BMEG 524 Project

Jack Chiang, Louis Lax-Roseman

2025-04-02

Contents

Introduction

More than 400 million years ago, plants evolved from algae to land plant lineages, where they started to survive on land (Weng & Chapple, 2010). During the evolution process, the genetic composition has varied drastically such that land plants can adapt and survive in a different environments (Panchy, Lehti-Shiu & Shiu, 2016). Transcription factors (TF) followed such changes in order to allow the plant to adapt and survive (Movahedi, Kadkhodaei & Yang, 2024). As plant species from different lineages were used as model organisms, researchers were gaining knowledge in the evolution pathway of TFs, yet, there was still missing linkage between the green algae lineage and the land plants.

In 2013, Sharma, Bhalla, & Singh first used *Marchantia polymorpha* (*M. polymorpha*), which is a liverwort, to perform RNA-sequencing (RNA-seq), hoping to find the TF families present in different stages in liverwort (Sharma ,Bhalla, & Singh, 2013). Liverworts are plants which first colonized land (Bowman et al., 2022). Being one of the earliest land plant lineages, it can thus be considered as one of the most ancient land plant lineages to elucidate the evolution to land plants, connecting the missing linkage between green algaes and land plants. After analysing the RNA-seq data, Sharma, Bhalla, & Singh compared the number and grouping of TF to other plant lineages so as to discover the evolution of TF families. They discovered that there are a number of TFs which are unique to land plants. On top of that, several TFs such as SAP only emerged in ferns and seed plants (Sharma ,Bhalla, & Singh, 2013). This revealed the evolution of TF families as well as their impact in adapting to the local environment when withstanding diverse stresses. They further confirmed the role of transcription through looking at the roles of various transcriptional factors during the development in *M. polymorpha*. Despite the novel discoveries of the evolutionary role of TF families, there were a few flaws which could be further improved. As more and more researchers investigate *M.polymorpha*, the whole genome sequencing of *M. polymorpha* has already been performed. Instead of using *de novo* assembly, we sought to map with the existing genome of *M. polymorpha* to increase the accuracy of the transcriptome from this data. On top of that, more downstream analysis can be performed to further provide evidence to support the significance of certain evolutions of TF families.

In this project, we reanalyzed the data set obtained from six different stages of *M. polymorpha*, including male plants in immature reproductive stage (IM_Tak1); female plants in immature reproductive stage (IM_Tak2); male plants in mature reproductive stage (M_Tak1); female plants in mature reproductive stage (M_Tak2); male plant in vegetative stage (V_Tak1); and female plants in vegetative stage (V_Tak2). We used a more robust reference genome for increasing accuracy, and performing more thorough downstream analysis including cross species comparison of TF families and GO-analysis. We hope to discover more information from this dataset such that we can further elucidate the functions of various TF families and their evolutionary relationship along the green lineage.

Our analysis aims to determine whether reference-based alignment improves transcript quantification compared to de novo assembly and to explore how TF family composition changes across developmental stages of *M. polymorpha*.

Scope of the Re-analysis

In this project, we re-analyze the dataset obtained from six different developmental stages of *M. polymorpha*, but with several key improvements:

1. **Reference-based alignment**: Instead of using de novo assembly, we map reads to the existing *Marchantia* reference genome to increase the accuracy of transcriptome analysis.
2. **Updated transcription factor classification**: We utilize the most recent version of PlantTFDB database to identify and classify transcription factors.
3. **Extended downstream analyses**: We perform additional analyses including gene ontology (GO) enrichment and expression pattern clustering to further explore the relationships of various transcription factor families. # Method

RNA-Seq Analysis

RNA-seq data from different stages and sexes of *M. polymorpha* were downloaded from SRA accession SRP029610 (Sharma, Bhalla, & Singh, 2013). Raw sequencing reads were first processed by trimming adapters using fastp v0.12.4 (Chen et al., 2018). The reads were then aligned to the *M. polymorpha* genome; the genome and GTF files were obtained from the *M. polymorpha* Tak-1 reference genome v5.1 revision 1 available at marchantia.info (https://marchantia.info) (Wang et al., 2023). The alignment and annotation were performed with STAR 2.7.9a (Dobin et al., 2013), using the following arguments:

> "–runMode alignReads –outFilterMultimapNmax 10 –outFilterMismatchNoverLmax 0.05 –quantMode GeneCounts –outSAMtype BAM SortedByCoordinate –outReadsUnmapped Fastx"

Choice of Parameters

- **–outFilterMultimapNmax 10:**
  This parameter allows up to 10 multiple alignments per read. It was selected due to the high repetitive content in plant genomes, ensuring that reads from conserved gene families are not discarded unnecessarily. This helps retain data from genes that are part of large, conserved families.
- **–outFilterMismatchNoverLmax 0.05:**
  This setting limits the number of allowable mismatches relative to the read length, ensuring high-confidence alignments while accommodating natural genetic variation.
- **–quantMode GeneCounts:**
  This option applies gene count quantification during alignment, providing a streamlined workflow in which read counts are directly associated with genomic features.

Transcript quantification, expressed in transcripts per million (TPM), was performed using the salmon v1.10.1 tool on the trimmed reads from fastp v0.12.4 (Patro et al., 2017). The quantification used the argument "–l A" (which auto-detects the library type), enabling accurate and bias-aware estimation of transcript abundance. The quantified transcripts were then loaded into RStudio v2024.09.0+375 via tximport v1.30.0 (Soneson et al., 2015). Finally, the *M. polymorpha* Tak-1 reference genome v5.1 revision 1 from marchantia.info (https://marchantia.info) was processed using GenomicFeatures v1.54.4 to generate an annotated transcript database that maps the quantified reads to corresponding transcript IDs in *M. polymorpha* (Lawrence et al., 2013).

Enrichment Analysis

An upset plot of transcription factor (TF) families shared among different stages of *M. polymorpha* was generated using the annotated reads from the RNA-seq analysis. This plot was created with the ComplexHeatmap package (v2.23.1), with the combination mode set to "Distinct" (Gu et al., 2016).

Gene counts from the annotation were subsequently used to perform gene set enrichment analysis (GSEA) for biological processes within the Gene Ontology (GO) framework, utilizing the GO.db package (v3.19.1) (Calson, 2024). The GO annotation files for *M. polymorpha* were downloaded from marchantia.info. Fisher's exact test was used to calculate p-values for the GO analysis, and the Benjamini-Hochberg correction was applied to control the false discovery rate (Benjamini & Hochberg, 1995).

Transcription Factor Family Selection and Evolutionary Comparative Analysis

To compare TF families across distinct evolutionary groups, TF annotations were obtained from the Plant Transcription Factor Database (PlantTFDB v5.0). This database provides extensive, standardized TF classifications across numerous plant and algal species (Jin et al., 2017), facilitating reliable evolutionary comparisons.

Representative species were selected from key evolutionary transitions spanning green algae through flowering plants, including algae (*Chlorella variabilis* NC64A, *Chlamydomonas reinhardtii*, *Coccomyxa subellipsoidea* C-169, *Micromonas pusilla* CCMP1545, *Ostreococcus lucimarinus*, *Ostreococcus* sp. RCC809, *Ostreococcus tauri*), bryophytes (*Marchantia polymorpha*, *Physcomitrella patens*), a lycophyte (*Selaginella moellendorffii*), gymnosperms (*Picea abies*, *Picea menziesii*, *Picea sitchensis*), and angiosperms (*Arabidopsis thaliana*, *Zea mays*, *Populus trichocarpa*, *Vitis vinifera*). TF family sizes for each species were directly sourced from PlantTFDB.

To visualize TF family distribution, a heatmap was generated using the R package **pheatmap** (Kolde, 2019), clearly identifying conserved TF families and potential lineage-specific expansions or losses. Log fold-changes in TF family sizes across evolutionary transitions (Algae → Liverwort → Moss → Lycophyte → Gymnosperm → Eudicot → Monocot) were calculated and displayed via an additional heatmap (Kolde, 2019), highlighting significant evolutionary shifts.

A stacked bar chart illustrating the proportional representation of the 15 major TF families across evolutionary stages (algae to angiosperms) was created using **ggplot2** (Wickham, 2016). This visualization complemented the heatmaps, providing quantitative insights into TF family dynamics. All TF family analyses were conducted in R (v4.3.1; R Core Team, 2024) using RStudio, with additional R packages **dplyr** and **tidyr** employed.

Results

RNA-Seq Processing

After adapter trimming with fastp v0.12.4, the reads were suitable for downstream analysis. The overall sequence length and duplication rate were below 50% (Fig. 1A–1D). The Phred scores of the reads were consistently over 20, with an overall quality peak around a Phred score of 36 (Fig.

