Oviedo, April 10th, 2025

Dear Editor,

We thank the opportunity to submit the revised version of the manuscript “**Pra-GE-ATLAS: empowering Pinus radiata stress and breeding research through a multi-omics database (Manuscript ID:** **JIPB-2024-0568)”.**

In this revised version, we have addressed all suggestions and questions raised during the review process, leading to an overall improvement in the quality of this work. According to the referees’ feedback, we have provided more metadata and clarification regarding data collection, conducted additional analyses, and deployed new applications on the Pra-GE-ATLAS website. The main text has been modified in response to the reviewers’ comments. Detailed responses to the reviewers are provided below in a point-by-point manner.

We believe that our manuscript is now well-prepared for consideration for publication in “Journal of Integrative Plant Biology (JIPB)”, and we hope that both the reviewers and the editorial committee share our opinion.

Sincerely yours,

Un dibujo de un ojo

El contenido generado por IA puede ser incorrecto.

**Dr. Luis Valledor**

Professor

Plant Physiology

Department of Organisms and Systems Biology

University of Oviedo

Spain

E-mail: [valledorluis@uniovi.es](mailto:valledorluis@uniovi.es)

**Manuscript ID:** **JIPB-2024-0568**

**Manuscript TITLE: Pra-GE-ATLAS: empowering Pinus radiata stress and breeding research through a multi-omics database**

***Editor's comments to the author:***

We regret to inform you that, although we find the work of interest, significant problems remain that will need to be resolved. In any event, we would be willing to reconsider a fully revised version that resolves the comments and concerns outlined below. Please try to address all the concerns raised by the reviewers at the authors’ earliest convenience.

# We would like to sincerely thank the Editor to give us this opportunity to submit our revised work and anonymous reviewers for their valuable feedback, which has helped us improve our manuscript. We have carefully modified the manuscript according to all the reviewers’ criticisms, and we hope that Pra-GE-ATLAS could be a useful resource for the integrative plant biology community.

----------

***Reviewers' comments:***

***Reviewer: 1***

*Comments to the Author:*

Compared with angiosperms, especially model plants, the number of databases for conifers, which play a crucial role in forestry and forest ecology, is incredibly scarce. I believe this study takes a crucial step forward. The researchers collected (including newly generated) RNA-seq data from 141 samples and proteomic data from 155 samples, and these data are well visualized. However, it must be said that in the era of big data, the amount of these data is actually not sufficiently abundant.

# First, we want to thank the reviewer for taking the time to review our manuscript. The insightful comments provided have undoubtedly contributed to enhancing our work. We appreciate the reviewer’s perspective on the crucial importance of establishing databases for conifers due to their ecological and molecular relevance, as well as the efforts in visualization. We acknowledge that further clarification regarding data collection, its value, and the outputs of different pipelines and analyses would be highly beneficial, and we sincerely thank the reviewer for highlighting this. We genuinely hope that the comments and analyses presented below will address the concerns raised by the reviewer.

*Points:*

More detailed information about the data is needed, such as the types of stresses and corresponding tissue types, sampling batches, sampling time, and the sequencing platform used to obtain the data.

# We fully agree with the reviewer. Regarding the transcriptomic data, we have expanded the information in **Table S1**, sheet “Samples\_MetaData”, in the revised manuscript. It now includes, in addition to the corresponding tissue ("Tissue"), the stress intensity ("StressIntensity") and the sample batch effects ("Batch1" and "Batch2" columns); type of stress ("StressType"), full name of the stressor ("Treatment"), sampling time ("StressTime"), and sequencing platform ("Platform"). Regarding the proteomic data, **Table S3** now includes a new sheet titled “Samples\_MetaData” which follows the same format as **Table S1**.

# Additionally, we have expanded the description of the data collection in the Methods section to cover the time points of each stress experiment.

**[ Methods, Page 14-15, Lines 487-500]**

The transcriptomic data collection covered five tissues (bud, xylem, phloem, needle, and megagametophyte), one abiotic stress (heat[, with three time points]), and three biotic stresses (*Fusarium circinatum*, *Dothistroma septosporum*, and *Phytophthora pluvialis*[, with five, three, and two time points, respectively]).

[…]

The proteomic data collection covered three tissues analyzed in this study (root, needle, and bud), one biotic stress (*F. circinatum*[, with two time points]), and two abiotic stresses (heat and ultraviolet (UV)), over three different subcellular locations (total proteins, nucleus, and chloroplast). [The heat and UV abiotic stress experiments included three and five time points, respectively, for total proteomes; six and three time points for nucleus proteomes; and four time points in both cases for chloroplast proteomes.]

It is a pity that a high-quality reference genome for this important conifer has not been released yet, for its huge genome, this may be a considerable challenge. The authors attempted to address this issue by constructing a reference transcriptome, but we may need more details, such as whether this assembly comes from long-read or short-read sequencing? How to ensure the accuracy of the assembly (such as redundancy)? How to determine whether the AS is truly present rather than a variation caused by multiple gene copies?

# We thank the reviewer for these questions. As the comment contains several questions, we have decided to address them in a point-by-point manner.

**It is a pity that a high-quality reference genome for this important conifer has not been released yet, for its huge genome, this may be a considerable challenge. The authors attempted to address this issue by constructing a reference transcriptome, but we may need more details, such as whether this assembly comes from long-read or short-read sequencing?**

# We agree with the reviewer’s observation regarding the absence of reference genomes. Long life cycles, large genome sizes, and approximately 70% transposon content are among the main challenges plant scientists face when selecting *Pinus* species as model organisms and prioritizing them for sequencing efforts. The increasing availability of long-read DNA sequencing holds great promise for addressing this issue. However, genome size remains a computational bottleneck, and the slower progress in conifer research continues to limit their broader integration and traction within the field.

# That said, all available transcriptomic evidence from *P. radiata* is based on short-read sequencing, with no reported datasets utilizing long-read RNA sequencing. This absence can likely be attributed to three main factors:

# 1) It is challenging to invest in and accumulate long-read sequencing efforts within the community, both in terms of time and resources, for a species that lacks a well-established short-read reference and a centralized database hub. As the reviewer rightly points out, this is unfortunate, especially considering that exceptional efforts by individual research groups, such as those working on *P. tabuliformis*, have provided remarkable insights - not only for plant science and evolutionary biology but also for the broader scientific audience (Niue, et al. The Chinese pine genome and methylome unveil key features of conifer evolution. **Cell** 185(1), 204-217 (2022)).

