Norwegian University
of Life Sciences

Master's Thesis 2019 60 ECTS
Department of Chemistry, Biotechnology and Food Science

Comparison Between Gene Expression and Protein Abundance in *Populus tremula* Wood Development

Eivind Kjeka Broen
Bioinformatics and applied statistics

Acknowledgements

This thesis is written as a part of my master's degree in bioinformatics at the Norwegian University of Life Sciences, Department of Chemistry, Biotechnology and Food Sciences. The work with this thesis was carried out from fall 2018 to spring 2019.

I would like to thank my supervisor Torgeir Rhoden Hvidsten for introduction to relevant topics, helpful advice and guidance. I would also like to thank Francesco Delogu and Lish Earnest for reading through my thesis and giving valuable comments.

Ås May 2019

Eivind Kjeka Broen

Abstract

This thesis compares the gene expression and protein abundance across a series spanning the wood forming developmental stages in *Populus tremula* (common aspen): phloem, cambium and xylem. The comparison was based on two data sets provided by Obudulu et al. (2016) and Sundell et al. (2017).

Data treatments, such as moving average calculation, successfully elevated the subpar proteomics data set and improved its correlations with the transcriptomics data set.

Correlation coefficients were calculated between the two full data sets (full correlation), by gene and corresponding protein (row correlation) and by sample number in the series (correlation by sample). The full correlation yielded correlation coefficients ranging from 0.256 to 0.347 based on the extent of data treatments. The moving average treated summed isoform data correlated with the corresponding transcript yielded a correlation coefficient of 0.395. The correlation by sample suggested that there were more post-transcriptional regulations in samples in the phloem and the late xylem than in the other samples.

By comparing presence of molecules in the two data sets it was found that in 20% of the entries, both protein abundance and gene expression above 0 were found. In 3.3% of the entries, both protein abundance and gene expression were 0. In 76% of the entries, gene expression was above 0, while protein abundance was 0. In 0.18% of the entries, protein abundance was above 0, while the corresponding gene expression was 0. This indicated that protein abundance is strongly dependent on presence of gene expression. It was also shown that the likelihood of protein abundance in an entry increased significantly with increased levels of gene expression.

By superimposing the protein abundance series on the gene expression series for single genes, dynamics between the transcripts and the proteins were revealed. Most notably delays between transcription and translation between some proteins and genes and “translation on demand” relationship between some other proteins and genes.

GO enrichment analysis of proteins, which protein abundance series correlated well with their corresponding gene expression, was performed. The enrichment indicated that that many GO terms may be related to proteins that are easier to study with certain protein profiling methods.

Sammendrag

Denne masteroppgaven sammenligner genuttrykk og protein mengder i en serie som spenner over seksjoner i *Populus tremula* (osp) som danner ved: silvev, kambium og vedvev. Sammenligningen var basert på to artikler skrevet av Obudulu et al. (2016) og Sundell et al. (2017).

Data behandlinger, som for eksempel «moving average» beregning, forbedret det mangelfulle proteomikk datasettet og forbedret settets korrelasjon med transkriptomikk datasettet.

Korrelasjonskoeffisienter ble beregnet mellom de to hele datasettene («full correlation»), ut ifra gen og tilsvarende protein («row correlation») og ut ifra prøvenummer i tidsseriene («correlation by sample»). Korrelasjonen mellom de hele datasettene ga korrelasjonskoeffisienter imellom 0,256 og 0,347 basert på omfanget av databehandlinger. Korrelasjonen mellom det «moving average» behandlede datasettet summert ut ifra isoformer og den tilsvarende transkriptomikk datasettet var 0,395. Korrelasjonen basert på prøvenummer indikerte at det var flere post-transkripsjonelle reguleringer i prøvene i silvev og sent i vedvev enn i de andre prøvene.

Ved å sammenligne forekomst av molekyler i de to datasettene ble det funnet ut at i 20% av oppføringene ble det funnet både protein mengder og genekspresjon i verdier over 0. I 3,3% av oppføringene var både protein mengder og genuttrykk 0. I 76% av oppføringene var genuttrykk over 0 mens protein mengder var 0. I 0,18% av oppføringene var proteinmengden over 0 mens det tilsvarende gen ikke ble uttrykt. Dette indikerer at forekomst av proteinmengde er avhengig av forekomst av genuttrykk. Det ble også vist at sannsynligheten for forekomst av proteinmengder i en oppføring økte i betydelig grad med økt genuttrykk.

Ved å sammenligne proteinmengdeseriene med genuttrykksseriene for spesifikke gener, ble spesielle dynamikker mellom transkripsjon og protein tydeliggjort. Spesielt forsinkelse mellom transkripsjon og translasjon mellom noen proteiner og gener, og "translation on demand"-forhold mellom noen andre proteiner og gener.

“GO enrichment"-analyse av proteiner, som hadde proteinmengdeserie som korrelerte godt med deres tilsvarende genuttrykk, ble utført. Analysen indikerte at mange GO-termer kan være relatert til proteiner som er enklere å studere med proteinprofileringsmetoder brukt her.

Contents

1 Introduction

To what extent are protein abundance levels dependent on gene expression levels in different developmental stages, such as xylem and phloem in aspen trees? Developments in Next-generation sequencing and protein identification techniques have enabled high-quality analyses of the genome, transcriptome, and proteome of cells (Steen & Pandey, 2002).This has provided an opportunity for analyzing and comparing all these together to observe the relationship between them.

This thesis will explore the relationship between mRNA and protein expression across different developmental stages in wood tissues of common aspen (*Populus tremula* in Latin). The basis for this thesis was two articles, the first providing expression profiles of proteins (Obudulu et al., 2016) and the second with gene expression profiles (Sundell et al., 2017) both from the same tissues in *Populus tremula*. *Populus tremula* is especially interesting because of the wood's usefulness in renewable energy and the production of goods and because it is one of the most important carbon sinks in northern Europe and Asia. Aspen trees are becoming the model organism for woody plants due to characteristics that makes them suitable for research (Hertzberg et al., 2001) and because of the full genome sequencing of *Populus trichocarpa (Jansson & Douglas, 2007).*

The matching data sets provide the opportunity to compare transcriptomics and proteomics in the same samples across a continuous spatial series with differential expression in the various developmental stages of the tree, enabling insight into post-transcriptional regulation and modification and strengths and weaknesses of both techniques. By comparing continuous patterns of gene expression and protein abundance phenomenon like "translation on demand" and the delay between transcription and translation may be uncovered (Liu et al., 2016).

Since gene expression and protein abundance rarely correlate perfectly, proteomics is an increasingly useful field (Liu et al., 2016). The comparison between mRNA and protein in a single cell or a series may uncover details of their relationship. Translation is regulated by many factors and different proteins relate to mRNA in different ways. Furthermore, through closer inspection of the proteins that correlate well with their transcript, it is possible to identify characteristics of the proteins that are more detectable by proteomics methods.

In this thesis the relationship between gene expression and protein abundance in a specific wood formation series in *Populus tremula* will be studied through several methods in the coding language R. Previous studies have compared gene expression and protein abundance; however, the two overlapping data sets provides a unique opportunity to study the relationship between proteins and mRNA in different wood forming cells. This thesis has conducted and will discuss varied analyses, including correlations, clustering and Gene Ontology enrichment analysis. Since the proteomics data provided was limited compared to the transcriptomics data, data treatments such as moving average calculations and transformations were conducted to diminish the effects of uneven and lacking protein abundance. The effect of these were evaluated.

2 Theory

2.1 The Aspen Tree

Populus tremula (common aspen), is native to the colder regions of Europe. It is an angiosperm (flowering) tree. Aspen is widely distributed across Europe and Asian Russia (Figure 1) and trees represent one of the major CO2 sinks on earth (Sundell et al., 2017). Its wood is an important resource and can be used as lumber/renewable energy or in the production of paper, plywood and matchsticks. The usefulness of the *Populus tremula* as a carbon sink and wood producer could be increased by developing elite varieties (Sundell et al., 2017).

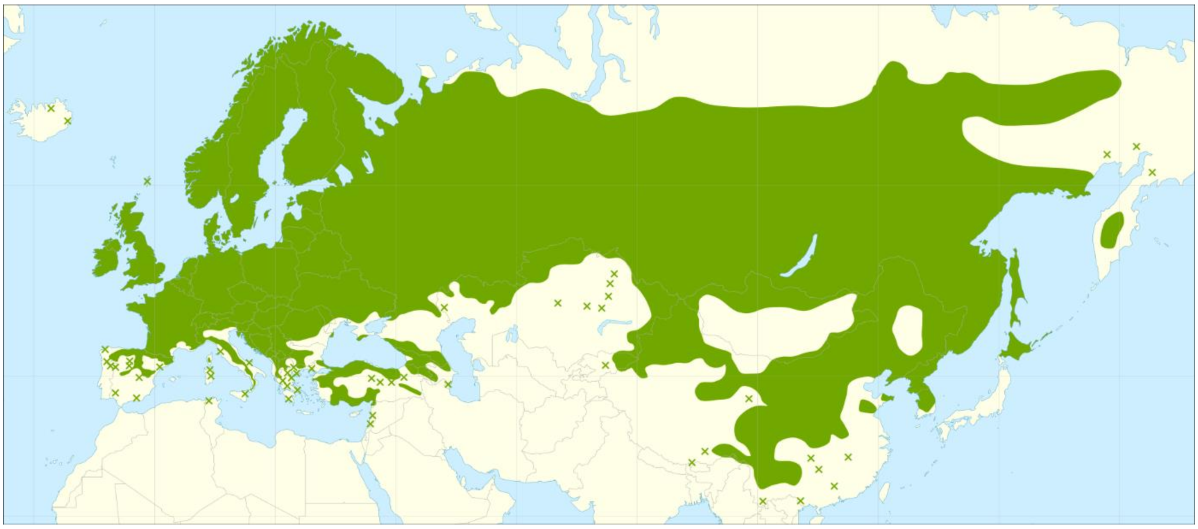

Figure 1: The distribution of Populus tremula (Caudullo et al., 2017).

While being a species that is geographically widespread and that provides a useful type of wood, *Populus tremula* also has aspects making it a good candidate for research. Aspen trees have a physically large meristem and the different developmental stages of the trees are easily distinguished (Hertzberg et al., 2001).

2.2 Tissues in Trees

Trees are usually defined as woody plants with secondary growth. Primary growth is growth at the meristem of the plant, which provides elongation in the branches and roots. Secondary growth is a lateral growth procured by cell division by the cambium. Through lateral growth, trees can produce wood that provides rigidity and support, enabling the trees to grow higher and sturdier than other plants (Thomas, 2003).

Wood is a rigid, porous, organic material used to maintain the structure of trees (Thomas, 2003). It is created by xylem cells depositing cellulose and lignin fibers into the secondary cell wall (Mellerowicz et al., 2001). Vascular cambium initiates wood development. The cambium is the cylindrical sheet of cells that divides, creating new partially undifferentiated cells for plant growth. Stem cells located here divide into more specialized cells in both directions. It forms parallel rows of cells, which result in secondary tissues, cork cambium and vascular cambium (Thomas, 2003). Outwards cells gradually differentiate into phloem cells and inward cells specialize into xylem cells (Figure 2), the two vascular tissues plant that are used to transport fluids and nutrients across the (Hertzberg et al., 2001). In trees like the *Populus tremula* the inward secondary tissues become woody and intermingled with cellulose fibers embedded in a lignin network.

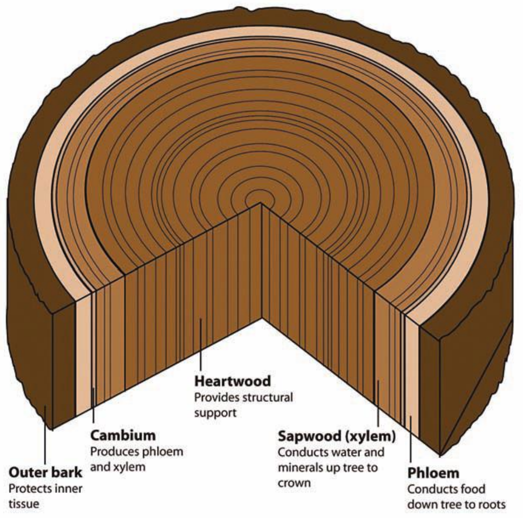
Heartwood
Provides structural support
Cambium
Produces phloem and xylem
Sapwood (xylem)
Conducts water and minerals up tree to crown
Outer bark
Protects inner tissue
Phloem
Conducts food down tree to roots

Figure 2: Illustration of the cross-section and location of xylem and phloem in trees (Hood, 2010).

Xylem's main task is to transport water and micronutrients from the roots and provide structural support (Thomas, 2003). They are water conducting cells that are elongated, thin and rigid. The xylem cells gradually become the long, dead tubes that transport water passively through capillary mechanisms, as shown in Figure 3 (Bollhoner et al., 2012). The differentiation stages of xylem cells are in general: cell division in the cambium, cell expansion, deposition of the secondary cell wall and cell death (Fukuda, 1996). In angiosperm trees, like *Populus tremula*, the xylem consists of both vessel elements and tracheid elements (Bollhoner et al., 2012). Vessel elements and tracheid elements are intermixed with fiber cells to increase rigidity and mechanical strength (Figure 4). Fiber cells do not transport water or nutrients but provide structural support long after their death.

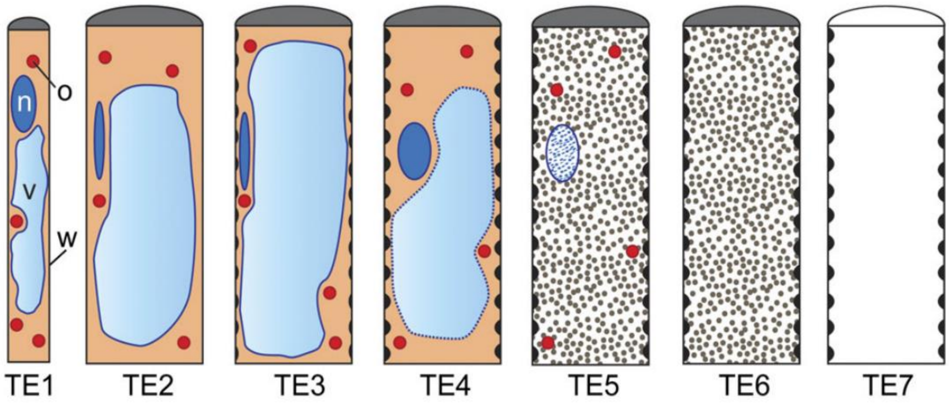

Figure 3: Illustrations of the differentiation of xylem tracheary elements from an article on xylem cell death by Bollhoner et al. (2012). In the figure: n indicates nucleus, v indicates vacuole, o indicates organelles and w indicates cell wall. It shows from left to right differentiation, expansion, secondary cell wall formation, vacuole rupture, degradation of DNA, final enzymatic breakdown and partial breakdown of primary cell walls.

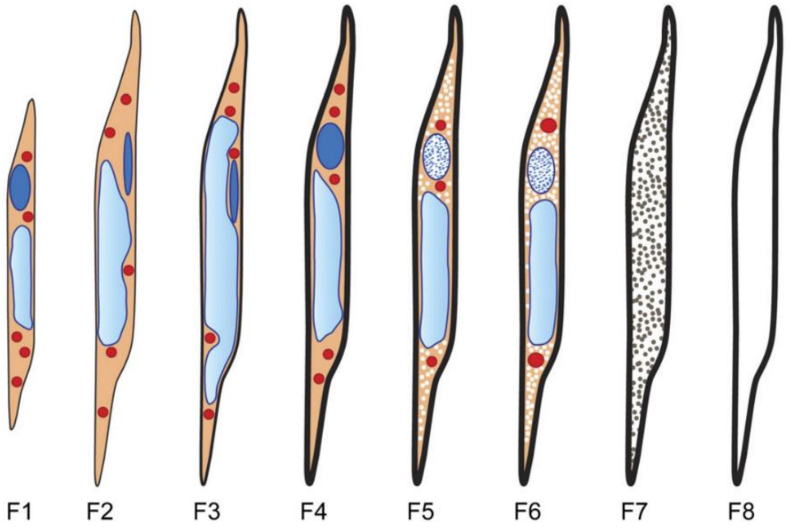

Figure 4: Illustrations the differentiation of fiber cells from the article on xylem cell death by Bollhoner et al. (2012). Shows from left to right, cambium differentiation, cell expansion, secondary wall formation, loss of turgor in the vacuole, breakdown of organelles, breakdown of DNA and proteins, swelling of organelles, vacuole rupture and continued autolysis, cleared cell.

Phloem transports photosynthates, which are the soluble products produced in the leaves during photosynthesis (usually sugars). The phloem tissue layer lies just under the bark, far from the center of the stem (Pate & Atkins, 1983).

2.3 The Flow of Genetic Information

“The central dogma of molecular biology” is a framework for illustrating the flow of information in genetics, how different macromolecules may transfer information between each other. It states that the main flow of genetic information goes from DNA to RNA to protein and that there are other special transfers such as RNA to DNA (Crick, 1970). DNA also replicate itself to create new DNA in a process called replication.

Regions in the DNA called genes, are transcribed into RNA by RNA polymerase with the aid of several other proteins which can enable or increase the transcription rate. In eukaryotes a wide array of transcription activators binds to enhancer sites in the DNA and together they determine the rate of RNA transcription. Through a mediator they assemble with the RNA polymerase at the promoter region of the gene. The raw mRNA needs to be processed before entering cytosol for translation. The raw mRNA is processed into mature mRNA through splicing. During splicing, introns are removed, and the remaining exons are pasted together. The exons could be pasted together in alternative ways, thus giving rise to different types of mRNA (splice variants) which in turn translate into different proteins (isoforms). The messenger RNA is transported out of the cell to be translated by ribosomes. The end products are proteins, which are macromolecules that provide the majority of functions for living organisms, including enzymatic catalysis and structural support and movement in the cells (Alberts, 2014).

There are many factors that complicate this linear formula of how a DNA strand encodes an RNA strand, which in turn translates into a protein. Proteins, genes and mRNA are intricately interconnected and may affect each other in several ways, therefore there is almost never a one-to-one relationship between mRNA and its corresponding protein in a cell. Processing of the RNA may determine how many proteins will be translated before it is degraded. In many cases, the mRNA may degrade before it even translates a single protein (Alberts, 2014). Factors such as RNA interference may inhibit either transcription or translation. Proteins, like transcription factors, may affect transcription rates. Other proteins may change the structure and function of proteins. Prions may even misshape other protein in a way that can be replicated to other proteins.

2.3.1 Post-Transcriptional Regulations

Post-transcriptional regulations are all the regulatory mechanisms performed after the transcription of RNA and is a crucial part of gene regulation. It can infer the correct level of translation or even abort it entirely. Post-transcriptional regulations are especially important for quick adaptation of cells in new environments or into new roles (Liu et al., 2016). In many cases, stopping and degrading all mRNA molecules would be a quite slow process and quick post-translational mechanisms enable a more dynamic control of protein expression.

“Translation on demand” is a term describing situations where mRNA is transcribed regularly and protein is expressed only when needed (Liu et al., 2016). Transcript is always readily available in cytosol available for translation, and when the protein is needed, the necessary signal is sent initiating the translation of the protein in demand.

Protein translation can be terminated, up-regulated or down-regulated at many stages between RNA transcription and protein synthesis. Splicing regulates which protein isoform will be expressed before it exits the nucleus. Control of the mRNA abundance in cytosol can be done by regulating degradation of mRNA. For instance, long polyadenylated tails often means long half-lives. Upstream Open Reading Frames provide translation control. They are open reading frames in the leader sequence, the untranslated region of the mRNA upstream of the initiation codon. mRNA may have a hairpin structure downstream of the uOPS that terminates translation when translation is initiated at the uOPS (Wethmar et al., 2010). Internal ribosome entry sites are located on the mRNA and recruits the ribosome to initiate translation. Furthermore, translation itself can be modulated. Proteins can bind to regulatory elements on the mRNA molecule inhibiting synthesis. In many cases protein synthesis can be suppressed while mRNA is being expressed, which can lead to detection mRNA, but not of proteins.

2.3.2 Protein Isoforms

Protein isoforms are proteins with the same genetic origin, either from the same gene or gene family (Stastna & Van Eyk, 2012). One single gene can produce proteins that are different in folded structure and composition of amino acids and domains. Different isoforms of a protein often arise from splicing, variable promoter usage or other post-transcriptional modifications. Splicing is the main post-transcriptional process that produces protein isoforms. In eukaryotes, differential splicing can result in different types of proteins by removing introns and assembling exons in different ways. The splicing process occurs during or after transcription in the nucleus and the mature mRNA exits to the nucleus. While the resulting isoform proteins are related and usually function similarly, isoform proteins may also have vastly different structures and functions.

In practice the process of distinguishing between different protein isoforms is difficult. Protein isoforms have similar sequence and may be difficult to tell apart when the protein is fragmented into peptides (Stastna & Van Eyk, 2012).

2.3.3 Post-Translational Regulations

Post- translational regulations of proteins are common in cells. To save resources, amino acids can be made available from proteins through protein hydrolysis. Which is a non-reversible form of protein regulation. One method used for intracellular proteins is ubiquitination, where several ubiquitins are attached to a protein and the protein is subsequently degraded by a proteasome complex (Alberts, 2014). The ubiquitinated proteins and amino acids are released and can be reused (Glickman & Ciechanover, 2002). Reversible post-translational regulation is possible, and these are called post-translational regulation. Histone modifications where the histone tails are phosphorylation, methylation and acetylation to control the openness of the chromatin is an example of post-translational regulation. The openness of chromatin enables transcription of RNA, which is important transcriptional regulation. This makes histone modification both an example of post-translational modification of proteins and a transcriptional regulation of genes. In the end, protein levels are not only dependent on gene expression levels, but also translation regulation and post-translational regulation (Steen & Pandey, 2002).

2.4 RNA-Sequencing

Through RNA-seq one may study the transcriptome of a cell or a sample at a given time. The transcriptome is defined as "*the complete complement of mRNA molecules generated by a cell or population of cells"* (McGettigan, 2013). The typical RNA-seq process first isolates the RNA molecules from the cell and then selects a subset of the RNA, for example mRNAs. If long coding RNA is going to be sequenced, the molecules need to be fragmented by the shotgun method before it is reverse transcribed to cDNA (alternatively the whole RNA is reverse-transcribed into cDNA and then fragmented) (Hrdlickova et al., 2017). The cDNA is afterwards sequenced by a next-generation sequencer. It could be sequenced by a single end method or a pair end method. Pair end methods sequences the fragment from both ends, and this could be useful in detecting paralogs with similar sequences. The reads, the products of the sequencing, are digitally trimmed for areas of high error. After trimming, the RNA-Seq reads can be de novo assembled to a reference transcriptome or mapped to an existing transcriptome or annotated genome (Kukurba & Montgomery 2015).

RNA-Seq uses next-generation sequencing (NGS) to directly sequence RNA from cDNA, the technique has many benefits. RNA-seq does not rely on a corresponding genomic sequence. RNA-seq is especially attractive for organisms without a reference genome since it is possible to assemble the reads de novo (Wang et al., 2009). Either way, it is possible to find novel genes through RNA-seq. In the transcriptomics article by Sundell et al. (2017), 78 novel genes were found. RNA-seq has less background noise as it maps RNA to its site unambiguously. Additionally, RNA-seq is the favored method of measuring expression levels and is highly reproducible (Wang et al., 2009). RNA-Seq can be applied to all RNA in the cell, not just protein-coding transcripts (Kukurba & Montgomery, 2015).

RNA-Seq may be used to determine the structure and locations of splice sites and how exons are connected (Liu et al., 2016). Reads of can be mapped to a reference genome. This way the exons and introns in the genes can be identified, as intron areas will have less mapped reads than exon sites. It might also be possible to identify the exact splice sites as the introns are spliced at a specific base sequence: GU at the 5' splice site and AG at the 3' splice site. “Mate pair” is a method usually used to mitigate challenges due to long repeats when sequencing DNA. If uilized in RNA-Seq, the distance between the pair ends will be known and therefore the size of the gap can be determined. In RNA-seq pair end sequencing could be useful for de novo sequencing or to better understand the exon-intron structures of the RNA.

RNA-seq may identify putative genes. Putative genes are predicted to be genes based on their open reading frame. While the putative genes have an open reading frame, they have no corresponding identified protein. For additional reliability, they are required to share sequence similarity with other identified genes.

RNA sequencing has some challenges. Fragmentation of larger molecules is necessary for most sequencers as many of them have a maximum read length. The error rate usually increases towards the end of the reads (Del Fabbro et al., 2013). Longer sequences increase the risk of going “out of phase”, which is when a base pair is skipped, and this leads to incorrect sequenced bases in the rest of the read. The fragmentation method might be biased and may fragment the molecules in a non-random way. Polymerase chain reaction (PCR) artifacts could be another challenge where one RNA fragment has an unconventionally large number of duplicates (Acinas et al., 2005). The most relevant challenge is however that the relationship between mRNA and protein is not 1:1 and the actual relation between them is not fully determined, RNA-seq cannot accurately predict the corresponding protein content in the cell. Therefore, protein profiling is necessary to obtain the full picture.

RNA-seq can be used to investigate how cells of the same species express RNA in different situations. The purpose of RNA-seq in the transcriptomics article by Sundell et al. (2017) was to observe how the genes are differentially expressed in different developmental stages in Populus tremula, especially in wood formation areas of the tree. Since the genome is mostly identical for all cells in the same organism, DNA sequencing may not be used for elucidating differences between different cells from the same individual in the same way as RNA-Seq can. Whether an RNA molecule will be transcribed, and in what quantity, is reliant on genes, epigenetics (the heritable changes that does not involve alterations in DNA) (Dupont et al., 2009) and on external stimulus. These provides cells in different developmental stages with different tasks and unique phenotypes (Kukurba & Montgomery, 2015).

2.5 Proteomics

Proteomics is the study of function and structure of proteins on a large scale (Chandramouli & Qian, 2009). The proteome is defined as the whole set of proteins found in a system or organism. As information on the function of a cell or an organism is difficult to elucidate only based on genes and mRNA, proteomics is a useful and increasingly necessary field of research. For each gene in the genome, there may be several distinct proteins and these proteins may have many different functions. Recently, many new techniques have been developed to enable detection, identification and quantification of proteins.

Protein profiling is the identification and quantification of the total protein content in a tissue or cell at a specific time. Often the protein profile is assigned to a reference genome (Graves & Haystead, 2002). There are many methods available for conducting protein profiling. The basic common process includes preparation of the protein by splitting it into peptide fragments and quantifying them, and thereafter identifying these proteins based on the peptide sequences, the latter is usually done with the help of computer technology. Optionally the resulting identified proteins may be compared to a reference genome.

Mass spectroscopy is often used to identify and quantify peptides and it is a key tool in proteomics. A mass spectrometer has three main components: the ion source, the mass analyzer and a detector that registers the number of ions per m/z level (Han et al., 2008). The ion source ionizes the sample by bombarding it with electrons to produce gas-phase ions. These ions are separated based on mass-to-charge (m/z) ratio, the ions are detected and quantified based on this ratio. (Han et al., 2008). Mass spectroscopy supports both relative and absolute protein measurement on a large scale, without need of generating antibodies (Liu et al., 2016). Antibodies are used in immunoassays, where they bind to specific macromolecules and gives off a detectable signal (Wingren 2016) .

Due to the many complicated steps, protein profiling is prone to errors and mass spectrometry techniques are usually the main bottleneck of the process (Chandramouli & Qian, 2009).

2.6 Relationship Between Gene Expression and Protein Abundance

Protein abundance in a cell is dependent on many factors and the existence of mRNA is likely the main one. According to the "the central dogma of molecular biology" proteins must be transcribed by RNA or be brought in from outside the cell. Calculating the variance explained in protein abundance by the variance in gene expression using Pearson's correlation score squared is a common way to quantify associations between mRNA and protein (explained chapter 3.4.4). The calculated score is often very different depending on the way the data is obtained, matched and treated and depending on the organism providing the data. One study calculated the association between mRNA and protein expression in mouse dendritic cells (Jovanovic et al., 2015). The correlation showed that 27% of the variation in protein levels was explained by the raw mRNA data. Through different data analysis strategies that score increased to 52%. A study of mammalian cells indicated that the variance in protein abundance explained by gene expression was 40% (using Pearson's correlation coefficient) (Schwanhausser et al., 2011) and a restudy using a different model concluded that the variance explained was between 56%–84% (Li et al., 2014). Another study analyzed gene expression and protein abundance in *Saccharomyces cerevisiae,* which showed that the variance in gene expression levels explained 80% of the variation in protein levels in yeast (Lee et al., 2011).