1E, 1F). The base pairs were of good quality, displaying an appropriate GC content and a negligible proportion of bases called as N (Fig. 1G, 1H). Less than 1% of the reads were overrepresented, and adapter contamination was below 0.1% (Fig. 1I, 1J). The mapping quality was also robust. The number of uniquely mapped reads exceeded 5,500,000, which is over 85% of the total reads. The mismatch rate was less than 0.5%, indicating that the mapped reads were suitable for downstream analysis (Fig. 1K).

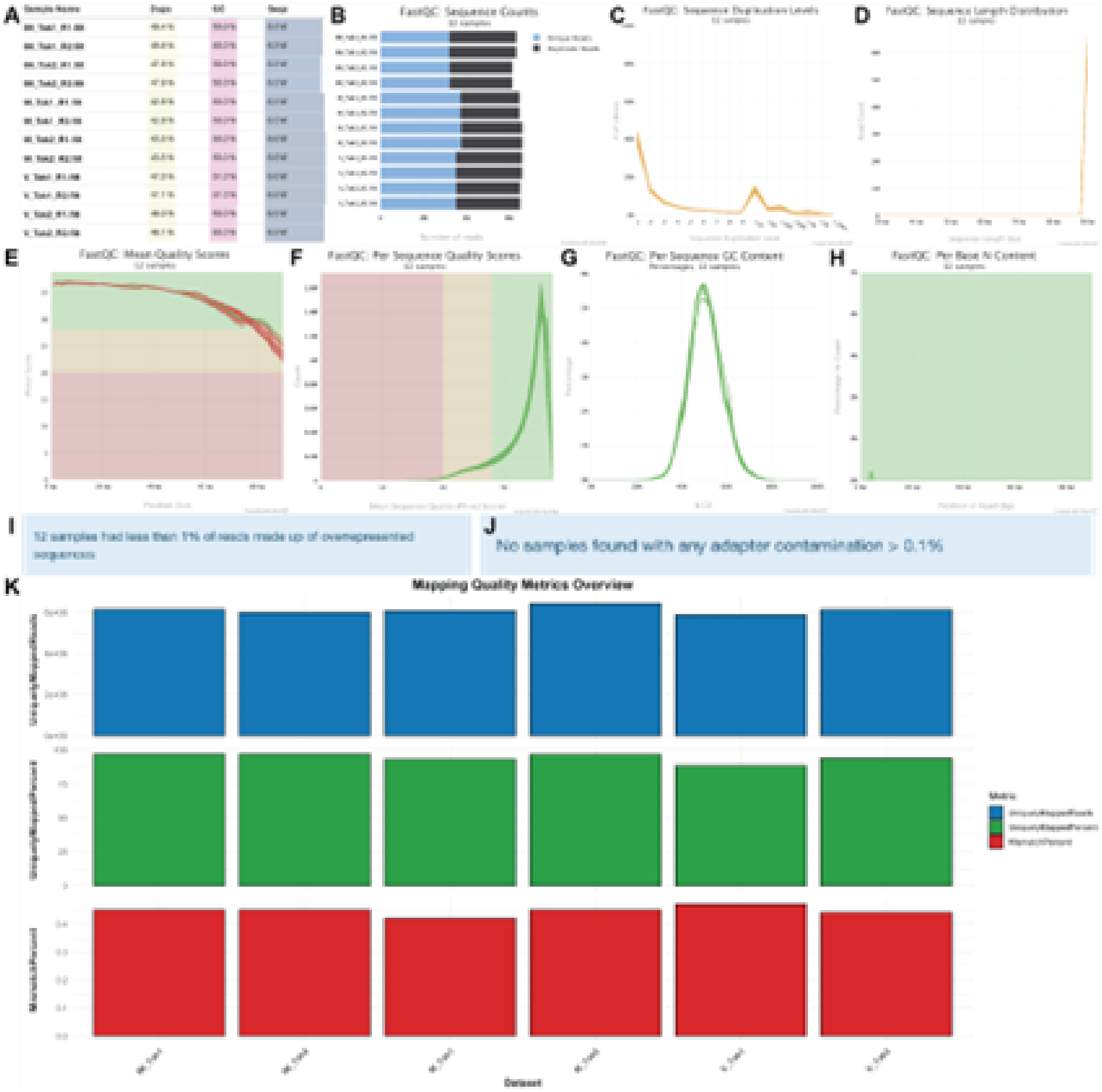

Figure 1. The Quality Control on Raw Sequence Reads and Mapping Efficiency

Fig. 1A–1J show the quality control of the raw sequence reads after processing with fastp v0.12.4, fastqc, and multiqc (obtained from file:///Users/jackchiang/Desktop/BMEG524/multiqc_processed/multiqc_report.html). Fig. 1K shows the mapping quality after alignment with STAR 2.7.9a.

IM_Tak1 refers to male plants in the immature reproductive stage; IM_Tak2 refers to female plants in the immature reproductive stage; M_Tak1 refers to male plants in the mature reproductive stage; M_Tak2 refers to female plants in the mature reproductive stage; V_Tak1 refers to male plants in the vegetative stage; V_Tak2 refers to female plants in the vegetative stage.

Transcription Factor Family Comparison in Different Stages of *M. polymorpha*

Most TFs were co-expressed across the different stages, with 258 TFs common to all stages tested (Fig. 2). Female plants in the mature reproductive stage had the highest number of uniquely expressed TFs, with 25 in this group (Fig. 2). The remaining intersections contained a limited number of TFs, with no group comprising more than five TFs (Fig. 2). In the reproductive stage unique to females, predominantly bHLH and MYB TFs were expressed. In contrast, the unique TF expressed in male plants in the immature reproductive stage encoded the protein MpERF19. In the mature reproductive stages of both sexes, two TFs—MpERF23 and MpASLBD10—were uniquely expressed. In the overall reproductive stage (immature and mature, male and female), three TFs (MpbHLH22, MpbHLH34, and MpMYB13) were uniquely expressed.

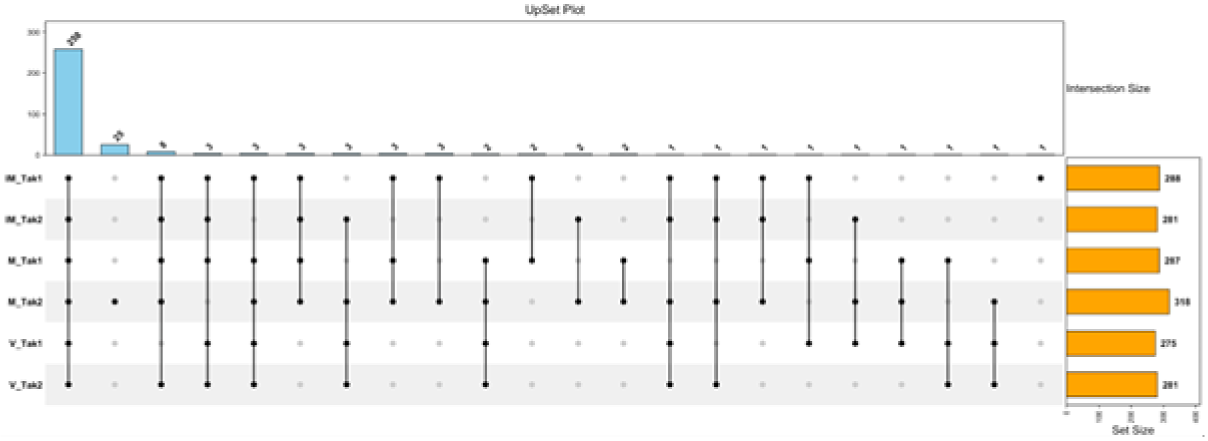

Figure 2. Upset Plot of Transcription Factors in 6 Different Stages of *M. polymorpha*

Figure 2 shows the upset plot of TFs in six different stages.
IM_Tak1 refers to male plants in the immature reproductive stage; IM_Tak2 refers to female plants in the immature reproductive stage; M_Tak1 refers to male plants in the mature reproductive stage; M_Tak2 refers to female plants in the mature reproductive stage; V_Tak1 refers to male plants in the vegetative stage; V_Tak2 refers to female plants in the vegetative stage.

In the GO analysis plot, the six datasets predominantly shared the same GO terms for biological processes, possibly because many TFs are commonly expressed within the same group (Fig. 3). The shared GO terms with significant enrichment were related to DNA binding and the regulation of proteins through interactions with other proteins or metal ions (Fig. 3). This is predictable since the primary function of TFs is to bind to DNA and regulate transcriptional activities. Additionally, processes such as protein kinase activity, hormone response, and cell cycle regulation were common across all six datasets (Fig. 3). However, certain datasets contained unique biological processes; for example, in the vegetative stage of both male and female plants, zinc ion binding was observed (Fig. 3). This may be because these groups mainly contain more transcripts related to zinc finger TFs, such as the C2H2 family (involved in stress tolerance) or the GATA family (involved in plant development) (Han et al., 2020; Schwechheimer, Schröder & Blaby-Haas, 2022).