# 2) While RNA long-read sequencing is undoubtedly superior in terms of full-length isoform recovery, structural variants, and RNA modifications, most biological hypotheses tested in conifer research have focused on gene expression and SNP genotyping—applications for which short-read sequencing remains the state-of-the-art. This is due to the lower throughput of long-read sequencing (fewer millions of reads) for expression analysis and higher error rates for SNP genotyping compared to short-read sequencing. As an expression atlas, Pra-GE-ATLAS shares this primary focus.

# However, to achieve the best splicing results with the resources available and considering the limitations of short-read sequencing, we opted for a *de novo* splicing detection approach in Pra-GE-ATLAS, followed by restrictive downstream parameters. This decision was made to prioritize consistency, even if it meant sacrificing some true information. Unlike other popular methods, such as SUPPA and rMATS, which rely on pre-existing annotations and/or focus on expression quantification at the isoform level in the reference, our approach is designed to identify splicing events without such dependencies. This will be further discussed below, along with the last question of the same comment.

# 3) As a non-model species, research progress is relatively slow, resulting in a wide temporal range for the release dates of many of these useful transcriptomic datasets. Given this, when most of these studies were conducted, long-read sequencing was neither widely available nor cost-effective. This has been highlighted as one of the limitations of Pra-GE-ATLAS in the Discussion (**Page 13, Lines 436-441**). To clarify this point further, we have explicitly added the term long-read sequencing to that section.

**[ Discussion, Page 13, Lines 438-441]**

Therefore, newly reported datasets, taking advantage of recent technological improvements, increased analytical resolution of MS[, and both long-read sequencing and enhanced sequencers performance], will significantly improve the resources presented, owing to higher throughputs.

# Taken together, although the data may seem limited, Pra-GE-ATLAS consolidates and manually curates all available multi-stress, multi-tissue proteomics and transcriptomics data for this important species, addressing a critical caveat in conifer research—a clade of broad significance to plant science. To compensate, we leveraged data diversity through comprehensive integrative analyses, revealing new biological insights, such as the overlap of stress responses across regulatory layers and intergenerational memory. We believe Pra-GE-ATLAS will enhance conifer research and help bridge key knowledge gaps in plant science.

**How to ensure the accuracy of the assembly (such as redundancy)?**

#.As the reviewer points out, to ensure and demonstrate the accuracy of the assembly, we have included all state-of-the-art transcriptome assembly evaluation metrics in **Table S1**, under the sheet titled "Assembly\_Evaluation." Among these metrics, the duplication level ("D" in the BUSCO column) is commonly used as a proxy for redundancy in the assembly, as higher redundancy typically leads to increased duplication levels within this framework.

# As notably highlighted by the reviewer, redundancy is a inherent aspect of the transcriptome assembly process,and its assessment largely depends on the intended downstream application. Therefore, we would like to provide additional context on this matter. Currently, the most commonly used methods to reduce redundancy in transcriptome assembly involve clustering sequences based on similarity and/or filtering out lowly expressed transcripts. Inspired by recent research in *Marchantia polymorpha*, an emerging model for testing the conservation of molecular mechanisms in land plants, we envision the application of *P. radiata* to assess the degree of conservation in seed plants (Spermatophyta). Therefore, we aim to retain as much sequence diversity as possible to support future evolutionary studies and a wide-range of user-driven downstream analyses.

# In Pra-GE-ATLAS, we carefully balanced sequence space retention and redundancy removal. We intentionally avoided redundancy removal through low-expressed filtering, as key regulatory transcripts, such as transcription factors, are ofter expressed at low levels. Eliminating them would risk losing crucial biological information for evolutionary applications. First, we used EvidentialGene to classify sequences as main and alternate. Sequences classified as main were retained unchanged and designated as “Consensus”. Alternate sequences underwent an additional similarity clustering step using CD-HIT -c 0.905. The reduced alternate sequences were then merged with the consensus set to form the final assembly, termed “Final”.

# Building on the excellent summary of Embryophyta\_odb10 BUSCO scores across gymnosperms assemblies by Visser, et al. Multispecies comparison of host response to *Fusarium circinatum* challenge in tropical pines show consistency in resistance mechanisms. **Plant, Cell & Environment** 46(5), 1705-1725 (2023); **Data S2 – Table S5**, our consensus assembly, despite containing a high number of sequences, remains low in redundancy. Among the gymnosperms transcriptome assemblies with lower redundancy, only five other species surpass 90% completeness, and all fall short of our assembly’s completeness. The slight reduction in completeness and redundancy can be attributed to the application of low-expression filters, which were specifically avoided during the evaluation step in Pra-GE-ATLAS because of abovementioned reasons.

# We would like to highlight that, for scalability and resource usability, some non-evolutionary applications in the database (Search: Expression – Transcriptome, Tools: Heatmap, Tools: Regulation Browser) use a version of the “Final” assembly that has been gently filtered for low expression, consisting of 78,320 transcripts. However, the full “Consensus” set of transcripts is used for evolutionary-related tools: (DIAMOND – BLASTP and orthology inference).

# To further satisfy reviewer’s concerns, we performed one more analysis on the evaluated consensus assembly. Specifically, we assessed the extent of sequence removal after two stringent CD-HIT clustering rounds (-c 0.98 -b 3, as recommended by Trinity assembler developers to remove redundancy; and -c 0.95 -b 3). Overall, the results from BUSCO and the minor sequences removal (~ 6.3-8.9%) confirm the low level of redundancy in the consensus assembly. This new metric has been added to the “Assembly\_Evaluation” sheet in **Table S1**. With the information available in the download section and Zenodo repository, we provide future users all the necessary resources to perform low-expression filtering if needed for their specific applications.