Measurements are often divided into absolute and relative quantities. Absolute quantities, meaning the actual number of mRNA or protein in the cell in question (or a quantity that reflects this number). Relative quantities are dependent on other quantities for reference. For example, gene expression is calculated as the number of reads aligned to a gene divided by the total number of reads sequenced in the sample. The type of data used for correlation is important to consider when analyzing the relationship between mRNA levels and protein levels as not all data types scale and since absolute quantities of protein and mRNA is not interchangeable with relative quantities.

It is theorized that in a cell in steady-state, mRNA quantitates explain protein levels rather well (Liu et al., 2016) (Figure 5 A). In the study on mammalian cells by Schwanhausser et al. (2011), 40% of the variance in protein levels were due to variations in mRNA levels in steady-state cells. A cell in steady-state has a degradation rate that is approximately equal to the synthesis rate for proteins and mRNA. However, a cell is rarely in a perfect steady-state. Protein levels and gene expression levels in the cell are in constant fluctuation. The cell responds to stimuli such as nutrients, cell signals, chemicals etc. These stimuli may initiate transcription or translation to prepare the cell for new environments or a new role.

Gene expression and protein abundance at one moment in a cell may not correlate well due to a delay between mRNA transcription and protein translation (Figure 5, B). Before mRNA may translate a protein, it undergoes maturation processes including splicing, 5′ capping, 3′ cleavage and polyadenylation. The synthesized protein may be transported to other locations than where the mRNA was sampled for example outside the cell. The translation process itself is in some cases very slow. Translation can be regulated by upstream open reading frames (Wethmar et al., 2010) and internal ribosome entry sites (Liu et al., 2016) altering the rate of translation.

Protein abundance are dependent on many biological factors. Protein lifespan can be affected by post-translational regulation. Some protein is quickly degraded and therefore will be in lower numbers compared to proteins that are translated at the same rate but have longer lifespans. Proteins with signal delay between transcription and translation and “Translation on demand” proteins usually also correlate badly with gene expression as the mRNA will always be expressed and the protein only when necessary (Figure 5, B, C).

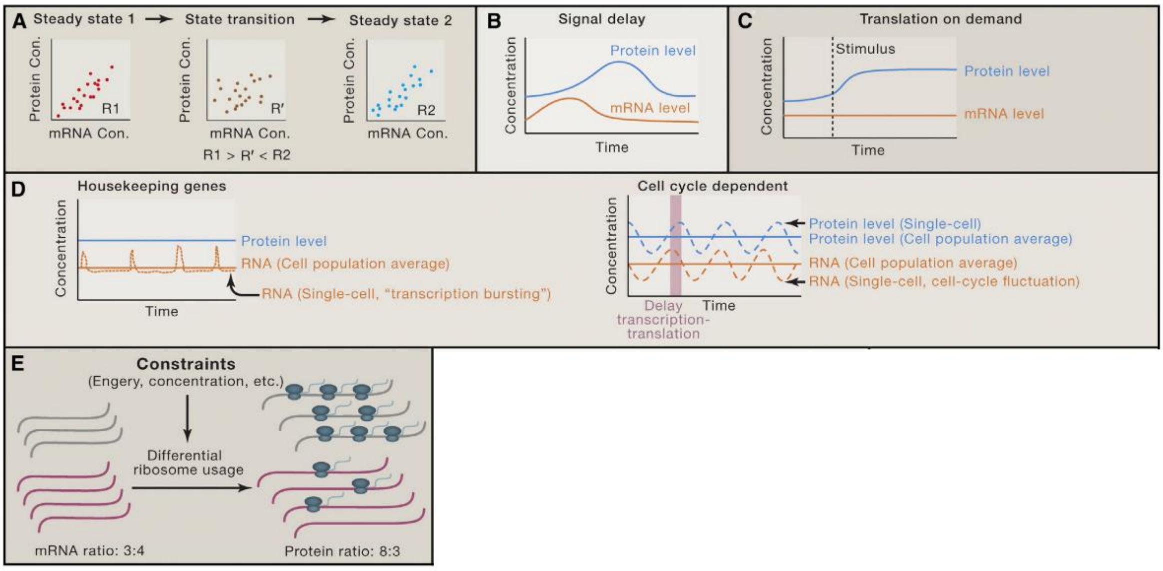
A
Steady state 1
State transition
Steady state 2
Protein Con.
mRNA Con.
R1
R′
R2
R1 > R′ < R2
B
Signal delay
Concentration
Protein level
mRNA level
Time
C
Translation on demand
Stimulus
Protein level
mRNA level
Time
D
Housekeeping genes
Protein level
RNA (Cell population average)
RNA (Single-cell, "transcription bursting")
Cell cycle dependent
Protein level (Single-cell)
Protein level (Cell population average)
RNA (Cell population average)
RNA (Single-cell, cell-cycle fluctuation)
Delay transcription-translation
E
Constraints
(Engery, concentration, etc.)
Differential ribosome usage
mRNA ratio: 3:4
Protein ratio: 8:3

Figure 5: The relationship between mRNA and protein under different dynamics by Liu et al. (2016). Delayed synthesis between mRNA between steady states (A), mRNA is first produced in response to a signal (B), translation on demand where mRNA is stable and translation levels are increased due to a signal (C), housekeeping genes and difference due to cell cycle stages (D), energy levels and ribosome numbers affect the translation capabilities and priorities of the cell (E).

Translation rates can vary and are dependent on many factors, but to synthesize proteins mRNA needs to be present. Cells that newly synthesized a protein should have the corresponding mRNA present. Therefore, instead of a strict numeric relation between protein and gene expression, one study has suggested that expression of mRNA could rather be treated as an on-off switch rather than numeric relation to each other (Vogel & Marcotte, 2012). This study of yeast indicated that if mRNA expression is over a certain threshold is a much higher chance of the corresponding protein being present (Figure 6).

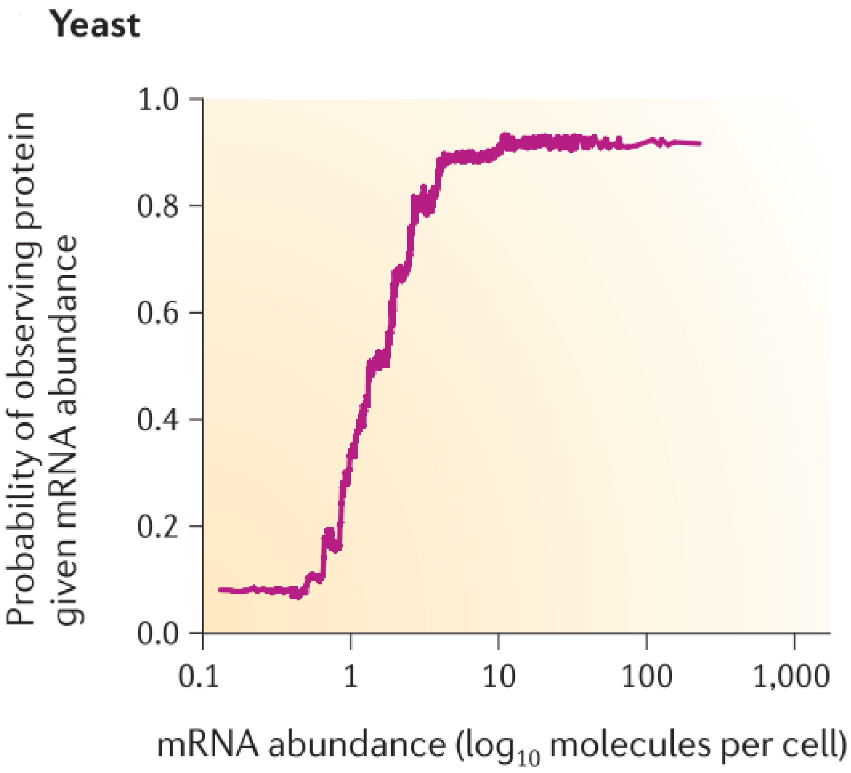

Figure 6: The relationship between mRNA abundance and the likelihood of observing protein abundance in yeast (Vogel & Marcotte, 2012).

2.7 Proteins of Interest

Presented here are some proteins of special interest in the context of wood development. Some of these proteins define the cell's developmental stage. Sucrose synthase is especially important in phloem. Cellulose synthase is found in large relative quantities in developmental stages where the cell wall is deposited. Peptidases are important in programmed cell death, especially for the later xylem as xylem tissues undergo apoptosis during maturation.

Sucrose synthases reversibly catalyze/cleaves sucrose from glucose and fructose (Zheng et al., 2011). The process is nearly energy neutral. Sucrose is highly mobile in plants and is the main soluble component of phloem sap in many trees. Since sucrose is non-reducing and not prone to metabolism compared to glucose and many other sugars, it is the selected carbohydrate for transportation around the tree (Lemoine, 2000).

Cellulose synthases catalyze the reversible production of cellulose mainly from beta-1,4-linked glucose residues. It is a large family of proteins and many different types with little genetic relation to each other seem to occur in any higher plant, indicating that they are conserved. In *Arabidopsis thaliana* ten different types of cellulose synthases are found

(Richmond, 2000). Cellulose is a polysaccharide consisting of glucose units bound together with a beta one to four position bond. The molecule may be up to several thousand units long. The secondary cell wall receives their structure through cellulose and bound together with lignin they are the main component of the wood tissues in trees.

Programmed cell death is a process necessary in both maturation of fiber cells and maturation of xylem tissues. The components needed to facilitate cell death are often procured early in xylem differentiation. The components are prohibited from initiating cell death by inhibitors or they may be stored vacuoles until needed. The timing of programmed cell death is different for each cell in the xylem developmental stages based on what function the cell will fulfill in the mature xylem (Bollhoner et al., 2012).

3 Material and Methods

3.1 The Data

In this thesis, data from two studies were used. The first data set was a transcriptomics data set consisting of identified gene transcript and their expression estimated using the VST (variance stabilized transformation) method across 106 samples from 4 different trees (Sundell et al., 2017). VST is used on data to either simplify presentation or some statistical procedures like regression or ANOVA. In some data sets, the statistics are determined mainly by the largest values. This can usually be solved by utilizing a log transformation, but the adverse effect of this is that the variance of lower values becomes disproportionately large. VST is a calculation that aims to normalize the variance of the lower values while scaling the higher values (Love et al., 2015). The other data set was a proteomics data set consisting of identified proteins and their expression profile for 111 samples across four trees (Obudulu et al., 2016). Both data sets were extracted from cryosections from the same four trees. All measures of gene expression in plots and tables are VST values.

The trees were four mature, wild *Populus tremula* growing in an uncontrolled environment in Northern Sweden. Cross-sections were from each stem about 3 meters above the soil level. The samples encompass the phloem through the vascular cambium to the xylem within one growth ring (Obudulu et al., 2016).

The samples in the data set can loosely be separated into four different developmental stages. The samples from the transcriptomic data were clustered into four developmental stages, roughly corresponding to those shown in Figure 7. The separations between these stages were characterized by three transcriptome reprogramming events according to the transcriptomics article by Sundell et al. (2017). These reprogramming events mark positions in the series where the cells are in the process of becoming distinctly differentiated. The first reprogramming event was in the middle of the specialization of cells into the phloem and xylem cells, i.e. The vascular cambium. Therefore, mainly phloem cells were in the first developmental stage and xylem cells in the second developmental stage. The second reprogramming event marked the end of stem cell expansion and where the cell begins depositing the secondary cell wall. The third and last reprogramming marks the end of secondary cell wall deposition and the start of apoptosis and transformation of cells into dead wood tissue (Sundell et al., 2017). The samples right of cambium mainly contains xylem tissues at different developmental stages. The proteomics article by Obudulu et al. (2016)

divided the samples into phloem, cambium, expansion and xylem. Xylem was further divided into four stages. An overview of the stages is given in Figure 8. In this thesis the separation of the samples in the series will mostly be based on how the transcriptomics article by Sundell et al. (2017) separated the samples. The developmental stages featured in this thesis are phloem (sample 1 to 5), expanding xylem (sample 6 to 12), secondary cell wall (SCW) forming xylem (sample 13 to 19) and late xylem (sample 20 to 25).

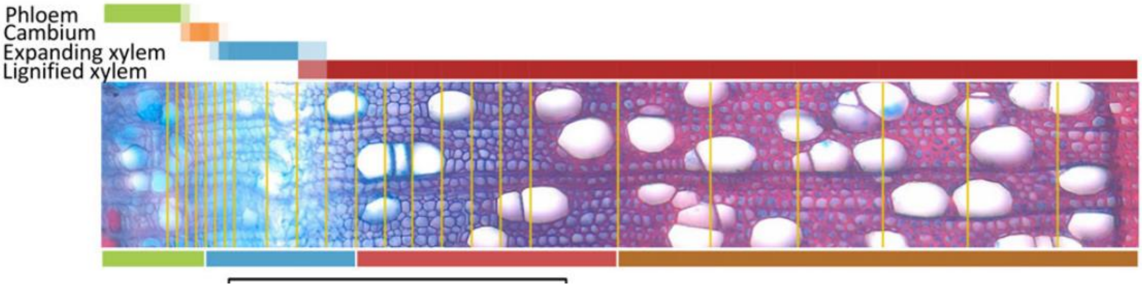

Figure 7: Illustration of the different sections of the cryosection from the transcriptomics article by Sundell et al. (2017). The developmental stages were identified by microscope observation.

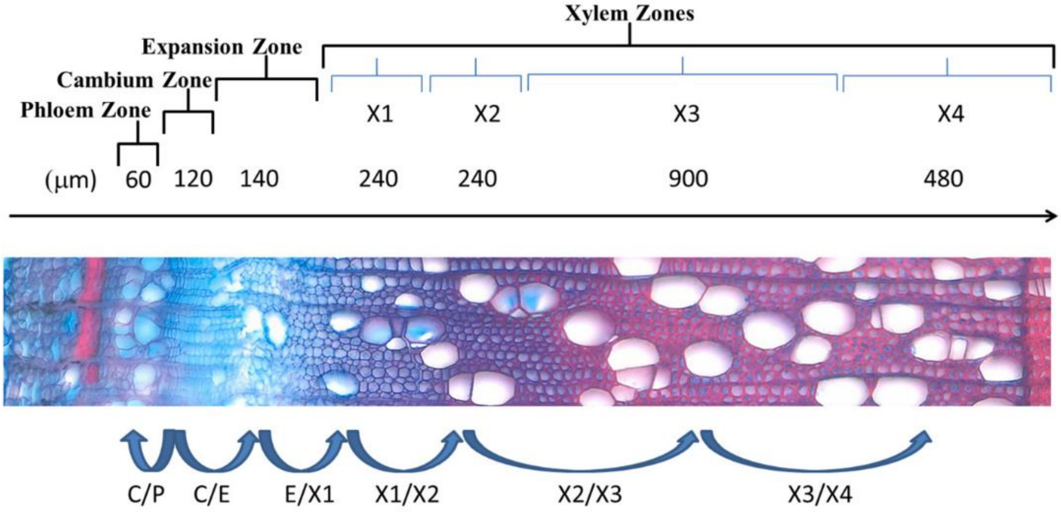

Figure 8: "Schematic overview of transverse sections prepared from a specimen in tree 1". From the proteomics article by Obudulu et al. (2016). The blue arrows indicate the direction of cell expansion and differentiation.

The protein expression data set was from a proteomics article by Obudulu et al. (2016). 3,082 proteins were identified in the study. Expression profiles of these proteins were given across 111 samples from four trees. 27 samples from tree number one, 28 samples from tree number two, 28 samples from tree number three and 28 samples from tree number four.

The transcriptomics data set Sundell et al. (2017) included expression profiles of 28,294 genes measured across 106 samples from four trees (25 samples in tree number one, 26 in tree number two, 28 in tree number three and 27 in tree number four). The whole genome of *Populus tremula* has not yet been sequenced, so instead *Populus trichocarpa* was used as the reference genome for both the transcriptomics study and proteomics study. As they were closely related one may assume that they have similar genetic qualities. In the NCBI database a recent full genome sequencing of *Populus trichocarpa* was made available with a total sequence length of 434,289,848 base pairs (bps), which is considered a modest genome size for a tree (Kainer et al., 2015). The full protein count for the tree was 51,717. The genome is organized in 19 chromosomes. *Populus tricharpa* was the first tree to be sequenced (Tuskan et al., 2006), making the genus *Populus* a model organism candidate for woody trees.

Information on protein function and isoforms were obtained from the proteomics data set.

3.2 Proteomics

There were three main steps for obtaining the protein data. The digestion of proteins and extraction of peptides, analysis of peptide content and protein identification. Methods for extracting protein from the samples are described in the proteomics article by Masuda et al. (2008) in greater detail. The process utilized trypsin to cut the proteins into enzymes with the aid of phase-transfer surfactants in this case sodium deoxycholate (SDC). The resulting peptides were analyzed using reversed-phase liquid chromatography-electro spray ionization mass spectrometry (LC-ESI-MS). The resulting data from the (LC-ESI-MS) was processed with Protein Lynx Global Server v.3.0 and the resulting spectra were searched against *Populus trichocarpa*, together with sequences for human keratin and rabbit glycogen phosphorylase. The JGI Comparative Plant Genomics Portal database provided the reference. The process for the search and quantification was provided in the transcriptomics article (Obudulu et al., 2016).

In this thesis, identification and quantification of the proteins in heterogeneous mixes of cells are the basis together with the transcriptomics data. While absolute quantification through mass spectroscopy is possible, MS techniques usually do not provide the full quantitative protein levels in a cell. Instead, the techniques approximate the abundance of the protein in a cell instead (Steen & Pandey, 2002). The quantity of the protein will be referred to as protein abundance throughout this thesis and which is a relative measurement.

3.3 Transcriptomics

RNA-seq (described in chapter 2.4) was employed to identify protein-encoding RNA transcript and their expression profile (Sundell et al., 2017). The gene expression levels are the fraction of reads mapped to the reference genome and therefore they are relative measurements. The RNA quantification will be referred to as gene expression throughout the thesis.

3.4 The R-Coding Processes

The main workload of the thesis was preparing and analyzing the data in R. An overview of the R scripts used in the thesis is provided in Appendix A.

3.4.1 Standardize the Data Set

The data sets were quite different in format. The transcriptomics data covered 28,294 genes, while the proteomics data covered merely 3,082 proteins. The number of samples in the series also differed in the two data sets. The goal was to coerce the data sets into having the same dimensions. To achieve this, some genes from both data sets had to be removed so that only proteins with a corresponding gene expression and vice versa remained. Additionally, the number of samples varied amongst the trees and therefore some samples were cut out of the data set. The gene expression data was in log2 scale due to the VST method used. Therefore, the proteomics expression values were log2-transformed. This was achieved using the R function log2(). Before the transformation the protein entries were given +1 in value, so that zero entries would not become negative infinity entries, but instead remain zeros. The non-zero entries for the protein values were between 505.4 and 1,280,000.0. Therefore, an addition of 1 is insignificant. A single proteomic entry containing the value 1 was edited to be 0 instead, since it likely were an error.

Indexing and the match() function in R was used to find which genes the two data sets had in common. Duplicates were made for each isomorph protein, so that each protein had its own corresponding gene expression series. The transcript data frame and the protein data frame, then had the same number of rows. All protein abundance rows which sums were zero and their corresponding rows in the transcription data frame were removed. After those procedures, 2,029 expression series remained in each data set.

The four trees from the transcriptomics data matched the four trees from the proteomics data, but the number of samples differed in each study and by each tree. Specifics are shown in

Table 1 and Table 2. For simplicity, both data sets were cut so that contained exactly 100 samples each. 25 samples from each tree. The subsets containing 25 samples from each tree were selected by a simple maximizing correlation method: all combinations of 25 continuous samples of the transcriptomics data and the proteomics data were correlated. The continuous series that yielded the highest correlation was saved. The scheme for cutting is given in Table 1 for protein samples and Table 2 for transcript samples. After the removal of some samples, both data sets contained the same number of rows and the same number of columns.

Table 1: Scheme for selecting the subset of the proteomics data.

Tree number	Number of samples	Samples included in the data
1	27	1 to 25
2	28	29 to 53
3	28	56 to 80
4	28	86 to 110
Total	111	100

Table 2: Scheme for selecting the subset of the transcriptomics data.

Tree number	Number of samples	Samples included in the data
1	25	1 to 25
2	26	26 to 51
3	28	53 to 77
4	27	81 to 105
Total	106	100

The plots and correlations may have been affected by the vague boundaries between the different developmental stages, differences between the four trees and the removal of some samples.

3.4.2 Moving Average Calculations

A moving average calculation can be used to smooth time series with high fluctuations/noise. Using a moving average calculation on a time series will shift the focus from local fluctuations to more long-term trends.

The moving average series for the protein data was calculated to provide a better correlation fit with the transcriptomics data. Each expression value in a series was summed with its two adjacent samples (only one if on either edge of the series) and divided by three (or two at the edges). This calculation smoothed out the expression series. The moving average data set was used alongside the normal protein data set for many correlations and plotting purposes. The moving average treatment was not necessary for the transcriptomics data since these expression series were in most cases smooth and the moving average treated transcriptomics data did not yield sufficiently different correlation results when correlating them with the proteomics data.

3.4.3 The “Best Method”

The protein data was in some cases heavily incomplete. To remedy this a method was developed where samples from each tree were in turn correlated with the average of the transcriptomics data. The tree which correlated the highest with the average of the gene expression series was saved. In this way the tree offering the “best” data could be used so that the correlations were not burdened by trees with incomplete or missing data.

3.4.4 Correlations

Correlation in statistics is a measure of the relationships between two variables. Calculated correlation coefficients extend from -1 to +1. Values of either +1 or -1 indicates a perfect relationship between the two variables, while values close to 0 indicate no or little relation between the variables. Negative values indicate a negative relationship between the variables, increase in variable A leads to a decrease in variable B and *vice versa*. There are several ways of measuring correlation and three main ways are possible through the R function cor(). These are Pearson, Kendall rank and Spearman. In this thesis, Pearson’s correlation was used. To investigate the amount of variance in Y explained by X, the coefficient of determinants or R^2 may be utilized. This can be calculated by squaring the correlation coefficient.

Protein abundance values across individual series was summed, meaning the total abundance for one single protein was calculated. The same was done for the transcriptomics data. The correlation between them was dubbed “Gene correlation”.

Full correlation was calculated across both the whole data sets. The expression values for the protein expression and the gene expression levels were saved in two separate vectors. The correlation between them was calculated. Full correlations using the moving average calculated data and the best data were similarly calculated. The vectors for the untreated data set and the data set using the moving average calculated protein data were 202,900 entries long. The vectors for the "best method" treated data set and the combined "best method" and moving average treated data were 50,725 entries long, due to "best method" only utilizing the protein samples from one tree. The correlations were also calculated between the data sets with zero entries removed.

Row correlation, meaning that the expression a single gene across the series was correlated with its corresponding protein across the same time series. The correlation was calculated for between each protein abundance series and gene expression series across the data sets. Since the samples were edited so that each tree for both the protein and the transcriptomics data had 25 values, the whole series could be correlated together.

The data was correlated by sample series to uncover areas of potentially high post-transcriptional regulation. If there was a high correlation across protein and gene pairs for one sample, it would indicate less post-transcriptional regulation. The series used for these correlations consisted of expression values for all proteins or gene in each sample from one tree. The same method was used as in the row correlations, but only the moving average calculated proteomics data set was used. The correlation by sample was in addition done based on the "best method".

3.4.5 The On-Off Switch Method

Since mRNA and protein abundance do not always correlate well, some studies have suggested that mRNA works more like an on-off switch (Vogel & Marcotte, 2012). To test this on the data set, all numeric entries in both data sets was set to 1 and all zero entries were kept as 0s. Then the data was compared to detect if genes were expressed when protein abundance was above zero and vice versa. This was done for the whole dataset, providing a matrix showing how often mRNA and protein were expressed or not and the relation between them. Additionally, the number of matches was counted for each transcript-protein pair and plotted. The moving average treated proteomics data was used. Lastly the likelihood of protein abundance above zero for different gene expression levels was calculated. All the samples above the given gene expression level was used to calculate the likelihood.

3.4.6 Protein Isoforms

When the proteomics data set and the transcriptomics data sets were matched, copies of the transcriptomics series were made to accommodate the different isoforms in the proteomics data. I.e. If a protein had three isoforms there would be three equal gene expression series corresponding to that protein. The edited proteomics data set contained 296 proteins that had two or more isoforms. In total, there were 650 rows in the data set that represented the isoform of another protein (this means that there were 650-296=354 gene expression series in the transcriptomics data set that were copies). The expression series of the isomorph proteins were summed by base protein and correlated with the transcription expression series. Furthermore, some isomorph proteins were plotted together and with the corresponding gene expression to explore whether there was a delay between the isoforms or if one isomorph was expressed in developmental stages and other isomorphs in others. The moving average treated data set was used for this analysis

3.4.7 Clustering and Heat Maps

In addition to correlation, heat maps and dendrograms were produced from the data . The distance used for the dendrograms was calculated with the following equation: 1 subtracted by the correlation between the series. The dendrograms were built using the “ward.D” method, referring to Ward’s criterion. The heatmap.2() R function plotted the dendrogram together with a heat map of all the entries of the data sets. The data was scaled by row. Dendrograms and heat maps were produced from the raw transcriptomics data, the average of trees transcriptomics data, the moving average calculated proteomics data and the combined “best method” and moving average calculated proteomics data.

The clusters in the average by trees transcriptomics data and the moving average calculated proteomics data were compared and a relation score was calculated. The relation score followed this equation: the number of intersecting proteins/genes divided by the total number of unique protein/genes. The combined clusters that had the highest correlation coefficients were reported.

3.4.8 ANOVA

ANOVA was utilized to test the difference in expression between the developmental stages. A linear model was created where developmental stage was the explanatory variables were the correlations by sample values and the response values where the developmental stages: phloem, expanding xylem, SCW forming xylem and late xylem. The ANOVA table was calculated in R and a Tukey’s test was reported.

3.5 GO Enrichment Analysis

Gene Ontology (GO) enrichment analysis was utilized to detect GO terms that were overrepresented in gene products in which protein abundance series correlated highly with their corresponding gene expression. The Gene Ontology project aims to provide a controlled vocabulary describing gene products (Ashburner et al., 2000). GO terms are separated into three categories: biological process, molecular function and cellular component (Ashburner et al., 2000). Furthermore, each category has GO terms that can be broader (possessed by many gene products) or more specific and rarer. GO terms in the biological process category are defined by the known objectives a gene product has. Often these are chemical or physical transformations. Synthesis of a of sucrose would be a narrow biological process while “Signal transduction” would be an example of a broader GO term. Molecular function GO terms are related to biochemical activity i.e. “Enzyme”, “transporter”, “ligand” etc. Cellular component GO terms are related to the area of the cell where the gene product is active.

A GO enrichment analysis was conducted of the proteins which abundance series correlated highly with their corresponding gene expression (correlation coefficient > 0.17, using moving average calculated and “best method” treated proteomics data), in total 1233 proteins. A Gene Ontology enrichment analysis can compare a subset of the proteome in an organism with its complete proteome and find GO terms that are overrepresented the subset. The enrichment tools and background genome (*Populus tricharpa)* was provided by popgenie.org (Sundell et al., 2015). The GO enrichment analysis reported the GO terms that were overrepresented, the False discovery rate adjusted P-value and the GO terms’ frequency rate in the subset and in the background genome.