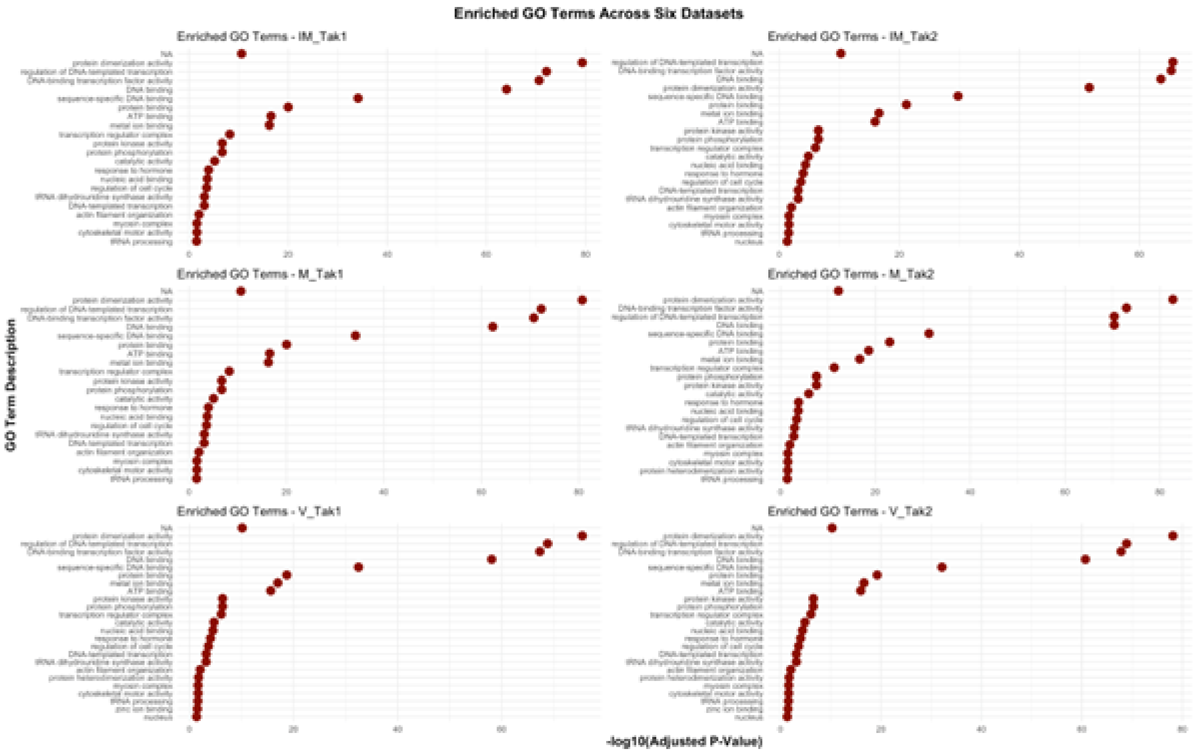

Figure 3. GO Analysis of Transcription Factors in Different Stages of *M. polymorpha*

Figure 3 shows the GO analysis of the TFs in different stages of *M. polymorpha.*
IM_Tak1 refers to male plants in the immature reproductive stage; IM_Tak2 refers to female plants in the immature reproductive stage; M_Tak1 refers to male plants in the mature reproductive stage; M_Tak2 refers to female plants in the mature reproductive stage; V_Tak1 refers to male plants in the vegetative stage; V_Tak2 refers to female plants in the vegetative stage.

Discussion of Transcriptomes

Unlike Sharma, Bhalla, & Singh, we used a reference genome to map raw reads for higher precision and coverage. The use of a reference genome increased the number of mapped reads significantly; while the *de novo* assembly identified approximately 45,000 unique transcripts, the reference-based approach located around 6,000,000 unique transcripts (Sharma, Bhalla, & Singh, 2013). Using a reference genome greatly increased the read counts, thereby reducing the loss of information from the raw reads.

Furthermore, instead of merely counting the number of reads per TF, we performed a more thorough analysis of TF profiles across different stages of *M. polymorpha.* We first identified both commonly and uniquely expressed TFs across all six stages. Although most TFs were commonly expressed, certain stages exhibited unique TF expression. For example, in the reproductive stage, bHLH and MYB TF families were predominantly expressed (Fig. X). Both families are large and functionally diverse, playing roles in growth and development, metabolism, hormone signaling, and responses to biotic and abiotic stress (Wu et al., 2024). In the vegetative stage, zinc finger TFs appear to dominate, with families in this group highly expressed (Fig. X, Fig. X). Zinc finger TFs are also an extensive family, with roles in stress tolerance and development (Han et al., 2020; Schwechheimer, Schröder & Blaby-Haas, 2022). These examples indicate that although similar biological processes occur in different stages, the distinct internal and external environments require different sets of TFs. Through this analysis, we not only unraveled the TF profiles in different stages but also predicted which TFs are more prevalent in each stage.

Unfortunately, the public dataset contains only one sample per condition, without any replicates. Consequently, the results we obtained cannot be claimed as statistically significant, even though generating pseudo-replicates for further analysis is possible.

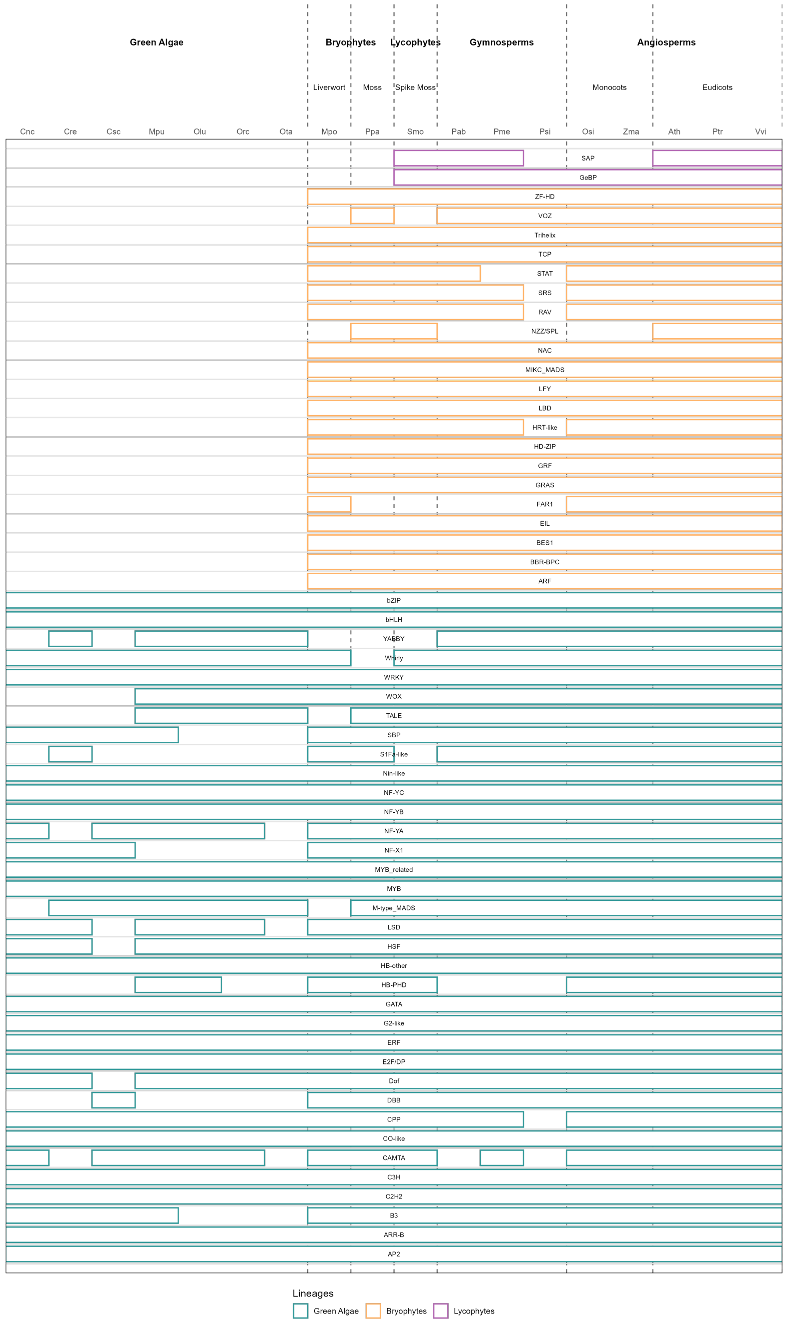

Figure 4. Comparative Distribution of Gene/Transcription Factor Families Across Multiple Plant Species

Each column corresponds to a different plant or algal species (Chlorella variabilis NC64A [Cnc], Chlamydomonas reinhardtii [Cre], Coccomyxa subellipsoidea C-169 [Csc], Micromonas pusilla CCMP1545 [Mpu], Ostreococcus lucimarinus [Olu], Ostreococcus sp. RCC809 [Orc], Ostreococcus tauri [Ota], Marchantia polymorpha [Mpo], Physcomitrella patens [Ppa], Selaginella moellendorffii [Smo], Picea abies [Pab], Picea menziesii [Pme], Picea sitchensis [Psi], Oryza sativa subsp. indica [Osi], Zea mays [Zma], Arabidopsis thaliana [Ath], Populus trichocarpa [Ptr], Vitis vinifera [Vvi]), arranged roughly in evolutionary order from algae to angiosperms. Each row indicates a distinct transcription factor (TF) family (e.g., WRKY, MYB, GRF, ARF, etc.). Gaps in the bars indicate the absence of a TF family in one or more species.

