Received: 17 April 2024 | Revised: 8 August 2024 | Accepted: 14 August 2024

DOI: 10.1111/ppl.14537

RESOURCE ARTICLE

Physiologia Plantarum

# A resource of identified and annotated lincRNAs expressed during somatic embryogenesis development in Norway spruce

Camilla Canovi[1] | Katja Stojkovič[2] | Aarón Ayllón Benítez[1] | Nicolas Delhomme[2] | Ulrika Egertsdotter[2,3] | Nathaniel R. Street[1,4]

[1]Umeå Plant Science Centre, Department of Plant Physiology, Umeå University, Umeå, Sweden

[2]Umeå Plant Science Centre, Department of Forest Genetics and Plant Physiology, Swedish University of Agricultural Sciences, Umeå, Sweden

[3]Renewable Bioproducts Institute, Georgia Institute of Technology Atlanta, USA

[4]SciLifeLab, Umeå University, Umeå, Sweden

**Correspondence**
Nathaniel R. Street,
Email: nathaniel.street@umu.se

**Present address**
Aarón Ayllón Benítez, BASF Digital Solution S.L, Madrid, Spain.

**Funding information**
Trees and Crops for the Future; Kempestiftelserna, Grant/Award Number: SMK1340

**Edited by** A.-J. van Dijk

## Abstract

Long non-coding RNAs (lncRNAs) have emerged as important regulators of many biological processes, although their regulatory roles remain poorly characterized in woody plants, especially in gymnosperms. A major challenge of working with lncRNAs is to assign functional annotations, since they have a low coding potential and low cross-species conservation.

We utilised an existing RNA-Sequencing resource and performed short RNA sequencing of somatic embryogenesis developmental stages in Norway spruce (*Picea abies* L. Karst). We implemented a pipeline to identify lncRNAs located within the intergenic space (lincRNAs) and generated a co-expression network including protein coding, lincRNA and miRNA genes.

To assign putative functional annotation, we employed a guilt-by-association approach using the co-expression network and integrated these results with annotation assigned using semantic similarity and co-expression. Moreover, we evaluated the relationship between lincRNAs and miRNAs, and identified which lincRNAs are conserved in other species. We identified lincRNAs with clear evidence of differential expression during somatic embryogenesis and used network connectivity to identify those with the greatest regulatory potential.

This work provides the most comprehensive view of lincRNAs in Norway spruce and is the first study to perform global identification of lincRNAs during somatic embryogenesis in conifers. The data have been integrated into the expression visualisation tools at the PlantGenIE.org web resource to enable easy access to the community. This will facilitate the use of the data to address novel questions about the role of lincRNAs in the regulation of embryogenesis and facilitate future comparative genomics studies.

## 1 | INTRODUCTION

Widespread adoption of RNA-Sequencing (RNA-Seq) and the ability to generate sequencing libraries at deep coverage for increasingly lower costs has revolutionised understanding of RNA in the past decade (Stark et al., 2019; Szakonyi et al., 2019; Tang and Tang, 2019; Zhao et al., 2019; Micheel et al., 2021; Rich-Griffin et al., 2019). This has included the discovery of previously unknown

classes of RNA as well as an appreciation of how much more diverse existing, known classes of RNA can be (Santer et al., 2019; Szakonyi and Duque, 2018; Bedre et al., 2019; Zhang et al., 2020; Micheel et al., 2021). Among the newly defined classes of RNA are various categories of long non-coding RNA (lncRNAs), which can either overlap with protein-coding gene regions in sense or antisense or can be entirely intergenic (lincRNAs). The suite of known lncRNAs now includes those derived from introns (ilncRNAs) or natural antisense transcripts (NATS), which can partially or completely span the protein-coding region of genes, and circular non-coding RNAs (ciRNAs).

It is common for an RNA-Seq analysis to identify hundreds to many thousands of putative lncRNAs. A major challenge in genomics is, therefore, to determine the signal from noise and to assign biological function to the genuine lncRNA transcripts. Unlike protein-coding genes, where the presence of defined protein domains or similarity to previously annotated proteins can indicate function, there is currently no such information for lncRNAs, rendering them an enigmatic class of RNA. Indeed, lncRNAs tend to have low cross-species homology and conservation (Deng et al., 2018; Ma et al., 2019; Jha et al., 2020) with little selective constraint and weak selection (positive or negative/purifying)(Palazzo and Koonin, 2020), suggesting that they are often evolutionarily young, representing a highly dynamic pool of transcripts with a diverse range of regulatory potential. These characteristics, alongside the failure to identify phenotypic effects resulting from disrupted expression for many lncRNAs (Wierzbicki et al., 2021), has raised questions about their biological significance (Palazzo and Koonin, 2020) or the validity of many reported lncRNAs (Lee et al., 2019). These challenges make assigning functional descriptions to lncRNAs challenging, particularly as analyses commonly identify large numbers of putative lncRNAs, making it hard to prioritise which putative lncRNA genes should be functionally characterised. Despite these challenges, there are now well-validated examples of functionally important lncRNAs, including *COOLAIR* (Swiezewski et al., 2009) and *COLDAIR*, which both affect flowering time via their effects on the expression of *Flowering Locus C* (*FLC*) in *Arabidopsis thaliana*. There are also validated examples where the transcription of a lncRNA (rather than the specific sequence of that transcript) represents the functional mechanism via *cis*-acting effects on the transcription of a proximal protein coding gene (Kindgren et al., 2018; Ali and Grote, 2020), a functional mechanism that inherently results in low selective constraint of the RNA transcript sequence but positional constraint for transcriptional activity. As such, functional validation of the biological mode of action for lncRNAs can be challenging and can require non-conventional approaches rather than relying on naïve knock-down/out or over-expression assays (Wang and Chekanova, 2017; Wierzbicki et al., 2021).

To date, several genetic and epigenetic regulatory mechanisms have been assigned to lncRNAs, including *cis* effects on transcription of proximal genes or via a range of *trans*-acting effects, such as functioning as micro RNA (miRNA) precursors, as endogenous target mimics (eTM) of miRNAs (also referred to as sponges), inducing RNA directed DNA Methylation (RdDM) and directing histone modifications (Liu et al., 2015; Wang and Chekanova, 2017; Lucero et al., 2020; Chen et al., 2020). Some lncRNAs can act to form RNA-protein scaffolds, which can either interact with a DNA element to guide the RNA-protein complex to target loci or can act independently of DNA to form macromolecular complexes (Chen et al., 2020). While the majority of lncRNAs are transcribed by RNA polymerase II (Pol II), those involved in the canonical RdDM pathway are transcribed by Pol IV and V, functioning to produce 24 nt short RNAs (sRNAs) and to facilitate DNA methylation (Pol V) directed by the 24 nt sRNAs. Although lncRNAs are not typically conserved across species, there are consistent characteristics, including lower average expression levels than protein-coding genes and often highly specific expression domains (Wang and Chekanova, 2017; Budak et al., 2020; Sang et al., 2021). There has been less consideration of whether the presence of a lncRNA at a certain physical position of a genome, for example, in relation to a protein-coding gene, or within a co-expression network, for example, co-expression with a miRNA, is more consistently conserved even when sequence homology and conservation is lacking.

LncRNAs have been identified in a range of sample types and conditions, including during developmental processes (Amor et al., 2009; Ariel et al., 2014; Jiang et al., 2019; Yan et al., 2020), abiotic and biotic stress responses (Qin et al., 2017; Seo et al., 2017; Bazin and Bailey-Serres, 2015; Hou et al., 2020; Zamora-Ballesteros et al., 2022; Ma et al., 2019; Jha et al., 2020) from a diverse range of annual herbaceous plants through to woody tree angiosperms (Tu et al., 2021; Xiao et al., 2020; Lemos et al., 2020; Patturaj et al., 2022; Yan et al., 2020) and gymnosperms (Nystedt et al., 2013; Wang et al., 2018; Jiang et al., 2019; Wu et al., 2019; Zamora-Ballesteros et al., 2022). However, exceedingly few reported putative lncRNAs have been functionally validated, and few studies have attempted to assign functional descriptions to identified putative lncRNAs. There remains extensive uncertainty as to how many reported lncRNAs represent transcriptional noise with no accepted method or established best practice for defining a *bona fide* lncRNA. One useful and commonly applied method of assigning functional descriptors to unannotated protein-coding genes is the use of guilt-by-association evidence derived from co-expression networks (Depuydt and Vandepoele, 2021). This approach can be used to associate genes with functional descriptions that are common to a set of co-expressed protein-coding genes and to assign lncRNAs within biological categorisations such as Gene Ontology. While a number of studies have taken this guilt-by-association approach to assign tentative descriptions of function to protein-coding genes, it has infrequently been applied to lncRNAs, even when co-expression to protein-coding genes has been analysed.

While an ever-increasing number of studies have catalogued lncRNAs in angiosperm species, there remain few studies of their expression in gymnosperms. Nystedt et al. (2013) reported the presence of a large pool of putative Pol II derived lncRNAs in Norway spruce (*Picea abies*), identified using an RNA-Seq expression atlas comprising different tissue types across a range of seasonal time

points or developmental stages. However, the results must be viewed with caution due to the highly fragmented nature of the genome assembly, which can substantially inflate estimated gene numbers. Given the extremely high content of transposable elements (TEs) in conifer genomes and the tendency for TEs to contain transcription factor binding sites and transcription start sites, it can be expected that many non-coding transcripts may be present (Palazzo and Koonin, 2020). While many of these may represent promiscuous transcription, there is still great potential that a subset will have acquired function. Indeed, an open question in conifer and gymnosperm genomics is whether a greater proportion of the genome has regulatory potential than typical angiosperm genomes and whether coniferous species contain a greater number of functional non-coding RNAs as a result of TE activity. We were therefore interested in identifying lincRNAs in Norway spruce and employing a guilt-by-association approach to assign putative functional descriptions. We reasoned that a recently described RNA-Seq data resource profiling transcript expression during somatic embryogenesis (SE) would be ideal for this due to the extensive remodelling of the transcriptome that was shown to occur during the SE process. SE in conifers results in multiplication of the early-stage embryos for extended periods, generating an effectively unlimited number of clonally identical somatic embryos. This clonal multiplication makes SE a valuable tool for the forestry industry to obtain clonally propagated trees from elite varieties (Egertsdotter et al., 2019) and a powerful model system for studying the process of embryogenesis.