**How to determine whether the AS is truly present rather than a variation caused by multiple gene copies?**

# We thank the reviewer for raising this intriguing question. As discussed above, our data is based on short reads, and we employ a *de novo* approach with the KisSplice software. KisSplice is designed to reconstruct local sequence variation bubbles and classify them into different types of structural variations. Its primary purpose, therefore, is not to assemble complete transcript isoforms, less suited for short reads, but rather to capture as much local sequence variation as possible, applying strict thresholds for later assignment. We leveraged KisSplice’s local assembly approach to optimize the use of short-read data, enabling high-precision analysis, albeit at the cost of more complex event-level interpretations and downstream analyses. To ensure the presence of alternative splicing (AS), we have carefully applied a set of stringent criterio:

# 1) Understanding and generealizing multiple gene copies as paralogs within the same gene family, two possible scenarios must be considered when calling splicing events using de Brujin graphs in KisSplice for local sequence variation assembly:

# 1.1) If gene copies have diverged significantly, KisSplice will not detect a sequence variation bubble, meaning no SNPs, repeats, or AS events will be called. To reinforce this, we consciously increased the k-mer length to 51 (-k 51) and required a minimum relative coverage of 5%. This choice prioritizes specificity, resolving complex regions and distinguishing divergent gene copies, albeit at the expense of sensitivity (e.g. losing number of AS events and missing rare isoforms).

# 1.2) If the paralogs have not diverge too much (e.g. recent duplications), they generate long stretches of SNP-like bubbles. KisSplice classifies these under the “inexact repeats” category, (termed as Type 0). According to KisSplice authors, long Type 0 bubbles serve as strong indicators of multiple gene copies. In our analysis, we explicitly considered only Type 1 sequence variation bubbles, which KisSplice specifically designates as AS events. To further enhance specificity, we excluded bubbles classified as Type 0 (substitution erros, SNPs, and similar gene copies), Type 2(inexact tandem repeats), Type 3 (short indels) and Type 4 (all other cases, such as complex AS patterns with shorter paths than twice the k-mer length that are not SNPs).

# In conclusion, with appropiate parameter setting and output selection, KisSplice effectively differentiates between multiple gene copies and AS events. This increased specificity comes at the cost of sensitivity and loss of truly transcriptomic variation information. However, this trade-off, along with its local assembly nature, was a key factor in selecting KisSplice for Pra-GE-ATLAS’s *de novo* splicing framework.

# 2) While KisSplice does not confound AS events with variation caused by multiple gene copies – an issue further addressed by mapping to *P. taeda* genome – another possible scenario is that at least one member of a gene family with highly similar gene copies is alternatively spliced. In this case, the detected event represents genuine splicing biological variation rather than an artifact of multiple gene copies. However, as a limitation of using a local approach, determining which specific member contributes to the AS event would be difficult. Since such cases still represent true AS variation, they should be considered instances where highly similar gene regions from the same gene family are “collectively spliced,” as proposed by the KisSplice developers. To address this, we cautiously conducted all splicing analyses at the “event-level” rather than the “gene-level”. To ensure a fair representation of these AS events in downstream analyses, we retained annotations from all associated gene copies, as they are expected to be highly similar in the regions where splicing occurs. To further support this notion, we performed an additional analysis. Specifically, for all such cases, we computed protein domain string similarity (based on the Jaccard distance of PFAM and SMART hits; ranging from 0 to 1, where 1 indicates identical protein domain profiles) and the CDS nucleotide global/local identity across all possible gene copies to which these AS events could be mapped. Among the 21,280 unique AS events reported after all processing steps, 6.95 % corresponded to such cases. As hypothesized above, these cases exhibited very high protein domain similarity (mean: 0.93, sd: 0.2, median: 1) and nucleotide sequence identity (global – mean: 83.7, sd: 17.3, median: 93.1; local – mean: 98.6, sd: 2.6, median: 100). Therefore, analyzing these “collectively spliced” AS events at the event level, while retaining annotations from the gene copies presenting the highly similar regions, ensures that genuine AS variation is captured without misrepresenting redundancy or losing valuable information.

# To further clarify this point, we have updated the Methods section to detail our approach in these cases:

**[ Methods, Page 15-16, Lines 521-527]**

*De novo* splicing events were identified, classified, and quantified using KisSplice v2.6.2 (-k 51 -C 0.05) (Sacomoto et al., 2012), Kiss2refgenome v2.0.8, and kissDE v1.4.0, respectively. [For AS events mapped to different genes due to at least one member of a gene family being alternatively spliced in a highly similar region, all annotations were retained].sva v3.48.0 (Leek et al., 2012) was employed to remove raw counts unwanted variation derived from study/sequencing-type.

A total of 7697 proteins were identified from over one hundred multi-tissue samples, which is a tremendous effort. However, it has to be said that the number of proteins is still has a gap from the state-of-the-art technical standards. For example, using the Orbitrap Astral platform, a single sample may detect more than 10,000 proteins. These proteomic are rich in many high-abundance proteins, such as photosynthesis proteins, RNA processing proteins, histones, and heat shock proteins, while potentially missing many important regulatory protein information. Nevertheless, I still admit this is an important endeavor.

# We sincerely appreciate the reviewer's recognition of our proteomic efforts. As noted earlier, we acknowledge recent technological advancements, such as increased depth and resolution of mass spectrometry (MS), as a limitation when compiling all available information for an expression atlas of a non-model species. This is discussed in the Discussion section (**Page 13, Lines 436–441**). Here, we would like to provide additional insights into this challenge within the context of pine proteomics.

# 1) We fully share the reviewer’s enthusiasm for the Orbitrap Astral platform and its promising capabilities in terms of depth and speed, which are driving an exciting new era of deep single-cell proteomics. This marks a clear inflection point in the proteomics field. While now considered state-of-the-art, the Orbitrap Astral platform was only recently introduced, publicly showcased, and officially released in June 2023. Since then, it has been adopted by various companies and universities, offering it as a publicly available service.

# Working with underrepresented biological systems is highly relevant to the scientific community, but it often comes with challenges. These non-model organisms are not always ideal candidates for the early adoption of new technological breakthroughs due to both biological and technical constraints. In this particular case, it is important to note that most experiments involving *P. radiata* seedlings typically use plants that are about one year old. As a result, aligning the generation of one-year-old seedlings with the implementation of this technology in June 2023 would have been highly challenging, especially considering that Pra-GE-ATLAS was submitted in early July 2024, which required significant time for data curation and integration.

# Additionally, we examined the number of proteins reported using this new platform in plant systems, as documented in Thermo Fisher’s official document, “Plant Proteomics Orbitrap Astral Publication” (which compiles studies up to February 2025, available at: https://assets.thermofisher.com/TFS-Assets/CMD/brochures/eb-003665-ov-orbitrap-astral-plantproteomics-eb003665-na-en.pdf). Among the nine articles cited, only four reported the number of proteins detected with this platform. A notable limitation given the significance of this metric for the proteomics community. The highest proteins counts reported were 7,127 proteins in Chinese chive seedlings subjected to five different drought treatments (N = 15) and 7,286 proteins in *Agropyron mongolicum* seedlings sampled from two different tissues under drought stress (N = 12). Overall, Pra-GE-ATLAS falls within the same range as these early adopters of the platform.