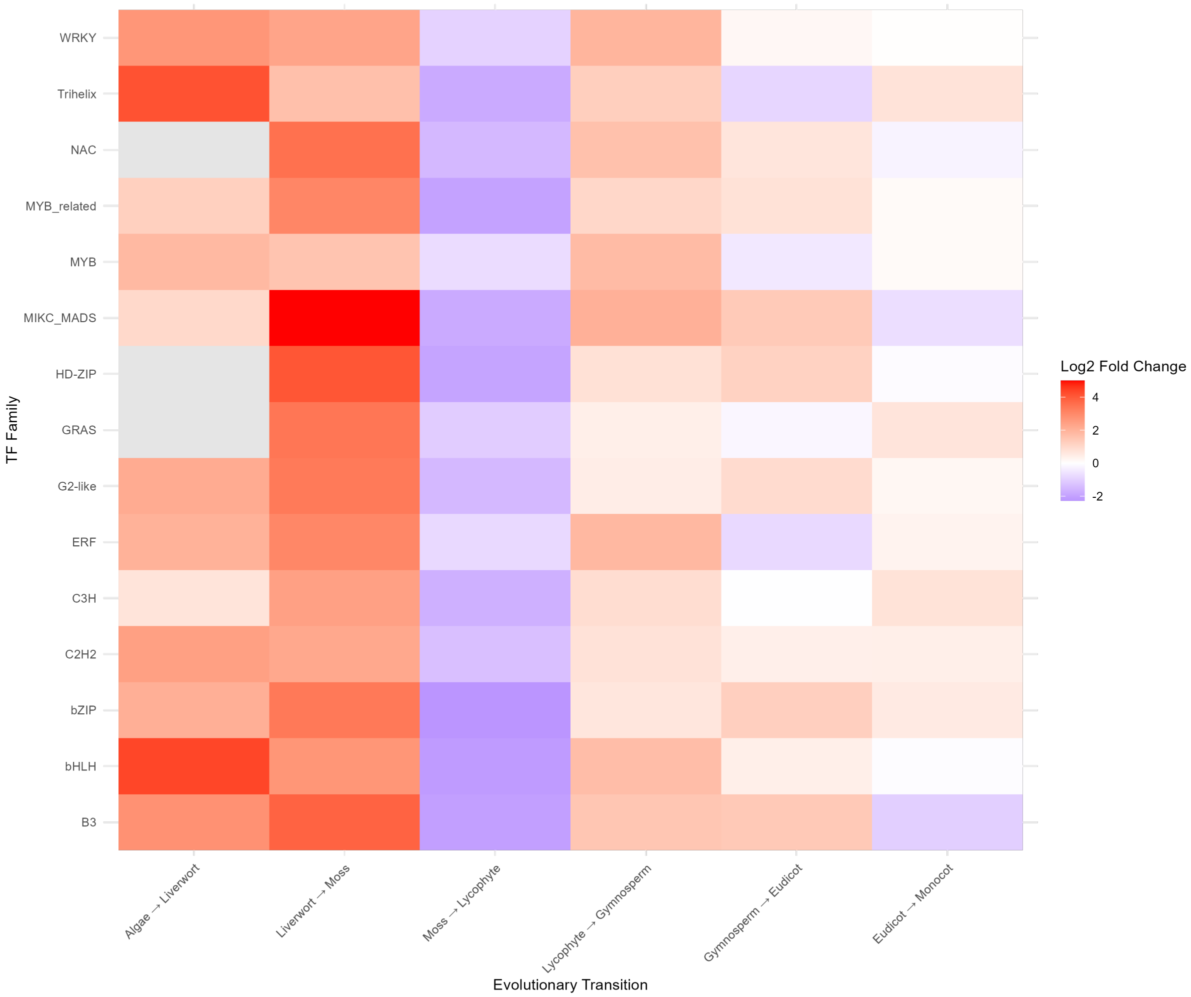

Figure 5. Heatmap showing the log fold change in transcription factor family sizes across major plant evolutionary transitions (Algae → Liverwort → Moss → Lycophyte → Gymnosperm → Eudicot → Monocot). Warmer (red) cells indicate expansion, cooler (blue) cells indicate contraction, white signifies minimal change, and gray denotes missing data or an uncomputable ratio.

Transcription Family Results

Our analysis of transcription factor (TF) families across plants and algae reveals patterns that illustrate the evolutionary history of these genes and their roles in shaping plant adaptation to terrestrial environments.

TF Family Distribution Across Plant Lineages (Figure 4)

Figure 4 summarizes the distribution of TF families over evolutionary time. Certain families—such as ARF, Trihelix, GRAS, GRF, and MIKC_MADS—emerged with the initial colonization of land, as evidenced in early-diverging lineages like liverworts (*Marchantia polymorpha*) and mosses (*Physcomitrella patens*). Their appearance indicates a critical role in helping plants cope with terrestrial conditions.

In contrast, families such as bZIP, bHLH, MYB-related, ERF, and GATA are present across nearly all groups—from algae to flowering plants—highlighting their ancient origins and continued importance. This conservation suggests that these families rapidly acquired essential functions that remained vital throughout subsequent evolutionary stages. For bHLH, previous research by Carretero-Paulet et al. (2010) examined the stability and significance of the bHLH family throughout the plant kingdom.

Finally, two TF families, such as SAP and GeBP, appear only from lycophytes onward. The GeBP family, consistently present from lycophytes to flowering plants, likely played a crucial role in the evolution of advanced vascular structures (Wu et al., 2024).

Patterns of TF Family Expansion and Contraction (Figure 5)

Notably, in Figure 5, MIKC_MADS, Trihelix, and bHLH are shown to expand during the transition from algae to liverworts. This supports the idea that early land plants developed new genes or duplicated existing ones to adapt to terrestrial environments (Shiu et al., 2005).

Families like NAC and MYB expanded during later evolutionary stages, such as the transition from mosses to lycophytes or from gymnosperms to flowering plants. These expansions may be associated with whole-genome duplications that introduce greater complexity to plant genomes (Vanneste et al., 2014).

Some TF families experienced periods of contraction. For example, the Trihelix and HD-ZIP families contracted during the transition from liverworts to mosses, while MIKC_MADS genes decreased between gymnosperms and eudicots, suggesting that evolution sometimes favored streamlining gene sets.

Quantitative Changes in TF Family Abundance (Figure 6)

From algae to liverworts, there's a notable decrease in the abundance of the C3H family and a slight reduction in MYB-related genes, while the bHLH and Trihelix families expanded significantly—likely assuming greater roles in terrestrial adaptation. These proportional shifts agree with the trends observed in Figure 6 and indicate that certain TF families became increasingly important during the water-to-land transition.

