Investigating evolution of gene regulation with cross-species comparative transcriptomics

Undersøkelser av genreguleringsevolusjon ved hjelp av genuttrykksdata fra flere arter

Philosophiae Doctor (PhD) Thesis

Lars Grønvold

Norwegian University of Life Sciences
Faculty of Chemistry, Biotechnology and Food Science

Ås (2017)

Thesis number 2017:101
ISSN 1894-6402
ISBN 978-82-575-1486-0

Summary

An essential ability of all lifeforms is to regulate the activity of its genes. This ability is what allows multicellular organisms to generate hundreds of completely different cells despite having the same genome. It is also this ability that allows organisms to adjust dynamically to external conditions, including how plants in temperate climates adjusts their metabolism in preparation for the coming winter as we study in paper 1 and 2 of this thesis.

Quantification of gene expression by RNA sequencing or microarrays is a common method for investigating gene regulation. However, comparing gene expression across species to study evolution of gene regulation presents many challenges. In the first paper, we investigated the evolution of cold adaption in grass by comparing the transcriptomic response to cold in five diverse species of temperate grasses (Pooideae). Two of these species, Barley and *Brachypodium distachyon*, are well characterized and their cold-response has been studied before. The three other species, which are less studied, belong to early diverging Pooideae tribes. By comparing the cold-response from both early and later diverging species it was possible to investigate to what extent the cold-response has been conserved since their common ancestor, or if it evolved gradually within the Pooideae. Although we observed a significant common response in all species, there were extensive differences. These differences did not follow the phylogeny, as we would expect if there was a gradual evolution. These results indicate that any conserved cold-response either involved very few genes or that the cold-response evolved independently in each lineage.

In the second paper, instead of looking at the entire transcriptome, we look specifically at orthologs of a handful of well-known cold-tolerance genes. By inspecting their phylogeny, we see that gene duplication have played an important role both early and later in the evolution of some of the cold-tolerance genes. This includes the CBF gene family, which is known to contain master regulators of cold-response. By investigating the protein sequences we also find that some of the functional protein motifs was missing in the early diverging species, suggesting a gradual and lineage specific evolution of cold-tolerance functions.

One of the challenges when comparing gene regulation across species is to acquire directly comparable samples. While we designed the experiment in paper 1 so that the samples should be as similar as possible across species, most experiments found in public databases are not suitable for direct comparison. In the third paper we investigate the use of co-expression to indirectly compare the expression similarities of orthologs. This approach makes it possible to use samples that are not directly comparable. The current methods for co-expression based comparison are not so well studied and we find that we can develop an improved method. Applying our new method to public gene expression data from five plant species, we investigate the link between gene duplication and expression divergence.

Acknowledgement

I would like to thank my main supervisor Torgeir for giving me this chance and for sticking with me to the end. Torgeir and my co-supervisors Simen and Siri have been very generous with sharing their knowledge and helping with editing the papers. It has also been a pleasure working together with Marian, who probably did more than his fair share of the work. Together with our supervisors, we had some good times. I am especially thankful for being able to go on the trips to London and to Vermont.

I also thank the biostatics group for a good social work environment. Especially Navreet, Hilde and Guro, i.e. the party planners. And of course the guys Niklas, Lars-Erik, Raju, Gareth and Francesco.

Lastly I would like to thank the tax-payers, who financed everything. This is what your hard-earned money is spent on. I hope you find it useful.

List of papers

1. Lars Grønvold*, Marian Schubert*, Simen R. Sandve, Siri Fjellheim and Torgeir R. Hvidsten. **Comparative transcriptomics provides insight into the evolution of cold response in Pooideae.** *bioRxiv* doi:10.1101/151431, 2017.
2. Marian Schubert*, Lars Grønvold*, Simen R. Sandve, Torgeir R. Hvidsten and Siri Fjellheim. **Evolution of cold acclimation in temperate grasses (Pooideae).** *bioRxiv* doi:10.1101/210021, 2017.
3. Lars Grønvold and Torgeir R. Hvidsten. **Cross species comparative transcriptomics with co-expression correlation.** Manuscript, 2017.

*Contributed equally

Contributions of the candidate:

Paper 1: The candidate contributed to the experiment design, was responsible for designing and creating the bioinformatics workflow, performed the data analysis and contributed to writing the manuscript.

Paper 2: The candidate aided in extracting the relevant data by creating software to search and browse the data from paper 1, helped with the bioinformatic analysis and contributed to writing the manuscript.

Paper 3: The candidate came up with the new method, implemented it, did all the analyses and wrote the manuscript.

Contents

Introduction

Gene regulation

While any given organism has only a fixed set of genes, not all of the genes are expressed at all times. The full set of genes can be thought of as a list of parts, while there is a regulatory network that decides what parts to produce. This allows cells to adapt dynamically to changing conditions and signals. Yeast for example, have the required genes for metabolizing galactose, but these genes are regulated so that they are only expressed when galactose is available. In more complex multicellular organisms, cells can differentiate to take on specialized roles, such as neuronal cells or liver cells, by turning on or off expression of the cell type specific genes. Correct regulation of growth and differentiation is essential during development as this decides the form and structure of the organism.

The first stage at which a gene can be regulated is at transcription initiation. In order for a gene to be transcribed, the transcriptional machinery must first be assembled at correct positions along the chromosome, i.e. the transcription start sites (TSS). This is achieved by DNA binding transcription factors (TFs) that recognize and bind to specific sequence motifs at TF binding sites (TFBS) in the promoter. These TFs are proteins that, in addition to the DNA binding domain, typically have an activation domain that facilitates transcription by interacting with the transcription machinery. TFs can also be repressors, which reduce transcription. In addition to the promoter, which is situated around the TSS, there can also be enhancers, distal clusters of TFBSs, that can activate transcription despite being far away from the TSS. There are also nucleosomes which add another layer of complexity to the transcription regulation, but that is beside the point. The point is that the regulation of transcription of a specific gene is defined by the regulatory sequences, i.e. TFBSs, in the promoter and enhancers around the gene. These regulatory sequences are termed cis-regulatory elements. The TFs and co-factors that interact with the cis-elements and modulate the transcriptional activity are called trans-factors. In theory, it should be possible to predict the transcription from the cis-regulatory sequences and the concentration levels of the trans-factors, but this remains elusive (Wilczynski *et al.* 2012; Huminiecki & Horbańczuk 2017).

Besides regulating transcription initiation, there are several other mechanisms to regulate the final gene expression, such as alternative splicing which in addition to producing alternative isoforms can lead to nonsense-mediated decay (Ge & Porse 2014). MicroRNAs also serve as trans-acting regulatory factor by degrading target mRNAs and repressing translation (Dalmay 2013). Translation can also be regulated by RNA binding proteins (Babitzke *et al.* 2009).

In addition to all these mechanisms, there is also post-translational modifications and protein-protein interactions of the signaling network (Papin *et al.* 2005) that together make up the regulatory network that controls gene expression.

Examples of cross species gene expression comparison methods

Mutations in cis-regulatory sequences or in any part of the sequence of proteins that participate in the regulatory network, such as the DNA binding domain of a TF, can potentially alter the regulatory network and consequently alter the expression pattern of the genes. RNA sequencing and microarrays are powerful tools as they enable the measurement of gene expression for every gene in the entire genome all at once. Such measurements can be used to study how gene expression patterns have evolved by comparing across species.

Depending on the experiment design and the aspect of gene regulatory evolution being studied, there are different strategies that can be applied. For example, phylogenetic methods that originally was used to study evolution of quantitative traits (Felsenstein 1985; Martins & Hansen 1997) has been adopted to study gene expression evolution (e.g. Rohlfs & Nielsen 2015). Such methods are applicable to a single tissue/condition at a time, with comparable and replicated samples from three or more species whose phylogeny and branch lengths must be known. The advantage is that they are based on theoretical models that can be used to test for significantly diverged gene expression in either the entire phylogeny, or in specific lineages, e.g. to test whether a gene is significantly higher expressed in one of the species.

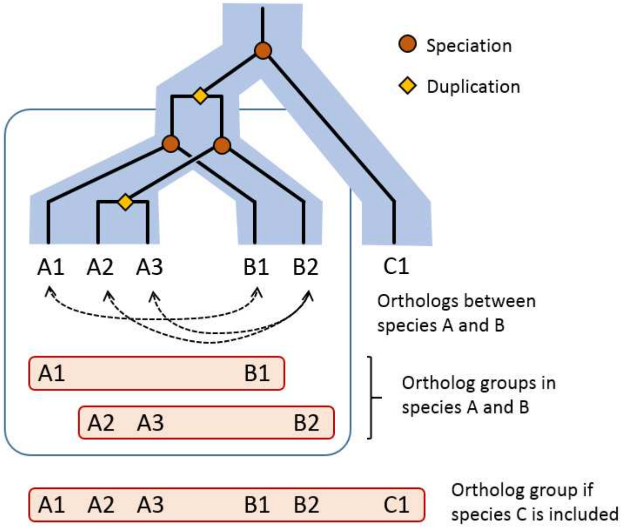
In cases where there is a contrast between two different conditions in each species, it is common to perform differential expression analysis in each species and then compare the sets of differentially expressed genes or fold changes (e.g. Zhao *et al.* 2015; Hefer *et al.* 2015). If the goal is not to compare but to combine the results from each species to get a more powerful test, it is possible to use Fisher's combined probability test to combine the P-values (Fisher 1948) or more advanced methods specifically developed for gene expression (Kristiansson *et al.* 2013).

For experimental designs with multiple comparable tissues or conditions in several species there are other analyses that can be performed. For example, principal component analysis of samples has been used to show that expression pattern differences between tissues dominate over differences between mammalian species (Brawand *et al.* 2011). However, this method has been criticized for just showing the dominant effect of a subset of genes, and it is therefore advised to investigate if the expression variance is dominated by the species or tissue differences in a gene-specific manner (Breschi *et al.* 2016). Multispecies gene expression profile clustering is another method that can be applied when the samples are directly comparable. This can involve correlating the expression profiles of genes both within and between species and then make clusters of co-expressed genes (Davidson *et al.* 2012). Orthologous genes can then be considered to have conserved or diverged expression pattern depending on whether they end up in the same cluster or not.

If comparable samples across species are unavailable, such as often is the case for expression data in public repositories, it is possible to do indirect comparison via co-expression. The principle is that if a gene in one species and its ortholog in another is co-expressed with the corresponding (i.e. orthologous) genes, then they have conserved co-expression and probably also similar/conserved expression patterns. This is sometimes referred to as co-expression network alignment and will be introduced in a later section. There are several different methods. The result can be a consensus network with only the conserved links or nodes; or it can result in consensus clusters.

Gene duplication and orthology

In addition to single nucleotide polymorphism and other small-scale mutations, evolution involves larger structural changes of the DNA that can lead to duplication or loss of genes. Duplications can be of various sizes, from single gene duplication, duplication that include long stretches of genes, to whole genome duplications (WGDs) that result in an extra copy of every chromosome. WGDs have played a significant role in the evolution of plants (Soltis *et al.* 2009). Although some duplicates have survived for hundreds of millions of years, the vast majority of duplicates tend to be lost over time (Maere *et al.* 2005). The genome wide half-life of duplicates in Arabidopsis has been estimated to be about 17.3 million years (Lynch & Conery 2003).



● Speciation
◇ Duplication
A1
A2
A3
B1
B2
C1
Orthologs between species A and B
A1 B1
A2 A3 B2
Ortholog groups in species A and B
A1 A2 A3 B1 B2 C1
Ortholog group if species C is included

Figure 1: Orthology and ortholog groups. *Phylogeny of a gene family in three species shown within the species tree.*

Because gene duplication and gene loss events occur independently in different species, each species will contain different sets of genes. To know which genes to compare across species it is necessary to identify the orthologous genes. Orthologous genes are by definition a pair of genes in different species that originate from a single gene in the most recent common ancestor (Fitch 2000). Ortholog identification is based on sequence similarity of the genes. There are several available methods for ortholog inference. For a pair of species, Inparalog (Sonnhammer & Östlund 2015) finds orthologs by doing an all-against-all blast and first uses the two-way best hits to identify the seed orthologs and then adds in-paralogs (duplicates occurring after the species split). When performing a comparative study across three or more species it is helpful to identify ortholog groups. Ortholog groups can be defined as all genes descending from a single gene in the most recent common ancestor of all the included species (Figure 1). OrthoMCL (Li *et al.* 2003) and OrthoFinder (Emms & Kelly 2015) are graph based methods that find ortholog groups using MCL clustering (Van Dongen 2008) on a graph based on all-against-all pair-wise sequence similarity. Another class of orthology inference is phylogeny based methods. Such methods perform multiple alignment of homologous clusters of genes and infers the phylogeny. The homologous clusters can be based on ortholog groups, e.g. eggnog (Powell *et al.* 2014), or by more loosely defined clustering, such as TribeMCL (Enright *et al.* 2002) which is used in PLAZA (Proost *et al.* 2009). Once the phylogeny is available, gene trees from each gene family is then reconciled with the species tree to identify duplication and speciation nodes. If two genes from different species can be traced back to a speciation node in the gene tree they are orthologs.

For most sequenced genomes, orthology information can be downloaded from orthology databases, e.g. Ensembl Compara (Vilella *et al.* 2009), TreeFam (Ruan *et al.* 2008), OMA (Altenhoff *et al.* 2015), PLAZA (Proost *et al.* 2009).

Evolution of duplicated genes

Even though the fate of most duplicates is degeneration into pseudo-genes and eventual loss, some manage to be retained in the long run. However, retention of gene duplicates have to bring some selective advantage. As duplication creates redundancy, it allows for mutations that would otherwise be detrimental to persist. This creates an opportunity for novel adaptive functions to evolve in the duplicate while maintaining the original gene, a process called neo-functionalization (Ohno 1970). It is also possible that the original gene had several functions, which is divided between the duplicates so that they become non-redundant. This process is called sub-functionalization. The duplication-degeneration-complementation model explains how sub-functionalization occurs by loss of complementary functions (Force *et al.* 1999). A duplicate gene with enhancers that activate expression in different tissues can for example lose a different enhancer in each copy, resulting in sub-functionalization at the expression level.

Another mechanism that may explain why duplicates are retained is that there can also be an advantage of the extra dosage resulting from having a duplicate. Also, each copy could acquire reduced expression so that both genes are required for sufficient dosage. After a WGD, dosage balance sensitive genes, like sub-units of a protein complex, tend to be retained because the loss of a single sub-unit disrupts the balance (Ouedraogo *et al.* 2012).

Co-expression network alignment

Since duplicate genes tend to diverge in functions over time, it is of interest to identify which of the duplicates that has retained the original function (or if any of them has). This is important if functional annotations are to be inferred from orthologs. Comparison of expression levels can help to identify if an ortholog has a conserved expression pattern, which would indicate that the function is also conserved. An ortholog with conserved expression pattern is sometimes referred to as the expression ortholog or expressolog (Patel *et al.* 2012). One way to identify expressologs is by co-expression network alignment (Netotea *et al.* 2014). Co-expression networks for the two species being compared (e.g. the model species and the unannotated species of interest) are compared by aligning nodes, i.e. genes that are orthologous. A pair of orthologous genes are said to be conserved if a significant number of their co-expressed neighbors are also orthologs, in other words, if there is a significant overlap in the network neighborhood (Figure 2). The significance of such an overlap can be calculated using the hypergeometric distribution.

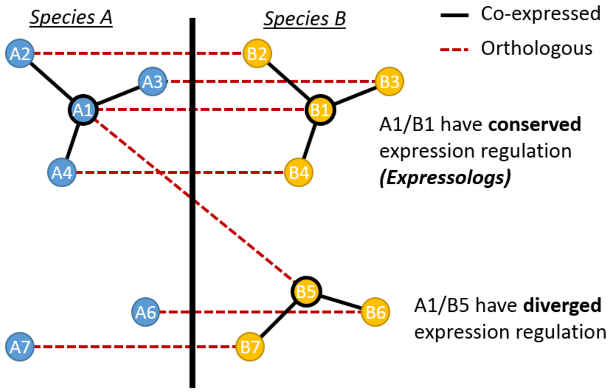

Species A
Species B
Co-expressed
Orthologous
A1
A2
A3
A4
A6
A7
B1
B2
B3
B4
B5
B6
B7
A1/B1 have **conserved** expression regulation ***(Expressologs)***
A1/B5 have **diverged** expression regulation

Figure 2: Example of co-expression network alignment. *The gene A1 has two orthologs, B1 and B5. The co-expressed neighbors of A1 and B1 are orthologs, indicating that A1 and B1 have conserved expression patterns and are more likely to have retained the same function. B5, the paralog of B1, is co-expressed with a different set of genes whose orthologs in species A are not co-expressed with A1, indicating that B5 has a diverged expression pattern and possibly performs a different function.*

Co-expressed genes tend to be functionally related (Usadel *et al.* 2009). Using the guilt-by-association approach it is therefore possible to generate hypotheses about the function of co-expressed genes (Wolfe *et al.* 2005). Co-expression can be conserved across species and combining co-expression networks from several species can improve estimates of biological functions (Stuart *et al.* 2003). There are several online tools that enable comparative co-expression analysis, including PlaNet (Mutwil *et al.* 2011) and ComPlEx (Netotea *et al.* 2014). While ComPlEx performs pairwise comparison between two species, PlaNet compares across several species at once, making the associations more robust.

Paper summaries and discussion

Paper 1

In the first paper, we compare the transcriptomic response to cold treatment in five species of temperate grass i.e. Pooideae. Unlike the tropical sister species rice and bamboo, a large portion of Pooideae thrive in temperate climates, which suggests that the Pooideae at some point adapted to cold climates while the sister clades remained in the tropics. This makes Pooideae a potential model to study cold adaptation. The molecular mechanisms that are required for grass to survive in cold have been extensively studied in a few species, such as barley and the model grass Brachypodium. From these studies, and studies in other cold adapted species, we know that an important aspect of cold adaption is the correct timing of transcriptional programs in anticipation of the changing seasons. One example is cold-acclimation, where plants increase production of frost-protection molecules in response to non-freezing temperatures (Thomashow 1999). Other processes must be inhibited, such as flowering and growth, as they make the plant more susceptible to frost damage. To discover which genes that are involved in the cold adaptive transcriptional programs, it is normal to perform gene expression studies where the plants are subjected to autumn-like conditions, such as lower temperature and shorter day length, and test which genes that are differentially expressed. In our study we perform such an experiment on five grass species, which includes the well-studied brachypodium and barley but also three less studied species representative of three early branching tribes among the Pooideae. The idea is that as a part of their adaptation to colder climates the Pooideae evolved a controlled cold-response program that did not exist in their common tropical ancestor. By comparing their responses, we found that it varies a lot between the species, indicating that very few of the many cold-responsive genes have a regulatory program conserved since early pooid evolution. Also, we could not see any signs of gradual evolution, as the two most closely related species did not have a more similar expression profile than any other species. Although it is hard to say anything conclusive with this data, it does tell us something about the extent of variation in cold-response that we can expect when comparing such distantly related species. Our observations may also indicate that a large part of the functional cold-response has evolved independently in each of the species as opposed to being inherited from a common ancestor that already had a functional cold-response.

One limitation in this study came from the strict filtering of ortholog groups, which excluded complex gene trees with duplications on any of the inner branches (i.e. only recent lineage duplications where allowed). The intention of this filtering was to exclude potential miss-assembled transcripts caused by the *de novo* transcriptome assembly and to keep the analysis simple. However, when comparing our data with a set of known cold-responsive barley genes identified in earlier studies we found that a large portion of these genes belong to gene families with complex gene trees. This indicates that gene duplication played an important role in cold adaptation. One of the caveats of *de novo* transcriptomics is that it makes it hard to analyze gene duplication events because only expressed genes are detectable and that miss-assembly result in false duplicates (as we found when comparing our *de novo* assembly against the barley reference genome assembly).

Paper 2

In the second paper, we looked specifically at a selection of gene families that were already known to be important in cold tolerance. Their gene tree phylogenies and their transcriptional response were analysed using the data acquired in paper 1. Some of the genes seems to have changed very little, having no duplications and responding to cold in all five species, indicating that these parts of the cold-tolerance machinery were already available in the Pooideae ancestor. On the other hand, some of the gene families have undergone drastic expansion, highlighting the role of gene duplication in cold adaptation, while some genes seem to have acquired novel protein functions that is missing in the early diverging species. Together it gives a picture of a gradual and lineage specific evolution of grass cold adaptation, involving both novel protein functions and gene duplication. Regarding the role of gene regulation, there is a trend that the conserved genes are limited to a short-term response while the long-term term response is common in duplicated and later evolving genes. In addition, the CBF transcription factor family, which is central in the regulation of cold-stress, has undergone repeated duplications both early and late in the Pooideae evolution.

Paper 3

While the experiment performed for paper 1 and 2 was specifically designed to get directly comparable samples in each species, the vast majority of the data available in public databases is not suitable for direct inter-species comparison. In the third paper, we developed a new method for comparing expression patterns across species indirectly via co-expression. The principle of comparison via co-expression is that genes with conserved expression pattern will be co-expressed with the same genes (i.e. orthologs) in the compared species, therefore making it possible to get a measure of expression pattern similarity despite not having directly comparable samples. While previous studies have used the same principle, these can be divided into two categories, each with its limitations: 1) The correlation approach is based on calculating the correlation between co-expression matrices. This approach has not received much attention, and all implementation to our knowledge, use only Pearson correlation for calculating the co-expression matrix. 2) The network overlap approach is based on first generating a co-expression network and then calculating the overlap between the neighbors of a given ortholog pair. Studies using this approach have taken advantage of state of the art methods for co-expression network generation, but uses a threshold to get a binary network, which limits the power of the cross-species comparison. We tested several of the common state-of-the-art methods for generating co-expression matrix in combination with correlation to calculate the cross-species similarity. The resulting score is a correlation value between -1 and 1 which indicates the expression pattern similarity between a pair of orthologs. Also, for each gene with a 1:1 ortholog we calculate the proportion of non-orthologous genes that gets a lower score than the actual ortholog. This proportion indicates the significance of the score, and was used to evaluate the performance of the different methods. Testing the different methods on public expression-data from five different plants, we found that including a normalization step such as “mutual ranks” improved the results over just using the correlation matrix. The power of the method was demonstrated by comparing the cross-species expression similarity of 1:1 orthologs with 1:N orthologs, confirming that duplicated genes have more diverged expression patterns. Interestingly, we observed that duplicate genes in soybean, which has experienced a relatively recent whole genome duplication (WGD), are more conserved than would have been expected. As a possible explanation, we suggest that dosage balance has a significant effect on which genes that are retained after a WGD.

Conclusion and future perspectives

Not all gene families are equal, some tend to diverge faster in expression pattern than others and some tend to duplicate more often than others do. Also, the history of duplications of a gene is predictive of how conserved the expression is between species. It seems obvious that there is a lot you can say about the regulation of a gene by studying expression of orthologous genes in other species. As more genomes are sequenced and more expression data is being released, there will be increased interest in methods that can leverage this data.

Methods like the one developed in paper 3, that can measure the level of expression divergence of any pair of genes between species, has many potential uses. One area that could be explored further is the link between gene duplication and expression divergence. It has been shown that the different modes of duplication affects expression divergence differently (Wang *et al.* 2011). With more high quality genomes available, it is possible to better classify the duplication event that each duplicate gene originated from by using collinearity information such as in PLAZA (Proost *et al.* 2009). Since most studies of expression divergence of duplicates has been performed by comparing the expression between the duplicates within species, it would be interesting to apply the method from paper 3 to investigate how different WGD events and other types of duplicates affect expression divergence across species.

While we only applied our method to plant genomes so far, there is no reason why the same method cannot be applied to other species. There is plenty of available expression and other experimental data for humans and model species like mice. To be able to learn more about evolution of regulation it is necessary to go beyond just expression levels and combine it with changes in regulatory sequences. Technologies such as DNase-seq have been used to detect active TF-binding sites in enhancers and promoters in mouse and human, giving a highly detailed image of cis-regulatory evolution (Stergachis *et al.* 2014). This kind of data might tell us more about the mechanism behind the diverged expression patterns.

Another useful new technology is single-cell RNA-seq, which makes it possible to measure the expression in individual cells. Co-expression based single-cell expression data may reveal some co-expression patterns that is not detectable in normal expression data from pools of cells (Crow *et al.* 2016). In paper 3 we argue that more varied samples give better results, so it would be interesting to see what adding some single-cell data to the mix could do.

References

Altenhoff AM, Škunca N, Glover N *et al.* (2015) The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. *Nucleic Acids Research*, **43**, D240–D249.

Babitzke P, Baker CS, Romeo T (2009) Regulation of Translation Initiation by RNA Binding Proteins. *Annual Review of Microbiology*, **63**, 27–44.

Brawand D, Soumillon M, Necsulea A *et al.* (2011) The evolution of gene expression levels in mammalian organs. *Nature*, **478**, 343–348.

Breschi A, Djebali S, Gillis J *et al.* (2016) Gene-specific patterns of expression variation across organs and species. *Genome biology*, **17**, 151.

Crow M, Paul A, Ballouz S, Huang ZJ, Gillis J (2016) Exploiting single-cell expression to characterize co-expression replicability. *Genome Biology*, **17**, 101.

Dalmay T (2013) Mechanism of miRNA-mediated repression of mRNA translation. *Essays in biochemistry*, **54**, 29–38.

Davidson RM, Gowda M, Moghe G *et al.* (2012) Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution. *The Plant Journal*, **71**, no-no.

Van Dongen S (2008) Graph Clustering Via a Discrete Uncoupling Process. *SIAM Journal on Matrix Analysis and Applications*, **30**, 121–141.

Emms DM, Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, **16**, 157.

Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic acids research*, **30**, 1575–84.

Felsenstein J (1985) Phylogenies and the Comparative Method. *The American Naturalist*, **125**, 1–15.

Fisher RA (1948) Answer to question 14 on combining independent tests of significance. *The American Statistician*, **2**, 30.

Fitch WM (2000) Homology: a personal view on some of the problems. *Trends in Genetics*, **16**, 227–231.

Force A, Lynch M, Pickett FB *et al.* (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, **151**, 1531–45.

Ge Y, Porse BT (2014) The functional consequences of intron retention: Alternative splicing coupled to NMD as a regulator of gene expression. *BioEssays*, **36**, 236–243.

Hefer CA, Mizrachi E, Myburg AA, Douglas CJ, Mansfield SD (2015) Comparative interrogation of the developing xylem transcriptomes of two wood-forming species: *Populus trichocarpa* and *Eucalyptus grandis*. *New Phytologist*, **206**, 1391–1405.

Huminiecki Ł, Horbańczuk J (2017) Can We Predict Gene Expression by Understanding Proximal Promoter Architecture? *Trends in Biotechnology*, **35**, 530–546.

Kristiansson E, Österlund T, Gunnarsson L *et al.* (2013) A novel method for cross-species gene expression analysis. *BMC Bioinformatics*, **14**, 70.

Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research*, **13**, 2178–89.

Lynch M, Conery JS (2003) The evolutionary demography of duplicate genes. *Journal of structural and functional genomics*, **3**, 35–44.

Maere S, De Bodt S, Raes J *et al.* (2005) Modeling gene and genome duplications in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 5454–9.

Martins EP, Hansen TF (1997) Phylogenies and the Comparative Method: A General Approach to Incorporating Phylogenetic Information into the Analysis of Interspecific Data. *The American Naturalist*, **149**, 646–667.

Mutwil M, Klie S, Tohge T *et al.* (2011) PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. *The Plant Cell*, **23**, 895–910.

Netotea S, Sundell D, Street NR, Hvidsten TR (2014) ComPlEx: conservation and divergence of co-expression networks in A. thaliana, Populus and O. sativa. *BMC Genomics*, **15**, 106.

Ohno S (1970) *Evolution by Gene Duplication*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Ouedraogo M, Bettembourg C, Bretaudeau A *et al.* (2012) The Duplicated Genes Database: Identification and Functional Annotation of Co-Localised Duplicated Genes across Genomes (RK Aziz, Ed,). *PLoS ONE*, **7**, e50653.

Papin JA, Hunter T, Palsson BO, Subramaniam S (2005) Reconstruction of cellular signalling networks and analysis of their properties. *Nature Reviews Molecular Cell Biology*, **6**, 99–111.

Patel R V, Nahal HK, Breit R, Provart NJ (2012) BAR expressolog identification: expression profile similarity ranking of homologous genes in plant species. *The Plant journal : for cell and molecular biology*, **71**, 1038–50.

Powell S, Forslund K, Szklarczyk D *et al.* (2014) eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Research*, **42**, D231–D239.

Proost S, Van Bel M, Sterck L *et al.* (2009) PLAZA: a comparative genomics resource to study gene and genome evolution in plants. *The Plant cell*, **21**, 3718–31.

Rohlfs R V, Nielsen R (2015) Phylogenetic ANOVA: The Expression Variance and Evolution Model for Quantitative Trait Evolution. *Systematic biology*, **64**, 695–708.

Ruan J, Li H, Chen Z *et al.* (2008) TreeFam: 2008 Update. *Nucleic acids research*, **36**, D735-40.

Soltis DE, Albert VA, Leebens-Mack J *et al.* (2009) Polyploidy and angiosperm diversification. *American Journal of Botany*, **96**, 336–348.

Sonnhammer ELL, Östlund G (2015) InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Research*, **43**, D234–D239.

Stergachis AB, Neph S, Sandstrom R *et al.* (2014) Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature*, **515**, 365–70.

Stuart JM, Segal E, Koller D, Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–55.

Thomashow MF (1999) Plant Cold Acclimation: Freezing Tolerance Genes and Regulatory Mechanisms. *Annual review of plant physiology and plant molecular biology*, **50**, 571–599.