# 2) Our research group has long focused on the biology of this prominent species, and our long-term goal is to continue these efforts. That is why we are still generating proteomic datasets for this species (which will be added to Pra-GE-ATLAS in the future, once fully released and analyzed, as stated in Discussion section (**Page 13, Lines 444-446**)), now leveraging new technological advancements. As an internal observation, one of our newly generated datasets, obtained using the MS timsTOF Pro 2 in DiaPASEF mode (released in 2021, just one generation before the aforementioned Astral Orbitrap platform), analyzed proteomes under combined stressors (N = 33) and yielded approximately 6,000 proteins after pre-processing. Remarkably, Pra-GE-ATLAS still falls within the same range of detected proteins, underscoring the high value of the diverse proteomes collected and curated.

# 3) We agree with the reviewer’s observation that proteomes are rich in high-abundance proteins. That said, while acknowledging the advancements brought by the aforementioned recent technological breakthroughs, this is typically considered a trade-off of shotgun proteomics rather than a specific feature of Pra-GE-ATLAS proteomes. Although shotgun proteomics allows the simultaneous identification of a broad range of proteins, including regulatory ones, it is not always the most suitable approach for detecting low-abundance proteins. Briefly, shotgun proteomics relies on high-throughput MS to detect peptides from highly complex mixtures, where low-abundance proteins can be masked by more abundant ones.

# While we partially attempted to address this limitation by incorporating proteomes based on enrichment techniques to explore less abundant proteins in specific subcellular locations, we intentionally avoided making fine-grained protein-protein interaction regulatory claims, being conscious of this trade-off. Instead, we leveraged our large dataset to tackle broad-scale biological questions without oversimplification. Our findings demonstrate that variations in proteomes rich in high-abundance proteins are robust and well-suited to uncover solid biological insights when analyzed through system-level approaches, as exemplified by evolutionary proteomics and multi-group latent factor analyses, proving their wide utility.

# Lastly, we would like to highlight that among the 7,697 proteins reported in Pra-GE-ATLAS, 45 proteins were identified as transcription factors from 21 different families (e.g. ZF-HD, TCP and LBD), and 53 proteins were identified as transcription regulators from 18 different families (e.g. SWI/SNF, HMG and PHD), according to PlantRegMap and TapScan v4 assignment rules, respectively. This provide further evidence that important regulatory information is not completely missing in Pra-GE-ATLAS.

I do not think that different tissue samples from different batches are suitable for WGCNA analysis, because you will always find that different tissues or batches contribute to the majority of the total variance. This result are obvious and may cause bias.

# We apologize for any confusion. While we fully understand and acknowledge the reviewer’s concerns, we emphasize that batch effects were carefully managed using well-established and widely validated methods from the bioinformatics community, specifically the sva and limma packages. These methods allow for the effective distinction between biological signals (tissue) and technical variation (batch), preventing batch effects from dominating the total variance and being mistaken for true tissue differences.

# To further address the reviewer’s concerns, we will first present recent studies on similar cases, followed by additional analyses using our datasets to demonstrate that batch effects do not impact the reported results.

# In Liesen et al. Developmental pathways underlying sexual differentiation in the U/V sex chromosome system of gian kelp. **Developmental Cell** In Press (2025), the authors introduced genetically male lines exhibiting a feminized phenotype, independently subjected to colchicine treatment from a previous study. As a result, different genetic lines originated from different batches. They addressed batch effects using limma package utilities and conducted WGCNA on the concatenated RNA-seq dataset, successfully extracting biological insights while accounting for technical variation.

# In Wu et al. Multiplexed transcriptomic analyzes of the plant embryonic hourglass. **Nature Communications**. 16(802) (2025), the authors integrated bulk RNA-seq data from four different studies to align it with their organ initiation network and assess it using phylotranscriptomics. Specifically, the proembryo stage, stage 2 embryo, and seedling stages were generated in separate studies/batches, while their single-cell RNA-seq datasets were also produced in two independent batches. As a result, different developmental stages, tissues, and potentially distinct cell types originated from different batches. To correct batch effects, they applied Harmony package utilities and performed both WGCNA and hdWGCNA analyses on the integrated dataset. The resulting co-expression modules reflected developmental biology trends rather than technical artifacts.

# In Yano et al. Comparative genomics of muskmelon reveals a potential role for retrotransposons in the modification of gene expression. **Communications Biology.** 3(1), 2399-3642 (2020), the authors generated new transcriptomic data for anthers, shoot apex, germination, and fruit ripening, which they combined with the Melonet-DB tissue-wide transcriptomes developed by the same group in 2018. Consequently, different tissues originated from different batches. The authors did not mention batch effect removal, as all datasets were generated by the same research group using a consistent RNA extraction protocol and sequencing with closely related Illumina platforms. WGCNA was performed on the combined dataset, successfully identifying ripening-specific modules enriched in well-established ethylene-related genes, demonstrating that biological signals, rather than batch effects, were effectively captured in their co-expression modules.

# In response to reviewer’s comment, we performed additional analyses. Specifically, we compared the impact of batch effects and tissue differences on the total variance by conducting Principal Component Analyses (PCA) on our batch-corrected matrices (which were also used for WGCNA analyses). We then evaluated whether the scores of the top five principal components (PC1-PC5), which capture the majority of the variance (gene expression – 36 %; protein abundance – 94 %), varied significantly across these factors using statistical tests. Since the PCA scores did not follow a normal distribution in all cases, Kruskal-Wallis tests were performed, and p-values were adjusted using the Holm method. If a factor (tissue or batch) significantly altered the PC scores, it indicates that the factor is associated with the variation in those scores, implying it contributes to the variance captured by the evaluated PC.