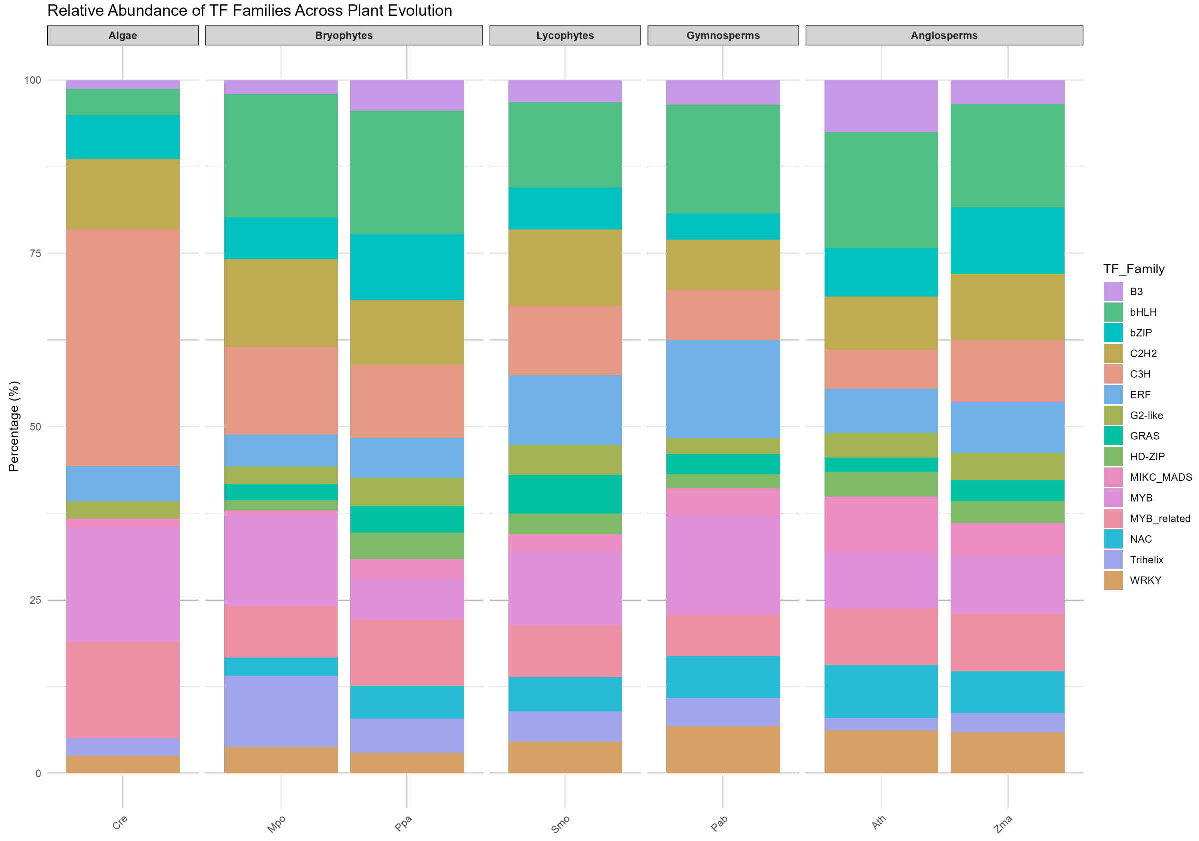

Figure 6. Stacked bar chart illustrating the relative abundance (percentage) of 15 transcription factor (TF) families in representative plant/algal species spanning key evolutionary groups. Each bar depicts the proportion of TF families for one species, grouped by broader phylogenetic category. The species abbreviations (in order shown) are: Cre = *Chlamydomonas reinhardtii* (alga); Mpo = *Marchantia polymorpha* (liverwort); Ppa = *Physcomitrella patens* (moss); Smo = *Selaginella moellendorffii* (lycophyte); Pab = *Picea abies* (gymnosperm); Ath = *Arabidopsis thaliana* (eudicot); Zma = *Zea mays* (monocot). Reductions in C3H and MYB-related TFs from algae to *M. polymorpha* contrast with substantial expansions in bHLH and Trihelix families, indicating shifts in TF composition as plants adapted to terrestrial environments.

Summary

Our figures illustrate how TF family evolution involved early innovations and later refinements. Early gene duplications in MIKC_MADS, Trihelix, and GRAS likely equipped ancestral plants with the genetic tools necessary to manage terrestrial stresses. Broadly conserved families like bHLH and bZIP reflect a balance between preserving essential functions and facilitating new adaptations. These molecular innovations provide plants with the flexibility required to colonize and thrive in diverse terrestrial habitats.

Conclusion

In this project, we reanalyzed a public dataset concerning the transcript profiles of TFs in six different stages of *M. polymorpha*. We found that using a reference genome greatly increased the mappability of raw reads, thereby reducing information loss. We also plotted the potential relationships among certain TF families across the different stages of *M. polymorpha* using an upset plot and subsequent GO analysis—an aspect not fully investigated by the original authors.

Unfortunately, since the public dataset lacks replicates, the results we obtained cannot be claimed as statistically significant, even though generating pseudo-replicates for further analysis is possible.

Nonetheless, by reanalyzing the public data generated by Sharma, Bhalla, & Singh, we achieved higher coverage of TF activity, reconstructed the emergence of TFs across plant lineages from red and green algae to flowering plants, and uncovered potential relationships between TF families and various developmental stages of *M. polymorpha.* Generating replicates—even pseudo-replicates—could enhance statistical significance and allow for additional analyses (including differential expression analysis) to provide a more comprehensive understanding of both the evolutionary and physiological significance of various TF families.

In summary, our study confirms that improved mapping using a reference genome not only recovers more transcript information but also provides novel insights into TF family evolution that were not fully explored in the original work. Future efforts incorporating replicates and advanced statistical analyses will further enhance our understanding of the complex regulatory networks that have shaped plant adaptation throughout evolutionary history.

References

1. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B, 57(1), 289–300.

2. Bowman, J. L., Arteaga-Vazquez, M., Berger, F., Briginshaw, L. N., Carella, P., Aguilar-Cruz, A., Davies, K. M., Dierschke, T., Dolan, L., Dorantes-Acosta, A. E., Fisher, T. J., Flores-Sandoval, E., Futagami, K., Ishizaki, K., Jibran, R., Kanazawa, T., Kato, H., Kohchi, T., Levins, J., Lin, S. S., … Zachgo, S. (2022). The renaissance and enlightenment of Marchantia as a model system. The Plant Cell, 34(10), 3512–3542. https://doi.org/10.1093/plcell/koac219

3. Carlson, M. (2024). GO.db: A set of annotation maps describing the entire Gene Ontology (Version 3.19.1) [R package].

4. Carretero-Paulet, L., Galstyan, A., Roig-Villanova, I., Martínez-García, J. F., Bilbao-Castro, J. R., & Robertson, D. L. (2010). *Genome-Wide Classification and Evolutionary Analysis of the bHLH Family of Transcription Factors in Arabidopsis, Poplar, Rice, Moss, and Algae.* Plant Physiology, 153(3), 1398–1412. https://doi.org/10.1104/pp.110.153593

5. Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. Bioinformatics, 34, i884–i890.

6. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics, 29, 15–21.

7. Fisher, R. A. (1922). On the interpretation of ^{2} from contingency tables, and the calculation of P. Journal of the Royal Statistical Society, 85(1), 87–94.

8. Gu, Z., Gu, L., Eils, R., Schlesner, M., & Brors, B. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics, 32(18), 2847–2849.

9. Han, G., Lu, C., Guo, J., Qiao, Z., Sui, N., Qiu, N., & Wang, B. (2020). C2H2 zinc finger proteins: Master regulators of abiotic stress responses in plants. Frontiers in Plant Science, 11, 115. https://doi.org/10.3389/fpls.2020.00115

10. Jin, J., Tian, F., Yang, D. C., Meng, Y. Q., Kong, L., Luo, J., & Gao, G. (2017). PlantTFDB 4.0: Toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Research, 45(D1), D1040–D1045. https://doi.org/10.1093/nar/gkw982

11. Kolde, R. (2019). pheatmap: Pretty heatmaps (Version 1.0.12) [R package]. Retrieved from https://github.com/raivokolde/pheatmap

12. Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., & Carey, V. J. (2013). Software for computing and annotating genomic ranges (A. Prlic, Ed.). PLoS Computational Biology, 9, e1003118.

13. Movahedi, A., Kadkhodaei, S., & Yang, L. (2024). Editorial: Transcriptional regulation and posttranslational modifications in plant growth and development under abiotic stresses. Frontiers in Plant Science, 15, 1454335. https://doi.org/10.3389/fpls.2024.1454335

14. Panchy, N., Lehti-Shiu, M., & Shiu, S. H. (2016). Evolution of gene duplication in plants. Plant Physiology, 171(4), 2294–2316. https://doi.org/10.1104/pp.16.00523

15. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods, 14, 417–419.

16. R Core Team. (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

17. Schwechheimer, C., Schröder, P. M., & Blaby-Haas, C. E. (2022). Plant GATA factors: Their biology, phylogeny, and phylogenomics. Annual Review of Plant Biology, 73, 123–148. https://doi.org/10.1146/annurev-arplant-072221-092913

18. Sharma, N., Bhalla, P. L., & Singh, M. B. (2013). Transcriptome-wide profiling and expression analysis of transcription factor families in a liverwort, Marchantia polymorpha. BMC Genomics, 14, 915. https://doi.org/10.1186/1471-2164-14-915

19. Soneson, C., Love, M. I., & Robinson, M. D. (2015). Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences. F1000Research, 4, 1521.