Usadel B, Obayashi T, Mutwil M *et al.* (2009) Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant, cell & environment*, **32**, 1633–51.

Vilella AJ, Severin J, Ureta-Vidal A *et al.* (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome research*, **19**, 327–35.

Wang Y, Wang X, Tang H *et al.* (2011) Modes of Gene Duplication Contribute Differently to Genetic Novelty and Redundancy, but Show Parallels across Divergent Angiosperms (SR Proulx, Ed,). *PLoS ONE*, **6**, e28150.

Wilczynski B, Liu Y-H, Yeo ZX, Furlong EEM (2012) Predicting Spatial and Temporal Gene Expression Using an Integrative Model of Transcription Factor Occupancy and Chromatin State (E Segal, Ed,). *PLoS Computational Biology*, **8**, e1002798.

Wolfe CJ, Kohane IS, Butte AJ (2005) Systematic survey reveals general applicability of &quot;guilt-by-association&quot; within gene coexpression networks. *BMC Bioinformatics*, **6**, 227.

Zhao L, Wit J, Svetec N, Begun DJ (2015) Parallel Gene Expression Differences between Low and High Latitude Populations of Drosophila melanogaster and D. simulans (S V. Nuzhdin, Ed,). *PLOS Genetics*, **11**, e1005184.

Paper 1

Comparative transcriptomics provides insight into the evolution of cold response in Pooideae

Lars Grønvold[1*], Marian Schubert[2*], Simen R. Sandve[3], Siri Fjellheim[2] and Torgeir R. Hvidsten[1,4,†]

[1]*Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, NO-1432, Ås, Norway.*

[2]*Department of Plant Sciences, Norwegian University of Life Sciences, Ås NO-1432, Norway.*

[3]*Centre for Integrative Genetics (CIGENE), Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, NO-1432, Ås, Norway.*

[4]*Umeå Plant Science Centre, Department of Plant Physiology, Umeå University, SE-90187, Umeå, Sweden.*

[*]*Contributed equally*

[†]*Author for correspondence: Email: torgeir.r.hvidsten@nmbu.no*

Abstract

Background: Understanding how complex traits evolve through adaptive changes in gene regulation remains a major challenge in evolutionary biology. Over the last ~50 million years, Earth has experienced climate cooling and ancestrally tropical plants have adapted to expanding temperate environments. The grass subfamily Pooideae dominates the grass flora of the temperate regions, but conserved cold-response genes that might have played a role in the cold adaptation to temperate climate remain unidentified.

Results: To establish if molecular responses to cold are conserved throughout the Pooideae phylogeny, we assembled the transcriptomes of five species spanning early to later diverging lineages, and compared short- and long-term cold response in orthologous genes based on gene expression data. We confirmed that most genes previously identified as cold responsive in barley also responded to cold in our barley experiment. Interestingly, comparing cold response across the lineages using 8633 high confidence ortholog groups revealed that nearly half of all cold responsive genes were species specific and more closely related species did not share higher numbers of cold responsive genes than more distantly related species. Also, the previously identified cold-responsive barley genes displayed low conservation of cold response across species. Nonetheless, more genes than expected by chance shared cold response, both based on previously studied genes and based on the high confidence ortholog groups. Noticeable, all five species shared short-term cold response in nine general stress genes as well as the ability to down-regulate the photosynthetic machinery during cold temperatures.

Conclusions: We observed widespread lineage specific cold response in genes with conserved sequence across the Pooideae phylogeny. This is consistent with phylogenetic dating and historic temperature data which suggest that selection pressure resulting from dramatic global cooling must have acted on already diverged lineages. To what degree lineage specific evolution acted primarily through gain or loss of cold response remains unclear, however, phylogeny-wide conservation of certain genes and processes indicated that the last common ancestor may have possessed some cold response.

Key words: regulatory evolution, cold response, comparative transcriptomics, Pooideae, adaptation

Background

Adaptation to a changing climate is essential for long term evolutionary success of plant lineages. During the last ~50 million years of climate cooling, several plant species adapted to temperate regions. A key step in this transitioning was the integration of novel temperate climate cues, such as seasonal fluctuations in temperature, in the regulatory network controlling cold stress responses. Here we used the temperate grass subfamily Pooideae as a model system for studying the evolution of gene expression responses to cold stress.

The temperate grass flora is dominated by members of the subfamily Pooideae [1], and the most extreme cold environments are inhabited by Pooideae species. The ancestors of this group were, however, most likely adapted to tropical or subtropical climates [2, 3]. Many Pooideae species experience cold winters (Fig. 1ab) and although a recent study inferred adaptation to cooler environments at the base of the Pooideae phylogeny [4], it is still not known whether the Pooideae's most recent common ancestor (MRCA) already was adapted to cold stress, or if adaptation to cold evolved independently in the Pooideae lineages.

Pooideae is a large subfamily comprising 4200 species [5], amongst them economically important species such as wheat and barley. Given the commercial importance of this group, various aspects of adaptation to temperate climate such as flowering time, cold acclimation, and frost and chilling tolerance have been studied (reviewed by [6–13]). These studies are, however, confined to a handful of species in the species rich clade "core Pooideae" [14] and recently also to its sister clade, containing the model grass *Brachypodium distachyon* [15–17]. It is thus unknown how adaptation to temperate climate evolved in earlier diverging Pooideae lineages and if conserved cold response genes could have promoted the success of this subfamily in temperate regions.

Environmental stress is assumed to be a strong evolutionary force, and the colonization of temperate biomes by Pooideae was likely accompanied by adaptation to cold conditions. A MRCA already adapted to cold (the ancestral hypothesis) offers a plausible basis for the ecological success of the Pooideae subfamily in the northern temperate regions [1]. However, paleoclimatic reconstructions infer a generally warm climate, and a very limited abundance of temperate environments, during the time of Pooideae emergence, around 50 million years ago (Mya) [18–22]. Indeed, it was not before ca. 33.5 Mya, during the Eocene-Oligocene (E-O) transition, that the global climates suddenly began to cool [23, 24] (Fig. 1c). Climate cooling at the E-O transition coincided with the emergence of many temperate plant lineages [25] and may have been an important selection pressure for improved cold tolerance in Pooideae [26, 27]. If the E-O cooling event has been the major evolutionary driving force for cold adaptation in Pooideae grasses, those findings lend support for lineage specific evolution of cold adaptation (the lineage specific hypothesis), as all major Pooideae lineages had already emerged by the time of the E-O transition [2, 28] (Fig. 1ac).

A restricted number of plant lineages successfully transitioned into the temperate region, suggesting that evolving the coordinated set of physiological changes needed to withstand low temperatures is challenging [29]. During prolonged freezing, plants need to maintain the integrity of cell membranes to avoid osmotic stress [30]. Cold and freezing tolerance is associated with the

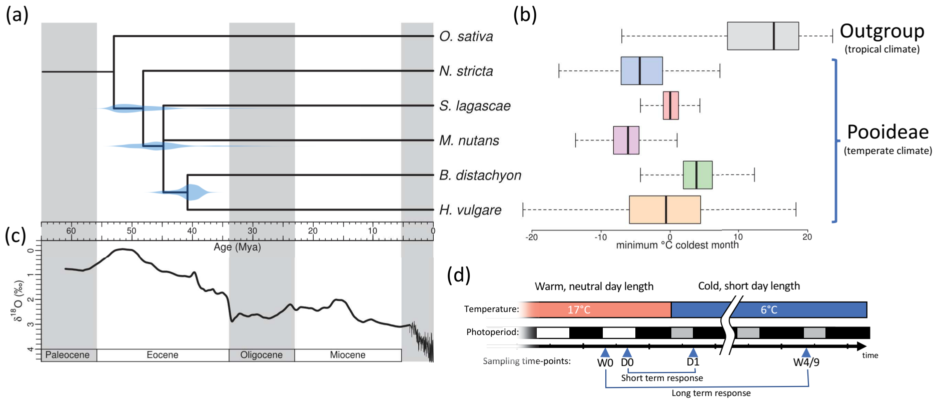


***Figure 1. The Pooideae phylogeny, present and historic temperature data and the experimental design of this study. (a)** Dated phylogenetic tree of the five Pooideae investigated in this study with O. sativa as an out-species. The species phylogeny was inferred from gene trees, with the distribution of mean gene-tree node ages shown in blue. (b) The range of the minimum temperature of the coldest month (WorldClim v1.4 dataset, Bioclim variable 6, 2.5 km^2 resolution [98]) at the species geographical distribution (source: www.gbif.org). (c) Oxygen isotope ratios as a proxy for historic global temperature [18, 22] (d) Experimental design. Plants from five species of Pooideae were subjected to a drop in temperature and shorter days to induce cold response. Leaf material was sampled on the day before the onset of cold (W0 and D0), once 8 hours after cold (D1) and two times after 4 and 9 weeks (W4/W9). Short-term response was identified by contrasting gene expression in time points D0 and D1, while long-term response was identified contrasting W0 and W4/W9.*

ability to cold acclimate, which is achieved through a period of extended, non-freezing cold triggered by the gradually lower temperature and day-length in the autumn. During cold acclimation, a suite of physiological changes governed by diverse molecular pathways results in an increase in the sugar content of cells, change in lipid composition of membranes and synthesis of anti-freeze proteins [13, 31]. Also, low non-freezing temperatures may affect plant cells by decreasing metabolic turnover rates, inhibiting the photosynthetic machinery and decreasing stability of biomolecules (e.g. lipid membranes) [10, 12]. Several studies have used transcriptomics to compare cold stress response, however, they focused on closely related taxa or varieties within model species [17, 32–36]. As such, these studies were not able to investigate evolutionary mechanisms underlying adaptation to cold climates of entire clades.

Here, we used *de novo* comparative transcriptomics across the Pooideae phylogeny to study the evolution of cold adaptation in Pooideae. Specifically, we aim to establish if molecular responses to cold are conserved in the Pooideae subfamily or if they are the result of lineage specific evolution. The transcriptomes of three non-model species (*Nardus stricta*, *Stipa lagascae* and *Melica nutans*), which belong to early diverging lineages, were compared to the transcriptomes of the model grass *Brachypodium distachyon* and the core Pooideae species *Hordeum vulgare* (barley). We found that only a small number of genes were cold responsive in all the investigated species, and that lineage specific evolution has been prominent in the different Pooideae lineages.

Results

To investigate the evolution of cold response in Pooideae, we sampled leaf material in five species before and after subjecting them to a drop in temperature and shorter days (Fig. 1d). RNA-sequencing (RNA-Seq) was used to reveal the short and long term cold response of transcripts, and the conservation of these responses was analyzed in the context of ortholog groups.

De novo transcriptome assembly identified 8633 high confidence ortholog groups

The transcriptome of each species was assembled *de novo* resulting in 146k-282k contigs, of which 68k-118k were identified as containing coding sequences (CDS, Table S1). Ortholog groups (OGs) were inferred by using the protein sequences from the five *de novo* assemblies, as well as the reference genomes of *L. perenne*, *H. vulgare*, *B. distachyon, Oryza sativa*, *Sorghum bicolor* and *Zea mays*. The five assembled Pooideae species were represented with at least one transcript in 24k-33k OGs (Table S1).

A set of 8633 high confidence ortholog groups (HCOGs) was identified after filtering based on gene tree topology and species representation (Table S1, Table S2). We then created a single cross species expression matrix, with HCOGs as rows and samples as columns, by summing the expression of paralogs and setting the expression of missing orthologs to zero (Table S3).

A dated species tree of the Pooideae

A dated species tree were generated based on ortholog groups with exactly one sequence from each of the five Pooideae and rice, and using prior knowledge about the divergence times of *Oryza*-Pooideae [37] and *Brachypodium*-*Hordeum* [28] (Fig. 1a). In the most common gene tree topology, *S. lagascae* or *M. nutans* formed a monophyletic clade, but topologies where either *S. lagascae* or *M. nutans* diverged first were also common (Fig. S1). Due to this uncertainty regarding the topology, *S. lagascae* and *M. nutans* branches were collapsed to a polytomy in the consensus species tree.

Expression clustering indicated a common global response to cold

To investigate broad scale expression patterns in cold response, we clustered all samples (including replicates) after scaling the expression values of each gene to remove differences in mean expression levels between species (Fig. 2a). This clustering revealed the differential effects of the treatments and resulted in a tree with replicates, and then time points, clustering together. Before scaling, the samples clustered by species (data not shown). An exception was time points W4 and W9, which tended to cluster together and by species, indicating that responses after 4 and 9 weeks were very similar. The fact that time points mostly clustered together before species indicated a common response to cold across species. We also observed a clear effect of the diurnal rhythm, with time points sampled in the morning (W0, W4 and W9) forming one cluster and time points sampled in the afternoon (D0 and D1) forming another. This diurnal effect might have resulted in

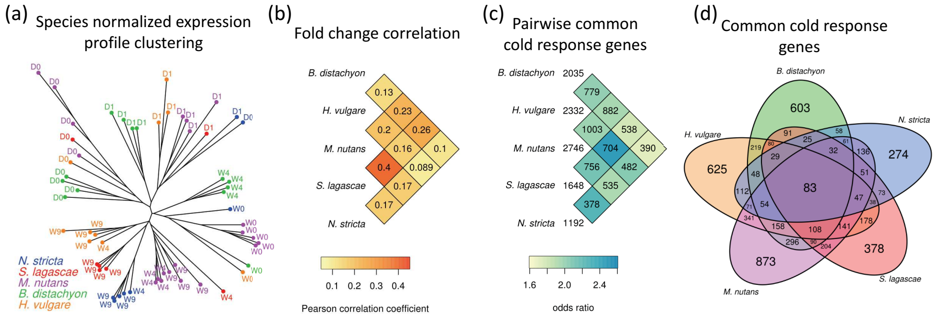


Figure 2. ***Comparison of cold response across the Pooideae.*** *(a) Expression clustering of the samples. The tree was generated by neighbor-joining of Manhattan distances given as the sum of log fold changes between all highly expressed genes after subtracting the mean expression per species. Each tip corresponded to one sample. (b) The Pearson correlations of log fold expression changes (only short-term cold response is shown) between pairs of species. The correlations were computed based on the high confidence ortholog groups (HCOG). (c) The number of differentially expressed genes per species and shared between pairs of species. The statistical significance of the overlaps between pairs of species were indicated with odds ratios. (d) The number of differentially expressed genes in each species (FDR adjusted p-value < 0.05 and absolute fold change > 2 in either short- or long-term cold response) and overlap between species.*

more unreliable estimates of the long term cold response in *S. lagascae* since for this species the afternoon sample (D0) was used to replace the missing morning sample (W0).

Many cold responsive genes were species specific

We next examined similarities in short and long term cold response between species by analysing changes in gene expression from before cold treatment to eight hours and 4-9 weeks after cold treatment (Fig. 1d). For all species pairs, there was a low, but statistically significant, correlation between the expression fold changes of orthologs in HCOGs (Fig. 2b). A similar pattern was observed when investigating the number of orthologs classified as differentially expressed in pairs of species (FDR adjusted p-value < 0.05 and fold change > 2, Table S4, see Methods): these numbers were low compared to the number of differentially expressed genes (DEGs) in individual species, but higher than expected by chance (Fisher´s exact test p < 0.05, Fig. 2c). Finally, the number of orthologs with differential expression in more than two species were very low (Fig. 2d), with only 83 DEGs common to all five species. Noticeably, neither the similarities in differential expression nor the fold change correlations reflected the phylogenetic relationship between the species, that is, the cold responses of related species were not more similar than that of distantly related species (Fig. 2bc).

Table 1. High confidence ortholog groups with conserved cold response in all five Pooideae. *S = short-term response, L = long-term response. ↗= up regulated, ↘= down regulated. Annotations were inferred from literature using orthologs. These 16 genes were the subset of the 83 genes in Fig. 2d with the same type of cold response (short- or long-term) in the same direction (up- or down-regulation) in all five species.*

Bd ortholog	Description and relation to stress response	Response
Bradi2g39230	**HyperOSmolality-gated CA2+ permeable channel (OSCA)** - Stress-activated calcium channels [72] that are highly conserved in eukaryotes [73]. In *Oryza sativa*, OSCA genes are differentially expressed in response to osmotic stress [74].	S↗
Bradi2g06830	**Calcium-binding EF-hand containing calcium exchange channel (EF-CAX)** - Calcium ions are important mediators of abiotic stress in plants [75, 76]. Expression of calcium binding proteins correlates with exposure to cold stress in several plants, e.g. *Arabidopsis thaliana* [3[30], *Musa × paradisiaca* [41] and *Hordeum vulgare* [38].	S↗
Bradi2g05226	**GIGANTEA** - Promotes flower development in plants [77]. In *Arabidopsis thaliana*, this gene is involved in CBF-independent freezing tolerance [78, 79], and is responsive to cold in *Zea mays* [80]. Also part of the circadian clock.	S↗
Bradi4g24967	**Arabidopsis Pseudo-Response Regulator 3-like (AtPRR3-like)** - AtPRR3 is a member of the circadian clock quintet AtPRR1/TOC1 [81, 82]. No association to stress response found in literature. However, AtPRR3-like might be closer related to AtPRR5/9 than to AtPRR3 (See Bradi4g36077, PRR95).	S↗
Bradi2g09060	**Triacylglycerol lipase, alpha/beta-Hydrolase superfamily** - Studies in *Arabidopsis thaliana* [83] and *Ipomoea batatas* [84] suggest that genes with alpha/beta-Hydrolase domains respond to osmotic stress. In *Triticum monococcum*, atriacylglycerol lipase was induced by pathogen stress [85].	S↗
Bradi2g07480	**Late-Embryogenesis-Abundant protein 14 (LEA-14)** - Responsive to drought, salt and cold stress in *Arabidopsis thaliana* [86, 87], Betula pubescens [88] and *Brachypodium distachyon* [89].	S↗
Bradi1g04150	**SNAC1-like / NAC transcription factor 67** - NAC transcription factors mediate abiotic stress responses. Osmotic stress increases the expression of SNAC1 in *Oryza sativa* [43], NAC68 in *Musa × paradisiaca* [41, 42] and NAC67 *Triticum aestivum* [44].	S↗
Bradi4g36077	**Pseudo-Response Regulator 95 (PRR95)** - Homologous to conserved circadian clock gene AtPRR5/9 [90, 91]. AtPRR5 gene is cold regulated in Arabidopsis thaliana [92] and PRR95 is cold responsive in Zea mays [80].	S↗
Bradi2g43040	**DnaJ chaperon protein** - DnaJ co-chaperons are vital in stress response and has been found to be involved in maintenance of photosystem II under chilling stress and enhances drought tolerance in tomato [93, 94].	S+L↗
Bradi3g33080	**Glycogenin GlucuronosylTransferase (GGT)** - GGT belongs to the GT8 protein family [95]. In *Oryza sativa*, OsGGT transcripts are induced in submerged plants and respond to various abiotic stresses except cold [96, 97].	L↗
Bradi1g04500	**Major facilitator superfamily transporter** - Association to stress response unknown.	L↗
Bradi3g14080	**Glycosyl transferase** - Association to stress response unknown.	L↗
Bradi1g35357	**Uncharacterized membrane protein** - Association to stress response unknown.	S↗
Bradi2g48850	**Uncharacterized protein** - Association to stress response unknown.	S↗
Bradi1g33690	**Uncharacterized protein** - Association to stress response unknown.	S↗
Bradi1g07120	**Putative S-adenosyl-L-methionine-dependent methyltransferase** - Association to stress response unknown.	L↘

Shared cold response genes included known abiotic stress genes

Sixteen genes shared the same cold response (short- or long-term) in the same direction (up or down) in all five Pooideae species, thus representing a response to cold that might have been conserved throughout the evolution of Pooideae (Table 1). Nine of these shared cold responsive

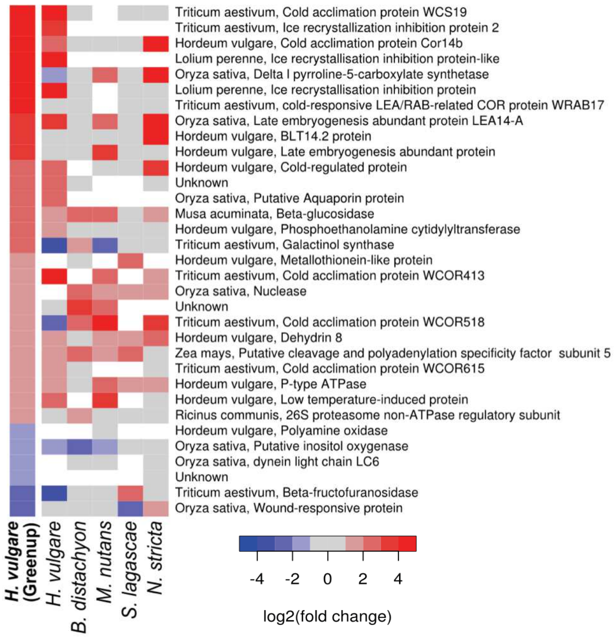


Figure 3. Comparison of cold response to previous studies*. A reference set of H. vulgare genes independently shown to respond to cold in several studies [38] is compared to our data using short-term log fold change values. White cells represent missing orthologs. Grey cells represent orthologs that were not differentially expressed (not DEGs).*

genes belonged to families known to be involved in cold stress or other abiotic stress responses in other plant species. The most common type of response was short-term up regulation, indicating that stress response, as opposed to long-term acclimation response, is potentially more conserved.

Identified cold response genes confirmed previous findings

We compared the cold response genes from our data to a compilation of *H. vulgare* genes shown to be responsive to low temperature in several previous microarray studies, subsequently referred to as the Greenup genes (table S10 in [38]). We could map 33 of these 55 genes to unique OGs, of which 11 were HCOGs. We observed significant similarity in cold response between the 33 Greenup genes and the short-term cold response observed in our data (Fig. 3); for all five species ($p < 0.05$). However, this similarity was noticeably larger in *H. vulgare* than in the other four species. This comparison showed that our transcriptome data was consistent with previous findings in *H. vulgare*, and that cold response genes identified in *H. vulgare* exhibits some cold response in other Pooideae.

Photosynthesis was down-regulated under cold stress

To identify biological processes that evolved regulation during different stages of Pooideae evolution, we targeted gene sets that were exclusively differentially expressed in all species within a clade in the phylogentic tree (i.e. branch specific DEGs), and tested these for enrichment of Gene Ontology (GO) biological process annotations (Fig. 4a). For the genes that were differentially

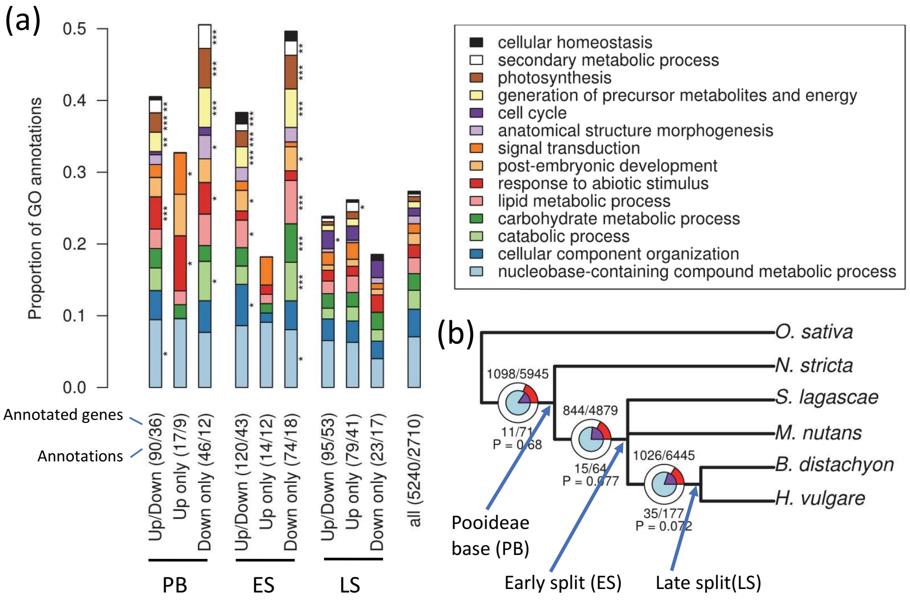


Figure 4. Gene Ontology enrichment and positive selection in branch specific cold responsive genes. *(a) Gene ontology enrichment analysis of high confidence ortholog groups (HCOG) that were differentially expressed (DEGs) in all species (PB), only in species after N. stricta split off (ES) or only in B. distachyon and H. vulgare (LS) (* P < 0.05, ** P < 0.01, *** P < 0.005, Fisher's exact test). Both the number of annotated genes and the number of annotations were indicated for each set of branch specific DEGs. (b) Positive selection at different stages in Pooideae evolution. The circles represent the high confidence ortholog groups that were tested for positive selection at each split (see Methods for the criteria). The inner blue circle represented HCOGs with branch specific differential expression (i.e. with genes that were cold responsive exclusively in the species under the respective branch) while the outer circle represented all other HCOGs. The purple and red pie-slices represented the proportions of HCOGs with positive selection (P < 0.05). The P-value indicated the overrepresentation of positive selection among the branch specific DEGs (Hypergeometric test).*

expressed in all our species (Pooideae base [PB] in Fig 4b), we found enrichments for annotations related to response to abiotic stimulus, photosynthesis and metabolism. Dividing the branch specific DEGs into up- or down-regulated genes revealed up-regulation of signal transduction (two pseudo response regulators and diacylglycerol kinase 2 (DGK2)) and abiotic stimulus (Gigantea, LEA-14, DnaJ and DGK2), and down-regulation of photosynthesis and metabolism. For the genes that were exclusively differentially expressed in all species except *N. stricta* (early split [ES] in Fig. 4b), down-regulated genes were again enriched for GO annotations related to metabolism and photosynthesis.

Cold response genes were associated with positive selection on amino acid content

For each HCOG, we tested for positive selection in coding sequences at each of the internal branches of the species tree. The tests were only performed on the branches where the gene tree topology was compatible with the species tree topology. 16-18% of the HCOGs showed significant signs of positive selection ($P < 0.05$) depending on the branch (Fig. 4b). Next, we tested for overrepresentation of positive selection among the branch specific DEGs. There was a tendency that gain of cold response was associated with positive selection at the early split (ES) and late split (LS) branches ($P = 0.077$ and $P = 0.072$, respectively) (Fig. 4b).

Discussion

The ecological success of the Pooideae subfamily in the northern temperate regions must have critically relied on adaptation to colder temperatures. However, it is unclear how this adaptation evolved within Pooideae. To test whether molecular responses to cold are conserved in the Pooideae subfamily, we applied RNA-seq to identify short- and long-term cold responsive genes in five Pooideae species ranging from early diverging lineages to core Pooideae species. Since three of the species lacked reference genomes, we employed a *de novo* assembly pipeline to reconstruct the transcriptomes. We showed that this pipeline could recover a set of *H. vulgare* genes previously identified as cold responsive and that most of these genes were also cold responsive in our data (Fig. 3). In order to compare the five transcriptomes, we compiled a set of 8633 high confidence ortholog groups (HCOGs) with resolved gene tree topologies. Clustering of scaled expression data based on these ortholog groups arranged samples according to replicates, then time points and finally species, indicating that cold response represents a distinct signal in the data and confirming the soundness of the approach (Fig. 2a).

Lineage specific adaptations to cold climates

Based on conserved genes in the 8,633 HCOGs, substantial portions of the individual Pooideae transcriptomes responded to cold (1000-3000 genes, ~10-30%). Although pairs of species shared a statistically significant number of cold responsive genes, nearly half of all responsive genes were species specific, with closely related species sharing approximately the same fraction of cold responsive genes as more distantly related species (no phylogenetic pattern, Fig. 2c-d). Moreover, only a small number of genes responded to cold in all the investigated species (83 genes, Fig. 2d). Even fewer genes responded similarly to cold in all species (e.g. short-term up-regulation, 16 genes, Table 1) and these shared cold response genes primarily included general abiotic stress genes clearly not representative of all the different molecular pathways constituting a fully operational cold response program. We also observed low (although statistically significant) correlations in expression fold changes between pairs of species, a result that was independent of the ability to correctly classify genes as differentially expressed. These results were based on conserved, high confidence ortholog groups that excluded complex families with duplication events shared by two or more species. Since many of the previously described *H. vulgare* cold responsive genes belonged to such complex families, we specifically investigated the regulation of these previously described genes using all ortholog groups, and again found that few genes displayed shared cold response across all species (Fig. 3). Taken together, our findings appear consistent with the lineage specific hypothesis of Pooideae cold adaptation, where cold induced regulation to a substantial degree has been gained or lost independently in the different lineages. However, it is unclear to what extent the most recent common ancestor (MRCA) of the Pooideae possessed response to cold, since we cannot determine whether lineage specific gain of cold response was more (or less) prominent than lineage specific loss.

The drastic cold stress during the E-O transition was likely an important cause for the evolution of cold adaptation in Pooideae. Previous studies have shown that many temperate plant lineages emerged during the E-O transition [25] and that the expansion of well-known cold responsive gene families in Pooideae coincided with this transition [15, 26]. From the dated phylogeny (Fig. 1a) as well as from earlier studies of the Pooideae phylogeny [2, 28], it is clear that all major Pooidaee lineages, including the core Pooideae, had emerged by the late Eocene. Hence, the five lineages studied here experienced the E-O transition as individual lineages (Fig. 1c). The observation that the five Pooideae lineages emerged during a relatively warm period before the E-O transition, that these species harbored high numbers of cold responsive genes specific to only one or two species and that the shared cold response genes showed no phylogenetic pattern, together suggests that a significant portion of the cold response in Pooideae lineages were gained during the last 40 M years. During this period, temperatures were constantly decreasing and dramatic cooling events took place, such as the E-O transition and the current Quaternary Ice Age. This must have placed a very strong selection pressure on plants favoring the evolution of cold adaptations, and would have acted on already diverged lineages and not on any ancestor of these lineages. However, although the independent gain of cold response would imply that the ancestor of the Pooideae possessed limited molecular cold response, further studies involving more species would be needed to definitively conclude on this issue.