4 Results

4.1 Matching the Data

The identified proteins which had a corresponding gene expression were found and matched in R. There were 3,082 identified proteins and 28,294 genes (567 of which were putative genes) were identified in the transcriptomics article by Sundell et al. (2017). 2,860 of the proteins had a match in the transcriptomics data. Furthermore, 920 of the proteins had zero expression across all the samples. There were 2,029 protein-gene pairs remaining in the data after removing protein series summing zero. No gene expression series summed to zero after removing samples. A Venn-diagram of the counts is provided in Figure 9. The total number of identified transcripts would likely be higher if splice variants were considered. The proteomics data considered isoforms.

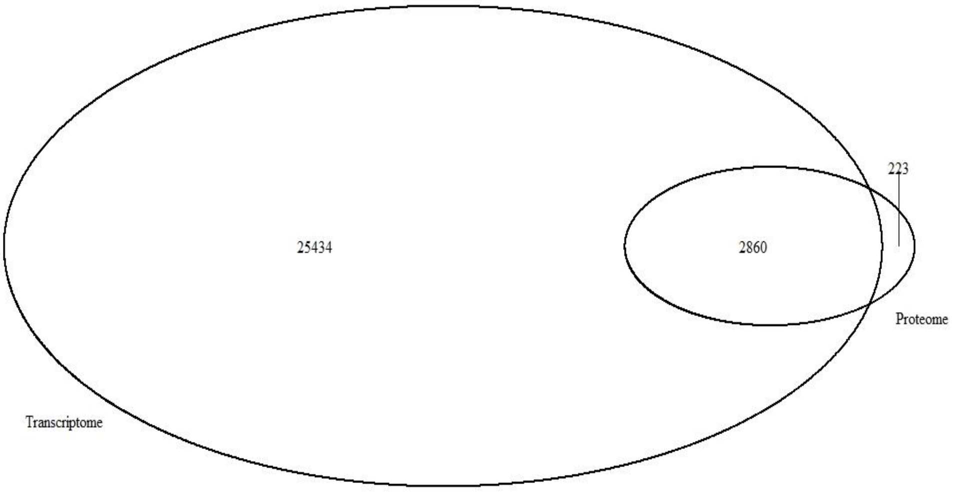

Figure 9: Venn-diagram showing the number of identified protein-coding transcript and expressed proteins, before zero-sum rows were removed.

4.2 Cases

This thesis has included some cases of correlations between gene expression and protein abundance. All the plots in this section correlated the standard transcript data with the moving average protein data unless stated otherwise.

4.2.1 Marker Genes in the Transcriptomics Article by Sundell et al. (2017)

In the transcriptomics article, five genes were used as markers for the different developmental stages: Potri.004G081300, Potri.016G142800, Potri.001G240900, Potri.004G059600, Potri.011G044500. Potri.004G081300, Potri.016G142800 and Potri.004G059600 were found in the proteomics data as one or more protein isoforms. Potri.001G240900 and Potri.011G044500 were not found in the proteomics data.

Potri.004G081300 was expressed as the proteins Potri.004G081300.1 and Potri.004G081300.2, which were both sucrose synthases. Potri.004G081300.1 was expressed in one sample in tree 1. Potri.004G081300.2 protein expression correlated highly with its gene expression with a correlation value of 0.745. It was expressed in the phloem developmental stage and similarly for all trees (Figure 10).

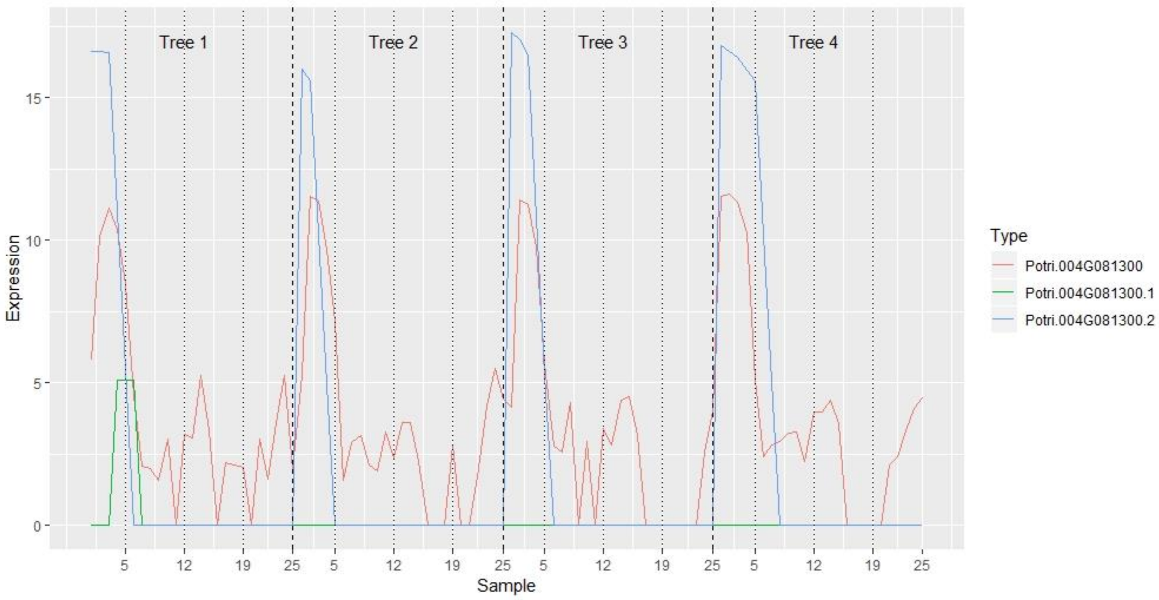

Figure 10: Plot of expression of the gene encoding Potri.004G081300 a sucrose synthase, together with its corresponding protein abundance series (Potri.004G081300.1and Potri.004G081300.2).

Potri.004G059600 encodes a cellulose synthase family protein (figure 11). The protein was expressed in the third developmental stage, where the secondary cell wall is deposited. The protein Potri.004G059600.1 was only expressed for tree number four in expanding xylem and SCW forming xylem and once for tree number two in SCW forming xylem.

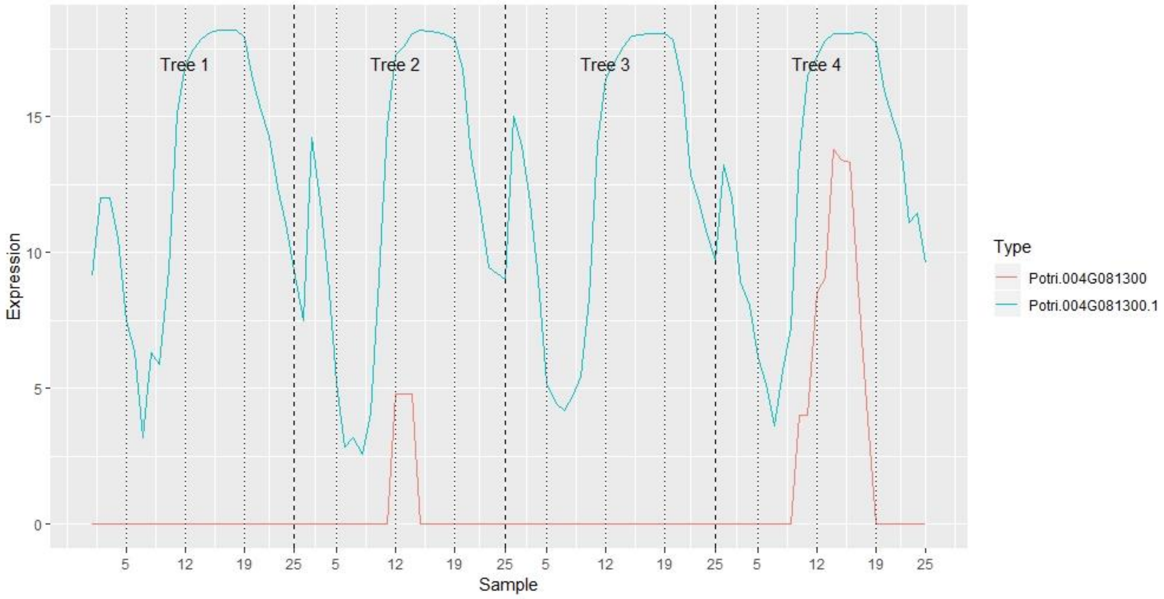

Figure 11: Plot of expression of the gene encoding Potri.004G059600 a cellulose synthase family protein together with its corresponding protein abundance series (Potri.004G059600.1).

Potri.016G142800, a cyclin-dependent kinase. The gene was expressed in the phloem and xylem with the peak in the first reprogramming event (in the middle of the cambium). Protein abundance was found in late xylem in tree number three and not near the gene expression peak (Figure 12).

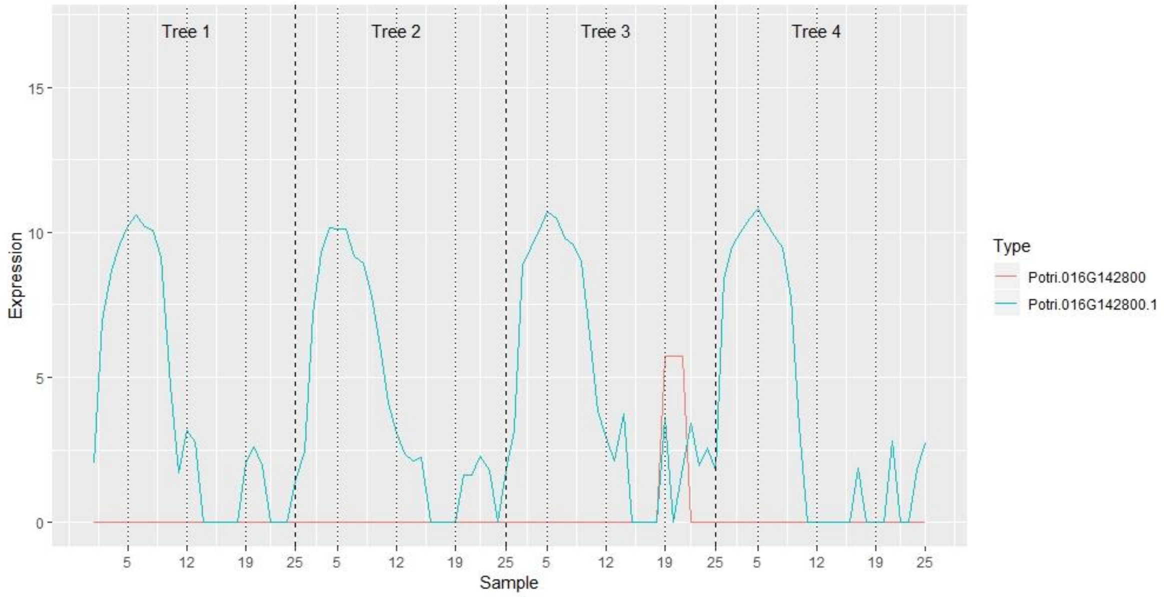

Figure 12: Plot of expression of the gene encoding Potri.016G142800 a cyclin-dependent kinase together with its corresponding protein abundance series (Potri.016G142800.1).

4.2.2 Sucrose Synthases

Sucrose synthase family proteins have been classified as potential regulators of phloem functions and it seems that the protein expression was regulated both at transcription level and post-transcription level. The genes were often expressed in many samples in different developmental stages but were nearly always up-regulated in the phloem. The exception being the gene encoding sucrose synthase 3, Potri.002G202300.1 (Figure 13, A), which was expressed at approximately the same level in all samples. While sucrose synthases were transcribed across almost all developmental stages, it was mainly translated into proteins in the phloem (Figure 13).

The sucrose synthases where amongst the proteins that had the highest correlation coefficients with their corresponding gene expression. The proteins Potri.004G081300.2 (C), Potri.012G037200.1 (D), Potri.015G029100.1 (F) and Potri.017G139100.3 (H) were expressed in a similar way following the fluctuation in gene expression levels: high expression in phloem samples and less or not at all in other samples. They were also the protein/gene pairs that had the highest correlation in the group. Exceptions from this were found in the protein abundance profile of Potri.012G037200.1 (D) where the protein abundance was found in other developmental stages than phloem in tree number three and tree number four. Potri.002G202300.1 (A) was only expressed in the phloem in tree number one and nowhere else. Potri.004G081300.1 (B) was expressed only in one sample in tree number three, but also in the phloem, and appears to be a translation on demand protein. Potri.012G037200.2 (E), an isoform of Potri.012G037200.1 (D), was expressed in the developmental stage number four in tree number three. Potri.017G139100.2 (G) was expressed in the phloem, but only in tree number one and tree number three.

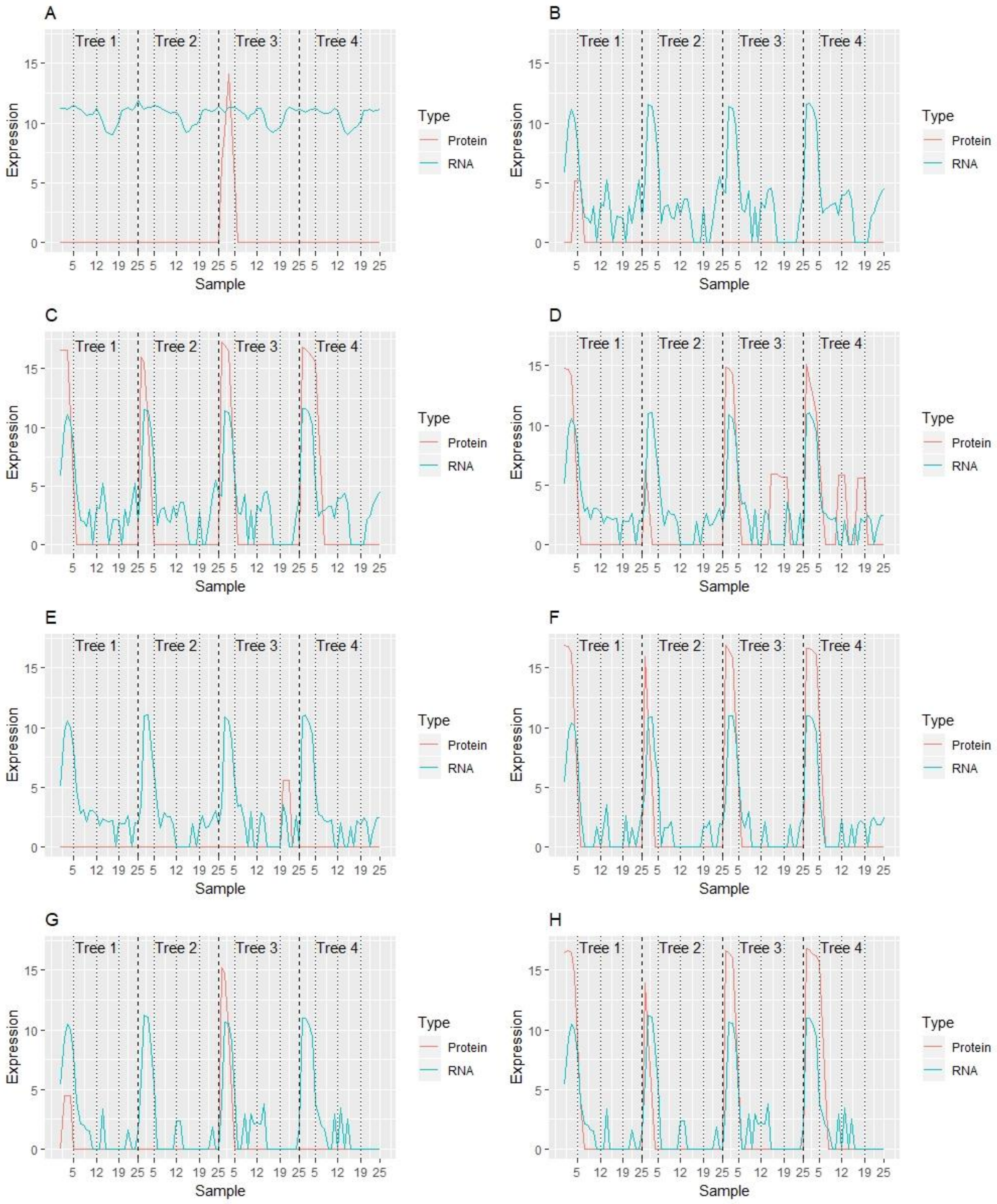

Figure 13: Protein abundance series (marked as "Expression") of various sucrose synthase identified in the proteomics article and corresponding gene expression from the transcriptomics article. Potri.002G202300.1 (A), Potri.004G081300.1 (B), Potri.004G081300.2 (C), Potri.012G037200.1 (D), Potri.012G037200.2 (E), Potri.015G029100.1 (F), Potri.017G139100.2 (G), Potri.017G139100.3 (H).

4.2.3 Cellulose Synthases

Cellulose synthase proteins are key in wood formation, as cellulose is deposited in the secondary cell wall. These proteins were expected to be expressed in phloem and SCW forming xylem. They were usually expressed in either of these developmental stages, but not for all trees (Figure 14). The proteins Potri.002G257900.1 (B), Potri.004G059600.1 (C), Potri.006G181900.2 (F) and Potri.011G069600.1 (I) had similar gene expression patterns with a large peak in SCW forming xylem and a smaller peak in phloem. Protein abundance appears in SCW forming xylem, but it was never reproduced across all trees. The protein Potri.002G066600.1 (A), Potri.006G052600.2 (E), Potri.006G251900.6 (G), Potri.007G076500.5 (H), Potri.007G076500.6 (J) and Potri.011G069600.1 (I) had more even gene expression patterns. Protein abundance above zero was found in SCW forming xylem but reproduced across the different trees. Potri.006G052600.1 (D) had an even gene expression pattern and was the only cellulose expressed in phloem, but only in tree number one.

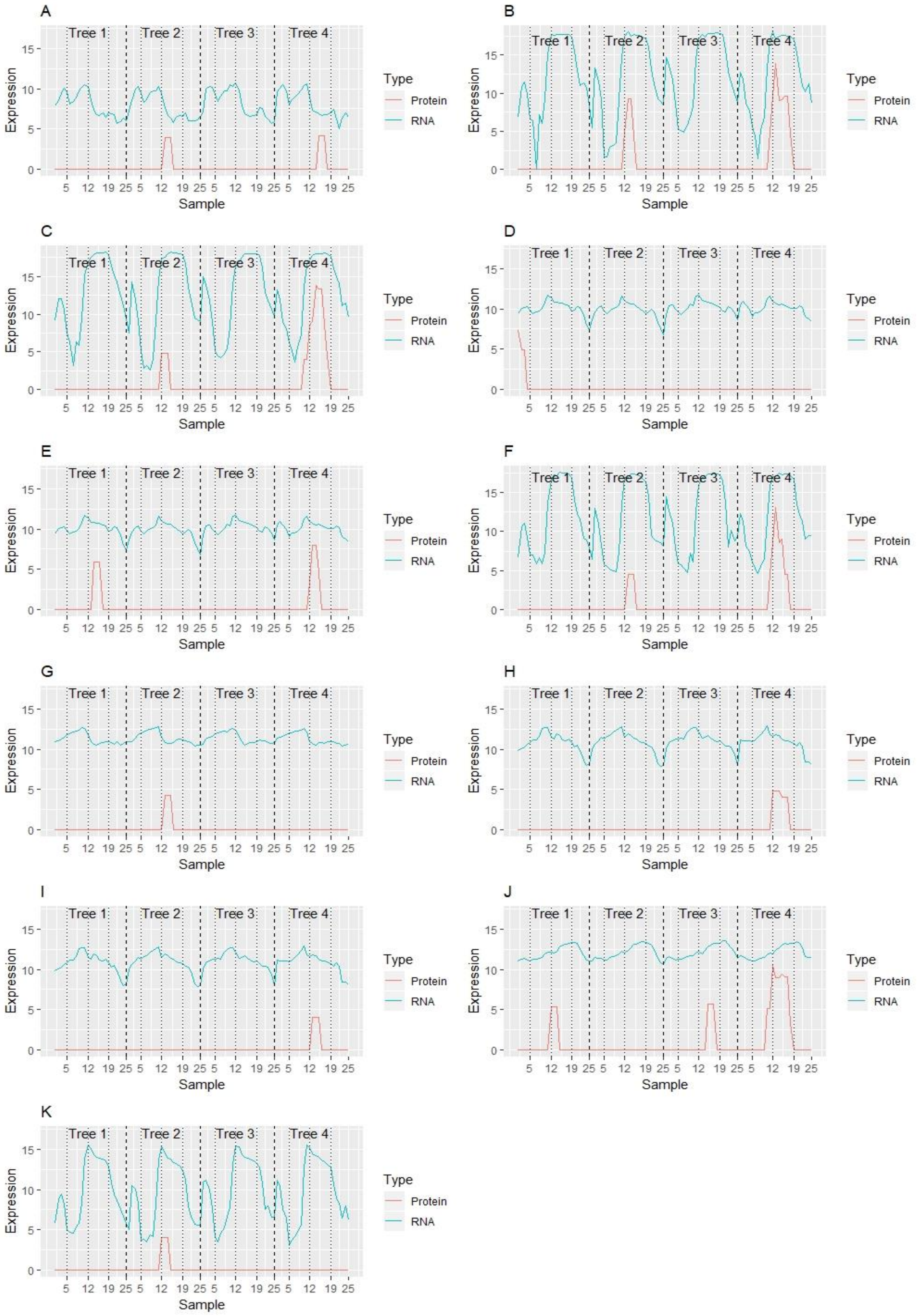

Figure 14: Protein abundance series (marked as "Expression") of cellulose synthase proteins identified in the proteomics article and corresponding gene expression from the transcriptomics article. Potri.002G066600.1 (A), Potri.002G257900.1 (B), Potri.004G059600.1 (C), Potri.006G052600.1 (D), Potri.006G052600.2 (E), Potri.006G181900.2 (F), Potri.006G251900.6 (G), Potri.007G076500.5 (H), Potri.007G076500.6 (J), Potri.009G060800.4 (K), Potri.011G069600.1 (I).

4.2.4 Xylem Related Peptidases

Xylem related peptidases are thought to be related to apoptosis in expanding xylem and lignified xylem. They were expected to be expressed in SCW forming xylem and perhaps further inwards the stem. The expression/abundance patterns of xylem related peptidases are found in Figure 15. Potri.002G005700.1 (A) and Potri.005G256000.2 (C) both were xylem cysteine peptidase 2 and they correlated well with the gene expression. The gene expression and the protein abundance of both proteins peaked in the SCW forming xylem developmental stage. The gene expression for Potri.002G005700.1 (A) was also quite high in the late xylem developmental stage compared to the gene expression of Potri.005G256000.2 (C) which sharply plummeted after reprogramming event number 3. Potri.002G005700.2 (B), a xylem cysteine peptidase 1 which is an isoform of Potri.002G005700.1 (A) was only expressed in SCW forming xylem in tree number three. Xylem serine peptidase 1 Potri.014G074500.2 (D) had a different gene expression pattern signifying that it had a larger importance in the late xylem developmental stage. The protein was only expressed in tree number two, with a peak in the late xylem developmental stage.

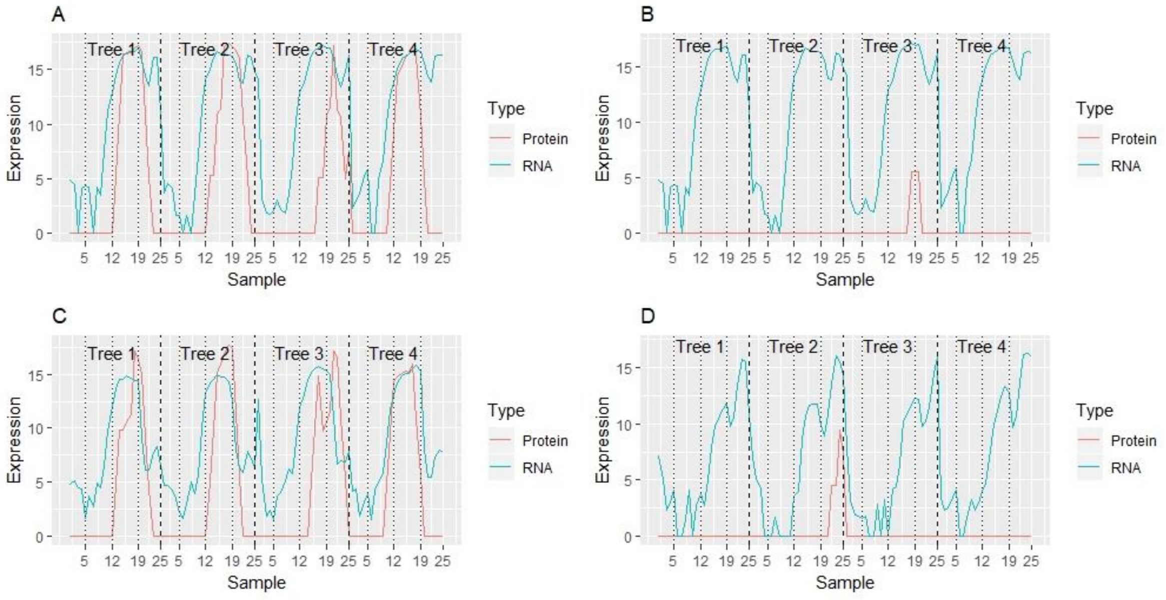

Figure 15: Protein abundance series (marked as "Expression") of various xylem related peptidases identified in the proteomics article and corresponding gene expression from the transcriptomics article. Potri.002G005700.1 (xylem cysteine peptidase 2) (A), Potri.002G005700.2 (xylem cysteine peptidase 1) (B), Potri.005G256000.2 (xylem cysteine peptidase 2) (C), Potri.014G074500.2 (xylem serine peptidase 1) (D).

4.3 Expression Distribution

The distribution of gene expression and protein expression values were plotted in Figure 16. The proteomics data had mainly zero entries. There was an island with values between 9 and 20.3 that has a peak at 15. The transcriptomics data's gene expressions were more evenly distributed and has a smaller peak at zero and a larger peak around expression value 11 (Figure 16). See Table 3 for a summary of statistics on the expression levels. The moving average calculated data was not used here.

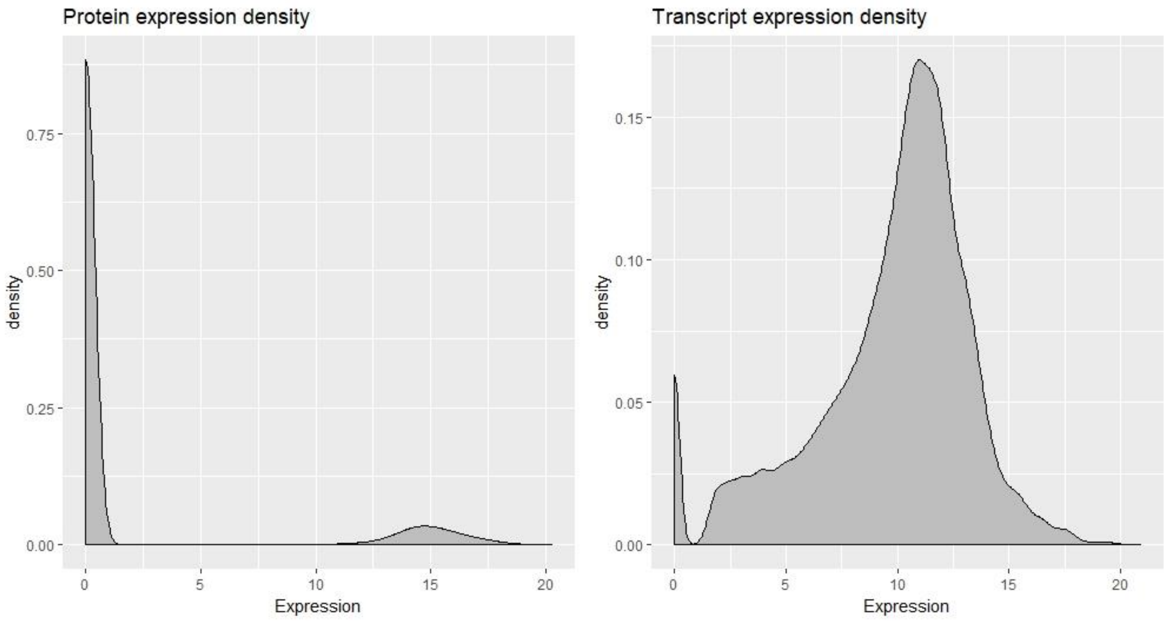

Figure 16: Density plot of the protein abundances to the left and of the gene expression to the right.