20. Wang, L., Wan, M. C., Liao, R. Y., Xu, J., Xu, Z. G., Xue, H. C., Mai, Y. X., & Wang, J. W. (2023). The maturation and aging trajectory of Marchantia polymorpha at single-cell resolution. Developmental Cell, 58(15), 1429–1444.e6. https://doi.org/10.1016/j.devcel.2023.05.014

21. Weng, J. K., & Chapple, C. (2010). The origin and evolution of lignin biosynthesis. The New Phytologist, 187(2), 273–285. https://doi.org/10.1111/j.1469-8137.2010.03327.x

22. Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

23. Wu, J., Liu, R., Xie, Y., Zhao, S., Yan, M., Sun, N., Zhan, Y., Li, F., Yu, S., Feng, Z., & Li, L. (2024). *Association of GhGeBP genes with fiber quality and early maturity related traits in upland cotton.* BMC Genomics, 25(1), 1058. https://doi.org/10.1186/s12864-024-10983-y

24. Wu, X., Xia, M., Su, P., Zhang, Y., Tu, L., Zhao, H., Gao, W., Huang, L., & Hu, Y. (2024). MYB transcription factors in plants: A comprehensive review of their discovery, structure, classification, functional diversity, and regulatory mechanism. International Journal of Biological Macromolecules, 282(Pt 2), 136652. https://doi.org/10.1016/j.ijbiomac.2024.136652

25. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686. https://doi.org/10.21105/joss.01686

26. Vanneste, K., Baele, G., Maere, S., & Van de Peer, Y. (2014). Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. Genome Research, 24(8), 1334–1347. https://doi.org/10.1101/gr.168997.113

Appendix

Appendix A: RNA-seq QC and Alignment

```
\SpecialCharTok{{-}{-}{-}}
\NormalTok{title}\SpecialCharTok{:} \StringTok{"01\_RNAseq\_QC\_and\_Alignment"}
\NormalTok{author}\SpecialCharTok{:} \StringTok{"Jack Chiang, Louis Lax{-}Roseman"}
\NormalTok{date}\SpecialCharTok{:} \StringTok{"2025{-}04{-}04"}
\NormalTok{output}\SpecialCharTok{:}\NormalTok{ html\_document}
\SpecialCharTok{{-}{-}{-}}

\StringTok{\textasciigrave{}\textasciigrave{}\textasciigrave{}}\AttributeTok{\{r setup,
↪  include=FALSE\}}
\AttributeTok{knitr::opts\_chunk$set(echo = TRUE, eval = FALSE, message = FALSE, warning
↪  = FALSE)}
```

1. Introduction

This document outlines the RNA-seq data processing pipeline for Marchantia polymorpha. The steps include:

```
Quality Control (QC): using fastp, FastQC, and MultiQC.

Alignment: of trimmed reads using STAR.

Summarizing alignment logs.
```

Adjust file paths and sample names as needed for your environment.

2. Data Preparation and Quality Control

2.1. Locating Raw FASTQ Files

Assume your raw FASTQ files (e.g., IM_Tak1_1.fastq, IM_Tak1_2.fastq, etc.) are located in data/raw_fastq/. 2.2. Trimming and Filtering with fastp

The following shell command demonstrates how to trim a paired-end sample using fastp. Adjust sample names and paths accordingly.

````
{r} fastp \   -i data/raw_fastq/IM_Tak1_1.fastq \   -I data/raw_fastq/IM_Tak1_2.fastq \
-o data/fastp_processed/IM_Tak1_R1.filt.fastq \   -O data/fastp_processed/IM_Tak1_R2.filt.fastq
\   --detect_adapter_for_pe \   --thread 4 \   --html data/fastp_processed/IM_Tak1_fastp.html
````

Repeat the above command for other samples (e.g., IM_Tak2, M_Tak1, etc.).

2.3. Running FastQC and MultiQC

After trimming, run FastQC on the filtered FASTQ files:

````
{r} fastqc data/fastp_processed/*_R1.filt.fastq data/fastp_processed/*_R2.filt.fastq -o
data/fastqc_processed/
````

Then, combine the FastQC reports with MultiQC: “‘{r} multiqc data/fastqc_processed/ -o data/multiqc_processed/

~~~

MultiQC will generate a comprehensive report (e.g., multiqc_report.html).
3. Alignment with STAR
3.1. Genome Index Generation

If a STAR index has not yet been built, run:
```{r}
STAR --runThreadN 8 \
     --runMode genomeGenerate \
     --genomeDir data/star_index \
     --genomeFastaFiles data/reference/MpTak1v5.1.fasta \
     --sjdbGTFfile data/reference/MpTak1v5.1_r1.gtf \
     --sjdbOverhang 99
~~~

3.2. Aligning Trimmed Reads

Below is an example alignment command for one sample (IM_Tak1): “‘{r} STAR –runThreadN 8
–genomeDir data/star_index
–readFilesIn data/fastp_processed/IM_Tak1_R1.filt.fastq
data/fastp_processed/IM_Tak1_R2.filt.fastq
–outFileNamePrefix data/aligned/IM_Tak1_
–outFilterMultimapNmax 10
–outFilterMismatchNoverLmax 0.05
–quantMode GeneCounts
–outSAMtype BAM SortedByCoordinate
–outReadsUnmapped Fastx

~~~

Repeat for all samples to generate sorted BAM files and alignment log files.

3.3. Summarizing STAR Alignment Logs

The following R code reads STAR log files and extracts basic metrics.
```{r}
library(dplyr)
library(readr)
~~~

```
# Define log files (adjust paths as needed)
log_files <- c(
  "data/aligned/IM_Tak1_Log.final.out",
  "data/aligned/IM_Tak2_Log.final.out",
  "data/aligned/M_Tak1_Log.final.out",
  "data/aligned/M_Tak2_Log.final.out",
  "data/aligned/V_Tak1_Log.final.out",
  "data/aligned/V_Tak2_Log.final.out"
)

# Function to extract metrics from a STAR log file
extract_star_metrics <- function(file) {
  lines <- readLines(file)
  uniquely_mapped <- as.numeric(gsub(".*\\|\\s*", "", lines[grep("Uniquely mapped reads number", lines)]
  mapped_percent <- as.numeric(gsub("%", "", gsub(".*\\|\\s*", "", lines[grep("Uniquely mapped reads %"

  data.frame(
    sample = gsub("_Log.final.out", "", basename(file)),
    uniquely_mapped_reads = uniquely_mapped,
    uniquely_mapped_percent = mapped_percent
  )
}

star_log_data <- do.call(rbind, lapply(log_files, extract_star_metrics))
print(star_log_data)
```

Generate a bar plot to visualize uniquely mapped reads per sample: “‘{r} library(ggplot2)

ggplot(star_log_data, aes(x = sample, y = uniquely_mapped_reads, fill = sample)) + geom_bar(stat = “identity”) + theme_minimal() + labs(title = “Uniquely Mapped Reads per Sample”, x = “Sample”, y = “Uniquely Mapped Reads”)

```
4. Next steps

The outputs from this pipeline (BAM files, QC reports, and alignment summaries) will be used in downstre


# References
- Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.
- Chen, S. et al. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor.
- Dobin, A. et al. (2013). STAR: ultrafast universal RNA-seq aligner.
```

Appendix B: Annotation Expression

```
\SpecialCharTok{{-}{-}{-}}
\NormalTok{title}\SpecialCharTok{:} \StringTok{"02\_TF\_Annotation\_Expression"}
\NormalTok{author}\SpecialCharTok{:} \StringTok{"Jack Chiang, Louis Lax{-}Roseman"}
\NormalTok{date}\SpecialCharTok{:} \StringTok{"2025{-}04{-}02"}
\NormalTok{output}\SpecialCharTok{:}\NormalTok{ html\_document}
\SpecialCharTok{{-}{-}{-}}
```

```
\StringTok{\textasciigrave{}\textasciigrave{}\textasciigrave{}}\AttributeTok{\{r setup,
 ↪  include=FALSE\}}
\AttributeTok{knitr::opts\_chunk$set(echo = TRUE, eval = FALSE, message = FALSE, warning
 ↪  = FALSE)}
```

1. Introduction

This document imports transcript quantifications from Salmon, annotates genes using a reference GTF and PlantTFDB, and summarizes transcription factor (TF) families with basic plots.

