R., Baethgen, W., Cane, M., e You, L. (2019). Synchronous crop failures and climate-forced production variability. *Sci Adv*, 5(7), eaaw1976.
- Ash, M. (2017). U.S. Soybean Shipments Surge but New Sales are Slowing. *Economic Research Service, USDA*.
- Barbazuk, W. B., Fu, Y., e McGinnis, K. M. (2008). Genome-wide analyses of alternative splicing in plants: opportunities and challenges. *Genome Res*, 18(9), 1381-1392.
- Belliény-Rabelo, D., De Oliveira, E. A., Ribeiro, E. S., Costa, E. P., Oliveira, A. E., et al. (2016). Transcriptome analysis uncovers key regulatory and metabolic aspects of soybean embryonic axes during germination. *Scientific Reports*, 6, 36009.
- Belliény-Rabelo, D., Oliveira, A. E., e Venancio, T. M. (2013). Impact of whole-genome and tandem duplications in the expansion and functional diversification of the F-box family in legumes (Fabaceae). *PLoS One*, 8(2), e55127.
- Berger, S., Sinha, A. K., e Roitsch, T. (2007). Plant physiology meets phytopathology: plant primary metabolism and plant-pathogen interactions. *J Exp Bot*, 58(15-16), 4019-4026.
- Black, R. J. (2000). Complexo soja: fundamentos, situação atual e perspectiva. *Soja: tecnologia de produção II. Piracicaba: ESALQ*, 1-18.
- Bray, N. L., Pimentel, H., Melsted, P., e Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*, 34(5), 525-527.
- Bueno, R. C., Bueno-Ade, F., Moscardi, F., Parra, J. R., e Hoffmann-Campo, C. B. (2011). Lepidopteran larva consumption of soybean foliage: basis for developing multiple-species economic thresholds for pest management decisions. *Pest Manag Sci*, 67(2), 170-174.
- Chung, W. H., Jeong, N., Kim, J., Lee, W. K., Lee, Y. G., et al. (2014). Population structure and domestication revealed by high-depth resequencing of Korean cultivated and wild soybean genomes. *DNA research*, 21(2), 153-167.
- Cober, E. R., e Morrison, M. J. (2010). Regulation of seed yield and agronomic characters by photoperiod sensitivity and growth habit genes in soybean. *Theor Appl Genet*, 120(5), 1005-1012.
- CONAB. (2020). Companhia nacional de Abastecimento – Acompanhamento da safra brasileira de grãos v. 4 Safra 2019/20. *Segundo levantamento, Brasília*, p. 1-171 julho 2020.
- Copeland, L. O., e McDonald, M. B. (2001). Seed germination *Principles of Seed Science and Technology* (pp. 72-123): Springer.
- da Graca, J. P., Ueda, T. E., Janegitz, T., Vieira, S. S., Salvador, M. C., et al. (2016). The natural plant stress elicitor cis-jasmone causes cultivar-dependent reduction in growth of the stink bug, Euschistus heros and associated changes

- in flavonoid concentrations in soybean, *Glycine max.* *Phytochemistry*, *131*, 84-91.
- Dall'Agnol, A. (2016). *A Embrapa Soja no contexto do desenvolvimento da soja no Brasil: histórico e contribuições*: Brasília, DF: Embrapa, 2016.
- de Oliveira, T. B., de Azevedo Peixoto, L., Teodoro, P. E., de Alvarenga, A. A., Bhering, L. L., *et al.* (2018). The number of measurements needed to obtain high reliability for traits related to enzymatic activities and photosynthetic compounds in soybean plants infected with *Phakopsora pachyrhizi*. *PLoS One*, *13*(2), e0192189.
- Dobin, A., e Gingeras, T. R. (2015). Mapping RNA-seq Reads with STAR. *Curr Protoc Bioinformatics*, *51*, 11 14 11-11 14 19.
- Dong, Y. S., Zhao, L. M., Liu, B., Wang, Z. W., Jin, Z. Q., *et al.* (2004). The genetic diversity of cultivated soybean grown in China. *Theor Appl Genet*, *108*(5), 931-936.
- dos Santos, J. V. M., Valliyodan, B., Joshi, T., Khan, S. M., Liu, Y., *et al.* (2016). Evaluation of genetic variation among Brazilian soybean cultivars through genome resequencing. *BMC Genomics*, *17*(1), 110.
- Fait, A., Angelovici, R., Less, H., Ohad, I., Urbanczyk-Wochniak, E., *et al.* (2006). Arabidopsis seed development and germination is associated with temporally distinct metabolic switches. *Plant Physiol*, *142*(3), 839-854.
- Fuganti-Pagliarini, R., Ferreira, L. C., Rodrigues, F. A., Molinari, H. B. C., Marin, S. R. R., *et al.* (2017). Characterization of Soybean Genetically Modified for Drought Tolerance in Field Conditions. *Front Plant Sci*, *8*, 448.
- Gazzoni, D., e Dall'Agnol, A. (2018). The soybean saga, from 1050 BC to 2050 AD. *The soybean saga, from 1050 BC to 2050 AD*.
- Ge, Y., Li, Y., Zhu, Y. M., Bai, X., Lv, D. K., *et al.* (2010). Global transcriptome profiling of wild soybean (*Glycine soja*) roots under NaHCO₃ treatment. *BMC Plant Biol*, *10*, 153.
- Grainger, C. M., e Rajcan, I. (2014). Characterization of the genetic changes in a multi-generational pedigree of an elite Canadian soybean cultivar. *Theor Appl Genet*, *127*(1), 211-229.
- Granier, C., Aguirrezabal, L., Chenu, K., Cookson, S. J., Dauzat, M., *et al.* (2006). PHENOPSIS, an automated platform for reproducible phenotyping of plant responses to soil water deficit in *Arabidopsis thaliana* permitted the identification of an accession with low sensitivity to soil water deficit. *New Phytologist*, *169*(3), 623-635.
- Gu, E. J., Kim, D. W., Jang, G. J., Song, S. H., Lee, J. I., *et al.* (2017). Mass-based metabolomic analysis of soybean sprouts during germination. *Food Chem*, *217*, 311-319.
- Hakeem, K. R., Chandna, R., Ahmad, P., Iqbal, M., e Ozturk, M. (2012a). Relevance of proteomic investigations in plant abiotic stress physiology. *OMICS*, *16*(11), 621-635.
- Hakeem, K. R., Khan, F., Chandna, R., Siddiqui, T. O., e Iqbal, M. (2012b). Genotypic variability among soybean genotypes under NaCl stress and proteome analysis of salt-tolerant genotype. *Appl Biochem Biotechnol*, *168*(8), 2309-2329.

- Han, Y., Zhao, X., Liu, D., Li, Y., Lightfoot, D. A., *et al.* (2016). Domestication footprints anchor genomic regions of agronomic importance in soybeans. *New Phytol*, 209(2), 871-884.
- Hayashi, M., Saeki, Y., Haga, M., Harada, K., Kouchi, H., *et al.* (2012). Rj (rj) genes involved in nitrogen-fixing root nodule formation in soybean. *Breed Sci*, 61(5), 544-553.
- Hershman, D. E., Vincelli, P., e Kaiser, C. A. (2011). Foliar fungicide use in corn and soybeans. *Plant pathology fact sheet PPFs-MISC-05, UK Cooperative extension service, University of Kentucky, College of Agriculture*.
- Hirakuri, M. H., e Lazzarotto, J. J. (2011). Evolução e perspectivas de desempenho econômico associadas com a produção de soja nos contextos mundial e brasileiro. *Londrina, PR: EMBRAPA*.
- Jaillon, O., Aury, J. M., Noel, B., Policriti, A., Clepet, C., *et al.* (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 449(7161), 463-467.
- Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., *et al.* (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature*, 473(7345), 97-100.
- Jin, J., Liu, X., Wang, G., Mi, L., Shen, Z., *et al.* (2010). Agronomic and physiological contributions to the yield improvement of soybean cultivars released from 1950 to 2006 in Northeast China. *Field Crops Research*, 115(1), 116-123.
- Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., *et al.* (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*, 14(4), R36.
- Lake, I. R., Hooper, L., Abdelhamid, A., Bentham, G., Boxall, A. B., *et al.* (2012). Climate change and food security: health impacts in developed countries. *Environ Health Perspect*, 120(11), 1520-1526.
- Landis, J. B., Soltis, D. E., Li, Z., Marx, H. E., Barker, M. S., *et al.* (2018). Impact of whole-genome duplication events on diversification rates in angiosperms. *Am J Bot*, 105(3), 348-363.
- Langmead, B., e Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 9(4), 357-359.
- Lee, G., Crawford, G. W., Liu, L., Sasaki, Y., e Chen, X. (2011). Archaeological soybean (*Glycine max*) in East Asia: does size matter? *PLoS One*, 6(11), e26720.
- Lersten, N. R., e Carlson, J. B. (2004). Vegetative morphology. *Soybeans: Improvement, production, and uses*, 16, 15-57.
- Levin, J., Adiconis, X., Yassour, M., Thompson, D., Guttman, M., *et al.* (2010). Development and evaluation of RNA-Seq methods. *Genome Biology*, 11(S1), P26.
- Libault, M., Farmer, A., Joshi, T., Takahashi, K., Langley, R. J., *et al.* (2010). An integrated transcriptome atlas of the crop model *Glycine max*, and its use in comparative analyses in plants. *Plant J*, 63(1), 86-99.
- Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., *et al.* (2020). Pan-Genome of Wild and Cultivated Soybeans. *Cell*, 182(1), 162-176 e113.
- Machado, F. B., Moharana, K. C., Almeida-Silva, F., Gazara, R. K., Pedrosa-Silva, F., *et al.* (2020). Systematic analysis of 1298 RNA-Seq samples and construction of a comprehensive soybean (*Glycine max*) expression atlas. *Plant J*.

- Manavalan, L. P., Guttikonda, S. K., Tran, L. S., e Nguyen, H. T. (2009). Physiological and molecular approaches to improve drought resistance in soybean. *Plant Cell Physiol*, 50(7), 1260-1276.
- Mandal, S., e Mandal, R. K. (2000). Pattern of Compliance with Treatment and Follow-up of Cervical Cancer Patients at Chittaranjan National Cancer Institute, Calcutta. *Asian Pac J Cancer Prev*, 1(4), 289-292.
- Matsumura, H., Reich, S., Ito, A., Saitoh, H., Kamoun, S., *et al.* (2003). Gene expression analysis of plant host-pathogen interactions by SuperSAGE. *Proc Natl Acad Sci U S A*, 100(26), 15718-15723.
- Meyer, R. S., e Purugganan, M. D. (2013). Evolution of crop species: genetics of domestication and diversification. *Nat Rev Genet*, 14(12), 840-852.
- Miraeiz, E., Chaiprom, U., Afsharifar, A., Karegar, A., J, M. D., *et al.* (2020). Early transcriptional responses to soybean cyst nematode HG Type 0 show genetic differences among resistant and susceptible soybeans. *Theor Appl Genet*, 133(1), 87-102.
- Mishra, S., e Verma, V. (2010). Soybean genetic resources. *The Soybean: Botany, Production and Uses*. London: CAB International, 74-91.
- Moharana, K. C., e Venancio, T. M. (2020). Polyploidization events shaped the transcription factor repertoires in legumes (Fabaceae). *Plant J*, 103(2), 726-741.
- Morrison, M. J., Voldeng, H. D., e Cober, E. R. (2000). Agronomic changes from 58 years of genetic improvement of short-season soybean cultivars in Canada. *Agronomy Journal*, 92(4):, 780-784.
- Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., e Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, 5(7), 621-628.
- Nogueira, A. P. O., Sedyama, T., e Gomes, J. D. (2015). Avanços no melhoramento genético da cultura da soja nas últimas décadas. *Doenças da Soja: Melhoramento genético e técnicas de manejo*, 159-178.
- One Thousand Plant Transcriptomes, I. (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*, 574(7780), 679-685.
- Palmer, R. G., Gai, J., Dalvi, V. A., e Suso, M. J. (2011). 13 Male Sterility and Hybrid Production Technology. *Biology and breeding of food legumes*, 193.
- Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., e Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods*, 14(4), 417-419.
- Pettersson, D., e Pontoppidan, K. (2013). Soybean meal and the potential for upgrading its feeding value by enzyme supplementation. *Soybean-Bio-Active Compounds*, 288-307.
- Prior, S. A., Torbert, H. A., Runion, G. B., Rogers, H. H., Ort, D. R., *et al.* (2006). Free-air carbon dioxide enrichment of soybean: influence of crop variety on residue decomposition. *J Environ Qual*, 35(4), 1470-1477.
- Qiu, L., e Chang, R. (2010). The origin and history of soybean. *The soybean: botany, production and uses*, 1-23.
- Ray, D. K., Gerber, J. S., MacDonald, G. K., e West, P. C. (2015). Climate variation explains a third of global crop yield variability. *Nat Commun*, 6, 5989.
- Reddy, A. S. (2007). Alternative splicing of pre-messenger RNAs in plants in the genomic era. *Annu Rev Plant Biol*, 58, 267-294.

- Reddy, A. S., Marquez, Y., Kalyna, M., e Barta, A. (2013). Complexity of the alternative splicing landscape in plants. *Plant Cell*, 25(10), 3657-3683.
- Reddy, A. S., Rogers, M. F., Richardson, D. N., Hamilton, M., e Ben-Hur, A. (2012). Deciphering the plant splicing code: experimental and computational approaches for predicting alternative splicing and splicing regulatory elements. *Front Plant Sci*, 3, 18.
- Reddy, A. S., e Shad Ali, G. (2011). Plant serine/arginine-rich proteins: roles in precursor messenger RNA splicing, plant development, and stress responses. *Wiley Interdiscip Rev RNA*, 2(6), 875-889.
- Salvagiotti, F., Cassman, K. G., Specht, J. E., Walters, D. T., Weiss, A., et al. (2008). Nitrogen uptake, fixation and response to fertilizer N in soybeans: A review. *Field Crops Research*, 108(1), 1-13.
- Sara, S. J., Dyon-Laurent, C., e Herve, A. (1995). Novelty seeking behavior in the rat is dependent upon the integrity of the noradrenergic system. *Brain Res Cogn Brain Res*, 2(3), 181-187.
- Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., et al. (2010). Genome sequence of the palaeopolyploid soybean. *Nature*, 463(7278), 178-183.
- Sedivy, E. J., Wu, F., e Hanzawa, Y. (2017). Soybean domestication: the origin, genetic architecture and molecular bases. *New Phytol*, 214(2), 539-553.
- Semon, M., e Wolfe, K. H. (2007). Consequences of genome duplication. *Curr Opin Genet Dev*, 17(6), 505-512.
- Severin, A. J., Woody, J. L., Bolon, Y. T., Joseph, B., Diers, B. W., et al. (2010). RNA-Seq Atlas of Glycine max: a guide to the soybean transcriptome. *BMC Plant Biol*, 10, 160.
- Shinozaki, K., e Yamaguchi-Shinozaki, K. (2000). Molecular responses to dehydration and low temperature: differences and cross-talk between two stress signaling pathways. *Curr Opin Plant Biol*, 3(3), 217-223.
- Shinozaki, K., e Yamaguchi-Shinozaki, K. (2007). Gene networks involved in drought stress response and tolerance. *J Exp Bot*, 58(2), 221-227.
- Shinozaki, K., Yamaguchi-Shinozaki, K., e Seki, M. (2003). Regulatory network of gene expression in the drought and cold stress responses. *Curr Opin Plant Biol*, 6(5), 410-417.
- Silva, D. F., Raimundo, E. K. M., e Forti, V. A. (2019). Nodulação em plantas de soja, Glycine max L. Merrill, submetidas a diferentes adubações. *Revista Verde de Agroecologia e Desenvolvimento Sustentável*, 14(3), 470-475.
- Singh, G. (2010). *The soybean: botany, production and uses*: CABI.
- Sun, L., Miao, Z., Cai, C., Zhang, D., Zhao, M., et al. (2015). GmHs1-1, encoding a calcineurin-like protein, controls hard-seededness in soybean. *Nat Genet*, 47(8), 939-943.
- Thakare, K., Chore, C., Deotale, R., Kamble, P., Pawar, S., et al. (2006). Influence of nutrients and hormones on biochemical and yield and yield contributing parameters of soybean. *Journal of Soils and Crops*, 16(1), 210-216.
- Tian, Z., Wang, X., Lee, R., Li, Y., Specht, J. E., et al. (2010). Artificial selection for determinate growth habit in soybean. *Proc Natl Acad Sci U S A*, 107(19), 8563-8568.
- Urano, K., Kurihara, Y., Seki, M., e Shinozaki, K. (2010). 'Omics' analyses of regulatory networks in plant abiotic stress responses. *Curr Opin Plant Biol*, 13(2), 132-138.