Table 3: The summary of the two data sets

	Proteomics	Transcriptomics
Zero rate	0.87	0.035
Mean	1.9	9.8
Median	0	10.5
Max	20.3	20.9

The sum of the total gene expression in each sample series, and total protein abundance in each sample series was calculated. There was a lot of variance in protein abundance in the samples and much less variation in the sum of gene expression in the different samples (Figure 17).

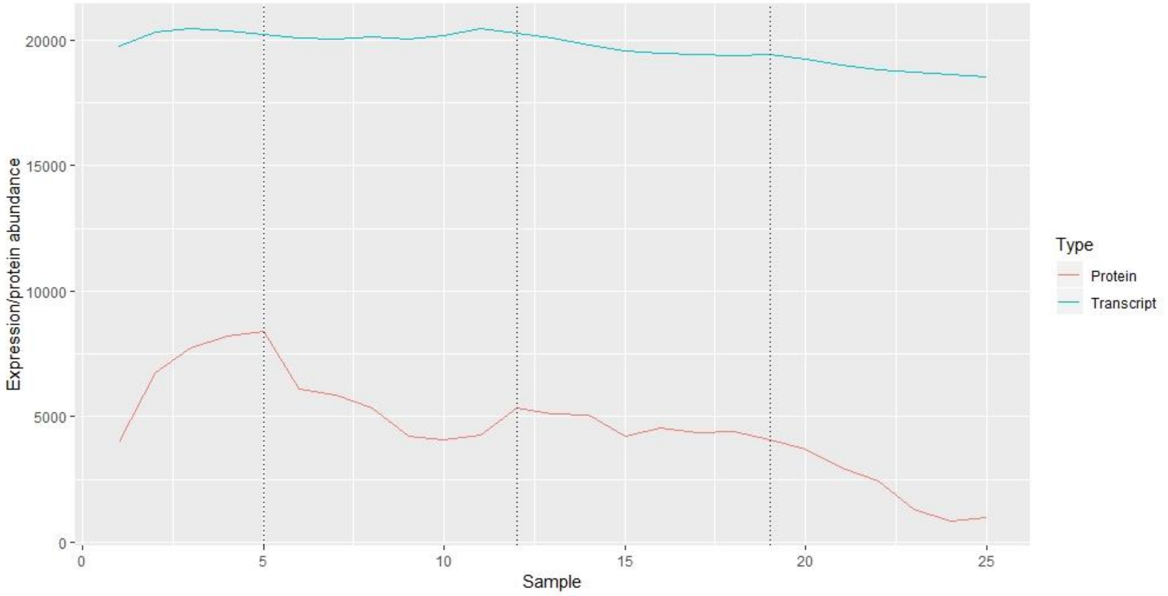

Figure 17: Total gene expression/protein abundance in each sample. Protein abundances provided by "best method" and moving average calculated proteomics data. Gene expressions were provided by the average of the 4 trees.

4.4 Full Correlation

Full correlation across the whole data set yielded a correlation coefficient of 0.256. Using the moving average calculated data, the correlation score was 0.295. Using the “best method”, the correlation score was 0.298. Using the “best method” together with the moving average calculation, gave a correlation score of 0.347. When all the samples were summed together by protein abundance and transcript the correlation was 0.321. All correlations were significant, according to the Pearson’s correlation test (Table 4). Table 5 shows the correlation score with zero values removed. The data with zero entries removed were plotted in Figure 18.

Table 4: Overview of correlations across all samples using different methods.

Correlation type	**Correlation scores**	**R^2**	**95 % Confidence interval**	**P-value**
Raw data	0.256	0.066	0.252 - 0.260	< 2.2e-16

Moving average	0.295	0.087	0.291 - 0.299	< 2.2e-16
“Best method”	0.298	0.089	0.290 - 0.306	< 2.2e-16
Combined “Best method” and Moving average	0.347	0.120	0.339 - 0.354	< 2.2e-16
Gene correlation	0.321	0.103	0.281 - 0.359	< 2.2e-16

Table 5: Overview of correlations across all samples using different methods with zero entries removed.

Correlation type	**Correlation scores**	**R^2**	**95 % Confidence interval**	**P-value**
Raw	0.340	0.116	0.329 - 0.350	< 2.2e-16
Moving average	0.318	0.101	0.309 - 0.326	< 2.2e-16
“Best method”	0.324	0.105	0.303 - 0.343	< 2.2e-16
Combined “Best method” and Moving average	0.325	0.106	0.309 - 0.341	< 2.2e-16

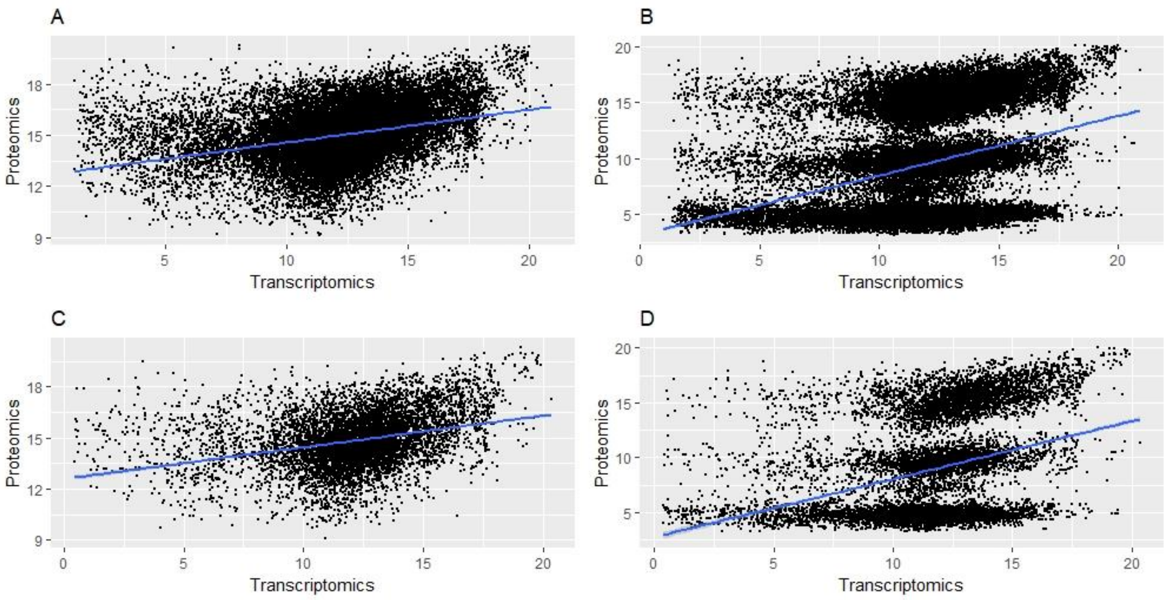

Figure 18: Dot plots showing the relation between the protein abundance and the gene expression. Raw protein data (A), moving average (B) “best method” (C), combined “best method” and moving average (D).

Through the R^2 values it is shown that between 6 and 12 percent of the variation in protein abundance was explained by the gene expression levels depending on the data treatments used.

The correlation between the summed protein abundance series and the summed gene expression series was 0.32. A large portion of summed protein abundance series had a low total abundance regardless of the corresponding summed gene expression; however, a substantial portion of the summed protein abundances correlates quite well with summed gene expression (Figure 19).

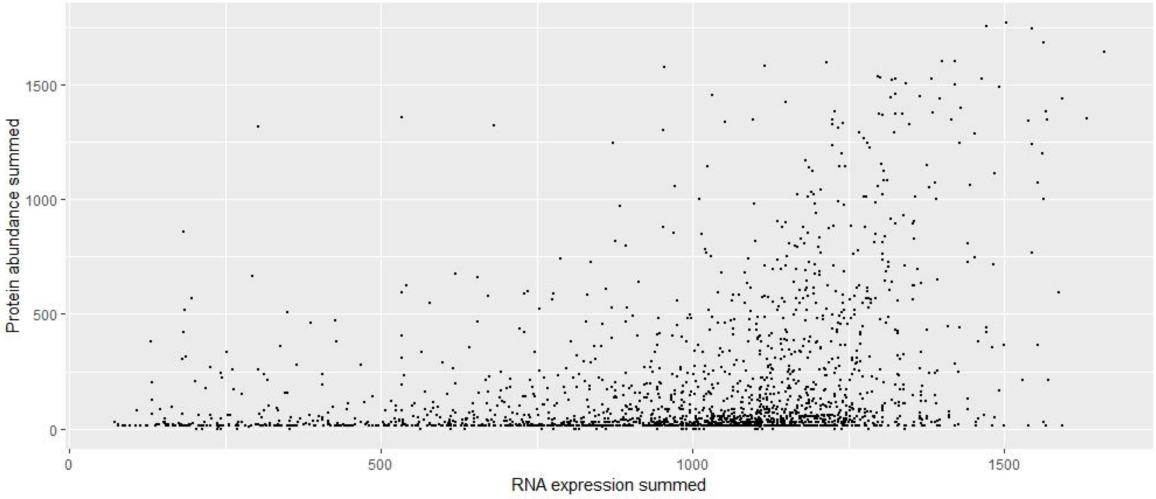

Figure 19: Dot plot of the summed RNA expression against the summed protein abundance.

4.5 Row Correlation

Density plots were made of the correlation coefficients using the different data treatments (Figure 20). The correlation by row (expression of a single gene across the gene expression series against corresponding protein abundance series) using the raw data yielded low correlations. Correlation using the raw proteomics data had a median value of 0.11 and the best-correlated gene had a score of 0.91. Using the moving average calculations to smooth protein expressions gave a median value of 0.17 and the best-correlated gene had a correlation score of 0.94. Using the “best method” virtually eliminated negative values since given a choice between a negative score with one tree and a neutral score, the algorithm would choose a neutral score. The median value was also improved to 0.27 and the best-correlated gene had a correlation score of 0.96. Combining the moving average calculations and the “best method”, the score was further improved, yielding a median score of 0.44 and the best-correlated gene had a correlation score of 0.98.

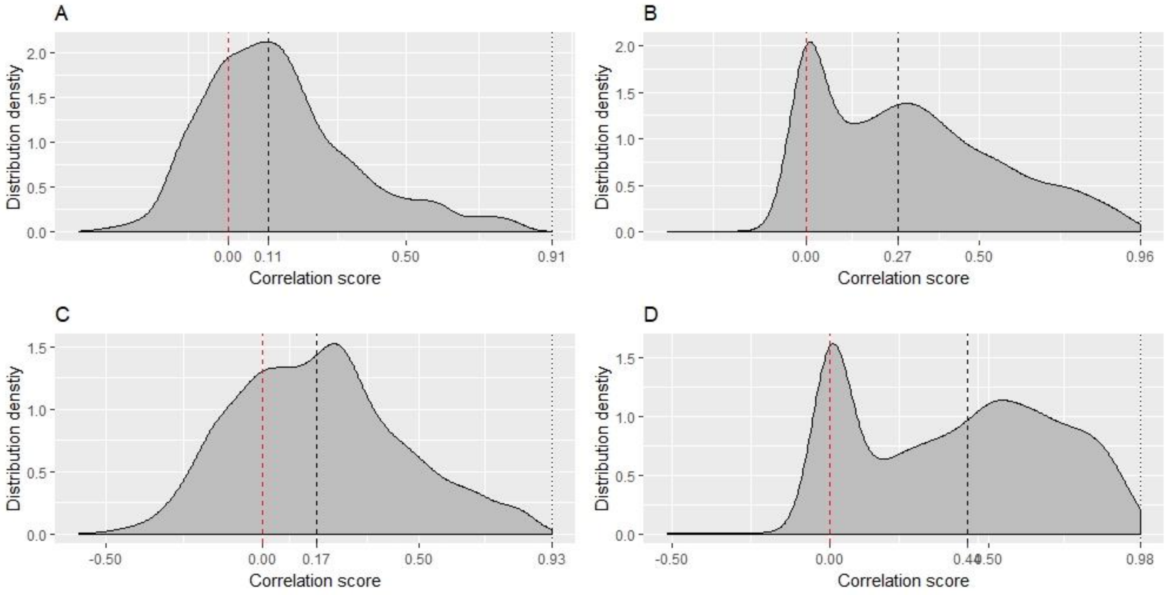

Figure 20: The figure shows density plots of the correlation score between gene expression and protein expression series. Raw proteomics data (A),"best method" (B), moving average (C), "best method" and moving average calculations combined (D). The vertical dashes show zero (red) on the x-axis and the median and the max correlation.

4.6 The Best-Correlated Genes

The 13 highest correlating genes were inspected in more detail. Eight of these genes and corresponding proteins were mainly expressed in phloem (Figure 21: B, C, F, G, H, J, L, M). Potri.001G340300.1 (B), Potri.001G340500.1 (C), Potri.002G175400.1 (F), Potri.015G029100.1 (L), Potri.017G139100.3 (M) were all similarly expressed with both the gene expression peak and protein abundance peak in phloem. The proteins corresponding to these genes were almost exclusively expressed in the phloem. The corresponding gene expression of these proteins were occasionally observed in other developmental stages, but at a much lesser extent. The protein abundance and gene expression patterns for these four protein/gene pairs were similarly reproduced in all four trees. Potri.015G029100.1 (L) and Potri.017G139100.3 (M) were identified as sucrose synthases the other three proteins were of unknown function. Potri.004G044700.1 (G), a “Pollen Ole e 1 allergen and extensin family protein”, peaked in phloem in tree number one and tree number four, but in tree number one and tree number four protein abundance peaked closer to the cambium and was expressed in the expanding xylem and phloem developmental stages. Potri.002G175400.1 (H) a protein of unknown function was only expressed only expression in tree number tree and tree number four. The protein abundance series had a high correlation with its gene expression series in tree number 3 and was probably selected due to the “best method”. Potri.012G095200.1 (J), a Zn-dependent exopeptidases superfamily protein, could be a translation on demand protein as its corresponding gene was evenly expressed throughout all developmental stages, but peaked slightly in phloem. Protein abundance was only found in phloem for Potri.012G095200.1.

Potri.001G054400.1 (A), Potri.002G029100.1 (D), Potri.011G110700.4 (I), Potri.014G071700.2 (K) were all expressed around cambium in phloem and expanding xylem. The corresponding gene expressions of these four proteins were expressed in all developmental stages and peaks around cambium. The gene expressions encoding Potri.001G054400.1 (A), Potri.011G110700.4 (I) were especially even across the developmental stages, suggesting they might have been “translation on demand” proteins. Potri.001G054400.1 (A) was identified as heat shock protein 60, Potri.002G029100.1 (D) was identified as a Walls Are Thin 1 protein, Potri.011G110700.4 (I) was identified as phosphoenolpyruvate carboxylase 3 and Potri.014G071700.2 (K) was identified as FASCICLIN-like arabinogalactan-protein 10.

Potri.002G099200.1 (E), a Class-II DAHP synthetase family protein, was expressed in later xylem stages. Its corresponding gene expression was expressed in all developmental stages, but peaks in later xylem along with Potri.002G099200.1.

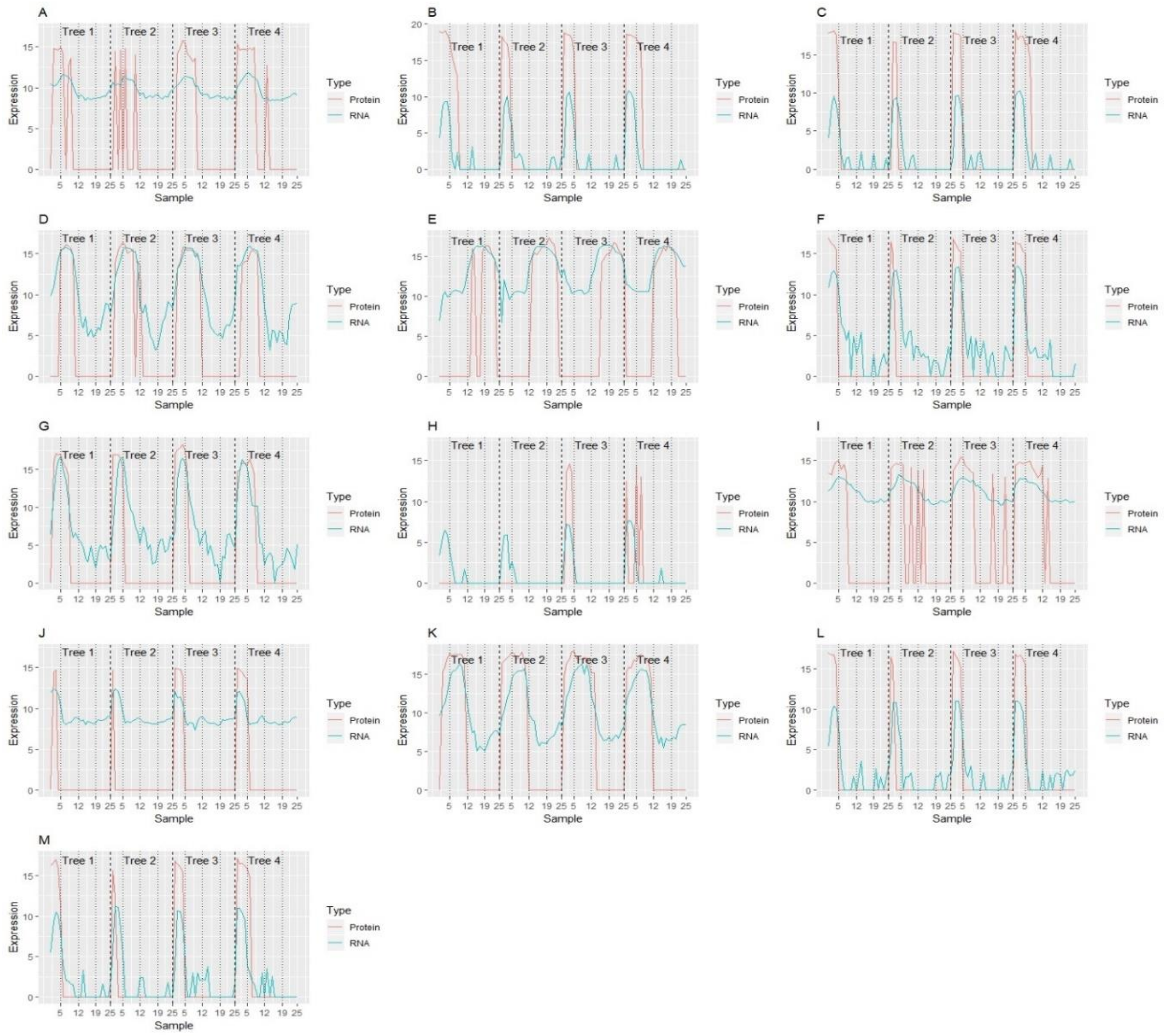

Figure 21: Protein abundance pattern and gene expression patterns of the 13 highest correlating gene/protein pairs for the four trees (using proteomics data treated with "best method" and moving average calculations). Potri.001G054400.1 (A), Potri.001G340300.1 (B), Potri.001G340500.1 (C), Potri.002G029100.1 (D), Potri.002G099200.1 (E), Potri.002G175400.1 (F), Potri.004G044700.1 (G), Potri.006G171200.1 (H), Potri.011G110700.4 (I), Potri.012G095200.1 (J), Potri.014G071700.2 (K), Potri.015G029100.1 (L), Potri.017G139100.3 (M).

4.7 Correlation by Sample

To correlate by sample number, one series of protein abundances and one series for gene expressions was made of each sample. The resulting series were correlated (Figure 22). There was a distinct decrease in the correlations at the endpoints of the series. In developmental stage one (phloem) and developmental stage four (late xylem), the level of correlation was lower than in expanding xylem and SCW forming xylem. The difference in correlation coefficients was significant in an ANOVA test. A box plot comparing the values in the different developmental stages was provided in Figure 23 and the corresponding ANOVA table provided in Table 6. According to the 95% Tukey test, there was a significant difference in correlation scores in the xylem cell death developmental stage and all other developmental stages (Figure 24). There was also a significant difference in the correlation scores in phloem and lignified xylem.

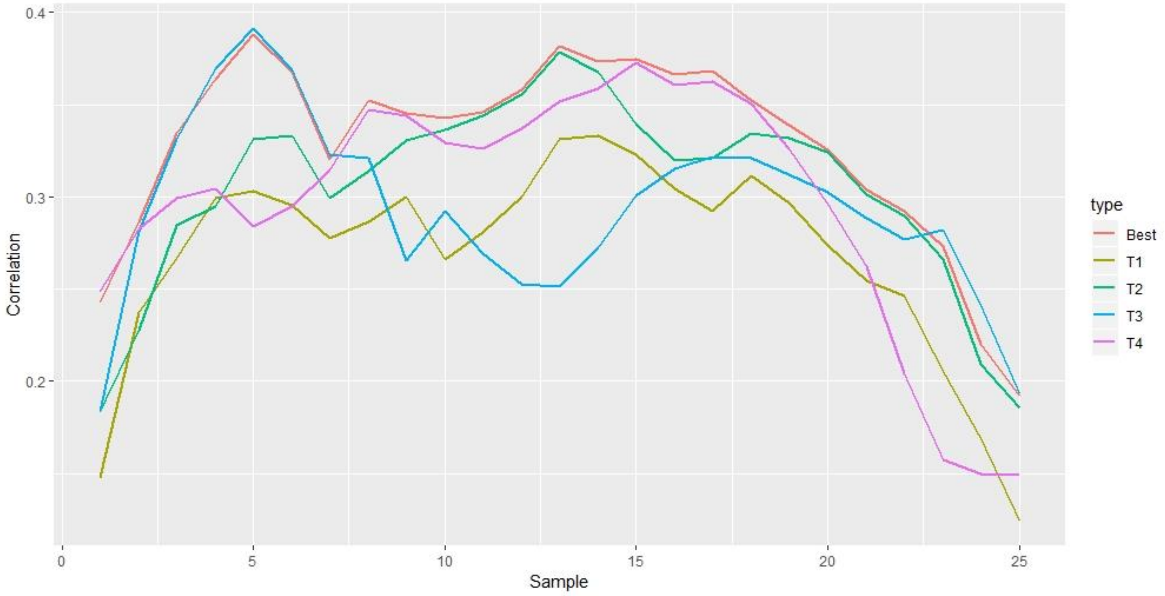

Figure 22: Plot of the correlation by sample for each tree and the "best method".

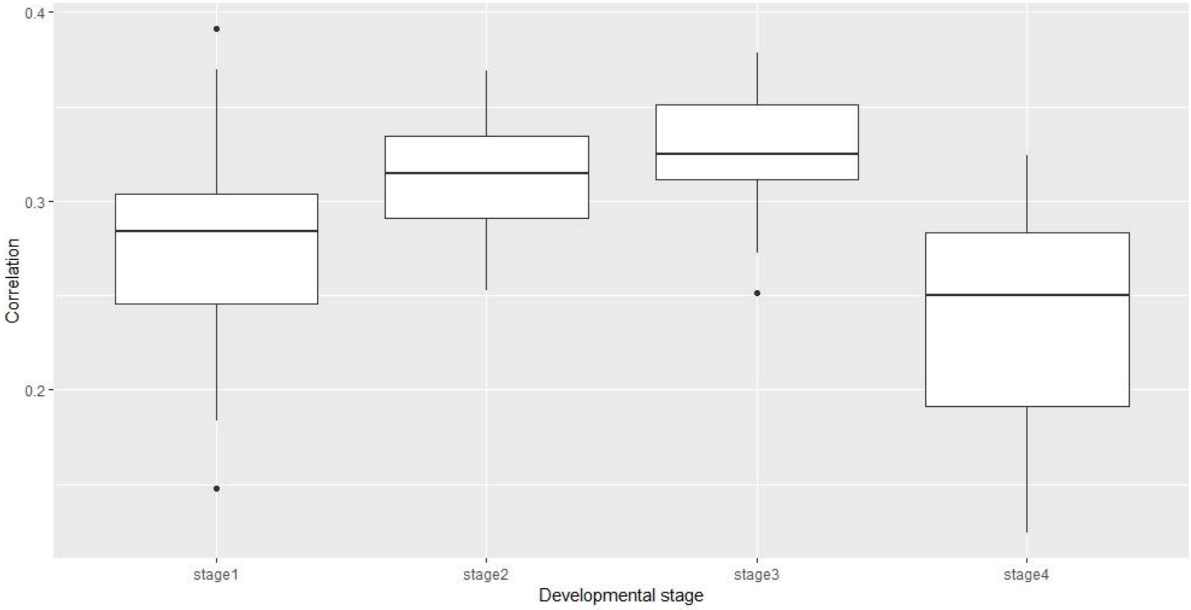

Figure 23: Boxplot of the correlation by sample values by developmental stage.

Table 6: ANOVA table describing the variance between the four developmental stages.

	Degrees of freedom	Sum of squares	Mean sum of squares	F-value	P-value
Developmental stage	3	0.125	0.042	20.4	2.62e-10
Residuals	96	0.196	0.002		

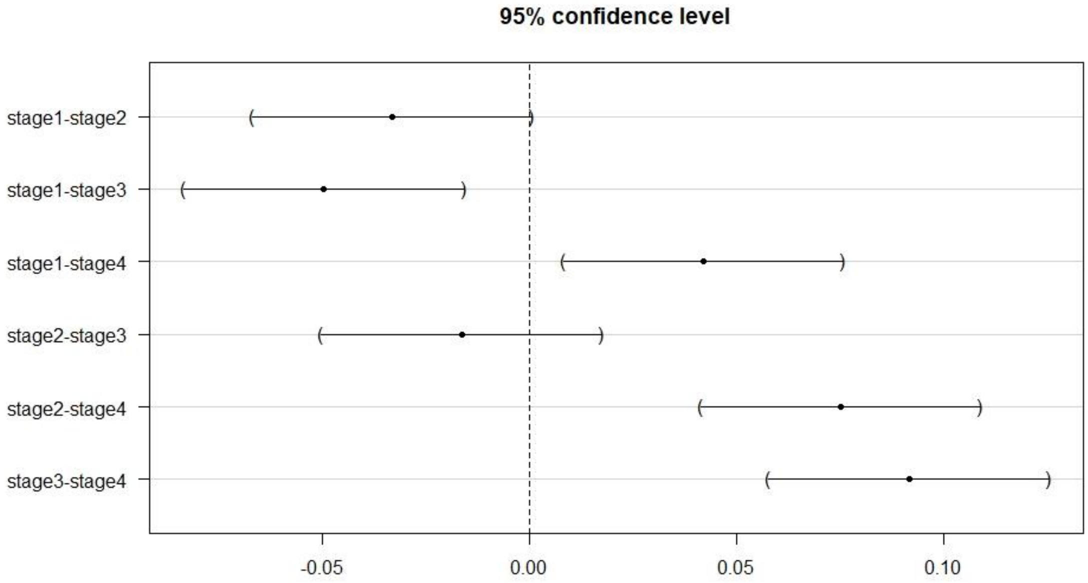

Figure 24: Plot of Tukey test for the different developmental stages. Phloem (stage1), expanding xylem (stage 2), SCW forming xylem (stage 3), late xylem (stage 4).

4.8 Clustering and Heat Maps

To see if the proteins abundance series and gene expressions series would be clustered similarly, a heat map and dendrogram were produced from the transcriptomics data and the proteomics data separately. For the protein data, the moving average calculated data set was used. The dendrograms and heat maps for the proteomics data and transcriptomics data are shown in Figure 25 and Figure 26 respectively.

The dendrograms were cut into clusters and then matched by protein names in the clusters from the other data set. Matching these clusters showed that the dendrograms in the transcriptomic data and proteomics data did not cluster similarly (Table 7).

The heat maps accompanying the gene expression cluster showed how the genes neatly divided into different expression patterns and four clusters were distinct. Most of the gene expression series were similarly reproduced in all the four trees. The heat map produced from the proteomics data also showed clear clusters but in almost all cases the pattern was not reproduced for all the trees. The heat map of the proteomics data also shows the lack of detected protein abundance in many of the samples.