Our results suggest that the Pooideae lineages evolved cold response programs that included a significant fraction of non-orthologous genes. This conclusion was reached both based on the conserved, high confidence ortholog groups, and based on a set of previously identified cold response genes (Fig. 3), and implies that these genomes contain many functionally redundant genes that can be co-opted in different combinations into the functional cold response program of the Pooideae. Gain and loss of protein coding genes might also have played an important role in the evolution of cold response of these species, however, here we focused on the gain and loss of regulatory response to cold in genes that were widely conserved across the phylogeny. It is worth noticing that although we observed many species specific cold response genes, all species pairs displayed a statistically significant correlation in cold response across all HCOGs (Fig. 2b) and a statistically significantly overlap in cold responsive genes (Fig. 2c). This could reflect that some genes are associated with biochemical functions more important or more suited for cold response than other genes [39, 40], and that different species thus have ended up co-opting orthologous genes into their cold response program more often than expected by chance.

An adaptive potential in the Pooideae ancestor

Multiple independent origins of cold adaptation raise the question whether connecting traits exists in the evolutionary history of the Pooideae that can explain why the Pooideae lineages were able to shift to the temperate biome. The genes that showed a conserved cold response across all five species (Table 1) might have gained cold responsiveness in the MRCA. Subsequently, these genes might have increased the potential of Pooideae lineages to adapt to a cold temperate climate. Nine of these conserved cold response genes are involved in response to abiotic stresses in other plants, such as osmotic stress and drought. The *SNAC1-like / NAC transcription factor 67* is one example, with homologs induced by osmotic stress in the three monocts *Musa × paradisiaca* [41, 42]*, Oryza sativa* [43] and *Triticum aestivum* [44]. Co-option of such genes into a cold-responsive pathway might have been the key to acquire cold tolerance. In fact, other studies have implied that drought tolerance might have facilitated the shift to temperate biomes [26, 45, 46]. Interestingly, most of the conserved genes were short-term cold responsive (Table 1) and this observation strengthens the hypothesis that existing stress genes might have been the first to be co-opted into the cold response program.

Also, three of the conserved cold responsive genes (GIGANTEA, PRR95 and AtPRR3-like) were associated with the circadian clock that is known to be affected by cold [47–49]. This might suggest that clock genes have had an important function in the Pooideae cold adaptation, for example by acting as a signal for initiating the cold defense. More generally, transcripts involved in photosynthesis and response to abiotic stimuli were significantly enriched among the genes with cold response in all species (Fig. 4a). An expanded stress responsiveness towards cold stress and the ability to down-regulate the photosynthetic machinery during cold temperatures to prevent photoinhibition might have existed in the early evolution of Pooideae. In conclusion, the conserved stress response genes discussed here may have represented a fitness advantage for the Pooideae ancestor in the newly emerging environment with incidents of mild frost, allowing time to evolve the more complex physiological adaptations required to endure the temperate climate with strong seasonality and cold winters that emerged following the E-O transition [23]. Consistent with this, Schubert et al. (unpublished) showed that the fructan synthesis and ice recrystallization inhibition protein gene families known to be involved in cold acclimation in core Pooideae species [10] evolved around the E-O split, whereas also earlier evolving Pooideae species show capacity to cold acclimate.

Evolution of coding and regulatory sequences

The molecular mechanisms behind adaptive evolution are still poorly understood, although it is now indisputably established that novel gene regulation plays a crucial role [50]. The evolution of gene regulation proceeds by altering non-coding regulatory sequences in the genome, such as (cis-) regulatory elements [51], and has the potential to evolve faster than protein sequence and function. The high number of species specific cold response genes observed in this study is thus most consistent with the recruitment of genes with existing cold tolerance functions by means of regulatory evolution. However, previous studies have also pointed to the evolution of coding sequences [27] as underlying the acquisition of cold tolerance in Pooideae. To investigate possible coding evolution, we tested for the enrichment of positive selection among branch specific cold responsive genes (Fig. 4b). Although not statistically significant, there was a tendency for positive selection in genes gaining cold response in a period of gradual cooling preceding the E-O event. Thus, we saw evidence of both coding and regulatory evolution playing a role in cold adaptation in Pooideae, and that these processes may have interacted. Finally, gene family expansion has previously been implied in cold adaptation in Pooideae [15, 26]. As previously discussed, the conservative filtering of ortholog groups employed in this study removed complex gene families containing duplication events shared by two or more species. Interestingly, out of the 33 previously described *H. vulgare* cold responsive genes (Fig. 3), as many as 22 were not included in the high confidence ortholog groups, the main reason being that they belonged to gene families with duplications involving two or more species. This observation thus confirms that duplication events are a relatively common feature of cold adaptation. Although *de novo* assembly of transcriptomes from short-read RNA-Seq data is a powerful tool that has vastly expanded the number of target species for conducting transcriptomic analysis, the approach has limited power to distinguish highly similar transcripts such as paralogs. Further insight into the role of duplication events in Pooideae cold adaptation would therefore benefit immensely from additional reference genomes.

Conclusion

Here we investigated the cold response of five Pooideae species, ranging from early diverging lineages to core Pooideae species, to elucidate evolution of adaptation to cold temperate regions. We observed extensive lineage specific cold response that seems to have evolved chiefly after the *B. distachyon* lineage and the core Pooideae diverged, possible initiated by the drastic temperature drop during the E-O transition. However, we do also see signs of conserved response that potentially represents a shared potential for cold adaptation that explain the success of Pooideae in temperate regions. This included several general stress genes with conserved short-term response to cold as well as the conserved ability to down-regulate the photosynthetic machinery during cold temperatures. Taken together, our observations are consistent with a scenario where the biochemical functions needed for cold response were present in the Pooideae common ancestor, and where different Pooideae lineages have assembled, in parallel, different overlapping subsets of these genes into fully functional cold response programs through the relatively rapid process of regulatory evolution. Answering whether gain or loss of cold response was most prominent in the evolution of cold adaption in these species, and thus whether the last common ancestor of the Pooideae possessed extensive cold response or not, will require further studies.

Methods

Plant material, sampling and sequencing

To address our hypothesis, we selected five species to cover the phylogenetic spread of Pooideae. The selected species also represent major, species rich lineages or clades in the Pooideae subfamily, or belong to very early diverging lineages [5]. Seeds were collected either in nature or acquired from germplasm collections for the following five species:

Nardus stricta, (2n=2x=26) is a perennial species, distributed in Europe, western parts of Asia, and North Africa and introduced to New Zealand and North America. Seeds were collected in July 2012 in Romania, [46.69098, 22.58302].

Melica nutans L. (2n=2x=18) is a perennial species distributed in temperate parts of Eurasia [52–54]. Seeds were collected in Germany, [50.70708, 11.23838], in June 2012.

S*tipa lagascae* Roem & Schult. (2n=2x=22) is a perennial species that is distributed in temperate regions around the Mediterranean Sea and parts of temperate West Asia. Seeds from accession PI 250751 were acquired from U.S. National Plant Germplasm System (U.S.-NPGS) via Germplasm Resources Information Network (GRIN)*,*

Brachypodium distachyon (L.) P. Beauv. (2n=2x=10) is an annual species natively distributed in Europe, East Africa and temperate parts of West Asia. Seeds for accession 'Bd1-1' (W6 46201) were acquired from U.S.-NPGS via GRIN.

Hordeum vulgare L. (2n=2x=14) seeds for cultivar 'Igri' were provided by Prof. Åsmund Bjørnstad, Department of Plant Sciences, Norwegian University of Life Sciences, Norway.

Seeds were germinated and initially grown in a greenhouse at a neutral day length (12 hours of light), 17°C and a minimum artificial light intensity of 150 µmol/m2s. To ensure that individual plants were at comparable developmental stages at sampling, the onset of treatment for different species were based on developmental stages rather than absolute time. Most importantly, none of the plants had transitioned from vegetative to generative phase, as this could have affected the cold response. Plants were grown until three to four leaves had emerged for *M. nutans*, *S. lagascae*, *B. distachyon* and *H. vulgare*, or six to seven leaves for *N. stricta* (which is a cushion forming grass that produces many small leaves compared to its overall plant size). Depending on the species, this process took one (*H. vulgare*), three (*B. distachyon* and *S. lagascae*), six (*M. nutans)* or eight (*N. stricta*) weeks from the time of sowing. Subsequently, plants from all species were randomized and distributed to two cold chambers with short day (8 hours of light), constantly 6°C and a light intensity of 50 µmol/m2s. Plants were kept in cold treatment for the duration of the experiment. Leaf material for RNA isolation was collected i) in the afternoon (at zeitgeber (ZT) 8) on the day before cold treatment (D0) and in the afternoon (ZT 8) on the first day after cold treatment (after 8 hours of cold treatment, D1) and ii) in the morning (ZT 0) before cold treatment (W0), 4 weeks after cold treatment (W4) and 9 weeks after cold treatment (W9) (Fig. 1d). The sampling time points were chosen to be able to separate chilling stress responses (first day of treatment) and long-term responses that represent acclimation to freezing temperatures (4 and 9 weeks). Flash frozen leaves were individually homogenized using a TissueLyser (Qiagen Retsch) and total RNA was isolated (from each leaf) using RNeasy Plant Mini Kit (Qiagen) following the manufacturer's instructions. The purity and integrity of total RNA extracts was determined using a NanoDrop 8000 UV-Vis Spectrophotometer (Thermo Scientific) and 2100 Bioanalyzer (Agilent), respectively. For each time point, RNA extracts from five leaves sampled from five different plants were pooled and sequenced as a single sample. In addition, replicates from single individual leaves were sequenced for selected timepoints (see Table S1 and "Differential expression" below). Two time points lacked expression values: W9 in B. distachyon (RNA integrity was insufficient for

RNA sequencing) and W0 in S. lagascae (insufficient supply of plant material). Samples were sent to the Norwegian Sequencing Centre, where strand-specific cDNA libraries were prepared and sequenced (paired-end) on an Illumina HiSeq 2000 system. The raw reads are available in the ArrayExpress database (www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-5300.

Transcriptome assembly and ortholog inference

Using Trimmomatic v0.32 [55], all reads were trimmed to a length of 120 bp, Illumina TruSeq adapters were removed from the raw reads, low quality bases were trimmed using a sliding window of 40 bp and an average quality cut-off of 15 and reads below a minimum length of 36 bp were discarded. Read quality was checked using fastqc v0.11.2. For each species, transcripts were assembled *de novo* with Trinity v2.0.6 [56] (strand specific option, otherwise default parameters) using reads from all samples. Coding sequences (CDS) were identified using TransDecoder rel16JAN2014 [57]. Where Trinity reported multiple isoforms, only the longest CDS was retained. Ortholog groups (OGs) were constructed from the five *de novo* transcriptomes and public reference transcriptomes of *H. vulgare* (barley_HighConf_genes_MIPS_23Mar12), *B. distachyon* (brachypodium v1.2), *O. sativa* (rap2), *Z. mays* (ZmB73_5a_WGS)*, S. bicolor* (sorghum 1.4) and *L. perenne* (GenBank TSA accession GAYX01000000) using OrthoMCL v2.0.9 [58]. All reference sequences except *L. perenne* were downloaded from http://pgsb.helmholtz-muenchen.de/plant/plantsdb.jsp. A summary of the results is provided in Table S1.

High confidence ortholog groups

To compare gene expression across Pooideae, we required ortholog groups containing one gene from each species that all descended from a single gene in the Pooideae ancestor. As the ortholog groups (OGs) inferred using orthoMCL sometimes cluster more distantly related homologs as well as include both paraphyletic and monophyletic paralogs, we further refined the OGs by phylogenetic analysis. Several approaches to phylogenetic refinement has been proposed previously (see e.g. [59]). Here we first aligned protein sequences within each OG using mafft v7.130 [60] and converted to codon alignments using pal2nal v14 [61]. Gene trees were then constructed from the codon alignments using Phangorn v1.99.14 [62] (maximum likelihood GTR+I+G). Trees with apparent duplication events before the most recent common ancestor of the included species were split into several trees. This was accomplished by identifying in-group (Pooideae) and out-group (*Z. mays*, *S. bicolor* and *O. sativa*) clades in each tree, and then splitting the trees so that each resulting sub-tree contained a single out-group and a single in-group clade. Finally, we only retained the trees were all species in the tree formed one clade each (i.e. only monophyletic paralogs), *B. distachyon* and *H. vulgare* formed a clade and at least three of the five studied species were included. These trees constituted the high confidence ortholog groups (HCOGs).

De novo assembly followed by ortholog detection resulted in higher numbers of monophyletic species-specific paralogs than the number of paralogs in the reference genomes of *H. vulgare* and *B. distachyon*. This apparent overestimation of paralogs was most likely the result of the *de novo* procedure assembling alleles or alternative transcript isoforms into separate contigs. We also observed some cases where the number of paralogs were under-estimated compared to the references, which may be due to low expression of these paralogs or the assembler collapsing paralogs into single contigs. Since the *de novo* assembly procedure did not reliably assemble paralogs, we chose to represent each species in each HCOG by a single read-count value equal to the sum of the expression of all assembled paralogs. By additionally setting counts for missing orthologs to zero, we created a single cross species expression matrix with HCOGs as rows and samples as columns (Table S3).

Species tree

Ortholog groups with a single ortholog from each of the five *de novo* Pooideae species and *O. sativa* (after splitting the trees, see "High confidence ortholog groups") were used to infer dated gene trees. To this end, BEAST v1.7.5 [63] was run with an HKY + Γ nucleotide substitution model using an uncorrelated lognormal relaxed clock model. A Yule process (birth only) was used as prior for the tree and the monophyly of the Pooideae was constrained. Prior estimates for the *Oryza*-Pooideae (53 Mya [SD 3.6 My], [37]) and *Brachypodium-Hordeum* (44.4 Mya [SD 3.53 My], [28]) divergence times were used to define normally distributed age priors for the respective nodes in the topology. MCMC analyses were run for 10 million generations and parameters were sampled every 10.000 generation. For each gene tree analysis, the first 10 percent of the estimated trees were discarded and the remaining trees were summarized to a maximum clade credibility (MCC) tree using TreeAnnotator v1.7.5. The topology of the species tree was equal to the most common topology among the 3914 MCC trees, with internal node ages set equal to the mean of the corresponding node age distributions of the MCC gene trees.

Differential expression

Reads were mapped to the *de novo* transcriptomes using bowtie v1.1.2 [64], and read counts were calculated with RSEM v1.2.9 [65]. In HCOGs, read counts of paralogs were summed (analogous to so called monophyly masking [66]) and missing orthologs were assumed to not be expressed (i.e. read counts equal to zero). To identify conserved and diverged cold response across species, we probed each HCOG for differentially expressed genes (DEGs). Specifically, DEGs were identified using DESeq2 v1.6.3 [67] with a model that combined the species factor and the timepoint factor (with timepoints W4/9 as a single level). Pooled samples provided robust estimates of the mean expression in each time point. To also obtain robust estimates of the variance, the model assumed common variance across all timepoints and species within each HCOG, thus taking advantage of both biological replicates available for individual time points within species and the replication provided by analysing several species. For each species, we tested the expression difference between D0 and D1 (short-term response) and the difference between W0 and W4/9 (long-term response) (Fig. 1d). *B. distachyon* lacked the W9 samples and long-term response was therefore based on W4 only. *S. lagascae* lacked the W0 sample and long-term response was therefore calculated based on D0. As a result, the observed diurnal effect (Fig. 2a) might have resulted in more unreliable estimates of the long term cold response in *S. lagascae* since for this species the afternoon sample (D0) was used to replace the missing morning sample (W0). Genes with a false discovery rate (FDR) adjusted p-value < 0.05 and a fold change > 2 were classified as differentially expressed.

Sample clustering

Sample clustering was based on read counts normalized using the variance-stabilizing transformation (VST) implemented in DESeq2 (these VST-values are essentially log transformed). HCOGs that lacked orthologs from any of the five species, or that contained orthologs with low expression (VST < 3), were removed, resulting in 4981 HCOGs used for the clustering. To highlight the effect of the cold treatment over the effect of expression level differences between species, the expression values were normalised per gene and species: First, one expression value was obtained per timepoint per gene by taking the mean of the replicates. Then, these expression values were centered by subtracting the mean expression of all timepoint. Distances between all pairs of samples were calculated as the sum of absolute expression difference between orthologs in the 4981 HCOGs (i.e. manhattan distance). The tree was generated using neighbor-joining [68].

Comparison with known cold responsive genes

A set of *H. vulgare* genes independently identified as cold responsive were acquired from supplementary table S10 in [38]. These genes were found to be responsive to cold in three independent experiments with Plexdb accessions BB64 [69], BB81 (no publication) and BB94 [38]. The probesets of the Affymetrix Barley1 GeneChip microarray used in these studies were blasted (blastx) against all protein sequences in our OGs. Each probe was assigned to the OG with the best match in the *H. vulgare* reference. If several probes were assigned to the same OG, only the probe with the best hit was retained. Correspondingly, if a probe matched several paralogs within the same OG, only the best match was retained. DESeq2 was used to identify short-term response DEGs for all transcripts in all OGs (i.e. this analysis was not restricted to the HCOGs), and these were compared to DEGs from [38]. The statistical significance of the overlap between our results and those reported in [38] was assessed for each species by counting the number of genes that had the same response (up- or down-regulated DEGs) and comparing that to a null distribution. The null distribution was obtained from equivalent counts obtained from 100 000 trials where genes were randomly selected from all expressed genes (mean read count > 10) with an ortholog in *H. vulgare*.

Gene ontology enrichment tests

Gene Ontology (GO) annotations for *B. distachyon* were downloaded from Ensembl Plants Biomart and assigned to the HCOGs. The TopGO v2.18.0 R package [70] was used to calculate statistically significant enrichments (Fisher's exact test, $p < 0.05$) of GO biological process annotations restricted to GO plant slim in each set of branch specific DEGs using all annotated HCOGs as the background. Branch specific DEGs were those genes that were exclusively differentially expressed in all species within a clade in the phylogenetic tree.

Positive selection tests

Each of the HCOGs were tested for positive selection using the branch-site model in codeml, which is part of PAML v4.7 [71]. We only tested branches for positive selection in HCOGs meeting the following criteria: (i) The tested branch had to be an internal branch also in the gene tree (i.e. there was at least two species below the branch). (ii) The species below and above the tested branch in the gene tree had to be the same as in the species tree or a subset thereof. (iii) The first species to split off under the branch had to be the same as in the species tree (for the early split, either *S. lagascae* or *M. nutans* was allowed). We then used the Hypergeometric test to identify statistically significant overrepresentation of positive selection among branch specific DEGs (see "Gene ontology enrichment tests") at the Pooideae base (PB), the early split (ES) and the late split (LS) branches.

Declarations

Availability of data and material

The raw reads, the assembled transcripts and the raw read counts are available in the ArrayExpress database (www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-5300.

Funding

The research was funded by grants from the Nansen Foundation to SF and the TVERRforsk program at the Norwegian University of Life Sciences (NMBU) to SF, TRH and SRS. This work was part of the PhD projects of MS and LG funded by NMBU.

Authors´ contributions

All authors designed the experiment. M.S. performed the growth experiments, sampled and prepared RNA for sequencing, helped designing the data analysis pipeline, contributed to the positive selection analysis and performed the phylogenetic analyses. L.G. developed, implemented and conducted the data analysis. All authors interpreted the results. M.S. and L.G. wrote the manuscript with input from S.F., T.R.H., and S.R.S.

Acknowledgements

We thank Åsmund Bjørnstad and USDA-NPGS GRIN for providing seeds of *H. vulgare*, and *B. distachyon* and *S. lagascae*, respectively. For technical assistance handling plants during growth experiments we thank Øyvind Jørgensen. We are grateful to Erica Leder, Thomas Marcussen, Ursula Brandes and Camilla Lorange Lindberg for helpful comments on earlier versions of this manuscript.

References

1. Hartley W. Studies on Origin, Evolution, and Distribution of Gramineae. 5. The Subfamily Festucoideae. Aust J Bot. 1973;21:201–34.

2. Bouchenak-Khelladi Y, Verboom AG, Savolainen V, Hodkinson TR. Biogeography of the grasses (Poaceae): a phylogenetic approach to reveal evolutionary history in geographical space and geological time. Bot J Linn Soc. 2010;162:543–57.

3. Strömberg CAE. Evolution of Grasses and Grassland Ecosystems. Annu Rev Earth Planet Sci. 2011;39:517–44.

4. Edwards EJ, Smith SA. Phylogenetic Analyses Reveal the Shady History of C_4 Grasses. Proc

Natl Acad Sci U S A. 2010;107:2532–7.

5. Soreng RJ, Peterson PM, Romaschenko K, Davidse G, Zuloaga FO, Judziewicz EJ, et al. A worldwide phylogenetic classification of the Poaceae (Gramineae). J Syst Evol. 2015;53:117–37.

6. Tsvetanov S, Atanassov A, Nakamura C. Gold Responsive Gene/Protein Families and Cold/Freezing Tolerance in Cereals. Biotechnol Biotechnol Equip. 2000;14:3–11.

7. Galiba G, Vágújfalvi A, Li C, Soltész A, Dubcovsky J. Regulatory genes involved in the determination of frost tolerance in temperate cereals. Plant Sci. 2009;176:12–9.

8. Thomashow MF. Molecular Basis of Plant Cold Acclimation: Insights Gained from Studying the CBF Cold Response Pathway. Plant Physiol. 2010;154:571–7.

9. Kosová K, Vítámvás P, Prášil IT. Expression of dehydrins in wheat and barley under different temperatures. Plant Sci. 2011;180:46–52.

10. Sandve SR, Kosmala A, Rudi H, Fjellheim S, Rapacz M, Yamada T, et al. Molecular mechanisms underlying frost tolerance in perennial grasses adapted to cold climates. Plant Sci. 2011;180:69–77.

11. Tondelli A, Francia E, Barabaschi D, Pasquariello M, Pecchioni N. Inside the *CBF* locus in Poaceae. Plant Sci. 2011;180:39–45.

12. Crosatti C, Rizza F, Badeck FW, Mazzucotelli E, Cattivelli L. Harden the chloroplast to protect the plant. Physiol Plant. 2013;147:55–63.

13. Preston JC, Sandve SR. Adaptation to seasonality and the winter freeze. Front Plant Sci. 2013;4 June:167.

14. Davis JI, Soreng RJ. Phylogenetic Structure in the Grass Family (Poaceae) as Inferred from Chloroplast DNA Restriction Site Variation. Am J Bot. 1993;80:1444–54.

15. Li C, Rudi H, Stockinger EJ, Cheng H, Cao M, Fox SE, et al. Comparative analyses reveal potential uses of *Brachypodium distachyon* as a model for cold stress responses in temperate grasses. BMC Plant Biol. 2012;12:65.

16. Priest HD, Fox SE, Rowley ER, Murray JR, Michael TP, Mockler TC, et al. Analysis of Global Gene Expression in Brachypodium distachyon Reveals Extensive Network Plasticity in Response to Abiotic Stress. PLoS One. 2014;9:e87499.

17. Colton-Gagnon K, Ali-Benali MA, Mayer BF, Dionne R, Bertrand A, Do Carmo S, et al. Comparative analysis of the cold acclimation and freezing tolerance capacities of seven diploid *Brachypodium distachyon* accessions. Ann Bot. 2014;113:681–93.

18. Zachos J, Pagani M, Sloan L, Thomas E, Billups K. Trends, rhythms, and aberrations in global climate 65 Ma to present. Science. 2001;292:686–93.

19. Eberle JJ, Greenwood DR. Life at the top of the greenhouse Eocene world--A review of the Eocene flora and vertebrate fauna from Canada's High Arctic. Geol Soc Am Bull. 2011;124:3–23.

20. Schubert BA, Jahren AH, Eberle JJ, Sternberg LSL, Eberth DA. A summertime rainy season

in the Arctic forests of the Eocene. Geology. 2012;40:523–6.

21. Pross J, Contreras L, Bijl PK, Greenwood DR, Bohaty SM, Schouten S, et al. Persistent near-tropical warmth on the Antarctic continent during the early Eocene epoch. Nature. 2012;488:73–7.

22. Mudelsee M, Bickert T, Lear CH, Lohmann G. Cenozoic climate changes: A review based on time series analysis of marine benthic δ 18 O records. Rev Geophys. 2014;52:333–74.

23. Eldrett JS, Greenwood DR, Harding IC, Huber M. Increased seasonality through the Eocene to Oligocene transition in northern high latitudes. Nature. 2009;459:969–73.

24. Hren MT, Sheldon ND, Grimes ST, Collinson ME, Hooker JJ, Bugler M, et al. Terrestrial cooling in Northern Europe during the eocene-oligocene transition. Proc Natl Acad Sci U S A. 2013;110:7562–7.

25. Kerkhoff AJ, Moriarty PE, Weiser MD. The latitudinal species richness gradient in New World woody angiosperms is consistent with the tropical conservatism hypothesis. Proc Natl Acad Sci U S A. 2014;111:8125–30.

26. Sandve SR, Fjellheim S. Did gene family expansions during the Eocene-Oligocene boundary climate cooling play a role in Pooideae adaptation to cool climates? Mol Biol. 2010;19:2075–88.

27. Vigeland MD, Spannagl M, Asp T, Paina C, Rudi H, Rognli O-A, et al. Evidence for adaptive evolution of low-temperature stress response genes in a Pooideae grass ancestor. New Phytol. 2013;199:1060–8.

28. Marcussen T, Sandve SR, Heier L, Spannagl M, Pfeifer M, Jakobsen KS, et al. Ancient hybridizations among the ancestral genomes of bread wheat. Science. 2014;345:1250092.

29. Donoghue MJ. Colloquium paper: a phylogenetic perspective on the distribution of plant diversity. Proc Natl Acad Sci U S A. 2008; Suppl 1:11549–55.

30. Thomashow MF. Plant Cold Acclimation: Freezing Tolerance Genes and Regulatory Mechanisms. Annu Rev Plant Physiol Plant Mol Biol. 1999;50:571–99.

31. Janská A, Maršík P, Zelenková S, Ovesná J. Cold stress and acclimation - what is important for metabolic adjustment? Plant Biol. 2010;12:395–405.

32. Carvallo MA, Pino M-T, Jeknic Z, Zou C, Doherty CJ, Shiu S-H, et al. A comparison of the low temperature transcriptomes and CBF regulons of three plant species that differ in freezing tolerance: Solanum commersonii, Solanum tuberosum, and Arabidopsis thaliana. J Exp Bot. 2011;62:3807–19.

33. Zhang T, Zhao X, Wang W, Pan Y, Huang L, Liu X, et al. Comparative transcriptome profiling of chilling stress responsiveness in two contrasting rice genotypes. PLoS One. 2012;7:e43274.

34. Lindlöf A, Chawade A, Sikora P, Olsson O. Comparative Transcriptomics of Sijung and Jumli Marshi Rice during Early Chilling Stress Imply Multiple Protective Mechanisms. PLoS One. 2015;10:e0125385.

35. Yang Y-W, Chen H-C, Jen W-F, Liu L-Y, Chang M-C. Comparative Transcriptome Analysis of Shoots and Roots of TNG67 and TCN1 Rice Seedlings under Cold Stress and Following Subsequent Recovery: Insights into Metabolic Pathways, Phytohormones, and Transcription Factors. PLoS One. 2015;10:e0131391.

36. Abeynayake SW, Byrne S, Nagy I, Jonavičienė K, Etzerodt TP, Boelt B, et al. Changes in Lolium perenne transcriptome during cold acclimation in two genotypes adapted to different climatic conditions. BMC Plant Biol. 2015;15:250.

37. Christin P-A, Spriggs E, Osborne CP, Stromberg CAE, Salamin N, Edwards EJ. Molecular Dating, Evolutionary Rates, and the Age of the Grasses. Syst Biol. 2014;63:153–65.

38. Greenup AG, Sharyar S, Oliver SN, Walford SA, Millar AA, Trevaskis B. Transcriptome Analysis of the Vernalization Response in Barley (Hordeum vulgare) Seedlings. PLoS One. 2011;6:e17900.

39. Christin P-A, Boxall SF, Gregory R, Edwards EJ, Hartwell J, Osborne CP. Parallel Recruitment of Multiple Genes into C4 Photosynthesis. Genome Biol Evol. 2013;5:2174–87.

40. Christin P-A, Arakaki M, Osborne CP, Edwards EJ. Genetic Enablers Underlying the Clustered Evolutionary Origins of C4 Photosynthesis in Angiosperms. Mol Biol Evol. 2015;32:846–58.

41. Yang Q-S, Gao J, He W-D, Dou T-X, Ding L-J, Wu J-H, et al. Comparative transcriptomics analysis reveals difference of key gene expression between banana and plantain in response to cold stress. BMC Genomics. 2015;16:446.

42. Negi S, Tak H, Ganapathi TR. Expression analysis of MusaNAC68 transcription factor and its functional analysis by overexpression in transgenic banana plants. Plant Cell, Tissue Organ Cult. 2015;125:59–70.

43. Nakashima K, Takasaki H, Mizoi J, Shinozaki K, Yamaguchi-Shinozaki K. NAC transcription factors in plant abiotic stress responses. Biochim Biophys Acta. 2012;1819:97–103.