- USDA. (2020). UNITED STATES DEPARTMENT OF AGRICULTURE. . *Production, Supply and Distribution (PSD) on line.*, Disponível em: <https://apps.fas.usda.gov/psdonline/app/index.html#/app/advQuery>. Acesso em: 31 ago. 2020.
- Velculescu, V. E., Zhang, L., Vogelstein, B., e Kinzler, K. W. (1995). Serial analysis of gene expression. *Science*, 270(5235), 484-487.
- Wally, O., e Punja, Z. K. (2010). Genetic engineering for increasing fungal and bacterial disease resistance in crop plants. *GM Crops*, 1(4), 199-206.
- Warrington, C. V., Abdel-Haleem, H., Hyten, D. L., Cregan, P. B., Orf, J. H., et al. (2015). QTL for seed protein and amino acids in the Benning x Danbaekkong soybean population. *Theor Appl Genet*, 128(5), 839-850.
- Weber, H., Borisjuk, L., e Wobus, U. (2005). Molecular physiology of legume seed development. *Annu Rev Plant Biol*, 56, 253-279.
- Wu, F., Sedivy, E. J., Price, W. B., Haider, W., e Hanzawa, Y. (2017). Evolutionary trajectories of duplicated FT homologues and their roles in soybean domestication. *Plant J*, 90(5), 941-953.
- Yang, S., Tang, F., Gao, M., Krishnan, H. B., e Zhu, H. (2010). R gene-controlled host specificity in the legume-rhizobia symbiosis. *Proc Natl Acad Sci U S A*, 107(43), 18735-18740.
- Yin, X., e Struik, P. C. (2010). Modelling the crop: from system dynamics to systems biology. *J Exp Bot*, 61(8), 2171-2183.
- Yuan, S. L., Li, R., Chen, H. F., Zhang, C. J., Chen, L. M., et al. (2017). RNA-Seq analysis of nodule development at five different developmental stages of soybean (*Glycine max*) inoculated with *Bradyrhizobium japonicum* strain 113-2. *Sci Rep*, 7, 42248.
- Zhang, H., He, L., e Cai, L. (2018). Transcriptome Sequencing: RNA-Seq. *Methods Mol Biol*, 1754, 15-27.
- Zhang, Y., Zhao, J., Xiang, Y., Bian, X., Zuo, Q., et al. (2011). Proteomics study of changes in soybean lines resistant and sensitive to *Phytophthora sojae*. *Proteome Sci*, 9, 52.
- Zhao, S., Fung-Leung, W. P., Bittner, A., Ngo, K., e Liu, X. (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS One*, 9(1), e78644.
- Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., et al. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nature biotechnology*, 33(4), 408-414.

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

Autores: Fabricio Brum Machado^{1,#}, Kanhu C. Moharana^{1,#,*}, Fabricio Almeida-Silva¹, Rajesh K. Gazara¹, Francisnei Pedrosa-Silva¹, Fernanda S. Coelho¹, Clícia Grativol¹, Thiago M. Venancio^{1,*},

¹ Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro; Campos dos Goytacazes, Brazil.

These authors contributed equally to this work.

* Corresponding authors.

2.1. SUMMARY

Soybean (*Glycine max* [L.] Merr.) is a major crop in animal feed and human nutrition, mainly for its rich protein and oil contents. The remarkable rise in soybean transcriptome studies over the past five years generated an enormous amount of RNA-seq data, encompassing various tissues, developmental conditions, and genotypes. In this study, we have collected data from 1,298 publicly available soybean transcriptome samples, processed the raw sequencing reads, and mapped them to the soybean reference genome in a systematic fashion. We found that 94% of the annotated genes (52,737/56,044) had detectable expression in at least one sample. Unsupervised clustering revealed three major groups, comprising samples from aerial, underground, and seed/seed-related parts. We found 452 genes with uniform and constant expression levels, supporting their roles as housekeeping genes. On the other hand, 1,349 genes showed heavily biased expression patterns towards particular tissues. A transcript-level analysis revealed that 95% (70,963/74,490) of the assembled transcripts have intron chains exactly matching those from known transcripts, whereas 3,256 assembled transcripts represent potentially novel splicing isoforms. The dataset compiled here constitute a new

resource for the community, which can be downloaded or accessed through a user-friendly web interface at <http://venanciogroup.uenf.br/resources/>. This comprehensive transcriptome atlas will likely accelerate research on soybean genetics and genomics.

2.2 INTRODUCTION

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

To date, several studies reported spatiotemporal changes occurring in various soybean tissues using RNA-seq. The two first soybean RNA-Seq studies were published by Libault et al. (Libault et al., 2010) and Severin et al. (Severin et al., 2010). The former reported the sequencing of 14 (mainly root and nodule) tissues, whereas the latter evaluated several tissues and seed developmental stages. Dozens of other studies followed, such as those addressing different life cycle stages (Jones and Vodkin, 2013, Bellieny-Rabelo et al., 2016, Gazara et al., 2019), conditions (Belamkar et al., 2014), and cultivars/lines (Goettel et al., 2014). The accumulation of plant transcriptomic data in public repositories [e.g. Sequence Read Archive (SRA) at the National Center for Biotechnology Information (NCBI)] inspired the development of unified collections or atlases, such as those found for

Arabidopsis thaliana (Fucile et al., 2011), *Medicago truncatula* (He et al., 2009), *G. max* (Supplementary Table S1), as well as multi-species atlases (Dash et al., 2012), which are often reused by the scientific community. Specifically in soybean, Kim et al. constructed the SoyNet (www.inetbio.org/soynet) database using 734 microarrays and 290 RNA-seq samples (Kim et al., 2017), while Wu et al. uncovered a nodulation-related co-expression module by analyzing 1,270 microarray samples generated with Affymetrix gene chips (Wu et al., 2019).

Despite the previous efforts to integrate soybean transcriptomes, there is a massive amount of soybean RNA-Seq data that remain largely unexplored. Here, we have collected data from 1,298 publicly available soybean RNA-seq samples from the NCBI SRA database. We systematically processed and mapped sequencing reads to the soybean reference genome (assembly version: Gmax_275_Wm82.a2.v1). Transcriptional levels were estimated to allow a systematic global gene expression analysis, aiming to elucidate the dynamics of transcriptional regulation across this broad range of samples, tissues, and cultivars. Further, the collected and processed data are readily available to allow both, automatic analysis and single-gene investigations through an easy-to-use interface at our lab website (<http://venanciogroup.uenf.br/resources/>).

2.3 RESULTS AND DISCUSSION

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

We performed an extensive literature mining process to gather as many as possible soybean RNA-seq datasets. A total of 1,742 raw read sequencing files were downloaded from the NCBI SRA database (Supplementary Table S2). Reads obtained from the same biological sample were combined in a single FASTQ file (or in two files, for paired-end data; *_1.fq and *_2.fq). This resulted in 1,298 samples (65% single-end and 35% paired-end) from 84 BioProjects comprising sixteen different broad tissue categories in various developmental stages (Supplementary Table S3). Approximately 35% (458/1298) of the samples lacked cultivar/genotype information in SRA. Among the other 840 samples, we found 157 different soybean

cultivar names, although this is likely an overestimation because of authors calling the same cultivars with slightly different names during data submission. The cultivar Williams 82, which had the genome sequenced, represented 23% (302/1,298) of the total samples. Leaf was the most abundant tissue, representing 46% (603/1,298) of the samples (Figure 1). Three libraries from unknown tissue sources were excluded. We have also found that 76% (986/1,295) of the libraries were unstranded (Supplementary Table S3).

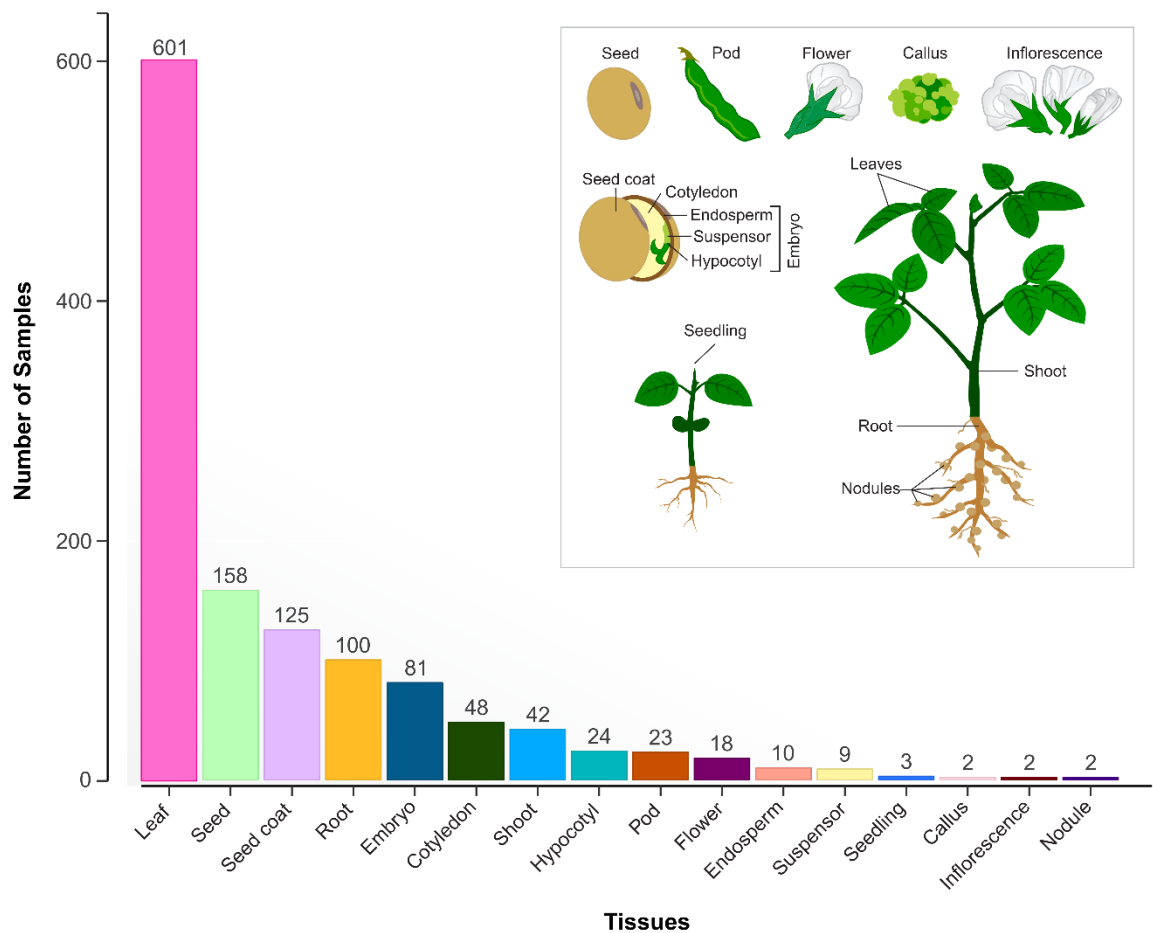


Figure 1: Number of samples analyzed in this study and a graphical representation of each tissue.

Reads from each RNA-seq library were mapped to the reference genome, assembled, and used for estimating gene expression (Figure 2). Whenever present, adapter sequences were trimmed. Reads with average quality lower than 20 were excluded. An average of 32,210,805 million reads pairs per sample with paired-end data and 29,579,316 million reads per sample with single- end data were used for read mapping. Mapped and uniquely mapped reads correspond to an average of 87.9% and 81%, respectively (Supplementary Table S4 and Supplementary Figure S1). Further, we excluded 47 samples for which: i) 50% or more of the reads failed to map or; ii) 40% or more of the reads failed to uniquely map. After these exclusions, 1,248 samples were kept for further downstream analysis.

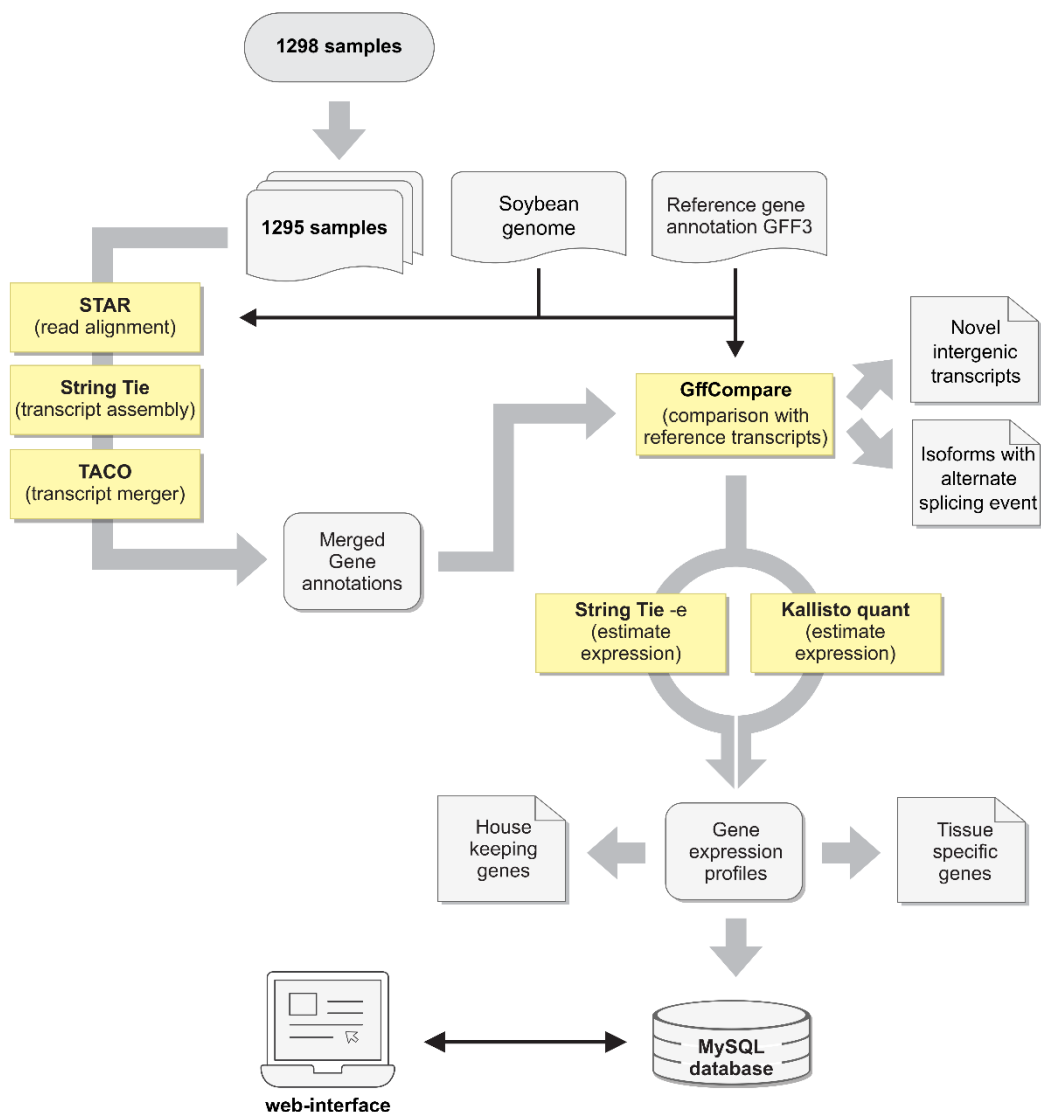


Figure 2: Pipeline used to create the soybean RNA-Seq atlas.