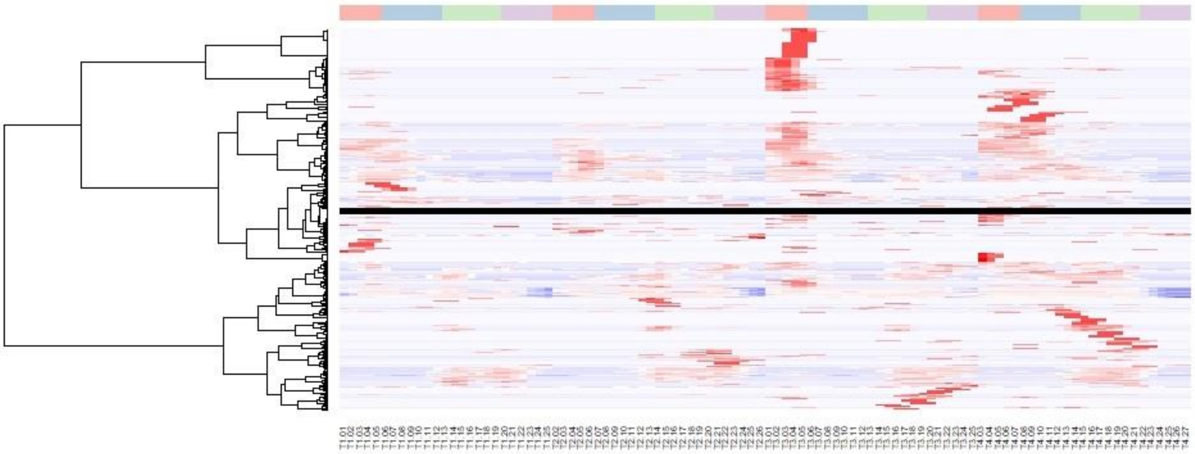

Figure 25: Heat map and dendrogram of the proteomics data treated with moving average calculations. The y-axis indicates the proteins and the x-axis indicates the sample series. In the heat map, red indicates up-regulation, blue indicates down-regulation. The different development stages are indicated in the color bar on top of the heat map. Phloem (red), early xylem (blue), SCW forming xylem (green), late xylem (purple).

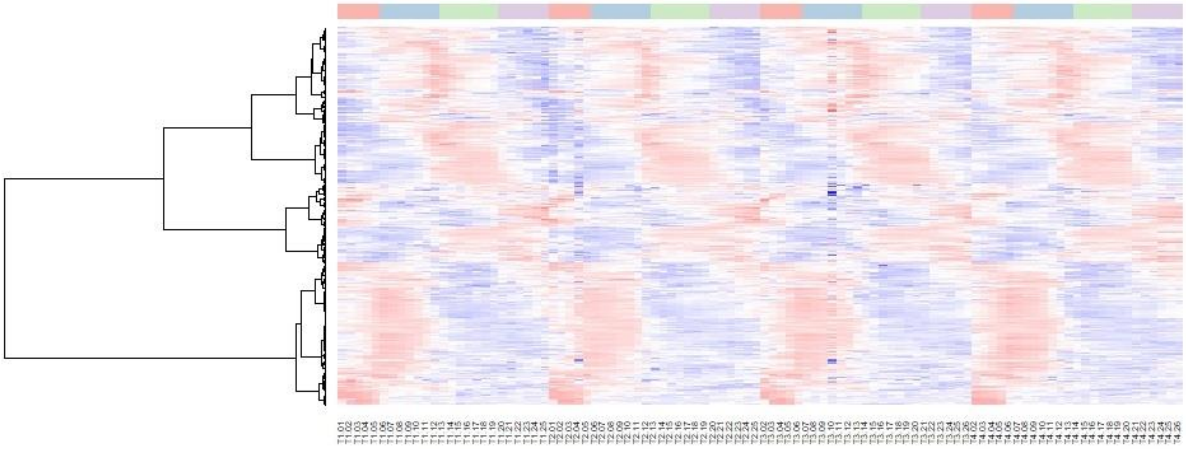

Figure 26: Heat map and dendrogram of the transcriptomics data. The y-axis indicates the genes and the x-axis indicates the sample series. In the heat map, red indicates up-regulation, blue indicates down-regulation. The

different development stages are indicated in the color bar on top of the heat map. Phloem (red), early xylem (blue), SCW forming xylem (green), late xylem (purple).

Table 7: Tables of the clusters and corresponding match score. Each protein cluster has been matched with the gene expression clusters and the intersection divided by the total number of unique genes in both clusters has been calculated and the protein was matched with the gene expression cluster with the highest score. This was done for the raw proteomics data set and the proteomics data set treated with moving average calculations.

Raw data			**Using moving average treated proteomics data**		
Proteomics cluster number	Transcriptomics cluster number	Intersect / Total gene number	Proteomics cluster number	transcriptomics cluster number	Intersect / Total gene number
1	4	0.125	1	4	0.094
2	7	0.124	2	4	0.058
3	8	0.195	3	2	0.085
4	4	0.072	4	4	0.092
5	7	0.078	5	4	0.165
6	4	0.050	6	4	0.181
7	4	0.189	7	8	0.119
8	2	0.134	8	7	0.069
9	4	0.025	9	8	0.169
10	7	0.052	10	7	0.076

The heat maps were also produced with average transcriptomics data set and the moving average and “best method” treated proteomics data. The purpose of this was to eliminate some of the incomplete data in the proteomics data set and better cluster the protein series. Here, the proteomics data were clustered into two distinct groups (Figure 27). One group with more expression in early developmental stages (phloem and cambium) and the other with more expression in later xylem developmental stages. The transcriptomics data (average of the four trees) were clustered into two similar main groups (Figure 28). The two dendrograms were cut into two clusters. The intersection between the clusters divided by the total number of genes in each cluster was calculated to 0.63 and 0.61 for the clusters with the highest matches.

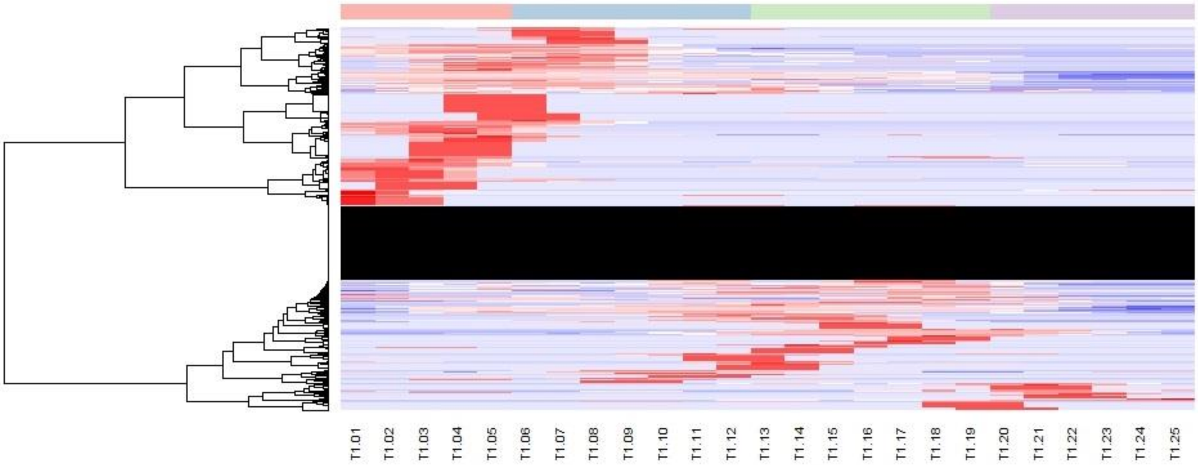

Figure 27: Heat map and dendrogram of the "best method" proteomics data treated with moving average calculations. The y-axis indicates the proteins and the x-axis indicates the sample series. In the heat map, red indicates up-regulation, blue indicates down-regulation. The different development stages are indicated in the color bar on top of the heat map. Phloem (red), early xylem (blue), SCW forming xylem (green), late xylem (purple).

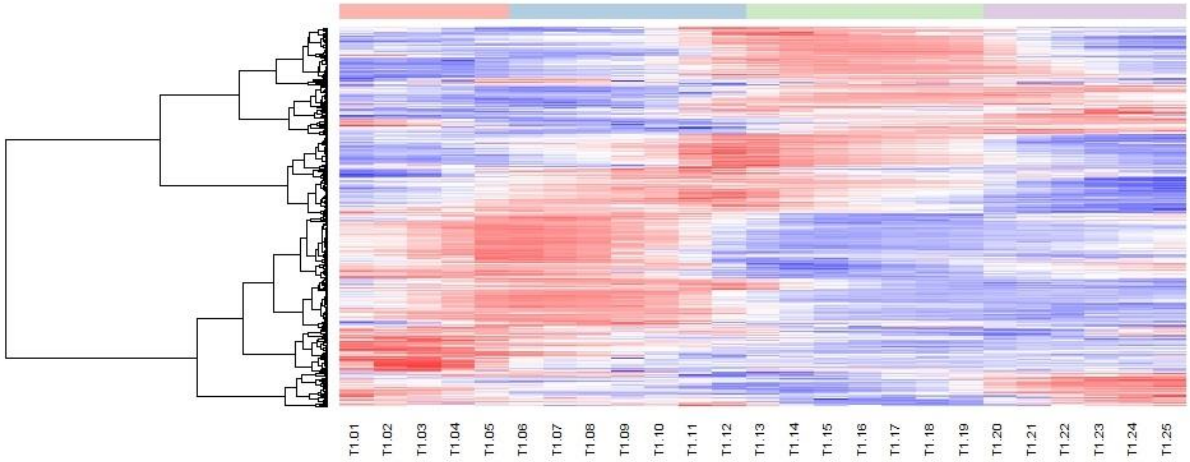

Figure 28: Heat map and dendrogram of the "best method" proteomics data treated with moving average calculations. The y-axis indicates the proteins and the x-axis indicates the sample series. In the heat map, red indicates up-regulation, blue indicates down-regulation. The different development stages are indicated in the color bar on top of the heat map. Phloem (red), early xylem (blue), SCW forming xylem (green), late xylem (purple).

4.9 The On-Off Switch Method

The numeric result of the on-off switch method was shown in Table 8. 202,900 entries were considered in this analysis. In 0.2 percent of entry pairs, or 41,161 entries, both gene expression and protein abundance were above 0. In 0.033 percent of entry pairs, or 6,665 entries, neither gene expression nor protein abundance were above 0. In 0.0018 percent of entry pairs, or 365 entries, gene expression were 0 and protein abundance were above 0. In 0.793 percent of entry pairs, or 161,374 entries, both gene expression were above 0 and protein abundance were 0.

Table 8: Overview of the relation of protein and gene expression by existence of expression.

Number of entries:	Protein abundance > 0	Gene expression = 0	Total
Gene expression > 0	41,161	154,709	195,870
Protein abundance = 0	365	6,665	7,030
Total	41,526	161,374	32,539 (diagonal sum)

Fractions:	Protein abundance > 0	Gene expression = 0	Total
Gene expression > 0	0.2	0.76	0.96
Protein abundance = 0	0.0018	0.033	0.0348
Total	0.2018	0.793	0.233 (diagonal sum)

Overall total number of entries:	202,900		

The number of matches (on and on or off and off) were counted for each protein/gene pair. I.e. the sum of matches for a gene expression/protein abundance series could be between 1 and 100. A density plot (Figure 29) was made from the results. The number of matches sharply peaks at one and then rapidly decreases.

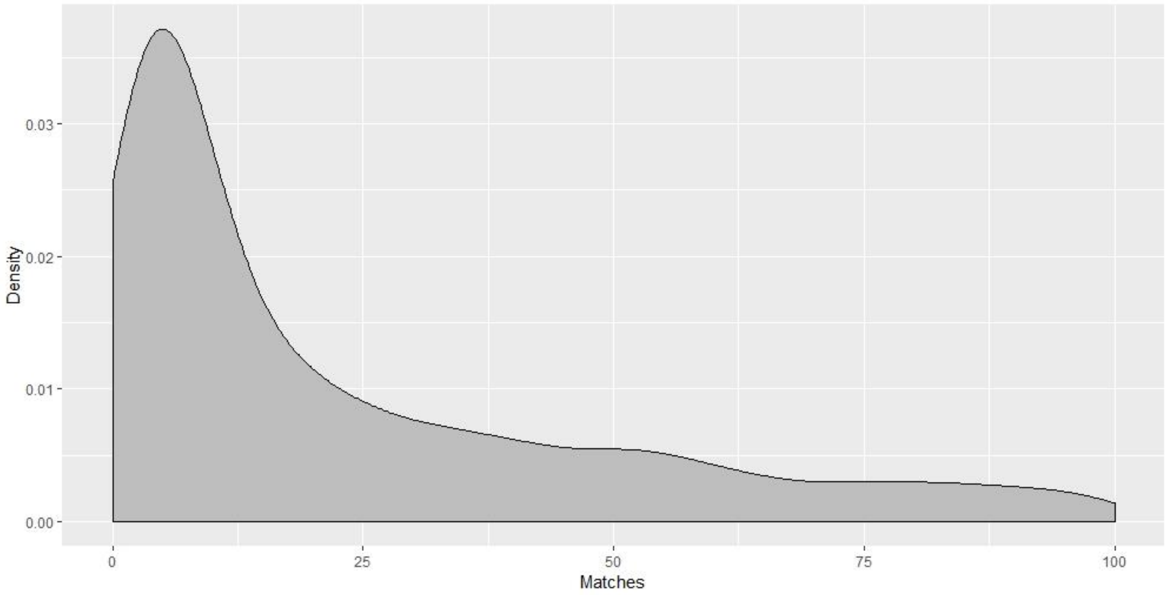

Figure 29: Density of number of matches for protein- gene expression pairs.

The number of matches by sample gives very different results for each tree, but for all trees the number of matches decreases at the edges of the series (Figure 30). There was also a peak in the match-numbers of all trees at cambium and a peak for tree 1, 2 and 3 in reprogramming event 3 between expanding xylem and lignified xylem.

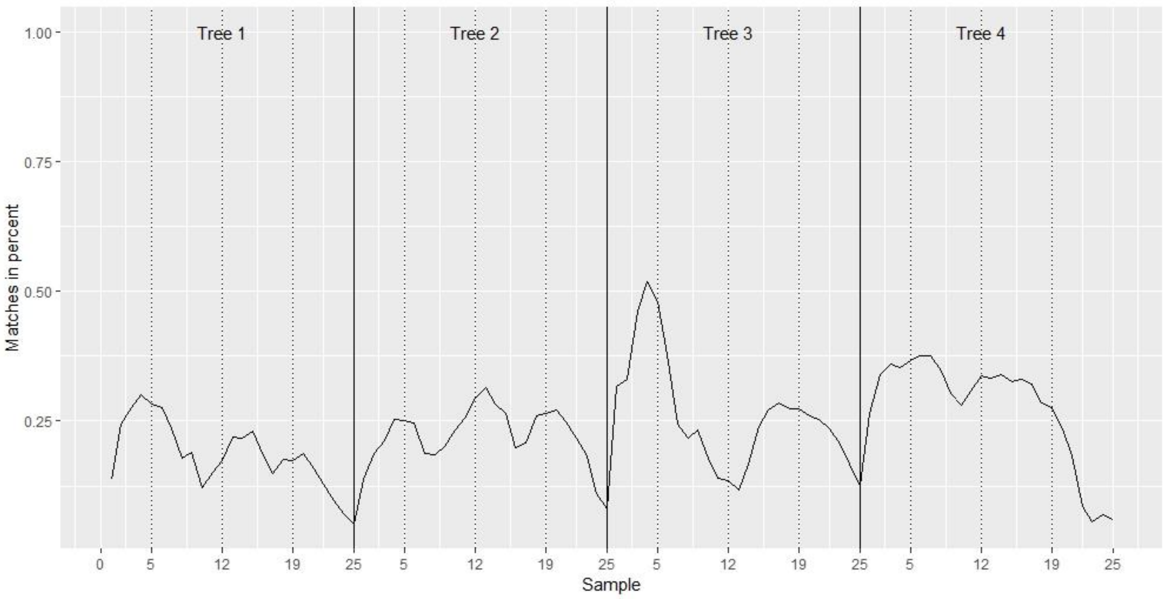

Figure 30: Plot of the number of matches of protein abundance (on-off) and gene expression (on-off) across all genes by sample.

The likelihood of finding any value of protein abundance above 0 was calculated for different levels of gene expression (Figure 31). Here a gene expression level considers all expressions values of the given level or higher. At the higher levels of gene expression there was a higher chance of finding protein abundance. There was a gentle slope of increased likelihood until approximately gene expression level 17 and above where the likelihood of finding protein abundance above zero was 68%. The likelihood dropped afterward and at expression level 18.4 and above the chance of protein abundance above 0 was 54%. Afterwards the likelihood rapidly ascended towards 100%. As the gene level threshold rises fewer genes are considered when calculating likelihood of finding proteins. The results are therefore less reliable for higher gene levels.

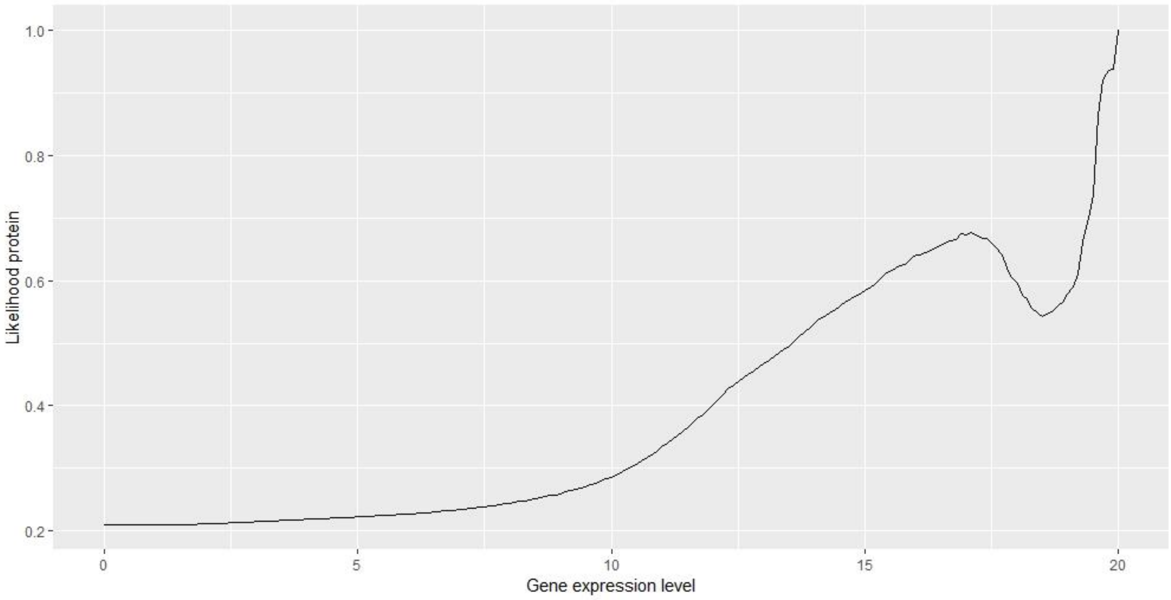

Figure 31: Plot of the relationship of gene expression level (VST) (considering gene expressions of the given level or higher) and the likelihood of finding protein abundance larger than 0 in these samples.

4.10 Protein Isoforms

Since 296 of the proteins had isoform variants, it was speculated that the sum of isoforms protein abundance correlated better with the transcriptomics data. Figure 32 shows a density plot of the correlations between the sum of the isoform protein abundance and the gene expression. The summed isoforms were saved into one single vector and correlated with the corresponding transcriptomics vector, yielding a correlation coefficient of 0.395. In comparison the full correlation between the transcriptomics data and the proteomics data was 0.295.

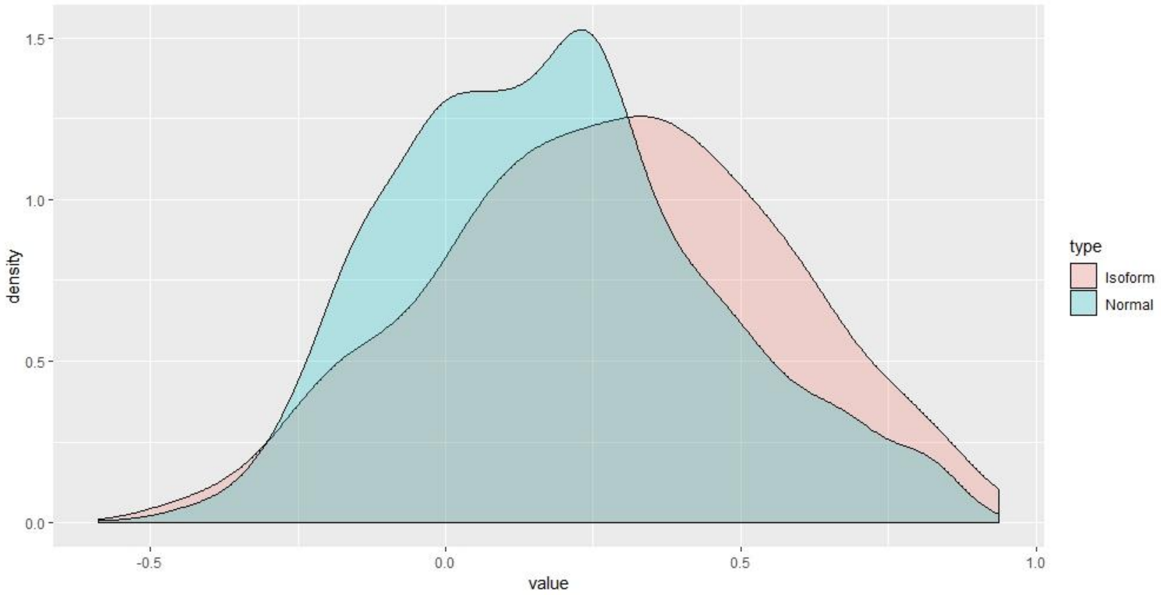

Figure 32: Density plot of correlations between sums of isoform abundance and gene expression compared with the row correlations. Both utilized the moving average calculated proteomics data.

4.10.1 Isoform Proteins with Sequential Expression

Some protein isoforms were expressed in sequence, meaning the protein abundance of two isoforms was detected in different samples or varying amounts in the same samples (one being reduced while the other increases). The gene for Potri.016G014500 was expressed as Potri.016G014500.1 and Potri.016G014500.2, two UDP-glucosyl transferase proteins. Potri.016G014500.1 was expressed in phloem and in tree number three and tree number four, it was partially replaced by Potri.016G014500.2 (Figure 33). Another gene Potri.014G068200 was the expression basis for the Eukaryotic aspartyl protease family proteins: Potri.014G068200.1 and Potri.014G068200.3. The gene expression peaked in the cambium and protein abundance was found in the phloem and expanding xylem (Figure 34).

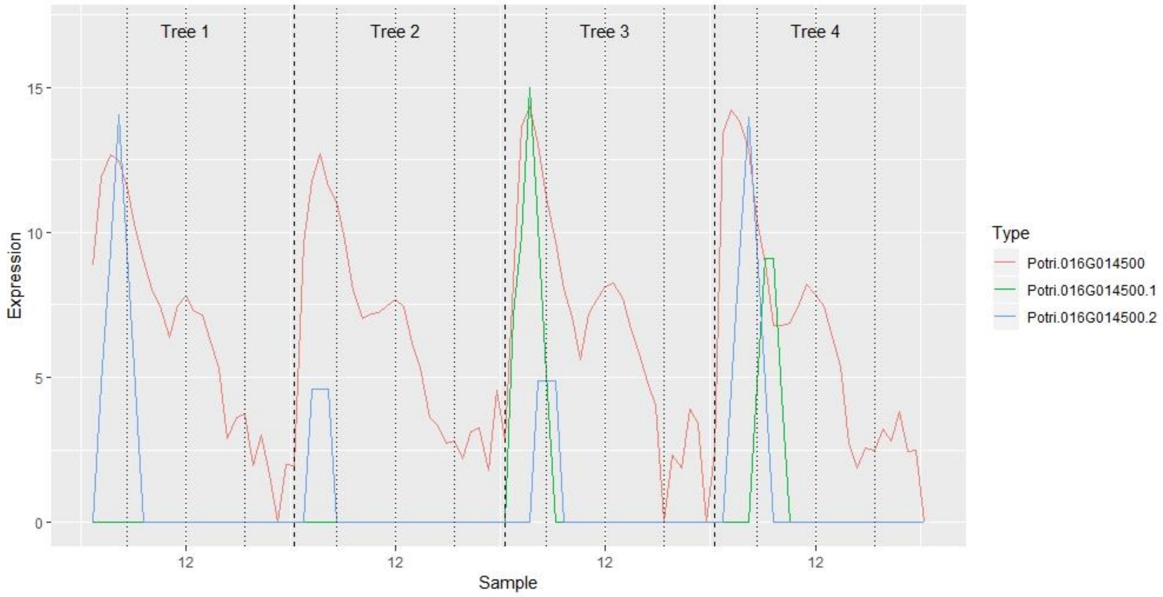

Figure 33: Expression profile of the gene encoding Potri.016G014500 and corresponding UDP-glucosyl transferase proteins (Potri.016G014500.1 and Potri.016G014500.2).

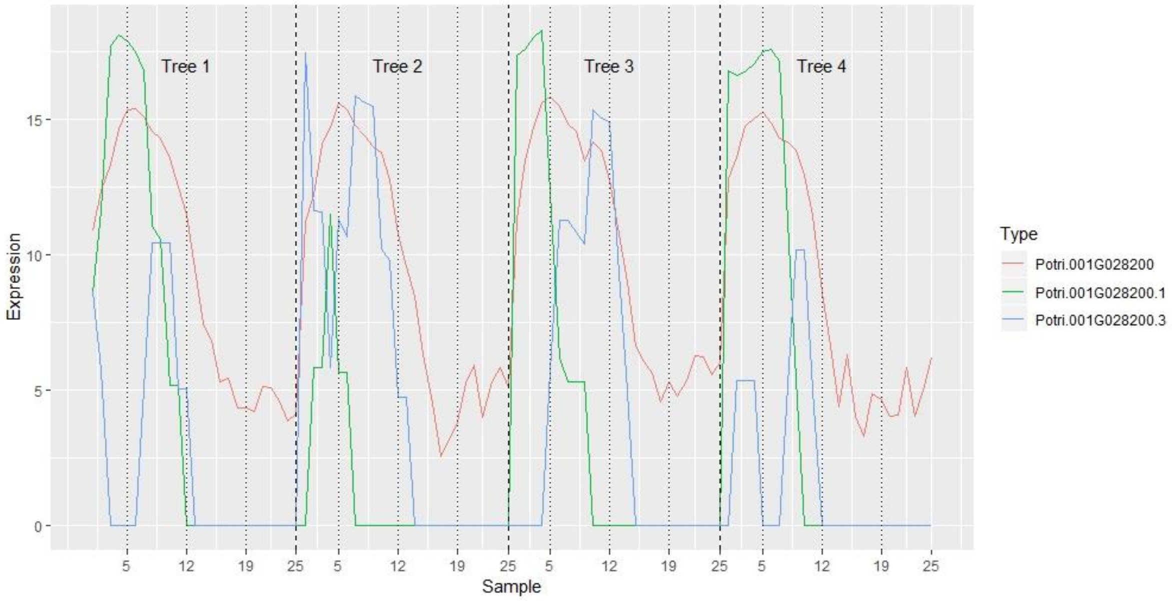

Figure 34: Expression profile of the gene encoding Potri.014G068200 and corresponding Eukaryotic aspartyl protease family proteins (Potri.014G068200.1 and Potri.014G068200.3).

4.10.2 Isoform Proteins with Overlapping Expression

Some proteins had isoforms that were expressed overlapping in the same samples. Sucrose synthase Potri.017G139100.2 and Potri.017G139100.3 overlap completely in tree 3 (Figure 35). Several proteins were expressed similarly as the proteins encoded by Potri.018G145900 (Figure 36) where the isoform proteins seem to overlap randomly and not in a clear sequence.