2. Building the Expression Matrix

2.1. Importing Salmon Quantifications

“‘{r} library(tximport) library(readr) library(dplyr) library(tidyr)

Define sample names and file paths (adjust to match your directory structure)

samples <- c(“IM_Tak1”, “IM_Tak2”, “M_Tak1”, “M_Tak2”, “V_Tak1”, “V_Tak2”) files <- file.path(“data/salmon_quant”, samples, “quant.sf”) names(files) <- samples

Import transcript-level quantifications

txi_tx <- tximport(files, type = “salmon”, txOut = TRUE) head(txi_tx$abundance)

~~~
2.2. Creating a Transcript-to-Gene Map
```{r}
library(GenomicFeatures)

# Path to reference GTF file (adjust as needed)
gtf_file <- "data/reference/MpTak1v5.1_r1.gtf"
txdb <- makeTxDbFromGFF(file = gtf_file, format = "gtf")
tx2gene <- select(txdb,
                  keys = keys(txdb, keytype = "TXNAME"),
                  columns = c("TXNAME", "GENEID"),
                  keytype = "TXNAME")
head(tx2gene)
write.csv(tx2gene, "data/tx2gene.csv", row.names = FALSE)
~~~

If the file tx2gene.csv is available, import gene-level quantifications:

“‘{r} if (file.exists(“data/tx2gene.csv”)) { tx2gene <- read.csv(“data/tx2gene.csv”) txi_gene <- tximport(files, type = “salmon”, tx2gene = tx2gene) expression_matrix <- as.data.frame(txi_gene$abundance) head(expression_matrix) }

~~~
3. Merging TF Annotations
```{r}
# Load PlantTFDB data for M. polymorpha (assumes file "Mpo_TF_list.txt" with columns Gene_ID and Family
marchantia_tfs <- read.delim("data/Mpo_TF_list.txt", sep = "\t", header = TRUE)
head(marchantia_tfs)

# Merge expression data with TF annotation. Adjust matching columns as needed.
if (exists("expression_matrix")) {
  expression_matrix$GeneID <- rownames(expression_matrix)
  tf_expression <- merge(expression_matrix, marchantia_tfs,
                         by.x = "GeneID", by.y = "Gene_ID", all.x = TRUE)
  head(tf_expression)
}
~~~

4. Summarizing TF Families “‘{r} library(dplyr)

tf_only <- tf_expression %>% filter(!is.na(Family)) non_tf <- tf_expression %>% filter(is.na(Family))

Count TF families

tf_family_counts <- tf_only %>% group_by(Family) %>% summarize(Count = n_distinct(GeneID)) %>% arrange(desc(Count)) knitr::kable(tf_family_counts, caption = “TF Family Counts in Marchantia Polymorpha”)

~~~
4.3. Basic Bar Chart of TF Family Distribution
```{r}
library(ggplot2)

ggplot(tf_family_counts, aes(x = reorder(Family, -Count), y = Count)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Distribution of TF Families in Marchantia", x = "TF Family", y = "Gene Count")
~~~

5. (Optional) Basic Expression Patterns

“‘{r} library(tidyr)

long_tf_exp <- tf_only %>% select(GeneID, Family, IM_Tak1, IM_Tak2, M_Tak1, M_Tak2, V_Tak1, V_Tak2) %>% pivot_longer(cols = c(IM_Tak1, IM_Tak2, M_Tak1, M_Tak2, V_Tak1, V_Tak2), names_to = “Sample”, values_to = “TPM”)

ggplot(long_tf_exp, aes(x = Family, y = TPM)) + geom_boxplot() + theme_minimal() + theme(axis.text.x = element_text(angle = 45, hjust = 1)) + labs(title = “TF Expression Levels Across Samples”, x = “TF Family”, y = “TPM”)

```
# References
- Patro, R. et al. (2017). Salmon provides fast and bias-aware quantification of transcript expression.
- Lawrence, M. et al. (2013). Software for Computing and Annotating Genomic Ranges.
- Jin, J. et al. (2017). PlantTFDB 4.0: toward a central hub for transcription factors and regulatory i
```

Appendix C: GO Analysis

```
\SpecialCharTok{{-}{-}{-}}
\NormalTok{title}\SpecialCharTok{:} \StringTok{"03\_GO\_Analysis\_and\_Advanced\_Plots"}
\NormalTok{author}\SpecialCharTok{:} \StringTok{"Jack Chiang, Louis Lax{-}Roseman"}
\NormalTok{date}\SpecialCharTok{:} \StringTok{"2025{-}04{-}02"}
\NormalTok{output}\SpecialCharTok{:}\NormalTok{ html\_document}
\SpecialCharTok{{-}{-}{-}}

\StringTok{\textasciigrave{}\textasciigrave{}\textasciigrave{}}\AttributeTok{\{r setup,
 ↪  include=FALSE\}}
\AttributeTok{knitr::opts\_chunk$set(echo = TRUE, eval = FALSE, message = FALSE, warning
 ↪  = FALSE)}
```

1. Introduction

This document performs downstream analyses including GO enrichment, advanced visualizations (upset plot, PCA, and correlation analysis), and interprets TF expression and evolution in Marchantia polymorpha.

2. GO Enrichment Analysis

“‘{r} library(dplyr) library(tidyr) library(ggplot2) library(GO.db) library(CopmlexHeatmap)

~~~
Load TF expression data with GO annotations (assumed to be in tf_expression_GO.csv):
```{r}
tf_go_data <- read.csv("data/tf_expression_GO.csv", stringsAsFactors = FALSE)
head(tf_go_data)
~~~

If necessary, merge external GO annotations here. Assume go_annotations is a data frame with columns: Gene_ID and GO_Term.

“‘{r} go_annotations <- read.delim(“data/go_annotation.tsv”, sep = “”̂, header = TRUE, stringsAsFactors = FALSE)

~~~
Define a function to perform GO enrichment analysis:
```{r}
perform_go_enrichment <- function(gene_set, go_annotations, background_genes) {
  go_counts <- go_annotations %>%
    filter(Gene_ID %in% gene_set) %>%
    group_by(GO_Term) %>%
    summarise(Count = n())

  background_counts <- go_annotations %>%
    group_by(GO_Term) %>%
    summarise(Background_Count = n())

  enrichment <- merge(go_counts, background_counts, by = "GO_Term", all.x = TRUE)
  enrichment <- enrichment %>% mutate(Total_Genes = length(gene_set), Total_Background = length(backgrou

  enrichment <- enrichment %>% rowwise() %>%
    mutate(P_Value = fisher.test(matrix(c(Count, Total_Genes - Count, Background_Count, Total_Backgroun
~~~

~~~
    ungroup() %>%
    mutate(Adjusted_P_Value = p.adjust(P_Value, method = "BH"))

  sig_go <- enrichment %>% filter(Adjusted_P_Value < 0.05) %>% arrange(Adjusted_P_Value)
  return(sig_go)
}
~~~

Example usage for a condition (e.g., IM_Tak1): `{r} im_tak1_genes <- tf_go_data %>% filter(IM_Tak1 > 0) %>% pull(GeneID) background_genes <- unique(tf_go_data$GeneID) sig_go_terms <- perform_go_enrichment(im_tak1_genes, go_annotations, background_genes) head(sig_go_terms)`

Visualize enriched GO terms if significant terms are found: “‘{r} if(nrow(sig_go_terms) > 0){ ggplot(sig_go_terms, aes(x = reorder(GO_Term, -log10(Adjusted_P_Value)), y = -log10(Adjusted_P_Value))) + geom_point(size = 3, color = “darkred”) + coord_flip() + labs(title = “Enriched GO Terms (IM_Tak1)”, x = “GO Term”, y = “-log10(Adjusted P-Value)”) + theme_minimal() } else { message(“No significant GO terms found for IM_Tak1.”) }

~~~
3. Upset Plot for TF Presence/Absence
```{r}
library(ComplexHeatmap)
~~~

Create a binary presence/absence matrix for TF genes:

“‘{r} tf_binary <- tf_go_data %>% mutate(across(c(IM_Tak1, IM_Tak2, M_Tak1, M_Tak2, V_Tak1, V_Tak2), ~ ifelse(. > 0, 1, 0))) %>% select(GeneID, IM_Tak1, IM_Tak2, M_Tak1, M_Tak2, V_Tak1, V_Tak2)

~~~
Define the condition list:

```{r}
condition_list <- list(
  IM_Tak1 = tf_binary %>% filter(IM_Tak1 == 1) %>% pull(GeneID),
  IM_Tak2 = tf_binary %>% filter(IM_Tak2 == 1) %>% pull(GeneID),
  M_Tak1  = tf_binary %>% filter(M_Tak1  == 1) %>% pull(GeneID),
  M_Tak2  = tf_binary %>% filter(M_Tak2  == 1) %>% pull(GeneID),
  V_Tak1  = tf_binary %>% filter(V_Tak1  == 1) %>% pull(GeneID),
  V_Tak2  = tf_binary %>% filter(V_Tak2  == 1) %>% pull(GeneID)
)
~~~

Create the combination matrix and generate the Upset plot:

“‘{r} comb_mat <- make_comb_mat(condition_list, mode = “distinct”) UpSet(comb_mat, set_order = c(“IM_Tak1”, “IM_Tak2”, “M_Tak1”, “M_Tak2”, “V_Tak1”, “V_Tak2”), comb_order = order(comb_size(comb_mat), decreasing = TRUE), top_annotation = HeatmapAnnotation(“Intersection Size” = anno_barplot(comb_size(comb_mat), add_numbers = TRUE)), right_annotation = rowAnnotation(“Set Size” = anno_barplot(set_size(comb_mat), add_numbers = TRUE)), column_title = “Upset Plot: TF Expression Across Conditions”, row_title = NULL)

~~~
4. Advanced Visualizations: PCA and Correlation Analysis

```{r}
library(ggrepel)
~~~

Perform PCA on TF expression data:

"'{r} species_cols <- c("IM_Tak1", "IM_Tak2", "M_Tak1", "M_Tak2", "V_Tak1", "V_Tak2") tf_expression_matrix <- tf_go_data %>% select(all_of(species_cols)) %>% as.matrix() pca_result <- prcomp(tf_expression_matrix, scale. = TRUE) pca_df <- as.data.frame(pca_result$x)pca_df$GeneID <- tf_go_data$GeneID

ggplot(pca_df, aes(x = PC1, y = PC2)) + geom_point(alpha = 0.7, color = "darkblue") + geom_text_repel(aes(label = GeneID), size = 2) + labs(title = "PCA of TF Expression", x = paste0("PC1 (", round(summary(pca_result)$importance[2, 1]*100, 1), "y = paste0("PC2(", round(summary(pca_result)$importance[2,2] * 100, 1), "%)")) + theme_minimal()

~~~
Perform correlation analysis between TF family count and average TPM:

```{r}
avg_tpm <- tf_go_data %>% group_by(Family) %>%
  summarize(Avg_TPM = mean(c(IM_Tak1, IM_Tak2, M_Tak1, M_Tak2, V_Tak1, V_Tak2), na.rm = TRUE))
cor_data <- merge(tf_family_counts, avg_tpm, by = "Family")

ggplot(cor_data, aes(x = Count, y = Avg_TPM, label = Family)) +
  geom_point(color = "forestgreen", size = 3) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  geom_text_repel() +
  labs(title = "Correlation Between TF Family Count and Average TPM",
       x = "TF Family Count",
       y = "Average TPM") +
  theme_minimal()
~~~

References

- Ashburner, M. et al. (2000). Gene Ontology: tool for the unification of biology.
- Fisher, R. A. (1922). On the interpretation of ^{2} from contingency tables.
- Gu, Z. et al. (2016). Complex heatmaps reveal patterns in genomic data.

~~~
## Appendix D: Visualizations
```r
---
title: "04_Advanced_Visualizations"
author: "Jack Chiang, Louis Lax-Roseman"
date: "2025-04-02"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE, eval = FALSE, message = FALSE, warning = FALSE)
~~~

1. Introduction

In this document we generate advanced visualizations to explore transcription factor (TF) family distributions and expression across species. The analyses include:

````
Clustered stacked bar plots for TF family distributions.

Normalized heatmaps of TF family abundance.

An improved PCA biplot with loadings.

A correlation network of TF families.

Hierarchical clustering heatmaps.
````

2. Clustered Stacked Bar Plot for TF Family Distribution

“‘{r} library(dplyr) library(tidyr) library(ggplot2) library(colorspace)

~~~
Load TF family summary data (adjust file path as needed):

```{r}
tf_summary <- read.delim("data/tf_family_summary.tsv", sep = "\t", header = TRUE)
~~~

Pivot data into long format for plotting:

“‘{r} stacked_data <- tf_summary %>% select(TF_Family, Cre, Mpo, Ppa, Smo, Pab, Ath, Zma) %>% pivot_longer(cols = c(Cre, Mpo, Ppa, Smo, Pab, Ath, Zma), names_to = “Species”, values_to = “Gene_Count”)

~~~
Generate a custom color palette:

```{r}
n_colors <- 15
my_palette <- rainbow_hcl(n_colors, c = 60, l = 70)
my_palette <- sample(my_palette, n_colors)
~~~

Create the stacked bar plot:

“‘{r} ggplot(stacked_data, aes(x = Species, y = Percentage, fill = TF_Family)) + geom_bar(stat = “identity”) + facet_grid(. ~ Species, scales = “free_x”) + theme_minimal() + labs(title = “Clustered Stacked Bar Plot: TF Family Distribution”, x = “Species”, y = “Percentage (%)”) + theme(axis.text.x = element_text(angle = 45, hjust = 1)) + scale_fill_manual(values = my_palette)

~~~
3. Normalized Heatmap of TF Family Abundance

```{r}
library(pheatmap)
library(viridis)
library(tibble)
~~~

Assume tf_summary has TF_Family and species columns (Cre, Mpo, Ppa, Smo, Pab, Ath, Zma):

“‘{r} species_cols <- c(“Cre”, “Mpo”, “Ppa”, “Smo”, “Pab”, “Ath”, “Zma”) species_totals <- colSums(tf_summary[, species_cols]) norm_data <- tf_summary %>% select(TF_Family, all_of(species_cols)) %>% column_to_rownames(“TF_Family”)

for(col in species_cols) { norm_data[, col] <- norm_data[, col] / species_totals[col] * 100 }

~~~
Create the normalized heatmap:
```{r}
pheatmap(as.matrix(norm_data),
         display_numbers = TRUE,
         number_format = "%.1f%%",
         color = viridis(100),
         main = "Normalized TF Family Abundance Across Species",
         fontsize = 12,
         fontsize_number = 10)
~~~

4. Improved PCA Biplot for TF Family Expression “‘{r} library(ggrepel)

~~~
Use tf_summary for PCA (rows = TF families, columns = species counts):
```{r}
pca_data <- tf_summary %>%
  select(TF_Family, all_of(species_cols)) %>%
  column_to_rownames("TF_Family") %>%
  t()

pca_result <- prcomp(pca_data, scale. = TRUE)
pca_df <- as.data.frame(pca_result$x)
pca_df$Species <- rownames(pca_df)
~~~

Extract loadings and calculate contributions: “‘{r} loadings <- pca_result$rotationloadings_df < -as.data.frame(loadings)loadings_df$TF_Family <- rownames(loadings_df) loadings_df$PC1_contrib < -loadings_df$PC1^2 loadings_df$PC2_contrib < -loadings_df$PC2^2 loadings_df$total_contrib < -loadings_df$PC1_contrib + loadings_df$PC2_contrib

Filter loadings for plotting

loadings_plot <- loadings_df %>% filter(PC1_contrib > 0.03 | PC2_contrib > 0.03) loadings_plot$is_strong < -loadings_plot$total_contrib > median(loadings_plot$total_contrib) scaling_factor <- 6

~~~
Base PCA plot:
```{r}
p1 <- ggplot() +
  geom_point(data = pca_df, aes(x = PC1, y = PC2, color = Species), size = 4) +
  geom_text_repel(data = pca_df, aes(x = PC1, y = PC2, label = Species), size = 3) +
  theme_minimal() +
  labs(x = paste0("PC1 (", round(summary(pca_result)$importance[2,1]*100, 1), "%)"),
       y = paste0("PC2 (", round(summary(pca_result)$importance[2,2]*100, 1), "%)"),
       title = "PCA Biplot of TF Family Expression")
~~~

Add arrows for loadings: “‘{r} p_biplot <- p1 + geom_segment(data = loadings_plot, aes(x = 0, y = 0, xend = PC1 * scaling_factor, yend = PC2 * scaling_factor, size = total_contrib), arrow = arrow(length

= unit(0.2, “cm”)), color = “darkblue”) + scale_size_continuous(range = c(0.5, 2), guide = “none”) + geom_text_repel(data = loadings_plot, aes(x = PC1 * scaling_factor, y = PC2 * scaling_factor, label = TF_Family), size = 3, box.padding = 0.5, color = “black”) p_biplot

~~~
5. Correlation Network of TF Families

```{r}
library(igraph)
library(ggraph)
~~~

Calculate correlation between TF families across species using the transposed normalized data: “‘{r} tf_cor <- cor(t(as.matrix(norm_data)), method = “spearman”) rownames(tf_cor) <- rownames(norm_data) colnames(tf_cor) <- rownames(norm_data)

Remove self-correlations and set a threshold to filter weak correlations

tf_cor[abs(tf_cor) < 0.6] <- 0 diag(tf_cor) <- 0

Build a graph object from the correlation matrix

graph <- graph_from_adjacency_matrix(tf_cor, mode = “undirected”, weighted = TRUE) E(graph)$width < -abs(E(graph)$weight) * 5 E(graph)$color < -ifelse(E(graph)$weight > 0, “blue”, “red”)

set.seed(123) # For reproducibility plot(graph, vertex.size = 20, vertex.label.color = “black”, vertex.color = “lightblue”, vertex.frame.color = “gray”, vertex.label.cex = 0.8, edge.curved = 0.2, layout = layout_with_fr(graph))

~~~
6. Hierarchical Clustering Heatmap

Save the clustered heatmap as a PNG file:

```{r}
png("TF_combined_clustered_heatmap.png", width = 1200, height = 1000, res = 150)
pheatmap(as.matrix(norm_data),
         display_numbers = TRUE,
         number_format = "%.1f%%",
         color = viridis(100),
         main = "Hierarchical Clustering of TF Families and Species",
         fontsize = 12,
         fontsize_number = 10,
         clustering_distance_rows = "euclidean",
         clustering_distance_cols = "euclidean",
         clustering_method = "ward.D2")
dev.off()
~~~

References

- Kolde, R. (2019). pheatmap: Pretty Heatmaps. R package version 1.0.12.
- Jolliffe, I. T. (2002). Principal Component Analysis. Springer.
- Csardi, G. & Nepusz, T. (2006). The igraph software package for complex network research.

“‘