44. Mao X, Chen S, Li A, Zhai C, Jing R. Novel NAC transcription factor TaNAC67 confers enhanced multi-abiotic stress tolerances in Arabidopsis. PLoS One. 2014;9:e84359.

45. Kellogg EA. Evolutionary History of the Grasses. Plant Physiol. 2001;125:1198–205.

46. Schardl CL, Craven KD, Speakman S, Stromberg A, Lindstrom A, Yoshida R. A novel test for host-symbiont codivergence indicates ancient origin of fungal endophytes in grasses. Syst Biol. 2008;57:483–98.

47. Bieniawska Z, Espinoza C, Schlereth A, Sulpice R, Hincha DK, Hannah MA. Disruption of the Arabidopsis circadian clock is responsible for extensive variation in the cold-responsive transcriptome. Plant Physiol. 2008;147:263–79.

48. Nakamichi N, Kiba T, Henriques R, Mizuno T, Chua NH, Sakakibara H. PSEUDO-RESPONSE REGULATORS 9, 7, and 5 Are Transcriptional Repressors in the Arabidopsis Circadian Clock. Plant Cell. 2010;22:594–605.

49. Johansson M, Ramos-Sánchez JM, Conde D, Ibáñez C, Takata N, Allona I, et al. Role of the Circadian Clock in Cold Acclimation and Winter Dormancy in Perennial Plants. In: Advances in Plant Dormancy. Cham: Springer International Publishing; 2015. p. 51–74.

50. Romero IG, Ruvinsky I, Gilad Y. Comparative studies of gene expression and the evolution of gene regulation. Nat Rev Genet. 2012;13:505–16.

51. Wittkopp PJ, Kalay G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat Rev Genet. 2012;13:59–69.

52. Tutin TG. Flora europaea. Cambridge University Press; 1980.

53. Hultén E, Fries M. Atlas of North European vascular plants (North of the Tropic of Cancer), Vols. I-III. Königstein, Federal Republic of Germany: Koeltz scientific books; 1986.

54. Clayton, W.D., Vorontsova, M.S., Harman, K.T. and Williamson H. GrassBase - The Online World Grass Flora. 2006.

55. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;:btu170.

56. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644.

57. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8:1494–512.

58. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13:2178–89.

59. Yang Y, Smith SA. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. Mol Biol Evol. 2014;31:3081–92.

60. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013;30:772–80.

61. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 2006;34 suppl 2:W609–12.

62. Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2011;27:592–3.

63. Drummond AJ, Rambaut A, Huelsenbeck J, Ronquist F, Beaumont M, Drummond A, et al. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 2007;7:214.

64. Langmead B, Trapnell C, Pop M, Salzberg SL, Down T, Rakyan V, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25.

65. Li B, Dewey CN, Wang Z, Gerstein M, Snyder M, Katz Y, et al. RSEM: accurate transcript

quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.

66. Smith SA, Wilson NG, Goetz FE, Feehery C, Andrade SCS, Rouse GW, et al. Resolving the evolutionary relationships of molluscs with phylogenomic tools. Nature. 2011;480:364–7.

67. Love MI, Huber W, Anders S, Lönnstedt I, Speed T, Robinson M, et al. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014 1512. 2014;15:31–46.

68. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4:406–25.

69. Svensson JT, Crosatti C, Campoli C, Bassi R, Stanca AM, Close TJ, et al. Transcriptome Analysis of Cold Acclimation in Barley Albina and Xantha Mutants. PLANT Physiol. 2006;141:257–70.

70. Alexa A, Rahnenfuhrer J. topGO: Enrichment analysis for Gene Ontology. R package version 2.18.0. October. 2010.

71. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol. 2007;24:1586–91.

72. Yuan F, Yang H, Xue Y, Kong D, Ye R, Li C, et al. OSCA1 mediates osmotic-stress-evoked Ca2+ increases vital for osmosensing in Arabidopsis. Nature. 2014;514:367–71.

73. Hou C, Tian W, Kleist T, He K, Garcia V, Bai F, et al. DUF221 proteins are a family of osmosensitive calcium-permeable cation channels conserved across eukaryotes. Cell Res. 2014;24:632–5.

74. Li Y, Yuan F, Wen Z, Li Y, Wang F, Zhu T, et al. Genome-wide survey and expression analysis of the OSCA gene family in rice. BMC Plant Biol. 2015;15:261.

75. Day IS, Reddy VS, Shad Ali G, Reddy ASN. Analysis of EF-hand-containing proteins in Arabidopsis. Genome Biol. 2002;3:RESEARCH0056.

76. Bose J, Pottosin II, Shabala SS, Palmgren MG, Shabala S. Calcium Efflux Systems in Stress Signaling and Adaptation in Plants. Front Plant Sci. 2011;2:85.

77. Andrés F, Coupland G. The genetic basis of flowering responses to seasonal cues. Nat Rev Genet. 2012;13:627–39.

78. Cao S, Ye M, Jiang S. Involvement of GIGANTEA gene in the regulation of the cold stress response in Arabidopsis. Plant Cell Rep. 2005;24:683–90.

79. Xie Q, Lou P, Hermand V, Aman R, Park HJ, Yun D-J, et al. Allelic polymorphism of GIGANTEA is responsible for naturally occurring variation in circadian period in Brassica rapa. Proc Natl Acad Sci U S A. 2015;112:3829–34.

80. Sobkowiak A, Jończyk M, Jarochowska E, Biecek P, Trzcinska-Danielewicz J, Leipner J, et al. Genome-wide transcriptomic analysis of response to low temperature reveals candidate genes

determining divergent cold-sensitivity of maize inbred lines. Plant Mol Biol. 2014;85:317–31.

81. Murakami-Kojima M. The APRR3 Component of the Clock-Associated APRR1/TOC1 Quintet is Phosphorylated by a Novel Protein Kinase Belonging to the WNK Family, the Gene for which is also Transcribed Rhythmically in Arabidopsis thaliana. Plant Cell Physiol. 2002;43:675–83.

82. Murakami M. Characterization of Circadian-Associated APRR3 Pseudo-Response Regulator Belonging to the APRR1/TOC1 Quintet in Arabidopsis thaliana. Plant Cell Physiol. 2004;45:645–50.

83. Wang Z-Y, Xiong L, Li W, Zhu J-K, Zhu J. The plant cuticle is required for osmotic stress regulation of abscisic acid biosynthesis and osmotic stress tolerance in Arabidopsis. Plant Cell. 2011;23:1971–84.

84. Liu D, Wang L, Zhai H, Song X, He S, Liu Q. A novel α/β-hydrolase gene IbMas enhances salt tolerance in transgenic sweetpotato. PLoS One. 2014;9:e115128.

85. Guan W, Ferry N, Edwards MG, Bell HA, Othman H, Gatehouse JA, et al. Proteomic analysis shows that stress response proteins are significantly up-regulated in resistant diploid wheat (Triticum monococcum) in response to attack by the grain aphid (Sitobion avenae). Mol Breed. 2015;35:57.

86. Kimura M, Yamamoto YY, Seki M, Sakurai T, Sato M, Abe T, et al. Identification of Arabidopsis Genes Regulated by High Light-Stress Using cDNA Microarray¶. Photochem Photobiol. 2003;77:226–33.

87. Singh S, Cornilescu CC, Tyler RC, Cornilescu G, Tonelli M, Lee MS, et al. Solution structure of a late embryogenesis abundant protein (LEA14) from Arabidopsis thaliana, a cellular stress-related protein. Protein Sci. 2005;14:2601–9.

88. Rinne P, Welling A, Kaikuranta P. Onset of freezing tolerance in birch (Betula pubescens Ehrh.) involves LEA proteins and osmoregulation and is impaired in an ABA-deficient genotype. Plant, Cell Environ. 1998;21:601–11.

89. Gagné-Bourque F, Mayer BF, Charron J-B, Vali H, Bertrand A, Jabaji S. Accelerated Growth Rate and Increased Drought Stress Resilience of the Model Grass Brachypodium distachyon Colonized by Bacillus subtilis B26. PLoS One. 2015;10:e0130456.

90. Murakami M. The Evolutionarily Conserved OsPRR Quintet: Rice Pseudo-Response Regulators Implicated in Circadian Rhythm. Plant Cell Physiol. 2003;44:1229–36.

91. Campoli C, Shtaya M, Davis SJ, von Korff M. Expression conservation within the circadian clock of a monocot: natural variation at barley Ppd-H1 affects circadian expression of flowering time genes, but not clock orthologs. BMC Plant Biol. 2012;12:97.

92. Lee B, Henderson DA, Zhu J-K. The Arabidopsis cold-responsive transcriptome and its regulation by ICE1. Plant Cell. 2005;17:3155–75.

93. Wang G, Cai G, Kong F, Deng Y, Ma N, Meng Q. Overexpression of tomato chloroplast-

targeted DnaJ protein enhances tolerance to drought stress and resistance to Pseudomonas solanacearum in transgenic tobacco. Plant Physiol Biochem. 2014;82:95–104.

94. Kong F, Deng Y, Zhou B, Wang G, Wang Y, Meng Q. A chloroplast-targeted DnaJ protein contributes to maintenance of photosystem II under chilling stress. J Exp Bot. 2014;65:143–58.

95. Yin Y, Mohnen D, Gelineo-Albersheim I, Xu Y, Hahn MG. Glycosyltransferases of the GT8 Family. In: Annual Plant Reviews Volume 41: Plant Polysaccharides, Biosynthesis and Bioengineering. WILEY-BLACKWELL; 2011. p. 167–211.

96. Qi Y, Kawano N, Yamauchi Y, Ling J, Li D, Tanaka K. Identification and cloning of a submergence-induced gene OsGGT (glycogenin glucosyltransferase) from rice (Oryza sativa L.) by suppression subtractive hybridization. Planta. 2005;221:437–45.

97. Uddin MI, Kihara M, Yin L, Perveen MF, Tanaka K. Expression and subcellular localization of antiporter regulating protein OsARP in rice induced by submergence, salt and drought stresses. African Journal of Biotechnology. 2012;11:12849–55.

98. Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. Very high resolution interpolated climate surfaces for global land areas. Int J Climatol. 2005;25:1965–78.

Supporting Information

Table S1. **Summary statistics for the sampling, the transcriptome assembly, coding sequence detection and ortholog group inference**. Isoforms were not included in the counts. Only ortholog groups with at least one coding transcript from the five studied species were included.

Table S2. High confidence ortholog groups (HCOGs). HCOGs generated by filtering and splitting ortholog groups (see Methods). Ortholog groups were stored as a table with the ortholog group IDs as rows and species as columns. Each cell contains sequence IDs separated by ",". Groups that were the result of splitting larger ortholog groups were marked by a number suffix in the group ID.

Table S3. A cross species expression matrix. Combined read counts for high confidence ortholog groups. Column represents samples and rows represents HCOGs. The sample IDs in the column header consists of the species ID, the time point, indication of whether the sample is pooled from five individual plants ("mix") or just a single individual plant ("ind") and the replicate number.

Table S4. Differential expression results for the high confidence ortholog groups (HCOGs). A table with rows representing HCOGs and columns representing differential expression results including log2 fold changes, P-values and FDR adjusted p-values for short- and long-term responses.

Figure S1: The four most common gene tree topologies. We generated gene trees from a selected set of 3914 ortholog groups (see Methods). This figure depicts the four most common topologies.

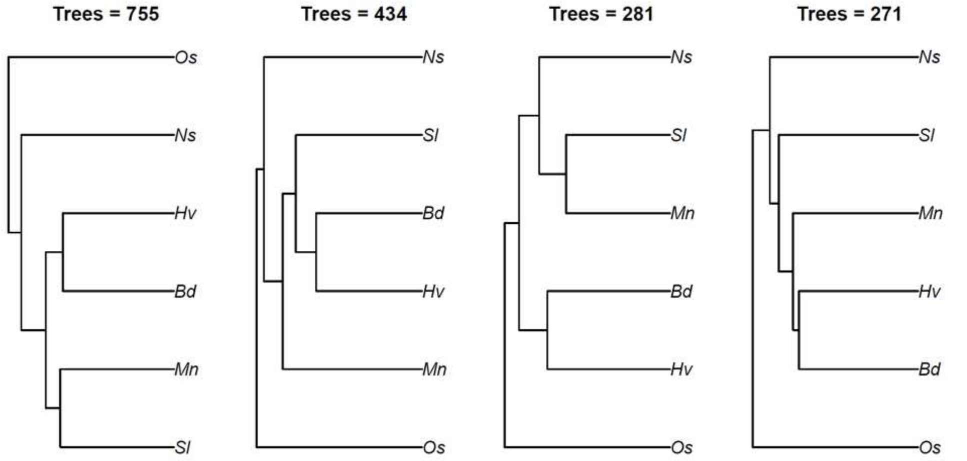


Figure S1: The four most common gene tree topologies. We generated gene trees from a selected set of 3914 ortholog groups (see Methods). This figure depicts the four most common topologies.

Paper 2

Evolution of cold acclimation in temperate grasses (Pooideae)

Marian Schubert[*,1], Lars Grønvold[*,2], Simen R. Sandve[3], Torgeir R. Hvidsten[2,4] and Siri Fjellheim[1,†]

[1]Faculty of Biosciences, Norwegian University of Life Sciences, Ås NO-1432, Norway.

[2]Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Ås NO-1432, Norway.

[3]Centre for Integrative Genetics (CIGENE), Faculty of Biosciences, Norwegian University of Life Sciences, Ås NO-1432, Norway.

[4]Umeå Plant Science Centre, Department of Plant Physiology, Umeå University, Umeå SE-90187, Sweden.

**Contributed equally*

† Author for correspondence: siri.fjellheim@nmbu.no

Abstract

In the past 50 million years climate cooling has triggered the expansion of temperate biomes. During this period, many extant plant lineages in temperate biomes evolved from tropical ancestors and adapted to seasonality and cool conditions. Among the Poaceae (grass family), one of the subfamilies that successfully shifted from tropical to temperate biomes is the Pooideae (temperate grasses). Subfamily Pooideae contains the most important crops cultivated in the temperate regions including wheat (*Triticum aestivum*) and barley (*Hordeum vulgare*). Due to the need of well-adapted cultivars, extensive research has produced a large body of knowledge about the mechanisms underlying cold adaptation in cultivated Pooideae species. Especially cold acclimation, a process which increases the frost tolerance during a period of non-freezing cold, plays an important role. Because cold adaptation is largely unexplored in lineages that diverged early in the evolution of the Pooideae, little is known about the evolutionary history of cold acclimation in the Pooideae. Here we test if several species of early diverging lineages exhibit increased frost tolerance after a period of cold acclimation. We further investigate the conservation of five well-studied gene families that are known to be involved in the cold acclimation of Pooideae crop species. Our results indicate that cold acclimation exists in early diverging lineages, but that genes involved in regulation of cold acclimation are not conserved. The investigated gene families show signs of lineage-specific evolution and support the hypothesis that gene family expansion is an important mechanism in adaptive evolution.

Introduction

Temperate biomes expanded during global cooling throughout the Eocene and the Eocene to Oligocene (E-O) transition, *ca.* 34 million years ago (Mya) (Potts and Behrensmeyer 1992, Donoghue 2008, Stickley et al. 2009, Mudelsee et al. 2014). In plants, biome shifts are rare and restricted to a small number of lineages (Crisp et al. 2009) and only a few, ancestrally tropical angiosperm lineages colonized the expanding temperate biomes (Judd et al. 1994, Wiens and Donoghue 2004, Kerkhoff et al. 2014). The successful colonizers faced several stresses and adapted to frost, increased temperature seasonality, and short growing seasons (Zachos et al. 2001, Eldrett et al. 2009, Mudelsee et al. 2014).

The grass (Poaceae) flora of temperate biomes is dominated by members of the subfamily Pooideae which account for up to 90% of the grass species in Northern biomes (Hartley 1973; Clayton 1981). Since their emergence in the late Paleocene or early Eocene (Bouchenak-Khelladi et al. 2010, Christin et al. 2014, Spriggs et al. 2014), the grass subfamily Pooideae successfully shifted their distribution from the tropical/subtropical biomes of their ancestors (Edwards and Smith 2010; Strömberg 2011) to temperate biomes. Except for a few hundred species in early diverging Pooideae tribes most of the *ca.* 4200 Pooideae species belong to the species-rich ‘core Pooideae’ clade (*sensu* Davis and Soreng 1993; Soreng and Davis 1998) and its sister tribe Brachypodieae, which contains the model grass *Brachypodium distachyon* (Soreng et al. 2015). Because core Pooideae contain economically important species like wheat (*Triticum aestivum*) and barley (*Hordeum vulgare*) as well as several forage grasses like ryegrass (*Lolium perenne*), the need for well-adapted cultivars has provided extensive knowledge about molecular mechanisms underlying adaptation to temperate environments (Thomashow 1999; Sandve et al. 2008; Galiba et al. 2009; Sandve et al. 2011, Preston and Sandve 2013; Fjellheim et al. 2014). Key molecular responses to low temperatures in Pooideae species may represent adaptations to temperate climate (McKeown et al. 2016, Woods et al. 2016, McKeown et al. 2017), whose expansion coincided with the early evolution of the Pooideae.

Plants in temperate biomes respond to frost and periods of prolonged cold with a suite of physiological and biochemical adaptations controlled by complex regulatory programs. The main challenge of exposure to frost is maintaining the integrity of the cellular membranes to avoid dehydration (Pearce et al. 2001). This is accomplished by adjusting the lipid composition of membranes to increase membrane stability (Uemura et al. 1995; Danyluk et al. 1998;) and to accumulate sugars and anti-freeze proteins (Griffith and Yaish 2004; Sandve et al. 2011). Additionally, accumulation of reactive oxygen species (ROS) damages lipid membranes and increases protein degradation (Murata et al. 2007; Crosatti et al. 2013) to which plants react by synthesizing proteins that decrease ROS-mediated stress (Crosatti et al. 2013). Through the process of cold acclimation – a period of cold, non-freezing temperatures – temperate plants can increase their frost tolerance to prepare for freezing during winter (Thomashow 1999; Thomashow 2010).

Five of the best studied cold acclimation gene families code for C-repeat binding factors (CBF), dehydrins (DHN), chloroplast-targeted cold-regulated proteins (ctCOR), ice recrystallization inhibition proteins (IRIP) and fructosyl transferases (FST). All of these are known to be induced by cold and play important roles during cold-stress response and cold acclimation in core Pooideae (*CBF*: Badawi et al. 2007; Li et al. 2012. *DHN*: Olave-Concha et al. 2004; Rorat 2006; Kosová et al. 2007, 2014. *ctCOR*: Gray et al. 1997; Crosatti et al. 1999, 2013; Tsvetanov et al. 2000. *IRIP*: Antikainen and Griffith 1997; Hisano et al. 2004; Kumble et al. 2008; Sandve et al. 2008; John et al. 2009; Zhang et al. 2010; Sandve et al. 2011. *FST*: Hisano et al. 2004; Tamura et al. 2014). The CBFs are transcription factors and function as “master-switches” of cold regulation and cold acclimation (Sarhan et al. 1998; Thomashow 1999) and are involved in various kinds of stress response (Agarwal et al. 2006; Akhtar et al. 2012). Two groups of *CBF* genes – *CBFIII* and *CBFIV* – are especially important for cold acclimation in Pooideae and are restricted to this subfamily (Badawi et al. 2007, Li et al. 2012). The *DHN* gene family encodes hydrophilic proteins that share a lysine-rich sequence, the “K-segment”, which interacts with membranes (Koag et al. 2003; Koag et al. 2009) and protects against dehydration-related cellular stress and possibly also acts as cryoprotectant (Close 1997; Danyluk et al. 1998; Houde et al. 2004). ctCOR proteins (i.e. COR14 and WCS19; Gray et al. 1997; Crosatti et al. 1999; Tsvetanov et al. 2000) are thought to alleviate damage by ROS that accumulate during cold-induced overexcitation of the photosystems (Crosatti et al. 2013). IRIPs bind to the edges of microscopic intracellular ice grains and restrict the formation of large hazardous ice crystals (Griffith and Ewart 1995; Antikainen and Griffith 1997; Sidebottom et al. 2000; Griffith and Yaish 2004; Tremblay et al. 2005; Sandve et al. 2011). Lastly, frost tolerance correlates with the accumulation and degree of polymerization of fructans which are synthesized by FSTs (Hisano et al. 2004; Tamura et al. 2014). Fructans are the major carbohydrate storage in model Pooideae species (Chalmers et al. 2005), but they have also been shown to improve membrane stability during freezing stress (Hincha et al. 2000). Interestingly, Li et al. (2012) identified cold responsive *IRIP* and *CBFIII* homologs in the model grass *B. distachyon*, which is sister to the core Pooideae lineage, while homologous *CBFIV* and *FST* genes were absent.

Because our knowledge about cold adaptations in Pooideae mostly stems from studies restricted to a handful species in the core Pooideae-Brachypodieae clade, it is unknown when and how cold acclimation and responsible genes evolved. A prerequisite for a successful biome shift lies in a lineage’s adaptive potential, which is controlled by the genetic toolkit of its ancestor (Edwards and Donoghue 2013; Christin et al. 2015). Therefore, functional knowledge about the five gene families provides an excellent basis to investigate their importance for the evolution of cold acclimation in the Pooideae. If those genes have been part of the cold response in the most recent common ancestor (MRCA) of the Pooideae, we expect to find a high degree of conservation and same expression patterns among Pooideae species. Alternatively, if cold adaptation evolved independently in the main Pooideae lineages, we expect a lower degree of conservation and a more diverse expression pattern.

Here we test the hypotheses that cold acclimation is conserved in Pooideae and that key cold acclimation genes known from core Pooideae are cold responsive in early diverging lineages. To test the conservation of cold acclimation we performed a classic growth experiment and investigated the effect of cold acclimation across the three core Pooideae species, three Brachypodieae species and three early diverging Pooideae species *Nardus stricta*, *Melica nutans* and *Stipa lagascae*. To test conservation of cold acclimation genes we compared the expression of *CBF*, *DHN*, *ctCOR*, *IRIP* and *FST* homologs between *H. vulgare*, *B. distachyon* and the three early diverging Pooideae species. All species increase their frost tolerance following cold acclimation, but only two out of the ten studied genes exhibited completely conserved expression pattern across the investigated species. Nevertheless, we suggest that gene families *DHN*, *ctCOR*, *CBFIIId* and *CBFIV* were instrumental in the Pooideae's shift to temperate biomes. The Pooideae MRCA might not have been cold tolerant, but those gene families were likely part of its genetic toolkit that steered the evolutionary trajectory of its descendants towards cold tolerance.

Material and methods

Plant material and freezing tests

Seeds from nine Pooideae species were acquired from germplasm collections or collected in nature. From the core Pooideae we included *Hordeum vulgare* L. (cultivar 'Sonja', provided by Professor Åsmund Bjørnstad, Department of Plant Sciences, Norwegian University of Life Sciences, Ås NO-1432, Norway), *Lolium perenne* L. (cultivar 'Fagerlin', provided by Dr. Kovi, Department of Plant Sciences, Norwegian University of Life Sciences, Ås NO-1432, Norway) and *Elymus repens* (L.) Gould (collected in October 2015, Norway [59.66111, 10.89194]). From the tribe Brachypodieae we included *Brachypodium distachyon* (L.) P. Beauv. (accession 'Bd1-1' [W6 46201] acquired from U.S. National Plant Germplasm System [U.S.-NPGS] via Germplasm Resources Information Network [GRIN]), *B. pinnatum* (collected in October 2015, Norway [59.71861, 10.59333]) and *B. sylvaticum* (collected in October 2015, Norway [59.68697, 10.61012]). From early diverging Pooideae lineages we included *Melica nutans* L. (collected in June 2012, Germany [50.70708, 11.23838]), S*tipa lagascae* Roem & Schult. (accession PI 250751, acquired from U.S.-NPGS via GRIN), and *Nardus stricta* (collected in July 2012, Romania [46.69098, 22.58302]).

The seeds were germinated and plants grown in the green house at 20°C under natural day light. Each individual was divided into four clones, one for each treatment and control. The plants were acclimated at 4°C and short (8h) days for three weeks. Control conditions were short days and 20°C. The light intensity was 50 µmol m^{-2} s^{-1}. At the end of the cold acclimation period, plants were subjected to freezing at three different temperatures (-4, -8 and -12°C) following Alm et al. (2011). For each temperature we used 15 acclimated and 15 non-acclimated individuals per species. Additional 15 individuals per species were kept at control conditions. After freezing, plants were cut down to approximately 3 cm and grown at 20°C under long days in a greenhouse with natural light conditions. Two and three weeks after the plants were moved into 20°C and long days they were assessed for regeneration ability and scored from 0 (dead) to 9 (growth without damage). Differences between acclimated and non-acclimated individuals within each species were tested with a one-tailed *t* test in R (R Core Team 2016) using the 'stats' package.

Transcriptomic data and differential gene expression

In another study (Grønvold et al., 2017) we compared the molecular cold response from the five Pooideae species *Nardus stricta, Stipa lagascae, Melica nutans, Brachypodium distachyon* (same populations/accessions as described above) and *Hordeum vulgare* (cultivar 'Igri', provided by Prof. Åsmund Bjørnstad). The transcriptomic data produced in Grønvold et al. 2017 was used as basis for the investigation of cold acclimation genes.

In short, seeds from the five species were germinated and initially grown in greenhouse at a neutral day length (12 hours of light), 17°C and a minimum light intensity of 150 µmol m^{-2} s^{-1}. After an initial growth period plants were randomly distributed to two growth chambers and subjected to cold treatment (6°C) under short days (8 hours of light) and a light intensity of 50 µmol m^{-2} s^{-1}. Leaf material for RNA isolation was collected after eight hours, four weeks and nine weeks of cold treatment. The sampling points were chosen to separate between short-term and long-term cold responsive expression patterns. For each time point, flash-frozen leaves from five individuals were individually homogenized using a TissueLyser (Qiagen Retsch, Haan, Germany) and total RNA was isolated (from each individual) using RNeasy Plant Mini Kit (Qiagen Inc., Germany), following the manufacturer's instructions. Purity and integrity of total RNA extracts was determined using a NanoDrop 8000 UV-Vis Spectrophotometer (Thermo Scientific, Wilmington, DE, USA) and 2100 Bioanalyzer (Agilent, Santa Clara, CA, USA), respectively. Pooled RNA extracts were delivered to the Norwegian Sequencing Centre (NSC), Centre for Evolutionary and Ecological Synthesis (CEES), Department of Biology, University of Oslo, Norway, where strand-specific cDNA libraries were prepared and paired end sequenced on a HiSeq 2000-system (Illumina, San Diego, CA, USA). *De novo* transcriptomes were assembled using Trinity v2.0.6 (Grabherr et al. 2011). Ortholog groups were inferred with orthoMCL (Li et al. 2003) using protein sequences translated from the assembled transcriptomes in addition to reference genomes from *H. vulgare*, *Lolium perenne* and *B. distachyon* as well as *Oryza sativa*, *Sorghum bicolor* and *Zea mays*. For each *de novo* transcriptome differential gene expression was estimated using DESeq2 (Love et al. 2014) for transcripts belonging to an ortholog group. We used an in-house computational pipeline to identify high confidence ortholog groups based on phylogenetic analysis of the orthoMCL ortholog groups.

Identification of cold acclimation candidate genes

For five cold acclimation gene families (*C-repeat binding factor* genes (*CBFIII* and *CBFIV*), *dehydrin* genes (*DHN*), *chloroplast targeting cold-regulated* genes (*ctCOR*), *ice-recrystallization inhibition protein* genes (*IRIP*) and *fructosyltransferase* genes (*FST*)) we extracted previously identified *H. vulgare* genes (Table S1) and downloaded translated amino acid (AA) from GenBank. These reference sequences were used in protein BLAST searches against translated *de novo* transcript sequences produced by Grønvold et al. 2017. Potential cold acclimation gene transcripts were identified by discarding protein BLAST results with a maximum bitscore < 90, and e-value > 1E-021. When at least one transcript met those criteria, all transcripts of the respective high confidence ortholog group were defined as candidate transcripts. The high confidence ortholog groups also contained genomic coding sequences (CDS) from the Pooideae species *H. vulgare, L. perenne* and *B. distachyon* as well as from the outgroup species

O. sativa, *S. bicolor* and *Z. mays*. *De novo* transcripts not part of a high confidence ortholog group were included when they met a maximum bitscore > 110 and an e-value < 1E-31. Since seven of the *H. vulgare* transcripts were found that did not belong to any ortholog groups, differential expression had to be re-run for *H. vulgare* after including these transcripts. MUSCLE v.3.8.31 (Edgar 2004) was used to create multiple sequence alignments by aligning all nucleotide candidate transcripts with *H. vulgare* reference coding sequences (CDS) and best CDS nucleotide BLAST hits for *Triticum sp.* acquired from GenBank. To reduce redundancy, genomic *H. vulgare* reference sequences of the ortholog groups were removed from the alignments when they were identical to *H. vulgare* sequences from GenBank.

To ensure proper tree root resolution, further outgroups were included where necessary. Alignments were manually trimmed and optimized using AliView v1.7.1 sequence editor (Larsson 2014). Transcripts were excluded from subsequent analyses when i) cross contamination was identified, i.e. very low total expression and sequence nearly identical to highly expressed contamination source or ii) the transcript was too fragmented/truncated to contain sufficient phylogenetically informative characters.