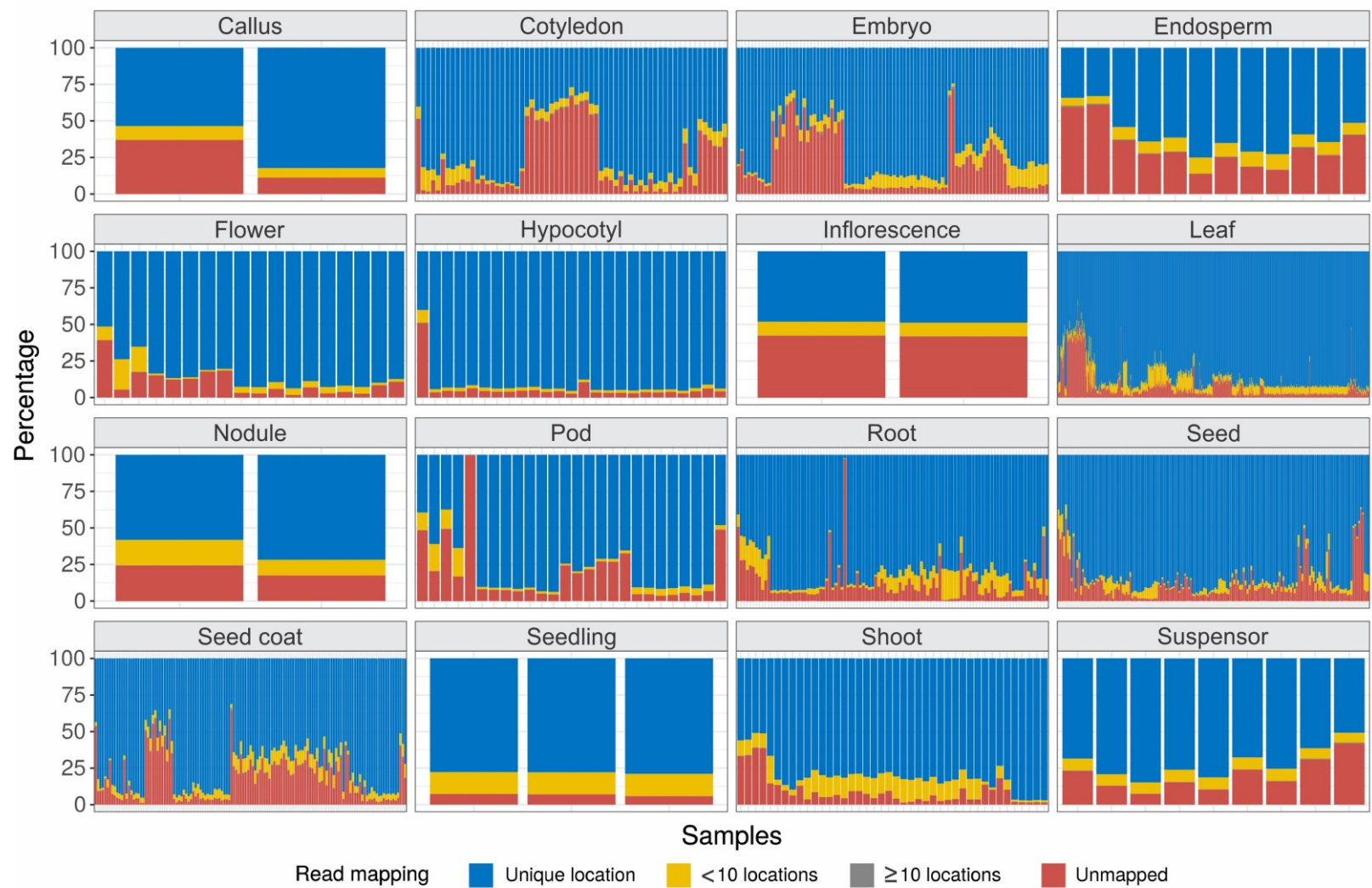


Figure S1: Stacked histograms of read mapping statistics across tissues.

Several methods used to analyze RNA-seq data (e.g. differential gene expression) rely on read count normalization strategies (Robinson and Oshlack, 2010, Po-Yen et al., 2011), such as Reads Per Kilobase Million (RPKM) (Mortazavi et al., 2008), Fragments Per Kilobase Million (FPKM), and Transcripts Per Million (TPM) (Wagner et al., 2012), out of which the latter has been proposed to be more consistent across technical replicates (Wagner et al., 2012, Conesa et al., 2016, Li and Li, 2018). Here, we normalized data using TPM for most of the downstream analysis. Nevertheless, log₂ transformed raw read counts are more commonly used for quality control steps such as unsupervised sample clustering (Jordan et al., 2015). In addition, many popular tools used for differential gene expression analysis (e.g. DESeq2, edgeR) require raw read counts instead of normalized read counts. Therefore, after read mapping, we estimated transcript abundances in the form of raw read counts per transcript and TPM. Transcript-level expression values were also aggregated to estimate expression at gene level. To check the robustness of our TPM estimations, we compared gene-level TPM values from StringTie and Kallisto. We observed that 98.1% (1263/1287) and 92.6% (1192/1287) of the samples showed at least 0.9 and 0.95 Spearman's rank correlation coefficient between the two methods (Supplementary Figure S2), strongly suggesting that both methods performed equally well in our dataset. Since Kallisto does not support SOLiD reads, which are present for some important samples, we kept StringTie TPM estimates in the downstream analyses described below, unless stated otherwise. Importantly, TPM quantification with both methods are available for download as well as for visualization in our website

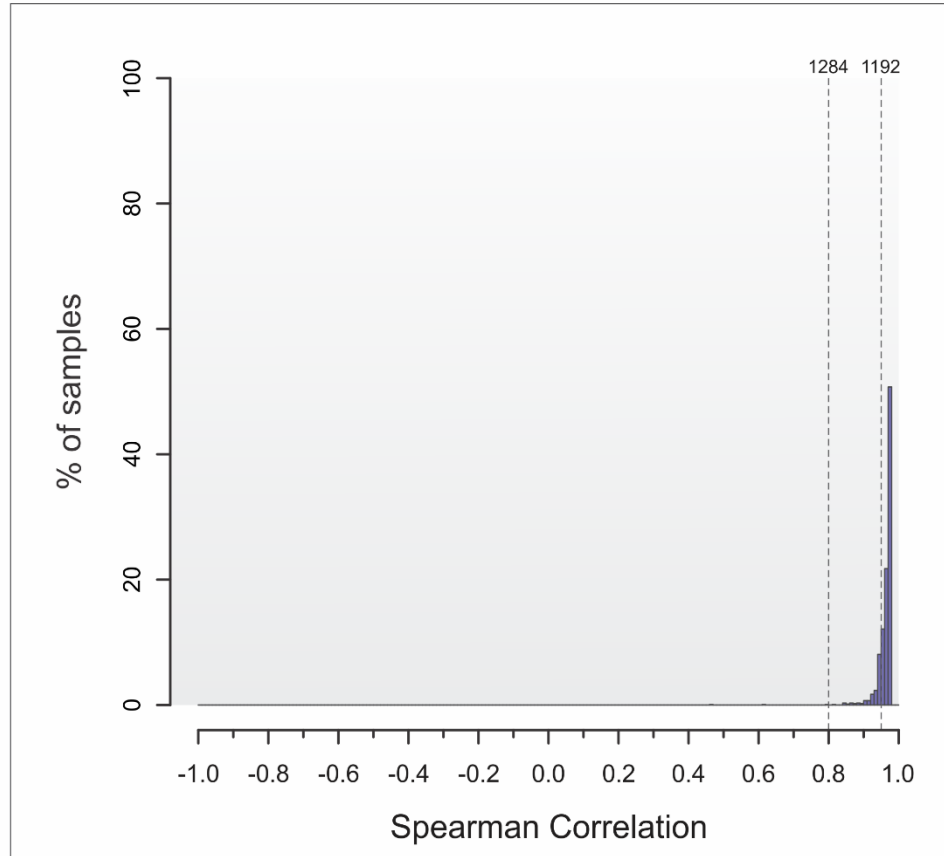


Figure S2: Distribution of Spearman's rank correlation coefficients between normalized expression values (in TPM) estimated with kallisto and stringtie across samples.

2.3.1 Unsupervised sample clustering reveals three major clades comprising underground, aerial, and seed tissues

In transcriptomics studies, genes and samples are often clustered to identify sub-groups with similar transcriptional profiles (Liu and Si, 2014, Marini and Binder, 2019). While gene clustering helps identify co-expressed genes, sample clustering is instrumental to detect broad transcriptional similarities between samples, as well as to identify potential technical artifacts and mislabeled samples. Among several methods, distance-based hierarchical clustering, K-means clustering, and dimensionality-reduction based methods (e.g. principal component analysis, PCA) are commonly used. Recently, t-Distributed Stochastic Neighbor Embedding (t-SNE) has been shown to provide a better global structure of sample sub-groups than several other methods (Dey et al., 2017). Here, we employed three sample clustering methods to identify outliers and overall pairwise sample similarity. We

used a gene expression matrix as input to perform hierarchical clustering, K-means clustering, and t-SNE analysis. These analyses uncovered three major groups comprising samples from aerial, underground, and developmental or seed tissues (Figure 3) (Severin et al., 2010). Interestingly, however, we found an additional cluster comprising leaf and shoot samples from drought-stress-related and leaf senescence conditions. Although not entirely novel, these results are part of an important step to check for technical issues or biases that could, for example, result in the clustering of samples from the same sequencing batch or research group. Four shoot samples and one root sample clustered with seed-embryo samples. After confirming this result with the t-SNE and K-means clustering, we excluded these samples. Overall, sample clustering supports a high quality level of the publicly available RNA-Seq samples analyzed here, as only 0.4% (5/1248) of the samples were excluded after the clustering analysis.

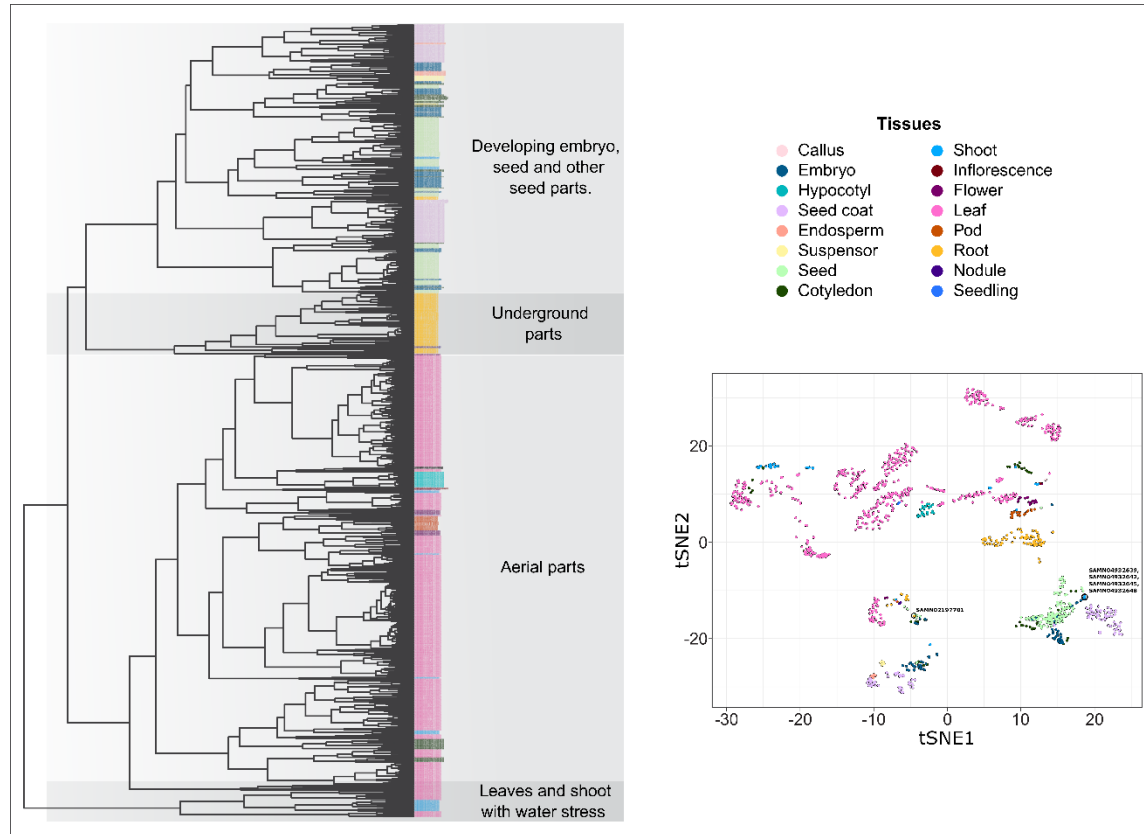


Figure 3: Hierarchical clustering of samples using their transcriptional profiles. Per gene raw read counts were used to perform hierarchical clustering using the R function `hclust()` with default parameters. Samples were grouped into three major clades: aerial, underground, and seed-embryo related.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

2.3.2 Systematic analysis of hundreds of RNA-Seq libraries support the expression of the vast majority of the soybean genes

After comparing the reference transcript annotations (for 56,044 genes) with the merged consensus transcript assembly, we excluded 1.3% (759/56,044) of the genes because of overlapping gene predictions. Next, we applied a minimum TPM threshold of 1 to define a gene as expressed and found that 92.1% (51,644/56,044) of the known soybean protein-coding genes were expressed in at least one sample. The remaining genes had their TPM values set to zero and classified as not expressed. An average of 31,063 genes were expressed per sample. The tissues with the greatest numbers of expressed genes were inflorescence (37,108 genes) and flower (average of 36,051 genes) (Supplementary Figure S3A), whereas nodules had the lowest number of expressed genes (average of 25,718 genes). We also found 16,916 genes expressed in at least 1,150 samples (Supplementary Figure S3B), including 1,758 genes that are expressed in all 1,243 samples. On the other hand, 6% (3,233/56,044) of the genes were not expressed (TPM < 1) in any sample, out of which 82% had coding regions comprising less than 500 codons (Supplementary Figure S4). As a final data quality check, we analyzed the top 1,000 expressed genes from each tissue category using MapMan pathway bins (see Methods). For example, by contrasting gene expression profiles of root and leaf samples, we uncovered several expected transcriptional patterns of photosynthesis genes in the latter (Supplementary Figure S5).