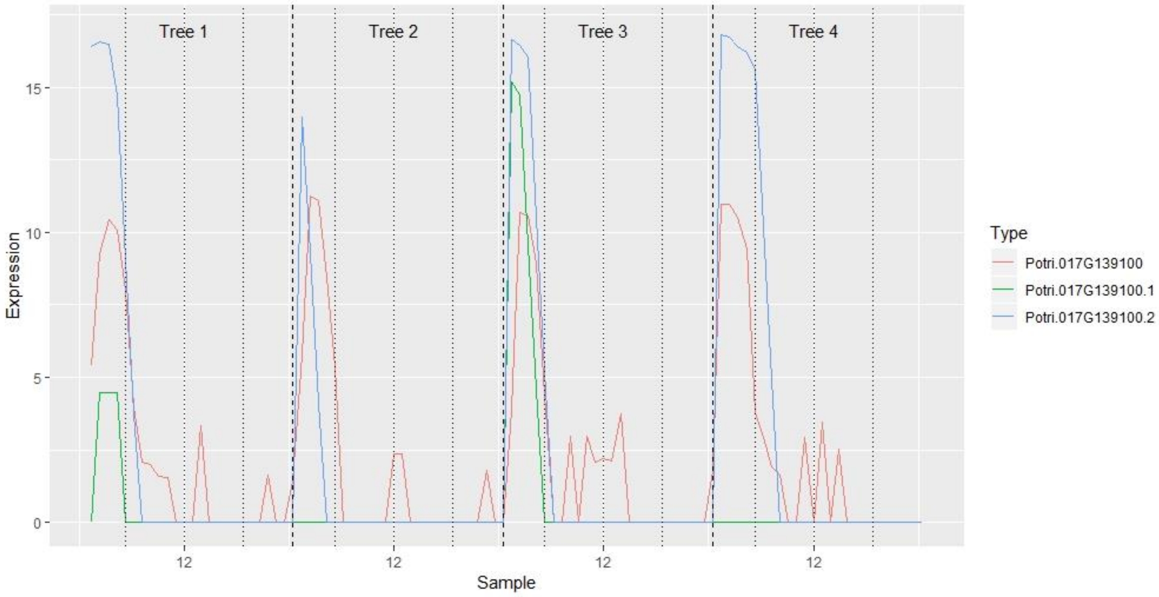

Figure 35: Expression profile of the gene encoding Potri.017G139100 and corresponding sucrose synthase 5 proteins.

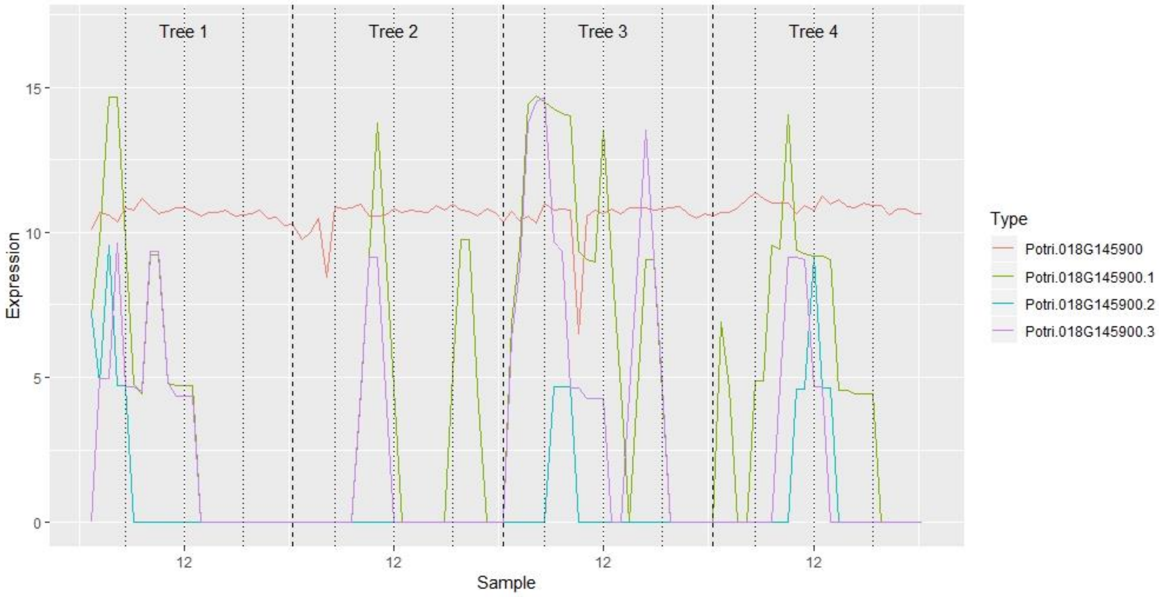

Figure 36: Expression profile of the gene encoding Potri.018G145900 and corresponding N-terminal nucleophile aminohydrolases (Ntn hydrolases) superfamily proteins.

4.11 GO Enrichment Analysis

The GO enrichment analysis of the proteins which abundance series correlated highly with their corresponding gene expression identified 246 GO terms in the category biological processes, 126 in the category molecular function and 48 in the category cellular component.

Many of these with very low p-value. The ten GO terms with the lowest p-value in each category are reported in Table 9, Table 10 and Table 11. The full report is provided in Appendix B.

Table 9: Top ten GO terms of the biological process category that are overrepresented in the highly correlating protein subset

GO ID	P-value (FDR)	Statistics	Description
GO:0006195	4.825e-10	20/829 | 36/14903	purine nucleotide catabolic process
GO:0009207	4.825e-10	20/829 | 36/14903	purine ribonucleoside triphosphate catabolic process
GO:1901658	4.825e-10	20/829 | 36/14903	glycosyl compound catabolic process
GO:0072523	4.825e-10	20/829 | 36/14903	purine-containing compound catabolic process
GO:0009146	4.825e-10	20/829 | 36/14903	purine nucleoside triphosphate catabolic process
GO:0046130	4.825e-10	20/829 | 36/14903	purine ribonucleoside catabolic process
GO:0006152	4.825e-10	20/829 | 36/14903	purine nucleoside catabolic process
GO:0042454	4.825e-10	20/829 | 36/14903	ribonucleoside catabolic process
GO:0009143	4.825e-10	20/829 | 36/14903	nucleoside triphosphate catabolic process
GO:0009154	4.825e-10	20/829 | 36/14903	purine ribonucleotide catabolic process

Table 10: Top ten GO terms of the Molecular function category that are overrepresented in the highly correlating protein subset

GO ID	P-value (FDR)	Statistics	Description
GO:0004455	**2.663e-09**	11/967 | 13/19622	ketol-acid reductoisomerase activity
GO:0016614	3.341e-09	55/967 | 392/19622	oxidoreductase activity, acting on CH-OH group of donors
GO:0070003	3.361e-09	16/967 | 34/19622	threonine-type peptidase activity
GO:0004298	3.361e-09	16/967 | 34/19622	threonine-type endopeptidase activity
GO:0003735	3.541e-09	103/967 | 487/19622	structural constituent of ribosome
GO:0005525	4.032e-09	72/967 | 321/19622	GTP binding
GO:0032561	4.032e-09	72/967 | 321/19622	guanyl ribonucleotide binding
GO:0005198	4.097e-09	117/967 | 548/19622	structural molecule activity
GO:0003924	4.144e-09	56/967 | 185/19622	GTPase activity
GO:0051082	4.340e-09	22/967 | 88/19622	unfolded protein binding

Table 11: Top ten GO terms of the cellular component category that are overrepresented in the highly correlating protein subset

GO ID	P-value (FDR)	Statistics	Description
GO:0005839	8.286e-10	16/360 | 34/6017	proteasome core complex
GO:0005874	9.274e-10	18/360 | 33/6017	microtubule
GO:0044464	1.620e-09	305/360 | 3667/6017	cell part
GO:0043229	1.683e-09	149/360 | 1580/6017	intracellular organelle
GO:0043226	1.683e-09	149/360 | 1580/6017	organelle
GO:0044424	1.782e-09	262/360 | 2754/6017	intracellular part
GO:0043234	1.819e-09	101/360 | 777/6017	protein complex

GO:0032991	1.852e-09	208/360 | 1363/6017	macromolecular complex
GO:0044444	2.067e-09	155/360 | 1027/6017	cytoplasmic part
GO:0005622	2.143e-09	134/360 | 1144/6017	intracellular

5 Discussion

5.1 Data treatment Improved the Correlation Scores

Two treatment methods were used to improve the relationship between the transcriptomics data and the proteomics data. These were moving average calculations which smoothed the protein abundance series and the “best method” which removed suboptimal trees from the proteomics data and only used the highest correlating tree series. Both these data treatments were used only on the proteomics data set. The transcriptomics data did not need those treatments, utilizing them did not improve the result significantly. The expression patterns of the transcriptomics data generally had smooth transitions and the patters were usually reproduced in all four trees.

Best method improved the full correlation between gene expression and protein abundance from 0.256 to 0.298 (Table 4). In correlation by row (expression of a single gene across the series was correlated against the corresponding protein abundance series for every gene (row) in the data sets), the “best method” improved the median and the max correlation from 0.11 to 0.17 and 0.91 to 0.93 respectively (Figure 20). The “best method” successfully eliminated the incomplete proteomics data provided by some trees and therefore improved the correlation score. According to Figure 20 it seems that a very large portion of the data set was improved slightly by the “best method”.

Moving average mainly improved a large subset of the proteomics data, indicating that a group of proteins were more receptive to the data treatment. Moving average improved the full correlation between gene expression and protein abundance from 0.256 to 0.295. In correlation by row it improved the median and the max correlation from 0.11 to 0.27 and from 0.91 to 0.96 respectively (Figure 20). By smoothing the protein abundance series, moving average calculations improved the relationship between protein abundance and gene expression substantially for a subset of genes. However, another large subset seemed to be largely unaffected by the treatment. These two subsets were clearly visualized in a density distribution of the correlation scores of all the individual genes and their corresponding protein abundance (Figure 20). In the figure, it seemed that a section of the protein/gene pairs “left” the group distributed around zero and instead moved towards a distribution around 0.28 correlation. In Figure 18 B (a dot plot of all the gene expression values against their corresponding moving average treated protein abundance values with zero entries removed) the protein abundance values were divided into three main groups with different protein abundance values. This is different from the division shown in Figure 20 since Figure 20 is based on correlation coefficients between series and the Figure 18 dot plots are based on all the entries of the data by themselves. They are separated this way due to the method of the moving average calculations. Some proteomics abundance entries were divided by three, some were divided in half, while some remained around their original value splitting the proteomics entries in three groups.

The two methods cumulated well without much diminishing returns when they were combined, hinting that they fixed two separate problems with the proteomics data set. Namely, high fluctuation/noise in general and incomplete data in some trees. They individually increased the full correlation score by approximately 5 percentage points (from 0.256 to 0.295 and 0.298) and together by approximately 10 percentage points (from 0.256 to 0.347). This proved their usefulness for improving the relationship between protein abundance and gene expression.

When zero entries were removed, the treatments did not improve the full correlation (Table 5). This means that the main contribution from the two data treatments were increasing the correlation coefficient by diminishing effect of proteomics zero entries. When the zero entries were already removed from the data sets, the data treatments instead diminished the correlation coefficient, both by themselves and combined.

The correlations of this thesis, both before and after data treatments, were low compared to other studies. Results in this thesis showed a very low correlation between gene expression and protein abundance which might suggest that there was a high amount of post-transcriptional regulations or inferior proteomics data, or a combination of both. Studies suggest that mRNA levels determine around 30-50% of the variation in protein levels (Csárdi et al., 2015; de Sousa Abreu et al., 2009). These studies are usually global correlations and not correlations across samples in a series. The Pearson’s correlation test suggested that 7% of the variance in protein levels was explained by mRNA levels in the raw data and 12% using the treated protein data. The gene correlation, where the molecule number across the series was summed before correlation, yielded similarly a 10% correlation coefficient. One thorough study suggests an even higher explanation of variance in protein based on expression levels in RNA. In (Csárdi et al., 2015) 24 studies of budding yeast were analyzed, and it seems to be that 85% of the variation in protein levels was explained by variations in mRNA levels.

5.2 Different Levels of Correlation in the Different Developmental Stages

Correlation by sample yielded significantly different scores for samples in the different developmental stages (Figure 22). Correlation scores were in general higher in expanding xylem developmental stages and lower in late xylem stages and in phloem. The correlation scores plummet for all trees in samples near the two edges of the series. Using the best method protein data set, there was a sharp peak in cambium with a correlation coefficient close to 0.4. This could indicate different levels of post-transcriptional regulation in the different developmental stages. If the difference was due to post-transcriptional regulation, there was less post-transcriptional regulation in cambium and more at the edges of the sample series.

These results should be observed in the context of total expression by sample (Figure 17). The gene expression levels by sample were very stable across all developmental stages while the pattern of protein expression was more variable, showing a similar pattern to that of the correlation by samples. The variations in correlation therefore seem to be mainly caused by variable protein abundances. See also the on-off switch results (Figure 30) which shows fewer matches at the samples on the edges of the series.

5.3 The ON-OFF Switch Showed that Protein Abundance was Rarely Registered Without Gene Expression

Transforming the data set to 1s and 0s enabled easier comparison of the two data sets by considering whether gene expression or protein abundance was present in a specific entry. Only in 0.18% of the comparisons were protein abundance found while no gene expression was found. In total, there were no gene expression in 3.48% of the entries, meaning 5.2% of the time gene expression was not present, the corresponding protein abundance was found. This shows that protein abundance in the proteomics data was largely dependent on gene expression.

Additionally, it was shown that increased levels of gene expression increased the likelihood of protein abundance being present (Figure 31). It has been suggested in other studies that not only mRNA presence is important but also the concentration of a specific mRNA makes a difference in the likelihood of observing protein abundance for the corresponding protein

(Figure 6) (Vogel & Marcotte, 2012). In yeast, increased gene expression increased the likelihood of observing the corresponding protein. The relationship between those two were not linear. Instead, the likelihood of finding protein abundance was stably low for gene expression until about an mRNA abundance of 10, where it rapidly rose and plateaued at a 90% chance of finding protein abundance which seemed to be the maximum chance. Similarly, with the *Populus tremula* data, increased gene expression increased the likelihood for protein abundance. The protein abundance series of the sucrose synthases, cellulose synthases and xylem related peptidases illustrate this trend well as protein abundance was mainly found near the gene expression peaks (Figure 15).

5.4 Isoforms Correlate Better with the Gene Expression Patterns When Their Protein Abundances were Summed Together

The proteomics data set contained 296 unique proteins that had one or more isoform variants. In total, these counted to 650 proteins. The correlation coefficient obtained when correlating the summed isoforms data against their corresponding gene expression was 0.395 (full correlation). This was a score substantially higher than the correlation score obtained when correlating the moving average proteomics data with the transcriptomics data (0.295) (full correlation) (Figure 32). The reason for this was likely that in many cases there were different isoforms expressed in different samples or in amounts summed together better compared to the gene expression (Figure 33 and Figure 34). The transcriptomics data did not distinguish between splice variants, and therefore summing the different isoforms provided a more accurate comparison. Still, the isoform data was only a small subset of the whole proteomics data and the increased correlation could also be due to fortuitous sampling.

In practice, distinguishing protein isoforms is difficult. In the proteomics study by Obudulu et al. (2016) the proteins were fragmented by enzymatic digestion before further identification procedures. A study by Stastna and Van Eyk (2012) suggests that identification and quantification of isoforms is better done with intact proteins. There was a chance that many isoform protein variants were not identified in the proteomics data, as they were digested before MS analysis. The isoform data set was a small subset (contained information from 650 abundance series) of the complete data set (contained information from 2029 abundance series). With more complete proteomics data, a better general analysis of isoforms in *Populus tremula* could have been conducted.

5.5 Clustering and Heat Maps

The full proteomics heat map/dendrogram plot (Figure 25) was very uneven compared to the full transcriptomics heat map/dendrogram plot (Figure 26) and it seems that some groups of proteins were more easily profiled in trees number three and four. Two separate sections seem to have protein abundance patterns that were repeated for each tree. The first of these sections has protein expression upregulated mainly in phloem and the second section has protein up regulated in SCW forming xylem and late xylem. In the rest of the heat map, the protein abundance patterns were mostly not repeated for each tree. There seems to be more series with protein abundance in tree number three and tree number four. There was a large group of proteins that were upregulated in phloem for tree number three that was not reproduced in any other trees. In tree number four there was a lot of up-regulated protein around the cambium in one section and in another section, there were many proteins that were up-regulated in expanding xylem and SCW forming xylem. The up-regulations in those two sections were also not reproduced in the other trees.

The reduced proteomics dendrogram and the reduced transcriptomics dendrogram both clustered into two main clusters. The reduced proteomics data dendrogram were split approximately in the middle of expanding xylem, while the reduced transcriptomics dendrogram were split in the second reprogramming event. The clusters in the proteomics dendrogram were paired up with their corresponding cluster in the transcriptomics dendrogram. Both pairs only had around 60% genes in common showing that the transcriptomics data and the proteomics data did not cluster similarly. The one cluster in the transcriptomics dendrogram that contained a subset of genes that were expressed both in phloem and in late xylem had no clear equivalent in the proteomics data.

5.6 Data Quality

To study the relationship between protein abundance and gene expression, data quality was imperative. Producing the data used for this thesis would not be possible if not for drastic, recent improvements in both RNA-seq and proteomics. Unfortunately, this proteomics data still had many limitations.

Differences in rows (protein and genes) and columns (sample numbers) between the data sets caused information loss. The transcriptomics had 25,434 identified genes that had no corresponding protein in the proteomics data set, meaning that the proteomics methods in the study failed to identify around 90% of the proteome, assuming the transcriptomics study managed to identify most of the genes in the genome. In a similar study of *Drosophila* brain tumor cells 6,200 tissue-specific proteins were identified. This corresponded to approximately 70% of all the protein-coding mRNA (Jüschke et al., 2013). *Populus trichocarpa* has a protein count of 51,717 according to the NCBI (Institute, 2006) and in this study of the proteome and transcriptome of *Populus tremula*, 28,294 protein-encoding genes and 3,083 proteins were identified. Six samples were removed from the transcriptomics data and eleven from the protein data. Choosing which to remove was based on an optimization considering only the untreated proteomics data and perhaps more sophisticated methods would align the data set more precisely regarding the different data treatments. After matching the two data sets by genes, the proteomics dataset still had 920 proteins with series in which all entries had zero protein abundance that needed to be removed.

The proteomics data had mostly zero entries when considering the 2860x100 (proteins x samples) matrix (Table 3). It was expected that cells in many cases are conservative when it comes to the translation of proteins compared to transcription. However, with a zero rate of 87% in protein abundance in the matched data with full zero series were removed seems excessive.

Comparing the marker genes from the transcriptomics article by Sundell et al. (2017) with their corresponding protein abundance provided in the proteomics article by Obudulu et al. (2016) gives an impression of the proteomics data's quality. In the transcriptomics article, the five marker genes were used to illustrate how differential expression occurred in the different developmental stages (Figure 37). The marker genes were expected to be expressed in certain patterns based on their corresponding proteins' function and the actual gene expression patterns largely reflected those expectations (Figure 37). The sucrose synthase (Potri.004G081300) was expected in the phloem tissue to maintain the correct ratio of sucrose, glucose and fructose. Cyclin-dependent kinase (Potri.016G142800) and was thought to be involved in cell cycle stages. It should therefore be highly expressed in the cambium. Potri.001G240900, an expansin-encoding gene, was highly expressed in the xylem expansion zone. Potri.004G059600 encodes a cellulose synthase family protein. Cellulose is primarily needed for secondary cell wall formation and the cellulose synthase is expected to be highly expressed in expanding xylem. Potri.011G044500 encodes a protein similar to endonuclease (Sundell et al., 2015) and is expected to be found at elevated levels in xylem undergoing apoptosis. Only three of those five were found in the proteomics data: synthase, cellulose synthase and Cyclin-dependent kinase. The sucrose synthase protein and the cellulose synthase was expressed in its expected pattern.

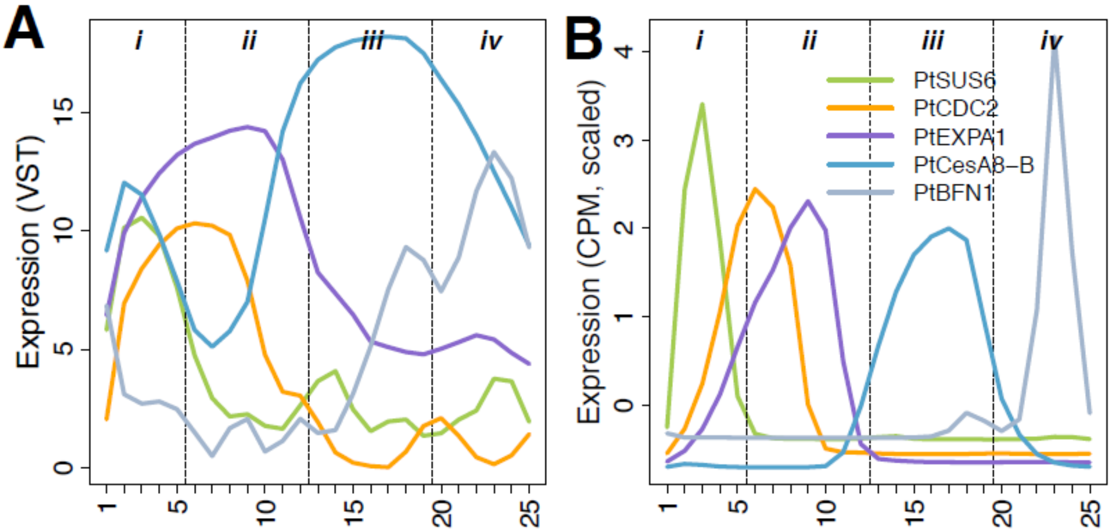

Figure 37: Expression of marker genes from the transcriptomics article by Sundell et al. (2017). "Expression is shown with (A) the variance stabilized transformation (VST) and (B) scaled counts per million (CPM, calculated as 2VST, scaled: mean centered and normalized by the standard deviation of each gene)." Gene explanation: PtSUS6/Potri.004G081300: Sucrose Synthase (SUS), PtCDC2/Potri.016G142800: cyclin kinase, PtEXPA1/Potri.001G240900: expansin, PtCesA8-B/Potri.004G059600: cellulose synthase family protein, PtBFN1/Potri.011G044500: Xylem specific proteases and nucleases.

While the gene expression patterns in most cases were reproduced in all trees, this was almost never the case for the proteomics data. The reproduction in the gene expression patterns strengthens the claim that the patterns were close to the true gene expression pattern in the series and the lack of this in the proteomics data made any claim of true protein abundance levels in these samples unreliable. This is especially apparent when comparing the full heat maps of the two data sets (Figure 25 and Figure 26). The heat map of the transcriptomics shows some color in almost every field and there were smooth transitions and the patterns were largely repeated four times for each tree. The patterns in the proteomics-based heat map were mostly not repeated for each tree. Furthermore, the protein abundance series in the cases section (chapter 4.2) gives examples of protein abundance series that are not reproduced. Especially the cellulose synthases (Figure 14) were lacking.

5.7 Protein Abundance and Gene Expression are Difficult to Compare due to the Macromolecules' Biological Attributes

It is far easier to obtain good data from transcriptomics methods than protein profiling methods. The hybridization capabilities of RNA and DNA simplifies the process drastically. Processing proteomics data is a highly difficult multi-step procedure.

The method used to characterize the lipids of the proteins was the ultra-performance liquid chromatography/quadrupole time-of-flight mass spectrometry system. Mass spectroscopy-based proteomics have developed quickly in recent years. Unfortunately, eukaryotic plants are highly complex organisms with many protein variants, and they are still difficult to study. Plant cells are in addition resistant to degradation and require especially potent license to access all the proteins. This could cause issues for proteome analysis, since the required process may unintentionally harm the enzymatic process and halt the analysis (Abraham et al., 2013). Several approaches have been done to limit the adverse effects of strong detergent or mechanical lysis, but the optimum for plant cells may not have been found yet. The ionization process of mass spectroscopy is affected by many factors possibly complicating the protein profiling, such as chemical and physical qualities of the amino acids, the other components in the sample such as solvents present in the samples while they are being ionized.

The cell is rarely in a steady state. This is especially true for cells undergoing continuous proliferation, which is a good description of the cells described in this thesis. Both RNA-seq and the proteomics techniques provided in the article by Obudulu et al. (2016) measures gene expression levels and protein levels at one specific moment for each sample in the series. This complicates the comparison since translational rates are often slow and degradation rates can be variable for both protein and gene, they may be difficult to compare as mRNA levels will rise much faster than protein levels.

In some sample series, there were delays between increased gene expression and increased protein abundance. Examples of delay in translation may include two xylem related peptidases: Potri.002G005700.1 and Potri.005G256000.2 (Figure 15: A, C, D). Delay affects the correlation scores detrimentally. A method which considered delays in translation might better illustrate the relationship between gene expression and protein abundance. In the study by Jovanovic et al. (2015) on mouse dendritic cells, they found that after an LPS (Lipopolysaccharide) treatment meant to induce immune response the mRNA levels and protein levels correlated best when they sampled the mRNA after 5 hours of the LPS treatment and the protein levels after 12 hours.

“Translation on demand” proteins generally correlates badly with the gene expression, as they are only sporadically expressed while the gene was expressed at all times at regular rates (Beyer et al., 2004). Examples of possible translation on demand proteins may be found amongst the sucrose synthases (Figure 13 A), amongst the cellulose synthases (Figure 14: D, E, G, H, I, J) and probably among many other proteins in the data set. It can also be observed on the transcriptomics heat map that there are many gene expression that are translated in all samples and some are translated quite evenly across all samples. In the proteomics heat maps, small spikes are abundant.

5.8 Concerning Wood Development

Wood formation initiates in cambium and it is likely that proteins that are highly expressed in developing xylem are important for wood formation in trees (Vander Mijnsbrugge et al., 2000). In the heat map (Figure 27) it is shown that most proteins were highly expressed in around cambium in expanding xylem or phloem. In the rest of xylem developmental stages, the different proteins do not seem to cluster, but rather “slide” across the different tissues (figure 27).

Cellulose synthase and sucrose synthase are thought to be important in the secondary cell wall synthesis (Kalluri et al., 2009). Protein profiles of sucrose synthase documents that that the sucrose synthase proteins were mainly expressed in phloem and they correspond to their gene expression (Figure 13). They are not highly expressed in cells that directly contribute to the central wood. However, they facilitate transport and uptake of photosynthates needed for energy and secondary cell wall formation they are nonetheless vital in wood production. Cellulose is the main component of the cell wall which most cells has. The amount of cellulose in plant cells varies depending on the cell type (Zhong & Ye, 2007). Many of the genes encoding cellulose synthases were highly expressed across all developmental stages, while the protein abundance was more less consistent (Figure 14). The cellulose synthases were highly expressed in the SCW forming xylem and less expressed in expanding xylem and later xylem.

5.9 GO Enrichment Analysis

GO enrichment analysis was performed on the group of protein abundance series that were more responsive to the moving average treatment. This was to see if there were GO terms that were associated with the gene product that were easier to analyze with the proteomics methods used to obtain the proteomics data used in this thesis. The row correlation coefficients calculated between average transcript series and protein abundance series treated with moving average calculation and the "best method" were split into two groups main groups. The proteins largely unresponsive to data treatment and the ones that responded well. The former group had a mean of approximately 0 correlation and the latter had a mean approximately 0.5 correlation (Figure 20 D). The boundary between them seems to be approximately at 0.17 and the 1,233 proteins above that threshold were selected. Many GO terms were found to be overrepresented (full lists in Appendix B), indicating that gene product with specific GO terms were more responsive to data treatment and/or the protein profiling method used.

As the full proteomics data set described only specific developmental stages in the trees it is no wonder that GO terms related to wood production are overrepresented in the set analyzed. Examples of this is the "proteasome core complex" and "microtubule" GO term from the cellular component category. "Proteasome core complex" is likely overrepresented since protein degradation is vital in xylem maturation (Bollhoner et al., 2012). Previously in the thesis (chapter 4.2.4) xylem related peptides have been shown to correlate well with their gene expression series. Microtubules are polymers of tubulin and organized in xylem they are an important component of wood (Oda et al., 2010).

