Comparative transcriptomics of flower development in barley and wheat

Marosi, Vanda Beáta

Munich 2020

Comparative transcriptomics of flower development in barley and wheat

Marosi, Vanda Beáta

Master's Thesis

for the Faculty of Biology

at Ludwig-Maximilians-University

Munich

Written by

Marosi, Vanda Beáta

from Budapest, Hungary

Munich, 23rd of November 2020

External supervisor: Prof. Dr. Klaus F. X. Mayer

External, direct supervisor: Dr. Daniel Lang

Helmholtz Zentrum Munich

German Research Center for Environmental Health (GmbH)

Plant Genome and Systems Biology

Ingolstädter Landstraße 1, 85764 Neuherberg

Internal supervisor: Prof. Dr. Korbinian Schneeberger

Ludwig-Maximilians-University Munich

Department of Genetics

Großhaderner Str. 2-4, 82152 Martinsried

Contents

Abstract **vii**

Zusammenfassung **viii**

Introduction **1**

1.1 Comparative transcriptomics in the *Triticeae* 1

1.2 Flower anatomy of the *Poaceae* family 3

1.3 Objectives 3

Materials & Methods **5**

2.1 Data availability 5

2.2 Text mining 6

2.3 Selection of reference genes 8

2.3.1 Mapping to reference genomes and ortholog detection 9

2.4 Collection of openly available RNA-sequencing samples 9

2.4.1 Metadata compilation 10

2.4.2 Preprocessing and mapping to reference genomes 11

2.4.3 Exploratory analysis and data normalization 12

2.4.4 Differential Gene Expression Analyses 13

Results **14**

3.1 Text mining for flower development related publications 15

3.2 Identification and projection of reference gene sets 17

3.3 Comparative transcriptomics among barley and wheat 17

3.3.1 Attributes of RNA-sequencing sample collection 18

3.3.2 Preprocessing, mapping and exploratory analysis 21

3.3.3 Comparative Differential Gene Expression Analysis 25
3.4 Assessment of reference gene set with DEG-sets 28

Discussion **31**
4.1 Comparative meta-analysis of flower development transcriptomics in barley and wheat 31
4.1.1 Transcriptome profiling of inflorescence development 32
4.1.2 Insights into anther development and male sterility 33
4.2 Reproducibility and data mining as standards for life sciences 34
4.2.1 Evaluation of the reference gene set 35
4.3 Perspectives on a large-scale comparative study in *Triticeae* 36

GitHub Repository **38**

Acknowledgments **41**

Abbreviations **42**

Bibliography **45**

Supplementary Materials **52**

Abstract

Bread wheat (*Triticum aestivum*) and barley (*Hordeum vulgare*) are major cereal crops, with close evolutionary relationship as a result of a common ancestor 13 million years ago. Having had their genomes recently completed and continuously perfected, only a few studies focuses on comparative transcriptome analyses in the *Triticeae* tribe, most of them applying outdated approaches. Therefore, it is important to establish a semi-automated and reproducible method to incorporate earlier findings into the quickly evolving field of transcriptomics.

The inflorescence of grasses is a highly specialized organ of reproduction, whose development is responsible for production of food grains and most of its regulatory pathways still remain unclear.

In this research, we investigated each scientific article concerning *Triticeae*, and determined a set of 898 publications inspecting flower development, corresponding to 118 gene subfamilies shared between wheat and barley. Then we repeated the analysis of 455 publicly available RNA-seq samples of barley and wheat to perform combined differential gene expression of several datasets targeting different aspects of flower development. Transcript changes in four developmental or experimental factor contrasts were used to identify significantly different up-, or down-regulated gene-sets, and their corresponding gene subfamilies across barley and wheat subgenomes. By intersecting these differentially expressing genes (DEGs) with results of the text mining, we found 87 shared gene subfamilies, which represents the reliability of text-mining based gene identification and reproducibility of published studies.

Our comparative meta-analysis of transcriptomics provides insights for the complexity of investigating inflorescence development in barley and wheat, and how such an integrative approach can be a precedent for future larger scale analyses.

Zusammenfassung

Brotweizen (*Triticum aestivum*) und Gerste (Hordeum vulgare) sind wichtige Getreidepflanzen, die aufgrund eines gemeinsamen evolutionären Ursprungs vor etwa 13 Millionen Jahren in einer engen Beziehung zueinander stehen. Da ihre Genome erst vor Kurzem vollständig sequenziert und seither in ständiger Bearbeitung sind, gibt es nur wenige Studien, die sich auf vergleichende Transkriptomanalysen des Stammes der Triticeae fokussieren und die zudem oft veraltete Ansätze anwenden. Daher ist es wichtig, eine halbautomatisierte und reproduzierbare Methode zu etablieren, um vorherige Erkenntnisse in das sich schnell entwickelnde Gebiet der Transkriptomik einzubeziehen.

Der Blütenstand von Gräsern ist ein hochspezialisiertes Fortpflanzungsorgan, dessen Entwicklung für die Produktion von Nahrungsmittelkörnern verantwortlich ist und dessen Regulationsmechanismen noch unklar sind.

In dieser Studie haben wir wissenschaftliche Artikel zur Familie der Triticeae untersucht und insgesamt 898 Veröffentlichungen identifiziert, die sich mit der Blütenentwicklung von insgesamt 118 Genunterfamilien von Weizen und Gerste auseinandersetzen. Wir haben die Analyse mit 455 öffentlich verfügbaren RNA-seq-Proben von Gerste und Weizen wiederholt, mit dem Ziel differenzierte Genexpression in mehreren Datensätzen über verschiedene Aspekte der Blütenentwicklung zu ermitteln. Transkriptomunterschiede, die sich in vier Entwicklungs- oder experimentellen Faktorkontrasten widerspiegeln, waren die Grundlage, um signifikant unterschiedliche hoch- oder runterregulierte Gensätze (DEGs) und ihre entsprechenden Genunterfamilien innerhalb von Gersten- und Weizensubgenome zu identifizieren. Durch Vergleiche der DEGs mit den Ergebnissen des Text Mining konnten wir 87 gemeinsame Genunterfamilien feststellen, was die Zuverlässigkeit der Text Mining-basierten Genidentifikation und Reproduzierbarkeit veröffentlichter Studien wider-

spiegelt.

Unsere vergleichende Metaanalyse der Transkriptomik liefert Einblicke in die Komplexität der Forschung der Blütenstandsentwicklung von Gerste und Weizen und zeigt wie ein solcher integrativer Ansatz ein Modell für zukünftige Analysen in größerem Maßstab darstellen kann.

1. Introduction

1.1 Comparative transcriptomics in the *Triticeae*

In the *Pooideae* subfamily of grasses, the *Triticeae* tribe includes two worldwide cultivated staple crops, wheat and barley, both of them originate from the Fertile Crescent (Mochida and Shinozaki 2013). The immense economical and anthropological relevance has given them considerable attention in research, which led to the barley and wheat complete genomes sequenced in 2011 and 2014, respectively (Mayer et al. 2011, Consortium et al. 2014).

The genome of domesticated bread wheat (*Triticum aestivum*) is hexaploid, comprising 21 pairs of chromosomes, which equals to three homeologous sets of seven chromosomes for each of the A, B and D subgenomes (AABBDD). Modern cultivated wheat has originated from two hybridization events, first between cultivated tetraploid emmer wheat (AABB, *Triticum dicoccoides*) and then with diploid goat grass (DD, *Aegilops tauschii*) approximately 10,000 years ago (Dubcovsky and Dvorak 2007, Pont et al. 2019). Its large and highly repetitive genome with more than 85% transposable element content, caused difficulties in assembly and consequently delayed advancing in its genomics research. Since its first genome version has been published, continuous improvement led to the latest assembly of the Chinese Spring cultivar in 2018, which comprises of 14.5-gigabase genetic information and 107,891 assigned high-confidence protein-coding loci (Appels et al. 2018).

Barley (*Hordeum vulgare*) has a diploid genome (HH) with 14 pairs of chromosomes and was domesticated from its wild barley relative (*Hordeum vulgare subsp. spontaneum*) again about 10,000 years ago (Badr et al. 2000). Its genome comprises of 5.1-gigabase genetic material in the latest reference genome assembly of the Morex cultivar with 32,787 high-confidence gene models (Monat et al. 2019).

The close evolutionary relationship between polyploid wheat and diploid barley with their divergence 13 million years ago (Gaut 2002), is reflected in their similar morphological and developmental patterns, and it promises particularly interesting comparison among their transcriptomes. Such study could bring insights into the consequences of speciation and polyploidization or explain the functional diversification of homoeologous genes. While the overall architecture of gene regulatory pathways often can remain conserved across close relatives, in some cases the orthologous genes do not undertake equivalent functions, which limits their direct translation from one crop to another (Borrill et al. 2015).

Sonnhammer and Koonin explains clearly the subtle differences among homology related expressions (Sonnhammer and Koonin 2002). According to them, the original description of orthologs defines two genes from two different species that are derived from a single gene in the last common ancestor of the species. Similarly, paralogs are defined as genes that are derived from a single gene that was duplicated within a genome. However, paralogs may arise from a duplication that occurred either before or after speciation, if the duplication occurred first, the genes resulting from the duplication cannot be orthologs, but called outparalogs. Moreover, if the duplication happened after the speciation, the resulting genes can be considered co-orthologs, and so such genes are called inparalogs. Based on the previous assumptions, in our study, a subfamily is part of a gene family where all members are either co-orthologs or inparalogs, which means that they correspond to the same ancestral gene in the last common ancestor of the taxon set of interest.

While there has been individual focus to identify tissue-specific gene expression and coexpression networks in both wheat (Uauy 2017, Ramírez-González et al. 2018), and barley (Druka et al. 2006, Saisho and Takeda 2011), there seems to be an alarming gap in research of their combined investigation. To date, only a few studies aimed to decipher a novel question or contrasted barley and wheat gene expression discoveries with comparative transcriptomics (Schreiber et al. 2009, Mochida and Shinozaki 2013).

Besides developing more resistant traits to withstand the altering environment of climate change, the demands of growing human population risking global food security and sustainable agriculture are all calling for urgent scientific advancements

in crop breeding.

1.2 Flower anatomy of the *Poaceae* family

The inflorescence of grasses is a highly specialized organ of reproduction, whose development is responsible for production of food grains and most of its regulatory pathways still remain unclear. Inflorescences are composed of stems, stalks, bracts, and flowers, as it is described in cereal development guides (Fettell et al. 2010, Larsen and Smith 2010), as follows in the next paragraphs. *Poaceae* is one of the largest families within the monocotyledonous flowering plants, and inflorescences of this family are characterized by their spike shapes, complex branches with unique spikelets, as well as flowers without obvious petals and sepals. The wheat or barley inflorescence is called the spike, which consists of many spikelets, every one of them contain one or more florets, sitting on rachis nodes. A mature floret owns a single pistil with its ovary, style, and stigma, three anthers attached to the base through the filaments, two lodicules, and it is enclosed by bract-like organs called lemma and palea, surrounded by a glume ending in its thin awn (Figure 5E).

Floral organs grow from a specialized structure called the shoot apical meristem, from which the young inflorescence develops ridges composed of bract primordia, followed by the development of spikelet meristems as axillary buds, commonly referred as the double ridge stage. Each spikelet meristem initiates a glume primordium followed by a series of floral meristems, which have the potential to differentiate into three stamen primordia and one pistil primordium. The subsequent stage of reproductive growth before flowering, or anthesis, involves the differentiation of the floral organs and pollen cell meiosis and develops together with the elongation process of the spike from less than 1 cm to its final length.

1.3 Objectives

The aim of this project was to determine a set of reference genes, reportedly being involved in inflorescence development. With this end in view, we have investigated all *Triticeae* tribe related peer-reviewed scientific articles for text mining.

In the second phase of the project, a collection of published RNA-sequencing stud-

ies of various barley and wheat tissues were re-analysed, following the necessary preprocessing. From the transcriptomics analysis, flowering related differentially expressed gene sets were determined and compared with the text mining based reference genes.

By addressing gene expression differences in floral tissues of two major crops we aimed for an improved understanding of the evolution of the inflorescence. Furthermore, the repeated assessment of published studies allowed us to investigate the reproducibility of peer-reviewed articles.

2. Materials & Methods

We have used a text-mining approach to perform a meta-study allowing the comparative analysis of barley and wheat. The general workflow is represented in Figure 1. Primarily, we focused on the creation of a reference gene set, with a proposed role in flower development, for later application in investigation of DEGs.

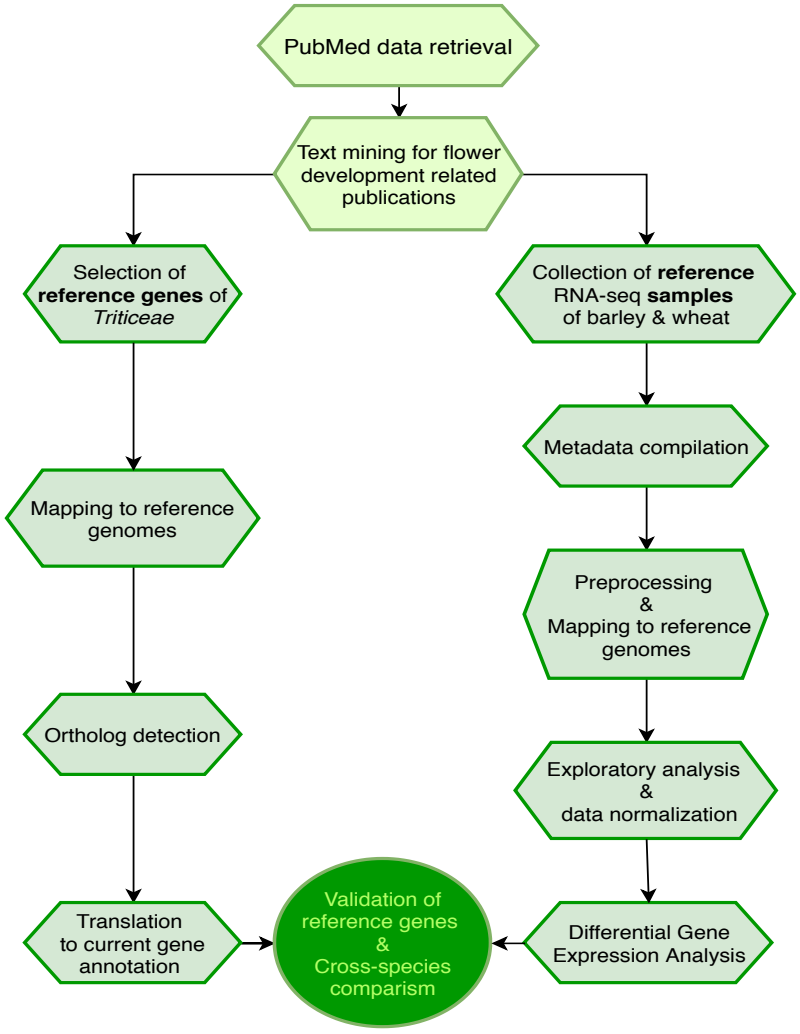

Figure 1: Flow-chart of general data processing workflow. Steps in creation and assessment of reference gene set and RNA-seq sample collections.

2.1 Data availability

All data analysing scripts as well as information about the downloaded datasets are openly available in a GitHub Repository which will be cited for the corresponding scripts in the next sections (the link is also listed in the Bibliography as Marosi 2020). The digital version of this thesis including the LaTeX-source code is included in the repository as well. The Anaconda open-source package and environment management system was used for all of the analysis (*Anaconda Software Distribution* 2016). The main programming language for performing statistical analysis and data wrangling was *R 3.6 & 4.0* (R Core Team 2019), and across all R-scripts the "meta" package of *tidyverse 1.2.1* was the primary data organizing and visualization tool (Wickham et al. 2019). For reproducible analysis, R-scripts were uniformly recorded using Jupyter notebooks (Kluyver et al. 2016).

2.2 Text mining

The corpus for the text mining was obtained from the NCBI (National Center for Biotechnology Information) PubMed database (Sayers et al. 2019). The search query for retrieving publications included the *Triritceae* tribe and its roughly 300 child taxa listed in the NCBI Taxonomy Browser until genus level (NCBI 2020*c*), furthermore comprised the common names of *Triticum aestivum* and *Hordeum vulgare* such as bread wheat or barley. The following Boolean word combination is an example for the inquired search: `"(common + wheat) +OR+ (Triticum + aestivum)"`. The purpose of the query was to obtain in XML format all ever published scientific articles related to the *Triticeae* tribe available through the PubMed database, using the *easyPubMed 2.11* package (Fantini 2018). The accessible data for text mining consisted of 80,826 articles that were published between 1965 and the 2nd of April, 2020 and was retrieved with title, abstracts and further metadata. The articles in PubMed are well-structured objects, meaning their search-fields contain predefined features, from which the following were used in this analysis: PMID (PubMed ID), Title, Abstract, Year, Journal. The two items with the richest, albeit unstructured information, were the Abstract and Title, and were used together for text mining.

Other fields (PMID, Year, Journal) had predefined features and were included in the analysis as metadata. The aim of the text mining was to establish a list of articles involved in flower-development research and as part of the second phase of my project, to determine a subset of publications involving transcriptomics research, with the use of the *tm 0.7-6* package (Feinerer and Hornik 2019).

After converting the downloaded abstracts and titles into a single corpus, text preprocessing led to clean and compressed data. This involved converting all characters to lower case, removing punctuation marks, numbers, and excess white-space. Furthermore, in the filtering process, a list of 174 most frequent "stopwords" in English were removed, provided by the *tm* package. This collection was expanded with additional common scientific words that are not relevant for this analysis (such as "enhanced" or "increasing"), and the most frequent words are represented in a diagram and a word-cloud in S1.

Before the final stage, the corpus was re-filtered with the original query used to retrieve data from PubMed. This step was necessary due to the original search included all openly available text on PubMed, which in some cases meant the entire article, and thus implicated any minor mention of the *Triticeae* tribe, which could lead to not relevant publications.

Finally, fifteen keywords were selected for filtering of the corpus, based on frequently used flower-development related words of representative studies, as follows: `"anther", "fertility", "floral", "floret", "flower", "inflorescence", "lemma", "palea", "panicle", "pistil", "pollen", "spike", "stamen", "stigma", "style"`. For ensuring the exclusion of the false-positive hits from the selection procedure, only those articles were chosen that contained at least two out of the fifteen words. Results of the keyword-based filtering were visualized using the *UpsetR 1.4.0* package (Conway et al. 2017), as it is represented on S2.

The UpSet technique visualizes quantitative analysis of sets, by displaying set intersections in a matrix layout and can represent collections based on groupings and queries (Lex et al. 2014). The left horizontal bars represent the groups, whose intersections are illustrated on the right half of the plot. Vertical columns display the unique number of elements in each intersection, indicated by dots in the matrix layout.

In the next stage, these resulting 898 articles served as a new corpus to construct a transcriptomics related publication list. Based on frequently used terms in RNA-sequencing studies, six expressions were chosen for another selection procedure: `"differentially expressed", "comparative transcriptom", "rnaseq", "rna seq", "differential gene expression", "transcriptome"`. Repeatedly, each of the 32 publications of the final collection contained at least two of the six transcriptomics-keywords, as it is shown in S3. Thereafter, PMID-s were extracted from both publication lists by accessing corpus metadata and saved for further investigations.

All of the above mentioned analysis are recorded in the *01-Text-mining-download.ipynb* and *02-Text-mining-analysis.ipynb* notebooks with the use of the *r.yml* package environment file as listed in the GitHub Repository.

2.3 Selection of reference genes

By intersecting the "Gene2Pubmed" database, curated by the NCBI Entrez Gene database (Maglott et al. 2011), with the previously created PMID-list, we were able to generate direct connection between flower-development related research articles and their gene subject in question. Additionally, with the *taxize 0.9.94* package of R (Chamberlain et al. 2020), we restricted the taxonomic identifiers to the 389 child taxa of the *Triticeae* tribe, by which we ensured the inclusion of the species only in our interest.

We were able to associate 165 unique gene identifiers (GeneID-s) to our initial list of 898 publications, which in the following steps had to be converted into submitted gene sequences of the GenBank and Entrez Nucleotide databases. Gene sequences were downloaded via the *rentrez 1.2.2* package, an R interface to the NCBI databases (Winter 2017). As each GeneID is assigned to several nucleotide sequences uploaded by individual researchers, the GeneIDs were interchanged to 254 FASTA nucleotide files, which comprised of four types of sequences: "genomic DNA", "unassigned DNA", "transcribed RNA", "mRNA"; and belonged to three organisms, *Triticum aestivum*, *Hordeum vulgare* and *Aegilops tauschii*.

All of the above mentioned analysis are available in the *03-Ref-genes-download.ipynb*

notebook with using the *r.yml* package environment file as listed in the GitHub Repository.

2.3.1 Mapping to reference genomes and ortholog detection

The downloaded FASTA sequences were mapped to their respective reference genome using *gmap 2020.06.01* (Genomic Mapping and Alignment Program) (Wu and Watanabe 2005), based on the latest wheat cultivar Chinese Spring (Ramírez-González et al. 2018), goatgrass (Luo et al. 2017), and barley cultivar Morex reference genomes (Monat et al. 2019). We applied or detached splicing detection on genomic or RNA sequences, respectively, and we treated "unassigned DNA" as RNA sequences, as their labeling was ambiguous.

As outcome of the mapping, 242 genes had to be projected to the respective reference genome annotations, which operations together with the next steps of summarizing the results of ortholog detection, were executed by Dr. Daniel Lang.

Ortholog detection on the selected genes among species of the *Triticeae* tribe was performed by Nico van Gessel using *OrthoFinder* (Emms and Kelly 2015). Orthofinder orthologous gene pairs were used to build an orthology graph from which Daniel Lang extracted subfamilies comprised of (co)orthologs and inparalogs that were used to represent the ancestral gene when comparing barley and wheat.

Bash scripts containing the mapping parameters are available in the GitHub Repository, using the *seqtools.yml* package environment file, and additional details are included in the *03-Ref-genes-download.ipynb* notebook using the *r.yml* package environment file as listed in the GitHub Repository.

2.4 Collection of openly available RNA-sequencing samples

All the samples analysed in this study were selected both from the initial 32 PMIDs obtained via text mining and were extended with further manual search.

For the analysis, 22 publicly available datasets were downloaded from the NCBI SRA database (NCBI 2020*a*), using *SRA Toolkit* (NCBI 2020*b*). Further two datasets were accessed from the GSA (Genome Sequence Archive) sequencing data reposi-

tory of the China Genomic Data Sharing Initiative (Wang and Song 2017), and one remaining dataset was reached from a provided website.
Download of all samples are recorded in the *04-Ref-samples-download-preprocess.ipynb* notebook using the *r.yml* package environment file as listed in the GitHub Repository.

2.4.1 Metadata compilation

Driven by the intention to detect batch-effects as a result of the diverse experimental background via multivariate statistical methods, and normalize according to them, the final metadata tables contain the following variables for both species: "Dataset", "PMID", "Run ID", "BioProject ID", "SRA sample", "BioSample ID", "Sample name", "Batch", "Treatment", "RNA extraction", "RNA enrichment", "RNA input (µg)", "PCR purification", "Growth condition", "Day-night condition", "Temperature day-night", "Growth location", "Seq location", "Instrument", "Library layout"; and biological variables: "Organism", "Cultivar", "GM" (sample is a mutant or wild-type), "Genotype", "Dev stage", "Zadok scale", "Intermediate age", "High level age", "Intermediate tissue", High level tissue", "PO dev general", "PO dev narrow", "PO tissue", "Additional". The distribution of variables and their modalities among the samples and datasets are presented on Figures 3., 4., 5. and S4., S5., S6.
Furthermore, the original number of tissues types were 18, with very few samples for some categories, which made it difficult to see differences among them. Hence we organized the 18 "Intermediate tissue" levels into 7 groups of the "High level tissue" variable, as it is described in Table S2. For an example, the tissues named "ear" and "lemma" were both assigned to the inflorescence category of "High-level tissue".
Similarly, the "Dev stage" category had too diversified developmental stage definitions and needed to be united into a single type of developmental scale. Therefore, we created a developmental stage conversion table which allowed us to compare samples among the various reported developmental scale definitions for cereals such as the Waddington-, Zadok-, Feekes- and the BBCH-scales (Waddington et al. 1983, Zadoks et al. 1974, Large et al. 1954, Lancashire et al. 1991). The originally de-

veloped table by Nadia Kamal was adopted and modified to have Zadok-scale as a consensus for the present study. Furthermore the 99 categories of Zadok-scale needed to be assigned to "Intermediate age" and "High-level age", for a more transparent and easily understandable portrayal of the entire sample collection, as it is shown in Table S1. In this manner, an example barley sample described as "Waddington 0.5 - leaf at coleoptile tip" in the "Dev stage", was translated as 9 in a Zadok-scale, which was converted further into "Germination" as "Intermediate age", and even further as "Seedling" in the "High-level age" category. Finally, we assigned Plant Ontology (PO) terms (Cooper et al. 2018) for each tissue and developmental stage for later analysis.

Within the GitHub Repository, the *Metadata-Zadoc-PO.ods* table contain all dataset and sample information, including developmental stage and tissue conversion tables. Visualization is recorded in the 06-Ref-samples-metadata-visualization.ipynb notebook using the *r.yml* package environment file.

2.4.2 Preprocessing and mapping to reference genomes

All datasets were produced using Illumina platforms, and all raw data were checked for quality using *FastQC 0.11.9*, and summarized with *MultiQC 1.9* before and after trimming (Andrews et al. 2010, Ewels et al. 2016). Adaptors were trimmed using *Trimmomatic 0.39* with default settings to preserve a minimum Phred score of Q20 over 60 base-pairs (Bolger et al. 2014). The changes in the mean quality scores after trimming for paired and single reads in both species are represented in S7.

Transcript abundance was calculated using *kallisto 0.46.2* (Bray et al. 2016), based on the latest wheat cultivar Chinese Spring (Ramírez-González et al. 2018) and barley cultivar Morex (Monat et al. 2019) transcript references, respectively. Estimated read counts were summarised to gene level using *tximport 1.16.0* (Soneson et al. 2015). Additionally, the minimum Pearson's correlation between different samples was 0.005 for barley and 0.0005 for wheat, thereby indicating positive correlation across all samples.

Bash scripts applying arrays were adopted from Maxim Messerer and their modified versions are available in the GitHub Repository using the *seqtools.yml* package environment file and additional details are included in the *04-Ref-samples-download-*

preprocess.ipynb notebook using the *r.yml* package environment file. The import of transcript-level abundance, and summarizes into matrices was recorded in the *05-Ref-samples-mapping.ipynb* notebook using the *tximport.yml* package environment file.

2.4.3 Exploratory analysis and data normalization

We determined a gene-wise expression matrix for further analysis, and transcriptome analysis revealed the expression of 42,775 transcripts for barley and 120,297 transcripts for wheat at levels greater than 0.5 transcripts-per-million (TPM) in at least two libraries. In certain exploratory analyses, the filtered counts were transformed into log2-transcripts-per-million (log2TPM), as `(log2(tpm+1))`.

Principal Component Analysis (PCA) was calculated using functions of *FactoMineR 2.3* (Lê et al. 2008) and *factoextra 1.0.7* (Kassambara and Mundt 2020). For exploring the effects of different statistical transformations, four types of PCAs were inspected: 1, PCA on non-transformed and non-filtered TPM (further referred as raw counts), 2, PCA with centering and scaling on raw counts, 3, PCA on filtered log2TPM, 4, PCA on filtered log2TPM with centering and scaling, as it is shown for the barley dataset in S8.

As another exploratory test, Expression Level Category (ELC) analysis was performed on the filtered gene-wise expression matrix. The method was adopted from Daniel Lang, and based on K-mer clustering towards five distinct groups, followed by log-transformation and ordering of factors into E0-E4 categories, which comprises genes from lowly to highly expressed levels. In addition, PCAs were calculated on ELC-transformed and filtered TPMs.

Uniform Manifold Approximation and Projection (UMAP) was used for visualization of dimension reduction of the various groups of biological and technical variables in the sample collection with the *umap 0.2.6* package (Konopka 2020). Repeatedly, for exploratory purposes, three types of UMAPs were inspected: 1, UMAP on non-transformed and non-filtered counts; 2, UMAP on non-transformed and filtered counts; and 3, UMAP on log2-transformed and filtered counts, as it is shown for both barley and wheat datasets in S9.

For comparison among the methods, PCAs of log2TPM and of ELC, along with

UMAPs are summarized on different technical and biological variables for both datasets (S10., S11., S12., S13.). Finally, ELC analysis was further used for visualizing expression bias via summarizing the number of E4 category genes for each sample and dividing them by the library size in barley (S14.) and in wheat (S15).
All of the above mentioned analysis are available in the *07-Ref-samples-PCA-barley.ipynb*, *07-Ref-samples-PCA-wheat.ipynb* and *08-Ref-samples-ELC-barley.ipynb*, *08-Ref-samples-ELC-wheat.ipynb* notebooks with using the *r.yml* package environment file as listed in the GitHub Repository.

2.4.4 Differential Gene Expression Analyses

Differential gene expression (DGE) analyses between different contrast groups were performed with likelihhood ratio test (LRT) and with Wald-test (WT) using *sleuth 0.30.0* (Pimentel et al. 2017). The resulting p-value from LRT was adjusted using the Benjamini-Hochberg multiple testing correction for controlling the false discovery rate. An adjusted p-value was computed for each gene and those with an adjusted p-value <0.01 were assigned as diferentially expressed genes (DEGs), as listed in Table 2.
Translation of DEGs to subfamilies of wheat and barley species and their intersections with earlier defined reference gene sets based on text mining, are represented with UpsetR, in Figure 6.
DGE analyses are available in the *09-Ref-samples-DGE-barley.ipynb*, *09-Ref-samples-DGE-wheat.ipynb* and *10-Ref-samples-comparism.ipynb* notebooks with using the *r.yml* package environment file as listed in the GitHub Repository.

3. Results

3.1 Text mining for flower development related publications

In this study, the NCBI PubMed database served as source for collecting publications for text mining. Our aim was to obtain all available studies related to the *Triticeae* tribe. We retrieved 80,826 articles associated with the tribe and published between 1965 and the 2nd of April, 2020. For text mining analysis, the abstracts and titles of these studies have been converted into a combined corpus object, which allowed us various investigations. Our purpose was to define a subset of studies involving research on flower development.

Following the preprocessing, the size of the corpus decreased considerably. With this step the number of elements in the corpus reduced from the initial 80,826 to 68,753 publications, described as *"Triticeae"* set in Figure 2A. Among the *"Triticeae"* list of publications, more than 3/4th of the papers mentioned scientific or common name of wheat (56,257 elements in the *"Wheat"* set), and 1/4th of them included barley (17,252 elements in the *"Barley"* set), which indicates the distribution of attention in research among the two major species of the tribe.

In addition, in Figure 2A represented as *"Flower-selection"*, the result of the 15 flowering related keywords-based filtering, from which 898 scientific papers were selected as a list of *Triticeae* and inflorescence-development related articles and were used in further analyses. Full list of the applied search words and their distribution is given in S2, which also illustrates how at least two out of the 15 selected words had to be present in each of the 898 articles. Such as in the case of "stigma", a single search word could indicate misleading studies, but the presence of more than

one search word considerably reduced the possibility of ambiguous cases.

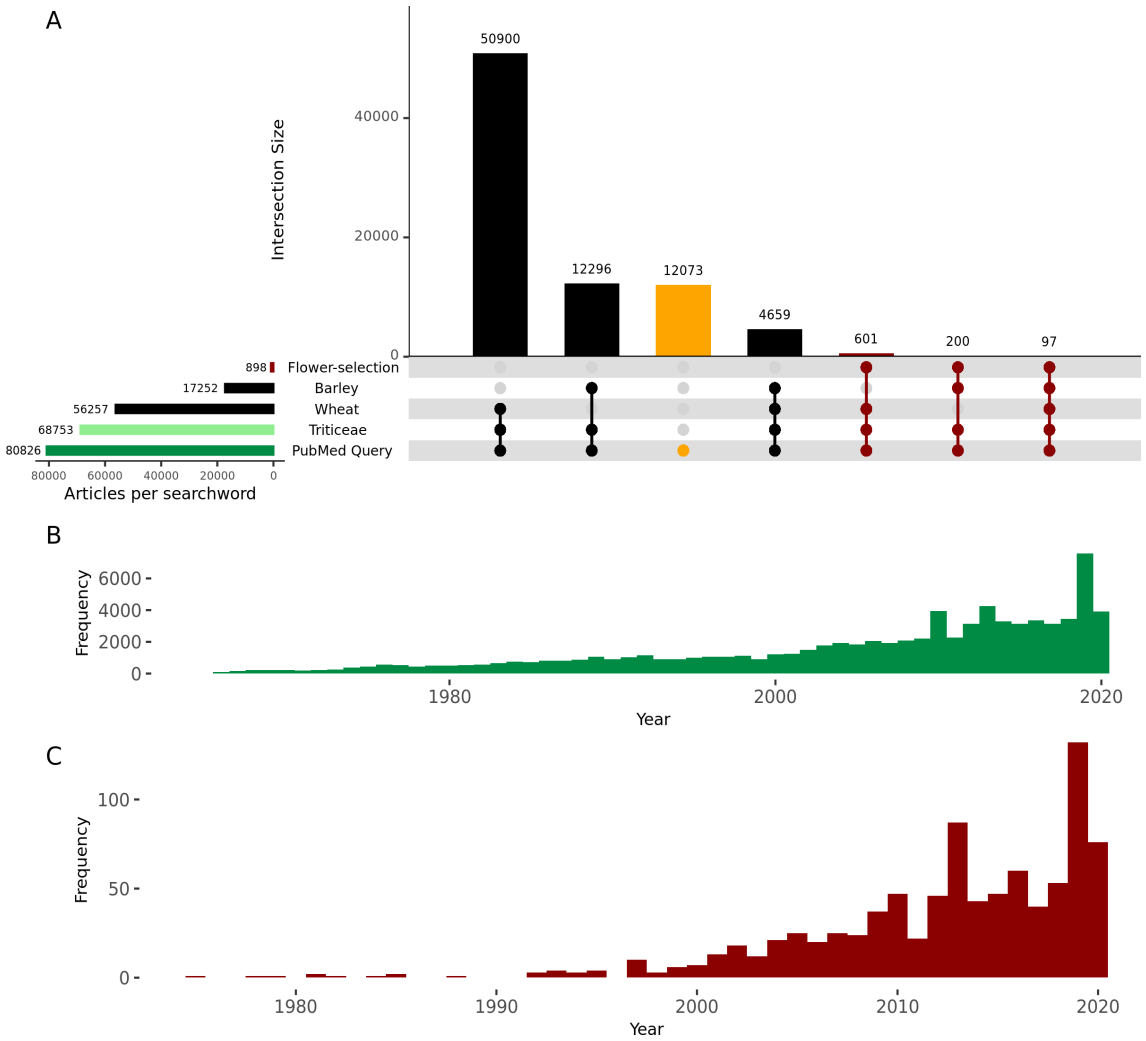

Figure 2: UpSet-plot of sub-sets in the *"PubMed Query"* text mining corpus, where (A) describes the intersections of elements among different filtering conditions, (B) illustrates the distribution of the elements across years of the full *"PubMed Query"* set, and (C) of the *"Flower-selection"* subset. Orange color represents the empty-intersection of the *"PubMed Query"* with all the other sets.

Looking at the division between the two species in the "Triticeae" set in Figure 2A, 50,900 papers deal exclusively with wheat, and 12,296 with barley, but only 4659 of the studies had mentioned them both. In the 898 studies of the *"Flower-selection"*, only 97 articles had both species discussed, which is a small number compared to the 601 publications on wheat or the 200 barley studies.

The magnitude of change in the amount of annual publications from 1965 to early 2020 allowed us to see the growth in research projects related to *Triticeae* and to

flower-development, as it is illustrated in Figure 2B and 2C, respectively.

3.2 Identification and projection of reference gene sets

Text mining of the *Triticeae* tribe related publications of the PubMed database, resulted in 898 supposedly flower-development associated studies. Employing the "Gene2Pubmed" database of the NCBI, publication identification numbers were linked to the respective taxonomic and gene identification numbers. By restricting "Gene2Pubmed" database to the 389 child taxa of *Triticeae*, we were able to associate 165 unique gene identifiers (GeneID-s) to our initial list of publications. These 165 GeneIDs comprised 254 unique Genbank entries, which where exported to FASTA format of the GenBank and Entrez Nucleotide databases and mapped to their respective reference genome as described in page 9.

The mapping resulted in 242 genes with 99-100 % identity with corresponding GenBank entries, from that 215 of them could be assigned to 149 Uniprot entries. As from the 242 mapped Genbank Nucleotide entries, many included independent submissions for the same locus, they could be collapsed into 215 unique GenBank entries. The 242 entries were intersected with respective reference genome annotations and projected onto latest versions of annotations, resulting in 289 locus to Genbank entry mappings. Five genes had no match in the new annotations and had to be excluded from the projection, and 18 genes overlapped with distinct homeologs in wheat. Through these manipulations we obtained 289 GenBank identifiers corresponding to 149 annotated Uniprot entries, corresponding to 175 loci in the barley and wheat genomes.

In the next phase of our analysis, ortholog detection on the selected genes among species of the *Triticeae* tribe allowed us to identify 118 subfamilies from 101 gene families (as defined by OrthoFinder) comprising 869 genes and corresponding to 145 Uniprot entries, represented in Table 1.

Thus we identified protein-coding genes for 242 Genbank Entries, representing 149 Uniprot Entries that could be mapped to 869 distinct loci in the wheat and barley genomes. These loci correspond to 118 subfamilies from 101 gene families.

Genome	Uniprot	Gene	Subfamily	Family
Hordeum vulgare	136	415	101	87
Triticum aestivum A	138	130	103	88
Triticum aestivum B	138	152	104	90
Triticum aestivum D	140	153	109	94
Triticum aestivum Un	15	19	9	9
Summary	**145**	**869**	**118**	**101**

Table 1: Results of identified reference gene sets across barley and wheat subgenomes and their projection to Uniprot Entries, genes, gene subfamilies and families.

3.3 Comparative transcriptomics among barley and wheat

By applying text mining on the flower-publication list, we identified 32 transcriptomic studies and further curated this list manually. Joining of this set of articles and further manual selection resulted in a final list of 22 publications and 3 additional NCBI SRA (Sequence Read Archive) RNA-sequencing datasets. Our collection comprised 25 RNA-sequencing datasets in total and consisted of 455 samples. In details, 19 datasets with 215 samples were gathered for wheat, and 6 datasets with 240 samples for barley, as represented in Figure 3A.

3.3.1 Attributes of RNA-sequencing sample collection

Our data selection resulted both publications and publicly available datasets to reach a large and comparable sample size for the two implied species, *Hordeum vulgare* and *Triticum aestivum*. Major selection criteria was to include exclusively openly available subjects with comprehensive metadata documentation. Unfortunately for three further potential publications neither the data was provided, nor the authors were available for supporting information.

Among the collection criteria, we focused on experiments with different developmental stages of several types of tissues rather than experiments investigating the effect of environmental variables or infection of pathogens. Besides gathering tissue-samples other than inflorescence, to employ them as control, we also targeted studies with natural mutations in flower-development regulatory genes, as their sample distribution represented in Figure 3B. The set of mutant-involving studies were especially important, because investigating deregulated genes in flowering gene pathways can provide an alternative angle highlighting genes that are important in the developmental process.

It was necessary to manually extract experimental variables from all of the selected articles and datasets, due to the lack of established agreement in the scientific community on publishing detailed metadata information. For the reason that several experiments with various design were combined into a single analysis, we aimed to mine as much comparable metadata as possible from each experimental layout as it is shown in Figure 3., S4, and S5. Unfortunately, the absence of uniform practice on metadata provision did not allow us to have complete collection for many of the selected variables, thus "unknown" values were assigned where the information was missing.

We aimed to select samples with as similar technical features as possible, and so we focused on Illumina sequenced paired read library layouts, apart from a few exceptional datasets with single reads, as it is shown in Figure 3A.

For all samples, the reported developmental stage definitions were extremely diverse and needed unification into a single type of developmental scale. Hence we established a conversion table across frequently cited growth scales, and selected the most commonly cited Zadoks-scale as the primary unifier staging method (Zadoks et al. 1974), based on the detailed sources of cereal staging field guides of Larsen and Smith 2010, and Fettell et al. 2010. While various phenological developmental scales are in practice for cereals, most of them either focuses on a narrow phase of the plant life-cycle, or on the contrary, describes only a few main growth stages vaguely. Namely, the Waddington-scale distinguishes development of the spike meristem and pistil from the transition of apex until the beginning of flowering (Waddington et al. 1983). Feekes-scale on the other hand, depicts a crop's entire life in 23 aspects

(Large et al. 1954). BBCH-scale is the latest developmental staging method, and bases mainly on the widely accepted decimal Zadoks-scale (Lancashire et al. 1991). Thus, both of them are described by a two-digit code between 00 and 99, where the first refers to one of the 10 primary growth stages, numbered from 0 to 9, and the second number indicates the secondary stage of development, allowing precise staging. It is important to note that several stages of the Zadoks-scale occur together, therefore a plant can be described with more than one decimal code at the same time.

For our study, each sample has been assigned to a single stage of the Zadoks-

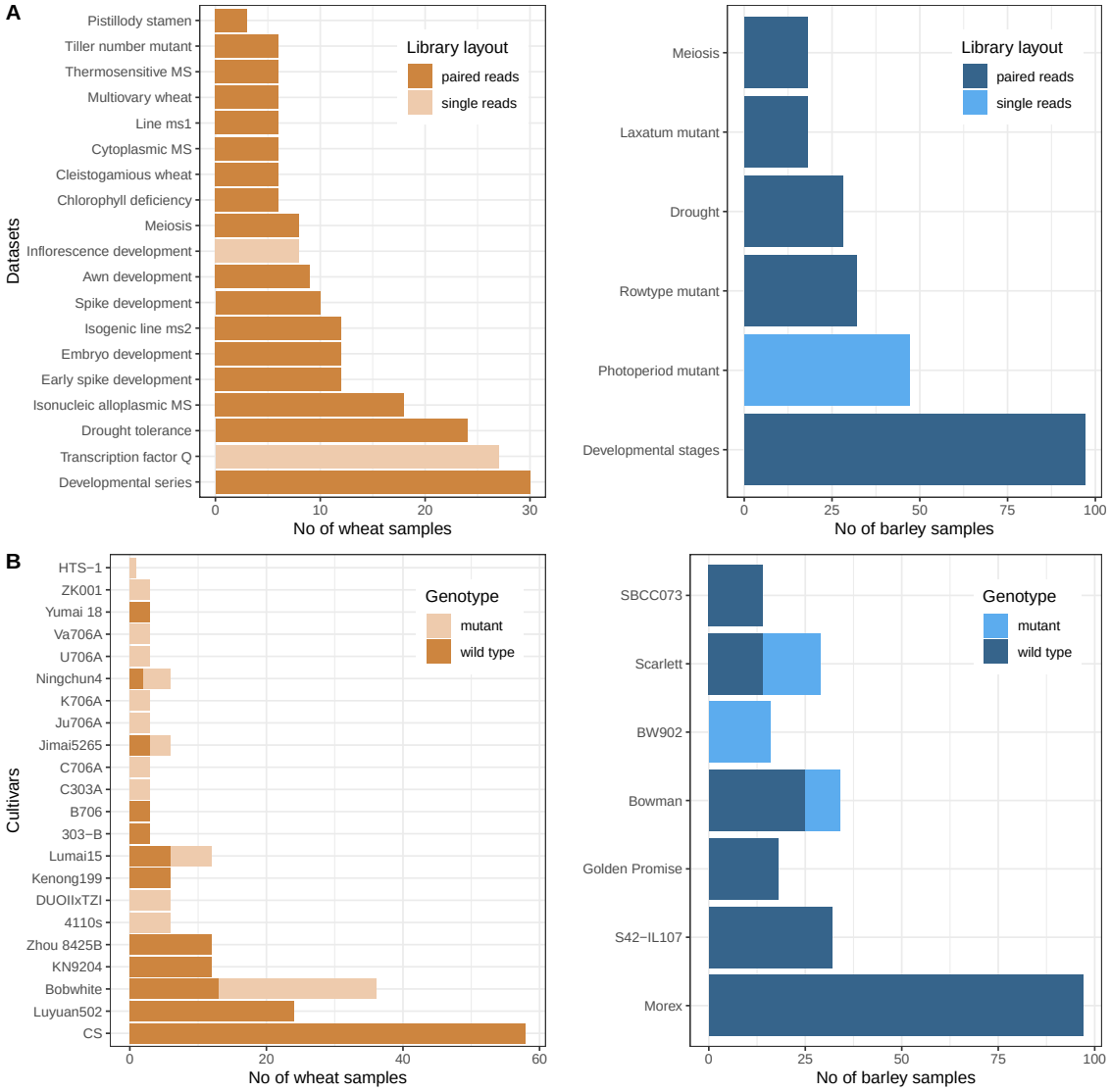

Figure 3: Distribution of datasets (A) and cultivars (B) of the RNA-seq samples of wheat and barley representing their genotype and library layout.

scale, based on the reported description. Furthermore the 100 categories of the scale were grouped into 11 "Intermediate age", and 5 "High-level age" categories, following the original description of the decimal scale, as it is represented in Figure 4A-C. Additionally, in our study, the stages between 13-69 are described as *"extended reproductive phase"*, which includes flower primordium development stages within 13-30. The grouping of the Zadoks-scale stages into higher levels are reported in Table S1. An illustration of the barley life cycle in Figure 4D supports the understanding of diverse staging definitions.

Likewise, the various types of tissues (referred as "Intermediate tissue"), illustrated in Figure 5E, were organized into 7 categories of the "High level tissue" variable, based on which plant organ they belong to, as it is detailed in Table S2. This reorganization was also necessary to have comparable sized groups of functionally similar tissues, as the distribution of samples across different tissue categories is compared among the two species in Figure 5A and 5C. Detailed representation with developmental stage and tissue type dimensions is shown for wheat and barley samples on a heat map in S6.

3.3.2 Preprocessing, mapping and exploratory analysis

The trimmed and quality controlled datasets comprised 734 GB data for barley and 742 GB for wheat, which taken together resulted in a 1.44 TB quality cleaned data. In total, mapping revealed the expression of 49,281 transcripts for barley and 123,075 transcripts for wheat samples, and it was performed on the most frequent cultivars of each species's dataset, namely on Chinese Spring for wheat and Morex for barley.

Dimensionality reduction methods for data visualization help to explore structural patterns of complex datasets, which can explain causes of batch effects, and detect potential outliers. With the quickly expanding openly available data-sources and the new standard of large scale data analysis of the past decade, recently formulated and potentially better performing statistical approaches emerged to serve these purposes. In this study, two unsupervised machine learning methods have been compared to explore our datasets. First, we tested Principal Component Analysis (PCA), which was invented in 1901 by Karl Pearson (Pearson 1901), is a linear projection of the

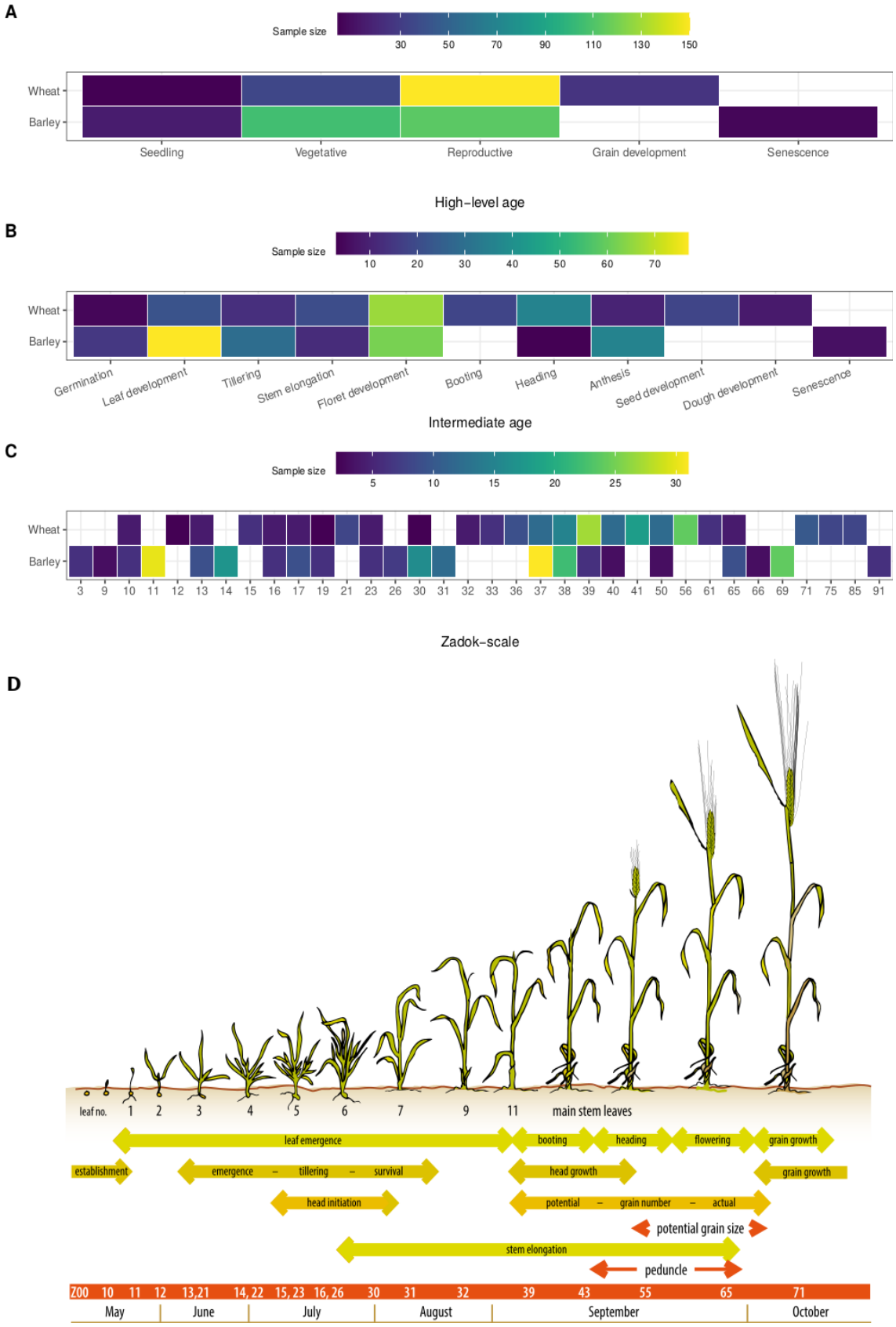

Figure 4: Distribution of barley and wheat samples across (A) high-level age, (B) intermediate age, and (C) Zadok-scale categories, along with (D) representation of the barley life cycle, adopted from Fettell et al. 2010.

data, and it describes directions in a high-dimensional space, and creates a low-dimensional representation that maximizes the variance.

First, we performed 4 kinds of PCA to find the best fit to explain variation in our gene expression matrix as it is represented for the barley datasets in S8. We compared the effects of statistical normalizations, such as scaling, centering and log2-transformation on non-expression cutoff filtered or non-filtered gene expression matrix. Auto-scaling for the matrix computed scaling to zero mean and unit variance for each gene. Centering removed the mean from each unit, and log-transformation spread out the clusters and made the data closer to be linear.

Scaling and centering clearly reduced the differences between genes and consequently samples, but could not substitute for the effect of log-transformation, to reduce the strong influence of outliers on skewness of our data. Therefore log2-transformation alone was chosen to correct our data for PCA as in S8C, and was used to describe various technical and biological variables shown in S10., S11., S12., and S13. PCA visualized strong division between single and paired read library type samples as shown in S10G and S12G, suggesting to correct for library type before DGE analysis. Furthermore, two distinct outliers of the barley dataset could be spotted through all of the PCA plots, which was further investigated.

Because PCA aims to describe as much variance in the data as possible, it cannot capture all aspects of variation, and so more delicate patterns of the data will be lost with its application (Diaz-Papkovich et al. 2020). Manifold learning, or also called nonlinear dimensionality reduction methods have been invented to overcome these limitations. The second dimensional reduction approach discussed in our research, is Uniform Manifold Approximation and Projection (UMAP). UMAP was created in 2018, and beside preserving global structure of the data using nearest neighbour algorithm, it is also effective for visualizing clusters of data and their relative proximities (McInnes et al. 2018). Its main feature is to keep similar data points close to each other while distancing dissimilar ones.

Like PCA, we explored the effects of applying a non-expression cutoff and log2-transformation on our expression matrix prior to perform UMAP, as it is shown in S9. From these comparisons, it was clear that filtering the non-expression cutoff brings clusters closer to each other, by excluding unnecessary noise from our

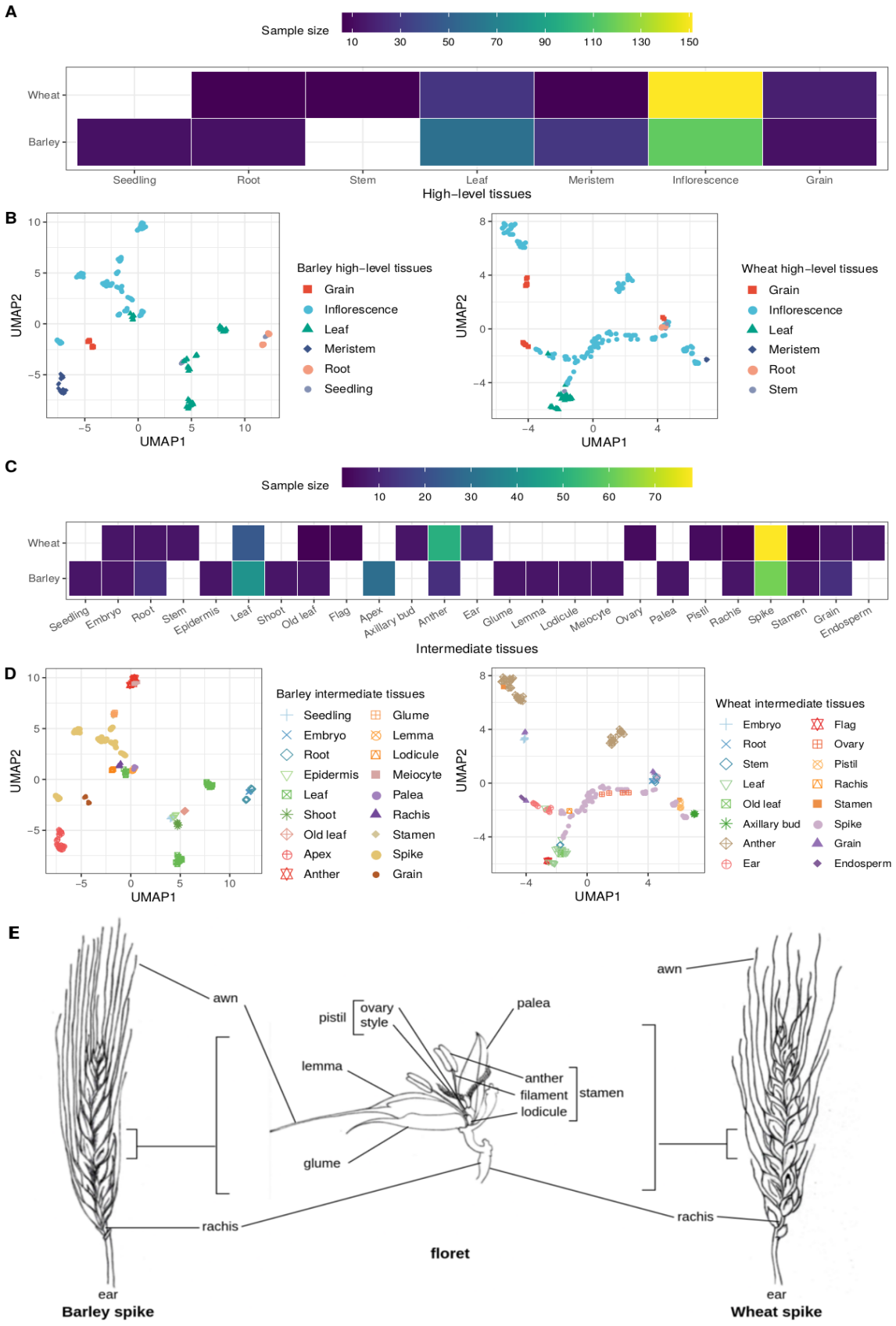

Figure 5: Distribution of RNA-seq samples across species and tissue categories. Heat maps depict sample size differences of (A) high-level, and (C) intermediate tissues. UMAP analyses show clustering of (B) high-, and (D) intermediate-level of tissues. (E) Illustration of anatomy of barley and wheat spike.

data. Interestingly, we found major differences in the effect of log2-transformation between the barley and wheat datasets, as in S9C and S9F. In the barley dataset, log2-transformation preserved and enhanced the clustering of similar tissue types together, while in wheat it dissolved the distinct groups that were present before. Furthermore, Figure 5B and 5D illustrates how clusters of UMAP analysis are explained by high-level and intermediate tissue types in both datasets.

Additionally to the multivariate statistical approaches, we conducted Expression Level Category (ELC) analysis to determine the abundance of transcripts in five distinct groups of the gene expression matrix. The E0-E4 groups defined genes, whose expression level is similar to each other within each sample. More details on the exact method are described on page 12. We performed PCA on the 5 ELCs and found that it compressed variation among datasets but showed more distinct clustering between library types, in the barley dataset as represented in S10B and S10H. Moreover, we wanted to capture samples where the abundance of highly expressing genes (E4 category) is significantly different from other samples as in S14 and S15. The reason was that in some of the investigated tissues it is difficult to isolate RNA for sequencing, and such technical difficulties can lead to misinterpretations in gene expression analysis. As an example, wheat pistils are rich in polysaccharides which can co-precipitate with the isolated RNA (Manickavelu et al. 2007), and so the low-quality RNA samples can cause over-representation of strongly expressing genes, which appears as an especially high proportion of E4-category. Thus, ELC analysis for expression bias helped in outlier detection in consort with the PCA results, as barley samples ERR781040 from the "Photoperiod mutant" dataset, and ERR1457187 from the "Developmental stages" dataset had to be excluded from further analysis (S14). Exclusion of these samples was consistent with their original studies (Consortium et al. 2012, Digel et al. 2015).

All the above described methods are summarized with different technical and biological variables for both datasets in S10., S11., S12., and S13.

3.3.3 Comparative Differential Gene Expression Analysis

To investigate the expression patterns of genes in flower development, we tested four types of differential analyses. Depending on the involved set of samples, 20,302

- 21,153 genes were included for barley, and 60,955 - 65,540 genes for the wheat dataset in the tests. Our main strategy in the design of contrasts was to primarily normalize for dataset batch effects, including the differences between single and paired reads, and build "spatiotemporal" contrasts of samples on top of that. In this manner, in order to test for differences between tissues, which could be comprehended as spacial constrain, we had to normalize for developmental stage differences as a time constrain, and vice versa. As final part of our analysis we projected the identified gene sets into orthologous gene subfamilies between the barley and wheat genomes. Furthermore, for the current analyses, mutant genotypes were excluded except for the fourth test. Results of the created DGE contrasts and their projection to subfamilies are summarized in Table 2.

In the first DEG test, referred as *"Inflorescence"*, we compared inflorescence versus all other classes of tissues against reproductive or non-reproductive developmental stage samples. We identified 2,987 subfamilies for barley and 2,765, 2,582, and 2,811 subfamilies for the wheat A, B and D subgenomes, respectively. Of the DEGs that were regulated during the reproductive developmental stage, proportionally more showed upregulation then downregulation across all genomes. The magnitude difference between up- and downregulated genes appeared especially strongly in wheat subgenomes. Additionally, the distribution of subfamilies were equivalent over barley and wheat A, B and D subgenomes.

Next in the *"Primordium"* test, we identified the critical stage of flower primordium development as Zadoks-stages 13-30 and contrasted inflorescence and meristem samples from this developmental window with mature flower tissues of Zadoks-scales 31-69. 7,363 subfamilies for barley, and 2,211, 2,059, 2,357 subfamilies for the wheat A, B and D subgenomes accordingly, were detected to be significantly differently expressed. In primordium stage samples, more DEGs were downregulated than upregulated across all genomes. In contrast to *"Inflorescence"*, in the *"Primordium"* test the barley genome showed more then three times higher number of subfamilies compared to the wheat A, B and D subgenomes.

In our third contrast design, assigned to *"Anther"*, we investigated expression differences between one specific flower organ, the anther, against all the other types of inflorescence tissues, correcting for developmental differences with the intermediate

Genome	Category	Contrast			
		"Inflorescence"	*"Primordium"*	*"Anther"*	*"MS"*
Barley	Subfamily	2987	7363	3466	-
	DEGs	3050	7610	3644	-
	DEGs up	1694	3591	1745	-
	DEGs down	1356	4019	1899	-
Wheat A	Subfamily	2765	2211	4912	3341
	DEGs	2840	2263	5058	3425
	DEGs up	1878	636	2871	1566
	DEGs down	962	1627	2187	1859
Wheat B	Subfamily	2582	2059	4719	3165
	DEGs	2666	2105	4898	3222
	DEGs up	1734	542	2889	1552
	DEGs down	932	1563	2009	1670
Wheat D	Subfamily	2811	2357	4950	3328
	DEGs	2878	2402	5079	3405
	DEGs up	1959	711	2667	1577
	DEGs down	919	1691	2412	1828
Wheat Un	Subfamily	79	71	168	115
	DEGs	87	78	186	127
	DEGs up	65	14	107	55
	DEGs down	22	64	79	72

Table 2: Results of DGE-analyses. Transcript changes in four developmental or experimental factor contrasts were used to identify significantly different up, or down-regulated gene-sets (DEGs), and their corresponding subfamilies across barley and wheat subgenomes (chromosome A, B, D and Unknown). Contrast *"Inflorescence"* represents inflorescence versus non-inflorescence tissues, *"Primordium"* indicates primordium versus mature flower tissues, *"Anther"* stands for anther versus flower tissues and *"MS"* implies mutant male sterile versus wild type fertile anther samples.

age categories. The 3,466 subfamilies identified for barley were substantially less than the 4,912, 4,719 and 4,950 subfamilies determined significantly differentially expressed in the wheat A, B and D subgenomes, respectively. Among anther samples, slightly more genes were upregulated in wheat subgenomes, in contrast with barley, where more genes showed downregulation.

In the last *"MS"* contrast, we tested male sterile mutants against fertile wild type anther samples using intermediate age categories for developmental stage correction, exclusively in the wheat dataset. Among the identified subfamilies, 3,341, 3,165, 3,328 of them came from the A, B and D subgenomes, respectively, with no indication of subgenome preference. In the sterile mutant samples, more genes showed downregulation across all subgenomes.

To visualize the uniquely shared proportions of the identified subfamilies among barley and wheat genomes, we illustrated their first 50 largest intersections in Figure 6B. We found that the largest differentially expressed subfamily set is from the *"Primordium"* contrast of the barley genome. Within the DGE tests of barley samples, 524 subfamilies are shared between the *"Primordium"* and the *"Anther"* contrasts, 271 subfamilies of *"Inflorescence"* and *"Primordium"*, and only 55 subfamilies among the *"Inflorescence"* and the *"Anther"* contrasts. Furthermore among all 3 of types of test within the barley genome 64 shared subfamilies were identified. In the wheat A, B and D subgenomes, the widest overlay was detected in the 332 subfamilies within the *"Anther"* contrast and 266 subfamilies within the *"MS"* contrast. Throughout the entire analysis, chromosome Unknown included very few subfamilies compared to A, B and D subgenomes and thus there is little overlay among them which is not included in the largest intersections in Figure 6B.

When comparing shared intersections between DGE subfamily sets of barley and wheat, the largest overlap is 131 subfamilies between the *"Primordium"* test of barley and *"Anther"* contrast of wheat A, B and D subgenomes. Including the *"Anther"* contrast of barley to the previous combination, increases the shared number of subfamilies with 80 more. Additionally, 67 subfamilies are uniquely shared among the *"MS"* contrast of the wheat A, B and D subgenomes and the barley *"Primordium"* contrast.

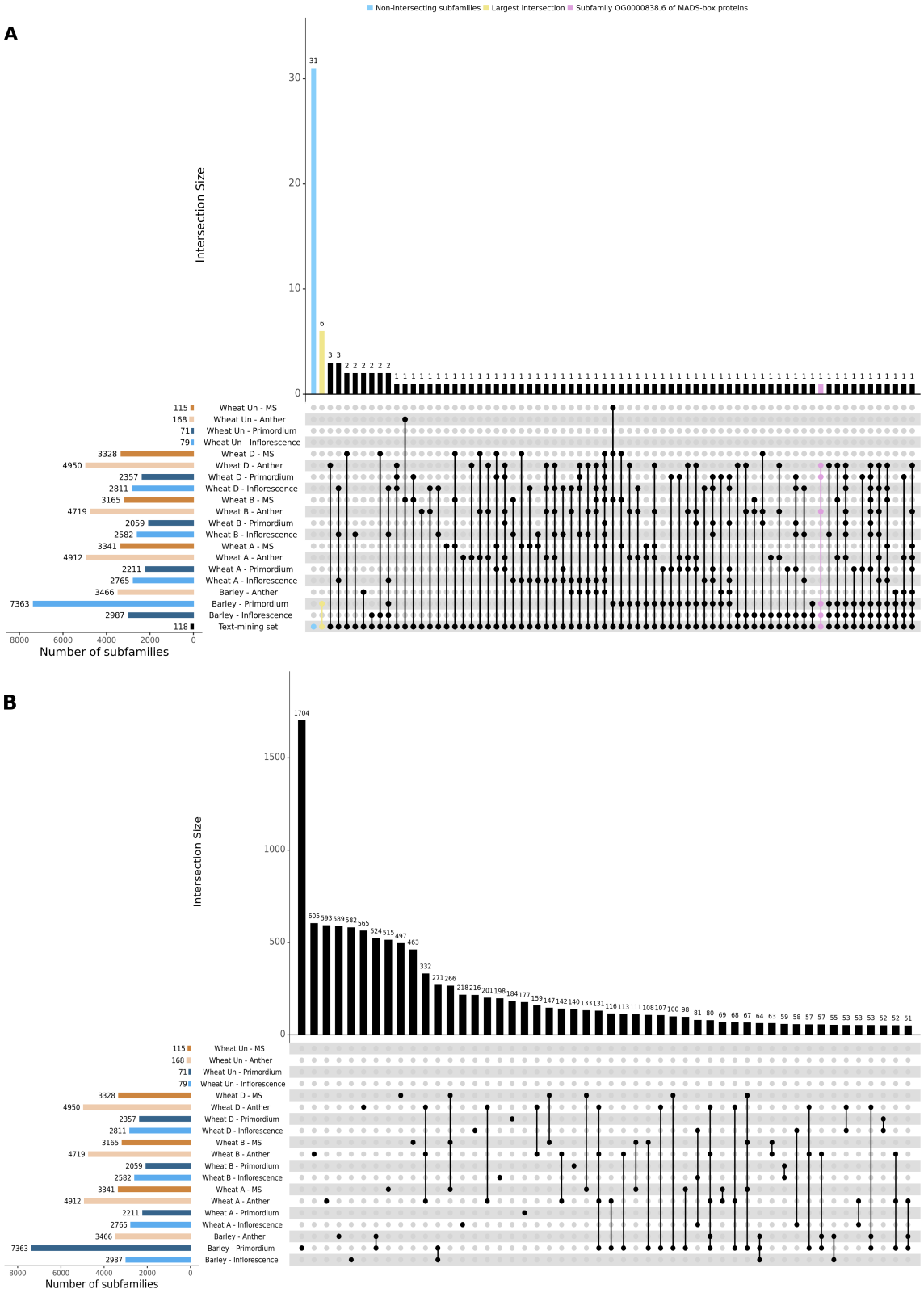

Figure 6: Subfamily intersections across different genomes and DGE contrasts, (A) describes all intersections between reference gene set ("Text-mining set") and DEG-sets, and (B) illustrates first 50 intersections between DEG-sets.

3.4 Assessment of reference gene set with DEG-sets

With an aim at identifying shared elements of reference genes (referred as *"Text-mining set"*) and differentially expressed subfamilies of barley and wheat genomes, Figure 6A illustrates the respective intersections among them. In our study, we performed DGE to determine gene sets responsible for changes in the flower development gene regulatory network of barley and wheat genomes. Similarly, we used text-mining to define a reference gene set that have been published in relation to inflorescence development. As the blue column in Figure 6A shows, apart from 31 non-intersecting subfamilies, 87 out of 118 subfamilies of the *"Text-mining set"* has been detected through the four types of DGE analyses. Additionally, as it is marked with yellow in Figure 6A, 6 subfamilies are uniquely shared between the *"Primordium"* contrast of barley genome and the *"Text-mining set"*. Each of these 6 subfamilies translates to a single Uniprot Entry, such as MutS homolog 7 (*Q8RVT1*), Putative MAP kinase phosphatase (*Q70AJ6*), Putative ribosomal protein (*Q70AI2*), Putative AP2-like protein (*Q70JR4*), PHD-type domain-containing protein (*Q70JP3*) and DREB transcription factor 3A (*Q2TN83*).
However, 6 subfamilies are shared exclusively among the *"Primordium"* contrast of barley genome and the *"Text-mining set"*, in total they share 36 subfamilies, from which 30 subfamilies are common with other DEG identified subgenomes as well. Interestingly, many of these 30 subfamilies pointing at similar types of proteins, namely MADS-box proteins. One example is subfamily OG0000838.6, marked as lilac column in Figure 6A, which translates into three Uniprot Entries, such as MADS-box protein (*Q1G169*), Putative MADS-box protein 7 (*Q70JR1*) and MADS-box transcription factor TaAGL24 (*Q1G184*).

4. Discussion

4.1 Comparative meta-analysis of flower development transcriptomics in barley and wheat

To date, no systematic analyses of gene expression profiles during the important process of flower development have been conducted for comparisons among barley and wheat genomes. Here, we repeated the analyses of previously reported RNA-sequencing samples from 22 peer-reviewed publications, and two additional SRA datasets. Our data collection comprised of 240 samples for barley, and 215 samples for wheat (Figure 3.). Preprocessing revealed 42,775 transcripts for barley and 120,297 transcripts for wheat samples at levels greater than 0.5 TPM in at least two libraries. We collected extensive metadata and defined a conversion table across several developmental stage definitions to translate each samples developmental age into Zadoks-scale (Figure 4.). Furthermore, exploratory statistics allowed us to validate the integrated analysis of datasets from multiple experiments. Several multivariate statistical methods are available for dimension reduction and outlier detection, from which we performed PCA and UMAP analyses.

Interestingly, barley and wheat datasets showed distinct behaviour for log2- transformation prior UMAP (S9.). We assume these differences are due to the genome size characteristics, where the wheat gene expression matrix had almost 2,5 times more genes than barley. Besides, for barley, we obtained samples from 6 datasets, which can cause skewness in the data collection, on the contrary to wheat, where 19 separate experiments contributed to our sample collection. Considering these facts and the differences in the UMAP plots, we can conclude that the wheat dataset was closer to be in normal distribution and thus does not need log-transformation, compared

to the probably highly skewed data of barley datasets, where log-transformation was beneficial.

We employed ELC analysis for detecting samples with excessive proportion of highly expressing genes. However, we could capture the same samples, that showed up as outliers in the PCAs for the barley dataset (S14.), the ELC analysis for outlier detection needs more development for consistent results comprising whole datasets.

In conclusion, we demonstrated the integrity of our sample selection and visualized how different technical and biological variables contribute to the variance of the data (S10., S11., S12., and S13.). From these plots, it was clear that merging samples of different cultivars, genotypes, and growth conditions did not cause batch effect in our datasets. Contrarily, the need for correction of experimental set differences are clearly shown, especially in the barley dataset. UMAP plots indicated that most of the similarity in gene expression among different clusters was a consequence of tissue types, rather then developmental stage (Figure 5., S11. and S13.).

To summarize our exploratory analyses, an important feature of PCA is that it is highly influenced by excess data points, which is advantageous in the process of identifying outliers, but fall short in exploring global data structure. UMAP on the other hand, is excellent in preserving global structure, but fails to identify outliers causing the highest variation in a dataset. Moreover, with clustering similar samples together, UMAP can capture mislabelling, and correct for such human errors. Thus we find both UMAP and PCA important exploratory methods during preprocessing of gene expression data.

4.1.1 Transcriptome profiling of inflorescence development

To identify genes underlying inflorescence developmental regulation, RNA-seq samples of different tissues and developmental stages were analyzed using a comparative transcriptomics approach. The genes identified to be significantly different during reproductive stage showed upregulation, proportionally twice as much as presented downregulation (Table 2., *"Inflorescence"* contrast). This observation indicated that more genes are activated during reproductive phase than other developmental stages such as germination, or vegetative growth.

To reveal the characteristics of floral meristem development, in our second DGE test,

we targeted early reproductive development and found that extensively more genes are upregulated than downregulated (Table 2., *"Primordium"* contrast). Together with our conclusion from the first DGE test, we suggest that mature flower tissues have higher number of actively expressing genes than primordium stage samples.
Our hypothesis is supported by a recent study conducted by Feng and his colleagues, who investigated expression differences between meristematic tissues at different early spike developmental stages in wheat (Feng et al. 2017). In their DGE results, they identified substantially more upregulated genes in flower tissues that had passed primordial development, compared to meristematic samples.
Furthermore, the distributions of subfamilies between barley and wheat subgenomes were equivalent in the first DGE test. On the contrary, we observed striking differences of the numbers of identified subfamilies between barley and wheat subgenomes in the other applied tests for primordium stage and anther development (Table 2., *"Inflorescence"*, *"Primordium"* and *"Anther"* contrasts). Differences were considerably large not only between the two species, but also within wheat subgenomes, which indicates differential expression profiles in terms of subgenome orthologs. Consistently with our observation, functional diversification just as dynamic expression of orthologs in the polyploid wheat genome associated to inflorescence development has been shown in previous studies (Feng et al. 2017, Ma et al. 2018).

4.1.2 Insights into anther development and male sterility

Having economic interest in hybrid wheat research, proportionally more transcriptomics studies are available from recent years investigating male sterility. Such as within the 19 resources of our wheat dataset, four publications discussed exclusively male sterility concerning stamen or anther RNA-seq samples (Wang et al. 2017, Liu et al. 2020, Liu et al. 2020 and Yang et al. 2019). Heterosis, or hybrid vigor is an important genetic mechanism when a hybrid progeny outperforms its parents, with increased yield, adaptability or resistances to biotic and abiotic stressors (Fu et al. 2015). Despite some well-established genetic crop improvement systems in rice or corn, a stable trait for wheat remains a challenge, as much as resolving its molecular background.
In our third DGE analysis, fertile anthers compared to inflorescence tissues presented

more differentially regulated genes in wheat, than in barley (Table 2., *"Anther"* contrast) with slightly more upregulated genes in wheat, and more downregulated genes in barley. While it seems controversial for the first review, the result can be easily explained by the developmental stage differences between the samples of the two datasets. Wheat anther samples comprised of more mature booting (Zadoks-scale 41-49) and heading stages (Zadoks-scale 50-60), whereas barley samples were from younger age of stem elongation (Zadoks-scale 31-36) and floret-development (Zadoks-scale 37-40), as it is shown in S6. and in Table S1. To sum up, in the above DGE, wheat samples were more specialized than barley samples, which were closer to the primordium stage. Taking into consideration our established conclusion, according to which the further matured stages of inflorescence tissues present more upregulated DEGs, the result of this DGE contrast confirms our observations.
We investigated sterile mutant anthers against fertile wild types as our last DGE analysis (Table 2., *"MS"* contrast), and found that sterile samples had consistently more downregulated than upregulated genes. This observation aligns with one of the original studies, where Liu and her colleagues presented that considerably more genes are downregulated in sterile anthers compared to fertile ones (Liu et al. 2020).

4.2 Reproducibility and data mining as standards for life sciences

In the current study, we performed repeated transcriptomic meta-analyses of 455 RNA-seq samples. From our results, it is clear that combining several datasets can lead to a comprehensive analysis across more species. By presenting consistent conclusions to the original publications, we proved that integration of data from distinct experiments can increase the robustness of such analyses. An important requirement for such investigations is an established metadata documentation that especially allows uniform reporting of detailed tissue type and developmental stage description, possibly in a widely used consensus such as the Zadoks-scale. Considering that more than 2400 RNA-seq samples for wheat and more than 1600 samples for barley are already openly available in the SRA database, future integrative meta-analyses are

only possible with improved documentation standards.

It is well acknowledged that novel scientific discoveries are being generated at nearly an exponentially increasing rate. Several databases are stakeholders of the growing information, such as wheat research specific URGI, Wheat eFP Browser, WheatExp, PlantGDB, Wheat Expression Browser, EnsemblPlants, the most recent KnetMiner or just as one of the earliest resource, the NCBI. However, many of them are well-maintained data reserves, peer-reviewed publications dominate the primary communication among international scientific community. Besides, such curated databases can serve as additional, supporting information sources, but cannot substitute for the analysis of raw data to engage in novel research questions.

The PubMed database of NCBI has become one of the most prominent collecting source that openly provides over 11 million citations and abstracts of the natural sciences. Moreover, the database has its own preprocessing of articles, by extracting their meta information into distinct search-fields, which makes them suitable for text mining analysis. Using PubMed, we investigated all peer-reviewed publication concerning Triticeae, and determined a set of 898 publications inspecting flower development, corresponding to 118 gene subfamilies shared between wheat and barley. Text mining further allowed us to picture globally the progress of research in *Triticeae* and even closer in flower development. While the number of yearly publications has been growing gradually as in Figure 2B-C, the average seems to be doubled in every 10 years since 2000. This observation aligns with the recent achievement of wheat genome sequencing in 2014 and its continuous improvement along with the more affordable and thus widely available methods of NGS sequencing. Having both the reference genome and transcriptome sequencing as prerequisites available, the vastly expanding scientific information demands urgency for computational data mining and processing.

4.2.1 Evaluation of the reference gene set

Establishment of references are crucial for large-scale computational analyses to ensure reliability of novel discoveries. In the present study, we targeted genes that have been investigated in scientific publications and thus presumably had been validated for their functionality. Our attempt was to create a reference gene set dedicated

to one subject, in our case, inflorescence development, and to validate its accuracy with transcriptomics analysis as gene expression patterns are good indicators of gene functions.

With an aim at detecting shared elements among our reference gene collection and differentially expressing subfamilies, both of them involved in flower development in the barley and wheat genomes, we intersected the identified differentially expressed subgenomes with the reference gene set. As a result, we found 87 shared gene subfamilies, and 31 subfamilies which had no overlap with any of the DGE-sets.

It is important to underline that we did not expect all of our reference genes to be present in flower development targeting differential gene expression sets. While the process of inflorescence development is under strict tempo-spatial regulatory control (Rutley and Twell 2015), there can be several reasons why a gene is present in our reference set but missing from DGE tests. Such as, studies can introduce genes that are essential for flower development but their point mutation in a crucial domain, like in a miRNA or histone binding site could lead to disruption of flowering. Another further possibility is that some of the reported genes, upon mutation, can cause a special flowering phenotype, but in their wild type, their presence does not necessarily reflect on expression level differences. Furthermore, our goal was to explore the abundance of available literature in this field and how well genes are characterized related to flower development.

Taking this into consideration, our results complemented well with studies involved in wheat and barley floral development, as almost half of them reported MADS transcription factors regulating floral meristem development (Ma et al. 2018). In our intersection, we found more than one shared subfamilies that translated to the protein class of MADS-box transcription factors. Moreover, out of 22 original studies that has been included in our analyses, more than half of them mentioned MADS-box proteins in their DGE analyses, such as (Liu et al. 2020, Bull et al. 2017, Li et al. 2018).

4.3 Perspectives on a large-scale comparative study in *Triticeae*

Despite newly appearing comprehensive databases on wheat transcriptome and gene developmental atlases, a systematic study of comparative transcriptomics in *Triticeae* is needed for global insights into gene evolutionary network development. Transcriptome profiling of several key developmental stages between closely related species such as barley and wheat, is our next goal to achieve. With this project, we aimed to explore the demands and difficulties of such a larger-scale analysis.

In the future, our project should undergo several improvement in methodology, such as DEG-sets should be more comprehensively investigated, performing K-means clustering and functional annotation on the identified sets and compare them directly with DEG-sets of the original studies to evaluate their reliability. In the next steps Weighted Gene Coexpression Network Analysis (WGCNA) should provide more insight in gene regulatory networks across developmental stages. Further goals are to explore the interaction between different homeologs and the overall expression level of all homeologs of key regulatory genes. Last but not least, semi-automation of the entire pipeline is a future goal as well.

The current study brings a novel approach to the field with integrating various resources of RNA-seq samples into a meta-analysis of comparative transcriptomics, while it attempts to include peer-reviewed discoveries for a particular question, flower development.

GitHub Repository

Main site to the: https://github.com/vanda-marosi/master_thesis repository

1. Data-tables:

- Metadata-Zadoc-PO.ods
- barley-samples.csv
- wheat-samples.csv

2. R scripts:

- 01-Text-mining-download.ipynb
- 02-Text-mining-analysis.ipynb
- 03-Ref-genes-download.ipynb
- 04-Ref-samples-download-preprocess.ipynb
- 05-Ref-samples-mapping.ipynb
- 06-Ref-samples-metadata-visualization.ipynb
- 07-Ref-samples-PCA-barley.ipynb
- 07-Ref-samples-PCA-wheat.ipynb
- 08-Ref-samples-ELC-barley.ipynb
- 08-Ref-samples-ELC-wheat.ipynb
- 09-Ref-samples-DGE-barley.ipynb

- 09-Ref-samples-DGE-wheat.ipynb
- 10-Ref-samples-comparism.ipynb

2. Bash scripts:

- Preprocessing, mapping for barley samples
- Preprocessing, mapping for wheat samples
- GMAP mapping for reference genes

3. Conda environment files:

- r.yml
- tximport.yml
- base.yml
- seqtools.yml

Acknowledgments

First and foremost, I am indebted to my direct supervisor, Dr. Daniel Lang, for offering me the opportunity to work on this project, and more importantly, for his continuous support and inspiration in the process of my study. His always encouraging feedback, patient guidance and fruitful discussions showed me various sides of computational biology, and eventually helped me to choose further direction in my academic career.

I am sincerely grateful to Prof. Dr. Klaus Mayer for giving me the possibility to join his research group for my master thesis, and for introducing me to my internal advisor Prof. Dr. Korbinian Schneeberger. I am also grateful to Prof. Schneeberger for supporting my ambitions and undertaking this project to his supervision.

I especially thank Nadia Kamal, Maxim Messerer and Leon van Ess for their involvement and advice. In addition, I appreciated the great research environment established by all members of the Plant Genomics and Systems Biology group of the Helmholtz Zentrum and would like to thank for all the special insights into their ongoing projects and passionate debates on science.

Furthermore, I would like to thank to the BAYHOST organization, having supported my ambitions for two years long as well as helped me immensely to concentrate on my studies through their financial support.

I would like to thank Amit Fenn for his constant and continuous support in every day of the last months, also for his encouragement, to be always ready to share ideas, and to grow together to become bioinformaticians.

I am grateful to my mother and father for always encouraging me in all my decisions.

Last, but not least, I would like to thank my friends, Virág Kocsis, Chit Tong Lio, Mirjam Plaschke, Lorenz Oberkofler, Yagya Chadha, Timotheus Fischer and Furkan Tunc for their support during my master thesis.

Abbreviations

BBCH	Biologische Bundesantstalt, Bundessortenamt und Chemische Industrie
DEGs	Differentially Expressing Genes
DGE	Differential Gene Expression
ELC	Expression Level Category
GB	Gigabyte
GeneID	Gene Identifier
GSA	Genome Sequence Archive
LRT	Likelihood Ratio Test
MS	Male Sterility
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PMID	PubMed ID
PO	Plant Ontology
SRA	Sequence Read Archives
TB	Terabyte
TPM	Transcripts Per Million
UMAP	Uniform Manifold Approximation and Projection
WGCNA	Weighted Gene Coexpression Network Analysis
WT	Wald-test
XML	Extensible Markup Language

Bibliography

Anaconda Software Distribution 2016.
URL: *https://anaconda.com/*

Andrews, S. et al. 2010, 'Fastqc: a quality control tool for high throughput sequence data'.

Appels, R., Eversole, K., Stein, N., Feuillet, C., Keller, B., Rogers, J., Pozniak, C. J., Choulet, F., Distelfeld, A., Poland, J. et al. 2018, 'Shifting the limits in wheat research and breeding using a fully annotated reference genome', *Science* **361**(6403).

Badr, A., Rabey, H. E., Effgen, S., Ibrahim, H., Pozzi, C., Rohde, W. and Salamini, F. 2000, 'On the origin and domestication history of barley (hordeum vulgare)', *Molecular biology and evolution* **17**(4), 499–510.

Bolger, A. M., Lohse, M. and Usadel, B. 2014, 'Trimmomatic: a flexible trimmer for illumina sequence data', *Bioinformatics* **30**(15), 2114–2120.

Borrill, P., Adamski, N. and Uauy, C. 2015, 'Genomics as the key to unlocking the polyploid potential of wheat', *New Phytologist* **208**(4), 1008–1022.

Bray, N. L., Pimentel, H., Melsted, P. and Pachter, L. 2016, 'Near-optimal probabilistic rna-seq quantification', *Nature biotechnology* **34**(5), 525–527.

Bull, H., Casao, M. C., Zwirek, M., Flavell, A. J., Thomas, W. T., Guo, W., Zhang, R., Rapazote-Flores, P., Kyriakidis, S., Russell, J. et al. 2017, 'Barley six-rowed spike3 encodes a putative jumonji c-type h3k9me2/me3 demethylase that represses lateral spikelet fertility', *Nature communications* **8**(1), 1–9.

Chamberlain, S., Szoecs, E., Foster, Z., Arendsee, Z., Boettiger, C., Ram, K., Bartomeus, I., Baumgartner, J., O'Donnell, J., Oksanen, J., Tzovaras, B. G., Marchand, P., Tran, V., Salmon, M., Li, G. and Grenié, M. 2020, *taxize: Taxonomic information from around the web*. R package version 0.9.95.
URL: *https://github.com/ropensci/taxize*

Consortium, I. B. G. S. et al. 2012, 'A physical, genetic and functional sequence assembly of the barley genome', *Nature* **491**(7426), 711–716.

Consortium, I. W. G. S. et al. 2014, 'A chromosome-based draft sequence of the hexaploid bread wheat (triticum aestivum) genome', *Science* **345**(6194).

Conway, J. R., Lex, A. and Gehlenborg, N. 2017, 'Upsetr: an r package for the visualization of intersecting sets and their properties', *Bioinformatics* **33**(18), 2938–2940.

Cooper, L., Meier, A., Laporte, M.-A., Elser, J. L., Mungall, C., Sinn, B. T., Cavaliere, D., Carbon, S., Dunn, N. A., Smith, B. et al. 2018, 'The planteome database: an integrated resource for reference ontologies, plant genomics and phenomics', *Nucleic acids research* **46**(D1), D1168–D1180.

Diaz-Papkovich, A., Anderson-Trocmé, L. and Gravel, S. 2020, 'A review of umap in population genetics', *Journal of Human Genetics* pp. 1–7.

Digel, B., Pankin, A. and von Korff, M. 2015, 'Global transcriptome profiling of developing leaf and shoot apices reveals distinct genetic and environmental control of floral transition and inflorescence development in barley', *The Plant Cell* **27**(9), 2318–2334.

Druka, A., Muehlbauer, G., Druka, I., Caldo, R., Baumann, U., Rostoks, N., Schreiber, A., Wise, R., Close, T., Kleinhofs, A. et al. 2006, 'An atlas of gene expression from seed to seed through barley development', *Functional & integrative genomics* **6**(3), 202–211.

Dubcovsky, J. and Dvorak, J. 2007, 'Genome plasticity a key factor in the success of polyploid wheat under domestication', *Science* **316**(5833), 1862–1866.

Emms, D. M. and Kelly, S. 2015, 'Orthofinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy', *Genome biology* **16**(1), 157.

Ewels, P., Magnusson, M., Lundin, S. and Käller, M. 2016, 'Multiqc: summarize analysis results for multiple tools and samples in a single report', *Bioinformatics* **32**(19), 3047–3048.

Fantini, D. 2018, 'Package "easypubmed"'.
URL: *https://www.data-pulse.com/dev_site/easypubmed/*

Feinerer, I. and Hornik, K. 2019, *tm: Text Mining Package*. R package version 0.7-7.
URL: *https://CRAN.R-project.org/package=tm*

Feng, N., Song, G., Guan, J., Chen, K., Jia, M., Huang, D., Wu, J., Zhang, L., Kong, X., Geng, S. et al. 2017, 'Transcriptome profiling of wheat inflorescence development from spikelet initiation to floral patterning identified stage-specific regulatory genes', *Plant Physiology* **174**(3), 1779–1794.

Fettell, N., Bowden, P., McNee, T. and Border, N. 2010, 'Barley growth development'.
URL: *https://www.dpi.nsw.gov.au/__data/assets/pdf_file/0003/516180/Procrop-barley-growth-and-development.pdf*

Fu, D., Xiao, M., Hayward, A., Jiang, G., Zhu, L., Zhou, Q., Li, J. and Zhang, M. 2015, 'What is crop heterosis: new insights into an old topic', *Journal of applied genetics* **56**(1), 1–13.

Gaut, B. S. 2002, 'Evolutionary dynamics of grass genomes', *New phytologist* **154**(1), 15–28.

Kassambara, A. and Mundt, F. 2020, 'Package "factoextra"'.
URL: *https://rpkgs.datanovia.com/factoextra/index.html*

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J. B., Grout, J., Corlay, S. et al. 2016, Jupyter notebooks-a publishing format for reproducible computational workflows., *in* 'ELPUB', pp. 87–90.

Konopka, T. 2020, 'Package "umap"'.
URL: *https://CRAN.R-project.org/package=umap*

Lancashire, P. D., Bleiholder, H., Boom, T. v. d., Langelüddeke, P., Stauss, R., WEBER, E. and Witzenberger, A. 1991, 'A uniform decimal code for growth stages of crops and weeds', *Annals of applied Biology* **119**(3), 561–601.

Large, E. C. et al. 1954, 'Growth stages in cereals. illustration of the feekes scale.', *Plant pathology* **3**, 128–129.

Larsen, J. and Smith, P. 2010, 'A field guide to cereal staging'.

Lê, S., Josse, J., Husson, F. et al. 2008, 'Factominer: an r package for multivariate analysis', *Journal of statistical software* **25**(1), 1–18.

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. and Pfister, H. 2014, 'Upset: visualization of intersecting sets', *IEEE transactions on visualization and computer graphics* **20**(12), 1983–1992.

Li, Y., Fu, X., Zhao, M., Zhang, W., Li, B., An, D., Li, J., Zhang, A., Liu, R. and Liu, X. 2018, 'A genome-wide view of transcriptome dynamics during early spike development in bread wheat', *Scientific reports* **8**(1), 1–16.

Liu, Z., Li, S., Li, W., Liu, Q., Zhang, L. and Song, X. 2020, 'Comparative transcriptome analysis indicates that a core transcriptional network mediates isonuclear alloplasmic male sterility in wheat (triticum aestivum l.)', *BMC plant biology* **20**(1), 1–19.

Luo, M.-C., Gu, Y. Q., Puiu, D., Wang, H., Twardziok, S. O., Deal, K. R., Huo, N., Zhu, T., Wang, L., Wang, Y. et al. 2017, 'Genome sequence of the progenitor of the wheat d genome aegilops tauschii', *Nature* **551**(7681), 498–502.

Ma, L., Ma, S.-W., Deng, Q., Yuan, Y., Wei, Z., Jia, H. and Ma, Z. 2018, 'Identification of wheat inflorescence development-related genes using a comparative transcriptomics approach', *International journal of genomics* **2018**.

Maglott, D., Ostell, J., Pruitt, K. D. and Tatusova, T. 2011, 'Entrez gene: gene-centered information at ncbi.', *Nucleic acids research* **39**(Database issue), D52–7.

Manickavelu, A., Kambara, K., Mishina, K. and Koba, T. 2007, 'An efficient method for purifying high quality rna from wheat pistils', *Colloids and Surfaces B: Biointerfaces* **54**(2), 254–258.

Marosi, V. B. 2020, 'GitHub Repository for the master thesis of "Comparative transcriptomics of flower development in barley and wheat"', `https://github.com/vanda-marosi/master_thesis`.

Mayer, K. F., Martis, M., Hedley, P. E., Šimková, H., Liu, H., Morris, J. A., Steuernagel, B., Taudien, S., Roessner, S., Gundlach, H. et al. 2011, 'Unlocking the barley genome by chromosomal and comparative genomics', *The Plant Cell* **23**(4), 1249–1263.

McInnes, L., Healy, J. and Melville, J. 2018, 'Umap: Uniform manifold approximation and projection for dimension reduction', *arXiv preprint arXiv:1802.03426* .

Mochida, K. and Shinozaki, K. 2013, 'Unlocking triticeae genomics to sustainably feed the future', *Plant and Cell Physiology* **54**(12), 1931–1950.

Monat, C., Padmarasu, S., Lux, T., Wicker, T., Gundlach, H., Himmelbach, A., Ens, J., Li, C., Muehlbauer, G. J., Schulman, A. H. et al. 2019, 'Tritex: chromosome-scale sequence assembly of triticeae genomes with open-source tools', *Genome biology* **20**(1), 284.

NCBI 2020*a*, 'Database', `https://www.ncbi.nlm.nih.gov/sra`. Accessed: 2020-05-01.

NCBI 2020*b*, 'SRA Toolkit', `https://ncbi.github.io/sra-tools/`. Accessed: 2020-05-01.

NCBI 2020*c*, 'Triticeae tribe', `https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=147389`. Accessed: 2020-04-01.

Pearson, K. 1901, 'Liii. on lines and planes of closest fit to systems of points in space', *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**(11), 559–572.

Pimentel, H., Bray, N. L., Puente, S., Melsted, P. and Pachter, L. 2017, 'Differential analysis of rna-seq incorporating quantification uncertainty', *Nature methods* **14**(7), 687.

Pont, C., Leroy, T., Seidel, M., Tondelli, A., Duchemin, W., Armisen, D., Lang, D., Bustos-Korts, D., Goué, N., Balfourier, F. et al. 2019, 'Tracing the ancestry of modern bread wheats', *Nature genetics* **51**(5), 905–911.

R Core Team 2019, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Ramírez-González, R., Borrill, P., Lang, D., Harrington, S., Brinton, J., Venturini, L., Davey, M., Jacobs, J., Van Ex, F., Pasha, A. et al. 2018, 'The transcriptional landscape of polyploid wheat', *Science* **361**(6403).

Rutley, N. and Twell, D. 2015, 'A decade of pollen transcriptomics', *Plant reproduction* **28**(2), 73–89.

Saisho, D. and Takeda, K. 2011, 'Barley: emergence as a new research material of crop science'.

Sayers, E. W., Agarwala, R., Bolton, E. E., Brister, J., Canese, K., Clark, K., Connor, R., Fiorini, N., Funk, K., Hefferon, T., Holmes, J., Kim, S., Kimchi, A., Kitts, P. A., Lathrop, S., Lu, Z., Madden, T. L., Marchler-Bauer, A., Phan, L., Schneider, V. A., Schoch, C. L., Pruitt, K. D. and Ostell, J. 2019, 'Database resources of the National Center for Biotechnology Information', *Nucleic Acids Research* **47**(Database issue), D23–D28.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6323993/*

Schreiber, A. W., Sutton, T., Caldo, R. A., Kalashyan, E., Lovell, B., Mayo, G., Muehlbauer, G. J., Druka, A., Waugh, R., Wise, R. P. et al. 2009, 'Comparative transcriptomics in the triticeae', *BMC genomics* **10**(1), 1–17.

Soneson, C., Love, M. I. and Robinson, M. D. 2015, 'Differential analyses for rna-seq: transcript-level estimates improve gene-level inferences', *F1000Research* **4**.

Sonnhammer, E. L. and Koonin, E. V. 2002, 'Orthology, paralogy and proposed classification for paralog subtypes', *TRENDS in Genetics* **18**(12), 619–620.

Uauy, C. 2017, 'Wheat genomics comes of age', *Current opinion in plant biology* **36**, 142–148.

Waddington, S., Cartwright, P. and Wall, P. 1983, 'A quantitative scale of spike initial and pistil development in barley and wheat', *Annals of Botany* **51**(1), 119–130.

Wang, Y. and Song, F. 2017, 'Gsa: genome sequence archive', *Genomics, proteomics & bioinformatics* **15**(1), 14–18.

Wang, Z., Li, J. and Chen, S. 2017, 'Poaceae-specific ms1 encodes a phospholipid-binding protein for male fertility in bread wheat', *Proceedings of the National Academy of Sciences* **114**(47), 12614–12619.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K. and Yutani, H. 2019, 'Welcome to the tidyverse', *Journal of Open Source Software* **4**(43), 1686.

Winter, D. J. 2017, 'rentrez: an r package for the ncbi eutils api', *The R Journal* **9**, 520–526.

Wu, T. D. and Watanabe, C. K. 2005, 'Gmap: a genomic mapping and alignment program for mrna and est sequences', *Bioinformatics* **21**(9), 1859–1875.

Yang, X., Ye, J., Zhang, L. and Song, X. 2019, 'Blocked synthesis of sporopollenin and jasmonic acid leads to pollen wall defects and anther indehiscence in genic male sterile wheat line 4110s at high temperatures', *Functional & Integrative Genomics* pp. 1–14.

Zadoks, J. C., Chang, T. T., Konzak, C. F. et al. 1974, 'A decimal code for the growth stages of cereals.', *Weed research* **14**(6), 415–421.

Supplementary Materials

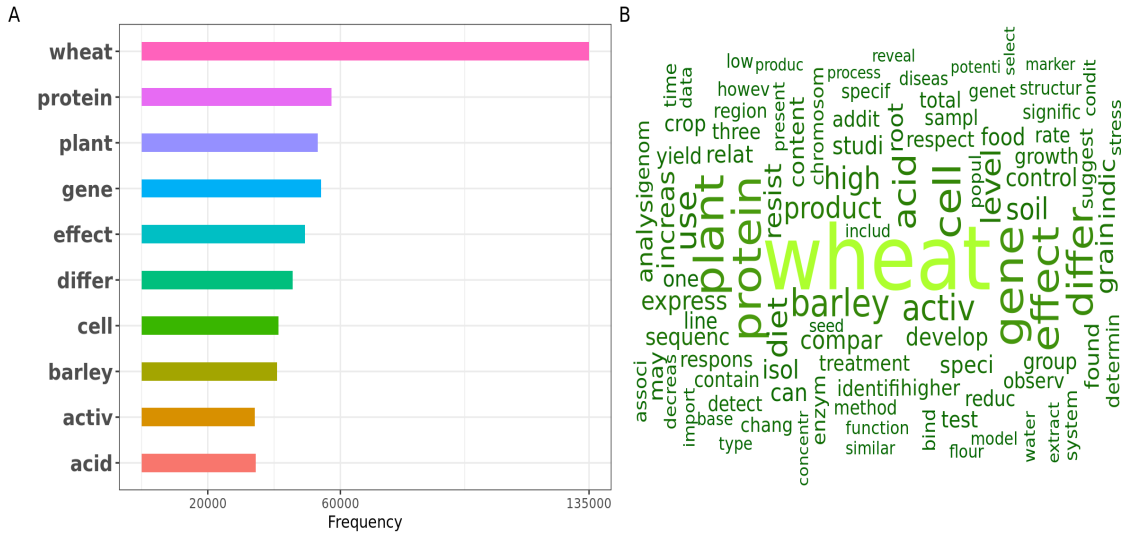

Figure S1: Visualization of the text-mining corpus in (A) a diagram of the ten most frequent words and (B) a word-cloud representing the 100 most frequent words.

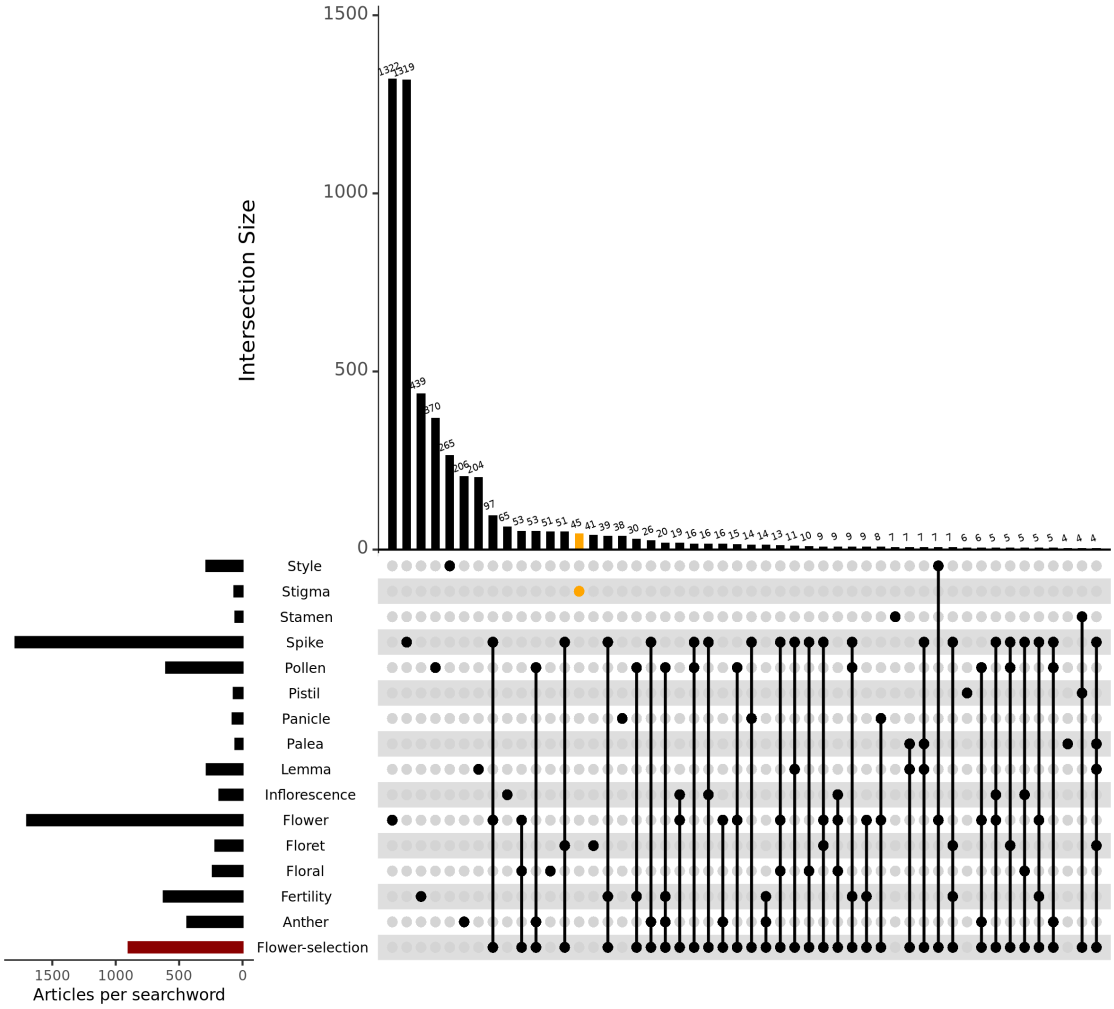

Figure S2: UpSet-plot of sub-sets in the "PubMed Query" text mining corpus, illustrating the intersections of elements across the 15 flower-development related keywords. Orange color represents the empty-intersection of the "Stigma" set with the "Flower-selection".

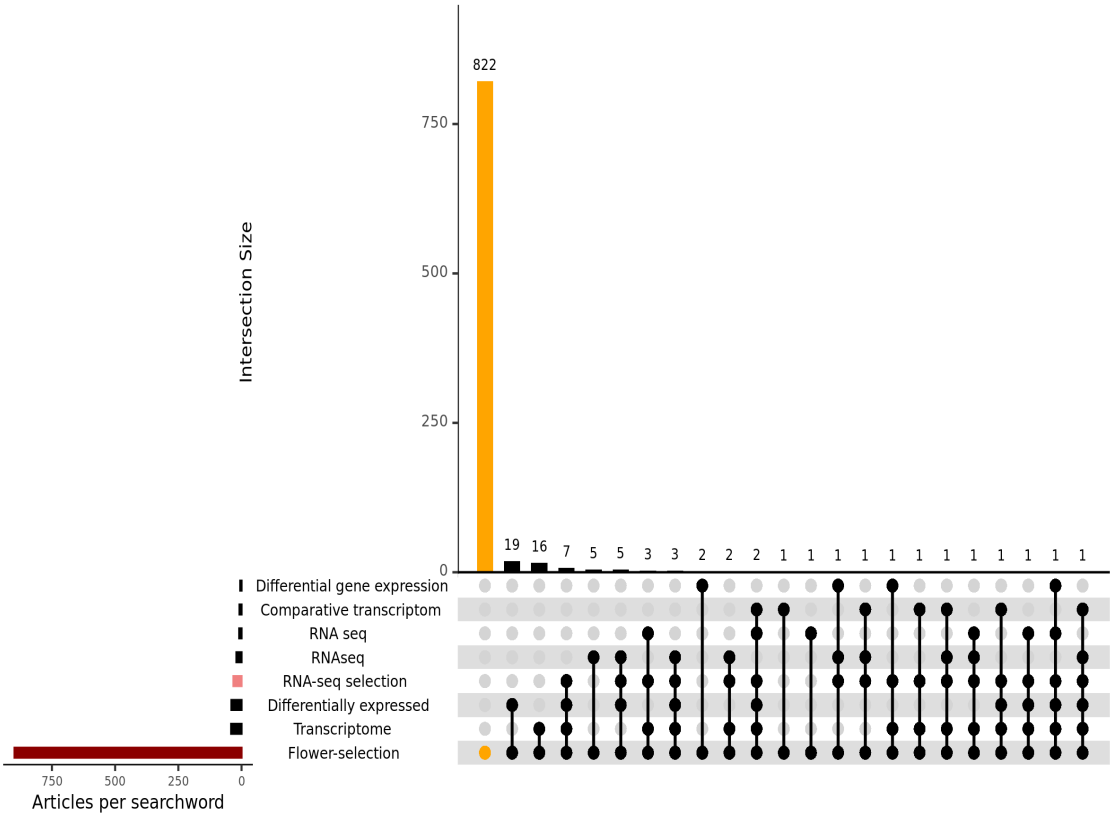

Figure S3: UpSet-plot of sub-sets in the "Flower-selection" corpus, illustrating the intersections of elements across the 6 transcriptomic-research related keywords. Orange color represents the empty-intersection of the "Flower-selection" with the "RNA-seq selection" set.

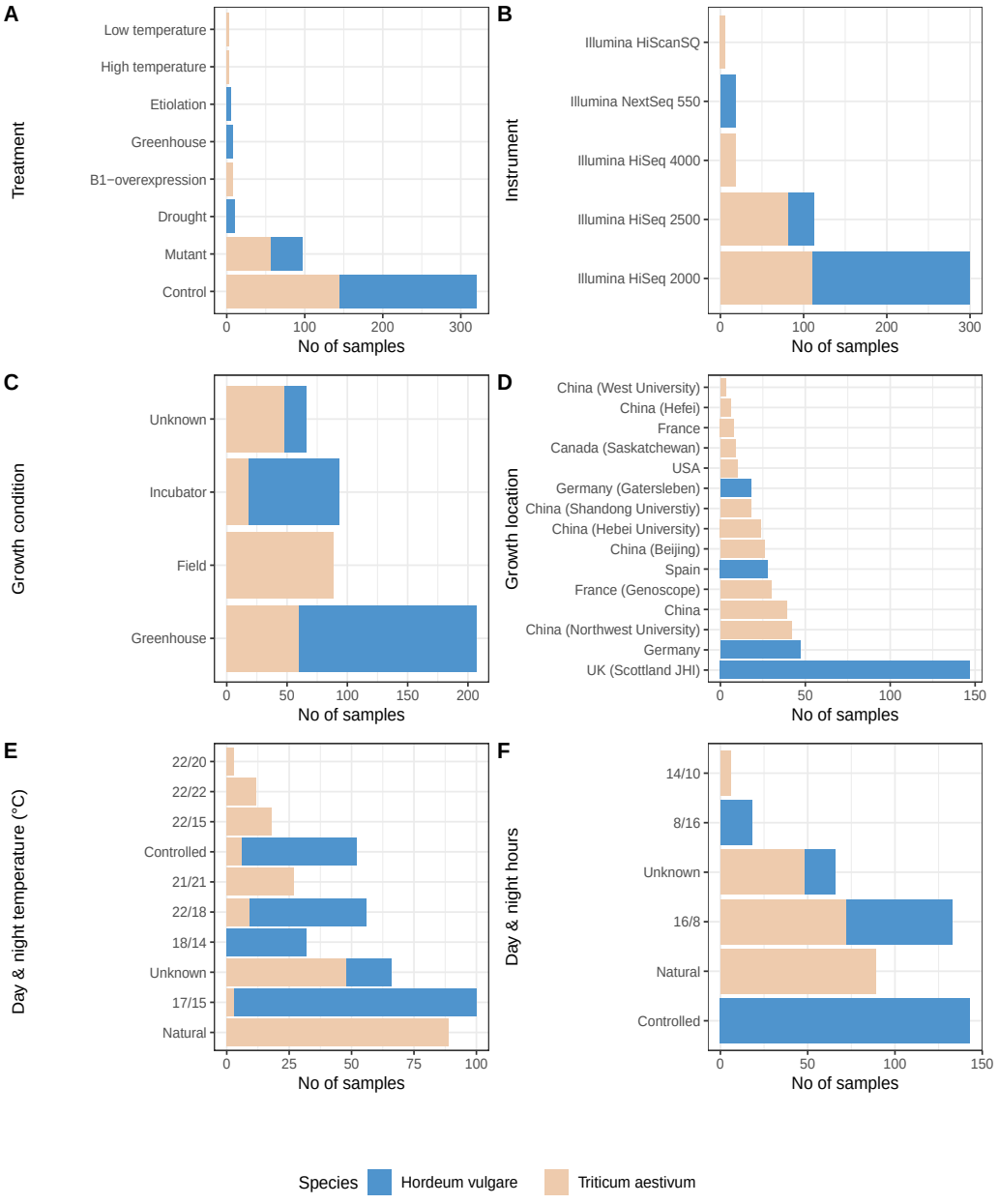

Figure S4: Distribution of barley and wheat samples across different technical variables.

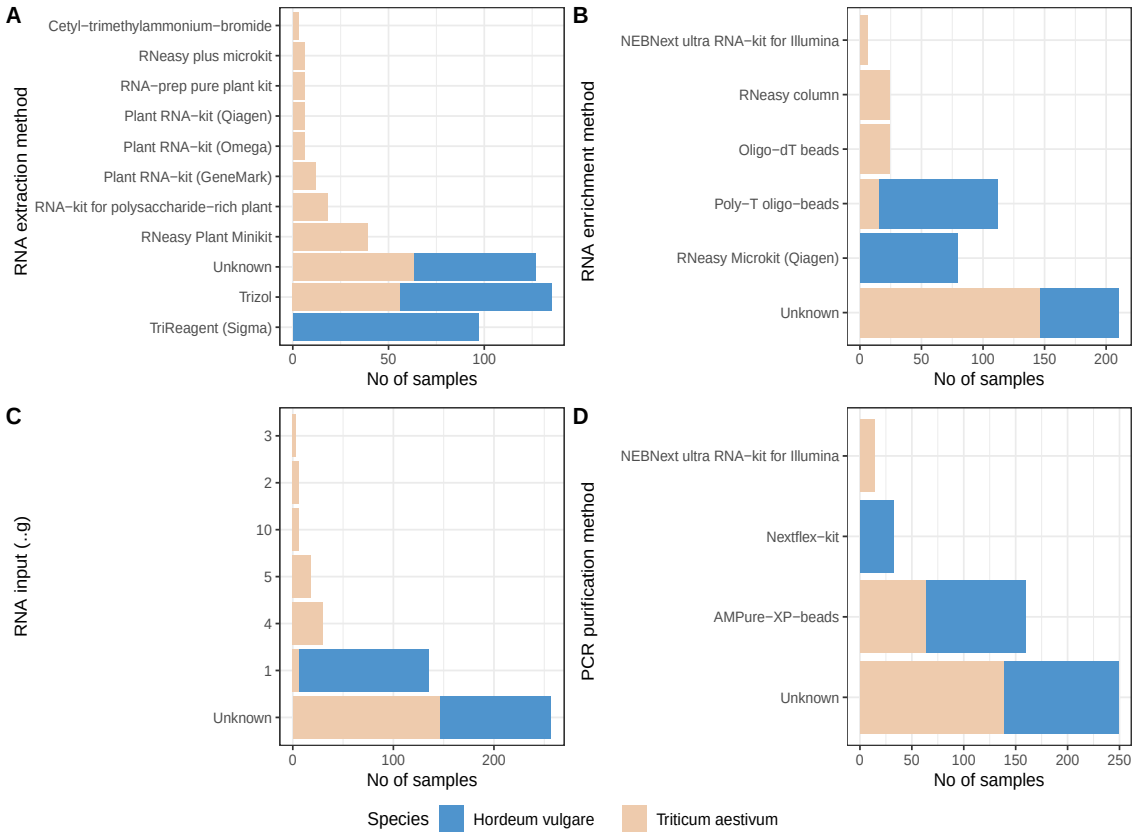

Figure S5: Distribution of barley and wheat samples across different categorical variables.

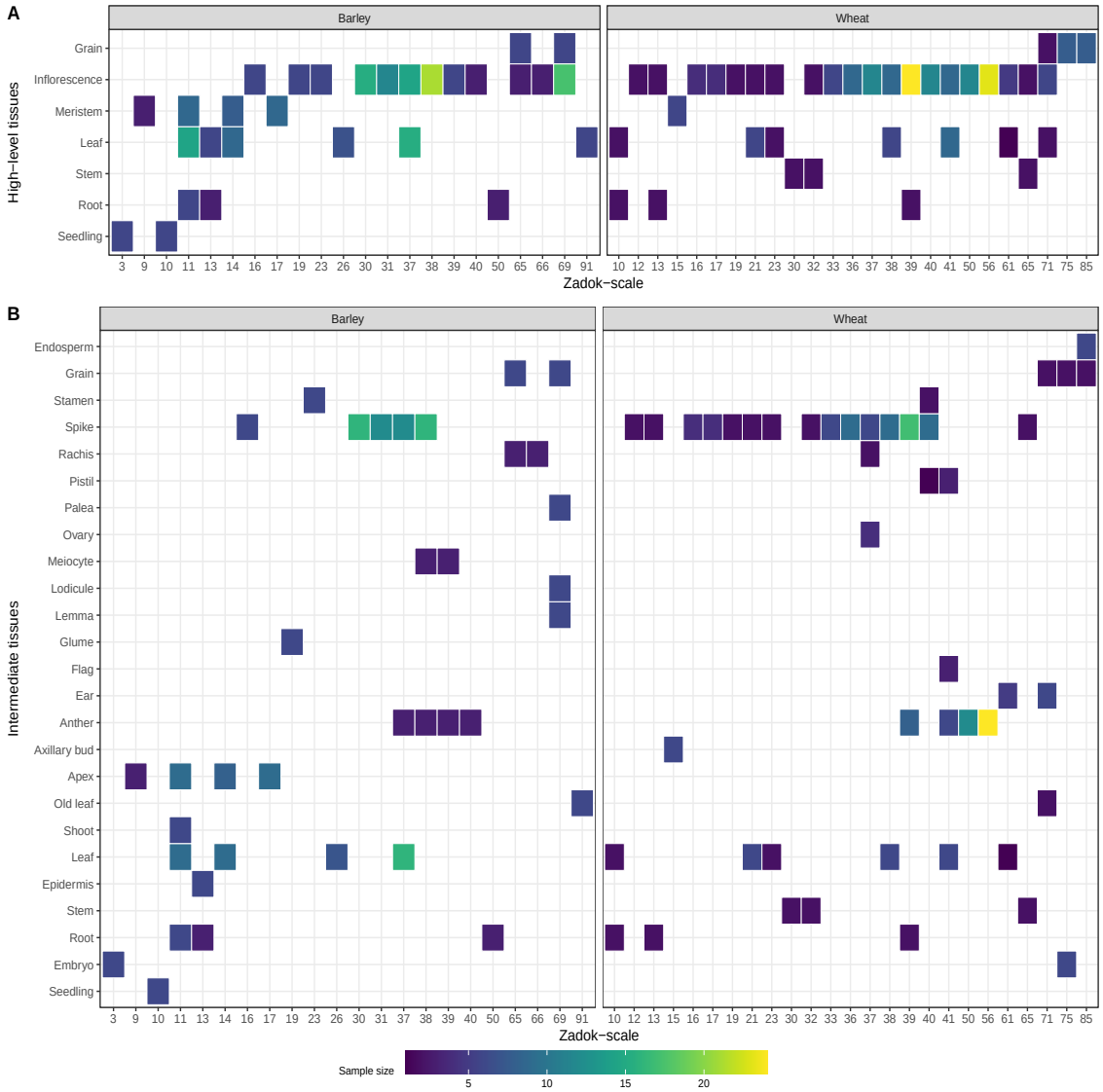

Figure S6: Distribution of barley and wheat (A) high-level tissue and (B) intermediate tissue categories across the Zadok-scale.

Zadok-scale	High-level age category	Intermediate age category	Extended description of reproductive phase (Z13-69)*			
0-10	Seedling	Germination	-			
11-19	Vegetative	Leaf development	13-16	Floral initiation	14-16	Double ridge stage
					17-18	Triple mound stage
20-30		Tillering	17-29	Spikelet initiation	19-20	Glume primordium stage
					21-22	Lemma primordium stage
					23-24	Stamen primordium stage
			30	"Head at 1 cm"	30	Awn primordium stage
31-36	Reproductive	Stem elongation	31-36	Stem elongation		
37-40		Floret development	37-40	Floret development	38	White anther stage
					39	Green anther stage
					40	Yellow anther stage
41-49		Booting	41-49	Booting		
50-60		Heading	50-60	Heading	50	Anther later uninucleate stage
					56	Anther binucleate stage
					58	Anther trinucleate stage
61-69		Anthesis	61-69	Anthesis		
70-79	Grain development	Seed development	-			
80-89		Dough development	-			
90-99	Senescence	Senescence	-			

Table S1: Developmental stage conversion among Zadok-scale, High-level age and Intermediate age categories and their relations to the extended reproductive phase.(*Numbers indicate the Zadok-scale.)

39

Intermediate tissue	High level tissue	Shortened definition from PO_plant_anatomy*	PO_plant_anatomy*
endosperm	grain	Maximal portion of nutritive plant tissue in a seed.	PO:0009089
grain	grain	Fruit which develops from a gynoecium, comprises a dry exocarp, mesocarp, and endocarp fused to a seed coat.	PO:0030104
anther	inflorescence	A collective plant organ structure that is the pollen-bearing part of a stamen.	PO:0009066
ear	inflorescence	Inflorescence that is highly compacted and bears the ear spikelets on a lateral inflorescence axis.	PO:0020136
glume	inflorescence	One of a pair of inflorescence bracts that is part of a spikelet and subtends the two florets.	PO:0009039
lemma	inflorescence	The lower, usually larger, of the pair of flower bracts enclosing a spikelet floret.	PO:0009037
lodicule	inflorescence	A phyllome that is part of a grass floret and is one of two or three tiny scales or flaps of tissue outside the stamens.	PO:0009036
meiocyte	inflorescence	A native plant cell which is diploid (2n), and undergoes meiosis to produce four haploid (1n) plant spores.	PO:0006204
ovary	inflorescence	A plant structure that is the basal portion of a carpel or group of fused carpels and encloses the plant ovule(s).	PO:0009072
palea	inflorescence	The upper (distal), usually smaller, of the pair of flower bracts enclosing a spikelet floret.	PO:0009038
pistil	inflorescence	(Gynoecium used as PO term) Collective phyllome structure composed all of the carpels in a flower.	PO:0009062
rachis	inflorescence	(Inflorescence axis used as PO term) A shoot axis that is part of an inflorescence.	PO:0020122
spike	inflorescence	(Spikelet used as PO term) Reproductive shoot system, the higher order inflorescence axis of the grasses.	PO:0009051
stamen	inflorescence	A microsporophyll bearing one or more microsporangia.	PO:0009029
epidermis	leaf	Portion of plant tissue of epidermal cells that develops from the protoderm and covers the surface of a plant structure.	PO:0005679
flag	leaf	The last mature leaf before the inflorescence in a cereal crop plant.	PO:0020103
leaf	leaf	(Juvenile vascular leaf used as PO term)	PO:0006339
old leaf	leaf	(Adult vascular leaf used as PO term)	PO:0006340
shoot	leaf	(Shoot system used as PO) Collective organ, produces shoot-borne portions of meristem and structures arise from it.	PO:0009006
apex	meristem	The most distal part of a shoot system and has as parts a shoot apical meristem and the youngest primordia.	PO:0000037
axillary bud	meristem	A bud that develops from an axillary bud meristem.	PO:0004709
root	root	A plant axis that lacks shoot axis nodes and usually grows indeterminately.	PO:0009005
embryo	seedling	A whole plant that participates in the plant embryo stage.	PO:0009009
seedling	seedling	(Seedling cotyledon used as PO term) A cotyledon that is part of a whole plant in the seedling development stage.	PO:0025471
stem	stem	A shoot axis that is the primary axis of a plant.	PO:0009047

Table S2: Tissue conversion between Intermediate tissue and High-level tissue categories (*Definition source: `http://planteome.org/`).

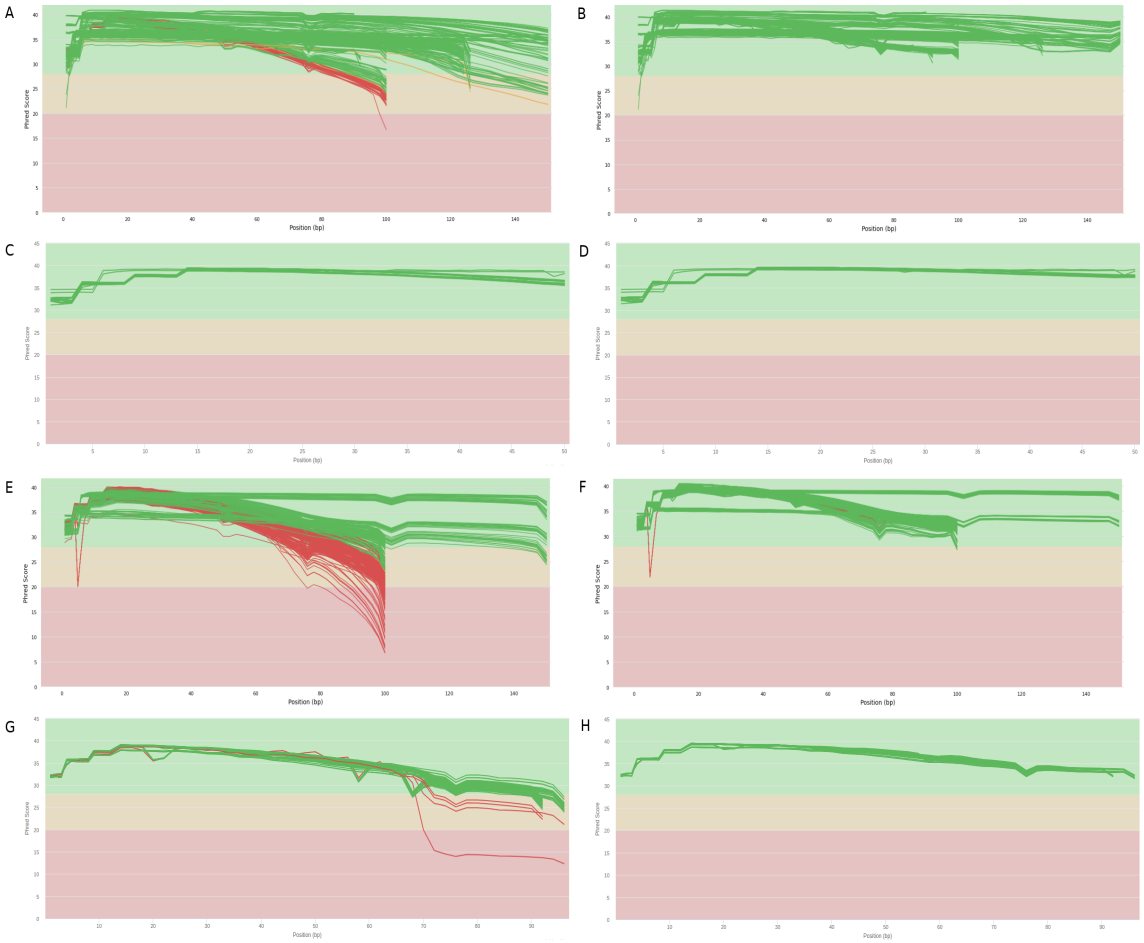

Figure S7: Sequence quality histograms of (A, B, C, D) wheat and (E, F, G, H) barley samples of (A, B, E, F) paired and (C, D, G, H) single reads (A, C, E, G) before and (B, D, F, H) after trimming of adapters and applying quality trimming. Histograms are displaying the mean quality value across each base position in the read.

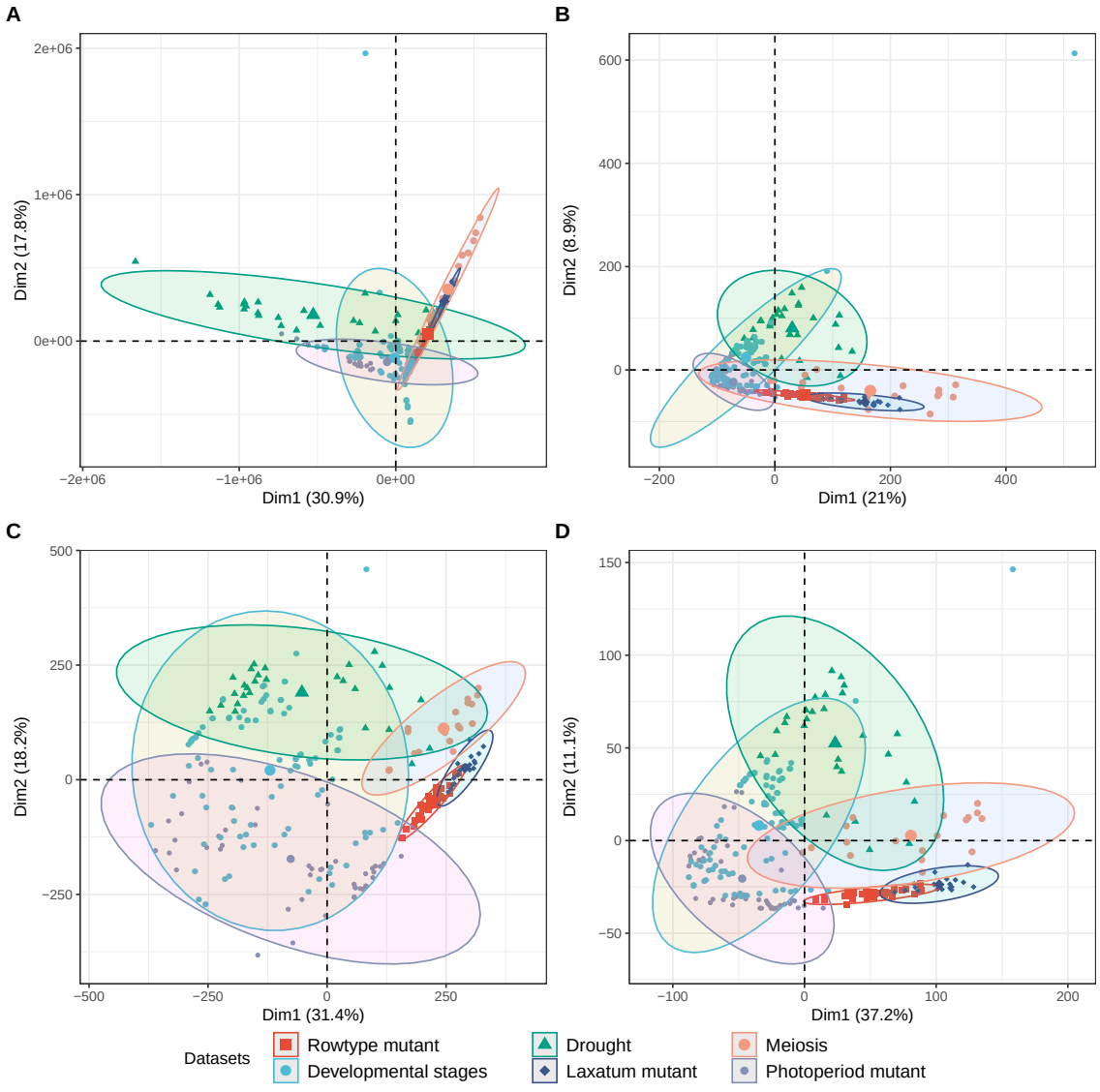

Figure S8: Different principal component analyses on barley datat-sets, (A) depicts PCA on not filtered TPM counts, (B) displays scaled and centered PCA on not filtered TPM counts, (C) shows PCA on log2-transformed and filtered TPM counts, and (D) represents scaled and centered PCA on log2-transformed and filtered TPM counts.

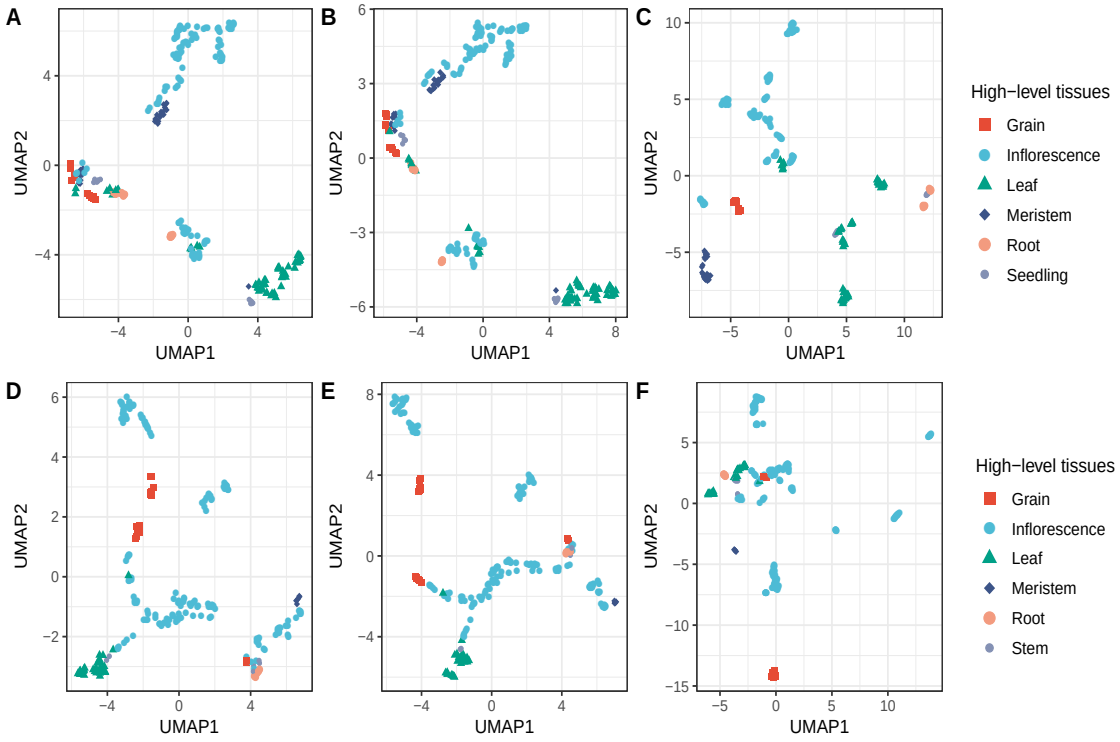

Figure S9: Different UMAP analyses on barley (A, B, C) and wheat (D, E, F) datat-sets, where (A, D) shows UMAP on not filtered TPM counts, (B, E) UMAP on filtered TPM counts, and (C, F) represents UMAP on log2-transformed and filtered TPM counts.

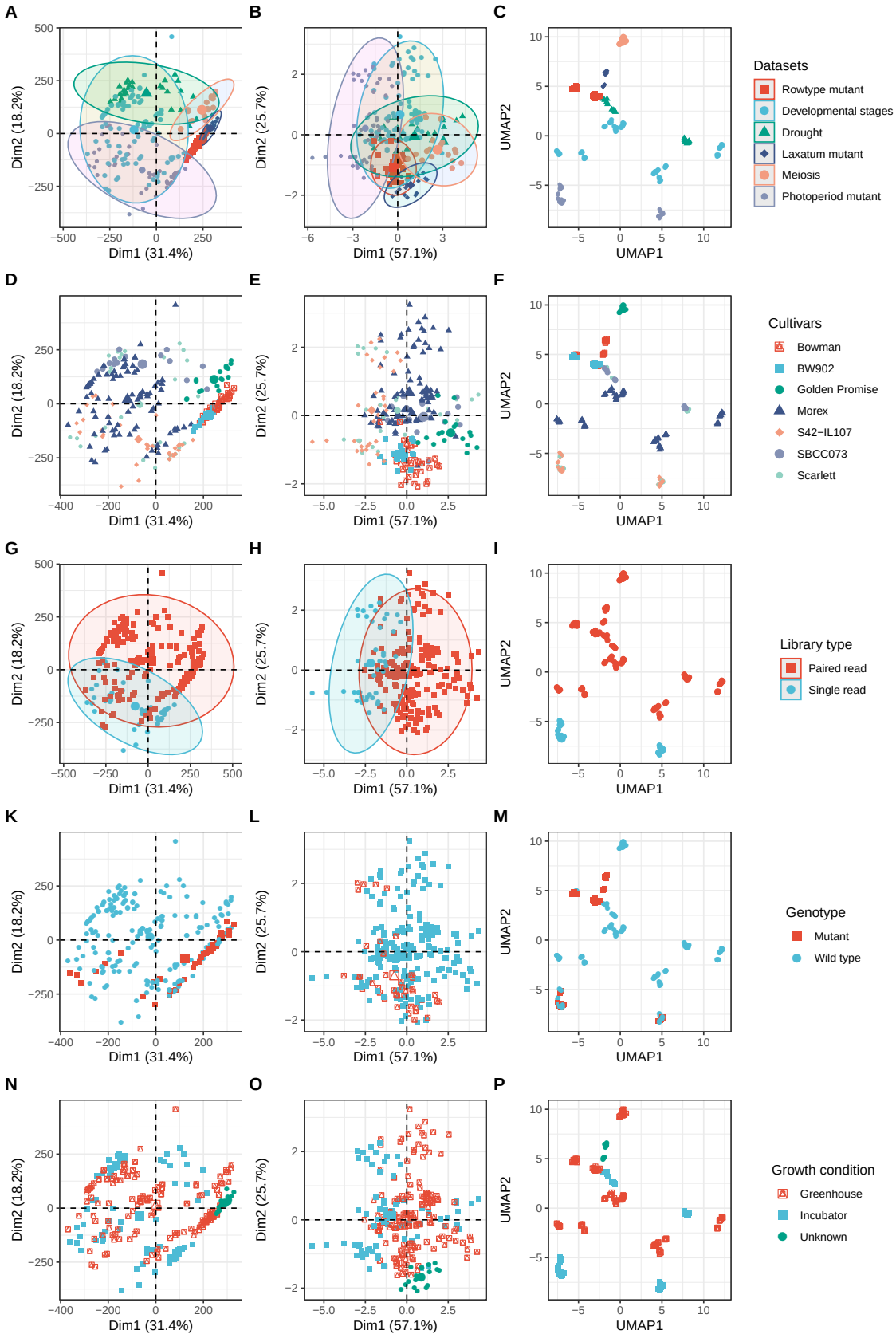

Figure S10: Exploratory analyses with different dimension reduction methods in the barley dataset on technical variables, representing (A, D, G, J, M) PCAs on log2TPM, (B, E, H, K, N) PCAs on the five different ELCs, and (C, F, I, L, O) UMAP on log2TPM.

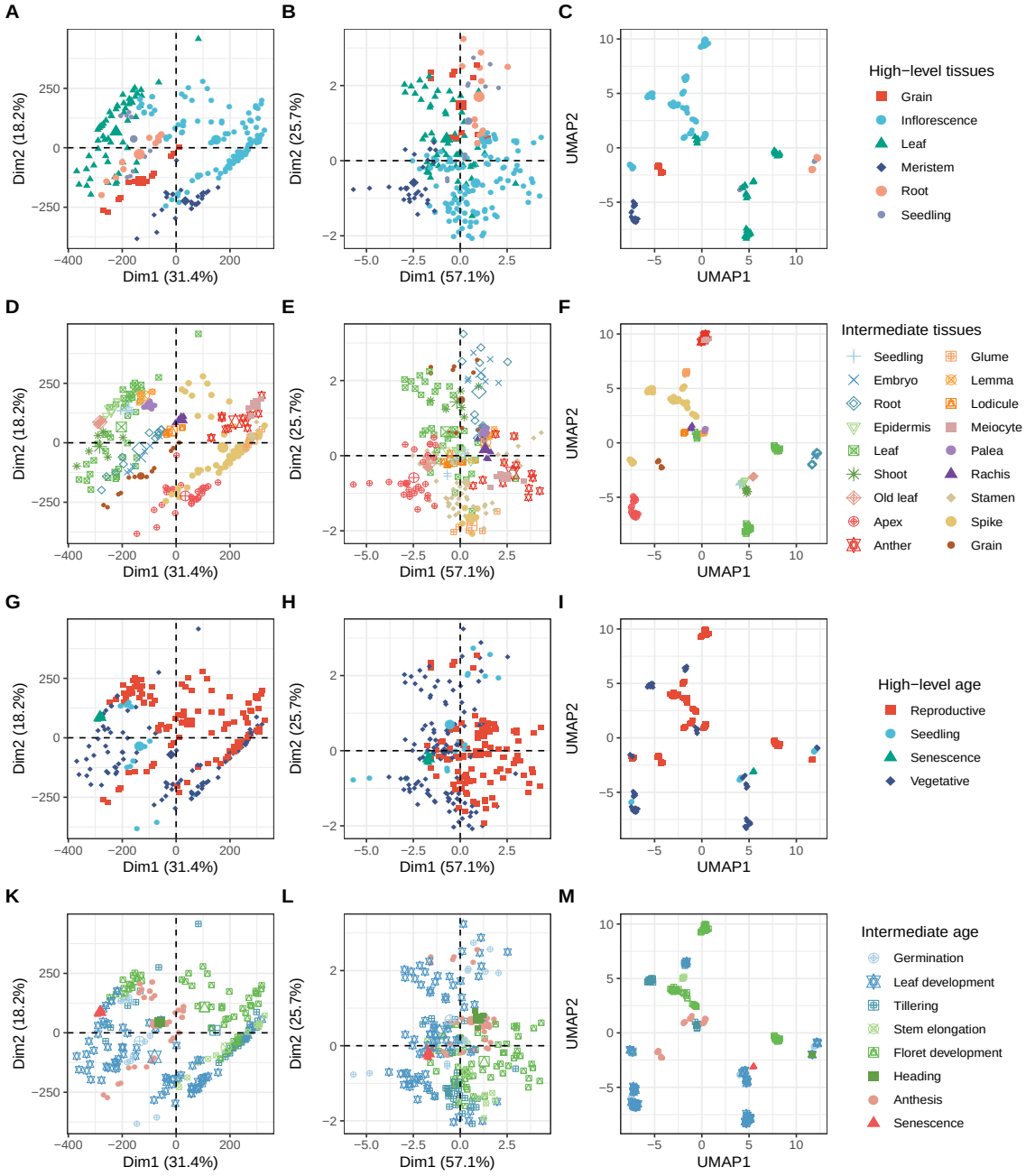

Figure S11: Exploratory analyses with different dimension reduction methods in the barley dataset on biological variables, representing (A, D, G, J, M) PCAs on log2TPM, (B, E, H, K, N) PCAs on the five different ELCs, and (C, F, I, L, O) UMAP on log2TPM.

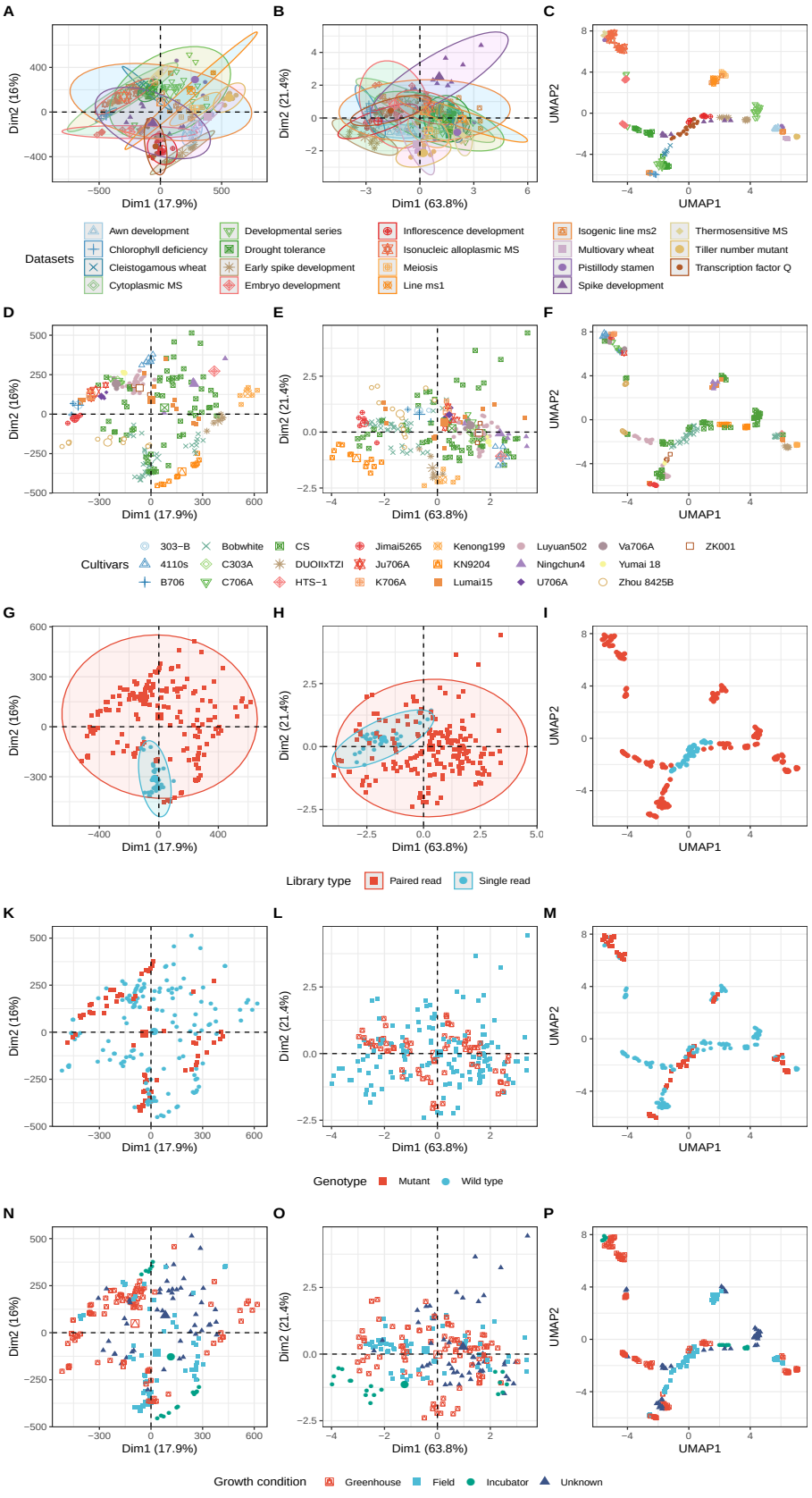

Figure S12: Exploratory analyses with different dimension reduction methods in the wheat dataset on technical variables, representing (A, D, G, J, M) PCAs on log2TPM, (B, E, H, K, N) PCAs on the five different ELCs, and (C, F, I, L, O) UMAP on raw-filtered counts.

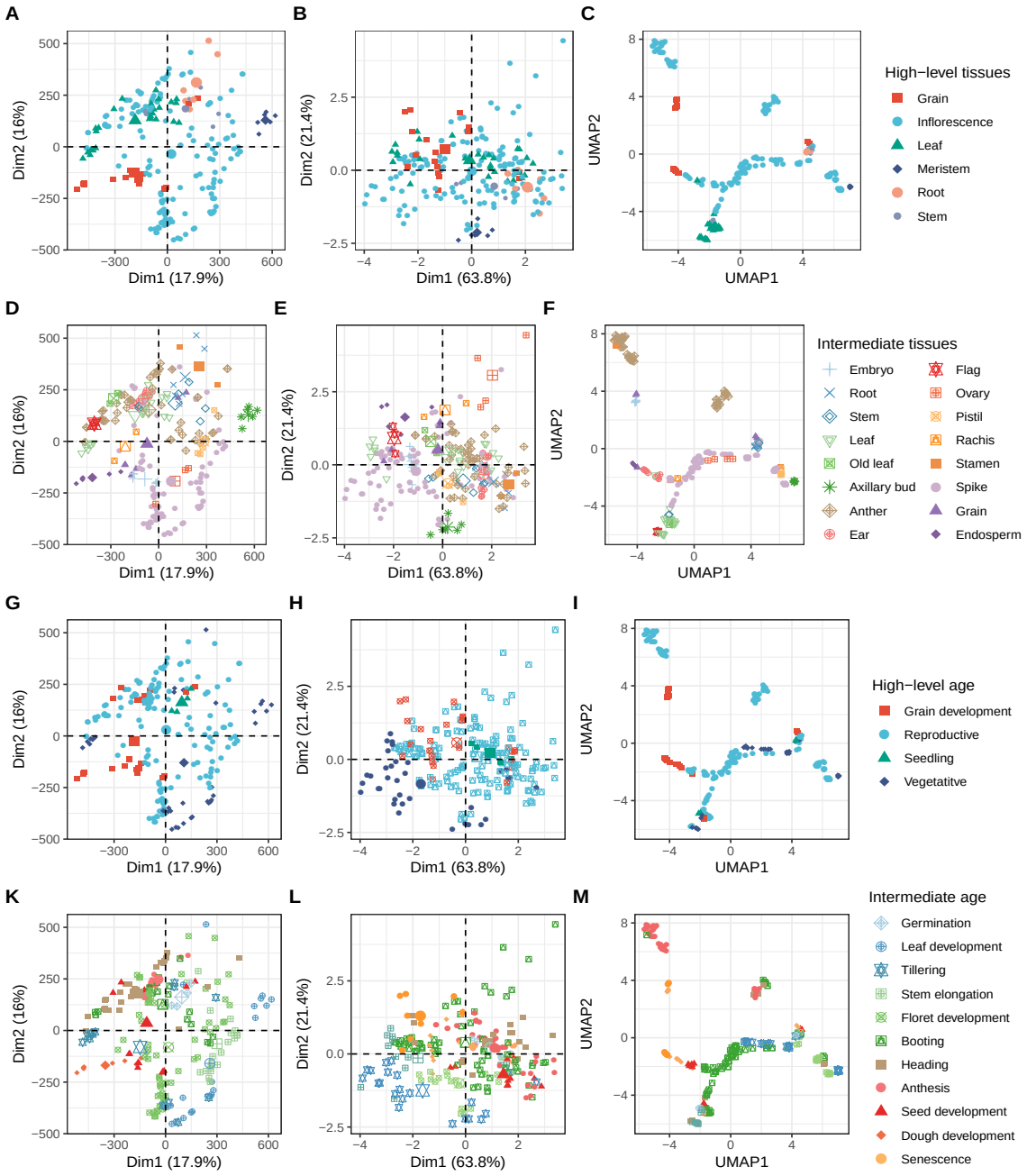

Figure S13: Exploratory analyses with different dimension reduction methods in the wheat dataset on biological variables, representing (A, D, G, J, M) PCAs on log2TPM, (B, E, H, K, N) PCAs on the five different ELCs, and (C, F, I, L, O) UMAP on raw-filtered counts.

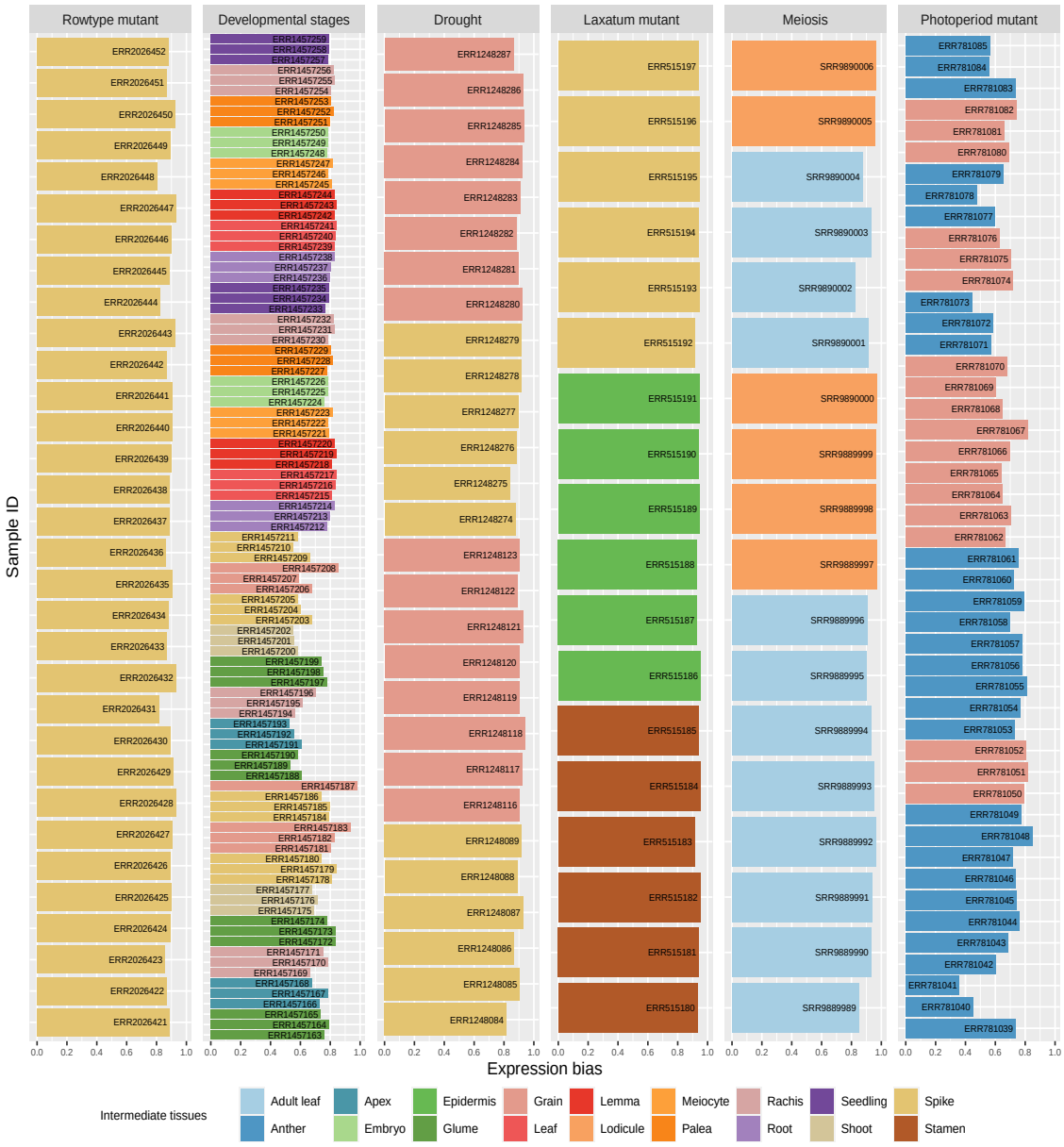

Figure S14: ELC analysis representing expression bias of the E4 category in the barley dataset.

Figure S15: ELC analysis representing expression bias of the E4 category in the wheat dataset.

Erklärung zur Bachelorarbeit/Masterarbeit

Hiermit versichere ich, dass die vorliegende Arbeit von mir selbstständig verfasst wurde und dass keine anderen als die angegebenen Quellen und Hilfsmittel benutzt wurden.
Diese Erklärung erstreckt sich auch auf in der Arbeit enthaltene Graphiken, Zeichnungen, Kartenskizzen und bildliche Darstellungen.

Bachelor's/Master’s thesis statement of originality

I hereby confirm that I have written the accompanying thesis by myself, without contributions from any sources other than those cited in the text and acknowledgements.
This applies also to all graphics, drawings, maps and images included in the thesis.

……	……..
Ort und Datum	Unterschrift
Place and date	Signature