Gene tree reconstruction and calibration

For each multiple sequence alignment, the best model of nucleotide substitution (either HKY + Γ, or GTR + Γ) was chosen based on estimations of jModelTest v2.1.7 (Darriba et al. 2012). Using BEAST v1.8.2 (Drummond et al. 2012), we estimated gene trees under an uncorrelated lognormal relaxed clock model with the prior mean substitution rate uniformly distributed between 0 and 0.06 and a Yule tree prior (birth only) for 100 Million MCMC generations while model parameters were logged every 100000 generations. Tracer v1.6 (http://tree.bio.ed.ac.uk/software/tracer/) was used to assure that all parameters had an effective sample size (ESS) above 200. Ten percent of all trees were discarded (burn-in). The remaining trees were concatenated to the maximum clade credibility tree by TreeAnnotator v2.3.0 (Drummond and Rambaut 2012) and visually adjusted using FigTree v1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/). Posterior probabilities equal to or greater than 0.8 were considered to be a significant node support. During BEAST analyses, two normally distributed node age priors were used to calibrate gene trees. Prior for *H. vulgare* and *B. distachyon* divergence was set to 44.4 My (3.53 standard deviation) according to estimates from Marcussen et al. (2014) and priors for *O. sativa* and *B. distachyon* divergence was set to 53 Mya (3.6 standard deviation) according to estimates from Christin et al. (2014). To ensure correct rooting of the gene trees, we constrained Pooideae transcripts to form a monophyletic clade with their closest *O. sativa* sequences.

Results

Freezing tests

Freezing tests revealed that cold acclimation, i.e. increased frost tolerance through exposure to cold non-freezing periods, exists in species of early diverging lineages. All acclimated plants from those lineages exhibited a higher survival rate at -4 and -8°C compared to non-acclimated plants (Fig. 1), although the *t* test was not significant at -4°C for *S. lagascae* and *N. stricta*. Interestingly, non-acclimated plants of early diverging Pooideae species performed better at -4°C than non-acclimated *Brachypodium* species and *H. vulgare*, comparable with the survival rates of non-acclimated, perennial core Pooideae species *L. perenne* and *Elymus repens*.

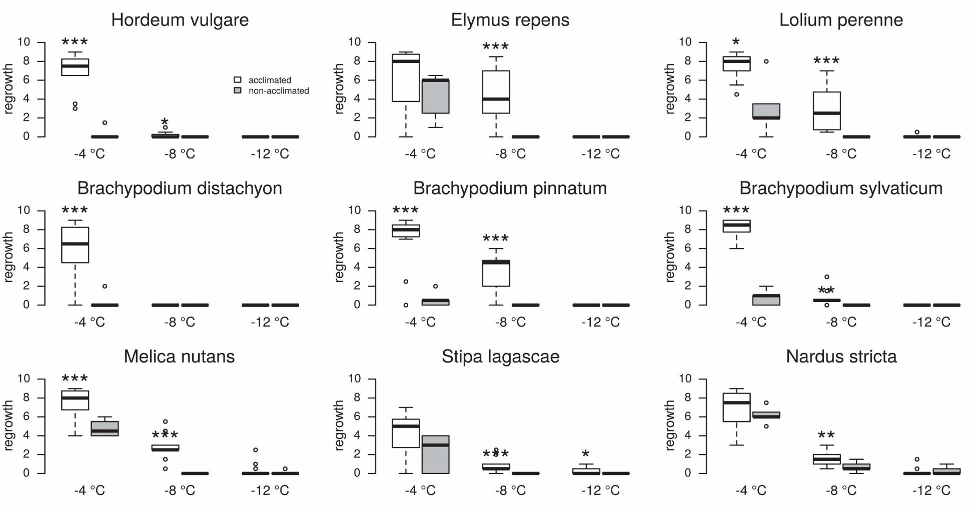


Figure 1: Frost tolerance after cold acclimation. *Regrowth of nine acclimated and non-acclimated Pooideae species after exposure to three freezing temperatures (-4°C, -8°C and 12°C). Significant differences in regrowth between acclimated and non-acclimated plants are indicated by asterisks (*** p ≤ 0.001, ** p ≤ 0.01, * p ≤ 0.05).*

CBFIIIc/d and *CBFIV* gene family

Consistently with previous results (Badawi et al. 2007), we reconstructed two monophyletic, Pooideae-specific *CBFIII* clades (Fig. 2). In the *CBFIIIc* clade, only *de novo* transcripts from *H. vulgare* were represented but none of them were differentially expressed. The *CBFIIId* clade contained numerous *de novo* transcripts from all studied species, but did not clearly reflect the Pooideae's species phylogeny. All but three *de novo* transcripts were differentially expressed and either induced in short-term cold (*N. stricta* and H. *vulgare*) or in short- and long-term cold (*B. distachyon* and *M. nutans*).

The CBFIV gene tree (Fig. 3) reconstructed the four known *CBFIV* clades (Badawi et al. 2007) but no homologous transcripts from *N. stricta* or *B. distachyon* were identified. Two short-term cold induced transcripts from *S. lagascae* and *M. nutans* formed a monophyletic clade together with the core Pooideae *CBFIV* homologs. In accordance with the fact that there are no known *CBFIVb* homologs in *H. vulgare*, we only identified cold induced *H. vulgare* transcripts in the three remaining *CBFIV* clades.

Interestingly the two *de novo* transcripts from *M. nutans* and *S. lagascae* were induced by short-term cold.

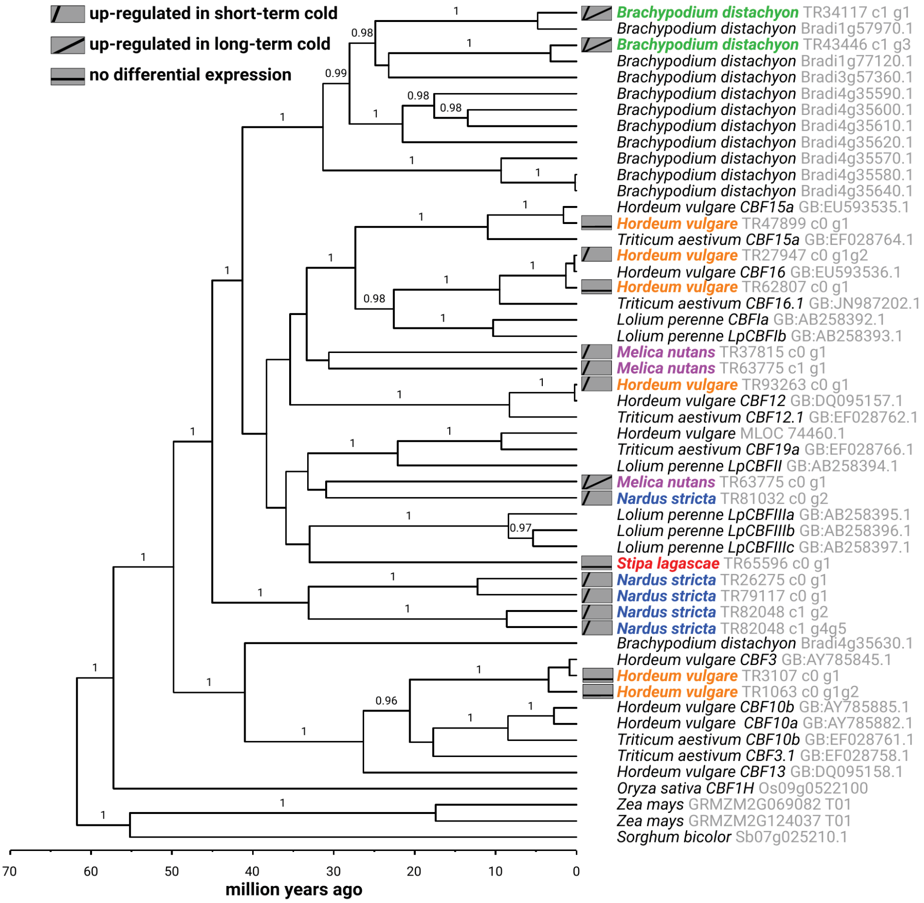

up-regulated in short-term cold
up-regulated in long-term cold
no differential expression
Brachypodium distachyon TR34117 c1 g1
Brachypodium distachyon Bradi1g57970.1
Brachypodium distachyon TR43446 c1 g3
Brachypodium distachyon Bradi1g77120.1
Brachypodium distachyon Bradi3g57360.1
Brachypodium distachyon Bradi4g35590.1
Brachypodium distachyon Bradi4g35600.1
Brachypodium distachyon Bradi4g35610.1
Brachypodium distachyon Bradi4g35620.1
Brachypodium distachyon Bradi4g35570.1
Brachypodium distachyon Bradi4g35580.1
Brachypodium distachyon Bradi4g35640.1
Hordeum vulgare CBF15a GB:EU593535.1
Hordeum vulgare TR47899 c0 g1
Triticum aestivum CBF15a GB:EF028764.1
Hordeum vulgare TR27947 c0 g1g2
Hordeum vulgare CBF16 GB:EU593536.1
Hordeum vulgare TR62807 c0 g1
Triticum aestivum CBF16.1 GB:JN987202.1
Lolium perenne CBFIa GB:AB258392.1
Lolium perenne LpCBFIb GB:AB258393.1
Melica nutans TR37815 c0 g1
Melica nutans TR63775 c1 g1
Hordeum vulgare TR93263 c0 g1
Hordeum vulgare CBF12 GB:DQ095157.1
Triticum aestivum CBF12.1 GB:EF028762.1
Hordeum vulgare MLOC 74460.1
Triticum aestivum CBF19a GB:EF028766.1
Lolium perenne LpCBFII GB:AB258394.1
Melica nutans TR63775 c0 g1
Nardus stricta TR81032 c0 g2
Lolium perenne LpCBFIIIa GB:AB258395.1
Lolium perenne LpCBFIIIb GB:AB258396.1
Lolium perenne LpCBFIIIc GB:AB258397.1
Stipa lagascae TR65596 c0 g1
Nardus stricta TR26275 c0 g1
Nardus stricta TR79117 c0 g1
Nardus stricta TR82048 c1 g2
Nardus stricta TR82048 c1 g4g5
Brachypodium distachyon Bradi4g35630.1
Hordeum vulgare CBF3 GB:AY785845.1
Hordeum vulgare TR3107 c0 g1
Hordeum vulgare TR1063 c0 g1g2
Hordeum vulgare CBF10b GB:AY785885.1
Hordeum vulgare CBF10a GB:AY785882.1
Triticum aestivum CBF10b GB:EF028761.1
Triticum aestivum CBF3.1 GB:EF028758.1
Hordeum vulgare CBF13 GB:DQ095158.1
Oryza sativa CBF1H Os09g0522100
Zea mays GRMZM2G069082 T01
Zea mays GRMZM2G124037 T01
Sorghum bicolor Sb07g025210.1
70
60
50
40
30
20
10
0
million years ago

Figure 2: Time calibrated phylogeny for the Pooideae-specific CBFIIIc/d gene family. *The phylogeny was estimated with BEAST v1.8.2 using a HKY+Γ and an uncorrelated, lognormal distributed molecular clock model. Significant posterior probabilities are shown as branch labels. CBFIIIc and CBFIIId form to distinct clades.*

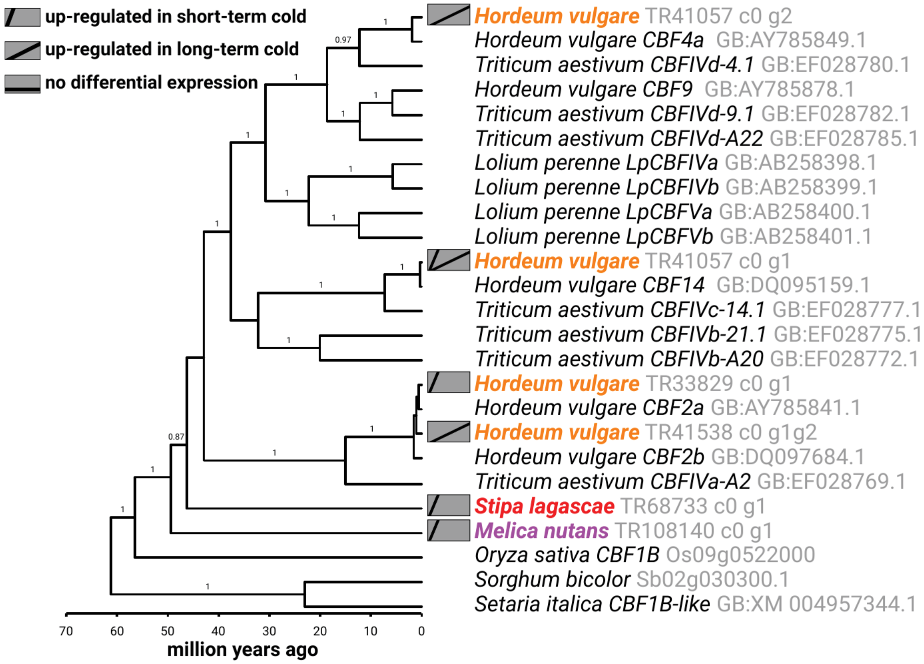

up-regulated in short-term cold
up-regulated in long-term cold
no differential expression
Hordeum vulgare TR41057 c0 g2
Hordeum vulgare CBF4a GB:AY785849.1
Triticum aestivum CBFIVd-4.1 GB:EF028780.1
Hordeum vulgare CBF9 GB:AY785878.1
Triticum aestivum CBFIVd-9.1 GB:EF028782.1
Triticum aestivum CBFIVd-A22 GB:EF028785.1
Lolium perenne LpCBFIVa GB:AB258398.1
Lolium perenne LpCBFIVb GB:AB258399.1
Lolium perenne LpCBFVa GB:AB258400.1
Lolium perenne LpCBFVb GB:AB258401.1
Hordeum vulgare TR41057 c0 g1
Hordeum vulgare CBF14 GB:DQ095159.1
Triticum aestivum CBFIVc-14.1 GB:EF028777.1
Triticum aestivum CBFIVb-21.1 GB:EF028775.1
Triticum aestivum CBFIVb-A20 GB:EF028772.1
Hordeum vulgare TR33829 c0 g1
Hordeum vulgare CBF2a GB:AY785841.1
Hordeum vulgare TR41538 c0 g1g2
Hordeum vulgare CBF2b GB:DQ097684.1
Triticum aestivum CBFIVa-A2 GB:EF028769.1
Stipa lagascae TR68733 c0 g1
Melica nutans TR108140 c0 g1
Oryza sativa CBF1B Os09g0522000
Sorghum bicolor Sb02g030300.1
Setaria italica CBF1B-like GB:XM 004957344.1
1
0.97
1
1
1
1
1
1
1
1
1
1
1
1
0.87
1
1
1
70
60
50
40
30
20
10
0
million years ago

Figure 3: Time calibrated phylogeny for the Pooideae CBFIV gene family. *The phylogeny was estimated with BEAST v1.8.2 using a HKY+Γ and an uncorrelated, lognormal distributed molecular clock model. Significant posterior probabilities are shown as branch labels.*

Dehydrin genes

Dehydrin (*DHN*) genes are well studied in *H. vulgare* and to date there are 13 known dehydrin homologs. Structurally they can be grouped into four distinct types based on the presence of amino acid segments (Y, K and S): SK$_3$-type (Hv*DHN8*), KS-type (Hv*DHN13*), K$_n$-type (Hv*DHN5*) and Y$_n$SK$_n$-type dehydrins (the 10 remaining Hv*DHN*-genes) (Kosová et al. 2007). Because these four groups represent phylogenetically distinct clades (Karami et al. 2013), we reconstructed individual gene trees for each group.

Cold induced accumulation of the gene *HvDHN5* and its wheat ortholog *WCS120* is regarded as marker for frost tolerant plants in *H. vulgare and T. aestivum* (Kosová et al., 2007). Beside a short- and long-term cold induced *DHN5* transcript in *H. vulgare*, we did not identify any *HvDHN5* homologs in the other investigated species. That is in accordance with the fact that K_n-type dehydrins have previously only been described in *Triticeae* species.

Short-term cold induced transcripts homologous to *HvDHN8* and *HvDHN13* were present in all investigated Pooideae species (Fig. S1 and S2). Altough the gene trees did not reflect the species phylogenies, the Pooideae homologs formed monophyletic clades. The literature contains no reports of *HvDHN8* and *HvDHN13* homologs in species of the *Poeae* tribe, e.g. *L. perenne,* and no homologous sequences have been deposited to GenBank (confirmed with blast searches). Contrary to this expectation, we identified one *L. perenne* homolog for each gene respectively.

The remaining barley dehydrin genes (*HvDHN1*, *HvDHN2*, *HvDHN3*, *HvDHN4*, *HvDHN6*, *HvDHN7*, *HvDHN9*, *HvDHN10*, *HvDHN11* and *HvDHN12*) belong to the Y_nSK_n-type. All homologous *de novo* transcripts belonged to four clades (*HvDHN1-2, HvDHN9-12, HvDHN3-4-7* and *DHN10),* which were specific to Pooideae (Fig. 4) and formed a clade with homologs from *O. sativa*, *Z. mays* and *S. bicolor*. Three of the Pooideae-specific clades contained transcripts from early diverging Pooideae species, with *DHN1-2* being specific to the core group and *B. distachyon*. While all transcripts of *N. stricta* and *M. nutans* were induced by short- and long-term cold treatment, *DHN* transcripts of *H. vulgare, B. distachyon* and *S. lagascae* were not differentially expressed.

For the last two Y_nSK_n-type clades – containing *HvDHN6* and *HvDHN11* – we did not identify any homologous, cold-induced transcripts, which was expected given that *HvDHN6* and *HvDHN11* are known to be expressed in embryos and caryopses only (Choi and Close 2000; Tommasini et al. 2008). These two clades were clearly less related to the rest of Y_nSK_n-type *DHN* homologs, because they were nested outside the monophyletic, Poaceae-specific Y_nSK_n-type clade (data not shown).

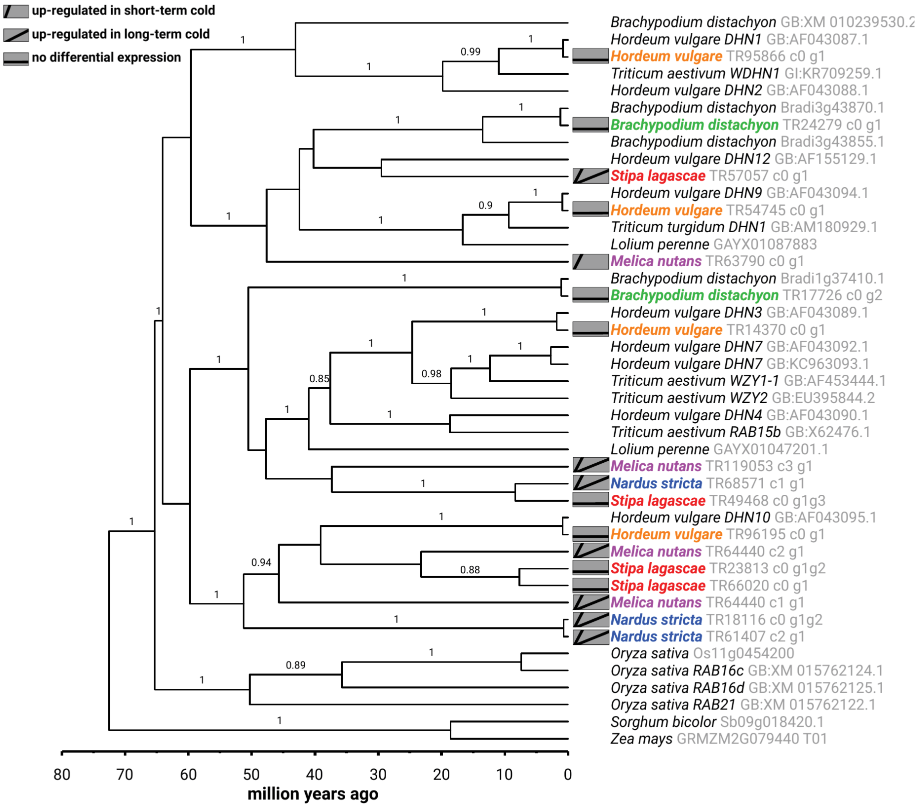

up-regulated in short-term cold
up-regulated in long-term cold
no differential expression
Brachypodium distachyon GB:XM 010239530.2
Hordeum vulgare DHN1 GB:AF043087.1
Hordeum vulgare TR95866 c0 g1
Triticum aestivum WDHN1 GI:KR709259.1
Hordeum vulgare DHN2 GB:AF043088.1
Brachypodium distachyon Bradi3g43870.1
Brachypodium distachyon TR24279 c0 g1
Brachypodium distachyon Bradi3g43855.1
Hordeum vulgare DHN12 GB:AF155129.1
Stipa lagascae TR57057 c0 g1
Hordeum vulgare DHN9 GB:AF043094.1
Hordeum vulgare TR54745 c0 g1
Triticum turgidum DHN1 GB:AM180929.1
Lolium perenne GAYX01087883
Melica nutans TR63790 c0 g1
Brachypodium distachyon Bradi1g37410.1
Brachypodium distachyon TR17726 c0 g2
Hordeum vulgare DHN3 GB:AF043089.1
Hordeum vulgare TR14370 c0 g1
Hordeum vulgare DHN7 GB:AF043092.1
Hordeum vulgare DHN7 GB:KC963093.1
Triticum aestivum WZY1-1 GB:AF453444.1
Triticum aestivum WZY2 GB:EU395844.2
Hordeum vulgare DHN4 GB:AF043090.1
Triticum aestivum RAB15b GB:X62476.1
Lolium perenne GAYX01047201.1
Melica nutans TR119053 c3 g1
Nardus stricta TR68571 c1 g1
Stipa lagascae TR49468 c0 g1g3
Hordeum vulgare DHN10 GB:AF043095.1
Hordeum vulgare TR96195 c0 g1
Melica nutans TR64440 c2 g1
Stipa lagascae TR23813 c0 g1g2
Stipa lagascae TR66020 c0 g1
Melica nutans TR64440 c1 g1
Nardus stricta TR18116 c0 g1g2
Nardus stricta TR61407 c2 g1
Oryza sativa Os11g0454200
Oryza sativa RAB16c GB:XM 015762124.1
Oryza sativa RAB16d GB:XM 015762125.1
Oryza sativa RAB21 GB:XM 015762122.1
Sorghum bicolor Sb09g018420.1
Zea mays GRMZM2G079440 T01
0.99
0.9
0.85
0.98
0.94
0.88
0.89
80
70
60
50
40
30
20
10
0
million years ago

Figure 4: Time calibrated phylogeny for the Pooideae Y_nSK_n-type DHN gene family. *The phylogeny was estimated with BEAST v1.8.2 using a GTR+Γ and an uncorrelated, lognormal distributed molecular clock model. Significant posterior probabilities are shown as branch labels.*

Chloroplast cold regulated genes

The phylogenetic analysis of the *ctCOR* gene family reconstructed a monophyletic origin for all Pooideae *COR14*/*WCS19* homologs. Within this Pooideae-specific clade, two previously reported (Crosatti et al. 2013) sister clades contained core Pooideae homologs of *WCS19* and *COR14* (Fig. 5). However, their sister relationship was not well supported by the BEAST analysis. Those two clades were nested within a monophyletic clade, containing homologous transcripts from *B. distachyon*, *M. nutans* and *S. lagascae*. Except of one *S. lagascae* transcript, all other transcripts in this clade were induced by short- and long-term cold. The remaining homologs formed an unsupported clade, which contained one *N. stricta* transcript that was induced by short- and long-term cold.

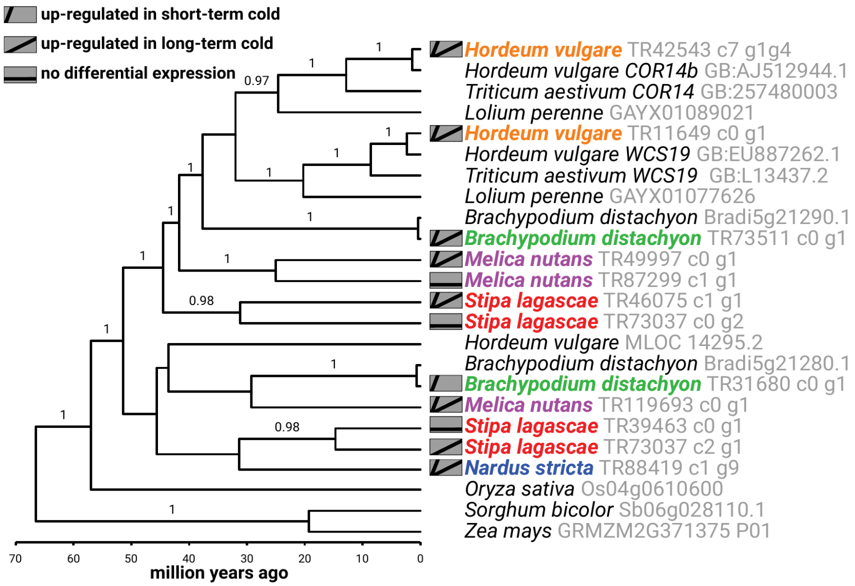

up-regulated in short-term cold
up-regulated in long-term cold
no differential expression
Hordeum vulgare TR42543 c7 g1g4
Hordeum vulgare COR14b GB:AJ512944.1
Triticum aestivum COR14 GB:257480003
Lolium perenne GAYX01089021
Hordeum vulgare TR11649 c0 g1
Hordeum vulgare WCS19 GB:EU887262.1
Triticum aestivum WCS19 GB:L13437.2
Lolium perenne GAYX01077626
Brachypodium distachyon Bradi5g21290.1
Brachypodium distachyon TR73511 c0 g1
Melica nutans TR49997 c0 g1
Melica nutans TR87299 c1 g1
Stipa lagascae TR46075 c1 g1
Stipa lagascae TR73037 c0 g2
Hordeum vulgare MLOC 14295.2
Brachypodium distachyon Bradi5g21280.1
Brachypodium distachyon TR31680 c0 g1
Melica nutans TR119693 c0 g1
Stipa lagascae TR39463 c0 g1
Stipa lagascae TR73037 c2 g1
Nardus stricta TR88419 c1 g9
Oryza sativa Os04g0610600
Sorghum bicolor Sb06g028110.1
Zea mays GRMZM2G371375 P01
1
1
0.97
1
1
1
1
1
1
1
1
0.98
1
1
1
0.98
1
1
70
60
50
40
30
20
10
0
million years ago

Figure 5: Time calibrated phylogeny for the Pooideae ctCOR gene family. *The phylogeny was estimated with BEAST v1.8.2 using a GTR+Γ and an uncorrelated, lognormal distributed molecular clock model. Significant posterior probabilities are shown as branch labels.*

Ice-recrystallization inhibition protein genes

We identified six *H. vulgare*, three *B. distachyon* and one *S. lagascae* transcripts that formed a monophyletic clade (Fig. S3) with previously identified Pooideae *IRIP* homologs (Li et al. 2012). The *S. lagascae* transcript and one *H. vulgare* transcript were truncated and did not contain a characteristic IRI motif (see Li et al. 2012), even though the transcript from *S. lagascae* was induced by short- and long-term cold. The remaining transcripts conatained IRI motifs and were but one induced by short- and long-term cold. No IRIP homologs were identified for *M. nutans* and *N. stricta*.

Fructosyltransferase genes

Phylogenetic analyses of the *FST* gene family reconstructed a monophyletic origin of the Pooideae *FST* homologs that contained two well supported main clades (Fig. S4). In the clade containing all known *FST* genes (Huynh et al. 2012) we identified two *H. vulgare* transcripts that were induced by long-term cold and contained the diagnostic FST motif (see Lasseur et al. 2008). Also part of that putative *FST* clade were one *H. vulgare*, one *B. distachyon* and one *M. nutans* transcript that lacked the FST motif. The other main clade contained putative vacuolar invertase 3 homologs of *H. vulgare*, *B. distachyon* and *N. stricta*.

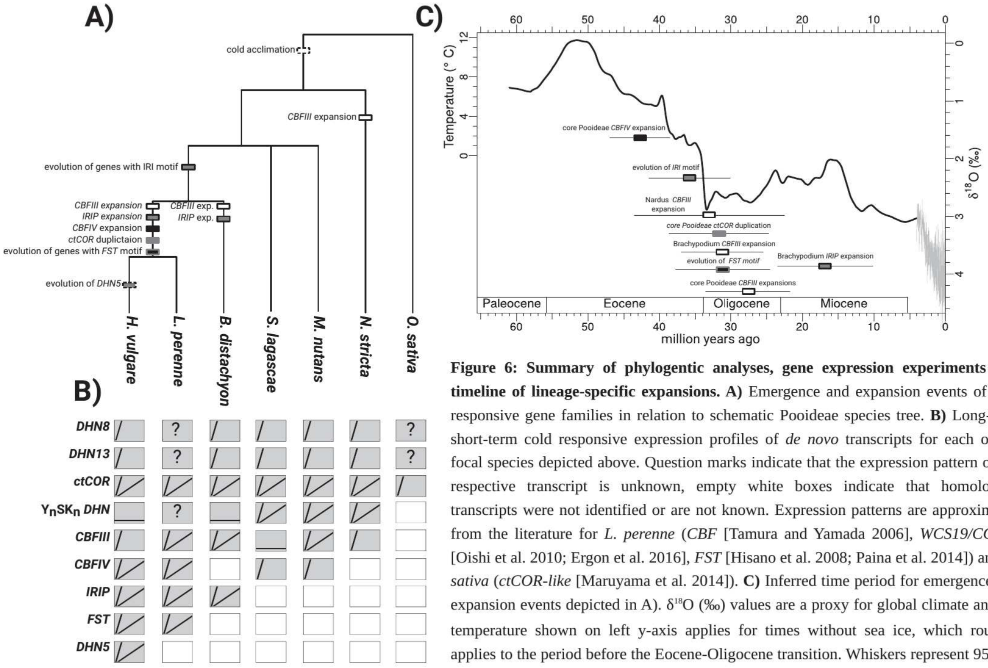


Figure 6: Summary of phylogentic analyses, gene expression experiments and timeline of lineage-specific expansions. A) Emergence and expansion events of cold responsive gene families in relation to schematic Pooideae species tree. **B)** Long- and short-term cold responsive expression profiles of *de novo* transcripts for each of the focal species depicted above. Question marks indicate that the expression pattern of the respective transcript is unknown, empty white boxes indicate that homologous transcripts were not identified or are not known. Expression patterns are approximated from the literature for *L. perenne* (*CBF* [Tamura and Yamada 2006], *WCS19/COR14* [Oishi et al. 2010; Ergon et al. 2016], *FST* [Hisano et al. 2008; Paina et al. 2014]) and *O. sativa* (*ctCOR-like* [Maruyama et al. 2014]). **C)** Inferred time period for emergence and expansion events depicted in A). δ^{18}O (‰) values are a proxy for global climate and the temperature shown on left y-axis applies for times without sea ice, which roughly applies to the period before the Eocene-Oligocene transition. Whiskers represent 95% of the highest posterior density (HPD) distribution.