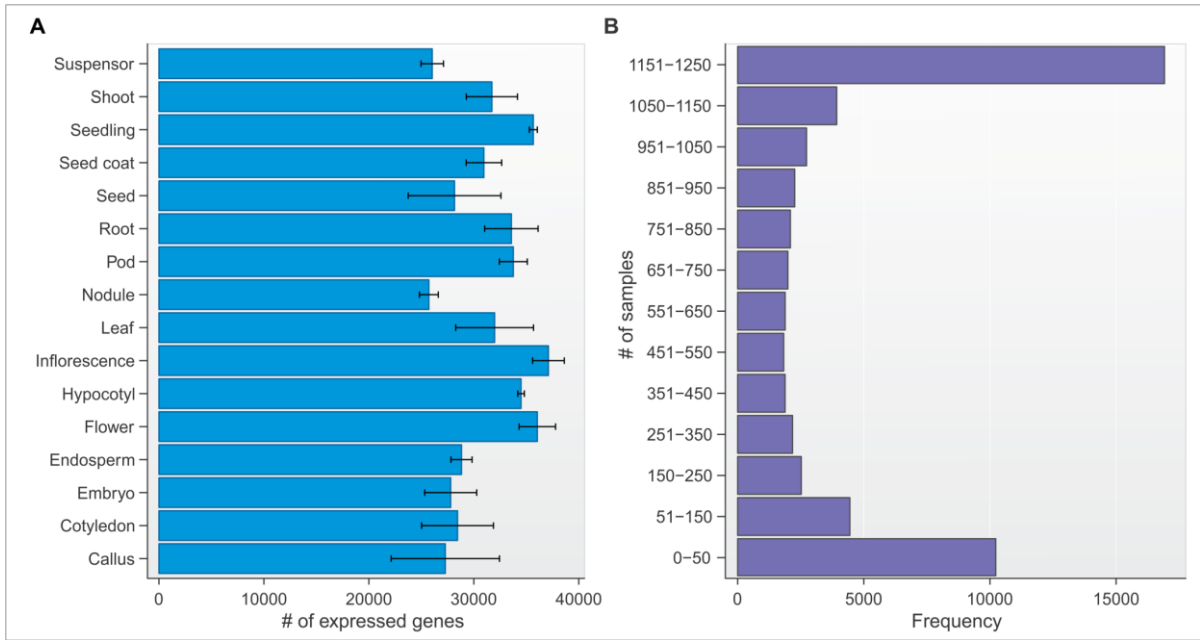


Figure S3: Tissue- and sample-wise distribution of expressed genes (TPM \geq 1). A. Number of expressed genes in each tissue. B. Number of samples in which genes are expressed.

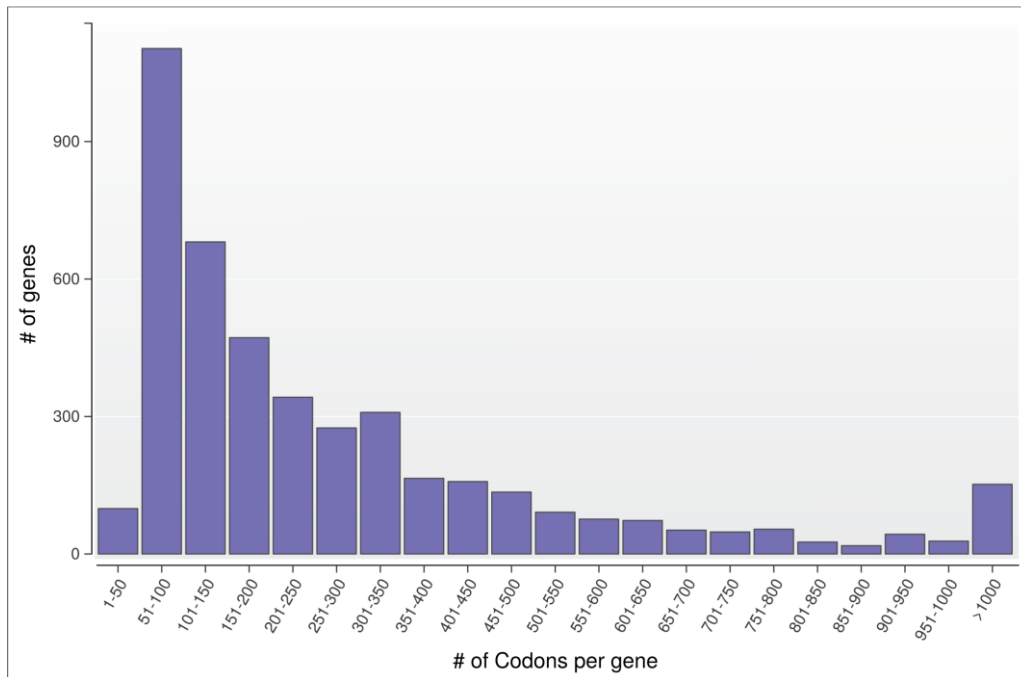


Figure S4: Length of coding regions with undetectable expression levels (TPM $<$ 1).

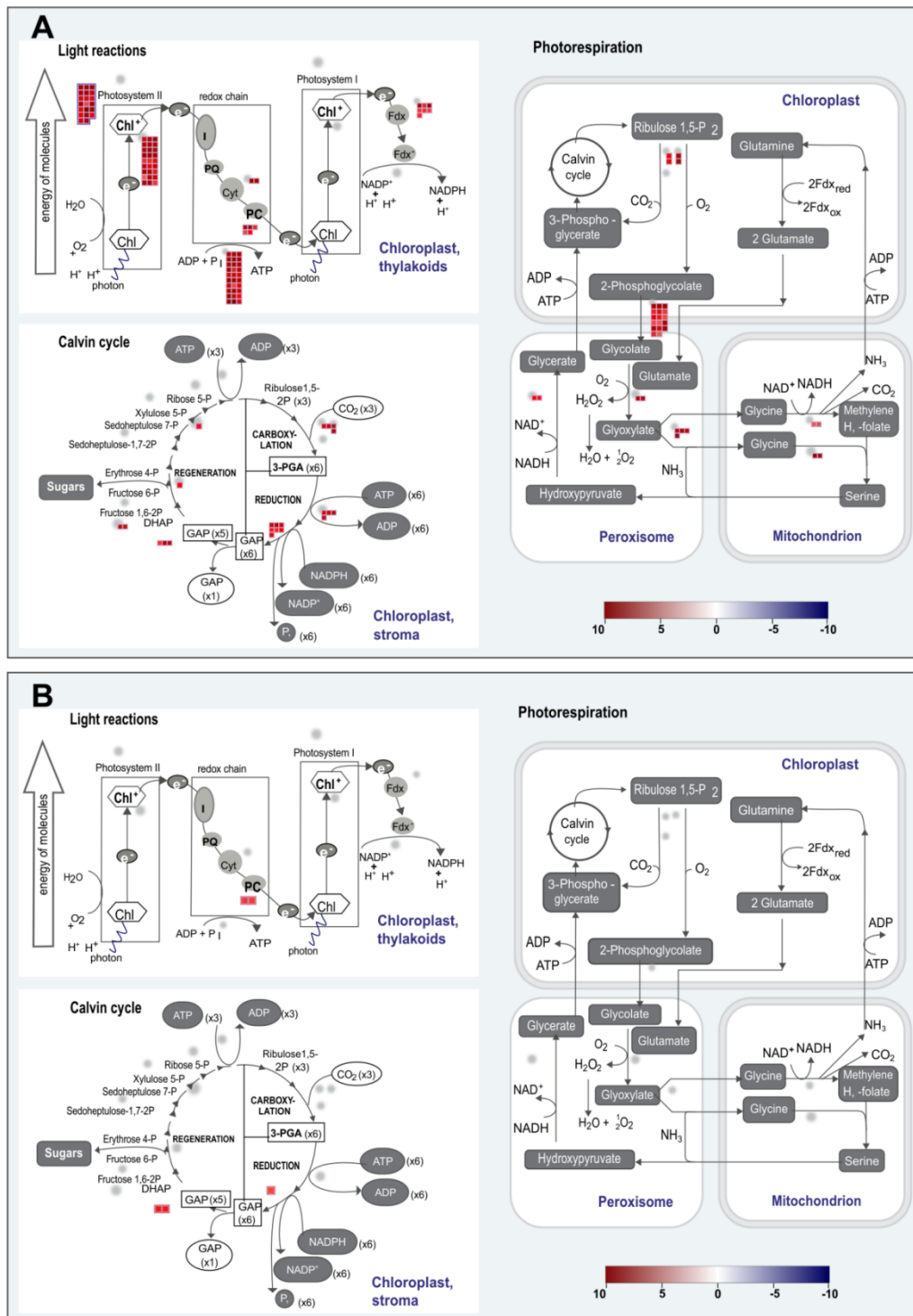


Figure S5: Pathway analysis of the top 1,000 highest expressed genes in leaves and roots. As expected, we found that photosynthesis genes are enriched in the former. The small groups of boxes in each pathway represent genes involved in that process. The color of these boxes ranges from dark red to dark blue representing extremely high expression and low expression, respectively.

2.3.3 Housekeeping genes

Given the wide coverage of tissues and conditions, we also sought to identify housekeeping (HK) genes based on the assumption that these genes are constitutively and robustly expressed across broad conditions (Czechowski et al., 2005, Hu et al., 2009). Further, several of these genes have also been used as references in real-time quantitative polymerase chain reaction (RT-qPCR) assays (Supplementary Table S5). Hence, by using a large collection of RNA-Seq datasets as the one presented here, one can not only evaluate commonly used reference genes, but also propose new ones. By employing a previously developed method (Hoang et al., 2017), we inferred 452 HK genes (Supplementary Table S6). We evaluated expression levels of each gene in tissues with at least 10 samples and found that HK genes had very low expression variation (Figure 4A). To identify HK genes, we used a score that consists of the product of the Coefficient of Variation and ratio of the maximum to the minimum expression level (see methods for details). Genes with scores within the 1st quartile were classified as HK genes. Further, we used a tissue-specificity index Tau (τ) (Yanai et al., 2004, Kryuchkova-Mostacci and Robinson-Rechavi, 2017) to estimate tissue specificity and verify whether our predicted HK genes were broadly expressed or not. The τ values scale from 0 to 1, where low and high values indicate widely expressed and more tissue-specific genes, respectively. The τ scores of the HK genes ranged from 0.053 to 0.379, supporting their stable expression level (Figure 5).

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

Pathway enrichment analysis of the 452 putative HK genes revealed that these genes are involved in various biological processes such as RNA degradation, mRNA surveillance, and TCA cycle (Figure 4B). We also found an enrichment of orthologs of Arabidopsis essential genes (Meinke, 2019) among the HK genes

(Fisher's Exact test; p-value = 1.76e-2). Given their roles in basic biological processes, we also verified the conservation of the HK genes in other 14 species on Phytomine and found that 85% (385/452) of them have orthologs in at least 10 other species (Supplementary Table S6), as opposed to an average of 181.6 (\pm 11.6) in 5 random lists of 452 non-HK genes.

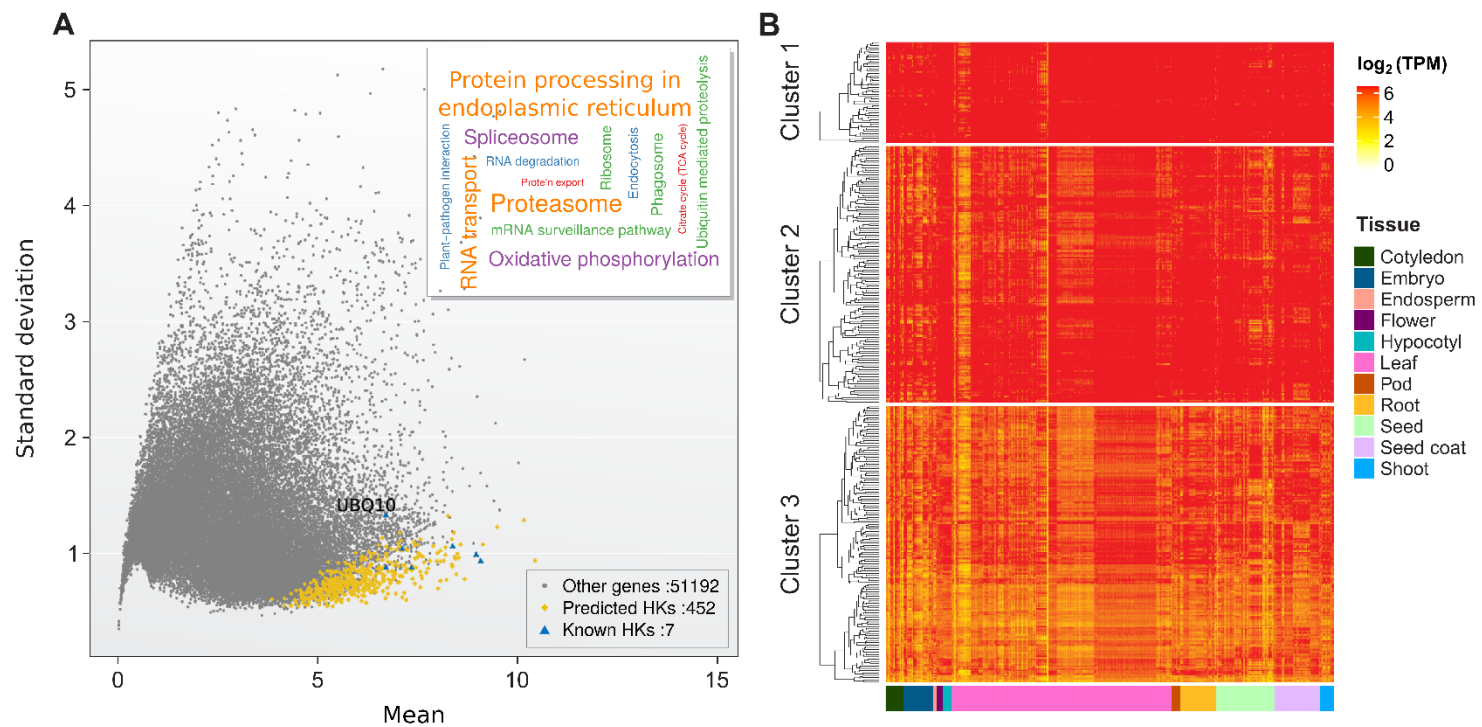


Figure 4: Global gene expression patterns of the housekeeping (HK) genes. A. Scatter plot of mean vs standard deviation showing uniform and stable expression of 452 HK genes. The gray dots represent all the non-HK expressed genes (TPM \geq 1 in at least one sample). The word cloud represents KEGG pathways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