5.10 Future Work

This thesis aimed at analyzing gene expression and protein abundance together through different methods to uncover the dynamics between gene expression and protein synthesis and content in cells. The contemporary digital toolbox offers near endless opportunities for comparing two data sets and only a select few were utilized for this thesis.

Even though the data was separated into two sections, separating the data further into the alternative sections based on the correlation between the protein abundance series and the gene expression series, or by some other criteria, could yield additional information.

The proteins and their characteristics could be inspected more closely to see if they possibly had anything in common (as was done in the GO enrichment analysis). The proteins and expressed genes that correlated highly had more overrepresented GO terms which could be revisited more closely as they could provide some new insight in proteomics and valuable information on which type of proteins are more detectable. Additionally, GO enrichment could provide insight in protein attributes of the proteins that were less detectable and maybe how proteins with attributes more easily be quantified and profiled.

Utilizing technology that differentiates the splice variants would enable more correct comparisons between transcriptomics data and proteomics data where isoforms are distinguished. When comparing gene expression levels and protein abundance, it is important to note that one mRNA molecule may encode several isoforms. In the transcriptomics data set, splice variants encoding different isoforms were not differentiated. This could have skewed the comparison, since in some cases only a fraction of the mRNA will ultimately encode a single isoform. In cases where the detected isoforms were not summed this may have caused issues for the comparisons. Utilizing technology that differentiates the splice variants based on which isoform it will translate into could improve correlations.

Many proteins in the data set were of unknown function, identifying the function of the protein could lead to a greater understanding of the proteome of *Populus tremula* and its closely related species. Since species in the *Populus* genus is becoming the model organisms for woody plants more insight into its proteome will be valuable for future studies.

Bibliography

Abraham, P., Giannone, R. J., Adams, R. M., Kalluri, U., Tuskan, G. A. & Hettich, R. L. (2013). Putting the pieces together: high-performance LC-MS/MS provides network-, pathway-, and protein-level perspectives in Populus. *Mol Cell Proteomics*, 12 (1): 106-19. doi: 10.1074/mcp.M112.022996.

Acinas, S. G., Sarma-Rupavtarm, R., Klepac-Ceraj, V. & Polz, M. F. (2005). PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Applied and environmental microbiology*, 71 (12): 8966-8969. doi: 10.1128/AEM.71.12.8966-8969.2005.

Alberts, B. e. a. (2014). Molecular Biology of the Cell. 6. edition: 1464.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25 (1): 25-9. doi: 10.1038/75556.

Beyer, A., Hollunder, J., Nasheuer, H. P. & Wilhelm, T. (2004). Post-transcriptional expression regulation in the yeast Saccharomyces cerevisiae on a genomic scale. *Mol Cell Proteomics*, 3 (11): 1083-92. doi: 10.1074/mcp.M400099-MCP200.

Bollhoner, B., Prestele, J. & Tuominen, H. (2012). Xylem cell death: emerging understanding of regulation and function. *J Exp Bot*, 63 (3): 1081-94. doi: 10.1093/jxb/err438.

Caudullo, G., Welk, E. & San-Miguel-Ayanz, J. (2017). Chorological maps for the main European woody species. *Data in brief*, 12: 662-666. doi: 10.1016/j.dib.2017.05.007.

Chandramouli, K. & Qian, P.-Y. (2009). Proteomics: challenges, techniques and possibilities to overcome biological sample complexity. *Human genomics and proteomics : HGP*, 2009: 239204. doi: 10.4061/2009/239204.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227 (5258): 561-3.

Csárdi, G., Franks, A., Choi, D. S., Airoldi, E. M. & Drummond, D. A. (2015). Accounting for Experimental Noise Reveals That mRNA Levels, Amplified by Post-Transcriptional Processes, Largely Determine Steady-State Protein Levels in Yeast. *PLOS Genetics*, 11 (5): e1005206. doi: 10.1371/journal.pgen.1005206.

de Sousa Abreu, R., Penalva, L. O., Marcotte, E. M. & Vogel, C. (2009). Global signatures of protein and mRNA expression levels. *Mol Biosyst*, 5 (12): 1512-26. doi: 10.1039/b908315d.

Del Fabbro, C., Scalabrin, S., Morgante, M. & Giorgi, F. M. (2013). An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. *PLOS ONE*, 8 (12): e85024. doi: 10.1371/journal.pone.0085024.

Dupont, C., Armant, D. R. & Brenner, C. A. (2009). Epigenetics: definition, mechanisms and clinical perspective. *Seminars in reproductive medicine*, 27 (5): 351-357. doi: 10.1055/s-0029-1237423.

Fukuda, H. (1996). XYLOGENESIS: INITIATION, PROGRESSION, AND CELL DEATH. *Annu Rev Plant Physiol Plant Mol Biol*, 47: 299-325. doi: 10.1146/annurev.arplant.47.1.299.

Glickman, M. H. & Ciechanover, A. (2002). The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction. *Physiol Rev*, 82 (2): 373-428. doi: 10.1152/physrev.00027.2001.

Graves, P. R. & Haystead, T. A. J. (2002). Molecular biologist's guide to proteomics. *Microbiology and molecular biology reviews : MMBR*, 66 (1): 39-63. doi: 10.1128/MMBR.66.1.39-63.2002.

Han, X., Aslanian, A. & Yates, J. R., 3rd. (2008). Mass spectrometry for proteomics. *Current opinion in chemical biology*, 12 (5): 483-490. doi: 10.1016/j.cbpa.2008.07.024.

Hertzberg, M., Aspeborg, H., Schrader, J., Andersson, A., Erlandsson, R., Blomqvist, K., Bhalerao, R., Uhlén, M., Teeri, T. T., Lundeberg, J., et al. (2001). A transcriptional roadmap to wood formation. *Proceedings of the National Academy of Sciences of the United States of America*, 98 (25): 14732-14737. doi: 10.1073/pnas.261293398.

Hood, S. (2010). *Mitigating Old Tree Mortality in Long-Unburned,Fire-Dependent Forests: A Synthesis*.

Hrdlickova, R., Toloue, M. & Tian, B. (2017). RNA-Seq methods for transcriptome analysis. *Wiley interdisciplinary reviews. RNA*, 8 (1): 10.1002/wrna.1364. doi: 10.1002/wrna.1364.

Institute, U. D. J. G. (2006). *NCBI Genome Populus trichocarpa (black cottonwood)*. Available at: https://www.ncbi.nlm.nih.gov/genome/?term=txid3694[orgn] (accessed: 09/14).

Jansson, S. & Douglas, C. J. (2007). Populus: a model system for plant biology. *Annu Rev Plant Biol*, 58: 435-58. doi: 10.1146/annurev.arplant.58.032806.103956.

Jovanovic, M., Rooney, M. S., Mertins, P., Przybylski, D., Chevrier, N., Satija, R., Rodriguez, E. H., Fields, A. P., Schwartz, S., Raychowdhury, R., et al. (2015). Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens. *Science*, 347 (6226): 1259038. doi: 10.1126/science.1259038.

Jüschke, C., Dohnal, I., Pichler, P., Harzer, H., Swart, R., Ammerer, G., Mechtler, K. & Knoblich, J. A. (2013). Transcriptome and proteome quantification of a tumor model provides novel insights into post-transcriptional gene regulation. *Genome biology*, 14 (11): r133-r133. doi: 10.1186/gb-2013-14-11-r133.

Kainer, D., Lanfear, R., Peñalba, J., Foley, W. & Külheim, C. (2015). *Targeted repeat reduction in whole tree genomes prior to sequencing*.

Kalluri, U. C., Hurst, G. B., Lankford, P. K., Ranjan, P. & Pelletier, D. A. (2009). Shotgun proteome profile of Populus developing xylem. *Proteomics*, 9 (21): 4871-80. doi: 10.1002/pmic.200800854.

Kukurba, K. R. & Montgomery, S. B. (2015). RNA Sequencing and Analysis. *Cold Spring Harbor protocols*, 2015 (11): 951-969. doi: 10.1101/pdb.top084970.

Lee, M. V., Topper, S. E., Hubler, S. L., Hose, J., Wenger, C. D., Coon, J. J. & Gasch, A. P. (2011). A dynamic model of proteome changes reveals new roles for transcript alteration in yeast. *Molecular systems biology*, 7: 514-514. doi: 10.1038/msb.2011.48.

Lemoine, R. (2000). Sucrose transporters in plants: update on function and structure. *Biochim Biophys Acta*, 1465 (1-2): 246-62.

Li, J. J., Bickel, P. J. & Biggin, M. D. (2014). System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ*, 2: e270. doi: 10.7717/peerj.270.

Liu, Y., Beyer, A. & Aebersold, R. (2016). On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell*, 165 (3): 535-50. doi: 10.1016/j.cell.2016.03.014.

Love, M. I., Anders, S., Kim, V. & Huber, W. (2015). RNA-Seq workflow: gene-level exploratory analysis and differential expression. *F1000Research*, 4: 1070-1070. doi: 10.12688/f1000research.7035.1.

Masuda, T., Tomita, M. & Ishihama, Y. (2008). Phase transfer surfactant-aided trypsin digestion for membrane proteome analysis. *J Proteome Res*, 7 (2): 731-40. doi: 10.1021/pr700658q.

McGettigan, P. A. (2013). Transcriptomics in the RNA-seq era. *Curr Opin Chem Biol*, 17 (1): 4-11. doi: 10.1016/j.cbpa.2012.12.008.

Mellerowicz, E. J., Baucher, M., Sundberg, B. & Boerjan, W. (2001). Unravelling cell wall formation in the woody dicot stem. *Plant Mol Biol*, 47 (1-2): 239-74.

Obudulu, O., Bygdell, J., Sundberg, B., Moritz, T., Hvidsten, T. R., Trygg, J. & Wingsle, G. (2016). Quantitative proteomics reveals protein profiles underlying major transitions in aspen wood development. *BMC Genomics*, 17 (1): 119. doi: 10.1186/s12864-016-2458-z.

Oda, Y., Iida, Y., Kondo, Y. & Fukuda, H. (2010). Wood cell-wall structure requires local 2D-microtubule disassembly by a novel plasma membrane-anchored protein. *Curr Biol*, 20 (13): 1197-202. doi: 10.1016/j.cub.2010.05.038.

Pate, J. S. & Atkins, C. A. (1983). Xylem and Phloem transport and the functional economy of carbon and nitrogen of a legume leaf. *Plant physiology*, 71 (4): 835-840.

Richmond, T. (2000). Higher plant cellulose synthases. *Genome biology*, 1 (4): REVIEWS3001-REVIEWS3001. doi: 10.1186/gb-2000-1-4-reviews3001.

Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W. & Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature*, 473 (7347): 337-42. doi: 10.1038/nature10098.

Stastna, M. & Van Eyk, J. E. (2012). Analysis of protein isoforms: can we do it better? *Proteomics*, 12 (19-20): 2937-2948. doi: 10.1002/pmic.201200161.

Steen, H. & Pandey, A. (2002). Proteomics goes quantitative: measuring protein abundance. *Trends Biotechnol*, 20 (9): 361-4.

Sundell, D., Mannapperuma, C., Netotea, S., Delhomme, N., Lin, Y.-C., Sjödin, A., Van de Peer, Y., Jansson, S., Hvidsten, T. R. & Street, N. R. (2015). The Plant Genome Integrative Explorer Resource: PlantGenIE.org. *New Phytologist*, 208 (4): 1149-1156. doi: 10.1111/nph.13557.

Sundell, D., Street, N. R., Kumar, M., Mellerowicz, E. J., Kucukoglu, M., Johnsson, C., Kumar, V., Mannapperuma, C., Delhomme, N., Nilsson, O., et al. (2017). AspWood: High-Spatial-Resolution Transcriptome Profiles Reveal Uncharacterized Modularity of Wood Formation in <em>Populus tremula</em&gt. *The Plant Cell*, 29 (7): 1585.

Thomas, B. e. a. (2003). *Encyclopedia of Applied Plant Science*, vol. 1 edition: Academic Press.

Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., et al. (2006). The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science*, 313 (5793): 1596-604. doi: 10.1126/science.1128691.

Vander Mijnsbrugge, K., Meyermans, H., Van Montagu, M., Bauw, G. & Boerjan, W. (2000). Wood formation in poplar: identification, characterization, and seasonal variation of xylem proteins. *Planta*, 210 (4): 589-98.

Vogel, C. & Marcotte, E. M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet*, 13 (4): 227-32. doi: 10.1038/nrg3185.

Wang, Z., Gerstein, M. & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10 (1): 57-63. doi: 10.1038/nrg2484.

Wethmar, K., Smink, J. J. & Leutz, A. (2010). Upstream open reading frames: molecular switches in (patho)physiology. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 32 (10): 885-893. doi: 10.1002/bies.201000037.

Zheng, Y., Anderson, S., Zhang, Y. & Garavito, R. M. (2011). The structure of sucrose synthase-1 from Arabidopsis thaliana and its functional implications. *J Biol Chem*, 286 (41): 36108-18. doi: 10.1074/jbc.M111.275974.

Zhong, R. & Ye, Z. H. (2007). Regulation of cell wall biosynthesis. *Curr Opin Plant Biol*, 10 (6): 564-72. doi: 10.1016/j.pbi.2007.09.001.

Appendix A: Details of the R-scripts

0-load_info

This script loaded the excel file “Protein_profilling_data.xlsx". The loaded data was adapted to a data frame with the following columns: “GO” showing the GO terms of each protein, “Function” where the basic function description was saved for each protein, “ID” where the whole POTRI code including isomorph suffix was saved. Another column was made, where the suffix of the POTRI code was excluded with substring. As follows: Potri.001G001600.1 to Potri.001G001600. A mock entry called New_Var3083 was removed from the data frame.

0-load_protein

This script loaded the text file “Protein_profilling_data.txt” and sourced “0-load_info”. It scaled the data using a logarithmic transformation. The function log2() was used for this transformation on the data +1. The +1 made sure that zero entries stay zero instead of becoming negative infinity. The data frame from “0-load_info” and the data frame made from “Protein_profilling_data.txt” were pasted together. Lastly, an anomaly entry was coerced to a zero entry. The dimension of the resulting data frame was now 3082x115.

0-load_transcript

This script loaded the text file “TPC2017-LSB-00153R2_Supplemental_Data_Set_2.txt”. The dimensions of the data frame loaded was 28294x107.

1-match_rows

This script sourced “0-load_protein” and “0-load_transcript”. The purpose of this script was to find which genes in the transcriptomics data corresponded to the proteins in the proteomics data. Index based coding and the match() function was used to achieve this. Afterward, the transcript data frame and the protein data frame had the same number of rows. Further, all rows with zero expression in the protein data were cut out and the corresponding rows in the transcription data frame were also cut out. Lastly the columns in the protein data frame containing GO terms, Function and full ID were removed from the frame and instead saved in a separate data frame to be used as an info resource.

2-match_columns

This script sourced “1-match_rows”. The purpose of this script was to equalize the number of columns in both data sets and for all trees since the number of samples for the transcript data and the protein data differed. The protein data had 111 entries divided among four trees and the transcript data had 106 divided by four trees. The script removed sample entries so that the data frames had a uniform 25 entries per tree for a total of 100 samples for each data set. The samples included are described in the Table 1 and Table 2. After the removals the dimensions of both data sets were 2029 x 101. The first column indicated the protein names.

Table 1: protein cutting scheme.

Tree	Number of samples in data	Samples Included
1	27	1 to 25
2	28	29 to 53
3	28	56 to 80
4	28	86 to 110
Total	111	100

Table 2: transcript cutting scheme.

Tree	Number of samples in data	Samples Included
1	25	1 to 25
2	26	26 to 51
3	28	53 to 77
4	27	81 to 105
Total	106	100

3-moving_average

This script sourced “2-match_columns” and a function made for calculating the moving average. This script calculated the moving average for the proteomics data. The function takes in a vector and calculates the average of each entry with both adjacent entries (or one at the start and end of the vector). The protein data frame was split into the four trees before looping through the function and then pasted together as a separate “moving average” data frame.

4-best_method

This script sourced “3-moving_average” and a function made for calculating the “best method”. This script calculated which of the protein series (from the four trees) correlated best with the average transcript. The “best method” took two data frames. Both were split into four sections by columns. The average of the four sections was calculated for the second data frame. The function looped through every row and calculated the correlation between the four sections of the first data frame with the average in the other in turn. The section that correlated the highest was saved in a different data frame which was returned. The function ran for both the raw proteomics data and for the moving average treated protein data. Additionally, the average transcript data frame was calculated.

5-clustering

This script sourced “4-best_method”. Three functions were built into this script that were used to cluster the two data sets and compare the content of the clusters in the proteomics data set and the transcriptomics data set. The first function clustered the data sets based on correlation distances using the “ward.d” method and saved the result as a data frame. Number of clusters could be specified. The second function took two series of protein names calculated the total number of entries in the two and divided them by the number of unique protein names in the two sets. The third function looped through all the clusters in one data frame and used the second function to calculate the scores for each possible cluster pair in the two data frames. These scores were calculated for the moving average calculated protein data set and the transcriptomics data set, and for the “best method” moving average calculated protein data and the average by trees transcriptomics data set.

5-isoforms

This script sourced “4-best_method”. The proteins with isoforms were found in the moving average calculated data frame and the different isoforms of a protein were summed together based on samples. These sums were saved in a data frame.

5-on_off_switch

This script sourced “3-moving_average”. This script creates new data frames for the proteomics data, moving average calculated proteomics data and the transcriptomics data where the non-zero entries are saved as ones and zero entries remain zeros. The zeros and ones are compared between the resulting data sets.

5-molecules_per_sample

This script sourced “1-best_method”. The protein abundance was summed by sample for the proteomics data and the transcript was summed by sample for the transcriptomics data. The resulting data was saved in a data frame. Secondly the scores for the combined isoform proteins correlated with their corresponding transcript was calculated

6-correlations

This script sourced “5-isoforms”. The script was divided into four sections. The first section correlated the proteomics data sets with the transcriptomics data set by protein name (row correlation). This was done for the raw, the moving average calculated, the “best method“ and the combined proteomics data sets. Correlation by protein names was also calculated for the individual trees and the average of the trees. This was repeated using the moving average calculated series. A data frame containing the correlations was saved. Second, the combined isoform proteins were correlated with their corresponding transcript series. Third, the data sets were transposed, and correlations were calculated by sample, meaning the vector of all protein abundance entries for sample one was correlated with the corresponding transcript vector and so on. Lastly, full correlation was done, meaning the whole data sets were saved as single vectors and then correlated.

7-results

This script sourced “6-correlations”, 5-on_off_switch, 5-clustering, 5-cor_by_sample and the correlation function. This script produced the all plots (using mainly the ggplot() function), the statistical tests and related figures and the heat maps and dendrograms (using the heatmap.2() function).

Appendix B: Results from Gene Ontology Enrichment Analysis

GO Terms of the Biological Process Category That are Overrepresented in the Highly Correlating Protein Subset

GO ID	P-value (FDR)	Statistics	Description
GO:0006195	4.825e-10	20/829 | 36/14903	purine nucleotide catabolic process
GO:0009207	4.825e-10	20/829 | 36/14903	purine ribonucleoside triphosphate catabolic process
GO:1901658	4.825e-10	20/829 | 36/14903	glycosyl compound catabolic process
GO:0072523	4.825e-10	20/829 | 36/14903	purine-containing compound catabolic process
GO:0009146	4.825e-10	20/829 | 36/14903	purine nucleoside triphosphate catabolic process
GO:0046130	4.825e-10	20/829 | 36/14903	purine ribonucleoside catabolic process
GO:0006152	4.825e-10	20/829 | 36/14903	purine nucleoside catabolic process
GO:0042454	4.825e-10	20/829 | 36/14903	ribonucleoside catabolic process
GO:0009143	4.825e-10	20/829 | 36/14903	nucleoside triphosphate catabolic process
GO:0009154	4.825e-10	20/829 | 36/14903	purine ribonucleotide catabolic process
GO:0009164	4.825e-10	20/829 | 36/14903	nucleoside catabolic process
GO:0009166	4.825e-10	20/829 | 36/14903	nucleotide catabolic process
GO:0009203	4.825e-10	20/829 | 36/14903	ribonucleoside triphosphate catabolic process
GO:1901292	4.825e-10	20/829 | 36/14903	nucleoside phosphate catabolic process
GO:0009261	4.825e-10	20/829 | 36/14903	ribonucleotide catabolic process
GO:0006006	5.480e-10	32/829 | 100/14903	glucose metabolic process
GO:0046700	5.532e-10	20/829 | 53/14903	heterocycle catabolic process
GO:0044724	5.532e-10	30/829 | 93/14903	single-organism carbohydrate catabolic process
GO:0046434	5.952e-10	20/829 | 42/14903	organophosphate catabolic process
GO:1901361	6.476e-10	20/829 | 54/14903	organic cyclic compound catabolic process
GO:0051258	6.571e-10	19/829 | 37/14903	protein polymerization
GO:0006461	6.624e-10	19/829 | 48/14903	protein complex assembly
GO:1901565	6.671e-10	23/829 | 80/14903	organonitrogen compound catabolic process
GO:1901605	7.483e-10	32/829 | 151/14903	alpha-amino acid metabolic process
GO:0007264	7.925e-10	32/829 | 143/14903	small GTPase mediated signal transduction
GO:0009144	8.071e-10	30/829 | 127/14903	purine nucleoside triphosphate metabolic process
GO:0009199	8.071e-10	30/829 | 127/14903	ribonucleoside triphosphate metabolic process
GO:0009205	8.071e-10	30/829 | 127/14903	purine ribonucleoside triphosphate metabolic process
GO:0072521	8.297e-10	36/829 | 170/14903	purine-containing compound metabolic process
GO:0005996	8.510e-10	34/829 | 144/14903	monosaccharide metabolic process
GO:0009653	8.720e-10	16/829 | 25/14903	anatomical structure morphogenesis
GO:0000902	8.720e-10	16/829 | 25/14903	cell morphogenesis
GO:0032989	8.720e-10	16/829 | 25/14903	cellular component morphogenesis
GO:0048869	8.720e-10	16/829 | 25/14903	cellular developmental process
GO:1901657	8.950e-10	44/829 | 198/14903	glycosyl compound metabolic process
GO:0009116	8.950e-10	44/829 | 198/14903	nucleoside metabolic process
GO:0006096	9.115e-10	25/829 | 62/14903	glycolysis
GO:0034655	9.154e-10	20/829 | 44/14903	nucleobase-containing compound catabolic process
GO:0006007	9.477e-10	30/829 | 90/14903	glucose catabolic process
GO:0019320	9.477e-10	30/829 | 90/14903	hexose catabolic process
GO:0046365	9.477e-10	30/829 | 90/14903	monosaccharide catabolic process
GO:0044723	1.003e-09	68/829 | 439/14903	single-organism carbohydrate metabolic process

GO:0043623	1.004e-09	19/829 | 45/14903	cellular protein complex assembly
GO:0046128	1.017e-09	36/829 | 156/14903	purine ribonucleoside metabolic process
GO:0042278	1.017e-09	36/829 | 156/14903	purine nucleoside metabolic process
GO:0009056	1.021e-09	67/829 | 385/14903	catabolic process
GO:0006184	1.022e-09	18/829 | 34/14903	GTP catabolic process
GO:1901069	1.022e-09	18/829 | 34/14903	guanosine-containing compound catabolic process
GO:0055086	1.027e-09	48/829 | 269/14903	nucleobase-containing small molecule metabolic process
GO:1901136	1.034e-09	21/829 | 63/14903	carbohydrate derivative catabolic process
GO:0008652	1.037e-09	40/829 | 164/14903	cellular amino acid biosynthetic process
GO:0009119	1.037e-09	36/829 | 164/14903	ribonucleoside metabolic process
GO:0044270	1.042e-09	20/829 | 51/14903	cellular nitrogen compound catabolic process
GO:1901575	1.049e-09	65/829 | 351/14903	organic substance catabolic process
GO:0019318	1.066e-09	33/829 | 131/14903	hexose metabolic process
GO:0044767	1.070e-09	16/829 | 29/14903	single-organism developmental process
GO:0016052	1.085e-09	30/829 | 107/14903	carbohydrate catabolic process
GO:0006412	1.118e-09	112/829 | 567/14903	translation
GO:0006520	1.163e-09	69/829 | 355/14903	cellular amino acid metabolic process
GO:0046039	1.198e-09	21/829 | 40/14903	GTP metabolic process
GO:0043436	1.209e-09	83/829 | 522/14903	oxoacid metabolic process
GO:0019752	1.209e-09	83/829 | 522/14903	carboxylic acid metabolic process
GO:0009141	1.213e-09	30/829 | 132/14903	nucleoside triphosphate metabolic process
GO:0006082	1.254e-09	83/829 | 523/14903	organic acid metabolic process
GO:0019439	1.283e-09	20/829 | 52/14903	aromatic compound catabolic process
GO:1901068	1.325e-09	21/829 | 46/14903	guanosine-containing compound metabolic process
GO:0034645	1.376e-09	152/829 | 948/14903	cellular macromolecule biosynthetic process
GO:0044281	1.566e-09	134/829 | 886/14903	small molecule metabolic process
GO:1901564	1.802e-09	115/829 | 771/14903	organonitrogen compound metabolic process
GO:0009059	1.859e-09	153/829 | 1019/14903	macromolecule biosynthetic process
GO:0044249	1.861e-09	205/829 | 1631/14903	cellular biosynthetic process
GO:0005975	1.897e-09	105/829 | 998/14903	carbohydrate metabolic process
GO:0009058	1.907e-09	234/829 | 1980/14903	biosynthetic process
GO:1901576	1.911e-09	216/829 | 1757/14903	organic substance biosynthetic process
GO:0044710	2.196e-09	272/829 | 3299/14903	single-organism metabolic process
GO:0071704	2.243e-09	481/829 | 6836/14903	organic substance metabolic process
GO:0071822	2.260e-09	19/829 | 61/14903	protein complex subunit organization
GO:0008152	2.285e-09	659/829 | 9859/14903	metabolic process
GO:0044237	2.316e-09	407/829 | 5635/14903	cellular metabolic process
GO:0044238	2.861e-09	457/829 | 6540/14903	primary metabolic process
GO:0016053	2.922e-09	43/829 | 273/14903	organic acid biosynthetic process
GO:0046394	2.922e-09	43/829 | 273/14903	carboxylic acid biosynthetic process
GO:0006886	2.978e-09	35/829 | 194/14903	intracellular protein transport
GO:0035556	6.159e-09	32/829 | 171/14903	intracellular signal transduction
GO:0009150	7.442e-09	30/829 | 154/14903	purine ribonucleotide metabolic process
GO:0006163	1.018e-08	30/829 | 156/14903	purine nucleotide metabolic process
GO:0019693	1.371e-08	30/829 | 158/14903	ribose phosphate metabolic process
GO:0009259	1.371e-08	30/829 | 158/14903	ribonucleotide metabolic process
GO:1901607	2.644e-08	23/829 | 101/14903	alpha-amino acid biosynthetic process
GO:0032502	4.155e-08	16/829 | 51/14903	developmental process
GO:0034622	4.894e-08	26/829 | 130/14903	cellular macromolecular complex assembly
GO:1901566	4.924e-08	52/829 | 398/14903	organonitrogen compound biosynthetic process
GO:0016043	6.124e-08	52/829 | 402/14903	cellular component organization
GO:0065003	7.920e-08	26/829 | 133/14903	macromolecular complex assembly
GO:0009117	9.287e-08	34/829 | 211/14903	nucleotide metabolic process
GO:0046907	1.092e-07	36/829 | 233/14903	intracellular transport