## 2 | MATERIALS AND METHODS

### 2.1 | Transcriptome sequencing and data pre-processing

RNA isolated from the eight developmental stages during somatic embryogenesis (SE) was sequenced on Illumina HiSeq 2500 at Science for Life Laboratory, Sweden (SciLifeLab) using 2x126bp paired-end reads to an average read number of 24.1 ± 2.1 M reads per sample. Three biological replicates were provided for each developmental stage. Full experimental details are available in Stojkovič et al. (2024), and the data is available at the European Nucleotide Archive (ENA) as accession PRJEB72619.

The quality of the raw sequence data was assessed using FastQC (v. 11.4; Andrews, 2012). Sequence reads originating from ribosomal RNAs (rRNA) were identified and removed using SortMeRNA (v. 2.1b;Kopylova et al., 2012; settings--log--paired_in--fastx--sam--num_alignments 1) using the rRNA sequences provided with SortMeRNA (rfam-5 s-database-id98.fasta, rfam-5.8 s-database-id98.fasta, silva-arc-16 s-database-id95.fasta, silva-bac-16 s-database-id85.fasta, silva-euk-18 s-database-id95.fasta, silva-arc-23 s-database-id98.fasta, silva-bac-23 s-database-id98.fasta and silva-euk-28 s-database-id98.fasta). Data were then filtered to remove adapters and trimmed for quality using Trimmomatic (v. 0.46; Bolger et al., 2014; settings TruSeq3-PE-2.fa:2:30:10 SLIDINGWINDOW:5:20 MINLEN:50). After both filtering steps, FastQC was run again to ensure that no technical artefacts were introduced. Read counts were obtained using Salmon (v. 0.11.2; Patro et al., 2017).

### 2.2 | Identification of long intergenic non-coding RNAs (lincRNAs)

We implemented a pipeline to identify putative lincRNAs on the pre-processed data, where default settings were used unless specified. We first *in silico* normalised the reads to reduce data redundancy and then reconstructed the transcriptome using the *de novo* assembler, Trinity (v. 2.8.3;Grabherr et al., 2011; Haas et al., 2013).

On the set of transcripts assembled by Trinity, we ran the following programs, which are detailed below: TransRate (v. 1.0.3; Smith-Unna et al., 2016), Detonate (v. 1.8.1; Li, Fillmore, et al., 2014), TransDecoder (version 2.8.3; https://github.com/TransDecoder/TransDecoder/wiki; Haas et al., 2013), GMAP (Genomic Mapping and Alignment Program; v. 2020-11-15; settings-i 70000; Wu & Watanabe, 2005), Salmon Index (v. 0.11.2; Patro et al., 2017), PLEK (***p***redictor of ***l***ong non-coding RNAs and m***e***ssenger RNAs based on an improved ***k***-mer scheme; v. 1.2; settings-minlength 200; Li, Zhang, et al., 2014), CNCI (Coding-Non-Coding Index; v. 2; Sun et al., 2013), CPC2 (Coding Potential Calculator version 2; v. 2.0 beta; settings-r TRUE; Kang et al., 2017), PLncPRO (**P**lant **L**ong **N**on-**C**oding RNA **P**rediction by **R**andom f**O**rest; v. 1.2; Singh et al., 2017) and BEDTools closest (v. 2.30.0; https://bedtools.readthedocs.io/en/latest/content/tools/closest.html; Quinlan & Hall, 2010).

To assess the quality of the *de-novo* transcript assembly, we used Detonate and TransRate. TransDecoder was run to evaluate transcript coding potential. GMAP was used to map the transcriptome to the genome reference. Salmon Index was used to subsequently run Salmon (Patro et al., 2017), a pseudoalignment tool used to map RNA-Seq reads to the assembled transcripts for expression quantification. PLEK, CNCI, CPC2, and PLncPRO were run to classify transcripts as coding or non-coding. GMAP results were pre-sorted by chromosome and start position after which the BEDTools ´closest´ option was used to exclude transcripts with any overlap to the reference annotation. Diamond (v. 0.9.40;Buchfink et al., 2014) was run using the TransDecoder results to check sequence similarities with protein-coding regions of other species. The last step in the analyses was to run a custom R script to find tissue-specific lncRNAs. This analysis generates a score of specificity ranging from 0 to 1, where 1 is specific and 0 is non-specific. We then filtered all those results, considering lincRNAs characteristics. We focused on transcripts that were identified as having no coding potential by the four classification programs (PLEK, CNCI, CPC2, and PLncPRO) and retained transcripts longer than 200 nt. We used the TransDecoder results and kept only transcripts identified as having no coding potential and having a distance >1000 nt from the nearest coding gene on the reference genome, which was identified using the BEDTools closest function. We removed transcripts with an expression value of NA in all stages of the SE process.

## 2.3 | Differential expression analysis of lincRNAs

Initially, the salmon abundance values of all the lncRNAs were imported into R (v. 4.3.1; R Core Team 2015) using the tximport package (v. 1.28.0;Soneson et al., 2015). Subsequently, only the lincRNAs previously identified by the pipeline were kept for further analysis. For the data quality assessment (QA) and visualisation, the read counts were normalised using zinbwave (v. 1.22.0;Risso et al., 2018). The biological relevance of the data (e.g. biological replicates similarity) was assessed by Principal Component Analysis (PCA) and other visualisations (e.g. heatmaps), using custom R scripts. We normalised the raw read counts in zinbwave with the following parameters: K = 0, epsilon = 1e12, X = "~Stages", obrvationalWeights = TRUE. The weights from the ZINB model were used for differential expression analysis using the DESeq2 package (v. 1.40.2; Love et al., 2014) with the following settings: sfType = "poscounts", useT = TRUE, minmu = 1e-6. The formula used in DESeq2 included the factor ´stage´ and this formula was used to identify DE lincRNAs between consecutive stages of the experiment. The sets of DE lincRNAs were extracted using the function 'results', provided with the optional filter ´rowMedians(counts(dds))´. DE lincRNAs considered for further analysis were filtered by fold change ($\log_2$FC) ≥ 0.5 and *P*-values adjusted for multiple testing ($P_{adj}$ <0.01), as suggested by Schurch et al. (2016).

## 2.4 | sRNA sequencing and data pre-processing

RNA was isolated from eight developmental stages during somatic embryogenesis (SE) from two embryogenic cultures, initiated and captured from seeds of the same tree. Cultures from a cell line K11-35, initiated in 2011, were proliferated for embryo maturation and germination by the same protocol as described for the cell line K14-03 in the section Plant material.

Isolated RNA was sequenced on Illumina HiSeq 2500 using 2x101 bp and 2x126 bp paired-end reads (in the first and second sequencing batch, respectively) at SciLifeLab to an average read number of 22.3 ± 11.2 M reads per sample. Only forward reads were used for sRNA analysis. Four biological replicates were provided for each developmental stage/time. Sequencing was performed in two batches corresponding to two separate experiments. The raw data is available at ENA as part of accession PRJEB72619.

The quality of the sequencing data was assessed using FastQC (v0.11.7; Andrews, 2012) and reads were manipulated using Kraken (v13-274), a set of tools for quality control and analysis of high-throughput sequence data (Davis et al., 2013). 3′ adapter sequences were identified, trimmed off providing the first 15 bp of the adapter sequence, and reads containing ambiguous bases, not meeting the quality threshold or having low complexity were discarded (−3pa TGGAATTCTCGGGTG-geom no-bc-tri 40-qqq-check 20/5). Length (18–24 nt) and quality filtered reads were mapped to rRNA and tRNA sequences from genus *Picea*, downloaded from the RNAcentral database (The RNAcentral Consortium 2017) using Bowtie (v 1.2.2; Langmead et al., 2009) with 2 allowed mismatches. Reads matching rRNA and tRNA sequences were removed from further analysis.

## 2.5 | Identification of miRNAs

Clean reads were used to identify small RNA clusters in the genome of *Picea abies* (v1.0; Nystedt et al., 2013) using ShortStack (v3.8.5; Johnson et al., 2016). Reads were mapped to the reference with no mismatches allowed and with the placing of multi-mapping reads (with not more than 100 mapping sites) guided by uniquely mapped reads. A minimum of 0.5 read per million (RPM) was required to call a cluster and stranded clusters, shorter than 300 nt, were checked for folding requirements and miRNA features using default settings and dicermin set to 18. After miRNA clusters were identified, the number of primary alignments corresponding to the 5′ and 3′ mature miRNA sequence from each miRNA cluster was counted in the merged alignment file for each sample. Only miRNA counts in the samples from the cell line K14-03 were used for further analysis as there was only one sample available for each stage from the cell line K11-35.

## 2.6 | Differential expression analysis of miRNAs

miRNAs having a minimum of one raw count in at least 2 replicates in any of the time points were retained for further analysis. The raw read counts were normalised using the main function in zinbwave (v1.10.0; Risso et al., 2018) with parameters K = 2 and epsilon set to the number of miRNAs (340). Variation in the samples was explored using principal component analysis (PCA). The weights from the ZINB model were used for differential expression analysis using DESeq2 (v1.28.1). Wald tests were performed using model ' ~ Stage' (settings sfType = "poscounts", useT = TRUE, minmu = 1e-6), and differentially expressed (DE) miRNAs between consecutive stages of the experiment were extracted using function 'results', provided with option 'filter = rowMedians(counts(dds))'. DE miRNAs considered for further analysis were filtered by $\log_2$FC ≥0.5 and *P*-values adjusted for multiple testing ($P_{adj}$ <0.01).

## 2.7 | Annotation of known and novel miRNAs

Predicted miRNAs and their precursors were mapped to precursor sequences from miRBase release 22 (Kozomara and Griffiths-Jones, 2014) using BWA (v0.7.17 aln) with default settings using the samse algorithm with the option-n 200 specified (Li and Durbin, 2009) to report multiple primary alignments. Beforehand, all uridines were changed to thymines. miRNA sequences often had more than one primary alignment, therefore mapping results of predicted miRNA precursors and 5′ and 3′ mature miRNAs were compared. When there was a miRBase hit common to all three sequences or at least two of them, it was chosen as the preferred hit. In all other

instances, miRBase hits from all primary alignments of the miRNA sequences were reported.