# Batch effects at the gene expression level were removed using sva package utilities. For illustration purposes, we refer to the stronguest batch effect (Batch2 column) in “Samples\_MetaData” of **Table S1**. The adjusted p-values for the association of tissue and batch with the top five PCs scores are as follows:

**# PC1** (11.2 % variance explained) - Tissue (**7.52E-07**), Batch (**1**)

# **PC2** (8.7 % variance explained) - Tissue (**1**), Batch (**0.45**)

# **PC3** (5.9 % variance explained) - Tissue (**0.078**), Batch (**0.12**)

# **PC4** (5.6 % variance explained) - Tissue (**1.59E-27**), Batch (**0.82**)

# **PC5** (4.5 % variance explained) - Tissue (**0.022**), Batch (**0.15**)

# Only the tissue factor appears to be significantly associated with the variation in PCs scores, particularly in **PC1**, **PC4** and **PC5**, while no significant associations were found for the batch effect. This suggests that tissues differences contribute to the majority of the variance, while batch effects have been effectively mitigated. Furthermore, these results indicate that tissue differences are not confounded by/with batch effects. As cited in the Results section, M03—one of the largest WGCNA co-expression modules associated with tissue differences—includes a tissue that is unevenly distributed across batches (vascular). If batch effects played a major role, this module would be expected to be correlated only in the tissue represented by that batch. However, the shared positive correlation between vascular and bud tissues (which are evenly distributed across batches), while remaining negatively correlated with *Fusarium* stress (also evenly distributed) within this module, further suggests that batch effects are not driving the observed co-expression patterns.

# At the protein abundance level, we do not face the issue described above due to two main reasons. 1) Although not all tissue samples are in the same batch, each tissue has at least some samples represented in the same batch. This ensures that the batch effects are not entirely confounded with tissue type. This allows us to control for batch effects and reduces risk of bias that would occur if tissues were unevenly distributed across batches. 2) Data generation was consistent and standardized. All proteomic datasets were obtained under the direct supervision of our research group and primarily conducted in our department. Protein extraction, pre-fractionation, cleaning and digestion was performed uniformly using always standardized protocols develovep in our laboratory (Valledor et al. A universal protocol for the combined isolation of metabolites, DNA, long RNAs, small RNAs, and proteins from plants and microorganisms. **The Plant Journal.** 79(1), 173-180 (2014); Valledor and Weckwerth. An improved detergent-compatible gel-fractionation LC-LTQ-Orbitrap-MS workflow for plant and microbial proteomics. **Plant Proteomics**. Methods in Molecular Biology, 347-358 (2014)). The stronguest batch effect in these datasets is likely introduced by the Orbitrap platform used. Additionally, we identified an error in the submitted manuscript regarding the software used for batch effect removal in protein data. Batch effects at the protein abundance level were corrected using the limma package (not sva), which has been corrected in the revised manuscript and highlighted accordingly. The adjusted p-values for the association of tissue and batch with the top five PCs are as follows:

**# PC1** (61.3 % variance explained) - Tissue (**0.004**), Batch (**0.14**)

# **PC2** (17.9 % variance explained) - Tissue (**3.36E-06**), Batch (**5.40E-09**)

# **PC3** (8.8 % variance explained) - Tissue (**1**), Batch (**0.0002**)

# **PC4** (4.3 % variance explained) - Tissue (**0.69**), Batch (**1**)

# **PC5** (1.5 % variance explained) - Tissue (**1**), Batch (**0.44**)

# Although both tissue and batch showed significant associations with some PCs scores, PC1 (which captures the majority of variance) scores were significantly associated with tissue, but not with batch. This supports that tissue differences are the primary contributors to the majority of the total variance. While batch effects are linked to minor sources of variation, they are unlikely to confound our WGCNA analyses, as the dominant biological signal comes from tissues, as illustrated above.

# Overall, batch effects in our data exhibited minimal or no associations with the majority of the total variance, especially when compared to the effect of tissue differences. These results strongly support that our key WGCNA findings were not influenced by confounding batch effects, ensuring that the reported results are not biased and remain biologically meaningful.

# To further clarify this, we have updated the Methods section:

**[ Methods, Page 18, Lines 599-612]**

Abundance values were normalised by sample-centric approach and multiplied by the average intensity of all samples. Protein abundances were transformed with a log10(+1.1) for subsequent analyses. [limma v3.56.2 (Ritchie et al., 2015)] was employed to remove abundance unwanted variation. GenEra v1.4 was employed, as mentioned above, using P. radiata proteins as query.

[…]

Weighted Gene Co-expression Analysis (WGCNA) was conducted using WGCNA v1.72-1 (Langfelder and Horvath, 2008) to identify highly co-expressed genes ([sva batch-removed] DESeq2 VST) and proteins ([limma batch-removed] log10(+1.1)).

Please clarify the specific meanings of "juvenile" and "adult" in the MS, such as whether "juvenile" refers to the seedling stage with single-needle fascicles or the vegetative growth stage before entering the reproductive development stage. Usually, the juvenile stage of conifers is very short, and it is often confused with the adult vegetative growth stage.

# We apologize for any confusion and appreciate the reviewer’s helpful comments. The reviewer is absolutely correct that further clarification would be beneficial. Juvenile single-needle fascicles (growth period of 1 cm in length) were sampled from seedlings, while adult multiple-needle fascicles (greater than 12 cm in length) were sampled from adult trees. We have also provided additional context regarding the buds used for the splicing validation, as we noticed this could be also missleading.

**[ Methods, Page 14, Lines 454-466]**

To generate the tissue proteomic dataset, we sampled seedlings (one-year-old) and adult trees, both maintained under routine management at the Plant Physiology Laboratory of the University of Oviedo. The following tissues were collected: roots (growing tips in seedlings), juvenile needles (single-needle fascicles, with a growth period of 1 cm in length, from seedlings), adult needles (mature multiple-needle fascicles, > 12 cm, from trees), and stem apical floral buds (less lignified and mature, from trees). For the tissue splicing evaluation, we additionally collected vegetative buds: juvenile buds (with leaf primordia differentiated into photosynthetically active primary needles around the shoot apical meristem in seedlings) and adult buds (with leaf primordia differentiating into scale leaves, and photosynthetic activity shifting to long needles in more distal positions, in trees). In all cases, three biological replicates for each tissue were constituted by pooling samples from two different plants.

This study pays a lot of attention to alternative splicing in conifers. But, the number of these alternative splicing events seems to be fewer than expected (Akhter, S., et al., Front Plant Sci, 2018. 9: 1625.). Moreover, I am very curious about how much protein-level evidence there is to prove that these different transcripts are truly translated into proteins?