Discussion

Cold acclimation evolved independently in different Pooideae lineages

Cold acclimation as adaptation to temperate climates is not restricted to core Pooideae species and *B. distachyon.* Our results (Fig. 1) show that cold acclimation is part of the cold adaptation in early-diverging Pooideae lineages. Cold acclimated plants of all tested species showed increased frost tolerance relative to non-acclimated plants. Interestingly, non-acclimated plants of the three early diverging Pooideae species exhibited high frost tolerance at 4°C relative to most tested core Pooideae and *Brachypodium* species. This might indicate that species of early diverging Pooideae possess a high tolerance against sudden frost shocks. This hypothesis however remains to be tested.

In core Pooideae, the gene *DHN5*, and genes from the families *FST*, *IRIP*, *CBFIV* and *CBFIII* are known to be crucially involved in cold acclimation and induced during short- and long-term cold (Choi et al. 2002; Vágújfalvi et al. 2003; Badawi et al. 2007; Knox et al. 2008; Zhang et al. 2010; Livingston et al. 2009; Knox et al. 2010; Soltesz et al. 2013; Jeknić et al. 2014; Todorovska et al. 2014; Marozsán-Tóth et al. 2015). The expression patterns of respective homologous *H. vulgare* transcripts presented here are in line with this research. However, the lack of conserved expression patterns of these genes and gene families among the tested Pooideae species indicates that cold acclimation is regulated differently in the five Pooideae lineages. These results contradict the hypothesis that key cold acclimation genes were recruited early in the Pooideae lineage. We therefore propose that regulation of cold acclimation evolved independently in Pooideae lineages and different genes were recruited.

For the *H. vulgare* cold acclimation gene *DHN5* and its ortholog in *T. aestivum WCS120*, no homologous transcripts were identified outside the Triticeae tribe and are neither reported in the literature (Kosová et al. 2012). Both genes belong to the K_n-type dehydrins, that are characterized by multiple copies of the dehydrin specific K-segment. The lack of any K_n-type dehydrins in other Pooideae species suggest that this specific family evolved relatively recently in the Triticeae – likely by expanding the K-segments of a dehydrin-like gene – and gained a crucial role in cold acclimation of Triticeae species (reviewed in Kosová et al. 2007, Kosová et al. 2012).

Two of the gene families whose function in cold acclimation is well described for core Pooideae are the *FST* and *IRIP* gene families (reviewed by Sandve et al. 2011), but no homologs were expressed in species of early-diverging lineages (Fig. 6). Although we identified *FST* homologs in *M. nutans* and *B. distachyon* (Fig. S4) and an *IRIP* homolog in *S. lagascae* (Fig. S3), the lack of diagnostic FST or IRI motifs suggests that functional genes are restricted to core Pooideae and core Pooideae-Brachypodieae, respectively. Hence, both gene families were recruited into cold acclimation in those two clades and represent more recent cold adaptations.

Genes of the *CBFIII* and *CBFIV* families code for transcription factors that are important for regulation of cold acclimation in core Pooideae and in case for *CBFIII* for *B. distachyon* (Badawi et al. 2007, Li et al. 2012). Our analyses could not resolve the species topology within the *CBFIII* gene tree (Fig. 2), but we identified homologous transcripts in all early diverging Pooideae species that were cold induced. Except for one *M. nutans* transcript however, we suggest that *CBFIII* homologs are not involved in cold acclimation of early diverging Pooideae species, because they are not induced by long-term cold treatment. Homologs closely related to *CBFIV* have never been reported for *B. distachyon* and also our results lack evidence for such transcripts. Although we identified *CBFIV* transcripts (Fig. 3) in *S. lagascae* and *M. nutans,* they were not induced by long-term cold and are unlikely to be involved in cold acclimation. We therefore suggest that *CBFIV* genes were recruited into cold acclimation in core Pooideae only, correlating with the expansion of this gene family.

While Y_nSK_n-type *DHN* homologs are known to be expressed during frost in core Pooideae – which is in line with our findings for *H. vulgare* – transcripts from *M. nutans, S. lagascae* and *N. stricta* are induced by long-term cold (Fig. 4). These results suggest that those transcripts might be involved in cold acclimation exclusively in early-diverging Pooideae lineages. Functional studies are necessary to investigate their role in cold acclimation.

The only tested genes whose function in cold acclimation might be conserved throughout the Pooideae are homologs of the *ctCOR* gene family (Fig. 5), because transcripts in all investigated species were induced by long-term. The gene tree failed to reconstruct the species topology of the Pooideae, due to an unsupported clade containing the one *N. stricta* transcript. Two duplication events – one after the divergence of *N. stricta* and one in the core Pooideae giving rise to the *COR14* and *WCS120* clade – could explain the abundance and relationship of *ctCOR* homologs in a parsimonious way. The conserved expression pattern of the *ctCOR* homologs renders this gene family a very interesting candidate for further studies of the regulation of cold acclimation in Pooideae. Our results suggest that *ctCOR* genes might have been involved in a putative cold acclimation of the Pooideae MRCA.

Opposed to cold acclimation genes which are induced by long-term cold. Genes involved in response to short-term cold seem to be more conserved. Among dehydrin genes homologs of *HvDHN8* (Fig. S1) and *HvDHN13* (Fig. S2) exhibited the same expression pattern in all species, in addition to highly conserved nucleotide sequences and a lack of duplication events. Because all other gene families showed signs of lineage specific evolution, the findings for *DHN8* and *DHN13* suggest that their cold-responsive function has been conserved since the MRCA.

Pooideae lineages evolved specific cold adaptation by expanding gene families

Using transcriptomic data to reconstruct phylogenetic relationships between genes is inherently problematic. Homologous genes might exist in early diverging Pooideae species, but might not be expressed. In addition, coding sequences might not contain sufficient informative substitutions to reconstruct the true species topology. Nonetheless we were able to identify homologous transcripts and argue that a lack of long-term cold induced transcripts correlates with lack of function in cold acclimation.

Furthermore the gene trees displayed several duplication events and gene family expansions. Based on our results we propose that gene family expansion was an important mode of cold adaptation in the Pooideae subfamily. Our results corroborate findings from Sandve and Fjellheim (2010), who identified an increase of gene copy number in *CBFIV*, *FST* and *IRIP* gene families as an evolutionary force of cold climate adaptation of core Pooideae and *Brachypodium* species. Although we lack sufficient genomic data for early diverging Pooideae species, we found evidence for Pooideae specific expansions in *CBFIII*, *Y_nSK_n-type DHN* and *ctCOR* gene families.

Brachypodium distachyon possesses long-term cold induced transcripts for the *IRIP* (Fig. S2) and also *CBFIIId* (Fig. 2) gene families. Both gene families expanded specifically in the *B. distachyon* lineage and we therefore argue that their function in cold acclimation evolved independently from the core Pooideae. The Y_nSK_n-type *DHN* family (Fig. 4) expanded into at least three distinct clades early in the Pooideae history, possibly in the MRCA. Another Y_nSK_n-type *DHN* clade (containing *HvDHN1* and *2*) seems to have emerged in the Brachypodieae and core Pooideae. Due to an unresolved gene tree it is difficult to assess where in the Pooideae phylogeny *CBFIII* gene family began to expand (Fig. 2), but it is apparent that there were at least three independent expansions in the Nardeae tribe, Brachypodieae tribe and the core Pooideae that lead to several gene copies. Expansion of gene families may have lead to functional specialization or novel functions of the various gene copies. Other studies have shown that stress related gene families tend to expand via tandem duplications (Hanada et al. 2008), which may lead to lineage-specific expansion of the gene family (Lespinet et al. 2002).

Our estimates placed the evolution of FST and IRI motifs as well as the core Pooideae specific duplication of the *ctCOR* family and *CBFIII* expansions in the period of the E-O transition (Fig. 6). Most interestingly, two independent expansions of *CBFIII* the lineage of *N. stricta* and *B. distachyon* happened during the same time period. Those findings partly confirm analyses by Sandve and Fjellheim (2010), who suggested that cold responsive gene families expanded due to increased selection pressure for improved cold adaptation during the E-O transition. In contrast to Sandve and Fjellheim (2010), we were unable to correlate initial core Pooideae-specific expansions of the *CBFIV* family with the E-O transition. In the case of *FST* and *IRIP* gene families the increased cold stress during the E-O transition might have led to the evolution of novel protein domains in the core Pooideae and Brachypodieae (Sandve et al. 2008; Li et al.

2012). It remains to be tested if similar novelties evolved in other lineages. Based on our findings, there is evidence that the E-O transition affected the molecular evolution of cold adaptive mechanisms in all investigated Pooideae lineages. By the time of the E-O transition, all major Pooideae lineages had already diverged (Marcussen et al. 2014; Grønvold et al. 2017) and this supports the indications in our data that cold adaptation largely evolved separately in different Pooideae lineages.

Drought tolerance – an evolutionary basis for cold tolerance?

Parts of the early cold tolerance in Pooideae might have been derived from ancestral drought tolerance. Cold stress may cause dehydration due to decreased membrane stability as well as lead to accumulation of ROS (Kratsch and Wise 2000; Murata et al. 2007; Crosatti 2013), similar to stresses occurring during drought (Mahajan and Tutej 2005). Following this, several authors have speculated about the importance of drought tolerance for the early evolution of the Pooideae (Kellogg 2001; Schardl et al. 2008; Vigeland et al. 2013).

Due to their molecular functions in membrane stabilizing and ROS scavenging, ancestrally drought-responsive genes were suitable candidates for cold responsive pathways when ancestral Pooideae species faced temperate conditions. It has been shown that genes with suitable, pre-existing molecular functions seem to be recruited into certain molecular pathways preferentially (True and Carroll 2003; Christin, Osborne et al. 2013; Christin et al. 2015). Some of our findings lend support to this scenario. Firstly, early-diverging lineages possess cold-induced transcripts of *CBFIII* (Fig. 2) and *CBFIV* (Fig. 3) and many *CBF* genes are known to be involved in drought tolerance in several angiosperms (Agarwal et al. 2006; Akhtar et al. 2012). Secondly, *ctCOR* (Fig. 5) transcripts are cold induced in all species. The *ctCOR* genes *COR14* and *WCS19* are thought to be involved ROS-mediated stress (Crosatti et al. 2013), which is both beneficial during drought and cold stress. Thirdly, *DHN8* (Fig. S1) is an ancestral drought-responsive gene, that gained a function in cold tolerance. Orthologs of *DHN8* are involved in the protection of plasma membrane during cold (Yang et al. 2014) and drought stress (Danyluk et al. 1998; Houde et al. 2004), due to a putative ROS scavenging function (Kumar et al. 2014). Since its drought responsiveness seems to be conserved outside the Pooideae (Lee et al. 2005; Badicean et al. 2012), it is likely that *DHN8* gained its cold responsiveness first in the Pooideae MRCA.

Conclusion

Taken together, our results provide valuable insights in the early adaptive evolution of Pooideae and contribute to the understanding how cold acclimation evolves in plants. We found signs of conserved cold response that likely existed in the Pooideae MRCA and increased the Pooideae's potential to evolve cold tolerance. The conserved fraction of cold response might have enabled early Pooideae members to survive the first encounters with temperate conditions, making subsequent cold adaptation possible. Even though all species were able to increase their frost tolerance in response to cold acclimation, we observed a trend of lineage specific evolution of the regulation of cold acclimation. This has led to increased gene family complexity by gene family expansion, particularly within the core Pooideae lineage (Fig. 6).

Due to the scarce fossil record, divergence times of early-diverging Pooideae lineages are still under dispute and so is the stem age of the Pooideae. In the absence of reliable fossil calibration points our results contribute to a better understanding of climatic changes, that influenced molecular evolution of certain gene families in Pooideae subfamily. However, more reliable calibration points and a resolved genome-level phylogeny of the Pooideae subfamily will help us to improve the dating of the evolutionary history of cold adaptation. This will enable us to confidently correlate paleoclimatic events, like the E-O transition, with molecular innovations in order to reconstruct the colonization of temperate biomes by the Pooideae subfamily. And lastly, higher resolution, i.e. higher phylogenetic coverage, of cold adaptation evolution in the core Pooideae lineage will be valuable to understand the molecular traits that might have contributed to their putative rapid radiation, species richness and expansion into extreme habitats.

Acknowledgments

The presented research was funded by the Nansen Foundation and through the TVERRforsk grant provided to SF, TRH and SRS by the Norwegian University of Life Sciences (NMBU). This work was part of the PhD project of MS and LG funded by NMBU. We thank Åsmund Bjørnstad, Morten Lillemo, Mallikarjuna Rao Kovi and USDA-NPGS for providing seeds of *H. vulgare*, *T. aestivum*, *L. perenne* and *S. lagascae*, respectively. For technical assistance handling plants during growth experiments we thank Øyvind Jørgensen. We are grateful to Pascal-Antoine Christin for invaluable advice and to Erica Leder, Thomas Marcussen, Ursula Brandes, Camilla Lorange Lindberg and Martin Paliocha for helpful comments on earlier versions of this manuscript.

Literature

Agarwal PK, Agarwal P, Reddy MK, Sopory SK. 2006. Role of DREB transcription factors in abiotic and biotic stress tolerance in plants. Plant Cell Rep. 25:1263–1274.

Akhtar M, Jaiswal A, Taj G, Jaiswal JP, Qureshi MI, Singh NK. 2012. DREB1/CBF transcription factors: their structure, function and role in abiotic stress tolerance in plants. J. Genet. 91:385–395.

Alm V, Busso CS, Ergon Å, Rudi H, Larsen A, Humphreys MW, Rognli OA. 2011. QTL analyses and comparative genetic mapping of frost tolerance, winter survival and drought tolerance in meadow fescue (Festuca pratensis Huds.). Theor. Appl. Genet. 123:369–382.

Antikainen M, Griffith M. 1997. Antifreeze protein accumulation in freezing-tolerant cereals. Physiol. Plant. 99:423–432.

Badawi M, Danyluk J, Boucho B, Houde M, Sarhan F. 2007. The *CBF* gene family in hexaploid wheat and its relationship to the phylogenetic complexity of cereal *CBF*s. Mol. Genet. Genomics 277:533–554.

Badicean D, Scholten S, Jacota A. 2012. Transcriptional profiling of *Zea mays* genotypes with different drought tolerances – new perspectives for gene expression markers selection. Maydica 56.

Battaglia M, Olvera-Carrillo Y, Garciarrubio A, Campos F, Covarrubias AA. 2008. The Enigmatic LEA Proteins and Other Hydrophilins. Plant Physiol. 148:6–24.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120.

Bouchenak-Khelladi Y, Verboom AG, Savolainen V, Hodkinson TR. 2010. Biogeography of the grasses (Poaceae): a phylogenetic approach to reveal evolutionary history in geographical space and geological time. Bot. J. Linn. Soc. 162:543–557.

Chalmers J, Lidgett A, Cummings N, Cao Y, Forster J, Spangenberg G. 2005. Molecular genetics of fructan metabolism in perennial ryegrass. Plant Biotechnol. J. 3:459–474.

Choi D-W, Close TJ. 2000. A newly identified barley gene, *Dhn12* , encoding a YSK 2 DHN, is located on chromosome 6H and has embryo-specific expression. TAG Theor. Appl. Genet. 100:1274–1278.

Choi D-W, Rodriguez EM, Close TJ. 2002. Barley *Cbf3* gene identification, expression pattern, and map location. Plant Physiol. 129:1781–1787.

Christin P-A, Arakaki M, Osborne CP, Edwards EJ. 2015. Genetic enablers underlying the clustered evolutionary origins of C4 photosynthesis in angiosperms. Mol. Biol. Evol. 32:846–858.

Christin P-A, Boxall SF, Gregory R, Edwards EJ, Hartwell J, Osborne CP. 2013. Parallel recruitment of multiple genes into C4 photosynthesis. Genome Biol. Evol. 5:2174–2187.

Christin P-A, Osborne CP. 2013. The recurrent assembly of C4 photosynthesis, an evolutionary tale. Photosynth. Res. 117:163–175.

Christin P-A, Osborne CP, Chatelet DS, Columbus JT, Besnard G, Hodkinson TR, Garrison LM, Vorontsova MS, Edwards EJ. 2013. Anatomical enablers and the evolution of C4 photosynthesis in grasses. Proc. Natl. Acad. Sci. U. S. A. 110:1381–1386.

Christin P-A, Spriggs E, Osborne CP, Strömberg CAE, Salamin N, Edwards EJ. 2014. Molecular dating, evolutionary rates, and the age of the grasses. Syst. Biol. 63:153–165.

Clayton WD. 1981. Evolution and distribution of grasses. Ann. Missouri Bot. Gard. 68:5–14.

Close TJ. 1997. Dehydrins: A commonalty in the response of plants to dehydration and low temperature. Physiol. Plant. 100:291–296.

Crisp MD, Arroyo MTK, Cook LG, Gandolfo MA, Jordan GJ, McGlone MS, Weston PH, Westoby M, Wilf P, Linder HP. 2009. Phylogenetic biome conservatism on a global scale. Nature 458:754–756.

Crosatti C, Laureto PP de, Bassi R, Cattivelli L. 1999. The interaction between cold and light controls the expression of the cold-regulated barley gene cor14b and the accumulation of the corresponding protein. Plant Physiol. 119:671–680.

Crosatti C, Rizza F, Badeck FW, Mazzucotelli E, Cattivelli L. 2013. Harden the chloroplast to protect the plant. Physiol. Plant. 147:55–63.

Danyluk J, Perron A, Houde M, Limin A, Fowler B, Benhamou N, Sarhan F. 1998. Accumulation of an acidic dehydrin in the vicinity of the plasma membrane during cold acclimation of wheat. Plant Cell 10:623–638.

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat. Methods 9:772.

Davis JI, Soreng RJ. 1993. Phylogenetic structure in the grass family (Poaceae) as inferred from chloroplast DNA restriction site variation. Am. J. Bot. 80:1444–1454.

Donoghue MJ. 2008. A phylogenetic perspective on the distribution of plant diversity. Proc. Natl. Acad. Sci. U. S. A. 105:11549–11555.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29:1969–1973.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Edwards EJ, Donoghue MJ. 2013. Is it easy to move and easy to evolve? Evolutionary accessibility and adaptation. J. Exp. Bot. 64:4047–4052.

Edwards EJ, Smith SA. 2010. Phylogenetic analyses reveal the shady history of C_4 grasses. Proc. Natl. Acad. Sci. U. S. A. 107:2532–2537.

Eldrett JS, Greenwood DR, Harding IC, Huber M. 2009. Increased seasonality through the Eocene to Oligocene transition in northern high latitudes. Nature 459:969–973.

Ergon Å, Melby TI, Höglind M, Rognli OA. 2016. Vernalization Requirement and the Chromosomal *VRN1*-Region can Affect Freezing Tolerance and Expression of Cold-Regulated Genes in *Festuca pratensis*. Front. Plant Sci. 7:207.

Fjellheim S, Boden S, Trevaskis B. 2014. The role of seasonal flowering responses in adaptation of grasses to temperate climates. Front. Plant Sci. 5:431.

Galiba G, Vágújfalvi A, Li C, Soltész A, Dubcovsky J. 2009. Regulatory genes involved in the determination of frost tolerance in temperate cereals. Plant Sci. 176:12–19.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29:644–652.

Gray GR, Chauvin LP, Sarhan F, Huner NPA. 1997. Cold acclimation and freezing tolerance - A complex interaction of light and temperature. Pant Physiol. 114:467–474.

Griffith M, Ewart KV. 1995. Antifreeze proteins and their potential use in frozen foods. Biotechnol. Adv. 13:375–402.

Griffith M, Yaish MWF. 2004. Antifreeze proteins in overwintering plants: a tale of two activities. Trends Plant Sci. 9:399–405.

Grønvold, L., Schubert, M., Sandve, S.R., Fjellheim, S., Torgeir, R., 2017. Comparative transcriptomics reveals lineage specific evolution of cold response in Pooideae. bioRxiv 1–32. doi:10.1101/151431

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8:1494–1512.

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu S-H. 2008. Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant Physiol. 148:993–1003.

Hartley W. 1973. Studies on origin, evolution, and distribution of Gramineae. 5. The subfamily Festucoideae. Aust. J. Bot. 21:201–234.

He X, Zhang J. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics 169:1157–1164.

Hincha DK, Hellwege EM, Heyer AG, Crowe JH. 2000. Plant fructans stabilize phosphatidylcholine liposomes during freeze-drying. Eur. J. Biochem. 267:535–540.

Hisano H, Kanazawa A, Kawakami A, Yoshida M, Shimamoto Y, Yamada T. 2004. Transgenic perennial ryegrass plants expressing wheat fructosyltransferase genes accumulate increased amounts of fructan and acquire increased tolerance on a cellular level to freezing. Plant Sci. 167:861–868.

Hisano H, Kanazawa A, Yoshida M, Humphreys MO, Iizuka M, Kitamura K, Yamada T. 2008. Coordinated expression of functionally diverse fructosyltransferase genes is associated with fructan accumulation in response to low temperature in perennial ryegrass. New Phytol. 178:766–780.

Houde M, Dallaire S, N'Dong D, Sarhan F. 2004. Overexpression of the acidic dehydrin WCOR410 improves freezing tolerance in transgenic strawberry leaves. Plant Biotechnol. J. 2:381–387.

Jeknić Z, Pillman KA, Dhillon T, Skinner JS, Veisz O, Cuesta-Marcos A, Hayes PM, Jacobs AK, Chen THH, Stockinger EJ. 2014. *Hv-CBF2A* overexpression in barley accelerates *COR* gene transcript accumulation and acquisition of freezing tolerance during cold acclimation. Plant Mol. Biol. 84:67–82.

John UP, Polotnianka RM, Sivakumaran KA, Chew O, Mackin L, Kuiper MJ, Talbot JP, Nugent GD, Mautord J, Schrauf GE, et al. 2009. Ice recrystallization inhibition proteins (IRIPs) and freeze tolerance in the cryophilic Antarctic hair grass *Deschampsia antarctica* E. Desv. Plant. Cell Environ. 32:336–348.

Judd WS, Sanders RW, Donoghue MJ. 1994. Angiosperm family pairs: preliminary phylogenetic analyses. Harvard Pap. Bot. 1:1–51.

Karami A, Shahbazi M, Niknam V, Shobbar ZS, Tafreshi RS, Abedini R, Mabood HE. 2013. Expression analysis of dehydrin multigene family across tolerant and susceptible barley (*Hordeum vulgare* L.) genotypes in response to terminal drought stress. Acta Physiol. Plant. 35:2289–2297.

Kellogg EA. 2001. Evolutionary history of the grasses. Plant Physiol. 125:1198–1205.

Kerkhoff AJ, Moriarty PE, Weiser MD. 2014. The latitudinal species richness gradient in New World woody angiosperms is consistent with the tropical conservatism hypothesis. Proc. Natl. Acad. Sci. U. S. A. 111:8125–8130.

Knox AK, Dhillon T, Cheng H, Tondelli A, Pecchioni N, Stockinger EJ. 2010. *CBF* gene copy number variation at *Frost Resistance-2* is associated with levels of freezing tolerance in temperate-climate cereals. Theor. Appl. Genet. 121:21–35.

Knox AK, Li C, Vágújfalvi A, Galiba G, Stockinger EJ, Dubcovsky J. 2008. Identification of candidate CBF genes for the frost tolerance locus Fr-Am2 in Triticum monococcum. Plant Mol. Biol. 67:257–270.

Koag M-C, Fenton RD, Wilkens S, Close TJ. 2003. The binding of maize DHN1 to lipid vesicles. Gain of structure and lipid specificity. Plant Physiol. 131:309–316.

Koag M-C, Wilkens S, Fenton RD, Resnik J, Vo E, Close TJ. 2009. The K-segment of maize DHN1 mediates binding to anionic phospholipid vesicles and concomitant structural changes. Plant Physiol. 150:1503–1514.

Kosová K, Vítámvás P, Prášil IT. 2007. The role of dehydrins in plant response to cold. Biol. Plant. 51:601–617.

Kosová K, Vítámvás P, Prášil IT. 2014. Wheat and barley dehydrins under cold, drought, and salinity - what can LEA-II proteins tell us about plant stress response? Front. Plant Sci. 5:343.

Kosová K, Vítámvás P, Prášilová P, Prášil IT. 2012. Accumulation of WCS120 and DHN5 proteins in differently frost-tolerant wheat and barley cultivars grown under a broad temperature scale. Biol. Plant. 57:105–112.

Kratsch HA, Wise RR. 2000. The ultrastructure of chilling stress. Plant, Cell Environ. 23:337–350.

Kumar M, Lee S-C, Kim J-Y, Kim S-J, Aye SS, Kim S-R. 2014. Over-expression of dehydrin gene, *OsDhn1*, improves drought and salt stress tolerance through scavenging of reactive oxygen species in rice (*Oryza sativa* L.). J. Plant Biol. 57:383–393.

Kumble KD, Demmer J, Fish S, Hall C, Corrales S, DeAth A, Elton C, Prestidge R, Luxmanan S, Marshall CJ, et al. 2008. Characterization of a family of ice-active proteins from the Ryegrass, *Lolium perenne*. Cryobiology 57:263–268.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10:R25.

Larsson A. 2014. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics 30:3276–3278.

Lasseur, B., Schroeven, L., Lammens, W., Le Roy, K., Spangenberg, G., Manduzio, H., Vergauwen, R., Lothier, J., Prud'homme, M.-P., Van den Ende, W., 2008. Transforming a Fructan:Fructan 6G-Fructosyltransferase from Perennial Ryegrass into a Sucrose:Sucrose 1-Fructosyltransferase. Plant Physiol 149:327–339.

Lee S-C, Lee M-Y, Kim S-J, Jun S-H, An G, Kim S-R. 2005. Characterization of an Abiotic Stress-inducible Dehydrin Gene, *OsDhn1*, in Rice (*Oryza sativa* L.). Mol. Cells 19:212–218.

Lespinet O, Wolf YI, Koonin E V., Aravind L. 2002. The Role of Lineage-Specific Gene Family Expansion in the Evolution of Eukaryotes. Genome Res. 12:1048–1059.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12:323.

Li C, Rudi H, Stockinger EJ, Cheng H, Cao M, Fox SE, Mockler TC, Westereng B, Fjellheim S, Rognli OA, et al. 2012. Comparative analyses reveal potential uses of *Brachypodium distachyon* as a model for cold stress responses in temperate grasses. BMC Plant Biol. 12:65.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Res. 13:2178–2189.

Livingston DP, Hincha DK, Heyer AG. 2009. Fructan and its relationship to abiotic stress tolerance in plants. Cell. Mol. Life Sci. 66:2007–2023.

Love MI, Huber W, Anders S, Lönnstedt I, Speed T, Robinson M, Smyth G, McCarthy D, Chen Y, Smyth G, et al. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15:550.

McKeown M, Schubert M, Marcussen T, Fjellheim S, Preston JC. 2016. Evidence for an early origin of vernalization responsiveness in temperate Pooideae grasses. Plant Physiol. 172:416–426.

McKeown M, Schubert M, Preston JC, Fjellheim S. 2017. Evolution of the miR5200-FLOWERING LOCUS T flowering time regulon in the temperate grass subfamily Pooideae. Mol. Phylogenet. Evol. 114:111–121.

Mahajan S, Tuteja N. 2005. Cold, salinity and drought stresses: an overview. Arch. Biochem. Biophys.

Marcussen T, Sandve SR, Heier L, Spannagl M, Pfeifer M, Jakobsen KS, Wulff BBH, Steuernagel B, Mayer KFX, Olsen O-A. 2014. Ancient hybridizations among the ancestral genomes of bread wheat. Science 345:1250092.

Marozsán-Tóth Z, Vashegyi I, Galiba G, Tóth B. 2015. The cold response of *CBF* genes in barley is regulated by distinct signaling mechanisms. J. Plant Physiol. 181:42–49.

Maruyama K, Urano K, Yoshiwara K, Morishita Y, Sakurai N, Suzuki H, Kojima M, Sakakibara H, Shibata D, Saito K, et al. 2014. Integrated analysis of the effects of cold and dehydration on rice metabolites, phytohormones, and gene transcripts. Plant Physiol. 164:1759–1771.

Mudelsee M, Bickert T, Lear CH, Lohmann G. 2014. Cenozoic climate changes: A review based on time series analysis of marine benthic δ 18 O records. Rev. Geophys. 52:333–374.