## 2.8 | Gene co-expression network inference

The lincRNAs expression data were transformed to homoscedastic, asymptotically $\log_2$ counts using the variance stabilising transformation as implemented in DESeq2. We then merged the three different classes of RNAs to be able to build the network. Then, ten network inference methods (aracne, clr, genie3, llr-ensemble, narromi, pcor, pearson, plsnet, spearman, and tigress; Haury et al., 2012; Guo et al., 2016; Schäfer and Strimmer, 2005; Zhang et al., 2013; Ruyssinck et al., 2014; Faith et al., 2007; Huynh-Thu et al., 2010; Margolin et al., 2006) were run using the Seidr toolkit (Schiffthaler et al., 2023). The networks were aggregated using the inverse rank product (IRP) method (Zhong et al., 2014) and edges were filtered according to the noise-corrected backbone (Coscia and Neffke, 2017) at multiple backbone values. We used Receiver Operating Characteristics (ROC) curves to assess the specificity and sensitivity of each individual backbone network (Allen et al., 2012). ROC makes curves based on probability, where the True Positive Rate (TPR) is plotted against the False Positive Rate (FPR) (Davis and Goadrich, 2006). These values are calculated using a Gold standard based on KEGG pathways (Sferra et al., 2017). We used these methods to threshold the network and selected a backbone value of 1 for downstream analyses. Network partitions were identified using InfoMap (Rosvall and Bergstrom, 2008) with default settings. The network was further visualised and processed using Cytoscape (v. 3.10.1; Shannon et al., 2003).

## 2.9 | Conservation and clustering analysis

To evaluate if the lincRNAs belonging to the co-expression network were conserved, we ran a blastn (v. 2.11.0+; Altschul et al., 1990) against a set of genomes that were downloaded from the PlantGenIE.org (Sundell et al., 2015) resource: *Amborella trichocarpa*, *Arabidopsis thaliana*, *Eucalyptus grandis*, *Ginkgo biloba*, *Gnetum montanum*, *Nicotiana tabacum*, *Physcomitrella patens*, *Pinus taeda*, *Populus trichocarpa*, *Populus tremula*, and *Vitis vinifera*. Moreover, cd-hit-est (v. 4.8.1; W. Li & Godzik, 2006) was used to identify potential lincRNAs families by identifying transcripts with >80% identity, which were defined as a cluster/family.

## 2.10 | Functional enrichment analysis

An in-house tool (gopher2; https://github.com/bschiffthaler/gofer2 executed using the script 'gopher.R' from https://doi.org/10.5281/zenodo.10391673) was used for Gene Ontology, MapMan and Pfam enrichment. The gopher tool implements the Parent Child test from Grossmann et al. (2007) for Gene Ontology enrichment and a Fischer exact test for other enrichment tests and applies Benjamini-Hochberg multiple testing correction. All the enrichments with a Benjamini-Hochberg adjusted *P*-value lower than 0.05 were considered significative and used for analyses.

## 2.11 | Prediction of miRNA targets and miRNA sponges

Target predictions were performed for all the miRNA present in the network using psRNATarget (Dai et al., 2018) against all the coding transcripts present in the network. Default settings from scoring schema V2 were used, only hsp size was changed to be equal to the size of the miRNA sequences in the query. A strict Expectation value 3 was used to select the targets.

miRNA sponges were identified using psRNATarget on the set of miRNAs and lincRNAs present in backbone1, following the rules described by Wu et al. (2013) and having a strict Expectation value of 3.

## 2.12 | Prediction of lincRNAs acting as miRNA precursors

miRNA precursors (pre-miRNA) were aligned against the lincRNAs using blastn. The lincRNAs homologous to miRNA precursors with e-value = 1e-5 were defined as miRNA precursors.

## 2.13 | Guilt-by-association functional annotation of lincRNAs

The backbone of the co-expression network, which was generated by Seidr, served as input for New Gene Ontology Annotation (NewGOA), as proposed by Yu et al. (2018). NewGOA was cloned on June 1st 2021 from https://github.com/Ayllonbe/gni_predictors. The method constructs a hybrid network combining both the co-expression network and the GO network. NewGOA employs a bi-random walks algorithm to traverse the hybrid network, navigating through the interconnected nodes and edges. Through this iterative process, the algorithm dynamically explores the network topology, identifying potential functional associations between genes and their corresponding GO terms. We filtered NewGOA predictions based on their PredictionScore to keep only those with the score in the upper quantile (PredictionScore > = 2^-13), which retained 731 lincRNAs.

## 2.14 | Resource overview

The resource comprises global identification of lincRNAs and miRNAs during the process of somatic embryogenesis development. All raw sequencing data is available at the ENA as accession PRJEB72619

and all derived (normalised) expression values used for the analyses presented are available at the SciLife (Science for Life Laboratory, Sweden) FigShare resource at doi:10.17044/scilifelab.25315867. In addition, the FigShare resource includes the transcript assembly fasta file and details of the filtered set of lincRNAs used for network analysis. The gene expression data and co-expression network have been integrated within the PlantGenIE.org resource (Sundell et al., 2015) to enable easy visual exploration. Within PlantGenIE.org, the expression levels of lncRNAs can be visualised using the exImage, exPlot and exHeatmap tools. The lncRNAs are also included in the co-expression network for this dataset in the exNet tool. Gene information pages for lncRNAs have also been included. Gene lists can be created that contain a combination of protein-coding and lncRNA genes (or either type exclusively). All scripts used to perform the presented analyses are available at the Git repository DOI: 10.5281/zenodo.10716226. This resource was developed to identify whether lincRNAs are differentially expressed during the development process of somatic embryogenesis and to ascertain the subset that has the highest co-expression network connectivity as an indication of their regulatory potential to direct future studies.

## 3 | RESULTS

### 3.1 | Long intergenic non-coding RNA identification and differential expression during somatic embryogenesis

We utilised an existing RNA-Seq dataset profiling transcript expression during somatic embryogenesis (SE) of Norway spruce to perform a *de novo* transcript assembly from which we identified putative intergenic long non-coding RNAs (hereafter lincRNAs). Previous analysis of protein-coding genes using these data revealed extensive changes in the transcriptome at the assayed stages of SE (Stojkovič et al., 2024). We, therefore, considered that this dataset was suitable for identifying lincRNAs with a clear signal of active regulation.

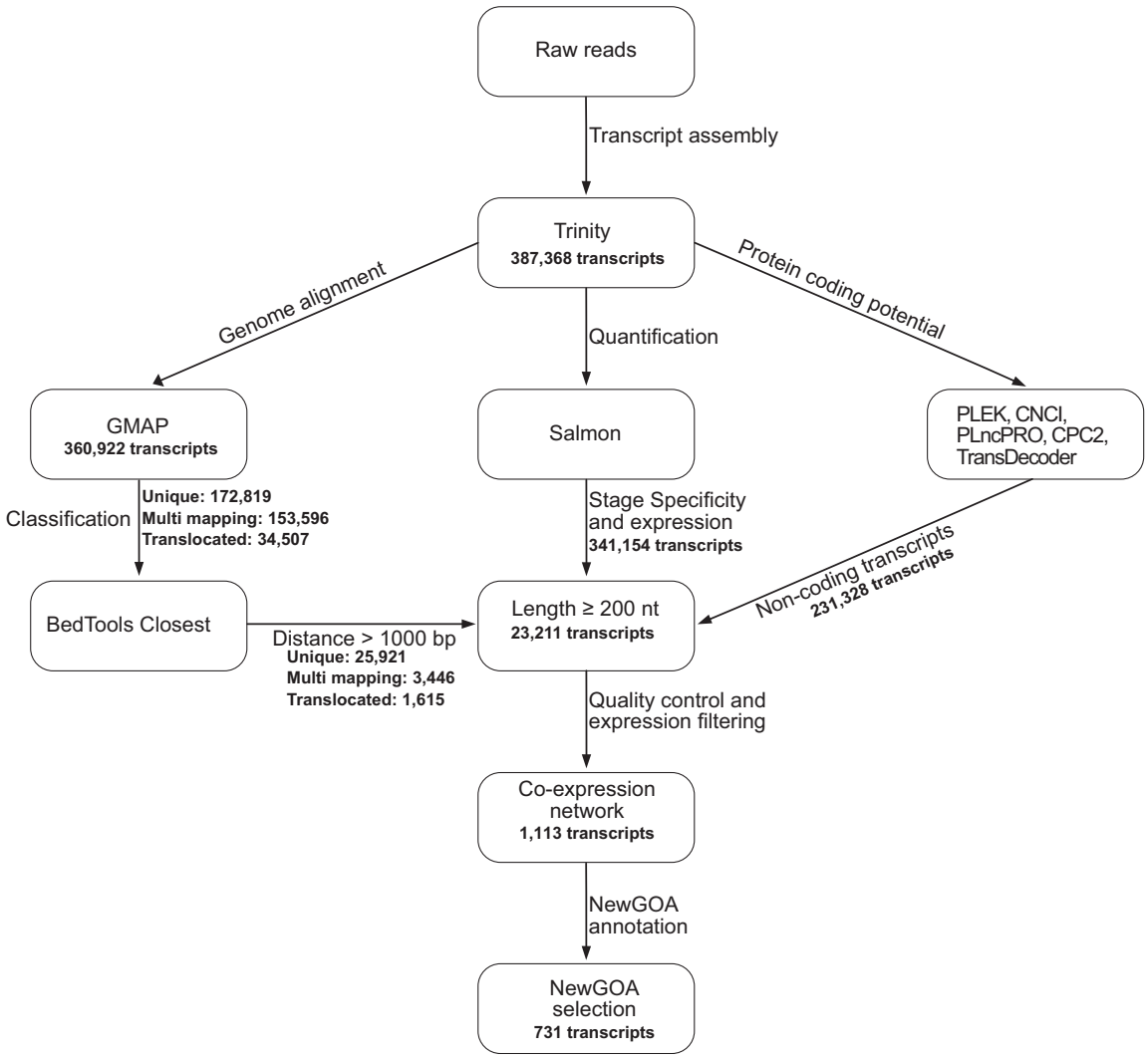



**FIGURE 1** An overview of the implemented pipeline including details of the number of transcripts remaining after certain analysis and filtering steps.

The transcript assembly comprised 387,368 sequences (Figure 1), which we filtered to remove those with any overlap to annotated genes (188,052 transcripts overlapping 30,916 annotated genes). Of the remaining 199,316 transcripts, 5,835 were removed as they had evidence of protein-coding potential, potentially representing coding genes missing in the current genome annotation, expressed pseudogenes or lncRNAs with small open reading frames producing short peptides. We then applied expression filters to the remaining sequences to identify the subset of lincRNAs that had consistent detectable expression (among replicates) in at least one sampled developmental stage, identifying 1,774 that were considered in further downstream analyses. This filtered set largely contained unique sequences (1,492), although 480 clustered at 80% identity into 198 clusters containing between two and 14 sequences, representing potential lincRNA families or cases of duplication, with most (148) containing two sequences. Within clusters, most transcripts originated from multiple loci, but 68 clusters contained potential splicing isoforms from a single locus (this number is likely an underestimate as some transcripts,8.9% in total, were split across scaffolds in the assembly).