GO:0006807	1.314e-07	139/829 | 1580/14903	nitrogen compound metabolic process
GO:0006564	1.358e-07	6/829 | 6/14903	L-serine biosynthetic process
GO:1901135	1.636e-07	45/829 | 334/14903	carbohydrate derivative metabolic process
GO:0071840	1.738e-07	53/829 | 427/14903	cellular component organization or biogenesis
GO:0044283	1.974e-07	43/829 | 310/14903	small molecule biosynthetic process
GO:0044711	2.243e-07	43/829 | 313/14903	single-organism biosynthetic process
GO:0006753	2.243e-07	34/829 | 219/14903	nucleoside phosphate metabolic process
GO:0006555	2.332e-07	7/829 | 9/14903	methionine metabolic process
GO:0044248	2.836e-07	37/829 | 253/14903	cellular catabolic process
GO:0006414	2.852e-07	12/829 | 32/14903	translational elongation
GO:0044262	3.485e-07	38/829 | 266/14903	cellular carbohydrate metabolic process
GO:0006457	4.865e-07	30/829 | 185/14903	protein folding
GO:0043933	5.023e-07	26/829 | 146/14903	macromolecular complex subunit organization
GO:0051649	1.088e-06	37/829 | 267/14903	establishment of localization in cell
GO:0022607	1.645e-06	26/829 | 155/14903	cellular component assembly
GO:0006563	1.899e-06	9/829 | 20/14903	L-serine metabolic process
GO:0051170	2.945e-06	6/829 | 8/14903	nuclear import
GO:0006606	2.945e-06	6/829 | 8/14903	protein import into nucleus
GO:0055114	3.234e-06	162/829 | 2019/14903	oxidation-reduction process
GO:0006091	4.351e-06	25/829 | 153/14903	generation of precursor metabolites and energy
GO:0033692	4.826e-06	25/829 | 154/14903	cellular polysaccharide biosynthetic process
GO:0000271	4.826e-06	25/829 | 154/14903	polysaccharide biosynthetic process
GO:0009084	4.876e-06	9/829 | 22/14903	glutamine family amino acid biosynthetic process
GO:0009069	6.820e-06	11/829 | 35/14903	serine family amino acid metabolic process
GO:0015988	7.038e-06	12/829 | 42/14903	energy coupled proton transmembrane transport, against electrochemical gradient
GO:0015991	7.038e-06	12/829 | 42/14903	ATP hydrolysis coupled proton transport
GO:0045184	7.116e-06	36/829 | 277/14903	establishment of protein localization
GO:0015031	7.116e-06	36/829 | 277/14903	protein transport
GO:0007017	1.626e-05	23/829 | 144/14903	microtubule-based process
GO:0030243	2.648e-05	13/829 | 55/14903	cellulose metabolic process
GO:0030244	2.648e-05	13/829 | 55/14903	cellulose biosynthetic process
GO:0016192	2.829e-05	25/829 | 170/14903	vesicle-mediated transport
GO:0009987	2.899e-05	519/829 | 8214/14903	cellular process
GO:0015992	3.304e-05	17/829 | 91/14903	proton transport
GO:0006818	3.304e-05	17/829 | 91/14903	hydrogen transport
GO:0009073	3.311e-05	8/829 | 21/14903	aromatic amino acid family biosynthetic process
GO:0044264	3.429e-05	27/829 | 194/14903	cellular polysaccharide metabolic process
GO:0009086	3.491e-05	5/829 | 7/14903	methionine biosynthetic process
GO:0051169	3.620e-05	6/829 | 11/14903	nuclear transport
GO:0006913	3.620e-05	6/829 | 11/14903	nucleocytoplasmic transport
GO:0000096	4.023e-05	7/829 | 16/14903	sulfur amino acid metabolic process
GO:0051273	4.234e-05	14/829 | 66/14903	beta-glucan metabolic process
GO:0051274	4.234e-05	14/829 | 66/14903	beta-glucan biosynthetic process
GO:0045226	4.290e-05	11/829 | 42/14903	extracellular polysaccharide biosynthetic process
GO:0046379	4.290e-05	11/829 | 42/14903	extracellular polysaccharide metabolic process
GO:0007010	6.801e-05	8/829 | 23/14903	cytoskeleton organization
GO:0043648	6.859e-05	11/829 | 44/14903	dicarboxylic acid metabolic process
GO:0006542	8.324e-05	5/829 | 8/14903	glutamine biosynthetic process
GO:0009250	8.345e-05	14/829 | 70/14903	glucan biosynthetic process
GO:0034637	9.248e-05	25/829 | 183/14903	cellular carbohydrate biosynthetic process
GO:0016051	9.945e-05	27/829 | 203/14903	carbohydrate biosynthetic process
GO:0005976	1.319e-04	27/829 | 208/14903	polysaccharide metabolic process
GO:0009064	1.952e-04	11/829 | 49/14903	glutamine family amino acid metabolic process
GO:0009070	2.967e-04	6/829 | 15/14903	serine family amino acid biosynthetic process
GO:0009066	2.969e-04	7/829 | 21/14903	aspartate family amino acid metabolic process

GO:0006099	4.073e-04	4/829 | 6/14903	tricarboxylic acid cycle
GO:0009072	5.390e-04	8/829 | 30/14903	aromatic amino acid family metabolic process
GO:0006418	6.091e-04	13/829 | 74/14903	tRNA aminoacylation for protein translation
GO:1901360	6.103e-04	111/829 | 1421/14903	organic cyclic compound metabolic process
GO:0043038	6.925e-04	13/829 | 75/14903	amino acid activation
GO:0043039	6.925e-04	13/829 | 75/14903	tRNA aminoacylation
GO:0046500	8.749e-04	4/829 | 7/14903	S-adenosylmethionine metabolic process
GO:0006556	8.749e-04	4/829 | 7/14903	S-adenosylmethionine biosynthetic process
GO:0000097	8.988e-04	5/829 | 12/14903	sulfur amino acid biosynthetic process
GO:0006073	1.093e-03	16/829 | 110/14903	cellular glucan metabolic process
GO:0044042	1.093e-03	16/829 | 110/14903	glucan metabolic process
GO:0006694	1.102e-03	18/829 | 132/14903	steroid biosynthetic process
GO:0019637	1.172e-03	35/829 | 337/14903	organophosphate metabolic process
GO:0006790	1.199e-03	11/829 | 60/14903	sulfur compound metabolic process
GO:0008202	1.293e-03	18/829 | 134/14903	steroid metabolic process
GO:0018193	1.644e-03	6/829 | 20/14903	peptidyl-amino acid modification
GO:0005985	1.966e-03	5/829 | 14/14903	sucrose metabolic process
GO:0044272	2.371e-03	9/829 | 46/14903	sulfur compound biosynthetic process
GO:0006108	2.834e-03	6/829 | 22/14903	malate metabolic process
GO:0051603	3.270e-03	18/829 | 142/14903	proteolysis involved in cellular protein catabolic process
GO:0017038	3.567e-03	6/829 | 23/14903	protein import
GO:0018196	6.203e-03	4/829 | 11/14903	peptidyl-asparagine modification
GO:0018279	6.203e-03	4/829 | 11/14903	protein N-linked glycosylation via asparagine
GO:0009128	7.297e-03	2/829 | 2/14903	purine nucleoside monophosphate catabolic process
GO:0033559	7.297e-03	2/829 | 2/14903	unsaturated fatty acid metabolic process
GO:1901568	7.297e-03	2/829 | 2/14903	fatty acid derivative metabolic process
GO:0043094	7.297e-03	2/829 | 2/14903	cellular metabolic compound salvage
GO:0043101	7.297e-03	2/829 | 2/14903	purine-containing compound salvage
GO:0043174	7.297e-03	2/829 | 2/14903	nucleoside salvage
GO:0009158	7.297e-03	2/829 | 2/14903	ribonucleoside monophosphate catabolic process
GO:0019370	7.297e-03	2/829 | 2/14903	leukotriene biosynthetic process
GO:0006166	7.297e-03	2/829 | 2/14903	purine ribonucleoside salvage
GO:0006200	7.297e-03	2/829 | 2/14903	ATP catabolic process
GO:0006206	7.297e-03	2/829 | 2/14903	pyrimidine nucleobase metabolic process
GO:0006207	7.297e-03	2/829 | 2/14903	'de novo' pyrimidine nucleobase biosynthetic process
GO:0046456	7.297e-03	2/829 | 2/14903	icosanoid biosynthetic process
GO:1901570	7.297e-03	2/829 | 2/14903	fatty acid derivative biosynthetic process
GO:0019856	7.297e-03	2/829 | 2/14903	pyrimidine nucleobase biosynthetic process
GO:0006636	7.297e-03	2/829 | 2/14903	unsaturated fatty acid biosynthetic process
GO:0006690	7.297e-03	2/829 | 2/14903	icosanoid metabolic process
GO:0006691	7.297e-03	2/829 | 2/14903	leukotriene metabolic process
GO:0009125	7.297e-03	2/829 | 2/14903	nucleoside monophosphate catabolic process
GO:0006730	7.297e-03	2/829 | 2/14903	one-carbon metabolic process
GO:0009169	7.297e-03	2/829 | 2/14903	purine ribonucleoside monophosphate catabolic process
GO:0009067	7.481e-03	5/829 | 19/14903	aspartate family amino acid biosynthetic process
GO:0009148	7.815e-03	3/829 | 6/14903	pyrimidine nucleoside triphosphate biosynthetic process
GO:0046036	7.815e-03	3/829 | 6/14903	CTP metabolic process
GO:0046051	7.815e-03	3/829 | 6/14903	UTP metabolic process
GO:0009208	7.815e-03	3/829 | 6/14903	pyrimidine ribonucleoside triphosphate metabolic process
GO:0006165	7.815e-03	3/829 | 6/14903	nucleoside diphosphate phosphorylation

GO:0006183	7.815e-03	3/829 | 6/14903	GTP biosynthetic process
GO:0006228	7.815e-03	3/829 | 6/14903	UTP biosynthetic process
GO:0006241	7.815e-03	3/829 | 6/14903	CTP biosynthetic process
GO:0046939	7.815e-03	3/829 | 6/14903	nucleotide phosphorylation
GO:0009209	7.815e-03	3/829 | 6/14903	pyrimidine ribonucleoside triphosphate biosynthetic process
GO:0006725	8.855e-03	94/829 | 1264/14903	cellular aromatic compound metabolic process
GO:0006139	1.065e-02	85/829 | 1137/14903	nucleobase-containing compound metabolic process
GO:0006102	1.158e-02	3/829 | 7/14903	isocitrate metabolic process
GO:1901070	1.158e-02	3/829 | 7/14903	guanosine-containing compound biosynthetic process
GO:0006544	1.422e-02	5/829 | 22/14903	glycine metabolic process
GO:0006399	1.475e-02	17/829 | 153/14903	tRNA metabolic process
GO:0046129	1.637e-02	13/829 | 103/14903	purine ribonucleoside biosynthetic process
GO:0042451	1.637e-02	13/829 | 103/14903	purine nucleoside biosynthetic process
GO:0009082	1.736e-02	3/829 | 8/14903	branched-chain amino acid biosynthetic process
GO:0006536	1.761e-02	4/829 | 15/14903	glutamate metabolic process
GO:0042398	1.977e-02	7/829 | 43/14903	cellular modified amino acid biosynthetic process
GO:1901659	2.251e-02	13/829 | 111/14903	glycosyl compound biosynthetic process
GO:0042455	2.251e-02	13/829 | 111/14903	ribonucleoside biosynthetic process
GO:0009163	2.251e-02	13/829 | 111/14903	nucleoside biosynthetic process
GO:0006541	2.395e-02	5/829 | 25/14903	glutamine metabolic process
GO:0015977	2.736e-02	4/829 | 17/14903	carbon fixation
GO:0006575	2.781e-02	7/829 | 46/14903	cellular modified amino acid metabolic process
GO:0072522	2.899e-02	13/829 | 116/14903	purine-containing compound biosynthetic process
GO:0071702	2.901e-02	36/829 | 429/14903	organic substance transport
GO:0046131	3.195e-02	3/829 | 10/14903	pyrimidine ribonucleoside metabolic process
GO:0046132	3.195e-02	3/829 | 10/14903	pyrimidine ribonucleoside biosynthetic process
GO:0046134	3.195e-02	3/829 | 10/14903	pyrimidine nucleoside biosynthetic process
GO:0006213	3.195e-02	3/829 | 10/14903	pyrimidine nucleoside metabolic process
GO:0009218	3.195e-02	3/829 | 10/14903	pyrimidine ribonucleotide metabolic process
GO:0009220	3.195e-02	3/829 | 10/14903	pyrimidine ribonucleotide biosynthetic process
GO:0006487	3.231e-02	4/829 | 18/14903	protein N-linked glycosylation
GO:0019725	3.465e-02	17/829 | 169/14903	cellular homeostasis
GO:0051130	3.471e-02	2/829 | 4/14903	positive regulation of cellular component organization
GO:0043243	3.471e-02	2/829 | 4/14903	positive regulation of protein complex disassembly
GO:0043244	3.471e-02	2/829 | 4/14903	regulation of protein complex disassembly
GO:0045901	3.471e-02	2/829 | 4/14903	positive regulation of translational elongation
GO:0045905	3.471e-02	2/829 | 4/14903	positive regulation of translational termination
GO:0006449	3.471e-02	2/829 | 4/14903	regulation of translational termination
GO:0006452	3.471e-02	2/829 | 4/14903	translational frameshifting
GO:0042592	3.480e-02	17/829 | 170/14903	homeostatic process
GO:0016482	3.838e-02	8/829 | 61/14903	cytoplasmic transport
GO:0009147	4.023e-02	3/829 | 11/14903	pyrimidine nucleoside triphosphate metabolic process
GO:0009132	4.023e-02	3/829 | 11/14903	nucleoside diphosphate metabolic process
GO:0015672	4.244e-02	18/829 | 184/14903	monovalent inorganic cation transport
GO:0045454	4.961e-02	16/829 | 164/14903	cell redox homeostasis

GO Terms of the Molecular Function Category that are Overrepresented in the Highly Correlating Protein Subset

GO ID	P-value (FDR)	Statistics	Description
GO:0004455	**2.663e-09**	11/967 | 13/19622	ketol-acid reductoisomerase activity
GO:0016614	3.341e-09	55/967 | 392/19622	oxidoreductase activity, acting on CH-OH group of donors
GO:0070003	3.361e-09	16/967 | 34/19622	threonine-type peptidase activity
GO:0004298	3.361e-09	16/967 | 34/19622	threonine-type endopeptidase activity
GO:0003735	3.541e-09	103/967 | 487/19622	structural constituent of ribosome
GO:0005525	4.032e-09	72/967 | 321/19622	GTP binding
GO:0032561	4.032e-09	72/967 | 321/19622	guanyl ribonucleotide binding
GO:0005198	4.097e-09	117/967 | 548/19622	structural molecule activity
GO:0003924	4.144e-09	56/967 | 185/19622	GTPase activity
GO:0051082	4.340e-09	22/967 | 88/19622	unfolded protein binding
GO:0016616	4.369e-09	55/967 | 372/19622	oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor
GO:0017111	4.387e-09	103/967 | 902/19622	nucleoside-triphosphatase activity
GO:0016818	4.575e-09	106/967 | 948/19622	hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides
GO:0016817	4.618e-09	108/967 | 972/19622	hydrolase activity, acting on acid anhydrides
GO:0016462	4.741e-09	106/967 | 923/19622	pyrophosphatase activity
GO:0019001	4.880e-09	72/967 | 329/19622	guanyl nucleotide binding
GO:0019904	5.345e-09	10/967 | 14/19622	protein domain specific binding
GO:0016491	5.578e-09	181/967 | 2275/19622	oxidoreductase activity
GO:0004617	1.927e-07	6/967 | 6/19622	phosphoglycerate dehydrogenase activity
GO:0003872	1.984e-07	8/967 | 12/19622	6-phosphofructokinase activity
GO:0005507	2.543e-07	29/967 | 181/19622	copper ion binding
GO:0019200	9.345e-07	8/967 | 14/19622	carbohydrate kinase activity
GO:0008443	9.345e-07	8/967 | 14/19622	phosphofructokinase activity
GO:0046961	2.276e-06	11/967 | 32/19622	proton-transporting ATPase activity, rotational mechanism
GO:0048037	2.455e-06	68/967 | 715/19622	cofactor binding
GO:0042085	3.081e-06	5/967 | 5/19622	5-methyltetrahydropteroyltri-L-glutamate-dependent methyltransferase activity
GO:0003871	3.081e-06	5/967 | 5/19622	5-methyltetrahydropteroyltriglutamate-homocysteine S-methyltransferase activity
GO:0019829	7.512e-06	11/967 | 36/19622	cation-transporting ATPase activity
GO:0044769	7.512e-06	11/967 | 36/19622	ATPase activity, coupled to transmembrane movement of ions, rotational mechanism
GO:0008466	1.007e-05	6/967 | 9/19622	glycogenin glucosyltransferase activity
GO:0004807	1.525e-05	5/967 | 6/19622	triose-phosphate isomerase activity
GO:0016841	1.525e-05	5/967 | 6/19622	ammonia-lyase activity
GO:0003746	2.255e-05	8/967 | 20/19622	translation elongation factor activity
GO:0050662	3.127e-05	52/967 | 536/19622	coenzyme binding
GO:0035251	3.179e-05	14/967 | 67/19622	UDP-glucosyltransferase activity
GO:0046527	3.179e-05	14/967 | 67/19622	glucosyltransferase activity
GO:0008831	3.329e-05	11/967 | 42/19622	dTDP-4-dehydrorhamnose reductase activity
GO:0003779	5.029e-05	10/967 | 36/19622	actin binding

GO:0003824	5.146e-05	591/967 | 10616/19622	catalytic activity
GO:0008092	1.322e-04	10/967 | 40/19622	cytoskeletal protein binding
GO:0016787	1.795e-04	205/967 | 3172/19622	hydrolase activity
GO:0016209	1.913e-04	24/967 | 192/19622	antioxidant activity
GO:0008964	1.923e-04	4/967 | 5/19622	phosphoenolpyruvate carboxylase activity
GO:0004618	1.923e-04	4/967 | 5/19622	phosphoglycerate kinase activity
GO:0016774	1.923e-04	4/967 | 5/19622	phosphotransferase activity, carboxyl group as acceptor
GO:0046933	2.917e-04	8/967 | 28/19622	proton-transporting ATP synthase activity, rotational mechanism
GO:0004812	5.069e-04	13/967 | 76/19622	aminoacyl-tRNA ligase activity
GO:0016875	5.645e-04	13/967 | 77/19622	ligase activity, forming carbon-oxygen bonds
GO:0016876	5.645e-04	13/967 | 77/19622	ligase activity, forming aminoacyl-tRNA and related compounds
GO:0016829	6.368e-04	29/967 | 274/19622	lyase activity
GO:0003723	7.086e-04	42/967 | 461/19622	RNA binding
GO:0051287	7.283e-04	11/967 | 59/19622	NAD binding
GO:0003854	7.650e-04	17/967 | 125/19622	3-beta-hydroxy-delta5-steroid dehydrogenase activity
GO:0033764	7.650e-04	17/967 | 125/19622	steroid dehydrogenase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor
GO:0016229	7.650e-04	17/967 | 125/19622	steroid dehydrogenase activity
GO:0070011	8.152e-04	51/967 | 599/19622	peptidase activity, acting on L-amino acid peptides
GO:0004743	8.181e-04	6/967 | 18/19622	pyruvate kinase activity
GO:0031420	8.181e-04	6/967 | 18/19622	alkali metal ion binding
GO:0030955	8.181e-04	6/967 | 18/19622	potassium ion binding
GO:0016861	8.355e-04	7/967 | 25/19622	intramolecular oxidoreductase activity, interconverting aldoses and ketoses
GO:0019205	8.457e-04	8/967 | 33/19622	nucleobase-containing compound kinase activity
GO:0004478	9.124e-04	4/967 | 7/19622	methionine adenosyltransferase activity
GO:0004611	9.124e-04	4/967 | 7/19622	phosphoenolpyruvate carboxykinase activity
GO:0004175	1.003e-03	37/967 | 392/19622	endopeptidase activity
GO:0016211	1.253e-03	5/967 | 13/19622	ammonia ligase activity
GO:0016880	1.253e-03	5/967 | 13/19622	acid-ammonia (or amide) ligase activity
GO:0004356	1.253e-03	5/967 | 13/19622	glutamate-ammonia ligase activity
GO:0008135	1.685e-03	13/967 | 88/19622	translation factor activity, nucleic acid binding
GO:0008172	1.826e-03	5/967 | 14/19622	S-methyltransferase activity
GO:0008233	2.207e-03	51/967 | 632/19622	peptidase activity
GO:0016860	2.208e-03	7/967 | 30/19622	intramolecular oxidoreductase activity
GO:0016615	2.244e-03	6/967 | 22/19622	malate dehydrogenase activity
GO:0009378	3.397e-03	10/967 | 62/19622	four-way junction helicase activity
GO:0008134	3.877e-03	4/967 | 10/19622	transcription factor binding
GO:0003849	4.341e-03	3/967 | 5/19622	3-deoxy-7-phosphoheptulonate synthase activity
GO:0016776	5.628e-03	4/967 | 11/19622	phosphotransferase activity, phosphate group as acceptor
GO:0008565	6.279e-03	11/967 | 79/19622	protein transporter activity
GO:0030060	7.960e-03	4/967 | 12/19622	L-malate dehydrogenase activity
GO:0004550	7.969e-03	3/967 | 6/19622	nucleoside diphosphate kinase activity

GO:0042625	8.135e-03	11/967 | 83/19622	ATPase activity, coupled to transmembrane movement of ions
GO:0003678	8.135e-03	11/967 | 83/19622	DNA helicase activity
GO:0016801	8.142e-03	2/967 | 2/19622	hydrolase activity, acting on ether bonds
GO:0004489	8.142e-03	2/967 | 2/19622	methylenetetrahydrofolate reductase (NADPH) activity
GO:0004001	8.142e-03	2/967 | 2/19622	adenosine kinase activity
GO:0004013	8.142e-03	2/967 | 2/19622	adenosylhomocysteinase activity
GO:0004590	8.142e-03	2/967 | 2/19622	orotidine-5'-phosphate decarboxylase activity
GO:0004375	8.142e-03	2/967 | 2/19622	glycine dehydrogenase (decarboxylating) activity
GO:0004107	8.142e-03	2/967 | 2/19622	chorismate synthase activity
GO:0004148	8.142e-03	2/967 | 2/19622	dihydrolipoyl dehydrogenase activity
GO:0016642	8.142e-03	2/967 | 2/19622	oxidoreductase activity, acting on the CH-NH2 group of donors, disulfide as acceptor
GO:0008553	8.142e-03	2/967 | 2/19622	hydrogen-exporting ATPase activity, phosphorylative mechanism
GO:0016802	8.142e-03	2/967 | 2/19622	trialkylsulfonium hydrolase activity
GO:0016830	9.437e-03	12/967 | 97/19622	carbon-carbon lyase activity
GO:0016840	9.813e-03	5/967 | 21/19622	carbon-nitrogen lyase activity
GO:0004448	1.149e-02	3/967 | 7/19622	isocitrate dehydrogenase activity
GO:0004450	1.149e-02	3/967 | 7/19622	isocitrate dehydrogenase (NADP+) activity
GO:0019843	1.346e-02	6/967 | 32/19622	rRNA binding
GO:0016853	1.355e-02	18/967 | 178/19622	isomerase activity
GO:0004579	1.575e-02	4/967 | 15/19622	dolichyl-diphosphooligosaccharide-protein glycotransferase activity
GO:0016836	1.621e-02	7/967 | 44/19622	hydro-lyase activity
GO:0005839	8.286e-10	16/360 | 34/6017	proteasome core complex
GO:0005874	9.274e-10	18/360 | 33/6017	microtubule
GO:0044464	1.620e-09	305/360 | 3667/6017	cell part
GO:0043229	1.683e-09	149/360 | 1580/6017	intracellular organelle
GO:0043226	1.683e-09	149/360 | 1580/6017	organelle
GO:0044424	1.782e-09	262/360 | 2754/6017	intracellular part
GO:0043234	1.819e-09	101/360 | 777/6017	protein complex
GO:0032991	1.852e-09	208/360 | 1363/6017	macromolecular complex
GO:0044444	2.067e-09	155/360 | 1027/6017	cytoplasmic part
GO:0005622	2.143e-09	134/360 | 1144/6017	intracellular

GO terms of the cellular component category that are overrepresented in the highly correlating protein subset

GO ID	P-value (FDR)	Statistics	Description
GO:0005839	8.286e-10	16/360 | 34/6017	proteasome core complex
GO:0005874	9.274e-10	18/360 | 33/6017	microtubule
GO:0044464	1.620e-09	305/360 | 3667/6017	cell part
GO:0043229	1.683e-09	149/360 | 1580/6017	intracellular organelle
GO:0043226	1.683e-09	149/360 | 1580/6017	organelle
GO:0044424	1.782e-09	262/360 | 2754/6017	intracellular part
GO:0043234	1.819e-09	101/360 | 777/6017	protein complex
GO:0032991	1.852e-09	208/360 | 1363/6017	macromolecular complex
GO:0044444	2.067e-09	155/360 | 1027/6017	cytoplasmic part

GO:0005622	2.143e-09	134/360 | 1144/6017	intracellular
GO:0043232	2.414e-09	104/360 | 529/6017	intracellular non-membrane-bounded organelle
GO:0043228	2.414e-09	104/360 | 529/6017	non-membrane-bounded organelle
GO:0030529	2.472e-09	103/360 | 521/6017	ribonucleoprotein complex
GO:0030117	3.080e-09	18/360 | 48/6017	membrane coat
GO:0005840	3.323e-09	101/360 | 469/6017	ribosome
GO:0005737	1.808e-08	53/360 | 379/6017	cytoplasm
GO:0030120	2.960e-07	12/360 | 30/6017	vesicle coat
GO:0005945	3.020e-07	8/360 | 12/6017	6-phosphofructokinase complex
GO:0019773	6.670e-07	9/360 | 17/6017	proteasome core complex, alpha-subunit complex
GO:0044430	9.442e-07	18/360 | 74/6017	cytoskeletal part
GO:0033178	6.544e-06	12/360 | 39/6017	proton-transporting two-sector ATPase complex, catalytic domain
GO:0044433	6.544e-06	12/360 | 39/6017	cytoplasmic vesicle part
GO:0030054	1.269e-05	6/360 | 9/6017	cell junction
GO:0044445	1.320e-05	11/360 | 35/6017	cytosolic part
GO:0030130	2.700e-05	6/360 | 10/6017	clathrin coat of trans-Golgi network vesicle
GO:0030125	2.700e-05	6/360 | 10/6017	clathrin vesicle coat
GO:0044422	3.438e-05	60/360 | 571/6017	organelle part
GO:0044446	3.438e-05	60/360 | 571/6017	intracellular organelle part
GO:0016469	3.565e-05	7/360 | 15/6017	proton-transporting two-sector ATPase complex
GO:0044431	6.776e-05	12/360 | 49/6017	Golgi apparatus part
GO:0030132	8.880e-05	6/360 | 12/6017	clathrin coat of coated pit
GO:0030118	8.880e-05	6/360 | 12/6017	clathrin coat
GO:0015935	4.113e-04	5/360 | 10/6017	small ribosomal subunit
GO:0030126	4.700e-04	4/360 | 6/6017	COPI vesicle coat
GO:0016023	5.262e-04	3/360 | 3/6017	cytoplasmic membrane-bounded vesicle
GO:0031982	5.262e-04	3/360 | 3/6017	vesicle
GO:0031410	5.262e-04	3/360 | 3/6017	cytoplasmic vesicle
GO:0031988	5.262e-04	3/360 | 3/6017	membrane-bounded vesicle
GO:0005853	1.760e-03	4/360 | 8/6017	eukaryotic translation elongation factor 1 complex
GO:0046930	1.881e-03	6/360 | 20/6017	pore complex
GO:0005643	1.881e-03	6/360 | 20/6017	nuclear pore
GO:0044391	1.926e-03	9/360 | 43/6017	ribosomal subunit
GO:0044459	4.310e-03	7/360 | 31/6017	plasma membrane part
GO:0005798	7.578e-03	2/360 | 2/6017	Golgi-associated vesicle
GO:0008250	7.578e-03	2/360 | 2/6017	oligosaccharyltransferase complex
GO:0030119	3.101e-02	4/360 | 17/6017	AP-type membrane coat adaptor complex
GO:0030131	3.101e-02	4/360 | 17/6017	clathrin adaptor complex
GO:0000275	3.736e-02	2/360 | 4/6017	mitochondrial proton-transporting ATP synthase complex, catalytic core F(1)

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway