Institut für Biochemie und Biologie
AG Bioinformatik

Integrative analysis of heterogeneous plant cell wall related data

Dissertation
zur Erlangung des akademischen Grades
"doctor rerum naturalium"
(Dr. rer. nat.)
in der Wissenschaftsdisziplin "Bioinformatik"

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
der Universität Potsdam

von
Dhivyaa Rajasundaram

Potsdam, im März 2015

Published online at the
Institutional Repository of the University of Potsdam:
URN urn:nbn:de:kobv:517-opus4-77652
http://nbn-resolving.de/urn:nbn:de:kobv:517-opus4-77652

Abstract

Plant cell walls are complex structures that underpin plant growth and are widely exploited in diverse human activities thus placing them with a central importance in biology. Cell walls have been a prominent area of research for a long time, but the chemical complexity and diversity of cell walls not just between species, but also within plants, between cell-types, and between cell wall micro-domains pose several challenges. Progress accelerated several-fold in cell wall biology owing to advances in sequencing technology, aided soon thereafter by advances in omics and imaging technologies. This development provides additional perspectives of cell walls across a rapidly growing number of species, highlighting a myriad of architectures, compositions, and functions. Furthermore, rather than the component centric view, integrative analysis of the different cell wall components across system-levels help to gain a more in-depth understanding of the structure and biosynthesis of the cell envelope and its interactions with the environment.

To this end, in this work three case studies are detailed, all pertaining to the integrative analysis of heterogeneous cell wall related data arising from different system-levels and analytical techniques. A detailed account of multiblock methods is provided and in particular canonical correlation and regression methods of data integration are discussed. In the first integrative analysis, by employing canonical correlation analysis - a multivariate statistical technique to study the association between two datasets - novel insight to the relationship between glycans and phenotypic traits is gained. In addition, sparse partial least squares regression approach that adapts Lasso penalization and allows for the selection of a subset of variables was employed. The second case study focuses on an integrative analysis of images obtained from different spectroscopic techniques. By employing yet another multiblock approach - multiple co-inertia analysis, *insitu* biochemical composition of cell walls from different cell-types is studied thereby highlighting the common and complementary parts of the two hyperspectral imaging techniques. Finally, the third integrative analysis facilitates gene expression analysis of the Arabidopsis root transcriptome and translatome for the identification of cell wall related genes and compare expression patterns of cell wall synthesis genes. The computational analysis considered correlation and variation of expression across cell-types at both system-levels, and also provides insight into the degree of co-regulatory relationships that are preserved between the two processes.

The integrative analysis of glycan data and phenotypic traits in cotton fibers using canonical methods led to the identification of specific polysaccharides which may play a major role during fiber development for the final fiber characteristics. Furthermore, this analysis provides a base for future studies on glycan arrays in case of developing cotton fibers. The integrative analysis of images from infrared and Raman spectroscopic approaches allowed the coupling of different analytical techniques to characterize complex biological material, thereby, representing various facets of their chemical properties. Moreover, the results from the co-inertia analysis demonstrated that the study was well adapted as it is relevant for coupling data tables in a symmetric way. Several indicators are proposed to investigate how the global and block scores are related.

In addition, studying the root cells of *Arabidopsis thaliana* allowed positing a novel pipeline to systematically investigate and integrate the different levels of information available at the global and single-cell level. The conducted analysis also confirms that previously identified key transcriptional activators of secondary cell wall development display highly conserved patterns of transcription and translation across the investigated cell-types. Moreover, the biological processes that display conserved and divergent patterns based on the cell-type-specific expression and translation levels are identified.

Abstrakt

Pflanzliche Zellwände sind komplexe Strukturen, die wichtig für das Zellwachstum und auch nützlich für den Menschen sind, weshalb sie eine wichtige zentrale Rolle in der Biologie haben. Zellwände sind schon seit einiger Zeit ein bedeutsames Untersuchungsgebiet, jedoch stellen Fragen nach der chemischen Komplexizität und Diversität nicht nur zwischen Zellwänden verschiedener Spezies, sondern auch innerhalb von Pflanzen, zwischen verschiedenen Zelltypen und auch zwischen dem Mikro-Bereich von Zellwänden eine Herausforderung dar. Ein großer Fortschritt in der Forschung konnte durch die Weiterentwicklung der Sequenziertechniken erzielt werden, sowie auch durch Fortschritte der "omik"-Technologien und Imaging-Technologien. Dieser Fortschritt ermöglicht eine zusätzliche Perspektive auf Zellwände über die stark wachsende Anzahl verschiedener Spezies, die eine Vielzahl von Architekturen, Zusammensetzung und Funktionen hervorhebt. Des Weiteren werden statt einer Komponenten-zentrierten Sichtweise eine integrative Analyse der verschiedenen Zellwandkomponenten über unterschiedliche Systemebenen genutzt, um ein tieferes Verständnis über die Struktur und Biosynthese der Zellhülle und ihrer Wechselwirkung mit der Umgebung zu erlangen.

Zu diesem Zweck werden in dieser Arbeit drei Fallstudien ausführlich beschrieben, die sich alle auf die integrative Analyse von heterogenen zellwandbezogenen Daten beziehen, die von unterschiedlichen Systemebenen und analystischen Techniken stammen. Eine detailierte Darstellung von Multiblock-Methoden wird verschafft, wobei besonders kanonische Korrelationsanalyse und Regressionsmethoden der Datenintegration diskutiert werden. In der ersten Studie werden unter Einsatz kanonischer Korrelationsanalyse - einer multivariaten statistischen Technik, um Zusammenhänge zwischen zwei Datensätzen zu ermitteln - angewendet, um neue Erkenntnisse in Bezug auf die Beziehungen zwischen Glycanen und phenotypischen Merkmalen zu erhalten. In der zweiten integrativen Analyse wird ein Sparse Partial Least Square Regressionsansatz verwendet, der Lasso Penalization anwendet und die Auswahl von einem Sub-Set von Variablen erlaubt. Außerdem fokussiert sich die zweite Studie auch auf integrative Analyse von Bildern, die von zwei verschiedenen spektroskopischen Techniken aufgenommen wurden. Zunächst werden die zwei Sets von Bildern vor-bearbeitet und so aufbereitet, dass sie eine Blockdaten-Struktur bilden. Durch Anwendung eines weiteren Multiblock-Verfahrens, der multiple Co-Inertia Analyse, wird die *in-situ* biochemische Zusammensetzung der Zellwände von verschiedenen Zelltypen untersucht und die Gemeinsamkeiten und Unterschiede der zwei hyperspektralen Imaging-Techniken hervorgehoben. Zuletzt ermöglicht die dritte Studie eine integrative Genexpressions-Analyse des Arabidopsis Wurzeltranskriptoms und -translatoms zur Identifikation von zellwandbezogenen Genen und dem Vergleich von Expressionsmustern von Zellwandsynthese-Genen. Die numerische Analyse zieht sowohl Korrelation als auch Variation der Genexpression verschiedener Zelltypen auf den beiden Systemebenen in Betracht und liefert so einen Einblick in den Grad der ko-regulierten Beziehungen, die zwischen den beiden Prozessen konserviert sind.

Die integrative Analyse der Glycandaten und den phenotypischen Merkmalen in Baumwollfasern unter Benutzung der kanonischen Methoden führte zur Identifikation von spezifischen Polysacchariden, welche eine wesentliche Rolle für die Entwicklung der finalen Fasereigenschaften spielen könnten. Weiterhin stellt diese Analyse eine Basis für zukünftige Studien über Glycanarrays von sich in der Entwicklung befindlichen Baumwollfasern dar. Die integrative Analyse von Bildern von Infrarot- und Raman-spektroskopischen Methoden erlaubt die Verknüpfung von verschiedenen analytischen Techniken, um komplexes biologisches Material zu charakterisieren, und somit eine Vielzahl ihrer chemischen Eigenschaften darzustellen. Darüber hinaus zeigen die Ergebnisse der Co-Inertia Analyse, dass die Studie gut adaptiert ist, was relevant für die symmetrische Verknüpfung von Datentabellen ist, aber auch weil mehrere Indikatoren vorgestellt wurden, um zu untersuchen, in wie fern die globalen und Block-Scores in Beziehung stehen. Außerdem konnte durch die Untersuchung der Wurzelzellen von *Arabidopsis thaliana* eine neue Pipeline zur systematischen Untersuchung und Integration verschiedener Informationsebenen auf globaler und Einzelzellebene zuzüglich Identifikation von zellwandbezogenen Genen postuliert werden. Die ausgeführte Analyse bestätig auch, dass vorherige identifizierte Schlüssel-Transkriptionsfaktoren der sekundären Zellwandentwicklung hoch-konservierte Transkripions- und Translationsmuster in den untersuchten Zelltypen zeigen. Dazu kommt, dass biologischen Prozesse mit konservierten und divergenten Mustern auf zelltypspezifischer Expressions- und Translationebene identifiziert werden konnten.

Acknowledgment

I am deeply appreciative of the many individuals who have continually encouraged me through the writing of this dissertation. Without their time, attention, encouragement, thoughtful feedback, and patience, I would not have been able to see it through.

I would like to express my special appreciation and thanks to my supervisor Prof. Dr. Joachim Selbig for his valuable guidance, scholarly inputs and consistent encouragement. I consider it as a great opportunity to do my doctoral research under his guidance and to learn from his research expertise.

I would especially like to thank Dr. Sebastian Klie for his brilliant comments and insightful discussions throughout my PhD.

Thanks to Prof. Dr. Staffan Persson, Dr. Dirk Walther and Dr. Marek Mutwil for being a part of my thesis committee and providing assistance at all levels of the research project.

A special thanks to Dr. Marie-Christine Ralet and Alexandre Thébaud for the wonderful network "Wall-TraC", and for all the training events and workshops.

I would also like to thank Dr. Frank Meulwaeter, and Dr. Jean-Luc Runavot, Bayer CropSciences for the collaborative project. I am deeply thankful to Dr. Marie-Francoise Devaux, Prof. Dr. Fabienne Guillon, and all other staff and friends at INRA, Nantes for a fruitful three month secondment.

Thanks to AG Selbig for the nice working environment and a special thanks to Dr. Stefanie Hartman, Jacqueline Novak, and Dr. Sebastian Proost for proof-reading my thesis.

I have amazing parents and a brother, unique in many ways, and the stereotype of a perfect family in many others. Their support has been unconditional all these years and they have cherished with me every great moment.

I am indebted to all my friends who have supported me over the last few years. Thanks to Arun and his family for their encouraging support, endless skype sessions, and for the many precious memories along the way.

Contents

Chapter 1

Introduction

Plants as living organisms arguably hold an exceptional place as dynamic components of our world that shape and are, in turn, shaped by the environment. The evolution of plants is an important chapter in the history of life and forms the backbone of human well-being on earth. Throughout their life, plants typically remain in one location and harvest the physical energy of the sunlight through photosynthesis, which is stored chemically in the form of carbohydrate-based polymers. These carbohydrates serve as the sole source of energy as well as the source of building blocks of a protective extra-cellular matrix, called the cell wall (McCann and Rose, 2010, Sørensen et al., 2010, Chapelle and Carpita, 1998). Without the ability to evolve locomotive mechanisms, plants face many challenges to survive predation, unfavorable climatic conditions, and scarcity of resources. During the course of evolution, they have adapted to their respective niche resulting in a wide variety of morphological changes. This specialization of function is reflected at the molecular level in the specific designs of cell walls. Thus cell walls not only changed throughout evolution but are constantly remodeled and reconstructed in response to environmental stress during the development of an individual plant (Purbasha et al., 2009, Popper, 2008, Sørensen et al., 2010).

Plant cell walls are complex and dynamic structures composed mainly of polysaccharides such as cellulose, hemicellulose, pectins, and a variety of other minor components, including proteins and lignin, thus allowing them to perform various functions (Heredia et al., 1995, Keegstra, 2010). The physical aspects of cell walls provide tensile strength to the plant body and acts as a physical barrier against biotic and abiotic stresses. The biochemical functions include many aspects of growth and development, such as cell division, growth, cell-cell communication, and differentiation. Nutritionally, plant cell wall polysaccharides are the major food and feed products that are used globally. Further-

more, these glycan rich cell walls are of commercial importance, and are widely used as gelling and thickening agents in the food industry, fiber products in textile industry, pharmaceuticals, and as materials for fuel and composite manufacture (McCann and Carpita, 2008, Chapelle and Carpita, 1998, Thakur et al., 1997). This wide reach of issues pertaining to cell walls and their components places them with a central importance in biology. Therefore, research on cell walls is essential to gain new insights into the role of cell walls during evolution and investigate how the observed diversity varies between different plant lineages, within a single plant, between cell-types, and even within individual walls.

1.1 Understanding the functional architecture of plant cell walls

Since the early 1970's cell wall biology has been an area of prominent research, but the plant community has been facing many challenges in understanding cell wall structure, function, and biosynthesis. Increasingly, genetic manipulation and conventional breeding techniques are being applied in attempts to modify the quality and quantity of individual cell wall components for broader commercial applications. The best candidates for this manipulation are the genes, enzymes, and biochemical pathways involved in the cell wall biosynthesis process, which includes multiple cellular components in different locations that are assembled into a functional wall matrix (Somerville et al., 2004, Somerville, 2006). Biochemical analyses have revealed that all plant cell walls share common features, which form the mechanical framework of the cell. However, cell walls exhibit diversity with respect to chemical composition and are modulated according to functional requirements, thereby, limiting our knowledge on cell wall design and maintenance (Roberts, 2001, Pilling and Höfte, 2003). Although biochemical analysis yields the composition and stoichiometry of cell wall components, fractionation and purification of intact polymers is quite complicated (McCann et al., 2001, Minorsky, 2002). Biochemical analyses complemented by genetic analyses help to identify the genes that are required for the synthesis and metabolism of cell wall synthesis (Reiter, 2002, Somerville et al., 2004, Mutwil et al., 2008, Ruprecht and Persson, 2012). It has been estimated that well over 2000 different gene products are involved in making and maintaining the cell wall (Somerville, 2006). In contrast, limited knowledge on the temporal and spatial patterns of the identified gene products together with a limited understanding of the physical and chemical interactions between wall components has hampered our understanding of the mechanisms and control of the bio-synthetic steps. Another goal is to examine the structural details of the

cell walls of various plants, and thereby determine how their primary, secondary, and higher-order structures vary from tissue to tissue and from plant to plant.

Although genetics and genomics have proved to be key tools in the past decade, new biophysical methods such as imaging and nanoscale interrogation are increasingly being used for characterizing the machinery that underlies the formation and growth of the cell wall (McCann and Carpita, 2008, Somerville et al., 2004, Minorsky, 2002). In order to develop a coherent picture of this complex process, it is necessary to combine information regarding the biosynthetic mechanisms and chemical structures of cell wall polysaccharides, the physical bases of molecular conformations for their assembly, the nature of their covalent and non-covalent interconnections, the specificity and regulation of enzymes that catalyze the formation of these interconnections, the overall topology of interconnected polysaccharide networks, and the rheological consequences of these interacting factors (Somerville et al., 2004, Purbasha et al., 2009). This is a formidable problem demanding a multidisciplinary approach.

1.2 Technology-driven opportunities and data generation

All of our exalted technological progress, civilization for that matter, is comparable to an axe in the hand of a pathological criminal

-Albert Einstein

These puzzling words by Einstein are certainly true in the field of plant sciences - thanks to the technological advances in the past few decades, we live in an exciting and progressive era. Recent game-changing technologies have led to the development of new techniques, including the adoption and adaptation of instruments from other fields at the highest levels of measurement resolution to query unchartered frontiers in the knowledge of plant growth, metabolism, and response to environmental cues. Current technologies have facilitated the measurements of cell content and activity at the whole plant, tissue and organ, cell layer, single cell, and even compartment level (Yuan et al., 2008b, Mochida and Shinozaki, 2011, 2010).

Of all the high-speed analytical techniques developed, two of the most commonly used ones are next-generation sequencing and microarray technology which enable a comprehensive understanding of complex biological systems. The first description of microarrays dates back to 1995 when Schena and co-workers used it for transcriptome analyses, and has a profound impact on gene expression research with a broad range of applications

(Schena et al., 1995, Peeters and Van der Spek, 2005). There exist several microarray platforms based on differences in array fabrication, density and length of the spotted probes, selection and number of dyes used. As microarray technology became widely adopted in profiling gene expression, the advent of next generation sequencing technologies (NGS) and their relatively low cost has made it possible to sequence genomes and profile transcriptomes, thereby widening the biological questions that scientists can investigate. Several NGS platforms have been released by various companies, and a few representative platforms are HiSeq and MiSeq (Illumina), and PacBio (Pacific Biotechnology) (Egan et al., 2012, Mochida and Shinozaki, 2010, 2011). In each of these platforms, distinct methods of template preparation and signal detection are used. The whole genome sequencing of *Arabidopsis thaliana* completed in 2000 was a milestone in plant biology and made Arabidopsis one of the most popular species for basic plant research. Following this, genome sequences for around 55 plant species have been completed as of 2013 (Michael and Jackson, 2013). Large-scale sequence analysis technology has opened a wide range of opportunities beyond genome sequencing, and it has contributed substantially to recent advances in omics research which spans a wide range of fields ranging from 'genomics' (complete set of DNA within an organism), 'transcriptomics' (RNA and gene expression), 'translatomics' (study of polysomal mRNA), 'proteomics' (focussing on protein abundance), 'metabolomics' (metabolites), 'glycomics' (study of glycans), and 'phenomics' (study of plant phenotypic traits).

Higher plants contain organs and tissues that comprise interspersions of different cell-types. In addition, the plant developmental signals differ across cell-types within an organ and in response to environmental stimuli. To this end, advances in omics technologies produce cell-type specific transcript profiles that allow functional annotation of genes, as well as elucidate many gene networks that were masked at the organ level due to restricted expression (Mochida and Shinozaki, 2010, Fukushima et al., 2009). Moreover, cell-type-specific metabolic profiles take into account variability in the concentration and chemical properties of the metabolites, which could otherwise not be obtained from intact organ metabolomics (Rogers et al., 2012). This could be done using the state-of-the-art technologies in mass spectrometry which pinpoints details at the single cell and sub-cellular resolutions (Kueger et al., 2012, Oikawa and Saito, 2012). In addition to our knowledge of genomics and metabolomics of plant tissues, exploration of the functional and structural role of specific classes of chemicals such as lipids is possible. Technological advances to dissect the lipid composition of extract, as well as the possibility to visualize lipids in a sub-cellular context are increasingly available to researchers (Horn and Chapman, 2012). For applications of mass spectrometry imaging technologies in plants, as for example in

surface lipids and secondary metabolites, usually chemical images are formed by compiling spectra of ions of interest (Lee et al., 2012). As of now, this could be done mostly up to the single-cell resolution level, because increasing the spatial resolution reduces the sensitivity of chemical analyses. Of particular note in this respect is that all of those approaches described above are at the macro-scale of investigating plant species, but the field of plant cell imaging has been developed extensively allowing micro-scale analysis.

It is fascinating to witness how technological advances in the optical imaging of cellular structures are operating over a range of spatial scales, spanning from tissues to single molecules, monitoring activities of sub-cellular compartments to levels that were not even imaginable just over a decade ago (Ehrhardt and Frommer, 2012). Parallel advances in fluorescent protein technology (FPT) image sub-cellular dynamics of plant cell organelles at a spatial and temporal resolution. In addition, FPT also manipulates the distribution of fluorescent markers to identify the genes responsible for the inner activities of plant cells by coupling light microscopy with genetics and genomics (Sparkes and Brandizzi, 2012). Innovative approaches based on laser trap microscopy manipulate the position of organelles and probe fundamental and exciting questions about inter-organelle relationships in live cells (Okumoto, 2012). In addition, adaptations of spectroscopic techniques could be used to query the composition and the activity of sub-cellular compartments at a scale of few micrometers (Gierlinger et al., 2012, Gierlinger, 2014).

Ehrhardt and Frommer (2012) provides an excellent account of the technical advances in the field of plant science. Thanks to such advancements - plus the breadth of techniques which offer a wide spectrum of data resources increasingly available to researchers for data mining and hypothesis generation.

1.3 Heterogeneous data analysis and challenges

Overall, large, heterogeneous data is composed of several data-types. The first category is nucleic acid sequence data, which mainly consists of genomic DNA sequences and several subgroups of RNA sequences. Gene expression data at the RNA level constitutes the second category. To date, this category contains data mostly from microarray and RNA-seq to effectively infer gene activities in a spatially-defined and time-resolved manner. The third category includes protein data such as protein-level expression, protein sequences, and even secondary and three dimensional structures of proteins that help to translate static genetic information to dynamic carriers of cellular activities. The fourth category of data includes interactions such as protein-protein interaction, protein-DNA/RNA interactions and genetic interactions. Metabolite data comprising the intermediary and

end products of metabolism corresponds to the fifth category of data. This category of data captures the cellular physiology and complements the big picture from genes to gene expression to cellular activities. The sixth category comprises of annotation data from various sources such as standard biological annotation terms such as gene ontology (GO) and Kyoto encyclopedia of genes and genomes (KEGG). Specifically, this category of data is useful for biologists to learn about each aspect of the biological system under study and helpful when leveraging the study from the known to the unknown. The last category of data includes the data from imaging and spectroscopic approaches characterizing the study of different cell-types at a scale of few micrometers.

Our repertoire of functional omics tools is steadily increasing, however, analysis of this data requires significant statistical and computational efforts. From a computational point of view, a great deal of information regarding cellular metabolism has been acquired through application of individual omics approaches. However, it is also becoming clear that any single omics approach may not be sufficient to characterize the complexity of biological systems. For instance, the expression level of a given gene does not indicate the amount of protein produced, nor its location, biological activity or functional relationship with metabolites. Moreover, in cells many levels of regulation occur after genes have been transcribed, such as post-transcriptional, and post-translational regulation, and all forms of biochemical control such as allosteric or feedback regulation (Mochida and Shinozaki, 2011, Ehrhardt and Frommer, 2012). For example, in a study on how cells of different types arise from a homogeneous cell pool during development of an organism, connections must be made based on the interplay of genes, gene products, hormone pathways, metabolites, and signaling pathways to determine how the components work together as a system. Taking this view into account, it is hard to believe that functional genomics can stop at the mRNA level or any other single level of information. To this end, integration of multiple layers of information to understand the functional principles and total dynamics of cellular systems is essential.

1.4 Thesis outline

The general aim of this thesis is an understanding of various cell wall related aspects by an integrative analysis of high-throughput data arising from modern systems biology experiments. The centrality of the biological question addressed in this thesis, i.e., an understanding of plant cell walls will undoubtedly create new opportunities for the development of crops with enhanced productivity, nutritional value, and biotechnological potential.

Specifically, Chapter 2 introduces multivariate analysis of heterogeneous data using multiblock methods, a statistical technique for the integration of two or more datasets. This chapter covers the fundamental principles and types of multiblock methods, of which the mathematical aspects and hypothesis testing of canonical correlation and regression based methods are provided. As it will be described in the subsequent chapters, only a short account of the three case studies is provided of which the first and third study focus on the integrative analysis of different system-levels. The second case study focuses on cross-platform comparisons using an integrative approach.

In Chapter 3, the integration of system-level data from cotton fibers is discussed. The relationship between the system-levels was established in an integrative manner using linear correlation and regression methods. The conducted analysis demonstrated the usefulness of regression based approaches in establishing a relationship between polysaccharide-rich cell walls and their phenotypic characteristics. In addition, the analysis also identified specific polysaccharides which may play a major role during fiber development for the final fiber characteristics (cf. Rajasundaram et al. (2014a)).

Chapter 4 describes the comparison and integration of data that estimates the cell-type composition of maize stem cell walls using two different spectroscopic techniques. The analysis of different hyperspectral images was done initially using exploratory approaches and then compared to results obtained from coupling of the images from both techniques using multiblock methods. The integrative analysis of the hyperspectral images helps to interpret and analyze on one hand the common structure revealed by the two imaging techniques and on the other hand the independent contribution of each technique.

Chapter 5 details a novel analysis pipeline for integrating different system-levels from the roots of Arabidopsis. Using cell-type-specific datasets of the root transcriptome and translatome of Arabidopsis, a systematic assessment was made of the degree of co-ordination and divergence between these two levels of cellular organization. The computational study focuses on comparing these two system-levels in the context of cell wall biogenesis by considering correlation and variation of expression across cell-types in addition to the degree of co-regulatory relationships. The importance of the translatome as the intermediate level at which reprogramming of biological processes propagate to changes in protein synthesis and finally the phenotype is elaborated here. (cf. Rajasundaram et al. (2014b)).

Finally, Chapter 6 provides a general conclusion of this thesis and highlights the contributions of the integrative analysis towards an understanding of the different aspects of plant cell walls. In addition, this chapter also provides an outlook, thus outlining a method for knowledge-transfer across other plant species.

Chapter 2

Multivariate analysis of multi-source data

The term 'systems biology' has emerged recently to describe the frontier of cross-disciplinary research in biology. Interest in systems biology has increased owing to the rapid advancements in high-throughput technologies and the need to assimilate fast-growing volumes of biological data into biologically meaningful interpretations (Kitano, 2002a). Nevertheless, the definition of systems biology is still contentious, reflecting the difficulty in defining a heterogeneous school of thought by a comprehensive, yet concise, definition. However, the consensus view of systems biology revolves around a fundamental understanding of biological systems by studying the underlying component interactions, which otherwise cannot be studied by a reductionist methodology (Hood, 2003, Kitano, 2002b). Past biological research has taught us about how individual biological units are structured and function, and the future lies in focusing on a system-level understanding emerging from dynamic interactions of individual components in a biological system.

Modern systems biology is a rapidly evolving discipline invoked in the context of a variety of systems and overlaps with several emerging, post-genomics fields such as systems microbiology, systems biotechnology, integrative biology, and metagenomics. While many systems biology approaches involve mathematical and computational modeling, there is often no clear boundary between bioinformatics and systems biology. As a discipline, bioinformatics is also increasingly merging and contributing to system approaches which include the development of software tools for analysis and visualization of a system, sophisticated data processing, statistical analysis for high-throughput molecular profiling technologies, and maintenance of biological databases to account for the inherent complexity of biological systems (Ideker et al., 2001, Likić et al., 2010).

As discussed in the previous chapter, there is an ever-growing effort in plant biology to generate data from experiments measuring multiple cellular levels across a wide range of species. In addition, data from advances in microscopy to study the chemical and structural information is fast accumulating.

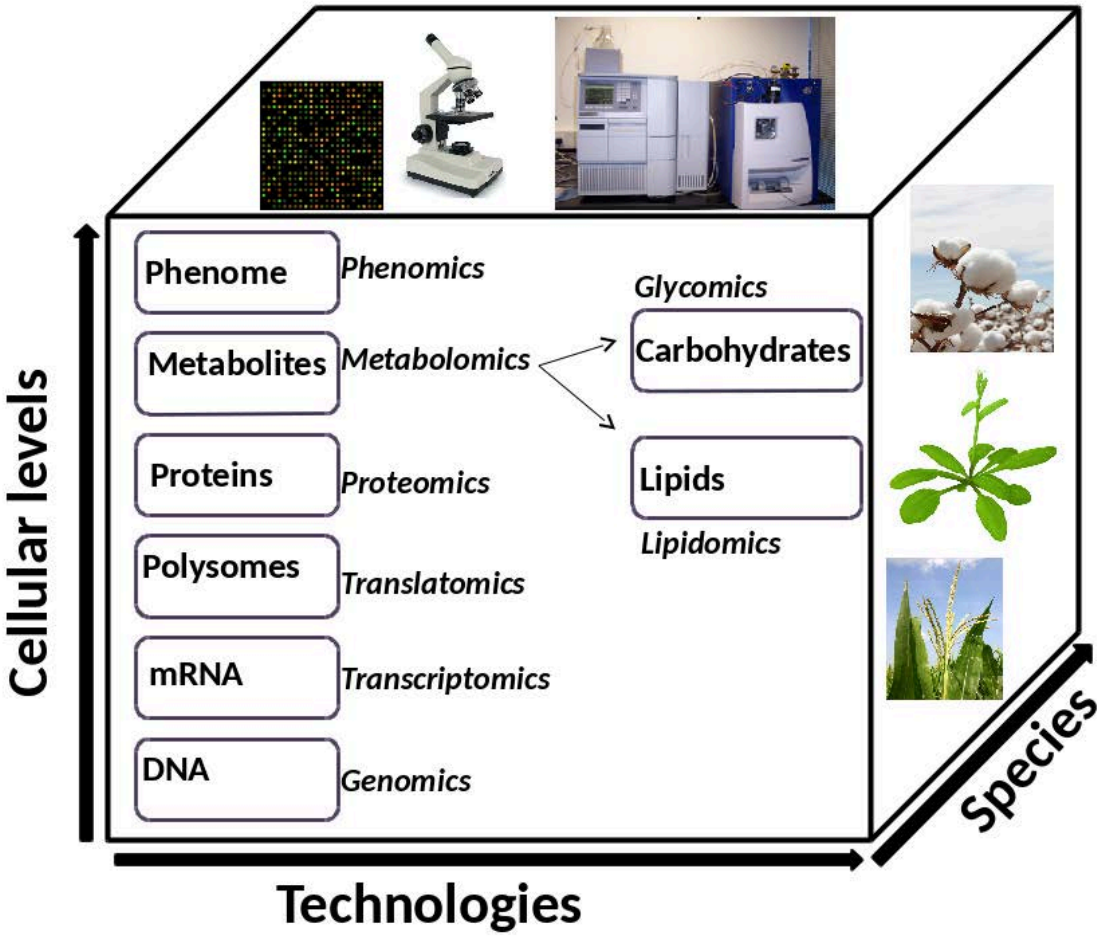


Figure 2.1: Hierarchical organization of biological complexity. The complexity of biological data is multi-dimensional owing to the advances in high-throughput technologies. The heterogeneity of the data sources is attributed to several factors ranging from the use of different technologies to studying multiple cellular levels across different species. Along the axes of the cellular levels, some of the possible system levels and their corresponding 'omics' is highlighted. There has been a dramatic increase in high-throughput technologies enabling the measurement of several omics levels starting from the genome to determining the phenotype. The short arrows within this figure depict that lipidomics is a direct subset of the metabolomics research area. Moreover, it also depicts that glycomics has the potential to gain insight into the biosynthesis of novel glyco-conjugate structures. This is done by probing the metabolome for substrates that are known to be involved in bio-synthetic processes.

Clearly, the multidimensional nature of the generated data (Figure 2.1) need systematic approaches that inherently require integration of heterogeneous information. The multi-source nature of the data is attributed to the use of different techniques and studying one or more plant species/population across different cellular levels at the macroscopic and microscopic scale.

2.1 Multiblock data analysis

With the availability of data from several fields, often problems occur that have to deal with analyzing several blocks of generated data. This has generated the need for two new classes of data analysis methods. The first class of methods include multiblock methods - the aim of which is to find underlying, i.e. latent relationships between several blocks, or matrices under the hypothesis that they are related. From the point of view of statistical methods, two approaches of multiblock methods can be contrasted; the multiblock component problem and the multiblock regression problem (De Roover et al., 2012). In the multiblock component problem, it is assumed that there is at least one mode in common between all the matrices involved, and component vectors that summarize the information in all the matrices simultaneously are sought (Smilde et al., 2003). In a multiblock regression problem, the models are designed to predict a certain response in a multiblock setting. The second class of methods namely the multiway analysis methods pertain to the analysis of datasets that can be arranged into a multiway array (Smilde et al., 2005, Coppi, 1994). Precisely, many problems in biology generate three-way data and this three-dimensional data matrix could be analyzed using two alternative strategies, a bi-linear or a tri-linear approach. Smilde et al. (2000) have also developed methods for analyzing multiway multiblock problems.

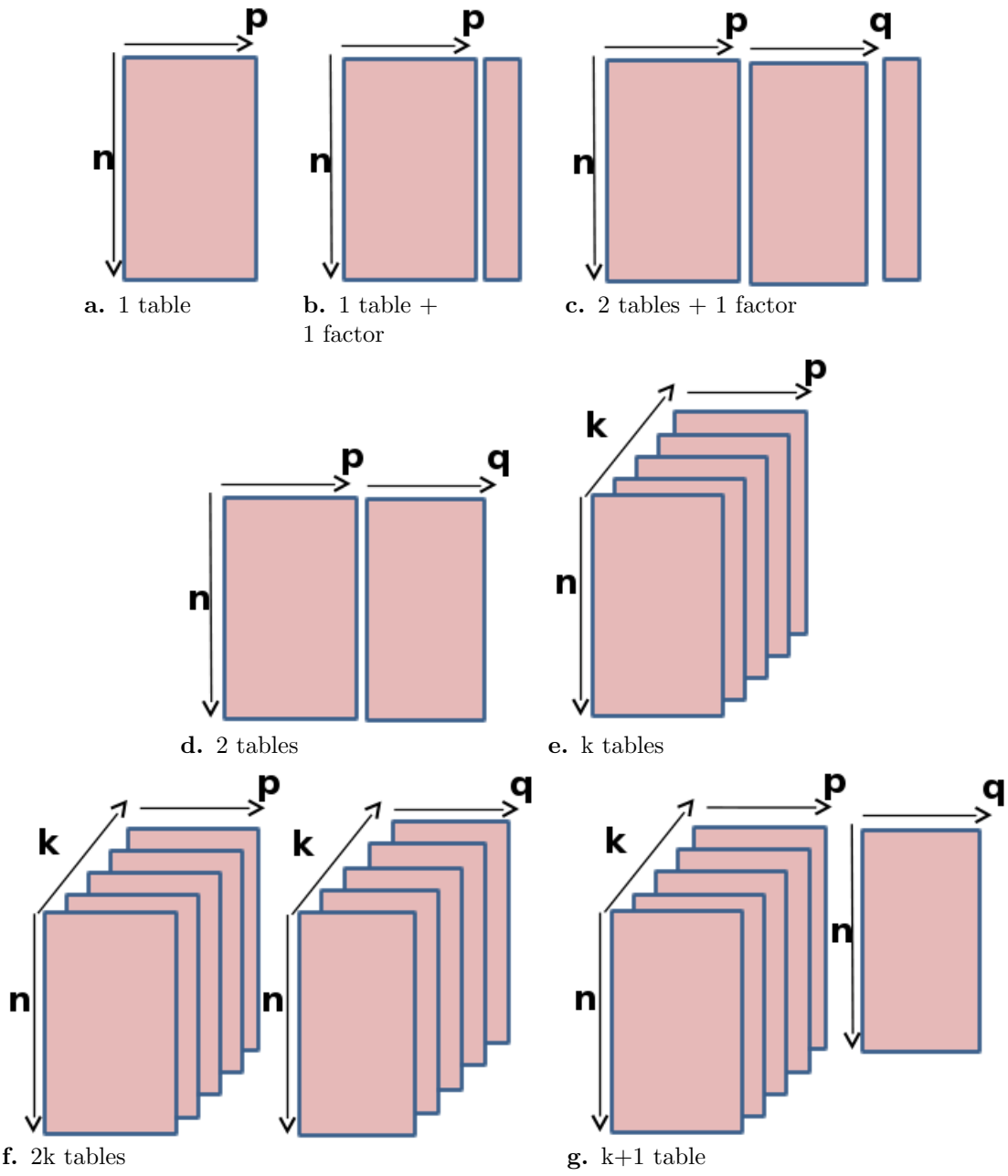
Multiblock analysis methods already have a long history in psychometrics and are still the subject of active research in the field of ecology, food quality assessment, chemometrics, and computational statistics (Kettenring, 1966, Moyon et al., 2012, Thioulouse et al., 2004, Thioulouse, 2011, Skov et al., 2014). In ecological data analysis, these methods have been developed for studying species-environment relationships. Exploring species-environment relationships is quite complicated owing to spatial and temporal influences and multiblock methods have contributed a necessary step towards the comprehension of ecosystem functioning. In addition, it is also used in the field of veterinary epidemiology to assess risk factors for animal health issues (Bougeard et al., 2011a). In sensory product development, this method gives a reliable basis for creating products which meets customer expectations by using sensory profiles to model consumer liking (Stéphanie and

Mireille, 2014). Other contributions in this field include understanding the relative contribution of oral food processing events in cheese consumption (Feron et al., 2014) and serve as powerful tools in analyzing spectroscopic data.

From a statistical perspective, all the data involving multiple omics data-types, multiple tissues, multiple sources, or any of their combination constitute separate data blocks and can be formulated as a multiblock framework problem. The resulting multivariate datasets gathered from biological entities (e.g., genes) under various conditions (e.g.,cellular levels) can be investigated by application of multiblock methods which account for the dependence between the data tables while capturing the common and unique information across them. This perception of high-throughput data analysis proposes a great variety of methods to integrate, summarize and visualize the data blocks so that the systems picture can be drawn.

Depending on the structure of the datasets under study, there are different classifications of multiblock problems. In Figure 2.2, we provide a schematic representation of some of the types of multiblock data representation wherein a number of variables have been measured on some samples, and the data for each subject constitute a different data block or data table.

The general representation of a single block is shown in Figure 2.2a and the data is obtained from 'n' samples and 'p' variables. The second class of problems is encountered in cases where we have '1 table + 1 factor' (Figure 2.2b) or '2 tables + 1 factor' (Figure 2.2c). The third class (Figure 2.2d) is where we have a pair of tables and usually the variable mode is different among the data blocks. In yet another case (Figure 2.2e), we have a 'k table structure' wherein the samples are the same but the measured variables are different. In this case 'k' refers to the different conditions (e.g., timepoints) in which the data is obtained. Stacking such datasets on top of each other gives a three-way array and hence multiblock methods can be seen as a generalization of multiway models. In another case, we have a series of pairs of '2k tables' (Figure 2.2f) although measured across different conditions. In Figure 2.2g, we have the 'k+1 table' data block structure consisting of 'k' explanatory blocks and a response dataset to be explained or predicted. It is not directly evident from the structure of the data if there exists a relationship between the different blocks of data. In the subsequent sections of this chapter, the focus is on detailing the analysis of two block data tables.



a. 1 table **b.** 1 table + 1 factor **c.** 2 tables + 1 factor

d. 2 tables **e.** k tables

f. 2k tables **g.** k+1 table

Figure 2.2: Schematic representation of possible types of data structures available from different experiments. (a) In a general block framework, each block is a n × p data matrix as represented here corresponding to a set of 'p' variables observed on a set of 'n' observations. (b and c) Both data representations are tailored to study data tables and their relation to factors like categorical variables or physico-chemical characteristics. (d) The '2 table' analysis deals with 'n' observations across variables 'p' and 'q'. (e and f) The 'k tables' or the '2k tables' representations are used to analyze series of tables where 'k' refers to experimental conditions such as timepoints or batches measured on 'n' observations across variables 'p' or 'q'. (g) This pertains to the k+1 setting addressed for the analysis of multiple 'k' tables which aims at explaining a response data block.

2.2 Data pre-processing

Data pre-processing of large and complex datasets produced by high-throughput biological experiments is one of the prime challenges in statistical analysis of the data. Complex biological, experimental and technical processing inherent to the technology introduces sources of variations in the data that alters the true signals (Mühlberger et al., 2011, Zhang et al., 2010, Kohl et al., 2014). Although there is no standard protocol for data pre-processing, normalization and missing value imputation generally occur as pre-processing steps followed by statistical inference to answer questions of primary scientific interest. In an effort to enable comparisons between samples, normalization techniques remove any excess technical variability whilst preserving biological information. In this thesis, it is emphasized that pre-processing is an important step of the analysis pipeline, and that the assumptions and limitations of the pre-processing method should always be taken into account.

Systematic bias refers to any non-biological signal which may occur due to many factors including variations in sample processing conditions, instrument calibrations, and changes in temperature over the course of an experiment. Expression levels of RNA transcripts are widely monitored using microarrays and the measured expression values are typically unitless. There are a number of reasons to normalize a microarray, which includes type of the microarray used, unequal starting quantities of starting RNA, differences in labeling, differences in use of fluorescent dyes and resulting detection efficiencies, hybridization artifacts, and accurate comparisons of expression levels between and within samples (Durinck, 2008, Fujita et al., 2006, Quackenbush, 2002). Recent transcriptomic approaches based on sequencing of transcripts result in sequence reads which are then mapped onto a reference genome. Often, sources of systematic variation that must be taken into account include library size, technical variation amongst samples, biases in sequencing e.g., longer fragments are sampled more frequently, preferential enrichment of specific sequences (ChIP-seq), within-sample gene-specific effects related to gene length, and GC-content. Most importantly, before starting the analysis on any dataset it is good to examine the biases present and choose a normalization method that does not mask the real biological differences between samples. A number of normalization approaches to treat RNA-seq data is available and relevant ones must be adopted prior to statistical analysis (Goncalves et al., 2011, Oshlack et al., 2010, Dillies et al., 2013). In case of metabolomic approaches, samples have large amounts of biological variability, variability from the analytical method itself, and diverse physical properties that makes quantification of large numbers of structurally diverse metabolites challenging. In such cases, data

pre-processing is essential to minimize or eliminate batch-batch data variation (De Livera et al., 2012, Wishart, 2010). In yet another example, high dimensional data arising from microscopic or spectroscopic techniques require pre-processing for proper band selection, spectral and temporal resolution, and to enhance the richness of the displayed data (Vidala and Manuel, 2012, Burger and Gowen, 2011). The aforementioned examples are only a few instances of why high-throughput data should be normalized and, importantly, it pertains to how specific methods should be used for particular data-types.

In the course of data analysis, standard methods for analyzing data lead to biased estimates and a loss of statistical power due to missing values (Liew et al., 2011). A value may be missing due to several reasons, including: (i) the sample truly is present at an abundance the instrument should be able to detect, but is not detected or is incorrectly identified, (ii) the sample truly is present but at an abundance below the instrument's detection limits, and (iii) the sample is not present. Most of the statistical methods require complete datasets, and furthermore missing values imply a certain loss of information. For this reason, the validity of results of a study with missing values has to be rated less than in a case where all data had been available. Common statistical procedures that are implemented in most of the statistical standard software packages or tools are based on a complete dataset, and in case of a missing value, the regarding observation will be excluded from the analysis population. The omission of whole observations may lead to a drastic reduction of the number of observations, and a reduced validity of the study (Gromski et al., 2014, Oh et al., 2011). There is no guideline yet that explicitly controls the handling of missing data and the usage of the most appropriate imputation method is highly dependent on several aspects. It mostly depends on the cause of the missing value in the data, dependence between observed and missing values, distribution, amount of the missing values in the dataset, and most importantly on the experimental settings and question under study. Some of the methods which have been reported in the literature include: replacing missing values by half of the minimum value found in the dataset; missing value imputation using probabilistic principal component analysis (PPCA), Bayesian PCA (BPCA) or singular value decomposition imputation (SVDImpute); replacing missing value by means of K nearest neighbors (KNN impute algorithm); simple row average (for gene expression analysis) or replacing the missing values with zeros (Aittokallio, 2010, Troyanskaya et al., 2001). Filling missing values with zeros or with average values over the cases are far from optimal solutions, and generally lead to serious biases.

Another important issue to be considered in multiblock analysis is replicate filtering. Replication is essential to identify and reduce the variation in any experimental assay design. Biological replicates are derived from distinct biological sources and provide a

measure of natural biological variability in the system under study, as well as any random variation in sample preparation. Technical replicates include replicated elements of a particular sample and accounts for the natural and systematic variability that occur in performing the experiment. To reduce the complexity of the dataset the replicate measures maybe averaged. The averaging is not justified if there is poor concordance between samples and the variance in each sample is not similar. This could be tested using a multivariate correlation estimator.

However, in the context of integrative analysis, the input biological data frequently originate from different experimental platforms, different levels of biological organization, and from a wide range of cells, tissues and organs (Palsson and Zengler, 2010). Moreover, it is not straight-forward to compare data from different experimental designs, and, therefore, normalization procedures must be applied carefully depending on the choice of data to be analyzed. The three different case studies considered in this thesis details the normalization methods used to facilitate cross-data comparison.

2.3 An overview of multiblock data analysis methods

With the unprecedented amount of information from high-throughput experiments, reliable and robust methods for integrating heterogeneous data have been developed and some of these methods have been covered in excellent reviews (Gonzalez et al., 2012, Lê Cao et al., 2009, Hamid et al., 2009, Sánchez et al., 2012). Table 2.1 gives an overview of some of the multiblock methods available for integrative analysis and the application of these methods in different contexts. However, it is beyond the scope of this thesis to discuss in detail all the available methods for data integration.

Method	**Type**	**R-package**	**References**
1 table + 1 factor			
Within-group analysis	Predictive approach	ADE4	Doldec and Chessel (1987)
Linear discriminant analysis	Predictive approach	ADE4	Venables and Ripley (2002)
Between-group analysis	Descriptive approach	ADE4	Culhane et al. (2002)
2 table + 1 factor			

Table 2.1 – *Continued from previous page*

Method	**Type**	**R-package**	**References**
Between group co-inertia analysis	Predictive approach	ADE4	Thioulouse (2011)
2 tables			
Canonical correspondence analysis	Predictive approach	Vegan	ter Braak (1986)
Multiple correspondence analysis	Descriptive approach	FactomineR	Tenenhaus and Young (1985)
Multiple coinertia analysis	Descriptive approach	omicade4	Chessel and Hanafi (1996)
Canonical correlation analysis	Descriptive approach	mixOmics	Weenink (2003), Hardoon (2004)
Partial least squares regression	Predictive approach	pls	Geladi and Kowalski (1986)
Sparse partial least squares regression	Predictive approach	mixOmics	Lê Cao et al. (2009)
k tables			
Multiple factor analysis	Descriptive approach	FactoMineR	Abdi et al. (2013)
STATIS	Descriptive approach	ADE4, MExPosition	Abdi et al. (2012), Lavit et al. (1994)
Partial triadic analysis	Descriptive approach	ADE4	Mendes et al. (2010)
dual-STATIS	Descriptive approach	ADE4.	Abdi et al. (2012)
COVSTATIS	Descriptive approach	ADE4, MExPosition	Abdi et al. (2012)
DISTATIS	Descriptive approach	ADE4, MExPosition, Distatis	Abdi et al. (2012)
ANISOSTATIS	Descriptive approach	ADE4, MExPosition	Abdi et al. (2012)
power-STATIS	Descriptive approach	ADE4	Abdi et al. (2012)

Table 2.1 – *Continued from previous page*

Method	Type	R-package	References
CANOSTATIS	Descriptive approach	ADE4, MExPosition	Abdi et al. (2012)
Regularized generalized canonical correlation analysis	Descriptive approach	RGCCA, mixOmics	Tenenhaus and Tenenhaus (2011)
Sparse regularized generalized canonical correlation analysis	Predictive approach	SGCCA, mixOmics	Tenenhaus et al. (2014)
2k tables			
STATICO	Descriptive approach	ADE4	Simier et al. (1999)
COSTATIS	Descriptive approach	ADE4	Thioulouse (2011)
DO-ACT	Descriptive approach	MExPosition	Abdi et al. (2012), Vivien and Sabatier (2004)
STATIS4	Descriptive approach	ADE4	Sabatier and Vivien (2008)
Multiblock pls	Predictive approach	ADE4, PLS-2.1.0	Westerhuis et al. (1998)
K+1 tables			
k+1 STATIS	Predictive approach	ADE4, MExPosition	Abdi et al. (2012)
Multiblock redundancy Analysis	Predictive approach	ADE4	Bougeard et al. (2011b)

Table 2.1: Overview of the available multiblock methods. Each of the methods is distinguished based on their predictive or descriptive nature. The R packages implementing the specific method is also listed.

Mathematical aspects

Of the multiblock methods presented in Table 2.1, the mathematical details of three particular methods namely canonical correlation analysis, sparse partial least squares, and co-inertia analysis are presented here. The focus is on these particular methods that

are used in this thesis to analyze two-block data matrices of the dimension X ($n \times p$) and Y ($n \times q$).

Canonical correlation analysis

Canonical correlation analysis (CCA) has attracted considerable attention as the importance of integrating multiple data sources has been noticed. Proposed by Hotelling in 1936 (Hotelling, 1936), CCA is a method of correlating linear relationships between two multidimensional variables and is closely related to several methods like multiple regression, discriminant analysis, and principal component analysis (PCA). As such, multiple regression can predict the value of a single dependent variable from a linear function of a set of independent variables, whereas CCA seeks to quantify the strength of the relationship between two sets of independent and dependent variables. CCA resembles discriminant analysis in its ability to determine independent dimensions similar to discriminant functions for each variable set. When compared to these two methods, CCA has the added advantage of handling multiple dependent variables which can be metric or non-metric. PCA attempts to explain the linear relationship among a set of observed variables and an unknown number of variates whereas CCA focuses on the linear relationship between two variates (Gonzalez et al., 2012, Lê Cao et al., 2009). Clearly, CCA is the generalized member of the family of multivariate statistical techniques and the availability of statistical tools or packages contributed to its role as an integration tool for data arising from different system-levels.

Two datasets are represented by matrices X and Y of dimension $n \times p$ and $n \times q$ respectively where 'n' denotes the number of observations, 'p' corresponds to the variables in matrix X and 'q' to the variables in matrix Y. It can be seen as the problem of finding basis vectors for two sets of variables such that the correlation between the projections of the variables onto these basis vectors are mutually maximized. We denote the two basis vectors as $a^1 = (a^1_1, ..., a^1_p)^T$ and $b^1 = (b^1_1, ..., b^1_q)^T$ such that the correlation between the projection of the variables-columns in X and Y-onto these basis vectors are given by the derived linear projections $U_1 = a^1X$ and $V_1 = b^1Y$ which maximize the correlation $\rho = \text{corr}(a^1X, b^1Y)$. Here, the derived linear projections are called the first canonical variates and constrained to be of unit variance. The correlation between the projections of the variables-columns in X and Y- onto the basis vectors is given by a linear combination of X variables

$$U_1 = a_1^1 X_1 + a_2^1 X_2 + ... + a_p^1 X_p, \tag{2.1}$$

and a linear combination of Y variables

$$V_1 = b_1^1 Y_1 + b_2^1 Y_2 + ... + b_q^1 Y_q, \tag{2.2}$$

where a_i^1 and b_j^1 are the canonical coefficients, U_1 and V_1 are the canonical variates. The first canonical correlation is the maximum correlation coefficient between U_1 and V_1, for all U_1 and V_1.

$$\rho = cor(U_1, V_1) = \max_{a_1, b_1} cor(U_1, V_1). \tag{2.3}$$

The subsequent pairs are found by eigenvalues of decreasing magnitudes and orthogonality is guaranteed by the symmetry of the correlation matrices. Thus, we can seek vectors maximizing the same correlation and successive canonical functions can be found as a step-wise problem by estimating pairs of canonical variates based on residual variance from the previous canonical functions.

The mathematical aspects of CCA are shown below but a more detailed treatment of the derivation of the CCA equations is given in Hardoon (2004), Weenink (2003). For the variables included in X and Y, we define the total covariance matrix

$$C = \begin{bmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{bmatrix}$$

as a block matrix where C_{XX} and C_{YY} are the within-group covariance matrices of variables in X and Y, respectively. Correspondingly, $C_{XY} = C_{YX}^T$ denotes the between-group covariance matrix between variables in X and Y.

Employing the definition of the Pearson correlation coefficient, as well as the definition of the covariance matrix C, we can then rewrite the generalized equation (2.3):

$$\rho = \max_{a,b} \frac{a^T C_{XY} b}{\sqrt{a^T C_{XX} a b^T C_{YY} b}} \tag{2.4}$$

Note, that ρ is not affected by re-scaling of the vectors a and b, i.e., the multiplication by the same scalar α does not change the value for ρ in equation (2.4):

$$\frac{a^T C_{XY} b}{\sqrt{a^T C_{XX} a b^T C_{YY} b}} = \frac{\alpha a^T C_{XY} b}{\sqrt{\alpha^2 a^T C_{XX} a b^T C_{YY} b}} \tag{2.5}$$

The optimization problem formulated in equation (2.4) is equivalent to maximizing the numerator subject to the two constraints $a^T C_{XX} a = 1$ and $b^T C_{YY} b = 1$. By incorporating these constraints and writing the Lagrangian form, we obtain:

$$L(\lambda_X, \lambda_Y, a, b) = a^T C_{XY} b - \lambda_X (a^T C_{XX} a - 1) - \lambda_Y (b^T C_{YY} b - 1),$$

where λ_X and λ_Y denote the Lagrange multipliers. Taking the derivatives with respect to a and b by considering that $a^T a = a^2$ and $b^T b = b^2$, respectively, we arrive at:

$$\frac{\partial L}{\partial a} = C_{XY} b - 2\lambda_X C_{XX} a = 0 \tag{2.6}$$

and

$$\frac{\partial L}{\partial b} = C_{YX} a - 2\lambda_Y C_{YY} b = 0 \tag{2.7}$$

Furthermore, by subtraction of equation (2.7) multiplied with b^T from equation (2.6) multiplied with a^T, one obtains:

$$0 = a^T C_{XX} b - a^T 2\lambda_X C_{XX} a - b^T C_{YX} a - b^T 2\lambda_Y C_{YY} b = 2\lambda_Y b^T C_{YY} b - 2\lambda_X a^T C_{XX} a$$

Considering the initial constraints $a^T C_{XX} a = 1$ and $b^T C_{YY} b = 1$, it can be concluded that $2\lambda_Y - 2\lambda_X = 0$ and further define $\lambda = 2\lambda_X = 2\lambda_Y$. In the case of C_{YY} being invertible we get from equation (2.7):

$$b = \frac{C_{YY}^{-1} C_{YX} a}{\lambda} \tag{2.8}$$

and by substitution in equation (2.6):

$$C_{XY}C_{YY}^{-1}C_{YX}a = \lambda^2 C_{XX}a \tag{2.9}$$

Note that equation (2.9) is a generalized eigenproblem of the form $A_x = \lambda B_x$, where $A = C_{XY}C_{YY}^{-1}C_{YX}$, $B = C_{XX}$, $a = x$ and λ corresponding to the squareroot of the eigenvalues. By solving the eigenproblem, we obtain a solution for the canonical correlation ρ as the root of the derived eigenvalues λ^2 and the eigenvector 'a' corresponds to the normalized canonical basis vectors initially defined in equation (2.3). Finally, by inserting 'a' in equation (2.8), we were able to find the corresponding vector 'b'. Alternatively, to the outlined procedure, one can also arrive at the following generalized eigenproblem, directly by combining equations (2.6) and (2.7):

$$\begin{bmatrix} 0 & C_{XY} \\ C_{YX} & 0 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \lambda \begin{bmatrix} C_{XX} & 0 \\ 0 & C_{YY} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$$

where $\lambda = 2\lambda_X = 2\lambda_Y$.

If the eigenvalues are $r_1, r_2, ..., r_m$ where m is the number of canonical correlations, we test the hypothesis that there is no (linear) relationship between the two variable sets. This is equivalent to the statement that none of the correlations $r_1, r_2, ..., r_m$ is significant. Some of the measures for assessing the statistical significance of the found canonical correlations is listed below. The following statistics test the multivariate hypothesis in various ways, and their p-values can be approximated by F distributions.
Wilks' (Wilks, 1932) lambda statistic

$$\Lambda_1 = \prod_{i=1}^{m} (1 - r_i^2) \tag{2.10}$$

is a likelihood-ratio statistic. This statistic is distributed as the Wilks Λ-distribution. Rejection of the null hypothesis is for small values of $\Lambda 1$.
Pillai's (Pillai, 1955) trace for canonical correlations is

$$V^{(m)} = \sum_{i=1}^{m} (1 - r_i^2) \tag{2.11}$$

Thus, the null hypothesis is rejected if this test statistic is large.
The Lawley-Hotelling trace (Lawley, 1938, Hotelling, 1951) is

$$V^{(m)} = \sum_{i=1}^{m} \frac{r_i^2}{(1 - r_i^2)} \tag{2.12}$$

Thus, the null hypothesis is rejected if this test statistic is large.
Roy's (Roy, 1939, Kuhfeld, 1986) largest root is given by

$$\Theta = r_i^2 \tag{2.13}$$

Rejection of the null hypothesis is for large values of Θ.

In this thesis, the application of CCA to biological datasets is described in Chapter 3 wherein CCA is employed for integrating data arising from cotton fiber measurements across two different system-levels. Here, the matrix X is defined by 'p' probes (monoclonal antibodies) whose specificity for a particular polysaccharide is measured for a wide range of cotton lines 'n'. The matrix Y describes the fiber characteristics (set of 'q' measurements) measured on the same 'n' cotton lines. The obtained results serve as an exploratory tool to display associations between the probes and the fiber measurements which would not be obtained by investigating the linear relationship using Pearson correlation. This is an example of classical CCA wherein $n > p + q$. However, when the number of variables is larger than the number of experimental observations, a form of regularization must be considered. Such a regularization in this context was first proposed by Vinod (1976), then developed by Leurgans et al. (1976).

Sparse partial least squares

There are two important statistical problems that commonly arise in regression problems, the first being the selection of a set of important variables among a large number of predictors. The second problem is related to the fact that such variable selection often arises as an ill-posed problem where the sample size is much smaller than the total number of variables or the covariates are highly correlated. Partial least squares (PLS) regression was introduced by Wold in 1966 (Wold, 1966) and has been used as an alternative approach to the ordinary least squares (OLS) regression in ill-conditioned linear regression models. PLS is clearly superior to CCA in its stability property to face collinear matrices and, in addition, has an edge over multiple linear regression, ridge regression or other regression techniques. The PLS regression method aims to describe linear relationships between two

sets of variables, and the prediction factor is achieved by extracting a set of orthogonal factors called latent variables (Boulesteix and Strimmer, 2007). The basic concept in this procedure is that the weights used to determine these linear combinations of the original variables are proportional to the maximum covariance among input and response variables. The noise and the multicollinearity of the original data are removed by compressing the p-dimensional X-space into the H-dimensional latent variable space (Commonly, H<<p, where p is the number of the original variables and H is the number of latent variables). Given the standardized predictor (X) and response variables (Y), there are many methods for extracting the latent variables from X and Y. The algorithm for PLS which uses the singular value decomposition (SVD) is provided below as a follow-up to comprehend the working principle behind sparse partial least squares regression (sPLS) (Boulesteix and Strimmer, 2007). In this algorithm, steps (e) and (f) compute the regression coefficients of the matrices on the latent variables, whereas steps (g) and (h) compute the deflated residual matrices.

Algorithm 1 PLS

1: $X_0 = X$, $Y_0 = Y$

2: For h = 1..H (where H is the number of latent variables):

(a) Set $M_{h\text{-}1} = X_{h\text{-}1}^{T} Y_{h\text{-}1}$

(b) Using SVD, decompose M_{h-1} and extract the first pair of singular vectors $u = u_1$ and $v = v_1$ corresponding to the eigenvalue with the maximum absolute value.

(c) $t_h = \frac{X_{h-1}u}{u'u}$ where t is the latent variable vector of X.

(d) $w_h = \frac{Y_{h-1}v}{v'v}$ where w is the latent variable vector of Y.

(e) $c_h = \frac{X_{h-1}^{T}t_h}{t_h't_h}$ where c is the loading vector of X.

(f) $d_h = \frac{Y_{h-1}^{T}t_h}{t_h'w_h}$ where d is the loading vector of Y.

(g) $X_h = X_{h-1} - t_h c_h$

(h) $Y_h = Y_{h-1} - t_h d_h$

3: Return X_h and Y_h

Although dimension reduction via PCA or PLS is a principled way of dealing with ill-posed problems, it does not automatically lead to the selection of relevant variables. Penalized partial least squares method proposed by Huang et al. (2004) imposes sparsity on the final PLS estimates by using a soft thresholding rule but it does not necessarily lead to sparse linear combinations of the original predictors. This is because the sparsity principle

is imposed on the solution and is not incorporated on the dimension reduction step. To overcome this issue, Chun and Kele (2010) proposed the sparse partial least squares which imposes sparsity on the dimension reduction step of PLS so that the sparsity can play a direct principled role.

Sparse partial least square approach deals with integration problems which cannot be solved with usual feature selection approaches. The main principle of this methodology is to impose sparsity within the context of partial least squares and thereby carry out dimension reduction and variable selection simultaneously. sPLS regression exhibits good performance even when (1) the sample size is much smaller than the total number of variables; and (2) the covariates are highly correlated. One additional advantage of sPLS regression is its ability to handle both univariate and multivariate responses (Lê Cao et al., 2009). sPLS is a combination of two different penalties: the continuous penalty is a LASSO penalty and discrete penalization is achieved by PLS. Variable selection is achieved by LASSO, and dimension reduction by PLS. The respective hyper-parameters i.e. the number of PLS components and the size of the LASSO penalty are optimized simultaneously by cross fold validation. As in normal PLS, each of the latent components is a linear combination of the original variables. The sparse PLS algorithm (sPLS) with its two deflation variants (Lê Cao et al., 2009), based on the iterative PLS algorithm:

Algorithm 2 sPLS

1: $X_0 = X$, $Y_0 = Y$

2: For h = 1..H

(a) Set $M_{h\text{-}1} = X_{h\text{-}1}^T Y_{h\text{-}1}$

(b) Using SVD decompose M_{h-1} and extract the first pair of singular vectors $u = u_1$ and $v = v_1$ corresponding to the eigenvalue with the maximum absolute value.

(c) Until convergence of u_{new} and v_{new} (in the first iteration $u_{old} = u_1$ and $v_{old} = v_1$)

(i) $u_{new} = g\lambda_1(M_{h-1}v_{old})$, re-normalize u_{new}

(ii) $v_{new} = g\lambda_2(M_{h-1}^T u_{old})$, re-normalize v_{new}

(iii) $u_{old} = u_{new}$, $v_{old} = v_{new}$

where $g_\lambda(y) = \text{sign}(y)(|y| - \lambda)_+$ is the soft-thresholding penalty function and λ is the penalty parameter.

(d) $t_h = \frac{X_{h-1}u}{u'u}$ where t is the latent variable vector of X.

(e) $w_h = \frac{Y_{h-1}v}{v'v}$ where w is the latent variable vector of Y.

(f) $c_h = \frac{X_{h-1}^T t_h}{t_h' t_h}$ where c is the loading vector of X.

(g) $d_h = \frac{Y_{h-1}^T t_h}{t_h' w_h}$ where d is the loading vector of Y.

(h) $X_h = X_{h-1} - t_h c_h$

(i) $Y_h = Y_{h-1} - t_h d_h$

3: Return X_h and Y_h

The re-normalization of the weighting vectors 'u' and 'v' in step (b) of the algorithm is very important as the comparatively uninformative elements of the weighting vectors are sucessfully forced to zero. Re-normalization of the variables lead to the re-valuation of their importance and as a result, the contributions of important variables having the larger absolute weighting values can be enhanced. In cases where there is no sparsity constraint ($\lambda_1 = \lambda_2 = 0$), we obtain the same results as in a classical PLS. The penalization parameters λ_1^h and λ_2^h can be simultaneously chosen by computing the root mean squared error prediction (RMSEP) with k-fold or leave-one out cross validation for each given dimension. In both PLS and sPLS, the optimal number 'H' of dimensions has to be determined. The parameter 'H' can be tuned by cross-validation as in the original PLS and as proposed by Chun and Kele (2010). In sPLS regression, in addition to 'H', the number of variables selected in each dimension of the model has to be fixed.

Q_h^2 is computed for validation of the choice of PLS dimension. It is a criteria that measures the marginal contribution of the latent variable to the predictive power of the

PLS model by performing cross-validation computations. Q^2_h is computed as

$$Q^2_h = 1 - \frac{\Sigma_{q=1} PRESS_{qh}}{\Sigma_{q=1} RSS_{q(h-1)}} \tag{2.14}$$

where $PRESS^q_h$ is the PRediction Error Sum of Squares and RSS^q_h is the Residual Sum of Squares for the variable 'q' and the PLS dimension. The value of $Q^2_h \geq (1–0.95^2) = 0.0975$ is said to contribute significantly to the prediction. The approach for evaluation of the predictive power include computing the RMSEP which is done for each variable 'q' in Y.

The application of this method is illustrated in Chapter 3 of this thesis for a regression based study of cotton fibers.

Multiple Co-inertia analysis

Multiple Co-inertia analysis (MCIA) is used to describe the similarities between two or more than two data matrices, also called blocks, observed on the same 'n' samples by recovering the maximum total variance from each matrix. It is a symmetrical method used widely in the field of hyperspectral data analysis and provides orthogonal loadings and score images for spectral and spatial interpretation, respectively. MCIA can be compared to other methods that are used to analyze multiblock data. When compared to the partial least squares approach which focuses on the prediction of one data table based on the subspace generated by the other one, MCIA gives both data tables the same importance. Moreover, in PLS, deflation is made on the global scores whereas MCIA deflates data tables using block scores. Deflation is the process that consists in removing the information described by a component from the initial data table before assessing the next one. MCIA is an extension of co-inertia analysis (CIA) to more than two data tables and has an added advantage in the deflation step. The first step in both MCIA and CIA is the same, whereas the major difference lies in the deflation step. CIA, just like PLS deflates data tables using global scores. Hence, MCIA method being symmetric and orthogonal has an advantage over PLS, CIA, and consensus principal component analysis in its ability to deflate data tables using the block scores.

When we consider X and Y data tables, the method is based on the analysis of the covariance matrix between the whole set of variables of the data tables. In order to do this, a common structure is required to study the covariances between the data tables. The steps involved in MCIA include determining a global component 'c_v' and block components such that the sum of the squared covariances is maximized (Hanafi et al., 2011,

Chessel and Hanafi, 1994).

Usually, the solution to the maximization problem is obtained by performing PCA of the concatenated data tables X and Y. The first global score $c_v^{(1)}$ is the first standardized principal component of the concatenated data tables. The loading $\widetilde{u}^{(1)}$ associated to $c_v^{(1)}$ is then fragmented into subsets $\widetilde{u}^{(1)} = [\widetilde{u_X}^{(1)}, \widetilde{u_Y}^{(1)}]$. Each of the subvector associated to the concatenated data table is normalized to obtain the block loadings

$$\widetilde{u_X}^{(1)} = \frac{\widetilde{u_1}^{(1)}}{\|\widetilde{u_X}^{(1)}\|} \tag{2.15}$$

and

$$\widetilde{u_Y}^{(1)} = \frac{\widetilde{u_1}^{(1)}}{\|\widetilde{u_Y}^{(1)}\|} \tag{2.16}$$

for X and Y data tables, respectively. The second step in MCIA is to calculate the scores and loadings of order higher than 1 by deflating each data table with respect to the block loadings. Deflation of each data tables with respect to its block loadings involves subtracting the information from the initial component before calculating the next component. The original X and Y blocks are updated and this is an iterative procedure where the new block loadings and scores are calculated as mentioned.
The covariant patterns revealed by MCIA are assessed using block loadings, block scores, global loadings and global scores. The common and specific information brought by each data table could be investigated. The relationship between the two data tables are calculated by the RV coefficient. This is a measure of global similarity between the data tables, and is a number between 0 and 1. The closer it is to 1, the greater the global similarity between the two data tables.

Application of multiple co-inertia analysis in integrative analysis is illustrated in Chapter 4. By applying this method to analyze hyperspectral data, we were able to analyze the common structure revealed by two different hyperspectral techniques and also the independent contribution which explains the variability under each block.

Difference between the approaches

The above explained canonical methods are used to determine the common structure between two sets of variables (p and q) but profoundly differ in their construction, and

hence their aims. CCA uses information from all the variables in both the exposure and outcome variable sets and maximizes the estimation of the relationship between the two datasets. Thus, CCA may offer a more efficient approach for assessing the relationship of two sets of variables than methods routinely used such as multiple linear regression. CCA starts with simultaneous consideration of both datasets, limiting the inefficiencies that may accompany conventional multiple testing, and thus, reducing type-1 error. However, canonical correlation based methods are statistically difficult to assess as they do not fit into a regression framework. In this context, penalized CCA adapted with elastic net (CCA-EN) could be used but the elastic net is similar to a Lasso soft-thresholding penalization and the algorithm uses partial least squares and not canonical correlation computations (Lê Cao et al., 2009). On the other hand, sPLS and MCIA aim at maximizing the covariance between the score vectors. From Lê Cao et al. (2009), it is evident that sPLS made a good compromise between all of these approaches and includes variable selection. Moreover, sPLS maximizes the covariance between the latent variables whereas the canonical correlation based methods maximize the correlation. The major advantage of coinertia analysis is its ability to deflate data tables using block scores.

However, in addition to the application of canonical methods for data integration, this thesis includes a novel analysis pipeline for two block data tables of higher complexity. Here, the focus is on joint analysis of gene expression datasets using network based approaches to study the correlation and co-expression patterns of 22000 genes across different cellular levels. Furthermore, Tukey Honest Significant Difference (HSD) tests determines the number of genes which show cell-type specific patterns across different cellular levels and classify them into particular motif occurrences. The benefits of this analysis pipeline are two-fold: using the concept of expression conservation score, it is possible to identify genes which exhibit altered expression levels. Moreover, the analysis pipeline detects co-expression relationships which are reflective of substantial rewiring of co-regulation of genes across different cellular levels. In addition, the pipeline takes into account the gene expression patterns and their specificity across multiple cell-types. Bootstrap procedures ensure the robustness of the analyses.

2.4 Case studies for integrative analysis of two block data tables

In this section, the data structure of the case studies used in this thesis is detailed. Three case studies involving heterogeneous data-types from various plant sources will be

presented. The underlying biological question for each case study will be explained in detail as subsequent chapters.

Case-study 1

This study was based on cotton fibers, one of the commercially important raw materials of the fiber industry. With the sequencing of the *Gossypium hirsutum* genome, studies on cotton fiber is increasingly becoming more important. In this case, two different datasets were generated across different omics levels, the glycome and the phenome. Specifically, the glycome refers to the full set of sugar molecules or the entirety of carbohydrates in a cell (Bertozzi and Sasisekharan, 2009, Campbell and Yarema, 2005). The dataset employed in this study is generated from high throughput screening of plant cell wall polysaccharides from cotton fibers. A key goal of biology is to understand the phenotypic variations in the environment and phenomics is pursued as an independent discipline to enable the development of high throughput phenotyping (Bilder et al., 2009, Houle et al., 2010, Joyce and Palsson, 2006). To this end, the second dataset is generated from a phenotyping experiment in cotton fibers. Experimental information related to these two experiments will be explained in detail in Chapter 3 of the thesis.

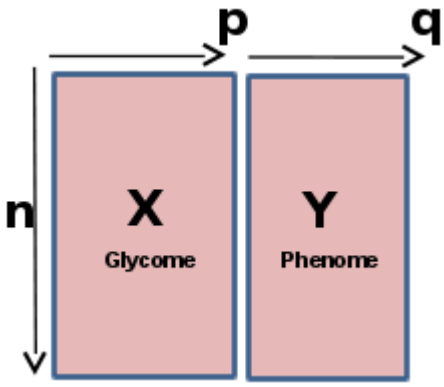


Figure 2.3: Integration two block data tables generated using different experiments. These two blocks of data were generated across different cellular levels from cotton fibers. There are 'n' observations that corresponds to the cotton lines. The variables 'p', and 'q' correspond to antibody probes and phenotypic traits, respectively.

The datasets from the glycan profiling and phenotyping experiments represent a two block data structure and correspond to matrices X and Y respectively (Figure 2.3). The number of observations 'n' in this case refer to the number of cotton lines used. The variables 'p' of the glycan profiling experiment correspond to the antibody probes used to study specific cell wall polysaccharides. Contrastingly, the variables 'q' refer to the measurements of the

cotton fiber characteristics for eg., elongation and strength of the fibers. This case study is different from the others in terms of the cellular levels and also the data structure. The number and type of variables studied in the two datasets are completely different whereas the number of observations still remain the same.

Case-study 2

Spectroscopic approaches is quite extensively applied in monitoring developmental and compositional changes of plant cell walls. In this case-study, we use data from spectroscopy techniques such as Fourier Transformed Infrared (FT-IR), and Raman. Cell wall polymers can be studied by these two types of spectroscopy with different sensitivity depending on the polymer structure. FT-IR spectra can reveal very slight modifications in the polysaccharide composition of the cell wall (Alonso-Simón et al., 2011, Sene et al., 1994, Kačuráková et al., 2000). Raman spectroscopy involves inelastic scattering with a photon and the structural characteristics of plant cell walls in the native state could be studied on the single cell level (Gierlinger et al., 2012).

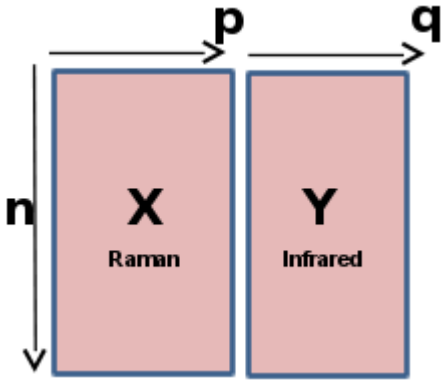


Figure 2.4: Data tables generated from spectroscopy experiments. These two blocks of data were generated to study the chemical composition of particular cell-types using two different techniques. There are 'n' observations, and the variables are represented by 'p', and 'q' across both datasets.

The data from the spectroscopy experiments represent a two way data table wherein the first data block X refers to the Raman dataset and Y to the infrared dataset (Figure 2.4). This study is an example of integrating two block data tables generated from two different techniques used to study the same set of samples. The observations in both datasets correspond to the paired spectral profile (number of common pixels) across both spectroscopy techniques. The variables 'p', and 'q' correspond to the wavenumbers used in the Raman and infrared experiment, respectively. In this case, the number of observations

across both experiments remain the same whereas the nature and number of variables is different.

Case-study 3

This study is on *Arabidopsis thaliana*, a widely used model plant organism in basic research. The availability of its genome sequence has made it an invaluable resource for the identification and characterization of genes encoding enzymes especially in the field of plant cell wall biosynthesis. Arabidopsis has 22800 genes and 14 non-overlapping cell-types. Recently, two microarray datasets of Arabidopsis root samples were generated to study the gene expression of multiple cell-types across two different omics levels. The omics level under study were the transcriptome and the translatome level characterizing cell-type specific gene expression and translation levels respectively. Briefly, the genome is made up of deoxyribonucleicacid (DNA) molecules that contains the information necessary for the construction and maintenance of a cell. Utilization of the biological information in a cell requires the coordinated activity of enzymes and proteins to produce a complex series of biochemical process called as genome expression. The initial product of genome expression is the transcriptome and is maintained by a process called transcription in which individual genes are copied into ribonucleicacid (RNA) molecules (Sharp, 2009). Translation of the individual RNA molecules into aminoacid chains forms proteins and the repertoire of proteins is referred to as the proteome. As an intermediate step between the transcriptome and the proteome is the translatome which describes the RNA which are attached to the ribosomes. In particular, the composition of the translatome is based primarily on translation initiation, i.e. the loading of ribosomes on messenger ribonucleoprotein particles (mRNPs) to form polysomes and secondarily on translation elongation (Gebauer and Hentze, 2004). More detailed information of the experiments pertaining to the generated datasets can be found in the material and methods section of Chapter 5.

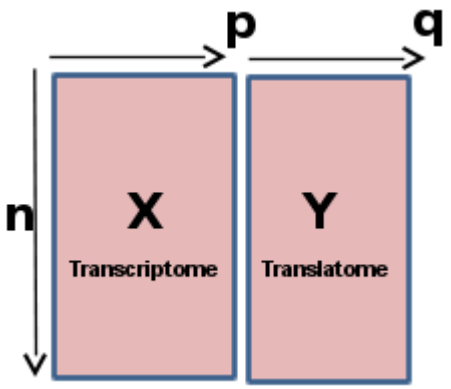


Figure 2.5: Two block data table representation. These two blocks of data were generated across different cellular levels from the same plant system. There are 'n' observations that correspond to the genes. The variables 'p', and 'q' correspond to the cell-types from the transcriptome and translatome level, respectively.

Clearly, the above mentioned datasets represent a block data structure wherein the first block refers to the transcriptome dataset and, second to the translatome dataset (Figure 2.5). The block structures are of the order two if both blocks refer to two-way data matrices wherein 'n' represent the number of observations, 'p', and 'q' represent the number of variables in matrices X and Y, respectively. Here, data matrices X and Y corresponds to the transcriptome and translatome datasets, respectively. In the transcriptome (X) and translatome (Y) data matrices, 'n' refer to the number of genes under study in Arabidopsis roots. The variables 'p' refer to the cell-types in which the gene expression studies were carried out in the transcriptome. The variables 'q' refer to the number of cell-types profiled in the translatome. The number of genes in the transcriptome and translatome dataset were the same albeit the number of profiled cell-types were different.

Statistical computation

In consecutive chapters of this thesis, specific case-based examples of various integrated approaches are elaborated with an aim to understand the architecture of plant cell walls. R version 3.1.2 (R Core Team, 2014) on a 64-bit Linux platform was used for computations in Chapters 3 and 5. MATLAB (MATLAB, 2013) was used for image processing and analysis in Chapter 4 of this thesis.

Chapter 3

Case study 1: Understanding the relationship between cotton fiber properties and non-cellulosic cell wall polysaccharides

3.1 Specific rationale and objectives

A detailed knowledge of cell wall glycans and complexity is crucial for understanding plant growth and development. However, glycans are not readily amenable to sequencing and existing biochemical methods for glycan analysis are usually low throughput. Microarrays are widely used in plant research for the high throughput analysis of nucleotides, proteins, and increasingly, carbohydrates (Schena, 1996, Wang, 2003). Carbohydrate microarrays also referred to as glycan arrays enable hundreds of glycans to be analyzed in parallel. Glycans on the arrays can include oligosaccharides, polysaccharides, glycoproteins, and glycolipids (Park et al., 2008, Wang et al., 2005). Glycan arrays have several biological, and medical applications which include glycoproteomic methods to identify new glycoproteins and glycans (Hanson et al., 2007, Hsu et al., 2007), characterization of glycan probes (Pedersen et al., 2012), profiling carbohydrate-lectin interactions (Uchiyama et al., 2006, Gupta et al., 2010), glycosaminoglycans-growth factor and cytokine interactions (Gama et al., 2006, De Paz et al., 2006), pathogen-induced antibody interaction (Ratner and Seeberger, 2007, Wang et al., 2004), cancer-antibody induced interaction (Huang et al., 2013, Lawrie et al., 2006), carbohydrate-virus interactions (Blixt et al., 2004), quantita-

tive carbohydrate-protein interactions (Liang et al., 2007), and drug discovery (Bryan and Wong, 2004, Disney and Barrett, 2007).

Comprehensive microarray polymer profiling (CoMPP), a microarray based glycan screening method is mostly used for high throughput characterization of plant cell walls. In this technique, generation of microarrays by sequential extraction of cell wall polysaccharides and screening samples against a large number of well-defined cell wall probes such as antibodies, carbohydrate binding proteins, and modules is done (Wang et al., 2005, Park et al., 2008, Pedersen et al., 2012). Despite the availability of glycan arrays from several experiments, computational analysis has mostly been restricted to collection of glycobiology information in databases, motif analysis of glycans, and oligosaccharide structure determination (von der Lieth, 2004, Marchal et al., 2003, Aoki-Kinoshita and Kanehisa, 2006).

One key challenge is to establish links between polysaccharide-rich cell walls, and their phenotypic characteristics. It is of particular interest for some plant material, like cotton fibers, which are of both biological, and industrial importance. To this end, the glycan array technology is used to study cotton fibers, one of the most important raw materials for the textile industry. There are four different domesticated species producing cotton fibers namely *Gossypium hirsutum* ('Upland cotton'), *Gossypium barbadense* ('Pima' or 'Egyptian' cotton), *Gossypium arboreum* ('Tree cotton'), and *Gossypium herbaceum* (Wendel et al., 2009). The development of cotton fibers occurs in four major stages: initiation, elongation, secondary wall synthesis, and maturation. Although much work has already been done on the cotton fiber transcriptome, the key question in cotton fiber research is to link the cell wall profile of different cotton types to the cotton fiber properties for a better understanding of fiber development.

Here, the aim is to study the relation between fiber properties and non-cellulosic polysaccharide composition using correlation and regression based approaches on a diverse set of cotton fibers. The two datasets used in this chapter include a glycan array profile and the physical properties of cotton fibers. Taking advantage of the comprehensive microarray polymer profiling technique (CoMPP), 32 cotton lines from different cotton species were studied. The glycan array was generated by sequential extraction of cell wall polysaccharides from mature cotton fibers and screening samples against eleven extensively characterized cell wall probes. Also, phenotypic characteristics of cotton fibers such as length, strength, elongation, and micronaire were measured. The conducted analysis highlights the usefulness of regression based approaches in establishing a relationship between glycan measurements, and phenotypic traits. In addition, the analysis also identified specific polysaccharides which may play a major role during fiber development for

the final fiber characteristics.

3.2 Materials and methods

The methods described under the subsection 3.2.1 was done by Dr. Jean-Luc Runavot from Bayer CropSciences and the work on glycan microarrays (3.2.2) was carried out by Xiaoyuan Guo in Copenhagen. They provided the datasets for the regression based analysis.

3.2.1 Plant material and evaluation of phenotypic traits

The plant material used in the analysis include 32 different cotton lines of which three are from *Gossypium arboreum*, three from *Gossypium barbadense*, two from *Gossypium herbaceum*, and 24 from *Gossypium hirsutum*. Deatils of the cotton lines, and their corresponding plant introduction number (PI number) from the USDA National plant germplasm system (http://www.ars-grin.gov/npgs/) are listed in Table S1 (Appendix B). Seeds were sown in soil compost and plants were grown at constant conditions in a greenhouse at 26-28 °C during a 16 h photoperiod. Mature cotton fibers were collected by harvesting all fully open bolls from several plants. The impact of boll position, and plant-to-plant variation was minimized by mixing the fiber from all harvested bolls. Two types of analyses were performed on these fibers, the first being the glycan array measurements and the second being fiber characteristics/phenotype measurements. For each line, High Volume Instrument (HVI) and Advanced Fiber Information System (AFIS) measurements were performed on 40 g of mature cotton fiber by CIRAD (France) according to the standard methods ASTM D3818-92 and D5867-05. These measurements were done on six and five replicates for HVI and AFIS, respectively, except for micronaire where only two replicates were performed. Five fiber characteristics which include length from HVI and AFIS, strength, elongation, and micronaire were selected for further analysis due to their importance for textile processing. Length HVI refers to the average fiber length of the longer 50 % of fibers in a given sample. Length AFIS (W) deduces length parameters from individual fiber measurements. Strength of the cotton fiber refers to the force required to break a bundle of fibers 1 tex in size (1 tex equals the weight in grams of 1000 meters of fibers). Elongation of the cotton fibers is the measurement of the elasticity of cotton fibers with a higher number indicating more elasticity. Micronaire is obtained by measuring the resistance of the fibers to airflow and depends on the fiber fineness and degree of maturation.

3.2.2 Comprehensive Microarray Polymer Profiling (CoMPP) of mature cotton fiber cell wall

CoMPP analysis was performed on mature cotton fibers as previously described by Singh et al. (2009) with minor modifications. Mature cotton fiber samples were extracted sequentially in 50 mM cyclohexanediamine tetraacetic acid (CDTA) and 4 M Sodium hydroxide (NaOH) with 1 % (v/v) sodium tetrahydridoborate (NaBH4). These two solvents were used to extract pectins and non-cellulosic polysaccharides, respectively. For each line, 300 µL of solvent was added to 10 mg of sample and incubated with shaking for 2 h. After centrifugation, supernatant from each extraction was printed in four replicates and four dilutions (1:2, 1:6, 1:18 and 1:54 [v/v] dilutions). Cadoxen extraction was omitted because it is mainly used to extract cellulose which we do not aim to analyze in our study. The array was probed with eleven monoclonal antibodies (mAbs) recognizing different carbohydrate epitopes as listed out in Table 3.1.

Probes used in the analysis	Specificity of the probes	Reference
BS-400-2	(1,3)-β-D-glucan (callose)	Meikle et al. (1991)
JIM5	Partially methyl-esterified homogalacturonan (HG)	Willats et al. (2000)
LM19	Un-esterified homogalacturonan (HG)	Verhertbruggen et al. (2009)
JIM13	Arabinogalactan (AGP)	Yates et al. (1996)
JIM20	Extensin glycoproteins	Smallwood et al. (1994)
LM11	Xylan	McCartney et al. (2005)
LM15	XXXG xyloglucans (XG) epitope	Marcus et al. (2008)
LM24	XXLG and XLLG xyloglucan (XG) epitopes	Pedersen et al. (2012)
LM25	XXLG and XLLG xyloglucan (XG) epitopes	Pedersen et al. (2012)
BS-400-4	Mannan	Pettolino et al. (2001)
LM21	Mannan	Marcus et al. (2010)

Table 3.1: List of probes used in the glycan array. The cell wall epitopes used in the glycan array experiment were kindly provided by Prof. Knox's lab, University of Leeds.

The dataset was generated to display the relative intensity of each signal to the maximum signal observed within each antibody detection. CoMPP is a semi-quantitative technique and should not be taken to obtain absolute amounts. Practically speaking, we set the maximum value in the whole dataset as 100 and the other values are divided by this maximum value and multiplied by 100 to obtain numbers comprised between 0 and 100. When the quantification is done, the arrays are manually checked to make sure that there

are clear dots on it and not only background or noise. The negative control is an array incubated with 5 % milk in phosphate buffer saline and probed with secondary antibody and then developed as the others.

3.2.3 Pre-processing of the data

The numerical values from both datasets were of different physical quantities and on different scales of magnitude. Moreover, there is no external knowledge that variables with higher numeric variation should be considered more important. Standardization of the raw data was done by computing Z-scores of the raw data. The Z-scores were calculated for each data point by subtracting the mean and dividing by the standard deviation of all data points.

3.2.4 Linear methods to delineate the relationship between the two datasets

Multiple regression models the relationship between a single scalar response variable and a set of explanatory (or independent) variables. Here, we used multiple regression analysis to model which of the cell wall probes were associated to the fiber characteristics. This allowed us to determine the overall fit (variance explained) of the model and the relative contribution of each of the cell wall probes to the total variance explained. The results from the analysis were reported in the coefficients and ANOVA tables. Summary of the fitted model object gave an account of the residuals, the estimates of the intercept, the slope (with the results of a t-test), the residual standard error, the R^2 statistic and the results of an F-test. Residual standard error is the standard deviation of the data about the regression line. The squared multiple correlation coefficient (R^2) is the proportion of variability in the response that is fitted in the model and the F value is a test statistic to decide whether the model as a whole has statistically significant predictive capability. The statistically significant predictive capability in the presence of other variables is given by the p-values (Schneider et al., 2010, Tabachnick, 2013). Based on this, five models were selected to determine which of the cell wall polysaccharides play an important role in determining that particular fiber characteristic.

In addition to the multiple regression analysis, relationships between multiple dependent and independent variables were investigated simultaneously using canonical correlation analysis (CCA). The two sets of data were represented by matrices X (dimension $n \times p$) and Y (dimension $n \times q$). The columns in X and Y denote the variables 'p' (glycan measurements) and 'q' (fiber characteristics), respectively. Classification of variables

as dependent or independent is of little importance for the statistical estimation of the canonical functions as canonical correlation finds linear combinations of sets of multiple dependent and independent variables which are maximally correlated (Lutz and Eckert, 1993, González et al., 2009).

The first step in CCA was to derive one or more canonical function between the glycan and phenotypic measurements. Each function consisted of a pair of variates, one representing the cell wall probes and the other representing the fiber characteristics. The maximum number of canonical variates (functions) that could be extracted from the sets of variables equals the number of variables in the smallest dataset, independent or dependent. As a result, the first pair of canonical variates was derived so as to have the highest inter-correlation possible between the glycan array and the fiber measurements. Technically, the second pair of canonical variates exhibits the maximum relationship between the two sets of variables (variates) not accounted for by the first pair of variates and successive pairs of canonical variates were based on residual variance. Therefore, each of the pairs of variates is orthogonal and independent of all other variates derived from the same set of data. The strength of the relationship between the pairs of variates obtained from both datasets was determined by the canonical correlation. An estimate of shared variance between the canonical variates was provided by the squared canonical correlations, also called canonical roots or eigenvalues. The statistical significance of each canonical function was assessed as detailed in Chapter 2 using multivariate tests of significance namely Wilk's lambda, Hotelling's trace, Pillai's trace and Roy's greatest characteristic criterion (Roy's gcr). The statistically significant canonical functions were then interpreted using canonical loadings, cross-loadings, and redundancy index (Witten and Tibshirani, 2009, Gonzalez et al., 2012, Rencher, 2002, Tenenhaus et al., 2014).

3.2.5 Sparse partial least square regression to predict the cell wall probes associated to fiber characteristics

Partial least squares (PLS), a well-known regression technique dealing with collinear matrices, clearly has an edge over other regression techniques (Boulesteix and Strimmer, 2007). Unlike CCA, the PLS latent variables are linear combinations of the variables based on the maximization of covariance but do not allow feature selection. There are many variants of PLS of which we focused on a sparse partial least squares approach (sPLS) which includes a built-in feature to select variables while integrating the data. A detailed algorithm for both the PLS and sPLS is available in Chapter 2 of this thesis. Specifically, we use a two block data setup, X be the $n \times p$ matrix and Y be the $n \times q$ ma-

trix where 'n' denotes the samples, variables 'p' and 'q' denote the glycan measurements and fiber characteristics, respectively. Sparse PLS, based on Lasso regression penalizes the loading vectors using singular value decomposition (SVD) and has an additional advantage to perform better even when the covariates are highly correlated. We used sPLS in the regression mode where the aim was to model the relationship between the variables and also predict one group of variables from the other (Gonzalez et al., 2012, Lê Cao et al., 2009, 2008).

3.3 Results

3.3.1 Standardization of the raw data

The analysis was done to assess the relationship between the cell wall polysaccharides and the physical fiber properties of mature cotton fibers. The glycan array values used for the regression analysis were the sum of the CDTA and the NaOH extractions as performing the analysis using the individual values gave the same correlations. For the fiber characteristics dataset, the values were in different units and scales such as mm (for length), g/tex (for strength), and percentage (for elongation). To make the fiber characteristics dataset compliant to the glycan array, the raw data were jointly standardized using Z-scores prior to the analysis.

3.3.2 Modeling the fiber properties using linear regression models

We investigated the linear relationship between the fiber properties and their corresponding array values by a series of regression analyses. Multiple regression models were built considering one fiber characteristic at a time as the dependent variable and multiple probes as the independent variables. Five such models were predicted for the phenotypic traits and the overall model prediction result (Table 3.2) shows that the model for length HVI, length AFIS and micronaire are statistically significant. The significant predictor variables of length HVI are BS-400-2, LM19 and the ones for length AFIS include BS-400-2, JIM5, JIM20, LM15, and LM19.

Fiber characteristics	Residual standard error	Multiple R-squared	Adjusted R-squared	F-statistic	p-value	Significant predictors
Length HVI	0.696	0.706	0.545	4.372 on 11 and 20 DF	0.002	BS-400-2, LM19
Length AFIS	0.632	0.720	0.566	4.677 on 11 and 20 DF	0.001	BS-400-2, JIM5, JIM20, LM15, LM19
Strength	0.940	0.378	0.036	1.107 on 11 and 20 DF	0.404	JIM20
Elongation	0.825	0.376	0.033	1.098 on 11 and 20 DF	0.410	-
Micronaire	0.469	0.851	0.769	10.400 on 11 and 20 DF	4.90E-006	LM15, LM19, LM24, LM25

Table 3.2: Summary statistics of the five possible multiple regression models. The F-statistic and p-value criterion were further used to assess the significant predictors.

Probes LM15, LM19, LM24 and LM25 are the significant predictor variables for the model predicting cotton fiber micronaire and the overall model has a p-value of 4.90E-006. The models for strength and elongation do not show any statistical significance.

3.3.3 Simultaneous assessment of the relationship between multiple probes and all of the fiber characteristics

The multiple regression analysis can predict the value of a single (metric) dependent variable from a linear function of a set of independent variables. However, to explore the relationship of sets of multiple predictor variables (probe measurements) to sets of multiple response variables (phenotypic traits) CCA was used (cf. Chapter 2). For the CCA analysis, the glycan array measurements (probed by eleven antibodies) are designated as the set of independent variables. The fiber characteristics namely length AFIS, length HVI, strength, elongation and micronaire were specified as the set of dependent variables (Figure 3.1). However, it is of little importance to classify the variables as independent or dependent as the technique aims to maximize the correlation between the two sets of variables. In Figure 3.1, the terms r_{x1} to r_{x11} represent the canonical loadings which reflect the variance that the eleven variables from the glycan array shares with the independent canonical variate U_1. Similarly the terms r_{y1} to r_{y5} represent the canonical loadings which reflect the variance that the five phenotypic variables share with the dependent canonical variate V_1.

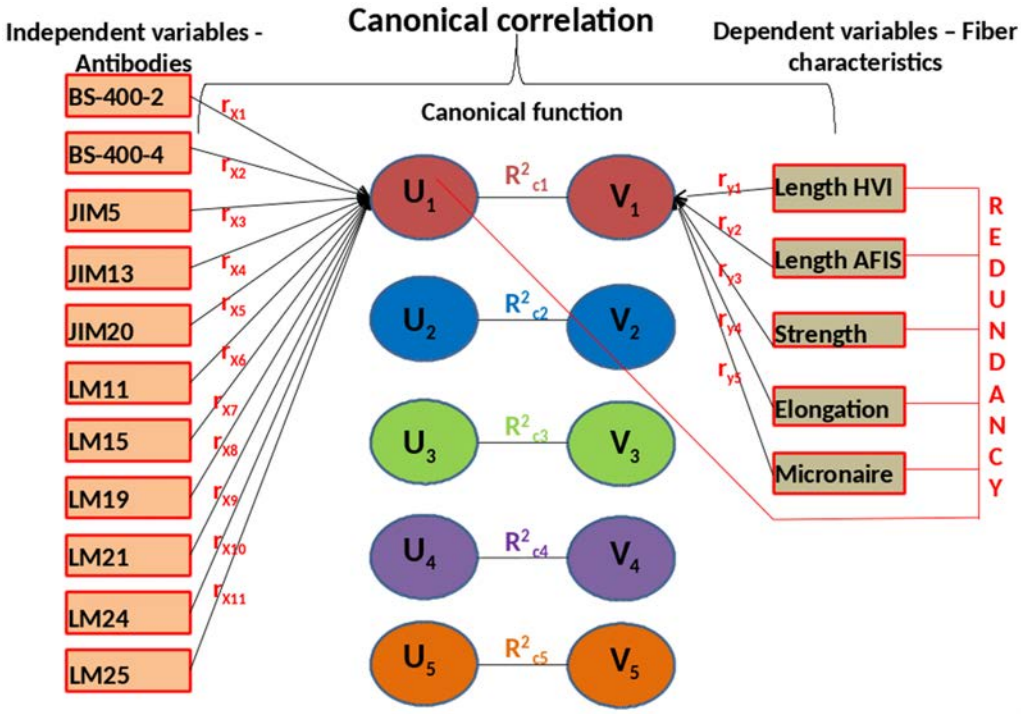


Figure 3.1: Canonical correlation analysis maximizes the correlation between the linear combination of the cell wall polysaccharides in the glycan array and the fiber properties. In this figure, given a linear combination of X variables and a linear combination of Y variables, the first canonical correlation is the maximum correlation coefficient between U_1 and V_1, for all U_1 and V_1.

The canonical correlation between the independent and dependent canonical variates is measured by the canonical functions which are represented by R^2_{c1} to R^2_{c5}. The statistical problem involved identifying any latent relationships (relationships between composites of variables rather than the individual variables themselves) between the glycan and the fiber measurements.

The canonical correlation which is based on the linear relationship of the glycan array data and fiber characteristics was computed to derive five canonical functions (Table 3.3). Each of these functions consists of a pair of variates, one for the glycan array data and the other for the fiber characteristics. Since the study includes eleven independent variables and 5 dependent variables, the maximum number of canonical functions which could be derived is five.

Canonical function	Canonical correlation	Canonical R^2	F statistics	p-value
1	0.945	0.883	2.850	1.573E-005
2	0.868	0.753	1.883	0.011
3	0.803	0.645	1.321	0.203
4	0.523	0.273	0.590	0.871
5	0.342	0.116	0.344	0.904

Table 3.3: Canonical correlation analysis relating probe signals and fiber characteristics with the measure of overall model fit. There are five canonical functions because the maximum number of canonical functions that could be extracted equals the number of variables in the smallest dataset.

In addition to tests of each canonical function separately, multivariate tests of these five functions simultaneously were also performed. The test statistics employed include Wilks' lambda, Pillai's criterion, Hotelling's trace, and Roy's gcr (cf. Chapter 2). Table 3.4 details the p-values from the multivariate test statistics, which all indicate that only the first canonical function, taken collectively, is statistically significant at 1 % level.

Canonical function	Wilks' Lambda	Hotelling-Lawley Trace	Pillai-Bartlett Trace	Roy's largest root
1	1.573E-005	2.666E-007	0	8.732E-012
2	0.011	0.001	0.042	
3	0.203	0.055	0.285	
4	0.871	0.836	0.801	
5	0.904	0.885	0.848	

Table 3.4: Multivariate tests of significance for the canonical functions. Four different test statistics were employed to indicate if the canonical functions are significant. Test statistics were computed as detailed in Section 2.3 (cf. Chapter 2).

From the results of these tests, we proceeded to interpret other aspects of the analysis based on the first canonical function. A redundancy index was calculated for the independent and dependent variates of the first function in Table 3.5. The redundancy index is calculated as the average loading squared times the canonical R^2. As can be seen, the redundancy index for the dependent (0.191) and independent variates (0.200) is quite low. The low values result from the relatively low shared variance in the dependent variates (0.214) and independent variates (0.225), not the canonical R^2. With such a small percentage, this is an example of a statistically significant canonical function that does not have practical significance because it does not explain a large proportion of the dependent variables' variance.

Parameters	Their own canonical variates			The opposite canonical variates	
	Percentage	Cumulative percentage	Canonical R^2	Percentage	Cumulative percentage
Standardized variance of the dependent variables explained by:	0.214	0.214	0.883	0.191	0.191
Standardized variance of the independent variables explained by:	0.225	0.225	0.883	0.200	0.200

Table 3.5: Redundancy analysis of dependent and independent variates for the first canonical function. Redundancy is the measure of how well the independent canonical variate predicts the values of the original dependent variables or vice versa. Standardized variance of the dependent and independent variables explained by their own canonical variates and their opposite canonical variates is listed here.

The interpretations involve examining the canonical functions to determine the relative importance of each of the original variables in deriving the canonical relationships (Table 3.6). The three methods for interpretation are (1) canonical weights (standardized coefficients), (2) canonical loadings (structure correlations), and (3) canonical cross-loadings.

Table 3.6 contains the standardized canonical weights for each canonical variate for both dependent and independent variables. As mentioned earlier, the magnitude of the weights represent their relative contribution to the variate. Based on the size of the weights, the order of contribution of independent variables to the first variate is LM19, LM25, JIM5, LM15, BS-400-4, LM21, LM24, JIM13, and JIM20. Similarly, the order of contribution of dependent variables to the first variate is micronaire followed by length AFIS, length HVI, strength and elongation. Because canonical weights are typically unstable, particularly in instances of multicollinearity, owing to their calculation solely to optimize the canonical correlation, the canonical loadings, and cross-loadings are considered more appropriate.

Table 3.6 also contains the canonical loadings for the dependent and independent variates for the first canonical functions. In the first dependent variates, all the five variables had different values of loadings resulting in low shared variance (0.214). This indicates a low degree of inter-correlation among the five dependent variables. Observing the independent variates, there is a different pattern and loading values ranged from 0.06 to 0.77. The variables with the highest loadings on the independent variate are LM25, LM19, LM15, and JIM5. We also observed some loadings with negative values which

include those of BS-400-4, JIM20, and LM11.

In case of the cross loadings, micronaire has a value of 0.890 and interestingly has a negative loading. Length AFIS to some extent has a loading value of 0.387 while those of the other variables is low. By squaring these terms, we find the percentage of the variance for each of the variables explained by function 1. The results show that 79.21 % of the variance in micronaire, 14.97 % of the variance in length AFIS is explained by function 1 whereas strength, elongation and length HVI have very low values. Similarly for the independent variables' cross loadings, variables LM25, LM19, LM15, JIM5 have high correlations of 0.73, 0.67, 0.61, and 0.61, respectively. From this information, approximately 51.8 % of the variance in LM25, 45.1% of the variance in LM19, 36.3% of the variance in LM15, and 35.7% of the variance in JIM5 is explained by the dependent canonical variates.

	Canonical weights	Canonical loadings	Canonical cross-loadings
Dependent variables			
Length HVI	-0.636	0.127	0.120
Strength	-0.226	0.040	0.038
Elongation	-0.033	0.056	0.053
Micronaire	-0.843	-0.941	-0.890
Length AFIS	0.810	0.409	0.387
Independent variables			
BS-400-2	0.184	0.362	0.342
BS-400-4	-0.487	-0.119	-0.113
JIM5	-0.823	0.632	0.598
JIM13	-0.290	0.288	0.272
JIM20	-0.204	-0.376	-0.355
LM11	-0.114	-0.362	-0.343
LM15	-0.719	0.638	0.603
LM19	1.243	0.712	0.672
LM21	0.356	0.275	0.265
LM24	-0.324	0.066	0.062
LM25	1.082	0.767	0.724

Table 3.6: Canonical weights, loadings, and cross-loadings for the first canonical function. Larger canonical weights contribute more to the canonical function. Canonical loading provides a direct assessment of each variable's contribution to it's respective canonical variate. Canonical cross-loading corresponds to the correlation of each observed dependent or independent variable with its opposite canonical variate.

The final step of interpretation is examining the signs of the cross-loadings. Examining the signs of the independent variables' cross loadings, those with high correlations have a positive direct relationship whereas BS-400-4, JIM20, and LM11 have an inverse relationship. The four highest cross-loadings of the first independent variate correspond to

the variables with the highest canonical loadings as well. Observing the cross loadings of the dependent variables, we see that micronaire has the highest canonical loading and an inverse relationship. Also, elongation is observed to have an inverse relation but since it is of very low value, it was not taken into account.

3.3.4 sPLS approach to predict specific cell wall polysaccharides involved in fiber properties

sPLS was computed in the regression mode and the input for the analysis included the eleven cell wall probes along with the five fiber characteristics. The number of dimensions to be retained was estimated with the Q^2_h criterion, for which a value below the threshold 0.0975 indicates a significant contribution for the prediction purpose. The Q^2_h values calculated as described in Chapter 2 showed that 2 dimensions were enough to capture the whole information. From Figure 3.2, we can interpret the results from the sPLS via the correlation circle plot where the predictor variables are in red and the response variables are represented in blue. A correlation circle plot is a graphic tool to represent variables of two different data-types and examine the relationships between the variables and variates. The relationship between these two data-types is approximated by the inner product between the associated vectors which is defined as the product of the two vector lengths and their cosine angle. For better interpretation, two circles of radii 0.5 and 1 are represented to visualize the variables. The longer the distance to the origin, the stronger is the relationship between the variables.

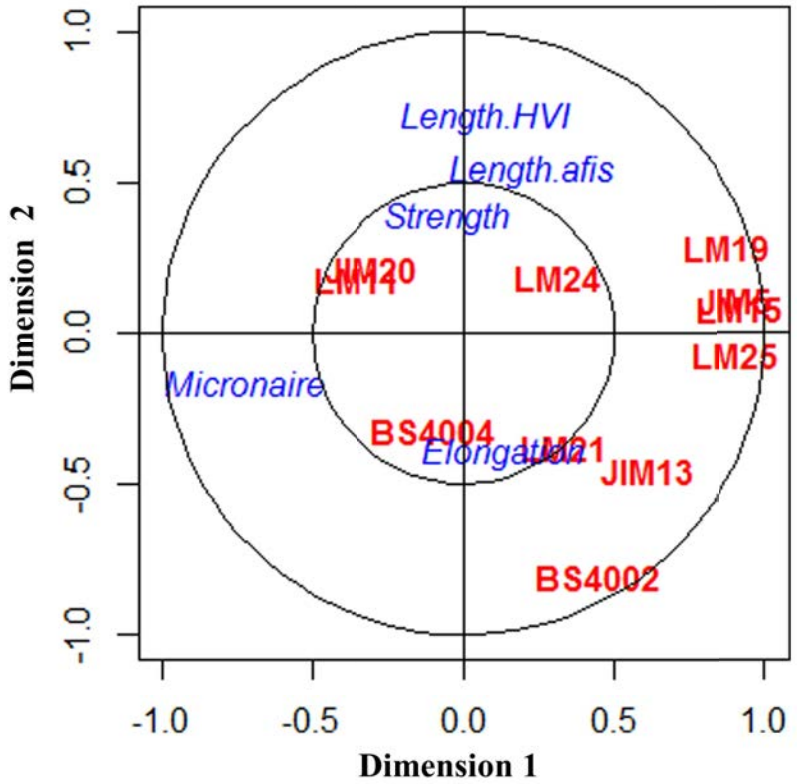


Figure 3.2: Graphical representation of the variables selected by sPLS on the first two dimensions predicts specific cell wall polysaccharides linked to the fiber properties. The coordinates of each variable are obtained by computing the correlation between the latent variable vectors and the original dataset. The selected variables are then projected onto correlation circles where highly correlated variables cluster together. These graphics help to identify association between the two datasets. The correlation between two variables is positive if the angle is sharp ($\cos(\alpha) > 0$), negative if the angle is obtuse ($\cos(\Theta) < 0$), and null if the vectors are perpendicular ($\cos(\beta) \sim 0$).

Using the interpretation which is detailed, we find that BS-400-4, LM21, and JIM13 share a positive relationship with elongation characteristic of cotton fibers. We were also able to attribute the strength of the cotton fibers to JIM20, LM11, and LM24. Interestingly, LM19, JIM5, LM15, and LM25 were projected diametrically opposite to that of the micronaire in the correlation circle, thereby indicating a strongly negative relationship. Length HVI and length AFIS share a negative relation to BS-400-2. To estimate the significance of the predicted relationships, the root mean squared error prediction (RMSEP) values were computed for each response variable (fiber properties) and ranked according to the absolute value of their loadings. The lower the RMSEP value, the better the prediction of the model is. In this case, the model for micronaire was the best one (RMSEP of 0.71), followed by that of length AFIS (1.13), strength (1.14), elongation (1.15), and length HVI (1.21).

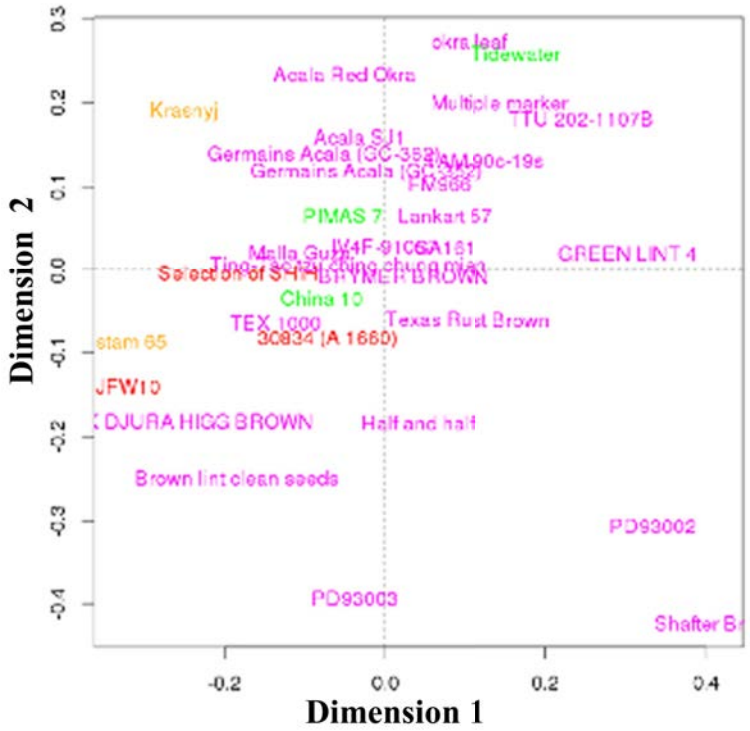


Figure 3.3: Graphical representation of the cotton lines on the first two sPLS dimensions shows the trend in clustering of specific cotton lines across different species. Four different species of cotton are shown in different colors. *Gossypium hirsutum* is colored in magenta, *Gossypium barbadense* in green, *Gossypium herbaceum* in orange and *Gossypium arboreum* in red.

Figure 3.3 displays the graphical representation of the cotton lines in dimension 1 and 2. This plot shows that some of the lines are clustered together, with Acala SJ1, Germains Acala (GC 352 and GC 362), TAM-90C-19 S, and FM966 forming one cluster, Acala red okra, okra leaf, multiple marker, Tidewater, and TTU 202-1107B forming a second cluster and PIMAS7, Lankart 57, IV4F-91057, GA161, Ting tao tzu ching chung mien, Brymer brown, Malla guza, Selection of SHIH, China 10, Texas rust brown, Tex 1000 and 30834 (A1660) forming a third cluster.

Interestingly, some of these clusters indicate that the variation in fiber characteristics and composition is clearly not species-specific. However, one should be careful in interpreting the results from the individual lines as the study was designed to discover correlations between fiber properties and composition and not to study properties of individual lines.

3.4 Discussion

Understanding the genetics and physiology of cotton fibers is of importance to the textile industry. There have been numerous studies, both profiling and sequencing based experiments to study cotton fiber development at the transcriptional level. The high degree of transcriptional complexity in the development of cotton fibers has been the focus of these studies (Singh et al., 2009, Gilbert et al., 2013, Bowman et al., 2013, Rambani et al., 2014, Lacape et al., 2012). We used the CoMPP technique in our analysis to study directly the glycan composition of cotton lines from different species. The work presented here demonstrates the potential of glycan microarrays in combination with multivariate statistical approaches for understanding the cell wall composition responsible for the fiber characteristics. Specifically, the use of regression based approaches in our study helps to predict models for each of the fiber trait under study.

We studied the association between glycan array measurements and their relation with fiber characteristics using linear approaches like multiple regression, CCA and sPLS. From the results of multiple regression (Table 3.2), we were able to predict three models for length HVI, length AFIS and micronaire of cotton fibers but not for strength and elongation characteristics. Moreover, to extend our understanding of the data to situations involving more than one fiber characteristic at a time, CCA was used as it simultaneously models effects of multiple independent variables on multiple dependent variables. CCA uses information from all the variables in both the exposure and outcome variable sets and maximizes the estimation of the relationship between the two sets. The resulting procedure gives a global view of association between indicators of both datasets. Thus, CCA could be used as a comprehensive approach to extract information from data simultaneously. Another major advantage of using the CCA to multiple regression analysis is to deal with the issue of multicollinearity. In multiple regression, the interpretation is usually based on the significance of weights, which is highly influenced by multicollinearity. If two variables have a high correlation one of them will be completely eliminated even if both have a high correlation to the outcome. In our analysis, this is illustrated by JIM5 and LM19 (both detecting homogalacturonan), with both showing a high correlation with micronaire in CCA but only LM19 being identified as a predictor of micronaire in the linear regression model. From the results of the CCA, we obtained an overall picture of associations between the glycan and phenotype measurements, with information about the relative contribution of the variables to that particular canonical variate through canonical loadings. The canonical analysis revealed that the canonical correlation was statistically significant at 1%. Additionally, we used the sPLS approach to be able

to predict specific cell wall polysaccharides linked with fiber characteristics.

There were both unique and common findings from the three types of regression analysis. The major and most significant finding in common to all these analyses is that micronaire is negatively correlated with the xyloglucan (XG) and homogalacturonan (HG) probes. One possible explanation for this observation is that cotton fiber with a high micronaire usually has a very thick secondary cell wall resulting in very high levels of cellulose and lower levels of the non-cellulosic components. However, we do not find a negative correlation of micronaire with other non-cellulosic compounds suggesting that increased cellulose levels of high micronaire fibers affect the XG and HG epitopes in a different way than the other non-cellulosic epitopes. For instance, it could specifically decrease extractability of the XG and HG epitopes. As micronaire measures a combination of fiber fineness and maturity, we wanted to understand whether the observed correlation is with maturity or fineness or a combination of both. We tested this using linear regression models once again and built models for fineness and maturity of the fibers. We observed that the regression models for fineness had an adjusted R^2 value of 0.803 with JIM5, LM19, and LM25 as significant predictors at a 1 % threshold. The regression model for maturity was also significant at the 1 % threshold but with no particular significant predictors thereby suggesting that the observed correlation is attributed to fiber fineness. This indicates that this correlation is linked to the thickness rather than the shape of the fiber, which is consistent with a link to the cellulose levels.

Since only the first canonical function of the CCA analysis is statistically significant and this function explains only for micronaire a large fraction of the variance, the results of the CCA analysis are not informative with respect to the other fiber properties. For these fiber properties, the correlation between fiber length and callose is the only one that was detected in both the linear regression and the sPLS analysis. Callose has been described to play a role in cotton fiber elongation. Indeed, it was reported that plasmodesmatal closure was positively correlated with the rapid fiber elongation and that callose was involved in the gating of these plasmodesmata (Ruan et al., 2004). However, this observation involves transient callose detection, only after 5 dpa and already significantly reduced at 20 dpa, what makes it unlikely to be detected in mature fibers. Another type of callose deposition was reported by Salnikov et al. (2003) wherein the callose is supposed to be deposited in the secondary cell wall and remains in the fiber. From the results of the multiple regression models (Table 3.2), a positive correlation between several of the homogalacturonan probes and length property of the fibers is apparent. The link between pectins and the elongation of cell walls is already observed in several plant systems (Goldberg et al., 1996) and studies in flax stems, pea stems and maize coleoptiles revealed a negative correlation

between pectin levels and cell elongation. In cotton fibers and trichomes, there exists a positive correlation between pectic sheath and elongation (Vaughn and Turley, 1999). Recent studies by Tokumoto et al. (2002) have established that pectic polysaccharides and xyloglucan containing uronic acids were the major polysaccharides extracted during elongation. Hence, our results are in agreement with various studies which state that pectin biosynthesis promotes fiber elongation (Haigler et al., 2012) and that the degree of esterification is a key factor in controlling the elongation (Singh et al., 2009, Wang et al., 2010a). The correlation between length and HG was not detected in the sPLS analysis most likely because the stronger (negative) correlation of HG with micronaire.

Furthermore, relationships between fiber strength or elongation and specific carbohydrate epitopes could be deduced from the results of the sPLS analysis (Figure 3.2). For instance, fiber strength was associated both with the xylan (LM11) and the extensin (JIM20) epitope. A role of xylan in fiber strength would be consistent with the function of heteroxylan in other cell-types which is commonly related to the strengthening of cell walls as revealed by defects in cellulose deposition in xylan mutants (Hao and Mohnen, 2014). A role of extensin in fiber strength is less expected and would need experimental validation. In the linear regression analysis, extensin was identified as a significant predictor for length AFIS but not for length HVI. A role for extensin in determining cotton fiber length would be more consistent with its role in other plant cell-types (Sadava et al., 1973). Finally, AGP glycan (JIM13) and mannan (BS-400-4 and LM21) epitopes were found to predict cotton fiber elongation from the sPLS model. Interestingly, studies have indicated that AGPs are important players during fiber development. Immunofluorescence assays by JIM 13 showed distinct patterns in developing fiber cells indicating that polysaccharide chains of AGPs are involved in initiation and elongation stages of cotton fibers (Bowling et al., 2011, Huang et al., 2013, Qin et al., 2013). However, it is not clear how these AGPs would affect the elongation property of the mature fiber. These unexpected correlations thus present interesting hypotheses for further structure-function relationship studies of the cotton fiber.

Overall, CoMPP assays of cell wall polysaccharides from cotton fibers suggest that it will be a powerful tool in detecting and quantifying the differences between large sets of cotton lines. With the use of predictive statistical approaches to integrate different kinds of datasets, this analysis has thus discovered some correlations that are in line with already known biological functions and others for which the biological relevance still has to be tested. Also, it confirmed the relevance of this type of analysis to enable a detailed understanding of the data from CoMPP assays of cell wall polysaccharides. However, the use of mature cotton fibers in this analysis only allows detecting relevant correla-

tions for components that are still present at maturity. In addition, many changes in polysaccharide composition occur between the fiber elongation stage and maturity. One would thus expect to identify only a fraction of the relationships between polysaccharide composition and fiber properties by analysis of mature fibers, especially for fiber properties such as length that are determined in the early stages of development. Hence it would be interesting to perform a similar kind of analysis using the polysaccharide composition of developing fibers to see whether additional relationships with fiber properties can be determined. The panel of cotton lines used in this study was selected to have maximal diversity in fiber properties and composition. Applying this type of analysis to commercially important cotton lines would allow to understand whether differences in polysaccharide composition affect properties of commercial cotton in the same way as observed in this study and to gain insight into the developmental polysaccharides that are essential to obtain high quality cotton fibers. With the sequencing of the G. hirsutum genome, cotton fiber research is an exciting field and the work presented here will provide a base for future studies, with potential to translate this study on the developing fibers.

Chapter 4

Case study 2: Understanding cell wall chemical composition by integrative analysis of infrared and Raman images from maize cross-sections

4.1 Specific rationale and objectives

Plant cell walls constitute the single largest source of renewable biomass in plants and play an increasingly important role in our energy and industrial future. The plant cell wall is mainly composed of cellulose, hemicellulose, pectins, including a wide variety of non-polysaccharide components like proteins, lipids, enzymes, and other aromatic compounds. The polysaccharides can be broken down to sugar monomers (saccharification) for potential conversion into biofuels, bioplastics, and other chemicals. However, the current biochemical conversion rate into biofuels is far below from first generation feedstocks such as corn and sugarcane. Although the reasons for biomass recalcitrance to biofuels remains yet to be fully elucidated, a better understanding of cell-type specific cell wall polysaccharide composition would greatly improve the saccharification process. In addition, this would provide information on the limits of cell wall degradation prior to alcoholic fermentation (Hansen et al., 2011, Jung and Casler, 2006).

Most traditional chemical analysis of cell walls require the disintegration of plant tissues, use of enzymes or chemicals to study the polysaccharides, and isolation from single layers or smaller regions of interest which is quite tedious. Hence, traditional methods of studying plant cell walls are destructive and much of the chemical and structural information is lost. Advances in microscopy tackle this problem and spectroscopic approaches is extensively applied in monitoring developmental and compositional changes of plant cell walls (McCann et al., 2001, Purbasha et al., 2009). Investigating the variability of cell wall composition according to cell-types requires techniques to determine the chemical composition at a maximum scale of a few micrometers. This can be achieved using microspectroscopy such as Fourier-transform infrared (FT-IR) or Raman. Cell wall polymers can be studied by these different techniques at varying sensitivity depending on the polymer structure. FT-IR spectra can reveal very slight modifications in the polysaccharide composition and has many applications in the study of cell wall architecture and plant development in general (Černá et al., 2003, Alonso-Simón et al., 2011, Hori and Sugiyama, 2003). Raman spectroscopy involves inelastic scattering with a photon from a laser light source in contrast to infrared spectroscopy which involves photon absorption. The structural characteristics of plant cell walls in the native state could be studied on the single cell level using Raman microspectroscopy (Gierlinger, 2014, Gierlinger et al., 2012). Details of 1×1 μm^2 can be revealed by Raman spectroscopy using a conventional source of light, and a pixel size around 5×5 μm^2 can be obtained by FT-IR microspectroscopy using a synchrotron source. Although both methods are based on discrete vibrational transitions and have been developed as important tools in plant cell wall research, they reveal complementary information in the compositional analysis of cell walls. Previous studies have focused on the use of one particular type of spectroscopy to reveal visualized differences in composition between and within cell wall layers. Besides providing both spectral and spatial information, chemical imaging method of vibrational spectroscopy creates a dataset with a huge amount of data. Methods, such as principal component analysis (PCA), independent component analysis (ICA), and correspondence analysis (COA) are the exploratory data analysis approaches that are usually applied to analyze the key information from the spectra.

Owing to the the complexity of cell wall polymers, it is expected that coupling several spectral domains will heighten differences in composition not observable using only one spectral range. To this end, we compare the cell wall composition of different cell-types by taking into account the information provided by the use of different hyperspectral imaging techniques. Here, the focus is to understand the biochemical composition of plant cell walls in maize stem cells using both Raman and infrared spectroscopy. Maize, a

plant model and a major source of biofuel production is used to understand the cell wall composition variability in different cell-types. In practice, each technique operates at its own spatial resolution, making the coupling of the different spectral domains not straightforward. Multiblock methods can be employed to jointly analyze the two data tables. However, the data to be analyzed was acquired from two different techniques and has to be appropriately pre-processed and normalized. We analyze the common and independent information from these hyperspectral images using multiple co-inertia analysis (MCIA). MCIA is used instead of the canonical correlation analysis (CCA) because the former maximizes the covariance and are efficient in determining main individual effects in paired dataset analysis. In contrast, CCA maximizes the correlation between datasets and tends to discover effects present in both datasets, but may omit to discover strong individual effects.

4.2 Materials and methods

4.2.1 Plant material: Maize stem cross-sections

The hyperspectral images used in this chapter were kindly provided by INRA, Nantes. Briefly, maize lines (F2 generation) were grown at INRA Lusignan and stems were harvested at female flowering stage.

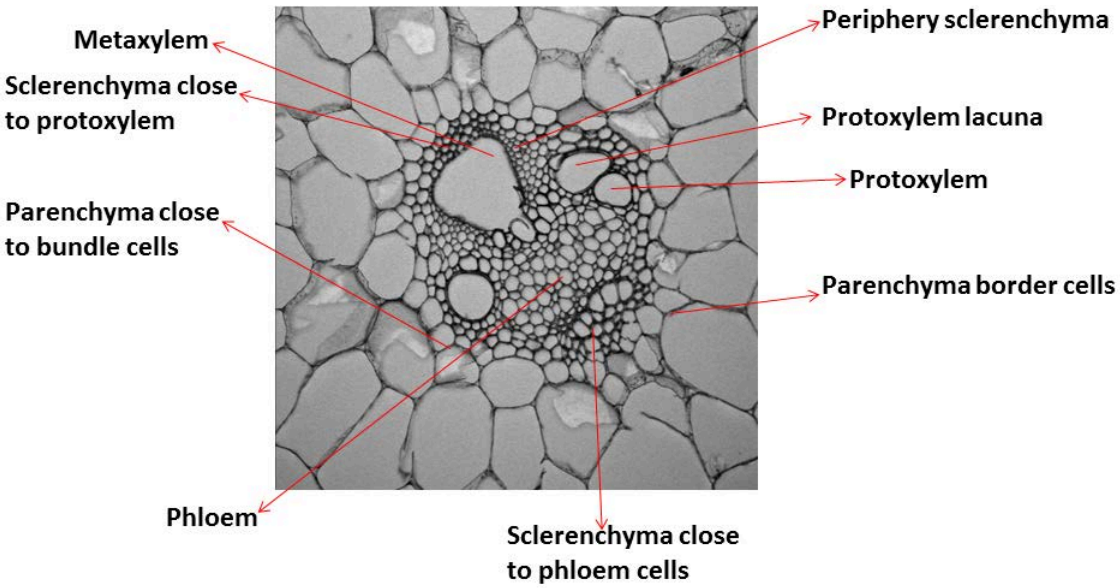


Figure 4.1: Available cell-types in maize stem cell cross-section. *Zea mays* (Maize), an important monocotyledonous crop resemble other grasses in the arrangement of tissues in the stem, leaf, and root.

The internodes under the ear were collected and stored in 70 % (v/v) ethanol/water. Sections of 10 µm from the middle of maize internodes were used as samples and embedded into paraffin. Sectioning was done using a microtome. Before spectral acquisition, starch was eliminated using alpha-amylase enzymes, paraffin was removed using the protocol described in Jamme et al. (2008), and proteins using a protease enzyme subtilisin A type VIII from *Bacillus licheniformis*. The different cell-types in maize are shown in Figure 4.1 of which spectral data from xylem, phloem, sclerenchyma, and parenchyma cells was acquired using FT-IR and Raman spectroscopy.

4.2.2 Infrared and Raman hyperspectral imaging

Infrared images were obtained using the FT-IR spectrometer Tensor 27 (Bruker optics) equipped with an Hyperion 2000 microscope. Infrared spectra were collected in the range of $1800 - 700\,\mathrm{cm}^{-1}$ using the X15 lens. 700 scans were co-added for the background and 500 scans for the sample. An infrared image consisting in 10 points per line and 25 points per column was acquired.

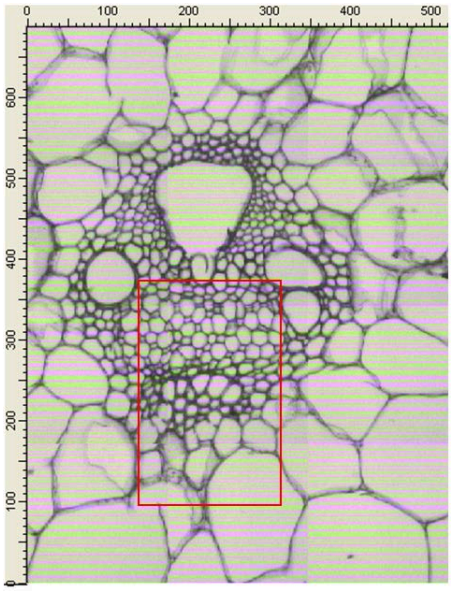


Figure 4.2: Infrared image of the maize stem cells cross-section. The mapped regions include xylem, phloem, sclerenchyma, and parenchyma cells highlighted by the red rectangle.

For each spectrum, the size of the infrared window was adapted to obtain a good signal/noise ratio. The aperture was at least 20×20 µm^2. The infrared images were obtained first (Figure 4.2) and the Raman images second (Figure 4.3) because the latter

has the possibility to destroy the samples during acquisition. Infrared image of size 10×25 μm^2 was recorded.

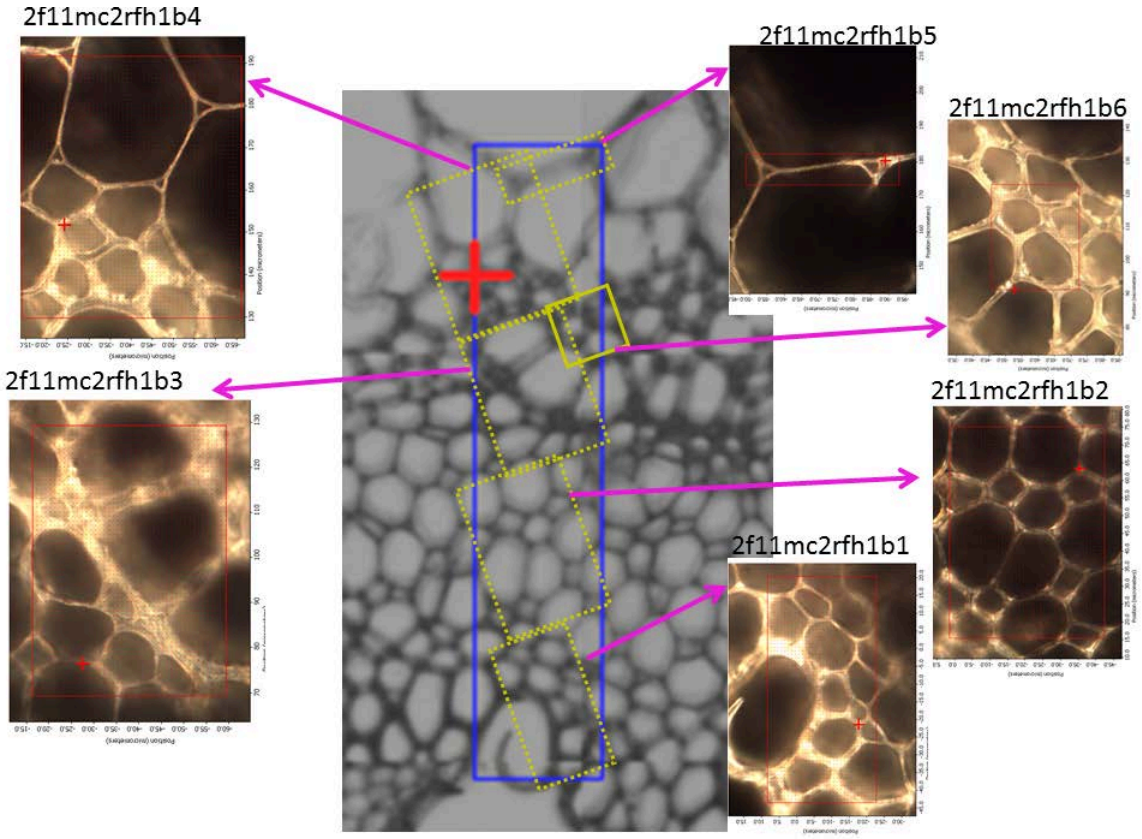


Figure 4.3: Raman images of the maize stem cells cross-section. The mapped regions are highlighted by yellow rectangles. Here the six Raman images labeled as 2f11mc2rfh1b1, 2f11mc2rfh1b2, 2f11mc2rfh1b3, 2f11mc2rfh1b4, 2f11mc2rfh1b5, and 2f11mc2rfh1b6 correspond to images acquired from xylem+phloem, phloem, sclerenchyma+phloem, sclerenchyma+parenchyma, single parenchyma wall, and sclerenchyma cell-types, respectively. The inset blue rectangle represents the region mapped by the infrared.

Raman images were acquired at the synchrotron SOLEIL (Gif sur Yvette, France). All spectra were acquired using a confocal DXR Raman microscope (Thermo Fisher Scientific, WI, USA) using a 532 nm exciting laser of a power of 10 mW. Raman spectra were recorded between 3500 and 50 cm^{-1} per step of 0.9642 cm^{-1}. Data collection, stage control, and baseline correction was performed using OMNIC software. The collection time for each spectrum was 5s and the spatial step size was 1×1 μm^2. Five Raman images of size 32×65, 45×61, 44×61, 53×63, 45×10, and 28×33 μm^2 were recorded to cover the same region mapped using infrared spectroscopy. The raw Raman spectra are

available as supplementary figures S1-S6 (Appendix A). The hyperspectral images used for this analysis were obtained from the same regions (cell-types) mapped using infrared. Together with the hyperspectral image, a visible image is also acquired, brightfield in case of the infrared image and darkfield in case of the Raman images. These images are referred to as infrared visible image and Raman visible image in all further instances in this chapter. In addition, a brightfield image of the whole cross-section was acquired using the facilities of biopolymers-structural biology platform (INRA, Nantes) using an inverted Nikon A1 confocal laser scanning microscope. This is used as the reference image for registration.

4.2.3 Pre-processing of Raman images

A hyperspectral image is a three-dimensional data cube with two dimensions corresponding to the spatial coordinates of the pixels and a third dimension corresponding to the wavelength. The obtained hyperspectral image is a wide collection of data stored in pixels and is composed of thousands or, sometimes, millions of data points (Schultz et al., 2001, Grahn and Geladi, 2007). In hyperspectral imaging, there is a spectral resolution that determines the amount of different information obtained for each pixel based on the different channels (wavelengths). An important problem for the analysis is the presence of erroneous data values which might be bad pixels (having either missing or zero signal values), unexpected spectral readings (extreme values), and outliers (observations inconsistent to the whole dataset). Hence, pre-processing of these hyperspectral images is mandatory (Vidal et al., 2012, Jones et al., 2012). Raw Raman spectra were pre-processed in several steps to reduce spectral/spatial artifacts and the procedure includes (Allouche et al., 2012b):

- Band selection was performed because the entire spectral range is not usable.
- Baselines can be described as the slowly varying curve going through the lower part of the spectra without the jumps of the peaks. Baseline variation is a problem encountered in spectral data. Typically, it is a linear or nonlinear addition to the spectra that causes expected zero measurements to attain a positive value. For better resolution and analysis of the spectra, the phenomenon of baseline drift was eliminated from the spectra during acquisition using OMNIC software (Gonzalez and Woods, 2006).
- Spikes and spectral noise were removed by considering each spectrum individually. Spikes are narrow peaks and the principle consists in identifying peaks narrower

than a window of given size and higher than a given threshold. This was done by applying a 1-dimensional top-hat transformation (Gonzalez and Woods, 2006).

- The noise in our data was removed using the Fourier transform as described in Gonzalez and Woods (2006). For each spectrum, the Fourier transform is calculated and a Gaussian filter is applied to retain only the lower frequency components of the signal and the inverse Fourier transform yields the smoothed spectrum.

Spatial and spectral normalization

Spectra are acquired by scanning adjacent points, which form a matrix with two spatial dimensions and one spectral dimension. There may be intensity variations in the spectra based on the morphology of the maize sections, i.e., high intensity spectra may be acquired on the junctions or in cell wall whereas low intensity spectra are acquired from the inside of the cells. Other issues such as the thickness of the sample, heterogeneity leading to spectral acquisition at different focal planes with respect to the surface, and variations of intensity related to the position in the image may result in different spectral intensity (Vovk et al., 2007, Tomazevic et al., 2002). Hence, it is essential to enhance both the spectral and spatial structures of the spectra in order to study the chemical information contained within.

Conventionally, in spectroscopy, the spectra are pre-processed individually giving the same overall intensity, i.e., by dividing each spectrum by total sum of intensities or the sum of intensities of a peak or by their standard. In this case, this would give the same importance to spectra within cells as spectra from walls and thus introduce spectra noise in the analysis. One way would be to eliminate the spectra corresponding to a no signal by defining a threshold intensity only above which a spectrum is considered valid. The problems associated with this include defining the threshold and the application of thresholds to areas of high and low contrast. To overcome this, we considered the methods which take into account the spatial neighborhood of the spectra as described in Devaux et al. (2010). The principle is similar to the classic normalization procedure used in spectroscopy except that the normalization factor applied to each spectrum was assessed by considering neighbor pixels. This was done by assessing the image of the sum of spectral intensities and low values corresponding to the cell lumens are replaced by the lowest values of the surrounding cell walls by adopting the hole fill image analysis procedure. Then the image of the sum of intensities of each spectrum after normalization was obtained. The result of this normalization is such that the spectra on cell walls show similar intensity, signal is low inside the cell and intermediate intensity on the edge is

preserved (Devaux et al., 2010).

4.2.4 Infrared image pre-processing

Spectra were pre-processed using the Unscrambler V 10.1 software. Linear baseline correction was applied between $1800 - 700\,\mathrm{cm}^{-1}$. A moving average of size 9 was applied for spectral smoothing. No intensity normalization was performed in order to preserve the variation of sample density between cell-types.

4.2.5 Principal component analysis of hyperspectral data

Principal component analysis (PCA) is a technique that is useful for the compression and classification of data. The purpose is to reduce the dimensionality of a dataset (sample) by finding a new set of variables, smaller than the original set of variables, that nonetheless retains most of the sample's information. Here, information refers to the variation present in the sample, given by the correlations between the original variables. The new variables called principal components (PCs) are uncorrelated, and are ordered by the fraction of the total information each retains. The initial investigation of hyperspectral images was done using PCA. Hyperspectral images can be seen as stack of images of dimension (a, b), as sets of spectral vectors of λ wavelengths or as cubes of data $(\mathrm{x}, \mathrm{y}, \lambda)$. The third dimension λ is referred to as the spectral way and are usually considered as variables in multivariate spectral data analysis.

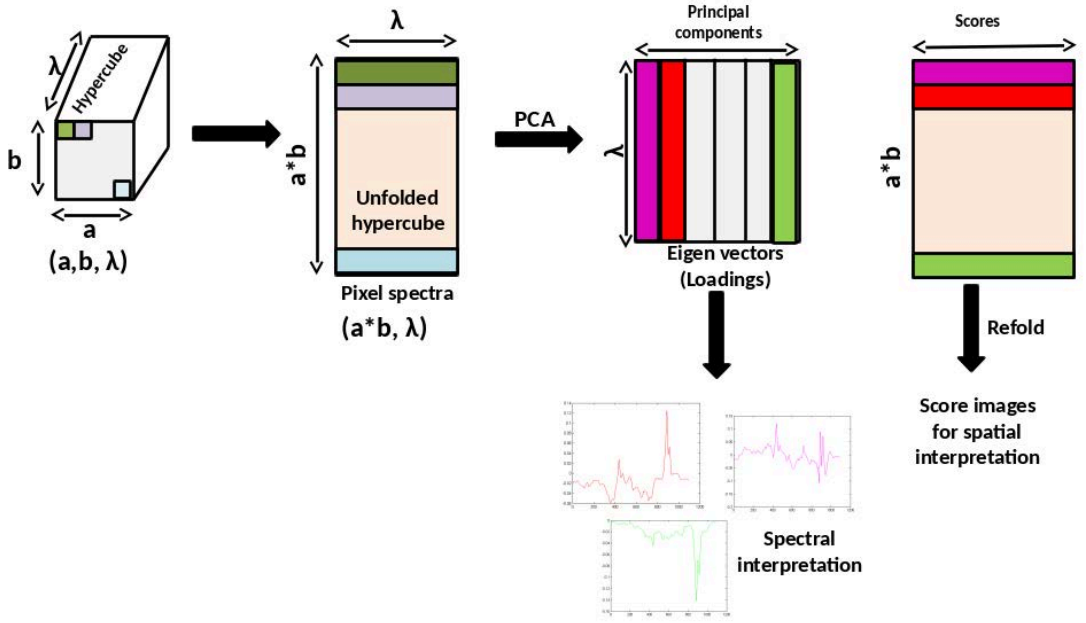


Figure 4.4: Steps involved in PCA of hyperspectral data. The hyperspectral cube is unfolded to obtain the matrix of the format $(a * b, \lambda)$ and after computing the PCA, scores and loadings are used to make the spectral and spatial interpretations, respectively.

Unlike the conventional way of computing PCA, here it is necessary to 'unfold' the hypercube into a two-dimensional matrix in which each row represents the spectrum of pixel (Hori and Sugiyama, 2003, Gowen et al., 2008, Lim et al., 2001). Unfolding in this context means reorganizing a hypercube into a two-dimensional data matrix. The hyperspectral cube with the dimensions (a, b, λ) shown in Figure 4.4 was unfolded to obtain the matrix of the format $(a * b, \lambda)$. Note that the notation denoted here is only for an understanding of how the PCA works for the case of a hypercube. After unfolding of the hypercube, PCA was applied to the hyperspectral data to obtain the scores and loading matrix. The scores contain the concentration variability of the pixels and the loadings denote the spectral variability. After the application of PCA, refolding of the scores give the score images which contain the individual contribution for each component of the original hypercube and help to visualize the distribution of the components.

4.2.6 Image registration

Infrared and Raman hyperspectral images were acquired independently using separate instruments. The raw data was therefore not paired and the multiblock data structure could not be obtained in a straightforward way. The first step to obtain a two block data table was to register the spatial position of the infrared and Raman pixels. In addition,

the two hyperspectral domains are heterogeneous in terms of spatial resolution and joint data analysis of Raman and infrared images therefore requires registration and data fusion techniques.

Image registration is the process of aligning a pattern image over a reference image so that pixels present in both images are disposed in the same location. In typical image registration problems, the reference image and the pattern image are expected to be related to each other in some way and have some elements in common. The source of differences could be of alignment, differences due to occlusion, differences due to noise, differences due to change and these may be significant for the interpretation of the mapped region (Gonzalez and Woods, 2006, Grahn and Geladi, 2007). The registration algorithm used was based on the combination of a template matching technique with a multi-resolution technique (Allouche et al., 2012b). The image to be registered is chosen as the template. A brightfield image of the whole cross-section was used as the reference image for registration. As the mapped infrared and Raman regions were small compared to the reference image, registration was performed in two steps (Figure 4.5):

- Each spectral image was registered to its associated visible image (cf. Section 4.2.2).
- Each visible image was registered to the reference image.

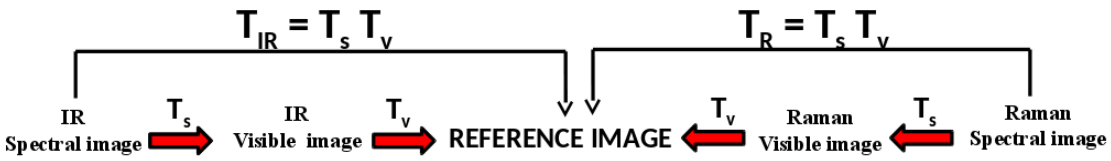


Figure 4.5: Registering a spectral image onto a reference image. The first part of the figure depicts how the registration of the spectral images onto the reference image is produced in two steps (from spectral image → visible image → reference image). The figure also represents the calculation of transformation matrices for the infrared ($T_{IR} = T_v T_s$) and Raman ($T_R = T_v T_s$). The subscripts 'v' and 's' refer to the visible and spectral image, respectively.

The procedure was followed as in Allouche et al. (2012b,a). A set of template images is assessed by scaling and rotating the template image with several scales, and with rotation combinations within a given interval defined from a rough initial estimates of scale and rotation. The solution was to test different values of scaling factor and rotation angles, and then the cross correlation translations were calculated for template matching. Values of scale, rotation and translation corresponding to the maximum correlation are retained. The resolution and the number of pixels of the reference images are very different than

those of the images to be registered. Search parameter transformations by applying a pyramid decomposition of images was considered. In order to speed up the search processing, reference and template images were considered at different resolutions using a Gaussian pyramidal decomposition. Briefly, the search process starts with a low resolution, and the whole procedure is iterated to refine values of scale and rotation until the maximum image resolution is reached. In this multi-resolution approach, the level of the pyramid must be chosen to preserve the cellular structure within the images. The algorithm developed by Allouche et al. (2012b) provides all possible solutions, and a suitable level must be chosen.
At the end of the registration procedure, we have the transformation matrices T_{IR} and T_{Raman} connecting different infrared spectral images, and Raman to the reference image. The calculation of the affine transformation 'T' to register the spectral image to the area of the reference image was obtained by multiplying affine transformations of two matrices (Figure 4.5):

- T_s to pass the spectral image to the visible image.
- T_v for moving from the visible image to the reference image.

At the end of the registration step, it is possible to find the spectra acquired at every location point for each spectral data image. However, Raman and infrared datasets are heterogeneous in terms of size and content. The final structure of the data is obtained through unfolding first the infrared data table. For each infrared data table, a set of Raman pixels will be paired. All pixels will not be paired and missing data can be observed. Each data block will be considered as partitioned with some parts completely paired with the other spectral modalities, and some parts with no pairing.

4.2.7 Application of multiple co-inertia analysis

Infrared pixels covered an area of 5×5 μm^2 while pixels from Raman imaging covered a 1×1 μm^2 area. Consequently, each infrared spectrum was paired to a set of Raman spectra. The data tables were built through unfolding first the infrared hyperspectral image. Each line of the infrared table was paired to the Raman spectra after unfolding the small images. The steps include unfolding multi-resolution pixels, pairing and stacking all pixels, obtaining a paired data structure, and obtaining the two data tables. The final data structure was a set of two data tables:

- A two-way Raman data table X, with 'n' being the number of Raman spectra for which the infrared spectra were also acquired and 'p' corresponds to the variable mode (λ).
- A two-way infrared data table Y with 'n' observations and a variable mode (q).

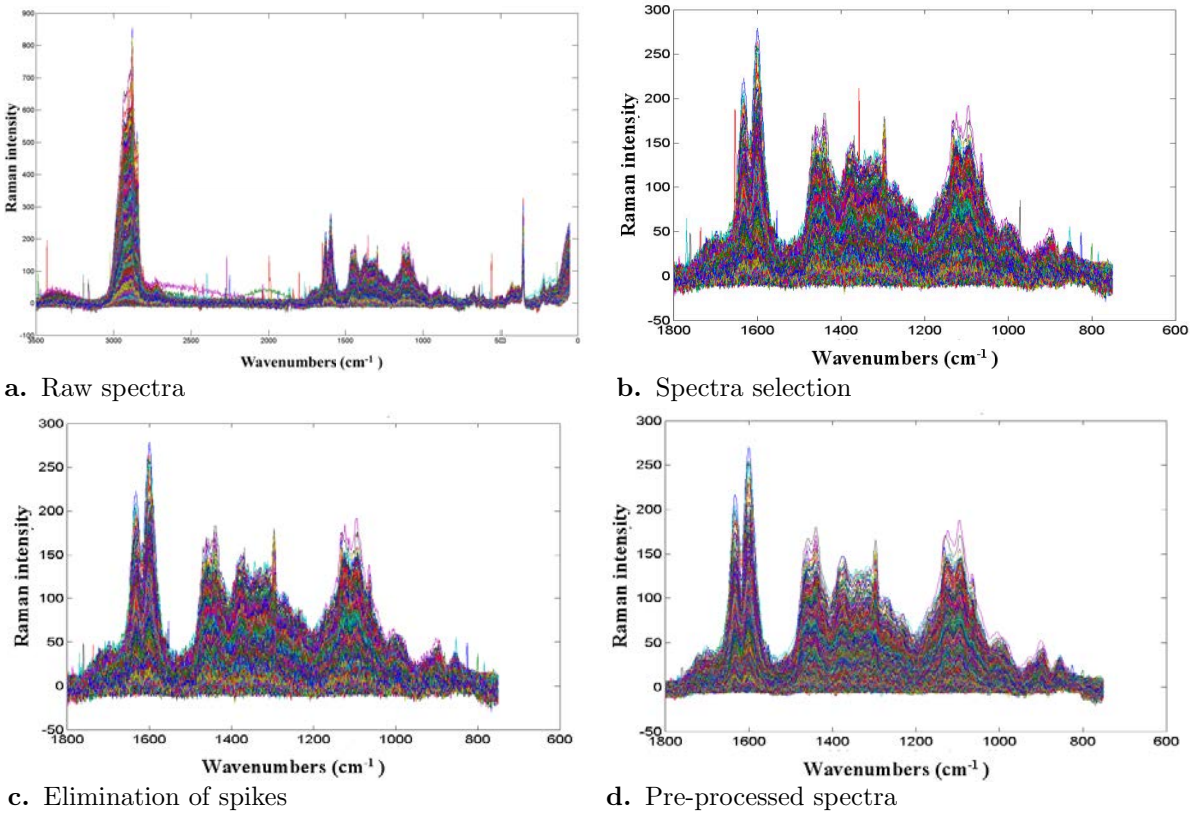
The working principle behind MCIA is detailed in Chapter 2. Briefly, we applied MCIA to the two data tables. This analysis will reveal covariant patterns between the multivariate data tables. Block loadings, scores, global loadings, scores were assessed and used as indicators to investigate the common and specific information brought by the infrared and Raman data table. In contrast to the CCA, this technique maximizes the covariance between the two data tables and explains the unique as well as common information between the data tables.

4.3 Results

Infrared image of the size 10×25 and Raman images of size 32×65, 45×61, 44×61, 53×63, 45×10, and 28×33 μm^2 were used. For each spectral domain, pre-processing consisted of the corrections specific to each spectroscopy. Raman spectra were automatically baseline corrected at acquisition. Pre-processing steps consisted in selecting spectral region of interest, spike removal, smoothing and are illustrated below. The normalization step takes into account both the spectral and spatial information.

4.3.1 Spectral pre-processing and normalization

Figure 4.6 shows the different pre-processing steps for analyzing one Raman image. Band selection was performed based on Raman hyperspectral data taken at 1800-600 cm^{-1} spectral range.



a. Raw spectra **b.** Spectra selection

c. Elimination of spikes **d.** Pre-processed spectra

Figure 4.6: Steps involved in pre-processing of one hyperspectral image. The steps involved in pre-processing is detailed using the second Raman image (corresponding to phloem cell-type). The process includes (a) visualizing the raw spectra (3500-0 cm^{-1}), (b) band selection of the raw spectra from 1800-600 cm^{-1}, (c) elimination of spikes, and (d) pre-processed spectra. Similar pre-processing was done for all the other Raman images.

Spikes observed in the Raman spectra were removed using the procedure described in methods section (Section 4.2.3). First, the most intense spikes of width less than 11 spectral points and greater than 200 in intensity were removed, then, the less intense spikes of less than 7 spectral points width and exceeding 50 in intensity were also removed. The noise was removed by Fourier transform using a Gaussian filter of size 151 spectral points. The spatial/spectral normalization was applied by considering a neighborhood of 11×11 pixels. Intensity normalization was performed taking into account both the spectral and spatial information as described in Devaux et al. (2010). Infrared spectra were pre-processed as described in the material and methods section of this chapter. The spatial/spectral normalization was not performed for the infrared spectra in order to keep the information that in some regions, only small cell walls (phloem or parenchyma) are observed while in other regions large cell walls are observed (sclerenchyma). The pre-processed and normalized Raman and infrared spectra are available as supplementary

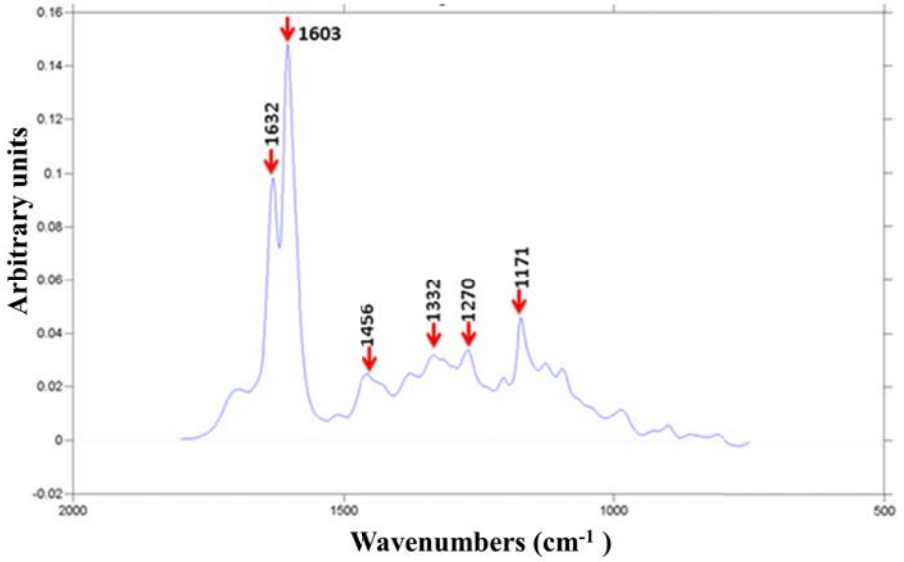
figures S7-S13 (Appendix A).

4.3.2 Variability according to cell-types

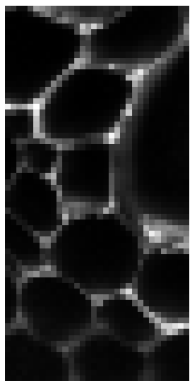
Spectral images from infrared (one image) and Raman (six images) were analyzed individually using principal component analysis. The objective is to identify the information available in each dataset before submitting it to joint analysis. Since the spectra were normalized preserving the differences of intensities between holes and cell walls, the first component describes the intensity variations between spectra. Score images are displayed as grey level images and were obtained by refolding the scores.

Xylem+phloem cell-types (2f11mc2rfh1b1)

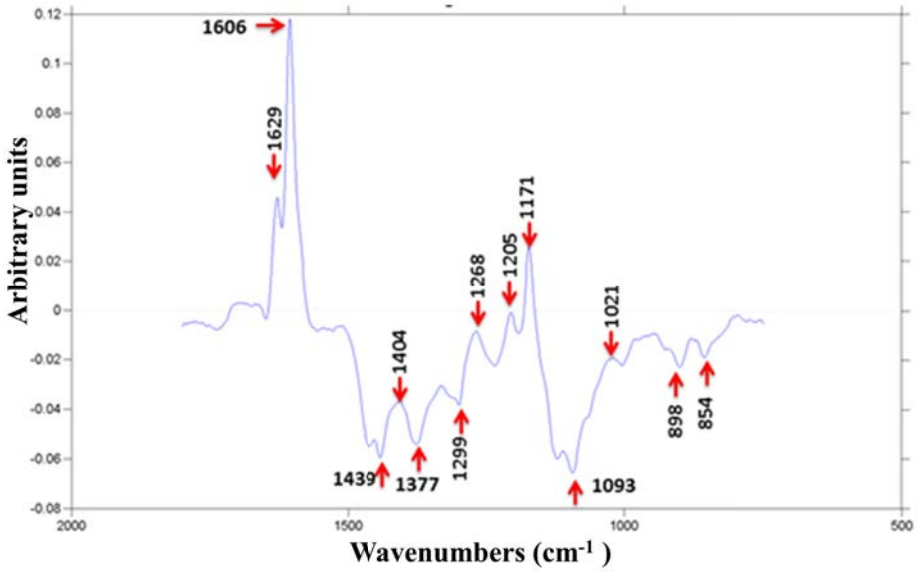
The first two components obtained for the first Raman image describe 94.97 % and 4.02, respectively (a cumulative of 98.99 %) of the total variability. Assignment and interpretation of the peaks in the spectrum was done using previously reported assignments from literature (Table S2, Appendix B). The first and second components are shown in Figure 4.7. In the corresponding score images, the cell walls are shown in white and the holes (inside of the cell) are in black. Again the first spectral profile was an intensity profile (Figure 4.7) and the positive peaks at 1171, 1270, 1332, 1603, and 1632 cm^{-1} correspond to lignin profile whereas the one at 1268 cm^{-1} correspond to arabinoxylan. In the second loading of Figure 4.7, the positive peaks at 1171, 1606, 1632 cm^{-1} correspond to lignin. Negative peaks at 898, and 1093 cm^{-1} correspond to arabinoxylan, and those at 1093, and 1377 cm^{-1} correspond to cellulose. In the second score image, xylem cell walls appear bright and phloem cell walls appear black consistent to the lignin/polyscaccharide content of xylem and phloem cell walls.



a. Intensity spectra



b. Score image



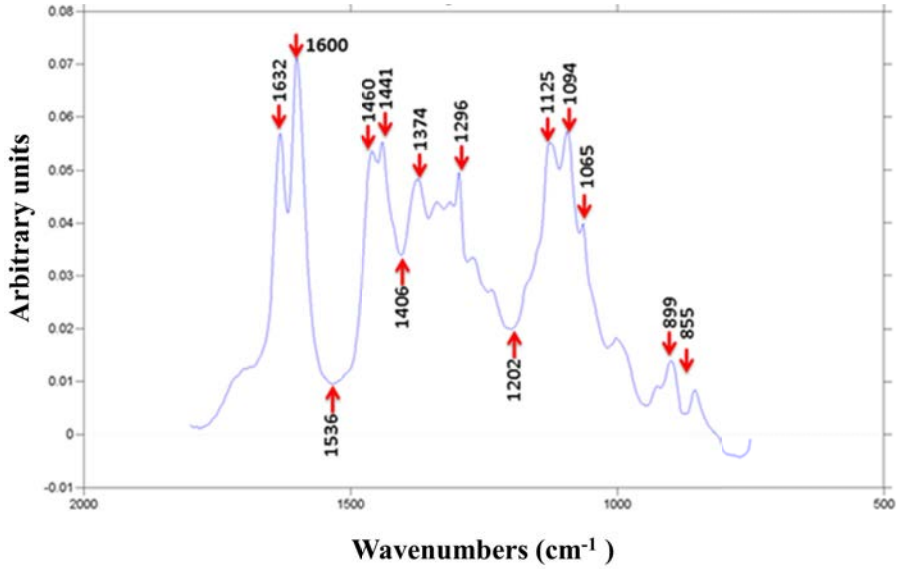
c. Second loading

d. Score image

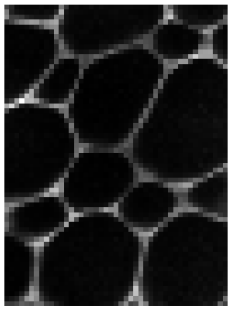
Figure 4.7: PCA analysis of the first Raman image that profiles xylem+phloem cell-type. (a) and (b) depicts the intensity spectra and its corresponding score image, respectively. (c) and (d) pertains to the second component and its corresponding score image, respectively.

Phloem cell-type (2f11mc2rfh1b2)

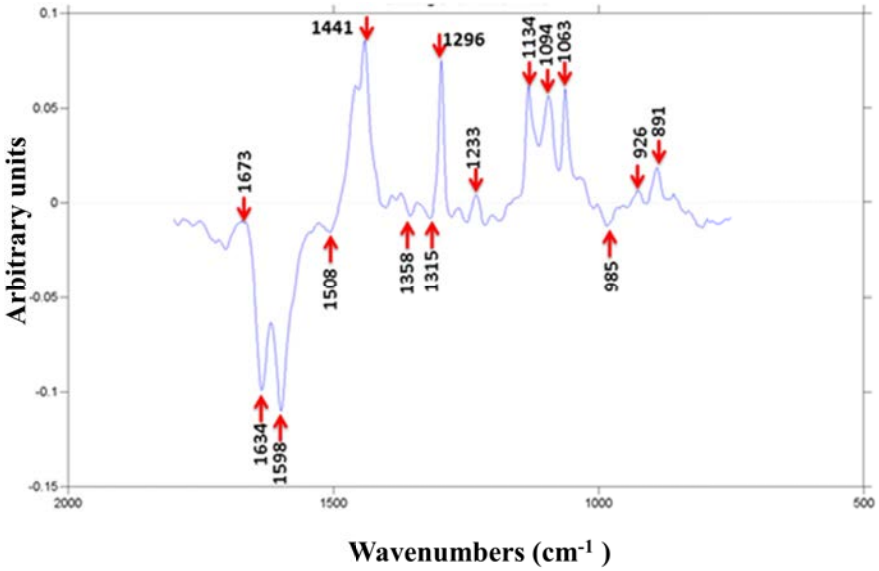
The first and second components describe 95.64 % and 1.29 % of the total variability, respectively. In the intensity profile of Figure 4.8, the positive peaks at 1094 correspond to cellulose whereas those at 1094, 1460 cm^{-1} correspond to arabinoxylans and 1600, 1632 cm^{-1} to a strong lignin profile.



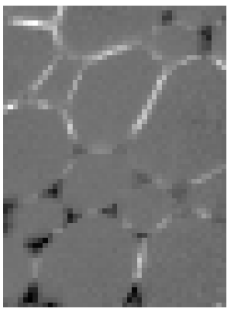
a. Intensity spectra



b. Score image



c. Second loading



d. Score image

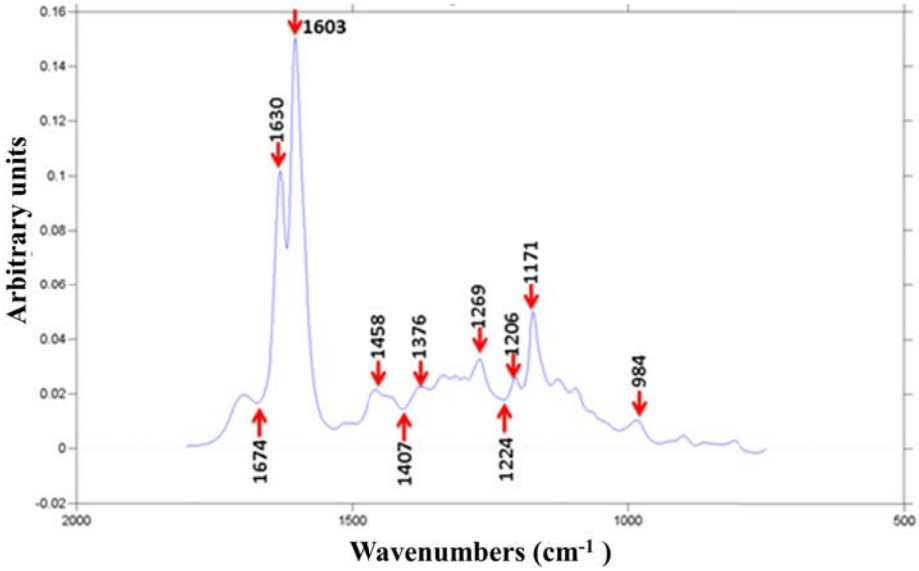
Figure 4.8: PCA analysis of the second Raman image that profiles phloem cell-type. (a) and (b) depicts the intensity spectra and its corresponding score image, respectively. (c) and (d) pertains to the second component and its corresponding score image, respectively.

Once again, the score image shows the cell wall in contrast to cell lumen. In the second loading of Figure 4.8, the positive peak at 1094 cm^{-1} correspond to both cellulose and arabinoxylans. Negative peaks at 1508, 1598, and 1634 cm^{-1}correspond to lignin. The corresponding score images show black spots corresponding to cell junctions mainly for the smaller cells of phloem. As this image correspond only to phloem cell types, this loading shows some enrichment maybe in lignin or phenolic compound in the junction

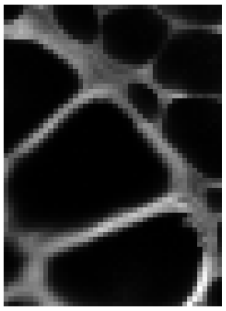
region and in polysaccharides in other cell walls. The unknown peaks at 1296, 1063 cm^{-1} were found mainly on the large cell walls.

Sclerenchyma+phloem cell-type (2f11mc2rfh1b3)

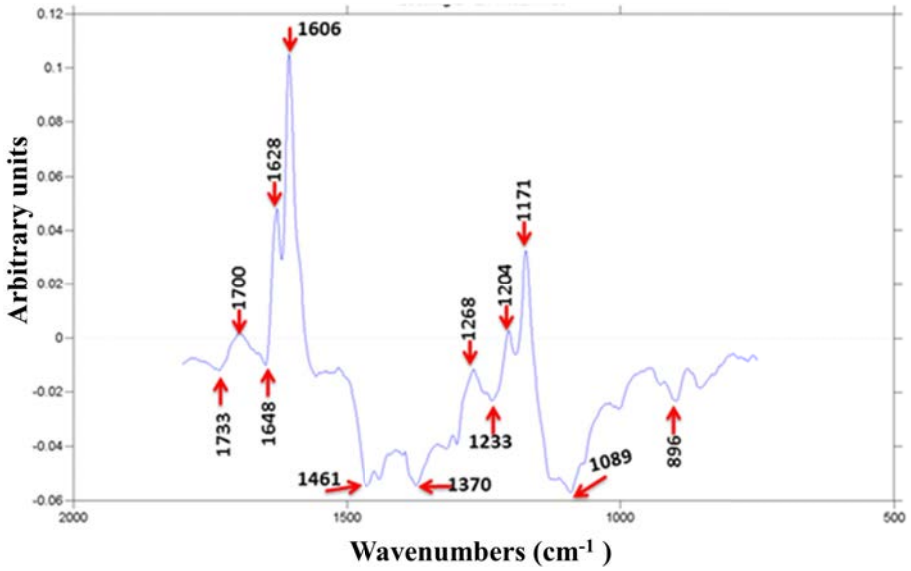
The first two components describe 97.70 % and 1.75 % of the total variability (a cumulative of 99.45 %).



a. Intensity spectra



b. Score image



c. Second loading

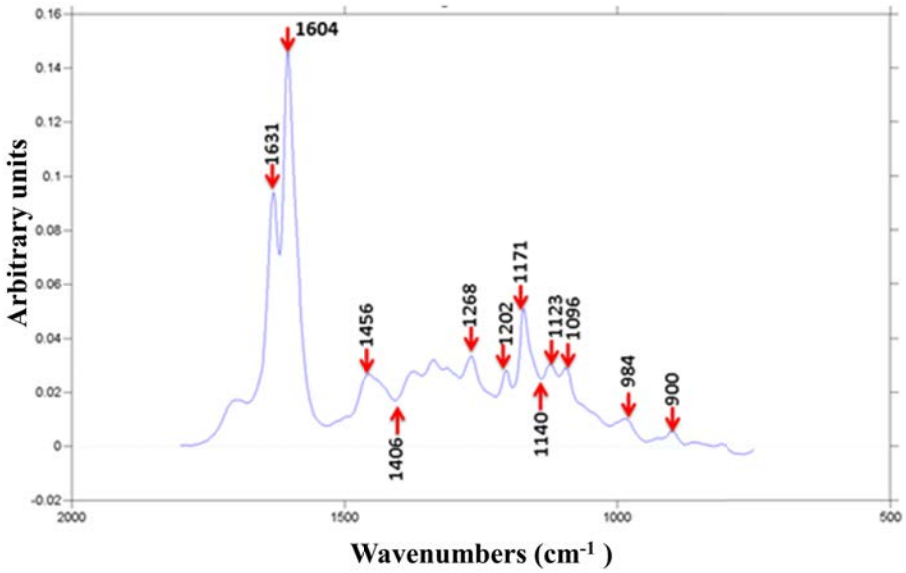
d. Score image

Figure 4.9: PCA analysis of the third Raman image that profiles sclerenchyma+phloem cell-type. (a) and (b) depicts the intensity spectra and its corresponding score image, respectively. (c) and (d) pertains to the second component and its corresponding score image, respectively.

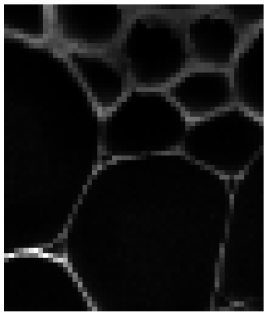
The first spectral profiles and score images were interpreted as for the other images as a description of global intensity variations as shown in Figure 4.9. The positive peaks at 1171, 1603, 1630 cm^{-1} correspond to a strong lignin profile. In this image, sclerenchyma cells were mainly observed with very few phloem cells. Hence, the first loading corresponds to a sclerenchyma spectrum. In the second spectral profile of Figure 4.9, the positive peaks at 1171, 1606, 1628 cm^{-1}, and negative peak at 1648 cm^{-1} correspond to a lignin profile. The positive peak at 1268 cm^{-1} and negative ones at 896, and 1461 cm^{-1} refer to arabinoxylans. The score image shows phloem/sclerenchyma walls in contrast.

Sclerenchyma+parenchyma cell-types (2f11mc2rfh1b4)

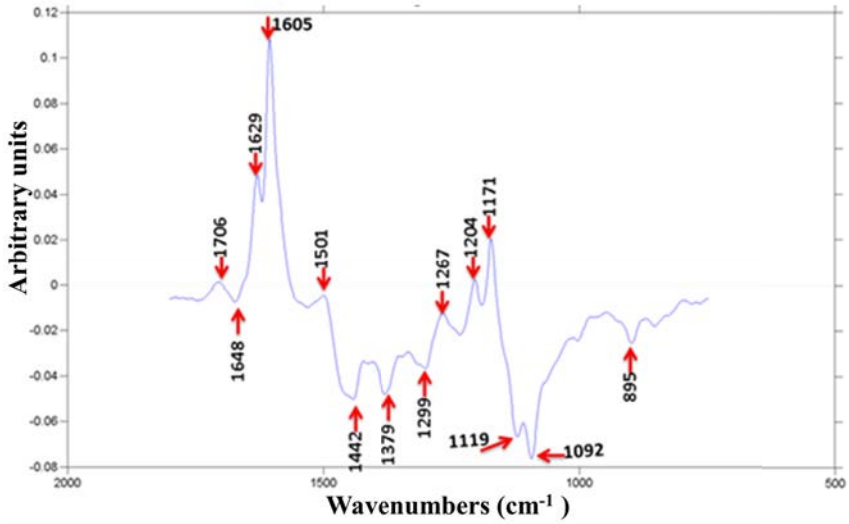
The first two components obtained for the fourth Raman image describe 91.79 and 6.28 % of the total variability, respectively. Positive profiles for lignin (1171, 1604, 1631 cm^{-1}), and arabinoxylans (1268 cm^{-1}) were found in the first profile. The second profile in Figure 4.10 has positive peaks referring to lignin and negative ones to cellulose and arabinoxylans. In the second score image, some parenchyma cell walls were contrasted with the sclerenchyma.



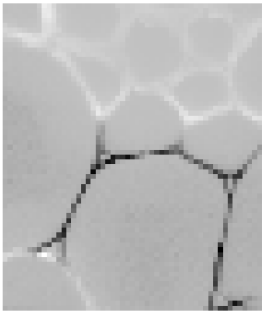
a. Intensity spectra



b. Score image



c. Second loading

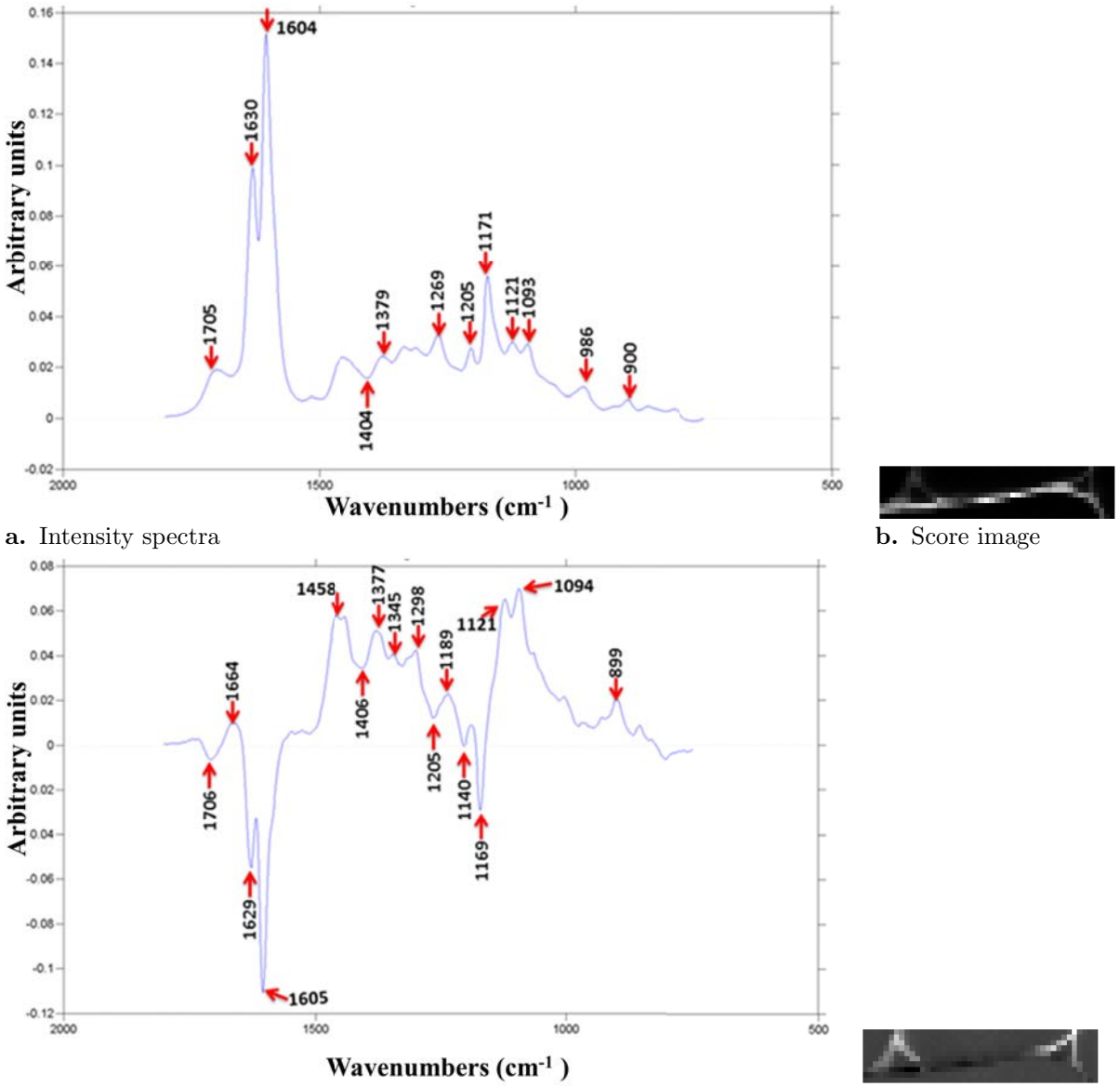


d. Score image

Figure 4.10: PCA analysis of the fourth Raman image that profiles sclerenchyma+parenchyma cell-types. (a) and (b) depicts the intensity spectra and its corresponding score image, respectively. (c) and (d) pertains to the second component and its corresponding score image, respectively.

Single parenchyma cell wall (2f11mc2rfh1b5)

The fifth image corresponded to a cell wall selected at the frontier between the two parenchyma cells highlighted in the preceding image. The first two components obtained for the fifth Raman image describe 99.77 % of the total variability. Again, the first loading and its corresponding score images reveal the variations in the global intensity. Positive peaks of the intensity profile at 1093, and 1379 cm^{-1} correspond to cellulose whereas those at 1171, 1604, and 1630 are lignin profiles (Figure 4.11).



a. Intensity spectra

b. Score image

c. Second loading

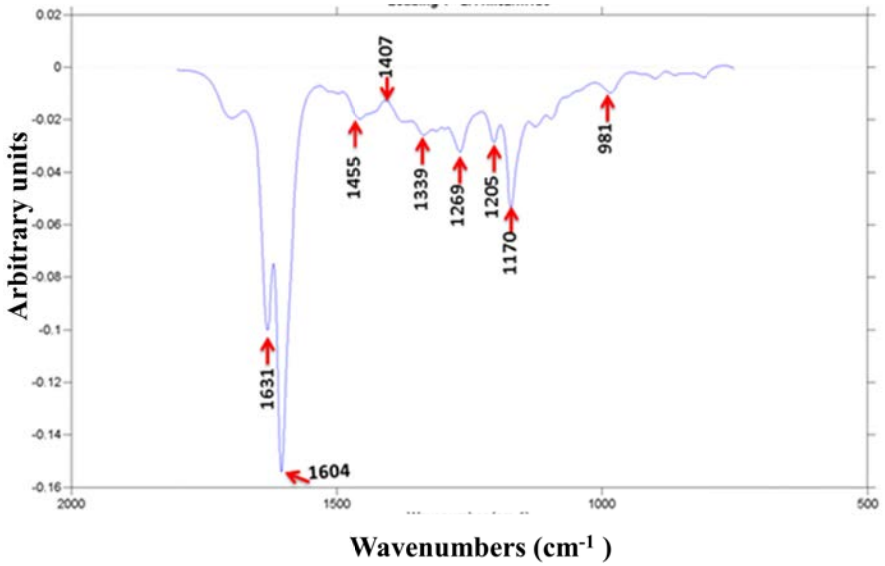
d. Score image

Figure 4.11: PCA analysis of the fifth Raman image that profiles a single parenchyma cell wall. (a) and (b) depicts the intensity spectra and its corresponding score image, respectively. (c) and (d) pertains to the second component and its corresponding score image, respectively.

Positive peaks of second spectral profile correspond to cellulose at 1094, and 1377 cm^{-1} and negative peaks to lignin (1169, 1605, and 1629 cm^{-1}). The second loading actually corresponds to the inverted loadings of Raman images 2f11mc2rfh1b1, 2f11mc2rfh1b3, and 2f11mc2rfh1b4 and thus revealed an opposition between lignin and polysaccharides. The cell junctions were rich in polysaccharides while the cell wall contained some lignin showing a strong variation of composition within a small sampled region.

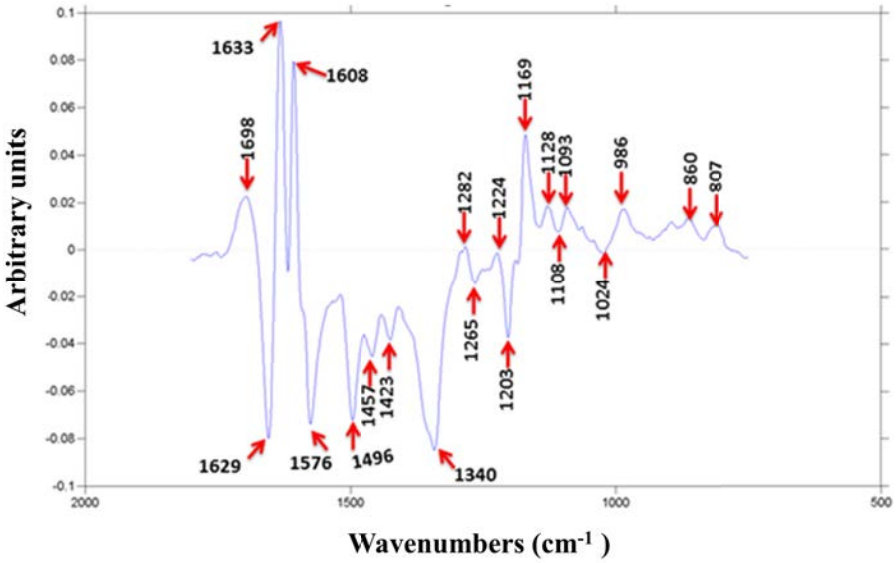
Sclerenchyma cell-type (2f11mc2rfh1b6)

This last image contained only the sclerenchyma cell walls.

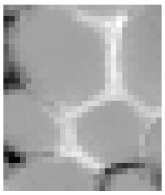


a. Intensity spectra

b. Score image



c. Second loading



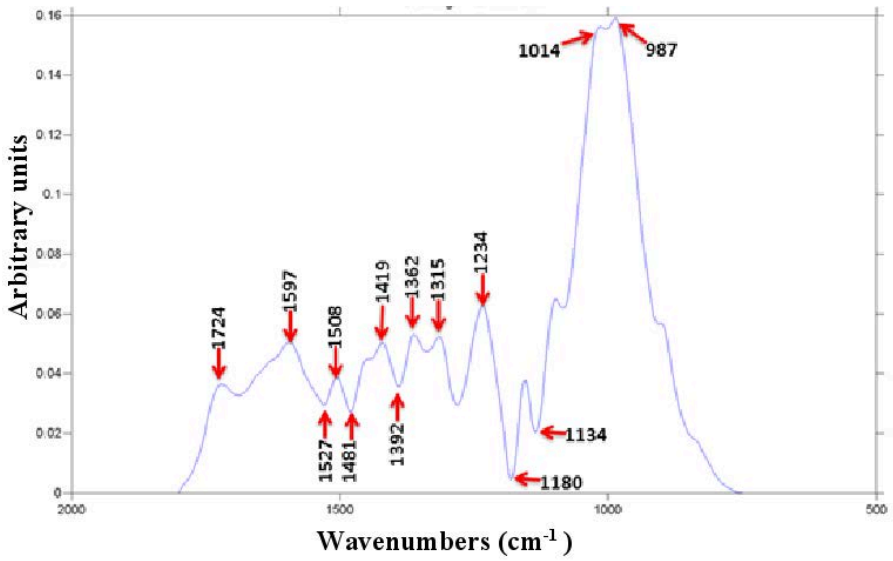
d. Score image

Figure 4.12: PCA analysis of the sixth Raman image that profiles border Sclerenchyma cell-types. (a) and (b) depicts the intensity spectra and its corresponding score image, respectively. (c) and (d) pertains to the second component and its corresponding score image, respectively.

The first loading and score image described intensity variations. The second loading exhibited many peaks that could not be assigned easily to known compounds. The first two components obtained for the sixth Raman image describe 99.80 % of the total variability.

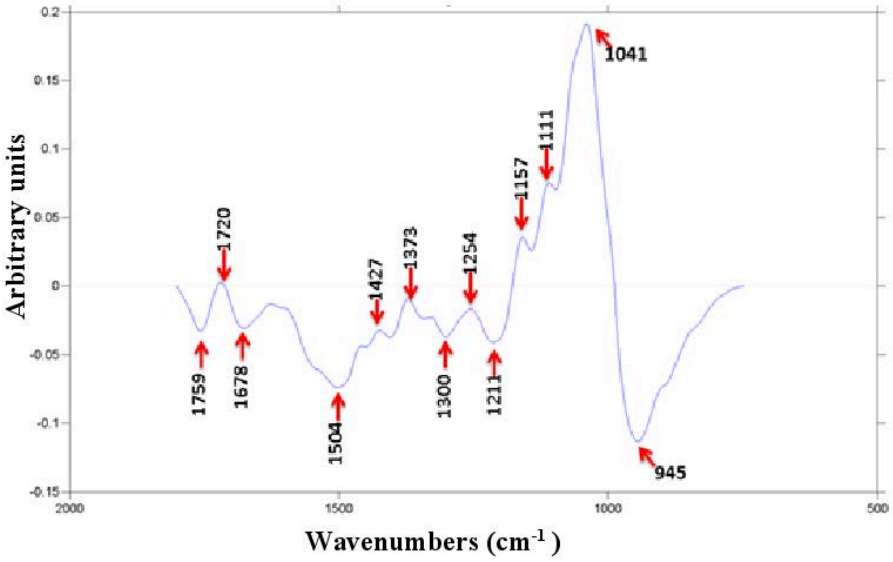
In the first spectral profile (Figure 4.12), the peaks correspond to cellulose (positive ones at 1407 cm^{-1}).

Infrared image (2f11mc2r)

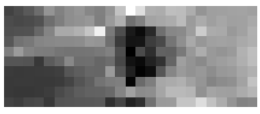


a. Intensity spectra

b. Score image



c. Second loading



d. Score image

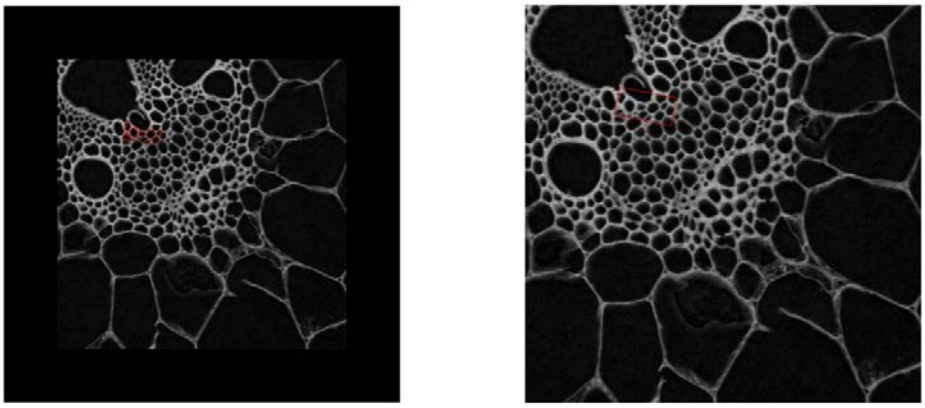
Figure 4.13: PCA analysis of the infrared image. (a) and (b) depicts the intensity spectra and its corresponding score image, respectively. (c) and (d) pertains to the second component and it's corresponding score image, respectively. The infrared image profiled the same cell-types as obtained from the 6 Raman images.

Similar to the peak assignments of the Raman spectra, the infrared peaks assignments

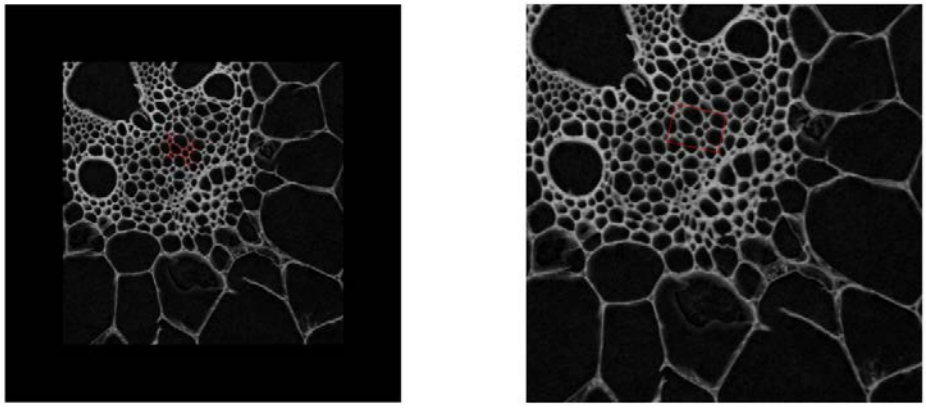
were collected from literature and tabulated in Table S3 (Appendix B). The first and second components obtained for the infrared image describe 91.97 %, and 5.44 % of the total variability, respectively. The peaks of the first spectral profile correspond to lignin at 1419, 1508, and 1597 cm^{-1}. The corresponding score image highlights the sclerenchyma region as the one with the most intense spectra. It also corresponds to the larger cell wall, and therefore a higher density of the material. The second profile corresponds to xylans at 1041 cm^{-1}, cellulose at 1111 cm^{-1}, and lignins at 1720 cm^{-1}. The negative peak at 945 cm^{-1} describes the shift between lignified and non-lignified tissue in the region between 1040 and 800 cm^{-1}. Together with the peak at 1504 cm^{-1}, it can be associated to a lignin response. The second loading mainly described the opposition between between lignified and non-lignified tissues. The score image is not easy to interpret in this case but the sclerenchyma region appeared in black compared to other cell-types.

4.3.3 Pairing of hyperspectral images

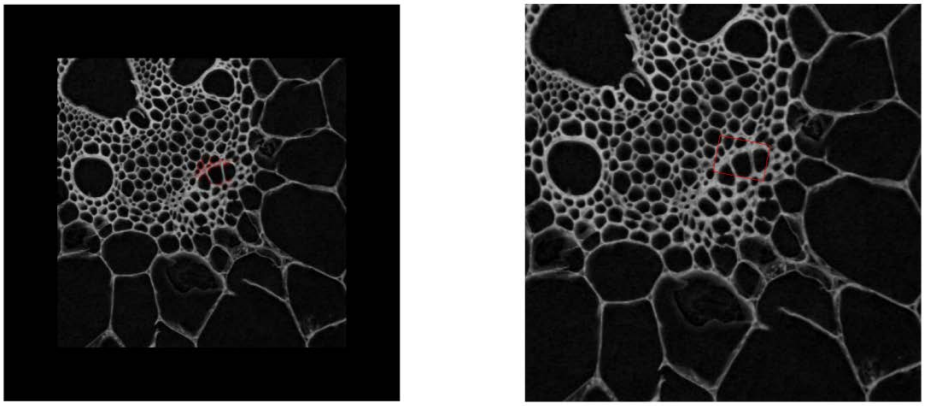
Based on the algorithm described in methods section (Section 4.2.5), pairing of the hyperspectral images was done. Briefly, each spectral image was registered to the reference image (bright-field image) in two steps. In Figure 4.14, pairing of the spectral to reference images is depicted for each of the Raman images and infrared image. The scale and rotation factors for pairing each of these images is listed in Table 4.1 including the processing parameters before and after registration. The level of the pyramid chosen to preserve the cellular structure within the images during registration is mentioned. In case of 2f11mc2rfh1b5 corresponding to the single parenchyma cell wall, the obtained Raman image was small and direct registration of the target image (spectral image) to the reference image was not feasible. To ensure accurate registration of this image to the reference, the target image was first registered to a portion of the reference image and then to the entire reference image (Figure 4.14). Hence, in Table 4.1, we have two sets of parameters for the registration of the target to the reference image. In contrast to Figure 4.14, Figure 4.15 depicts the registration of all the six Raman images and the infrared image onto the reference image. This was done using the computed affine transformation matrices for each of the images. More details on computing the affine transformation matrices is detailed in Section 4.2.5.



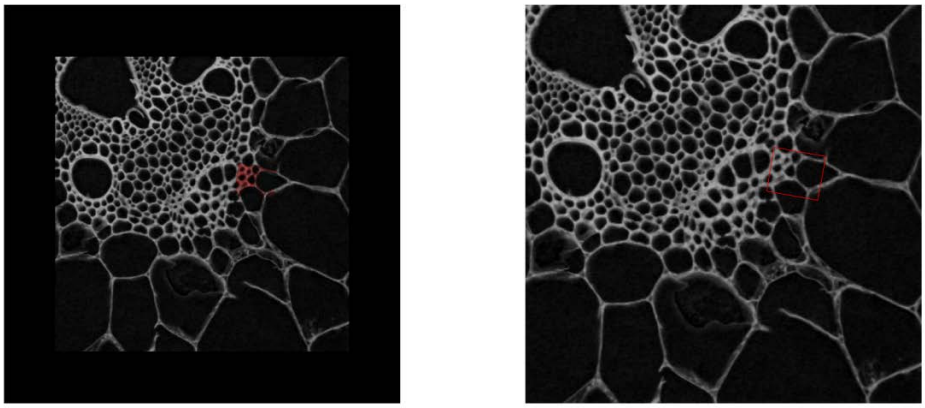
a. Xylem+phloem cell-type (2f11mc2rfh1b1)



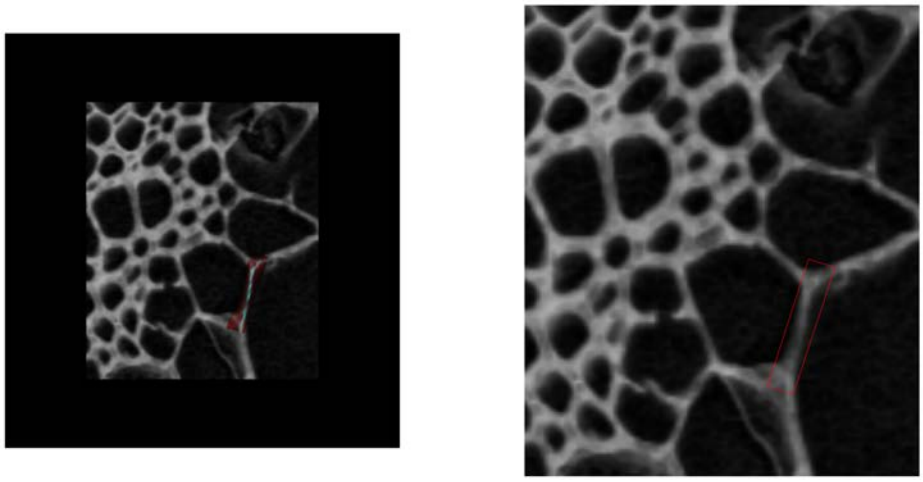
b. Phloem cell-type (2f11mc2rfh1b2)



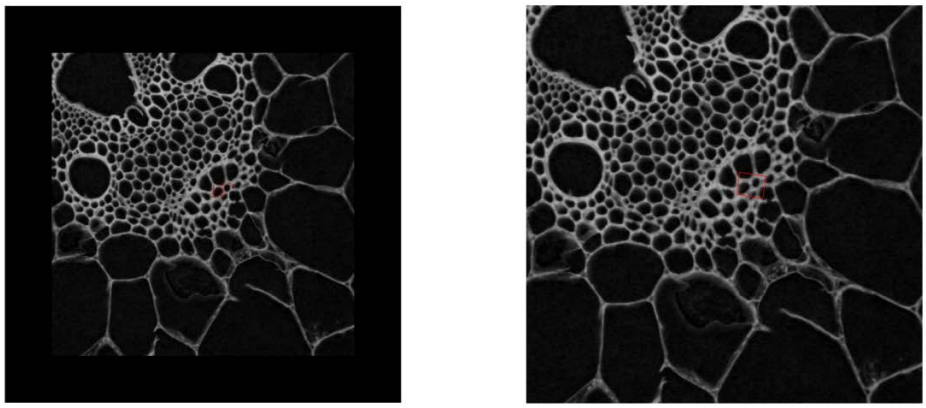
c. Sclerenchyma+phloem cell-type (2f11mc2rfh1b3)



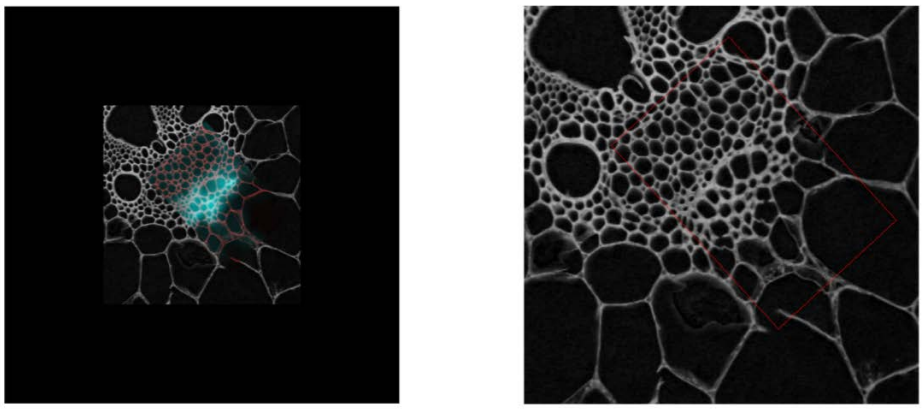
d. Sclerenchyma+parenchyma cell-type (2f11mc2rfh1b4)



e. Single parenchyma cell wall (2f11mc2rfh1b5)



f. Sclerenchyma cell-type (2f11mc2rfh1b6)



g. Infrared image (2f11mc2r)

Figure 4.14: The first six images show the registered Raman images while the last shows the registered infrared image. (a-f) Briefly, using the registration algorithm detailed in section 4.2.5 each of the Raman image (spectral image) is registered onto the reference image in two steps (spectral image → visible image → reference image). (g) Similarly, the infrared image is also registered to the reference image. The figure to the left shows the registration of spectral image → visible image and the figure to the right shows the visible image → reference image. The parameters used in this registration are listed out in Table 4.1.

Sample	PROCESS PARAMETERS				FINAL PARAMETERS				
	Pyramid level: Registrating image	Pyramid level: Target image	Initial scale factor	Initial rotation angle	Scale factor	Rotation angle	Translation Row	Translation Column	Cross-correlation coefficient
2f11mc2rfh1b1	1	2	0.660	-78	3.940	-79.200	483.241	387.114	0.520
2f11mc2rfh1b2	1	2	0.660	-75	3.850	-78.000	605.287	612.631	0.600
2f11mc2rfh1b3	1	2	0.660	-78	3.700	-79.200	737.272	822.220	0.750
2f11mc2rfh1b4	1	2	0.660	-78	3.700	-79.800	816.414	1054.592	0.54
2f11mc2rfh1b5	1	2	0.660	-78	3.700	-72.000	467.681	292.333	0.580
	1	2	0.660	-78	3.700	-96.000	392.104	1603.965	0.700
2f11mc2rfh1b6	1	2	0.660	-78	3.620	-81.000	831.721	918.751	0.730
2f11mc2r	1	5	0.880	-106	22.690	-132.510	1446.687	1119.794	0.410

Table 4.1: Registration parameters used for registering the Raman and infrared images to the reference image. The process parameters refer to choosing the pyramid level for template matching and the final parameters list the final registration parameters.

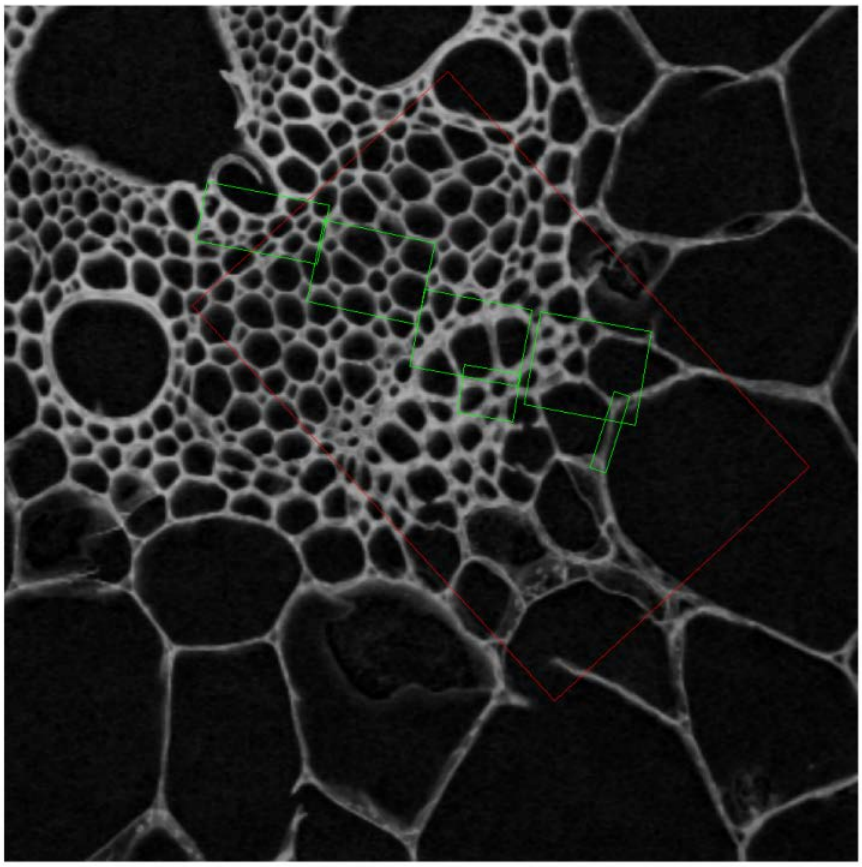


Figure 4.15: The six Raman images (shown in green) and the infrared image (shown in red) are projected together onto the reference image. In figure 4.14, each of the hyperspectral images is registered individually onto the reference image. But, in this case all the images from both techniques are registered onto the reference image.

After registration of the spectral images to the reference image, data tables were built from the common region of interest covered by the infrared and Raman images. The translation, scale and rotation factors were used to assess transformation matrices that allowed computing the coordinates of the infrared pixels corresponding to the Raman pixels. Each infrared pixel in the region of interest was considered and its corresponding Raman spectra were recovered. This way, we were able to identify 47 paired spectra across the infrared and Raman images. Figure 4.16 displays the 47 pixels of the Raman spectra that are also found in the infrared.

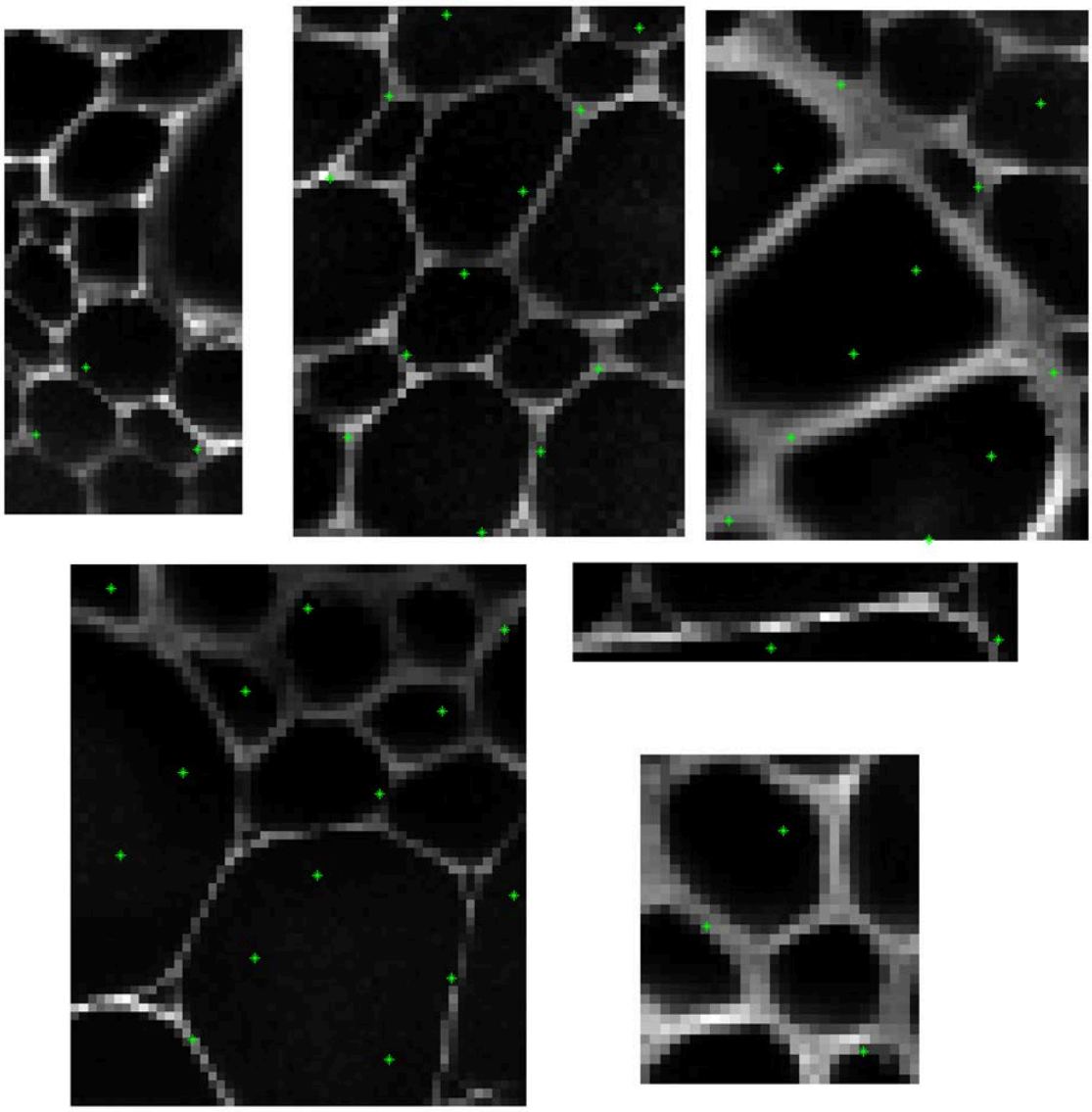


Figure 4.16: The pixels of the six Raman images which are common to the infrared image. The green star represents the pixels in the Raman images for which their corresponding infrared pixels were also recovered.

4.3.4 Percentage of variances and data table contributions

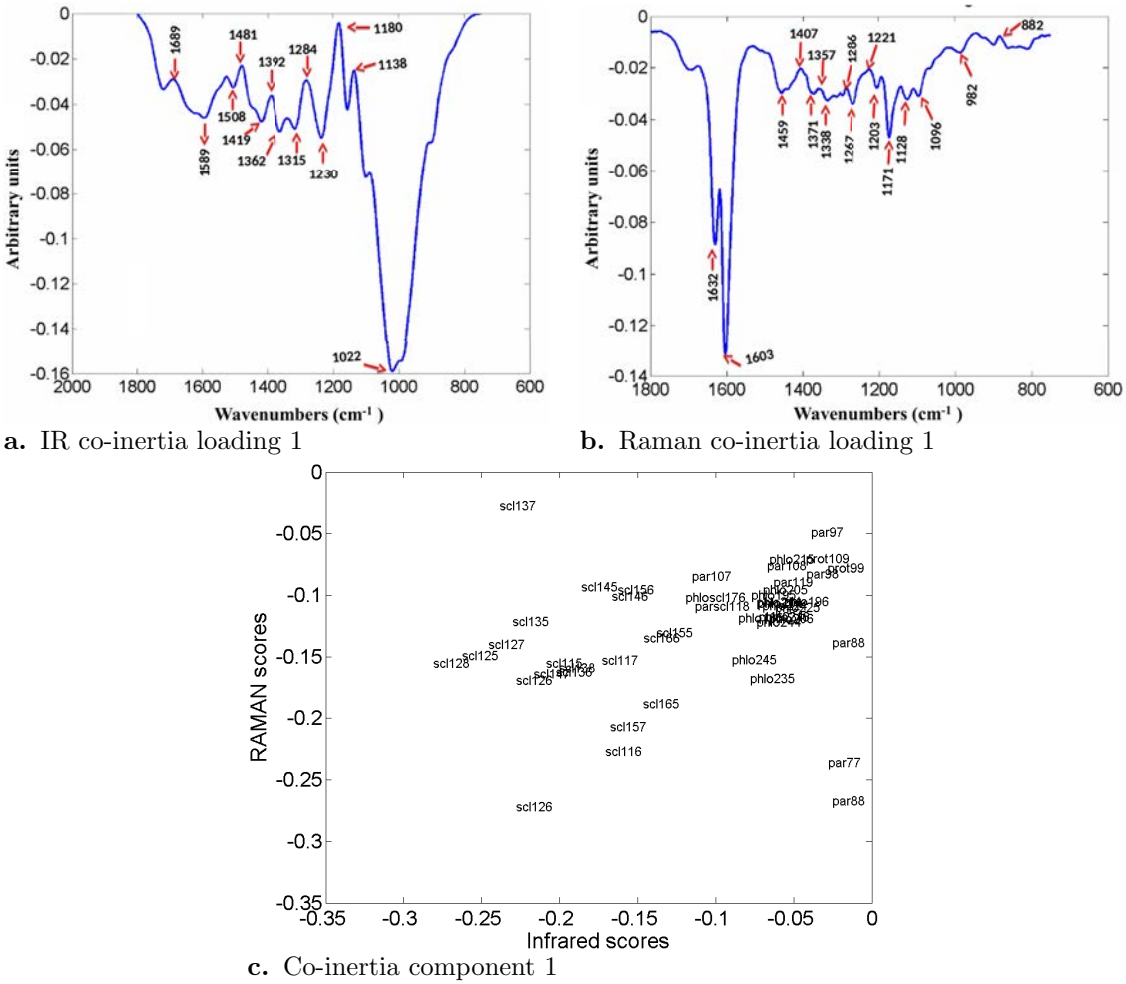
Co-inertia analysis was performed on the two data tables obtained: 47×1090 (Raman spectra× wavenumbers) and 47×274 (Infrared spectra × wavenumbers). Table 4.2 summarizes the contribution of each data table to the analysis using different indicators. For comparison, the first indicator is the percentage of variance described by principal components, and second is the variance explained by the block components. The correlation coefficient between the infrared and Raman block scores is also presented.

	Spectroscopy	Component 1	Component 2	Component 3	Component 4	Component 5
Principal component analysis	Infrared	96.7	2.6	0.6	0.1	0
	Raman	91.9	7.6	0.3	0.1	0
Multiple co-inertia analysis	Infrared	96.6	2.7	0.6	0	0.1
	Raman	91.8	7.7	0.2	0.3	0
Contribution of blocks to common component	Infrared	51	23	96	2	98
	Raman	49	77	4	98	2
Percentage of covariance	Co-inertia common component	87	85	49	53	41
Correlation coefficient	Between Infrared and Raman block	26	85	30	46	27

Table 4.2: Results comparing the Infrared and Raman data blocks. Percentage of the variance of each component is given for comparison with those of multiple-Co-inertia analysis. Indicators for the percentage of the total covariance described by the co-inertia common component between block scores is provided. Moreover, the contribution of each block to the global score and also the correlation coefficient between the block scores is listed here. The percentage of variance described by principal components or co-inertia components were largely similar except for a few differences. The contribution of the blocks differed significantly.

4.3.5 Spectral and spatial interpretation

Loadings were drawn for each spectroscopy and the first loading usually corresponds to intensity variation. The first, second and fourth components are interpreted and discussed as the peaks of the third component look deformed. Moreover, the block scores for each data table were assessed, back-folded and displayed. Such representation allowed interpreting score intensity variations with their spatial location i.e., their cell-type origin to be specific.



a. IR co-inertia loading 1

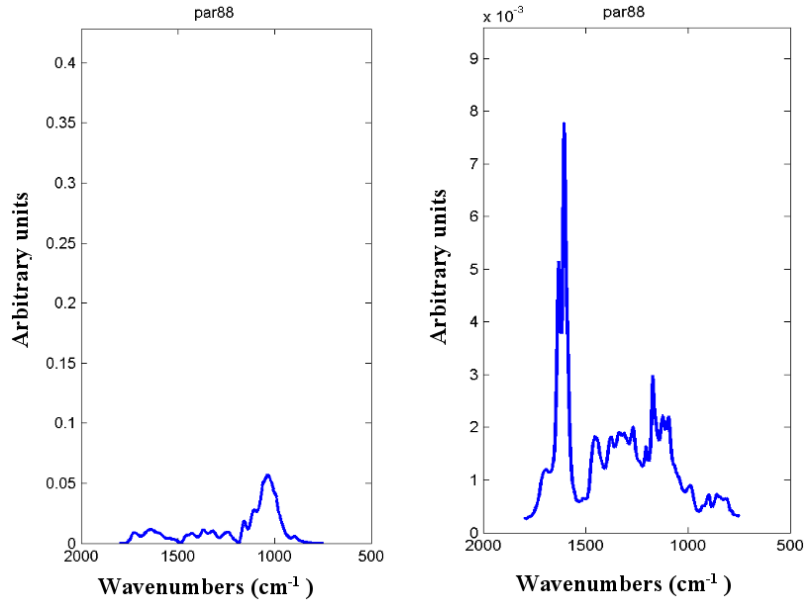
b. Raman co-inertia loading 1

c. Co-inertia component 1

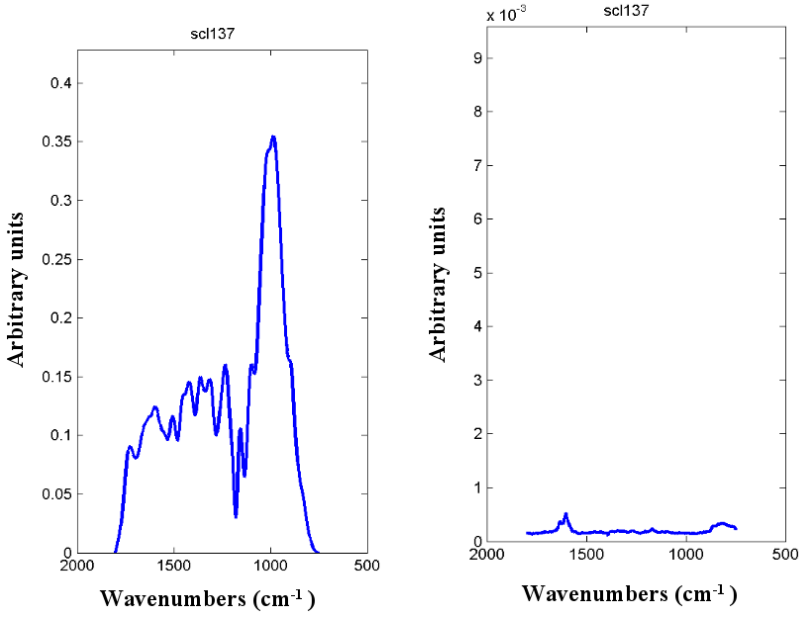
Figure 4.17: (a and b) Multiple co-inertia first loading plot for infrared and Raman spectroscopy. (c) In addition, score plots are also shown to facilitate interpretation of the peaks with respect to their cell-type.

The spectral profile (first loading) is intense for the sclerenchyma both in Raman and infrared (Figure 4.17). The negative peaks correspond to pectins (1022 cm^{-1}), esters (1230 cm^{-1}), and lignins (1419 and 1508 cm^{-1}). The positive Raman peaks correspond

to Ferulic acid (1221, and 1286 cm^{-1}), and cellulose (1407 cm^{-1}). The negative ones correspond to 1096 cm^{-1} (cellulose, xylans, glucomannans), 1171 cm^{-1} (lignins), 1267 cm^{-1} (arabinoxylan), 1338 cm^{-1} (cellulose), 1603 and 1632 cm^{-1} (lignins).



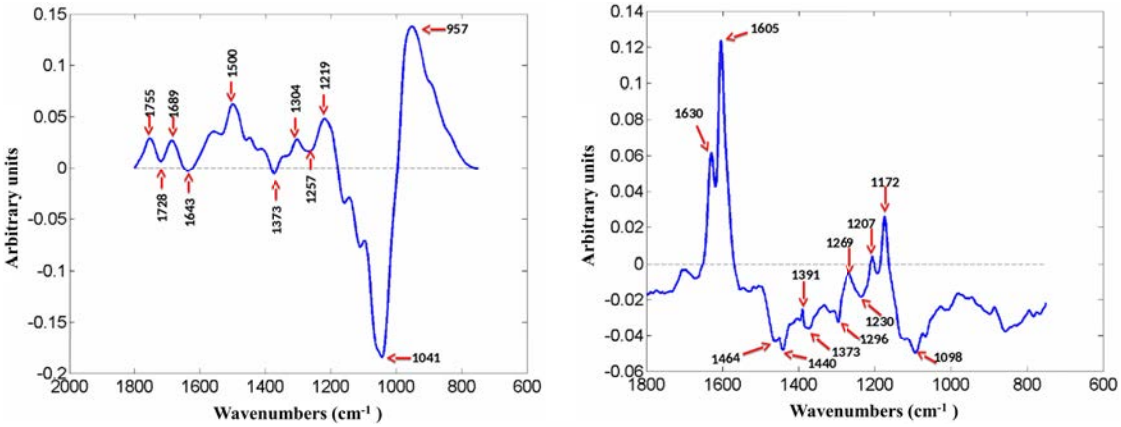
a. Infrared spectra **b.** Raman spectra



c. Infrared spectra **d.** Raman spectra

Figure 4.18: Spectral profiles of par88 and scl137 corresponding to the parenchyma and sclerenchyma region. (a and b) Infrared and Raman spectra corresponding to the parenchyma region. Clearly, there is a deformation in case of the obtained infrared spectra. (c and d) The profile from scl137 is deformed in case of the Raman spectra.

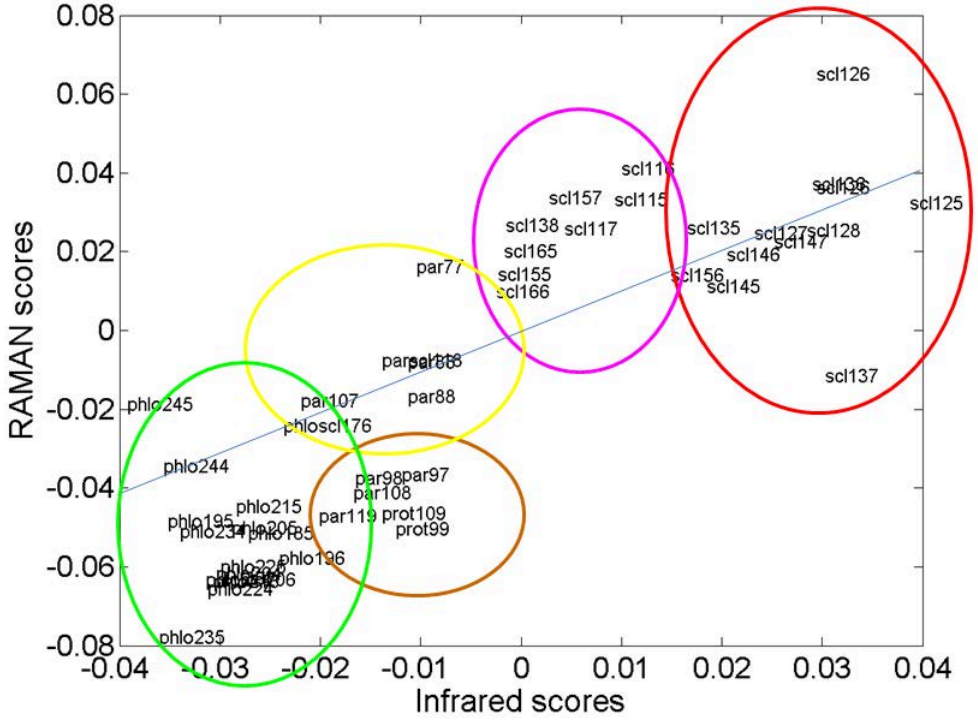
Spectra from the parenchyma and sclerenchyma region (par88 and scl137) show some discrepancies. When we observe the paired spectral profiles in Figure 4.18, we notice that spectra par88 varies between the Raman and infrared. Spectra (scl137) is a difficult region to obtain the Raman spectra and some deformation has occurred (Figure 4.18). Similarly, the spectral profiles of scl 125, 127, 128, and 135 were observed and attributes to the poor correlation between infrared and Raman scores.



a. IR co-inertia loading 2 **b.** Raman co-inertia loading 2

Figure 4.19: Multiple co-inertia second loading plot for infrared and Raman spectroscopy. In addition, score plots are also shown to facilitate interpretation of the peaks with respect to their cell-type.

In figure 4.19, we observe the second co-inertia component for the two spectroscopies and the correlation coefficient between the block scores is 0.85 (Table 4.2). The positive peaks of the infrared at 953, 1219, and 1500 cm^{-1} correspond to a lignin profile. Similarly for the Raman profile, peaks at 1603, 1630, 1701 and 1171 cm^{-1} correspond to a lignin profile. Contribution of Raman is stronger and it may be due to the strong lignin profile. Clustering of the spectral points were observed and marked in the score plots (Figure 4.20) to understand the strong correlation between the two techniques.



a. Co-inertia component 2

Figure 4.20: Multiple co-inertia score plot for the second component. The score plot was further interpreted by marking manually the observed clusters. The identified clusters are represented in green (phloem), yellow (parenchyma), brown (parenchyma+proteins), pink (sclerenchyma), and red (another set of sclerenchyma clusters).

In addition, we also studied the score spatial interpretations using the score images obtained by combining the high-resolution brightfield image with the refolded scores. From the score image (Figure 4.21), we see that the positive peaks correspond to sclerenchyma and the negative ones to phloem.

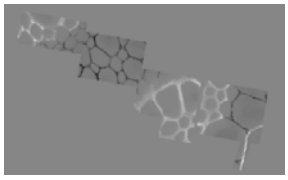
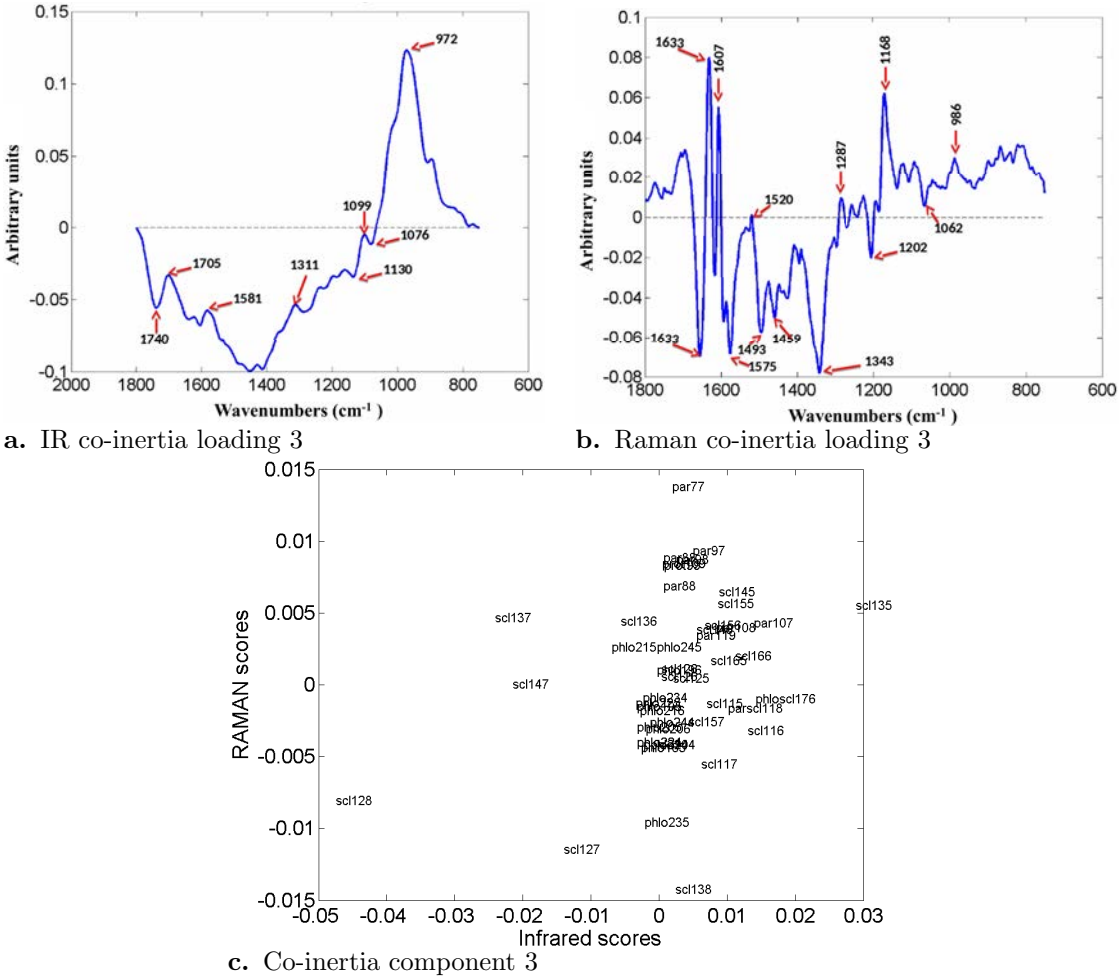


Figure 4.21: Score image of the multiple co-inertia second component.

In the third component (Figure 4.22), the correlation coefficient between the block scores is 0.30. The peak looks deformed and hence not interpreted.



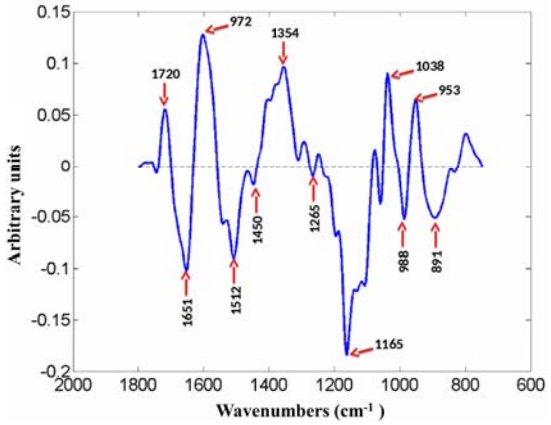
a. IR co-inertia loading 3

b. Raman co-inertia loading 3

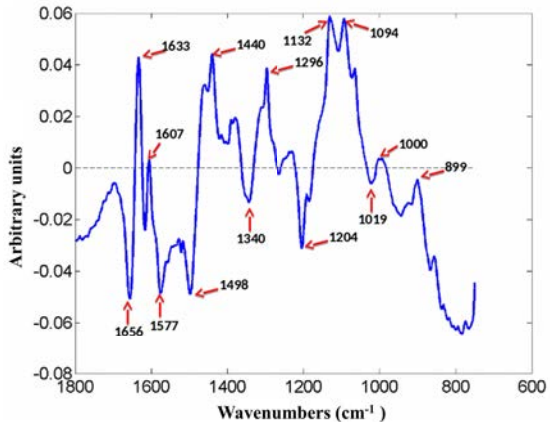
c. Co-inertia component 3

Figure 4.22: (a and b) Multiple co-inertia third loading plot for infrared and Raman spectroscopy. (c) In addition, score plots are also shown to facilitate interpretation of the peaks with respect to their cell-type.

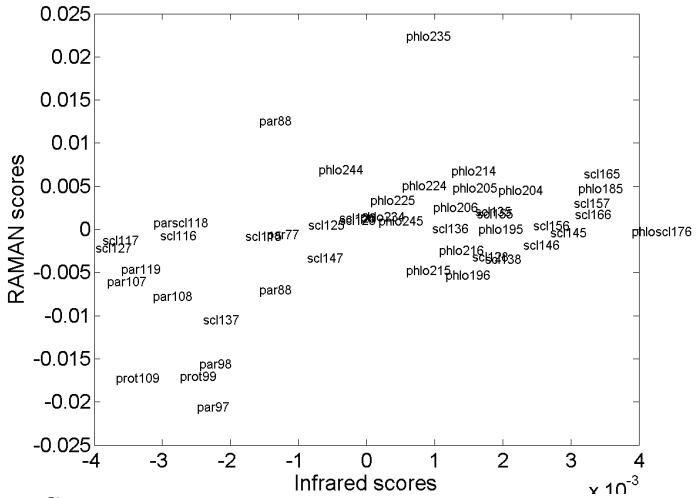
The fourth component has a correlation coefficient of 0.46 between the block scores (Figure 4.22). From the score plots, we see a clear separation between the parenchyma and phloem cells. The Raman peaks at 899 (arabinoxylans), 1094 (cellulose and arabinoxylans), 1132 (cellulose and arabinoxylans), 1607 (lignins), and 1633 (lignin) cm^{-1} have a positive profile. The negative peaks at 1340 and 1656 cm^{-1} correspond to cellulose and lignin respectively. The positive infrared peaks correspond to lignins at 1720 cm^{-1}. The negative infrared peaks correspond to 891(pectins), 1165 (cellulose), 1512(lignins) cm^{-1}.



a. IR co-inertia loading 4



b. Raman co-inertia loading 4



c. Co-inertia component 4

Figure 4.23: (a and b) Multiple co-inertia fourth loading plot for infrared and Raman spectroscopy. (c) In addition, score plots are also shown to facilitate interpretation of the peaks with respect to their cell-type.

4.4 Discussion

In the plant domain, investigating the structure and assembly of plant cell wall polymers is very complex. To this end, it is necessary to use methods which probe the cell walls *insitu* and preserve the chemical and structural information. Hyperspectral imaging is one such tool to perform *insitu* chemical analysis and it has been used to reveal the occurrence of particular compounds in the biological material (Chylińska et al., 2014, Sun et al., 2011, Agarwal, 2014). Hyperspectral images provide three dimensional data structures with two spatial dimensions and one spectral dimension. Moreover, the spectroscopic techniques could be either complementary (infrared and fluorescence) or partly redundant (infrared and Raman techniques).

Up to now, the most commonly used multivariate methods for the analysis of infrared and Raman spectra includes principal component analysis, vertex component analysis, and multivariate curve resolution (Chylińska et al., 2014, Bajorski, 2009, Jones et al., 2012, Gierlinger, 2014). Moreover, studies by Chylińska et al. (2014) and Gierlinger (2014) focused on the visualization and distribution of polysaccharides acquired from Raman spectroscopy. In other cases as in Szymanska-Chargot and Zdunek (2013), PCA was used to analyze the FT-IR spectra for the characterization of cell wall residues in fruits and vegetables. To take advantage of the different spatial resolutions of various spectroscopy techniques, a joint analysis of the data is essential. To this end, Allouche et al. (2012b) conducted a joint analysis of infrared and fluorescent spectroscopy data using multivariate inter-battery tucker analysis. The strong point of this analysis is that Allouche et al. (2012b) extended the multivariate inter-battery method to analyze jointly a three-way and a two-way data table. The three-way data table in this case is a typical example of multiway data analysis as detailed in Section 2.1 of this thesis. In addition, another study on the joint data analysis of three different spectroscopies (Infrared, Raman, and fluorescent) was also done. The main drawback of this study was the use of fluorescence spectroscopy as very few molecules are naturally luminescent. Moreover, the common regions between three different techniques had to be recovered. In addition, only three different cell-types were profiled namely, the xylem, phloem, and sclerenchyma cell-types from maize stem cross-sections (Allouche et al., 2012a). Hence, the joint analysis of different spectroscopic techniques is limited to the studies of Allouche et al. (2012a,b).

In this chapter, the focus is to estimate the raw cell wall content and composition from several tissue proportions at different levels of spatial resolution. Here, two different spectroscopy techniques were used (infrared and Raman) to jointly analyze the complexity of cell wall polymers, because coupling spectral domains could heighten differences in composition not observable using one spectral range. Infrared spectroscopy has proven to be relevant in identifying main cell wall components such as cellulose, pectins, xyloglucans, arabinose, and galactose. To complement the information obtained from FT-IR, Raman spectroscopy is used to identify compounds at a resolution of 1 μm^2.

The images used in the analysis were provided by INRA, Nantes and covered a region in the maize internode cross-section that contained mainly xylem, sclerenchyma, phloem, parenchyma and parenchyma border cells. For each spectral domain, pre-processing was done in two steps; the first being the corrections specific to each spectroscopy and the second is a normalization step that took into account both the spatial and spectral information. The procedure avoids normalizing the spectra recorded in the holes of the maize stem cells by taking the spatial neighborhood into account. This principle is important

to preserve the spatial resolution and in particular, the variations of intensities observed in cell junctions and holes when compared to walls were taken into account.

Initially, the application of spectroscopic techniques was also limited by the large amount of data obtained as well as overlapping band components. However, the development of computerized systems and the use of statistical tools have enabled easier processing of the information obtained from the spectra, thus increasing the interest and usefulness of these spectroscopic techniques. Principal components analysis (PCA) has been used to study the variability within each spectrum and is then represented as a smaller set of values (axes) termed principal components. In addition, refolding of the scores was done and represented as grey-level images. They are used to interpret the individual contribution for each component and helps in visualizing the distribution of the chemical composition in different cell-types. In the infrared spectra, intense and large peak were observed for polysaccharides such as xylans, cellulose and xyloglucans. In case of the Raman spectra, peaks in xylem cells correspond to an intense lignin profile. Phloem Raman spectrum correspond strongly to cellulose, and arabinoxylans whereas sclerenchyma Raman spectrum correspond mainly to lignins. Parenchyma close to the sclerenchyma had a polysaccharide profile while parenchyma cells correspond to lignin profiles. Raman and infrared results clearly revealed that lignin or polysaccharide profile exists depending on the investigated cell-types.

In addition to the use of PCA to investigate particular cell-types, we focused on jointly analyzing the hyperspectral images from both techniques using multiblock methods. Since both the spectroscopic techniques were of different resolutions, coupling of the different spectral domains was not straight-forward. The procedure consisted of identifying the region of interest where the spectra were jointly acquired by projecting each spectral image onto the reference image using image registration techniques. After pairing the infrared and Raman images, the regions mapped in common were recovered. The scale and rotation factors listed in Table 4.1 were used to assess the transformation matrices which allowed computing the coordinates of the Raman pixels homologous to the infrared pixels. The scale factor to register the spectral image onto the reference image ranged between 3.6 - 3.9 in case of the six Raman images and was 22.69 for the infrared image. Values of scale, rotation and translation corresponding to cross correlation coefficients values between 0.5 - 0.7 were obtained.

The recovered 47 pixels of the Raman images which had corresponding similar pixels in the infrared were used as 'n' observations in the data tables and were submitted to multiple co-inertia analysis. The two block data tables were of the dimension 47×1090 (Raman spectra $\times$ wavenumbers) and 47×274 (infrared spectra $\times$ wavenumbers), a typical

example of 'n < p+q' setting. Hence, the use of MCIA is more appropriate in the 'n < p+ q' setting when compared to methods like CCA. In addition, MCIA focuses on extracting the unique and common information between the two datasets unlike CCA which focuses on maximizing the correlation between two datasets. Once again, this emphasizes the use of appropriate methods based on the biological question under investigation. The percentage of variances and data table contributions detailed in Table 4.2 were described by principal components and block components. The sum of contributions to the first global loading is 100 over the two data tables wherein 51 % are due to the infrared table, and 49 % are due to the Raman data table. Similarly in case of the second global loading, 23 % and and 77 % are contributed by infrared and Raman, respectively. The sum of contributions of global scores over both data tables is equal to the first eigenvalue of the principal component analysis of the merged block scores table, and therefore less than 100 %. The first block component for infrared accounted for 96.6 % of the total variance and 91.8 % for Raman. As expected from the orthogonal loadings computed using multiple co-inertia analysis, the percentage of variance described by block components obtained by multiple co-inertia components follow the percentage of variance obtained by principal components. This was consistent for all the five components of infrared and Raman.

Furthermore, three components (first, second and fourth) were interpreted and discussed. The first component (Figure 4.17) described intensity variations caused by the occurrence of holes and cell walls in case of both spectroscopies. Interpreting the second component, we observed lignin signals opposed to polysaccharide signals (Figure 4.19 and 4.20). The score spatial interpretations of the lignin peaks correspond to the sclerenchyma region (Figure 4.20). This observation is consistent with previous studies that sclerenchyma cells have strong lignin signals. Here, contribution of Raman is stronger than that of the infrared probably due to strong lignin profile in the former. In the fourth infared loading (Figure 4.23), the positive peaks have strong lignin signals and the negative ones correspond to pectins, cellulose, and lignins. Whereas, in case of the Raman loading the positive peaks pertain to cellulose, arabinoxylans, and lignins. In addition, from the score images we were able to see a clear separation between the parenchyma and phloem cells (Figure 4.23).

Thus, with the use of multiblock methods and adequate interpretation tools, it is possible to prove that each of the different techniques used to study similar cell-types could contribute disproportionately to form the global components. The joint analysis of the hyperspectral image data is useful to characterize biological material on the basis of data tables representing various facets of their chemical properties.

Chapter 5

Case study 3: Co-ordination and divergence of cell-specific transcription and translation of genes in Arabidopsis root cells

5.1 Specific rationale and objectives

Arabidopsis thaliana has become one of the most widely used plant model organisms in basic research, largely due to the availability of resources (Hamilton and Buell, 2012, Mochida and Shinozaki, 2010). Recently, efforts have been made to monitor gene expression at the level of specific cell-types and across different developmental stages in Arabidopsis to obtain a deeper and systematic understanding of the underlying cellular processes (Edwards and Coruzzi, 1990, Birnbaum et al., 2005, Brandt, 2005, Shen-Orr et al., 2010, Wang and Jiao, 2011). These studies have resulted in the accumulation of distinct data-types which provide a different, partly independent and complementary view of the whole genome.

The relationship between information elements (genes/transcripts) and functional elements (metabolites) has been studied by integrated transcriptomic and metabolomic analyses (Tohge et al., 2005, Hannah et al., 2010, Osorio et al., 2012, Brink-Jensen et al., 2013). In addition, there are numerous integrative studies of transcriptomic and proteomic datasets (Kleffmann et al., 2004, Baginsky et al., 2010, Vogel and Marcotte, 2012, Pan et al., 2012). However, these mass spectrometry-based approaches only reveal a limited

coverage of the proteome and metabolome, i.e. allowing identification of a few hundred metabolites (Schauer et al., 2006) and a few thousand proteins in plants (Petricka et al., 2012). Despite the limited coverage of the proteome, recent combined transcriptomic and proteomic analyses have reported weak correlation of protein and mRNA abundances (Hack, 2004, Wang et al., 2010b). As an intermediate step in the flow of genetic information in a biological system, the level of translational control determines quantitative variation of the proteome together with protein degradation (Tebaldi et al., 2012). In particular, the composition of the translatome is based primarily on translation initiation, i.e. the loading of ribosomes on messenger ribonucleoprotein particles (mRNPs) to form polysomes, and secondarily on translation elongation (Tebaldi et al., 2012). Finally, correlation of levels of transcripts and polysomal-bound mRNA abundances allow inferences about gene activities and the conversion of its mRNA into a protein.

This chapter elaborates on the comparative study of cell-specific transcripts in plants. The relative simplicity of Arabidopsis root anatomy and availability of cell-specific expression profiling data from developmental zones has made it appropriate for this study (Bevan and Walsh, 2005, Benfey et al., 2010). Arabidopsis also serves as a powerful model system for plant cell wall research, such as the identification of cell wall biosynthesis-related genes. Moreover, Arabidopsis has also been extensively used to study the root cell wall biology and understand how cell walls are developmentally controlled in different cells (Milioni et al., 2002, Liepman et al., 2010). To complement these studies, we investigated coupled transcription and translation by use of publicly available root datasets. Using cell-type-specific datasets of the root transcriptome and translatome of Arabidopsis, a systematic assessment was made of the degree of co-ordination and divergence between these two levels of cellular organization. Although the previously described canonical methods of data integration are efficient in case of large datasets, here, we wanted to study the variation in cell-type specific gene expression patterns across the two system levels. Hence, the computational analysis considered correlation and variation of expression across cell-types at both system-levels, and also provided insights into the degree of co-regulatory relationships that are preserved between the two processes. We elucidated the genome-wide correlation of cell-specific transcription and translation for the majority of genes in the Arabidopsis genome. We present evidence for translational prioritization of transcripts of cell-wall-related gene families and root-related biological processes.

5.2 Materials and methods

5.2.1 Arabidopsis root transcriptome and translatome gene expression datasets

In this study, two microarray datasets of A. thaliana root samples were employed characterizing gene expression and translation levels of tissues and cell-types. While the first dataset comprises a discrete global map of total mRNA levels, i.e. the transcriptome (Birnbaum et al., 2005, Brady et al., 2007), the latter measured polysome-associated mRNAs, i.e. the translatome (Mustroph et al., 2009), in a variety of cell-types and developmental stages of the root.

The two datasets were generated using different experimental protocols and originate from two different laboratories. For the transcriptome dataset, fluorescence activated cell sorting (FACS; (Iyer-Pascuzzi and Benfey, 2010)) of Arabidopsis radial sections of root samples under the control of cell-type-specific promoters was used to profile transcript levels. Affymetrix ATH1 microarrays were used as the platform for gene expression profiling. Genome-wide transcriptomic profiles of eight green fluorescent protein (GFP)-marked cell populations with two-three replicates each were obtained, that in combination with a previous study of 11 microarray expression experiments yielded a transcriptomic expression atlas. The expression atlas profiles the expression of 14 non-overlapping Arabidopsis root cell-types targeted by 19 promoters (Brady et al., 2007, Birnbaum et al., 2003). For the transcriptome data available from Birnbaum et al. (2003), cell-type and tissue-specific expression was obtained by protoplasting of plant roots expressing GFP in specific cell-types. The raw data [accession number GSE ID 8934 available via Gene Expression Omnibus (GEO; (Barrett and Edgar, 2006))] from the radial root sections were downloaded in the form of CEL files for further analysis.

The translatome dataset was obtained by the immunopurification of ribosome-associated transcripts from Arabidopsis root cells and the immunopurified mRNAs were hybridized to the Affymetrix ATH1 microarray platform. In this study, the immunopurification was extended by using developmentally regulated promoters to drive the expression of FLAG-tagged RPL18 lines allowing the generation of 21 cell-specific populations in root and shoot (Mustroph et al., 2009). While the complete translatome data additionally include a stress condition (hypoxia), for the following computational analysis, raw data with an accession number GSE ID 14502 comprising only root control samples were used. An overview of all promoters used in the transcriptome and translatome data together with their intended tissue specificity can be found in Table S4 (Appendix B).

Furthermore, to avoid artefacts arising from different normalization techniques, the raw data from these two datasets were pre-processed using the same normalization strategy. Here, the robust multichip average (RMA) method was used to conduct pre-processing, e.g. removal of background noise and quality control, subsequent probe summarization and adjustment by quantile normalization (Irizarry et al., 2003). Applying the same normalization strategy, both datasets were jointly and independently normalized to study the influence of the normalization to the final results.

Lastly, given the availability of 19 promoters in the transcriptome and 10 different promoters in the translatome dataset (see Table S4, Appendix B), a common set of identical cell-types corresponding to identical promoters (see Appendix C: Supplementary text for the nucleotide sequences of the promoters) in both datasets - namely the phloem companion cells, root vasculature, quiescent center, cortex, and non-hair cells/root atrichoblast epidermis were identified. Moreover, because in some cases different promoters were used to drive gene expression of the same cell-type on the two system-levels, a mapping of promoter and target cell-type was conducted using a literature survey.

5.2.2 System-level analysis of cell specific mRNA levels

Genome-scale system-level comparisons between the transcriptome and translatome were conducted by quantifying the similarity of expression (total mRNA) and translation (polysome-associated mRNA) levels of genes across the same set of cell-types. Here, Pearson correlation coefficient (PCC) was used to assess this similarity (Stigler, 1989). The statistical significance of observed PCC values was further assessed by creating 1000 bootstrapped datasets of the transcriptome and the translatome, respectively. Using the available data for all 19 and ten promoters for transcriptome and translatome as background, a bootstrapped dataset of equal size was randomly selected without replacement. As each bootstrapped group comprises a random mixture of cell-types, any variations in deriving PCCs would exist mainly as a result of cell-type-specific expression rather than differences in translatome and transcriptome expression levels, thus resembling an adequate null model. For each of these bootstrapped data, and for all genes, PCC of translatome and transcriptome was calculated. Z-scores were calculated for each observed PCC value by subtracting the mean and dividing by the standard deviation of the corresponding gene's PCC value obtained from the bootstrapping analysis (Figure 5.1). Those genes which exhibit a high positive PCC value that corresponds to a Z-score ≥ 1.96 comprise the set of genes that exhibit a high degree of correlation between transcription and translation. Likewise, PCCs of high negative value, i.e. a corresponding Z-score of ≤ -1.96,

correspond to genes that display a high degree of uncoupling of cell-specific transcription and translation. Note that an absolute Z-score of 1.96 corresponds to a statistical significance level of 5 % in the case of a two-tailed test (Sokal and Rohlf, 1995). Additionally, following suggestions of Huttenhower et al. (2006), PCC values were adjusted using Fisher transformation, resulting in normal distributions of PCC values irrespective of the dataset analyzed, further allowing for cross-dataset comparisons.

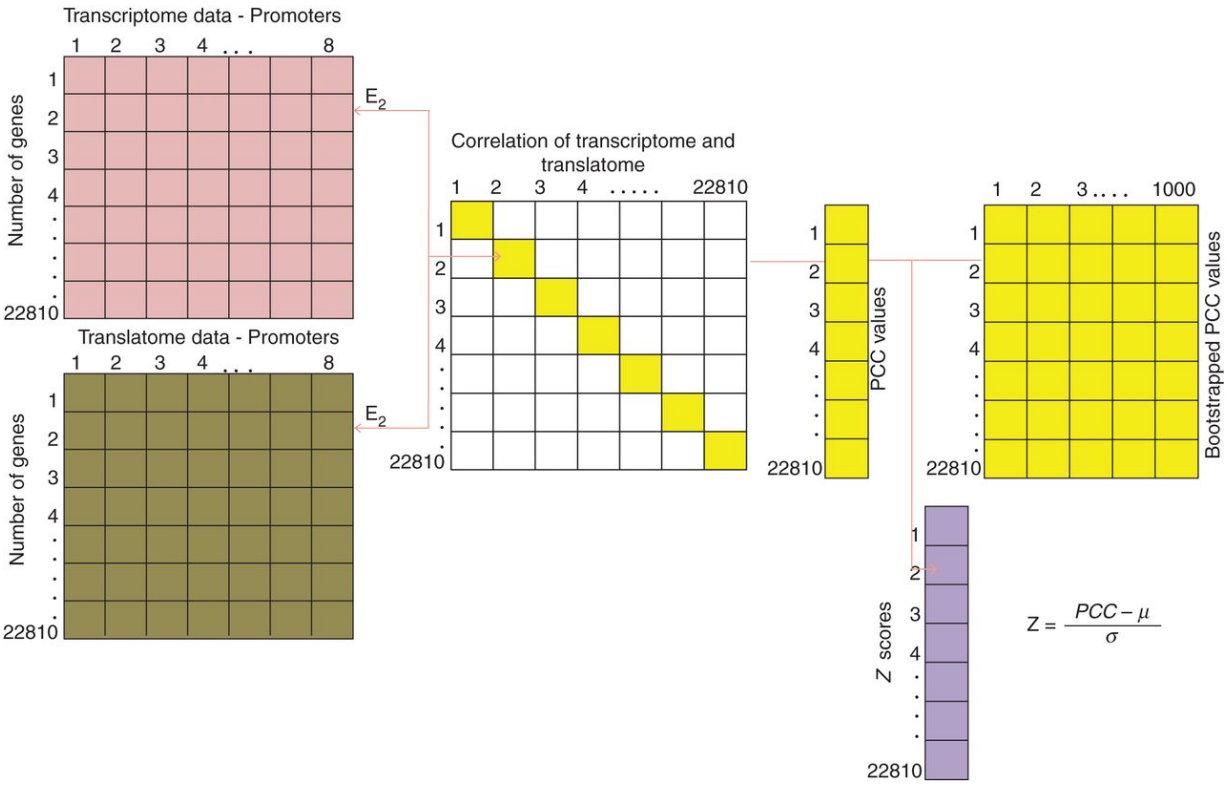


Figure 5.1: Analysis of similarity of cell-specific mRNA levels on the level of transcriptome and translatome in Arabidopsis root cells using identical and common cell-types for both datasets. The Pearson correlation coefficient (PCC) between expression and translation levels for each gene was computed. Eight promoters were identified that drive gene expression in common cell-types in both datasets. By comparing the observed PCC value of each gene with PCC values obtained from bootstrapped data, Z-scores were computed for each of the corresponding genes.

Finally, a characterization of the biological processes of these genes was conducted by gene set enrichment analysis (GSEA; (Subramanian et al., 2005)). Here, the gene ontology (GO) was used to obtain functional gene annotation used for GSEA (Ashburner et al., 2000). Specifically, the sub-ontology Biological Process (GO-BP) was used to derive overrepresented GO terms and further pre-processed following the considerations given by Klie and Nikoloski (2012). For statistical testing, the hypergeometric distribution was

used to test for the probability that a specific set of genes is annotated with the same GO term by considering the background distribution of GO terms (Rivals et al., 2007).

5.2.3 Identification of altered regulation of gene expression across system-levels

In addition to the analysis of cell-type-specific gene expression/translation similarity by means of PCC, the genome-wide co-expression structure for each gene was investigated by deriving co-expression networks (Butte and Kohane, 2003). Specifically, both a co-expression and a co-translation network were constructed based on the common cell-type-specific transcriptome and translatome data to further identify the co-expression/co-translation relationships of genes which differ between the two networks. In the co-expression and co-translation networks, nodes correspond to genes and edges (connections) are present between any two nodes that are significantly connected resulting in a fully connected network (Chartrand, 1985). As a consequence, the similarity of a gene's neighbourhood within the co-expression and co-translation in the network reflects the extent to which expression and translation relationships of groups of genes are coupled. Accordingly, changed network topology suggests altered regulation or regulatory uncoupling of co-expression and co-translation relationships.

All edges were weighted, where the weight of an edge adjacent to two nodes/genes corresponds to the value of the PCC between the corresponding mRNA levels of both genes. For the co-expression network this weight is defined by the PCC of expression levels; likewise, in the co-translation network, this weight is defined by the PCC of translation levels in the cell-types of interest. Furthermore, the concept of expression conservation (EC) was used to assess the similarity of co-expression relationships for all genes contrasting the translatome and transcriptome networks (Dutilh et al., 2006). Both networks can be represented by an adjacency matrix that can then be compared. Technically, this procedure simplifies the computation of the PCC of the same row corresponding to a gene in both matrices (cf. Figure 5.2).

The statistical significance of the difference of a gene's EC score was assessed by creating a series of n=1000 random co-expression networks for translatome and transcriptome by selecting two random equal-sized sets of cell-type-specific translatome and transcriptome data. Based on this procedure, an empirical null distribution of random EC scores, genes that exhibit a statistically significant low EC score can be derived by calculation of Z-scores: genes with a positive EC score and a Z-score ≥ 1.96 comprise genes with highly conserved co-expression relationships. Correspondingly, genes with a low EC

score (matching to Z-scores ≤ -1.96) exhibit low or no conservation of co-expression/co-translation. Again, to characterize processes over-represented within those genes with low and high EC scores, a subsequent characterization of biological processes, GO-BP by GSEA, was conducted.

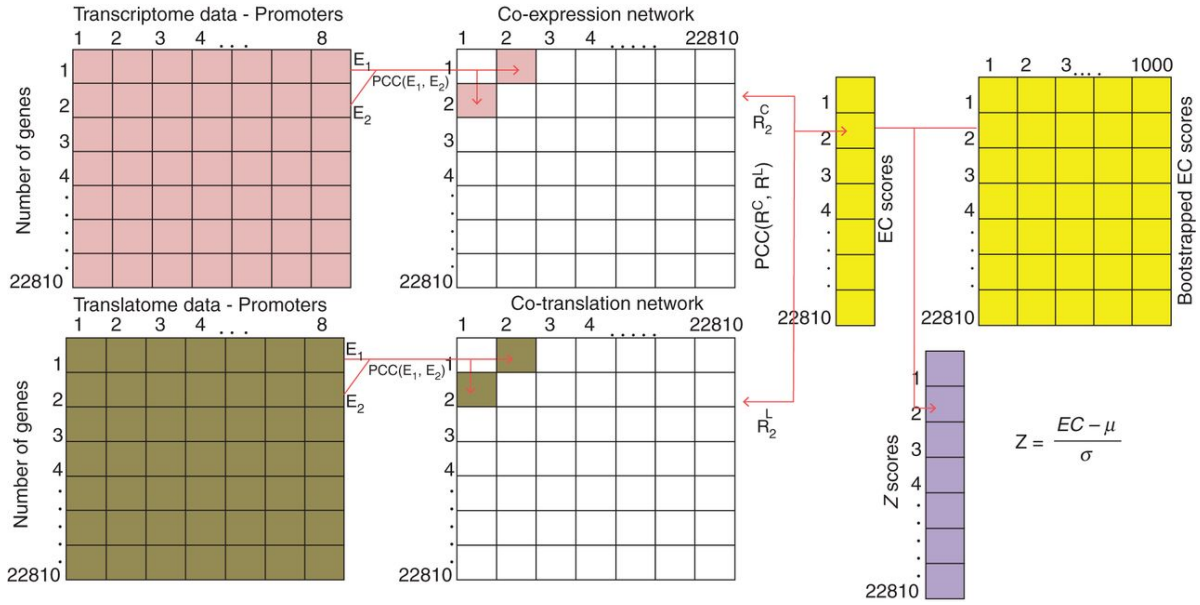
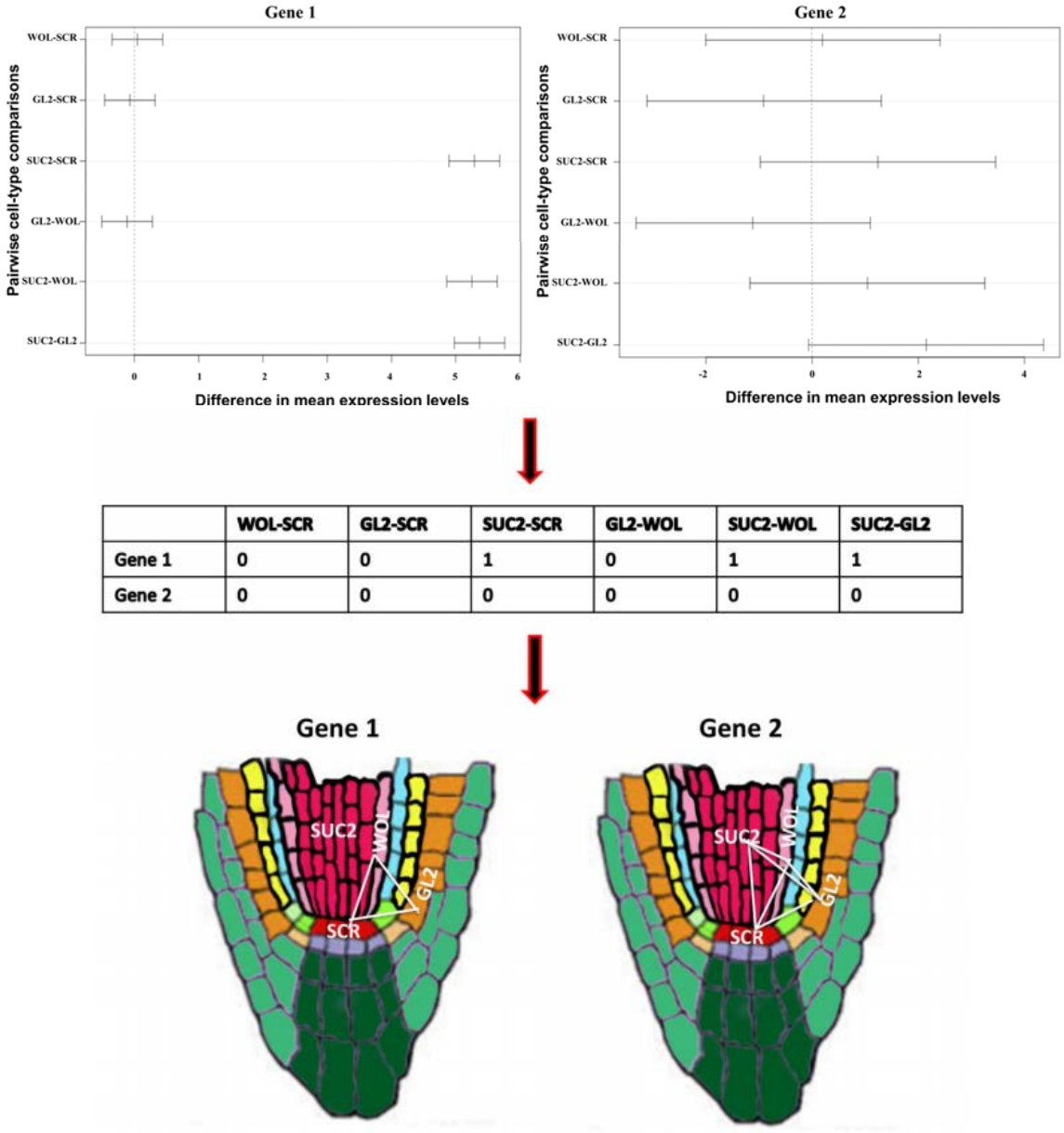


Figure 5.2: Co-expression/co-translation networks (represented by adjacency matrices) for five cell-types using the common eight promoters. Two fully connected co-expression/co-translation networks were generated separately using the transcriptome and translatome datasets. R_2^C and R_2^L represent the second row of the transcriptome and translatome adjacency matrix, respectively. An expression conservation score (EC score) was then defined as the PCC between the co-expression and co-translation networks. Further, the same analysis was performed on a set of 1000 bootstrapped co-expression/co-translation networks generated from bootstrapped datasets. Rewired genes were identified using Z-scores.

To investigate co-expression and co-translation relationships on a global scale, we computed the correlation of the adjacency matrices of both networks. The similarity of the co-expression and co-translation network is defined by their topology. As with PCC, the value of this full matrix (or network) correlation falls in the interval [-1,1] (Swanson-Wagner et al., 2012). Again, the statistical significance of the observed value of the full matrix correlation was assessed by selecting random cell-type expression and translation data to generate 1000 pairs of networks. Subsequently, the observed similarities were then compared with values obtained from these bootstrapped networks.

5.2.4 System-level analysis of cell-type specificity



	WOL-SCR	GL2-SCR	SUC2-SCR	GL2-WOL	SUC2-WOL	SUC2-GL2
Gene 1	0	0	1	0	1	1
Gene 2	0	0	0	0	0	0

Figure 5.3: Conversion of pairwise differences in cell-type-specific expression levels derived by Tukey's HSD test, to cell-type similarity networks. Gene 1 is overexpressed in phloem companion cells (SUC2), rendering this cell-type different from the remaining 3 cell-types (represented by no edge). In case of gene 2, there exists no significant difference in cell-type specific expression levels. As a result, all cell-types are similar, further represented by edges.

Differentially expressed/translated genes (for simplicity jointly referred to as DE genes) displaying statistically significant mean differences in expression levels across the common

cell-types were identified by performing an analysis of variance (ANOVA; (Kerr et al., 2000)). The ANOVA was performed independently for the transcriptome and translatome data. Moreover, while an ANOVA identifies genes exhibiting significant mean expression differences across all cell-types, post-hoc tests, such as Tukey's honest significant difference (HSD) were applied to further derive statistical significance of pairwise mean differences (Tukey and Braun, 1985). In this study, a series of Tukey's HSD tests were performed for those genes determined to be differentially expressed by ANOVA after a Benjamini and Hochberg false discovery rate (FDR) correction (Benjamini and Hochberg, 1995). The significance level was set to 5 % for both ANOVA and Tukey's HSD.

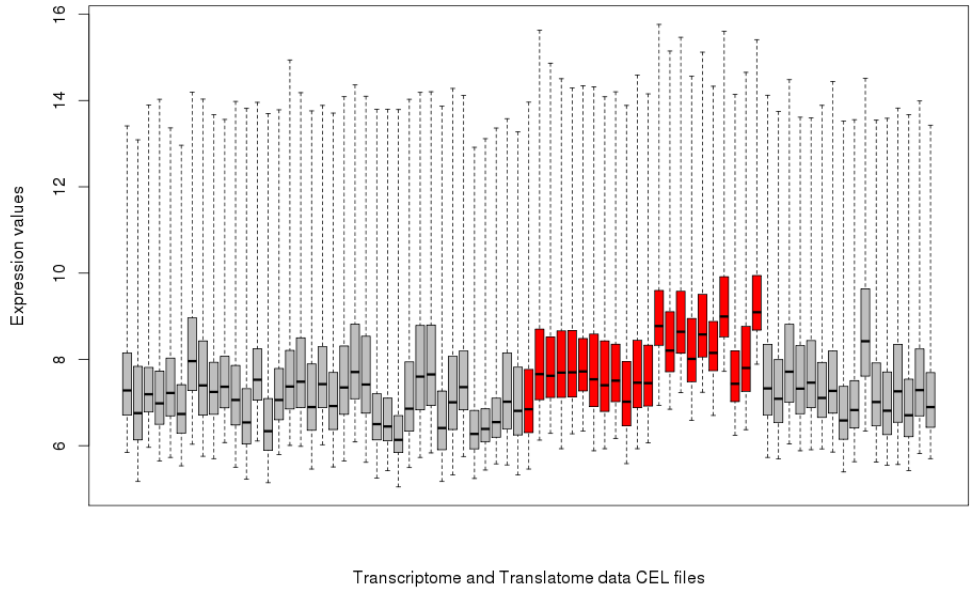
To summarize the pairwise differences of cell-type expression levels on translatome and transcriptome, cell-type similarity networks were constructed for each gene (Figure 5.3). In this cell-type similarity network, nodes correspond to the respective cell-type of the Arabidopsis root. An edge between two nodes indicates no significant mean difference of expression values of the investigated gene. Accordingly, the absence of an edge indicates a significant difference and thus dissimilarity of the adjacent nodes. Given a certain number of cell-types, it is possible to obtain a total of 2^{m} possible network topologies or configurations, which are defined as network motifs (Milo et al., 2002). Furthermore, the statistical significance of a particular number of occurrences, i.e. the number of genes that coincide with any particular network motif, was assessed empirically by permutating the data obtained for pairwise cell-type mean expression differences (i.e. the results of the Tukey HSD tests) for all DE genes.

5.3 Results

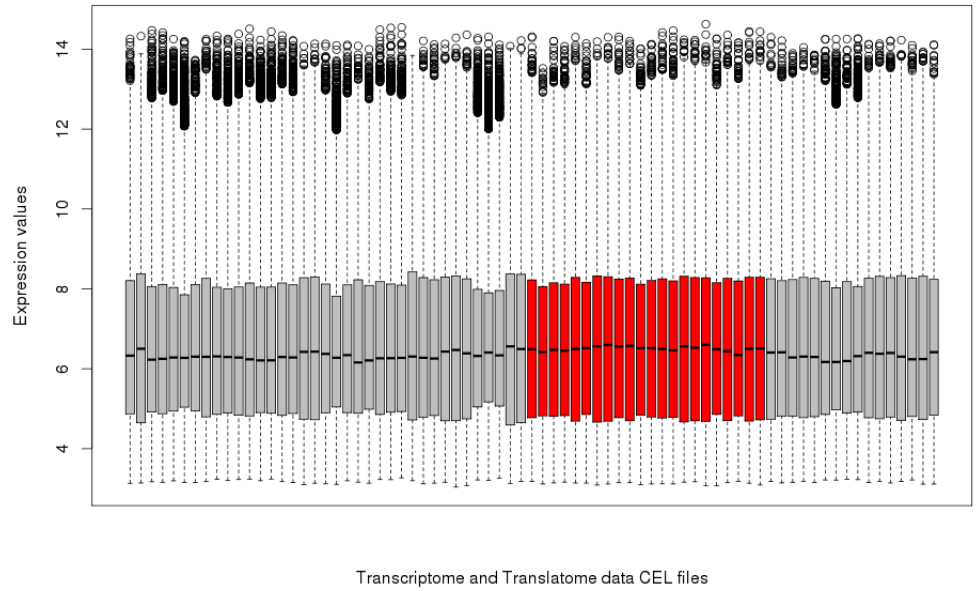
In this chapter, we attempted to compare and assess the relationship between cell-type-specific transcriptome and translatome data of Arabidopsis roots. In particular, we were interested in testing to what degree gene expression and translation patterns were conserved in these samples.

5.3.1 Normalization of datasets

The obtained raw data were jointly normalized using RMA (Figure 5.4) to allow comparisons between the gene expression and translation datasets (Irizarry et al., 2003).



a. Pre-normalization



b. Post-normalization

Figure 5.4: Normalization of the CEL files. (a) The transcriptome dataset has 19 promoters in total with replicates (53 CEL files in total) which are shown in grey and the translatome dataset has 10 promoters with replicates (22 CEL files in total) coloured in red. The pre-normalized CEL files are displayed here. (b) RMA normalization of the CEL files from the transcriptome and translatome was done together.

For the transcriptome datasets, data from Birnbaum et al. (2003) and Brady et al. (2007) were used. Here, 19 cell- or tissue-type-specific root promoters driving GFP had been used in combination with cell sorting to obtain a transcriptome map of root cells. For the translatome datasets, data from Mustroph et al. (2009) were used. In this experiment,

ten cell- or tissue-type-specific root promoters driving a FLAG-tagged RPL18 to achieve a translatome map of the root cells (the full list of promoters is available in Table S2, Appendix B).

First, we considered the effect of separate RMA normalization of the datasets. We observed only slight differences in the distributions of probe log-intensities over all microarrays belonging either to the transcriptome or to the translatome datasets indicating comparable average signal intensities between datasets and limiting the possibility of technical bias between the two datasets. We also affirmed high reproducibility of the biological replicates (correlation between replicates of 0.96 ± 0.03 in the transcriptome and 0.98 ± 0.01 in the translatome).

5.3.2 Promoter/cell-type mapping

Four identical promoters had been used to obtain the transcriptome and translatome data (Birnbaum et al., 2003, Brady et al., 2007, Mustroph et al., 2009) and these therefore served as a first platform for our study (Table 5.1).

Cell-type	Transcriptome	Translatome
Phloem companion cells	SUC2 (At1g22710), APL	SUC2 (At1g22710), SULTR2
Root vasculature	WOL (At2g01830)	WOL (At2g01830), SHR
Quiescent center	AGL42, J0571, SCR (At3g54220)	SCR (At3g54220)
Cortex	CORTEX	CO2, PEP (based on whether it is meristamatic, elongation or maturation zones)
Non-hair cells/ root atrichoblast epidermis	GL2 (At1g79840)	GL2 (At1g79840)

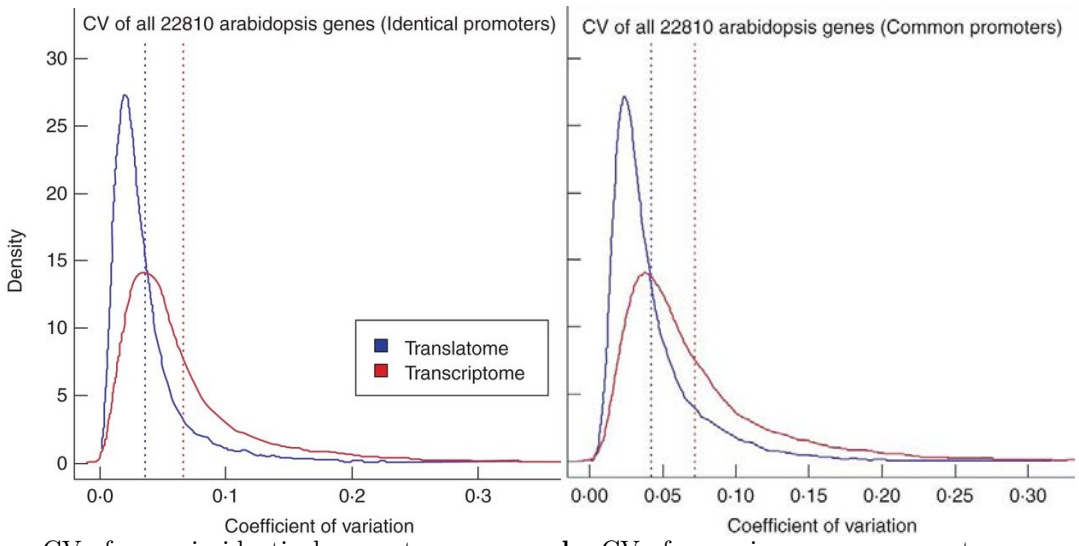
Table 5.1: List of promoters and cell-types common to the transcriptome and the translatome datasets. Identical promoters in both datasets are underlined. The promoters used in the analysis include SUC2 (Sucrose transporter 2), APL (Altered phloem development), SULTR2 (Sulfate transporter), WOL (Woodenleg), SHR (Shortroot), AGL42 (Agamous-like 42), JO571 (J0571), SCR (Scarecrow), CORTEX (Cortex), CO2 (Cortex specific transcript), PEP (Endopeptidase), and GL2 (Glabra2). In addition, the genomic coordinates of the identical promoters are also specified.

Additionally, eight different promoters that target the same five cell-types were also used in these two studies (Table 5.1). For example, WOL and SHR promoters are both indicated as vasculature-related; however, it is clear that the activity of these promoters may not exactly overlap. Nevertheless, these related promoters served as a second platform for

our study. Hence, two scenarios were considered: (1) only data from the four identical promoter sets were used in comparisons (referred to as 'identical'), and (2) combined data from the four identical promoters and the eight promoter sets that presumably target the same cell-types were used in comparisons (referred to as 'common'). Therefore, the 'identical' and 'common' datasets target four and five different cell-types, respectively.

5.3.3 Transcription and translation of cell-wall-related genes are highly correlated

We investigated the variability in total or polysome-associated mRNA levels for any given gene to test how expression or translation patterns change across cell-types. This is derived by employing the coefficient of variation (CV).



a. CV of genes in identical promoters **b.** CV of genes in common promoters

Figure 5.5: Coefficient of variation (CV) of ribosome-associated (translatome) and total mRNA (transcriptome) for the identical and common dataset. The distribution of obtained CV values for all 22810 genes is visualized using kernel density estimates. (a) In the identical dataset (n = 22810, bandwidth = 0.002265), the mean CV value is 0.066 and 0.036 for the transcriptome and translatome, respectively. (b) In the common dataset (n = 22810, bandwidth = 0.002568), the mean CV value is 0.043 and 0.0072 for the transcriptome and translatome, respectively. In both comparisons, the translatome displays a smaller degree of variation in cell-type expression levels.

Figure 5.5 shows the distribution of CVs for all genes (22810 on the ATH1 platform) for the 'identical', and 'common', cell-types of the translatome and transcriptome data. The transcriptome varies more (CV mean value: 0.066) than the translatome (CV mean value:

0.036) across the 'identical' promoter datasets. The 'common' promoter datasets revealed a similar scenario (CV mean value transcriptome: 0.072, translatome: 0.043).

To examine how similar a given gene's expression and translation patterns are across the different cell-types we used PCC. Figure 5.6 shows the PCCs between translatome and transcriptome for all genes across the 'identical' and 'common' datasets, respectively. In the case of the 'identical' promoter dataset, the distribution of PCCs is best characterized by an almost uniform distribution, with a slightly higher frequency of positive PCC values (mean/median: 0.08/0.12; Figure 5.6). When using the 'common' promoter dataset the distribution of observed gene-wise PCCs resembles a normal distribution (mean = median: 0.04) in which extreme absolute values of PCCs are less common (Figure 5.6).

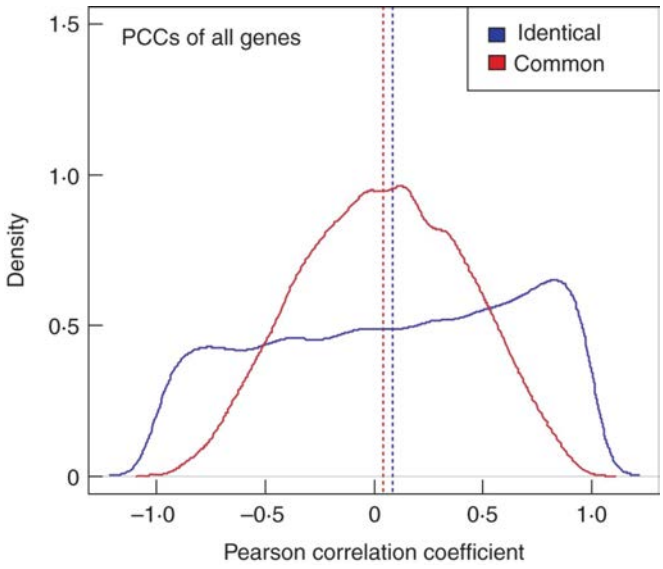


Figure 5.6: Pearson correlation coefficient (PCC) between ribosome-associated (translatome) and total mRNA (transcriptome) levels of the identical (red) and common promoter dataset (blue). The distribution of obtained PCC values for all 22810 genes is visualized using kernel density estimates. In the identical dataset, the PCC distribution is characterized by an almost uniform shape and has a higher frequency of positive PCC values. In the common dataset, the PCC distribution resembles a normal distribution.

To estimate whether the observed PCC for a gene, i.e. correlation of its expression and translation, is higher or lower than what may be observed by chance, bootstrapping was employed. Here, we re-computed PCCs using 1000 randomized datasets. Next, the observed PCC values for each gene were compared with an empirical null distribution derived from the randomized bootstrapping analysis. This null distribution of PCCs was derived by performing a bootstrap procedure randomly selecting four (for the 'identical' analysis corresponding to four cell-types) or eight (for the 'common' analysis corresponding to

five cell-types) promoters from the transcriptome and translatome dataset (in total 19 promoters and ten promoters, respectively, see Table S4, Appendix B). By computing Z-scores, the strength of the observed PCC value can be compared to what is randomly expected. In theory, genes with high positive or negative PCC values should therefore display high absolute Z-scores. Finally, based on PCC and Z-score, each gene can be classified into one of two groups: genes with coupled total and polysome-associated mRNA levels (high PCC and Z-score of ≥ 1.96) or genes in which the mRNA levels are uncoupled (low PCC and Z-score of ≤ -1.96).

A subsequent enrichment analysis of GO-BP terms allowed us to estimate if certain processes were enriched in either of the two groups of genes. Supplementary tables S5 and S6 (Appendix B) list enriched GO-BP terms found for genes with strong positive and negative PCC values, respectively. We found that 494 and 373 genes displayed uncoupled expression and translation for the 'identical' and 'common' promoters, respectively. These genes were enriched for GO-BP terms related to cell growth, root and meristem development, protein glycosylation, and cytoskeletal organization (Supplementary tables S5 and S6, Appendix B). However, it is important to remember that the two datasets, i.e. the transcript and translation datasets, were generated in two different labs with two different techniques, and it is therefore possible that some of the uncorrelated processes are due to these differences.

We found that 851 genes and 790 genes for the 'identical' and 'common' promoters displayed coupled expression and translation, respectively. GO-BP enrichment analyses showed that these genes are associated with regulation of transcription, post-translational modification (protein phosphorylation), and responses to various biotic and abiotic stimuli/stresses (Supplementary tables S5 and S6, Appendix B). Moreover, we found that genes associated with cell-wall-related processes and root tissue formation processes were common. This comprises GO-BP terms such as cell wall modification, secondary cell wall biogenesis, xylan biosynthetic process, xylem and phloem pattern formation, and meristem initiation. These data are in agreement with various co-expression approaches that have been undertaken for secondary wall synthesis, i.e. many secondary wall genes are transcriptionally and functionally coordinated (Persson et al., 2005).

5.3.4 Co-expressed relationships are not preserved across system-levels

So far, our analysis has focused on quantifying the degree of similarity in expression and translation for individual genes across different cell-types. However, one could also inves-

tigate whether larger contexts of genes are coordinated across the two levels. To assess whether genes that are transcriptionally coordinated, or co-expressed, are also coordinated on a translational level, we constructed co-expression and co-translation networks for the 'identical' and 'common' promoter datasets. Note that while a particular gene can exhibit changes between cell-specific transcription and translation, this does not exclude that the co-expression and co-translation neighborhoods of genes are preserved, i.e. one could imagine that certain co-expressed genes change their translational patterns in a coordinated fashion.

For one particular gene, an EC score is derived by calculating the PCC between the adjacent edge-weights, i.e. the co-expression relationship, of the two networks thus capturing similarity of gene neighborhoods (Figure 5.2). Accordingly, genes displaying low EC scores show different patterns of co-expression relationships in the two respective networks, while high EC values indicate the presence of highly similar co-expression relationships on the translatome and transcriptome.

The edges are weighted according to the similarity of expression/translation, which is defined as the PCC scores between the cell-specific expression/translation levels of the neighboring genes. For each gene, the EC, i.e. the similarity of the gene's genome-wide co-expression and co-translation relationships, was calculated. For this, differences in the edge-weights of a gene's incident edges, i.e. its network neighborhood, are compared between the co-expression and co-translation networks. The computation of EC score has been previously applied to elucidate the 'expression context' of orthologous genes in four Eukaryote species, successfully illustrating that co-expression neighborhoods of orthologues are highly conserved (Dutilh et al., 2006). Figure 5.7 shows the distributions of EC scores using the 'identical' and 'common' promoters, respectively. For both the 'identical' and the 'common' datasets, the range of EC scores lies in the interval -0.4 to 0.5.

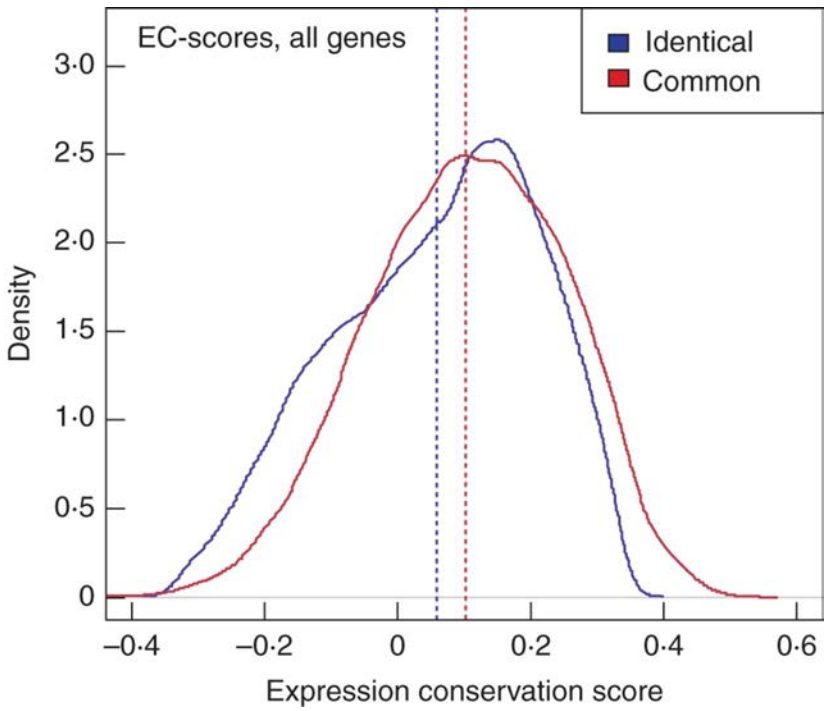
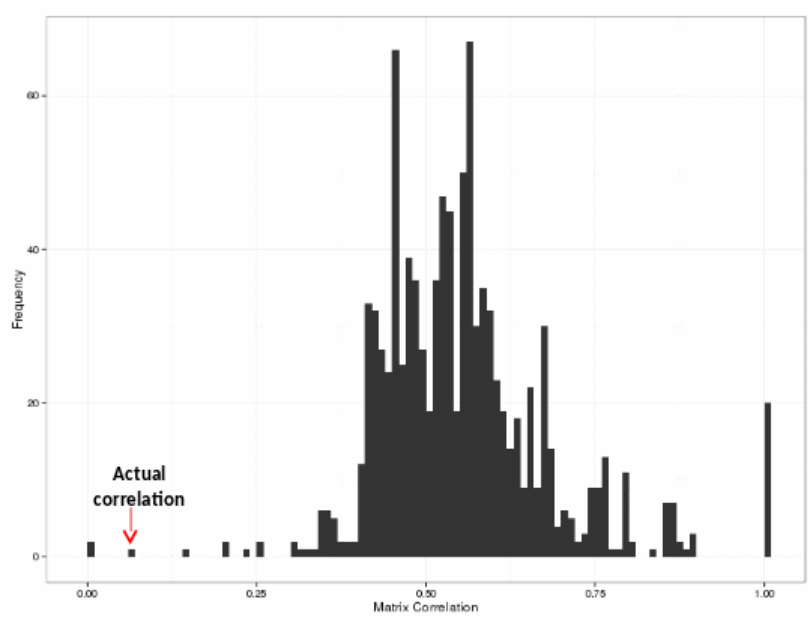
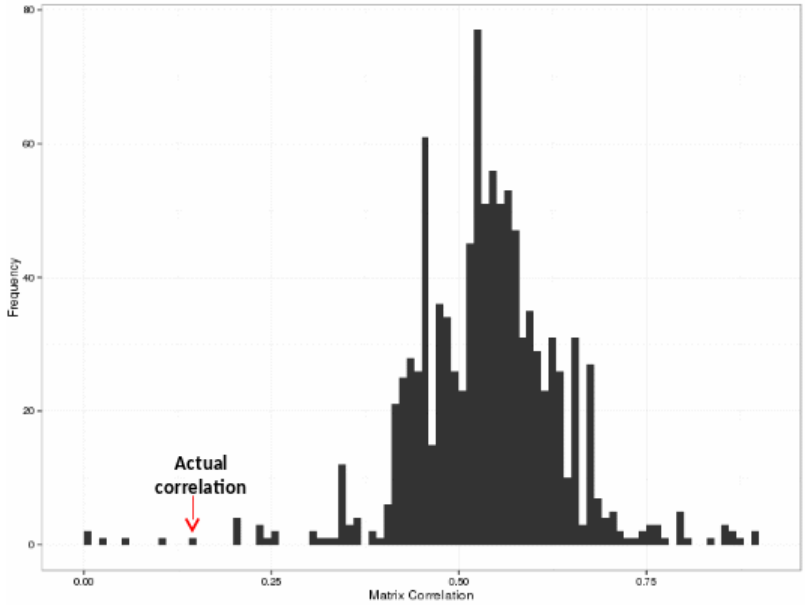


Figure 5.7: Expression conservation (EC) scores of co-expression relationships on the translatome and translatome within the identical (red) and common promoter dataset (blue). The distribution of obtained PCC values for all 22810 genes is visualized using kernel density estimates. For both the identical and the common dataset, EC score values lie in the interval 0.4 - 0.5.

To validate the observed relationships on the level of co-expression and co-translation networks, we derived the correlation on a global scale by considering the entire networks. The observed value of the full-matrix correlation was 0.06 and 0.10 between the co-expression and co-translation networks for the 'identical' and 'common' promoters, respectively. To assess whether these values are different from what could be expected by chance, we again selected random cell-type expression and translation data, and used 1000 bootstrap samples that then were compared against our observed similarities (Figure 5.8). For both the 'identical' and the 'common' promoter sets we found the values to be statistically significantly lower than expected by chance ($p < 0.01$). These data suggest that globally, or genome-wide, co-expressed gene patterns are dissimilar from co-translational patterns in Arabidopsis root cells.



a. Identical promoters



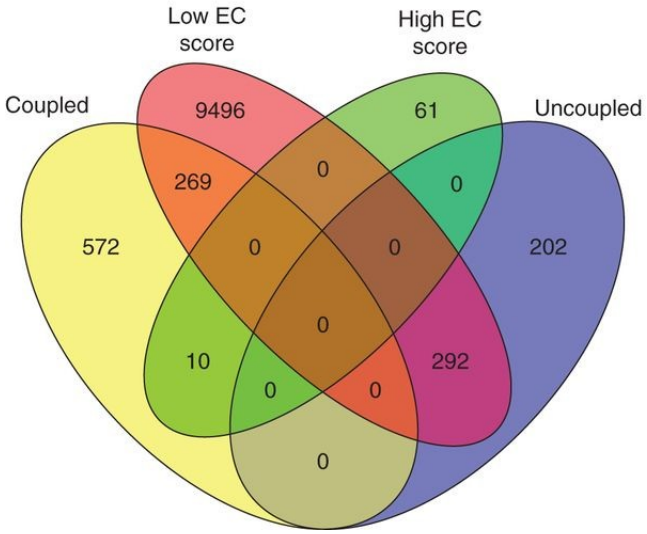
b. Common promoters

Figure 5.8: Similarity of the co-expression and co-translation network for the (a) identical and (b) common dataset. The similarity of both networks is determined by the PCC of the adjacency representation of the networks, i.e. a full matrix correlation. Only 0.001 and 0.005 % of the 1000 pairs of networks derived from bootstrapping procedure exhibit lower correlations than the observed transcriptome and translatome networks, respectively. This indicates a low degree of coupling of co-expression relationships between the two system-levels.

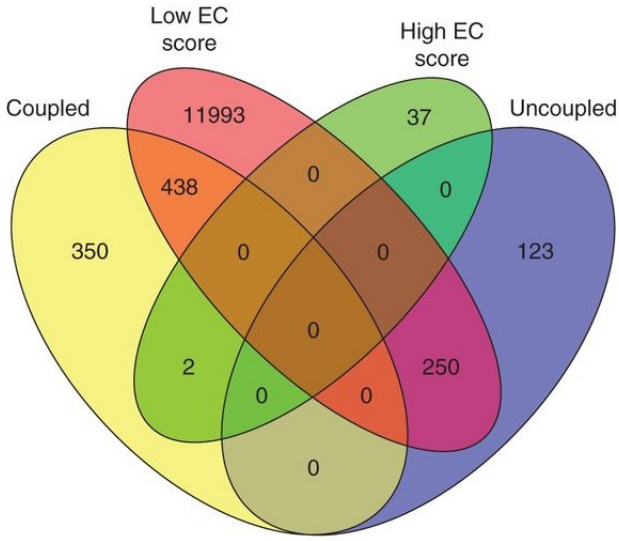
Returning to a per-gene analysis, we investigated which genes either significantly deviated or correlated in co-expression and co-translation relationships (i.e. EC scores) by

assessing statistical significance by Z-scores and bootstrapping. Hence, these genes either exhibited tight coupling (high EC score, and Z-scores $\geq$ 1.96) or uncoupling (low EC scores, and Z-scores $\leq$ –1.96; Figure 5.2) of their co-expression and co-translation to other groups of genes, thus reflected in conserved or changed network neighborhoods. Only 71 and 39 genes exhibited (statistically significant) high EC scores for the 'identical' and 'common' promoters, respectively. In contrast, 10057 and 12681 genes displayed (statistically significant) low EC scores for the two promoter sets, respectively. Subsequently, we tested if these sets of genes were enriched for certain GO-BP terms (Tables S7 and S8, Appendix B). For both groups of genes (high/low EC scores) and promoter sets ('identical'/'common'), we found enrichment of the GO-BP terms DNA-dependent regulation of transcription, cell wall biogenesis and organization, transmembrane transport, cell wall organization and cell growth, and signal transduction. Due to the high number of genes with uncoupled co-expression and co-translation relationships, i.e. low EC scores, we found numerous GO-BP term enrichments, including a wide range of metabolic and catabolic, as well as transport processes.

Finally, to assess what types of genes both display a good correlation between expression and translation (high PCC score) and retain a good correlation between co-expression and co-translation network neighborhoods (high EC score) we identified such genes for the 'identical' and 'common' promoter sets. Figure 5.9 shows that ten genes ('identical' promoters) and two genes ('common' promoters) have these characteristics. Remarkably, the majority of the identified 12 genes are transcription factors, or contain predicted DNA binding protein domains, e.g. ATHB-3, MYB46, VND7 and WRKY9. Several of the genes are associated with key regulatory roles in roots, either for developmental or for response processes (Table 5.2). For instance, WRKY9 is involved in mediating cell responses to nutrient deprivation (Shin and Schachtman, 2004, Shin et al., 2005), and the transcription factor MYB46 has a prominent role in the developmental programme of secondary wall biosynthesis (Zhong et al., 2007). In addition, VND7 has been characterized as transcriptional master switches for plant meta- and protoxylem formation in Arabidopsis (Kubo et al., 2005).



a. Identical promoters



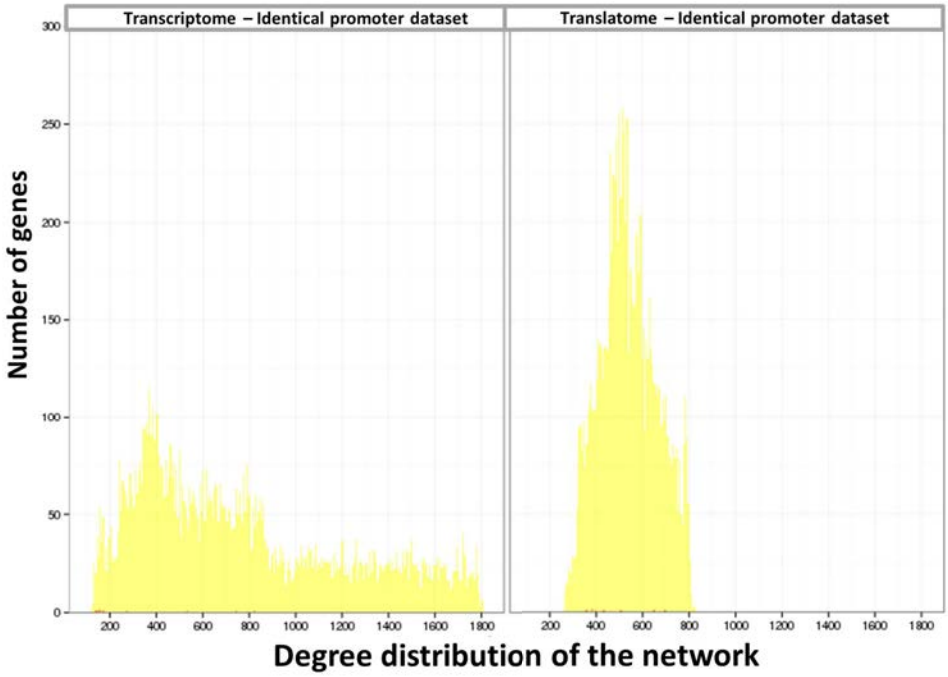
b. Common promoters

Figure 5.9: Venn diagrams illustrating the overlap of genes displaying conserved expression levels (PCC) and co-expression relationships (EC scores) across root cell-types in translatome and transcriptome. (a) Ten genes could be identified using the scenario considering identical promoters (four promoters) and (b) two genes considering common promoters (eight promoters).

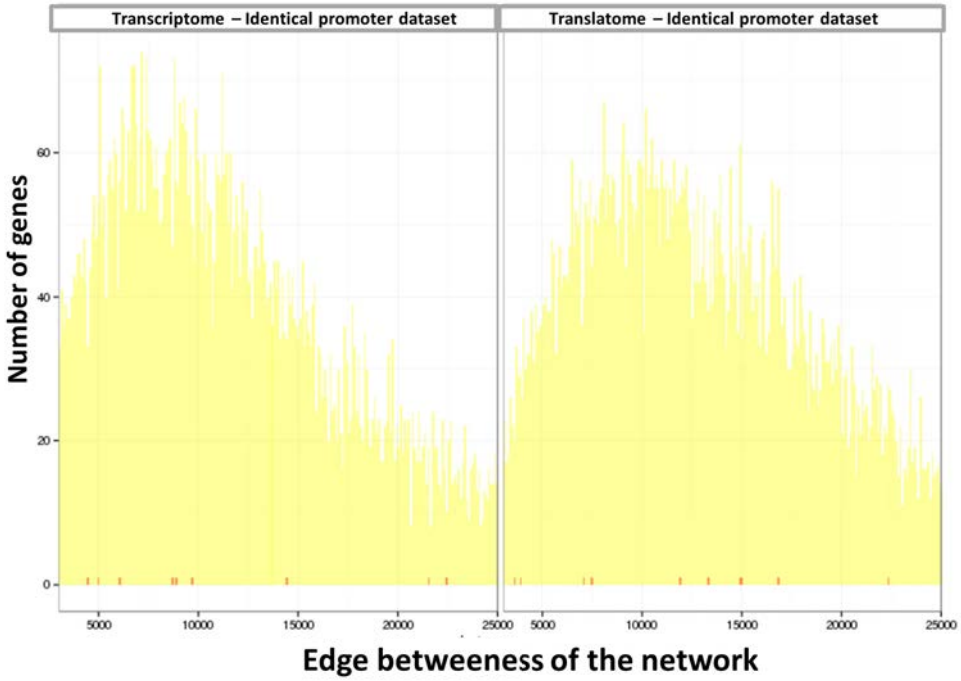
No. of cell-types	Array element	Locus identifier	Annotation
4	260067_at	AT1G73780	Protease inhibitor/seed storage/lipid transfer protein (LTP) family protein
4	250108_at	AT5G15150	ATHB-3 (*Arabidopsis thaliana* HOMEOBOX 3); DNA binding/transcription factor
4	250322_at	AT5G12870	AtMYB46/MYB46 (myb domain protein 46); DNA binding/transcription factor
4	251009_at	AT5G02640	Similar to unknown protein [Arabidopsis thaliana] (TAIR:AT3G46300.1); similar to hyp. protein [Vitis vinifera] (GB:CAN667791)
4	260173_at	AT1G71930	VND7 (VASCULAR RELATED NAC-DOMAIN PROTEIN 7); transcription factor
4	253076_at	AT4G36160	ANAC076/VND2 (VASCULAR-RELATED NAC-DOMAIN 2); transcription factor
4	267613_at	AT2G26700	Protein kinase family protein
4	253120_at	AT4G35790	ATPLDDELTA (Arabidopsis thaliana phospholipase D delta); phospholipase D
4	266342_at	AT2G01540	C2 domain-containing protein
4	260468_at	AT1G11100	SNF2 domain-containing protein/helicase domain-containing protein/zinc finger protein-related
5	255637_at	AT4G00750	Dehydration-responsive family protein
5	260432_at	AT1G68150	WRKY9 (WRKY DNA-binding protein 9); transcription factor

Table 5.2: Genes displaying conserved expression levels (PCC) and co-expression relationships (EC scores) across root cell-types in transcriptome and translatome. In total, ten genes could be identified using the scenario considering the 'identical' promoters and two genes considering the 'common' promoters.

Furthermore, a more detailed network analysis was conducted to analyze whether additional network properties of those 12 genes deviate from the majority of genes and further explain the conservation/similarity of neighborhoods. Thus, un-weighted, classical gene-relevance networks were created, using a threshold, $t = 0.9$, for the PCC of two genes to decide whether an edge ($\geq t$) or no edge ($< t$) is present. Based on this threshold, two networks were created dichotomously capturing the co-expression and co-translation properties of the twelve genes across the two system-levels.



a. Degree distribution of the network



b. Edge betweeness of the network

Figure 5.10: Degree distribution and edge betweenness of the co-expression and co-translation network. (a) This figure displays the distribution of the degree for the co-expression and co-translation network using the 4 identical promoters. Clearly, the difference of both network topologies becomes visible. However, the degree of the ten genes (indicated by small red bars at the x-axis) is not strongly confined to a particular range, i.e. low degree or high degree. (b) The other figure displays the distribution of the betweenness for the co-expression and co-translation network using the 4 identical promoters. In this case, both network-structures seem more similar, but again, the betweenness of the ten genes (indicated by small red bars at the x-axis) is evenly spaced across the whole range of obtained betweenness values.

The network properties considered here comprise of the degree, edge betweenness, closeness, eigen vector centrality, alpha centrality and transitivity. Figure 5.10 indicates the degree distribution and edge betweenness for the ten genes identified to show both conserved expression as well as co-expression relationships in case of the 'identical' promoters (Table 5.2). Similar plots were obtained for the other computed network properties of the ten genes but not shown here. Also, network properties of the 2 genes identified to show both conserved expression as well as co-expression relationships in case of the 'common' promoters were computed (data not shown). Interestingly, for all of the tested properties, the twelve genes (ten from the 'identical' and 2 from the 'common' promoters, cf. Table 5.2) display network properties that are well distributed across the whole range of the corresponding properties as compared with all genes. An enrichment analysis of these genes suggests that certain key genes for root development maintain a direct relationship between expression and translation (Table 5.3).

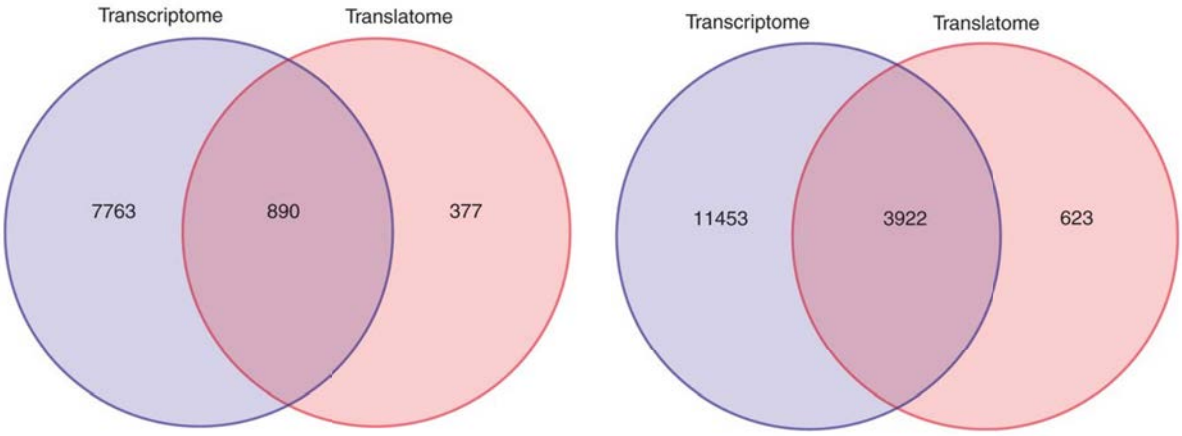
GO term	Term description	No. of genes	p-value
GO:0006355	Regulation of transcription, DNA-dependent	4	< 0.01
GO:0009741	Response to brassinosteroid stimulus	2	< 0.01
GO:0010089	Xylem development	2	< 0.01
GO:0010413	Glucuronoxylan metabolic process	2	< 0.01
GO:0045492	Xylan biosynthetic process	2	< 0.01
GO:0045893	Positive regulation of transcription, DNA-dependent	2	< 0.01

Table 5.3: GSEA of genes displaying conserved expression/translation levels and co-expression/co-translation relationships (EC scores) in the identical promoter dataset. Note that the corresponding two genes from the common promoter dataset did not result in a significant enrichment of GO-BP terms (cf. Table 5.2).

5.3.5 Root cell-type similarity based on transcriptome and translatome

To complement the gene 'centric' analysis of (un-)coupled expression and translation, a cell-type 'centric' analysis may reveal common themes among root cell-types. Here, we attempted to elucidate whether transcriptional and/or translational patterns (termed themes) may be conserved across multiple cell-types. To characterize a particular cell-type, we first identified genes that showed differential expression and translation across the datasets. These estimates were derived for both the 'identical' and the 'common'

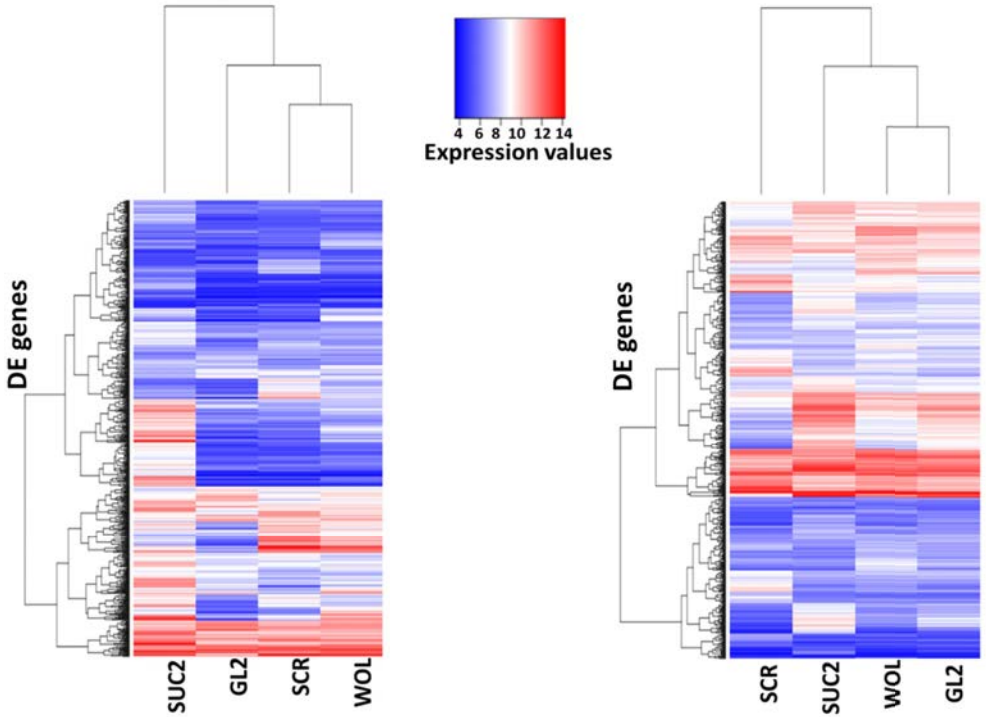
promoter sets using ANOVA [FDR of 5 % by Benjamini-Hochberg (BH) multiple testing correction]. For the 'identical' promoters, a set of 890 genes displayed both differential expression and translation across the cell-types (Figure 5.11). Moreover, for the five 'common' promoters, a set of 3923 genes showed differential expression and translation across the cell-types (Figure 5.11).



a. Identical promoters **b.** Common promoters

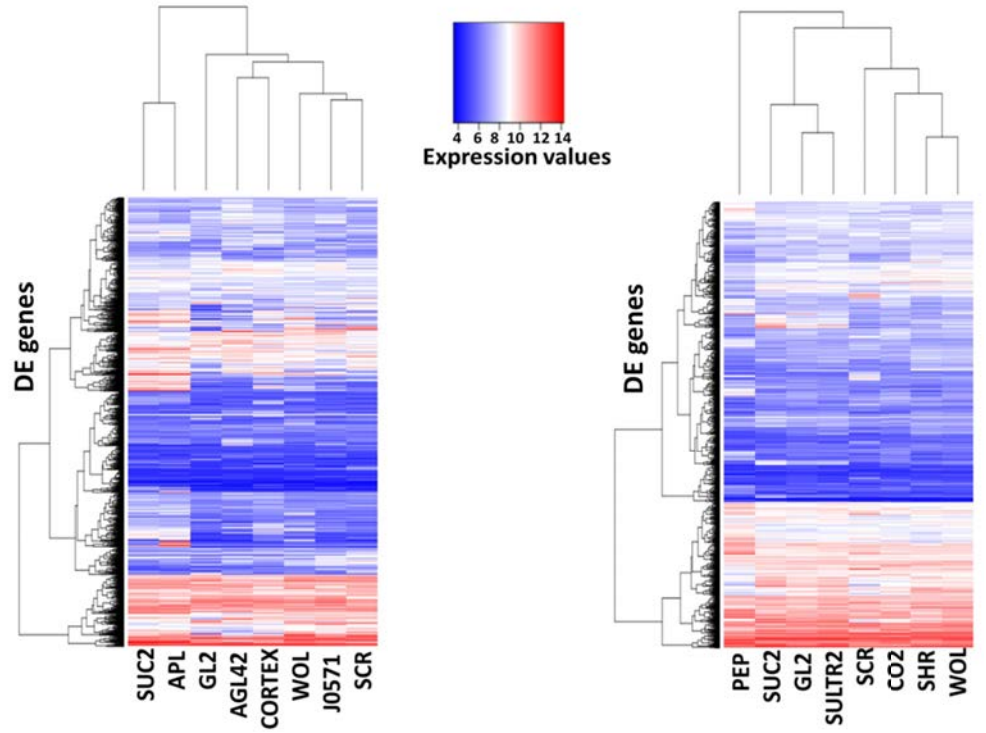
Figure 5.11: Visualization of the set of differentially expressed (DE) genes across the transcriptome and translatome of the identical (a) and common (b) promoter datasets. Differential expression was assessed by ANOVA at an FDR rate of 5 %. The numbers in the Venn diagram correspond to the number of DE genes found in each system-level and the intersection thereof.

Considering expression and translation separately, we found that most genes exhibit differential expression across the cell-types: $\approx$ 38 % ('identical' promoters) and $\approx$ 67 % ('common' promoters), while only $\approx$ 5 % ('identical' promoters) and $\approx$ 20 % ('common' promoters) of all genes display differential translation. Hierarchical clustering of the genes exhibiting differential expression on both levels, i.e. the 890 and 3923 genes above, revealed divergent patterns in cell-type-specificity (Figure 5.12).



a. Transcriptome-Identical promoters

b. Translatome-Identical promoters



c. Transcriptome-Common promoters

d. Translatome-Common promoters

Figure 5.12: Common DE genes show divergent patterns of cell specificity across the two system-levels. (a) Relative gene expression levels for the 890 DEG with significant differences across identical promoters in transcriptome and (b) translatome. (c) Relative gene expression levels for the 3922 DEG with significant differences across common promoters in transcriptome and (d) translatome.

We conducted a series of Tukey HSD tests ($p < 0.05$) for all genes displaying differential transcription and translation to further derive which cell-type-specific expression and translation levels differed significantly. Performed for each gene, the Tukey HSD post-hoc test allowed us to determine for which pair-wise cell-type comparisons there is a significant difference in cell-type expression/translation levels (Figure 5.3). When considering the four 'identical' promoters (representing cell-types; Table 5.1) in the Tukey HSD test, we obtained a characteristic pattern of six pairwise cell-type comparisons encoded for by a binary matrix (0 and 1). Then, for any given pair of cell-types and system-level, a significant difference in cell-type expression or translation profile is assigned the value 1, and a similar expression profile is assigned a value of 0.

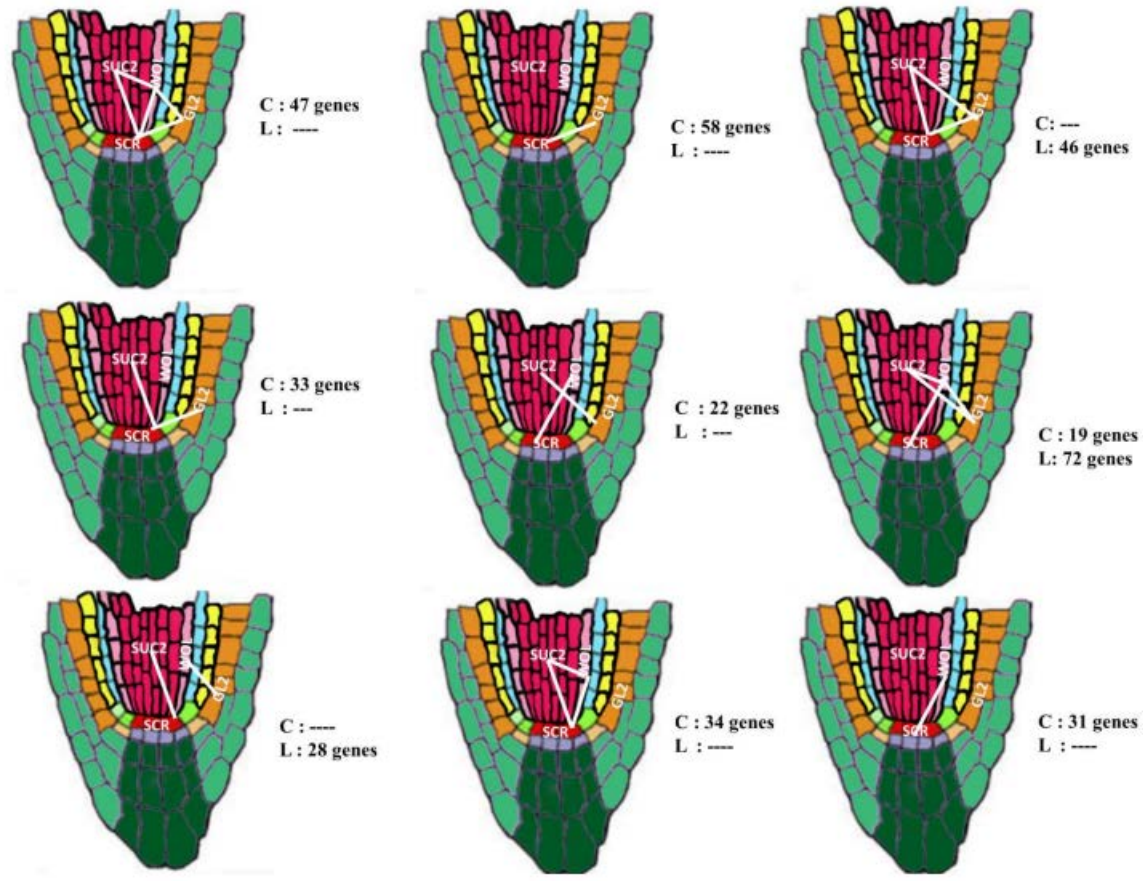
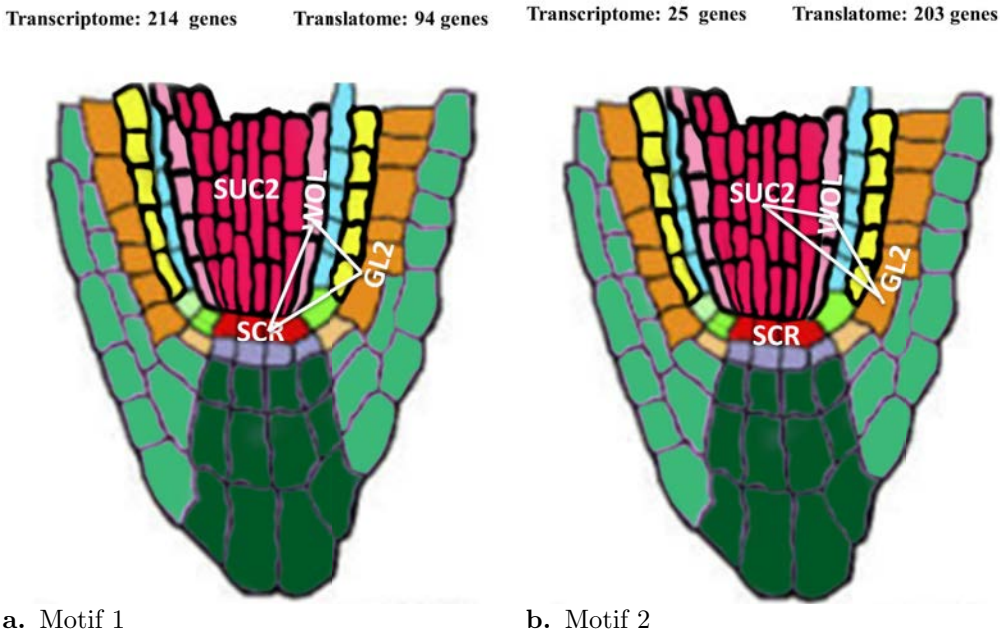


Figure 5.13: All possible motif occurrences across the identical promoter data of the transcriptome and translatome. Out of all possible network motifs (Supplementary table S9, Appendix B) for the transcriptome and translatome, only nine for the transcriptome and five for the translatome occur more often than expected by chance. The "C" and "L" used in the figure indicate the transcriptome and translatome, respectively and give the number of genes which coincide with each motif.

We displayed the results as networks, here referred to as network motifs, to show patterns of similar gene levels. In these small networks, the four cell-types ('identical' promoter dataset) correspond to nodes, and similarity between two cell-types (= 0) is indicated by an edge. From the 64 possible configurations of those networks (cf. Materials and Methods), we found that only five network motifs for the translatome and nine for the transcriptome occur more often than expected by chance (Figure 5.13).

The significance of these motifs was tested empirically by permuting the results from the Tukey HSD test for each gene (i.e. shuffling the 0 and 1 values) and comparing the observed occurrence counts of genes for a particular motif with those obtained randomly over n = 1000 permutations (Supplementary table S9, Appendix B). Furthermore, of all significantly occurring motifs only two are represented by a high number of genes [279 and 214 unique genes, motif 1 (Figure 5.14) and motif 2 (Figure 5.14), respectively] of the 890 differentially expressed genes. Both these motifs contain one isolated node, i.e. a cell-type that is not connected by edges to any other cell-types, indicating that this cell-type is dissimilar based on the gene's expression or translation profile.

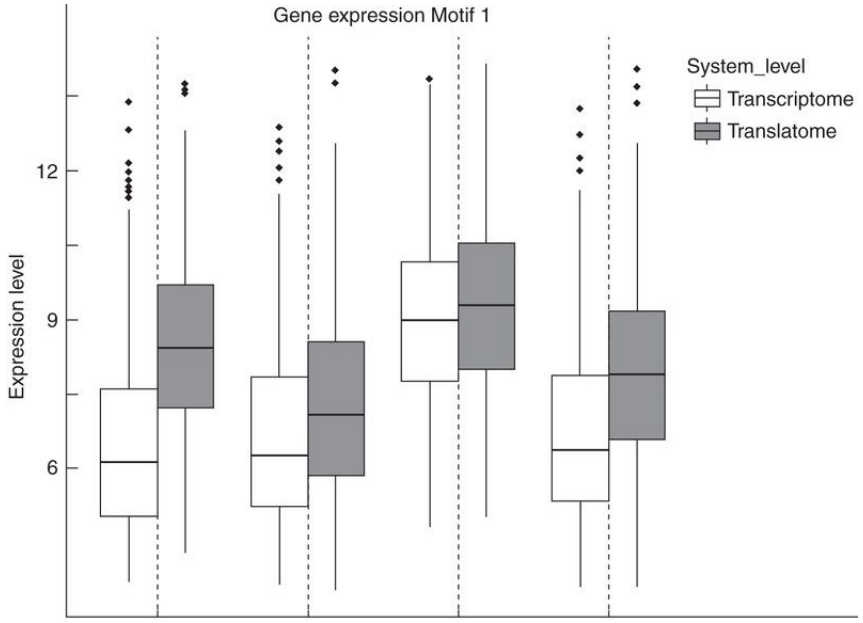


a. Motif 1 **b.** Motif 2

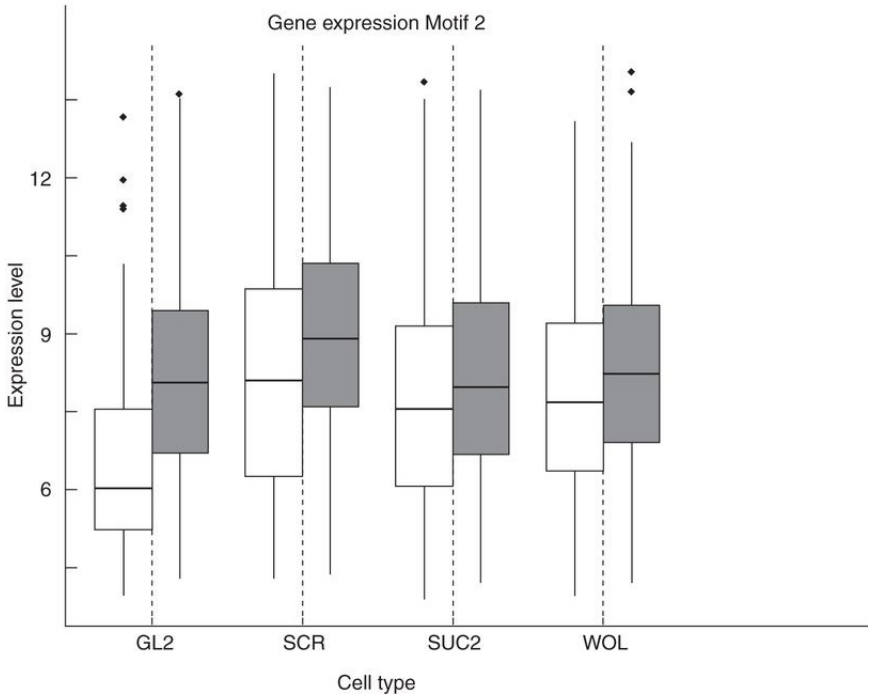
Figure 5.14: Two most commonly occurring network motifs for the transcriptome and translatome. A series of Tukey HSD tests were used on the common DE genes from the identical promoter dataset (890 genes) to detect significant pairwise cell-type differences across the transcriptome and translatome. Network motifs were constructed using cell-types as nodes and cell-type similarity indicated by a red edge. (a) Network motif 1 is represented by 279 unique genes. (b) Network motif 2 is represented by 214 unique genes.

Moreover, in both the motifs, the remaining three cell-types are fully connected by edges, indicating similar behavior of the relative levels of expression/translation of the genes included in the motif. The first motif (motif 1) shows that the majority of genes have dissimilar expression pattern in phloem companion cells (SUC2) as compared with the other three cell-types (WOL, SCR and GL2). In the second motif (motif 2), the root quiescent center (SCR) displays deviating expression patterns. Looking more closely at these differences shows that the major driving force behind the deviation in the phloem companion cells can be attributed to the transcriptomic datasets (214 differentially expressed genes), while the deviation of the quiescent root center is largely due to differences in the translatome profiles (203 differentially expressed genes). An enrichment analysis using GO-BP terms associated with the genes in the first motif (214 genes) reveals mainly transport processes (general and transmembrane), as well as responses to sugar stimuli (glucose, sucrose and fructose) to be over-represented (Table S10, Appendix B). These data are in agreement with a major function of phloem companion cells in sugar transport (Stadler et al., 1995, Oparka, 1999, Williams et al., 2000). When we looked at the relative expression levels of the genes associated with this motif (again 214; Figure 5.15) we found that the transcript levels were elevated. This is consistent with a role of the gene products in the function of these cells.

By contrast, genes associated with the second motif, i.e. where the root quiescent center showed dissimilar expression patterns, were enriched for cell wall modification, xylan biosynthetic process and root hair cell differentiation/elongation. More importantly, the GO-BP terms oxidative stress, oxidation-reduction processes and auxin polar transport were also enriched. The quiescent center cells typically accumulate high auxin levels that serve as a distal organizer (Sabatini et al., 1999). This is accompanied by the overproduction of reactive oxygen species. This is mediated by high levels of activity of ascorbate oxidase that cause reduction in the reduced form of ascorbic acid and glutathione and, simultaneously, an increase in the content of reactive oxygen species in the root quiescent center cells (Ivanov, 2007). Oxidative stress represses proliferation of these cells, thus maintaining the cells in a quiescent state (Jiang et al., 2003). Expression levels of genes corresponding to this second motif are shown in Figure 5.15, which indicates a relatively higher degree of translation in the root quiescent center.



a. Expression levels of genes associated to Motif 1



b. Expression levels of genes associated to Motif 2

Figure 5.15: Boxplots of expression levels separated by promoters individually for translatome and transcriptome for network motif 1 and motif 2. (a) In motif 1, 279 of 890 DE genes exhibit this characteristic cell-type-specific expression pattern. (b) Furthermore, 214 genes correspond to motif 2.

5.4 Discussion

A long-standing question in cellular biology is how well the transcriptome is coupled to the proteome (Zanetti et al., 2005). Profiling of mRNAs associated with polysomes can give a rough estimate of a cell's or tissue's proteome. Hence, by comparing cell-type-specific levels of total and polysomal mRNA in a global context, one can derive to what extent expression and translation are coupled. Based on the efforts of Brady et al. (2007) and Mustroph et al. (2009) the Arabidopsis root atlas allows us to analyze transcriptomic and translatomic datasets and to identify particular genes that show either a tight coupling, or an uncoupling, of expression and translation profiles over a collection of cell-types. On the computational level, this analysis represents an extension to the analysis of Mustroph et al. (2009), who used their root and shoot data in combination with hypoxia conditions to identify DE genes at the single cell-, region-, and organ-specific levels. Recently, Lin et al. (2014) have investigated the translatome of *in-vivo* grown pollen tubes from self-pollinated gynoecia of Arabidopsis. By using a pollen-specific promoter, epitope-tagged polysomal-RNA complexes could be affinity purified to obtain mRNAs undergoing translation. We also employed joint (RMA-) normalization to compare the translatome data with publicly available transcriptomics datasets. Set theory and analysis of the differential behavior of genes finally identified a group of genes important in *in-vivo* pollen tube biology.

Although canonical correlation based methods has previously been used in the analysis of the NCI60 datasets, it was largely limited to the identification of clusters of genes and their associated over-represented biological terms (Lê Cao et al., 2009). Moreover, Lê Cao et al. (2009) focused on analyzing the transcriptome data arising from two different microarray platforms (cDNA and affymetrix chips). In this case study, the number of profiled cell-types were similar i.e., four in case of the 'identical' promoters and '8' for the common promoters across the transcriptome and translatome. One of the biological questions under investigation in this chapter focuses on studying the correlation of expression patterns in multiple cell-types across system-levels. Using CCA in such a case will result in the problem of high dimensionality i.e., the observations will be the profiled cell-types, and the variables will be the gene expression values across the two system-levels. This is clearly a problem of '$n < p + q$' setting which CCA is not suited for. Although sparse CCA addresses this issue, the problem is that it obtains the weighted linear combinations of the variables from each dataset and hence interpretation would be quite tedious considering that there are 22000 genes. On the other hand, considering the observations as the number of genes, and the variables as profiled cell-types while using CCA is not suitable to answer the biological question under investigation. This is because, CCA will produce

a weighted linear combinations of the different cell-types and does not focus on the gene-expression measurements. In contrast, the novel analysis pipeline in this chapter takes into account two different system-levels (transcriptome and translatome) and focuses on extracting cell-type specific expression patterns of genes. Moreover, the pipeline takes into account the expression conservation of the genes across different system levels using appropriate computations.

Furthermore, we examined the variance in expression and translation levels using CVs and tested the similarity of gene expression/translation patterns across the root cell-types using PCC. The observed change in CVs (Figure 5.5) and the presence of negative PCC values (Figure 5.6) when globally comparing the translatome and transcriptome is similar to what was found by Tebaldi et al. (2012). Here, the authors concluded a general uncoupling of the translatome and transcriptome based on low correlations found using epidermal growth factor stimulation in mammalian HeLa cells. Uncoupling of transcriptome and translatome has also been documented in human and yeast cells in response to various stimuli and stresses (Mikulits et al., 2000, Grolleau et al., 2002). For example, yeast exposed to different stresses, such as amino acid depletion and fusel alcohol addition, show distinct translational profiles (Smirnova et al., 2005), suggesting there is a distinct role of translational regulation for rapid responses in cells to environmental stress. However, by further focusing the analysis to the level of individual genes, our results also revealed groups of genes displaying coupled transcription/translation involved in processes such as stress responses (e.g. wounding, bacteria, nitrogen starvation and osmotic stress). These findings are, on the other hand, in agreement with the study in yeast by Halbeisen and Gerber (2009), who found relatively high overall PCC values of 0.75-0.81 of the overall genomic (fold-) changes in expression upon different conditions of cellular stress, such as osmotic stress between transcriptome and translatome.

Genes that show correlated transcription and translation are enriched in cell-wall-related processes, which is in agreement with co-expression approaches that have successfully been undertaken for secondary wall synthesis (Persson et al., 2005). Here, many secondary wall genes are transcriptionally and functionally coordinated, which implies that the translation also would be coordinated with the transcription (Mutwil et al., 2009, Ruprecht and Persson, 2012). While these processes appear to be coupled, most of the genes displayed uncoupled transcription and translation in the cell-types considered in our analysis.

In addition, a high degree of uncoupling between transcription and translation was observed when investigating correlations in co-expression and co-translation relationships. Here, over 12000 genes displayed altered co-expression patterns in the eight 'common' pro-

moter datasets. This may reflect a re-wiring of co-regulation of genes during translation compared to transcription. In addition, cell-type-specific mRNA abundance appeared different on the two levels with 11453 genes differentially expressed exclusively in the transcriptome (Figure 5.11). Notably, large proportions of genes displaying conserved co-expression/co-translation neighborhoods are transcription factors or are putatively involved in regulation of transcription.

Note that bootstrapping procedures were carried out to ensure the robustness of our analyses. The benefits of this approach are two-fold: the random PCC and EC scores account for, first, sample size and, second, differences in cell-type promoter specificity and the presence of multiple promoters targeting the same cell-type in the case of the 'common' promoter dataset. As a consequence, one can robustly classify genes whose total and polysomal-associated mRNA levels are coupled (high PCC and Z-score of ≥ 1.96) as well as genes that display an uncoupling of both mRNA levels (low PCC and Z-score of ≤ -1.96). Accordingly, we employed the same statistical framework to confirm the similarity of co-expression and co-translation neighborhoods in the network analysis to ensure robustness. Nevertheless, the observed effects must be carefully interpreted given that the datasets originate from different labs and, moreover, rely on different extraction procedures. Here, the identified coupled attributes of gene expression on the transcriptional and translational level are therefore remarkable. Moreover, many of our observations are in close agreement with well-established characteristics of root cell-type function and development.

One of the limitations of this analysis is the available selection of promoters, i.e. cell-types, for the datasets. Clearly, in the case of the correlation analyses of transcription and translation of the individual genes, a greater sample size would have been desirable. Also, in the case of the 'common' five cell-types, artefacts may arise due to slight variation in promoter strength and specificity across the cell-types. Therefore, it is impossible to rule out deviations in transcription and translation based on promoter patterns. Nevertheless, we found correlation between transcription and translation for genes that we anticipated, such as for the secondary wall genes discussed above. These results are reassuring, and may provide a foundation for future efforts in this area. We propose that using more cell-type-specific promoters and performing the transcript and translatome analyses in one lab using the same methods will generate a robust and interesting data series that may be used to improve our results. Such datasets would be of immense interest to understand coupled and un-coupled gene regulation in Arabidopsis roots.

Chapter 6

Conclusion

Given the availability of sophisticated techniques, it is essential to adopt a multidisciplinary approach to understand the complexity and diversity of plant cell walls. Therefore, the contributions of this thesis are two-fold: first, novel applications of existing methods are illustrated for the integration of heterogeneous datasets to yield unprecedented views on different aspects of plant cell walls. Secondly, it was demonstrated using a novel analysis pipeline how integrative system-level analysis can be used to extract discernable information pertaining to cell wall related genes and functions. In addition to the discussion sections contained in the different case-studies of this thesis, a general conclusion concerning the integration of heterogeneous datasets shall be discussed here.

Contributions to integrative system-level analysis

The availability of high-throughput techniques revolutionize plant research and provide system-level measurements for virtually all types of cellular components in various plant species (cf. Chapter 1). However, the abundance of datasets generated from different system-levels pose challenges to understand the system as a whole. To this end, the concept of multiblock data analysis is introduced in Chapter 2 wherein each dataset generated from different system-levels or using different analytical techniques is considered as a data block. Multiblock methods have a long history in the field of behavioral research and ecological data analysis and a table of methods proposing the available methods in literature for different instances of combining data blocks is provided (cf. Table 2.1). All the three case-studies elaborated in this thesis focus on the integration of two data blocks and biological understanding of various aspects of cell walls are discussed.

Cotton fiber is one of the most useful systems for cell wall research. Further analysis of cotton fiber cell walls is relevant to improve this important natural textile fiber and to

create the next generation of crop plants for optimized production of biomaterials (Haigler et al., 2012). The development of cotton fibers is a very complicated biological process, and previous analysis in cotton fibers mainly focused on the identification of genes for fiber quality improvement (Gilbert et al., 2013, 2014, Al-Ghazi et al., 2009). Moreover, the results from these studies employed data from one particular system level i.e., the transcriptome or the metabolome. In chapter 3 of this thesis, the combined analysis of the glycome and phenome levels in cotton fibers established links between polysaccaharide rich cell walls and their phenotypic characteristics. The inherent potential of glycan arrays (CoMPP) for high throughput characterization of plant polysaccharides has mostly been restricted to the generation of quantitative information (Moller et al., 2008, Pedersen et al., 2012) and lacks efficient computational tools to extract relevant information. The comparative analysis of the glycan array with the phenotypic characteristics reveals the potential of glycan arrays in combination with multivariate statistical methods as a powerful approach for understanding cell wall composition and their effect on phenotypic characteristics. Specifically, correlation and regression based approaches were used to elucidate the relationship between the two datasets. Appropriate pre-processing of the datasets was done and the initial analysis by multiple regression analysis depicted that it is possible to predict only one fiber characteristic at a time. The application of canonical correlation analysis (CCA) provided a global view of association between the system-levels with information about the relative contribution of the variables to a particular canonical variate. In case of a high dimensional framework, other methods such as regularized canonical correlation analysis and sparse generalized canonical correlation analysis proposed in Table 2.1 could be used. Although CCA proves to be a universal tool to identify exploratory relationships between datasets, it is not suitable to predict models for each of the fiber trait under study. For this purpose, CCA with elastic net penalization was proposed (Lê Cao et al., 2009) and allowed variable selection as a one step procedure. The major drawbacks of using CCA-EN is that it requires intensive computations in cases when "p+q" is large, and moreover, it uses elastic net with a similar Lasso soft thresholding penalization and does not involve CCA computations. Instead, sparse partial least squares regression (sPLS) maximizes the covariance between the latent variables and is a highly recommended approach as it includes a built-in variable selection that captures subtle individual effects. The conducted analysis using sPLS approach deduced relationships between fiber strength, and specific carbohydrate epitopes. Furthermore, this association was attributed to the role of xylan epitopes consistent to previous experimental studies on the role xylans in fiber strength. It would be interesting to perform similar kinds of analysis as in Chapter 3 by relating time series glycan arrays

to that of phenotypic characteristics using 'k+1' table methods proposed in Table 2.1. Time series glycan arrays refers to those obtained from developing cotton fibers and such a kind of analysis would allow to understand differences in polysaccharide composition over time and their relation to the development of cotton fibers.

The conducted analysis using correlation and regression based methods in Chapter 3 emphasize statistical significance by adopting different tests of significance to validate the results. In case of validation of results from the CCA, measures such as Wilk's lambda, Pilai's Trace, Lawley Hotelling Trace, and Roy's criterion were used as detailed in Chapter 2. However, computing an empirical permutation tests on the original data would destroy the covariance structure both within and among each dataset thus leading to different canonical variates. Therefore, this would directly influence the observed canonical structure between the variable and the canonical variates. In case of the regression analysis, model validation involved determining how well the data fits the theoretically implied model and focuses on the model predictive ability. The performance of the model was reported in terms of root mean square standard error of prediction (RMSEP), and for simplicity the latent variables included were defined from that model with the lowest RMSEP. Computing the Q^2_h criterion is closely related to RMSEP, but it gives a more general insight of the model, whereas the RMSEP requires to be computed for each variable in the response dataset (cf. Section 2.3). The predictive ability of the model could also be assessed by the R^2, also called the determination coefficient.

Contributions to analyzing different analytical techniques

The presented integrated analysis in Chapter 4 relies on maximizing the sum of the squared covariance between scores of each data table by multiple co-inertia analysis (MCIA). The power of MCIA for cross-platform comparison is illustrated here. In this chapter, the variations in the biochemical composition of plant cell walls were analyzed combining two different kinds of spectroscopy. Spectroscopy has proven to be an effective tool for rapid estimation of the numerous polysaccharides and lignin components in unfractionated plant cell wall materials. Both FT-IR and Raman spectroscopy are highly suitable for the estimation of cell wall composition but each of the spectroscopic technique operates at different levels of resolution. Therefore, coupling of the spectral domains reflects the information provided by different techniques. Estimation of the cell wall composition at different spatial resolutions across several cell-types help in understanding enzymatic digestibility and differences in saccharification yield for biochemical conversion into ethanol (Chen and Dixon, 2007, Yuan et al., 2008a, Hisano et al., 2009, Van Acker

et al., 2013). For cross-comparisons and joint analysis of different analytical platforms, MCIA is preferred when compared to other methods such as CCA, PLS, and CIA. More specifically, MCIA is preferred over PLS because the former deflates data blocks using the block scores whereas the latter deflates the scores of one data table. Furthermore, MCIA has the advantage over CCA and PLS in it's ability to analyze both the common and specific information brought by different data tables. In addition, MCIA was more appropriate to analyze the 'n < p + q' condition of the joint analysis.

Pre-processing and normalization was an important step in the integrative analysis of different spectroscopic techniques. The pre-processing procedure was different than that of the other integrative studies in this thesis. The procedure includes band selection, spike and noise elimination taking into account corrections specific to each spectroscopy. In addition, the normalization step was assessed considering spatial neighborhood and intensity variations in the spectra based on the morphology of the maize sections. The pre-processing procedures involved in this integrative analysis were a vital step to enhance the spatial and spectral resolution of the spectra in order to study the chemical information within. Consequently, pairing the hyperspectral images to map the regions in common between the two spectroscopies was done and allows the joint analysis of the hyperspectral image data. Cross correlation coefficients were computed to assess the validity of the registration parameters used to map the infrared and Raman images to the reference images (Table 4.1). The joint analysis of the data tables obtained after pairing of the hyperspectral images were used for the MCIA, and adapted to compute and compare how well the global and block information are related. Different indicators were proposed to analyze the contributions of the two data tables and the percentage of variances described by the block scores. Adequate visualization tools were used to interpret the findings from the MCIA such as the comparison of the spectral plots as shown in Figure 4.18 which reveals particular spectral profile profiling the same cell-types. This information when combined with the infrared and Raman score plots allows interpreting why particular spectral profiles are not correlated across both techniques and the discrepancies in the information provided. Another kind of visualization is the score images of the MCIA components obtained by blending the bright field image and the refolded scores. These score images reveal information with respect to particular cell-types and can be used in the interpretation of the infrared and Raman loadings. Moreover, identification of the peaks from the PCA and MCIA loadings help to understand the biochemical composition of various cell-types.

The statistical significance of the relationship between the data blocks was assessed by means of the RV coefficient. The RV coefficient is the coefficient of correlation between

the two tables X and Y (cf. Section 2.3). This coefficient varies between 0 and 1: the closer the coefficient to 1, the stronger the relationship and the common information provided between the tables. This kind of joint analysis on maize stem cells allows studying the *insitu* chemical composition which aids in the degradation of cell walls for biofuel production. The joint analysis can be extended to more than two different kinds of spectroscopies using MCIA. Such kinds of joint analysis using MCIA can also be applied for an integrative analysis of multiple microarray or RNAseq platform comparisons.

Correlation and altered regulation of global transcript levels and translatome

One of the key challenges in biology is to analyze how strong the correlation between mRNA expression levels and protein abundance is. There is a growing body of literature reporting studies integrating either the transcriptome to the proteome or the transcriptome to the metabolome (Kleffmann et al., 2004, Baginsky et al., 2010, Tohge et al., 2005, Hannah et al., 2010, Osorio et al., 2012). Originally studied in yeast and mammalian cells (Halbeisen and Gerber, 2009, Tebaldi et al., 2012), translational control has been identified as the intermediate level in the flow of genetic information and might reflect why particular mRNA's do not necessarily correlate with those of the encoded proteins. Chapter 5 of this thesis addressed two sets of biological questions (1) first, using cell-specific datasets a systematic assessment of the variation of expression was made between the transcriptome and translatome levels, and (2) Secondly, the data used to study the two system levels was from Arabidopis root cells and hence components and processes associated to cell wall related genes were also studied. This is because Arabidopsis serves as a powerful plant model for the identification and functional characterization of genes encoding enzymes involved in cell wall biosynthesis.

Since the two datasets originated from different labs, appropriate normalization procedures were adopted to enable an integrative system-level comparison. Moreover, as detailed in Chapter 2 of the thesis averaging the replicates was done using the multivariate correlation estimator. The cell specificity of the two employed datasets was further classified into identical and common based on the promoters used to drive the expression of the particular cell-type under study. Careful comparison of the promoter sequences and the identification of the genomic coordinates helped to classify the promoters as identical or common (cf Table 5.1). In contrast to other studies investigating different system levels using only the differentially expressed genes, here genome scale system-level comparisons were adopted. Pearson correlation coefficient (PCC) was used in compari-

son to the Spearmans rank correlation coefficient to estimate the similarity as the former is sensitive to the actual expression values while the latter loses some of the precision in the data through ranking. Moreover, the investigation of co-expression relationships on a global scale allows the identification of altered regulation of gene expression levels across both system levels. The similarity of the co-expression and co-translation network was determined by the PCC of the adjacency representations of the networks, and a full matrix correlation showed that the whole-network correlation is significantly lower than expected by chance. This finding suggests a global, or genome-wide, dissimilarity of co-expression and co translation, which has previously been interpreted as an uncoupling of both system levels (Halbeisen and Gerber, 2009, Tebaldi et al., 2012). Moreover, this finding serves as a starting point to analyze individual gene neighborhoods in detail as it is done by the expression conservation score (EC score) approach. The computation of EC scores was used to assess the similarity of the co-expression relationships for all genes contrasting the translatome and transcriptome networks (cf. Figure 5.2). Twelve genes displaying both conserved expression levels and co-expression relationships (based on EC scores) across the two system-levels were identified to be transcription factors, or those containing predicted DNA binding protein domains (Table 5.2). Additional network properties of these twelve genes were computed to further explain the conservation/similarity of gene neighborhoods. Thus, the computational analysis considered a systematic assessment of the degree of co-ordination and divergence on the global level between the two levels of cellular organization. In addition, key insights into the biological processes associated to cell wall and root development that display conserved and divergent patterns of transcription and translation were identified.

To complement the analysis done on the entire system-level, cell-type centric analysis elucidates whether transcriptional and translational patterns are conserved across multiple cell-types. Through a series of Tukey Honest Significant tests, the differentially expressed genes from the identical promoter dataset reveal significant pair-wise cell-type differences across the two system-levels. Network motifs were constructed using cell-types as nodes and cell-type similarity indicated by an edge. These network motifs are used to display the significant differences in expression or translation. In addition, it also reflects the number of genes that coincide with any particular network motif, indicating that a particular cell-type is dissimilar based on the genes expression or translation profile.

All the approaches in Chapter 5 emphasize statistical significance at each level of the conducted analysis step. The outlined procedure for computing the PCC and EC scores across two system levels strongly relies on the employed bootstrapping procedures. As a consequence, it ensures robust classification of genes whose total and polysomal associ-

ated mRNA levels are coupled (Z-scores >1.96) or uncoupled (Z-scores < –1.96). With respect to the Fisher transformation of the PCC values by Z-transformation, normal distribution of PCC correlation values was ensured. This allows to limit effects that arise when applying the proposed approach on datasets with varying numbers of observations, here corresponding to promoters. Characterization of the biological processes of the genes identified in each step of the analysis was done by gene set enrichment analysis. Hypergeometric distribution was used to test for the probability that a specific set of genes annotates with the same gene ontology term. Finally, where applicable, derived p-values are adjusted to account for the testing of multiple hypotheses.

Although CCA has been used in the integration of large scale microarray datasets, Chapter 5 focuses on establishing a novel analysis pipeline which takes into account the cell-type specificity across the system-levels. A more detailed discussion on the need for a new analysis pipeline for this chapter instead of the CCA is emphasized in Section 5.4 of this thesis. Moreover, the use of EC scores elucidates the "expression content" of the rewired genes, successfully illustrating the co-expression/co-translation relationships between gene neighborhoods across the system-levels.

Synopsis

Case study	Plant species	Number of observations (n)	Number of variables (p)	Number of variables (q)
1	Cotton	32	11	5
2	Maize	47	1090	274
3	Arabidopsis	22810	4	4
			8	8

Table 6.1: An overview of the heterogeneous datasets used in this thesis. The complexity of the datasets in terms of the number of observations and variables is illustrated here. In case study 1, the observations correspond to the number of cotton lines whereas the variables 'p' and 'q' correspond to the monoclonal antibodies and fiber characteristics, respectively. In case study 2, the observations correspond to the common Raman and infrared spectra whereas the variables 'p' and 'q' corresponds to the wavenumbers (Raman) and wavenumbers (Infrared), respectively. The observations were obtained by mapping the common pixels between the Raman and infrared. In case study 3, the observations correspond to the number of genes whereas the variables 'p' and 'q' correspond to the profiled transcriptome and translatome cell-types, respectively. The 'p' and 'q' variables are provided for both the identical and common promoters.

As a summary, Table 6.1 illustrates the complexity and heterogeneity of the datasets used in this thesis. Clearly, each case study has different data dimensions and require appropriate statistical methods to investigate particular biological questions. The use of CCA in case study 1 illustrates the relationship between the two datasets in a descriptive manner. Moreover, use of CCA is appropriate for the 'n > p+q' condition of the datasets. For a predictive approach to deal with the two datasets, sPLS was more suitable as it takes into account variable selection and model prediction of the fiber characteristics. In case study 2, 'n < p + q' condition of the available datasets makes it appropriate to use MCIA instead of CCA. Case study 2 is a typical example of high dimensional data analysis, where the number of variables exceeds the observations. In such cases, CCA does not provide sparse linear solutions and may lack biological plausibility and interpretability. Thus, MCIA was used in order to extract both unique and common information between the two datasets. In addition, the results from MCIA were also compared to the results from principal component analysis. When compared to the other case studies, the data from Arabidopsis was complex and had 22000 genes. The biological question under investigation in this chapter focuses to identify the correlation of gene expression patterns between two different system-levels and also analyze cell-type specific expression patterns. Although CCA is a descriptive method which determines linear combinations of all variables of each type with maximal correlation between the two linear combinations, it is not suited in this particular context as discussed in Section 5.4 of this thesis. Hence, the novel analysis pipeline in Chapter 5 takes into account the expression conservation of the genes and identifies co-regulation of gene expression across system levels. Moreover, the established pipeline classifies the genes into coupled and uncoupled category across the transcriptome and translatome. Overall, it is justified that each of the methods employed in the different case-studies helps to answer different kinds of biological questions. Table 2.1 highlights some of the other methods which could be applied in two or more than two data block setting. In short, the essential steps to be highlighted in integrative data analysis are five-fold: (1) the centrality of the biological question; (2) predictive or descriptive nature of the statistical analysis; (3) choosing the right method based on the question under study; (4) appropriate pre-processing of the datasets; and (5) ensuring robustness of the analysis using suitable hypothesis testing and bootstrapping procedures.

Future perspectives

The systematic integrative analysis of heterogeneous data envisages the relationship between and within different biological layers for extensive knowledge discovery. The main findings presented in Chapter 3, 4, and 5 implies that integrative analysis is insightful than analysis of individual datasets and how inter-relationships between different datasets can be exploited to understand cell wall related biological questions in crop species like Cotton, Maize and Arabidopsis. Application of the outlined approaches to situations involving more than two data tables including times series datasets could help to capture the dynamics of the response of the biological system. However, cell wall related mechanisms are very complex and vary among the same plant species, different tissues or even the same tissue at different developmental stages. In cross-species translation type of studies, it is important to highlight that the role of the cell wall components need to be tested in diverse genotypes, species, and specific tissues.

Bibliography

Abdi, H., Williams, L. J., and Valentin, D. (2013). Multiple factor analysis: principal component analysis for multitable and multiblock data sets. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5:149–179.

Abdi, H., Williams, L. J., Valentin, D., and Bennani-Dosse, M. (2012). STATIS and DISTATIS: optimum multitable principal component analysis and three way metric multidimensional scaling. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4:124–167.

Agarwal, U. P. (2014). 1064 nm FT-Raman spectroscopy for investigations of plant cell walls and other biomass materials. *Frontiers in plant science*, 5:490.

Aittokallio, T. (2010). Dealing with missing values in large-scale studies: microarray data imputation and beyond. *Briefings in bioinformatics*, 11:253–264.

Al-Ghazi, Y., Bourot, S., Arioli, T., Dennis, E. S., and Llewellyn, D. J. (2009). Transcript profiling during fiber development identifies pathways in secondary metabolism and cell wall structure that may contribute to cotton fiber quality. *Plant & cell physiology*, 50:1364–81.

Allouche, F., Hanafi, M., Jamme, F., Robert, P., Barron, C., Guillon, F., and Devaux, M. (2012a). Coupling hyperspectral image data having different spatial resolutions using Multiple Co-inertia Analysis. *Chemometrics and Intelligent Laboratory Systems*, 117:200–212.

Allouche, F., Hanafi, M., Jamme, F., Robert, P., Guillon, F., and Devaux, M. (2012b). Coupling hyperspectral image data having different spatial resolutions by extending multivariate inter-battery Tucker analysis. *Chemometrics and Intelligent Laboratory Systems*, 113:43–51.

Alonso-Simón, A., García-Angulo, P., Mélida, H., Encina, A., Álvarez, J. M., and Acebes, J. L. (2011). The use of FTIR spectroscopy to monitor modifications in plant cell wall architecture caused by cellulose biosynthesis inhibitors. *Plant signaling & behavior*, 6:1104–1110.

Aoki-Kinoshita, K. F. and Kanehisa, M. (2006). Bioinformatics approaches in glycomics and drug discovery. *Current opinion in molecular therapeutics*, 8:514–520.

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., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25:25–29.

Baginsky, S., Hennig, L., Zimmermann, P., and Gruissem, W. (2010). Gene expression analysis, proteomics, and network discovery. *Plant physiology*, 152:402–410.

Bajorski, P. (2009). On the reliability of PCA for complex hyperspectral data. pages 1–5.

Barrett, T. and Edgar, R. (2006). Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods in enzymology*, 411:352–369.

Barron, C., Robert, P., Guillon, F., Saulnier, L., and Rouau, X. (2006). Structural heterogeneity of wheat arabinoxylans revealed by Raman spectroscopy. *Carbohydrate research*, 9:1186–1191.

Benfey, P. N., Bennett, M., and Schiefelbein, J. (2010). Getting to the root of plant biology: impact of the Arabidopsis genome sequence on root research. *The Plant journal : for cell and molecular biology*, 61:992–1000.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57:289 – 300.

Bertozzi, C. R. and Sasisekharan, R. (2009). Glycomics. In A, V., Cummings, R., Esko, J., Freeze, H., Stanley, P., Bertozzi, C., Hart, G., and Etzler, M., editors, *Essentials of glycobiology*. Cold Spring Harbor Laboratory Press, NewYork, USA.

Bevan, M. and Walsh, S. (2005). The Arabidopsis genome: a foundation for plant research. *Genome research*, 15:1632–1642.

Bilder, R., Sabb, F., Cannon, T., London, E., Jentsch, J., Parker, D. S., Poldrack, R., Evans, C., and Freimer, N. (2009). Phenomics: the systematic study of phenotypes on a genome-wide scale. *Neuroscience*, 164:30–42.

Birnbaum, K., Jung, J. W., Wang, J. Y., Lambert, G. M., Hirst, J. A., Galbraith, D. W., and Benfey, P. N. (2005). Cell type-specific expression profiling in plants via cell sorting of protoplasts from fluorescent reporter lines. *Nature methods*, 2:615–619.

Birnbaum, K., Shasha, D. E., Wang, J. Y., Jung, J. W., Lambert, G. M., Galbraith, D. W., and Benfey, P. N. (2003). A gene expression map of the arabidopsis root. *Science*, 302:1956–1960.

Blixt, O., Head, S., Mondala, T., Scanlan, C., Huflejt, M. E., Alvarez, R., Bryan, M. C., Fazio, F., Calarese, D., Stevens, J., Razi, N., Stevens, D. J., Skehel, J. J., van Die, I., Burton, D. R., Wilson, I. A., Cummings, R., Bovin, N., Wong, C.-H., and Paulson, J. C. (2004). Printed covalent glycan array for ligand profiling of diverse glycan binding proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 101:17033–17038.

Bougeard, S., Mostafa, E., Lupo, C., and Chauvin, C. (2011a). Multiblock redundancy analysis from a user's perspective. application in veterinary epidemiology. *Electronic journal of applied statistical analysis*, 4:203–214.

Bougeard, S., Qannari, E. M., Lupo, C., and Hanafi, M. (2011b). From Multiblock Partial Least Squares to Multiblock Redundancy Analysis. A Continuum Approach. *Informatica*, 22:11–26.

Boulesteix, A.-L. and Strimmer, K. (2007). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, 8:32–44.

Bowling, A. J., Vaughn, K. C., and Turley, R. B. (2011). Polysaccharide and glycoprotein distribution in the epidermis of cotton ovules during early fiber initiation and growth. *Protoplasma*, 248:579–590.

Bowman, M. J., Park, W., Bauer, P. J., Udall, J. A., Page, J. T., Raney, J., Scheffler, B. E., Jones, D. C., and Campbell, B. T. (2013). RNA-Seq transcriptome profiling of upland cotton (Gossypium hirsutum L.) root tissue under water-deficit stress. *PloS one*, 8:e82634.

Brady, S. M., Orlando, D. A., Lee, J.-Y., Wang, J. Y., Koch, J., Dinneny, J. R., Mace, D., Ohler, U., and Benfey, P. N. (2007). A high-resolution root spatiotemporal map reveals dominant expression patterns. *Science*, 318:801–806.

Brandt, S. P. (2005). Microgenomics: gene expression analysis at the tissue-specific and single-cell levels. *Journal of experimental botany*, 56:495–505.

Brink-Jensen, K., Bak, S., Jørgensen, K., and Ekstrøm, C. T. (2013). Integrative analysis of metabolomics and transcriptomics data: a unified model framework to identify underlying system pathways. *PloS one*, 8:e72116.

Bryan, M. C. and Wong, C.-H. (2004). Aminoglycoside array for the high-throughput analysis of small moleculeRNA interactions. *Tetrahedron Letters*, 45:3639–3642.

Burger, J. and Gowen, A. (2011). Data handling in hyperspectral image analysis. *Chemometrics and Intelligent Laboratory Systems*, 108:13–22.

Butte, A. and Kohane, I. (2003). *Relevance networks: a first step toward finding genetic regulatory networks within microarray data.* Springer, NewYork, USA.

Campbell, C. T. and Yarema, K. J. (2005). Large-scale approaches for glycobiology. *Genome biology*, 6:236.

Chapelle, C. and Carpita, N. (1998). Plant cell walls as targets for biotechnology. *Current Opinion in Plant Biology*, 1:179–185.

Chartrand, G. (1985). *Introductory graph theory.* Dover publishers, NewYork, USA.

Chen, F. and Dixon, R. (2007). Lignin modification improves fermentable sugar yields for biofuel production. *Nat Biotechnology*, 25:759–761.

Chessel, D. and Hanafi, M. (1994). Analyses de la co-inertie de K nuages de points. *Revue de Statistique Appliquée*, 44:35–60.

Chessel, D. and Hanafi, M. (1996). Analysis of the co-inertia of K tables Analyses de la co-inertie de K nuages de points. *Revue de statistique appliquée*, 44:35 – 60.

Chun, H. and Kele, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 72:3–25.

Chylińska, M., Szymaska-Chargot, M., and Zdunek, A. (2014). Imaging of polysaccharides in the tomato cell wall with Raman microspectroscopy. *Plant methods*, 10:14.

Coppi, R. (1994). An introduction to multiway data and their analysis. *Computational Statistics & Data Analysis*, 18:3–13.

Culhane, A. C., Perrière, G., Considine, E. C., Cotter, T. G., and Higgins, D. G. (2002). Between-group analysis of microarray data. *Bioinformatics*, 18:1600–8.

De Livera, A. M., Dias, D. A., De Souza, D., Rupasinghe, T., Pyke, J., Tull, D., Roessner, U., McConville, M., and Speed, T. P. (2012). Normalizing and integrating metabolomics data. *Analytical chemistry*, 84:10768–10776.

De Paz, J., Noti, C., and Seeberger, P. H. (2006). Microarrays of synthetic heparin oligosaccharides. *Journal of the American Chemical Society*, 128:2766–2777.

De Roover, K., Ceulemans, E., and Timmerman, M. E. (2012). How to perform multiblock component analysis in practice. *Behavior research methods*, 44:41–56.

Devaux, M., Allouche, F., Jamme, F., Robert, P., and Guillon, F. (2010). Spatial and spectral normalisation of hyperspectral images.

Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Lalo, D., Le Gall, C., Schaffer, B., Le Crom, S., Guedj, M., and Jaffrzic, F. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in bioinformatics*, 14:671–683.

Disney, M. D. and Barrett, O. J. (2007). An aminoglycoside microarray platform for directly monitoring and studying antibiotic resistance. *Biochemistry*, 46:11223–11230.

Doldec, S. and Chessel, D. (1987). Rythmes saisonniers et composantes stationnelles en milieu aquatique. i. description dun plan dobservations complet par projection de variables. *Acta Oecologica*, 8:403426.

Durinck, S. (2008). Pre-processing of microarray data and analysis of differential expression. In Keith, J., editor, *Bioinformatics*. Humana Press, Totowa, New Jersey, USA.

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

Edwards, J. W. and Coruzzi, G. M. (1990). Cell-specific gene expression in plants. *Annual review of genetics*, 24:275–303.

Egan, A. N., Schlueter, J., and Spooner, D. M. (2012). Applications of next-generation sequencing in plant biology. *American journal of botany*, 99:175–185.

Ehrhardt, D. W. and Frommer, W. B. (2012). New technologies for 21st century plant science. *The Plant cell*, 24:374–394.

Feron, G., Ayed, C., Qannari, E. M., Courcoux, P., Laboure, H., and Guichard, E. (2014). Understanding aroma release from model cheeses by a statistical multiblock approach on oral processing. *PloS one*, 9:e93113.

Fujita, A., Sato, J., Rodrigues, L. D. O., Ferreira, C. E., and Sogayar, M. C. (2006). Evaluating different methods of microarray data normalization. *BMC bioinformatics*, 7:469.

Fukushima, A., Kusano, M., Redestig, H., Arita, M., and Saito, K. (2009). Integrated omics approaches in plant systems biology. *Current opinion in chemical biology*, 13:532–538.

Gama, C. I., Tully, S. E., Sotogaku, N., Clark, P. M., Rawat, M., Vaidehi, N., Goddard, W. A., Nishi, A., and Hsieh-Wilson, L. C. (2006). Sulfation patterns of glycosaminoglycans encode molecular recognition and activity. *Nature chemical biology*, 2:467–473.

Gebauer, F. and Hentze, M. W. (2004). Molecular mechanisms of translational control. *Nature reviews. Molecular cell biology*, 5:827–35.

Geladi, P. and Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, 185:1–17.

Gierlinger, N. (2014). Revealing changes in molecular composition of plant cell walls on the micron-level by Raman mapping and vertex component analysis (VCA). *Frontiers in plant science*, 5:306.

Gierlinger, N., Keplinger, T., and Harrington, M. (2012). Imaging of plant cell walls by confocal Raman microscopy. *Nature protocols*, 7:1694–1708.

Gierlinger, N. and Schwanninger, M. (2006). Chemical imaging of poplar wood cell walls by confocal Raman microscopy. *Plant Physiology*, 140:12461254.

Gilbert, M., Kim, H., Tang, Y., Naoumkina, M., and Fang, D. (2014). Comparative Transcriptome Analysis of Short Fiber Mutants Ligon-Lintless 1 And 2 Reveals Common Mechanisms Pertinent to Fiber Elongation in Cotton (Gossypium hirsutum L.). *PLoS ONE*, 9:e95554.

Gilbert, M. K., Turley, R. B., Kim, H. J., Li, P., Thyssen, G., Tang, Y., Delhom, C. D., Naoumkina, M., and Fang, D. D. (2013). Transcript profiling by microarray and marker analysis of the short cotton (Gossypium hirsutum L.) fiber mutant Ligon lintless-1 (Li1). *BMC genomics*, 14:403.

Goldberg, R., Morvan, C., Jauneau, A., and Jarvis, M. (1996). Pectins and Pectinases, Proceedings of an International Symposium. In Aalbersberg, W., Hamer, R., Jasperse, P., de Jongh, H., de Kruif, C., Walstra, P., and de Wolf, F., editors, *Industrial Proteins in Perspective*. Elsevier, Wageningen, Netherlands.

Goncalves, A., Tikhonov, A., Brazma, A., and Kapushesky, M. (2011). A pipeline for RNA-seq data processing and quality assessment. *Bioinformatics*, 27:867–9.

González, I., Déjean, S., Martin, P. G. P., Gonçalves, O., Besse, P., and Baccini, A. (2009). Highlighting relationships between heterogeneous biological data through graphical displays based on regularized canonical correlation analysis. *Journal of Biological Systems*, 17:173–199.

Gonzalez, I., Lê Cao, K.-A., Davis, M., and Dejean, S. (2012). Visualising associations between paired 'omics' data sets. *BioData Mining*, 5:19.

Gonzalez, R. C. and Woods, R. E. (2006). *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc., Upper Saddle River, New Jersey, USA.

Gowen, A. A., O'Donnell, C. P., Taghizadeh, M., Cullen, P. J., Frias, J. M., and Downey, G. (2008). Hyperspectral imaging combined with principal component analysis for bruise damage detection on white mushrooms (*Agaricus bisporus*). *Journal of Chemometrics*, 22:259–267.

Grahn, H. and Geladi, P. (2007). *Techniques and Applications of Hyperspectral Image Analysis*. John Wiley & Sons, Ltd, Chichester, UK.

Grolleau, A., Bowman, J., Pradet-Balade, B., Puravs, E., Hanash, S., Garcia-Sanz, J. A., and Beretta, L. (2002). Global and specific translational control by rapamycin in T cells uncovered by microarrays and proteomics. *The Journal of biological chemistry*, 277:22175–22184.

Gromski, P. S., Xu, Y., Kotze, H. L., Correa, E., Ellis, D. I., Armitage, E. G., Turner, M. L., and Goodacre, R. (2014). Influence of missing values substitutes on multivariate analysis of metabolomics data. *Metabolites*, 4:433–452.

Gupta, G., Surolia, A., and Sampathkumar, S.-G. (2010). Lectin microarrays for glycomic analysis. *Omics : a journal of integrative biology*, 14:419–436.

Hack, C. J. (2004). Integrated transcriptome and proteome data: The challenges ahead. *Briefings in Functional Genomics and Proteomics*, 3:212–219.

Haigler, C. H., Betancur, L., Stiff, M. R., and Tuttle, J. R. (2012). Cotton fiber: a powerful single-cell model for cell wall and cellulose research. *Frontiers in plant science*, 3:104.

Halbeisen, R. and Gerber, A. (2009). Stress-dependent coordination of transcriptome and translatome in yeast. *PLoS biology*, 7:e1000105.

Hamid, J. S., Hu, P., Roslin, N. M., Ling, V., Greenwood, C. M. T., and Beyene, J. (2009). Data integration in genetics and genomics: methods and challenges. *Human genomics and proteomics (Online publication)*.

Hamilton, J. P. and Buell, C. R. (2012). Advances in plant genome sequencing. *The Plant journal : for cell and molecular biology*, 70:177–190.

Hanafi, M., Kohler, A., and Qannari, E.-M. (2011). Connections between multiple coinertia analysis and consensus principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 106:37–40.

Hannah, M. A., Caldana, C., Steinhauser, D., Balbo, I., Fernie, A. R., and Willmitzer, L. (2010). Combined transcript and metabolite profiling of Arabidopsis grown under widely variant growth conditions facilitates the identification of novel metabolite-mediated regulation of gene expression. *Plant physiology*, 152:2120–2129.

Hansen, M. A. T., Kristensen, J. B., Felby, C., and Jø rgensen, H. (2011). Pretreatment and enzymatic hydrolysis of wheat straw (Triticum aestivum L.)–the impact of lignin relocation and plant tissues on enzymatic accessibility. *Bioresource technology*, 102:2804–2811.

Hanson, S. R., Hsu, T.-L., Weerapana, E., Kishikawa, K., Simon, G. M., Cravatt, B. F., and Wong, C.-H. (2007). Tailored glycoproteomics and glycan site mapping using saccharide-selective bioorthogonal probes. *Journal of the American Chemical Society*, 129:7266–7277.

Hao, Z. and Mohnen, D. (2014). A review of xylan and lignin biosynthesis: foundation for studying Arabidopsis irregular xylem mutants with pleiotropic phenotypes. *Critical reviews in biochemistry and molecular biology*, 49:212–241.

Hardoon, D. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16:2639–2664.

Heredia, A., Jiménez, A., and Guillén, R. (1995). Composition of plant cell walls. *Zeitschrift für Lebensmittel-Untersuchung und -Forschung*, 200:24–31.

Himmelsbach, D. and Akin., D. (1998). Near-InfraredFourier-TransformRaman Spectroscopy of Flax (*Linum usitatissimum L.*) Stems. *Journal of Agriculture and Food Chemistry*, 46:991–998.

Hisano, H., Nandakumar, R., and Wang, Z.-Y. (2009). Genetic modification of lignin biosynthesis for improved biofuel production. *In Vitro Cellular & Developmental Biology-Plant*, 45:306–313.

Hood, L. (2003). Systems biology: integrating technology, biology, and computation. *Mechanisms of Ageing and Development*, 124:9–16.

Hori, R. and Sugiyama, J. (2003). A combined FT-IR microscopy and principal component analysis on softwood cell walls. *Carbohydrate Polymers*, 52:449–453.

Horn, P. J. and Chapman, K. D. (2012). Lipidomics in tissues, cells and subcellular compartments. *The Plant journal : for cell and molecular biology*, 70:69–80.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28:321–377.

Hotelling, H. (1951). *A Generalized T Test and Measure of Multivariate Dispersion*. University of California Press, Berkeley, California.

Houle, D., Govindaraju, D. R., and Omholt, S. (2010). Phenomics: the next challenge. *Nature reviews. Genetics*, 11:855–66.

Hsu, T.-L., Hanson, S. R., Kishikawa, K., Wang, S.-K., Sawa, M., and Wong, C.-H. (2007). Alkynyl sugar analogs for the labeling and visualization of glycoconjugates in cells. *Proceedings of the National Academy of Sciences of the United States of America*, 104:2614–2619.

Huang, G.-Q., Gong, S.-Y., Xu, W.-L., Li, W., Li, P., Zhang, C.-J., Li, D.-D., Zheng, Y., Li, F.-G., and Li, X.-B. (2013). A fasciclin-like arabinogalactan protein, GhFLA1, is involved in fiber initiation and elongation of cotton. *Plant physiology*, 161:1278–1290.

Huang, X., Pan, W., Park, S., Han, X., Miller, L. W., and Hall, J. (2004). Modeling the relationship between LVAD support time and gene expression changes in the human heart by penalized partial least squares. *Bioinformatics*, 20:888–94.

Huttenhower, C., Hibbs, M., Myers, C., and Troyanskaya, O. G. (2006). A scalable method for integration and functional analysis of multiple microarray datasets. *Bioinformatics*, 22:2890–2897.

Ideker, T., Galitski, T., and Hood, L. (2001). A new approach to decoding life: systems biology. *Annual review of genomics and human genetics*, 2:343–372.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4:249–264.

Ivanov, V. B. (2007). Oxidative stress and formation and maintenance of root stem cells. *Biochemistry*, 72:1110–1114.

Iyer-Pascuzzi, A. S. and Benfey, P. N. (2010). Fluorescence-activated cell sorting in plant developmental biology. *Methods in molecular biology*, 655:313–319.

Jamme, F., Robert, P., Bouchet, B., Saulnier, L., Dumas, P., and Guillon, F. (2008). Aleurone cell walls of wheat grain: high spatial resolution investigation using synchrotron infrared microspectroscopy. *Applied spectroscopy*, 62:895–900.

Jiang, K., Meng, Y. L., and Feldman, L. J. (2003). Quiescent center formation in maize roots is associated with an auxin-regulated oxidizing environment. *Development*, 130:1429–1438.

Jones, H. D., Haaland, D. M., Sinclair, M. B., Melgaard, D. K., Collins, A. M., and Timlin, J. A. (2012). Preprocessing strategies to improve MCR analyses of hyperspectral images. *Chemometrics and Intelligent Laboratory Systems*, 117:149–158.

Joyce, A. R. and Palsson, B. O. (2006). The model organism as a system: integrating 'omics' data sets. *Nature reviews. Molecular cell biology*, 7:198–210.

Jung, H. and Casler, M. (2006). Maize stem tissues: impact of development on cell wall degradability. *Journal of crop science*, 46:1801–1809.

Kačuráková, M., Capeka, P., Sasinková, V., Wellner, N., and Ebringerová, A. (2000). FT-IR study of plant cell wall model compounds: pectic polysaccharides and hemicelluloses. *Carbohydrate Polymers*, 43:195–203.

Kačuráková, M., Capeka, P., Sasinková, V., Wellner, N., Ebringerová, A., Hromadkova, Z., Wilson, R., and Belton, P. (1999). Characterisation of xylan-type polysaccharides and associated cell wall components by FT-IR and FT-Raman spectroscopies. *Food Hydrocolloid*, 13:3541.

Kačuráková, M., Smith, A., Gidley, M., and Wilson, R. (2002). Molecular interactions in bacterial cellulose composites studied by 1D FT-IR and dynamic 2D FT-IR spectroscopy. *Carbohydrate Research*, 337:11451153.

Keegstra, K. (2010). Plant cell walls. *Plant physiology*, 154:483–486.

Kerr, M. K., Martin, M., and Churchill, G. A. (2000). Analysis of variance for gene expression microarray data. *Journal of computational biology : a journal of computational molecular cell biology*, 7:819–837.

Kettenring, J. R. (1966). Simultaneous factor analysis ofseveral gramina matrices. *Psychometrika*, 31:413–419.

Kitano, H. (2002a). Computational systems biology. *Nature*, 420:206–10.

Kitano, H. (2002b). Systems biology: a brief overview. *Science*, 295:1662–1664.

Kleffmann, T., Russenberger, D., von Zychlinski, A., Christopher, W., Sjölander, K., Gruissem, W., and Baginsky, S. (2004). The Arabidopsis thaliana chloroplast proteome reveals pathway abundance and novel protein functions. *Current biology*, 14:354–362.

Klie, S. and Nikoloski, Z. (2012). The Choice between MapMan and Gene Ontology for Automated Gene Function Prediction in Plant Science. *Frontiers in genetics*, 3:115.

Kohl, M., Megger, D. A., Trippler, M., Meckel, H., Ahrens, M., Bracht, T., Weber, F., Hoffmann, A.-C., Baba, H. A., Sitek, B., Schlaak, J. F., Meyer, H. E., Stephan, C., and Eisenacher, M. (2014). A practical data processing workflow for multi-OMICS projects. *Biochimica et biophysica acta*, 1844:52–62.

Kubo, M., Udagawa, M., Nishikubo, N., Horiguchi, G., Yamaguchi, M., Ito, J., Mimura, T., Fukuda, H., and Demura, T. (2005). Transcription switches for protoxylem and metaxylem vessel formation. *Genes & development*, 19:1855–1860.

Kueger, S., Steinhauser, D., Willmitzer, L., and Giavalisco, P. (2012). High-resolution plant metabolomics: from mass spectral features to metabolites and from whole-cell analysis to subcellular metabolite distributions. *The Plant journal : for cell and molecular biology*, 70:39–50.

Kuhfeld, W. F. (1986). A note on Roy's largest root. *Psychometrika*, 51:479–481.

Lacape, J.-M., Claverie, M., Vidal, R. O., Carazzolle, M. F., Guimarães Pereira, G. A., Ruiz, M., Pré, M., Llewellyn, D., Al-Ghazi, Y., Jacobs, J., Dereeper, A., Huguet, S., Giband, M., and Lanaud, C. (2012). Deep sequencing reveals differences in the transcriptional landscapes of fibers from two cultivated species of cotton. *PloS one*, 7:e48855.

Lavit, C., Escoufier, Y., Sabatier, R., and Traissac, P. (1994). The ACT (STATIS method). *Computational Statistics & Data Analysis*, 18:97–119.

Lawley, D. N. (1938). A generalization of Fisher's z-test . *Biometrika*, 30:180–187.

Lawrie, C. H., Marafioti, T., Hatton, C. S. R., Dirnhofer, S., Roncador, G., Went, P., Tzankov, A., Pileri, S. A., Pulford, K., and Banham, A. H. (2006). Cancer-associated carbohydrate identification in Hodgkin's lymphoma by carbohydrate array profiling. *International journal of cancer*, 118:3161–3166.

Lê Cao, K.-A., Martin, P., Robert-Granie, C., and Besse, P. (2009). Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics*, 10:34.

Lê Cao, K.-A., Rossouw, D., Robert-Granié, C., and Besse, P. (2008). A sparse PLS for variable selection when integrating omics data. *Statistical applications in genetics and molecular biology*, 7:Article 35.

Lee, J.-Y., Colinas, J., Wang, J. Y., Mace, D., Ohler, U., and Benfey, P. N. (2006). Transcriptional and posttranscriptional regulation of transcription factor expression in Arabidopsis roots. *Proceedings of the National Academy of Sciences of the United States of America*, 103:6055–6060.

Lee, Y. J., Perdian, D. C., Song, Z., Yeung, E. S., and Nikolau, B. J. (2012). Use of mass spectrometry for imaging metabolites in plants. *The Plant journal : for cell and molecular biology*, 70:81–95.

Leurgans, S., Moyeed, R., and Silverman, B. (1976). Canonical Correlation Analysis when the Data are Curves. *Journal of the Royal Statistical Society*, 55:725–740.

Levesque, M., Vernoux, T., Busch, W., Cui, H., Wang, J., Blilou, I., Hassan, H., Nakajima, K., Matsumoto, N., Lohmann, J., Scheres, B., and Benfey, P. (2006). Whole-Genome

Analysis of the SHORT-ROOT Developmental Pathway in Arabidopsis. *PLoS biology*, 4:e249.

Liang, P.-H., Wang, S.-K., and Wong, C.-H. (2007). Quantitative analysis of carbohydrate-protein interactions using glycan microarrays: determination of surface and solution dissociation constants. *Journal of the American chemical society*, 129:11177–11184.

Liepman, A. H., Wightman, R., Geshi, N., Turner, S. R., and Scheller, H. V. (2010). Arabidopsis - a powerful model system for plant cell wall research. *The Plant journal*, 61:1107–1121.

Liew, A. W.-C., Law, N.-F., and Yan, H. (2011). Missing value imputation for gene expression data: computational techniques to recover missing data from available information. *Briefings in bioinformatics*, 12:498–513.

Likić, V. A., McConville, M. J., Lithgow, T., and Bacic, A. (2010). Systems biology: the next frontier for bioinformatics. *Advances in bioinformatics*, 2010:1–10.

Lim, S., Sohn, K. H., and Lee, C. (2001). Principal component analysis for compression of hyperspectral images. *Geoscience & Remote Sensing symposium*, 1:97–99.

Lin, S.-Y., Chen, P.-W., Chuang, M.-H., Juntawong, P., Bailey-Serres, J., and Jauh, G.-Y. (2014). Profiling of translatomes of in vivo-grown pollen tubes reveals genes with roles in micropylar guidance during pollination in Arabidopsis. *The Plant cell*, 26:602–618.

Lutz, J. G. and Eckert, T. L. (1993). The Relationship between Canonical Correlation Analysis and Multivariate Multiple Regression. *Educational and Psychological Measurement*, 54:666–675.

Marchal, I., Golfier, G., Dugas, O., and Majed, M. (2003). Bioinformatics in glycobiology. *Biochimie*, 85:75–81.

Marcus, S., Verhertbruggen, Y., Hervé, C., Ordaz-Ortiz, J., Farkas, V., Pedersen, H., Willats, W., and Knox, J. (2008). Pectic homogalacturonan masks abundant sets of xyloglucan epitopes in plant cell walls. *BMC plant biology*, 8:60.

Marcus, S. E., Blake, A. W., Benians, T. A. S., Lee, K. J. D., Poyser, C., Donaldson, L., Leroux, O., Rogowski, A., Petersen, H. L., Boraston, A., Gilbert, H. J., Willats, W. G. T., and Knox, J. P. (2010). Restricted access of proteins to mannan polysaccharides in intact plant cell walls. *The Plant journal : for cell and molecular biology*, 64:191–203.

MATLAB (2013). *Version 8.1.0 (R2013a)*. The MathWorks Inc., Natick, Massachusetts.

McCann, M. and Rose, J. (2010). Blueprints for building plant cell walls. *Plant physiology*, 153:365.

McCann, M. C., Bush, M., Milioni, D., Sado, P., Stacey, N. J., Catchpole, G., Defernez, M., Carpita, N. C., Hofte, H., Ulvskov, P., Wilson, R. H., and Roberts, K. (2001). Approaches to understanding the functional architecture of the plant cell wall. *Phytochemistry*, 57:811–821.

McCann, M. C. and Carpita, N. C. (2008). Designing the deconstruction of plant cell walls. *Current opinion in plant biology*, 11:314–20.

McCartney, L., Marcus, S. E., and Knox, J. P. (2005). Monoclonal antibodies to plant cell wall xylans and arabinoxylans. *The journal of histochemistry and cytochemistry*, 53:543–546.

Meikle, P. J., Bonig, I., Hoogenraad, N. J., Clarke, A. E., and Stone, B. A. (1991). The location of (13)-β-glucans in the walls of pollen tubes of Nicotiana alata using a (13)-β-glucan-specific monoclonal antibody. *Planta*, 185:1–8.

Mendes, S., Gómez, M. J. F., Pereira, M. J., Azeitero, U. M., and Villardón, M. P. G. (2010). The efficiency of the Partial Triadic Analysis method: an ecological application. *Biometrical Letters*, 47:83–106.

Michael, T. and Jackson, S. (2013). The first 50 plant genomes. *Plant genome*, 6:2.

Mikulits, W., Pradet-Balade, B., Habermann, B., Beug, H., Garcia-Sanz, J. A., and Müllner, E. W. (2000). Isolation of translationally controlled mRNAs by differential screening. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 14:1641–1652.

Milioni, D., Sado, P.-E., Stacey, N. J., Roberts, K., and McCann, M. C. (2002). Early gene expression associated with the commitment and differentiation of a plant tracheary element is revealed by cDNA-amplified fragment length polymorphism analysis. *The Plant cell*, 14:2813–2824.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science*, 298:824–827.

Minorsky, P. V. (2002). The wall becomes surmountable. *Plant physiology*, 128:345–53.

Mochida, K. and Shinozaki, K. (2010). Genomics and bioinformatics resources for crop improvement. *Plant and Cell Physiology*, 51:497–523.

Mochida, K. and Shinozaki, K. (2011). Advances in omics and bioinformatics tools for systems analyses of plant functions. *Plant & cell physiology*, 52:2017–2038.

Moller, I., Marcus, S. E., Haeger, A., Verhertbruggen, Y., Verhoef, R., Schols, H., Ulvskov, P., Mikkelsen, J. r. D., Knox, J. P., and Willats, W. (2008). High-throughput screening of monoclonal antibodies against plant cell wall glycans by hierarchical clustering of their carbohydrate microarray binding profiles. *Glycoconjugate journal*, 25:37–48.

Moyon, T., Le Marec, F., Qannari, E. M., Vigneau, E., Le Plain, A., Courant, F., Antignac, J.-P., Parnet, P., and Alexandre-Gouabau, M.-C. (2012). Statistical strategies for relating metabolomics and proteomics data: a real case study in nutrition research area. *Metabolomics*, 8:1090–1101.

Mühlberger, I., Wilflingseder, J., Bernthaler, A., Fechete, R., Lukas, A., and Perco, P. (2011). Computational analysis workflows for Omics data interpretation. *Methods in molecular biology*, 719:379–97.

Mustroph, A., Zanetti, M. E., Jang, C. J. H., Holtan, H. E., Repetti, P. P., Galbraith, D. W., Girke, T., and Bailey-Serres, J. (2009). Profiling translatomes of discrete cell populations resolves altered cellular priorities during hypoxia in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, 106:18843–18888.

Mutwil, M., Debolt, S., and Persson, S. (2008). Cellulose synthesis: a complex complex. *Current Opinion in Plant Biology*, 11:252–257.

Mutwil, M., Ruprecht, C., Giorgi, F. M., Bringmann, M., Usadel, B., and Persson, S. (2009). Transcriptional wiring of cell wall-related genes in Arabidopsis. *Molecular plant*, 2:1015–1024.

Nawy, T., Lee, J.-Y., Colinas, J., Wang, J. Y., Thongrod, S. C., Malamy, J. E., Birnbaum, K., and Benfey, P. N. (2005). Transcriptional profile of the Arabidopsis root quiescent center. *The Plant cell*, 17:1908–1925.

Oh, S., Kang, D. D., Brock, G. N., and Tseng, G. C. (2011). Biological impact of missing-value imputation on downstream analyses of gene expression profiles. *Bioinformatics*, 27:78–86.

Oikawa, A. and Saito, K. (2012). Metabolite analyses of single cells. *The Plant journal : for cell and molecular biology*, 70:30–38.

Okumoto, S. (2012). Quantitative imaging using genetically encoded sensors for small molecules in plants. *The Plant journal: for cell and molecular biology*, 70:108–117.

Oparka, K. J. (1999). Sieve Elements and Companion CellsTraffic Control Centers of the Phloem. *The Plant cell*, 11:739–750.

Oshlack, A., Robinson, M., and Young, M. (2010). From RNA-seq reads to differential expression results. *Genome biology*, 11:220.

Osorio, S., Alba, R., Nikoloski, Z., Kochevenko, A., Fernie, A. R., and Giovannoni, J. J. (2012). Integrative comparative analyses of transcript and metabolite profiles from pepper and tomato ripening and development stages uncovers species-specific patterns of network regulatory behavior. *Plant physiology*, 159:1713–1729.

Palsson, B. and Zengler, K. (2010). The challenges of integrating multi-omic data sets. *Nature Chemical Biology*, 6:787–789.

Pan, Z., Zeng, Y., An, J., Ye, J., Xu, Q., and Deng, X. (2012). An integrative analysis of transcriptome and proteome provides new insights into carotenoid biosynthesis and regulation in sweet orange fruits. *Journal of proteomics*, 75:2670–2684.

Park, S., Lee, M.-R., and Shin, I. (2008). Carbohydrate microarrays as powerful tools in studies of carbohydrate-mediated biological processes. *Chemical communications*, 37:4389–4399.

Pedersen, H., Fangel, J., McCleary, B., Ruzanski, C., Rydahl, M., Ralet, M., Farkas, V., Von-Schantz, L., Marcus, S., Andersen, M., Field, R., Ohlin, M., Knox, J., Clausen, M., and Willats, W. (2012). Versatile high resolution oligosaccharide microarrays for plant glycobiology and cell wall research. *The Journal of biological chemistry*, 287:39429–38.

Peeters, J. K. and Van der Spek, P. J. (2005). Growing applications and advancements in microarray technology and analysis tools. *Cell biochemistry and biophysics*, 43:149–166.

Persson, S., Wei, H., Milne, J., P.P, G., and , Christopher, S. (2005). Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. *Proceedings of the National Academy of Sciences of the United States of America*, 102:8633–8638.

Petricka, J. J., Schauer, M. A., Megraw, M., Breakfield, N. W., Thompson, J. W., Georgiev, S., Soderblom, E. J., Ohler, U., Moseley, M. A., Grossniklaus, U., and Benfey, P. N. (2012). The protein expression landscape of the arabidopsis root. *Proceedings of the National Academy of Sciences*, 109:6811–6818.

Pettolino, F. A., Hoogenraad, N. J., Ferguson, C., Bacic, A., Johnson, E., and Stone, B. A. (2001). A (1–¿4)-beta-mannan-specific monoclonal antibody and its use in the immunocytochemical location of galactomannans. *Planta*, 214:235–42.

Pillai, K. C. S. (1955). Some new test criteria in multivariate analysis. . *Some new test criteria in multivariate analysis.*, 26:17–121.

Pilling, E. and Höfte, H. (2003). Feedback from the wall. *Current opinion in plant biology*, 6:611–616.

Popper, Z. A. (2008). Evolution and diversity of green plant cell walls. *Current opinion in plant biology*, 11:286–92.

Purbasha, S., Elena, B., and Manfred, A. (2009). Plant cell walls throughout evolution: towards a molecular understanding of their design principles. *Journal of experimental botany*, 60:3615–3635.

Qin, L.-X., Rao, Y., Li, L., Huang, J.-F., Xu, W.-L., and Li, X.-B. (2013). Cotton GalT1 encoding a putative glycosyltransferase is involved in regulation of cell wall pectin biosynthesis during plant development. *PloS one*, 8:e59115.

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature genetics*, 32:496–501.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rajasundaram, D., Runavot, J.-L., Guo, X., Willats, W. G. T., Meulewaeter, F., and Selbig, J. (2014a). Understanding the Relationship between Cotton Fiber Properties and Non-Cellulosic Cell Wall Polysaccharides. *PloS one*, 9:e112168.

Rajasundaram, D., Selbig, J., Persson, S., and Klie, S. (2014b). Co-ordination and divergence of cell-specific transcription and translation of genes in arabidopsis root cells. *Annals of botany*, 114:1109–1123.

Rambani, A., Page, T., and Udall, J. (2014). Polyploidy and the petal transcriptome of Gossypium. *BMC plant biology*, 14:3.

Ratner, D. M. and Seeberger, P. H. (2007). Carbohydrate microarrays as tools in HIV glycobiology. *Current pharmaceutical design*, 13:173–183.

Reiter, W. D. (2002). Biosynthesis and properties of the plant cell wall. *Current opinion in plant biology*, 5:536–42.

Rencher, A. C. (2002). *Methods of Multivariate Analysis*. John Wiley & Sons, Inc., NewYork, NY, USA.

Rivals, I., Personnaz, L., Taing, L., and Potier, M.-C. (2007). Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics (Oxford, England)*, 23:401–407.

Roberts, K. (2001). How the cell wall acquired a cellular context. *Plant physiology*, 125:127–30.

Rogers, E. D., Jackson, T., Moussaieff, A., Aharoni, A., and Benfey, P. N. (2012). Cell type-specific transcriptional profiling: implications for metabolite profiling. *The Plant journal : for cell and molecular biology*, 70:5–17.

Roy, S. (1939). p-statistics or some generalizations in analysis of variance appropriate to multivariate problems. *Sankhya*, 4:381–396.

Ruan, Y.-L., Xu, S.-M., White, R., and Furbank, R. T. (2004). Genotypic and developmental evidence for the role of plasmodesmatal regulation in cotton fiber elongation mediated by callose turnover. *Plant physiology*, 136:4104–4113.

Ruprecht, C. and Persson, S. (2012). Co-expression of cell-wall related genes: new tools and insights. *Frontiers in plant science*, 3:83.

Sabatier, R. and Vivien, M. (2008). A new linear method for analyzing four-way multiblock tables: STATIS-4. *Journal of Chemometrics*, 22:399–407.

Sabatini, S., Beis, D., Wolkenfelt, H., Murfett, J., Guilfoyle, T., Malamy, J., Benfey, P., Leyser, O., Bechtold, N., Weisbeek, P., and Scheres, B. (1999). An auxin-dependent distal organizer of pattern and polarity in the Arabidopsis root. *Cell*, 99:463–472.

Sadava, D., Walker, F., and Chrispeels, M. J. (1973). Hydroxyproline-rich cell wall protein (extensin): Biosynthesis and accumulation in growing pea epicotyls. *Developmental Biology*, 30:42–48.

Salnikov, V. V., Grimson, M. J., Seagull, R. W., and Haigler, C. H. (2003). Localization of sucrose synthase and callose in freeze-substituted secondary-wall-stage cotton fibers. *Protoplasma*, 221:175–184.

Sánchez, A., Fernández-Real, J., Vegas, E., Carmona, F., Amar, J., Burcelin, R., Serino, M., Tinahones, F., de Villa, M., Minrro, A., and Reverter, F. (2012). Multivariate methods for the integration and visualization of omics data. In Freitas, A. and Navarro, A., editors, *Bioinformatics for Personalized Medicine*. Springer publishers, Berlin, Germany.

Schauer, N., Semel, Y., Roessner, U., Gur, A., Balbo, I., Carrari, F., Pleban, T., Perez-Melis, A., Bruedigam, C., Kopka, J., Willmitzer, L., Zamir, D., and Fernie, A. R. (2006). Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. *Nature biotechnology*, 24:447–454.

Schena, M. (1996). Genome analysis with gene expression microarrays. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 18:427–431.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467–470.

Schneider, A., Hommel, G., and Blettner, M. (2010). Linear regression analysis: part 14 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt international*, 107:776–782.

Schultz, R. A., Nielsen, T., Zavaleta, J. R., Ruch, R., Wyatt, R., and Garner, H. R. (2001). Hyperspectral imaging: a novel approach for microscopic analysis. *Cytometry*, 43:239–47.

Sene, C., McCann, M. C., Wilson, R. H., and Grinter, R. (1994). Fourier-Transform Raman and Fourier-Transform Infrared Spectroscopy (An Investigation of Five Higher Plant Cell Walls and Their Components). *Plant Physiology*, 106:1623–1631.

Sharp, P. A. (2009). The centrality of RNA. *Cell*, 136:577–80.

Shen-Orr, S. S., Tibshirani, R., Khatri, P., Bodian, D. L., Staedtler, F., Perry, N. M., Hastie, T., Sarwal, M. M., Davis, M. M., and Butte, A. J. (2010). Cell type-specific gene expression differences in complex tissues. *Nature methods*, 7:287–289.

Shin, R., Berg, R. H., and Schachtman, D. P. (2005). Reactive oxygen species and root hairs in Arabidopsis root response to nitrogen, phosphorus and potassium deficiency. *Plant & cell physiology*, 46:1350–1357.

Shin, R. and Schachtman, D. P. (2004). Hydrogen peroxide mediates plant root cell response to nutrient deprivation. *Proceedings of the National Academy of Sciences of the United States of America*, 101:8827–8832.

Simier, M., Blanc, L., Pellegrin, F., and Nandris, D. (1999). Approche simultanée de K couples de tableaux : application à létude des relations pathologie végétale - environnement. *Revue de Statistique Appliquée*, 47:31–46.

Singh, B., Avci, U., Eichler Inwood, S. E., Grimson, M. J., Landgraf, J., Mohnen, D., Sørensen, I., Wilkerson, C. G., Willats, W. G., and Haigler, C. H. (2009). A specialized outer layer of the primary cell wall joins elongating cotton fibers into tissue-like bundles. *Plant physiology*, 150:684–699.

Skov, T., Honoré, A. H., Jensen, H. M., Næ, S., Tormod, E., and Søren, B. (2014). Chemometrics in foodomics: Handling data structures from multiple analytical platforms. *Trends in Analytical Chemistry*, 60:71–79.

Smallwood, M., Beven, A., Donovan, N., Neill, S., Peart, J., Roberts, K., and Knox, J. (1994). Localization of cell wall proteins in relation to the developmental anatomy of the carrot root apex. *The Plant Journal*, 5:237–246.

Smilde, A., Bro, R., and Geladi, P. (2005). *Multi-way Analysis: Applications in the Chemical Sciences*. John Wiley & Sons, West Sussex, England.

Smilde, A. K., Westerhuis, J. A., and Boque, R. (2000). Multiway multiblock component and covariates regression models. *Journal of Chemometrics*, 14:301–331.

Smilde, A. K., Westerhuis, J. A., and de Jong, S. (2003). A framework for sequential multiblock component methods. *Journal of Chemometrics*, 17:323–337.

Smirnova, J. B., Selley, J. N., Sanchez-Cabo, F., Carroll, K., Eddy, A. A., McCarthy, J. E. G., Hubbard, S. J., Pavitt, G. D., Grant, C. M., and Ashe, M. P. (2005). Global gene expression profiling reveals widespread yet distinctive translational responses to different eukaryotic translation initiation factor 2B-targeting stress pathways. *Molecular and cellular biology*, 25:9340–9349.

Sokal, R. and Rohlf, F. (1995). *Biometry: the principles and practice of statistics in biological research.* W.H. Freeman, NewYork, USA.

Somerville, C. (2006). Cellulose synthesis in higher plants. *Annual review of cell and developmental biology*, 22:53–78.

Somerville, C., Bauer, S., Brininstool, G., Facette, M., Hamann, T., Milne, J., Osborne, E., Paredez, A., Persson, S., Raab, T., Vorwerk, S., and Youngs, H. (2004). Toward a systems approach to understanding plant cell walls. *Science*, 306:2206–2211.

Sørensen, I., Domozych, D., and Willats, W. G. T. (2010). How have plant cell walls evolved? *Plant physiology*, 153:366–72.

Sparkes, I. and Brandizzi, F. (2012). Fluorescent protein-based technologies: shedding new light on the plant endomembrane system. *The Plant journal : for cell and molecular biology*, 70:96–107.

Stadler, R., Brandner, J., Schulz, A., Gahrtz, M., and Sauer, N. (1995). Phloem Loading by the PmSUC2 Sucrose Carrier from Plantago major Occurs into Companion Cells. *The Plant cell*, 7:1545–1554.

Stéphanie, B. and Mireille, C. (2014). Multiblock modeling for complex preference study. Application to European preferences for smoked salmon. *Food Quality and Preference*, 32:56–64.

Stigler, S. M. (1989). Francis Galton's Account of the Invention of Correlation. *Statistical Science*, 4:73–79.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102:15545–15550.

Sun, L., Simmons, B. A., and Singh, S. (2011). Understanding tissue specific compositions of bioenergy feedstocks through hyperspectral Raman imaging. *Biotechnology and bioengineering*, 108:286–95.

Sun, R. C., Tomkinson, J., Zhu, W., and Wang, S. Q. (2000). Delignification of maize stems by peroxymonosulfuric acid, peroxyformic acid, peracetic acid, and hydrogen

peroxide, 1. Physicochemical and structural characterization of the solubilized lignins. *Journal of Agricultural and Food Chemistry*, 48:1253–1262.

Swanson-Wagner, R., Briskine, R., Schaefer, R., Hufford, M. B., Ross-Ibarra, J., Myers, C. L., Tiffin, P., and Springer, N. M. (2012). Reshaping of the maize transcriptome by domestication. *Proceedings of the National Academy of Sciences of the United States of America*, 109:11878–11883.

Szymanska-Chargot, M. and Zdunek, A. (2013). Use of FT-IR Spectra and PCA to the Bulk Characterization of Cell Wall Residues of Fruits and Vegetables Along a Fraction Process . *Food biophysics*, 8:29–42.

Tabachnick, B. (2013). *Using multivariate statistics*. Pearson Education, Boston, USA.

Tebaldi, T., Re, A., Viero, G., Pegoretti, I., Passerini, A., Blanzieri, E., and Quattrone, A. (2012). Widespread uncoupling between transcriptome and translatome variations after a stimulus in mammalian cells. *BMC genomics*, 13:220.

Tenenhaus, A., Philippe, C., Guillemot, V., Lê Cao, K., Grill, J., and Frouin, V. (2014). Variable selection for generalized canonical correlation analysis. *Biostatistics*, 15:569–583.

Tenenhaus, A. and Tenenhaus, M. (2011). Regularized Generalized Canonical Correlation Analysis. *Psychometrika*, 76:257–284.

Tenenhaus, M. and Young, F. W. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika*, 50:91–119.

ter Braak, C. J. F. (1986). Canonical correspondence analysis: A new eigenvector technique for multivariate gradient analysis. *Ecology*, 67:1167–1179.

Thakur, B. R., Singh, R. K., and Handa, A. K. (1997). Chemistry and uses of pectin–a review. *Critical reviews in food science and nutrition*, 37:47–73.

Thioulouse, J. (2011). Simultaneous analysis of a sequence of paired ecological tables: A comparison of several methods. *The Annals of Applied Statistics*, 5:2300–2325.

Thioulouse, J., Simier, M., and Chessel, D. (2004). Simultaneous analysis of a sequence of paired ecological tables. *Ecology*, 85:272–283.

Tohge, T., Nishiyama, Y., Hirai, M. Y., Yano, M., Nakajima, J.-i., Awazuhara, M., Inoue, E., Takahashi, H., Goodenowe, D. B., Kitayama, M., Noji, M., Yamazaki, M., and Saito, K. (2005). Functional genomics by integrated analysis of metabolome and transcriptome of Arabidopsis plants over-expressing an MYB transcription factor. *The Plant journal : for cell and molecular biology*, 42:218–235.

Tokumoto, H., Wakabayashi, K., Kamisaka, S., and Hoson, T. (2002). Changes in the sugar composition and molecular mass distribution of matrix polysaccharides during cotton fiber development. *Plant & cell physiology*, 43:411–418.

Tomazevic, D., Likar, B., and Pernus, F. (2002). Comparative evaluation of retrospective shading correction methods. *Journal of microscopy*, 208:212–223.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17:520–525.

Tukey, J. and Braun, H. (1985). *The collected works of John W. Tukey*. Chapman & Hall, NewYork, USA.

Uchiyama, N., Kuno, A., Koseki-Kuno, S., Ebe, Y., Horio, K., Yamada, M., and Hirabayashi, J. (2006). Development of a lectin microarray based on an evanescent-field fluorescence principle. *Methods in enzymology*, 415:341–51.

Van Acker, R., Vanholme, R., Storme, V., Mortimer, J., Dupree, P., and Boerjan, W. (2013). Lignin biosynthesis perturbations affect secondary cell wall composition and saccharification yield in arabidopsis thaliana. *Biotechnology for Biofuels*, 6:46.

Vaughn, K. C. and Turley, R. B. (1999). The primary walls of cotton fibers contain an ensheathing pectin layer. *Protoplasma*, 209:226–237.

Černá, M., Barros, A. S., Nunes, A., Rocha, S. M., Delgadillo, I., Čopková, J., and Coimbra, M. A. (2003). Use of FT-IR spectroscopy as a tool for the analysis of polysaccharide food additives. *Carbohydrate Polymers*, 51:383–389.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S-PLUS*. Springer, Newyork.

Verhertbruggen, Y., Marcus, S. E., Haeger, A., Ordaz-Ortiz, J. J., and Knox, J. P. (2009). An extended set of monoclonal antibodies to pectic homogalacturonan. *Carbohydrate research*, 344:1858–1862.

Vidal, M., Rubio, A., and Manuel, J. (2012). Pre-processing of hyperspectral images. essential steps before image analysis. *Chemometrics and Intelligent Laboratory Systems*, 117:138–148.

Vidala, M. and Manuel, J. (2012). Pre-processing of hyperspectral images. Essential steps before image analysis. *Chemometrics and Intelligent Laboratory Systems*, 117:138–148.

Vinod, H. D. (1976). Canonical ridge and econometrics of joint production. *Journal of Econometrics*, 4:147–166.

Vivien, M. and Sabatier, R. (2004). A generalization of STATIS-ACT strategy: DO-ACT for two multiblocks tables. *Computational Statistics & Data Analysis*, 46:155–171.

Vogel, C. and Marcotte, E. M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature reviews. Genetics*, 13:227–232.

von der Lieth, C.-W. (2004). Bioinformatics for glycomics: Status, methods, requirements and perspectives. *Briefings in Bioinformatics*, 5:164–178.

Vovk, U., Pernus, F., and Likar, B. (2007). A review of methods for correction of intensity inhomogeneity in MRI. *IEEE transactions on medical imaging*, 26:405–421.

Wang, D. (2003). Carbohydrate microarrays. *Proteomics*, 3:2167–2175.

Wang, H., Guo, Y., Lv, F., Zhu, H., Wu, S., Jiang, Y., Li, F., Zhou, B., Guo, W., and Zhang, T. (2010a). The essential role of GhPEL gene, encoding a pectate lyase, in cell wall loosening by depolymerization of the de-esterified pectin during fiber elongation in cotton. *Plant molecular biology*, 72:397–406.

Wang, H., Wang, Q., Pape, U., Shen, B., Huang, J., Wu, B., and Li, X. (2010b). Systematic investigation of global coordination among mrna and protein in cellular society. *BMC Genomics*, 11:364.

Wang, L.-X., Ni, J., Singh, S., and Li, H. (2004). Binding of high-mannose-type oligosaccharides and synthetic oligomannose clusters to human antibody 2G12: implications for HIV-1 vaccine design. *Chemistry & biology*, 11:127–134.

Wang, R., Liu, S., Shah, D., and Wang, D. (2005). A practical protocol for carbohydrate microarrays. *Methods in molecular biology*, 310:241–52.

Wang, Y. and Jiao, Y. (2011). Advances in plant cell type-specific genome-wide studies of gene expression. *Frontiers in Biology*, 6:384–389.

Weenink, D. (2003). Canonical correlation analysis. *In IFA proceedings, Institute of Phoentic Sciences, University of Amsterdam.*, 25:81–99.

Wendel, J., Brubaker, C., Alvarez, I., Cronn, R., and Stewart, J. M. (2009). Evolution and natural history of the cotton genus. In Paterson, A. H., editor, *Genetics and Genomics of Cotton*. Springer, NewYork, USA.

Westerhuis, J. A., Kourti, T., and MacGregor, J. F. (1998). Analysis of multiblock and hierarchical pca and pls models. *Journal of Chemometrics*, 12:301–321.

Wilks, S. S. (1932). Certain generalization in the analysis of variance . *Biometrika*, 24:471–494.

Willats, W. G., Limberg, G., Buchholt, H. C., van Alebeek, G. J., Benen, J., Christensen, T. M., Visser, J., Voragen, A., Mikkelsen, J. D., and Knox, J. P. (2000). Analysis of pectic epitopes recognised by hybridoma and phage display monoclonal antibodies using defined oligosaccharides, polysaccharides, and enzymatic degradation. *Carbohydrate research*, 327:309–20.

Williams, L. E., Lemoine, R., and Sauer, N. (2000). Sugar transporters in higher plants–a diversity of roles and complex regulation. *Trends in plant science*, 5:283–290.

Wishart, D. S. (2010). Computational approaches to metabolomics. *Methods in molecular biology*, 593:283–313.

Witten, D. M. and Tibshirani, R. J. (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical applications in genetics and molecular biology.*, 8:Article28.

Wold, H. (1966). Non-linear estimation by iterative least squares procedure. In David, F., editor, *Research papers in statistics*. Wiley, NewYork, USA.

Xiao, B., Sun, X., and Sun, R. (2001). Chemical, structural and thermal characterizations of alkali soluble lignins and hemicelluloses, and celluloses from maize stems, rye straw, and rice straw. *Polymer degradationa nd stability*, 74:307–319.

Yates, E. A., Valdor, J. F., Haslam, S. M., Morris, H. R., Dell, A., Mackie, W., and Knox, J. P. (1996). Characterization of carbohydrate structural features recognized by anti-arabinogalactan-protein monoclonal antibodies. *Glycobiology*, 6:131–139.

Yuan, J., Tiller, K., Al-Ahmad, H., Stewart, N., and Stewart, N. (2008a). Plants to power: bioenergy to fuel the future. *Trends in Plant Science*, 13:421–429.

Yuan, J. S., Galbraith, D. W., Dai, S. Y., Griffin, P., and Stewart, C. N. (2008b). Plant systems biology comes of age. *Trends in plant science*, 13:165–171.

Zanetti, M. E., Chang, I.-F., Gong, F., Galbraith, D. W., and Bailey-Serres, J. (2005). Immunopurification of polyribosomal complexes of Arabidopsis for global analysis of gene expression. *Plant physiology*, 138:624–635.

Zhang, W., Li, F., and Nie, L. (2010). Integrating multiple 'omics' analysis for microbial biology: application and methodologies. *Microbiology*, 156:287–301.

Zhong, R., Richardson, E. A., and Ye, Z.-H. (2007). The MYB46 transcription factor is a direct target of SND1 and regulates secondary wall biosynthesis in Arabidopsis. *The Plant cell*, 19:2776–2792.

Appendix A: Supplementary figures

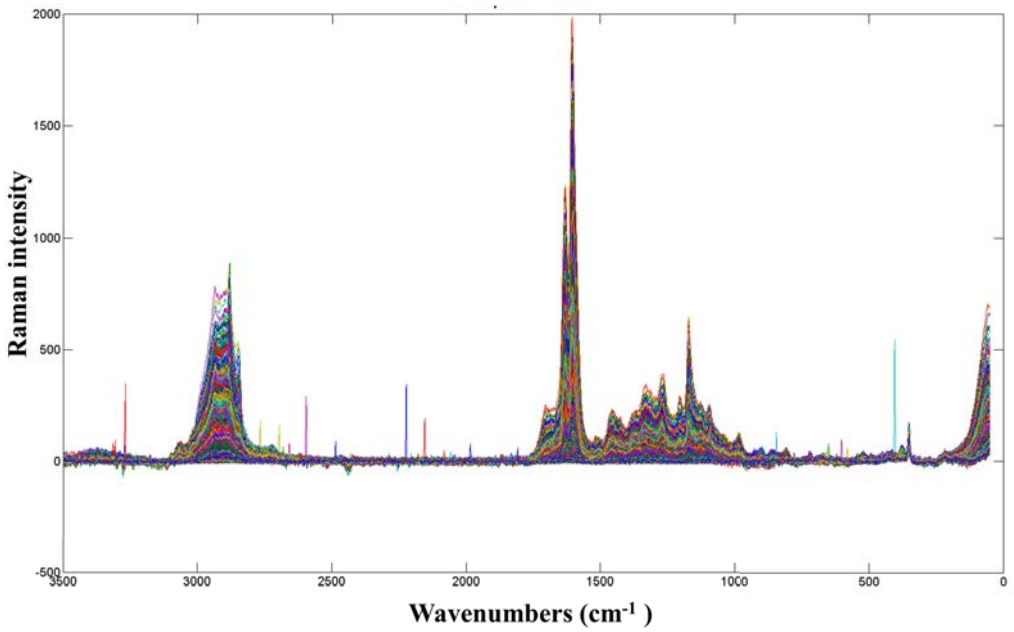


Figure S1: Raw Raman spectra corresponding to xylem cell-type. Spectral noise and spikes can be observed.

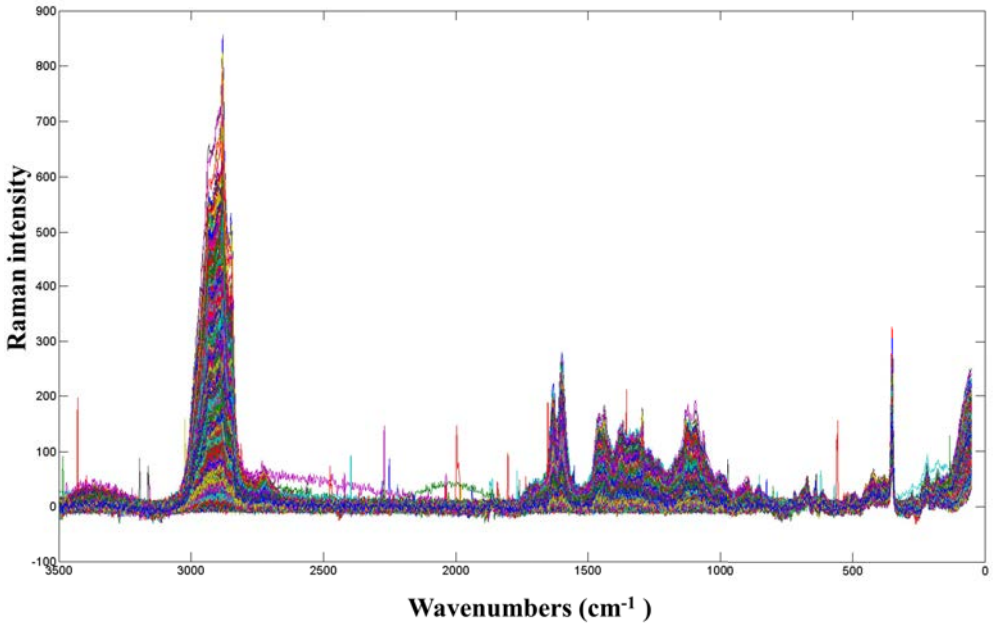


Figure S2: Raw Raman spectra corresponding to xylem+phloem cell-types. Raw spectra was acquired between 3500-0 cm^{-1}. In addition, baseline curves, spectral noise and spikes can be observed.

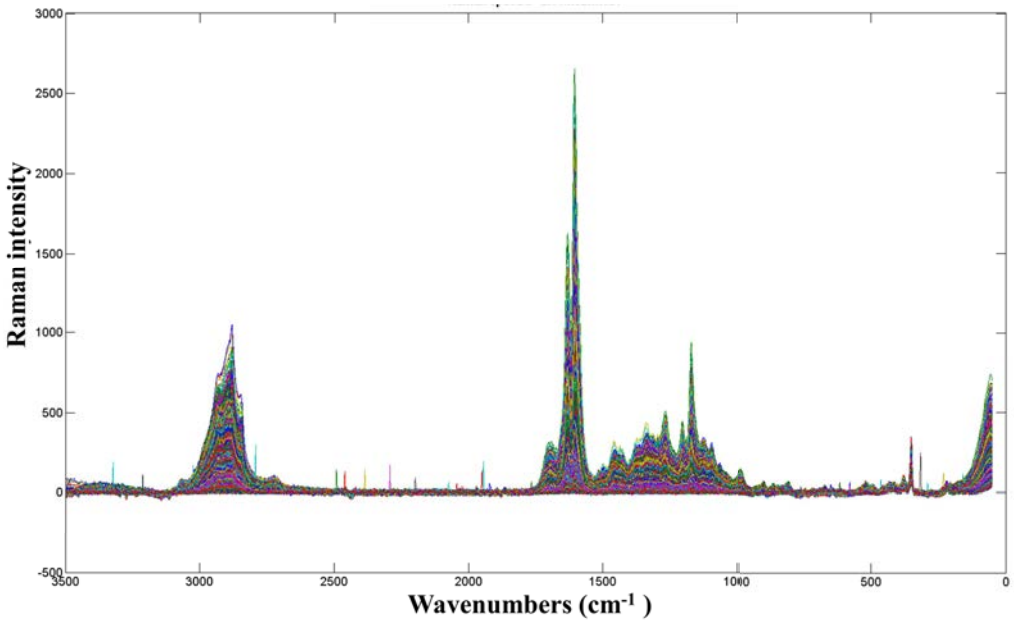


Figure S3: Raw Raman spectra corresponding to phloem cell-type. Raw spectra has the presence of spikes which need to be eliminated.

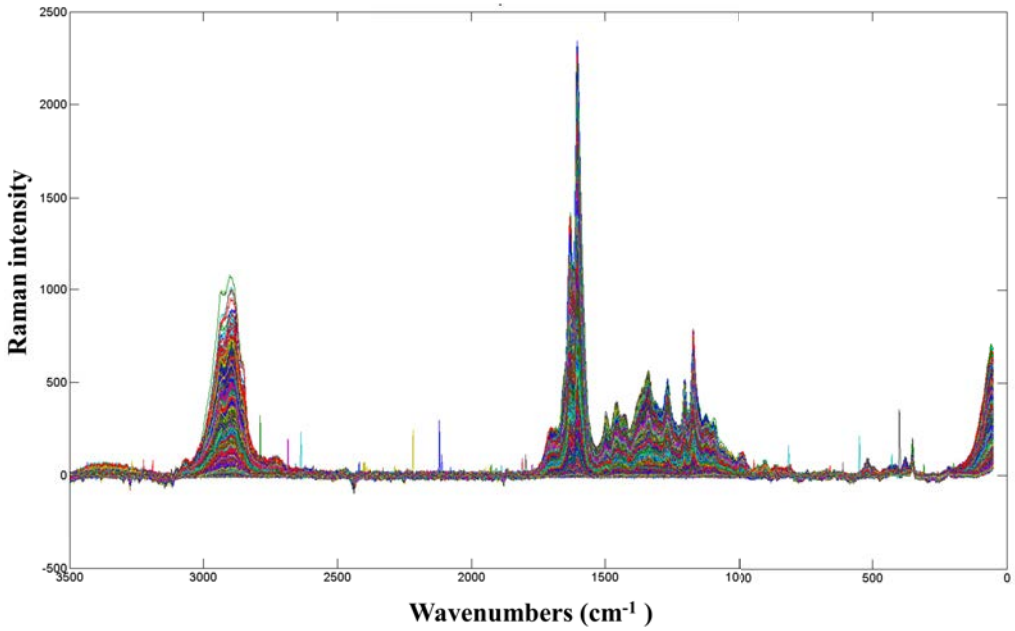


Figure S4: Raw Raman spectra corresponding to sclerenchyma+parenchyma border cell-types. Baseline curves can be observed and are slowly varying curves that are considered as linear or nonlinear addition to the spectra. In addition, spikes need to be eliminated for better resolution of the spectra.

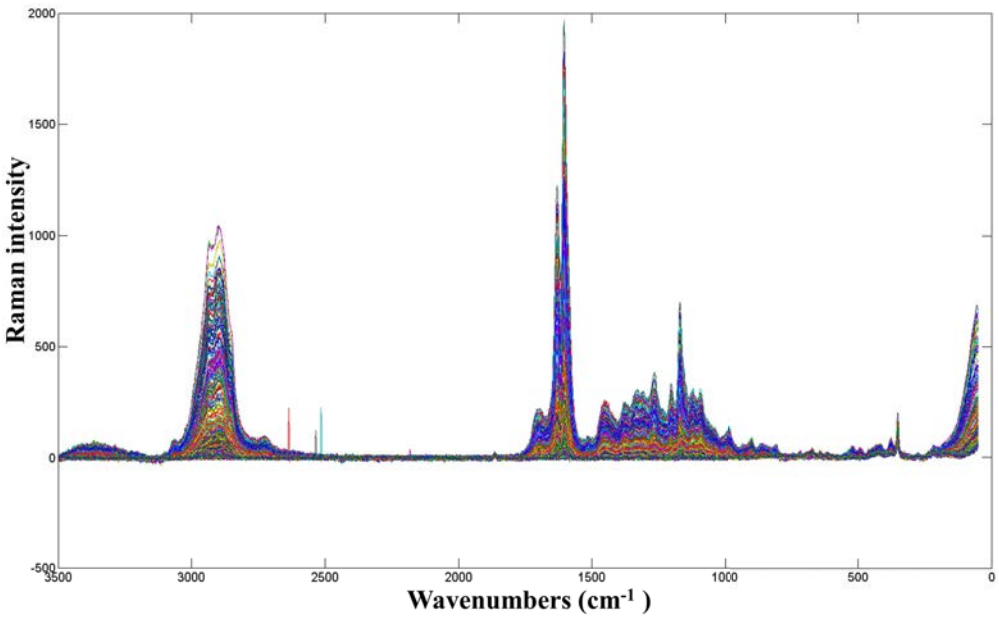


Figure S5: Raw Raman spectra corresponding to parenchyma border cell-type. Spikes can be observed in the region between 2000-2500 cm^{-1}.

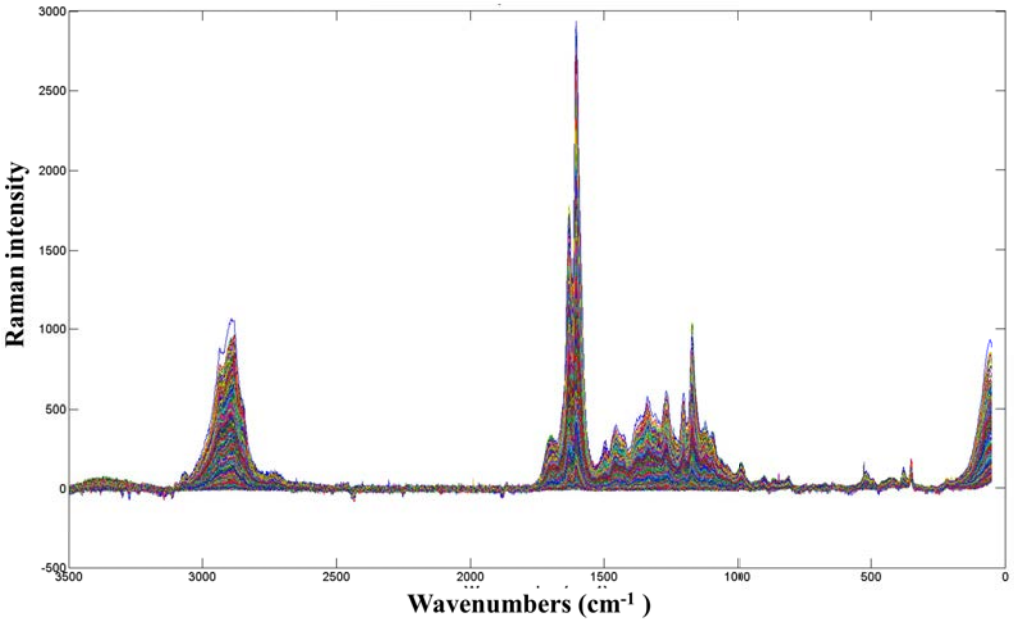


Figure S6: Raw Raman spectra corresponding to sclerenchyma cell-type. The raw spectra has the presence of spectral noise and baseline curves.

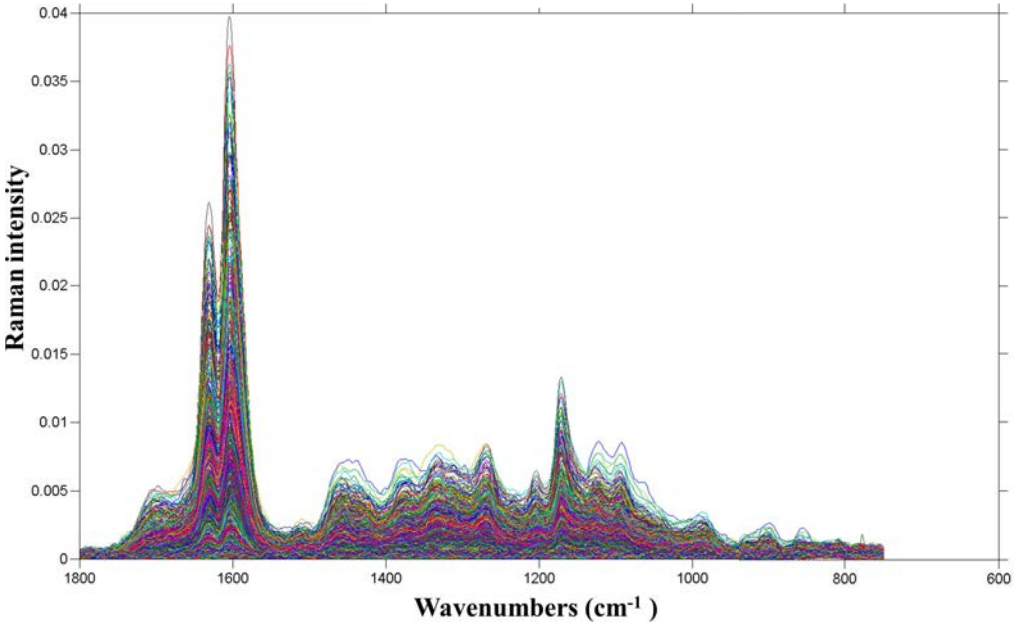


Figure S7: Pre-processed Raman spectra corresponding to xylem cell-type. Steps involved in pre-processing of the Raman spectra are detailed in Section 4.2.3.

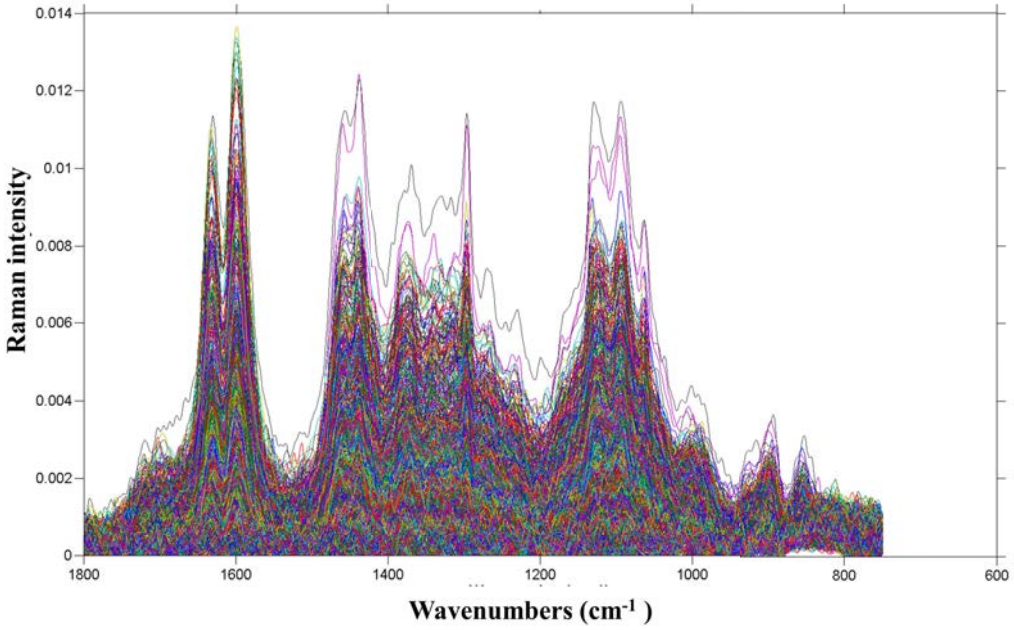


Figure S8: Pre-processed Raman spectra corresponding to xylem+phloem cell-types. Steps involved in pre-processing of the Raman spectra are detailed in Section 4.2.3.

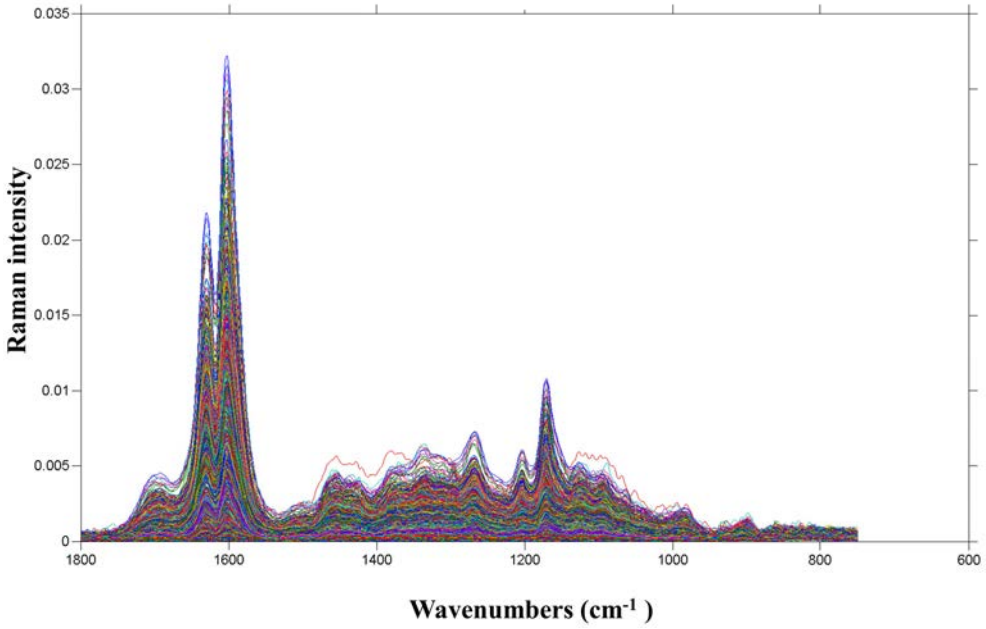


Figure S9: Pre-processed Raman spectra corresponding to phloem cell-type. Steps involved in pre-processing of the Raman spectra are detailed in Section 4.2.3.

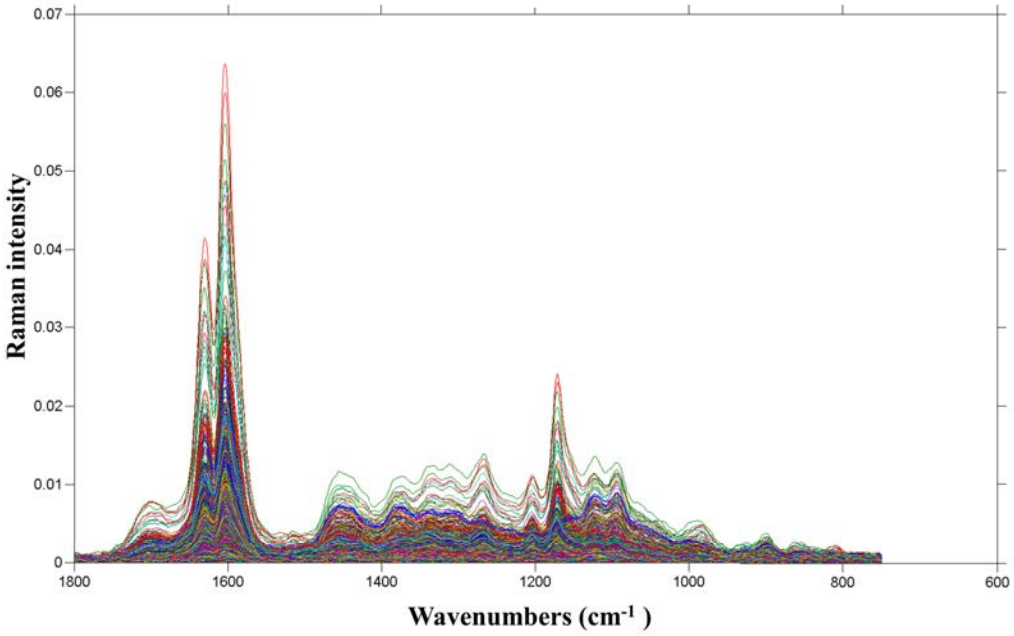


Figure S10: Pre-processed Raman spectra corresponding to sclerenchyma+parenchyma border cell-types. Steps involved in pre-processing of the Raman spectra are detailed in Section 4.2.3.

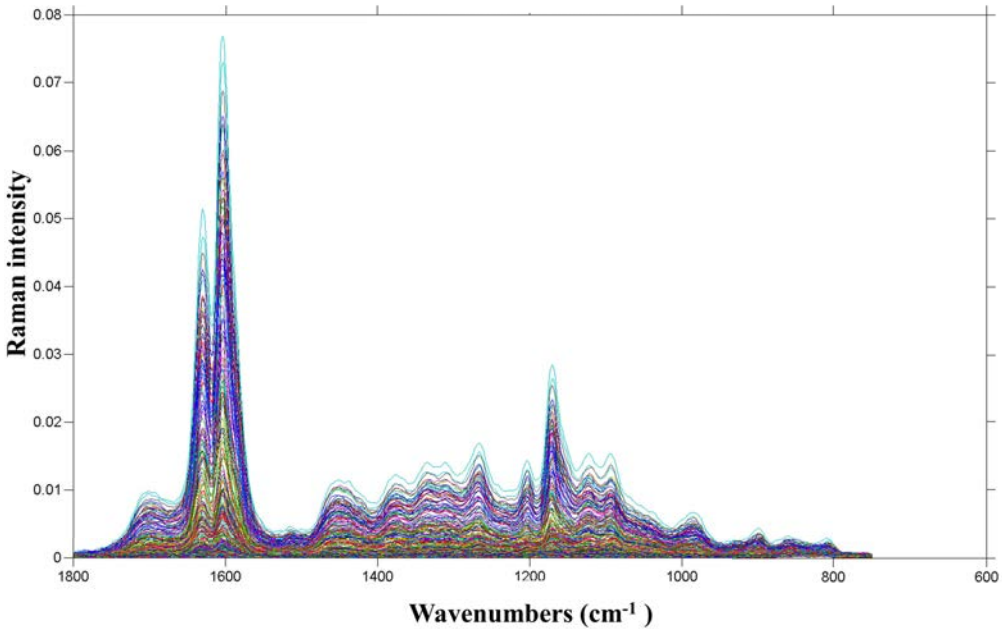


Figure S11: Pre-processed Raman spectra corresponding to parenchyma border cell-type. Steps involved in pre-processing of the Raman spectra are detailed in Section 4.2.3.

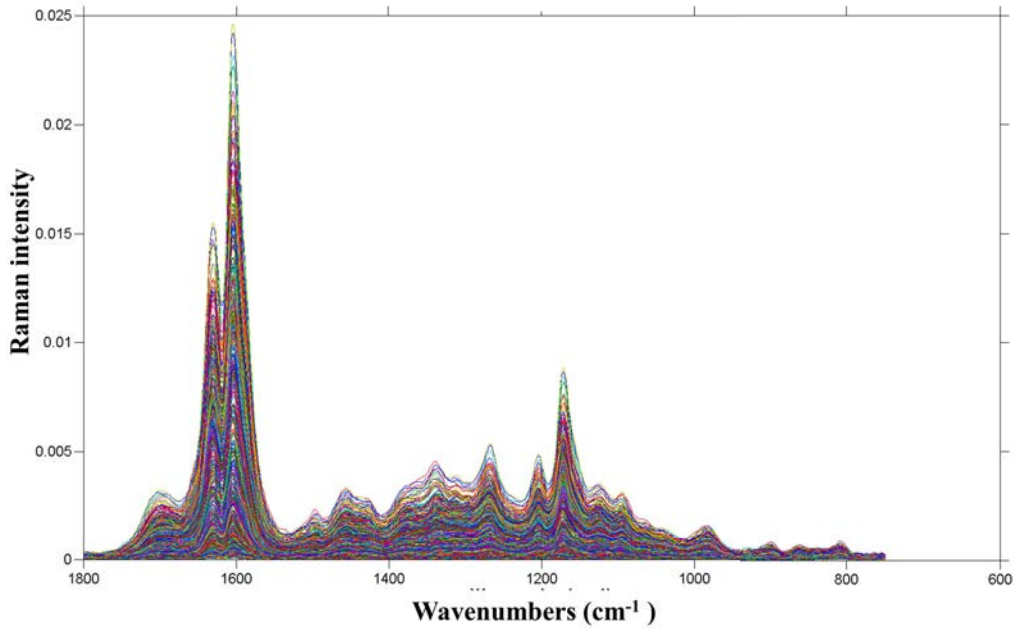


Figure S12: Pre-processed Raman spectra corresponding to sclerenchyma cell-type. Steps involved in pre-processing of the Raman spectra are detailed in Section 4.2.3.

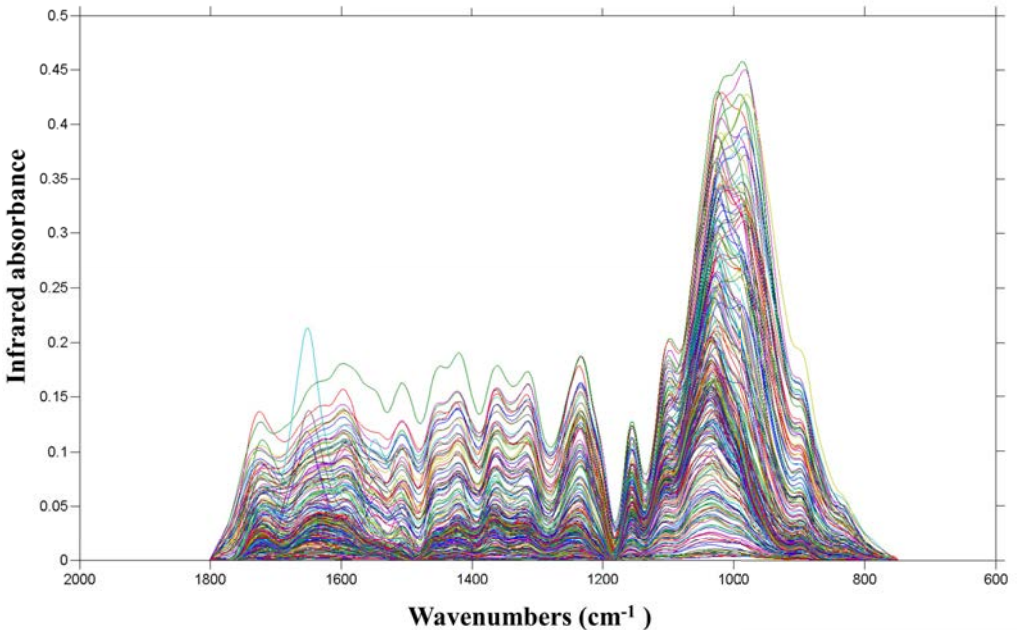


Figure S13: Pre-processed infrared spectra pertaining to different cell-types. The profiled cell-types are the same as that of the Raman experiment. Steps involved in pre-processing of the infrared spectra are detailed in Section 4.2.4.

Appendix B: Supplementary tables

Sample name	USDA PI number	Species
30834 (A 1660)	PI 629988	G. arboreum
JFW10	PI 629811	G. arboreum
Selection of SHIH	PI 529781	G. arboreum
China 10	PI 433738	G.barbadense
PIMAS 7	PI 560140	G.barbadense
Tidewater	PI 528642	G.barbadense
Krasnyj	PI 529661	G.herbaceum
Rustam 65	PI 529699	G.herbaceum
Acala Red Okra	PI 528608	G.hirsutum
Acala SJ1	PI 529540	G.hirsutum
AK DJURA HIGG BROWN	PI 529165	G.hirsutum
Brown lint clean seeds	PI 528476	G.hirsutum
BRYMER BROWN	PI 528452	G.hirsutum
FM966	PI 619097	G.hirsutum
GA161	PI 612959	G.hirsutum
Germains Acala (GC-352)	PI 601180	G.hirsutum
Germains Acala (GC-362)	PI 601142	G.hirsutum
GREEN LINT 4	PI 528787	G.hirsutum
Half and half	PI 528511	G.hirsutum
IV4F-91057	PI 566946	G.hirsutum
Lankart 57	PI 528822	G.hirsutum
Malla Guza	PI 529659	G.hirsutum
Multiple marker	PI 528950	G.hirsutum
okra leaf	PI 552560	G.hirsutum
PD93002	PI 573282	G.hirsutum
PD93003	PI573283	G.hirsutum
Shafter Brown	PI 528451	G.hirsutum
TAM 90c-19s	PI 614954	G.hirsutum
TEX 1000	PI 529888	G.hirsutum
Texas Rust Brown	PI 528453	G.hirsutum
Ting-Tao tzu ching chung mien	PI 451747	G.hirsutum
TTU 202-1107B	PI 613162	G.hirsutum

Table S1: Fiber characteristics/phenotype measurements for the 32 cotton lines used in the study. The plant introduction number (PI number) from the USDA national plant germplasm is included for each cotton line.

Raman Band intensity (cm^{-1})	Compound	Reference
896	Arabinoxylan	Himmelsbach and Akin. (1998)
903	Cellulose	Gierlinger and Schwanninger (2006)
970	Cellulose	Himmelsbach and Akin. (1998)
997	Cellulose	Himmelsbach and Akin. (1998), Gierlinger and Schwanninger (2006)
1045	Lignin	Gierlinger and Schwanninger (2006)
1093	Cellulose, arabinoxylan	Himmelsbach and Akin. (1998)
1096	Cellulose, Xylan, Glucomannan	Gierlinger and Schwanninger (2006)
1122	Cellulose, Xylan, Glucomannan	Gierlinger and Schwanninger (2006)
1131	Cellulose, arabinoxylan	Barron et al. (2006)
1143	Lignin	Gierlinger and Schwanninger (2006)
1150	Cellulose	Gierlinger and Schwanninger (2006)
1152	Cellulose, arabinoxylan	Barron et al. (2006)
1176	Ferulic acid	Himmelsbach and Akin. (1998)
1219	Ferulic acid	Himmelsbach and Akin. (1998)
1265	Ferulic acid	Himmelsbach and Akin. (1998)
1267	arabinoxylan	Barron et al. (2006)
1274	Lignin	Gierlinger and Schwanninger (2006)
1287	Ferulic acid	Himmelsbach and Akin. (1998)
1291	Cellulose	Himmelsbach and Akin. (1998)
1333	Cellulose	Gierlinger and Schwanninger (2006)
1337	Cellulose	Himmelsbach and Akin. (1998)
1376	Cellulose	Gierlinger and Schwanninger (2006)
1378	Cellulose	Himmelsbach and Akin. (1998)
1408	Cellulose	Himmelsbach and Akin. (1998)
1423	Lignin	Gierlinger and Schwanninger (2006)
1453	Ferulic acid lignin G	Himmelsbach and Akin. (1998)
1462	Arabinoxylan, Lignin and Cellulose	Gierlinger and Schwanninger (2006), Barron et al. (2006)
1465	Cellulose	Himmelsbach and Akin. (1998)

1508	Lignin	Gierlinger and Schwanninger (2006)
1600-1631	Lignin	Himmelsbach and Akin. (1998)
1599,1630	Lignin	Himmelsbach and Akin. (1998)
1641-1612	p-coumaric acid	Himmelsbach and Akin. (1998)
1599-1658	Lignin	Himmelsbach and Akin. (1998)
2897	Cellulose	Gierlinger and Schwanninger (2006)
2945	Lignin, Glucomannan, Cellulose	Gierlinger and Schwanninger (2006)

Table S2: Raman shift (cm^{-1}) and assignment of bands in the Raman spectra of cell wall polysaccharides based on the literature. This tabulated information was used to identify the peaks in the spectra acquired from different cell-types of the maize stem cross-section.

Infrared Band intensity (cm^{-1})	Compound	Reference
807	Arabinan	Kačuráková et al. (2000)
808	Arabinogalactan	Kačuráková et al. (2000)
810	Arabinogalactorhamnoglycan, Lignin	Kačuráková et al. (2000)
813	Galactoglucomannan	Kačuráková et al. (2000)
814	Glucomannan	Kačuráková et al. (2000)
834	Pectin	Kačuráková et al. (2000)
837	Arabinogalactorhamnoglycan, Lignin	Kačuráková et al. (2000)
840	Glucan, Lignin	Kačuráková et al. (2000)
842	Arabinogalactan	Kačuráková et al. (2000)
850	Starch	Kačuráková et al. (2000)
868	Arabinogalactan	Kačuráková et al. (2000)
879	Arabinogalactan (Type II)	Kačuráková et al. (2000)
880	Arabinogalactan	Kačuráková et al. (2000)
881	Arabinoglucoronoxylan+ Galactoglucomannan	Kačuráková et al. (2000)
883	Galactan	Kačuráková et al. (2000)
891	Pectin	Kačuráková et al. (2000)
892	Arabinogalactan (Type II)	Kačuráková et al. (2000)
893	Galactan	Kačuráková et al. (2000)
895	Arabinan	Kačuráková et al. (2000)

897	Xyloglucan, Arabinogalactan, Galactoglucomannan GX	Kačuráková et al. (2000)
898	Cellulose, Glucomannan, Arabinoglucoronoxylan+Galactoglucomannan	Kačuráková et al. (2000)
902	Rhamnogalacuronan	Kačuráková et al. (2000)
914	Arabinogalactorhamnoglycan	Kačuráková et al. (2000)
918	Arabinan	Kačuráková et al. (2000)
930	Cellulose	Kačuráková et al. (2000)
931	Starch	Kačuráková et al. (2000)
934	Galactoglucomannan	Kačuráková et al. (2000)
941	Glucomannan	Kačuráková et al. (2000)
944	Xyloglucan	Kačuráková et al. (2002)
945	Xyloglucan	Kačuráková et al. (2000)
951	Rhamnogalacuronan	Kačuráková et al. (2000)
953	Pectin	Kačuráková et al. (2000)
960	Galactoglucomannan	Kačuráková et al. (2000)
972	Pectin	Kačuráková et al. (2000)
985	Arabinogalactan GX	Kačuráková et al. (2000)
989	Rhamnogalacuronan	Kačuráková et al. (2000)
1004	Pectin	Kačuráková et al. (2000)
1017	Pectin	Kačuráková et al. (2000)
1022	Pectin	Kačuráková et al. (2000)
1026	Glucan, Starch	Kačuráková et al. (2000)
1030	Xylane	Kačuráková et al. (1999)
1033	Cellulose	Kačuráková et al. (2000)
1038	Galactan	Kačuráková et al. (2000)
1039	Arabinan	Kačuráková et al. (2000)
1040	Arabinogalactan (Type II)	Kačuráková et al. (2000)
1041	Xyloglucan, Glucan	Kačuráková et al. (2000)
1042	Xylane	Xiao et al. (2001)
1043	Rhamnogalacuronan, Arabinogalactan	Kačuráková et al. (2000)
1045	Arabinogalactan, Xylane	Kačuráková et al. (2000)
1047	Pectin GX	Kačuráková et al. (2000)
1049	Arabinogalactorhamnoglycan	Kačuráková et al. (2000)
1051	Pectin	Kačuráková et al. (2000)

1059	Cellulose	Kačuráková et al. (2000)
1061	Cellulose	Kačuráková et al. (2000)
1064	Glucomannan, Galactoglucomannan	Kačuráková et al. (2000)
1065	Xylane	Kačuráková et al. (1999)
1072	Galactan	Kačuráková et al. (2000)
1074	Arabinogalactan	Kačuráková et al. (2000)
1076	Glucan	Kačuráková et al. (2000)
1082	Pectin, Starch	Kačuráková et al. (2000)
1092	Glucomannan	Kačuráková et al. (2000)
1097	Arabinan	Kačuráková et al. (2000)
1100	Pectin	Kačuráková et al. (2000)
1104	Glucan	Kačuráková et al. (2000)
1109	Arabinoglucoronoxylan+ Galactoglucomannan	Kačuráková et al. (2000)
1110	Starch	Kačuráková et al. (2000)
1118	Xyloglucan	Kačuráková et al. (2000)
1120	Cellulose	Kačuráková et al. (2000)
1134	Galactan	Kačuráková et al. (2000)
1139	Arabinogalactan	Kačuráková et al. (2000)
1122	Rhamnogalacuronan	Kačuráková et al. (2000)
1141	Arabinan	Kačuráková et al. (2000)
1144	Pectin	Kačuráková et al. (2000)
1149	Galactoglucomannan	Kačuráková et al. (2000)
1150	Rhamnogalacuronan, Glucomannan	Kačuráková et al. (2000)
1152	Pectin	Kačuráková et al. (2000)
1153	Xyloglucan	Kačuráková et al. (2000)
1155	Galactan, Starch	Kačuráková et al. (2000)
1156	Arabinogalactan (Type II)	Kačuráková et al. (2000)
1161	Arabinoglucoronoxylan+ Galactoglucomannan	Kačuráková et al. (2000)
1162	Cellulose	Kačuráková et al. (2000)
1341	Lignin	Sun et al. (2000)
1378	Ferulic acid	Kačuráková et al. (1999)
1387	Lignin	Sun et al. (2000)
1420	Lignin	Sun et al. (2000)

1426	Lignin	Xiao et al. (2001)
1466	Lignin	Xiao et al. (2001)
1467	Lignin	Sun et al. (2000)
1638	Lignin	Xiao et al. (2001)
1665	Lignin	Kačuráková et al. (2002)

Table S3: Assignment of wavenumbers corresponding to the infrared band intensity. This tabulated information was used to identify the peaks in the spectra acquired from different cell-types of the maize stem cross-section.

Promoter	Targeting cell-type	Dataset	Dataset source
AGL 42	Quiescent center	Transcriptome	Nawy et al. (2005)
PET 111	Columella tier 2	Transcriptome	Nawy et al. (2005)
LRC	Lateral root cap	Transcriptome	Birnbaum et al. (2003)
GL2	Non-hair cells	Transcriptome	Birnbaum et al. (2003)
J0571	Ground endodermis +cortex+quiescent center	Transcriptome	Birnbaum et al. (2003)
S17	Phloem pole pericycyle	Transcriptome	Brady et al. (2007)
S32	Protophloem and Metaphloem	Transcriptome	Brady et al. (2007)
COBL9	Isolated hair cells	Transcriptome	Brady et al. (2007)
JO121	Xylem pole pericycle	Transcriptome	Brady et al. (2007)
S4	Isolated Protoxylem and Metaphloem	Transcriptome	Brady et al. (2007)
WOL	Stele	Transcriptome	Birnbaum et al. (2003)
SCR	Quiescent center	Transcriptome	Birnbaum et al. (2003)
SUC2	Phloem companion cells	Transcriptome	Brady et al. (2007)
J2501	Pericycle, protoxylem, metaxylem	Transcriptome	Brady et al. (2007)
RM1000	Lateral root primordia initials	Transcriptome	Brady et al. (2007)
J2661	Mature pericycle	Transcriptome	Levesque et al. (2006)
APL	Phloem sieve cells and companion cells	Transcriptome	Lee et al. (2006)
CORTEX	Cortex	Transcriptome	Lee et al. (2006)

S18	Differentiating xylem	Transcriptome	Lee et al. (2006)
35S	Root proliferating cells	Translatome	Mustroph et al. (2009)
SCR	Root endodermis and Quiescent center	Translatome	Mustroph et al. (2009)
SHR	Root vasculature	Translatome	Mustroph et al. (2009)
WOL	Root vasculature	Translatome	Mustroph et al. (2009)
GL2	Root atrichoblast epidermis	Translatome	Mustroph et al. (2009)
SULTR2	Root phloem companion cells	Translatome	Mustroph et al. (2009)
CO2	Root cortex meristamatic zone	Translatome	Mustroph et al. (2009)
PEP	Root Cortex elongation and maturation zone.	Translatome	Mustroph et al. (2009)
RPL11C	Root proliferating cells	Translatome	Mustroph et al. (2009)
SUC2	Sucrose transporter	Translatome	Mustroph et al. (2009)

Table S4: List of available cell-types and their corresponding promoters in the transcriptome and translatome dataset. Originally, there were 19 and 10 cell-types profiled across the transcriptome and translatome, respectively.

Coupled/ uncoupled	GO-term	Term description	No. of genes	p-value
C	GO:0000041	transition metal ion transport	9	3.26E-02
C	GO:0000278	mitotic cell cycle	4	1.36E-02
C	GO:0006184	GTP catabolic process	4	2.73E-02
C	GO:0006355	regulation of transcription, DNA-dependent	77	3.31E-02
C	GO:0006468	protein phosphorylation	51	1.71E-02
C	GO:0006633	fatty acid biosynthetic process	8	4.04E-02
C	GO:0006661	phosphatidylinositol biosynthetic process	8	2.35E-02
C	GO:0006807	nitrogen compound metabolic process	4	1.75E-02
C	GO:0006857	oligopeptide transport	10	8.20E-03
C	GO:0006970	response to osmotic stress	10	2.16E-02

C	GO:0006995	cellular response to nitrogen starvation	3	4.40E-02
C	GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway	10	2.51E-02
C	GO:0008610	lipid biosynthetic process	4	1.36E-02
C	GO:0009416	response to light stimulus	11	4.93E-02
C	GO:0009553	embryo sac development	5	3.15E-02
C	GO:0009611	response to wounding	20	1.92E-02
C	GO:0009617	response to bacterium	12	2.08E-02
C	GO:0009741	response to brassinosteroid stimulus	7	2.23E-02
C	GO:0009834	secondary cell wall biogenesis	5	4.00E-03
C	GO:0009862	systemic acquired resistance, salicylic acid mediated signaling pathway	17	2.35E-02
C	GO:0009886	post-embryonic morphogenesis	3	3.09E-02
C	GO:0009908	flower development	8	1.62E-02
C	GO:0009939	positive regulation of gibberellic acid mediated signaling pathway	3	3.26E-04
C	GO:0010014	meristem initiation	10	2.51E-02
C	GO:0010073	meristem maintenance	6	3.97E-02
C	GO:0010089	xylem development	9	3.99E-03
C	GO:0010150	leaf senescence	7	6.88E-03
C	GO:0010162	seed dormancy process	11	2.68E-02
C	GO:0010167	response to nitrate	15	9.78E-03
C	GO:0010260	organ senescence	4	1.75E-02
C	GO:0010440	stomatal lineage progression	5	4.87E-02
C	GO:0010583	response to cyclopentenone	12	1.30E-02
C	GO:0015706	nitrate transport	15	1.57E-02
C	GO:0016569	covalent chromatin modification	3	2.02E-02

C	GO:0042218	1-aminocyclopropane-1-carboxylate biosynthetic process	3	3.09E-02
C	GO:0042546	cell wall biogenesis	7	1.06E-02
C	GO:0043069	negative regulation of programmed cell death	13	1.28E-02
C	GO:0045454	cell redox homeostasis	12	3.69E-03
C	GO:0046855	inositol phosphate dephosphorylation	3	3.94E-02
U	GO:0006486	protein glycosylation	7	8.04E-03
U	GO:0007010	cytoskeleton organization	5	4.61E-02
U	GO:0009607	response to biotic stimulus	4	1.63E-03
U	GO:0009860	pollen tube growth	8	4.26E-02
U	GO:0009886	post-embryonic morphogenesis	3	3.41E-03
U	GO:0016579	protein deubiquitination	3	4.22E-02
U	GO:0048364	root development	7	3.89E-03
U	GO:0048366	leaf development	5	3.95E-02
U	GO:0048440	carpel development	4	3.43E-02
U	GO:0048507	meristem development	3	2.41E-02
U	GO:0048653	anther development	4	1.20E-02
U	GO:0048825	cotyledon development	5	6.47E-03
U	GO:0051301	cell division	5	1.22E-02

Table S5: Enriched GO-terms of coupled/uncoupled gene expression patters on translatome and transcriptome for the identical promoter data. Here, 'C' refers to those genes with coupled total and polysome-associated mRNA levels. 'U' refers to uncoupled genes across the transcriptome and the translatome.

Coupled/ uncoupled	GO-term	Term description	No. of genes	p-value
C	GO:0000023	maltose metabolic process	10	2.55E-02
C	GO:0000041	transition metal ion transport	9	1.53E-02
C	GO:0000956	nuclear-transcribed mRNA catabolic process	12	1.09E-04
C	GO:0006355	regulation of transcription, DNA-dependent	68	3.73E-02
C	GO:0006399	tRNA metabolic process	4	2.61E-02

C	GO:0006468	protein phosphorylation	45	2.13E-02
C	GO:0006817	phosphate ion transport	3	3.14E-02
C	GO:0006914	Autophagy	5	4.99E-02
C	GO:0007043	cell-cell junction assembly	3	2.21E-04
C	GO:0007165	signal transduction	28	1.26E-02
C	GO:0007186	G-protein coupled receptor signaling pathway	3	4.27E-03
C	GO:0007188	adenylate cyclase-modulating G-protein coupled receptor sign. Pathway	3	8.98E-05
C	GO:0007346	regulation of mitotic cell cycle	3	4.71E-02
C	GO:0007389	pattern specification process	5	3.21E-02
C	GO:0008219	cell death	5	3.44E-02
C	GO:0009416	response to light stimulus	11	2.17E-02
C	GO:0009630	Gravitropism	8	4.95E-02
C	GO:0009638	Phototropism	3	5.45E-03
C	GO:0009733	response to auxin stimulus	18	4.21E-02
C	GO:0009789	positive regulation of abscisic acid mediated signaling pathway	3	1.42E-02
C	GO:0009886	post-embryonic morphogenesis	3	2.18E-02
C	GO:0009902	chloroplast relocation	9	5.57E-03
C	GO:0010027	thylakoid membrane organization	13	1.91E-02
C	GO:0010051	xylem and phloem pattern formation	7	5.76E-03
C	GO:0010119	regulation of stomatal movement	4	1.96E-02
C	GO:0010413	glucuronoxylan metabolic process	16	4.81E-04
C	GO:0016114	terpenoid biosynthetic process	4	3.66E-02
C	GO:0016558	protein import into peroxisome matrix	7	3.55E-02
C	GO:0016569	covalent chromatin modification	3	1.42E-02
C	GO:0019252	starch biosynthetic process	14	4.14E-03

C	GO:0032957	inositol trisphosphate metabolic process	3	1.91E-02
C	GO:0034660	ncRNA metabolic process	9	1.26E-03
C	GO:0042545	cell wall modification	11	1.81E-03
C	GO:0045492	xylan biosynthetic process	16	4.81E-04
C	GO:0046777	protein autophosphorylation	9	3.70E-02
C	GO:0046855	inositol phosphate dephosphorylation	3	2.80E-02
C	GO:0048439	flower morphogenesis	6	2.27E-02
C	GO:0048519	negative regulation of biological process	5	1.34E-02
C	GO:0048527	lateral root development	7	1.15E-02
C	GO:0048574	long-day photoperiodism, flowering	3	1.20E-02
C	GO:2000067	regulation of root morphogenesis	3	8.98E-05
U	GO:0000902	cell morphogenesis	4	6.85E-03
U	GO:0006486	protein glycosylation	6	8.38E-03
U	GO:0006816	calcium ion transport	6	3.71E-03
U	GO:0006863	purine nucleobase transport	4	4.41E-02
U	GO:0006972	hyperosmotic response	4	3.14E-02
U	GO:0009734	auxin mediated signaling pathway	3	8.08E-03
U	GO:0010584	pollen exine formation	3	2.70E-02
U	GO:0015986	ATP synthesis coupled proton transport	3	2.82E-03
U	GO:0016049	cell growth	7	6.60E-04
U	GO:0030243	cellulose metabolic process	3	6.31E-03
U	GO:0044237	cellular metabolic process	3	3.83E-02
U	GO:0048868	pollen tube development	3	2.34E-02
U	GO:0080022	primary root development	4	3.32E-05

Table S6: Enriched GO-terms of coupled/uncoupled gene expression patters on translatome and transcriptome for the common promoters. Here, 'C' refers to those genes with coupled total and polysome-associated mRNA levels. 'U' refers to uncoupled genes across the transcriptome and the translatome.

AEC	GO-term	Term description	No. of genes	p-value
+	GO:0006355	regulation of transcription, DNA-dependent	12	4.47E-03
+	GO:0030968	endoplasmic reticulum unfolded protein response	3	1.52E-02
+	GO:0042546	cell wall biogenesis	3	8.20E-04
+	GO:0055085	transmembrane transport	4	4.53E-02
+	GO:0071555	cell wall organization	3	6.27E-03
-	GO:0000338	protein deneddylation	5	7.53E-03
-	GO:0000956	nuclear-transcribed mRNA catabolic process	57	1.07E-03
-	GO:0002679	respiratory burst involved in defense response	79	8.54E-07
-	GO:0006091	generation of precursor metabolites and energy	64	1.45E-21
-	GO:0006120	mitochondrial electron transport, NADH to ubiquinone	5	3.22E-02
-	GO:0006353	DNA-dependent transcription, termination	3	4.03E-02
-	GO:0006354	DNA-dependent transcription, elongation	91	2.25E-12
-	GO:0006511	ubiquitin-dependent protein catabolic process	146	3.62E-03
-	GO:0006569	tryptophan catabolic process	47	2.71E-03
-	GO:0006826	iron ion transport	62	2.08E-02
-	GO:0006863	purine nucleobase transport	87	9.65E-16
-	GO:0006873	cellular ion homeostasis	3	4.03E-02
-	GO:0006882	cellular zinc ion homeostasis	11	4.14E-02
-	GO:0006952	defense response	169	1.46E-02
-	GO:0006974	response to DNA damage stimulus	9	3.42E-02
-	GO:0007243	intracellular protein kinase cascade	10	3.82E-02

-	GO:0007275	multicellular organismal development	67	1.89E-02
-	GO:0007346	regulation of mitotic cell cycle	18	1.37E-02
-	GO:0008272	sulfate transport	10	1.97E-02
-	GO:0008643	carbohydrate transport	6	4.34E-02
-	GO:0009150	purine ribonucleotide metabolic process	3	4.03E-02
-	GO:0009414	response to water deprivation	147	1.06E-02
-	GO:0009607	response to biotic stimulus	22	3.73E-03
-	GO:0009611	response to wounding	158	9.79E-04
-	GO:0009612	response to mechanical stimulus	38	3.17E-05
-	GO:0009615	response to virus	16	4.97E-02
-	GO:0009620	response to fungus	59	3.73E-04
-	GO:0009646	response to absence of light	16	3.25E-02
-	GO:0009653	anatomical structure morphogenesis	9	6.16E-03
-	GO:0009684	indoleacetic acid biosynthetic process	59	5.07E-03
-	GO:0009693	ethylene biosynthetic process	70	3.19E-05
-	GO:0009695	jasmonic acid biosynthetic process	70	4.54E-03
-	GO:0009753	response to jasmonic acid stimulus	135	3.55E-04
-	GO:0009767	photosynthetic electron transport chain	11	2.28E-02
-	GO:0009769	photosynthesis, light harvesting in photosystem II	5	7.53E-03
-	GO:0009809	lignin biosynthetic process	25	7.65E-03
-	GO:0009853	Photorespiration	88	1.86E-04
-	GO:0009966	regulation of signal transduction	7	2.30E-02
-	GO:0010038	response to metal ion	7	2.30E-02
-	GO:0010043	response to zinc ion	30	2.59E-02
-	GO:0010106	cellular response to iron ion starvation	58	4.14E-02

-	GO:0010167	response to nitrate	104	1.75E-03
-	GO:0010200	response to chitin	232	1.39E-08
-	GO:0010254	nectary development	3	4.03E-02
-	GO:0010262	somatic embryogenesis	5	3.22E-02
-	GO:0010286	heat acclimation	43	3.67E-03
-	GO:0010289	homogalacturonan biosynthetic process	3	4.03E-02
-	GO:0010337	regulation of salicylic acid metabolic process	6	4.34E-02
-	GO:0010387	signalosome assembly	5	3.22E-02
-	GO:0010390	histone monoubiquitination	4	1.74E-02
-	GO:0010434	bract formation	3	4.03E-02
-	GO:0010438	cellular response to sulfur starvation	4	1.74E-02
-	GO:0010507	negative regulation of autophagy	3	4.03E-02
-	GO:0010599	production of lsiRNA involved in RNA interference	3	4.03E-02
-	GO:0015675	nickel cation transport	3	4.03E-02
-	GO:0015706	nitrate transport	111	7.94E-04
-	GO:0015866	ADP transport	3	4.03E-02
-	GO:0015979	Photosynthesis	140	3.08E-12
-	GO:0015986	ATP synthesis coupled proton transport	17	7.13E-03
-	GO:0015991	ATP hydrolysis coupled proton transport	16	3.25E-02
-	GO:0015992	proton transport	12	4.40E-02
-	GO:0016106	sesquiterpenoid biosynthetic process	3	4.03E-02
-	GO:0016441	posttranscriptional gene silencing	7	2.30E-02
-	GO:0018119	peptidyl-cysteine S-nitrosylation	11	1.12E-02
-	GO:0019684	photosynthesis, light reaction	80	5.03E-03
-	GO:0019761	glucosinolate biosynthetic process	88	1.73E-03

-	GO:0030010	establishment of cell polarity	3	4.03E-02
-	GO:0030029	actin filament-based process	5	7.53E-03
-	GO:0030048	actin filament-based movement	43	2.95E-02
-	GO:0030308	negative regulation of cell growth	4	1.74E-02
-	GO:0031930	mitochondria-nucleus signaling pathway	4	1.74E-02
-	GO:0034605	cellular response to heat	9	3.42E-02
-	GO:0034755	iron ion transmembrane transport	4	1.74E-02
-	GO:0035556	intracellular signal transduction	102	1.44E-03
-	GO:0042218	1-aminocyclopropane-1-carboxylate biosynthetic process	15	5.33E-03
-	GO:0042325	regulation of phosphorylation	3	4.03E-02
-	GO:0042742	defense response to bacterium	159	2.63E-02
-	GO:0042773	ATP synthesis coupled electron transport	8	6.06E-04
-	GO:0043069	negative regulation of programmed cell death	84	2.31E-02
-	GO:0043255	regulation of carbohydrate biosynthetic process	3	4.03E-02
-	GO:0045333	cellular respiration	20	2.99E-04
-	GO:0046168	glycerol-3-phosphate catabolic process	3	4.03E-02
-	GO:0046786	viral replication complex formation and maintenance	4	1.74E-02
-	GO:0048507	meristem development	24	3.62E-02
-	GO:0048510	reg. of timing of trans. from veg. to reproductive phase	14	1.72E-02
-	GO:0048523	negative regulation of cellular process	6	1.57E-02
-	GO:0048830	adventitious root development	4	1.74E-02

-	GO:0048869	cellular developmental process	7	1.40E-03
-	GO:0051026	chiasma assembly	8	2.92E-02
-	GO:0051090	regulation of sequence-specific DNA binding TF activity	3	4.03E-02
-	GO:0051127	positive regulation of actin nucleation	3	4.03E-02
-	GO:0051603	proteolysis involved in cellular protein catabolic process	17	1.75E-03
-	GO:0051788	response to misfolded protein	95	7.25E-03
-	GO:0051865	protein autoubiquitination	9	1.71E-03
-	GO:0052542	defense response by callose deposition	26	4.90E-02
-	GO:0055062	phosphate ion homeostasis	7	2.30E-02
-	GO:0070301	cellular response to hydrogen peroxide	3	4.03E-02
-	GO:0070734	histone H3-K27 methylation	3	4.03E-02
-	GO:0071456	cellular response to hypoxia	14	4.76E-02
-	GO:0071722	detoxification of arsenic-containing substance	3	4.03E-02
-	GO:0072334	UDP-galactose transmembrane transport	3	4.03E-02
-	GO:0080003	thalianol metabolic process	3	4.03E-02
-	GO:0080027	response to herbivore	7	1.40E-03
-	GO:0080086	stamen filament development	7	7.59E-03
-	GO:0080129	proteasome core complex assembly	72	8.44E-04

Table S7: GSEA of GO-BP terms for AEC genes for identical promoter data. A '+' indicates a high degree of conservation of co-expression relationships derived by Z-scores, while a '-' indicates divergences of co-expression relationships on the translatome and transcriptome.

AEC	GO-term	Term description	No. of genes	p-value
+	GO:0000271	polysaccharide biosynthetic process	3	3.26E-04
+	GO:0007165	signal transduction	4	8.68E-03

+	GO:0007389	pattern specification process	3	8.18E-05
+	GO:0008361	regulation of cell size	3	4.52E-05
+	GO:0009825	multidimensional cell growth	3	2.47E-04
+	GO:0009926	auxin polar transport	3	1.60E-04
+	GO:0009932	cell tip growth	3	2.20E-04
+	GO:0010015	root morphogenesis	3	3.68E-05
+	GO:0010075	regulation of meristem growth	4	9.70E-05
+	GO:0010817	regulation of hormone levels	3	1.40E-04
+	GO:0040007	Growth	3	7.75E-05
+	GO:0043481	anthocyanin accumulation in tissues in response to UV light	3	6.75E-04
+	GO:0048767	root hair elongation	3	2.45E-03
+	GO:0071555	cell wall organization	3	1.04E-03
-	GO:0000041	transition metal ion transport	78	2.12E-03
-	GO:0000059	protein import into nucleus, docking	11	1.51E-02
-	GO:0000302	response to reactive oxygen species	21	2.78E-02
-	GO:0000338	protein deneddylation	5	2.45E-02
-	GO:0000461	endonucleolytic cleavage	4	4.48E-02
-	GO:0002679	respiratory burst involved in defense response	90	7.35E-06
-	GO:0006071	glycerol metabolic process	12	2.52E-02
-	GO:0006091	generation of precursor metabolites and energy	65	1.30E-16
-	GO:0006352	DNA-dependent transcription, initiation	25	2.33E-02
-	GO:0006354	DNA-dependent transcription, elongation	94	3.06E-07
-	GO:0006355	regulation of transcription, DNA-dependent	893	5.33E-03
-	GO:0006511	ubiquitin-dependent protein catabolic process	179	3.31E-03
-	GO:0006569	tryptophan catabolic process	54	1.03E-02
-	GO:0006612	protein targeting to membrane	217	3.24E-02
-	GO:0006629	lipid metabolic process	122	2.97E-02

-	GO:0006636	unsaturated fatty acid biosynthetic process	46	1.55E-02
-	GO:0006680	glucosylceramide catabolic process	4	4.48E-02
-	GO:0006814	sodium ion transport	26	1.70E-02
-	GO:0006826	iron ion transport	83	2.61E-04
-	GO:0006857	oligopeptide transport	76	2.80E-04
-	GO:0006863	purine nucleobase transport	96	8.25E-15
-	GO:0006886	intracellular protein transport	109	3.21E-02
-	GO:0006952	defense response	219	4.33E-04
-	GO:0006984	ER-nucleus signaling pathway	11	3.83E-02
-	GO:0007243	intracellular protein kinase cascade	13	5.80E-03
-	GO:0008272	sulfate transport	11	3.83E-02
-	GO:0009228	thiamine biosynthetic process	7	3.46E-02
-	GO:0009395	phospholipid catabolic process	5	2.45E-02
-	GO:0009414	response to water deprivation	177	3.24E-02
-	GO:0009607	response to biotic stimulus	24	1.85E-02
-	GO:0009611	response to wounding	192	1.32E-03
-	GO:0009612	response to mechanical stimulus	44	2.13E-05
-	GO:0009620	response to fungus	69	7.64E-04
-	GO:0009637	response to blue light	64	4.24E-03
-	GO:0009646	response to absence of light	23	1.50E-04
-	GO:0009649	entrainment of circadian clock	5	2.45E-02
-	GO:0009653	anatomical structure morphogenesis	9	3.80E-02
-	GO:0009684	indoleacetic acid biosynthetic process	68	2.48E-02
-	GO:0009693	ethylene biosynthetic process	80	2.37E-04
-	GO:0009694	jasmonic acid metabolic process	25	2.33E-02
-	GO:0009695	jasmonic acid biosynthetic process	92	3.40E-05
-	GO:0009734	auxin mediated signaling pathway	26	1.70E-02

-	GO:0009751	response to salicylic acid stimulus	81	1.93E-02
-	GO:0009753	response to jasmonic acid stimulus	164	2.92E-04
-	GO:0009767	photosynthetic electron transport chain	13	1.64E-02
-	GO:0009768	photosynthesis, light harvesting in photosystem I	4	4.48E-02
-	GO:0009769	photosynthesis, light harvesting in photosystem II	5	2.45E-02
-	GO:0009773	photosynthetic electron transport in photosystem I	36	1.73E-02
-	GO:0009813	flavonoid biosynthetic process	40	1.41E-02
-	GO:0009867	jasmonic acid mediated signaling pathway	170	4.44E-03
-	GO:0009939	positive regulation of gibberellic acid mediated signaling pathway	5	2.45E-02
-	GO:0010043	response to zinc ion	37	1.31E-02
-	GO:0010093	specification of floral organ identity	24	1.01E-02
-	GO:0010103	stomatal complex morphogenesis	89	1.30E-02
-	GO:0010106	cellular response to iron ion starvation	78	9.66E-04
-	GO:0010107	potassium ion import	4	4.48E-02
-	GO:0010114	response to red light	59	2.56E-02
-	GO:0010167	response to nitrate	130	2.35E-04
-	GO:0010200	response to chitin	286	9.33E-11
-	GO:0010207	photosystem II assembly	111	2.51E-03
-	GO:0010231	maintenance of seed dormancy	4	4.48E-02
-	GO:0010286	heat acclimation	52	2.03E-03
-	GO:0010363	regulation of plant-type hypersensitive response	220	2.36E-02
-	GO:0010390	histone monoubiquitination	4	4.48E-02

-	GO:0010438	cellular response to sulfur starvation	4	4.48E-02
-	GO:0015706	nitrate transport	137	1.97E-04
-	GO:0015979	Photosynthesis	164	6.86E-13
-	GO:0018119	peptidyl-cysteine S-nitrosylation	12	2.52E-02
-	GO:0019684	photosynthesis, light reaction	105	4.07E-05
-	GO:0019761	glucosinolate biosynthetic process	101	1.96E-02
-	GO:0030308	negative regulation of cell growth	4	4.48E-02
-	GO:0031540	regulation of anthocyanin biosynthetic process	12	2.52E-02
-	GO:0031930	mitochondria-nucleus signaling pathway	4	4.48E-02
-	GO:0034755	iron ion transmembrane transport	4	4.48E-02
-	GO:0035556	intracellular signal transduction	125	6.82E-04
-	GO:0035725	sodium ion transmembrane transport	31	1.75E-02
-	GO:0042218	1-aminocyclopropane-1-carboxylate biosynthetic process	17	7.30E-03
-	GO:0042549	photosystem II stabilization	7	7.33E-03
-	GO:0042773	ATP synthesis coupled electron transport	7	3.46E-02
-	GO:0043161	proteasomal ubiquitin-dependent protein catabolic process	73	3.69E-02
-	GO:0043407	negative regulation of MAP kinase activity	12	9.39E-03
-	GO:0045333	cellular respiration	20	1.10E-02
-	GO:0045727	positive regulation of translation	13	1.46E-03

-	GO:0046786	viral replication complex formation and maintenance	4	4.48E-02
-	GO:0048235	pollen sperm cell differentiation	20	2.15E-02
-	GO:0048869	cellular developmental process	7	7.33E-03
-	GO:0050832	defense response to fungus	177	2.77E-02
-	GO:0050994	regulation of lipid catabolic process	4	4.48E-02
-	GO:0051603	proteolysis involved in cellular protein catabolic process	18	1.12E-02
-	GO:0051865	protein autoubiquitination	9	1.23E-02
-	GO:0052542	defense response by callose deposition	32	3.26E-02
-	GO:0055062	phosphate ion homeostasis	8	2.07E-02
-	GO:0055114	oxidation-reduction process	635	4.09E-02
-	GO:0070838	divalent metal ion transport	53	4.67E-02
-	GO:0071216	cellular response to biotic stimulus	11	3.83E-02
-	GO:0071577	zinc ion transmembrane transport	10	2.41E-02
-	GO:0071786	endoplasmic reticulum tubular network organization	5	2.45E-02
-	GO:0080027	response to herbivore	7	7.33E-03
-	GO:0080086	stamen filament development	7	3.46E-02
-	GO:0080129	proteasome core complex assembly	80	2.31E-02
-	GO:0090333	regulation of stomatal closure	7	3.46E-02
-	GO:2000037	regulation of stomatal complex patterning	5	2.45E-02
-	GO:2000038	regulation of stomatal complex development	5	2.45E-02
-	GO:2001141	regulation of RNA biosynthetic process	5	2.45E-02

Table S8: GSEA of GO-BP terms for AEC genes for the common promoters. A '+' indicates a high degree of conservation of co-expression relationships derived by Z-scores, while a '-' indicates divergences of co-expression relationships on the translatome and transcriptome.

System-level	Number of genes	Observed motif	p-value
Transcriptome	214	1110	0
Transcriptome	58	11111	0
Transcriptome	47	1000	0
Transcriptome	33	11101	0
Transcriptome	34	111000	0
Transcriptome	31	111110	0
Transcriptome	22	110110	0
Transcriptome	19	100100	0.04
Transcriptome	26	100000	0.2
Transcriptome	25	100101	0.02
Transcriptome	23	10011	0
Transcriptome	25	1010	0.99
Transcriptome	67	0	1
Transcriptome	25	1100	1
Transcriptome	16	11110	1
Translatome	203	100101	0
Translatome	94	1110	0
Translatome	72	100100	0
Translatome	46	10011	0
Translatome	28	101011	0
Translatome	29	110110	0.01
Translatome	45	101110	0.11
Translatome	37	110	0.11
Translatome	8	111011	0.92
Translatome	57	100110	0.95
Translatome	23	101111	0.99
Translatome	12	100	0.99

Translatome	10	111000	0.99
Translatome	21	111111	1

Table S9: All possible motif occurences across the identical promoters of transcriptome and translatome. The observed motifs are depicted in the order GL2-SCR, GL2-WOL, SUC2-GL2, SUC2-SCR, SUC2-WOL, WOL-SCR of pairwise cell-type comparison. A characteristic pattern of the pairwise differences are represented by 0 (no significant mean difference of expression values) and 1 (significant mean difference of expression values).

Motif	GO-term	Term description	No. of genes	p-value
1	GO:0001666	response to hypoxia	4	1.03E-02
1	GO:0005986	sucrose biosynthetic process	3	1.31E-03
1	GO:0006084	acetyl-CoA metabolic process	5	2.44E-03
1	GO:0006085	acetyl-CoA biosynthetic process	3	2.02E-04
1	GO:0006612	protein targeting to membrane	9	4.85E-02
1	GO:0006810	Transport	9	4.45E-02
1	GO:0006833	water transport	5	3.38E-02
1	GO:0008033	tRNA processing	3	4.79E-03
1	GO:0008219	cell death	3	2.59E-02
1	GO:0009684	indoleacetic acid biosynthetic process	4	3.74E-02
1	GO:0009691	cytokinin biosynthetic process	3	5.25E-03
1	GO:0009735	response to cytokinin stimulus	3	2.72E-02
1	GO:0009736	cytokinin mediated signaling pathway	3	4.74E-02
1	GO:0009741	response to brassinosteroid stimulus	4	1.19E-02
1	GO:0009744	response to sucrose stimulus	9	1.29E-03
1	GO:0009749	response to glucose stimulus	6	3.83E-04
1	GO:0009750	response to fructose stimulus	6	9.06E-03
1	GO:0009910	negative regulation of flower development	3	1.37E-02
1	GO:0009963	positive regulation of flavonoid biosynthetic process	6	1.79E-03
1	GO:0010075	regulation of meristem growth	5	3.93E-02

1	GO:0010264	myo-inositol hexakisphosphate biosynthetic process	3	3.37E-02
1	GO:0016036	cellular response to phosphate starvation	5	1.87E-02
1	GO:0016126	sterol biosynthetic process	7	4.71E-03
1	GO:0016132	brassinosteroid biosynthetic process	4	4.57E-02
1	GO:0019375	galactolipid biosynthetic process	5	7.54E-03
1	GO:0019745	pentacyclic triterpenoid biosynthetic process	3	1.06E-02
1	GO:0045454	cell redox homeostasis	9	3.59E-05
1	GO:0048653	anther development	7	6.87E-06
1	GO:0051645	Golgi localization	3	1.21E-02
1	GO:0051646	mitochondrion localization	3	1.21E-02
1	GO:0055085	transmembrane transport	15	8.51E-04
1	GO:0060151	peroxisome localization	3	1.21E-02
2	GO:0000041	transition metal ion transport	11	5.21E-08
2	GO:0006084	acetyl-CoA metabolic process	6	1.51E-04
2	GO:0006633	fatty acid biosynthetic process	5	3.98E-03
2	GO:0006826	iron ion transport	6	1.50E-03
2	GO:0006865	amino acid transport	4	4.78E-02
2	GO:0006979	response to oxidative stress	12	1.50E-05
2	GO:0007043	cell-cell junction assembly	5	1.00E-10
2	GO:0007169	transmembrane receptor protein tyrosine kinase sign. pathw.	7	4.11E-04
2	GO:0007389	pattern specification process	3	1.57E-02
2	GO:0008152	metabolic process	16	2.67E-02
2	GO:0008361	regulation of cell size	3	9.11E-03
2	GO:0009269	response to desiccation	3	4.55E-03
2	GO:0009611	response to wounding	10	1.92E-03
2	GO:0009664	plant-type cell wall organization	4	4.05E-02
2	GO:0009739	response to gibberellin stimulus	5	3.06E-03
2	GO:0009741	response to brassinosteroid stimulus	5	1.01E-03
2	GO:0009750	response to fructose stimulus	5	1.62E-02
2	GO:0009805	coumarin biosynthetic process	4	1.84E-03

2	GO:0009825	multidimensional cell growth	3	4.15E-02
2	GO:0009926	auxin polar transport	3	2.85E-02
2	GO:0009932	cell tip growth	4	7.52E-03
2	GO:0009963	positive regulation of flavonoid biosynthetic process	6	7.36E-04
2	GO:0010015	root morphogenesis	3	7.54E-03
2	GO:0010075	regulation of meristem growth	7	1.20E-03
2	GO:0010106	cellular response to iron ion starvation	6	1.20E-03
2	GO:0010167	response to nitrate	10	5.37E-05
2	GO:0010413	glucuronoxylan metabolic process	8	7.10E-04
2	GO:0010817	regulation of hormone levels	3	2.54E-02
2	GO:0015706	nitrate transport	10	8.54E-05
2	GO:0016126	sterol biosynthetic process	9	7.11E-05
2	GO:0016132	brassinosteroid biosynthetic process	10	4.23E-07
2	GO:0030003	cellular cation homeostasis	4	1.96E-02
2	GO:0040007	Growth	3	1.49E-02
2	GO:0042545	cell wall modification	8	3.30E-05
2	GO:0045492	xylan biosynthetic process	8	7.10E-04
2	GO:0048765	root hair cell differentiation	5	1.62E-02
2	GO:0048767	root hair elongation	6	1.13E-02
2	GO:0055114	oxidation-reduction process	22	7.23E-03
2	GO:0070838	divalent metal ion transport	4	9.36E-03

Table S10: Enriched GO-terms of genes with characteristic cell-type specific gene expression patters found for motif 1 and motif 2. More details on motifs 1 and 2 can be found in Figure 5.14.

Appendix C: Supplementary text

Transcriptome dataset- Promoter sequences
Promoter SUC2

>gi|1019750|emb|X79702.1| A.thaliana AtSUC2 gene
GGATCCCCTAAAATCTGGTTTCATATTAATTTCACACACCAAGTTACTTTCTATTATTA
ACTGTTATAATGGACCATGAAATCATTTGCATATGAACTGCAATGATACATAATCCACT
TTGTTTTGTGGGAGACATTTACCAGATTTCGGTAAATTGGTATTCCCCCTTTTATGTGA
TTGGTCATTGATCATTGTTAGTGGCCAGACATTTGAACTCCCGTTTTTTTGTCTATAAG
AATTCGGAAACATATAGTATCCTTTGAAAACGGAGAAACAAATAACAATGTGGACAAAC
TAGATATAATTTCAACACAAGACTATGGGAATGATTTTACCCACTAATTATAATCCGAT
CACAAGGTTTCAACGAACTAGTTTTCCAGATATCAACCAAATTTACTTTGGAATTAAAC
TAACTTAAAACTAATTGGTTGTTCGTAAATGGTGCTTTTTTTTTTTGCGGATGTTAGTA
AAGGGTTTTATGTATTTTATATTATTAGTTATCTGTTTTCAGTGTTATGTTGTCTCATC
CATAAAGTTTATATGTTTTTTCTTTGCTCTATAACTTATATATATATATGAGTTTACAG
TTATATTTATACATTTCAGATACTGATCGGCATTTTTTTTGGTAAAAAATATATGCATG
AAAAACTCAAGTGTTTCTTTTTTAAGGAATTTTTAAATGGTGATTATATGAATATAATC
ATATGTATATCCGTATATATATGTAGCCAGATAGTTAATTATTTGGGGGATATTTGAAT
TATTAATGTTATAATATTCTTTCTTTTGACTCGTCTGGTTAAATTAAAGAACAAAAAAA
ACACATACTTTTACTGTTTTAAAAGGTTAAATTAACATAATTTATTGATTACAAGTGTC
AAGTCCATGACATTGCATGTAGGTTCGAGACTTCAGAGATAACGGAAGAGATCGATAAT
TGTGATCGTAACATCCAGATATGTATGTTTAATTTTCATTTAGATGTGGATCAGAGAAG
ATAAGTCAAACTGTCTTCATAATTTAAGACAACCTCTTTTAATATTTTCCCAAAACATG
TTTTATGTAACTACTTTGCTTATGTGATTGCCTGAGGATACTATTATTCTCTGTCTTTA
TTCTCTTCACACCACATTTAAATAGTTTAAGAGCATAGAAATTAATTATTTTCAAAAAG
GTGATTATATGCATGCAAAATAGCACACCATTTATGTTTATATTTTCAAATTATTTAAT
ACATTTCAATATTTCATAAGTGTGATTTTTTTTTTTTTGTCAATTTCATAAGTGTGATT
TGTCATTTGTATTAAACAATTGTATCGCGCAGTACAAATAAACAGTGGGAGAGGTGAAA
ATGCAGTTATAAAACTGTCCAATAATTACTAACACATTTAAATWATCTAAAAAGAGTGT

TTCAAAAAAAATTCTTTTGAAATAAGAAAAGTGATAGATATTTTTACGCTTTCGTCTGA
AAATAAAACAATAATAGTTTATTAGAAAAATGTTATCACCGAAAATTATTCTAGTGCCA
CTCGCTCGGATCGAAATTCGAAAGTTATATTCTTTCTCTTTACCTAATATAAAAATCAC
AAGAAAAATCAATCCGAATATATCTATCAACATAGTATATGCCCTTACATATTGTTTCT
GACTTTTCTCTATCCGAATTTCTCGCTTCATGGTTTTTTTTTAACATATTCTCATTTAA
TTTTCATTACTATTATATAACTAAAAGATGGAAATAAAATAAAGTGTCTTTGAGAATCG
AAACGTCCATATCAGTAAGATAGTTTGTGTGAAGGTAAAATCTAAAAGATTTAAGTTCC
AAAAACAGAAAATAATATATTACGCTAAAAAAGAAGAAAATAATTAAATACAAAACAGA
AAAAAATAATATACGACAGACACGTGTCACGAAGATACCCTACGCTATAGACACAGCTC
TGTTTTCTCTTTTCTATGCCTCAAGGCTCTCTTAACTTCACTGTCTCCTCTTCGGATAA
TCCTATCCTTCTCTTCCTATAAATACCTCTCCACTCTTCCTCTTCCTCCACCACTACAA
CCACCGCAACAACCACCAAAAACCCTCTCAAAGAAATCTTTTTTTTCTTACTTTCTTGG
TTTGTCAAATATG

Associated genomic coordinate: AT1G22710

Promoter WOL

>gi|11595486|emb|AJ278528.1| Arabidopsis thaliana
AGTTGGAGCAAAGTTGCTTCTTTTGAGAACCATGCGTTTCTTTCTCTCTTTTGTTCTTG
AATTCGCAAAAACATGTCCTTTTTCGTCTACAGGTTTCTAGGGTTTGTTTCTGTACTAT
AAACTATGTTTATGCTCAGATATGAACTGGGCACTCAACAATCATCAAGAAGAAGAAGA
AGAGCCACGAAGAATTGAAATTTCTGATTCCGAGTCACTAGAAAACTTGAAAAGCAGCG
ATTTTTATCAACTGGGTGGTGGTGGTGCTCTGAATTCGTCAGAAAAGCCGAGAAAGATC
GATTTTTGGCGTTCGGGGTTGATGGGTTTTGCGAAGATGCAGCAGCAGCAACAGCTTCA
GCATTCAGTGGCGGTGAAGATGAACAATAATAATAATAACGATCTAATGGGTAATAAAA
AAGGGTCAACTTTCATACAAGAACATCGAGCATTGTTACCAAAAGCTTTGATTCTGTGG
ATCATCATTGTTGGGTTTATAAGCAGTGGGATTTATCAGTGGATGGATGATGCTAATAA
GATTAGAAGGGAAGAGGTTTTGGTCAGCATGTGTGATCAAAGAGCTAGAATGTTGCAGG
ATCAATTTAGTGTTAGTGTTAATCATGTTCATGCTTTGGCTATTCTCGTCTCCACTTTT
CATTACCACAAGAACCCTTCTGCAATTGATCAGGAGACATTTGCGGAGTACACGGCAAG
AACAGCATTTGAGAGACCGTTGCTAAGTGGAGTGGCTTATGCTGAAAAAGTTGTGAATT
TTGAGAGGGAGATGTTTGAGCGGCAGCACAATTGGGTTATAAAGACAATGGATAGAGGA
GAGCCTTCACCGGTTAGGGATGAGTATGCTCCTGTTATATTCTCTCAAGATAGTGTCTC
TTACCTTGAGTCACTCGATATGATGTCAGGCGAGGAGGATCGTGAGAATATTTTGCGAG
CTAGAGAAACCGGAAAAGCTGTCTTGACTAGCCCTTTTAGGTTGTTGGAAACTCACCAT
CTCGGAGTTGTGTTGACATTCCCTGTCTACAAGTCTTCTCTTCCTGAAAATCCGACTGT

CGAAGAGCGTATTGCAGCCACTGCAGGGTACCTTGGTGGTGCGTTTGATGTGGAGTCTC
TAGTCGAGAATTTACTTGGTCAGCTTGCTGGTAACCAAGCAATAGTTGTGCATGTGTAT
GATATCACCAATGCATCAGATCCACTTGTCATGTATGGTAATCAAGATGAAGAAGCCGA
CAGATCTCTCTCTCATGAGAGCAAGCTCGATTTTGGAGACCCCTTCAGGAAACATAAGA
TGATATGCAGGTACCACCAAAAGGCACCAATACCGTTGAATGTGCTCACAACTGTGCCA
TTGTTCTTTGCGATTGGTTTCTTGGTGGGTTATATACTGTATGGTGCAGCTATGCACAT
AGTAAAAGTCGAAGATGATTTCCATGAAATGCAAGAGCTTAAAGTTCGAGCAGAAGCTG
CTGATGTCGCTAAATCGCAGTTTCTTGCTACCGTGTCTCACGAGATCAGGACACCAATG
AATGGCATTCTCGGAATGCTTGCTATGCTCCTAGATACAGAACTAAGCTCGACACAGAG
AGATTACGCTCAAACCGCTCAAGTATGTGGTAAAGCTTTGATTGCATTGATAAATGAGG
TTCTTGATCGCGCCAAGATTGAAGCTGGAAAGCTGGAGTTGGAATCAGTACCATTTGAT
ATCCGTTCAATATTGGATGATGTCCTTTCTCTATTCTCTGAGGAGTCAAGGAACAAAAG
CATTGAGCTCGCGGTTTTCGTTTCAGACAAAGTACCAGAGATAGTCAAAGGAGATTCAG
GGAGATTTAGACAGATAATCATAAACCTTGTTGGAAATTCGGTTAAATTCACAGAGAAA
GGACATATCTTTGTTAAAGTCCATCTTGCGGAACAATCAAAAGATGAATCTGAACCGAA
AAATGCATTGAATGGTGGAGTGTCTGAAGAAATGATCGTTGTTTCCAAACAGTCAAGTT
ACAACACATTGAGCGGTTACGAAGCTGCTGATGGTCGGAATAGCTGGGATTCATTCAAG
CATTTGGTCTCTGAGGAGCAGTCATTATCGGAGTTTGATATTTCTAGCAATGTTAGGCT
TATGGTTTCAATCGAAGACACGGGTATTGGAATCCCTTTAGTTGCGCAAGGCCGTGTGT
TTATGCCGTTTATGCAAGCAGATAGCTCGACTTCAAGAAACTATGGAGGTACTGGTATT
GGTTTGAGTATAAGCAAGTGTCTTGTTGAACTTATGCGTGGTCAGATAAATTTCATAAG
CCGGCCTCATATTGGAAGCACGTTCTGGTTCACGGCTGTTTTAGAGAAATGCGATAAAT
GCAGTGCGATTAACCATATGAAGAAACCTAATGTGGAACACTTGCCTTCTACTTTTAAA
GGAATGAAAGCTATAGTTGTTGATGCTAAGCCTGTTAGAGCTGCTGTGACTAGATACCA
TATGAAAAGACTCGGAATCAATGTTGATGTCGTGACAAGTCTCAAAACCGCTGTTGTTG
CAGCTGCTGCGTTTGAAAGAAACGGTTCTCCTCTCCCAACAAAACCGCAACTTGATATG
ATCTTAGTAGAGAAAGATTCATGGATTTCAACTGAAGATAATGACTCAGAGATTCGTTT
ATTGAATTCAAGAACCAACGGAAACGTTCATCACAAGTCTCCGAAACTAGCTCTATTCG
CAACAAACATCACAAATTCGGAGTTCGACAGAGCTAAATCCGCAGGATTTGCAGATACG
GTAATAATGAAACCGTTAAGAGCAAGCATGATTGGGGCGTGTCTGCAACAAGTTCTCGA
GCTGAGAAAAACAAGACAACAACATCCAGAAGGATCATCACCCGCAACTCTCAAGAGCT
TGCTTACAGGGAAGAAGATTCTTGTGGTTGATGATAATATAGTTAACAGGAGAGTAGCT
GCAGGAGCTCTCAAGAAATTTGGAGCAGAAGTGGTTTGTGCAGAGAGTGGTCAAGTTGC
TTTGGGTTTGCTTCAGATTCCACACACTTTCGATGCTTGCTTCATGGATATTCAAATGC
CACAGATGGACGGATTTGAAGCAACTCGTCAGATAAGAATGATGGAGAAGGAAACTAAA
GAGAAGACAAATCTCGAATGGCATTTACCGATTCTAGCGATGACTGCGGATGTGATACA

CGCGACCTACGAGGAATGTCTGAAAAGTGGGATGGATGGTTACGTCTCCAAACCTTTTG
AAGAAGAGAATCTCTATAAATCCGTTGCCAAATCATTCAAACCTAATCCTATCTCACCT
TCGTCGTAATCCAATCTTCCGGCGAGTTTTTTTTCTCTCTCCGCAGCCGGAAGAGTGGA
CCGATTCTGCTGATTGATATGCATTTTGGTTTCTGTACATACAGTAGGTTCACAATCTA
GAGATTTTGAAGGTTTTTTTTTCTTTCACCGAAGTAATGTAGCTTGCCATGACTAGTGT
ATGTTGTTAAACGACAACGTCTAAGACGACGGTTCAGTGTTGATCTTAGCGTAAGTATT
AATCCCACGGGATCGTTTGTACTGTATCAGATTTGGTTAGTCGTTTAAACATTGTAATG
TTCTAATAATAACTTTTCCAT

Associated genomic coordinate: AT2G01830.2

Promoter GL2

>gi|334184031|ref|NM_001198514.1|
TTTATATCATTCCAACATAATTCATATTAAAGTTAGTAGCTGAAATTGGAAGGCTGATA
TATTTTCCATAATTCAAATTTGAATTTTGCTCATCATATATATATGTATATATTAAAAA
TCGAATATTAAGAAGAAAAATGAAGTCGATCGATGGCTGCCAATGCTGTAGCTGGCCAT
GTTTTAAACTACTCAATTCAAAGAAGCTAGCTAGGGACAGGATTTGTATGTCAATGGCC
GTCGACATGTCTTCCAAACAACCCACCAAAGACTTTTTCTCCTCTCCAGCCCTCTCTCT
ATCTCTCGCTGGGATATTCCGGAATGCATCCTCCGGCAGCACCAACCCTGAGGAGGATT
TCCTGGGCAGAAGAGTAGTTGACGATGAGGATCGCACTGTGGAGATGAGCAGCGAGAAC
TCAGGACCCACGAGATCCAGATCAGAGGAGGATTTGGAGGGTGAGGATCACGACGATGA
GGAGGAGGAAGAGGAGGACGGCGCAGCTGGAAACAAGGGCACTAATAAGAGAAAGAGGA
AGAAGTATCATCGTCACACCACCGATCAGATCAGACACATGGAAGCGCTATTCAAAGAG
ACACCACATCCGGACGAGAAGCAAAGACAGCAGCTGAGCAAGCAACTAGGGCTGGCCCC
TCGCCAGGTCAAGTTCTGGTTCCAAAACCGCCGCACACAGATCAAGGCTATTCAAGAAC
GGCACGAGAACTCCCTGCTCAAGGCGGAACTAGAGAAGCTGCGAGAGGAAAACAAAGCC
ATGAGGGAGTCTTTTTCCAAGGCTAATTCCTCCTGCCCCAACTGCGGAGGAGGCCCCGA
TGATCTCCACCTCGAAAACTCCAAACTGAAAGCCGAGCTCGATAAGCTTCGTGCAGCTC
TTGGACGCACTCCCTATCCCCTGCAGGCTTCATGCTCCGACGATCAAGAACACCGTCTC
GGCTCTCTCGATTTCTACACGGGCGTCTTTGCCCTCGAGAAGTCCCGTATTGCCGAGAT
TTCTAACCGAGCCACCCTTGAACTCCAGAAGATGGCCACCTCAGGCGAACCTATGTGGC
TCCGCAGCGTTGAGACTGGCCGTGAGATTCTCAACTACGATGAGTACCTCAAGGAGTTT
CCCCAAGCGCAAGCCTCTTCGTTTCCTGGAAGGAAAACCATCGAAGCATCTAGAGATGC
GGGGATTGTGTTTATGGACGCACATAAACTTGCCCAGAGTTTCATGGACGTGGGACAAT
GGAAAGAGACATTTGCATGCTTGATCTCAAAGGCTGCAACGGTCGATGTTATCCGGCAA
GGCGAAGGGCCTTCACGGATCGACGGGGCTATTCAGCTGATGTTCGGAGAGATGCAGCT

GCTCACTCCGGTCGTCCCCACAAGAGAAGTGTACTTCGTGAGAAGCTGCCGGCAGCTGA
GCCCTGAGAAATGGGCAATAGTGGACGTCTCGGTCTCCGTGGAGGACAGCAACACGGAG
AAGGAGGCTTCTCTTCTGAAATGTCGAAAACTCCCCTCCGGTTGCATCATCGAGGACAC
CTCCAACGGTCACTCCAAGGTCACCTGGGTGGAGCACCTCGACGTGTCTGCATCCACAG
TTCAGCCTCTCTTCCGCTCCTTAGTCAACACCGGTTTGGCCTTTGGGGCTCGACACTGG
GTCGCCACCCTTCAGCTCCATTGCGAACGCCTTGTCTTCTTCATGGCTACCAACGTCCC
CACCAAAGACTCTCTCGGAGTTACAACTCTTGCCGGGAGAAAGAGTGTGCTGAAGATGG
CTCAGAGAATGACACAAAGCTTCTACCGCGCCATTGCTGCATCAAGCTACCATCAATGG
ACCAAAATCACCACCAAAACTGGACAAGACATGCGGGTTTCTTCCAGGAAGAACCTTCA
TGATCCTGGCGAGCCCACGGGAGTCATTGTCTGCGCTTCTTCTTCGCTGTGGTTACCTG
TTTCTCCAGCTCTTCTCTTCGATTTCTTTAGAGATGAAGCTCGTCGGCATGAGTGGGAT
GCTTTGTCAAACGGAGCTCATGTTCAGTCTATTGCAAACTTATCCAAGGGACAAGACAG
AGGCAACTCAGTGGCAATCCAGACAGTGAAATCGAGAGAAAAGAGCATATGGGTGCTGC
AAGACAGCAGCACTAACTCGTATGAGTCGGTGGTGGTATACGCTCCCGTAGATATAAAC
ACGACACAGCTGGTGCTCGCGGGACATGATCCAAGCAACATCCAAATCCTCCCCTCTGG
ATTCTCAATCATACCTGATGGAGTAGAGTCACGGCCACTGGTAATAACGTCTACACAAG
ACGACAGAAACAGCCAAGGAGGGTCGCTCCTGACACTCGCCCTCCAAACCCTCATCAAC
CCTTCTCCTGCAGCAAAGCTGAATATGGAGTCTGTGGAATCCGTGACAAACCTCGTCTC
AGTCACACTACACAACATTAAGAGAAGTCTACAAATCGAAGATTGCTGATGACAAGTCA
CAGCAGATATTATTTACCTATATATAATTATATGATAATGTATAGCAGCAGTGCATTAA
AGTTTTGTACAAAAACGACCAGCTCTCTCTTCTCCAATCCTATTATTATCCAACACCTT
TTTGGTCCATTCCATTGGCAAATGAACCATAACAAGAGGAGCAAGAACCGTAGAATTAG
CAGAAACAAAAGTCGGATCACTGAGACCACAAGCACACAGTAGCAACAGAAACATATTA
ATTCACATTCTAAATGTAACTGGAGGTGAAGATGAAGTAAGAGCAAACAATTGGTAGTC
GGAAACAATCAGATTGAAAACACACTCATGGCATAAGCAATGAAATCACAAAAGCATTC
CATAAATTACACTGGTTCCGGGATACACAACAAACAAACAGAAGCAGAGCAAAAAAAAG
ACGG

Associated genomic coordinate: AT1G79840

Promoter SCR

>gi|1497986|gb|U62798.1|ATU62798 Arabidopsis thaliana
CCTTATTTATAACCATGCAATCTCACGACCAACAACCCTTCAATCTCCATGGCGGAATC
CGGCGATTTCAACGGTGGTCAACCTCCTCCTCATAGTCCTCTGAGAACAACTTCTTCCG
GTAGTAGCAGCAGCAACAACCGTGGTCCTCCTCCTCCTCCTCCTCCTTTAGTGATG
GTGAGAAAAAGATTAGCTTCCGAGATGTCTTCTAACCCTGACTACAACAACTCCTCTCG

TCCTCCTCGCCGTGTCTCTCACCTTCTTGACTCCAACTACAATACTGTCACACCACAAC
AACCACCGTCTCTTACGGCGGCGGCTACTGTATCTTCTCAACCAAACCCACCACTCTCT
GTTTGTGGCTTCTCTGGTCTTCCCGTTTTTCCTTCAGACCGTGGTGGTCGGAATGTTAT
GATGTCCGTACAACCAATGGATCAAGACTCTTCATCTTCTTCTGCTTCACCTACTGTAT
GGGTTGACGCCATTATCAGAGACCTTATCCATTCCTCAACTTCAGTCTCTATTCCTCAA
CTTATCCAAAACGTTAGAGACATTATCTTCCCTTGTAACCCAAATCTCGGTGCTCTTCT
TGAATACAGGCTCCGATCTCTCATGCTCCTTGATCCTTCCTCTTCCTCTGACCCTTCTC
CTCAAACTTTCGAACCTCTCTATCAGATCTCCAACAATCCTTCTCCTCCACAACAGCAA
CAGCAGCACCAACAACAACAACAACAGCATAAGCCTCCTCCTCCTCCGATTCAGCAGCA
AGAAAGAGAAAATTCTTCTACCGATGCACCACCGCAACCAGAGACAGTGACGGCCACTG
TTCCCGCCGTCCAAACAAATACGGCGGAGGCTTTAAGAGAGAGGAAGGAAGAGATTAAG
AGGCAGAAGCAAGACGAAGAAGGATTACACCTTCTCACATTGCTGCTACAGTGTGCTGA
AGCTGTCTCTGCTGATAATCTCGAAGAAGCAAACAAGCTTCTTCTTGAGATCTCTCAGT
TATCAACTCCTTACGGGACCTCAGCGCAGAGAGTAGCTGCTTACTTCTCGGAAGCTATG
TCAGCGAGATTACTCAACTCGTGTCTCGGAATTTACGCGGCTTTGCCTTCACGGTGGAT
GCCTCAAACGCATAGCTTGAAAATGGTCTCTGCGTTTCAGGTCTTTAATGGGATAAGCC
CTTTAGTGAAATTCTCACACTTTACAGCGAATCAGGCGATTCAAGAAGCATTTGAGAAA
GAAGACAGTGTACACATCATTGACTTGGACATCATGCAGGGACTTCAATGGCCTGGTTT
ATTCCACATTCTTGCTTCTAGACCTGGAGGACCTCCACACGTGCGACTCACGGGACTTG
GTACTTCCATGGAAGCTCTTCAGGCTACAGGGAAACGTCTTTCGGATTTCACAGATAAG
CTTGGCCTGCCTTTTGAGTTCTGCCCTTTAGCTGAGAAAGTTGGAAACTTGGACACTGA
GAGACTCAATGTGAGGAAAAGGGAAGCTGTGGCTGTTCACTGGCTTCAACATTCTCTTT
ATGATGTCACTGGCTCTGATGCACACACTCTCTGGTTACTCCAAAGGTAAAATAAACAT
TACCTTTTAATCACTCTTTATCTATAAATTATTTTAAGATTATATAGGAAAGATATGTT
CTAAAAAGCTGGCTTTTTTGGTTAATGATTGGGGAATGAACAGATTAGCTCCTAAAGTT
GTGACAGTAGTGGAGCAAGATTTGAGCCACGCTGGTTCTTTCTTAGGAAGATTTGTAGA
GGCAATACATTACTACTCTGCACTCTTTGACTCACTGGGAGCAAGCTACGGCGAAGAGA
GTGAAGAGAGACATGTCGTGGAACAGCAGCTATTATCGAAAGAGATACGGAATGTATTA
GCGGTTGGAGGACCATCGAGAAGCGGTGAAGTGAAGTTTGAGAGCTGGAGGGAGAAAAT
GCAACAATGTGGGTTTAAAGGTATATCTTTAGCTGGAAATGCAGCTACACAAGCGACTC
TACTGTTGGGAATGTTTCCTTCGGATGGTTACACTTTGGTTGATGATAATGGTACACTT
AAGCTTGGATGGAAAGATCTTTCGTTACTCACTGCTTCAGCTTGGACGCCTCGTTCTTA
GTTTTCTTCTCCTTTTTCACAAACAATGTGCCCATAAAT

Associated genomic coordinate: AT3G54220.1

Translatome dataset-Promoter sequences

Promoter SUC2

Primer sequences

SUC2 fw CACCAAGTTACTTTCTATTATTAACTGTTATAATGG

SUC2 rev ATTTGACAAACCAAGAAAGTAAGAAAAAAAAG

>gi|240254421:8032971-8035071 Arabidopsis thaliana chromosome 1

CACCAAGTTACTTTCTATTATTAACTGTTATAATGGACCATGAAATCATTTGCATATGAA
CTGCAATGATACATAATCCACTTTGTTTTGTGGGAGACATTTACCAGATTTCGGTAAATT
GGTATTCCCCCTTTTATGTGATTGGTCATTGATCATTGTTAGTGGCCAGACATTTGAACT
CCCGTTTTTTTGTCTATAAGAATTCGGAAACATATAGTATCCTTTGAAAACGGAGAAACA
AATAACAATGTGGACAAACTAGATATAATTTCAACACAAGACTATGGGAATGATTTTACC
CACTAATTATAATCCGATCACAAGGTTTCAACGAACTAGTTTTCCAGATATCAACCAAAT
TTACTTTGGAATTAAACTAACTTAAAACTAATTGGTTGTTCGTAAATGGTGCTTTTTTTT
TTTGCGGATGTTAGTAAAGGGTTTTATGTATTTTATATTATTAGTTATCTGTTTTCAGTG
TTATGTTGTCTCATCCATAAAGTTTATATGTTTTTTCTTTGCTCTATAACTTATATATAT
ATATGAGTTTACAGTTATATTTATACATTTCAGATACTTGATCGGCATTTTTTTTGGTAA
AAAATATATGCATGAAAAACTCAAGTGTTTCTTTTTTAAGGAATTTTTAAATGGTGATTA
TATGAATATAATCATATGTATATCCGTATATATATGTAGCCAGATAGTTAATTATTTGGG
GGATATTTGAATTATTAATGTTATAATATTCTTTCTTTTGACTCGTCTGGTTAAATTAAA
GAACAAAAAAAACACATACTTTTACTGTTTTAAAAGGTTAAATTAACATAATTTATTGAT
TACAAGTGTCAAGTCCATGACATTGCATGTAGGTTCGAGACTTCAGAGATAACGGAAGAG
ATCGATAATTGTGATCGTAACATCCAGATATGTATGTTTAATTTTCATTTAGATGTGGAT
CAGAGAAGATAAGTCAAACTGTCTTCATAATTTAAGACAACCTCTTTTAATATTTTCCCA
AAACATGTTTTATGTAACTACTTTGCTTATGTGATTGCCTGAGGATACTATTATTCTCTG
TCTTTATTCTCTTCACACCACATTTAAATAGTTTAAGAGCATAGAAATTAATTATTTTCA
AAAAGGTGATTATATGCATGCAAAATAGCACACCATTTATGTTTATATTTTCAAATTATT
TAATACATTTCAATATTTCATAAGTGTGATTTTTTTTTTTTTTGTCAATTTCATAAGTGT
GATTTGTCATTTGTATTAAACAATTGTATCGCGCAGTACAAATAAACAGTGGGAGAGGTG
AAAATGCAGTTATAAAACTGTCCAATAATTTACTAACACATTTAAATATCTAAAAAGAGT
GTTTCAAAAAAAATTCTTTTGAAATAAGAAAAGTGATAGATATTTTTACGCTTTCGTCTG
AAAATAAAACAATAATAGTTTATTAGAAAAATGTTATCACCGAAAATTATTCTAGTGCCA
CTCGCTCGGATCGAAATTCGAAAGTTATATTCTTTCTCTTTACCTAATATAAAAATCACA
AGAAAAATCAATCCGAATATATCTATCAACATAGTATATGCCCTTACATATTGTTTCTGA
CTTTTCTCTATCCGAATTTCTCGCTTCATGGTTTTTTTTTAACATATTCTCATTTAATTT
TCATTACTATTATATAACTAAAAGATGGAAATAAAATAAAGTGTCTTTGAGAATCGAACG
TCCATATCAGTAAGATAGTTTGTGTGAAGGTAAAATCTAAAAGATTTAAGTTCCAAAAAC

AGAAAATAATATATTACGCTAAAAAAGAAGAAAATAATTAAATACAAAACAGAAAAAAAT
AATATACGACAGACACGTGTCACGAAGATACCCTACGCTATAGACACAGCTCTGTTTTCT
CTTTTCTATGCCTCAAGGCTCTCTTAACTTCACTGTCTCCTCTTCGGATAATCCTATCCT
TCTCTTCCTATAAATACCTCTCCACTCTTCCTCTTCCTCCACCACTACAACCACCGCAAC
AACCACCAAAAACCCTCTCAAAGAAATTTCTTTTTTTTCTTACTTTCTTGGTTTGTCAAA
T

Associated genomic coordinate: At1g22710

Promoter WOL
Primer sequences
WOL fw CACCTACTGTCTCTAAGCGCACG
WOL rev CTGAGCTACAACAATAGAGAACAAAAGAAG

>gi|240254678:367432-369575 Arabidopsis thaliana chromosome 2
GTTCCTAAAATCCATTTGAATATTCAAAAACTTCTCTCAAATATCATGTAGTTATAGAAG
CTACTGTCTCTAAGCGCACGAGAGAAAGCTACACAACCCACGTCAGTTTCCATCTACACA
TATAAGGTAATAATAATATTTTCATGTATCTTTAATAATAGCTCTATGTTTTTTTCTGTA
TTTTTCATTATAAAACTCATAACTATGTTATCATTTAATATGGTACTAATTTAATGGGAT
TGATTTACTATTGCCTCAAACATGTAATAATTTAATGATTTTTTGTTTTTAACGTTTTTA
GAAATTCATGAGCATTTTAAATTTGTGGTTAGGTCATAACAATTTGCTATTACAAAAAAA
AGAAACACTCTAAATAATATAAAAAATAGTTTACCGTATAATACTAGTAGTAAATAAATA
ATTTGATTGTTATTCATAAATTTTGAATTCTAAAATCTCCTGAATCAACTCATGCAATTG
TCTTAAGAATTACACGTGGATAAATCATGGGCTTATGAGTCAGGCCCATTTAACCGGGGT
ATTTTCGTAGTTAAGAGACTAGAATGGTGGGTATTTCAGGTAAAAGGTCTATGGGGCCAG
ATCTGCGCTTTGTCGCGATGTCATTATCGCCAAAGATATGCGATAGCGACTCTCGTACAA
AGTCTCTCACTCACCTATATTTTTTGTTTTCTTATATTTCAACAAAAAAACGTTTTATTT
TCCTTTTGGTGTAAGTAAAAAAACAAAACAAAACGTTTTATTTCTAAAGTTCAGAAAACT
TATTTATACCAAGGAAAAAATAGATAATAAATTTTGAGAAGTTGGTGACTATATATTACT
TCACTTATTCAAGAAATTTAAACATGGTAAATGTTACTTTAAATGTTAAATGATGTATAA
GAAATGTAATGAAATTGAATAAATGTAGTTTTAAAGATGTTTTAATTAGTAAGACAAACC
TAGTTAGTGTCACAATAATTATATTTTTTTTTTTGTCATCCAAAATTATTAAAGCTCAAG
TAAACCAATCCTGAGGGATATTATTTACAAATGTGATATGATGCGGTTCGGTGCGGATCT
TCCGCGCCAAATTATACGCTTTTATATTAGCATTATAAAAAATTATAGATAAAGAGAAGT
TTGTGAATTCTTCATTGTCGCTTTGCAATTTCTCTAAATACACAGTAAATACCGACAATT
CGGTTAGAGAAAATATATCTATTTCGTATAATAATGTTAACTTTGAGGAGATTTTGGGTA
AAATAATAACTTTTGTTGGATGGATCATATCATGAGCCATTAAGAAAAAGTCCAAAACTT

TTCTTCTTCAAAGTTGGACTCAAGTTAGAAAAAGAAAAAAGAGCTAGAGAGATATAAAAA
TGAAAAGAAAGTTCATGGCAAAAAACTGATATAGACAGAGACACAGAGAGAGAGAAACGT
ATCTGAAGAAAATCTAAAAAATTCGATTCAATTTTTTTCTTACTTTTAAAAGCAAAAAAT
CTCACTAAAACAAAAGAAGAAGAAAGAAGAAAGAAAATGGAATACCTACATTTGAAGTGA
TGAGAAGAGATTTTGTGTATAATAATAATGCAATGTTCAATCCTCTCACAACTCATTACA
GGTAACTAAAATAATTTCTCCATGTGCTTGCTTATTAGTCGTTCTTCCTAATGTTATGTT
TCTCTCTGTGTTCTTTCTTTCTTTGGTCAAAGCTTTAATTTTTTTTCTATTGTTGGATTT
GAGACAGTGAACATAGCTATGTTCTTGTTCCAATAATAAACAATCACGCCTGTAAAGAGC
TTATGATTGATTAGTGTGTTTTTTAGTATTAATTAATTTCTCTGACAATAATTACTTAGT
TTTTAATTCTTCTCTGTAAGAAACCTTTGGAAACTGAGCAAAGTTGCTTCTTTTGAGAAC
CATGCGTTTCTTTCTCTCTTTTGTTCTTGAATTCGCAAAAACATGTCCTTTTTCGTCTAC
AGGTTTCTAGGGTTTGTTTCTGTACTATAAACTATGTTTATGGTAACATTCTTAATCATA
ACTACACTACCAATGCTTTTATGTTATATGTATGCAAAAAAGGCTCTAACTTTTGTTTTC
TTTCACTATTGTTTCTTCTTTTGTTCTCTATTGTTGTAGCTCAG

Associated genomic coordinate: At2g01830

Promoter GL2

Primer sequences

GL2 fw CACCGTTTCCTTCACTATACGTCTTCGTCC

GL2 rev CTGTCCCTAGCTAGCTTCTTTGC

>gi|240254421:30035467-30037518 Arabidopsis thaliana chromosome 1
GTTTCCTTCACTATACGTCTTCGTCCATTTACGTACGTATTATACGGACGGTTTAAGCT
ACTATATCTATATTGTTAACAATGTAACTGTTGAGATATATCTTGCAATAATATGTCAT
GGTGTATGCATACGATAATATGAATCAATGTTTGAAATCTTGACGTGCCCGTGATACAA
TAAGATGATCAAAATTTCAAATTTTGTCAAATATTAAAACAACATACACATACACATGT
GTCCAGGTGGCATTATAAAATGTATATATGGTGGATATAGAGAGAGAGGGAGATGCGTA
TAGTGAATAGGAAAGTAAGTAATAAAGAGAGGGTGGAGGAATTGGAAAGGGGTTGGAGG
CAAACCCATAAAGAGCATTCATTTCCTTTTAAGGTCGCTGAAATTAATGAGTAACGATC
GGTCAATGCCTCTCGCTGACCTTTTTCTTTTTTTACAACAACAAATAAAAATAAAATAA
ATTTCGACGTCTCTTTCCGCTGCTGAATTACATTTGTTGAATTAATTTTCTCTGCTTAC
GTACGTCTTCTAAACTTTCTCTATCCGAATTCTTTTTTAACTTTCTAACTTATATTCAA
CAACTCTTCTTTCCTGCCTTTACCGTTAGTCTAATTGTTTTCCTAATACTGCTACGTAC
ATACCCCTACTATACTAGTCAGTGTATTAGATTCGATTGGGATTAATCCAGGAATATAG
ATATCCCATTAGTTTTTATAAAAATATTGGAAGAGGACAAGTCTCAAGCAATTTAGGGT

TCCATGTAGCGCTGCAATATACTGTTAGTAACTCTCTCTTACCCATATATTGTATATGC
TAATTCTTATCAAATATATATATATGCTTCTCCCAGAGTCCCAGTTTCCTATAATCCTG
ACGCAATTATACTAATAGAGCCAAGTTTACATAATAAAGTATATATGATTAATAGATAG
GGTTTCTTATTAAGCCATATCTTAAATTAAGATGTGATGATAGCGTTTTGTATAAGTTA
CCAATTGTTTGAAAGAAGAGATCATCACAATAATAAATCATAAGTAGTAGTATATAGTA
ATAAATAAATACACAAGTCATAATAAGAGTAATGAGAGGATAATTAAGGAGGGAAGAAG
AAAGCAGAAAATGCGGTTGGAGAATTAGGTGCTAAAAGTTAGTTGAGTCCATCTCAGTA
TCTAACGGTCAACTCTCTCTCTCTCTAGAGAAAACAATTAAGAAATCTGACATACACAT
ATGTCTCTCTCTCTCTCTCTCTCTAGTCTATACACACAATTCAATTAAAGAAGAGACAG
AGAAGTTCGTCTTTTTTGTTTTTATACCCTTAAATCAATCATGCAATTGTAACCCTTCC
TTCTTATTCTCATTCCTTCCCCCCCTGTCTACAGTAATCTATAGCAACGCCATTATGTA
CTACTTTTAACGGATAATTTGCTCATGTTTCAATATGGCTTCATTGTATATATGTTCAA
GTTCTTCTCAATCCTTTATATCATTCCAACATAATTCATATTAAAGTTAGTAGCTGAAA
TTGGAAGGCTGATATATTTTCCATAATTCAAATTTGAATTTTGCTCATCATATATATAT
GTATATATTAAAAATCGAATATTAAGAAGAAAAATGAAGTCGATCGATGGCTGCCAATG
CTGTAGCTGGCCATGTTTTAAACTACTCAATTGTCGGATTGAAGTATAGCCAAAATATA
TAAAACCGTAAAAGGACTAAATATAATAATATAATAGGTATTAATTAATTAAAACTAAT
TAATTATAAAAGAAGCACCTAAAAGTCAAGAGCAGTAGAGAAATGGAAGAAATATCTGA
AAAACGACCGCTTATATATATATGTATCATTGGAATTGAAGAGGCTATATATATATATA
TATATATATATATCGATCTTAGCTTATATATTAATTGAAAGTACATTTTGGTGTATAAG
TAATTAAAGAAGAAAGAAAAAAAGAGAGATAATATATAAGGAAGAAGGAGTGCGAGGAG
AAGAGGGAAGAGATCATAATTAAGCAAAGAAGCTAGCTAGGGACAG

Associated genomic coordinate: At1g79840

Promoter SCR

Primer sequences

SCR fw CACCGGATAAGGGATAGAGGAAGAGG

SCR rev GGAGATTGAAGGGTTGTTGGTCG

>gi|240255695:20068432-20070549 Arabidopsis thaliana chromosome 3
GGAGATTGAAGGGTTGTTGGTCGTGAGATTGCATGGTTATAAATAAGGTAAGAAAAGGGT
TAAATCCAAAATCGAAAATTTCCAAACAGAAAAAGGTTTGGTAGTAGAAAAGCTAGACTC
TTTTATCAATGGTGGGATTCTTAAATTACTTGATCGTCTTCACGATCTCCGATGAGGAGA
AAAAGCAAAAACTCAGTGAGGGGGTGAGAATTTGGAAGGATGTGGGTTGGAGATGAATCC
CGGAAAAAGGGGTTAGGGGTAATTGAGGAGGACGAGGAGGACAGTGAAGCCTGGCCGTAT
CTAAGTCGTCTTCCACCTTTCGAATTATTGTGAATGTCGCTTCGTACATTTAATGACTCT

TTCTATCTCTATTCATCTTAACTCTTTCTCTCTCTGTTTCTGTTTTTTTCTTCTTAATTA
ATTTGAGTTTTGAAAAATAATAAAAATGGGAAGAGATTAGGTGGGTGATCGGAAAATGTT
CGTCGTTGTGTGTTTGACGTCTTCCGTTTGCCGATCCATTTCGTTTCTTCTTGACCTCTC
TCTCGAACTAACCACAAGAAGATAATATTTTGTTTGTTTGGTCTATAATAATAGAGCTCT
ATTTCGTGGAACCGGTACAATATATAACAAAATTTCAGAATATTATGATGTATGACAAAG
TTAATAAATTATGGAATATCAAAATTGTTGAATAATGAAAAAGATGATCTTGCTAATTAC
AAAATTGAGTAGTAAATGATGTAAATAAAATGTCAGATTCAGTTCATTATGTGAAATGAA
TGGGTTTCTATTCTATAGAGCTACATGAGTTGGATTCATAAACCGATTTATTAGTTATGG
TCATATGGAGTCAGCAGGTTGCACCACTAGTCATTTGATTTAGGTTTGGTATGAGAGATG
GGATGAAGAAGTCTCTCACACAATCAAAGTGTGGTACGATGTGCTTACTAAGATGGGGCA
CCAACTTGCAATTTTGTGCTAAAATTAGAAAGATGGTATTCCTTAAACCGTACTAATCTA
AAGAGAGGAGAAAAATAAAGGAATCCAAGTAAAGATTTAGACAAAACAAGGATAAGAAAA
GGGAAAAAGCATTAGTCCCCTTGTATATTAGTATCATATTAAAGGCATTTTACTTGAGAG
GAATTGGACAAGACGTAGAGCTAAAACAGTAAATAACAAAAAAGGAAAAAGATGCACATG
TTTTGATGAGAGGCCATTACATACATACACCAACTGTCAAACCCCAGTTCCCGATAAAAT
CTGAAACTAGTGTTTGACATTTTGCTTTCATCTGCATCCTCATCTTTTCCCTTTTGTTCT
TTTTTTAACAATTTTTAGAAAATTTGCTATTATTATATTATTTTTCTTACAAAGTTTTGG
ACCTGTTTGACTGACAATTTTTAGACTTTTGTACGTCAAACACTAGCCCCATGGCCAATT
TAATAATCGCATGGCAGTGAACCCAAGAAGAAACCATCCACGCTTTCTACGATCTTATTT
TTTCTTCTCCTTTTTCACTTCAAATTATTTTTTATTACTTCTCTTCACAACACAAACACA
CGGCTCAATATGAAAGTTTCTCAGCGTAAAAAGCTGAATGTCTTTCTCCAAGTCCATATT
TGAATAGTCTTTTCTGCCCTTTTAGACTTAATTTGTTAATTTTTGAGTGATCTTAACGAA
TACCATGTACTATTATACACAAAAAATGTGAAATGTATTTTACTACATAGTTTTTTTTAA
TACTCTTTATCTTCAATTTAACTGATATTAAGGATTTTCCCAAAATTTATCCAATTTTAA
GAAAGTATATAAAATATTTTAATATATCTTAGAACACCATTTATTAACATCAAAGTCTCT
ATAATAGAAATGCTCATTAAAACCAAAATAAAAATGAAATGTTTGTAAATCACAAATGCA
CTTAAACAATATCTAGCAATAGCATACATTATAAAATAATTCTAACATTACATAGCCCAA
ATGCAAATATCTGTTTGGCAACAAAAACCTTAAAAAGTCTCTTGTTGGCAAAGCGCTACA
GAGTTACAGTTTATAGGCCCATTAAGGCCCATCAAAAGGTTTCTGATAAACAAAGTCCTC
TTCCTCTATCCCTTATCC

Associated genomic coordinate: At3g54220

Appendix D: Abbreviations

AFIS	Advanced fiber information system
AGL42	Agamous-like 42
ANOVA	Analysis of variance
APL	Altered phloem development
BH	Benjamini-Hochberg
BPCA	Bayesian principal component analysis
CCA	Canonical correlation analysis
CCA – EN	Canonical correlation analysis-Elastic net
CDTA	Cyclohexanediamine tetraacetic acid
CIA	Co-inertia analysis
CO2	Cortex-specific transcript
COA	Correspondence analysis
CoMPP	Comprehensive microarray polymer profiling
CORTEX	Cortex
CV	Coefficient of variation
DEG	Differentialy expressed genes
DNA	Deoxyribonucleic acid
EC	Expression conservation
FACS	Fluorescence activated cell sorting
FDR	False discovery rate
FPT	Fluorescent protein technology
FT – IR	Fourier transformed infrared spectroscopy
GEO	Gene expression omnibus
GFP	Green fluorescent protein
GL2	Glabra2
GO	Gene ontology
GO – BP	Gene ontology biological process

GSEA	Gene set enrichment analysis
HVI	High volume instrument
ICA	Independent component analysis
KEGG	Kyoto encyclopedia of genes and genomes
KNN	K nearest neighbors
mAbs	Monoclonal antibodies
MCIA	Multiple co-inertia analysis
mRNA	Messenger ribonucleic acid
mRNP	Messenger ribonucleoprotein
NaBH4	Sodium tetrahydridoborate
NaOH	Sodium hydroxide
NGS	Next-generation sequencing
OLS	Ordinary least squares
PEP	Endopeptidase
PCs	Principal components
PCA	Principal component analysis
PCC	Pearson correlation coefficient
PLS	Partial least squares
PPCA	Probabilistic principal component analysis
RMA	Robust multichip average
RMSEP	Root mean squared error in prediction
GCR	Roy's greatest characteristic criterion
SCR	Scarecrow
SHR	Shortroot
sPLS	Sparse partial least squares
SUC2	Sucrose transporter 2
SULTR2	Sulfate transporter
SVD	Singular value decomposition
TukeyHSD	Tukey Honest significance difference
USDA	United States department of agriculture
WOL	Woodenleg

Curriculum Vitae

Name:	Dhivyaa Rajasundaram
Date of Birth:	6^{th} October, 1988
Place of Birth:	Pondicherry, India
Nationality:	Indian
Address:	Melanchthonstr 5, D-10557 Berlin
E-mail:	rajasund@uni-potsdam.de

Education

05/2012-05/2015	PhD student in Bioinformatics in the group of Prof. Dr. Joachim Selbig, University of Potsdam, & Max-Planck Institute for Molecular Plant Physiology in Potsdam, Germany. Marie Skłodowska-Curie (MC) early stage researcher in the FP7 Initial Training Network (ITN) - WallTraC. (`https://www.walltrac-itn.eu/`)
05/2011-05/2012	Master of technology in Plant Biotechnology, TamilNadu Agricultural University, India.
05/2010-05/2011	Master of professional studies in Plant Genetics, Cornell University, USA.
06/2006-04/2010	Bachelor of technology in Bioinformatics, TamilNadu Agricultural University, India.

Honors

April 2012	MC fellowship to pursue doctoral studies in Germany.
April 2010	Navajbhai Ratan Tata scholarship to pursue graduate studies in USA.

WallTraC secondment activities

03/2014-06/2014	Institut national de la recherche agronomique, France.
11/2013-12/2013	Bayer Cropsciences, Belgium.

Teaching activities

WiSe 2014/15	University master course: Statistical bioinformatics, University of Potsdam.

Selected seminars and presentations

11.11.2014	sMODIA (Statistical methods for omics data integration and analysis) 2014: Integrative analysis of multi-source data: Application and methodologies, Crete, Greece.
04.07.2014	Progress seminar II: Integrative analysis of plant cell wall related complex and heterogeneous data, Potsdam-Golm, Germany.
03.07.2014	PhD day: Integrative analysis of plant cell wall related complex and heterogeneous data, Potsdam, Germany.
19.05.2014	WallTraC symposium: Co-ordination and divergence of cell-specific transcription and translation in Arabidopsis root cells, Paris, France
12.05.2014	Affinity seminar: Integrative analysis of cell-related data using multiblock methods, Nantes, France.
12.04.2013	Progress seminar I: Integrative analysis of Transcriptome, Translatome and Proteome data from root cell-types of Arabidopsis thaliana, Potsdam-Golm, Germany.

Conference participation

10.11-12.11.2014	Oral presentation at *sMODIA (Statistical Methods for Omics Data Integration and Analysis)*, Heraklion, Crete.
19.07-23.07.2013	Poster presentation at the 21st *Annual International Conference on Intelligent Systems for Molecular Biology and 12th European Conference on Computational Biology*, Berlin, Germany.
07.07-12.07.2013	Poster presentation at the XIIIth *Cell Wall Meeting*, Nantes, France.
18.06-19.06.2013	Poster presentation at the 2nd *Plants and People Conference*, Potsdam-Golm, Germany.

WallTraC network meetings and training events

06.10-10.10.2014	Introduction to industrial pectin production and regulatory affairs, by CP Kelco Aps in Lille Skensved, Denmark.
21.05-23.05.2014	Mass spectrometry & standards for quality management systems, by INRA in Nantes, France.
16.11-20.11.2013	Cloning and characterization of recombinant proteins and grant proposal writing, by Newcastle University in Newcastle-upon-Tyne, United Kingdom.
13.06-17.06.2013	Systems biology-oriented bioinformatics and EU project management, by the University of Potsdam in Potsdam, Germany.
14.01-18.01.2013	Plant cell wall molecular probes and ethics, by the University of Leeds in Leeds, UK.
11.06-15.06.2012	Plant cell wall architecture and commercial exploitation of results, by Bayer CropScience NV in Gent, Belgium.

Publications

- **Rajasundaram D**, Selbig J, Persson S, Klie S. 2014. Co-ordination and divergence of cell-specific transcription and translation of genes in Arabidopsis root cells. *Annals of Botany*, 114:1109–1123.
- **Rajasundaram D**, Runavot JL, Guo X, Meulewaeter F, Willats W, Selbig J. 2014. Understanding the relation between cotton fiber properties and non-cellulosic cell wall polysaccharides. *PLoS ONE*, 9:e112168.

Selbständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig und unter Verwendung keiner anderen als den von mir angegebenen Quellen und Hilfsmitteln verfasst habe. Ferner erkläre ich, dass ich bisher weder an der Universität Potsdam noch anderweitig versucht habe, eine Dissertation einzureichen oder mich einer Doktorprüfung zu unterziehen.

Potsdam, 3.März 2015

_____

Dhivyaa Rajasundaram

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7 2007-2013) under Grant Agreement n°263916. This presentation reflects the author's views only. The European Community is not liable for any use that may be made of the information contained herein. More information about the WallTraC project at www.walltrac-itn.eu.