Protein coding genes had higher average expression levels (Figure 2A) and broader expression (*i.e.*, expressed in a larger proportion of stages and having a lower stage specificity score; Figure 2B) than lincRNAs. Stage specificity was used to divide each gene class into genes with broad or narrow (specific) expression. For both gene classes, expression was higher for genes with broad expression (Figure S1A), which suggests that broad expression was not merely a signature of low-level promiscuous transcription. Stages S7 and, most notably, S8 had the highest representation of lincRNAs with specific expression (Figure S1B). Protein coding genes contained a greater number of exons, with clear enrichment of lincRNAs having one or two exons (Figure 2C). There was also a notable difference in GC content of the two gene classes, with protein-coding genes having higher GC content and a narrower distribution (Figure 2D).

To ascertain whether the expression of the filtered set of lincRNAs captured among-sample relationships was similar to that represented by protein-coding genes, we performed a Principal Component Analysis (PCA), which revealed a highly similar structure in both datasets (Figure 3A,C). For both gene types, there was one group of samples

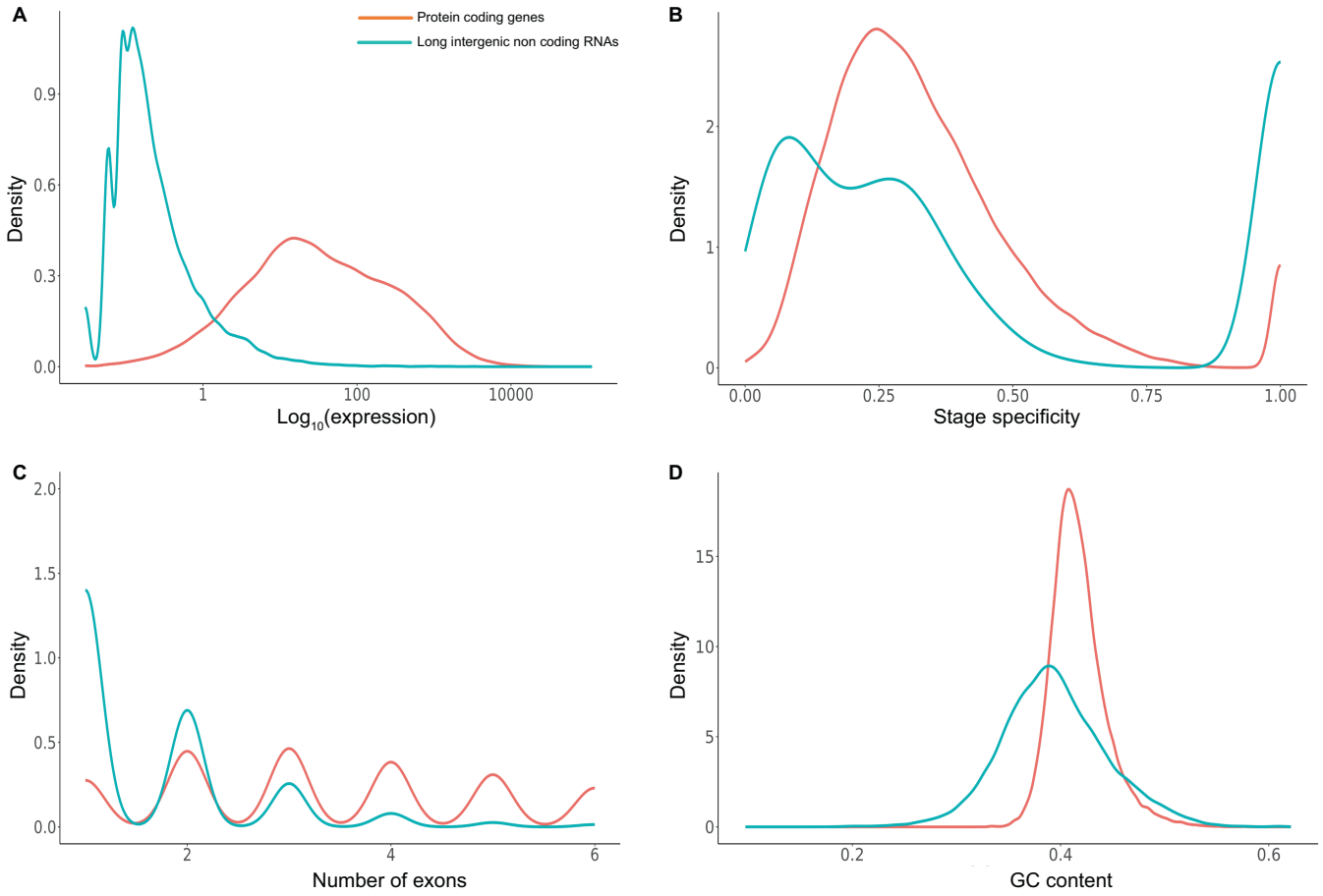



**FIGURE 2** Characteristics of protein coding genes and the identified long intergenic non-coding RNA (lincRNA) genes. Protein coding genes represent annotated genes from the v1 Norway spruce genome. **A** Density distribution of protein coding and lincRNA gene expression values. Values are $\log_{10}$ variance stabilising transformation (VST) normalised. **B** Stage specificity of protein coding and lincRNA genes. Stage specificity is a score ranging from 0 to 1 where a score of 1 indicates highly specific expression (i.e., expressed at only one sampled stage) and 0 indicates expression across all sampled stages. **C** The number of exons in protein coding and lincRNA genes. **D** GC content (proportion) in protein coding and lincRNA genes.

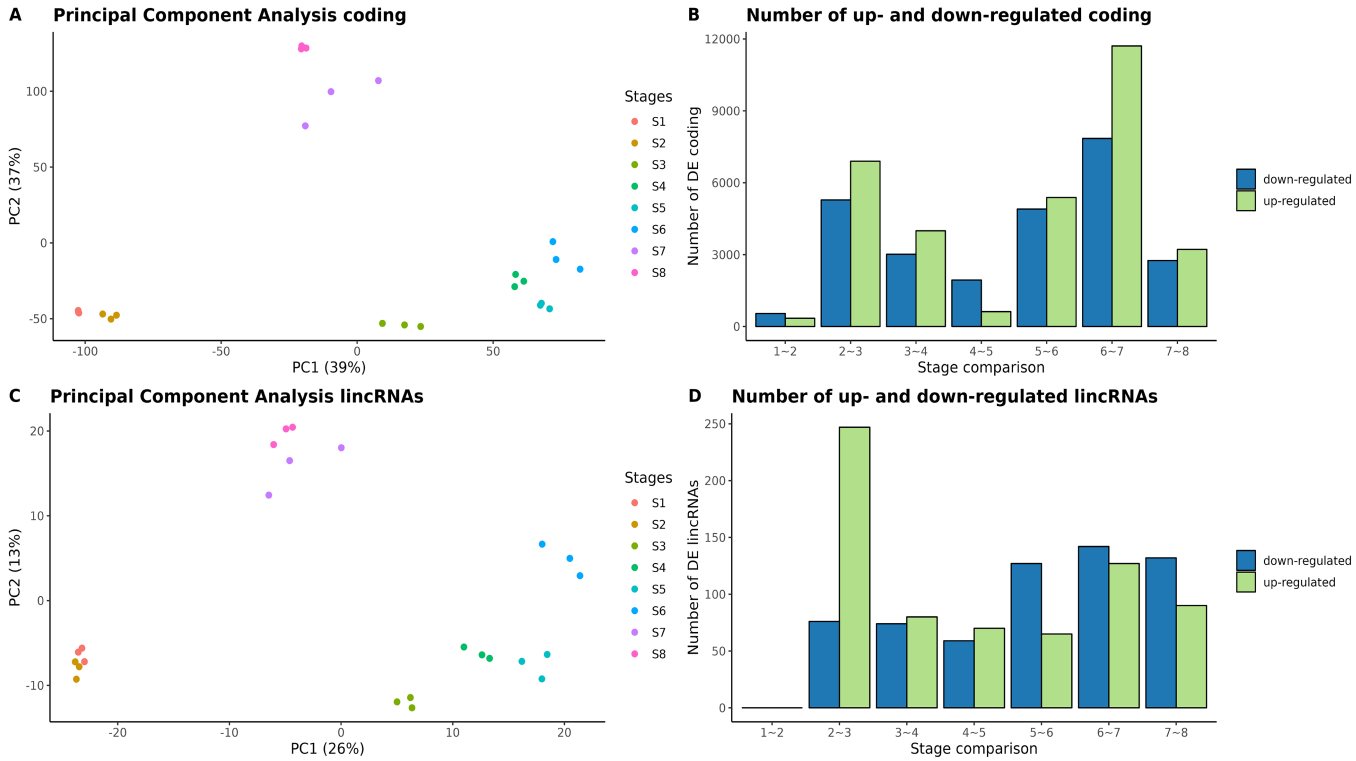



**FIGURE 3** Expression characteristics of protein coding (**A,B**) and the identified long intergenic non-coding RNA (lincRNA) genes (**C,D**). **A** Principal component analysis (PCA) plot of gene expression values for protein coding genes. The PCA was performed using variance stabilising transformation (VST) normalised expression values. Samples are coloured to indicate sample stage. **B** The number of up (green bars) and down (blue bars) regulated protein coding genes at stage transitions of the sampled somatic embryogenesis developmental process. Differential expression was defined by a false discovery rate corrected *P*-value <0.05 and $\log_2$ fold change <0.5. **C** PCA plot of normalised gene expression values of identified lincRNA genes. Samples are coloured to indicate sample stage. **D** The number of up (green bars) and down (blue bars) regulated lincRNA genes at stage transitions of the sampled somatic embryogenesis developmental process.

representing stages S1 and S2, a second group comprising samples from S3-S6 and a final group containing samples from S7-S8. Samples from stage S6 were more distinctly separated from those of S4 and S5 based on lincRNA expression than for protein-coding genes.