2.3.4 Tissue-specific gene expression

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

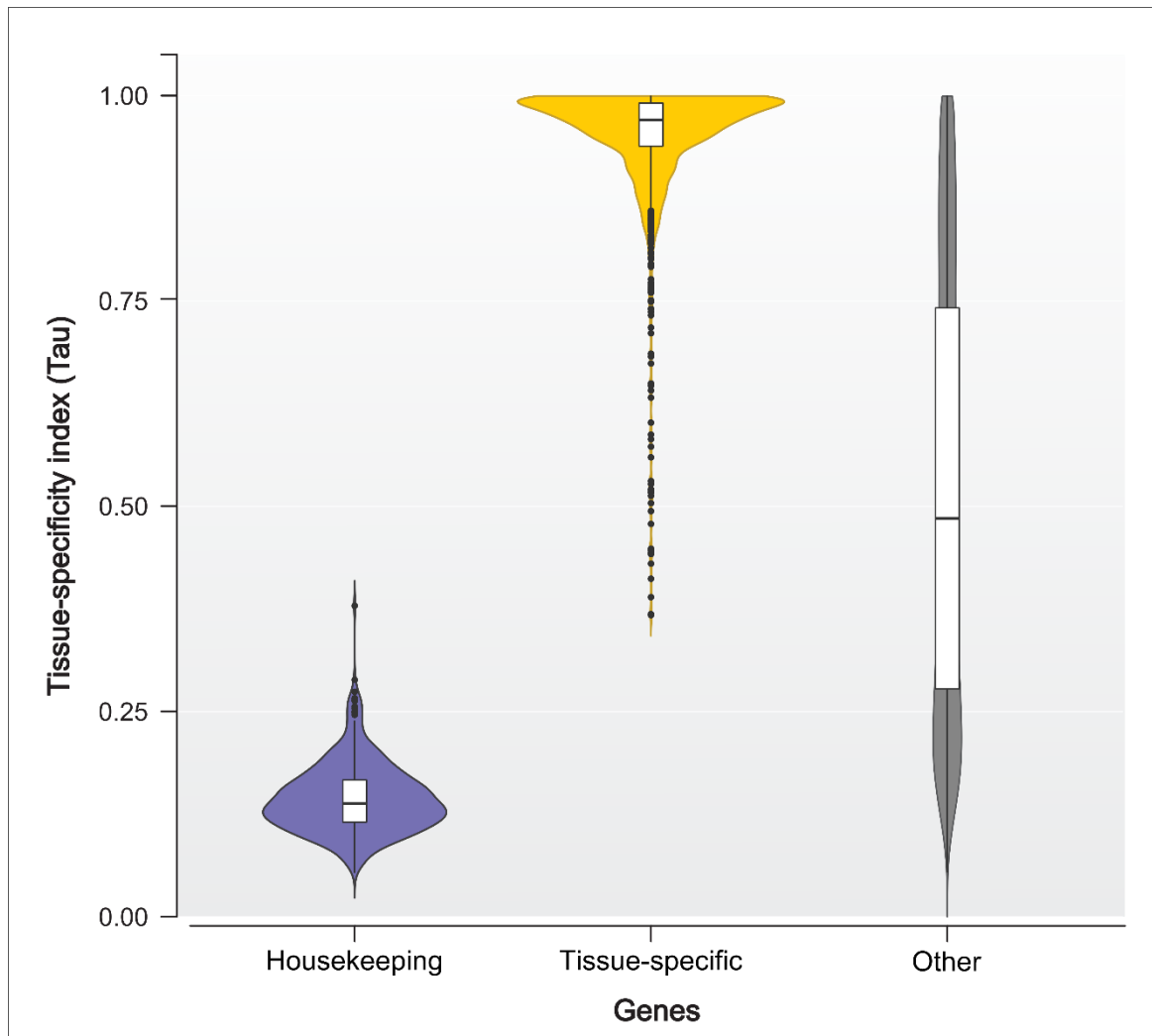


Figure 5: Violin plot showing the distribution of Tau indexes of housekeeping, tissue-specific, and the remaining genes. Tau values range between 0 and 1, with low values indicating a stable and constitutive expression and higher values supporting tissue-specificity.

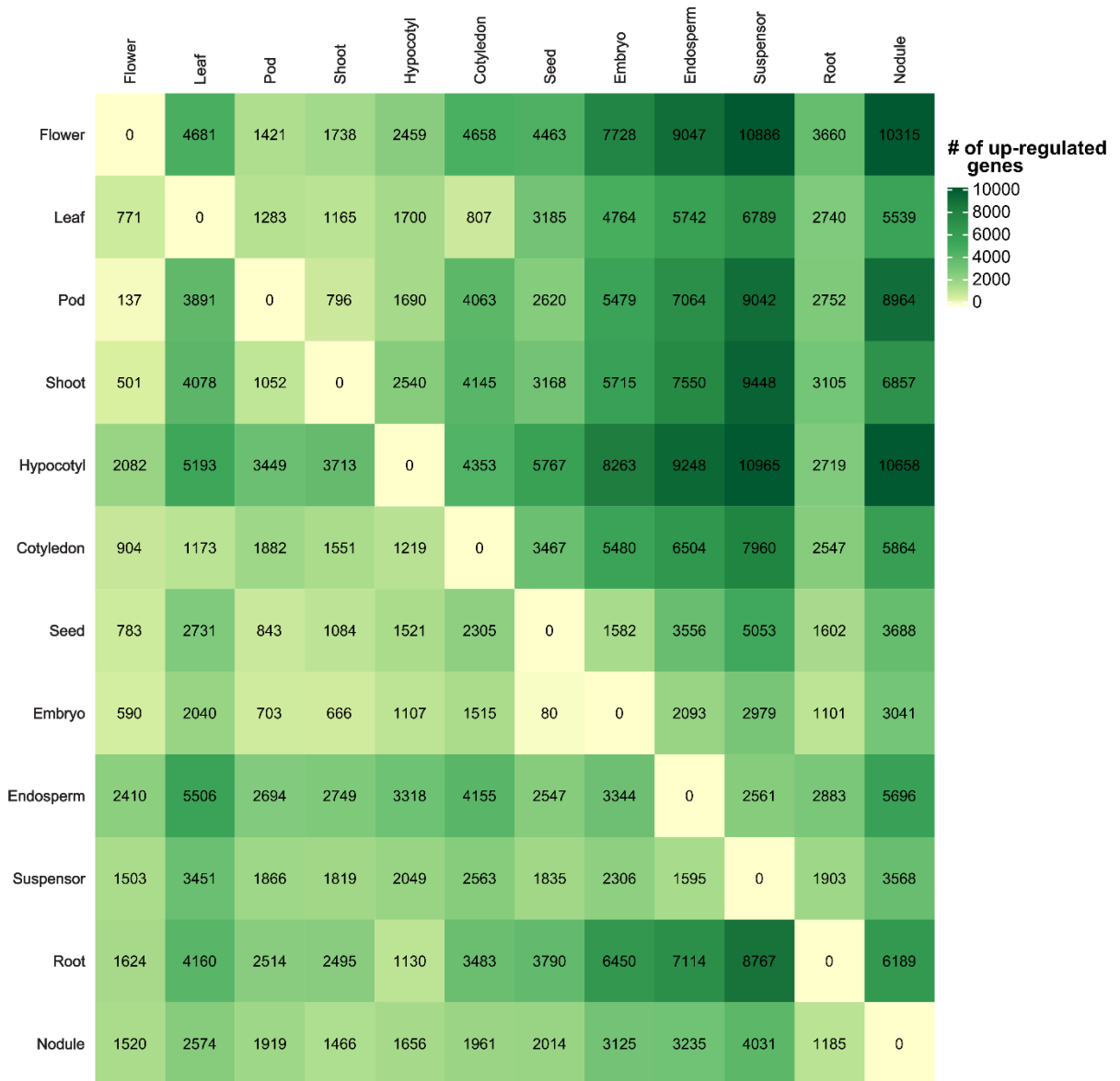


Figure 6: Heatmap showing the number of up-regulated genes in the tissues from the rows when compared with those from the columns. Gene up-regulation was determined by using a $\log_2(\text{foldchange}) \geq 2$ and adjusted $p\text{-value} \leq 0.05$ using the moderated t-statistic in the limma package.

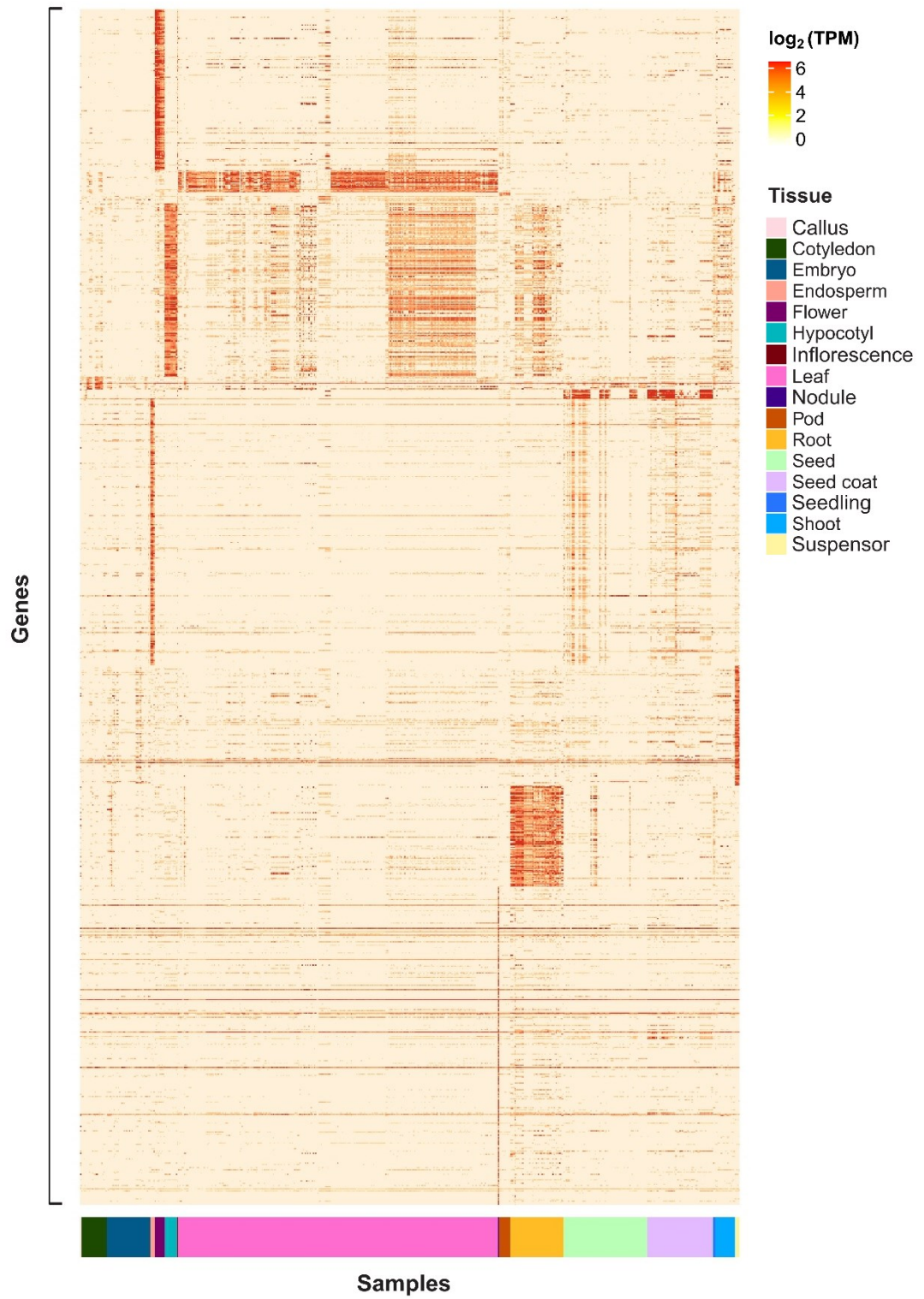


Figure 7: Global transcriptional patterns of tissue-specific genes. Expression values are represented as log₂(TPM) values in 1243 samples.

The number of tissue-specific genes ranged from 4 in pods to 358 in nodules. Collectively, nodule (26.5%) and endosperm (301; 22%) account for nearly half of the tissue-specific genes. The lower number of tissue-specific genes in leaf, shoot, cotyledon, and pod can be explained by the physiological or developmental relatedness of some samples (e.g. cotyledon and seed). Notably, 39% (520/1,349) of the tissue-specific genes identified here were also identified by Severin et al. (Severin et al., 2010) using a much smaller set of samples, supporting the general high quality and reproducibility of the publicly available soybean transcriptomes. Strikingly, nearly 12% (168/1,349) of the tissue-specific genes were transcription factors (TFs) (Table 1), which is a remarkable enrichment (Fisher's Exact Test, p-value = $2.94e-11$) considering the overall abundance of TFs in the soybean genome (Moharana and Venancio, 2020). Among the tissue-specific TFs, 27, 21, and 20 genes belong to the MYB, C2H2, and ERF families, respectively. Of the 27 MYB TFs, 20 were specific to flower (n=8), hypocotyl (n=7), and endosperm (n=5). Of the 21 C2H2 genes, 12 were specific to nodule (n=6) and endosperm (n=6). Ten out of 20 ERF genes and six out of 10 WRKY genes were specific to hypocotyl. Finally, 8 of 9 MIKC type MADS TFs were flower-specific. Several interesting tissue-specific genes are discussed in the sections below.

Table 1: Tissue-specific transcription factors.

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

2.3.5 Nodule-specific genes

Symbiotic N₂ fixation takes place in root nodules of several Fabaceae species. Nodulation had a single origin in the common ancestor of the N₂-fixing clade, followed by multiple independent losses (Griesmann et al., 2018). Among the genes

lost in non-nodulating species, Nodule Inception (NIN) and Rhizobium-Directed Polar Growth (RPG) were reported to be of paramount importance for the origin of root nodules (Griesmann et al., 2018). As mentioned above, nodule is the tissue with the greatest number of tissue-specific genes in soybean, a trend that has also been reported in other legumes (Benedito et al., 2008). Soybean nodules have been shown to correlate poorly with other tissues at the transcriptional level (Severin et al., 2010), a finding that we also corroborated here.

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

We found 12 soybean nodule-specific genes within the experimentally validated list of over 200 nodulins described previously (Roy et al., 2019). These 12 genes include the above mentioned ERF TFs, NIN (Glyma.04G000600), C2H2 (Glyma.07G135800), and GRAS (Glyma.16G008200). Next, we analyzed the 28 genes from a nodule-related module identified in a co-expression network derived from soybean microarray data (Wu et al., 2019). Notably, 9 of these 28 genes were identified as nodule-specific in our analysis: one leghemoglobin (Glyma.10G199000), two NIN-like TFs (Glyma.02G311000, Glyma.14G001600), two purine biosynthesis genes (Glyma.08G001000, Glyma.11G221100), one iron transporter (Glyma.05G121600), one zinc finger protein-related (Glyma.08G044700), one sulfate transporter (Glyma.18G018900), and a formyl transferase (Glyma.19G115900).

2.3.6 Endosperm-specific genes

The endosperm plays important roles during seed development. *Ar. thaliana* endosperm-specific genes are associated with cell cycle, DNA processing, chromatin assembly, protein synthesis,

cytoskeleton- and microtubule-related processes, and cell/organelle biogenesis and organization (Day et al., 2008). Out of the 301 endosperm-specific genes reported here, 9 (Glyma.19G040600, Glyma.09G194500, Glyma.01G147300, Glyma.19G058100, Glyma.19G044000, Glyma.04G187100,

Glyma.03G219800, Glyma.02G255900, and Glyma.08G129200) encode chromatin modifiers such as histone acetyltransferases, histone-lysine n-methyltransferases, histone deacetylases, and histone demethylases. Further, 17 endosperm-specific genes encode F-box proteins and 8 genes encode BTB-POZ and MATH domain proteins, which likely operate in the ubiquitin-proteasome pathway (Smalle and Vierstra, 2004, Figueroa, 2005). We also found 36 endosperm-specific TFs, including 6 and 5 C2H2 and MYB TFs, respectively. Together, these results clearly show a number of endosperm-specific genes as involved in transcriptional and post-transcriptional regulatory processes.

2.3.7 Flower-specific genes

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

suggested to be involved in soybean flower development, out of which 19 displayed flower-specific expression (Jung et al., 2012).

Recently, several studies reported transcriptional changes during flowering time in legumes (Weller and Ortega, 2015). We found 182 flower-specific genes, including at least 20 members of the plant invertase/pectin methylesterase inhibitor (PMEI) superfamily, which is involved in cell wall modification in *Ar. thaliana* (Zhao et al., 2015). Specific PMEIs are highly expressed in specific wheat

floral parts, such as anthers and pollen tubes (Rocchi et al., 2012), playing a significant role in flower development (Wormit and Usadel, 2018). In addition, we found 20 flower-specific TFs, mostly from the MYB (40%, 8/20) and MIKC-type MADS (40%, 8/20) families. Finally, out of these 8 MIKC genes, two AGAMOUS-like (Glyma.03G019400, Glyma.07G081300) and three PISTILLATA (Glyma.06G117600, Glyma.13G034100, Glyma.14G155100), corresponding to 36 transcripts, are represented among the 19 flower-specific genes reported by Jung et al. (Jung et al., 2012).

2.3.8 Identification of novel transcripts

We compared the genomic coordinates of the transcripts assembled in our atlas with those available in Phytozome and categorized them in nine classes (Table 2). We found that 95% (70,963/74,490) of the transcripts precisely matched the exon-intron splice junctions of known transcripts (class =). We also investigated class-J and class-U categories, which account for 3,256 and 23 transcripts, respectively. Class-J comprises multi-exon transcripts with at least one known exon junction, while class-U encompasses transcripts located in intergenic regions. While class-J transcripts include new isoforms of known genes, those from class-U are useful to identify potentially new genes. We found that 30% (983/3256) of the class-J transcripts and 17% (4/23) of the class-U transcripts had TPM \geq 1 in 907 and 1,207 samples, respectively. Only one of the four class-U expressed transcripts (TU4871, Chr02:12125821-12127123) encode a protein longer than 50 aa, which contains a reverse transcriptase-like RNase_H (PF13456) domain, suggesting that it is a

mobile element. In two of these expressed class-U transcripts (TU28093, TU56508), only one exon showed high read coverage (Supplementary Figure S6).

Table 2: Number of transcripts in each transcript-classification code defined by 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

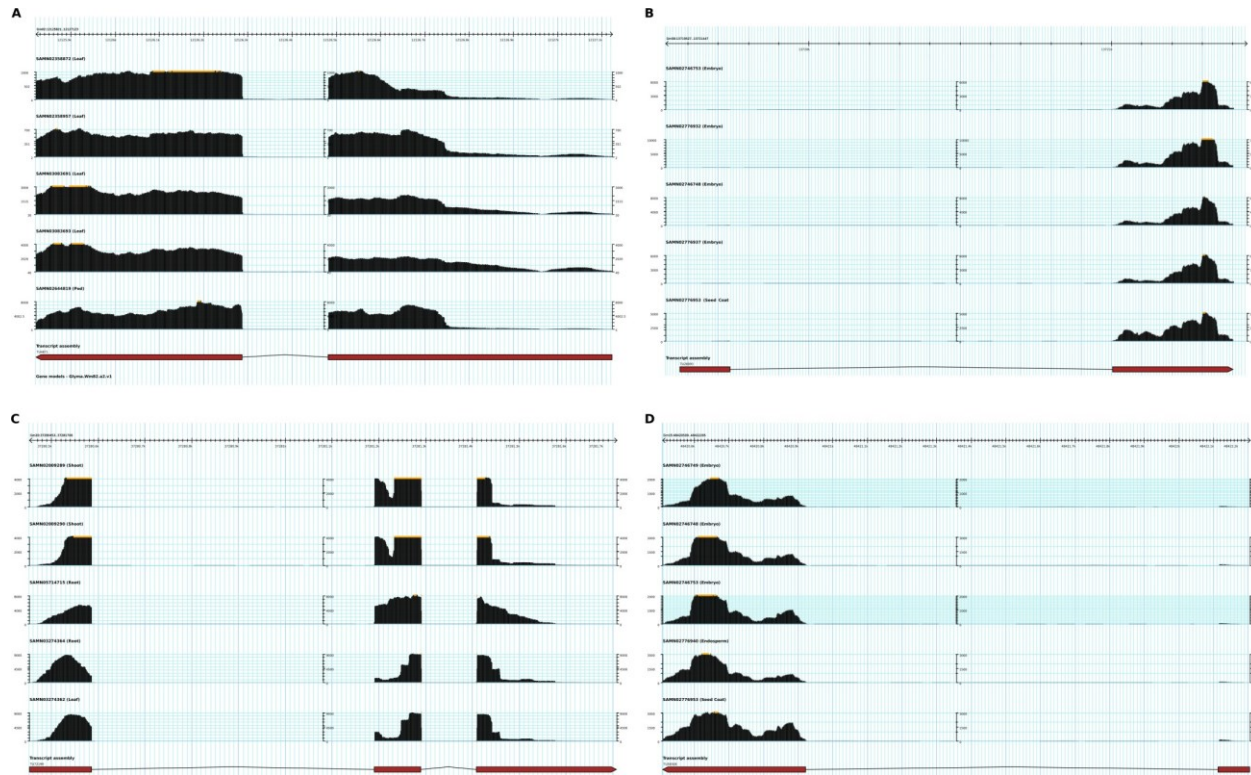








Figure S6: Wiggle plots showing read coverage of potentially novel genes at four unannotated loci in the soybean genome. We selected five samples in which these genes had the highest expression levels. A: TU4871, B: TU28093, C: TU72199, D: TU56508.

All the 3,256 class-J transcripts were further analyzed for alternate splicing (AS) events using ASprofile (Florea et al., 2013). AS events were categorized in one of six categories: (i) exon-skipping; (ii) multiple exon-skipping; (iii) alternative transcription start site (TSS); (iv) alternative transcription termination sites (TTS); (v) intron retention and; (vi) alternate 5' and/or 3' exon ends. We detected 6,582 AS events, mostly TSS and TTS (Table 3). Several novel AS events were supported by hundreds of split reads (Supplementary Figure S7-S9). For example, TU62356 from Glyma.17G195900 (CASEIN KINASE 1-LIKE PROTEIN 4) is a novel isoform with a skipped exon (Supplementary Figure S7). Interestingly, we found transcriptional support for this alternative isoform only in nodules.

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

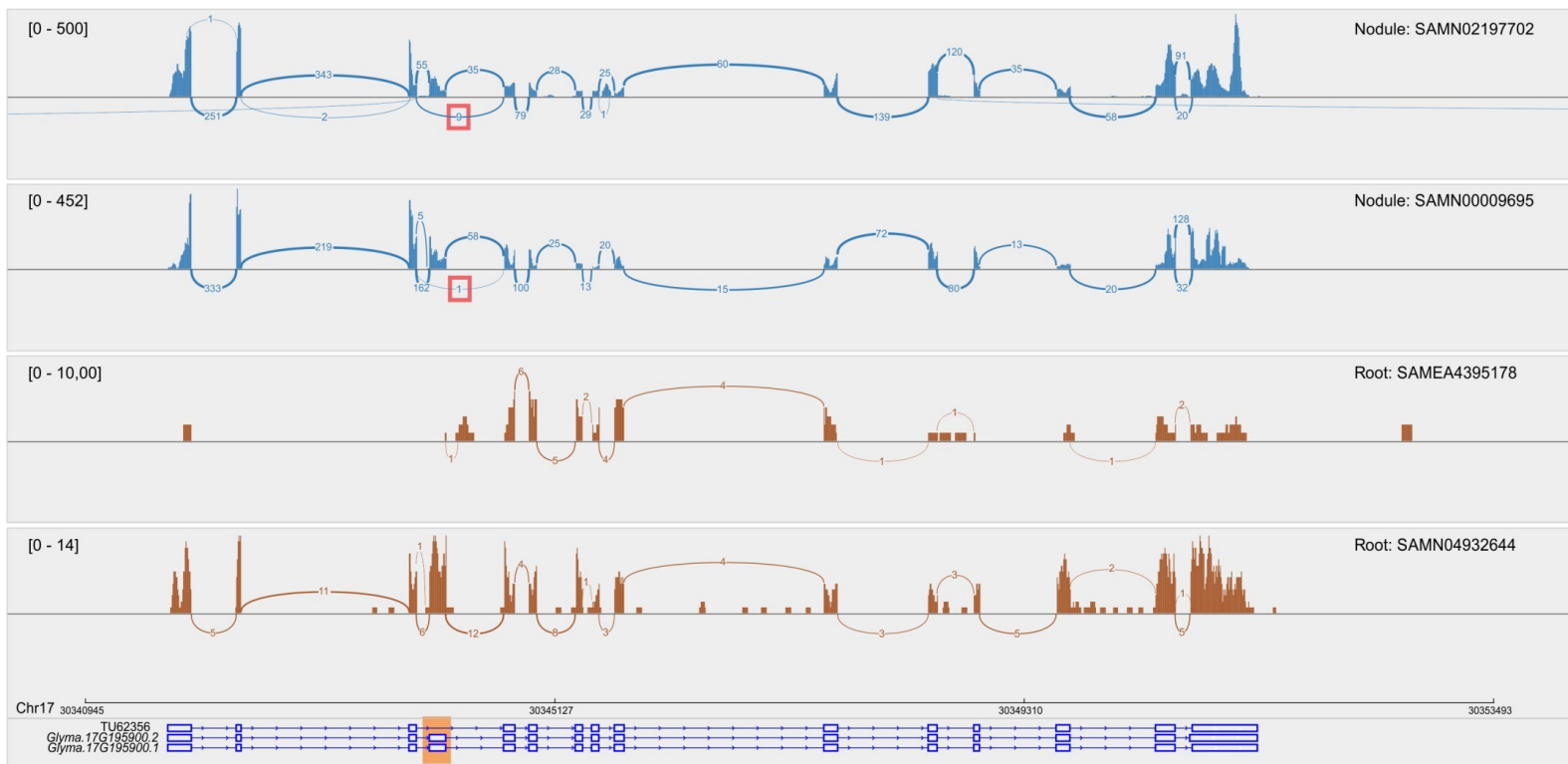


Figure S7: Sashimi plot of Glyma.17G195900 (CASEIN KINASE 1-LIKE PROTEIN 4) showing the number of reads supporting splice junctions in two nodule and two root samples. The tracks below the plot represent splicing isoforms. Exons within the highlighted region indicate variation in splicing patterns. The top track (TU62356) is the novel isoform, comprising an exon skipping event, which is supported by 10 reads from nodule samples.

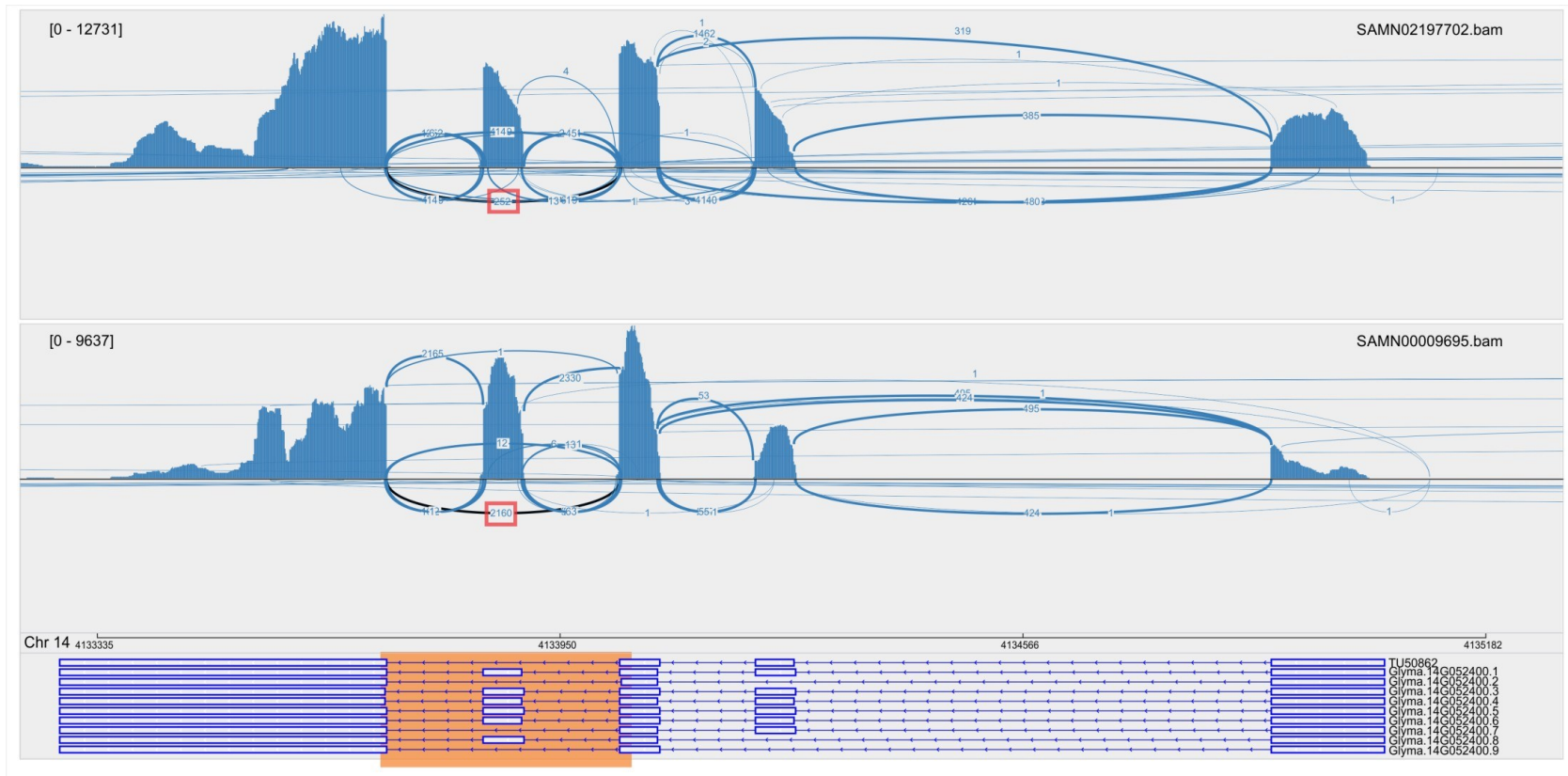


Figure S8: Sashimi plot of Glyma.14G052400 (Glycine rich protein family) showing the number of reads supporting the splice junctions in two nodule samples. The tracks below the plot represent splicing isoforms. The exon within the highlighted region indicates variation in splicing patterns due to the skipping of exon 2 in TU50862. This new isoform also has some small variations in exon junctions in exons 3 and 4. The skipping of exon 2 in the new isoform is supported by 2,412 reads from two nodule samples.

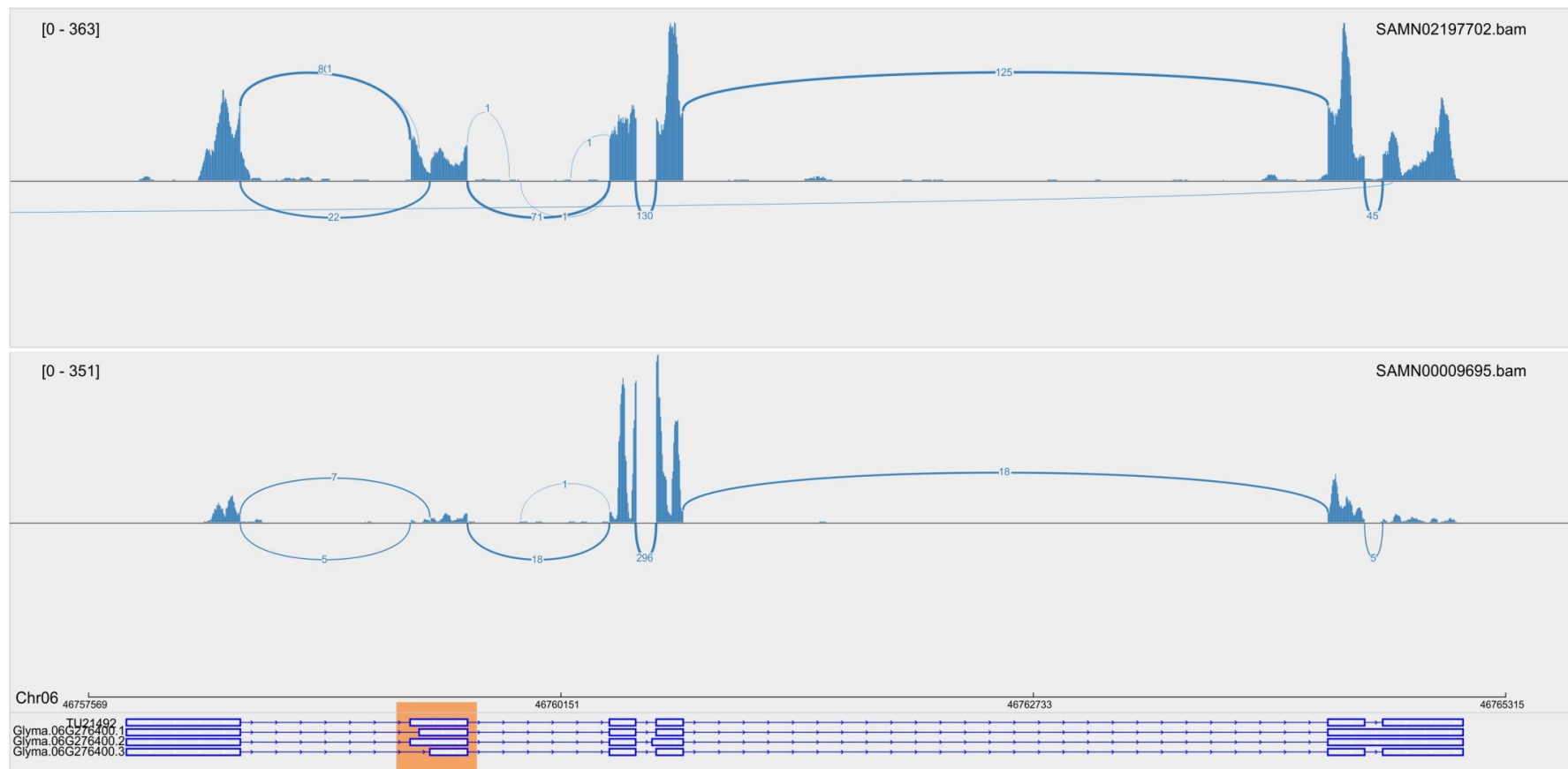
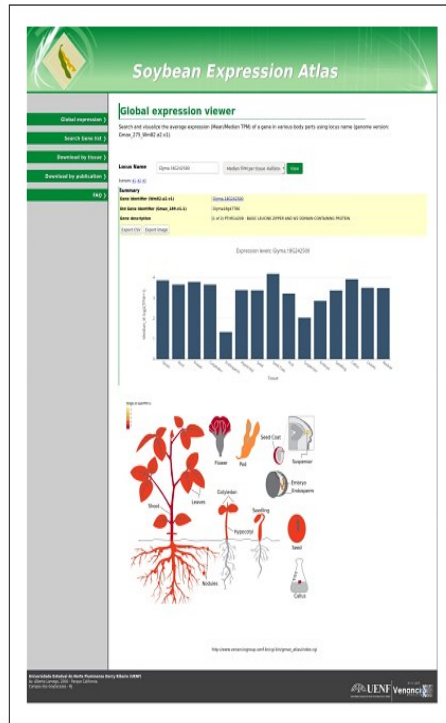


Figure S9: Sashimi plot of Glyma.06G276400 (Cysteamine dioxygenase/Persulfurase) showing the number of reads supporting the splice junctions in two nodule samples. The tracks below the plot represent splicing isoforms. Exons within the highlighted region indicate variation in splicing patterns. The top track (TU21492) is a novel isoform comprising a different length in exon 2, along with two terminal 3' exons instead of one, giving rise to a new combination of exons.

2.3.9 Data availability through a user-friendly web interface

We developed a simple user-friendly web interface to allow researchers to easily explore 1,243 soybean transcriptome samples. Through this interface (Figure 8), one can explore the expression of a particular gene in multiple tissues, with the aid of an image illustrating all the available tissues. Alternatively, users can also retrieve expression profiles of multiple genes in batch, with multiple filtering options (e.g. by tissue, BioProject, study). The outputs can be exported as plain text files. Heatmaps of gene selections can also be constructed in our website. We strongly believe that this website will optimize data reuse and help research groups in their own projects. This service can be freely accessed at <http://venanciogroup.uenf.br/resources/>. In addition, to allow long-term storage in a separate repository, reproducibility and data reuse, our analysis pipeline and TPM estimates calculated with StringTie and Kallisto are also available for download in Figshare (DOI: 10.6084/m9.figshare.12043188).

A



B



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

2.4 CONCLUSIONS

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

2.5 METHODS

2.5.1 Soybean genome and annotation data

Soybean genomic sequences and gene annotation data (assembly version: Gmax_275_Wm82.a2.v1) were obtained from Phytozome (Schmutz et al., 2010, Goodstein et al., 2012). The gene annotation file contained 56,044 and 88,647 genes and transcripts, respectively. The gene annotation file containing exon-intron boundaries (GFF3 format) was used as a reference guide in read mapping. We excluded 759 overlapping genes from the analysis. The gene description file was used to obtain various annotations such as GO, KEGG, KOG, and Arabidopsis ortholog descriptions.

2.5.2 Soybean RNA-Seq data

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

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

2.5.3 Preprocessing and quality control

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

2.5.4 Transcript assembly and gene expression estimation

We aligned the reads to the *G. max* reference genome (Gmax_275_Wm82.a2.v1) by using STAR (v.2.5.3a) (Dobin et al., 2013) with default parameters, along with the soybean gene annotation file containing exon-intron boundaries (in GFF3). When required, STAR also splits reads to find novel exon-intron boundaries or splice sites. The log files were processed to obtain read mapping statistics. Color-space reads generated with the SOLiD platform were mapped with the gmapper-cs tool from SHRiMP (v.2.2.3), using the parameters –local –strata (Rumble et al., 2009). Next, StringTie (v. 1.3.4) (Pertea et al., 2015) was used to assemble transcripts and estimate normalized gene expression. We performed transcriptome assemblies for each of the 16 tissues separately. In

StringTie, we set the following parameters: i) at least 5 reads with at least 25% of the total read length covering both sides of an exon junction boundary (`-j 5 -a 0.25*read_length`); ii) average read depth for a transcript of at least 10 (`-c 10`) and; iii) library strandedness, when applicable. The resulting 16 assembled transcript annotations from each tissue were combined with TACO v0.7.3 (Niknafs et al., 2017). GffCompare (v0.10.5) (<https://ccb.jhu.edu/software/stringtie/gffcompare.shtml>) was used to compare assembled and reference transcripts. Further, featureCounts (subread-v1.6.2) (Liao et al., 2014) was used to count the number of reads per feature at transcript and gene levels, while normalized expression was estimated in TPM using StringTie (`-e` option). In addition, we have also computed TPM values for all samples (except for 11 SOLiD samples) using Kallisto (v0.46.1) (Bray et al., 2016). The transcript-level TPM estimates generated by Kallisto were converted to gene-level estimates with tximport (Soneson et al., 2016). Spearman's rank correlation coefficient between TPM values from StringTie and Kallisto for each sample was calculated using `cor()` in R.

2.5.5 Sample clustering

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

2.5.6 Identification of novel genes and splicing isoforms

To identify novel genes and isoforms, we analyzed the GffCompare output files. Transcripts not overlapping with any known reference transcript were assigned to class-U. The nucleotide sequences of the class U transcripts were extracted and translated using TransDecoder (v. 3.0.1). Protein domains were predicted using HMMER3 v. 3.1b2 (all default parameters except domain e-value < 0.01) (hmmer.org) and the Pfam database (release 32.0) (El-Gebali et al., 2019). Read coverage of these novel genes were visualized with Gbrowse, available on Soybase (<https://soybase.org/gb2/gbrowse/gmax2.0>). Class-J transcripts were classified as putative novel isoforms. Splice junctions of these transcripts in GTF format were compared against all known splice junctions using ASprofile v.b-1.0.4 (Florea et al., 2013). The number of reads supporting a splice junction was visualized as sashimi plots using Integrated Genome Viewer (v2.5.10)(Robinson et al., 2011).

2.5.7 Analysis of the top 1000 highest expressed gene lists

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

2.5.8 Identification of housekeeping genes

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

- i. A gene with TPM < 1 in a given sample was considered as not expressed (these TPM values were set to 0);
- ii. Genes must be expressed in all 1,225 samples. This step resulted in 1,809 genes;

iii. The mean TPM of each gene was calculated by taking the average of the gene expression across all samples;

iv. The Coefficient of Variation (CoV) was computed by taking the standard deviation divided by the mean expression of a gene;

v. The ratio of the maximum to minimum (MFC) was calculated by dividing the largest by the smallest TPM value. A product score (MFC-CoV) was calculated based on the product of CoV and MFC for each gene;

vi. Genes with MFC-CoV scores within the 1st quartile were classified as HK genes.

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

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

where

x_i = expression of the gene in tissue i . n = number of tissues.

2.5.9 Assessment of tissue-specific expression

We used the \log_2 transformed TPM values for this analysis. Each of the 12 tissues was compared against each other (a total of 144 comparisons) to find significantly over-expressed genes using *limma* (Ritchie et al., 2015). We used \log_2 (fold-change) ≥ 2 and adjusted p-value ≤ 0.05 (moderated t-statistic) to identify significantly over-expressed genes. If a gene G is over-expressed in a tissue T in comparison to the other 11 tissues, G was considered as specifically expressed in T . We also used τ to assess tissue-specific expression by applying a minimum

threshold of 0.8, as previously recommended (Kryuchkova-Mostacci and Robinson-Rechavi, 2017).

2.5.10 Gene orthologs and enrichment tests

We obtained the gene descriptions from Phytomine (<https://phytozome.jgi.doe.gov/phytomine/begin.do>), which is an InterMine (Lyne *et al.*, 2015) interface to genomic data from Phytozome (Goodstein *et al.*, 2012). We used Phytomine to assess the conservation of HK genes in 14 different species (*Ph. vulgaris*, *Me. truncatula*, *Vigna unguiculata*, *Ar. thaliana*, *Oryza sativa*, *Gossypium raimondii*, *Carica papaya*, *Vitis vinifera*, *Sorghum bicolor*, *Zea mays*, *Amborella trichopoda*, *Selaginella moellendorffii*, *Physcomitrella patens*, and *Volvox carteri*). To estimate the conservation of non-HK genes, we created 5 sets of 452 randomly selected genes from the 55,592 non-HK genes. Each of these sets were searched for orthologs in the above mentioned 14 species. GO enrichment was performed on Phytomine (corrected p-value < 0.05). We performed Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment using KOBAS 3.0 (Ai and Kong, 2018). We used the Fisher's Exact test to assess the enrichment of essential genes and TFs in particular gene sets. The list of 510 *Arabidopsis* EMBRYO-DEFECTIVE (EMB) genes (Meinke, 2019) were searched on Phytomine and the corresponding 1,010 soybean orthologs were retrieved. The list of soybean TFs was obtained from a recently published work (Moharana and Venancio, 2020).

2.5.11 Web server

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

2.5.12 Data statement

All soybean RNA-Seq datasets used here are deposited in public repositories, as indicated in the Supplementary Table 2. The integrated atlas and visual results are available in the website indicated above. The analysis scripts and final expression values are available on our Figshare repository (DOI: 10.6084/m9.figshare.12043188).

2.6 ACKNOWLEDGEMENTS

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

2.7 REFERENCES

- Ai, C. and Kong, L. (2018) CGPS: A machine learning-based approach integrating multiple gene set analysis tools for better prioritization of biologically relevant pathways. *Journal of Genetics and Genomics*, 45, 489-504.
- Becker, A. (2003) The major clades of MADS-box genes and their role in the development and evolution of flowering plants. *Molecular Phylogenetics and Evolution*, 29, 464-489.
- Belamkar, V., Weeks, N.T., Bharti, A.K., Farmer, A.D., Graham, M.A. and Cannon, S.B. (2014) Comprehensive characterization and RNA-Seq profiling of the HD-Zip transcription factor family in soybean (*Glycine max*) during dehydration and salt stress. *BMC Genomics*, 15, 950.
- Belliény-Rabelo, D., De Oliveira, E.A., Ribeiro, E.S., Costa, E.P., Oliveira, A.E. and Venancio, T.M. (2016) Transcriptome analysis uncovers key regulatory and

metabolic aspects of soybean embryonic axes during germination. *Sci Rep*, 6, 36009.

- Benedito, V.A., Torres-Jerez, I., Murray, J.D., Andriankaja, A., Allen, S., Kakar, K., Wandrey, M., Verdier, J., Zuber, H., Ott, T., Moreau, S., Niebel, A., Frickey, T., Weiller, G., He, J., Dai, X., Zhao, P.X., Tang, Y. and Udvardi, M.K. (2008) A gene expression atlas of the model legume *Medicago truncatula*. *The Plant Journal*, 55, 504-513.
- Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30, 2114-2120.
- Bowman, J.L., Smyth, D.R. and Meyerowitz, E.M. (2012) The ABC model of flower development: then and now. *Development*, 139, 4095-4098.
- Bray, N.L., Pimentel, H., Melsted, P. and Pachter, L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*, 34, 525-527.
- Chi, Y., Wang, T., Xu, G., Yang, H., Zeng, X., Shen, Y., Yu, D. and Huang, F. (2017) GmAGL1, a MADS-Box Gene from Soybean, Is Involved in Floral Organ Identity and Fruit Dehiscence. *Frontiers in Plant Science*, 8.
- Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szczesniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X. and Mortazavi, A. (2016) A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17.
- Czechowski, T., Stitt, M., Altmann, T., Udvardi, M.K. and Scheible, W.R. (2005) Genome-wide identification and testing of superior reference genes for transcript normalization in *Arabidopsis*. *Plant Physiol*, 139, 5-17.
- Dash, S., Van Hemert, J., Hong, L., Wise, R.P. and Dickerson, J.A. (2012) PLEXdb: gene expression resources for plants and plant pathogens. *Nucleic Acids Res*, 40, D1194-1201.
- Day, R.C., Herridge, R.P., Ambrose, B.A. and Macknight, R.C. (2008) Transcriptome Analysis of Proliferating *Arabidopsis* Endosperm Reveals Biological Implications for the Control of Syncytial Division, Cytokinin

- Signaling, and Gene Expression Regulation. *Plant Physiology*, 148, 1964-1984.
- Deshmukh, R., Sonah, H., Patil, G., Chen, W., Prince, S., Mutava, R., Vuong, T., Valliyodan, B. and Nguyen, H.T. (2014) Integrating omic approaches for abiotic stress tolerance in soybean. *Frontiers in Plant Science*, 5.
- Dey, K.K., Hsiao, C.J. and Stephens, M. (2017) Visualizing the structure of RNA-seq expression data using grade of membership models. *PLoS Genet*, 13, e1006599.
- Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29, 15-21.
- El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A., Sonnhammer, E.L L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S.C E. and Finn, R.D. (2019) The Pfam protein families database in 2019. *Nucleic Acids Res*, 47, D427-D432.
- Figuroa, P. (2005) Arabidopsis Has Two Redundant Cullin3 Proteins That Are Essential for Embryo Development and That Interact with RBX1 and BTB Proteins to Form Multisubunit E3 Ubiquitin Ligase Complexes in Vivo. *The Plant Cell Online*, 17, 1180-1195.
- Florea, L., Song, L. and Salzberg, S.L. (2013) Thousands of exon skipping events differentiate among splicing patterns in sixteen human tissues. *F1000Research*, 2, 188.
- Fucile, G., Di Biase, D., Nahal, H., La, G., Khodabandeh, S., Chen, Y., Easley, K., Christendat, D., Kelley, L. and Provart, N.J. (2011) ePlant and the 3D data display initiative: integrative systems biology on the world wide web. *PLoS One*, 6, e15237.
- Garg, R. and Jain, M. (2013) Transcriptome Analyses in Legumes: A Resource for Functional Genomics. *The Plant Genome*, 6, 0.

- Gazara, R.K., de Oliveira, E.A.G., Rodrigues, B.C., Nunes da Fonseca, R., Oliveira, A.E.A. and Venancio, T.M. (2019) Transcriptional landscape of soybean (*Glycine max*) embryonic axes during germination in the presence of paclobutrazol, a gibberellin biosynthesis inhibitor. *Sci Rep*, 9, 9601.
- Goettel, W., Xia, E., Upchurch, R., Wang, M.L., Chen, P. and An, Y.Q. (2014) Identification and characterization of transcript polymorphisms in soybean lines varying in oil composition and content. *BMC Genomics*, 15, 299.
- Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N. and Rokhsar, D.S. (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res*, 40, D1178-1186.
- Griesmann, M., Chang, Y., Liu, X., Song, Y., Haberer, G., Crook, M.B., Billault-Penneteau, B., Laressergues, D., Keller, J., Imanishi, L., Roswanjaya, Y.P., Kohlen, W., Pujic, P., Battenberg, K., Alloisio, N., Liang, Y., Hilhorst, H., Salgado, M.G., Hocher, V., Gherbi, H., Svistoonoff, S., Doyle, J.J., He, S., Xu, Y., Xu, S., Qu, J., Gao, Q., Fang, X., Fu, Y., Normand, P., Berry, A.M., Wall, L.G., Ane, J.M., Pawlowski, K., Xu, X., Yang, H., Spannagl, M., Mayer, K.F.X., Wong, G.K., Parniske, M., Delaux, P.M. and Cheng, S. (2018) Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis. *Science*, 361.
- He, J., Bedito, V.A., Wang, M., Murray, J.D., Zhao, P.X., Tang, Y. and Udvardi, M.K. (2009) The *Medicago truncatula* gene expression atlas web server. *BMC Bioinformatics*, 10, 441.
- Hoang, V.L.T., Tom, L.N., Quek, X.C., Tan, J.M., Payne, E.J., Lin, L.L., Sinnya, S., Raphael, A.P., Lambie, D., Frazer, I.H., Dinger, M.E., Soyer, H.P. and Prow, T.W. (2017) RNA-seq reveals more consistent reference genes for gene expression studies in human non-melanoma skin cancers. *PeerJ*, 5, e3631.
- Hu, R., Fan, C., Li, H., Zhang, Q. and Fu, Y.F. (2009) Evaluation of putative reference genes for gene expression normalization in soybean by quantitative real-time RT-PCR. *BMC Mol Biol*, 10, 93.

- lizumi, T., Luo, J.-J., Challinor, A.J., Sakurai, G., Yokozawa, M., Sakuma, H., Brown, M.E. and Yamagata, T. (2014) Impacts of El Niño Southern Oscillation on the global yields of major crops. *Nat Commun*, 5.
- Jones, S.I. and Vodkin, L.O. (2013) Using RNA-Seq to profile soybean seed development from fertilization to maturity. *PLoS One*, 8, e59270.
- Jordan, I.K., Reeb, P.D., Bramardi, S.J. and Steibel, J.P. (2015) Assessing Dissimilarity Measures for Sample- Based Hierarchical Clustering of RNA Sequencing Data Using Plasmode Datasets. *PLoS One*, 10, e0132310.
- Jung, C.H., Wong, C.E., Singh, M.B. and Bhalla, P.L. (2012) Comparative genomic analysis of soybean flowering genes. *PLoS One*, 7, e38250.
- Kim, E., Hwang, S. and Lee, I. (2017) SoyNet: a database of co-functional networks for soybean *Glycine max*. *Nucleic Acids Res*, 45, D1082-D1089.
- Konishi, M. and Yanagisawa, S. (2013) Arabidopsis NIN-like transcription factors have a central role in nitrate signalling. *Nat Commun*, 4, 1617.
- Krijthe, J.H. (2015) Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation.
- Kryuchkova-Mostacci, N. and Robinson-Rechavi, M. (2017) A benchmark of gene expression tissue-specificity metrics. *Brief Bioinform*, 18, 205-214.
- Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 9, 357-359.
- Leinonen, R., Akhtar, R., Birney, E., Bonfield, J., Bower, L., Corbett, M., Cheng, Y., Demiralp, F., Faruque, N., Goodgame, N., Gibson, R., Hoad, G., Hunter, C., Jang, M., Leonard, S., Lin, Q., Lopez, R., Maguire, M., McWilliam, H., Plaister, S., Radhakrishnan, R., Sobhany, S., Slater, G., Ten Hoopen, P., Valentin, F., Vaughan, R., Zalunin, V., Zerbino, D. and Cochrane, G. (2010) Improvements to services at the European Nucleotide Archive. *Nucleic Acids Res*, 38, D39-45.
- Li, W.V. and Li, J.J. (2018) Modeling and analysis of RNA-seq data: a review from a statistical perspective. *Quantitative Biology*, 6, 195-209.

- Liao, Y., Smyth, G.K. and Shi, W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30, 923-930.
- Libault, M., Farmer, A., Joshi, T., Takahashi, K., Langley, R.J., Franklin, L.D., He, J., Xu, D., May, G. and Stacey, G. (2010) An integrated transcriptome atlas of the crop model Glycine max, and its use in comparative analyses in plants. *Plant J*, 63, 86-99.
- Liu, P. and Si, Y. (2014) Cluster Analysis of RNA-Sequencing Data. 191-217.
- Lyne, R., Sullivan, J., Butano, D., Contrino, S., Heimbach, J., Hu, F., Kalderimis, A., Lyne, M., N. Smith, R., Štěpán, R., Balakrishnan, R., Binkley, G., Harris, T., Karra, K., A. T. Moxon, S., Motenko, H., Neuhauser, S., Ruzicka, L., Cherry, M., Richardson, J., Stein, L., Westerfield, M., Worthey, E. and Micklem, G. (2015) Cross-organism analysis using InterMine. *genesis*, 53, 547-560.
- Marini, F. and Binder, H. (2019) pcaExplorer: an R/Bioconductor package for interacting with RNA-seq principal components. *BMC Bioinformatics*, 20.
- Meinke, D.W. (2019) Genome-wide identification of EMBRYO-DEFECTIVE (EMB) genes required for growth and development in Arabidopsis. *New Phytol*.
- Moharana, K.C. and Venancio, T.M. (2020) Polyploidization events shaped the transcription factor repertoires in legumes (Fabaceae). *Plant J*.
- Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5, 621-628.
- Niknafs, Y.S., Pandian, B., Iyer, H.K., Chinnaiyan, A.M. and Iyer, M.K. (2017) TACO produces robust multisample transcriptome assemblies from RNA-seq. *Nat Methods*, 14, 68-70.
- O'Rourke, J.A., Graham, M.A. and Whitham, S.A. (2017) Soybean Functional Genomics: Bridging the Genotype-to-Phenotype Gap. 151-170.

- Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T. and Salzberg, S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*, 33, 290- 295.
- Po-Yen, W., Phan, J.H., Fengfeng, Z. and Wang, M.D. (2011) Evaluation of normalization methods for RNA- Seq gene expression estimation. 50-57.
- Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. and Smyth, G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*, 43, e47-e47.
- Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G. and Mesirov, J.P. (2011) Integrative genomics viewer. *Nat Biotechnol*, 29, 24-26.
- Robinson, M.D. and Oshlack, A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11, R25.
- Rocchi, V., Janni, M., Bellincampi, D., Giardina, T. and D'Ovidio, R. (2012) Intron retention regulates the expression of pectin methyl esterase inhibitor (Pmei) genes during wheat growth and development. *Plant Biol (Stuttg)*, 14, 365-373.
- Roy, S., Liu, W., Nandety, R.S., Crook, A.D., Mysore, K.S., Pislariu, C.I., Frugoli, J.A., Dickstein, R. and Udvardi, M.K. (2019) Celebrating 20 years of genetic discoveries in legume nodulation and symbiotic nitrogen fixation. *Plant Cell*, tpc.00279.02019.
- Rumble, S.M., Lacroute, P., Dalca, A.V., Fiume, M., Sidow, A. and Brudno, M. (2009) SHRiMP: accurate mapping of short color-space reads. *PLoS Comput Biol*, 5, e1000386.
- Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J.,
- Cheng, J., Xu, D., Hellsten, U., May, G.D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M.K., Sandhu, D., Valliyodan, B., Lindquist, E., Peto, M., Grant, D., Shu, S., Goodstein, D., Barry, K., Futrell- Griggs, M., Abernathy,

- B., Du, J., Tian, Z., Zhu, L., Gill, N., Joshi, T., Libault, M., Sethuraman, A., Zhang, X.C., Shinozaki, K., Nguyen, H.T., Wing, R.A., Cregan, P., Specht, J., Grimwood, J., Rokhsar, D., Stacey, G., Shoemaker, R.C. and Jackson, S.A. (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, 463, 178-183.
- Schwacke, R., Ponce-Soto, G.Y., Krause, K., Bolger, A.M., Arsova, B., Hallab, A., Gruden, K., Stitt, M., Bolger, M.E. and Usadel, B. (2019) MapMan4: A Refined Protein Classification and Annotation Framework Applicable to Multi-Omics Data Analysis. *Molecular Plant*, 12, 879-892.
- Severin, A.J., Woody, J.L., Bolon, Y.T., Joseph, B., Diers, B.W., Farmer, A.D., Muehlbauer, G.J., Nelson, R.T.,
- Grant, D., Specht, J.E., Graham, M.A., Cannon, S.B., May, G.D., Vance, C.P. and Shoemaker, R.C. (2010) RNA-Seq Atlas of Glycine max: a guide to the soybean transcriptome. *BMC Plant Biol*, 10, 160.
- Sinharoy, S., Torres-Jerez, I., Bandyopadhyay, K., Kereszt, A., Pislariu, C.I., Nakashima, J., Benedito, V.A., Kondorosi, E. and Udvardi, M.K. (2013) The C2H2 transcription factor regulator of symbiosome differentiation represses transcription of the secretory pathway gene VAMP721a and promotes symbiosome development in *Medicago truncatula*. *Plant Cell*, 25, 3584-3601.
- Smalle, J. and Vierstra, R.D. (2004) The Ubiquitin 26s Proteasome Proteolytic Pathway. *Annual Review of Plant Biology*, 55, 555-590.
- Soltis, D.E., Chanderbali, A.S., Kim, S., Buzgo, M. and Soltis, P.S. (2007) The ABC Model and its Applicability to Basal Angiosperms. *Annals of Botany*, 100, 155-163.
- Soneson, C., Love, M.I. and Robinson, M.D. (2016) Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, 4, 1521.
- Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., Selbig, J., Müller, L.A., Rhee, S.Y. and Stitt,

- M. (2004) mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *The Plant Journal*, 37, 914-939.
- Vernié, T., Moreau, S., de Billy, F., Plet, J., Combier, J.-P., Rogers, C., Oldroyd, G., Frugier, F., Niebel, A. and Gamas, P. (2008) EFD Is an ERF Transcription Factor Involved in the Control of Nodule Number and Differentiation in *Medicago truncatula*. *Plant Cell*, 20, 2696-2713.
- Wagner, G.P., Kin, K. and Lynch, V.J. (2012) Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in Biosciences*, 131, 281-285.
- Wang, L., Wang, S. and Li, W. (2012) RSeQC: quality control of RNA-seq experiments. *Bioinformatics*, 28, 2184-2185.
- Weller, J.L. and Ortega, R.I. (2015) Genetic control of flowering time in legumes. *Frontiers in Plant Science*, 6. Wormit, A. and Usadel, B. (2018) The Multifaceted Role of Pectin Methyltransferase Inhibitors (PMEIs). *International Journal of Molecular Sciences*, 19, 2878.
- Wu, Z., Wang, M., Yang, S., Chen, S., Chen, X., Liu, C., Wang, S., Wang, H., Zhang, B., Liu, H., Qin, R. and Wang, X. (2019) A global coexpression network of soybean genes gives insights into the evolution of nodulation in nonlegumes and legumes. *New Phytologist*.
- Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-Saban, S., Safran, M., Domany, E., Lancet, D. and Shmueli, O. (2004) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, 21, 650- 659.
- Yim, A.K., Wong, J.W., Ku, Y.S., Qin, H., Chan, T.F. and Lam, H.M. (2015) Using RNA-Seq Data to Evaluate Reference Genes Suitable for Gene Expression Studies in Soybean. *PLoS One*, 10, e0136343.

- Zhao, S., Zhang, Y., Gordon, W., Quan, J., Xi, H., Du, S., von Schack, D. and Zhang, B. (2015) Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. *BMC Genomics*, 16.
- Zhao, T., Holmer, R., de Bruijn, S., Angenent, G.C., van den Burg, H.A. and Schranz, M.E. (2017) Phylogenomic Synteny Network Analysis of MADS-Box Transcription Factor Genes Reveals Lineage- Specific Transpositions, Ancient Tandem Duplications, and Deep Positional Conservation. *Plant Cell*, 29, 1278-1292.