Title: Alternative splicing enhances the transcriptome complexity in desiccating seeds.

Running title: Alternative splicing in seed desiccation

Arunkumar Srinivasan[1,2], José M. Jiménez-Gómez[1,3], Fabio Fornara[4], Wim J. J. Soppe[1], Vittoria Brambilla[1,4] *.

[1] Department of Plant Breeding and Genetics, Max Planck Institute for Plant Breeding Research, Cologne, Germany

[2]??? ARUN PRESENT ADDRESS

[3] Institut Jean-Pierre Bourgin, Institut National de la Recherche Agronomique, Versailles, France

[4] Università degli Studi di Milano, Dipartimento di Bioscienze, Milano, Italy

*Correspondence: Dipartimento di Bioscienze, via Celoria 26, 20133 Milano Italy vittoria.brambilla@unimi.it

## ABSTRACT

Commento [V1]: No more than 200 words.

Before being dispersed in the environment, a mature seed needs to be desiccated. Correct desiccation is essential for seed survival and it is achieved during the last phase of seed maturation. Some key seed maturation genes have been reported to be regulated by alternative splicing (AS). However, so far AS was described only for single genes and no comprehensive analyses of AS at different stages of seed maturation have been performed. We investigated gene expression and AS in Arabidopsis seeds, prior and after desiccation, at a global transcriptional level. We developed bioinformatics tools to identify differentially spliced regions within genes. Our data suggest the importance and also the peculiarity of AS during seed desiccation. We found that 34% of the genes expressed prior or after desiccation undergo AS and the transcript variants that we found in this tissues are predominantly novel. Comparing difference in gene expression and AS, we found that 6% of the transcripts were not transcriptionally regulated but only modified by AS. Among the AS genes some seed master regulators could be found. We suggest that AS should be more routinely taken into account once analyzing transcriptomic data and looking for potentially important regulators of a specific process.

**Keywords:**

Seed desiccation, alternative splicing, Arabidopsis

## INTRODUCTION

Seeds have a great ecological and economic importance. As plants' offspring, are deputed to ensure the species survival and ecological success. As the main source of the calories required for human nutrition worldwide, they have a primary economical value. Seed quality depends on seed development and maturation. Seed longevity depends on its dehydrated state. It is of great importance to understand how seed maturation and desiccation are regulated.

Seed development starts with ovule fertilization and ends when the seed is ready for dispersal. In *Arabidopsis thaliana*, it takes 20 days and it can be divided into three major phases (Baud et al. 2002) that are: 1) embryo morphogenesis (day 1 to 10), 2) early seed maturation (day 11 to 14) and 3) late seed maturation, seed ripening and desiccation (day 14 to 20). During the first 10 days after pollination (DAP) the embryo develops and the seed grows. Once morphogenesis is completed, the seed has reached its final shape and size. Nevertheless some crucial processes still need to take place before a viable seed can be dispersed. During early seed maturation from 11 to 13 DAP chlorophyll is degraded and storage compounds as fatty acids and proteins are accumulated. Later on, from 14 to 20 DAP the seed progressively dehydrates, the seed is fully ripened and dormancy is imposed. Freshly harvested 20 DAP seeds have less than 10% water content, that is far less than all other plant tissues (van Zanten et al. 2011). To survive this extreme dehydrated status the embryo cells are filled with storage protein vacuoles and oil bodies (Baud et al. 2002; Fait et al. 2006; Goldberg et al. 1994; Jenik et al. 2007). During seed maturation, embryonic nuclei chromatin is also increasingly compacted and nuclei size decreases (van Zanten et al. 2011). For these reasons dry seed cells are physiologically dissimilar from most plant tissues, but they have common features to vegetative tissues of desiccation tolerant plants. Several studies have elucidated a framework of seed development molecular regulators. Four transcription factors, *ABSCISIC ACID INSENSITIVE 3 (ABI3), FUSCA3 (FUS3), LEAFY COTYLEDON 1 (LEC1)* and *LEC2,* play a major role in seed maturation. *LEC1* and *LEC2* are expressed since early embryo morphogenesis followed by *FUS3* and finally by *ABI3,* which is primarily involved in seed maturation. *ABI3*, *FUS3* and *LEC2* belong to the AFL (ABI3-FUS3-LEC2) subgroup of the B3 transcription factor family (Holdsworth et al. 2008; Suzuki & McCarty 2008; Swaminathan et al. 2008). The B3 domain of AFL transcription factors recognize the RY motif of its target DNA, that has been shown to be involved in seed specific expression (Reidt et al. 2000; Mönke et al. 2004). Many direct, indirect and also common targets of LEC1, ABI3, FUS3 and LEC2 have been identified so far (Wang & Perry 2013). One common function of AFL genes during seed maturation is the control of storage compounds accumulation. Phytohormones, especially abscisic acid (ABA) and gibberellins (GA), have an important role in the seed development regulatory network (Wang & Perry 2013; Santos-Mendoza et al. 2008; Gazzarrini et al. 2004). Hormonal responses are in part mediated by the seed development master regulators. ABI3 is required, together with the bZIP transcription factor ABI5 and in parallel to ABA, to induce seed dormancy in an ABA-dependent manner (Lopez-Molina et al. 2002). FUS3 is required to induce ABA and repress GA biosynthesis by direct interaction with the promoter of GA biosynthetic genes (Tsai & Gazzarrini 2012). Due to their vital role in seed maturation, *ABI3, LEC1* and *FUS3* homologues have also been characterized in distantly related species as pea, rice, maize, barley and Selaginella (Miyoshi et al. 2002; Nakagawa et al. 1996; Kirkbride et al. 2013; Moreno-Risueno et al. 2008; Shen et al. 2010; Gagete et al. 2009). In Arabidopsis, tomato, pea, rice and wheat, *ABI3/VP1* genes have been shown to be regulated by AS (Gao et al. 2013; Gagete et al. 2009; Sugliani et al. 2010; Fan et al. 2007; McKibbin et al. 2002). In Arabidopsis *ABI3* AS is developmentally regulated. Two transcript variants have been detected: the full-length coding *ABI3-α* and the *ABI3-β* transcript that contains a cryptic intron within its first exon causing a frame-shift shortly after. At the beginning of seed maturation only *ABI3-α* transcript could be detected, but from 16 DAP *ABI3-β* was also detectable and became the most abundant protein AS isoform in mature seeds (Sugliani et al. 2010). Similarly in tomato, two *SlABI3* transcripts that encode for a full-length and a truncated protein are accumulated. Also in this case, the truncated SlABI3 protein results from the AS of a cryptic intron within the first exon of *SlABI3*. Only the full length SlABI3 protein is able to activate its target genes, while the truncated SlABI3 has possibly a regulatory function. Another interesting example comes from rice, where the AS of *ABI3* homologous *OsVIVIPAROUS 1 (OsVP1)* has been phenotypically linked to seed dormancy and pre-harvest-sprouting. Similar to Arabidopsis, OsVP1 can interact with OsABI5 in rice. Two OsABI5 distinct proteins derived from alternative spliced transcripts have different binding affinity to OsVP1 and transactivation of their targets (Zou et al. 2007).