We identified differentially expressed genes (DEGs) between progressive stages of the development series (Figure 3B,D). For protein-coding genes, there were major transitions in the transcriptome at stages S2-S3 and S6-S7 (as reported in Stojkovič et al., 2024), which were similarly reflected in the number of lincRNA DEGs. There were, however, differences in the balance of up-and down-regulated genes, with a higher proportion of up-regulated lincRNA between stages S2-S3 and a less pronounced representation of up-regulated genes at S6-S7. For both classes of gene there were very few DEGs between S1-S2 despite similar numbers of expressed genes at all stages, suggesting a steady-state transcriptome in these two stages.

## 3.2 | miRNA identification, differential expression and lincRNA sponge prediction

As one of the known roles of lincRNAs is their interaction with miRNAs to act as 'sponge' sequences, we analysed short RNA data generated from the same samples used to profile mRNA to identify and quantify miRNAs. We identified 422 mature miRNAs produced from 211 clusters (see FigShare resource for details), among which most had a predominant sRNA size of 21 nt or 22 nt (67% and 24% of clusters, respectively; Figure S2). Of sRNA reads that aligned to identified miRNAs, most (80%) aligned to miRNAs of length 21 nt, with a further 16% aligning to miRNAs of length 22 nt. 340 miRNAs reached the expression threshold used for filtering and were analysed further, of which 294 had been previously reported in miRbase (Kozomara and Griffiths-Jones, 2014) while the remaining 46 were novel to this study (see FigShare resource for details). Among this expressed set of miRNAs, 69% were of length 21 nt while 24% were 22 nt. Known and novel predicted miRNAs had similar and expected patterns of first base composition, with a dominance of sequences starting with a uracil (Figure S3).

In general, the pattern for DE miRNAs was similar to that of protein-coding and lincRNA genes (Figure S4), although the greatest number of DE miRNAs occurred between S2-S3 (65 miRNAs), followed by S7-S8 (50 miRNAs) and S6-S7 (44 miRNAs). There were 120 DE miRNAs representing 112 unique mature sequences originating from 103 unique precursor sequences (see FigShare resource for details). Similar to DEGs, few miRNAs were DE in mid-maturing

embryos compared to early-maturing embryos and during desiccation. All but five DE miRNAs had computationally predicted targets (see FigShare resource for details), with 87% (1442 genes) of 1658 unique predicted targets expressed in our dataset, of which 59% were DE (850 genes). DE target genes were detected for 101 DE miRNAs.

Among the predicted pre-miRNAs, 14 were sequence-similar to 12 predicted lincRNAs. Excluding those lincRNAs as they represented potential pre-miRNAs, there was one predicted miRNA sponge. Although the correlation of expression between the sponge and associated miRNA was not significant ($P$-value $= 0.1$, $r = -0.34$; Figure S5), there was a general pattern of lower miRNA expression during stages where the sponge lincRNA was more highly expressed. This may indicate that a linear correlation test is not indicative of the regulatory relationship. However, future work is needed to validate the role of this lincRNA as a miRNA sponge.

### 3.3 | Guilt-by-association annotation of lincRNAs

To explore the potential role of lincRNAs more comprehensively, we calculated a co-expression network that included protein coding, lincRNA genes and mature miRNAs (Figure 4). We filtered all three classes using the same expression criteria before network calculation. For protein-coding genes, 66,360 genes were classed as expressed, of

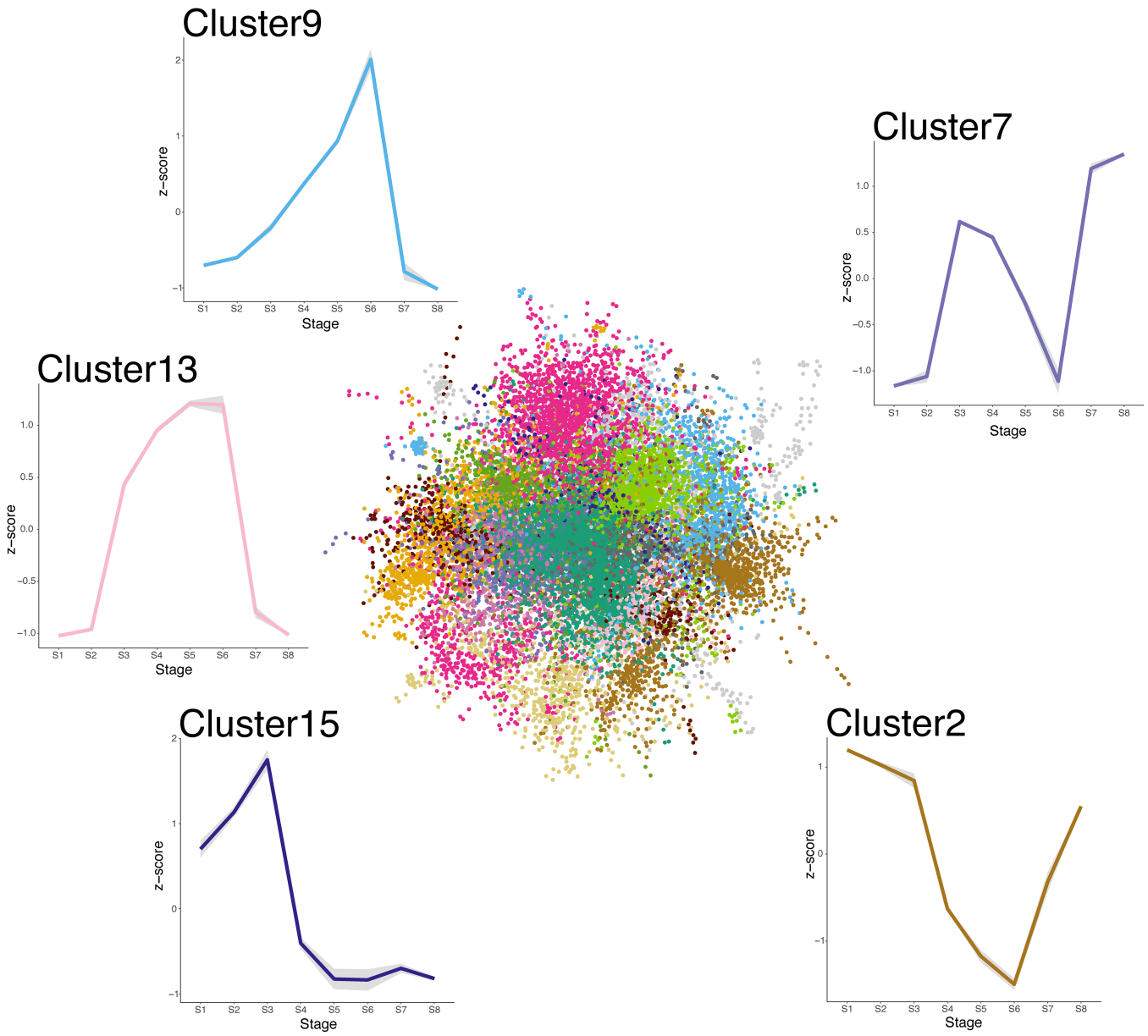



**FIGURE 4** Co-expression network comprising protein coding, long intergenic non-coding RNA and micro-RNA genes. Nodes within the network are coloured to indicate their assigned cluster membership. The eigengene expression profile for a subset of clusters is represented as inserts around the co-expression network. Edges within the network have been omitted for visual clarity. The associated network file is available at the FigShare resource detailed in the data availability statement.

which 28,697 were included in the network. We considered that inclusion in the co-expression network suggests that these genes are regulated during SE and of potential biological interest. For miRNAs, 329 were included, while for lincRNAs 1,774 were included. We examined network connectivity metrics of the different gene classes, which revealed that both lincRNAs and miRNAs had similar network connectivity metric distributions to protein-coding genes (Figure 5). This similarity was greater for lincRNAs, with several lincRNAs having higher connectivity than the subset of the protein-coding genes annotated as transcription factors.

As lincRNAs lack any functional annotation, and as classical tools for assigning functional annotations do not work for non-protein coding genes, we employed a guilt-by-association approach to assign lincRNAs functional descriptions based on their co-expression with expressed protein-coding genes. This was achieved using a method that simultaneously considers both a co-expression network graph and the graph relationship of assigned Gene Ontology (GO) terms (NewGOA; Yu et al., 2018). Protein coding genes had higher NewGOA score values than lincRNAs (Figure S6), reflecting the additional challenge of assigning putative functional descriptions to lincRNAs. We additionally performed sequence similarity searches against reference genomes to determine the conservation of lincRNAs, considering conservation as an additional signal of functional potential. Very few lincRNAs had any evidence of conservation, with 655 conserved among the considered gymnosperm species (see FigShare resource for details).

We performed functional enrichment tests of co-expression clusters using the pre-existing annotation of protein-coding genes (See FigShare resource for details) and focused on the fifteen largest clusters (Figure S7 depicts the eigengene expression profile for each of these 15 clusters). For each cluster, we examined the proportion of protein-coding, lincRNA and miRNA genes and the number of each type assigned an annotation using NewGOA (Table 1). In general, there were three dominant expression profiles comprising genes with high expression in proliferation stages (S1-S2; cluster 4), those with expression during maturation (S3-S5; cluster 1) and those with expression at the later desiccation and germination stages

**TABLE 1** The number of genes in different categories within co-expression network clusters.

| | Genes | lincRNAs | miRNAs |
|---|---|---|---|
| Cluster1 | 2876 (2771) | 193 (160) | 35 (33) |
| Cluster2 | 1226 (1189) | 40 (36) | 20 (20) |
| Cluster3 | 1433 (1380) | 17 (17) | 1 (1) |
| Cluster4 | 1031 (811) | 549 (335) | 188 (149) |
| Cluster5 | 639 (604) | 30 (23) | 4 (4) |
| Cluster6 | 965 (928) | 23 (19) | 2 (1) |
| Cluster7 | 719 (690) | 23 (19) | 1 (1) |
| Cluster8 | 674 (644) | 13 (10) | 0 |
| Cluster9 | 942 (879) | 5 (5) | 0 |
| Cluster10 | 492 (474) | 2 (2) | 0 |
| Cluster11 | 535 (516) | 12 (9) | 0 |
| Cluster12 | 603 (583) | 25 (22) | 0 |
| Cluster13 | 356 (334) | 9 (9) | 0 |
| Cluster14 | 439 (432) | 14 (14) | 0 |
| Cluster15 | 228 (204) | 40 (24) | 5 (1) |

Values in parentheses indicate the number of genes assigned annotation using NewGOA.