Murata N, Takahashi S, Nishiyama Y, Allakhverdiev SI. 2007. Photoinhibition of photosystem II under environmental stress. Biochim. Biophys. Acta - Bioenerg. 1767:414–421.

Oishi H, Takahashi W, Ebina M, Takamizo T. 2010. Expression and gene structure of the cold-stimulated gene *Lcs19* of Italian ryegrass (*Lolium multiflorum* Lam.). Breed. Sci.

Olave-Concha N, Ruiz-Lara S, Muñoz X, Bravo LA, Corcuera LJ. 2004. Accumulation of dehydrin transcripts and proteins in response to abiotic stresses in *Deschampsia antarctica*. Antarct. Sci. 16:175–184.

Paina C, Byrne SL, Domnisoru C, Asp T. 2014. Vernalization Mediated Changes in the *Lolium perenne* Transcriptome. PLoS One 9:e107365.

Pearce R. 2001. Plant Freezing and Damage. Ann. Bot. 87:417–424.

Potts R, Behrensmeyer A. 1992. Late Cenozoic terrestrial ecosystems. In: Behrensmeyer AK, Damuth JD, DiMichele WA, Potts R, Sues H-D, Wing SL, editors. Terrestrial ecosystems through time: evolutionary paleoecology of terrestrial plants and animals. 1st ed. Chicago: The University of Chicago Press. p. 419–451.

Prasad V, Strömberg CAE, Leaché AD, Samant B, Patnaik R, Tang L, Mohabey DM, Ge S, Sahni A. 2011. Late Cretaceous origin of the rice tribe provides evidence for early diversification in Poaceae. Nat. Commun.:480.

Preston JC, Sandve SR. 2013. Adaptation to seasonality and the winter freeze. Front. Plant Sci. 4:167.

R Core Team. 2016. R: A Language and Environment for Statistical Computing.

Rorat T. 2006. Plant dehydrins — Tissue location, structure and function. Cell. Mol. Biol. Lett. 11:536–556.

Sandve SR, Fjellheim S. 2010. Did gene family expansions during the Eocene-Oligocene boundary climate cooling play a role in Pooideae adaptation to cool climates? Mol. Biol. 19:2075–2088.

Sandve SR, Kosmala A, Rudi H, Fjellheim S, Rapacz M, Yamada T, Rognli OA. 2011. Molecular mechanisms underlying frost tolerance in perennial grasses adapted to cold climates. Plant Sci. 180:69–77.

Sandve SR, Rudi H, Asp T, Rognli OA. 2008. Tracking the evolution of a cold stress associated gene family in cold tolerant grasses. BMC Evol. Biol. 8:1–15.

Sarhan F, Danyluk J, Jaglo-Ottosen KR, Al. E, Sarhan F, Ouellet F, Vazquez-Tello A, Stockinger EJ, Gilmour SJ, Thomashow MF, et al. 1998. Engineering cold-tolerant crops—throwing the master switch. Trends Plant Sci. 3:289–290.

Schardl CL, Craven KD, Speakman S, Stromberg A, Lindstrom A, Yoshida R. 2008. A novel test for host-symbiont codivergence indicates ancient origin of fungal endophytes in grasses. Syst. Biol. 57:483–498.

Sidebottom C, Buckley S, Pudney P, Twigg S, Jarman C, Holt C, Telford J, McArthur A, Worrall D, Hubbard R, et al. 2000. Heat-stable antifreeze protein from grass. Nature 406:256.

Soltesz A, Smedley M, Vashegyi I, Galiba G, Harwood W, Vagujfalvi A. 2013. Transgenic barley lines prove the involvement of *TaCBF14* and *TaCBF15* in the cold acclimation process and in frost tolerance. J. Exp. Bot. 64:1849–1862.

Soreng RJ, Davis JI. 1998. Phylogenetics and character evolution in the grass family (Poaceae): Simultaneous analysis of morphological and Chloroplast DNA restriction site character sets. Bot. Rev. 64:1–85.

Soreng RJ, Peterson PM, Romaschenko K, Davidse G, Zuloaga FO, Judziewicz EJ, Filgueiras TS, Davis JI, Morrone O. 2015. A worldwide phylogenetic classification of the Poaceae (Gramineae). J. Syst. Evol. 53:117–137.

Spriggs EL, Christin PA, Edwards EJ. 2014. C4 photosynthesis promoted species diversification during the miocene grassland expansion.Allaby R, editor. PLoS One 9:e97722.

Stickley CE, St John K, Koç N, Jordan RW, Passchier S, Pearce RB, Kearns LE. 2009. Evidence for middle Eocene Arctic sea ice from diatoms and ice-rafted debris. Nature 460:376–379.

Strömberg CAE. 2011. Evolution of Grasses and Grassland Ecosystems. Annu. Rev. Earth Planet. Sci. 39:517–544.

Tamura K-I, Sanada Y, Tase K, Kawakami A, Yoshida M, Yamada T. 2014. Comparative study of transgenic *Brachypodium distachyon* expressing sucrose:fructan 6-fructosyltransferases from wheat and timothy grass with different enzymatic properties. Planta 239:783–792.

Tamura K, Yamada T. 2006. A perennial ryegrass *CBF* gene cluster is located in a region predicted by conserved synteny between Poaceae species. Theor. Appl. Genet. 114:273–283.

Thomashow MF. 1999. Plant cold acclimation: freezing tolerance genes and regulatory mechanisms. Annu. Rev. Plant Physiol. Plant Mol. Biol. 50:571–599.

Thomashow MF. 2010. Molecular basis of plant cold acclimation: insights gained from studying the CBF cold response pathway. Plant Physiol. 154:571–577.

Todorovska EG, Kolev S, Christov NK, Balint A, Kocsy G, Vágújfalvi A, Galiba G. 2014. The expression of CBF genes at *Fr-2* locus is associated with the level of frost tolerance in Bulgarian winter wheat cultivars. Biotechnol. Biotechnol. Equip. 28:392–401.

Tommasini L, Svensson JT, Rodriguez EM, Wahid A, Malatrasi M, Kato K, Wanamaker S, Resnik J, Close TJ. 2008. Dehydrin gene expression provides an indicator of low temperature and drought stress: transcriptome-based analysis of barley (*Hordeum vulgare* L.). Funct. Integr. Genomics 8:387–405.

Tremblay K, Ouellet F, Fournier J, Danyluk J, Sarhan F. 2005. Molecular characterization and origin of novel bipartite cold-regulated ice recrystallization inhibition proteins from cereals. Plant Cell Physiol. 46:884–891.

True JR, Carroll SB. 2002. Gene co-option in physiological and morphological evolution. Annu. Rev. Cell Dev. Biol. 18:53–80.

Tsvetanov S, Ohno R, Tsuda K. 2000. A cold-responsive wheat (*Triticum aestivum* L.) gene *wcor14* identified in a winter-hardy cultivar'Mironovska 808'. Genes Genet. Syst. 75:49–57.

Uemura M, Joseph RA, Steponkus PL. 1995. Cold Acclimation of Arabidopsis thaliana (Effect on Plasma Membrane Lipid Composition and Freeze-Induced Lesions). Plant Physiol. 109:15–30.

Vágújfalvi A, Galiba G, Cattivelli L, Dubcovsky J. 2003. The cold-regulated transcriptional activator *Cbf3* is linked to the frost-tolerance locus *Fr-2A* on wheat chromosome 5A. Mol. Genet. Genomics 269:60–67.

Vigeland MD, Spannagl M, Asp T, Paina C, Rudi H, Rognli O, Fjellheim S, Sandve SR. 2013. Evidence for adaptive evolution of low-temperature stress response genes in a Pooideae grass ancestor. New Phytol. 199:1060–1068.

Wiens JJ, Donoghue MJ. 2004. Historical biogeography, ecology and species richness. Trends Ecol. Evol. 19:639–644.

Woods, D.P., McKeown, M.A., Dong, Y., Preston, J.C., Amasino, R.M., 2016. Evolution of VRN2/Ghd7-like genes in vernalization-mediated repression of grass flowering. Plant Physiol. 170, 2124–2135.

Yang Y, Sun X, Yang S, Li X, Yang Y. 2014. Molecular cloning and characterization of a novel SK3-type dehydrin gene from *Stipa purpurea*. Biochem. Biophys. Res. Commun. 448:145–150.

Zachos J, Pagani M, Sloan L, Thomas E, Billups K. 2001. Trends, rhythms, and aberrations in global climate 65 Ma to present. Science 292:686–693.

Zhang C, Fei S, Arora R, Hannapel DJ. 2010. Ice recrystallization inhibition proteins of perennial ryegrass enhance freezing tolerance. Planta 232:155–164.

Zolotarov Y, Strömvik M. 2015. De Novo Regulatory Motif Discovery Identifies Significant Motifs in Promoters of Five Classes of Plant Dehydrin Genes. PLoS One 10:e0129016.

Zou C, Lehti-Shiu MD, Thomashow M, Shiu S-H. 2009. Evolution of stress-regulated gene expression in duplicate genes of Arabidopsis thaliana.Copenhaver GP, editor. PLoS Genet. 5:e1000581.

Supplementary figures

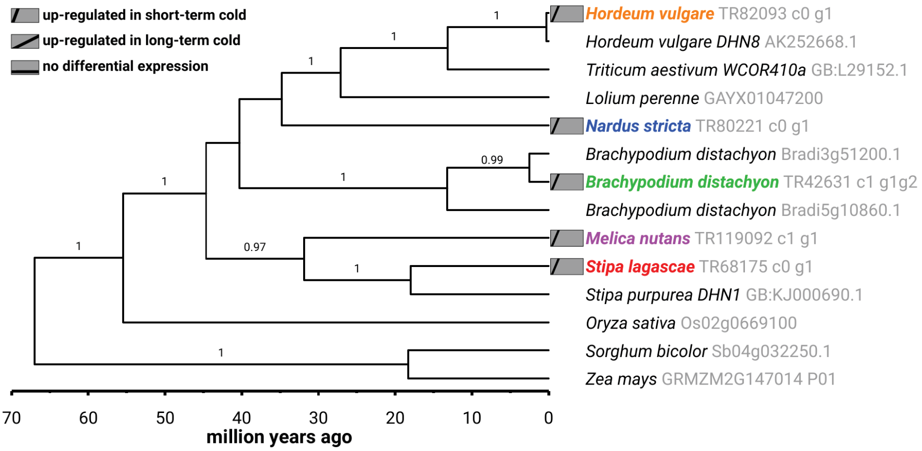

up-regulated in short-term cold
up-regulated in long-term cold
no differential expression
Hordeum vulgare TR82093 c0 g1
Hordeum vulgare DHN8 AK252668.1
Triticum aestivum WCOR410a GB:L29152.1
Lolium perenne GAYX01047200
Nardus stricta TR80221 c0 g1
Brachypodium distachyon Bradi3g51200.1
Brachypodium distachyon TR42631 c1 g1g2
Brachypodium distachyon Bradi5g10860.1
Melica nutans TR119092 c1 g1
Stipa lagascae TR68175 c0 g1
Stipa purpurea DHN1 GB:KJ000690.1
Oryza sativa Os02g0669100
Sorghum bicolor Sb04g032250.1
Zea mays GRMZM2G147014 P01
1
1
1
1
0.99
1
1
0.97
1
1
1
70
60
50
40
30
20
10
0
million years ago

Figure S1: Time calibrated phylogeny for the Pooideae DHN8 gene family. *The phylogeny was estimated with BEAST v1.8.2 using a GTR+Γ and an uncorrelated, lognormal distributed molecular clock model. Significant posterior probabilities are shown as branch labels.*

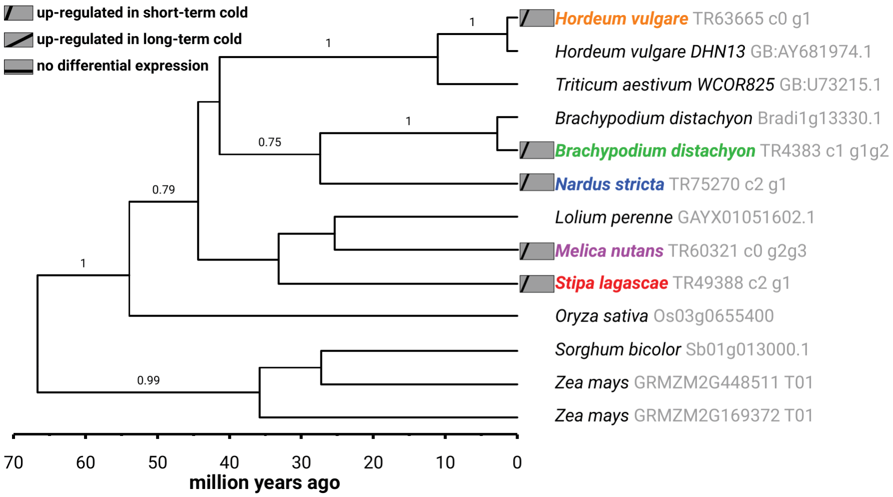


Figure S2: Time calibrated phylogeny for the Pooideae DHN13 gene family. *The phylogeny was estimated with BEAST v1.8.2 using a GTR+Γ and an uncorrelated, lognormal distributed molecular clock model. Significant posterior probabilities are shown as branch labels.*

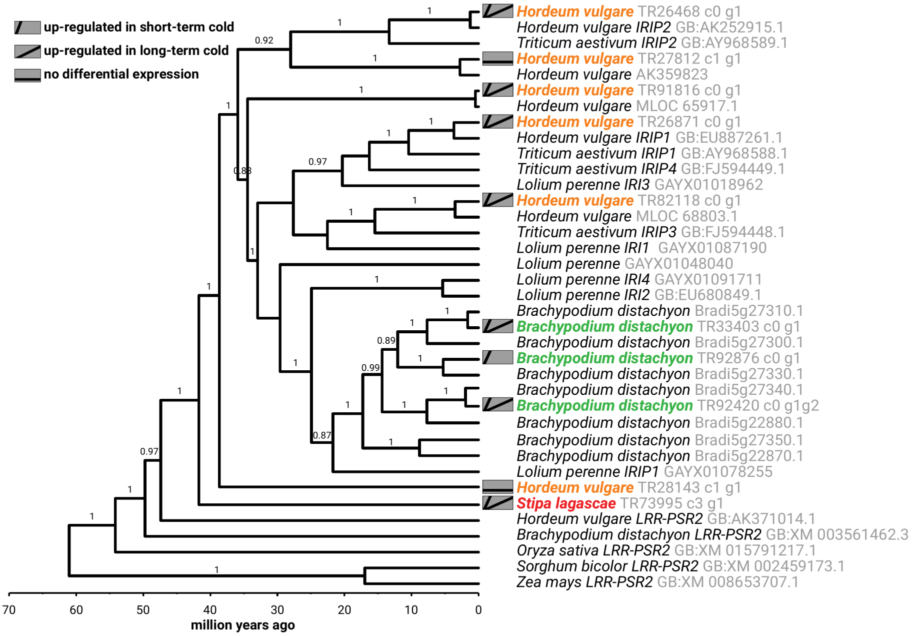
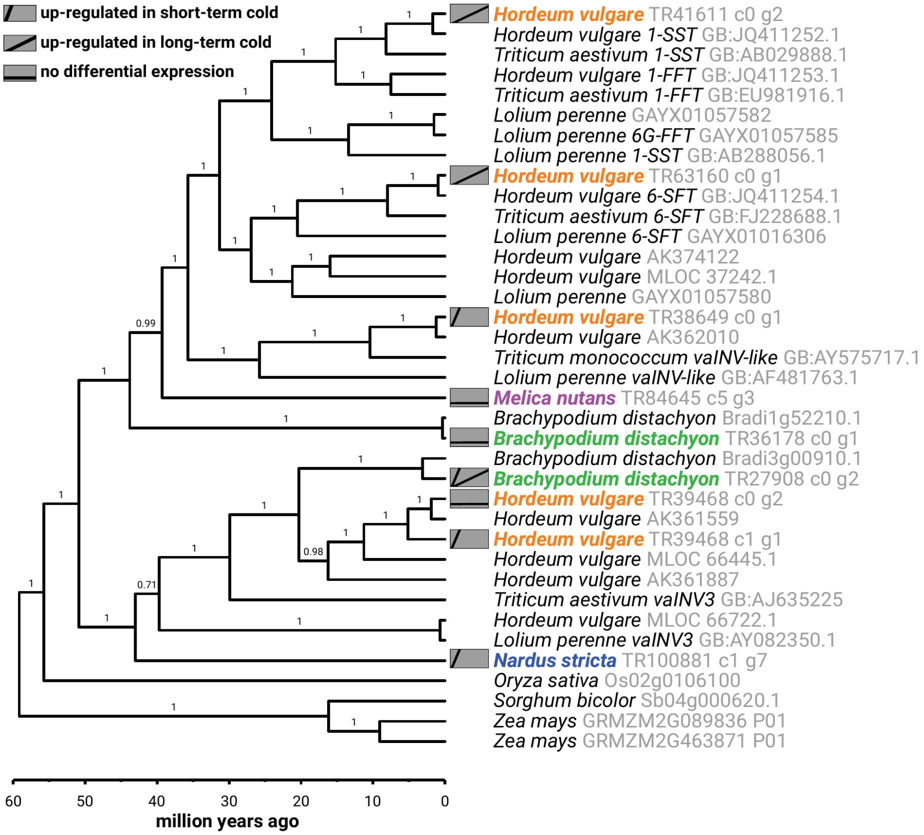

up-regulated in short-term cold
up-regulated in long-term cold
no differential expression
Hordeum vulgare TR26468 c0 g1
Hordeum vulgare IRIP2 GB:AK252915.1
Triticum aestivum IRIP2 GB:AY968589.1
Hordeum vulgare TR27812 c1 g1
Hordeum vulgare AK359823
Hordeum vulgare TR91816 c0 g1
Hordeum vulgare MLOC 65917.1
Hordeum vulgare TR26871 c0 g1
Hordeum vulgare IRIP1 GB:EU887261.1
Triticum aestivum IRIP1 GB:AY968588.1
Triticum aestivum IRIP4 GB:FJ594449.1
Lolium perenne IRI3 GAYX01018962
Hordeum vulgare TR82118 c0 g1
Hordeum vulgare MLOC 68803.1
Triticum aestivum IRIP3 GB:FJ594448.1
Lolium perenne IRI1 GAYX01087190
Lolium perenne GAYX01048040
Lolium perenne IRI4 GAYX01091711
Lolium perenne IRI2 GB:EU680849.1
Brachypodium distachyon Bradi5g27310.1
Brachypodium distachyon TR33403 c0 g1
Brachypodium distachyon Bradi5g27300.1
Brachypodium distachyon TR92876 c0 g1
Brachypodium distachyon Bradi5g27330.1
Brachypodium distachyon Bradi5g27340.1
Brachypodium distachyon TR92420 c0 g1g2
Brachypodium distachyon Bradi5g22880.1
Brachypodium distachyon Bradi5g27350.1
Brachypodium distachyon Bradi5g22870.1
Lolium perenne IRIP1 GAYX01078255
Hordeum vulgare TR28143 c1 g1
Stipa lagascae TR73995 c3 g1
Hordeum vulgare LRR-PSR2 GB:AK371014.1
Brachypodium distachyon LRR-PSR2 GB:XM 003561462.3
Oryza sativa LRR-PSR2 GB:XM 015791217.1
Sorghum bicolor LRR-PSR2 GB:XM 002459173.1
Zea mays LRR-PSR2 GB:XM 008653707.1
1
1
0.92
1
1
1
0.88
1
0.97
1
1
1
1
1
1
1
0.89
1
0.99
1
1
1
1
0.87
1
1
0.97
1
1
1
70
60
50
40
30
20
10
0
million years ago

Figure S3: Time calibrated phylogeny for the Pooideae IRIP gene family. *The phylogeny was estimated with BEAST v1.8.2 using a GTR+Γ and an uncorrelated, lognormal distributed molecular clock model. Significant posterior probabilities are shown as branch labels.*



up-regulated in short-term cold
up-regulated in long-term cold
no differential expression
Hordeum vulgare TR41611 c0 g2
Hordeum vulgare 1-SST GB:JQ411252.1
Triticum aestivum 1-SST GB:AB029888.1
Hordeum vulgare 1-FFT GB:JQ411253.1
Triticum aestivum 1-FFT GB:EU981916.1
Lolium perenne GAYX01057582
Lolium perenne 6G-FFT GAYX01057585
Lolium perenne 1-SST GB:AB288056.1
Hordeum vulgare TR63160 c0 g1
Hordeum vulgare 6-SFT GB:JQ411254.1
Triticum aestivum 6-SFT GB:FJ228688.1
Lolium perenne 6-SFT GAYX01016306
Hordeum vulgare AK374122
Hordeum vulgare MLOC 37242.1
Lolium perenne GAYX01057580
Hordeum vulgare TR38649 c0 g1
Hordeum vulgare AK362010
Triticum monococcum vaINV-like GB:AY575717.1
Lolium perenne vaINV-like GB:AF481763.1
Melica nutans TR84645 c5 g3
Brachypodium distachyon Bradi1g52210.1
Brachypodium distachyon TR36178 c0 g1
Brachypodium distachyon Bradi3g00910.1
Brachypodium distachyon TR27908 c0 g2
Hordeum vulgare TR39468 c0 g2
Hordeum vulgare AK361559
Hordeum vulgare TR39468 c1 g1
Hordeum vulgare MLOC 66445.1
Hordeum vulgare AK361887
Triticum aestivum vaINV3 GB:AJ635225
Hordeum vulgare MLOC 66722.1
Lolium perenne vaINV3 GB:AY082350.1
Nardus stricta TR100881 c1 g7
Oryza sativa Os02g0106100
Sorghum bicolor Sb04g000620.1
Zea mays GRMZM2G089836 P01
Zea mays GRMZM2G463871 P01
0.99
0.98
0.71
1
60
50
40
30
20
10
0
million years ago

Figure S4: Time calibrated phylogeny for the Pooideae FST gene family. *The phylogeny was estimated with BEAST v1.8.2 using a GTR+Γ and an uncorrelated, lognormal distributed molecular clock model. Significant posterior probabilities are shown as branch labels.*

Paper 3

Cross species comparative transcriptomics using co-expression networks

Lars Grønvold[1] and Torgeir R. Hvidsten[1,2]

[1]Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, NO-1432, Ås, Norway.
[2]Umeå Plant Science Centre, Department of Plant Physiology, Umeå University, SE-90187, Umeå, Sweden.

Introduction

Comparing protein sequence is very powerful and is used routinely when annotating newly sequenced genomes. This relies on the observation that similar/homologous sequences have the same or similar molecular functions which makes it possible to transfer the knowledge acquired from experiments on model species over to other species. However, the function of a gene is also determined by its transcriptional program. Changes in gene regulation may underlie much of the differences we observe between species (Chan *et al.* 2010), however, we know very little about the dynamics of gene regulatory evolution compared to what we know about sequence evolution. One way to learn more is to compare gene expression from several species.

Because of the dynamic nature of gene expression, direct comparison between species requires sampling from the same types of tissues at the same developmental stage and under the same conditions, something that can be difficult or impossible for distantly related species. To avoid this limitation, it is possible to use co-expression to indirectly compare the gene expression patterns between species (Tirosh *et al.* 2007). The principle is that if a gene has conserved regulation it will have retained the same co-expression partners. As this method does not require directly comparable samples, it can take advantage of the ever increasing amount of expression data available in databases.

Another prerequisite for comparing gene expression in different species is to identify the orthologous genes, i.e. genes in each species that descend from the same single gene in the most recent common ancestor. Because genes often are duplicated and/or lost during evolution, there will be orthologs that do not have a simple one-to-one relationship (1:1 orthologs), but instead have a one-to-many relationship (1:N), if the gene is duplicated in one of the species after speciation, or many-to-many relationship (N:N), if there are duplications in both species.

Gene-duplication is thought to be an important driver of the evolution of gene regulation. As most duplicated genes are lost in the long run, the ancient duplicates that are still retained must either have acquired a new function (neo-functionalization) or perform complementary sub-functions of the original gene function (sub-functionalization)(Ohno 1970). In terms of gene expression, sub-functionalization might imply that each copy is expressed in different tissues, thus allowing them to specialize, while neo-functionalization implies that one copy retains the ancestral expression pattern and the other copy acquires a novel expression pattern.

There are several different methods that compare co-expression across several species. Some of these methods mainly identify a cross-species consensus network or gene modules (Stuart *et al.* 2003; Oti *et al.* 2008; Mutwil *et al.* 2011; Zarrineh *et al.* 2011; Gerstein *et al.* 2014) while others specifically quantify the co-expression similarity between a pair of orthologous genes (Dutilh *et al.* 2006; Tirosh & Barkai 2007; Netotea *et al.* 2014; Monaco *et al.* 2015). There are two main approaches to comparing co-expression: overlap of co-expressed genes or correlation of co-expression values.

The overlap based methods first identifies a set of genes co-expressed with the gene of interest in each species. This co-expression set can either be the neighbors in an un-weighted co-expression network, typically generated by applying a threshold to the co-expression matrix, or can be based on clustering (e.g. Mutwil *et al.* 2011). The two sets of co-expressed genes, one set from each species, are then compared by first determining the number of orthologous genes they share and then by evaluating the statistical significance of this overlap (using e.g. the hypergeometric test). The methods also differ in how the orthologs are defined. Some methods analyze large gene families with ancient paralogs (e.g. Pfam), some use ortholog groups while others only consider 1:1 orthologs.

Correlation based methods have the advantage of being threshold-independent. They calculate the correlation between two co-expression vectors, one from each species, that contain the co-expression values of the compared orthologs to a set of 1:1 reference orthologs. The idea is that these reference orthologs are unduplicated genes performing the ancestral function and that they display conserved expression patterns. Compared to the overlap-methods, the correlation-based methods are relatively less studied, and all studies that use this method have relied on calculating the co-expression matrix using the Pearson correlation coefficient (PCC)(Dutilh *et al.* 2006; Tirosh & Barkai 2007; Wang *et al.* 2011). The overlap-based methods, on the other hand, also utilize alternative correlation measures, such as mutual information (MI), and often perform an additional normalization step, such as context likelihood of relatedness (CLR), highest reciprocal rank or mutual rank (MR) before applying a co-expression threshold.

In this study, we evaluate the correlation-based method by testing various methods for calculating the co-expression matrix. We use “co-expression correlation score” (CCS) to refer to the cross-species correlation value we obtain from comparing the co-expression patterns of two orthologs. To evaluate the performance of different methods, we apply a novel method that rank the score of the 1:1 orthologs among the scores obtained between one of the orthologs and all the genes in the other species. The idea is that the orthologous gene would tend to be the one with the most similar expression compared to all non-orthologous genes. This ortholog rank score (ORS) can

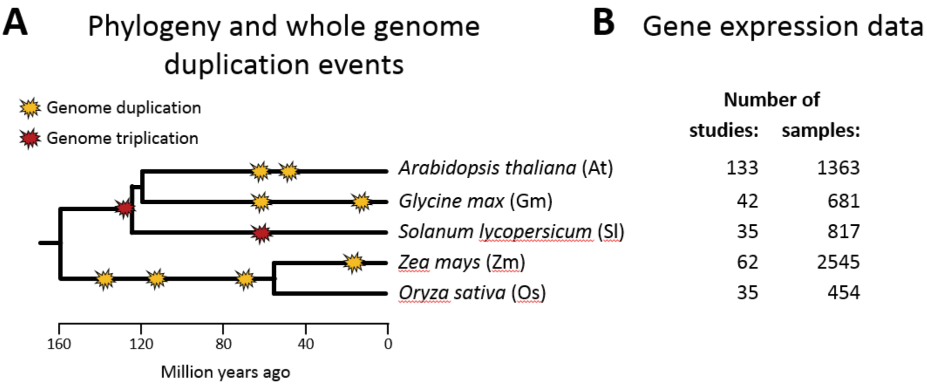


Figure 1: (A) Phylogeny of the included species and known whole genome duplication events (Vanneste et al. 2014). (B) Amount of public gene expression data used in this study.

also be considered a significance measure for the CCS and may be used as an indicator of conserved co-expression instead of using the CCS directly.

Using RNAseq data from five plant species, we identified Pearson correlation followed by mutual rank (PCC+MR) as the best method for comparing co-expression across species. We also investigated the effect that the number of samples has on the results and find that using samples from many diverse studies gives the best results. Although both the CCS and ORS measures are inevitably biased by the samples that are available for each species, we demonstrate that it can reliably detect trends for groups of genes as we can clearly observe a relationship between gene duplication and conservation of expression patterns.

Results

Method overview

Coexpression matrices were generated from public gene expression data for five plant species (Figure 1). For each pair of species, the co-expression correlation score (CCS) was then calculated between each pair of species by using 1:1 orthologs as a common reference (Figure 2A-D). This gives a measure of the co-expression similarity between orthologs i.e. to what degree orthologs are co-expressed with genes that are also orthologs.

As a measure of significance for the CCS values, we calculate the ortholog rank score (ORS) which indicate the fraction of all genes (not only orthologs) with a lower or equal CCS (Figure 2E). By taking $1 - ORS$, the resulting value can be interpreted as a P-value where the null hypothesis is that the CCS of the ortholog is no different than the CCS of a random gene and the alternative hypothesis is that it is higher. The ORS for orthologs tend to have a distribution heavily skewed towards 1, so we use a log transformed ORS (logORS) where a value >1 means the ortholog is among the top 10%, >2 means top 1%, >3 top 0.1% and 4 is the highest score.

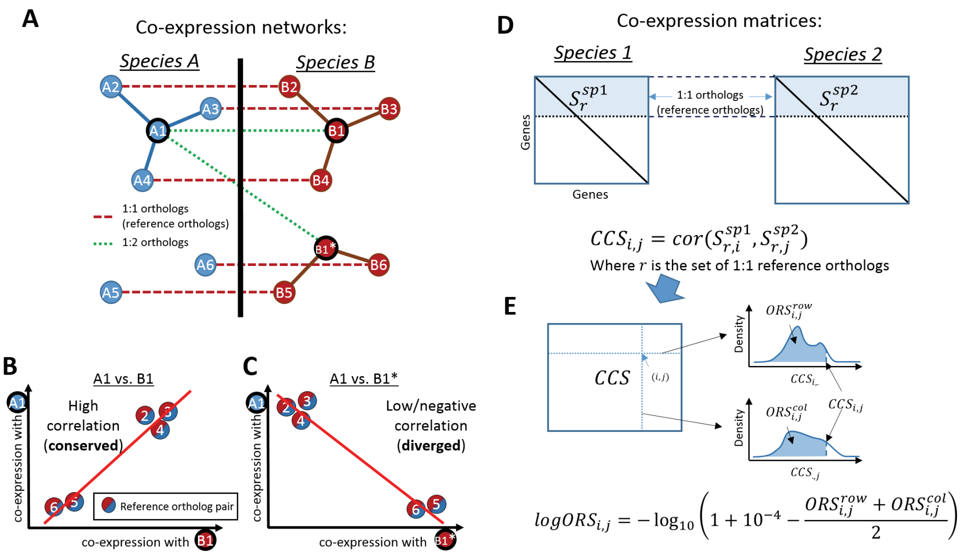


Figure 2: Co-expression correlation score (CCS) and ortholog rank score (ORS). (A) Cartoon example of two co-expression networks in two species aligned by their orthologs. (B) Cartoon example of how CCS is calculated for the ortholog pair A1 and B1. Correlation between co-expression with the reference orthologs in the respective species is high, indicating that the ortholog pair has a conserved gene expression pattern. (C) For the ortholog pair A1 and B1 the correlation of co-expression is negative as the B1* gene is co-expressed with a completely different set of genes, i.e. the expression pattern has diverged. (D) The rows of the co-expression matrices for two species are aligned by their 1:1 reference orthologs. Only these rows are used when the CCS is calculated by pearson correlation between all combinations of columns in the two co-expression matrices. (E) The ORS for an ortholog pair i and j is the proportion of values in the corresponding row or column in the CCS matrix which is lower or equal to the CCS of the ortholog pair itself, here illustrated as the area under the density curve. The mean of the column and row ORS is subtracted from 1.0001 and log transformed to get the logORS which is more suitable for visualization.*

Testing alternative co-expression methods

There are many approaches to calculating co-expression network, and here we evaluated several of the most common methods. We tested three correlation measures, Pearson correlation (PCC), Spearman correlation (SCC) and mutual information (MI), as well as two methods for normalizing the correlations, mutual rank (MR) or context likelihood of relatedness (CLR). These methods were applied to all pairs of species and evaluated using the median logORS of 1:1 orthologs (Figure 3).

With the exception of the *Gm-Os* and *Gm-Zm* comparisons, the PCC+MR method performed best and was therefore used in all subsequent analyses. Among the correlation methods, SCC performed slightly worse than PCC, while MI performed differently depending on the species

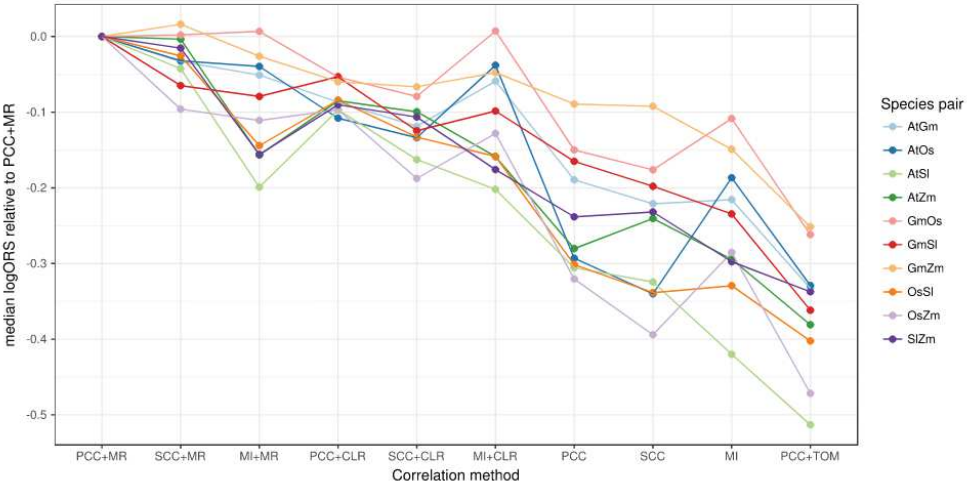


Figure 3: Evaluation of different methods for calculating co-expression matrices. The median logORS of all 1:1 orthologs between all pairs of species for each co-expression measure relative to the median logORS for PCC+MR.

pair. MR performed better than CLR, and both perform clearly better than using the correlation matrix directly. We also tested the topological overlap measure (TOM) from the commonly used WGCNA R package, which seemed to not be suitable for calculating CCS.

More samples gives higher ORS

The number of samples needed for robust inference of co-expression networks, and hence robust estimates of the conservation of networks in network comparison applications, is still debated. To test the effect of the number of samples on the conservation score, we picked subsets of various sizes from the 1363 *At* samples available and calculated ORS (PCC+MR) against the full set of *Os* samples (454 samples). To save computation time, only 1:1 orthologs were included (~5k genes). Samples were selected either randomly among all samples (individual samples) or by selecting all samples from the same study before selecting samples from another study (studies).

When selecting studies, adding more samples will, with few exceptions, result in a higher median ORS (Figure 4). This shows that in general it is a good idea to include as many samples as possible. On the other hand, when selecting individual samples, the ORS scores seem to reach the maximum and stay levelled after around one tenth of the samples have been included. This indicates that the studies tend to contain redundant samples (replicates or similar conditions) and illustrates the importance of diversity in the sampled conditions in order to get good ORSs.

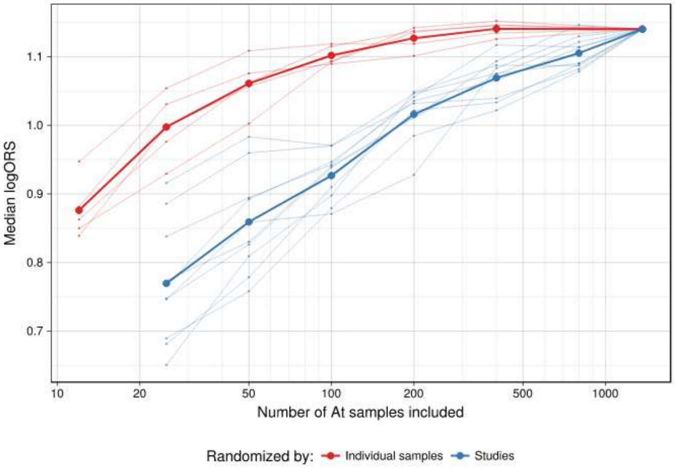


Figure 4: logORS between At and Os using various numbers of At samples. Each of the thin lines represent one selection of samples, where each point represents the same samples as the previous point plus addition samples. The thick lines show the mean of the permutations.

Sample subsets within species

The expression similarity between two species as measured by CCS or ORS will be affected by both the biological differences and by the bias caused by differences in the experimental conditions of the included samples (i.e. sampling bias) from each species. It is impossible to disentangle these factors when comparing different species, however, by comparing networks inferred from different sets of samples from within the same species it possible to study the effect of sampling bias alone (if we discount the biological differences within the sampled population which could contain mutants and variants).

The variation occurring from comparing different networks from the same species can also be considered to give rise to the background distribution for conserved gene expression (Meysman *et al.* 2013), i.e. the expected distribution of scores for an ortholog that has conserved expression pattern.

For each species, the samples were randomly split into two sets making sure that each half did not share any samples from the same study. The ORS was then calculated within species between networks inferred from the two halves of the samples. and this was repeated 10 times. When comparing networks within a species, any gene could be used as reference “orthologs”, however, to make these intra-species comparisons more similar to cross-species comparison, we used the genes which have 1:1 orthologs to the other species.

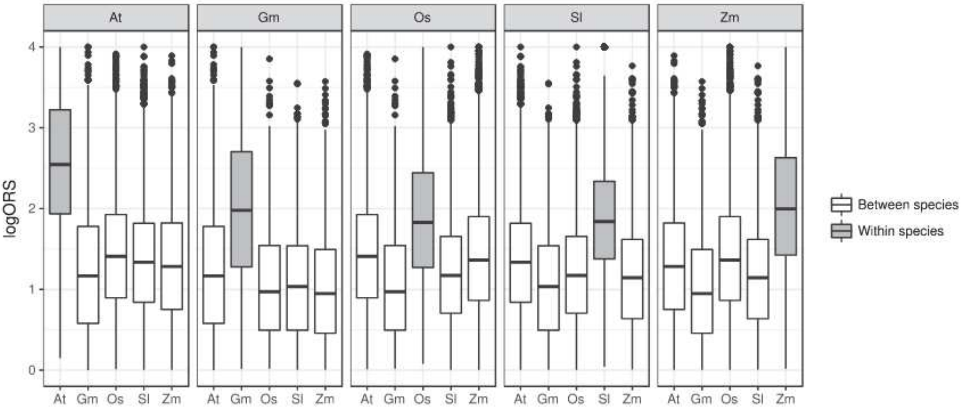


Figure 5: Distribution of ORSs of 1:1 orthologs between all pairs of species and between subsets of samples within species. Note that the results for each species pair is displayed twice for easier comparison.

The resulting ORS distribution within species are consistently higher than the ORS between species (Figure 5). Note that the between species ORS only partially reflect the phylogeny, e.g. the highest median ORS is between distantly related At and Os. One reason for this could be sample bias, e.g. the high within-species ORS for At is reflected in generally higher than expected ORS between any species and At. Another reason that the between-species ORS vary could be that the set of 1:1 orthologs also varies. This could be the reason for the low ORS for Gm, as it has relatively few 1:1 orthologs because of the recent whole genome duplication (WGD).

Expression divergence in duplicated genes

It has been shown before that duplicated genes have higher rate of expression divergence than single-copy genes (Gu *et al.* 2004; Huminiecki & Wolfe 2004; Assis & Bachtrog 2015). We therefore expect orthologs with duplications to have lower expression similarity than 1:1 orthologs. Indeed, when comparing the median logORS of 1:1 orthologs with one-to-many orthologs (i.e. 1:2, 1:3 and 1:4), there is a clear downwards trend as the number of duplicates increase (Figure 6).

An interesting exception is observed in *Gm* where the duplicates (i.e. 1:2 orthologs) seem to have an equally conserved expression pattern as the 1:1 orthologs. A possible explanation is that most of the duplicates in *Gm* originate from the relatively recent whole genome duplication (WGD) about 13 million years ago (Schmutz *et al.* 2010) and as such has not had enough time for gene expression to diverge. Note that, although much less distinct, a similar trend can be observed in *Zm*, which also experienced a WGD in about the same time frame. For comparison, the three other species haven’t experienced a WGD event in about 50-70 million years (Figure 1).

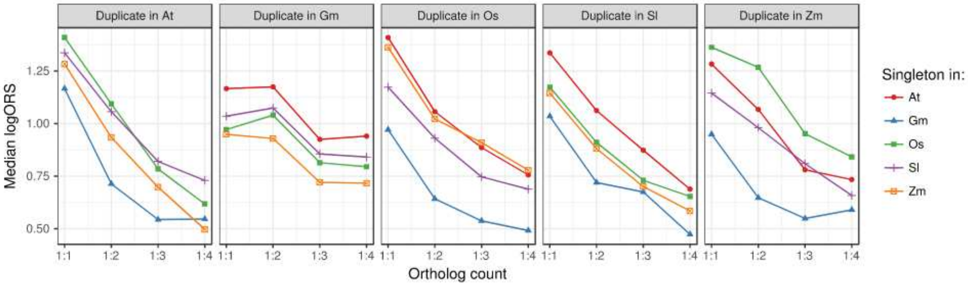


Figure 6: Median expression conservation of orthologs with 1:1-4 relationship for all pairs of species.

Even considering that a majority of the duplicates in *Gm* are recent, it would still be expected for the singleton genes to have a more conserved expression pattern. Since there is almost no difference between the ORS of the singletons and duplicates in *Gm*, there must be an additional mechanism at work. To explain the high ORS of *Gm* duplicates we hypothesize that there has been a preferential retention of genes that tend to have high ORS. To test this, we use the ORS distribution of 1:1 orthologs between two other species (*At* and *Os*) and compare the subset for which the corresponding Gm orthologs are duplicates (At1:Gm2) with the subset for which the corresponding Gm ortholog is a singleton (At1:Gm1). As predicted, the ORS tend to be higher (Wilcox rank-sum test, P=3.2e-7) when the Gm orthologs are duplicates (Figure 7).

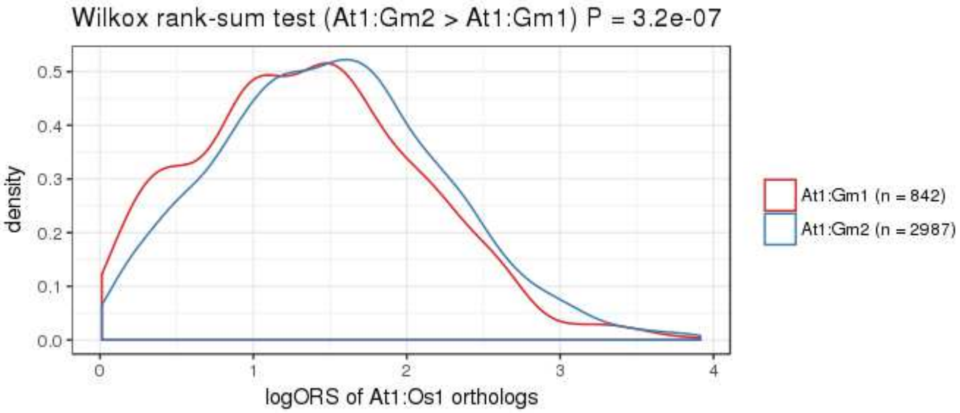


Figure 7: Expression similarity between At1:Os1 orthologs tend to be higher if the corresponding Gm ortholog is duplicate than if it is singleton. The density of logORS for two subsets of 1:1 orthologs between At and Os (At1:Os1). The red line denotes the subset where the At ortholog has a 1:1 relationship with a Gm ortholog, i.e. no duplicates in Gm. The blue line denotes the subset where the At ortholog has a 1:2 relationship with Gm, i.e.duplicates in Gm.

Discussion

Regulation of transcription is an essential function in all organisms and it is a central goal of biology to make sense of regulatory mechanisms and pathways. However, there is a lack of knowledge regarding how regulation differ between species and the evolutionary dynamics of regulatory pathways. Better methods for cross-species comparative transcriptomics would help increase our understanding of gene regulation and how it evolves. Using co-expression to indirectly compare gene transcription between species is a compelling method as it makes it possible to take advantage of the vast amount of expression data collected in public databases.

In this study we investigated several variations of co-expression measures as a basis for cross-species comparison and applied a novel scoring scheme that we termed ortholog rank score (ORS) to determine the best method. We find that by applying a normalization procedure to the co-expression matrix, such as mutual rank (MR), there is a clear improvement (figure 3) over using correlation only.

While it is not obvious why normalization of the co-expression matrix improves the cross species comparison, one possible explanation is that it reduces the bias towards large clusters in the co-expression network. For example in plants, there is a significant number of genes that are mainly expressed in tissues performing photosynthesis. Because plant expression datasets usually include both leaf samples and non-photosynthetic tissues, these genes tend to have highly correlated expression profiles (related discussion in Kinoshita & Obayashi 2009). When calculating CCS, the genes from larger co-expression clusters will get a high score because they are supported by a large number of co-expressed reference orthologs. The CCS would therefore be biased by co-expression cluster size. Both MR and CLR considers the correlation between two genes relative to the correlation with all other genes, thereby reducing this bias.

When comparing ORS of 1:1 orthologs with 1:N orthologs we found that, as expected, there is a clear tendency for duplicated genes to have more diverged expression (Figure 6). However, soybean (*Gm*) which has undergone a whole genome duplication (WGD) about 13 million years ago does not show any difference in ORS between the duplicate (1:2) and singleton (1:1) orthologs. Duplicated genes are thought to experience relaxed selection pressure that allows their expression regulation to diverge faster (Ohno 1970). The singleton genes must have lost their duplicates at some point after the WGD event. If loss of duplicates occurred randomly then a possible explanation for the observed pattern could be that a majority of the lost duplicates were lost recently, and they are consequently indistinguishable from duplicates. On the other hand, it could be that some genes are more likely to be retained than others. Supporting this, we found that the orthologs of the retained *Gm* duplicates tend to have higher ORS when comparing between other species (figure 7). However, there must have been a selective advantage to retain the genes that usually tend to have high ORS. A plausible explanation is the dosage balance hypothesis, which states that genes coding proteins that act in protein complexes, tend to be sensitive to change in relative dosage between individual subunits (Papp *et al.* 2003). The hypothesis predicts that genes sensitive to dosage balance are preferentially retained after WGDs and preferentially lost after small scale duplications. This has been suggested as an explanation for why certain categories of genes, such as transcription factors and kinases, are preferentially retained after WGDs in *Arabidopsis* (Blanc & Wolfe; Maere *et al.* 2005).

The use of ORS as a measure of performance facilitates the evaluation of different variations of methods and it is likely that there are untried alternatives that outperform PCC+MR. For example, one reason why mutual information (MI) didn't perform so well could have been that it doesn't differentiate between positive and negative correlations, which could be resolved by combining MI with PCC to generate a signed MI. There might also be something to gain by applying a transformation function, such as the soft-threshold used in WGCNA (Langfelder & Horvath 2008), to put more weight on the strong correlations. There are also alternatives to using correlation when comparing the co-expression matrices across species, such as the topological overlap measure. As it is only the imagination that limits the alternatives, future work should aim to get some theoretical understanding of why one method outperforms another.

In other studies that generate/compare co-expression networks from mixed public expression data, a sample filtering step (Mutwil *et al.* 2011) or weighting scheme (Obayashi & Kinoshita 2009) is often included. The purpose of this is to reduce bias caused by the redundancy from samples with similar expression profiles. We did not include such a step in this analysis but it is something that should be considered in future work.

The median ORS between two species can be considered to be an indicator of the total divergence in expression between them. It is expected that the expression divergence between species corresponds to the phylogenetic distance. However the observed median ORS does not reflect the phylogeny. For example, we see that the dicot *At* and monocot *Os* have a median ORS that is higher than between *At* and other dicots (Figure 5). There are several factors that could bias the ORS, such as the number of samples, how many different conditions that are sampled and to what extent the same conditions are sampled in the compared species. These biases can complicate comparisons that include more than two species. It might however be possible to compensate for some of these effects. The number of samples seems to have a rather predictable effect on the ORS (Figure 4) - and can thus be accounted for. The within-species ORS (Figure 5) might also be used as a normalization factor. However, neither would account for the sample condition compatibility between pairs of species.

Co-expression based methods can leverage the diverse data gathered in public repositories to quantify similarity of gene expression patterns across species. With improved methods such as presented in this study, combined with the ever increasing amount of available data, we believe it will be possible to investigate hypotheses about the evolution of gene regulation that previously was out of reach.

Method

Gene orthology

Ortholog information was downloaded from Ensembl plants using biomart. These are based on EnsemblCompara gene trees (Vilella *et al.* 2009). A pair of genes are defined as orthologs if the ancestor node in the gene tree is a speciation event.

Expression data

FPKM expression values was downloaded from the PODC website (http://plantomics.mind.meiji.ac.jp/podc/, Ohyanagi *et al.* 2015). Isoform level expression values were converted to per gene expression by summing the FPKMs of all isoforms for each gene. The Gm and Zm expression data at PODC were mapped to a different reference genome then we used for orthology information (Ensembl plants). The corresponding samples were therefore downloaded from EBI using their RNAseq-er API (http://www.ebi.ac.uk/fg/rnaseq/api/). FPKM values were log transformed using log2(1+FPKM).

PCC+MR similarity matrix

Co-expression for each gene pair of each species was first calculated using Pearson correlation between gene expression vectors.

$$S_{i,j}^{PCC} = PCC(E_i, E_j)$$

where E_i is the expression value vector of gene i for all samples and PCC is the Pearson correlation coefficient function. The similarity matrix was then transformed using the log mutual rank:

$$S_{i,j}^{PCC+MR} = 1 - \frac{log\left(\sqrt{R_{i,j}R_{j,i}}\right)}{log(n)}$$

where n is the number of genes and $R_{i,j}$is the rank of $S_{i,j}^{PCC}$in row i of the similarity matrix ordered from highest to lowest value. Or in other words, if all genes were sorted from highest to lowest co-expression value with gene i, then $R_{i,j}$would be the position of gene j. Note that ATTED-II uses $\sqrt{R_{i,j}R_{j,i}}$ as the mutual rank measure (Obayashi *et al.* 2009). The log transformation puts a higher weight on the most similar genes. Subtracting from 1 and dividing by $log(n)$ scales the value to a range between 0 and 1 but has no effect on the downstream analysis.

Co-expression correlation score (CCS)

For two species, the CCS between gene i in the first species and gene j in the second species is an indirect measure of expression similarity calculated by taking the Pearson correlation between the corresponding two columns of the co-expression similarity matrices. As each species has a different set of genes, only the rows with the corresponding one-to-one orthologs are correlated:

$$CCS_{i,j} = PCC(S_{r,i}^{sp1}, S_{r,j}^{sp2})$$

where S^{sp1} and S^{sp2} are the S^{PCC+MR} similarity matrices for the two species (*sp1* and *sp2*) and r are all the one-two-one ortholog pairs in the two species except pairs containing genes i or j.

Ortholog rank score (ORS)

The ortholog rank score (ORS) is derived from the CCS and can be viewed as a measure of significance of the CCS or as an alternative to CCS as a co-expression similarity measure. $ORS_{i,j}$(ORS between gene i in the first species and gene j in a second species) is calculated as the proportion of all genes in the second species which has a CCS with gene i that is lower or equal to that of gene j. In other words, $ORS_{i,j}$is the normalised rank of the $CCS_{i,j}$in row i of the CCS matrix ordered from lowest to highest value. For example, if $ORS_{i,j} = 1$then $CCS_{i,j} \geq CCS_{i,g}$ for any gene $g$in the second species, or, if $ORS_{i,j} = 0.90$ then $CCS_{i,j} \geq CCS_{i,g}$for 90% of the genes in the second species. By taking one minus the ORS, the resulting value can be interpreted as an empirical P-value (i.e. $P = 1 - ORS$) for the hypothesis that the ortholog pair i, jhas a more similar expression than expected by chance. Because the ORS is directional, i.e. $ORS_{i,j} \neq ORS_{j,i}$, we calculate an undirected ORS by taking the mean of $ORS_{i,j}$ and $ORS_{j,i}$. Furthermore, as the distribution of ORS for ortholog pairs is skewed towards 1 we use a logarithmic transformation when plotting to make comparisons easier:

$$logORS_{i,j} = -log_{10}\left(1 + 10^{-4} - \frac{ORS_{i,j} + ORS_{j,i}}{2}\right)$$

Adding the value 10^{-4} before log transforming ensures that the score gets a value between 0 and 4.

Alternative co-expression methods

Mutual information (MI) was calculated with B-Spline smoothed bins (Daub *et al.* 2004). Number of bins was set to 7 and spline order to 3.

Context likelihood of relatedness (CLR) is a background correction step that aims to remove random and indirect correlation (Faith *et al.* 2007). It involves calculating the Z-score for each row in the similarity matrix, i.e subtract the mean and divide by the standard deviation. Only positive Z-scores were used, while negative values were replaced with 0 as in (Netotea *et al.*; Madar *et al.* 2010):

$$z_{i,j} = max\left\{0, \frac{S_{i,j} - \underline{S_{i,\cdot}}}{\sigma_i}\right\}$$

Where $\sigma_i$is the standard deviation of row i in the co-expression matrix S, and $\underline{S_{i,\cdot}}$is the mean of row i. The CLR is then calculated by combining the row-wise and column-wise Z-scores:

$$CLR_{i,j} = \sqrt{{z_{i,j}}^2 + {z_{j,i}}^2}$$

We also tested the topological overlap measure (TOM) using the TOMsimilarity function in the WGCNA R package (Langfelder & Horvath 2008) after applying soft-threshold exponent of 6 to the PCC co-expression matrix.

References

Assis R, Bachtrog D (2015) Rapid divergence and diversification of mammalian duplicate gene functions. *BMC Evolutionary Biology*, **15**, 138.

Blanc WG, Wolfe KH Functional Divergence of Duplicated Genes Formed by Polyploidy during Arabidopsis Evolution.

Chan YF, Marks ME, Jones FC *et al.* (2010) Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science (New York, N.Y.)*, **327**, 302–5.

Daub CO, Steuer R, Selbig J, Kloska S (2004) Estimating mutual information using B-spline functions--an improved similarity measure for analysing gene expression data. *BMC bioinformatics*, **5**, 118.

Dutilh B, Huynen M, Snel B (2006) A global definition of expression context is conserved between orthologs, but does not correlate with sequence conservation. *BMC Genomics*, **7**, 10.

Faith JJ, Hayete B, Thaden JT *et al.* (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS biology*, **5**, e8.

Gerstein MB, Rozowsky J, Yan K-K *et al.* (2014) Comparative analysis of the transcriptome across distant species. *Nature*, **512**, 445–448.

Gu Z, Rifkin SA, White KP, Li W-H (2004) Duplicate genes increase gene expression diversity within and between species. *Nature Genetics*, **36**, 577–579.

Huminiecki L, Wolfe KH (2004) Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome research*, **14**, 1870–9.

Kinoshita K, Obayashi T (2009) Multi-dimensional correlations for gene coexpression and application to the large-scale data of Arabidopsis. *Bioinformatics (Oxford, England)*, **25**, 2677–84.

Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, **9**, 559.

Madar A, Greenfield A, Vanden-Eijnden E, Bonneau R (2010) DREAM3: network inference using dynamic context likelihood of relatedness and the inferelator. *PloS one*, **5**, e9803.

Maere S, De Bodt S, Raes J *et al.* (2005) Modeling gene and genome duplications in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 5454–9.

Meysman P, Sánchez-Rodríguez A, Fu Q, Marchal K, Engelen K (2013) Expression Divergence between Escherichia coli and Salmonella enterica serovar Typhimurium Reflects Their Lifestyles. *Molecular Biology and Evolution*, **30**, 1302–1314.

Monaco G, van Dam S, Casal Novo Ribeiro JL, Larbi A, de Magalhães JP (2015) A comparison of human and mouse gene co-expression networks reveals conservation and divergence at the tissue, pathway and disease levels. *BMC evolutionary biology*, **15**, 259.

Mutwil M, Klie S, Tohge T *et al.* (2011) PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. *The Plant cell*, **23**, 895–910.

Netotea S, Sundell D, Street NR, Hvidsten TR ComPlEx : Conservation and divergence of co-expression networks in A. thaliana, Populus and O. sativa.

Netotea S, Sundell D, Street NR, Hvidsten TR (2014) ComPlEx: conservation and divergence of co-expression networks in A. thaliana, Populus and O. sativa. *BMC Genomics*, **15**, 106.

Obayashi T, Hayashi S, Saeki M, Ohta H, Kinoshita K (2009) ATTED-II provides coexpressed gene networks for Arabidopsis. *Nucleic acids research*, **37**, D987-91.

Obayashi T, Kinoshita K (2009) Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression. *DNA research : an international journal for rapid publication of reports on genes and genomes*, **16**, 249–60.

Ohno S (1970) *Evolution by Gene Duplication*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Ohyanagi H, Takano T, Terashima S *et al.* (2015) Plant Omics Data Center: An Integrated Web Repository for Interspecies Gene Expression Networks with NLP-Based Curation. *Plant and Cell Physiology*, **56**, e9–e9.

Oti M, van Reeuwijk J, Huynen MA, Brunner HG (2008) Conserved co-expression for candidate disease gene prioritization. *BMC bioinformatics*, **9**, 208.

Papp B, Pál C, Hurst LD (2003) Dosage sensitivity and the evolution of gene families in yeast. *Nature*, **424**, 194–197.

Schmutz J, Cannon SB, Schlueter J *et al.* (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.

Stuart JM, Segal E, Koller D, Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science (New York, N.Y.)*, **302**, 249–55.

Tirosh I, Barkai N (2007) Comparative analysis indicates regulatory neofunctionalization of yeast duplicates. *Genome Biology*, **8**, R50.

Tirosh I, Bilu Y, Barkai N (2007) Comparative biology: beyond sequence analysis. *Current Opinion in Biotechnology*, **18**, 371–377.

Vanneste K, Maere S, Van de Peer Y (2014) Tangled up in two: a burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **369**, 20130353.

Vilella AJ, Severin J, Ureta-Vidal A *et al.* (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome research*, **19**, 327–35.

Wang Y, Wang X, Tang H *et al.* (2011) Modes of Gene Duplication Contribute Differently to Genetic Novelty and Redundancy, but Show Parallels across Divergent Angiosperms (SR Proulx, Ed,). *PLoS ONE*, **6**, e28150.

Zarrineh P, Fierro AC, Sánchez-Rodríguez A *et al.* (2011) COMODO: an adaptive coclustering strategy to identify conserved coexpression modules between organisms. *Nucleic Acids Research*, **39**, e41–e41.