# We sincerely thank the reviewer once again for these valuable remarks. As the comment includes multiple questions, we have addressed them in a point-by-point manner.

**This study pays a lot of attention to alternative splicing in conifers. But, the number of these alternative splicing events seems to be fewer than expected (Akhter, S., et al., Front Plant Sci, 2018. 9: 1625.).**

# As highlighted in previous comments, since we are working with a non-model organism without well-established references, we prioritized high precision over sensitivity in our *de novo* splicing framework. This decision was primarily driven by the fact that, although we conducted critical integrative analyses to uncover new system-level biological insights, the ultimate goal of Pra-GE-ATLAS is to serve as a resource for future user-driven research, enabling the testing of new hypotheses. Given this purpose, we considered it essential to favor precision in the precision-sensitivity trade-off to minimize the risk of false positives, which could mislead future molecular studies.

# While we acknowledge that this approach results in fewer detected events, our thorough review of the cited reference did not reveal a genome- or transcriptome-wide estimation of the splicing landscape or the total number of alternative splicing events. Instead, the study primarily focuses on the number of isoforms within specific transcripts and gene families. We greatly appreciate this work, as it further explores isoform diversity in gymnosperms; however, we believe it is essential to provide both biological and technical context to clarify that the estimations from Pra-GE-ATLAS and the cited article should not be directly compared.

# 1) The initial number of AS events detected in Pra-GE-ATLAS was 73,293 (Type 1 sequence variation bubbles). After applying stringent thresholds and processing, this was refined to 21,280 unique AS events. This number falls within the expected range when compared to recent studies on splicing landscapes. García-Pérez et al. The landscape of expression and alternative splicing variation across human traits. **Cell Genomics**. 3, 100244 (2023), reported 62,269 AS events in humans across a comprehensive collection of tissues, sexes, ages, and other factors. Guiomar et al. Alternative splicing landscapes in Arabidopsis thaliana across tissues and stress conditions highlight major functional differences with animals. **Genome Biology**. 22, 35 (2021), reported around 30,000 AS events in *Arabidopsis thaliana* based on 516 independent RNA-seq studies. Zhang et al. Population-level exploration of alternative splicing and its unique role in controlling agronomic traits of rice. **The Plant Cell**. 36, 10, 4372-4387 (2024), reported around 40,000 AS events across two tissues for 250 acessions from six subpopulations of rice. These comparisons underscore that the AS event stimations in Pra-GE-ATLAS are consistent with those reported in other species.

# 2) The cited reference and Pra-GE-ATLAS differ fundamentally in scope and application. While the cited study employs a highly sensitive, discovery-based approach to recover all possible isoforms for specific genes or gene families, Pra-GE-ATLAS is designed as a resource-based framework to provide a comprehensive view of splicing landscapes. For instance, the cited study allowed transcript isoforms to map to different genomic scaffolds, when these scaffols were very short. Such considerations are not feasible when describing the splicing landscape at the transcriptome- or genome-wide level. Additionally, our estimations, primarily due to our local *de novo* sequence variation approach using KisSplice, are reported at the event-level, meaning that different combinations of AS events can give rise to distinct transcript isoforms. Consequently, AS event counts and transcript isoform numbers are inherently non-comparable. Since multiple AS events can occur simultaneously, the total number of transcript isoforms is expected to exceed the number of AS events.

# 3) We were particularly intrigued by the approach employed in the cited study due to the high number of isoforms found for MADS-box genes: 933 transcript isoforms for 38 putative genes. While isoform numbers varied across gene family members, this would imply that most of the putative genes present over 24 isoforms on average (933/38). For context, the genomes of *A. thaliana* and *O. sativa* do not contain any gene with such high isoform numbers, and more complex genomes, such as humans and *Triticum aestivum*, contain 1,038 and 386 genes with at least 24 isoforms, respectively (plant estimations based on the latest genome version hosted in Phytozome, and human estimations based on the latest genome versions hosted in Ensembl, accessed 15/03/2025). Upon reviewing follow-up research from the cited reference, we found that the approach has recently been published: Westrin et al. ClustTrast: a short read de novo transcript isoform assembler guided by clustered contigs. **BMC Bioinformatics**. 25, 54 (2024). While we appreciate the idea of a highly sensitive approach to recover as many isoforms as possible, several key considerations should be noted. The same authors stated that the method used in the cited reference would not scale to whole-transcriptome assembly. While improvements in the latest publication have enabled scaling to single transcriptomes, memory usage and runtime analyses on lightweight datasets suggests scalability issues for an atlas-level dataset given our read depth. Although the recovery of transcript isoforms is impressive, it comes at the cost of precision, leading to an increased rate of false positives. Notably, most of the reconstructed isoforms (47-78 %) rely on polymorphic variants, such as SNPs. We made the restrictive decision not to include SNPs (see above comment on Type 0 sequence variation bubbles) to specifically assess the splicing landscape and prioritize true positives. Taken together, from a technical perspective, the number of transcript isoforms in the cited reference and the number of AS events described in Pra-GE-ATLAS should not be directly compared. This is because our approach does not consider complex AS events or SNPs and applies more stringent criteria to ensure high precision.

**Moreover, I am very curious about how much protein-level evidence there is to prove that these different transcripts are truly translated into proteins?**

# We would like to once again emphasize that the reviewer's curiosity aligns completely with ours, particularly regarding the proteomic impact of AS event variants. Initially, the inclusion of a proteogenomic module was explicitly mentioned as a future direction. However, due to space limitations and additional considerations discussed below, we chose to summarize it within the future variation module (Discussion section, **Page 13, Lines 444–448**). There are several biological and technical factors that influenced our decision not to address the proteomic impact of the detected AS variations using MS data, and we would like to share our perspective on this matter.

# 1) Trypsin digestion limitations. In shotgun proteomic analyses, proteins are first digested into smaller peptides using trypsin and then analyzed by LC-MS/MS. Trypsin, the most commonly enzyme in MS and widely utilized in our collection, cleaves at the C-terminus of lysine or arginine, generating peptides with optimal length and charge. Peptides spanning exon-exon junctions provide direct evidence of splice variants at the protein level. Interestingly, lysine and arginine are highly enriched at exon-ending or exon-exon junctions of transcripts, making these sites preferred for trypsin digestion. This preference hinders the detection of junction-specific peptides and limits the identification of alternative splicing peptides in proteogenomics analyses using trypsin (Chaudhary et al. Perspective on Alternative Splicing and Proteome Complexity in Plants. **Trends in Plant Science**. 24, 6, 496-506 (2019)). Recent evidence further supported this notion, showing that trypsin produces the lowest ratio of junction-spanning versus exon body peptides among tested proteases (~25% versus 28–32%) (Sinitcyn et al. Global detection of human variants and isoforms by deep proteome sequencing. **Nature Biotechnology**. 41, 1776-1786 (2023)). This observation confirms in silico predictions of trypsin’s limited utility for detecting spliced junction sequences in shotgun proteomics data. Therefore, future proteomic datasets using multiple proteases simultaneously would be highly beneficial for addressing this limitation and deepening our understanding of alternative splicing at the protein level.

# 2) Based on our results in **Figure 2D**, the predicted impact of AS variation on protein sequences—particularly its ability to produce fully functional proteins without introducing premature stop codons—is highly unlikely. While some disrupted proteins may retain partial function, and AS-induced variation could enhance catalytic flexibility by introducing intrinsically disordered regions (Chaudhary et al. **Trends in Plant Science** (2019)), the most probable scenario is that AS variation in this species does not directly contribute to proteome complexity. This aligns with the presence of exceptionally large introns in this clade, as full intron retention would likely prevent successful translation. However, the absence of a direct impact on protein diversity does not imply AS variation lacks functionality. Instead, it is more likely involved to play a role in regulatory processes, such as buffering the metabolic cost of translation during stochastic transcription under stress or modulating gene expression through decay-stability balances, particularly by affecting non-coding regions, such as untranslated regions.

# 3) Given the aforementioned technical and biological trade-offs, as well as the already comprehensive description of the transcriptomic and proteomic modules, we believe that a thorough representation of the proteogenomic impact of variations requires considering both AS events and SNPs together in the upcoming variation module. While we consciously excluded SNP-induced variation in this study to make clear claims about the splicing landscape, we believe that these types of sequence alterations and their impact on the proteome should be analyzed jointly. More specifically, intron expansion in pines appears to be driven by transposon insertions in these regions (Niu et al. **Cell**. (2022)). Since transposons can act as hubs for SNP accumulation, their presence in pine introns presents an intriguing case where transposon content, SNPs, and splicing should be examined together to fully understand their impact on protein variation.

# However, to satisfy the reviewer's interest, we conducted additional analyses. Assuming the limitations of trypsin digestion, we examined how many cases exhibited expression of both primary and alternative protein forms, given that we used the "Final" version of our assembly as the reference database. The classification of transcripts (and proteins) into primary or alternative forms was performed using EvidentialGene, as explained above and in the Methods section, based solely on transcriptome assembly information. Since these preliminary analyses, conducted to address the reviewer's curiosity, are independent of KisSplice, the different transcript isoforms could arise not only from AS events but also from polymorphic variants (SNPs and indels) or closely related paralogs. Although EvidentialGene’s clustering settings are designed to group isoforms together while keeping paralogs separate, this possibility should still be considered in these analyses, as a definitive distinction would require further verification using genome assemblies. Therefore, these results should be understood as reflecting the general impact of similar transcript sequence variation—including, but not limited to, splicing—on the proteome.

# In total, 446 cases exhibited identification and quantification of more than one proteoform, including the primary one. Among these cases, the mean number of alternative proteoforms per primary protein was 1.4, with a standard deviation of 0.92. The maximum number of proteoforms found for a single primary protein was 10. We then computed the Spearman correlation using our batch-corrected abundance data between each pair of primary and alternative proteoforms. Surprisingly, the distribution of correlation values varied widely (mean: 0.25, median: 0.21, SD: 0.5; for absolute values: mean: 0.47, median: 0.46, SD: 0.288). This underscores the fact that the detected proteoforms can play both redundant and divergent roles, with highly case-specific patterns. Among the known functional bins covered, approximately 50% was linked to enzyme activity (32%), protein biosynthesis and homeostasis (13.2%), and redox homeostasis (6.2%).

# We hope that these preliminary analyses satisfy the reviewer's curiosity. Future work involving the upcoming variation module should adopt more comprehensive peptide-level approaches, such as the Perseus plugin used in Sinitcyn et al., **Nature Biotechnology** (2023), which can distinguish the impact of SNPs and AS events on MS data while incorporating genome splice-junction information.

***Reviewer: 2***

*Comments to the Author:*

Roces et al developed Pra-GE-ATLAS to integrate multi-omics database for stress and breeding research of Pinus radiate stress. However, several points require clarification before publication.

# We sincerely appreciate the reviewer's time and effort in assessing our manuscript. We greatly value the attention given to the database limitations and hope that the newly included applications and comments adequately address the reviewer's observations.

As Pra-GE-ATLAS aims to provide multi-omics database for conifer communities. However, the current database version is still lagging behind the convenient use for researchers in the conifer communities. It is nice that you integrated many multi omics data, including transcriptome, proteomics, and alternative splicing. The current version database is more like an additional website for browsing the AS, hub gene of WGCA, stress, and tissue-specific gene expression. However, several functions are frequently and urgently used and not included in the current database version. For example, the gene ontology, KEGG, the prediction of transcription factors, the prediction of regulatory network according to transcription factors, or the prediction of transcription factors from an input of promoters, and etc. For more detailed user-friendly woody plants, including Pinus database, please access the PPGR (Resource for Perennial Plant Genomes and Regulation): https://ngdc.cncb.ac.cn/ppgr/.

# The reviewer is absolutely right. Pra-GE-ATLAS now includes new applications such as GO, KEGG, and gene age enrichment analysis, as well as transcription factor/regulator prediction using PlantRegMap and TAPScan v4 family assignment rules. Additionally, it features regulatory network prediction based on multiple input genes, integrating protein-protein and TF-target interactions from the valuable PPGR resource for *P. taeda*, with the co-expression data already hosted in Pra-GE-ATLAS. Other enhancements include motif scanning, enrichment, and prediction. We sincerely thank the reviewer for the opportunity to improve Pra-GE-ATLAS utilities—the cited PPGR resource was especially helpful as inspiration. Overall, the newly available applications in Pra-GE-ATLAS now encompass all the essential functions commonly required in research, broadening its relevance to a wider audience within the plant science community.

# We have updated the Methods section, Figure 1 and Figure S1 to illustrate the newly included applications:

**[ Methods, Page 19-20, Lines 656-667]**

We developed P(inus)ra(diata)-G(ene)E(xpression)-ATLAS database, a comprehensive multi-omics hub aimed to provide public access to the information generated in this work. Pra-GE-ATLAS features multiple tools: 1) Search section with interactive tables and heatmaps for quick retrieval of protein-, transcript-, splicing event-information. 2) Diamond BLASTP sequence alignment (Buchfink et al., 2021). 3) Fast computation of *P. radiata* orthologs and the mode and strength of selection (dN/dS ratio) based on our consensus assembly, using orthologr (Drost et al., 2015). 4) Functional enrichments. 5) Exploration of global co-expression modules and regulatory networks of target genes, leveraging data from the PPGR database (Yang et al., 2024). 6) Prediction of transcription factors and regulators, based on PlantRegMap and TAPscan v4 assignment rules (Tian et al., 2020; Petroll et al., 2024), along with motif analysis. Pra-GE-ATLAS is available at <https://rocesv.github.io/Pra-GE-ATLAS>.

**[ References]**

Romy, P., Varshney, D., Hiltemann, S., Finke, H., Schreiber, M., de Vries, J. and Rensing, S.A. (2025). Enhanced sensitivity of TAPscan v4 enables comprehensive analysis of streptophyte transcription factor evolution. *The Plant Journal, 121*, e17184.

Tian, F., Yang, D.C., Meng, Y.Q., Jin, J. and Gao, G. (2020) PlantRegMap: charting functional regulatory maps in plants. *Nucleic Acids Research,* *48*, D1104-D1113.

Yang, S., Zong, W., Shi, L., Li, R., Ma, Z., Ma, S., Si, J., Wu, Z., Zhai, J., Ma, Y., Fan, Z., Chen, S., Huang, H., Zhang, D., Bao, Y., Li, R. and Xie, J. (2024) PPGR: a comprehensive perennial plant genomes and regulation database. *Nucleic Acids Research, 52*, D1588-DD1596.

I am curious why the authors chose Pinus radiate instead of loblolly pine or another more widely distributed pine.

# We appreciate reviewer’s curiosity and, to address it, we outline several key reasons behind our choice of *P. radiata* as the reference species for Pra-GE-ATLAS:

# 1) Economic factors: While *P. taeda* dominates as a forestry resource in the southeastern United States, *P. radiata* is one of the most widely planted pine species globally. It has expanded on an international and intercontinental scale, particularly in New Zealand, Australia, Chile, Argentina, Uruguay, South Africa, Kenya and Spain, making it a crucial species in global forestry.

# 2) Ecological factors: *P. radiata* exhibits a relatively broad range of adaptability, particularly to diverse soil conditions, and has one of the fastest growth rates among pines outside its native range. Additionally, its suitability for plantation systems enables shorter harvesting cycles than many other pine species, making it highly accessible as a complementary organisms for comparative research.

# 3) Conservation and genetic interest: Despite its commercial success, *P. radiata* is endangered in its native range (California and Mexico) due to habitat loss and diseases susceptibility (primarily *Fusarium*). By providing Pra-GE-ATLAS as a public resource, we aim to support future research on its population decline and the genetic factors influencing its survival. Among the key conservation efforts for this species, research for immunity and resistance to pathogens is particularly emphasized.

# 4) Stress response studies. Among pines, *P. radiata* is highly sensitive to biotic and abiotic stressors, making it a valuable model species for studying gene regulation in stress tolerance and acclimation. This research is particularly relevant as *P. radiata* represents gymnosperms and conifers, which possess unique and intriguing molecular features. Additionally, *P. radiata* could serve as a suitable model for studying conifer/gymnosperms immunity, as it appears to display compatibility and susceptibility to a broad range of pathogens.

# 5) Data diversity and research group. Due to its high stress sensitivity and widespread availability, *P. radiata* has been extensively used over time for generating molecular datasets, resulting in a highly diverse collection. Additionally, most of the proteomics datasets were produced by our research group or conducted under our direct supervision. These datasets cover key factors of broad interest to the plant science community, such as disease-resistant genotypes and intergenerational memory responses, which, to our knowledge, are not available for other pine species. However, due to their lack of standardization and the absence of a holistic analytical framework, these datasets have received limited attention from the plant science community—an issue that Pra-GE-ATLAS aims to address.

The expression regulator browser is not accessed from my computer.

# We apologize for any issues this may have caused. All applications are hosted on shinyapps.io, a widely recognized and stable server, and none of our applications have encountered this issue before. Additionally, we find it unusual that only the Regulation Browser was inaccessible from the reviewer’s computer, as all applications are hosted on the same service. To further assess the Regulation Browser’s stability, we conducted extensive testing between the time we received the reviewer’s feedback and March 25th. During this period, we accessed and used the Regulation Browser 30 times from three different computers and across the five most commonly used web browsers (Mozilla Firefox, Safari, Opera, Chrome, and Internet Explorer). In all cases, we encountered no issues with either accessing or running the application.

# The only plausible explanation for the reviewer's inability to access the Regulation Browser could be temporary maintenance on shinyapps.io during the visit, which would have been an unfortunate coincidence.

In addition, I am not sure it works, but it is strongly recommended that your website have an IGV browser if it can afford it. This will help the audience visualize the expression pattern of their interested genes easily.

# We completely agree with the reviewer that a genome browser could be beneficial for future users of Pra-GE-ATLAS. In fact, in another of our published databases, PlantFUNCO (<https://rocesv.github.io/PlantFUNCO/>), we have successfully integrated genome browsers for three species with different tracks, and we have found them to be incredibly useful. However, as the reviewer anticipated, after testing, we determined that hosting a genome browser on our website is not feasible due to the high fragmentation and large size of the genome. Even in the resource cited by the reviewer, PPGR, which has extensive genome-hosting capabilities, genome visualization is only possible if the assembly reaches the chromosome level and contains no more than 36 scaffolds/contigs. Given these limitations, a potential workaround would be to provide a genome browser view for a small subset of scaffolds, but this would exclude the majority of genes from visualization. Since Pra-GE-ATLAS is currently designed as a gene-centric resource, we do not make strong claims regarding intergenic regions. Rather than providing incomplete genome browser functionality, we believe it is more effective to visualize expression patterns through the interactive tables, heatmaps, networks, and regulation browser available in the database, offering a more comprehensive view of the expression landscape.