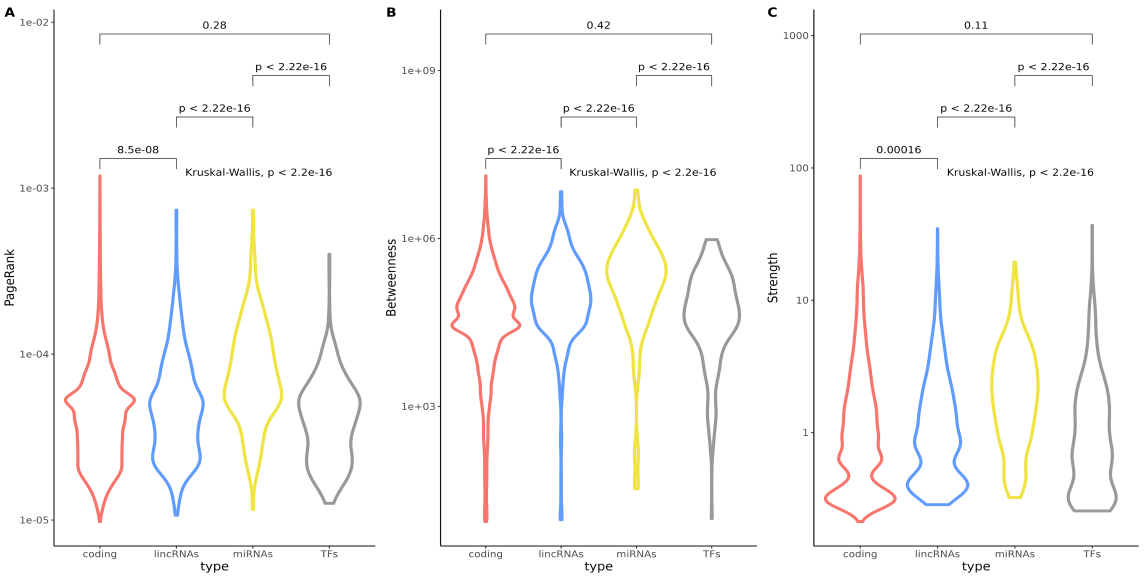



**FIGURE 5** Co-expression network connectivity measures for protein coding, long intergenic non-coding RNA, micro-RNA and transcription factor genes. **A** Distribution of PageRank scores. **B** Distribution of betweenness scores. **C** Distribution of strength scores. All parts are violin plot representations of the data distribution. The significance of distribution comparisons identified using a Kruskal-Wallis test is indicated above the distributions.

(S6-S8; cluster 6). Functional enrichment of these three clusters revealed significant enrichments including nitrogen compound metabolic process (GO:0006807) during proliferation (cluster 4); nutrient reservoir activity (GO:0045735) during maturation (cluster 1); and photosynthesis (GO:0015979), plastid (GO:0009536), chlorophyll-binding (GO:0016168), and thylakoid (GO:0009579) during germination (cluster 6). Cluster 2 had a distinctive expression profile with expression in all stages except S6, representing desiccation and embryo dormancy. This cluster had enrichment for categories including cell cycle (GO:0007049), cell division (GO:0051301) and DNA replication (GO:0006260), likely representing the repression of these processes.

We further considered the five lincRNAs with the highest network connectivity (PageRank score). To indicate the potential biological processes that these could be involved in, we examined their annotation assigned by NewGOA and the functional enrichment results of their first-degree network neighbours and network cluster. Of note, there were five lincRNAs in the 50 highest-ranked genes within the network (genes of all classes were ordered by rank).

## 4 | DISCUSSION

The Norway spruce genome is large (~20 Gb) and comprised primarily of repetitive sequences, yet is estimated to contain a similar number of protein-coding genes to other diploid plant species (Nystedt et al., 2013). As repetitive elements are known to be a source of both regulatory motifs and regulatory effects and to create various ncRNAs (Palazzo and Koonin, 2020) we were interested in exploring the diversity and extent of lncRNA expression in Norway spruce as a representative conifer species. It had previously been reported that there is extensive expression of lncRNA in Norway spruce (Nystedt et al. 2013), but that observation was not based on a detailed and dedicated analysis. To perform such a dedicated analysis, we utilised an existing data set profiling gene expression during the process of somatic embryogenesis (SE), which was shown to comprise extensive transcriptome changes with distinct transcriptome remodulation events and expression of a large proportion of all annotated protein-coding genes (Stojkovič et al., 2024). We reasoned that these characteristics would maximise the detection of any expressed lncRNAs and that the clearly defined expression profiles observed for protein-coding genes would enable similar distinction of clear regulation for lncRNAs. Evidence of regulated expression can be used as a signal to filter spurious, stochastic expression resulting from non-specific Pol-II binding and transcriptional initiation.

Our analysis identified 1,774 transcripts that passed the stringent set of structural and expression-based filters applied. We note, however, that there was a far more extensive set of lncRNAs detected prior to applying those filters (Figure 1; 23,211 transcripts) and it is likely that a proportion of the excluded transcripts represent genuine lncRNAs. In this study, we focused specifically on intergenic lncRNAs (lincRNAs) as we felt that the current quality of gene annotation in Norway spruce would result in a high error rate for defining lncRNAs with overlap to annotated protein-coding genes. The fragmented nature of the genome assembly also negates any inference of *cis* or *trans* relationships to other genes or genomic features. Taken together, the set of lincRNAs we present is, therefore, more likely to represent a lower bound estimate of the diversity of lncRNAs expressed in Norway spruce. Despite these caveats, we identified an extensive set of lincRNAs with clear differential expression profiles during SE (Figure 2C,D). Similar to previously reported studies of lncRNAs in plants, this set of identified lincRNAs had a lower average (Figure 2A) and more sample-specific (Figure 2B) expression than protein-coding genes. The lincRNAs were also shorter, with fewer multi-exonic genes (Figure 3C). Among the lincRNAs, 12 transcripts corresponded to precursors of 14 miRNAs. The number of lincRNAs we identified is consistent with other similar studies, however, we note that such comparisons are problematic as there is a lack of consistency in applied methodology and filtering. It can be seen from Figure 1 that even small changes to the filtering criteria applied at various steps of the pipeline could result in large changes in the number of lincRNAs considered, especially given the initially very high number of transcripts derived from the transcript assembly.

The expression profiles of the lincRNAs contained sufficient biological signals to clearly separate the sampled stages of SE, revealing similar among-sample relationships to those based only on protein-coding genes (Figure 3A,C). Similarly, the patterns of differential expression were similar for the two classes of genes with the same two major transitions in the transcriptome present between stages S2-S3 and S6-S7 (Figure 3B,D). These observations indicate that lincRNAs and protein-coding genes are under active regulation to a similar extent, with regulated expression providing circumstantial evidence of biological function. However, except for the few cases where a clear indication of function can be assigned, such as lincRNAs predicted to act as miRNA sponges, the functional role and significance of these lincRNAs remain enigmatic and unexplored. As such, an important first step in any species is to first catalogue and characterise the expression of the population of lincRNAs to enable subsequent selection of candidate genes for downstream hypothesis generation and functional validation studies.

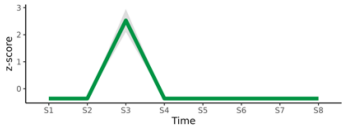
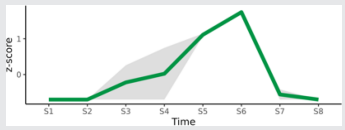
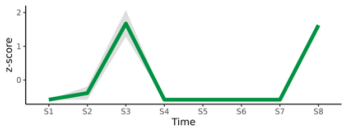
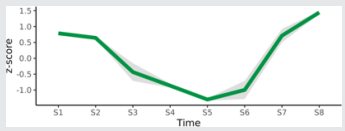
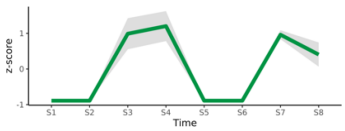
One of the major challenges in studying lncRNAs is a lack of methods to assign biological functions to non-coding transcripts. Unlike protein-coding genes, where sequence homology and the presence of conserved protein domains can be used to infer function, there are currently no sequence, structure, or context-based methods to associate lncRNAs to biological processes or functions. One method to overcome this barrier is to use a guilt-by-association approach to assign functions on the basis of annotations of protein-coding genes with similar expression profiles. This is commonly achieved using methods such as co-expression networks, clustering, or other correlation-based approaches to group lncRNAs with protein-coding genes. The same approach has previously been applied to protein-coding genes of unknown function, with evidence that this can achieve a high degree of reliability (Depuydt and

Vandepoele, 2021). There are also methods available to take into consideration both co-expression and functional annotation links between genes (Yu et al. 2018). We applied these concepts to our set of linRNAs by inferring a gene co-expression network using an ensemble method (Schiffthaler et al. 2023) that included protein coding, lincRNA genes and miRNAs (Figure 3). This network was used as input to NewGOA, a tool that infers functional annotation using both the expression network and Gene Ontology (GO) term relationship graphs. This enabled us to assign a putative function to 731 lincRNAs (Table 2). We were also interested in determining whether lincRNAs tended to be less centrally integrated into the gene expression network. Our expectation was that lincRNAs would be less central as they are evolutionarily young and typically have low and narrow expression. It has also previously been shown that central genes are enriched for high expression and signatures of purifying and positive selection (Mähler et al., 2017). However, the global distribution of several centrality measures for lincRNAs was highly similar to that of protein-coding genes (Figure 5), with some lincRNAs being ranked as highly for centrality or betweenness measures as transcription factor genes. It will be of interest to explore patterns of selection for these lincRNAs as population genetics data become available for conifer species. Similarly, as more reference quality gymnosperm genomes become available, it will be interesting to determine patterns of cross-species conservation to discover how this varies for lncRNAs of contrasting network centrality and to compare these patterns to those in genomes with contrasting degrees of repetitive element retention and activity.

## 5 | CONCLUSIONS

Using an extensive collection of RNA-Seq data profiling SE development in Norway spruce we identified lincRNAs with clear evidence of differential expression during the process of SE. The identified

**TABLE 2** Five most highly ranked long intergenic non-coding RNAs defined by PageRank.

| Expression profile | First Degree Neighbours | | | Functional Enrichment | | | PageRank score (Global rank) |
|---|---|---|---|---|---|---|---|
| | Total | Coding | Non-coding | MapMan | Pfam | GO | |
| TRINITY_DN96496_c0_g1_i1 (Cluster15)  | 101 | 92 | 9 | 16 | 10 | 0 | 0.000737942 (9) |
| TRINITY_DN12829_c0_g1_i5 (Cluster9)  | 30 | 30 | 0 | 7 | 12 | 0 | 0.000704327 (13) |
| **TRINITY_DN52747_c1_g1_i1** (Cluster2)  | 70 | 63 | 7 (2) | 18 | 19 | 1 | 0.000657588 (19) |
| **TRINITY_DN6985_c1_g1_i1** (Cluster13)  | 39 | 39 | 0 | 5 | 18 | 0 | 0.000531214 (29) |
| TRINITY_DN20571_c0_g1_i1 (Cluster7)  | 57 | 57 | 1 (1) | 18 | 28 | 1 | 0.000448974 (45) |

Gene IDs in bold indicate conservation in other plant genomes (detailed in Methods). MapMan terms, Pfam domains and GO terms represent significantly enriched terms within the set of first-degree neighbours ($q$-value <0.05).

lincRNAs represent interesting candidate genes for future characterisation studies to validate and elucidate their biological functions during SE development. To facilitate access to the resource, we have included the set of identified lincRNAs within the gene expression visualisation tools at the PlantGenIE.org resource and have made all presented data and analyses available in public repositories. Integration of the data in PlantGenIE.org offers easy and intuitive visualisation of expression profiles during SE and visual exploration of the co-expression network, for example, to identify sets of co-expressed genes. This work presents one of the most comprehensive explorations of lincRNAs in a conifer species and demonstrates that lincRNAs may serve important regulatory roles during developmental processes such as somatic embryogenesis.

## AUTHOR CONTRIBUTIONS

CC performed all analysis is lincRNAs, KS performed analysis of miRNAs, AAB advised on NewGOA analysis and interpretation of the results, ND co-supervised CC and performed BLAST analysis, UE supervised and conceived the experiment to produce the biological samples, NRS conceived the study, supervised CC, KS and AAB. NRS and CC wrote the manuscript.

## ACKNOWLEDGEMENTS

Ioana Gaboreanu and Sofie Johansson collected samples for the somatic embryogenesis series. K Stojkovič was supported by a grant from the Kempe Foundation (SMK1340). N Street, C Canovi and U Egertsdotter are supported by the Trees for the Future (T4F) project. This work was supported by grants from the Knut and Alice Wallenberg Foundation. The authors acknowledge support from the National Genomics Infrastructure in Genomics Production Stockholm funded by Science for Life Laboratory, the Knut and Alice Wallenberg Foundation and the Swedish Research Council, SNIC/Uppsala Multidisciplinary Center for Advanced Computational Science for assistance with massively parallel sequencing and access to the UPPMAX computational infrastructure, and the Umeå Plant Science Centre bioinformatics facility for support.

## CONFLICT OF INTEREST STATEMENT

N Street is a shareholder in Woodheads AB. Woodheads AB is a shareholder in SweTree Technologies, which has a commercial interest in somatic embryogenesis of Norway spruce.

## DATA AVAILABILITY AND FAIR COMPLIANCE

All raw sequencing data is available at the ERA as accession PRJEB72619 and all derived (normalised) expression values used for the analyses presented are available at the SciLife (Science for Life) FigShare resource at doi:10.17044/scilifelab.25315867. The transcript assembly fasta file is available at FigShare. All scripts used to perform the presented analyses are available at the Git repository DOI: 10.5281/zenodo.10716226.

## ORCID

*Nathaniel R. Street* https://orcid.org/0000-0001-6031-005X

## REFERENCES

Ali, T. and Grote, P. (2020) Beyond the RNA-dependent function of LncRNA genes. *Elife*, 9, 1–14.

Allen, J.D., Xie, Y., Chen, M., Girard, L. and Xiao, G. (2012) Comparing Statistical Methods for Constructing Large Scale Gene Networks. *PLoS One*, 7, e29348.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol*, 215, 403–410.

Amor, B. Ben, Wirth, S., Merchan, F., et al. (2009) Novel long non-protein coding RNAs involved in Arabidopsis differentiation and stress responses. *Genome Res*, 19, 57–69.

Andrews, S. (2012) FastQC: A quality control application for high throughput sequence data. *Babraham Institute Project page*: http://www. *bioinformatics. bbsrc. ac. uk/projects/fastqc*.

Ariel, F., Jegu, T., Latrasse, D., Romero-Barrios, N., Christ, A., Benhamed, M. and Crespi, M. (2014) Noncoding transcription by alternative rna polymerases dynamically regulates an auxin-driven chromatin loop. *Mol Cell*, 55, 383–396.

Bazin, J. and Bailey-Serres, J. (2015) Emerging roles of long non-coding RNA in root developmental plasticity and regulation of phosphate homeostasis. *Front Plant Sci*, 6, 144084.

Bedre, R., Irigoyen, S., Petrillo, E. and Mandadi, K.K. (2019) New era in plant alternative splicing analysis enabled by advances in high-throughput sequencing (HTS) technologies. *Front Plant Sci*, 10, 461357.

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

Buchfink, B., Xie, C. and Huson, D.H. (2014) Fast and sensitive protein alignment using DIAMOND. *Nature Methods 2014* 12:1, 12, 59–60.

Budak, H., Kaya, S.B. and Cagirici, H.B. (2020) Long Non-coding RNA in Plants in the Era of Reference Sequences. *Front Plant Sci*, 11, 441273.

Chen, L., Zhu, Q.H. and Kaufmann, K. (2020) Long non-coding RNAs in plants: emerging modulators of gene activity in development and stress responses. *Planta 2020* 252:5, 252, 1–14.

Coscia, M. and Neffke, F.M.H. (2017) Network backboning with noisy data. *Proc Int Conf Data Eng*, 425–436.

Dai, X., Zhuang, Z. and Zhao, P.X. (2018) psRNATarget: a plant small RNA target analysis server (2017 release). *Nucleic Acids Res*, 46, W49–W54.

Davis, J. and Goadrich, M. (2006) The relationship between precision-recall and ROC curves. *ACM International Conference Proceeding Series*, 148, 233–240.

Davis, M.P.A., Dongen, S. van, Abreu-Goodger, C., Bartonicek, N. and Enright, A.J. (2013) Kraken: A set of tools for quality control and analysis of high-throughput sequence data. *Methods*, 63, 41–49.

Deng, P., Liu, S., Nie, X., Weining, S. and Wu, L. (2018) Conservation analysis of long non-coding RNAs in plants. *Sci China Life Sci*, 61, 190–198.

Depuydt, T. and Vandepoele, K. (2021) Multi-omics network-based functional annotation of unknown Arabidopsis genes. *The Plant Journal*, 108, 1193–1212.

Egertsdotter, U., Ahmad, I. and Clapham, D. (2019) Automation and scale up of somatic embryogenesis for commercial plant production, with emphasis on conifers. *Front Plant Sci*, 10, 436563.

Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J.J. and Gardner, T.S. (2007) Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles. *PLoS Biol*, 5, e8.

Grabherr, M.G., Haas, B.J., Yassour, M., et al. (2011) Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat Biotechnol*, 29, 644.

Grossmann S, Bauer S, Robinson P.N, Vingron M. (2007) Improved detection of overrepresented Gene Ontology annotations with parent–child analysis. *Bioinformatics*, 23, 3024–3031.

Guo, S., Jiang, Q., Chen, L. and Guo, D. (2016) Gene regulatory network inference using PLS-based methods. *BMC Bioinformatics*, 17, 1–10.

Haas, B.J., Papanicolaou, A., Yassour, M., et al. (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols 2013* 8:8, 8, 1494–1512.

Haury, A.C., Mordelet, F., Vera-Licona, P. and Vert, J.P. (2012) TIGRESS: Trustful Inference of Gene REgulation using Stability Selection. *BMC Syst Biol*, 6, 1–17.

Hou, X., Cui, J., Liu, W., Jiang, N., Zhou, X., Qi, H., Meng, J. and Luan, Y. (2020) LncRNA39026 enhances tomato resistance to Phytophthora infestans by decoying miR168a and inducing PR gene expression. *Phytopathology*, 110, 873–880.

Huynh-Thu, V.A., Irrthum, A., Wehenkel, L. and Geurts, P. (2010) Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLoS One*, 5, e12776.

Jha, U.C., Nayyar, H., Jha, R., Khurshid, M., Zhou, M., Mantri, N. and Siddique, K.H.M. (2020) Long non-coding RNAs: emerging players regulating plant abiotic stress response and adaptation. *BMC Plant Biology 2020* 20:1, 20, 1–20.

Jiang, H., Jia, Z., Liu, S., Zhao, B., Li, W., Jin, B. and Wang, L. (2019) Identification and characterization of long non-coding RNAs involved in embryo development of Ginkgo biloba. *Plant Signal Behav*, 14.

Johnson, N.R., Yeoh, J.M., Coruh, C. and Axtell, M.J. (2016) Improved placement of multi-mapping small RNAs. *G3: Genes, Genomes, Genetics*, 6, 2103–2111.

Kang, Y.J., Yang, D.C., Kong, L., Hou, M., Meng, Y.Q., Wei, L. and Gao, G. (2017) CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res*, 45, W12–W16.

Kindgren, P., Ard, R., Ivanov, M. and Marquardt, S. (2018) Transcriptional read-through of the long non-coding RNA SVALKA governs plant cold acclimation. *Nature Communications 2018* 9:1, 9, 1–11.

Kopylova, E., Noé, L. and Touzet, H. (2012) SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, 28, 3211–3217.

Kozomara, A. and Griffiths-Jones, S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res*, 42, D68–D73.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10, 1–10.

Lee, H., Zhang, Z. and Krause, H.M. (2019) Long Noncoding RNAs and Repetitive Elements: Junk or Intimate Evolutionary Partners? *Trends in Genetics*, 35, 892–902.

Lemos, S.M.C., Fonçatti, L.F.C., Guyot, R., Paschoal, A.R. and Domingues, D.S. (2020) Genome-Wide Screening and Characterization of Non-Coding RNAs in Coffea canephora. *Non-Coding RNA 2020*, *Vol.* 6, *Page* 39, 6, 39.

Li, A., Zhang, J. and Zhou, Z. (2014) PLEK: A tool for predicting long noncoding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinformatics*, 15, 1–10.

Li, B., Fillmore, N., Bai, Y., Collins, M., Thomson, J.A., Stewart, R. and Dewey, C.N. (2014) Evaluation of de novo transcriptome assemblies from RNA-Seq data. *Genome Biol*, 15, 1–21.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25, 1754–1760.

Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22, 1658–1659.

Liu, J., Wang, H. and Chua, N.H. (2015) Long noncoding RNA transcriptome of plants. *Plant Biotechnol J*, 13, 319–328.

Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, 15, 1–21.

Lucero, L., Fonouni-Farde, C., Crespi, M. and Ariel, F. (2020) *Long noncoding RNAs shape transcription in plants*. https://doi.org/10.1080/21541264.2020.1764312, 11, 160–171.

Ma, J., Bai, X., Luo, W., Feng, Y., Shao, X., Bai, Q., Sun, S., Long, Q. and Wan, D. (2019) Genome-Wide Identification of Long Noncoding RNAs and Their Responses to Salt Stress in Two Closely Related Poplars. *Front Genet*, 10, 438010.

Mähler, N., Wang, J., Terebieniec, B.K.B.K., Ingvarsson, P.K.P.K., Street, N. R.N.R. and Hvidsten, T.R. (2017) Gene co-expression network connectivity is an important determinant of selective constraint N. M. Springer, ed. *PLoS Genet*, 13, e1006402.

Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R.D. and Califano, A. (2006) ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7, 1–15.

Micheel, J., Safrastyan, A. and Wollny, D. (2021) Advances in Non-Coding RNA Sequencing. *Non-Coding RNA 2021*, *Vol.* 7, *Page* 70, 7, 70.

Nystedt, B., Street, N.R., Wetterbom, A., et al. (2013) The Norway spruce genome sequence and conifer genome evolution. *Nature 2013* 497: 7451, 497, 579–584.

Palazzo, A.F. and Koonin, E. V (2020) Leading Edge Functional Long Noncoding RNAs Evolve from Junk Transcripts. *Cell*, 183, 1151–1161.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A. and Kingsford, C. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods 2017* 14:4, 14, 417–419.

Patturaj, M., Munusamy, A., Kannan, N. and Ramasamy, Y. (2022) Biologia Futura: progress and future perspectives of long non-coding RNAs in forest trees. *Biol Futur*, 73, 43–53.

Qin, T., Zhao, H., Cui, P., Albesher, N. and Xionga, L. (2017) A Nucleus-Localized Long Non-Coding RNA Enhances Drought and Salt Stress Tolerance. *Plant Physiol*, 175, 1321–1336.

Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26, 841–842.

Rich-Griffin, C., Stechemesser, A., Finch, J., Lucas, E., Ott, S. *and Schä Fer, P.* (2019) Single-Cell Transcriptomics: A High-Resolution Avenue for Plant Functional Genomics.

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. and Vert, J.P. (2018) A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications 2018* 9:1, 9, 1–17.

Rosvall, M. and Bergstrom, C.T. (2008) Maps of random walks on complex networks reveal community structure. *Proc Natl Acad Sci U S A*, 105, 1118–1123.

Ruyssinck, J., Huynh-Thu, V.A., Geurts, P., Dhaene, T., Demeester, P. and Saeys, Y. (2014) NIMEFI: Gene Regulatory Network Inference using Multiple Ensemble Feature Importance Algorithms. *PLoS One*, 9, e92709.

Sang, S., Chen, W., Zhang, D., Zhang, X., Yang, W. and Liu, C. (2021) Data integration and evolutionary analysis of long non-coding RNAs in 25 flowering plants. *BMC Genomics*, 22, 1–12.

Santer, L., Bär, C. and Thum, T. (2019) Circular RNAs: A Novel Class of Functional RNA Molecules with a Therapeutic Perspective. *Molecular Therapy*, 27, 1350–1363.

Schäfer, J. and Strimmer, K. (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat Appl Genet Mol Biol*, 4, 1–30.

Schiffthaler, B., Zalen, E. van, Serrano, A.R., Street, N.R. and Delhomme, N. (2023) SeiÃ°r: Efficient calculation of robust ensemble gene networks. *Heliyon*, 9, e16811.

Schurch, N.J., Schofield, P., Gierlin'ski, M., et al. (2016) How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?

Seo, J.S., Sun, H.X., Park, B.S., Huang, C.H., Yeh, S.D., Jung, C. and Chua, N.H. (2017) ELF18-INDUCED LONG-NONCODING RNA Associates with Mediator to Enhance Expression of Innate Immune Response Genes in Arabidopsis. *Plant Cell*, 29, 1024–1038.

Sferra, G., Fratini, F., Ponzi, M. and Pizzi, E. (2017) Phylo_dCor: Distance correlation as a novel metric for phylogenetic profiling. *BMC Bioinformatics*, 18, 1–7.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res*, 13, 2498–2504.

Singh, U., Khemka, N., Rajkumar, M.S., Garg, R. and Jain, M. (2017) PLncPRO for prediction of long non-coding RNAs (lncRNAs) in plants and its application for discovery of abiotic stress-responsive lncRNAs in rice and chickpea. *Nucleic Acids Res*, 45, e183–e183.

Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J.M. and Kelly, S. (2016) TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome Res*, 26, 1134–1144.

Soneson, C., Love, M.I. and Robinson, M.D. (2015) *Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences*. F1000Res, 4.

Stark, R., Grzelak, M. and Hadfield, J. (2019) RNA sequencing: the teenage years. *Nature Reviews Genetics 2019* 20:11, 20, 631–656.

Stojkovič, K., Canovi, C., Le, K.-C., Delhomme, N., Egertsdotter, U. and Street, N.R. (2024) A transcriptome atlas of zygotic and somatic embryogenesis in Norway spruce. *bioRxiv*, 2024.04.13.589382.

Sun, L., Luo, H., Bu, D., Zhao, G., Yu, K., Zhang, C., Liu, Y., Chen, R. and Zhao, Y. (2013) Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res*, 41, e166–e166.

Sundell, D., Mannapperuma, C., Netotea, S., Delhomme, N., Lin, Y.-C.Y.-C., et al. (2015) The Plant Genome Integrative Explorer Resource: PlantGenIE.org. *New Phytologist*, 208, 1149–1156.

Swiezewski, S., Liu, F., Magusin, A. and Dean, C. (2009) Cold-induced silencing by long antisense transcripts of an Arabidopsis Polycomb target. *Nature 2009 462:7274*, 462, 799–802.

Szakonyi, D., Confraria, A., Valerio, C., Duque, P. and Staiger, D. (2019) Editorial: Plant RNA biology. *Front Plant Sci*, 10, 469722.

Szakonyi, D. and Duque, P. (2018) Alternative splicing as a regulator of early plant development. *Front Plant Sci*, 9, 382146.

Tang, W. and Tang, A.Y. (2019) Biological significance of RNA-seq and single-cell genomic research in woody plants. *Journal of Forestry Research 2019* 30:5, 30, 1555–1568.

Tu, Z., Shen, Y., Wen, S., Liu, H., Wei, L. and Li, H. (2021) A Tissue-Specific Landscape of Alternative Polyadenylation, lncRNAs, TFs, and Gene Co-expression Networks in Liriodendron chinense. *Front Plant Sci*, 12, 705321.

Wang, H.L. V. and Chekanova, J.A. (2017) Long noncoding RNAs in plants. *Adv Exp Med Biol*, 1008, 133–154.

Wang, L., Xia, X., Jiang, H., Lu, Z., Cui, J., Cao, F. and Jin, B. (2018) Genome-wide identification and characterization of novel lncRNAs in Ginkgo biloba. *Trees-Structure and Function*, 32, 1429–1442.

Wierzbicki, A.T., Blevins, T. and Swiezewski, S. (2021) *Long Noncoding RNAs in Plants*. https://doi.org/10.1146/annurev-arplant-093020-035446, 72, 245–271.

Wu, H.J., Wang, Z.M., Wang, M. and Wang, X.J. (2013) Widespread Long Noncoding RNAs as Endogenous Target Mimics for MicroRNAs in Plants. *Plant Physiol*, 161, 1875–1884.

Wu, T.D. and Watanabe, C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21, 1859–1875.

Wu, Y., Guo, J., Wang, T., Cao, F. and Wang, G. (2019) Transcriptional profiling of long noncoding RNAs associated with leaf-color mutation in Ginkgo biloba L. *BMC Plant Biol*, 19, 1–13.

Xiao, L., Liu, X., Lu, W., Chen, P., Quan, M., Si, J., Du, Q. and Zhang, D. (2020) Genetic dissection of the gene coexpression network underlying photosynthesis in Populus. *Plant Biotechnol J*, 18, 1015–1026.

Yan, X., Ma, L. and Yang, M.F. (2020) Identification and characterization of long non-coding RNA (lncRNA) in the developing seeds of Jatropha curcas. *Scientific Reports 2020* 10:1, 10, 1–10.

Yu, G., Fu, G., Wang, J. and Zhao, Y. (2018) NewGOA: Predicting New GO Annotations of Proteins by Bi-Random Walks on a Hybrid Graph. *IEEE/ACM Trans Comput Biol Bioinform*, 15, 1390–1402.

Zamora-Ballesteros, C., Martín-García, J., Suárez-Vega, A. and Diez, J.J. (2022) Genome-wide identification and characterization of Fusarium circinatum-responsive lncRNAs in Pinus radiata. *BMC Genomics*, 23, 1–19.

Zhang, P., Li, S. and Chen, M. (2020) Characterization and Function of Circular RNAs in Plants. *Front Mol Biosci*, 7, 539771.

Zhang, X., Liu, K., Liu, Z.P., Duval, B., Richer, J.M., Zhao, X.M., Hao, J.K. and Chen, L. (2013) NARROMI: a noise and redundancy reduction technique improves accuracy of gene regulatory network inference. *Bioinformatics*, 29, 106–113.

Zhao, L., Zhang, H., Kohnen, M. V., Prasad, K.V.S.K., Gu, L. and Reddy, A.S.N. (2019) Analysis of transcriptome and epitranscriptome in plants using pacbio iso-seq and nanopore-based direct RNA sequencing. *Front Genet*, 10, 430951.

Zhong, R., Allen, J.D., Xiao, G. and Xie, Y. (2014) Ensemble-Based Network Aggregation Improves the Accuracy of Gene Network Reconstruction. *PLoS One*, 9, e106319.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Canovi, C., Stojkovič, K., Benítez, A.A., Delhomme, N., Egertsdotter, U. & Street, N.R. (2024) A resource of identified and annotated lincRNAs expressed during somatic embryogenesis development in Norway spruce. *Physiologia Plantarum*, 176(5), e14537. Available from: https://doi.org/10.1111/ppl.14537