Another example of AS affecting seed maturation and dormancy comes from the work of Penfield et al., 2009. They show that two splice variants of *PHYTOCROME INTERACTING FACTOR 6 (PIF6*) originate from an out of frame exon skipping AS event that creates a premature stop codon and a protein isoform where the DNA binding domain is abolished (Penfield et al. 2010). Overexpression of the PIF6 AS variant but not the full length coding sequence reduces seed dormancy.

These observations indicate the importance of AS in the regulation of seed maturation in both dicots and monocots. Transcriptomic analyses through RNA-seq and Tiling Arrays have been used to study the transcriptome complexity, also taking into account AS (Leviatan et al. 2013; Yoshimura et al. 2011). Thanks to these data, an increasing number of genes are predicted to be alternatively spliced. In plants, AS is currently estimated to affect about 61% of intron containing genes in Arabidopsis (Marquez et al. 2012) and 31% in rice (Filichkin et al. 2010). This number is very likely going to increase, as AS in different environmental conditions, developmental stages and tissues are investigated (Reddy et al. 2013; Naeem H. Syed et al. 2012; Loraine et al. 2013).

In spite of the importance of AS in seed maturation, this process has not been investigated at a transcriptomic level yet. This paper provides the first comprehensive description of AS during late seed maturation when desiccation is achieved. We developed a new pipeline to detect and annotate AS events form RNA-seq data. Our analysis highly expands the number of known splicing variants in the transcriptome in general and in particular in seeds.

## RESULTS

### Whole transcriptome profiling of Arabidopsis seeds prior and after desiccation.

In order to understand the seed transcriptome changes prior and after seed desiccation, we examined gene expression and splicing patterns using high-throughput RNA sequencing. Poly-adenylated RNA from Columbia-0 14 days after pollination (DAP) and mature dry seeds at 20 DAP was used to construct cDNA libraries. We retained information about the direction of transcription thanks to the application of the dUTP method to the second strand cDNA synthesis. (Levin et al. 2010). Our RNA-seq data is therefore strand-specific. In order to gain insight not only on gene expression but also on AS, we increased the number of reads, sequencing between 72.3 and 83.6 million 95 bp paired end (PE) reads per library, to have a higher genome coverage and information also on less abundant transcripts and splicing variants (FIG 1A). Three independent biological replicates were sequenced for each time point (named 14 DAP-1, 14 DAP-2, 14 DAP-3; 20 DAP-1, 20 DAP-2, 20 DAP-3). The splice junctions obtained during the PE run was provided as an additional argument to Tophat 2 when mapping the single end (SE) reads. Between 90.9% and 98.9% of the reads mapped uniquely to the genome in all libraries. Between 93% and 97% of the uniquely mapped reads from all the libraries were first strand specific (FIG 1B), as expected from the d-UTP strand specific selection applied (Levin et al. 2010). The reads mapping on the second strand could at least in part result from antisense transcription (Li et al. 2013). RPKM values (Mortazavi et al. 2008) were computed for each gene and library from the read counts obtained and with the gene lengths derived from the representative gene models. Those genes with RPKM greater than 1 in at least three out of the six samples were considered to be expressed and considered for differential gene expression analysis. With these parameters we obtained a total of 15670 expressed genes (SUPPLEMENTARY TABLE 1). 14519 genes were expressed at 14 DAP and 12925 genes at 20 DAP respectively (SUPPLEMENTARY TABLES 2-3), with 11774 genes, 75% of the

**Commento [V2]:** Cut?

total, being expressed in both time points. (FIG 1C; (SUPPLEMENTARY TABLE 4). 74% (8695) of the 11774 genes expressed at 14 DAP and 20 DAP were differentially expressed. 59% of the differentially expressed genes (5156) were down regulated between 14 and 20 DAP (SUPPLEMENTARY TABLE 5), while 41% (3539 genes) were up regulated at 20 DAP compared to 14 DAP (SUPPLEMENTARY TABLES 6). Therefore overall transcription was slightly reduced between 14 and 20 DAP. (FIG 1D).

**RNA-Seq identifies extensive alternative splicing during seed maturation**

We developed a pipeline to detect alternative splicing (AS) events, including not previously annotated (novel) AS events. We classified alternative splicing events into 6 classes: alternative 3' (A3P), exon skipping (ES), alternative 5' (A5P), intron retention (IR), cryptic intron (CI) and cryptic exon (CE) (FIG 2A). We obtained a total of 8927 AS events in 4875 genes (SUPPLEMENTARY TABLE 7), among which IR events were the most abundant (60.8% of all AS events), followed by A3P (21%) and A5P (10%, Figure 2B). The ratios between different types of AS events are similar to those reported for other plant tissues (Naeem H Syed et al. 2012; Marquez et al. 2012; Chardon et al. 2004; Kornblihtt et al. 2013). We also identified a 4% of CI, 1,8% of ES and 1,5% of CE (FIG 2B).

Out of the 8927 total AS events that we have found in the two seed developmental stages, 88 % (7856 AS events) were not annotated in the TAIR 10 gene model. In particular, 92% of the IR events detected, 89% of the A5P, 77% of the A3P and 66% of the ES events were not described in the annotation (FIGURE 2C). None of the CI and CE that we have identified were previously reported. Among the junctions described, the canonical splice site GT-AG was found in 97,75% of the total alternatively spliced transcripts (SUPPLEMENTARY TABLE 8). The second most abundant type of donor-acceptor sequence is GC-AG (1,8%) followed by AT-AC (0,45%) (SUPPLEMENTARY TABLE 8; FIGURE 2D). These numbers represent a 1 percent decrease in the novel GT-AG splice site compared to the TAIR10 annotation (98,76%), which is not observed when taking into account only the annotated events (FIGURE 2E). We also analyzed the type of di-nucleotide sequence in relation to the different type of AS (FIGURE 2F). The ratio of canonical vs. non-canonical splice sites varies among type of AS event taken into account and between novel and annotated splicing variants. 100% of CE and 92% of CI events had canonical AG-GT splice site junctions. IR events were the most abundant with 97,6% and 98% of novel and annotated splicing variants respectively having an AG-GT canonical splice site.

### Relationship between alternative splicing and differential gene expression during seed development

A total of 4723 and 4494 genes have AS variants in 14 and 20 DAP seeds, respectively. Since some genes are affected by more than one AS event, the total number of events per time point is 8567 at 14 DAP and 8250 at 20 DAP. We asked whether AS plays a role as a regulatory mechanism during seed maturation. We found a total of 1809 significantly differentially alternatively spliced (DAS) events in 1408 genes (SUPPLEMENTARY TABLE 9).These genes either present an alternative splicing variant only at 14 or at 20 DAP or they show a different ratio between the "canonical" (for our definition of "canonical" variant see methods) and an alternative variant at a specific splice site. This represents about 20% of the genes that are alternatively spliced in the seed at 14 or 20 DAP. The most common event again is IR, with 69.26% of total DAS events. A3P was 14.21%, followed by A5P at 8.46%, CI 6.69% and ES 1.33% ( FIG. 3A).

In order to assess the regulatory role of AS during seed maturation, we evaluated its relationship to differential gene expression.

We computed a total of 8695 differentially expressed genes between 14 and 20 DAP. Out of the 1408 significantly differentially alternatively spliced genes, 688 are also differentially expressed. In .particular 468 are up regulated and 220 are down-regulated from 14 DAP to 20 DAP (SUPPLEMENTARY TABLE 10-11). Interestingly, for 720 (51%) of the DAS genes found there is no significant difference in expression between 14 and 20 DAP (SUPPLEMENTARY TABLE 12; FIGURE 3B). This result suggests that DAS plays a prominent role in transcript regulation during seed maturation in Arabidopsis.

### AS impact on protein function

In order to evaluate the possible impact of AS on protein function, we predicted the protein sequence resulting from the AS events identified. We focused this analysis only on IR, which represents the most common type of AS (61% of the AS events, 5431 out of 8927). The majority of IR events resulted in the loss of frame and the formation of a premature termination of the proteins. Out of the 5431 IR events, only 420 (7.7%) did not end prematurely (SUPPLEMENTARY TABLE 13).

To see if some gene functional categories were preferentially regulated by AS during seed maturation, we analyzed the list of significantly DAS genes following the Gene Ontology (GO) biological process categorization (Ashburner et al. 2000; Gene et al. 2014). (FIGURE 4).

The most enriched (4.19 and 3.95%) functional categories include genes involved in mRNA catabolic processes. Genes specifically involved in mRNA splicing are enriched by 2.57%. These data are in agreement with the fact that genes related to RNA metabolism and in particular splicing factors are in general more affected by AS in response to development or environmental cues (Naeem H. Syed et al. 2012; Reddy et al. 2013).

**Experimental validation of the computational predictions**

In order to validate the results of our AS and DAS pipelines, we chose 21 events from all different types of AS taken into account. Twenty events were not previously reported in other tissues or experimental condition, while 1 was already annotated (IR in AT1G55350 3'UTR) (SUPPLEMENTARY TABLE 14). We designed specific primers for each splicing variant that we used to evaluate the relative transcript abundance in a qRT-PCR. All events were validated on 3 independent biological samples per time point. For all genes, the splicing variant computationally predicted could always be detected and the ratio between canonical and alternative variants was proportional to the predicted ratio (FIGURE 5A).

It's possible that for genes with low read counts, the pipeline was not able to detect the event as significant, yet AS events could be taking place and be biologically relevant. Therefore, to test the sensitivity of the method, we also tested an A3P event that was affecting *FUSCA*3, a master regulator of seed development. The event was not among the most statistically significant genes, but the read counts suggested a potential DAS event. *FUS3* expression decreased about 10 times from 14 to 20 DAP. In an opposite trend to the overall transcript, *FUS3* A3P variant is accumulated at 20 DAP, where it reaches about 35% of the canonic variant. Interestingly *FUS3* A3P encodes for a truncated version that lacks part of the B3 DNA-binding domain (Figure 5B-E), similarly to *ABI3* when the out-of-frame cryptic intron is removed (Sugliani et al. 2010) and to PIF6 exon skipping (Penfield et al. 2010).

DISCUSSION.

### Splicing dynamics in seed during acquisition of desiccation tolerance

AS is a powerful mechanism that controls gene expression and consent to rapid changes in transcriptome and proteome complexity during development and upon environmental changes (Reddy et al. 2013). Here, we presented an extensive description of AS during the last phase of seed maturation (14-20 DAP), when the fully developed seed desiccates and becomes dormant. This type of information was still lacking for this type of tissue and developmental stage and contributes to a growing list of AS events occurring during plant development (Barbazuk et al. 2008; Chang et al. 2014; Yoshimura et al. 2011; Leviatan et al. 2013). Additionally mature dry seeds have a unique cellular organization due to their very low water potentials (Terrasson et al. 2013). It is therefore interesting to study transcriptomic changes including AS in this specific environment. Besides seeds, AS has been reported to be involved in desiccation together with ABA, also in leaves and roots of desiccation tolerant plants (Xiao et al. 2015; Dinakar & Bartels 2013)

To better understand the transcriptome dynamics during seed dehydration, we also analyzed gene expression. Interestingly gene expression levels and AS have a slightly opposite trend. Overall transcription was reduced between 14 and 20 DAP, while AS was increased. Transcription is probably affected by increasing desiccation and by the vital but quiescent state in which the seed is set until germination is induced (Terrasson et al. 2013). Differently, AS affects a slightly higher percentage of genes at 20 DAP compared to 14 DAP, possibly to allow the rapid end of seed maturation once desiccation is achieved. As with many other biological processes, AS is likely to contribute largely to the correct progression towards seed desiccation, since it regulates some important regulators of this mechanism (Penfield et al. 2010; Sugliani et al. 2010; Gao et al. 2013).

Surprisingly, 88% of the AS events that we found in 14 and 20 DAP seeds were novel events. This might be due to the fact that no high-throughput studies had been performed on this tissue before. .

### Prediction of AS on proteasome complexity

AS has the potential to rapidly expand proteasome complexity and had a crucial role during plant evolution (Xiao et al. 2015; Li et al. 2015; Tack et al. 2014; Xu et al. 2014; Vitulo et al. 2014). To gain information about the possible effects of AS during seed desiccation, we predicted the sequences of the protein isoforms resulting from the IR AS transcripts. Opposed to animals, IR is far the most common type of AS event found in plants. This diversity reflects the different gene structure where plants introns are generally smaller and there are fewer exons per gene than in animals (Reddy et al. 2013). We predicted the protein sequence resulting from the IR event taken into account, when the rest of the gene would be “canonically” spliced. The majority of IR events resulted in premature termination of the proteins. Out of the 5431 IR events, only 420 (7.7%) did not end prematurely. Premature termination codons (PTC) IR events are not necessarily directed to degradation through NMD, as only a small percentage of them follows this fate(Kalyna et al. 2012). In addition it has been proposed that some of them can turn into Micro Proteins, that lack some of the functional domain of the entire protein and have regulatory functions (Staudt & Wenkel 2011; Graeff & Wenkel 2012; Brandt et al. 2014). To this extent, IR events might have a prominent regulatory role. The actual presence of the predicted isoforms was not verified. Nevertheless the fact that the ABI3 truncated cryptic intron isoform accumulated at 20 DAP, provides an example of an AS protein isoform having a regulatory function. Candidate genes with different predicted isoforms could be further characterized to verify their function.

It is well known that splicing can rapidly modify an mRNA sequence and function in response to stimuli. To promptly modify a cascade of several AS genes, splicing regulators are often the primary targets of AS themselves. We can conclude that this general observation is also valid in seeds, where the most represented category of AS genes are RNA-binding proteins and other various components of the spliceosome.

**Antisense transcription**

To enhance the accuracy of our prediction on AS, we paid attention to generate strand specific RNA-seq data. Indeed between 93% and 97% of the uniquely mapped reads from all the libraries were first strand specific. Nevertheless a small percentage that were mapping on the second strand could at least in part result from antisense transcription (Li et al. 2013). Deep RNA-seq technologies and strand selection allow investigation of also this aspect of RNA biology that is not well known but seemingly increasingly relevant.

## MATERIALS AND METHODS

### Plant growth conditions, RNA extraction and quantification of transcript abundance

Plants were grown under long day conditions (16h light). Siliques were collected at 14 and 20 days after pollination. Ripening siliques were homogeneous among different plants in our conditions. Plants for RNA-Seq or qRT-PCR validation were grown in independent experiments. Seeds were removed from the siliques in liquid nitrogen prior to RNA extraction. RNA was prepared according to the protocol from (Nakabayashi et al. 2012). 14 DAP and 20 DAP seeds were grinded in a mortar in liquid nitrogen and RNA was extracted with the Ambion RNAqueous® extraction kit supplemented with the RNA Isolation Aid® as previously described by Nakabayashi et al. 2012. Retro Transcription was performed using oligo dT and the SuperScript II® reverse transcriptase (Invitrogen). qRT-PCR was performed in an Eppendorf realplex$^2$. Expression of canonical and splicing variants was normalized to *ACTIN 8* (At1g49240) with the primers described in Sugliani et al., 2009. The ratio between 14 DAP and 20 DAP of the two variants was calculated for RNA-seq data using the average reads number from the three RNA-seq biological replicates. For qRT-PCR is the average of three independent experiments.

**Commento [V3]:** Should i also add the primers sequence? Mabe I could add a supplementary table

### Mapping and expression analysis

Sequencing was performed on the Illumina GAII platform. Each replica was run on a single flow cell. RNA-Seq reads were examined using fastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) to exclude lane-specific effects due to sequencing or contamination of individual libraries. Adapter sequences were clipped and reads with qualities of less than 30 were trimmed using in-house Perl scripts. The PE reads were filtered during pre-processing, and reads where only one pair passed filtering criteria were retained as single-end (SE) reads. Filtered reads shorter than 50 bp were discarded. After pre-processing, between 93.6% and 97.9% of the reads were retained. Reads were then mapped to the *Arabidopsis thaliana* reference genome TAIR10 using Tophat2 v2.0.9 (Trapnell et al. 2009) allowing a maximum of five mismatches, insertions and deletions each, with a total edit distance of 10.

Read counts for each of the six libraries were obtained using in-house R-package readCounts, which internally uses Bioconductor packages GenomicRanges, IRanges (Lawrence et al. 2013) and Rsamtools (http://bioconductor.org/packages/release/bioc/html/Rsamtools.html), and the CRAN package (M Dowle, T Short, S Lianoglou, A Srinivasan with contributions from R Saporta and E Antonyan). RPKM values

(Mortazavi et al. 2008) were computed for each gene and library from the read counts and using the gene lengths derived from the representative gene models (Extension of Data.frame. R package version 1.9.5 https://github.com/Rdatatable/data.table/wiki). Clcbio ® (www.clcbio.com) was also used for read mapping and visualization of the AS events.

## Alternative splicing

### Generation of non-overlapping introns

In order to efficiently detect also the novel AS events from RNA-seq data, we first computed a set of non-overlapping intron coordinates. This enabled us to detect splicing variants of genes with multiple transcripts that may contain introns (and exons) that overlap with one another across multiple transcripts. The set of non-overlapping intron coordinates are generated using inhouse R package gffutils. When multiple introns had identical start, end or both coordinates, only the shortest intron was retained. If introns were overlapping, but neither start nor end coordinate was identical, then all those introns were retained. We refer to these non-overlapping intron coordinates as canonical introns. A set of canonical introns can be constructed for each gene from TAIR10 Arabidopsis thaliana gene model. An example of this method is illustrated in FIGURE 6. Out of a total of 127896 unique introns, 124399 canonical introns were obtained using gffutils.

### Splice junctions

For each library splice junction coordinates were found from the reads obtained by mapping against the Arabidopsis *thaliana* genome using Tophat v2.0.9. At the first stage of filtering, splice junctions with >= 3 reads in at least four out of the six libraries were retained. The MMES score (Wang et al. 2010) of each of the retained spliced read were computed, and only those splice junctions with > 50% of the reads having MMES score > 5 were retained. If a read extended to more than one gene, and those genes were not *overlapping genes*, those reads were discarded as erroneous due to mapping inconsistencies. This resulted in a set of high quality splice junctions. Read counts across these filtered splice junctions were then calculated. Similarly *median coverage* across all the retained splice junction was also computed separately and only those junctions with a median coverage >= 3 in at least four out of six libraries were retained. These operations were accomplished using the in-house R package splicerutils. There were a total of 145550 splice junctions from all the six libraries combined. Out of those, 143093 junctions occurred within an annotated genic region, and the remaining 2457 junctions were in the intergenic region. Out of the 143093 splice junctions in the genic regions, 98666 junctions were annotated as an intron in the gene model. The remaining 44427 splice junctions were not present in the gene model and contained amongst them potential *novel* AS events.

**Alternative Splicing events**

All splice junctions with identical coordinates as canonical introns were classified as canonical (splice) junctions, CJ. The other splice junctions that overlap with CJ could be automatically classified as alternatively spliced AS. These AS junctions consist of union of the set of annotated AS junctions, i.e., present in the gene model, and novel AS junctions. They are classified as A3P, ES, A5P, IR, CE or CI. AS junctions where start coordinates matched a CJ, but not the end coordinates were classified as A3P events.

AS junctions where start coordinate matched a CJ, and the end coordinate matched another CJ within the same gene were classified as ES events. AS junctions where end coordinates matched a CJ, but not the start coordinates were classified as A5P events. AS junctions where unspliced reads across CJ had a median coverage >= 3 were considered to be IR events. AS junctions which occurred within an annotated exon are marked as CI events. They could be very well a result of introns not being annotated in the gene model. Two AS junctions where start of the first AS junction and end of the second correspond to a CJ implies that the end coordinate of the first AS and the start coordinate of the second occurred within that CJ. There must be therefore an exon within this CJ that has not yet been annotated. Those events are marked as CE events. Our pipeline was specifically designed to detect these rare types of events since ABI3 function and seed maturation was strongly affected by a CI AS event (Sugliani et al. 2010).

**Identification of differential alternative splicing (DAS) events**

The pipeline to detect DAS events is implemented in the in-house R package splicer. The filtered splice junctions are already classified into canonical (splice) junctions (CJ) and alternative spliced junctions (AS). For each AS event, the corresponding overlapping CJ is identified. Normalized read counts per library per time point were extracted for each AS event and its corresponding CJ. This resulted in a total of 12 values (read counts) per AS event - three each corresponding to AS and CJ in time points 14 DAP and 20 DAP.

Once the read counts were extracted, a negative binomial generalized linear model, NB-GLM, was fitted using the R package MASS (Venables & Ripley 2002) for each AS event by modeling a two-way interaction between *time point* and AS type (type) with *read counts* as the response variable. The binary variable TP takes two values corresponding to each of the time points, 14 DAP and 20 DAP. The binary variable type takes two values as well corresponding to whether the read counts come from alternatively spliced or canonical junctions, AS and CJ. P-values corresponding to the interaction term were extracted and adjusted to correct for multiple testing by using the Benjamini-Hochberg procedure using the R-package multtest (Pollard et al. 2005) at a false discovery rate (FDR) of 5%. AS events with $q \leq 0.05$ were considered as undergoing significant differential alternative splicing.

### Protein variant prediction

For each gene, we extracted and computed the length of the protein sequence for the representative gene model. Following that, we incorporated the IR event on to the representative gene model and computed the protein sequence under all six frames and chose the longest. Those events where the intron retained transcripts resulted in longer protein sequence than the representative model were considered not to end prematurely.

### Differential expression analysis

Raw read counts from genes that were retained after filtering using RPKM > 1 as explained above were inputted to the bioconductor package DESeq v1.14.0 to detect the genes that are differentially expressed between the time points TP14 and TP20. Genes with FDR corrected p-values (or q-values), $q <= 0.05$, were considered to be significantly differentially expressed.

### Gene onthology functional categorization

Gene lists were submitted to AmiGO 1.8 release 01.08.2015 (http://amigo.geneontology.org/amigo) and only those categories represented by more than 40 genes and enriched at least 2.5 times compared to occurrence in the entire genome were taken into account.

## ACKNOWLEDGEMENTS

**Commento [V4]:** please Arun add a legend to the supplementary tables.

We are grateful to prof. Maarten Koornneef for guidance and support.

This work was supported by the Alexander von Humboldt Foundation, Germany.

REFERENCES

Ashburner, M. et al., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1), pp.25–29.

Barbazuk, W.B., Fu, Y. & McGinnis, K.M., 2008. Genome-wide analyses of alternative splicing in plants: opportunities and challenges. *Genome research*, 18(9), pp.1381–92. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18669480 [Accessed January 24, 2014].

Baud, S. et al., 2002. An integrated overview of seed development in Arabidopsis thaliana ecotype WS. *Plant Physiology and Biochemistry*, 40(2), pp.151–160. Available at: http://linkinghub.elsevier.com/retrieve/pii/S098194280101350X.

Brandt, R. et al., 2014. Homeodomain leucine-zipper proteins and their role in synchronizing growth and development with the environment. *Journal of Integrative Plant Biology*, 56(6), pp.518–526. Available at: http://doi.wiley.com/10.1111/jipb.12185.

Chang, C.-Y., Lin, W.-D. & Tu, S.-L., 2014. Genome-Wide Analysis of Heat-Sensitive Alternative Splicing in Physcomitrella patens. *Plant physiology*, 165(2), pp.826–840. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4044832&tool=pmcentrez&rendertype=abstract [Accessed July 25, 2014].

Chardon, F. et al., 2004. Genetic architecture of flowering time in maize as inferred from quantitative trait loci meta-analysis and synteny conservation with the rice genome. *Genetics*, 168, pp.2169–2185.

Dinakar, C. & Bartels, D., 2013. Desiccation tolerance in resurrection plants: new insights from transcriptome, proteome and metabolome analysis. *Frontiers in Plant Science*, 4(November), pp.1–14. Available at: http://journal.frontiersin.org/article/10.3389/fpls.2013.00482/abstract.

Fait, A. et al., 2006. Arabidopsis seed development and germination is associated with temporally distinct metabolic switches. *Plant physiology*, 142(3), pp.839–854.

Fan, J. et al., 2007. Short, direct repeats (SDRs)-mediated post-transcriptional processing of a transcription factor gene OsVP1 in rice (Oryza sativa). *Journal of experimental botany*, 58, pp.3811–3817.

Filichkin, S.A. et al., 2010. Genome-wide mapping of alternative splicing in Arabidopsis thaliana. , pp.45–58.

Gagete, A.P. et al., 2009. Functional analysis of the isoforms of an ABI3-like factor of Pisum sativum generated by alternative splicing. *Journal of experimental botany*, 60(6), pp.1703–14. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2671620&tool=pmcentrez&rendertype=abstract [Accessed January 30, 2014].

Gao, Y. et al., 2013. Functional characterization of two alternatively spliced transcripts of tomato ABSCISIC ACID INSENSITIVE3 (ABI3) gene. *Plant Molecular Biology*, 82(1-2), pp.131–145.

Gazzarrini, S. et al., 2004. The transcription factor FUSCA3 controls developmental timing in Arabidopsis through the hormones gibberellin and abscisic acid. *Developmental cell*, 7(3), pp.373–85. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15363412.

Gene, T. et al., 2014. Gene Ontology Consortium: going forward. *Nucleic Acids Research*, 43(D1), pp.D1049–D1056. Available at: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gku1179.

Goldberg, R.B., de Paiva, G. & Yadegari, R., 1994. Plant embryogenesis: zygote to seed. *Science (New York, N.Y.)*, 266(5185), pp.605–614.

Graeff, M. & Wenkel, S., 2012. Regulation of protein function by interfering protein species. *BioMolecular Concepts*, 3(1).

Holdsworth, M.J., Bentsink, L. & Soppe, W.J.J., 2008. Molecular networks regulating Arabidopsis seed maturation, after-ripening, dormancy and germination. *The New phytologist*, 179, pp.33–54.

Jenik, P.D., Gillmor, C.S. & Lukowitz, W., 2007. Embryonic patterning in Arabidopsis thaliana. *Annual review of cell and developmental biology*, 23, pp.207–236.

Kalyna, M. et al., 2012. Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in Arabidopsis. *Nucleic acids research*, 40(6), pp.2454–69. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3315328&tool=pmcentrez&rendertype=abstract [Accessed January 21, 2014].

Kirkbride, R.C., Fischer, R.L. & Harada, J.J., 2013. LEAFY COTYLEDON1, a key regulator of seed development, is expressed in vegetative and sexual propagules of Selaginella moellendorffii. *PloS one*, 8(6), p.e67971. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3680378&tool=pmcentrez&rendertype=abstract [Accessed April 10, 2014].

Kornblihtt, A.R. et al., 2013. Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nature reviews. Molecular cell biology*, 14(3), pp.153–65. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23385723.

Lawrence, M. et al., 2013. Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology*, 9(8).

Leviatan, N. et al., 2013. Genome-Wide Survey of Cold Stress Regulated Alternative Splicing in Arabidopsis thaliana with Tiling Microarray. *PLoS ONE*, 8(6).

Levin, J.Z. et al., 2010. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature methods*, 7(9), pp.709–15. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3005310&tool=pmcentrez&rendertype=abstract [Accessed January 21, 2014].

Li, P., Tao, Z. & Dean, C., 2015. Phenotypic evolution through variation in splicing of the noncoding RNA COOLAIR. *Genes & development*, 29(7), pp.696–701. Available at: http://genesdev.cshlp.org/content/29/7/696.

Li, S. et al., 2013. Integrated detection of natural antisense transcripts using strand-specific RNA sequencing data. *Genome Research*, 23(10), pp.1730–1739.

Lopez-Molina, L. et al., 2002. ABI5 acts downstream of ABI3 to execute an ABA-dependent growth arrest during germination. *Plant Journal*, 32(3), pp.317–328.

Loraine, A.E. et al., 2013. RNA-seq of Arabidopsis pollen uncovers novel transcription and alternative splicing. *Plant physiology*, 162, pp.1092–109. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3668042&tool=pmcentrez&rendertype=abstract.

Marquez, Y. et al., 2012. Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome Research*, 22(6), pp.1184–1195. Available at: http://genome.cshlp.org/cgi/doi/10.1101/gr.134106.111.

McKibbin, R.S. et al., 2002. Transcripts of Vp-1 homeologues are misspliced in modern wheat and ancestral species. *Proceedings of the National Academy of Sciences of the United States of America*, 99, pp.10203–10208.

Miyoshi, K. et al., 2002. Temporal and spatial expression pattern of the OSVP1 and OSEM genes during seed development in rice. *Plant & cell physiology*, 43, pp.307–313.

Mönke, G. et al., 2004. Seed-specific transcription factors ABI3 and FUS3: molecular interaction with DNA. *Planta*, 219, pp.158–166.

Moreno-Risueno, M.A. et al., 2008. FUSCA3 from barley unveils a common transcriptional regulation of seed-specific genes between cereals and Arabidopsis. *The Plant journal : for cell and molecular biology*, 53, pp.882–894.

Mortazavi, A. et al., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5, pp.621–628.

Nakabayashi, K. et al., 2012. The Time Required for Dormancy Release in Arabidopsis Is Determined by DELAY OF GERMINATION1 Protein Levels in Freshly Harvested Seeds. *The Plant Cell*, 24, pp.2826–2838.

Nakagawa, H. et al., 1996. The seed-specific transcription factor VP1 (OSVP1) is expressed in rice suspension-cultured cells. *Plant & cell physiology*, 37, pp.355–362.

Penfield, S., Josse, E.-M. & Halliday, K.J., 2010. A role for an alternative splice variant of PIF6 in the control of Arabidopsis primary seed dormancy. *Plant Molecular Biology*, 73(1-2), pp.89–95. Available at: http://link.springer.com/10.1007/s11103-009-9571-1.

Pollard, K., Dudoit, S. & Laan, M., 2005. Multiple Testing Procedures: the multtest Package and Applications to Genomics. *Statistics for Biology and Health*, pp.249–271. Available at: http://dx.doi.org/10.1007/0-387-29362-0_15.

Reddy, A.S.N. et al., 2013. Complexity of the alternative splicing landscape in plants. *The Plant cell*, 25(10), pp.3657–83. Available at: http://www.plantcell.org/content/25/10/3657.full.

Reidt, W. et al., 2000. Gene regulation during late embryogenesis: the RY motif of maturation-specific gene promoters is a direct target of the FUS3 gene product. *The Plant journal : for cell and molecular biology*, 21, pp.401–408.

Santos-Mendoza, M. et al., 2008. Deciphering gene regulatory networks that control seed development and maturation in Arabidopsis. *The Plant journal : for cell and molecular biology*, 54, pp.608–620.

Shen, B. et al., 2010. Expression of ZmLEC1 and ZmWRI1 increases seed oil production in maize. *Plant physiology*, 153(3), pp.980–7. Available at:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2899924&tool=pmcentrez&rendertype=abstract [Accessed January 30, 2014].

Staudt, A.-C. & Wenkel, S., 2011. Regulation of protein function by “microProteins”. *EMBO reports*, 12, pp.35–42.

Sugliani, M. et al., 2010. The conserved splicing factor SUA controls alternative splicing of the developmental regulator ABI3 in Arabidopsis. *The Plant cell*, 22(6), pp.1936–1946. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2910958&tool=pmcentrez&rendertype=abstract [Accessed January 22, 2014].

Suzuki, M. & McCarty, D.R., 2008. Functional symmetry of the B3 network controlling seed development. *Current opinion in plant biology*, 11(5), pp.548–53. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18691932 [Accessed January 22, 2014].

Swaminathan, K., Peterson, K. & Jack, T., 2008. The plant B3 superfamily. *Trends in plant science*, 13(12), pp.647–55. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18986826 [Accessed January 24, 2014].

Syed, N.H. et al., 2012. Alternative splicing in plants – coming of age. *Trends in Plant Science*, 17(10), pp.616–623. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1360138512001276.

Syed, N.H. et al., 2012. Alternative splicing in plants--coming of age. *Trends in plant science*, 17(10), pp.616–23. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3466422&tool=pmcentrez&rendertype=abstract.

Tack, D.C., Pitchers, W.R. & Adams, K.L., 2014. Transcriptome Analysis Indicates Considerable Divergence in Alternative Splicing Between Duplicated Genes in Arabidopsis thaliana. *Genetics*, 198(December), pp.1473–1481. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25326238.

Terrasson, E. et al., 2013. An emerging picture of the seed desiccome: confirmed regulators and newcomers identified using transcriptome comparison. *Frontiers in Plant Science*, 4(December), pp.1–16. Available at: http://journal.frontiersin.org/article/10.3389/fpls.2013.00497/abstract.

Trapnell, C., Pachter, L. & Salzberg, S.L., 2009. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9), pp.1105–1111.

Tsai, A.Y.-L. & Gazzarrini, S., 2012. Overlapping and distinct roles of AKIN10 and FUSCA3 in ABA and sugar signaling during seed germination. *Plant Signaling & Behavior*, 7, pp.1238–1242.

Venables, W.N. & Ripley, B.D., 2002. Modern Applied Statistics with S Fourth edition by. *World*, 53(March), p.86. Available at: www.stats.ox.ac.uk/pub/MASS4/VR4stat.pdf.

Vitulo, N. et al., 2014. A deep survey of alternative splicing in grape reveals changes in the splicing machinery related to tissue, stress condition and genotype. *BMC Plant Biology*, 14(1), p.99. Available at: http://www.biomedcentral.com/1471-2229/14/99.

Wang, F. & Perry, S.E., 2013. Identification of Direct Targets of FUSCA3 , a Key Regulator of Arabidopsis Seed Development 1 [ C ][ W ][ OA ]. , 161(March), pp.1251–1264.

Wang, L. et al., 2010. A statistical method for the detection of alternative splicing using RNA-seq. *PloS one*, 5, p.e8529.

Xiao, L. et al., 2015. The resurrection genome of *Boea hygrometrica* : A blueprint for survival of dehydration. *Proceedings of the National Academy of Sciences*, 112(18), pp.5833–5837. Available at: http://www.pnas.org/lookup/doi/10.1073/pnas.1505811112.

Xu, P. et al., 2014. Conservation and functional influence of alternative splicing in wood formation of Populus and Eucalyptus. *BMC Genomics*, 15(1), p.780. Available at: http://www.biomedcentral.com/1471-2164/15/780.

Yoshimura, K. et al., 2011. Identification of alternative splicing events regulated by an arabidopsis serine/arginine-like protein, atsr45a, in response to high-light stress using a tiling array. *Plant and Cell Physiology*, 52(10), pp.1786–1805.

Van Zanten, M. et al., 2011. Seed maturation in Arabidopsis thaliana is characterized by nuclear size reduction and increased chromatin condensation. *Proceedings of the National Academy of Sciences*, 108(50), pp.20219–20224. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3250172&tool=pmcentrez&rendertype=abstract [Accessed January 24, 2014].

Zou, M. et al., 2007. Characterization of alternative splicing products of bZIP transcription factors OsABI5. *Biochemical and biophysical research communications*, 360(2), pp.307–13. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17604002 [Accessed January 30, 2014].