

Prediction of protein-protein complexes by combining size exclusion chromatography and mass spectrometric analysis

**Dissertation for Obtaining the Doctoral Degree of
Natural Sciences (Dr. rer. nat.)**

**Faculty of Natural Sciences
University of Hohenheim**

Institute of Biology

Department of Plant Systems Biology (190d), Professor Dr. Schulze

Submitted by

Max Gilbert

from Ludwigsburg, Germany, 2021

Dean:	Prof. Dr. Uwe Beifuss
First Reviewer:	Prof. Dr. Waltraud Schulze
Second Reviewer:	Prof. Dr. Andreas Schaller
Submission Date:	08.06.2021
Date of oral examination:	26.08.2021

Die vorliegende Arbeit wurde am 11.06.2021 von der Fakultät Naturwissenschaften der Universität Hohenheim als „Dissertation zur Erlangung des Doktorgrades der Naturwissenschaften“ angenommen.

Table of Contents

Eidesstattliche Versicherung	2
Curriculum Vitae	3
Acknowledgment.....	4
Summary	5
Zusammenfassung.....	7
Preface.....	9
General Introduction	12
Tandem-Mass spectrometry based proteomics	12
Proteomic workflow	16
Proteomic Systems Biology	19
Protein interaction	23
Relevance of protein interactions in plants and Nitrogen response.....	25
Research Question	27
Method establishment	27
Application to Nitrogen dynamics.....	28
First Publication.....	29
Second Publication	43
General discussion.....	56
Experimental procedure.....	56
Data processing strategy and data dynamics.....	58
Protein dynamics and database considerations.....	59
Network analysis	65
Dynamic induced changes.....	67
Data processing automatization	69
Closing remarks	71
Literature.....	72
Addendum	80
Script for initial processing	80
Script Pearson Correlation	85
Script “Confirmed Interactions”	93
Script “dF Evaluation”.....	98
Script “Abundance Rating”	103



Anlage 2 zur Promotionsordnung der Universität Hohenheim zum Dr. rer. nat.

**Eidesstattliche Versicherung gemäß § 7 Absatz 7 der Promotionsordnung der
Universität Hohenheim zum Dr. rer. nat.**

1. Bei der eingereichten Dissertation zum Thema

Prediction of protein-protein complexes by combining size exclusion chromatography and mass spectrometric analysis

handelt es sich um meine eigenständig erbrachte Leistung.

2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.
3. Ich habe nicht die Hilfe einer kommerziellen Promotionsvermittlung oder -beratung in Anspruch genommen.
4. Die Bedeutung der eidesstattlichen Versicherung und der strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt.

Die Richtigkeit der vorstehenden Erklärung bestätige ich: Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erkläre und nichts verschwiegen habe.

Ludwigsburg, 17.05.2020
Ort und Datum


Unterschrift

MAX GILBERT

| CURRICULUM VITAE |

OVERVIEW

Name: Max Gilbert
Birthday: 13.09.1990
Place of Birth: Ludwigsburg
Address: Im Weizen 6/1, 71636 Ludwigsburg

ACADEMIC EDUCATION

Academic Path

- 2016-2021: Promotion at the Institute for Plant Systems Biology at the University of Hohenheim, Germany
- 2013-2016: Masters in Biology, University of Hohenheim
- 2010-2013: Bachelor in Biology, University of Hohenheim

School

- 2001-2010: Mörike-Gymnasium, Ludwigsburg
Abitur (25.06.2010)
- 1997-2001: Primary School Pflugfelden, Ludwigsburg

CAREER

23.03.18 – ...

Scientific Employee as Ph.D Candidate at the Institute for Plant Systems Biology at the University of Hohenheim, Germany

01.03.16 – 31.03.18

Scholarship, Landesgraduiertendenförderung BaWü

17.05.2021, 
Datum, Unterschrift

Acknowledgment

I would like to express my sincere gratitude and appreciation to Prof. Dr. Waltraud Schulze for her excellent supervision, caring suggestions and the sharing of valuable experience and expertise throughout the years. Coming with nothing but good intend and ideas, your trust gave me room and confidence to grow and to undertake this research endeavor.

A very special thanks goes to my then colleagues and now friends with whom I shared uncountable memorable experiences and joy. Thank you Xuna, Lin, Zhi and Liangcui for your tutelage in the lab when I was still at the start and everything was new. Thanks goes also to Jiahui, Mingjie and Sven, who I value much and wish them all the best to carry the torch! A special award goes to Melina, our secretary and my best desk neighbor ever. Thank you also Susanne for all the assistance in lab and cooking advice! Thank you Zhaoxia and Thorsten for giving an independent perspective and the all the nice conversations. Another round of thanks goes to all the students visiting our lab which I truly enjoyed. You guys still teach me a lot that cannot be taught in the lab!

We had a good time, didn't we?

Extending my gratitude, I would also like to acknowledge all support received by Prof. Dr. Andreas Schaller, Dr. Annick Stinzi and the whole Institute of Plant Physiology (190 c – “c” as in “cool”). After all, your lectures started my interest for molecular plant physiology!

Thanks goes also to the ZMBP in Tübingen to Prof. Klaus Harter and Leander Rohr, with whom the opportunity for a fruitful collaboration arose.

Finally, I am eternally grateful for my family, who give me a loving home. I couldn't have done it without you.

...and also thanks to my friends for suffering my ramblings about this strange world of science!

Summary

Protein protein interactions (PPI) are required for all cellular processes and functions in all kingdoms of life. Protein complexes not only catalyze metabolic processes and maintain basic functionality, but they also serve as an interface to the outside world through signal perception and integration. To perceive signals from and react to a changing environment is key to survival, especially for plants which are multicellular sessile organisms directly subjected to their immediate environment. In the context of daily environmental fluctuations, adaptation cannot rely on evolutionary and genetic processes that take effect after generations. Because environmental perturbations are frequently sudden and brief, a response must be adequate and quick. Gene transcription and protein biogenesis are often too slow and cumbersome to respond to rapid changes, so proteins and their interactions bear the brunt of the initial response. Plants depend on a complex network of membrane-bound receptors and co-receptors to regulate signal response, and a great scientific effort is made to classify signaling proteins and to understand the mechanics of signal perception.

However, in order to fully comprehend signal response, a systematic approach is needed to place each individual components in context and to explain the spatial or temporal complex behaviors. Uncovering the network of protein interactions is also a prerequisite to the explanation of the plant phenotype. A better understanding of these mechanisms enables scientists to find solutions to emerging concerns of global importance such as increasing food production and exposure of crops to weather extremes in context of climate change. Nitrogen is an important macronutrient in terms of food security, crop yield stability, and plant nutrition. To meet global demand, nitrogen fertilization is being increased in order to access barren lands or increase crop yield – often at the expense of the local biome. A better understanding of nitrogen signaling in plants is therefore required to implement new strategies for long-term sustainable crop growth.

Two major objectives were pursued and met in this study.

First, the goal was to add to the scientific toolbox a diligent method for uncovering PPI dynamics on a proteomic scale, with a focus on plant membranes. There are large-scale or high-throughput approaches, but they rely on genetically modified proteins or heterologous expression systems to describe PPI outside of their natural context. Similarly, those methods are incapable of describing the dynamics of protein interactions. In course of this study, a co-elution based approach was combined with modern mass spectrometric label free quantification in order to investigate PPI and interaction dynamics on a proteomic scale. A rigorous data processing pipeline was developed to not only address known fallacies of using co-elution based methods (such as for example random co elution), but also to access and utilize meta-information in form of protein abundance and protein network connectivity to draw conclusions not only on proteomic scale, but also for individual proteins. In total, 6.928 individual proteins extracted from *Arabidopsis thaliana* root membranes were detected under different nutritional conditions (full nutrition, nitrogen starvation and nitrogen resupply). The data processing pipeline described in this study was used to predict and discover connectivity information for at least 2.058 of these proteins. Each step in data processing was validated by comparison to database confirmed interactions to improve filtering criteria. Protein abundance was evaluated through a unique ranking system, allowing a seamless integration as network attributes for each condition. From the suggested interaction data, an interactome network of the various nutritional conditions was reconstructed. Using different network parameters from graph theory, protein significance and dynamic conditional changes were described.

Second, this study applied the aforementioned approach to identify relevant proteins involved in nitrogen signaling in *Arabidopsis thaliana* root membranes. Through correlation analysis and network

reconstruction, receptor kinase AT5G49770 was identified as a component of the nitrogen signaling network that collaborates with co-receptor QSK1, BAK1, the nitrogen transporter NRT2.1 and proton pump AHA2. In response to nitrogen deficiency, the network parameters of AT5G49770 reacted strongly and its involvement was demonstrated by a phenotypic similarity to knock-out lines of NRT2.1, NRT1.1 and AHA2 during a root growth assay of Arabidopsis seedlings. The interaction between QSK1 and BAK1 was further confirmed using FRET/FLIM microscopy and pulldown assays.

These findings show that combining a co-elution based approach with a rigorous data processing pipeline and network analysis is suitable to study the protein interaction environment and signal response dynamics in plant root membranes. The modular experimental design allows for a simple adaptation to study different stimuli and the unbiased proteomic approach yields results for proteins regardless of the individual scientific focus. Meta-information such as protein abundance and network connectivity parameters can be used to prospect and identify important proteins involved in stress response dynamics.

The author of this study is confident that the proteomic data produced can be utilized in further research and contributes to the understanding of nitrogen signaling in plant root membranes. Through integration of the data processing pipeline and adaptation to different scientific scenarios, valuable information beyond protein interaction is gained. Thus, this work makes an important contribution to the advancement of proteomic analysis and data interpretation methodology.

Zusammenfassung

In allen Reichen des Lebens sind Protein-Protein-Interaktionen (PPI) eine entscheidende Voraussetzung für sämtliche zellulären Prozesse und Funktionen. Proteinkomplexe ermöglichen nicht nur Stoffwechselfvorgänge und eine Grundfunktionalität, sondern stellen über die Signalwahrnehmung und – Integration auch eine Schnittstelle zur Außenwelt dar. Signale aus einer sich verändernden Umwelt wahrzunehmen und darauf zu reagieren, ist der Schlüssel zum Überleben, insbesondere für Pflanzen, die als mehrzellige, sessile Organismen direkt ihrer unmittelbaren Umgebung ausgesetzt sind. Eine Anpassung an sich ständig verändernde Umwelteinflüsse kann nicht alleine von evolutionären oder genetischen Prozessen abhängig sein, da deren Effekte sich erst nach mehreren Generationen entfalten. Gentranskription und Proteinbiogenese sind oft zu langsam und schwerfällig, um auf diese schnellen Veränderungen zu reagieren. Die Hauptlast der unmittelbaren Anpassung und Reaktion wird stattdessen von Protein getragen. Ein großer wissenschaftlicher Schwerpunkt liegt dabei auf der Beschreibung von Mechanismen der Signalwahrnehmung und der Identifizierung von Signalproteinen. Ebenso wichtig ist es jedoch, diese individuellen Komponenten in einen gesamt einheitlichen Kontext zu rücken und über die Beschreibung der räumlichen und zeitlichen Zusammenhänge ein besseres Verständnis der Signalantwort herzustellen. Das Erkennen der Vernetzungen von Proteininteraktionen und –komplexen ist Voraussetzung für die Erklärung der molekularen Mechanismen, die zu einem Phänotyp führen. Ein besseres Verständnis dieser Zusammenhänge kann dazu genutzt werden, um Lösungen für aufkommende Probleme mit globaler Bedeutung zu finden, wie zum Beispiel die Nahrungsmittelproduktion und Exposition von Nutzpflanzen gegenüber zunehmenden Wetterextremen im Kontext des Klimawandels. In Hinblick auf Ernährungssicherheit, Stabilität der Ernteerträge und Pflanzenernährung ist Stickstoff der wichtigste Makronährstoff. Da der Bedarf an Dünger stetig zunimmt, um den globalen Bedarf zu erfüllen, kann ein besseres Verständnis des Stickstoffsignalweges in Pflanzen dazu genutzt werden, um Pflanzenernährung neu zu überdenken und nachhaltige Strategien zur Sicherung der Welternährung zu entwickeln.

Mit dieser Studie wurden zwei Hauptziele verfolgt und erreicht.

Zunächst war es dezidiertes Ziel dieser Arbeit das wissenschaftliche Methodenspektrum um eine sorgfältige Methode zur Aufdeckung von Protein-Protein Interaktionsdynamiken auf proteomischer Ebene unter besonderer Berücksichtigung der pflanzlichen Membranumgebung zu ergänzen. Zwar gibt es Ansätze für large-scale oder high-throuput Methoden, jedoch sind diese auf genetisch veränderte Proteine oder heterologe Expressionssysteme angewiesen und beschreiben PPI damit außerhalb des natürlichen Kontextes. Ebenso sind diese Methoden ungeeignet, um Dynamiken von Proteininteraktionen zu beschreiben. Zu diesen Zweck wurde in dieser Studie eine Methode entwickelt, die Vorteile eines Co-Elutions-basierten Ansatzes mit modernen label-freien massenspektrometrischen Quantifizierungsmethoden kombiniert. Dazu musste eine belastbare Datenverarbeitungspipeline entwickelt werden, um nicht nur bekannte Fallstricke von Co-Elutionsmethoden (wie z.B. zufällige Co-Elution) zu vermeiden, sondern auch um Metainformationen wie Proteinhäufigkeit oder Netzwerkkonnektivität zu erschließen und nutzbar zu machen. So können Schlussfolgerungen nicht nur auf proteomischer Ebene, sondern auch für individuelle Proteine gezogen werden. Im Rahmen dieser Studie wurden *Arabidopsis thaliana* Wurzelmembranproteine untersucht. Insgesamt konnten 6.928 individuelle Proteine extrahiert, identifiziert und unter verschiedenen Nährstoffverfügbarkeiten (Normalbedingung, Stickstoffmangel und Stickstoffzufuhr nach Mangel) untersucht werden. Die in dieser Studie beschriebene Datenverarbeitungspipeline wurde verwendet, um PPI für mindestens 2058 dieser Proteine vorherzusagen und zu beschreiben. Jeder Schritt der Datenverarbeitung wurde durch den Vergleich mit datenbankbestätigten Interaktionen kontrolliert, um die Filterkriterien zu verbessern. Aus

den vorhergesagten Interaktionen wurde ein Interaktom-Netzwerk der verschiedenen Ernährungszustände rekonstruiert, anhand dessen die Bedeutung der Proteine für den entsprechenden Ernährungszustand durch verschiedene Netzwerkparameter aus der Graphentheorie beschrieben werden konnte. Die Proteinabundanzen wurde durch ein einzigartiges Ranking-System bewertet, was eine nahtlose Integration als Netzwerkattribute für jede Bedingung ermöglichte.

Weiterhin zielte diese Studie darauf ab, das Verständnis des Stickstoff-Signalwegs anhand von *Arabidopsis thaliana* Wurzelmembranen als Modell zu vertiefen. Unter Verwendung des zuvor etablierten Co-Elutions-Workflows, der Korrelationsanalyse und Netzwerkrekonstruktion wurde die Rezeptorkinase AT5G49770 als wichtiger Einflussfaktor bei der Stickstoff-Signalübertragung in Verbindung mit dem Co-Rezeptor QSK1, dem Stickstofftransporter NRT2.1 und der Protonenpumpe AHA2 identifiziert. Die Netzwerkparameter von AT5G49770 zeigten die größten Veränderungen als Reaktion auf Stickstoffmangel. Die Beteiligung wurde durch eine phänotypische Ähnlichkeit zu Knock-out-Linien von NRT2.1, NRT1.1, QSK1 und AHA2 in einer phänotypischen Charakterisierung des Wurzelwachstumstests von *Arabidopsis* Keimlingen gezeigt. Die Interaktion zwischen QSK1 und AT5G49770 wurde weiterhin durch FRET/FLIM-Mikroskopie und Pulldown-Analyse bestätigt.

Die Ergebnisse zeigen, dass die Kombination eines auf Co-Elution basierenden Ansatzes mit einer rigorosen Datenverarbeitungs-Pipeline und der Netzwerkanalyse geeignet ist, um Proteininteraktionsnetzwerke und Signalantwortdynamik in Pflanzenwurzelmembranen zu untersuchen. Das modular konzeptionierte Versuchsdesign ermöglicht eine einfache Anpassung der Methodik zur Untersuchung verschiedener Stimuli, während die unvoreingenommene proteomische Analytik Ergebnisse unabhängig des individuellen wissenschaftlichen Fokus erzielt. Metainformationen wie Proteinabundanz und Netzwerkkonnektivitätsparameter sind geeignet, um Proteine, die maßgeblich an der Dynamik von Stressreaktionen beteiligt sind zu identifizieren.

Der Autor dieser Studie ist zuversichtlich, dass die erzeugten proteomischen Datensätze für weitere Forschung genutzt werden können und zu einem tieferen Verständnis der Stickstoffsignal-Regulierung in pflanzlichen Wurzelmembranen beitragen. Zusätzlich demonstriert diese Studie, dass über die Erschließung von Metadaten wichtige Informationen gewonnen werden können, um in verschiedenen wissenschaftlichen Szenarien einen Erkenntnisgewinn zu ermöglichen. Damit leistet diese Arbeit einen wichtigen Beitrag zur Weiterentwicklung der proteomischen Analytik und Datenauswertung.

Preface

Life emerged from cooperation.

Throughout history, scientists have debated the nature and origin of life. When discussing life as a metaphysical concept, various hypotheses about the origin and shape of living matter, as well as equally diverse explanations for the origin and shape of this universe are considered. May it be a god who creates by giving name, or the scientist in search of the indivisible, smallest particle of matter, the common denominator of those theorems is, that the world is made up from things and that those adhere to a set of rules, which allows the world to unfold through their dynamic interaction.

As the Age of Enlightenment fueled the flames of scientific progress in Europe from the 17th to the 19th centuries, our perception of life as an attribute bestowed by a creator shifted to the realization of life as an abstract concept in need of investigation. In 1749, French naturalist Georges-Louis Buffon first described “organic molecules” that constitute all organisms. In his concept, species were a collection of different organic molecules and different species are the result of an accumulation of different molecules through time [1]. The realization, that life is not only driven by small components but also a process of historic development was a novel concept which was later confirmed by Charles Darwin’s famous observations in “The Origin of Species” in 1859 and termed “evolution”. He showed, that the interaction of different species of birds with their environment induced dynamic hereditary changes over time. Yet, while evolution is a term coined in the context of hereditary changes only, organisms adapt their cellular constituents within seconds, minutes or hours in order to react to sudden changes in the environment. While the original idea of Buffon, that different organisms consist of different “molecules”, is true in essence only, we know now, that diversity comes not solely from the constituents of life, but also from their dynamic (and time-dependend) interaction with each other. This is evident, since our technical ability to discern the constituents of life increased spectacularly over the recent century without simplifying our descriptive models.

The foundation of cellular life rests on three pillars: DNA, which holds the hereditary information of the others, proteins as the working backbone of cellular function, facilitating chemical reactions and constructive or destructive processes and RNA as a link between that abstract layers of information and physicality, from which the other two emerged. Each pillar – DNA, RNA and proteins – is governed by a unique set of mechanisms for adaptation and regulation, while each exerts influence on each other.

Genes can be recombined during sexual reproduction, activation of transposable elements or altered by malfunction of repair mechanisms. They can also be repressed through physical occlusion by heterochromatization or direct chemical methylation.

RNA function is mainly regulated through its lifetime which is dependent on terminal modification or binding through complimentary RNA. Besides its prominent role as “messenger”-RNA in protein biogenesis, there are more regulatory functions taken by small, processed RNAs. An example would be the function of small RNAs in signal transduction and DNA methylation or the process of RNA-interference [2, 3] where small micro-RNAs (21 nucleotides long) specifically suppress endogenous RNAs or act as a defense mechanism against pathogens [3, 4].

Protein function, on the other hand, is the result of a far more diverse level of modulation. Proteins, at their most basic, are linear polymers of amino acids which are linked by peptide bonds. Through the unique chemical properties and the specific sequential order of amino acid arrangement, polypeptide chains can use disulfide- or water-bridges to fold into a macromolecule with a well-defined three dimensional structure to produce an environment of spatial organization in which the specific

arrangement of atoms can catalyze chemical reactions. A well characterized example is the catalytic triade as found in the reaction center of many proteases [5]. Yet, the three dimensional structure is not fixed and proteins can be present in different conformational states either caused by a changing milieu (e.g. pH-concentration or salinity) or allosteric interaction with other molecules or modification. The influence of the conformational state on enzymatic activity was first shown by Chris Anfinsen in 1957 [6]. However, protein structure enables a far more influential mode of operation than producing an environment for chemical reactivity. With a distinct structure comes the ability for physical protein-protein interaction and complex formation. Further, proteins can be chemically modified to modulate activity (e.g. phosphorylation), stability (e.g. ubiquitinylation) or other properties not primarily tied to protein function like localization or interaction [7].

The ability to recruit multiple proteins for a single complex task gave room for protein specialization to evolve and all processes in cellular biology require functionally specialized proteins and their coordinated interaction or cooperation. Signal perception and transduction, metabolic reactions and replication are all comprised of several sequential steps handled by individual proteins or protein complexes working in concert. Most notable are the protein complexes involved in the harvesting of energy from photons to synthesize sugars utilized by organisms such as algae or plants which are capable of photosynthesis. The light reaction is catalyzed by two major photosystems which are both protein supercomplexes comprising of a central reaction center enveloped by light-harvesting proteins. Photosystem 1 is a protein complex made up from about 15 subunits while Photosystem 2 encompasses around 30 protein subunits [8]. Both are tasked to harvest the energetic potential of light-excited electrons for the generation of a proton-gradient across the thylakoid membrane which is used by ATP-synthases to generate ATP (Adenosine-triphosphate). ATP is the primary energy-transfer currency of living cells and can drive reactions such as to synthesize sugars via carbon fixation. This reaction is catalyzed in part by the Ribulose-1,5-bisphosphat-carboxylase (RuBisCo), which is another mega-complex (eight big and eight small subunits) and supposedly the most abundant protein complex in the world [9].

Knowing the composition of protein-complexes, as well as the determinants of their spatial and temporal formation, is therefore critical and a challenge for modern molecular biology.

Historic timeline of molecular biology

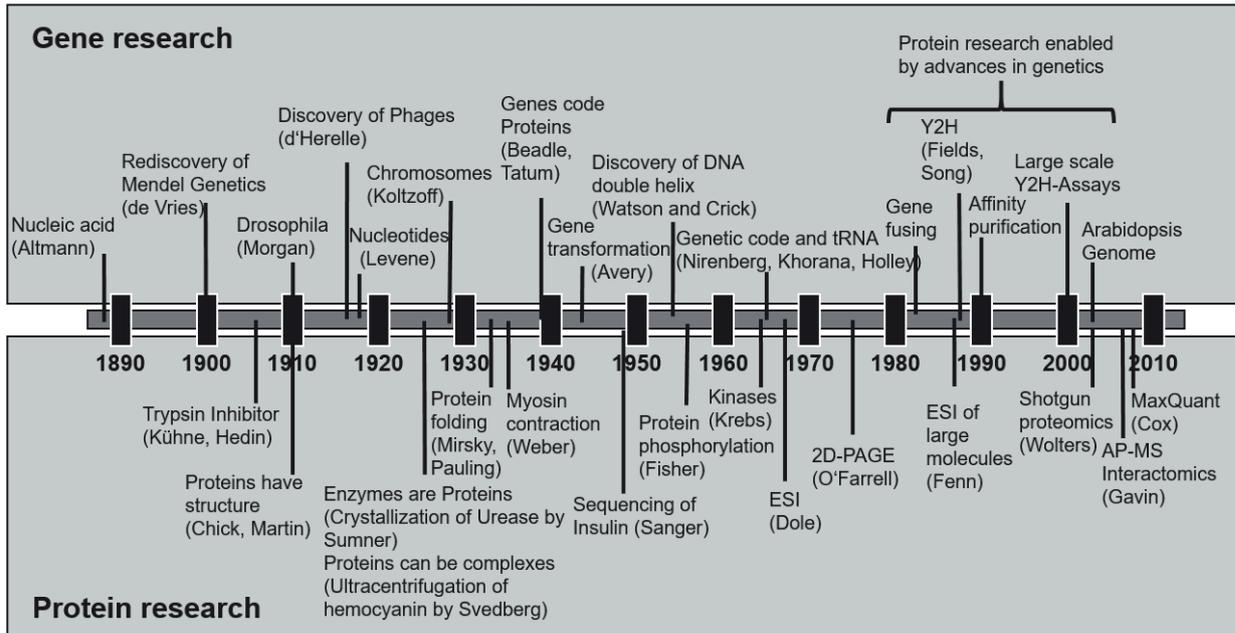


Figure 1: Historic development of molecular biology. Figure modified from Braun and Colleagues: PROTEOMICS, Volume: 12, Issue: 10, Pages: 1478-1498, First published: 19 June 2012, DOI: (10.1002/pmic.201100563) [10]. Additional references [11-39].

Although protein significance is well acknowledged, their chemical complexity delayed their research. A brief timeline of the historic development of modern molecular biology is depicted in Figure 1. The most important discoveries in both DNA and protein research occurred around the same time: DNA was found to contain the hereditary information as nucleotides in a DNA double helix [40] and protein primary structure as a concept was demonstrated in 1952 through the sequencing of insulin by Frederick Sanger [28]. Yet, protein studies were mainly limited to chemical characterization, centrifugation and sedimentation technique and enzymatic descriptions up until advances in genetic methods like molecular cloning or mutant screens enabled conceptual advances. Ironically, although protein interactions were shown before the concept of protein structure was formulated, early large scale protein interactions studies were highly independent of traditional protein research methods and drew conclusions either based on indirect genetic yeast hybrid screenings or tag-purified proteins [41]. A main obstacle to protein interaction research was the difficulty of protein identification – especially of unknown proteins for which no antibodies were readily available. Early identification methods were reliant on manual chemical sequencing (<1960), electrophoresis (<1975) or antibody hybridization (<1980) before the emergence of mass spectrometry based methods (>1990) facilitated routine analysis of whole protein extracts (>2000).

Mass spectrometric methods proved to be the key to open the field of proteomics.

General Introduction

Tandem-Mass spectrometry based proteomics

As the name suggests, the analytical procedure of mass spectrometry results in the determination of the mass to charge ratio of ions. In mass spectrometry, molecules are analyzed as ions in a gas and mass is obtained by observing the behavior of a charged molecule in an electrical field. A key principle of tandem-mass spectrometry is the fragmentation of molecules for identification. Thus, even for a single molecule, a whole spectrum of masses is acquired.

Each mass spectrometer consist of several, vital parts: Ion source, mass filter, ion trap, collision cell and mass analyzer as roughly depicted in Figure 2.

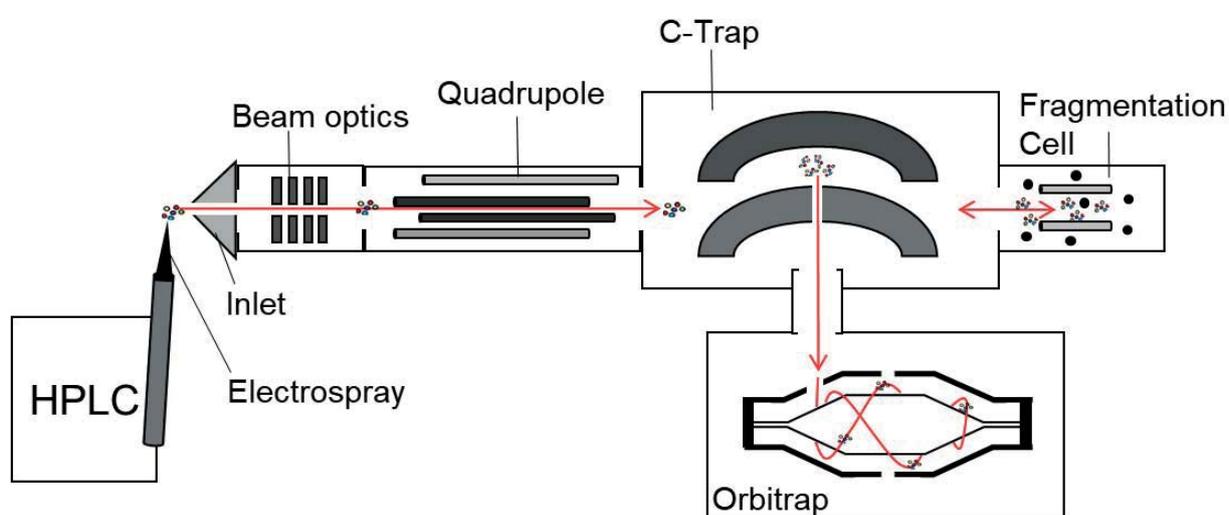


Figure 2: Schematic of a mass spectrometer as used in this study. HPLC-separated peptides are ionized via Electro spray-ionization and enter the mass-spectrometer through a set of beam-focusing optics. Peptide ions are sorted by the Quadrupole mass-selector and are transferred to the C-trap where they are stored for accumulation and transfer. A number of ions can be transferred to the Orbitrap detector unit or subjected to fragmentation in the fragmentation-cell before they are analyzed.

Since neutral molecules cannot interact with the electric field, they need to be ionized first. Ionization thereby refers to the method and process of producing ions. Several methods for ionization are suitable for different purposes. In proteomics, electro-spray-ionization (ESI) as introduced in 1989 by Fenn and colleagues [32] has been established as the predominant ionization method for protein research because it can be coupled to high pressure liquid chromatography (HPLC) and offers soft ionization without fragmentation. This allows presorting of peptides to reduce sample complexity. Reducing the amount of peptides analyzed per cycle leads to a much better sample coverage. In ESI, proteins or peptides are eluted through a thin needle using a strong organic solvent. At the tip of the needle, a strong electric potential is applied which transfers electrical charge to the solvent. As the solvent evaporates, electrical charge is transferred to the molecules which leads to ionization. Because of the high pressure and electrical charge, a thin and consistent stream of ions enter the mass spectrometer at a steady rate [33]. Ionization is facilitated by the use of acidic buffering conditions to maximize protonation and enable transfer of charge.

After ionization, ions enter the mass spectrometer through a set of electrical optic lenses into the first mass analyzer which can be used to analyze or sort the incoming stream of ions. In many devices, this role is filled out by a quadrupole mass analyzer (Figure 2, Quadrupole). In principle, it consists of four metal rods which are linked pairwise. The electrical charge of the rods can be adjusted and by changing the frequency of alternating the electrical field strength in the quadrupole, ions can be pushed out of their beam path to be excluded from analysis. Different ions respond to different frequencies because of their different masses. Quadrupoles are also one of the earliest mass analyzers developed. A simple mass analyzer consists only of a quadrupole and an electron multiplier array which gives a signal, if the ion for which the quadrupole actively filters is present. Other ions of different mass would not pass the mass filter and thereby not give a signal [42, 43].

After passing the quadrupole, the beam path of selected ions can be stopped in a trapping cell to allow for accumulation of a population of ions (Figure 2, C-Trap), or be allowed to further be guided into the mass detector (Figure 2, Orbitrap) or collision cell (Figure 2, HCD cell) directly.

Different types of mass detectors are used. In general, detectors can be divided into beam-type analyzers like Quadrupole- or Time-of-Flight (TOF) analyzers or trap-type analyzers like Fourier-transform ion cyclotron resonance (FT-ICR)- and Orbitrap-analyzers. Beam type analyzers are able to continuously scan ions, while trap-type analyzers analyze a certain ion population at a time. Each type of analyzer has different properties. Key parameters are resolution, accuracy, sensitivity and speed.

Resolution refers to the minimal difference of mass that can be detected by a given analyzer. It is calculated as the resolving power $R = \frac{m}{\Delta m}$, whereas m is the m/z value of a given mass and Δm is the difference of mass according to the full peak width at half-maximum height (FWHM-definition) or at low (5%-10%) peak width (valley-definition). Quadrupoles and TOF-analyzers feature good resolving powers because resolution is dependent of the length of the ions' flight path which can be increased without compromising analyzing speed. The resolution of trap-type analyzers like Orbitrap is potentially unlimited because it depends on time. In principle, Orbitrap analyzers trap ions by coercing them into an orbiting motion around and along a spindle-shaped electrode. The frequency of rotation around the central electrode is mainly affected by initial velocity and radius, while the frequency of oscillation along the axis of the spindle is only dependent of the m/z value [44]. By the ions' movement along the central spindle, currents are induced in the encasing shell which can be detected as a signal and decoded through Fourier transformation. By extending the time for analysis, the resolution is increased, as even smaller differences can be detected.

Accuracy is a metric describing the deviation of an experimentally divined mass to its real mass and is roughly dependent on mass resolution. Accurate mass measurements reduce the amount of ambiguity in data analysis of samples containing molecules with similar elemental formula. This is highly relevant in proteomics, where digested peptides are often similar in mass and identification of modified peptides is reliant on small mass differences that can only be detected with confidence if the estimated mass error is lower than the mass difference introduced by the modification. Mass resolution and mass accuracy are the reciprocal of each other and dependent on the mass range. Accuracy is usually calculated as ppm (parts per million) $ppm = 10^6 \times \frac{\Delta m}{m}$ whereas m is the true m/z value of an ion and Δm is the difference of true and observed mass. Advancements in resolution and accuracy therefore go hand in hand. A key advancement to the enhancement of accuracy was the lock-mass mediated real time calibration of measurement data [45]. During lock-mass calibration, certain ions from the environmental background (e.g. air or HPLC-column material) are admitted to the measurement. Because their masses are known, the measurement of lock-masses can be used for calibration and calculation of the mass-error.

Sensitivity refers to the ability of a mass analyzer to detect ions. Standard TOF-analyzers make use of electron multipliers as detectors which are highly sensitive and can detect single ions. This is a destructive measurement because the energy of the ion is completely transferred into the detector. In contrast, Orbitrap analyzers detect ions through the currents generated by orbiting ions in a non-destructive way. During analysis, ions are kept in an orbiting motion and are only discarded after the analysis is finished. This measurement process is affected by its signal-to-noise ratio which limits sensitivity and requires a population of ions to be present.

As pointed out, scan speed is correlated to resolution in trap-type analyzers which can compromise speed for resolution or vice versa. In TOF-analyzers, scan speed is almost instantaneous since it depends on the ions' travel time through the machine which follows the speed of light.

For proteomics, hybrid instruments are available. In this work, a quadrupole Orbitrap instrument by ThermoFisher Science [46] was used. Quadrupole Orbitrap instruments combine the different advantages of different analyzers. The Quadrupole can supply the Orbitrap with a fast and accurate mass selection, while the Orbitrap analyzer enables the non-destructive analysis of whole populations of ions with high resolution and speed.

Yet another key component for mass-spectrometry based proteomics is the fragmentation of ions. Typically, whole proteins and even peptides are simply too large or offer too many modification sites to be identified through their mass alone. To aid identification, proteins are digested into peptides and peptides are physically broken down in the analysis pipeline. By breaking down ions to fragment ions, different ions of similar mass can be distinguished through their fragment spectrums. Peptide fragmentation can be achieved through different methods, as for example Collision-induced Dissociation (CID) or Electron Capture Dissociation (ECD). In CID, peptide ions are accelerated and collided with an inert gas like Nitrogen which induces fragmentation [47, 48]. During ECD peptides are fragmented by irradiation with low-energy electrons [49]. In both cases, the peptide backbone is broken and ions are formed through internal electron rearrangement and proton capture summarized as the "Mobile Proton Hypothesis" [50, 51]. CID and ECD induce fragmentation in a consistent and predictable manner along the peptide backbone, retaining posttranslational modifications and rarely inducing fragmentation in the aminoacid-sidechain. A nomenclature for naming the resulting peptide fragments was introduced in 1984 by Roepstorff and Fohlman. Fragmentation can occur mainly at three different sites along the peptide backbone: Before the carboxy group, within the peptide bond or after the amino group. After fragmentation, the ions are called A, B, C if the charge is retained at the N-terminus, while they are referred to as X, Y and Z if the charge is retained at the C-terminus. Fragment ions are numbered according to their theoretical position in the amino acid chain. An example for the potential fragmentation of a peptide is given in Figure 3. Although each fragmentation method can produce any fragment ion, B/Y-ions are predominantly produced by CID while C/Z-ions are predominantly produced by ECD.

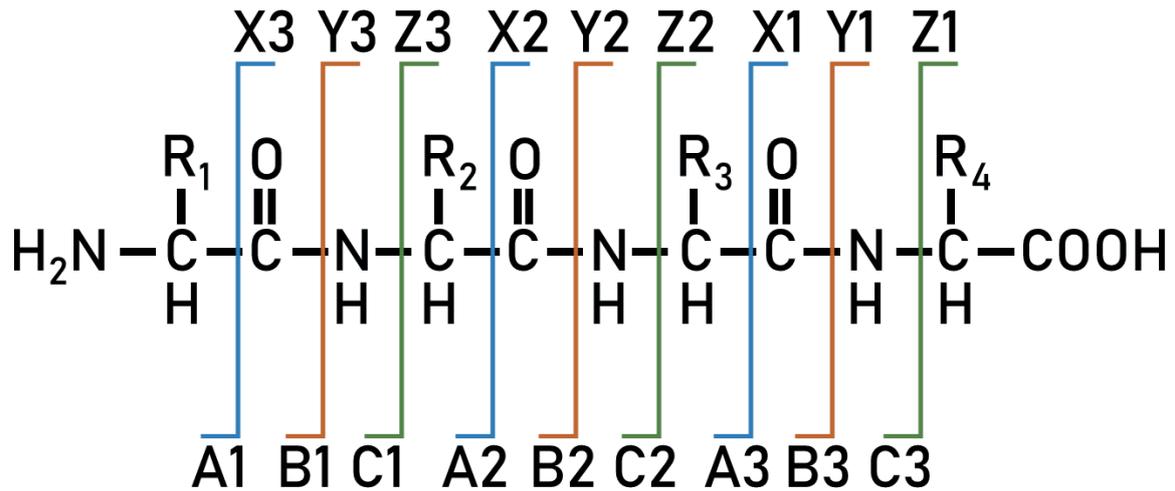


Figure 3: Nomenclature after Roepstorff and Fohlman. Figure after Roepstorff, P. and J. Fohlman, *Biomed Mass Spectrom*, 1984. 11(11): p. 601. [52]. A peptide consisting of four aminoacids (R1-R4) can be broken along the peptide backbone at three different sites with three sequential repetitions along the peptide backbone. The resulting fragments are labeled according to the fragmentation site and positioning of the remaining charge.

By applying stochastics, with a large enough population, every possible fragment ion of a peptide is produced. These fragment ions can then be transferred to the mass analyzer again and resulting mass spectra are used for amino-acid sequencing based on the mass difference of the found fragments. Full scan and fragment scan are often referred to as MS1- and MS2-scan respectively. MS2 information can be used to identify MS1-features during chromatography to enhance identification and quantification rates even across samples [53, 54].

Proteomic workflow

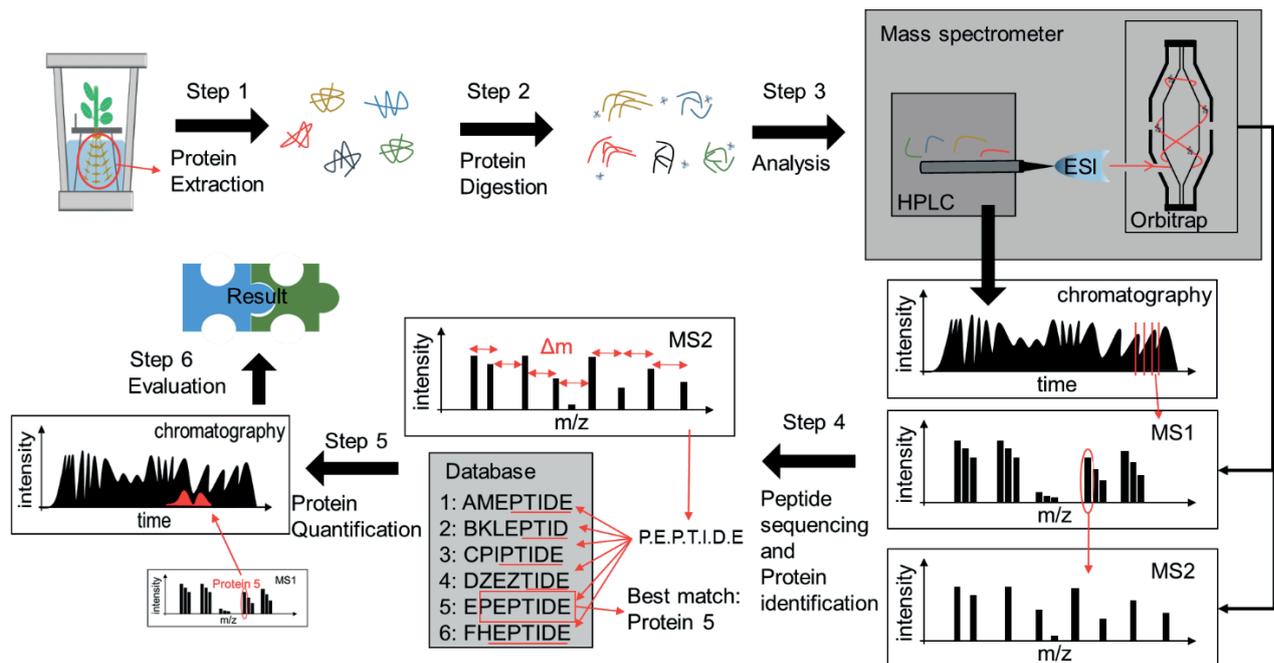


Figure 4: Mass spectrometry enabled proteomic workflow. **Step 1:** Proteins are extracted or enriched from any suitable tissue and/or cellular compartment (in this study: *Arabidopsis thaliana* root membrane fraction). **Step 2:** Proteins are digested with a proteinase e.g. Trypsin which cuts the proteins at specific sites. **Step 3:** Digested peptides are ionized and transferred into a mass spectrometer where the peptide ions are either analyzed directly (MS1) or subjected to fragmentation before mass determination (MS2). **Step 4:** The fragmentation spectra (MS2) of a peptide feature (MS1) can be used to determine the amino-acid sequence by calculating the mass differences of different fragment ions. Protein identity is inferred by referencing found peptides to a database which contains protein amino-acid sequences. **Step 5:** In a label-free quantification approach, identified MS1-features can be used to trace that specific feature during chromatography even if no MS2 information is given. Protein quantity is the integral of the area of the extracted ion current (XIC) and can be the result of a combination of peptide intensities.

The principal problem of proteomics is that the sample consists of many different proteins that often are similar in mass. Top-down analysis of whole proteins faces too many problems with protein ionization and fragmentation of whole proteins result in too complex mass spectra which are not easy to interpret. Therefore, top-down approaches are mainly used with purified proteins.

Instead, the most widespread workflow for proteomics follows a bottom-up or shotgun approach. A schematic of a typical workflow is depicted in Figure 4. After protein extraction (Figure 4, **Step 1**), specific proteases are used to cut proteins into peptides (Figure 4, **Step 2**). In this work, Trypsin – a digestive enzyme widely distributed in nature [55] – is used to cleave the C-terminal side of lysine and arginine residues. Ironically, this increases sample complexity even further. At the same time, protein identification is made possible since proteins with similar masses can be distinguished by their unique peptide composition. It is crucial to use a specific protease with a known cleavage motif as the computational identification relies on known peptide properties. Acquisition of peptide information is in most cases data dependent. In these cases, the mass spectrometer fully scans peptides that enter through HPLC with high resolution to get an overview of all present peptides (Figure 4, **Step 3**). In a following step, a collection of abundant peptides is selected for amino-acid sequencing via fragmentation (Figure 4, **Step 4**). In a data independent approach, all peptides are subjected to fragmentation which poses a challenge to the subsequent computational identification of peptides and proteins. In this work, a data dependent approach was used.

Upon obtaining MS1- and MS2 information, the experimental stage of mass spectrometry based proteomics is concluded and followed by the process of assigning protein identity through database comparison (Figure 4, **Step 5**). Many computational tools were developed to automate the process of spectral matching, peptide sequencing and validation of protein identifications. An overview of available methods and tools for mass spectrometric data analysis may be found in the review of Nesvizhskii et al. 2007 [56].

In principle, protein identity is discerned by comparing the experimentally obtained mass spectra to a database containing information about the calculated mass spectra of proteins (and peptides). This database is usually generated using a general database containing protein sequence information which virtually digested using a set of cleavage rules which conform to the experimental design. Miss(ed)-cleavages or aberrant fragmentation lead to peptide or fragment spectra that are more difficult to match and therefore are mostly discarded as low scoring results in the data analysis. Knowledge about the experimental design is critical for the database construction and *in silico* peptide/protein identification.

Many algorithms follow a probabilistic approach of scoring spectral matches. The main advantage of probabilistic based scoring is that it enables testing to prevent false positive identification events. In probability based scoring, each match between experimental spectrum and processed database entry is considered to be a chance event whereas spectral matches are ranked to a list of peptide matches that is compared to an expectation value. The expectation value can be derived from empirical derived fragmentation frequencies or the assumption that it follows a certain distribution (for example normal- or Poisson distribution). Decoy-databases are used to discriminate true positive matches from false positive matches and to estimate the false discovery rates (FDRs). First, spectra are matched and scored to a true database (“Target”) and then matched to a reversed or randomized version of the same database. (“Decoy”). It is assumed, that the distribution of false positive matches in the true database follows the distribution of matches within the false database. In this way, a FDR-based cutoff score can be used to enhance the accuracy of protein identification [57].

In this work, the MaxQuant software solution [58] was used to identify and quantify proteins present in the samples. The MaxQuant platform utilizes an advanced intensity determination and normalization procedure for identification of chromatographic features across multiple samples before the actual peptide-spectrum matching and subsequent protein identification is done by the integrated Andromeda-search engine [59]. Secondary, auxiliary information like retention time during HPLC separation of peptides can be used to improve peptide identification rate and enable a more reliable protein quantification across similar samples. MS/MS spectra are prepared for matching through centroiding, collapsing charge states, removal of isotope patterns and mass-window based top-n filtering of peaks.

Using peptide sequences to infer protein identity is also a major challenge in itself. Since peptide identification is a probabilistic result, larger proteins or proteins of high abundance produce a higher number of peptides that are identified with a higher chance than their smaller or less abundant counterparts. By increasing the generation of mass-spectra (either by a longer HPLC-gradient or a faster mass spectrometer), less abundant proteins may not necessarily be identified with greater confidence because the most abundant peptides will be measured repetitively. Obtaining a greater number of spectra translates to a higher chance of false positive peptide spectrum matches. Erroneous peptide spectrum matches propagate through the analysis pipeline and lead to an even higher accumulation of false positive protein identifications which would contradict the original intent to gather information about low abundant proteins [60, 61]. The MaxQuant platform features a false discovery rate estimation and correction at the level of peptide matching, protein grouping and posttranslational modification matching [54].

Another challenge is of biological nature. Proteins are the result of evolutionary processes like gene duplication, recombination or deletion. Different proteins of different functions may share peptide sequences to some extent and protein motives are often basis for protein classification. As a result, it is difficult to differentiate between homologous proteins, splicing variants or protein isoforms or other proteins with shared sequences. The identification of a single peptide that is shared within a group of proteins cannot prove the presence of all related proteins in the sample. The identification of a single peptide that is shared amongst a group of proteins proves only the possibility of the presence of the whole group while it cannot possibly confirm any member of that group with certainty. To resolve this paradox, protein grouping follows heuristic principles of parsimony famously dubbed as Occam's razor which states that "entities should not be multiplied without necessity" [62]. Following this maxim, MaxQuant limits the number of assumed protein identifications to the smallest set of proteins that can account for the identified peptides. Similarly, a peptide is only used for the identification of one protein. Multiple findings of the same peptides cannot prove the existence of different proteins and are assigned to the same protein or protein group. Therefore, protein identification is not so much dependent on sequence coverage or peptide count, but uniqueness of its peptide sequence.

For protein quantification, information gained from MS/MS-analysis can be used, too. Label-free quantitative proteomic approaches can be distinguished by the level of MS-information that is used for quantification. A straightforward method of protein quantification is to summarize MS2 spectra of each protein to determine the absolute protein quantity as the number of spectral counts in a sample [63]. This method is accurate for proteins of high abundance or those that generate a lot of peptides.

Another method is to analyze auxiliary information like chromatographic feature length or peak area to enable a better quantification of low abundant proteins. This is commonly termed "Extracted Ion Current" or "Extracted Ion Chromatogram" (XIC) and the main advantage is, that quantification is accomplished through evaluation of a chromatographic feature that consists of MS1-information. This higher-order information is more reliable because a weak MS/MS-result can be compensated to some extent. Here, MS2 information is only used for identification of the MS1-level. As long as the confirmed MS1-feature can be detected, a quantification is possible without the need for additional MS2 information. Yet, it is limited by sample complexity and chromatographic density. In modern high resolution mass-spectrometers, co-eluting peptides can be easily distinguished. Thus XIC is well suited for comparing proteins of low and high abundance. With the integrated quantification algorithm that makes use of XIC, MaxQuant enables a relative protein quantification even across samples by using a unique normalization approach dubbed MaxLFQ [53].

Label-free approaches for quantification of proteins are beneficial for most experimental designs since it is not required to subject the sample to various labeling procedures. Almost any kind of samples can be used while quantification results still remain within a window of reasonable accuracy [64]. For a proteomic systems biology approach, these advances in protein identification and quantification are most useful.

Proteomic Systems Biology

The classical approach of molecular proteomics focuses on the comprehensive characterization of proteins, their abundance and distribution in a sample and their posttranslational modifications. The dedicated goal of systems biology is to take a more integrative approach and combine a plethora of descriptive data with protein protein interaction (PPI) data to see the contextual relationship of all proteins as a network. In this network paradigm [65], players in a biological context (e.g. genes or proteins) are represented as nodes in a network, while the edge between nodes represents a relationship such as an interaction or enzymatic reaction. Such information is static and thereby only representative for the baseline conditions under which the experiments were performed. A full characterization of a phenotype or organism can be achieved by integrating primary connectivity information with secondary information like protein abundance change to describe the dynamic reaction of a network to perturbation. To combine different datasets- or types is the primary challenge of the systems biology approach and requires creative experimental design and the use of novel, computationally supported methods [65].

In case of model plant *Arabidopsis thaliana*, the *Arabidopsis* Information Resource TAIR [66] reported 27.655 protein coding genes, 952 pseudogenes, 3.901 transposable element genes and 5.178 non-coding genes with eighteen percent of *Arabidopsis* genes showing confirmed splice variants [67] coding for 39.345 estimated proteins [68]. The hundreds of millions of possible PPI combinations makes it impossible to research protein interaction networks on a pairwise basis. Hence, most proteomic studies heavily rely on mass spectrometric enabled methods which are advanced enough to provide a nearly full proteome coverage [69]. Advances in mass accuracy, sensitivity and development of new mass analyzers like Orbitrap [44] led to the development robust of workflows for label free quantification [53]. Comparative protein quantification is especially valuable for a systems biological approach as transitions from one state to another are seldom as simple as “on” and “off”. Often, adaptation is the adjustment of equilibria. Prominent examples are the multiple feedback loops which regulate circadian processes in *Arabidopsis* [70]. Multiple antagonizing components are regulating each other to produce an oscillating environment of gene expression. Only through external stimuli such as light, the balance of the equilibrium is tipped and a reaction is achieved.

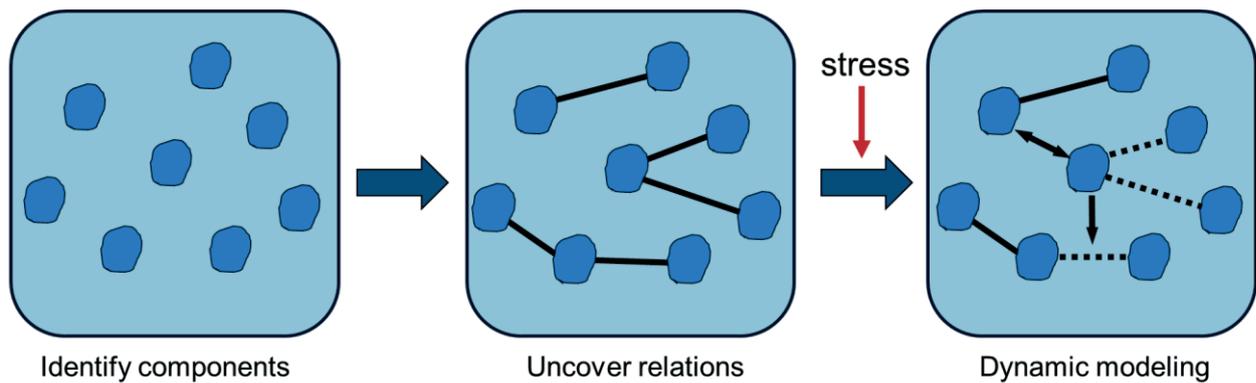


Figure 5: Proteomic systems biology workflow: A typical proteomic workflow in systems biology, adapted from Bensimon and colleagues [65]. To describe a biological system, both, the components that constitute that system and their relations need to be uncovered before the introduction of stress enables dynamic modeling of a phenotype.

A typical workflow of proteomic systems biology is following three steps as depicted in Figure 5. Along each step, different methodology may be employed. First, the components of a biological system must be explored. The context dictates the methods used to find components. As example, on the organism scale, proteomic components might be found through whole genome sequencing, while tissue specific components might be identified by tissue specific protein purification and mass spectrometry. Second, relations between proteomic components might be uncovered through an array of different methods as for example yeast-hybrid methods, affinity purification or co-elution methods. Lastly, to describe the dynamics of a system, a perturbation has to be introduced and the resulting changes need to be observed. The nature of the perturbation is dependent on the experimental context and can be as little as a loss-of-function mutation of a single constituent, or the introduction of environmental stresses like nutritional deficiency on a larger scale. At any time, information gained from multiple sources can be combined to complement the descriptive model. This also poses a sizable challenge since data from different sources may not necessarily be compatible or may only be valid in the specific context of the experiments.

When presented with a map of constituents and knowledge about their relations, a graph of a biological system may be drawn as a network of interaction partners. The analysis of graphs is a subfield of mathematics introduced by Euler in 1736 and advanced by Paul Erdős and Alfred Renyi in their hallmark study on random graphs in 1959 which paved the way for systematic analysis of networks and their topological features [71]. To analyze networks, the original data of constitutes (nodes or vertices) and connections (edges) need to be addressed and multiple layers of information can be described by a network. Information from protein-protein interactions can be described as an undirected network, while enzymatic reactions or signaling pathways are of directed nature. A network can combine both types of data which illustrates the integrative approach of network biology to paint a whole picture.

For the purpose of describing network structure (e.g. modularity or clusterization) and functional relevance of nodes and edges, different strategies have been suggested. A feature of biological networks that distinguishes them from random networks is their organized modularity [72, 73] and adherence to a scale-free topology [74]. This means, in comparison to random networks, biological networks tend to show a topology in which many constituents show few connections and few constituents show many connections. Instead of an even edge distribution across the whole network, biological networks tend to be organized in local clusters that are connected via key proteins. These properties are rooted in fundamental evolutionary processes like gene duplication and divergence [75] and mutation [76]. Still, the nature of real world network properties is up for debate as it is shown, that random sub-sample-networks of large artificial networks can also display a scale-free topology [77].

In principal, network parameters are founded in the intuitive notions of node connectivity and a node's influence is primarily defined by its direct or indirect neighborhood. To describe node significance, different centrality indices have been introduced for the use in biological networks [78] such as Degree, Shortest Path, Stress, Closeness Centrality, Betweenness Centrality, Clustering Coefficient and Topological Coefficient (see Figure 6). These indices can be distinguished by the metric which they are based on.

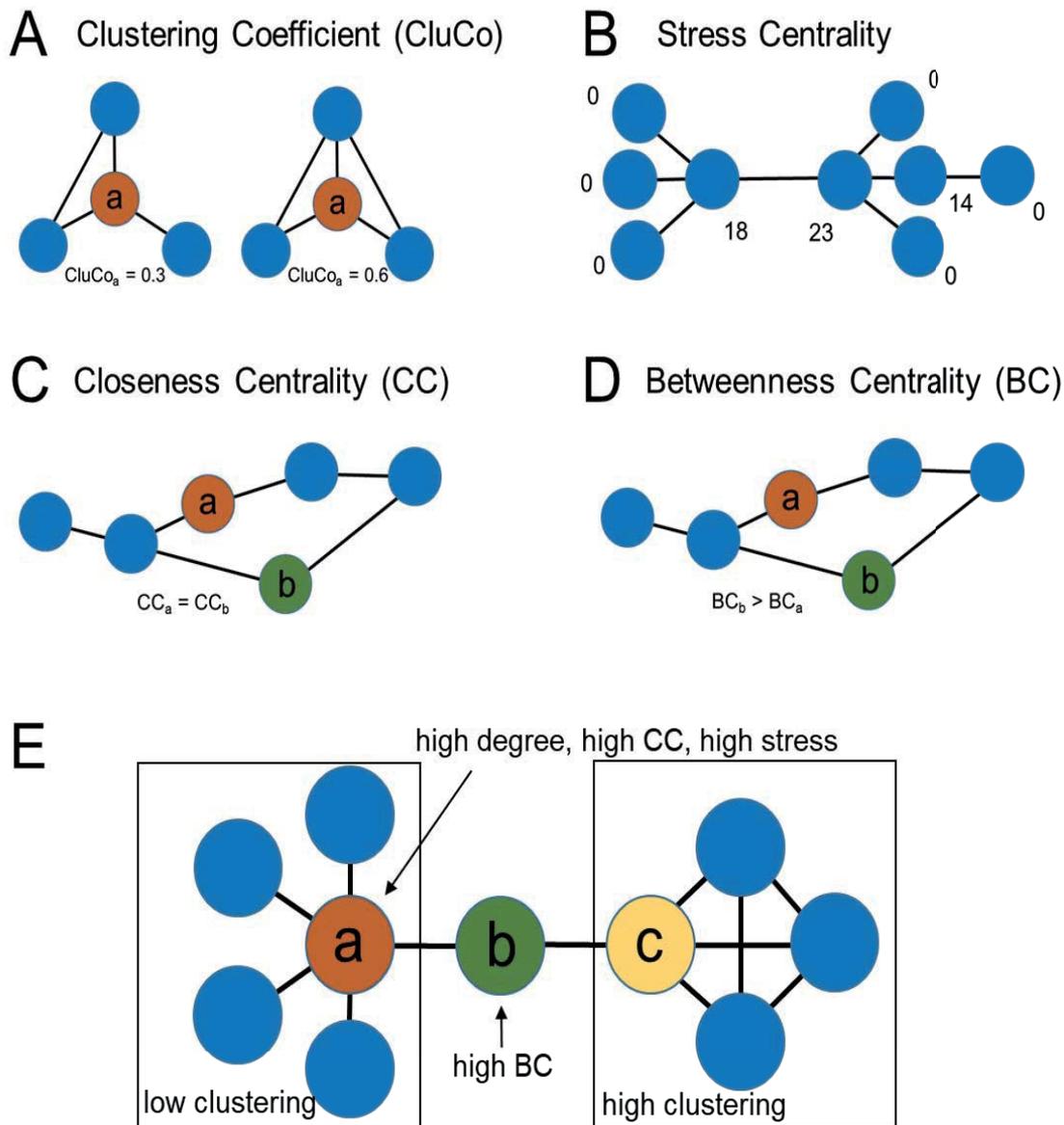


Figure 6: Illustration of several network parameter concepts: **A:** Clustering Coefficient (CluCo) is the quotient of the sum of edges between a node's neighbors and all possible edges between its neighbors; **B:** Stress Centrality is the sum of all shortest paths leading through a node; **C:** Closeness Centrality (CC) is the quotient of the count of network members and the sum of all shortest paths of a node; **D:** Betweenness Centrality (BC) is the sum of the quotients of the number of shortest paths leading through a node and all available shortest paths of a given node pairings; **E:** Example network indicating different network parameter properties.

Degree metrics are based on the number of edges linked to a node or its neighborhood, whereas path metrics describe node distance which is the number of edges between two nodes. The general assumption in network analysis is, that a node which lies within the center of a network, is also of high importance. Yet, the definition of a center varies depending on the network parameter used to describe centrality. With degree based metrics, the concept of centrality is that a node with many edges is more relevant to the network because it can easily access every part. Path based metrics try to describe a nodes influence on the flow of information through describing the position of a node and the distance to others. Although they are founded in similar metrics, there is little redundancy. Figure 6 exemplifies a selection of network parameters, while a detailed explanation is given in the second included publication. Network parameter definition accommodate the general idea that important proteins are clustered or stressed. Clustering is thereby described as the number of neighbors in a shared interaction (Figure 6, **A**) and stress is defined as the number of path accessions (Figure 6, **B**). Centrality indices describe importance based on the position of the node in the network and the distance to other nodes. For example, Closeness Centrality (CC, Figure 6, **C**) and Betweenness Centrality (BC, Figure 6, **D**) are both dependent on shortest path lengths and network size. As shown in Figure 6, **C** and **D**, while both nodes a and b have equal short access to the whole network (same CC-value), the BC value for node b is higher because it is positioned so that it lays between a higher number of shortest paths than node a. CC is thereby a measure to determine how quickly a node can distribute information while BC is an indicator for nodes that connect different network communities. Network parameters, especially when used in conjunction, can describe network topology. In Figure 6, **E**, both, nodes a and c feature a similar high degree but the direct neighborhood of node c shows a much higher clustering coefficient. Therefore, two distinct communities might be identified.

Identifying vital nodes to the network largely depends on the criteria which are used to define importance. In regard to biological networks, signaling pathways are often strictly regulated by a key regulator protein which controls signal transduction. Therefore, focusing on path-based centrality measures is appropriate since they can be used to define the nodes' ability to exert influence on the flow of information.

Protein interaction

Before the introduction of mass spectroscopy and discovery proteomics to network biology, protein interactions were mainly studied with genetic, immunological or affinity based methods.

Many large-scale protein-interaction studies in the early years 2000 were performed using the Yeast-2-Hybrid system (Y2H). It makes use of the genetic system of *Saccharomyces cerevisiae* [35]. To test for interaction, proteins are either fused to the DNA-binding domain (bait) or transcriptional activation domain (prey) of a transcription factor. Through interaction, the domains are recombined and transcription factor activity is restored. A restored transcription factor can drive the expression of a corresponding reporter-gene (e.g. beta-galactosidase) which results in a reaction that can be visually evaluated. Y2H methodology quickly gained popularity because of its modular design. DNA-binding domain and activator domain of the split transcription factor are available on different plasmids so that they can be easily used to construct the desired fusion proteins. Once a fusion protein containing plasmid is constructed, it can be combined with every complementary plasmid in an array based approach. A bait protein can even be screened against a cDNA prey library to find unknown interaction partners. In this way, both, individual protein interactions and entire interactomes can be assessed. For the yeast proteome, an array size of up to 6000 proteins was achieved [79]. The Y2H methodology was also adapted to test insoluble membrane proteins via split-ubiquitin assay [80] which utilizes the release of a linked transcription factor after reconstitution of a split ubiquitin after PPI. Characterization of the activation or binding domains of transcription factors is also feasible [81]. Advantages of this method clearly lie in the possibility for automated high-throughput testing. However, in classic methods, membrane proteins are under-represented.

Immunological methods for protein-protein interaction analysis make use of the particular function of antibodies to bind specific antigens [82]. Antibodies can be used to identify target proteins within a complex sample. Proteins are usually separated and transferred onto analytical membranes by electrophoresis and electro-blotting. Subsequent treatment with protein specific antibodies and secondary, antibody specific antibodies can reveal the presence of proteins and their partners. Upscaling immunological assays has also been done on a commercial basis because of their ease of use and rather low cost. For protein-protein interaction analysis though, pull-down and affinity based methods surpassed the classic blotting techniques since they can be integrated into mass-spectrometry based workflows more easily. Instead of membranes, antibodies are immobilized on small beads which can be mixed with the sample to bind the bait proteins and all primary and secondary interaction partners. Because of the high affinity of the proteins to their partners and the antibody, non-binding proteins can be washed of the beads. After solubilization, the identification of all interaction partners can be done with mass-spectrometric methods. Over time, the use of antibodies for purification of proteins and their complexes was made obsolete through the use of tandem affinity tags in which the protein of interest itself is immobilized. For examples, bait proteins can be fused to a streptavidin binding or protein G tag with a specific protease cleavage site as a spacer. This allows for efficient protein enrichment under non-denaturing conditions. Potential interaction partners can thereby be pulled down from a mixture of proteins if their association to the bait protein is strong enough to endure different, stringent washing steps. These affinity based purification methods can be up-scaled to proteome level through library cloning and potential interaction partners can be identified by mass spectrometry.

Despite their huge success and wide application, yeast 2 hybrid or affinity based methods for protein interaction screening are not without shortcomings. In large, these traditional methods are either using heterologous expression systems or in-vitro reactions. This can lead to false positive or false negative results. First, the construction of plasmids or the success of transformation and expression of proteins in

a heterologous system depends on protein eligibility. Some proteins might not be expressed due to differing codon usage, have toxic effects on the host organism or might fold incorrectly without supporting chaperone systems. Even with well constituted proteins, interactions could remain undetected due to missing or incorrect protein modifications that cannot be provided by host organisms or *in-vitro* settings. This leads to false-negatives. False positive results can occur by nonspecific activation of the reporter gene through the bait protein or unspecific binding of antibodies or bead substrates. Detected interactions in heterologous systems might be of non-biological relevance because the proteins might not be localized in the same compartments or expressed in different tissues or developmental stages *in vivo*. Yet most challenging are scale and scope. Scale is the overwhelming number of proteins and scope is the type of information gained. With a proteome of more than 30 thousand proteins in *Arabidopsis thaliana*, array based methods are at limit. It is reported, that high-throughput Y2H-datasets are predicted to contain false positive results for more than half of the data provided [83]. Likewise, information gained from such experiments is bereft of context. The scope is limited to those interactions that perform well under these particular experimental conditions. Since a lot of interactions are conditional and only occur in specific cell types or environments, even a combination of different methods may not fully reflect a phenotype. This is not primarily because of the limitations of current techniques, but also by the definition of what a phenotype is. Although strong key-interactions exist, most processes are not static and a multitude of partners modulate the cells processes and responses to external or internal stimuli. This concert of players cannot be adequately described using methods that provide binary (on – off) information taken out of the cellular context.

To that end – describing the interplay of partners in planta – new methodical approaches are necessary as for example protein network analysis.

Although protein interaction network analysis is on the verge of becoming more and more relevant and several efforts to compile large-scale protein interaction networks among different species have been made [84], proteomic coverage of protein interaction is still incomplete - even in yeast and human [85].

To fully understand cellular networks and machinery, it is the dedicated aim of this work to introduce a systemic method to protein-interaction research in the context of cellular membranes and changing nutritional conditions. It builds on a classic, chromatographic approach and - through the use of mass spectrometric analysis - expands it to proteome level.

Relevance of protein interactions in plants and Nitrogen response

In plants, just as in all organisms, perception of the environment, signaling pathways and adaptation to the environment is mediated through protein interactions. Secondary messengers like phytohormones or reactive oxygen species induce and are elicited by PPI in order to coordinate a physiological response on a systemic scale. As environmental conditions can change with great speed, adaptation cannot rely on slow genomic changes per se. A response to sudden environmental perturbation (e.g. herbivore attack or changes in light intensity) has to be enacted accordingly fast. This holds true, even for less obvious examples like root branching where a localized nutrient availability and nutrient perception is the driving force for root development complementing signals of the plants systemic status [86]. Plant fitness, which is the ability to thrive in an environment, is also of economic and ecological importance to modern day society. The main efforts to improve plant fitness, crop yield or food security are through selective breeding, hybrid crossings and trait selection. Those efforts focus on the genetic heritage and produce results without the need for an explanation of the underlying molecular mechanisms. Yet, recent advancements in targeted gene editing (eg. Crisper/Cas9 system) [87] allow for a more precise fine-tuning and thereby offer an even greater margin of return and greater influence on the desired effects. These approaches are only limited by our understanding of the phenotype and mechanisms of plant physiology. Thus, promoting the knowledge of the role of protein interactions and protein complex dynamics – which bear the brunt of the initial response to the environment – proves to be critical for finding a solution to the problems faced by the global community. To that purpose, scientists make use of plant model organisms since many underlying molecular mechanisms are conserved and shared across species.

Arabidopsis thaliana (“Thale cress”) is used as a model organism to study dicot plant genetics, evolution and development as reviewed by Woodward and colleagues [88]. In comparison to maize and tobacco, the rapid and widespread scientific use of *Arabidopsis* is not driven by its economic, nutritional or ecologic significance, but because it is well suited for molecular biological studies. Seeds and seedlings have a high germination and survival rate, grow small and fast and a generational cycle is completed in less than 2 months’ time. It was one of the first plants used for full genome sequencing in one of the first breakthrough whole-genome sequencing efforts by the *Arabidopsis* genome initiative, promoting international collaboration [89]. The genome annotation is extensive and updated regularly [90]. A vast resource of T-DNA insertion lines from over 225.000 insertion events compiled by Alonso and colleagues in 2003 [91] are available to enable genome wide knock-out analysis just three years after genome sequencing. Likewise, *Arabidopsis* was opened as a platform for functional genome analysis by researchers providing cDNA libraries [92], expression vectors systems [93] and two-hybrid vectors [94]. In recent development, *Arabidopsis* gene expression was mapped in great detail by incorporating advanced RNA-seq profiling methods [95], providing insights for genes without functional annotation. Recently, signaling processes like peptide hormones [96] or post-translational modification of proteins [97] are becoming more relevant in the scientific community.

Arabidopsis is also a suitable platform to test the effect of various external stimuli such as nitrogen availability. Nitrogen is a limiting factor in plant growth and does not only serve as nutrient, but also as a checkpoint as for example the control of developmental phase transition [98] or flowering time [99, 100]. Nitrogen is sensed as nitrate (NO₃) or ammonium (NH₄) in plant roots where it is taken up and assimilated. Since nitrate assimilation produces the toxic nitrite as intermediary product, nitrate uptake has to be tightly controlled and adjusted to not exceed demand. A detailed overview of the nitrate assimilation pathway and its regulation may be found for example in the reviews of Xu and Krapp [101, 102]. Nitrate is taken up and sensed at the root cell membrane by dual purpose nitrate transporters of the

NRT1 low affinity and NRT2 high affinity families [103]. It is known that these nitrogen transporters are modulated in their affinity and throughput by posttranslational modification [104] and serve a dual purpose as nitrogen sensors. Transporter activity regulation is well linked to several phosphorylation sites and corresponding kinases are characterized [105]. Moreover, the involvement of several auxiliary proteins is required for regular NRT function. For example, NRT2.1 interactor NAR2.1 was found by extensive transcriptomic analysis and is shown to promote NRT2 protein targeting, protein stability and transport activity [106]. Yet, besides kinase interactions that modify NRT function, little is known about their integration in the web protein interaction or association that facilitates crosstalk of signaling response.

Research Question

As established, protein interactions are crucial for cellular function. Although the role of PPi is acknowledged, information about dynamic changes of interaction networks is scarce. These dynamic changes can reveal the mechanisms of the immediate response to a changing nitrogen availability in plants.

It is the dedicated goal of this work to investigate protein-protein interaction *networks* in a natural context and to describe the dynamic signal response and identify key players involved in nitrogen signaling with minimal experimental effort. To this end, the primary methodology relies on the combination of native size-exclusion chromatography and mass spectrometric analysis.

Method establishment

Size exclusion chromatography (or gel filtration chromatography, short: SEC) was extensively used to characterize different macromolecules and polymers in early molecular biology research [107]. The first research goal was to adapt this methodology to a proteomic context. SEC uses superficially porous particles (e.g. methyl methacrylate gels or separeose, silicone or polystyrene particles) that slow down the movement of small molecules in a fluid phase. The performance depends on average pore size of the packing and relative size of macromolecules. In contrast to high-performance ion exchange or reverse phase chromatography, the mechanism of separation is obstructed diffusion. Therefore no direct interaction of the macromolecules with the column material is required. Conditions for an optimal obstructed diffusion are achieved with both, a high column volume and low pressure. Coincidentally this enables buffering conditions that allow proteins to remain in their active state since no high-ionic strength buffers or strong organic or inorganic solvents are required. This also allows the use of mild detergents like n-dodecyl maltoside that further preserve membrane protein constitution and function [108], since detergents do not interfere with the protein column interaction. With the active state and protein folding preserved, protein interaction can supposedly occur naturally even in an enriched sample during chromatography.

SEC was often used in the context of protein characterization, especially of single proteins or single purified complexes, where protein identification is of minor concern and constituents can be discerned using in-line UV-VIS spectroscopy. Adapting SEC for the analysis of protein complexes in a proteomic context is a challenge because of the quantity of data obtained. While mass-spectroscopic analysis of the elution fractions was following standard procedure, the primary goal of this study was to conceptualize a data processing strategy to draw meaning from the wealth of chromatographic data.

The data processing pipeline was built around the assumption that during SEC, proteins elute according to their molecular weight. An interaction between proteins would have a severe impact on the expected retention of complexes and thus, a multi protein complex would have an elevated retention speed because of its bigger size compared to the individual proteins.

To determine the expected retention times of proteins in SEC, molecular weight can be used as an easy and convenient substitute for protein size. Using a mix of know proteins for calibration, expected retention times for the whole proteome of *Arabidopsis thaliana* were calculated through a calibration curve.

After chromatography, it is critical to use mass spectrometry for identification and quantification of proteins in each chromatographic fraction, obtaining information about chromatographic peaks and deviance for each individual protein. Using other methods for protein identification is not feasible because of the proteomic scope and the requirement for accurate protein quantification. Proteins that showed a high deviation in their retention time were considered to be complex forming proteins. By comparing the time point of elution peaks of multiple complex forming proteins, proteins that participate in the same complex might be identified. In this work, similarity of retention time profiles is calculated following a Pearson Correlation analysis [109]. For this data analysis workflow, several methods of data filtering and refinement have been employed as described in the first research paper submitted to this work.

In short, fifty dynamic chromatographic fractions were analyzed. Proteins showing a greater than six fraction shift towards an earlier elution time (= increase in size) were considered to be complex forming. For protein correlation analysis, pairwise data was taken from the dataspace defined by the larger protein to exclude positive correlation results based on coincidence and a sufficient overlap of data points for each correlation event was required. An important benchmark for improving data processing efficiency and fidelity is published data of known protein interactions and complexes. Using pre-existing knowledge as reference, data filters and conditions for correlation were adjusted to improve the prediction of proteins and their associations. Thus, criteria for the definition of “high confidence”- and “low confidence” interactions were formulated to prospect the likelihood of a successful complex prediction. Through this combination of mass-shift- and correlation analysis, protein interactions and associations were predicted on a proteomic sample.

Application to Nitrogen dynamics

With the refined methodology of protein complex prediction, it was the second dedicated goal of this study to apply this method in the context of Arabidopsis plants under different nutritional conditions to identify key players involved in stress response. In this context, the response to nitrogen deprivation and short-term resupply was investigated as nitrogen is a key macronutrient for plant growth and development. Most plants are reliant on external nitrogen supply and have to adapt dynamically to changes of nitrogen availability. Wildtype Arabidopsis thaliana plants grown under full nutrition served to produce a reference-style dataset to investigate conditional changes of the proteomic network from plants grown under a different nutritional status that were nitrogen starvation and nitrogen resupply conditions. For each growth condition, a protein network was reconstructed containing proteins and interactions that meet the criteria for a predicted interaction.

Comparing those networks with the reference dataset, it was the goal to identify proteins involved stress induced perturbations and observe dynamic changes of the plant interactome. Several methods of network analysis were tested to find proteins regulating or responding to the signal as outlined in the second included publication. Network analysis made heavy use of auxiliary information like network connectivity and protein abundance to assign importance to individual proteins. Those proteins for which an important role was inferred were subjected to further experiments to confirm their involvement.

First Publication

Author contribution disclosure:

Max Gilbert: performed the proteomics experiments, data analysis, network analysis, conceived the study and wrote the manuscript

Prof. Waltraud Schulze: conceived the study, was involved in data interpretation and wrote the manuscript

Reprinted with the kind permission of the American Chemical Society and the Journal of Proteome Research.

Original publication:

Gilbert, M. and W.X. Schulze, *Global Identification of Protein Complexes within the Membrane Proteome of Arabidopsis Roots Using a SEC-MS Approach*. J Proteome Res, 2019. **18**(1): p. 107-119. [110].

 31.5.2021

Signature, Date

Prof. Waltraud Schulze; Supervisor

, 17.05.2021

Signature, Date

Max Gilbert; Researcher

Global Identification of Protein Complexes within the Membrane Proteome of *Arabidopsis* Roots Using a SEC-MS Approach

Max Gilbert and Waltraud X. Schulze*¹

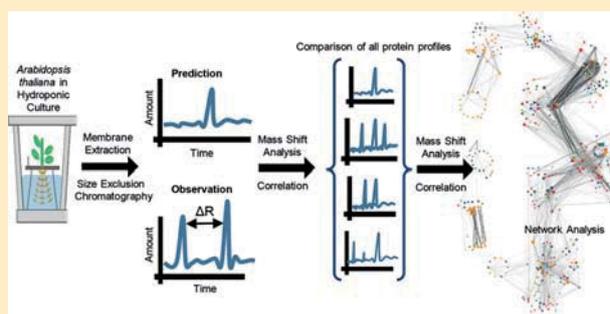
Department of Plant Systems Biology, Universität Hohenheim, 70593 Stuttgart, Germany

Supporting Information

ABSTRACT: Biological processes consist of several consecutive and interacting steps as, for example, in signal transduction cascades or metabolic reaction chains. These processes are regulated by protein–protein interactions and the formation of larger protein complexes, which also occur within biological membranes. To gain a large-scale overview of complex-forming proteins and the composition of such complexes within the cellular membranes of *Arabidopsis* roots, we use the combination of size-exclusion chromatography and mass spectrometry. First, we identified complex-forming proteins by a retention shift analysis relative to expected retention times of monomeric proteins during size-exclusion chromatography.

In a second step we predicted complex composition through pairwise correlation of elution profiles. As result we present an interactome of 963 proteins within cellular membranes of *Arabidopsis* roots. Identification of complex-forming proteins was highly robust between two independently grown root proteomes. The protein complex composition derived from pairwise correlations of coeluting proteins reproducibly identified stable protein complexes (ribosomes, proteasome, mitochondrial respiratory chain supercomplexes) but showed higher variance between replicates regarding transient interactions (e.g., interactions with kinases) within membrane protein complexes.

KEYWORDS: protein complex, interactome, membrane proteins, size-exclusion chromatography, protein mass spectrometry, root, *Arabidopsis thaliana*



INTRODUCTION

Within a cellular organism, all life-sustaining processes depend on functional proteins. Different protein functions during metabolism, defense, storage, and reproduction require specialization and coordinated interactions. Biological processes therefore consist of several consecutive and interacting steps as for example in signal transduction cascades or metabolic reaction chains. Protein–protein interactions and the formation of larger protein complexes are the rule rather than the exception. Therefore, knowledge of the composition of protein complexes and—in particular—their dynamics is of high importance if we want to understand cellular functions.¹

The analysis of proteins and their interactions has a long history in biochemical sciences.² Since the discovery of the first protein–protein interactions of the trypsin inhibitor in 1906, and the introduction of paper chromatography in the early 1940s, a vast array of methods for the analysis of protein–protein interaction was developed involving centrifugation and chromatographic strategies. Two of the most prominent and widely used techniques for the systematic and targeted identification of protein interaction partners are the yeast-two-hybrid system and the affinity purification of proteins. Both methods^{3,4} are now widely used in large-scale biology. Systematic large-scale yeast two hybrid studies were first carried out to characterize the interactome of *Saccharomyces cerevisiae*^{5–8} and became the standard large-scale

method for establishment of systematic interaction maps also within the plant proteome.⁹ With the development of the split-ubiquitin system,¹⁰ the yeast-two hybrid system became applicable also to large-scale interaction studies of membrane proteins.¹¹ However, at least for plant proteins, yeast two hybrid studies remain to be performed in a heterologous system. While yeast two hybrid assays allow high throughput and automated testing of 1000s of proteins, the result remains a pairwise interaction map. In contrast, affinity purifications enable the study of proteins in vivo and are often applied in combination with mass spectrometric analysis of proteins, and were initially used to analyze single protein complexes. More recently, affinity purifications were also applied in systematic large-scale studies¹² or to characterize specific cellular processes, such as the cell cycle.¹³ Affinity purifications enable the detection of more than pairwise interactions but are rather limited in the number of bait proteins tested.¹⁴

A comparative meta-analysis of high-throughput interactomes from yeast revealed that each method is biased toward different functional categories or subcellular compartments, and the overlap of identified interactions by more than one study did not cover more than 60% of the yeast proteome.¹⁵ Furthermore,

Received: May 29, 2018

Published: October 29, 2018

protein abundance was found to have higher impacts on the results gained from affinity purifications but not on yeast two hybrid assays.¹⁵ Protein complexes within a cell vary in size, number of partner proteins, their subcellular location, and time of interaction.¹⁶ Furthermore, the process of protein complex assembly and disassembly is often regulated by posttranslational modifications,^{17,18} which adds to the difficulty of investigating protein interactions in heterologous systems, e.g., in yeast-two hybrid methods.

A traditional biochemical method for the unbiased separation of protein complexes is the size exclusion chromatography (SEC). While size-exclusion chromatography or gradient centrifugations in combination with enzyme activity measurements has a long history in plant biochemistry,^{19,20} large-scale information on existing cellular protein complexes still is rather scarce. The protein complexes in chloroplasts were characterized using a blue-native gel separation and protein identification by mass spectrometry²¹ and were explored by promising approaches in top-down proteomics.²² Also, plant mitochondrial membrane complexes are well characterized by blue-native gel approaches coupled with mass spectrometric analyses.²³ In contrast, the systematic large-scale interactomes of plant proteins rather come from automated yeast-two hybrid assays of soluble proteins⁹ or membrane proteins through mating-based split-ubiquitin experiments.¹¹ Only more recently, in order to study the native status and dynamics of the cellular proteomes, separation of protein complexes by size-exclusion chromatography or blue-native gel electrophoresis was combined with high-throughput analysis of proteins by mass spectrometry. In this way, soluble protein complexes were characterized in human cell lines^{24–26} and temporal changes in protein complex composition were analyzed.²⁷ More recently, the SEC-MS approach was used to study protein complexes in *Arabidopsis* soluble proteins²⁸ and membrane proteins complexes from leaves.²⁹

In all organisms, systematic large-scale analyses of protein–protein interactions among membrane proteins are clearly under-represented despite the utmost importance of membrane protein complexes in transport processes as well as signal perception and transduction. In that regard, the large-scale split-ubiquitin interaction screen of plant membrane proteins set the scene revealing 12 102 high confidence interactions of membrane proteins.¹¹ In-depth characterization of individual plasma membrane protein complexes in plants suggested a highly modular and dynamic system in which protein–protein interactions are adjusted to the respective cellular conditions.^{30–32} To better understand this modular system of plasma membrane protein complexes, a systematic study of the native protein complexes, especially in membranes, becomes necessary.

Here, we use the combination of size-exclusion chromatography and mass spectrometry to perform a large-scale analysis of protein complexes in the cellular membranes of *Arabidopsis* roots. Thereby we predict complex composition through systematic analysis of retention time shifts relative to expected retention times of monomeric proteins and through pairwise correlation of elution profiles. We present an interactome of 963 proteins within the plasma membrane of *Arabidopsis* roots.

MATERIALS AND METHODS

Plant Material and Growth Conditions

All experiments were done using *Arabidopsis thaliana* (Col-0) grown in liquid half-strength Murashige and Skoog growth medium as described.³³

Membrane Protein Extraction

Frozen *Arabidopsis* roots (approximately 1–2 g of fresh weight) were smashed into small pieces and resuspended in ice-cold homogenization buffer (330 mM sucrose, 100 mM KCl, 1.5 mM EDTA, 50 mM Tris (tris(hydroxymethyl)aminomethane) adjusted with MES (2-(*N*-morpholino)ethanesulfonic acid) to pH 7.5, 5 mM DTT), and protease inhibitor cocktail (Sigma-Aldrich) was added from stock solutions (50 μ L per 10 mL of the homogenization buffer just before use). Tissue was homogenized with a Teflon Potter-Elvehjem-type homogenizer on ice. The homogenate was filtered through a 21 μ m nylon mesh and centrifuged at 7500g for 15 min at 4 °C. Finally, the supernatant was centrifuged at 48 000g for 80 min at 4 °C. The resulting pellet was the microsomal fraction, which was directly used for chromatographic separation.

Size-Exclusion Chromatography

A protein concentration of around 5 μ g/ μ L was obtained from at least 1 g roots. A total amount of 1 mg Protein in 200 μ L PBS (10 mM phosphate buffer, 2.7 mM potassium chloride and 137 mM sodium chloride, pH 7.4) with 0.1% DDM was protein was run on a Enrich SEC650 (BioRad) column using 1 column volume of phosphate buffered saline (PBS) to separate the protein complexes. A NGC Quest (BioRad, Germany) FPLC unit operated by Chromlab 3.3 software was used to perform the separation. Eluting proteins were collected in a total of 50 fractions. The volume of the collected fractions was set to 0.25 mL for the first 20 fractions, 0.5 mL for the following 10 fractions and 0.75 mL for the 10 remaining fractions. This was done to (i) account for the nonlinear elution properties of especially large proteins and (ii) to increase resolution of early fractions to better detect the expected mass-shift of proteins that are committed to an interaction. Expected elution fractions for each protein were calculated on the basis of a calibration curve with a protein size marker (BioRad Gel Filtration standard #151-1901) using logarithmic equations ($R^2 = 0.9064$, Supporting Table S1).

Immunoprecipitation of GFP-Tagged Proteins

Plants expressing GFP-tagged AHA2 in an *aha2* knockout background were kindly provided by the lab of Prof. Klaus Harter, Tübingen. 300 μ g microsomal proteins prepared from roots of AHA2-GFP expressing plants were incubated with 25 μ L of anti-GFP beads (Chromotek) for 2 h on a rotating wheel at 4 °C. After incubation, the beads were collected via centrifugation and washed two times with 500 μ L wash buffer I (10 mM Tris/Cl pH7.5, 150 mM NaCl 0.5 mM EDTA). For protein–protein interaction assay, the proteins were eluted from the beads with 100 μ L UTU (urea/thiourea), and then 100 μ L elutes were digested with trypsin and desalted with C18-Stage tip as described below (in-solution digestion).

Sample Preparation for Mass Spectrometric Analysis

Approximately 5 μ g protein from each fraction was predigested for 3 h with endoproteinase Lys-C (0.5 μ g/ μ L; Wako Chemicals, Neuss) at room temperature. After 4-fold dilution with 10 mM Tris-HCl (pH 8), samples were digested with 0.1 μ L Sequencing grade Modified trypsin (0.5 μ g μ L⁻¹; Promega) overnight at room temperature. After overnight digestion, trifluoroacetic acid (TFA) was added (until pH \leq 3) to stop digestion. Digested peptides were desalted over C18 tip.³⁴

LC–MS/MS

Tryptic peptide mixtures were analyzed by LC–MS/MS using nanoflow Easy-nLC1000 (Thermo Scientific) as an HPLC-system and a Quadrupole-Orbitrap hybrid mass spectrometer

(Q-Exactive Plus, Thermo Scientific) as a mass analyzer. Peptides were eluted from a 75 $\mu\text{m} \times 25\text{ cm}$ C18 analytical column (PepMan, Thermo Scientific) on a linear gradient running from 4 to 64% acetonitrile in 180 min and sprayed directly into the Q-Exactive mass spectrometer. Proteins were identified by MS/MS using information-dependent acquisition of fragmentation spectra of multiple charged peptides. Up to 12 data-dependent MS/MS spectra were acquired for each full-scan spectrum within the range of 300–1600 m/z acquired at 70 000 full-width half-maximum resolution with an AGC (targeted auto gain control) of 1×10^6 ions and the maximum IT (inclusion time) of 50 ms. Fragment spectra were acquired at a resolution of 17 500 within an isolation window of 1.6 m/z , a AGC of 5×10^5 ions and the maximum IT of 50 ms. The NCE (normalized collision energy, normalized to 300 m/z , $z = 1$) was set to 25. Peptides were dynamically excluded if they were analyzed within the last 20 s. Overall cycle time was approximately one second.

Mass Spectrometric Data Analysis and Statistics

Protein identification and ion intensity quantitation was carried out by MaxQuant version 1.5.3.8.³⁵ Spectra were matched against the *Arabidopsis* proteome (TAIR10, 35386 entries) using Andromeda.³⁶ Thereby, carbamidomethylation of cysteine was set as a fixed modification; oxidation of methionine was set as variable modifications. Mass tolerance for the database search was set to 20 ppm on full scans and 0.5 Da for fragment ions. Multiplicity was set to 1. For label-free quantitation, retention time matching between runs was chosen within a time window of 2 min. Peptide false discovery rate (FDR) and protein FDR were set to 0.01. For identification, the minimum peptide length was set to 7 with a maximum of 2 missed cleavages. Hits to contaminants (e.g., keratins) and reverse hits identified by MaxQuant were excluded from further analysis. A full list of all identified and quantified proteins is presented in [Supporting Table S1](#). Quantitative analysis of protein profiles across fractions was done based on LFQ values.³⁷ LFQ values were further standardized relative to the maximum value across the acquired fractions for each individual protein. Missing values were not imputed. Data processing, filtering and statistics (Pearson correlation analysis) were carried out using in-house Excel macro scripts. Raw files of both replicates were submitted to the ProteomeXChange database and are available under the accession number PXD009511.

Information from Public Databases

Functional classification of proteins was based on MAPMAN bins³⁸ and subcellular location of proteins was obtained from SUBA-database.³⁹ Confirmed interactions were obtained from public databases.^{9,11,40}

RESULTS

The experiments carried out here rely on size-exclusion chromatography combined with quantitative mass spectrometry. *Arabidopsis* roots were used as a model system, since roots have important functions in nutrient uptake and perception and thus require continuous adjustments of growth and development.

Concept for Identification of Interacting Proteins by Elution Shifts and Pairwise Correlation

In size exclusion chromatography intact protein complexes are separated based on their size, which in turn is mostly dependent on the molecular weight of the interacting proteins. Large particles have an earlier elution than small particles. For a single protein, the chromatogram shows a single major peak at the time of expected elution dependent on the protein's molecular weight.

These peaks could be flanked by some minor shoulders resulting from either protein degradation leading to later elution or resulting from protein modifications (e.g., glycosylation) which increase the size of the protein and leads to earlier elution. On the basis of these elution properties in size-exclusion chromatography, we defined proteins as a member of a protein complex (candidates), if their elution time was found to be earlier than expected from the molecular weight of the respective monomers ([Figure 1A](#)). Since the involved proteins merge into a new unit with a larger size than respective monomers, complex formation leads to an increase in apparent size and thus earlier elution from the size exclusion column. Thus, we assumed that large elution shifts of a given protein are due to this protein being member of a protein complex.

We then used pairwise correlations of the SEC-MS elution profiles of each candidate protein in order to identify which of the candidate proteins are in the same complex and interact with each other ([Figure 1C](#)). The prediction of interacting proteins by elution shift analysis and correlation of elution profiles was based on stringent data quality criteria: (i) Proteins were required to have at least four adjacent values with protein abundance values (LFQ-intensity) of at least 10% of the global maximum. (ii) The data window used for pairwise correlation was individually determined for each correlation pair on the basis of the expected elution of the monomeric proteins. In case of a pairwise correlation of a large and a small protein, the correlation window included all fractions (window size) earlier to the fraction that is expected to contain the monomeric large protein. Limiting the correlation to this dynamic window improved the informative value of the correlation results since the focus was on comparing the fractions that contained the apparently larger protein complexes without the monomeric peaks. (iii) Additionally, at least five paired values (sample size) were required within the correlation window ([Figure 1D](#)).

Candidates for Proteins in Protein Complexes

In total, the SEC-MS analysis of two independent biological replicates of 50 fraction of *Arabidopsis* root plasma membrane preparations identified 3889 proteins in replicate 1 (Rep1) and 5563 proteins in replicate 2 (Rep2). From these identified proteins, 1526 proteins of Rep1 and 3004 proteins from Rep2 fulfilled the criteria of being detected in at least four adjacent fractions and containing at least four LFQ-values greater 0.1 ([Figure 2A](#)). Out of these, proteins with elution from five fractions earlier to four fractions later than expected based on their monomeric masses were considered as static monomeric proteins (200 proteins in Rep1). This definition was based on a local minimum of proteins with these elution properties ([Figure 2B](#)). Thus, proteins with elution at least six fractions earlier than expected were considered as candidate proteins to be part of protein complexes (1300/1869 proteins in Rep1/Rep2). Proteins that eluted in later fractions than expected were considered as degraded proteins (26 proteins in Rep1). Most of these candidate proteins were found with a retention shift of 14 fractions earlier than predicted ([Figure 2B](#)). To define those candidate proteins that are part of protein complexes, we required an elution difference of at least six fractions earlier. The threshold of six fractions was defined on the basis of the biphasic distribution of the number of proteins within each elution shift class with a local minimum at five fractions earlier ([Figure 2B](#)). In general, elution shifts for individual proteins to earlier fractions than expected was very reproducible between the two biological replicates. A total of 1387 proteins identified in both replicates were

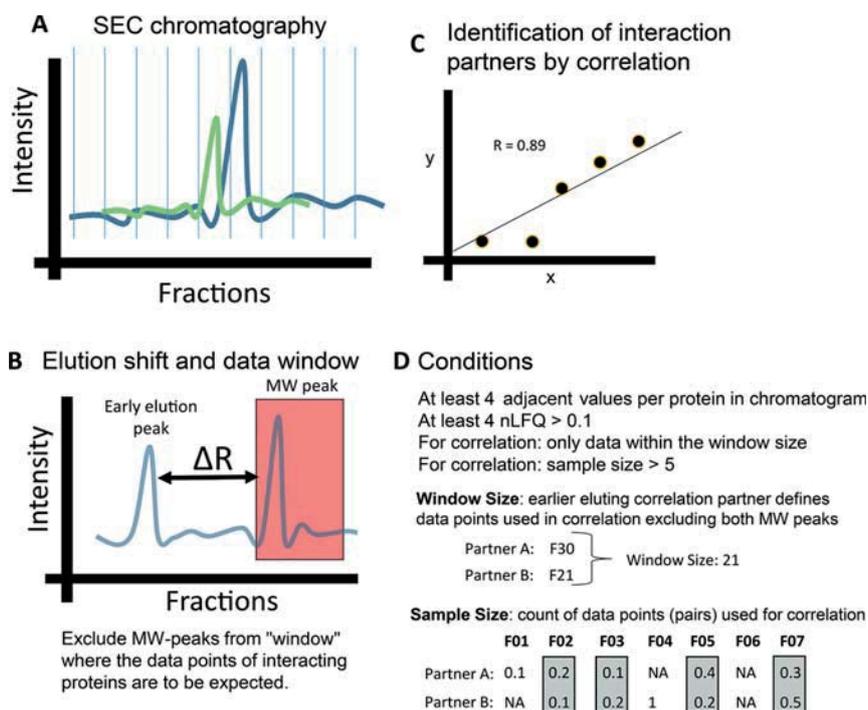


Figure 1. Workflow of identification of interacting proteins (candidates) and their interaction partners by a combination of a mass shift analysis combined with pairwise correlations.

used for quantitative analysis. Out of these commonly quantified proteins, 65% (965 proteins, Supporting Table S2) were classified as complex candidate independently in both replicates (Figure 2C). In general, the eluting fraction of over 50% of all quantified proteins showed a consistent elution pattern, which did not differ for more than five fractions between both replicates (Figure 2D). An elution difference of greater than six fractions elution difference was observed for 47% of the commonly identified proteins.

The overlap between the two biological replicates was very high (Figure 2E): 90% of the proteins identified in Rep1 were also identified in Rep2. However, Rep2 contained more proteins than Rep1; thus, only 63% of the proteins in Rep2 were also identified in Rep1. Altogether, 75% of the complex candidates defined in Rep1 were also identified as complex candidates in Rep2. However, there were another 928 proteins quantified in both replicates that were only defined as complex candidates in Rep1 (311 proteins) or Rep2 (617 proteins) (Figure 2C,E). Retention shift analysis was performed on 1387 proteins quantified in both replicates, resulting in the 965 overlapping proteins classified as complex candidates in both replicates. In addition, 220 of the commonly identified proteins only showed early elution in Rep1, and 44 proteins showed early elution only in Rep2 (Figure 2E). Out of the 965 candidate proteins identified in both replicates, interaction partners were identified for 963 proteins by pairwise correlation, and these were studied further (Supporting Table S2).

Only 256 proteins from our 965 complex candidates identified in both replicates were already defined as part of a protein complex in earlier work on *Arabidopsis* leaves²⁹ (Figure 3A). Most of the 965 complex candidates identified here from *Arabidopsis* roots were indeed localized to membrane compartments, such as plasma membrane (215 proteins), Golgi (43 proteins), and ER (125 proteins) (Figure 3B), and the fraction of proteins with membrane location was higher than in the published study on

Arabidopsis leaf membranes.²⁹ A large number of proteins (225 proteins) were also found with predominantly cytosolic localization. Most identified proteins with assigned functions were proteins of protein synthesis and modification (195 proteins in bin “protein”). A total of 84 candidates were signaling proteins (bin “signaling”), such as receptor kinases or calcium signaling proteins. Another 107 complex candidates were transporters (bin “transport”), such as the plasma membrane ATPases AHA1 and AHA2, aquaporins, calcium ATPases, and other channels (Figure 3C). Among the 43 proteins in bin “cell organization” were proteins with vesicle trafficking functions such as SYPs and other SNAREs.

Definition of Protein Complexes Composition by Pairwise Correlations

Pairwise correlations were used to find primary interaction partners for each complex candidate protein. We used confirmed protein–protein interactions from public databases to define optimal cutoff scores. For confirmed interactions the frequency for correlation scores $r > 0.7$ increased compared to unconfirmed interactions (Figure 4A). The static proteins and degraded proteins had rather few correlation events with $r > 0.7$, while proteins defined as complex candidates contributed the majority of pairwise correlations (Figure 4B). This finding suggests that the definition of protein complex based on elution shifts of at least six fractions was reasonable. Notably, bigger shifts to earlier fractions were associated with an increased count of correlation events per protein. This group of proteins with high elution shifts comprised the largest protein complexes, and thus it was not surprising to find high number of interactions per protein. As expected, the number of interactions per protein as derived from the number of correlations with $r > 0.7$, decreased for static proteins, but was high again also for degraded proteins (Figure 4B). However, average and median r -values were higher for pairwise correlations in large deviation groups (−26 to −16) of protein

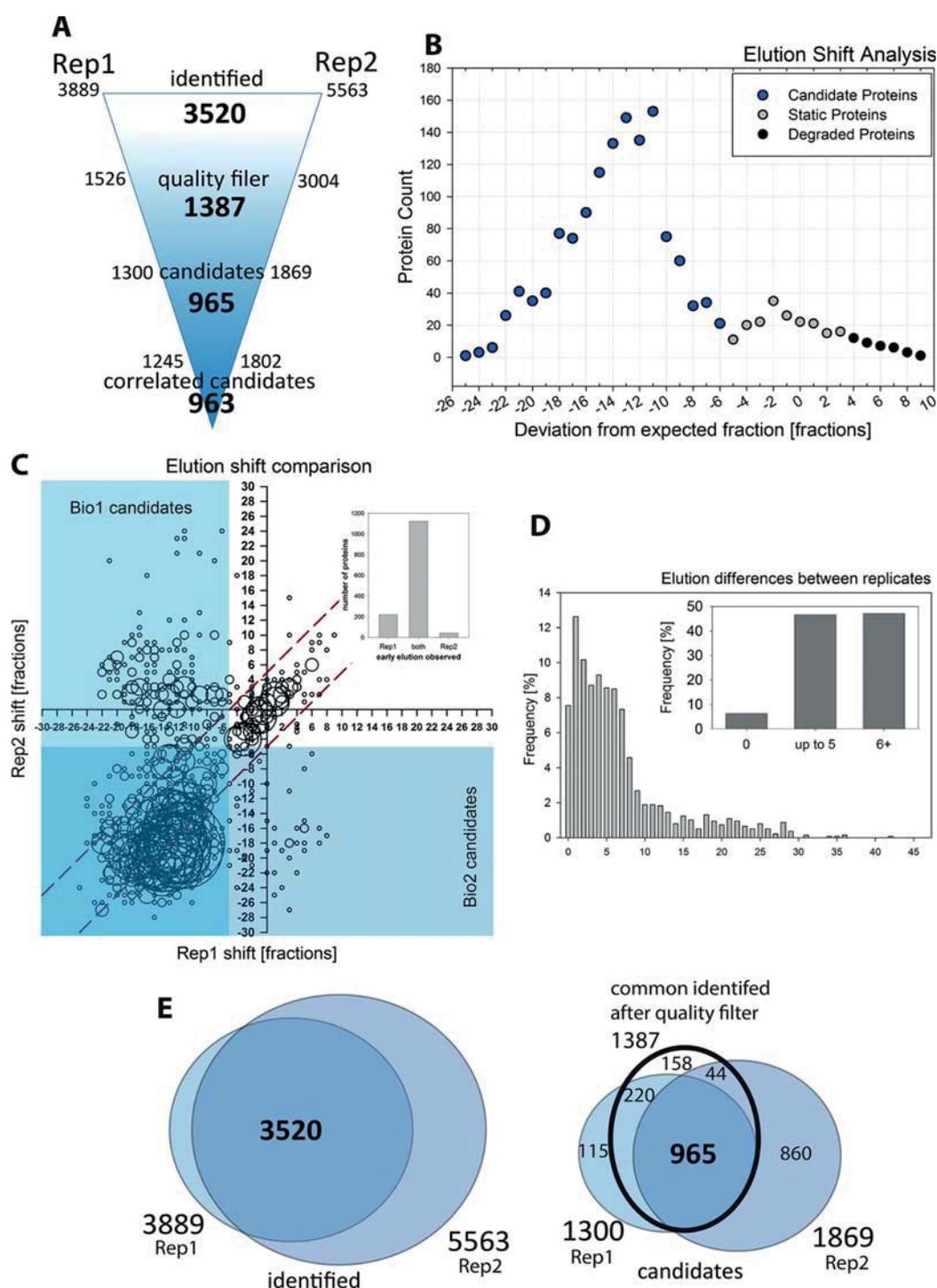


Figure 2. Definition of protein candidates to be part of protein complexes. (A) Number of identified proteins and complex candidates in two biological replicates (B) Proteins with retention time shift of at least 6 fractions earlier were defined as candidate proteins for protein complexes. Example plot from biological replicate 1. (C) Reproducibility of retention times for identified proteins in the two biological replicates. Blue shaded areas indicate regions with elution shifts of at least six fractions earlier in each replicate. The area within the dotted lines indicates the 5-fractions as tolerance for static proteins. (D) Classification of proteins based on their reproducibility in chromatography in both biological replicates. (E) Overlap of identified proteins and complex candidate proteins from both biological replicates.

complex candidates compared to static and degraded proteins (Figure 4C). Highest average correlation score of $r = 0.86$ was found in deviation group -24 fractions. Proteins defined as pairwise interaction partners showed a global maximum of elution

within the same fraction or not different for more than three fractions (Figure 4D). Most pairwise correlations were done with a window size of 20 fractions, and only very few proteins profiles were correlated over less than 12 fractions or more than

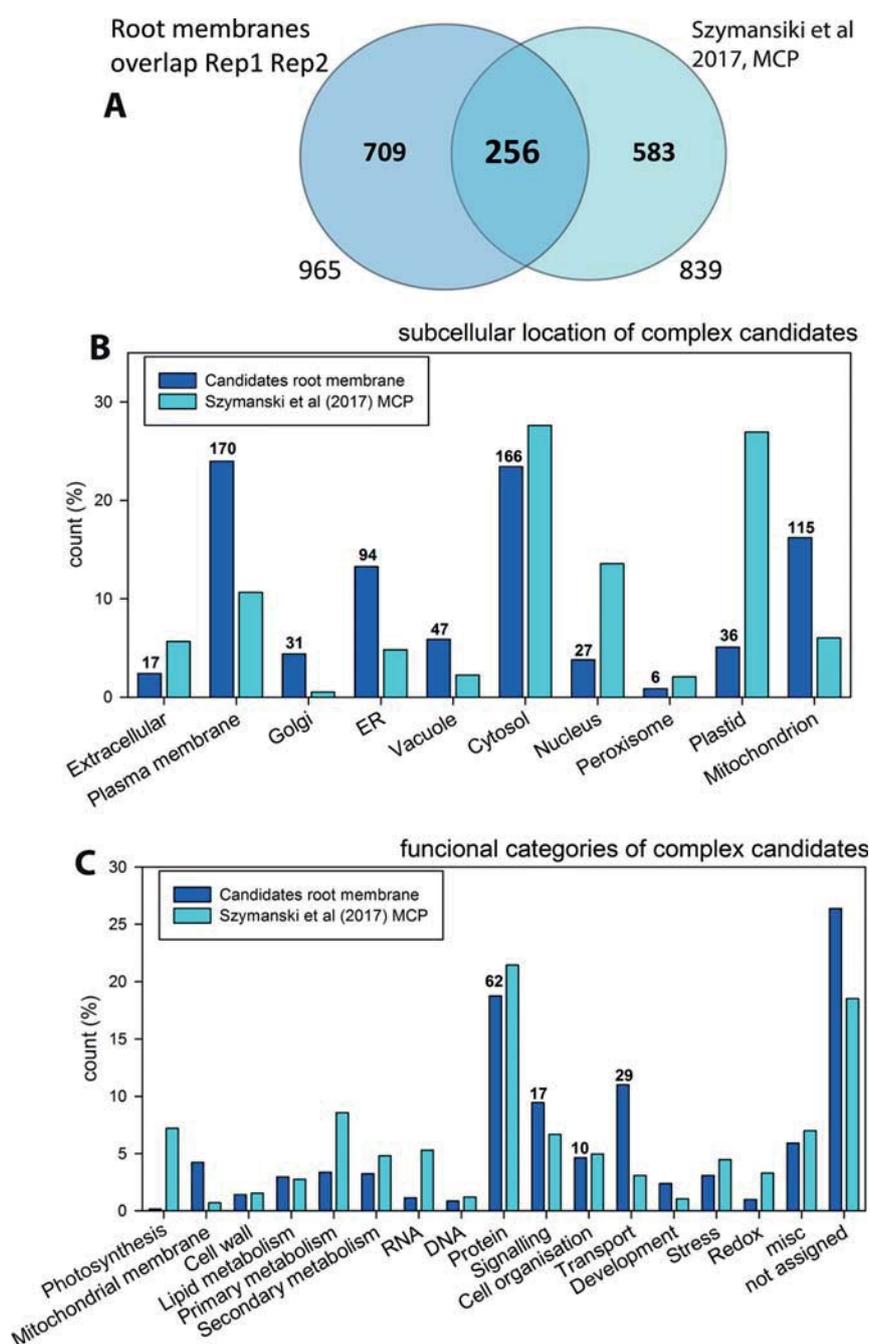


Figure 3. Functional characterization of protein complex candidate proteins. (A) Overlap of complex candidate proteins identified in two biological replicates from this study of *Arabidopsis* root membranes and previously published complex proteins from *Arabidopsis* leaf membranes.²⁹ (B) Subcellular location of identified protein complex candidates from root membranes in comparison to protein complexes identified from leaf membranes.²⁹ (C) Functional characterization of complex candidate proteins defined by the mass shift analysis in comparison to protein complexes identified from leaf membranes.²⁹

28 fractions (Figure 4E). More than half of all correlations were based on a sample size greater than 9 data points (Figure 4F).

The above analysis revealed a total of 102 361 and 290 266 correlation events were accepted with $r > 0.7$ in Rep1/Rep2. In order to more stringently assess the quality of those correlation events, we separated high confidence and low confidence interactions based on the correlation score distribution of confirmed and unconfirmed interactions. Confirmed interactions were retrieved from TAIR⁴¹ (2657 pairs), MIND (11) (12 103 pairs),

the *Arabidopsis* interactome⁹ ((11 374 pairs)) and the STRING database⁴⁰ (31 710 pairs). After removal of duplications, 73 929 confirmed *Arabidopsis* protein–protein interactions remained for comparison with correlation pairs identified here. Out of these published interactions, only 19 230 interactions were findable among the 1300 protein candidates of Rep1, and for Rep2, 30 323 published interactions were findable within the 1869 protein candidates. In total, 43 825 published interactions were found for 2074 out of 2084 unique proteins from Rep1 and Rep2

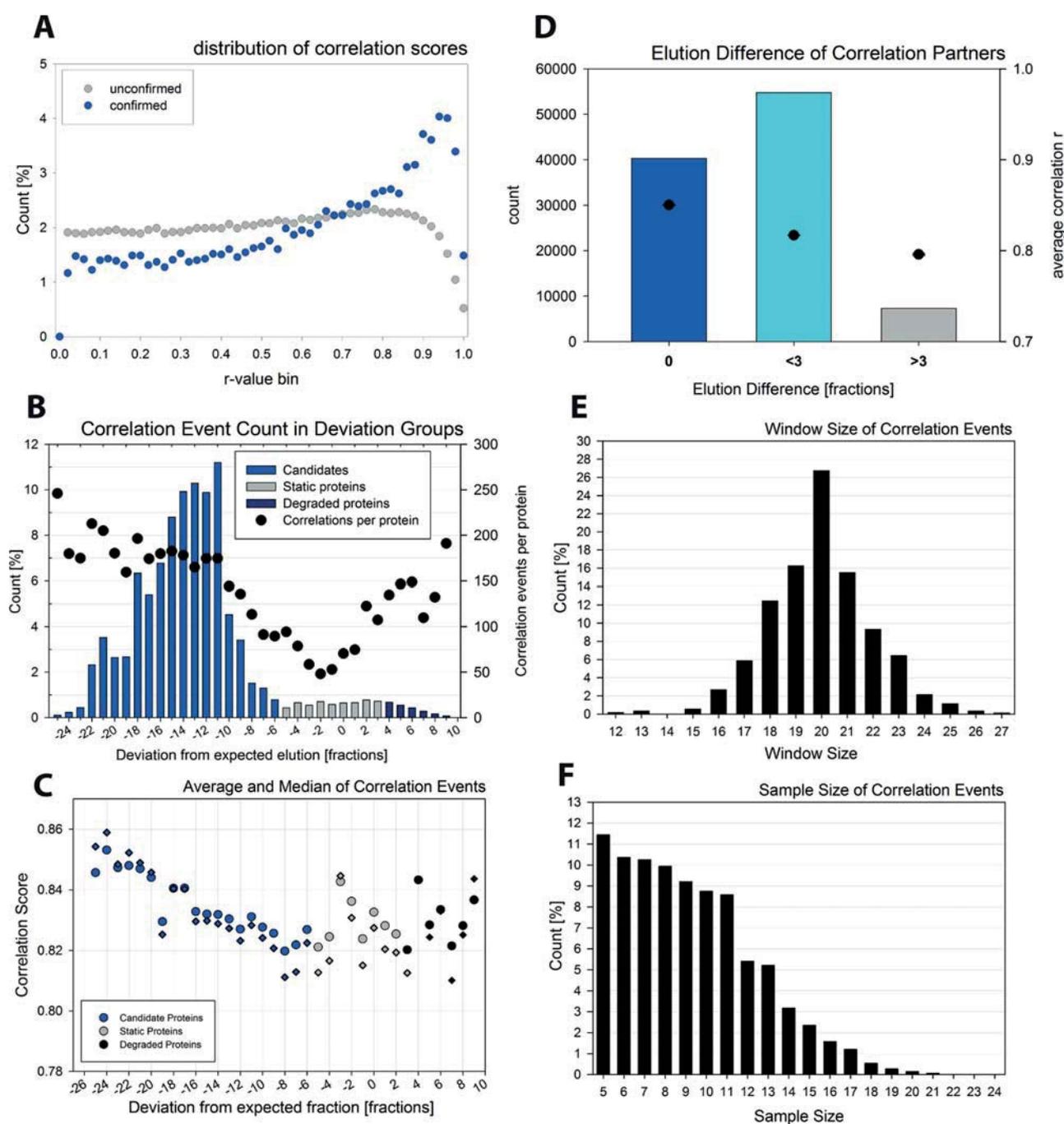


Figure 4. Identification of interaction partners to candidates using pairwise correlations. (A) Distribution of pairwise correlation scores for unconfirmed and confirmed interactions. Confirmed interactions were obtained from public databases.^{9,11,58} (B) Distribution of pairwise correlations with $r > 0.7$ across the deviation groups and average correlation events per protein within each deviation group. (C) Average (circles) and median (diamonds) correlation score for proteins within each deviation group. Each protein within a deviation group was correlated against all other proteins, and only correlation scores $r > 0.7$ were considered. (D) Elution differences of correlation partners. Most protein pairs with $r > 0.7$ showed a global maximum of elution in the same fraction or not more than 3 fractions different. (E) Distribution of data window sizes used for pairwise correlations. (F) Distribution of sample sizes used in pairwise correlations. All panels were exemplified on the basis of biological replicate 1.

used for correlations. Thus, there were 10 proteins for which no interactions were published.

The majority (>60%) of protein pairs undergoing such confirmed interactions indeed showed an elution behavior of both partners in the same fraction or with only one fraction difference (Figure 5A). In contrast, protein partners of unconfirmed

interactions eluted over a wider range of fractions. In addition, confirmed interactions had a tendency for high correlation r -values (Figure 4D). Thus, we defined high confidence interactions between proteins based on an elution difference of no more than 1 fraction and a correlation r -value greater than 0.89. A total of 26 491 pairwise interactions were identified overlapping

in both replicates. Out of these, after applying our high-confidence criteria described above, 7% (1752 interactions) were defined as high confidence interactions, and were found in both replicates, while another 69 534 interactions were classified as high-confidence in only one replicate (Figure 5B). In Rep1, 21% and in Rep2, 18% of all interactions were defined as high-confidence.

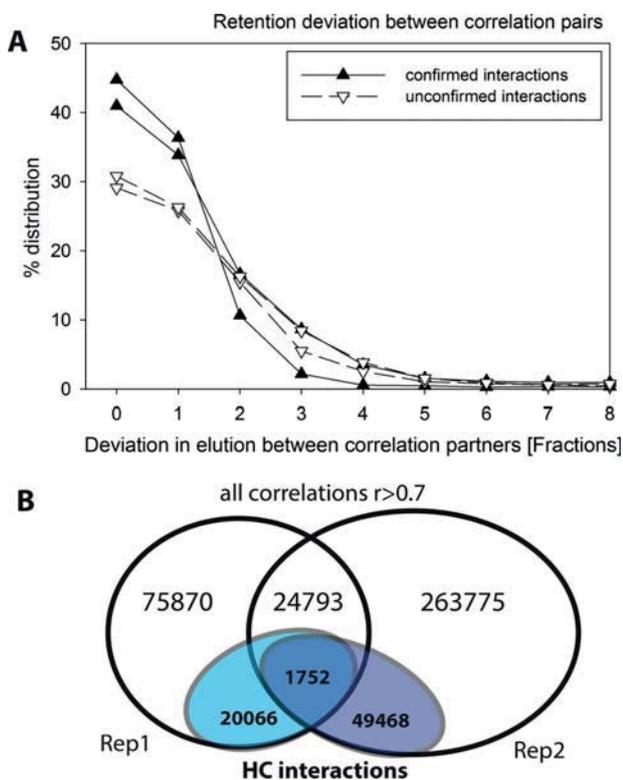


Figure 5. Confirmed and unconfirmed interaction pairs. (A) Distribution of elution fractions of correlation partners with r -values > 0.7 . Filled symbols indicate confirmed interaction pairs, open symbols indicate unconfirmed interactions. Unconfirmed interactions are a mixture of true interactions and false interactions. (B) Overview of identified interactions in the two replicates classified as high-confidence and low-confidence interactions.

While the identification of complex candidates was very robust between the two replicates, the composition of the complexes showed higher degree of variability with only 8% of the high-confidence interactions of Rep1 and 3.4% of the high-confidence interactions of Rep2 were identified in both replicates. The lower percentage of shared HC-interactions in Rep2 is most likely due to the fact that much more proteins were identified in Rep2.

The highest number of high confidence interactions (20.5%) was found among mitochondrial proteins, which were copurified with the microsomal membranes. High numbers of interactions (4.9% and 7.4%, 4.6%) were also found among plasma membrane proteins, among ER proteins, and between plasma membrane and ER proteins, respectively. For plasma membrane located transporters and channels (bin 34), most interactions were identified with yet uncharacterized proteins (bin 35; 5.6%), with other transporters (bin 34, 5.07%), with protein modifying proteins (bin 29, 3.65%) or signaling proteins (bin 30, 3.65%). As expected, several of the high-confidence interactions predicted here connect neighboring compartments, such as plasma

membrane and cytosol, as well as plasma membrane and ER. Surprisingly, we found 224 interaction pairs within the shared high-confidence interactions to occur between plasma membrane proteins and mitochondrial proteins. However, only two of these interactions are yet described in public database.

Reconstructed Interaction Networks

Next, we used the pairwise interaction data to reconstruct the protein complex networks for the complex candidate proteins. Thereby, only the 1752 high confidence interactions identified in both replicates were considered (Figure 6A, Supporting Table S3). The network was clustered according to the observed global maximum of protein retention in Rep1. Within each fraction several communities were identified, suggesting that indeed within each elution fraction, we identified several different distinct protein complexes.

Fractions 1 and 2 contained mainly ribosomal protein complexes. Fraction 4 contained a small cluster with signaling proteins BAK1. Fraction 1 and 2 contained also protein complexes with tubulins, motor proteins, and coatomers. Fraction 5 contained a large cluster of mitochondrial proteins around ATP-synthase in addition to a smaller plasma membrane complex around FER and another smaller complex containing ER proteins involved in lipid biosynthesis. Fraction 6 contained another mitochondrial membrane protein complex with NADH dehydrogenase, and two plasma membrane protein complexes, one containing PDR7 and ACA10 and the other around TPC1 and ACA1. Fraction 7 contained up to 10 distinct protein complexes, for example, one ER protein complex, a mitochondrial protein complex around cytochrome c, and a plasma membrane protein complex with plasma membrane ATPases AHA1 and AHA2 as well as calcium ATPase ACA8. Fraction 8 contained several ER protein complexes, a complex of vacuolar proteins with sugar transporter TMT2, and a small complex of plasma membrane proteins. In Fraction 9 we identified a signaling protein complex with protein kinase BSK1 and other receptor kinases in addition to another distinct receptor kinase complex with QSK1 and CERK1 in addition to several vesicle proteins (VAMPs and SYPs). Fraction 10 contained a complex with cellulose synthase like protein CSLE1. Fraction 11 was characterized by a heat shock protein complex, and Fractions 12/13 contained the proteasome complexes.

The plasma membrane H^+ -ATPase AHA2 is one of the most abundant proteins in roots. Therefore, we investigated the detected interactions of AHA2 in more detail (Figure 6B, Supporting Table S4). We found 11 interaction partners of AHA2 in both replicates. These were three members of the VHA-family, AHA1 (AT2G18960), ACA8 (AT5G57110), ACA11 (AT3G57330), two band 7 family members (HIR2, AT3G01290 and AT5G51570), a phospholipase C (ATPLC2, AT3G08510) and a member of the BSK signaling kinases (AT1G63500). However, most of the 174 high-confidence interaction partners were identified in only one replicate. AHA2 was found to interact with ammonium transporters, aquaporins, or members of the ABC transporter family (Supporting Table S4). Among the signaling proteins, several receptor-like kinases were identified as interactors for AHA2 in different replicates. The receptor kinase FERONIA, which was shown to be a negative regulator of AHA2,⁴² was identified as shared interaction partner, but did not meet our criteria for a high confidence interaction.

To confirm the interactions of AHA2 predicted by our methods, we performed pull-downs with GFP-tagged AHA2 under the same standard growth conditions as used in the complex

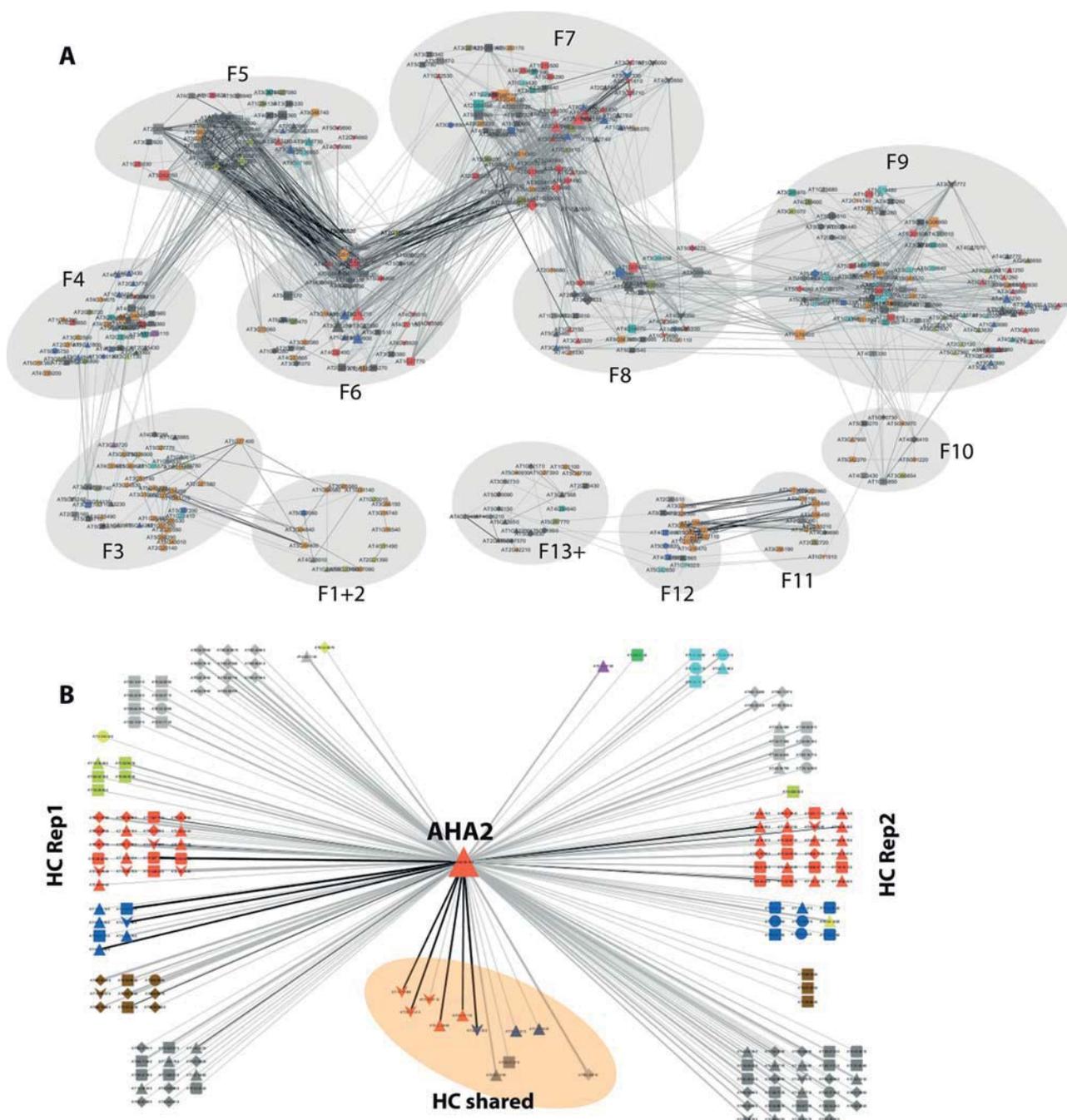


Figure 6. (A) Reconstructed network based 460 proteins undergoing 1752 high confidence interactions identified in both replicates. Node size is proportional to the degree of connection. A group-attribute sorted layout was used to separate the fractions and within each fraction an edge-weighted Allegro-Layout was used with the average r -value (Rep1 and Rep2) as weight attribute. Black edges: confirmed interactions. Diamond: mitochondrion, triangle: plasma membrane, square: endoplasmic reticulum, circle: cytosol, vee: vacuole, round square: default node shape. Red: transporters (bin 34), blue: signaling (bin 30), orange: protein (bin 29), beige: stress (bin 20) and redox (bin 21), yellow: vesicle trafficking (bin 31), pink: RNA (bin 27), cyan: metabolism general (bins 1 to 8, bins 13 to 19), dark gray: not assigned functions, light gray: default color. Black edges: confirmed interactions, gray edges: unconfirmed interactions. (B) Reconstructed network based on high confidence interactions of AHA2.

analysis. The pull-down analysis revealed 32 overlapping interaction partners (Supporting Table S5) with the 174 high-confidence interactions based on the SEC-MS analysis. Among the interactions confirmed by pull-down analysis were interactions of AHA2 with ACA2, PIP2F, TPC1 and receptor kinase AT3G02880. The kinase, PIP2F and AHA2 also found as being associated in interacting complexes in sucrose-stimulated seedlings (31).

DISCUSSION

We defined protein complex candidates based on the assumption that the elution profile of a protein is changed in the event of an interaction. The altered elution profile of a protein was determined by comparing the observed global maximum of protein retention with the expected fraction of maximum retention

based on the monomeric mass of the protein. Thereby, the altered retention behavior of proteins was not recorded as an actual “mass”-shift but rather as a “change in the fraction number”, which—of course—is directly related. Using and working with this original experimental information (“fractions”) eliminated the risk of introducing errors or inaccuracy by calculating virtual masses for each protein.

Definition of Protein Complex Candidates

A characteristic feature of size-exclusion chromatography is that separation of proteins occurs based on particle sizes. When drawing the standard curve, it can be observed that the relationship between molecular weight and retention time followed a logarithmic trend rather than a linear one. Consequently, each fraction covered a different range of molecular weights and that not every shift in the fractions observed was the result of the same increase of molecular weight. For example, early fractions covered a relatively large range of molecular weights: Fraction 10 covered a range of approximately 41 kDa while the average calculated protein size in fraction 6 was 142 kDa. In comparison, fraction 30 covered a range of 1.2 kDa with an average calculated molecular weight of around 4.4 kDa. Thus, the minimum increase in protein size that is necessary to result in a shift of the fraction number was much higher for large proteins than for smaller proteins. To account for this, smaller volume fractions were collected during earlier elution times and larger volumes were collected during later elution times.

Similarly, small proteins were more prone to show a greater shift of the fraction number when they commit an interaction. Thus, if a small protein was interacting with a big protein, then we may have only detected the shift for the small protein, while the fraction in which the big protein was detected remained unchanged because the mass gain was too low. Here, candidate proteins were defined as proteins that eluted at least six fractions earlier than expected. However, it was impossible to detect a shift of six fractions for the proteins eluting in the first five fractions. On the other hand, a shift of one fraction could be the result of a major increase in molecular weight for a big protein, but only a minor increase for a smaller protein. As a consequence, information about proteins that increase their size by complex formation was lost mainly for large proteins. This problem was partially compensated as the retention time shift analysis was combined with the Pearson correlation analysis. Even if one of the interaction partners did not show a shift in the elution fractions, the other interacting protein might display a significant retention time shift to be detected as complex candidate protein. However, these issues may be a source of false negatives. For example, the plasma membrane ATPase AHA2 (large protein) is known to interact with 14-3-3 proteins (small protein). While both AHA2 and 14-3-3 proteins were identified in the sample, an interaction between them was not predicted, possibly due to the problems described above.

An alternative approach is to define the change in protein size based on the apparent molecular weight of the protein compared to its monomeric molecular weight. This approach was followed in previous SEC-MS work on *Arabidopsis* proteins.^{28,29,43} Thereby, a ratio (R_{app}) was calculated of the apparent molecular weight of each protein (M_{app}) divided by the monomeric molecular weight (M_{mono}). A protein was defined to be part of a protein complex if $R_{app} > 2$. The definition of complex candidate proteins used here based on elution fraction shifts instead of apparent molecular weight indeed resulted in $R_{app} > 2$ for all our complex candidates with (Supporting Table S1). However, calibration curves are

necessary to convert fraction numbers for each protein back to apparent molecular weight. Thus, there is no increase in resolution by performing this calculation.²⁹ The primary information on protein identifications is the fraction number, and therefore we used this direct experimental evidence of shifted elution fraction to define an increase in apparent size of each protein.

Definition of Pairwise Interactions

Identification of complex candidate proteins based on earlier than expected elution revealed almost 1000 protein candidates that were suggested to be part of a protein complex. Here, we used pairwise correlations and network reconstruction to identify which proteins could be part of the same multiprotein-complex. In previously published work, complex composition was defined on the basis of proteins belonging to the same elution clusters.^{26,29} In that approach the number of identified protein complexes depended on the number of defined elution clusters. Results indicated a protein to be part of a complex but does not allow to derive information about pairwise interactions among the complex proteins. In contrast, the pairwise correlations applied here was more precise on pairwise protein interactions, while protein complex composition was derived based on these interactions. In future work, the view on pairwise interactions is most likely easier to handle when it comes to studying interaction dynamics, since it is based on quantitative correlation values.

Reproducibility

Our work indicates that the identification of protein complex candidates was very robust with high overlaps between the two replicates. In contrast, the underlying pairwise interactions showed far less overlap regarding the pairwise interactions identified in both replicates. Commonly identified interactions were 25.8% of all interactions in Rep1 and 9.1% of all interactions in Rep2. It needs to be noted that far more proteins were identified in Rep2 leading to a much higher number of interactions in Rep2 compared to Rep1. Both replicates were experimentally carried out in two different laboratory locations within six months, and this may have contributed to variability among interactions. However, the strength of our method in defining the complex candidates lies in the fact that we consider relative changes (here: elution shift) instead of using and comparing information on absolute eluting fractions. Thus, it is important that a given protein is bigger (i.e., earlier elution) relative to its monomeric mass, but it is not important that a protein elutes in the exact same fraction in the two replicates. On the basis of these criteria and within the 5-fraction tolerance applied (insert, Figure 2D), 80% commonly identified proteins were considered complex candidates in both replicates (insert, Figure 2C). We therefore consider the elution shift analysis as highly reproducible between replicates although the absolute chromatograms may differ.

There is more and more evidence that protein complexes within the plasma membrane are not static, but subject to dynamic composition depending on the status of the plant.³⁰ Especially the interaction of channel and transporter proteins with their regulators (i.e., kinases, phosphatases) are often transient and therefore be picked up in one replicate but not the other. Differential efficiency of plasma membrane vesicles isolated from different cell types of the root⁴⁴ may additionally contribute to the observed higher variability of complex composition compared to complex candidates. In contrast, highly stable complexes, such as ribosomes and protein complexes of the mitochondrial respiratory chain were identified with high reproducibility. Although we presented clear threshold for filtering possible interactions as high confidence interactions

based on known interaction pairs, this needs to be considered as a dynamic threshold depending on accepted false positive and false negative interactions. Our results suggest that transient, dynamic interactions are frequently found with lower confidence.

Biological Protein Complexes versus Random Associates

In our study we solubilized the membrane pellets in phosphate-buffered saline supplemented with 0.1% DDM. In contrast to previously published membrane complex characterizations from *Arabidopsis* leaves that used cholate to solubilize membranes,²⁹ our solubilization conditions were rather mild. However, the use of *n*-dodecyl β -*D*-maltoside (DDM) was essential for the experiment as DDM is known to enhance stability of membrane complexes and preserves the activity of membrane-proteins during purification.^{45,46} More recently, the effect of different detergents was explored by purification of a known protein complex around cadherin-11.⁴⁷ In fact, it was confirmed that dodecyl maltoside as a mild detergent preserved known interaction partners of Cad-11, which were lost when for example cholate was used.⁴⁷

Protein distributions within the membranes in general are not considered random and depend on active complex formation and protein recruitment,⁴⁸ targeted vesicle fusion,⁴⁹ or interactions with the cytoskeleton.⁵⁰ Highly polar localization of membrane proteins are extreme examples of these effects.⁵¹ To further avoid identification of random interactions, we trained our prediction of protein–protein interactions on 43 825 experimentally confirmed interaction pairs published for our identified protein complex candidates.^{41,11,9,40} Thus, we are convinced of identifying nonrandom protein associations that are the result of biologically induced dynamic processes within the respective membranes.^{52,53}

Despite the fact that we consider protein distribution within the membrane as not random due to complex formation, false-positive interactions based on copurification may have occurred when looking at protein interactions between different membrane types (e.g., plasma membrane and mitochondrion). It is well-known that the microsomal membrane preparation protocol applied here does include mitochondrial membranes as well as ribosomes, and copurification issues were pointed out previously.⁵⁴

Insights into the Interactome of the Root Membranes

We present an interactome of 963 candidate proteins undergoing at least 1752 high confidence interactions revealed in both replicates. Using AHA2 as a well characterize example, we could confirm known interactions of the AHA2 particularly with other members of the proton-ATPase family (AHAs), and different calcium-ATPase family members (ACAs). More surprisingly, among the high-confidence interactions of AHA2 we found rather few receptor-like kinases (five receptor kinases, one CDPK), which is in contrast to recent studies of the AHA interactome by immune precipitation, in which a high number of kinases were found.⁵⁵ We were unable to find known interactions of AHA2 for example with FERONIA,⁴² PKSS,⁵⁶ or PSK2.⁵⁷ 2077 interaction pairs involved receptor kinases. Among these, the interaction of BAK1 with BSK3 or BAK1 with ACA2 were among the confirmed interactions, and also the interaction of AT3G02880 with aquaporin PIP2F.³¹

It must be noted that the interactome defined here was established under nonstimulated conditions, and most of the interactions of published for AHA2 with different kinases were found at various stimulations. Thus, we may not have picked up such transient interactions, which are stabilized and thus noticed in our approach only under specific stress stimulations.

Furthermore, our interactome seems to be highly reproducible in identifying stable protein complexes, while identification of weaker, transient interactions (e.g., AHA2 with kinases) may require either respective stimulations or cross-linking strategies.⁵⁵

CONCLUSION

The method outlined here provides a robust identification of proteins in a complex, which we here named complex candidates. In addition we predict their pairwise interactions on the basis of correlation of elution profiles. We are convinced that the combination of retention shift analysis and pairwise correlation of elution profiles described here provides a very suitable approach for comparative analysis of protein complex dynamics in future.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.8b00382.

Descriptions of Supporting Tables (PDF)

Supporting Table S1 (XLSX)

Supporting Table S2 (XLSX)

Supporting Table S3 (XLSX)

Supporting Table S4 (XLSX)

Supporting Table S5 (XLSX)

AUTHOR INFORMATION

Corresponding Author

*E-mail: wschulze@uni-hohenheim.de.

ORCID

Waltraud X. Schulze: 0000-0001-9957-7245

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was performed within grant proposal DFG SCHU1544/11-1 to Waltraud Schulze.

REFERENCES

- (1) Bensimon, A.; Heck, A. J.; Aebersold, R. Mass spectrometry-based proteomics and network biology. *Annu. Rev. Biochem.* **2012**, *81*, 379–405.
- (2) Braun, P.; Gingras, A. C. History of protein-protein interactions: from egg-white to complex networks. *Proteomics* **2012**, *12* (10), 1478–1498.
- (3) Fields, S.; Song, O. A novel genetic system to detect protein-protein interactions. *Nature* **1989**, *340* (6230), 245–246.
- (4) Collins, M. O.; Choudhary, J. S. Mapping multiprotein complexes by affinity purification and mass spectrometry. *Curr. Opin. Biotechnol.* **2008**, *19* (4), 324–30.
- (5) Babu, M.; Vlasblom, J.; Pu, S.; Guo, X.; Graham, C.; Bean, B. D. M.; Buston, H. E.; Vizeacoumar, F. J.; Snider, J.; Phanse, S.; Fong, V.; Tam, Y. Y. C.; Davey, M.; Hnatshak, O.; Bajaj, N.; Chandran, S.; Punna, T.; Christiopolous, C.; Wong, V.; Yu, A.; Zhong, G.; Li, J.; Stagljar, I.; Conibear, E.; Wodak, S. J.; Emili, A.; Greenblatt, J. F. Interaction landscape of membrane-protein complexes in *Saccharomyces cerevisiae*. *Nature* **2012**, *489* (7417), 585–589.
- (6) Gavin, A.-C.; Bösch, M.; Krause, R.; Grandi, P.; Marzioch, M.; Bauer, A.; Schultz, J.; Rick, J. M.; Michon, A.-M.; Cruciat, C. M.; Remor, M.; Höfert, C.; Schelder, M.; Brajenovic, M.; Ruffner, H.; Merino, A.; Klein, K.; Hudak, M.; Dickson, D.; Rudi, T.; Gnau, V.; Bauch, A.; Rybin, V.; Drewes, G.; Raida, M.; Bouwmeester, T.; Bork, P.;

Seraphin, B.; Küster, B.; Neubauer, G.; Superti-Furga, G. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **2002**, *415*, 141–147.

(7) Ho, Y.; Grubler, A.; Heilbut, A.; Bader, G. D.; Moore, L.; Adams, S.-L.; Millar, A.; Taylor, P.; Bennett, K.; Boutlier, K.; Yang, L.; Wolting, C.; Donaldson, I.; Schandorff, S.; Shewnarane, J.; Vo, M.; Taggart, J.; Goudreaux, M.; Muskat, B.; Alfano, C.; Dewar, D.; Lin, Z.; Michalickova, K.; Willems, A. R.; Sassi, H.; Nielsen, P. A.; Rasmussen, K. J.; Andersen, J. R.; Johansen, L. E.; Hansen, L. H.; Jespersen, H.; Podtelejnikov, A.; Nielsen, E.; Crawford, J.; Poulsen, V.; Sorensen, B. D.; Matthiesen, J.; Hendrickson, R. C.; Gleeson, F.; Pawson, T.; Moran, M. F.; Durocher, D.; Mann, M.; Houge, C. W. V.; Figeys, D.; Tyers, M. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **2002**, *415*, 180–183.

(8) Yu, H.; Braun, P.; Yildirim, M. A.; Lemmens, I.; Venkatesan, K.; Sahalie, J.; Hirozane-Kishikawa, T.; Gebreab, F.; Li, N.; Simonis, N.; Hao, T.; Rual, J. F.; Dricot, A.; Vazquez, A.; Murray, R. R.; Simon, C.; Tardivo, L.; Tam, S.; Svrlkapa, N.; Fan, C.; de Smet, A. S.; Motyl, A.; Hudson, M. E.; Park, J.; Xin, X.; Cusick, M. E.; Moore, T.; Boone, C.; Snyder, M.; Roth, F. P.; Barabasi, A. L.; Tavernier, J.; Hill, D. E.; Vidal, M. High-quality binary protein interaction map of the yeast interactome network. *Science* **2008**, *322* (5898), 104–110.

(9) The Arabidopsis Interactome Mapping Consortium. Evidence for network evolution in an Arabidopsis interactome map. *Science* **2011**, *333* (6042), 601–607.

(10) Stagljar, I.; Korostensky, C.; Johnsson, N.; Te Heesen, S. A genetic system based on split-ubiquitin for the analysis of interactions between membrane proteins in vivo. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, S187–S192.

(11) Jones, A. M.; Xuan, Y.; Xu, M.; Wang, R.-S.; Ho, C. H.; Lalonde, S.; You, C. H.; Sardi, M. I.; Parsa, S. A.; Smith-Valle, E.; Su, T.; Frazer, K. A.; Pilot, G.; Pratelli, R.; Grossmann, G.; Acharya, B. R.; Hu, H. C.; Engineer, C.; Villiers, F.; Ju, C.; Takeda, K.; Su, Z.; Dong, Q.; Assmann, S. M.; Chen, J.; Kwak, J. M.; Schroeder, J. I.; Albert, R.; Rhee, S. Y.; Frommer, W. B. Border control - A membrane-linked interactome of Arabidopsis. *Science* **2014**, *344*, 711–716.

(12) Choi, H.; Kim, S.; Gingras, A. C.; Nesvizhskii, A. I. Analysis of protein complexes through model-based biclustering of label-free quantitative AP-MS data. *Mol. Syst. Biol.* **2010**, *6*, 385.

(13) Van Leene, J.; Stals, H.; Eeckhout, D.; Persiau, G.; Van Slijke, E.; Van Isterdael, G.; De Clercq, A.; Bonnet, E.; Laukens, K.; Remmerie, N.; Hendrickx, K.; De Vijlder, T.; Adbelkrim, A.; Pharazyn, A.; Van Onckelen, H.; Inze, D.; Witters, E.; De Jaeger, G. A tandem affinity purification-based technology platform to study the cell cycle interactome in Arabidopsis thaliana. *Mol. Cell. Proteomics* **2007**, *6* (7), 1226–1238.

(14) Braun, P.; Aubourg, S.; Van Leene, J.; De Jaeger, G.; Lurin, C. Plant protein interactomes. *Annu. Rev. Plant Biol.* **2013**, *64*, 161–187.

(15) von Mering, C.; Krause, R.; Snel, B.; Cornell, M.; Oliver, S. G.; Fields, S.; Bork, P. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **2002**, *417*, 339–403.

(16) Han, J. D.; Bertin, N.; Hao, T.; Goldberg, D. S.; Berriz, G. F.; Zhang, L. V.; Dupuy, D.; Walhout, A. J.; Cusick, M. E.; Roth, F. P.; Vidal, M. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **2004**, *430*, 88–93.

(17) Przybylski, C.; Jünger, M. A.; Aubertin, J.; Radvanyi, F.; Aebersold, R.; Pflieger, D. Quantitative analysis of protein complex constituents and their phosphorylation states on a LTQ-Orbitrap instrument. *J. Proteome Res.* **2010**, *9*, S118.

(18) Pflieger, D.; Jünger, M. A.; Müller, M.; Rinner, O.; Lee, H.; Gehrig, P. M.; Gstaiger, M.; Aebersold, R. Quantitative proteomic analysis of protein complexes: concurrent identification of interactors and their state of phosphorylation. *Mol. Cell. Proteomics* **2008**, *7* (2), 326–346.

(19) Berczi, A.; Fredlund, K. M.; Möller, I. M. Purification and characterization of an NADH-hexacyanoferrate(III) reductase from spinach leaf plasma membrane. *Arch. Biochem. Biophys.* **1995**, *320* (1), 65–72.

(20) Leonard, R. T.; Vanderwoude, W. J. Isolation of plasma membranes from corn roots by sucrose density gradient centrifugation: an anomalous effect of ficoll. *Plant Physiol.* **1976**, *57* (1), 105–114.

(21) Peltier, J. B.; Cai, Y.; Sun, Q.; Zabrouskov, V.; Giacomelli, L.; Rudella, A.; Ytterberg, A. J.; Rutschow, H.; van Wijk, K. J. The oligomeric stromal proteome of Arabidopsis thaliana chloroplasts. *Mol. Cell. Proteomics* **2006**, *5* (1), 114–133.

(22) Zabrouskov, V.; Giacomelli, L.; van Wijk, K. J.; McLafferty, F. W. A new approach for plant proteomics: characterization of chloroplast proteins of Arabidopsis thaliana by top-down mass spectrometry. *Mol. Cell. Proteomics* **2003**, *2* (12), 1253–60.

(23) Ramírez-Aguilar, S. J.; Keuthe, M.; Rocha, M.; Fedyayev, V. V.; Kramp, K.; Gupta, K. J.; Rasmussen, A. G.; Schulze, W. X.; van Dongen, J. T. The composition of plant mitochondrial supercomplexes changes with oxygen availability. *J. Biol. Chem.* **2011**, *286* (50), 43045–43053.

(24) Camacho-Carvajal, M. M.; Wollscheid, B.; Aebersold, R.; Steimle, V.; Schamel, W. W. A. Two-dimensional Blue Native/SDS gel electrophoresis of multi-protein complexes from whole cellular lysates. *Mol. Cell. Proteomics* **2004**, *3* (2), 176–182.

(25) Havugimana, P. C.; Hart, G. T.; Nepusz, T.; Yang, H.; Turinsky, A. L.; Li, Z.; Wang, P. I.; Boutz, D. R.; Fong, V.; Phanse, S.; Babu, M.; Craig, S. A.; Hu, P.; Wan, C.; Vlasblom, J.; Dar, V. U.; Bezginov, A.; Clark, G. W.; Wu, G. C.; Wodak, S. J.; Tillier, E. R.; Paccanaro, A.; Marcotte, E. M.; Emili, A. A census of human soluble protein complexes. *Cell* **2012**, *150* (5), 1068–1081.

(26) Kirkwood, K. J.; Ahmad, Y.; Larance, M.; Lamond, A. I. Characterization of native protein complexes and protein isoform variation using size-fractionation-based quantitative proteomics. *Mol. Cell. Proteomics* **2013**, *12* (12), 3851–3873.

(27) Kristensen, A. R.; Gsponer, J.; Foster, L. J. A high-throughput approach for measuring temporal changes in the interactome. *Nat. Methods* **2012**, *9* (9), 907–9011.

(28) Aryal, U. K.; Xiong, Y.; McBride, Z.; Kihara, D.; Xie, J.; Hall, M. C.; Szymanski, D. B. A proteomic strategy for global analysis of plant protein complexes. *Plant Cell* **2014**, *26* (10), 3867–82.

(29) McBride, Z.; Chen, D.; Reick, C.; Xie, J.; Szymanski, D. B. Global Analysis of Membrane-associated Protein Oligomerization Using Protein Correlation Profiling. *Mol. Cell. Proteomics* **2017**, *16* (11), 1972–1989.

(30) Ladwig, F.; Dahlke, R. I.; Stuhrowldt, N.; Hartmann, J.; Harter, K.; Sauter, M. Phytosulfokine regulates growth in Arabidopsis through a response module at the plasma membrane that includes CYCLIC NUCLEOTIDE-GATED CHANNEL17, H⁺-ATPase, and BAK1. *Plant Cell* **2015**, *27* (6), 1718–1729.

(31) Wu, X.; Sanchez-Rodriguez, C.; Pertl-Obermeyer, H.; Obermeyer, G.; Schulze, W. X. Sucrose-induced receptor kinase SIRK1 regulates a plasma membrane aquaporin in Arabidopsis. *Mol. Cell. Proteomics* **2013**, *12* (10), 2856–2873.

(32) Chinchilla, D.; Zipfel, C.; Robatzek, S.; Kemmerling, B.; Nurnberger, T.; Jones, J. D.; Felix, G.; Boller, T. A flagellin-induced complex of the receptor FLS2 and BAK1 initiates plant defence. *Nature* **2007**, *448* (7152), 497–500.

(33) Schlesier, B.; Breton, F.; Mock, H. P. A hydroponic culture system for growing Arabidopsis thaliana plantlets under sterile conditions. *Plant Mol. Biol. Rep.* **2003**, *21*, 449–456.

(34) Rappsilber, J.; Ishihama, Y.; Mann, M. Stop And Go Extraction tips for matrix-assisted laser desorption/ionization, nano-electrospray, and LC-MS sample pretreatment in proteomics. *Anal. Chem.* **2003**, *75* (3), 663–670.

(35) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26* (12), 1367–1372.

(36) Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **2011**, *10* (4), 1794–1805.

(37) Cox, J.; Hein, M. Y.; Lubner, C. A.; Paron, I.; Nagaraj, N.; Mann, M. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* **2014**, *13*, 2513–2526.

- (38) Thimm, O.; Bläsing, O.; Gibon, Y.; Nagel, A.; Meyer, S.; Kruger, P.; Selbig, J.; Muller, L. A.; Rhee, S. Y.; Stitt, M. MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.* **2004**, *37* (6), 914–939.
- (39) Tanz, S. K.; Castleden, I.; Hooper, C. M.; Vacher, M.; Small, I.; Millar, H. A. SUBA3: a database for integrating experimentation and prediction to define the SUBcellular location of proteins in Arabidopsis. *Nucleic Acids Res.* **2013**, *41*, D1185–91.
- (40) Szklarczyk, D.; Morris, J. H.; Cook, H.; Kuhn, M.; Wyder, S.; Simonovic, M.; Santos, A.; Doncheva, N. T.; Roth, A.; Bork, P.; Jensen, L. J.; von Mering, C. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* **2017**, *45* (D1), D362–D368.
- (41) Huala, E.; Dickerman, A. W.; Garcia-Hernandez, M.; Weems, D.; Reiser, L.; LaFond, F.; Hanley, D.; Kiphart, D.; Zhuang, M.; Huang, W.; Mueller, L. A.; Bhattacharyya, D.; Bhaya, D.; Sobral, B. W.; Beavis, W.; Meinke, D. W.; Town, C. D.; Somerville, C.; Rhee, S. Y. The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res.* **2001**, *29* (1), 102–105.
- (42) Haruta, M.; Sabat, G.; Stecker, K.; Minkoff, B. B.; Sussman, M. R. A peptide hormone and its receptor protein kinase regulates plant cell expansion. *Science* **2014**, *343*, 408–411.
- (43) Aryal, U. K.; McBride, Z.; Chen, D.; Xie, J.; Szymanski, D. B. Analysis of protein complexes in Arabidopsis leaves using size exclusion chromatography and label-free protein correlation profiling. *J. Proteomics* **2017**, *166*, 8–18.
- (44) 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.; Benfey, P. N. The protein expression landscape of the Arabidopsis root. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (18), 6811–6818.
- (45) Arnold, T.; Linke, D. The Use of Detergents to Purify Membrane Proteins. *Curr. Protoc. Protein Sci.* **2008**, *53* (1), 4.8.1–4.8.30.
- (46) Seddon, A. M.; Curnow, P.; Booth, P. J. Membrane proteins, lipids and detergents: not just a soap opera. *Biochim. Biophys. Acta, Biomembr.* **2004**, *1666* (1–2), 105–117.
- (47) Lee, Y. C.; Baath, J. A.; Bastle, R. M.; Bhattacharjee, S.; Cantoria, M. J.; Dornan, M.; Gamero-Estevez, E.; Ford, L.; Halova, L.; Kernan, J.; Kurten, C.; Li, S.; Martinez, J.; Sachan, N.; Sarr, M.; Shan, X.; Subramanian, N.; Rivera, K.; Pappin, D.; Lin, S. H. Impact of Detergents on Membrane Protein Complex Isolation. *J. Proteome Res.* **2018**, *17* (1), 348–358.
- (48) Demir, F.; Hontrich, C.; Blachutzik, J. O.; Scherzer, S.; Reinders, Y.; Kierszniowska, S.; Schulze, W. X.; Harms, G. S.; Hedrich, R.; Geiger, D.; Kreuzer, I. Arabidopsis nanodomain-delimited ABA signaling pathway regulates the anion channel SLAH3. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110* (20), 8296–8301.
- (49) Myers, C.; Romanowsky, S. M.; Barron, Y. D.; Garg, S.; Azuse, C. L.; Curran, A.; Davis, R. M.; Hatton, J.; Harmon, A. C.; Harper, J. F. Calcium-dependent protein kinases regulate polarized tip growth in pollen tubes. *Plant J.* **2009**, *59*, 528–539.
- (50) Szymanski, W. G.; Zauber, H.; Erban, A.; Wu, X. N.; Schulze, W. X. Cytoskeletal components define protein location to membrane microdomains. *Mol. Cell. Proteomics* **2015**, *14*, 2493.
- (51) Kleine-Vehn, J.; Wabnik, K.; Martinière, A.; Łangowski, Ł.; Willig, K.; Naramoto, S.; Leitner, J.; Tanaka, H.; Jakobs, S.; Robert, S.; Luschnig, C.; Govaerts, W.; Hell, S. W.; Runions, J.; Friml, J. Recycling, clustering, and endocytosis jointly maintain PIN auxin carrier polarity at the plasma membrane. *Mol. Syst. Biol.* **2011**, *7*, 540.
- (52) Luschnig, C.; Vert, G. The dynamics of plant plasma membrane proteins: PINs and beyond. *Development* **2014**, *141* (15), 2924–38.
- (53) Konrad, S. S.; Ott, T. Molecular principles of membrane microdomain targeting in plants. *Trends Plant Sci.* **2015**, *20* (6), 351–61.
- (54) Zauber, H.; Burgos, A.; Garapati, P.; Schulze, W. X. Plasma membrane lipid–protein interactions affect signaling processes in sterol-biosynthesis mutants in Arabidopsis thaliana. *Front. Plant Sci.* **2014**, *5*, 78.
- (55) Pertl-Obermeyer, H.; Schulze, W. X.; Obermeyer, G. In vivo cross-linking combined with mass spectrometry analysis reveals receptor-like kinases and Ca²⁺ signalling proteins as putative interaction partners of pollen plasma membrane H⁺ ATPases. *J. Proteomics* **2014**, *108*, 17.
- (56) Fuglsang, A. T.; Guo, Y.; Cuin, T. A.; Qiu, Q.; Song, C.; Kristiansen, K. A.; Bych, K.; Schulz, A.; Shabala, S.; Schumaker, K. S.; Palmgren, M. G.; Zhu, J. K. Arabidopsis protein kinase PKSS inhibits the plasma membrane H⁺-ATPase by preventing interaction with 14–3-3 protein. *Plant Cell* **2007**, *19* (5), 1617–1634.
- (57) Fuglsang, A. T.; Kristensen, A.; Cuin, T.; Schulze, W. X.; Persson, J.; Thuesen, K. H.; Ytting, C. K.; Oehlschlaeger, C.; Mahmood, K.; Sondergaard, T. E.; Shabala, S.; Palmgren, M. G. Receptor kinase mediated control of primary active proton pumping at the plasma membrane. *Plant J.* **2014**, *80* (6), 951–964.
- (58) Franceschini, A.; Szklarczyk, D.; Frankild, S.; Kuhn, M.; Simonovic, M.; Roth, A.; Lin, J.; Minguetz, P.; Bork, P.; Von Mering, C.; Jensen, L. J. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* **2013**, *41*, D808–D815.

Second Publication

Author contribution disclosure:

Max Gilbert: performed the proteomics experiments, data analysis, network analysis and wrote the manuscript

Zhi Li: contributed unpublished materials (construct and plasmid templates)

Dr. Xu Na Wu: contributed pull-down data sets

Leander Rohr: performed FRET/FLIM experiments and corresponding data evaluation

Gombos Sven: assisted with the performance of the root growth experiments

Prof. Klaus Harter: interpreted the FRET/FLIM experiments and participated in writing of the manuscript

Prof. Waltraud Schulze: conceived the study, was involved in data interpretation and wrote the manuscript

Reprinted under creative common licence.

Original publication as referenced:

Gilbert, M., et al., *Comparison of path-based centrality measures in protein-protein interaction networks revealed proteins with phenotypic relevance during adaptation to changing nitrogen environments*. Journal of Proteomics, 2021. **235**: p. 104114 [111].

 31.5.2021

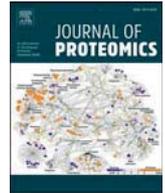
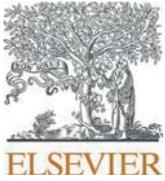
Signature, Date

Prof. Waltraud Schulze; Supervisor

 17.05.2021

Signature, Date

Max Gilbert; Researcher



Comparison of path-based centrality measures in protein-protein interaction networks revealed proteins with phenotypic relevance during adaptation to changing nitrogen environments

Max Gilbert^a, Zhi Li^a, Xu Na Wu^c, Leander Rohr^b, Sven Gombos^a, Klaus Harter^b, Waltraud X. Schulze^{a,*}

^a Department of Plant Systems Biology, University of Hohenheim, 70593 Stuttgart, Germany

^b Center for Plant Molecular Biology, University of Tübingen, 72076 Tübingen, Germany

^c School of Life Science, Yunnan University, 650091 Kunming, China

ARTICLE INFO

Keywords:

Dynamic protein complexes
Network path analysis
Nitrate starvation

ABSTRACT

Plants must rapidly adapt to changes in nutrient conditions. Especially adaptations to changing nitrogen environments are very complex involving also major adjustments on the protein level. Here, we used a size-exclusion chromatography-coupled to mass spectrometry approach to study the dynamics of protein-protein interactions induced by transition from full nutrition to nitrogen starvation. Comparison of interaction networks established for each nutrient condition revealed a large overlap of proteins which were part of the protein-protein interaction network, but that same set of proteins underwent different interactions at each treatment. Network topology parameter betweenness centrality (BC) was found to best reflect the relevance of individual proteins in the information flow within each network. Changes in BC for individual proteins may therefore indicate their involvement in the cellular adjustments to the new condition. Based on this analysis, a set of proteins was identified showing high nitrogen-dependent changes in their BC values: The receptor kinase AT5G49770, co-receptor QSK1, and proton-ATPase AHA2. Mutants of those proteins showed a nitrate-dependent root growth phenotype. Individual interactions within the reconstructed network were tested using FRET-FLIM technology. Taken together, we present a systematic strategy comparing dynamic changes in protein-protein interaction networks based on their network parameters to identify regulatory nodes.

Significance: Protein-protein interactions are known to be important in cellular signaling events, but the dynamic changes in interaction networks induced by external stimuli are still rarely studied. We systematically analyzed how changes in the nutrient environment induced a rewiring of protein-protein interactions in roots. We observed small changes in overall protein abundances, but instead a rewiring of pairwise protein-protein interactions. Betweenness centrality was found to be the optimal network topology parameter to identify protein candidates with high relevance to the information flow in the (dynamic) network. Predicted interactions of those relevant nodes were confirmed in FLIM/FRET experiments and in phenotypic analysis. The network approach described here may be a useful application in dynamic network analysis more generally.

1. Introduction

The dynamic assembly and disassembly of protein-protein interactions upon external and internal signals is a well-conceived concept. However, it is much more difficult to gain experimental evidence. Over the past decades the dynamic nature of individual protein-protein interactions has been elucidated. Especially transporters in the plasma membrane, which are involved in the flow of nutrients and ions

across the cellular barrier, undergo condition-dependent protein-protein interactions with regulatory proteins such as kinases and phosphatases. Well studied examples are the ion channel SLAH3, which was described to interact with an activating kinase CPK21 upon stimulation with the phytohormone abscisic acid, but in absence of the hormone forms an interaction with a phosphatase [1]. Also the plasma membrane ATPase AHA2 is known to interact with different kinases under different conditions to be phosphorylated at activating or inhibiting phosphorylation

* Corresponding author.

E-mail address: wschulze@uni-hohenheim.de (W.X. Schulze).

<https://doi.org/10.1016/j.jprot.2021.104114>

Received 20 October 2020; Received in revised form 21 December 2020; Accepted 6 January 2021

Available online 14 January 2021

1874-3919/© 2021 The Authors.

Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license

(<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

sites [2]. AHA2 is one of the most abundant proteins within the plasma membrane of roots and essential in control of the membrane potential and a proton gradient for nutrient uptake. Activating interactions include the interaction with receptor kinase PSKR1, which activates the proton pump in presence of the hormone phyto­sulfokine [3]. By contrast, interaction of AHA2 with FERONIA results in phosphorylation at an inhibitory serine and inactivation of the pump [4].

Two of the most prominent and widely used techniques for the systematic and targeted identification of protein interaction partners are the yeast-two-hybrid system and the affinity purification of proteins. Both methods [5,6] are now widely used in large-scale biology. However, the dynamic nature of protein-protein interactions *in planta* cannot be addressed using heterologous expression systems of the two-hybrid systems. Thus, researchers identify interaction partners to selected target proteins by affinity purification combined with quantitative mass spectrometry. By this method, for example, conditional interaction partners to receptor kinase SIRK1 were identified [7], or conditional interactors of cell cycle components were characterized [8]. The interactome of aquaporins was recently studied by enrichment of protein complexes from plants grown at different stresses, proposing new candidate proteins for regulation of the water channels [9]. However, the system-wide, *in planta* study of the dynamic nature of protein-protein interaction has not been widely addressed so far.

Dynamic re-arrangements of protein-protein interactions are especially to be expected when plants are exposed to a change in nutrient conditions, for example nitrogen. Nitrogen, as a component of protein, nucleic acids and thousands of metabolites, is an essential macronutrient for plants and is mainly taken up in the form of nitrate. The cellular responses to nitrate starvation or to the resupply of nitrate to starved plants was in the past especially well studied at the gene transcriptional level [10,11]. At the protein level, the change of phosphorylation status was quantified in nitrate starvation-resupply experiments [12], and in response to nitrate starvation [13]. However, global insights into dynamic changes at the level of protein-protein interactions are still missing. Therefore, in this work, we apply a recently published size-exclusion-chromatography coupled to mass spectrometry (SEC-MS) approach [14] to compare the protein-protein interaction network of plants grown under full nutrition and plants exposed to nitrogen starvation. We here aim to (i) define a workflow of comparing interaction networks and (ii) develop measures to identify key proteins involved in regulation of nitrogen starvation response.

2. Materials and methods

Plant growth conditions – All experiments were done using *Arabidopsis thaliana* (Col-0). Plant seedlings were grown in a hydroponic system [15]. Plants under full nutrition were grown in liquid half-strength Murashige and Skoog (½ MS) growth medium with 0.5% sucrose for three weeks before harvesting the full nutrition sample. Plants for nitrogen starvation were firstly grown in half-strength Murashige and Skoog growth medium supplemented with 2 mM KNO₃ for 14 days, and then transferred to 1 mM KNO₃ for 7 days. After three weeks of growth on nitrate supplemented medium, plants were then transferred to nitrogen-free half-strength MS-medium for two days before harvesting the N-starvation sample. For the nitrogen resupply condition, plants were firstly starved for nitrogen as described. Then after two days of starvation, nitrate was resupplied to a final concentration of 2 mM KNO₃ directly into the nitrogen-free growth medium for 15 min prior to harvest. In all conditions, the growth medium was supplemented with 0.5% sucrose.

Experimental design and statistical rationale – Two data sets were analyzed, acquired under full nutrition or nitrate starvation. Each data set consisted of two biological replicates of size exclusion chromatography followed by mass spectrometry analysis of the acquired fractions as described [14]. Only those proteins consistently identified in both replicates were considered for comparisons between conditions. A third

experimental condition, in which nitrate was resupplied to nitrate starved plants was included with one biological replicate. Statistical comparisons were carried out by pairwise *t*-tests (two sample comparisons) or ONEWAY ANOVA (multiple sample comparisons). We used the MaxQuant / Perseus data analysis platform for quantitative analysis in all proteomics experiments.

Microsomal protein enrichment – A total of 1 to 1.5 g of roots (fresh weight) was homogenized in 10 ml ice-cold extraction buffer (330 mM sucrose, 100 mM KCl, 1.5 mM EDTA, 50 mM Tris-MES, fresh 5 mM DTT, and 1 mM phenylmethylsulfonylfluoride, pH 7.5) [16] in the presence of 0.5% v/v proteinase inhibitor mixture (Sigma-Aldrich, Germany) and phosphatase inhibitors (50 mM NaF, 1 mM Na₃VO₄, 1 mM benzamidin, 4 μM proteinase inhibitor leupeptin). The homogenate was centrifuged for 15 min at 7500 ×g at 4 °C. The pellet was discarded, and the supernatant was centrifuged for 75 min at 48,000 ×g at 4 °C. Microsomal pellets were stored at –80 °C until further processing. The microsomal pellet was resuspended in 100 μl of membrane buffer (137 mM NaCl, 2.7 mM KCl, 10 mM Na₂HPO₄, 1.8 mM KH₂PO₄; pH = 7.4, 0.1% n-Dodecylmaltoside) or UTU (6 M urea, 2 M thiourea, pH 8) supplemented with 0.1% n-Dodecyl β-D-maltoside (DDM). Insolubilized membrane fragments were pelleted before further processing of solubilized membrane proteins.

For isolating membrane proteins in their native form, it is important to isolate membranes as intact vesicles. Sonication, French-Press or similar methods for plant-protein-purification use brute force to disrupt the cell wall and membranes to gain access to soluble proteins. This often results in a poor yield of membrane proteins as they are dependent of their membrane environment. This problem was avoided by using a Glass/PTFE Potter Elvehjem Tissue Grinder. Through careful grinding of the plant material, small membrane-vesicles form up which can be purified by differential centrifugation as described [17,18]. The use of DDM in solubilization of the membrane pellet was essential for the experiment as it was shown that DDM enhances stability and preserves activity of membrane-proteins during purification [19,20].

Size-exclusion chromatography followed by mass spectrometry – Protein-protein interactions were defined using a SEC-MS approach as described [14]. A total amount of 1 mg microsomal extract was run on an Enrich SEC650 (BioRad) column using 1 column volume of phosphate buffered saline (PBS). Eluting protein fractions were collected in 50 dynamic fractions. The volume of earlier fractions was 25 μL and 50 μL for later fractions to maximize functional chromatographic resolution and to correct for higher protein content in earlier fractions [14]. Equal protein amount from each fraction was digested by LysC/trypsin and desalted over C18 as described [7].

LC-MS/MS analysis of peptides – Peptides mixtures were analyzed by nanoflow Easy-nLC (Thermo Scientific) and Orbitrap hybrid mass spectrometer (Q-exactive, Thermo Scientific). Peptides were eluted from a 75 μm × 50 cm analytical C₁₈ column (PepMan, Thermo Scientific) on a linear gradient running from 4% to 64% acetonitrile over 135 min. Proteins were identified based on the information-dependent acquisition of fragmentation spectra of multiple charged peptides. Up to twelve data-dependent MS/MS spectra were acquired in the linear ion trap for each full-scan spectrum acquired at 70,000 full-width half-maximum (FWHM) resolution.

Protein identification and label-free quantitation – MaxQuant version 1.5.3.8 [21] was used for raw file peak extraction and protein identification against the *Arabidopsis* proteome (TAIR10, 35,386 entries). Protein quantification was performed in MaxQuant using the label free quantification (LFQ) algorithm [22]. The following parameters were applied: trypsin as cleaving enzyme; minimum peptide length of seven amino acids; maximal two missed cleavages; carbamidomethylation of cysteine as a fixed modification; N-terminal protein acetylation, oxidation of methionine as variable modifications. Peptide mass tolerance (4.5 ppm) and MS/MS tolerance (20 ppm) were used according to the standard settings in MaxQuant. Further settings were: “label-free quantification” marked, multiplicity set to 1; “match between runs”

marked with time window 2 min; peptide and protein false discovery rates (FDR) set to 0.01; common contaminants (trypsin, keratin, etc.) excluded. Label-free protein quantitation [22] was applied using the software Perseus and based on the “proteingroups.txt” output table of MaxQuant. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [23] partner repository with the dataset identifier PXD009511 for the full-nutrition data set, PXD015841 for the nitrate starvation and resupply data. Protein identification information is listed in Supporting Table S6.

Definition of protein-protein interactions – Interacting protein pairs were defined as described [14]. Briefly, proteins exhibiting an earlier-than expected retention pattern of at least six fractions during SEC-chromatography were defined as nodes in the protein-protein interaction network. Interaction partners to those node proteins were defined by pair-wise correlation of elution profiles among the node proteins. Thus, the edges in the protein-protein interaction networks were defined by pairwise comparison of elution profiles and the requirements for co-elution same fraction. We used publicly available databases STRING [24] and MIND [25] of confirmed protein-protein interactions to define optimal filtering criteria and cutoff scores. High confidence interactions were defined as those the interaction pairs which eluted with less than one fraction difference and stricter correlation score cutoff [14].

Protein abundance quantitation – LFQ-values [22] across multiple samples were used to compare protein abundance between different experiments. We normalize the data as follows: Firstly, within each replicate the relative abundance (relLFQ, Eq. [1]) for each protein was obtained by dividing the summarized protein LFQ-values by the total LFQ-sum of that replicate.

$$\text{relLFQ}_{\text{Protein}} = \frac{\text{LFQ}_{\text{Protein Sum}}}{\text{LFQ}_{\text{Experiment Sum}}} \quad (1)$$

Secondly, relative protein abundances were weighted against the respective protein mass in Dalton. This was done to compensate for differing identification probabilities due to different protein sizes (Eq. [2]).

$$\text{Abundance}_{\text{weighted}} = \frac{\text{relLFQ}_{\text{Protein}}}{\text{Molecular Weight [Dalton]}} \quad (2)$$

Thirdly, proteins were sorted according to their weighted abundances (AW_{LFQ}) and proteins were divided into 100 dynamic abundance categories (AC_{LFQ}) for each experiment to account for different numbers of proteins identified in different replicates. Category size depended on the total number of proteins in each replicate. By binning all proteins into a fixed number of abundance categories enabled the observation of protein abundance changes across experiments.

Network topology analysis – Networks were constructed with Cytoscape 3.7.1 and evaluated with the Cytoscape network analyzer tool regarding node parameters degree, betweenness centrality, closeness centrality, stress. Introduced in 1950 [26], Closeness Centrality (CC, Eq. [3]) can be used to describe the efficiency of a node for distributing information to other members of the network. It is defined as the reciprocal of the average shortest path length (σ) for a node i to all other nodes j , where n is the total number of nodes:

$$\text{CC}(i) = \frac{1}{\sum_{j \neq i}^n \sigma_{ij}} \quad (3)$$

Since CC scales with the total number of nodes, it was normalized by using the average length of the shortest paths instead of their sum (Eq. [4]):

$$\text{CC}(i)' = \frac{1}{\text{average} \left(\sum_{j \neq i}^n \sigma_{ij} \right)} \quad (4)$$

While a high CC is a good indicator for a node that has broad access

to all other nodes, it cannot be used to describe the potential influence on the flow of information. For that purpose, Stress Centrality can be used. It is defined as the number of shortest paths passing through a node i (Eq. [5]). By this definition, a node gains importance if it is facilitating the flow of information between other nodes. This parameter can only be defined for networks without multiple edges like protein-interaction networks.

$$\text{SC}(i) = \sum_{s \neq i \neq t} \sigma_{st}(i) \quad (5)$$

Betweenness Centrality (BC) [27] also assigns importance to nodes that are in the path of other nodes, by describing each node's power over the flow of information. It builds on the definition of Stress Centrality and is defined as the sum of all shortest paths (σ) between nodes j and t passing through node i ($\sigma_{jt}(i)$) divided by all possible shortest paths σ_{jt} (Eq. [6]).

$$\text{BC}(i) = \sum_{j \neq i \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}} \quad (6)$$

BC also scales linearly with network size. To account for scaling, the BC values can be normalized by dividing through the maximum value of a given network. For an undirected network with the size n , the division is further multiplied by the factor 2 (Eq. [7]):

$$\text{BC}(i)' = \frac{2 \sum_{s \neq i \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}}}{(n-1)(n-2)} \quad (7)$$

A high BC value is usually assigned to nodes that connect different network communities. In this way, BC offers more unique and sensitive information about local network topology than focusing on degree based metrics alone.

A metric to describe modular organization is the clustering coefficient [28]. For undirected networks, it is defined as the number of connected pairs between all neighbors of a node i , where e_n is the number of connected pairs and k_n is the number of neighbors (Eq. [8]).

$$\text{C}(n) = \frac{2 e_n}{(k_n(k_n - 1))} \quad (8)$$

Random networks – Random networks were generated containing the same numbers of nodes and edges as respective protein-protein interaction networks. Edges were randomly assigned between the nodes.

Root growth assay – Primary root length and lateral root density of wildtype and mutant plants were assessed by growth of seedlings on vertical plates of MES-buffered ½ MS-Agar (0.25 g l^{-1} MES, 0.39 g l^{-1} N-Free MS, 0.5% Sucrose, 8 g l^{-1} agar, supplemented with KNO_3 when needed). All plates were supplemented with 30 mM sucrose. After sterilization (incubation in 2% Sodium-Hypochlorite solution for 12 min and subsequent washing), the seeds (5 seeds per plate) were spread on ½ MS-agar plates and incubated for 2 days at 6 °C for vernalization. After vernalization, the seeds were germinated for 4 days at room temperature after which they were transferred to the treatment plates (5 mM KNO_3 , 0.3 mM KNO_3 , 0 mM KNO_3). Each plate was scanned with a desktop office scanner (600dpi) at the start of the treatment phase and after 5 days of growth. Images were evaluated with the ImageJ software v.1.52a (imagej.nih.gov/ij).

Pull-downs of GFP-tagged SIRK1 and QSK1 – Root microsomal proteins (150 μg) isolated as described above was incubated with 25 μl of anti-GFP agarose beads (Chromotek, Germany) for two hours on a rotating wheel at 4 °C [29]. After incubation, the beads were collected by centrifugation and washed two times with 500 μl wash buffer (10 mM Tris/HCl pH 7.5, 150 mM NaCl, 0.5 mM EDTA, 0.01% IGEPAL). For protein-protein interaction assays, the proteins were eluted from the beads with 100 μl UTU (6 M urea, 2 M thiourea), pH 8, before in-solution tryptic digestion.

FLIM/FRET – For FRET-FLIM analysis, the coding sequences of BAK1

and AHA2 as well as of QSK1 and At5g49770 were expressed as C-terminal fluorophore fusions in 2in1 vectors, namely pFRETcg-2in1-CC [30]. To obtain GFP-donor only controls, a coding sequence of gentamycin fused to mCherry was used. The binary vectors and p19 as gene silencing suppressor were transformed into *Agrobacterium tumefaciens* strain GV3101 and infiltrated into *Nicotiana benthamiana* leaves. The measurements were performed two to three days after infiltration using a SP8 laser scanning microscope (Leica Microsystems) with LAS AF and SymPhoTime software as described [31]. Before performing the FRET-FLIM measurement, the presence of the fluorophores was detected by using the 488 or 561 nm lasers for GFP or mCherry excitation, respectively. The fluorescence lifetime in nanoseconds of either the donor only expressing cells or the cells expressing the indicated combinations was measured with a pulsed laser as an excitation light source of 470 nm and a repetition rate of 40 MHz (PicoQuant Sepia Multichannel Picosecond Diode Laser, PicoQuant Timehart 300 TCSPC Module and Picosecond Event Timer). The acquisition was performed until 500 photons in the brightest pixel were reached. To obtain the GFP fluorescence lifetime, data processing was performed with SymPhoTime software and bi-exponential curve fitting as well as correction for the instrument response function. A total range of 23 ns was evaluated. Statistical analysis was carried out with JMP 14. To test for homogeneity of variance, a Brown-Forsythe test was applied. As the variances are unequal, a Kruskal-Wallis test followed by Dunn's post-hoc test were performed.

Public resources – Functional classification of proteins was done based on MAPMAN [32]. Information about subcellular location was derived from SUBA3 [33]. Detailed protein function was manually updated with the support of TAIR [34].

Data visualization – Functional classification of proteins was done based on MAPMAN [32]. Information about subcellular location was derived from SUBA3 [33]. Detailed protein function was manually updated with the support of TAIR [34]. Other statistical analyses were carried out with Sigma Plot (version 11.0) and Excel (Microsoft, 2013). Over-representation analysis was done via Fisher's exact test, *p* values were adjusted using Bonferroni correction. Networks were visualized using Cytoscape [35].

3. Results

3.1. Protein abundance dynamics

Different growth conditions can not only induce dynamics in protein-protein interactions, but can affect overall protein abundance. To address this, we used normalized LFQ-values to describe protein abundance changes. The normalized LFQ values were weighted by protein molecular weight, allowing abundance ranking of all proteins (weighted abundance, AW_{LFQ}). To be able to compare the different data sets and replicates which contained different total numbers of proteins, we defined 100 dynamic abundance categories (AC_{LFQ}) within each replicate. Each abundance category contained different numbers of proteins, but reflected the same relative abundance within the sample. For example: category 1 contained the 1% most abundant proteins, and this category contained 35 proteins in the full nutrition data set, and 22 proteins in the nitrogen starvation data set. To compare relative protein abundance changes between the two conditions, the change in AC_{LFQ} for each protein was taken as a readout.

Most proteins identified in two biological replicates of the same nutrient condition showed an abundance category variation of up to 10 categories, and over 80% of the proteins showed between replicate variations of less than 20 categories, independently of the nutrient condition (Fig. 1A). Based on this observation, we only considered those proteins that showed a change of less than 20 abundance categories between replicates for further comparison between conditions. In general, ribosomal proteins were among the most abundant proteins identified under both nutrient conditions (Supporting Table 1). The most abundant 10% of proteins were furthermore highly enriched for

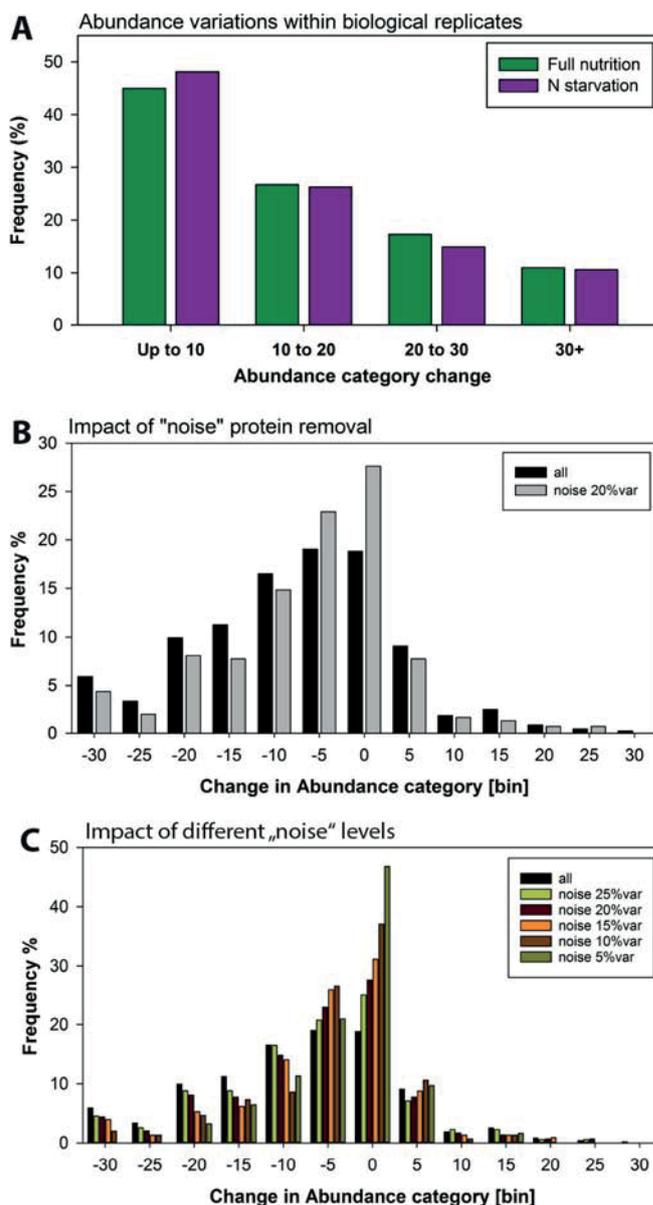


Fig. 1. Protein abundance in replicates. (A) Abundance variation within two biological replicates of each treatment, full nutrition and nitrogen starvation. Abundance variations of more than 20% were considered “noise”. (B) Changes in abundance category upon nitrogen starvation. Black bars represent the full data set, gray bars represent the data set corrected for those proteins with more than 20% variation between replicates (“noise”). (C) Changes in abundance category upon nitrogen starvation at different levels of variation between replicates (“noise”).

transporters ($p < 2.3 \cdot 10^{-7}$, Fisher Exact Test), particularly aquaporins, and the plasma membrane H^+ -ATPases. The least abundant 10% of proteins were enriched for signaling proteins ($4.4 \cdot 10^{-7}$, Fisher Exact Test), transporters, among them nitrate and ammonium transporters ($8.74 \cdot 10^{-6}$, Fisher Exact Test), and proteins involved in cell organization ($2.53 \cdot 10^{-3}$, Fisher Exact Test). Fig. 1:

In general, nitrogen starvation resulted in a decrease in protein abundance compared to full nutrition (Fig. 1B). Based on the MAPMAN [32] annotation categories (=“bins”), nitrogen starvation led to an increased abundance category for proteins of the “protein” bin (neogenesis, degradation and modification), and “RNA”-bins, while “stress”-, “signaling”- and “transport”- associated bins decreased in abundance categories. Excluding those proteins with change in abundance

categories between biological replicates of greater than 20 emphasized the skewed pattern of protein abundance changes under nitrogen starvation (Fig. 1B). Proteins with strongest change in abundance category (more than 30 abundance categories) upon nitrogen starvation were some transporters, and many other receptor kinases, among them FERONIA (FER, AT3G51550). Despite the tendency for reduced protein abundances under nitrate starvation, the abundance category of the majority of proteins remained unaltered. However, filtering the data set from proteins with abundance category changes of more than 20 categories between replicates (“noise”) reduced the total number of proteins with observed abundance change from 473 to 297 shared candidate proteins.

If the stringency of the “noise”-filter was increased (Fig. 1C), it became obvious that proteins with highly consistent abundance categories between replicates (i.e. change of less than 5 categories between replicates), were also those proteins that remained stable in their abundance category upon nitrogen starvation (Fig. 1C). Among this group we found many constituents of the ribosomes, essential proton-transporting ATPases (AHA2, AAC1/2), porins (e.g. PIP2B, VDAC3), and members of the mitochondrial electron transport chain (e.g. ATP Synthase, Cytochrome C Oxidase).

3.2. Protein-protein interaction networks under full nutrition and nitrogen starvation

Under full nutrition, 3520 proteins were identified in both replicates, and under nitrogen starvation 2280 proteins were found in both replicates (Fig. 2). Indeed, there was even a very high overlap of identified proteins in the two nutrient conditions. 90% of the proteins identified in the nitrogen starvation data set were also identified under full nutrition. After application of quality filter criteria and the quantitation workflow described previously [14], 965 and 677 proteins were proposed as nodes of the protein-protein interaction network under full nutrition and nitrogen starvation, respectively. Fig. 2:

These nodes were those proteins, which based on their elution properties in the SEC-MS experiments, displayed higher than expected

apparent mass and thus were considered to undergo an interaction. We show a high reproducibility of retention time shifts of individual proteins in the two biological replicates (Supporting Fig. S1 A,B) for full nutrition condition [14], as well as for nitrogen starvation condition. Since the definition of nodes in the interaction network was based on higher than expected retention times, the reproducible retention time shifts indicate robust detection of interaction nodes. Correlation analysis [14] revealed 1752 high confidence protein-protein interactions at full nutrition and 1375 interactions under nitrogen starvation. However, the overlap of identified high confidence interactions was rather low (7 to 9%) between the two nutrient conditions, while we found an overlap of identified interactions between biological replicates of up to 25% [14]. Rather, each condition showed a high number of exclusive interactions, the majority of which were not found in the other condition (Fig. 2). Thus, the same proteins (nodes) seemed to undergo different interaction (edges) in each of the nutritional environments. Overlapping high-confidence interactions were few, but well-characterized. These high-confidence interactions might therefore be an indicator for well-characterized very stable interactions, which were unaffected by the treatment (Supporting Table 2). Those known and stable interactions mainly related to larger protein complexes, such as ribosomes, or protein complexes of the mitochondrial respiration chain. Specifically, the interactions not affected by the nutrient conditions included constituents of the mitochondrial ATP synthase, chaperonin complexes, ion transporters (mitochondrial ADP/ATP carrier AAC2, AT5G13490 and VOLTAGE DEPENDENT ANION CHANNEL 2, VDAC2, AT5G67500) and redox systems (Cytochrome C1 and FRO1, AT1G01590). Since these interactions were detected independently of the treatment, they are likely to have been previously found in other experiments (i.e. confirmed by previous publications). By contrast, examples for novel (i.e. yet unconfirmed) high confidence interactions were among several LRR-receptor kinases and their substrates. Among these were co-receptor QIAN SHOU KINASE, QSK1 with CHITIN ELICITOR RECEPTOR KINASE 1 (CERK1, AT3G21630) or ABC-Transporter PEN3, AT1G59870. Since interactions of these proteins were found only under specific conditions, it is unlikely these interactions were described previously (i.e. they were

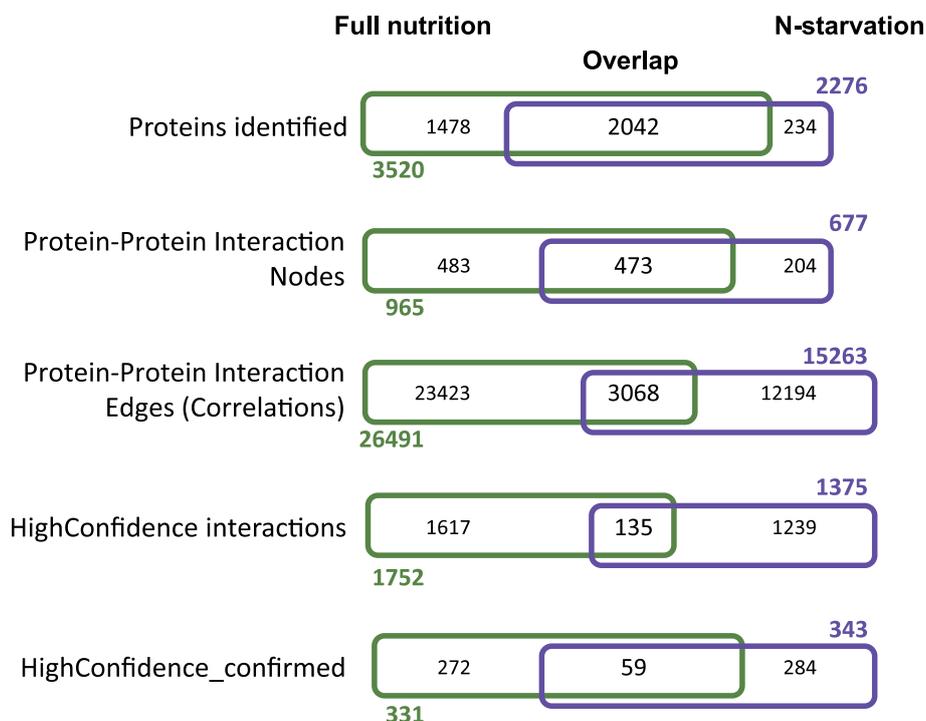


Fig. 2. Overlap of proteins identified and those considered as nodes of protein-protein interaction networks, as well as an overview of their edges in full nutrition and nitrogen starvation.

classified as unconfirmed). We conclude that high-confidence reproducible protein-protein interactions tend to depict those stable protein-protein interactions of larger protein complexes, while dynamic and conditional protein-protein interactions showed a tendency to affect the cellular signaling in pairwise interactions and smaller complexes. The

detection of stable and reproducible protein-protein interactions can be seen as an indication that the cell is able to maintain basic functions, such as protein synthesis, photosynthesis and respiration, or ion homeostasis also under unfavorable conditions of nitrogen starvation. Our results indicate that dynamic interaction rewiring upon nitrogen

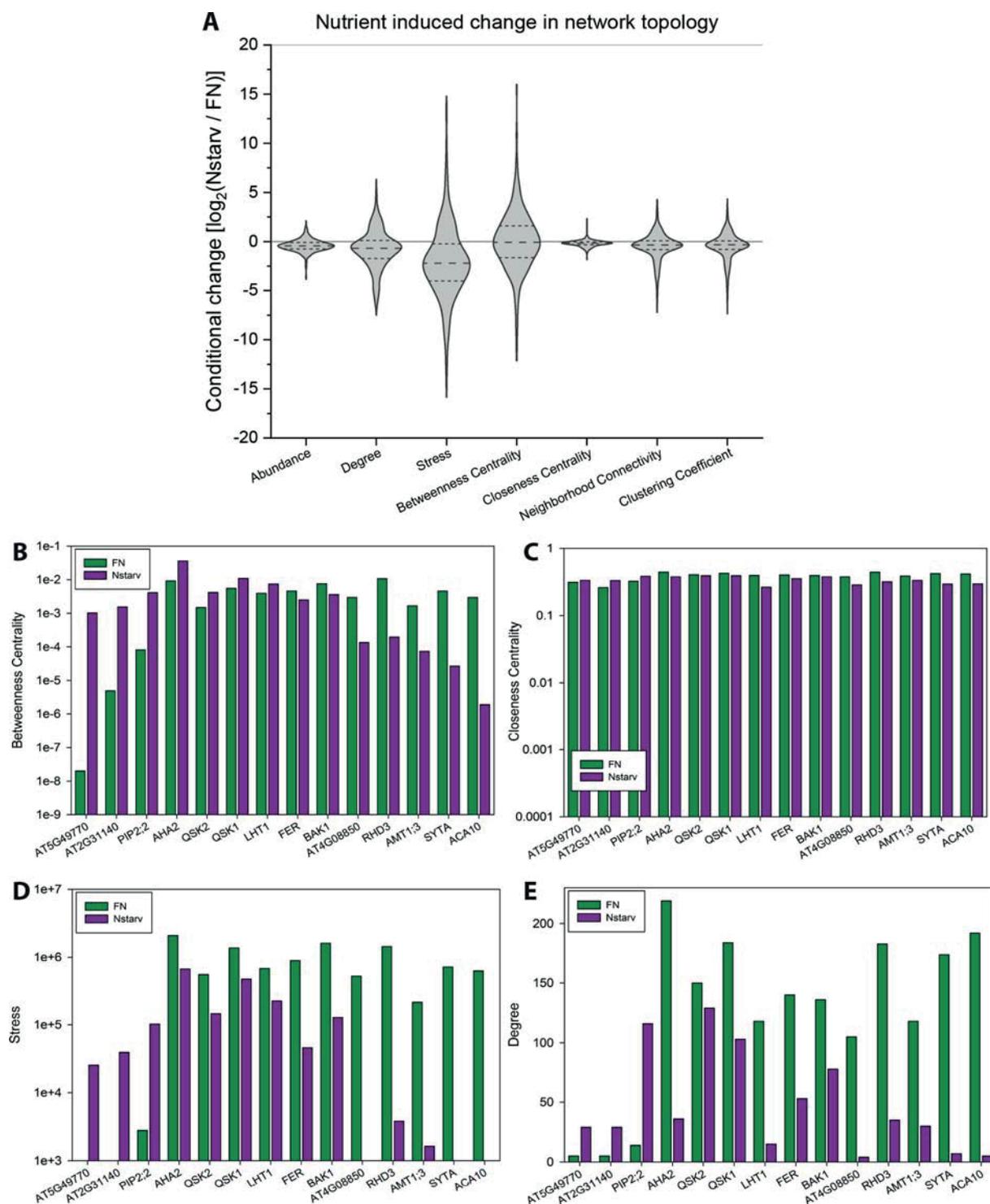


Fig. 3. Network dynamics full nutrition and nitrogen starvation. (A) Violin plots of nitrogen starvation-induced changes of network parameters. Each parameter is displayed as \log_2 ratio of the parameter in the nitrogen-starvation network to the parameter in the full nutrition network. For display of abundances, abundance categories (AC_{LFQ}) were inverse, since higher AC_{LFQ} indicates lower abundance. All network parameters beside Degree and Stress are normalized. Selected network parameters are displayed for individual proteins: (B) Betweenness Centrality, (C) Closeness Centrality, (D) Stresslevel, (E) Degree. Each parameter is displayed for the full nutrition network (green) and the nitrogen starvation network (purple). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

starvation seemed to more frequently affect interactions of proteins with signaling functions (MAPMAN [32] annotation category 'bin' 30) with proteins of cell organization (bin 31), within proteins of signaling functions (bin 30), within transporters (bin 34) and within protein of functions in protein modification, biosynthesis and degradation (bin 29) (Supporting Table S2). As expected, degree distribution of reconstructed protein-protein interaction networks under full nutrition as well as under nitrogen starvation followed a truncated power law [36] (Supporting Fig. S1C). Proteins of these stable complexes of respiratory chain and ribosomes were the proteins with highest degree in the protein-protein interaction network. (Supplementary Fig. S1)

3.3. Network dynamics induced by nitrogen starvation

As indicated above, the nodes of the protein-protein interaction network were defined based on their shifted retention behavior in the size-exclusion chromatography, and edges of the interaction network were derived from (i) pairwise correlations of elution profiles acquired by mass spectrometry and (ii) the requirement of being identified within a narrow window of neighboring elution fractions [37]. Application of this workflow to the full nutrition and nitrogen starvation data set resulted in a protein-protein interaction network in which nodes represent the proteins found with larger apparent size and edges indicate interactions present either at full nutrient supply (Supporting Fig. S2A, green edges) or at nitrogen starvation (Supporting Fig. S2A, purple edges). The network contains two large interaction domains, one dominated by plasma membrane proteins and the other dominated by proteins of cytosolic location. Interactions between proteins of the mitochondrial respiratory chain grouped more distantly from the plasma membrane and cytosolic proteins as revealed by the Mcode algorithm [38].

Examples for proteins with highly dynamic interaction partners are co-receptor SERK3/BAK1 (AT4G33430) (Supporting Fig. S2B), which was found to interact with receptor kinase FER under full nutrition, but under nitrogen starvation switched to an interaction preference with BAK1-INTERACTING RECEPTOR KINASE 2 (BIR2, AT3G28450), IMPAIRED OOMYCE TE SUSCEPTIBILITY 1 (IOS1, AT1G51800) and AT5G49760. (Supplementary Fig. S2)

To obtain a more unbiased view of the dynamic processes induced by nitrogen starvation, we specifically compared network topology parameters. Degree describes the number of interactions of each node. Average Shortest Path describes the average number of steps along the shortest paths for all possible pairs of each node. Closeness Centrality is a measure of centrality of a node within a network and Betweenness Centrality describes the frequency at which a given node occurs among all the possible shortest paths. Stresslevel describes the number of shortest paths leading through a given node. Comparing the nitrogen starvation network with the full nutrition network, it became apparent that nitrogen starvation induced large overall changes in Degree and Stresslevel, indicating a change in overall network size (Fig. 3A). Stresslevel and Betweenness Centrality were the two parameters with largest scatter in the magnitudes of changes induced by nitrogen starvation (Fig. 3A). This indicates that nitrogen starvation induced changes in the utilization of distinct paths during information flow through the network. Among the proteins with strongest increase in Betweenness Centrality in response to nitrogen starvation were protein kinase AT5G49770, a peptidase AT3G55600, respiratory burst oxidase RbohD (AT5G47910), aquaporin PIP2;2 (AT2G37170), and plasma membrane ATPase AHA2 (AT4G30190) (Fig. 3B). Proteins with a strong decrease in Betweenness Centrality upon nitrogen starvation were calcium ATPase ACA10 (AT4G29900), a synaptogamin SYTA (AT2G20990), ammonium transporter AMT1;3 (AT3G24300), GTP binding protein ROOT HAIR DEFICIENT (RHD3, AT3G13870), and receptor kinase MDIS1-INTERACTING RECEPTOR LIKE KINASE 2 (MIK2, AT4G08850) (Supporting Table 3). As already indicated above, Closeness Centrality was unaffected by the nutrient condition (Fig. 3C). Stresslevel was correlated with

Betweenness Centrality, thus proteins with a strong change in Betweenness Centrality also displayed an equally strong increase or decrease in Stresslevel (Fig. 3D). A strong decrease in degree was particularly observed for proteins which showed a decreased Betweenness Centrality upon nitrogen starvation, particularly ammonium transporter AMT1;3, SYNAPTOGAMIN A (SYTA), and calcium ATPase ACA10 (Fig. 3E). Fig. 3:

To test the hypothesis of dynamic adjustments in cellular networks by interaction rewiring and altered path utilization, we analyzed the protein-protein interaction network of nitrogen starved roots resupplied with 2 mM nitrate. By nitrogen re-supply, the network status is expected to restore similarly to the status found under full nutrition. Upon nitrate resupply, we identified 1746 nodes of the protein-protein interaction network. Interestingly, resupply of 2 mM nitrate (Supporting Table 4) to the roots of nitrate starved plants indeed resulted in partial recovery of the interaction network after 15 min. The overall pattern of protein-protein interactions and network parameters was again shifted towards the status at full nutrition. For example, there was a higher overlap of high-confidence interactions predicted for nitrogen resupply with interactions predicted for full nutrition than for interactions predicted under nitrogen resupply with interactions predicted at nitrogen starvation (Fig. 4A). On a more global scale, proteins with increased BC values under N-starvation were enriched for signaling proteins ($p = 0.023$), transport proteins ($p = 9.3 \times 10^{-3}$), lipid metabolism ($p = 0.041$) and proteins in TCA cycle ($p = 2.08 \times 10^{-3}$). Signaling proteins were also enriched ($p = 2.5 \times 10^{-3}$) in proteins with decreased BC valued upon nitrogen starvation, suggesting that indeed signaling proteins are frequently involved in rewiring of information flow in biological networks. Fig. 4:

Moreover, changes in betweenness centrality (BC) induced by nitrogen starvation were inversely proportional to changes induced by nitrate resupply (Fig. 4B). Thus, for a large number of proteins, opposite changes in network path utilization were induced by nitrogen starvation and nitrate resupply. Specifically, these changes affected vesicle trafficking processes, transporters and signaling proteins: For example, the interaction of different SNARE proteins with transporters, such as EARLY RESPONSE TO DEHYDRATION 6 (AT1G08930, ERD6) or SUGAR TRANSPORTER 1 (AT1G11260, STP1), the coatamer complex, or the confirmed interaction of two glutamine synthase proteins GSR2 (AT1G66200) and GLN1.3 (AT3G17820) (Supporting Table 4).

We used randomly generated networks of the same size as the full nutrition interaction network and nitrogen starvation network (Supporting Fig. S3A) to confirm that changes in BC induced by nitrogen starvation or resupply were due to non-random biological effects. In the random network the variation of BC values across proteins was far lower than in the biological network (Supporting Fig. S3B). We conclude that strong changes in BC for individual proteins, such as AT5G49770, were indeed due to different path usage and not due to random effects. (Supplementary Fig. S3)

In our analysis, the nitrate transporters NRT1.1 (AT1G12110) and NRT2.1 (AT1G08090) were among those proteins which were identified with low reproducibility between replicates, and were thus excluded from the large-scale comparative network analysis of centrality indices. NRT1.1 was found as a protein-protein interaction node in two replicates in the full nutrition data set, while NRT2.1 was found as a protein-protein interaction node only in one replicate under nitrate starvation. Nevertheless, nitrate transporters connected to the cellular network, for example through a high confidence interaction found between QSK1 and NRT1.1, or between PERK1 and NRT2.1 (Fig. 4C).

3.4. Specific nitrogen-starvation-induced regulatory modules

Receptor kinase AT5G49770 showed strongest changes in BC-values upon nitrogen starvation (Fig. 3B) and this was reverted upon nitrate resupply. Therefore, we hypothesized that this kinase could have a key role in the adjustment of cellular networks in the response to nitrogen

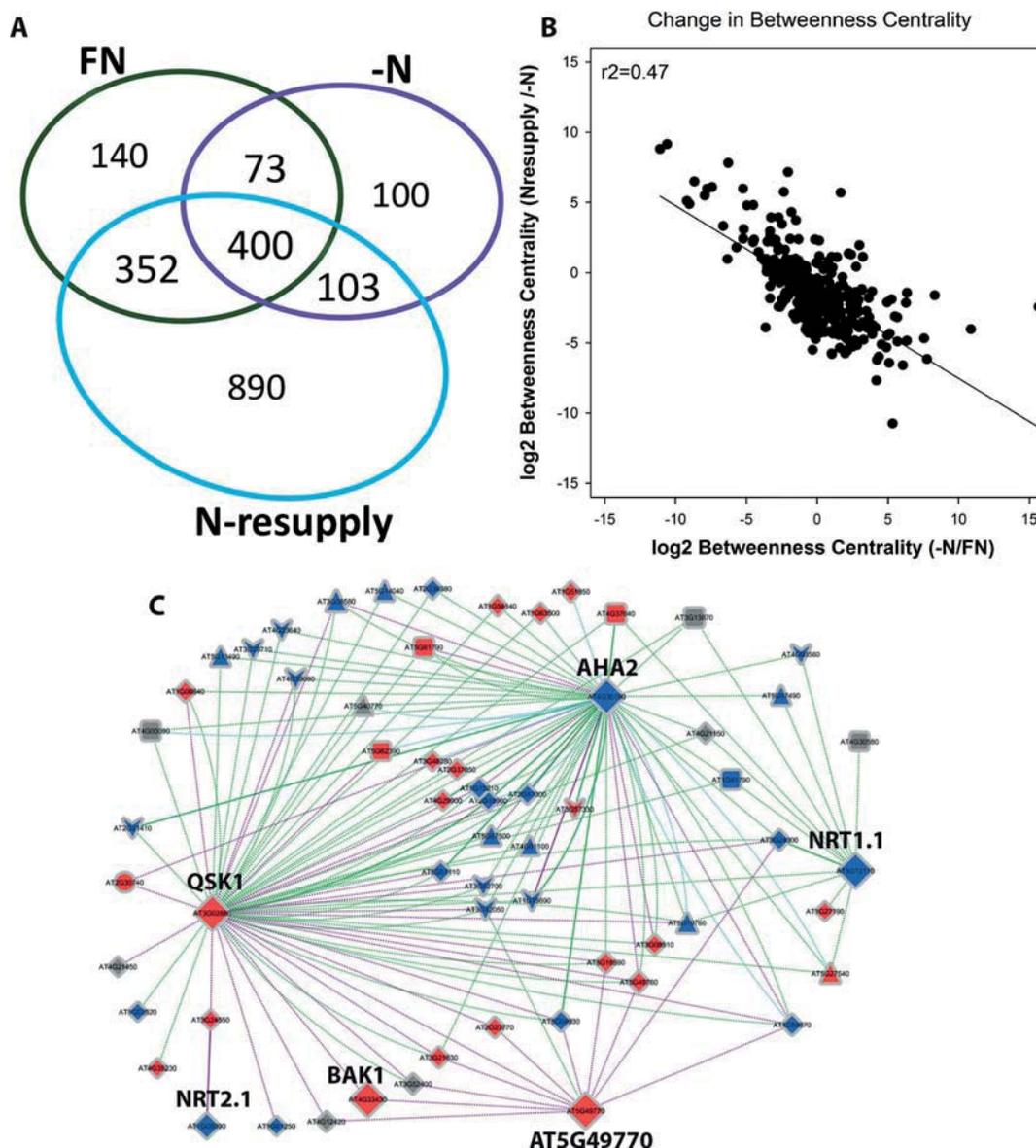


Fig. 4. Network dynamics in response to nutrient conditions. (A) Overlap of protein complex candidates predicted for the three nutrient conditions. In full nutrition (FN) and nitrogen starvation (–N) the complex candidates were derived from the overlap of two biological replicates, for N-resupply only one replicate was considered. (B) Change in Betweenness Centrality upon nitrogen starvation and nitrogen resupply. (C) Detailed network dynamics of nodes AHA2 (AT4G30190), receptor kinase AT5G49770, co-receptor QSK1 (AT3G02880) and nitrate transporters NRT1.1 and NRT2.1. Edge color represents nutrient status: green, full nutrition; purple, nitrogen starvation; cyan, nitrate resupply after starvation. Solid edges represent confirmed interactions based on public literature, dotted edges represent unconfirmed interactions. Node color represents protein function: blue, transporters; red, signaling proteins. Node size reflects betweenness centrality value changes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

starvation. The reconstructed interaction networks proposed an interaction of AT5G49770 with a co-receptor QSK1 (AT3G02880) specifically under nitrogen starvation. The plasma membrane ATPase AHA2 (AT4G30190), which is well known to be activated under nitrogen starvation to allow enhanced uptake of nitrate ions, was also connected to the two kinases AT5G49770 and QSK1 (Fig. 4C, Supporting Table 5). QSK1 as a coreceptor in the reconstructed networks showed direct high-confidence interactions with receptor kinase AT5G49770, NRT1.1, and BAK1. The interaction of QSK1 with AHA2 was considered as low-confidence interaction under full nutrition and the interaction of AT5G49770 with BAK1 as found as low-confidence interaction at nitrate starvation (Fig. 4C).

We next confirmed selected pairs of these proposed interactions in the network by co-immunoprecipitation or FLIM-FRET experiments. Firstly, pull-downs of QSK1-GFP in comparison to GFP only at re-supply

of 0.2 mM nitrate after nitrogen starvation revealed high-confidence interactions of QSK1 with AT5G49770, NRT1.1 and BAK1, but also with AHA2 (Fig. 5A). Secondly, FLIM-FRET experiments were performed to test interaction of QSK1 with AHA2 and BAK1, as well as of AT5G49770 with AHA2 (Fig. 5B, Supplementary Fig. S4). In cases where no interactions should take place (donor-GFP alone controls), the fluorescent lifetime is expected to be higher than for those protein pairs which show an interaction (Fig. 5B). QSK1 and BAK1 were found to interact with significant difference to the respective control, while for low-confidence interaction of QSK1 with AHA2, the fluorescent lifetime was found in between the control and the high-confidence interactor. The interaction of AT5G49770 with AHA2, which was not proposed as direct interaction in the network reconstruction, was used as a control of a negative interaction with fluorescent lifetime values not different from the respective control (Fig. 5B). The FLIM-FRET experiments very well

reflected the interactions proposed by our SEC-MS analytical workflow: QSK1, BAK1 and AT5G49770 show strong correlations in their elution profiles (Supplementary Fig. S5), while plasma membrane ATPase AHA2 showed a different elution profile which did not suggest direct interaction with QSK1 or AT5G49770. Fig. 5:

We next confirmed selected pairs of these proposed interactions in the network by co-immunoprecipitation or FLIM-FRET experiments. Firstly, pull-downs of QSK1-GFP in comparison to GFP only at re-supply of 0.2 mM nitrate after nitrogen starvation revealed high-confidence interactions of QSK1 with AT5G49770, NRT1.1 and BAK1, but also with AHA2 (Fig. 5A). Secondly, FLIM-FRET experiments were performed to test interaction of QSK1 with AHA2 and BAK1, as well as of AT5G49770 with AHA2 (Fig. 5B, Supplementary Fig. S4). In cases where no interactions should take place (donor-GFP alone controls), the fluorescent lifetime is expected to be higher than for those protein pairs which show an interaction (Fig. 5B). QSK1 and BAK1 were found to interact with significant difference to the respective control, while for low-confidence interaction of QSK1 with AHA2, the fluorescent lifetime was found in between the control and the high-confidence interactor. The interaction of AT5G49770 with AHA2, which was not proposed as direct interaction in the network reconstruction, was used as a control of a negative interaction with fluorescent lifetime values not different from the respective control (Fig. 5B). The FLIM-FRET experiments very well reflected the interactions proposed by our SEC-MS analytical workflow: QSK1, BAK1 and AT5G49770 show strong correlations in their elution profiles (Supplementary Fig. S5), while plasma membrane ATPase AHA2 showed a different elution profile which did not suggest direct interaction with QSK1 or AT5G49770. Fig. 5:

Based on their strong N-starvation induced changes betweenness centrality, we propose receptor kinase AT5G49770, and to some extent also co-receptor QSK1, and proton ATPase AHA2 to be directly or indirectly involved in rewiring of the cellular signaling network in response to nitrogen starvation. Therefore, we studied the knock-out mutants of these genes for nitrogen related root growth phenotypes. Knock-out mutants of high-affinity nitrate transporter *NRT2.1* and dual-affinity transporter *NRT1.1* were used as controls. Primary root growth and lateral root density was observed on growth assay under nitrogen starvation (0 mM KNO₃), low nitrate supply (0.3 mM KNO₃) or high nitrate supply (5 mM KNO₃).

As expected, primary root growth in wild type was reduced by nitrate starvation and low nitrate supply (Fig. 6A). The nitrate transporter mutants *nrt1.1* and *nrt2.1* showed higher than wild type root growth under nitrate starvation and significantly lower than wild type primary root growth at high nitrate supply. The general pattern of higher than wild type primary root growth at nitrate starvation and lower than wild type primary root growth at high nitrate supply was observed also in knock-out mutants of receptor kinase *at5g40770* and co-receptor *qsk1*. By contrast, *aha2* knock-out mutants had lower than wild type primary root growth under all conditions (Fig. 6B). Fig. 6:

For lateral roots, in wild type an increased lateral root density was found at low nitrate supply and lower lateral root densities were observed under nitrate starvation and high nitrate (Fig. 6C). The *nrt2.1* mutant displayed very high lateral root density at low nitrate supply and low lateral root density at high nitrate supply. This is in agreement with literature, at least when sucrose is supplied to the agar plates [39]. Mutants of receptor kinase *at5g40770* and co-receptor *qsk1* also regarding lateral root density resembled the phenotype of *nrt2.1* suggesting involvement in regulation of nitrate uptake through NRT2.1 (Fig. 6D). Mutants of *aha2* had lower than wild type lateral root density especially at high nitrate supply. These root growth experiments suggest, that although AT5G49770 can interact with QSK1 and BAK1, and co-receptor QSK1 can interact with AHA2 (Fig. 5A), the specific roles of the three proteins in nitrate responses may be quite different. We propose AT5G49770 to act together with QSK1 and BAK1, possibly as receptor-co-receptor pairs to regulate nitrate uptake through NRT2.1 and NRT1.1, and to regulate proton export necessary for nitrate uptake

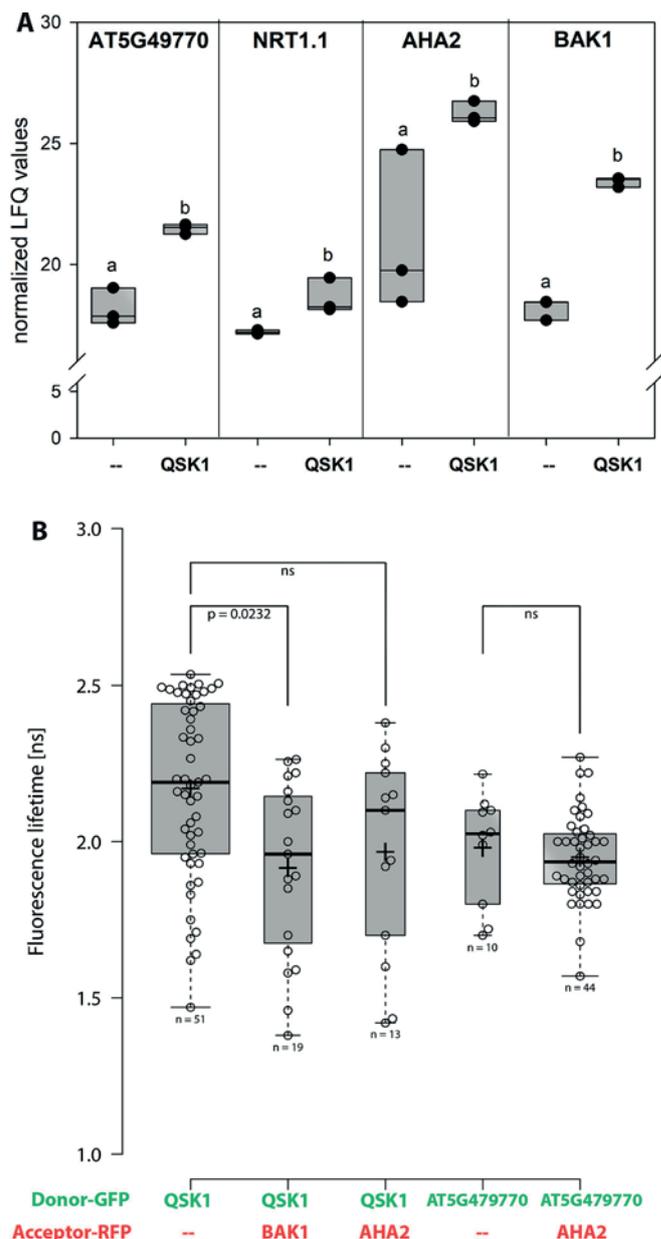


Fig. 5. Confirmation of proposed interaction pairs. (A) Pull-downs of bait QSK1 (QSK1-GFP) identified AT5G49770, NRT1.1, AHA2 and BAK1 as interactors to QSK1 compared to pull-downs with GFP alone (--). Small letters indicate significant differences ($p < 0.05$) of protein LFQ values in pull-downs with QSK1-GFP compared to GFP-only control. Boxplots of three independent experiments are shown, individual LFQ-values of three replicates are shown as black dots. (B) FRET-FLIM experiments for selected protein pairs. Box plots of fluorescent lifetime for each donor-GFP and acceptor-RFP pair are shown. Donor-GFP constructs without acceptor-RFP (--) were used as negative controls.

through AHA2. However, root growth in general is a complex trait involving not only direct responses to nutrients and the respective signaling pathways, but also proton pump activity [40] and gating of water channels [41]. Thus, it is to no surprise that the knock-out mutant of the most abundant H⁺-ATPase in roots, *aha2*, shows a more severe root growth phenotype irrespective of the nitrate environment.

4. Discussion

We studied the response of cellular protein-protein interaction networks to changes in nitrogen availability, regarding protein abundance

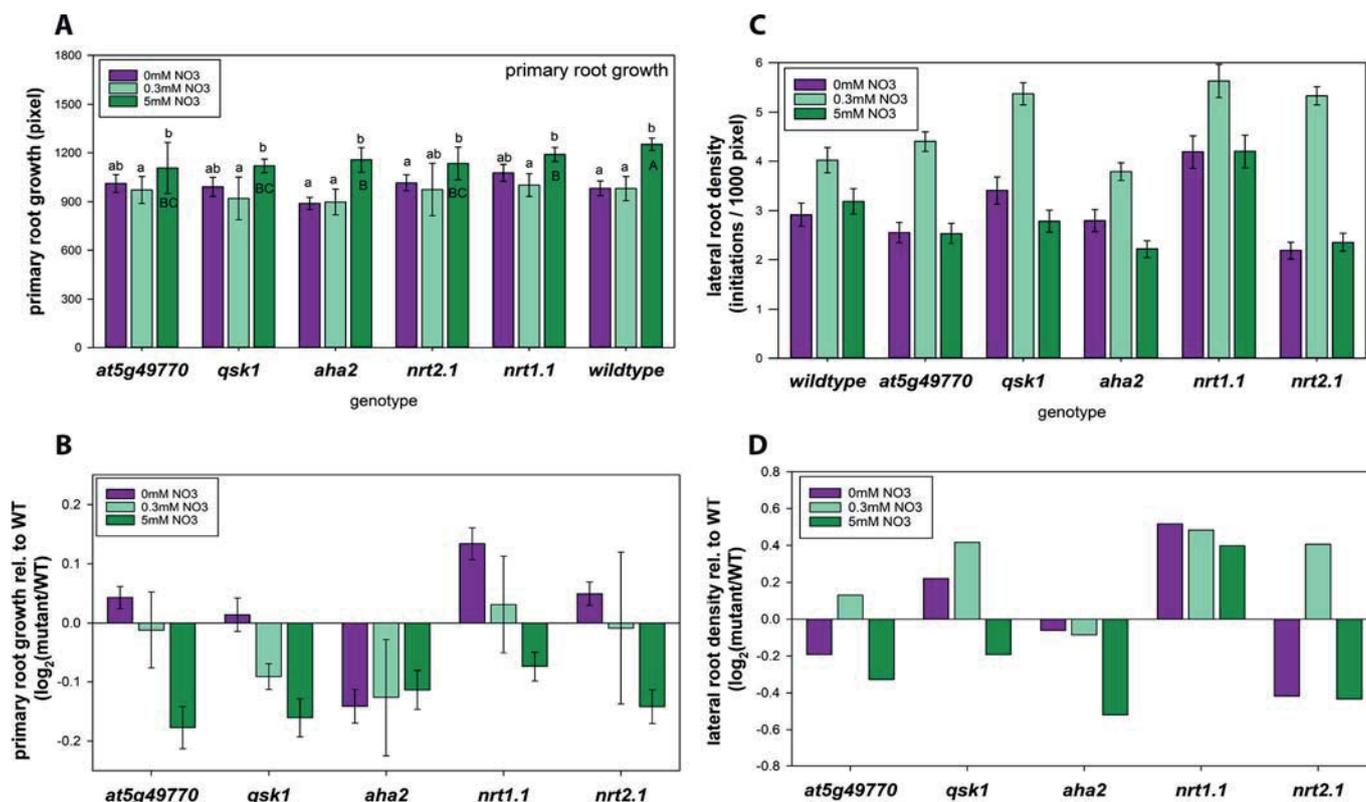


Fig. 6. Root growth phenotypes of selected mutants based on their betweenness centrality value changes upon nitrogen starvation. (A) Primary root growth, (B) Primary root growth relative to wild type, (C) Lateral root growth, (D) Lateral root growth relative to wild type. Seedlings were grown on agar plates without nitrate for one week and transferred to new plates without nitrogen (0 mM, nitrogen starvation), low nitrate (0.3 mM) or high nitrate (5 mM). All plates contained 30 mM sucrose. Averages of at least 40 seedlings are shown with standard deviations. For ratios, standard deviations were obtained by 10 bootstrapping iterations of random ratios.

as well as rewiring of protein-protein interactions. We used network topology parameters to define the most relevant proteins within the interaction network governing these changes.

4.1. Protein abundances and their variation

Abundance variations of proteins were thoroughly investigated regarding technical and biological variations, confirmed substantial quantitative variability for some proteins [42]. However, variability in protein abundance between biological replicates has been described to be slightly lower than variations introduced by the technical component, and the majority of proteins was not affected by treatments [42]. In our experiments, for most proteins nitrogen starvation induced a reduction in protein abundance category, even when considering only those proteins with low variability between replicates. This is to be expected from generally lower protein content in nitrate starved plants [43,44]. We observed that proteins with highly reproducible abundance levels between biological replicates were also those proteins which showed no strong abundance changes upon nitrogen starvation. It must be kept in mind that we compare changes in abundance categories. Thus, it can happen that a protein remains within the same abundance category, despite its actual change in abundance. In turn, proteins may change their abundance category mainly due to large abundance changes of other proteins.

Thus, the requirement for reproducibility between biological replicates filtered for those proteins, which seemed to have stable abundances under different cellular conditions. In turn, proteins with showed stronger abundance variability between replicates were tentatively filtered out as “non-reproducible”. However, these “variable” proteins were also those, which showed high response to the nitrogen starvation

treatment. This indicates that the cellular protein network indeed is in a dynamic equilibrium of protein abundances and their interactions. Environmental changes – such as nitrogen starvation – can push the system into a new equilibrium, in which protein abundances and their interactions are adjusted to the new conditions. Since these adjustments are a constant process depending on external and internal status of the plant, regulatory proteins showed a tendency to be among the more variable proteins also in comparisons between biological replicates. Protein abundances were also previously reported to respond to nitrate starvation to a similar magnitude as transcriptional changes [13].

4.2. Network topology parameters to identify key nodes in network rewiring

Following workflows established in social sciences [45,46], we used network topology parameters to identify nodes which could have a key role in regulation of information flow within the networks. In network analysis, certain parameters emerged aiming to describe overall network structure (e.g. modularity or clusters) and functional relevance of nodes and edges. These network parameters are based on intuitive conceptions of node connectivity, since the influence of each node is mainly defined by the direct or indirect neighborhood. The general assumption then is that a node, which lies within the center of a network, is also of high relevance to network structure and/or information flow.

To describe node centrality, different centrality indices have been introduced for the use in biological networks [47], which are mainly distinguished by their metric. Neighborhood based centralities such as “Degree Centrality” or “Clustering Centrality” are based on the degree of a node. Other measures of centrality like “Closeness Centrality” or “Betweenness Centrality” are path-based metrics, which rely on the

distances of nodes. Identifying vital nodes of the network largely depends on the criteria that are used to define “importance”. In biological networks, signaling pathways contain key regulator proteins which control signal transduction processes. Therefore, we decided to focus mainly on path-based centrality measures since they can be used to define the ability of each node to exert influence on the flow of information. However, the original definition of centrality measures have some shortcomings when trying to apply these principles to biological networks. Mainly, in biological networks, there might be some hubs or sub-networks that are not connected to each other. In this case, CC cannot be calculated, since $\sigma_{ij} = \infty$. These separated hubs can occur as a result of biological relevance (e.g. multiple self-sufficient protein-protein interaction complexes) or missing data. In addition, because BC is sensitive towards local topology, BC is more affected by missing data or random errors in edge-distribution than degree-based centrality measures [48]. On the other hand, because of its sensitivity, BC is most suited to observe true changes in network connectivity and node importance. Although the centrality indices like closeness centrality and betweenness centrality are prone to systematical error, the actual performance is also dependent on the underlying dataset. In direct comparison, CC and BC are quite robust in randomly generated (“sampled”) and perturbed networks, while the variability is higher in real life networks with randomly introduced errors [49]. This can be explained by the fact that perturbations of network topology have far more influences on CC and BC than sampling level alone. In consequence, performance is quite robust in real life networks which only showed errors in randomly selected sub-samples. In our experiments, the “true” networks show clear non-random changes in BC upon changes of nitrogen conditions (Supporting Fig. S3).

This indicates that network properties can still be used to describe dynamic processes in networks, even if there are inaccuracies in the data or general variability between samples exists [49,50]. We here used BC as a measure of the relevance of nodes in network rewiring induced by nitrate starvation or nitrate resupply. A member of the LRR receptor kinase family (AT5G49770) was identified as a candidate key regulator at low nitrate supply and nitrate starvation. Nitrate-dependent root growth analysis confirmed the involvement of AT5G49770 in nitrate-related processes. In that regards, it is important to note that nitrate resupply for 15 min did not fully recover the proteome status to the full nutrition condition. In general, based on previous nitrogen starvation experiments, despite a large overlap of responsive (phospho)proteins, the response to nitrate starvation was not found fully opposite to the response to nitrate resupply after starvation [13].

Furthermore, our data suggest a close connection between protein function and protein interaction stability. Stable and reproducible protein-interactions found in several experimental conditions seemed biased towards proteins that were not associated to the specific stress response but had basic homeostatic functions in processes such as protein synthesis, photosynthesis and respiration. Proteins associated to signaling and cellular organization showed higher variability in their interactions. One possible explanation could be that proteins associated with basic cellular function are often stable also regarding long term gene expression homeostasis which is why they are often used as reference genes/proteins (“house-keeping genes”). Signaling processes on protein level are more short lived and dynamic protein interactions, and their abundance is often also subject to control by gene expression.

In conclusion, we here present a strategy involving path-based centrality measures to describe dynamic rewiring of cellular protein-protein interaction networks in response to external stimuli. This approach identified AT5G49770 as a novel player in nitrate-induced responses, which was corroborated by phenotypic analysis. We expect the use of path-based centrality measures to be a powerful strategy also in the analysis of dynamic responses of cellular networks to other stresses and stimuli.

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jprot.2021.104114>.

Data availability

Data has been made available in public repositories mentioned in the methods section

Acknowledgements

Research in our laboratories was supported by the German Research Foundation (DFG) with grants to MG, WS and KH (HA 2146/23-1 and SCHU1533/11-1).

Author contributions

MG performed the proteomics experiments, data analysis, and network analysis, and wrote the manuscript. ZL contributed unpublished materials (constructs). XNW contributed pull-down data sets. LR performed FRET/FLIM experiments. SG performed root growth assays. KH interpreted the FRET/FLIM experiments and participated in writing of the manuscript. WXS conceived the study, was involved in data interpretation and wrote the manuscript.

References

- [1] F. Demirci, C. Hontrich, J.O. Blachutzik, S. Scherzer, Y. Reinders, S. Kierszniowska, et al., Arabidopsis nanodomain-delimited ABA signaling pathway regulates the anion channel SLAH3, *Proc. Natl. Acad. Sci. U. S. A.* 110 (2013) 8296–8301.
- [2] M. Haruta, W.M. Gray, M.R. Sussman, Regulation of the plasma membrane proton pump (H-ATPase) by phosphorylation, *Curr. Opin. Plant Biol.* 28 (2015) 68–75.
- [3] A.T. Fuglsang, A. Kristensen, T. Cuin, W.X. Schulze, J. Persson, K.H. Thuesen, et al., Receptor kinase mediated control of primary active proton pumping at the plasma membrane, *Plant J.* 80 (2014) 951–964.
- [4] M. Haruta, G. Sabat, K. Stecker, B.B. Minkoff, M.R. Sussman, A peptide hormone and its receptor protein kinase regulates plant cell expansion, *Science*. 343 (2014) 408–411.
- [5] S. Fields, O. Song, A novel genetic system to detect protein-protein interactions, *Nature*. 340 (1989) 245–246.
- [6] M.O. Collins, J.S. Choudhary, Mapping multiprotein complexes by affinity purification and mass spectrometry, *Curr Opin Biotech.* 19 (2008) 324–330.
- [7] X.N. Wu, L. Chu, L. Xi, H. Pertl-Obermeyer, Z. Li, K. Sklodowski, et al., Sucrose-induced receptor kinase 1 is modulated by an interacting kinase with short extracellular domain, *Molecular & cellular proteomics : MCP.* 18 (2019) 1556–1571.
- [8] J. Van Leene, H. Stals, D. Eeckhout, G. Persiau, E. Van Slijke, G. Van Isterdael, et al., A tandem affinity purification-based technology platform to study the cell cycle interactome in Arabidopsis thaliana, *Mol. Cell. Proteomics* 6 (2007) 1226–1238.
- [9] J. Bellati, C. Champeyroux, S. Hem, V. Rofidal, G. Krouk, C. Maurel, et al., Novel aquaporin regulatory mechanisms revealed by Interactomics, *Molecular & cellular proteomics : MCP.* 15 (2016) 3473–3487.
- [10] W.R. Scheible, R. Morcuende, T. Czechowski, C. Fritz, D. Osuna, N. Palacios-Rojas, et al., Genome-wide reprogramming of primary and secondary metabolism, protein synthesis, cellular growth processes, and the regulatory infrastructure of Arabidopsis in response to nitrogen, *Plant Physiol.* 136 (2004) 2483–2499.
- [11] G. Krouk, P. Mirowski, Y. LeCun, D.E. Shasha, G.M. Coruzzi, Predictive network modeling of the high-resolution dynamic plant transcriptome in response to nitrate, *Genome Biol.* 11 (2010) R123.
- [12] W.R. Engelsberger, W.X. Schulze, Nitrate and ammonium lead to distinct global dynamic phosphorylation patterns when resupplied to nitrogen starved Arabidopsis seedlings, *Plant J.* 69 (2012) 978–995.
- [13] J. Menz, Z. Li, W. Schulze, U. Ludewig, Early nitrogen-deprivation responses in Arabidopsis roots reveal distinct differences on transcriptome and (phospho-) proteome levels between nitrate and ammonium nutrition, *Plant J.* 88 (2016) 717–734.
- [14] M. Gilbert, W.X. Schulze, Global identification of protein complexes within the membrane proteome of Arabidopsis roots using a SEC-MS approach, *J. Proteome Res.* 18 (2019) 107–119.
- [15] B. Schlesier, A. Berna, F. Bernier, H.P. Mock, Proteome analysis differentiates between two highly homologous germin-like proteins in Arabidopsis thaliana ecotype Col-0 and Ws-2, *Phytochem.* 65 (2004) 1565–1574.
- [16] H. Pertl, M. Himly, R. Gehwolf, R. Kriechbaumer, D. Strasser, W. Michalke, et al., Molecular and physiological characterisation of a 14-3-3 protein from lily pollen grains regulating the activity of the plasma membrane H⁺ ATPase during pollen grain germination and tube growth, *Planta.* 213 (2001) 132–141.
- [17] H. Pertl-Obermeyer, O. Trentmann, K. Duscha, H.E. Neuhaus, W.X. Schulze, Quantitation of vacuolar sugar transporter abundance changes using QconCAT Synthetic peptides, *Front. Plant Sci.* 7 (2016) 411.
- [18] X.N. Wu, W.X. Schulze, Phosphopeptide profiling of receptor kinase mutants, *Methods Mol. Biol.* 1306 (2015) 71–79.
- [19] T. Arnold, D. Linke, The use of detergents to purify membrane proteins, *Current protocols in protein science* 53 (2008) 4.8.1–4.8.30.

- [20] A.M. Seddon, P. Curnow, P.J. Booth, Membrane proteins, lipids and detergents: not just a soap opera, *Biochim. Biophys. Acta* 2004 (1666) 105–117.
- [21] J. Cox, M. Mann, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification, *Nature Biotechnol.* 26 (2008) 1367–1372.
- [22] J. Cox, M.Y. Hein, C.A. Lubner, I. Paron, N. Nagaraj, M. Mann, Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ, *Mol. Cell. Proteomics* 13 (2014) 2513–2526.
- [23] J.A. Vizcaíno, R.G. Côté, A. Csordas, J.A. Dianes, A. Fabregat, J.M. Foster, et al., The PRoteomics IDentifications (PRIDE) database and associated tools: status in 2013, *Nucleic Acids Res.* 41 (2013). D1063-D9.
- [24] D. Szklarczyk, J.H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, et al., The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible, *Nucleic Acids Res.* 45 (2017). D362-D8.
- [25] A.M. Jones, Y. Xuan, M. Xu, R.-S. Wang, C.H. Ho, S. Lalonde, et al., Border control - a membrane-linked interactome of Arabidopsis, *Science*. 344 (2014) 711–716.
- [26] A. Bavelas, Communication patterns in task-oriented groups, *The Journal of the Acoustical Society of America*. 22 (1950) 725–730.
- [27] L.C. Freeman, Centrality in social networks conceptual clarification, *Soc. Networks* 1 (1978) 215–239.
- [28] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, A.L. Barabasi, Hierarchical organization of modularity in metabolic networks, *Science*. 297 (2002) 1551–1555.
- [29] X. Wu, C. Sanchez-Rodriguez, H. Pertl-Obermeyer, G. Obermeyer, W.X. Schulze, Sucrose-induced receptor kinase S1RK1 regulates a plasma membrane aquaporin in Arabidopsis, *Mol. Cell. Proteomics* 12 (2013) 2856–2873.
- [30] A. Hecker, N. Wallmeroth, S. Peter, M.R. Blatt, K. Harter, C. Grefen, Binary 2in1 vectors improve in planta (co)localization and dynamic protein interaction studies, *Plant Physiol.* 168 (2015) 776–787.
- [31] M. Veerabagu, K. Elgass, T. Kirchler, P. Huppenberger, K. Harter, C. Chaban, et al., The Arabidopsis B-type response regulator 18 homomerizes and positively regulates cytokinin responses, *The Plant journal : for cell and molecular biology*. 72 (2012) 721–731.
- [32] O. Thimm, O. Bläsing, Y. Gibon, A. Nagel, S. Meyer, P. Kruger, et al., MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes, *Plant J.* 37 (2004) 914–939.
- [33] S.K. Tanz, I. Castleden, C.M. Hooper, M. Vacher, I. Small, H.A. Millar, SUBA3: a database for integrating experimentation and prediction to define the SUBcellular location of proteins in Arabidopsis, *Nucleic Acids Res.* 41 (2013) D1185–D1191.
- [34] R.L. Poole, The TAIR database, *Methods Mol. Biol.* 406 (2007) 179–212.
- [35] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, et al., Cytoscape: a software environment for integrated models of biomolecular interaction networks, *Genome Res.* 13 (2003) 2498–2504.
- [36] J. Alstott, E. Bullmore, D. Plenz, Powerlaw: a Python package for analysis of heavy-tailed distributions, *PLoS One* 9 (2014), e85777.
- [37] G.A. Gilbert, M.V. Gadush, C. Wilson, M.A. Madore, Amino acid accumulation in sink and source tissues of *Coleus blumei* Benth. During salinity stress, *J. Exp. Bot.* 49 (1998) 107–114.
- [38] G.D. Bader, C.W. Hogue, An automated method for finding molecular complexes in large protein interaction networks, *BMC Bioinformatics*. 4 (2003) 2.
- [39] D.Y. Little, H. Rao, S. Oliva, F. Daniel-Vedele, A. Krapp, J.E. Malamy, The putative high-affinity nitrate transporter NRT2.1 represses lateral root initiation in response to nutritional cues, in: *Proceedings of the National Academy of Sciences of the United States of America* 102, 2005, pp. 13693–13698.
- [40] E. Młodzinska, G. Klobus, M.D. Christensen, A.T. Fuglsang, The plasma membrane H(+) -ATPase AHA2 contributes to the root architecture in response to different nitrogen supply, *Physiol. Plant.* 154 (2015) 270–282.
- [41] B. Péret, G. Li, J. Zhao, L.R. Band, U. Voß, O. Postaire, et al., Auxin regulates aquaporin function to facilitate lateral root emergence, *Nat. Cell Biol.* 14 (2012) 991–998.
- [42] M.R. Al Shweiki, S. Monchgesang, P. Majovsky, D. Thieme, D. Trutschel, W. Hoehenwarter, Assessment of label-free quantification in discovery proteomics and impact of technological factors and natural variability of protein abundance, *J. Proteome Res.* 16 (2017) 1410–1424.
- [43] G. Krouk, N.M. Crawford, G.M. Coruzzi, Y.F. Tsay, Nitrate signaling: adaptation to fluctuating environments, *Curr. Opin. Plant Biol.* 13 (2010) 266–273.
- [44] A. Melzer, G. Gebauer, H. Rehder, Nitrate content and nitrate reductase activity in *Rumex obtusifolius* L. II. Responses to nitrate starvation and nitrogen fertilization, *Oecologia*. 63 (1984) 380–385.
- [45] M.E. Sardiú, J.M. Gilmore, B. Groppe, L. Florens, M.P. Washburn, Identification of topological network modules in perturbed protein interaction networks, *Sci. Rep.* 7 (2017) 43845.
- [46] M.E. Sardiú, J.M. Gilmore, B.D. Groppe, A. Dutta, L. Florens, M.P. Washburn, Topological scoring of protein interaction networks, *Nat. Commun.* 10 (2019) 1118.
- [47] M. Jalili, A. Salehzadeh-Yazdi, S. Gupta, O. Wolkenhauer, M. Yaghmaie, O. Resendis-Antonio, et al., Evolution of centrality measurements for the detection of essential proteins in biological networks, *Front. Physiol.* 7 (2016) 375.
- [48] J.M. Bolland, Sorting out centrality: an analysis of the performance of four centrality models in real and simulated networks, *Soc. Networks* 10 (1988) 233–253.
- [49] E. Costenbader, T.W. Valente, The stability of centrality measures when networks are sampled, *Soc. Networks* 25 (2003) 283–307.
- [50] Q. Niu, A. Zeng, Y. Fan, Z. Di, Robustness of centrality measures against network manipulation, *Physica A: Statistical Mechanics and its Applications*. 438 (2015) 124–131.

General discussion

Experimental procedure

Numerous strategies for studying protein-protein interactions are established and readily available. For a meta-level analysis, traditional methods like AP-MS or Y2H severely lack throughput to reconstruct networks on a global scale. The competitive benefit of a co-elution strategy as introduced is that many hundreds of proteins can be simultaneously analyzed. Through measuring abundance across the elution fractions, a co-elution strategy scales much more readily and can be adapted to a multitude of experimental designs almost without effort. In general, the focus lies not on particular bait proteins for which preparatory steps need to be taken. This enables the identification of many interactors of many proteins regardless of scientific focus. Adding to that, proteins are investigated in a more physiological relevant environment because there is no need for extensive protein tagging, overexpression or existence of a specific antibody or heterologous expression system. Yet, in contrast to Y2H or AP-MS, co-elution strategies rely on the physical separation of proteins and their complexes.

In the scientific community, co-elution strategies were employed in different contexts with different methodology and variations of experimental procedure are used. On the most fundamental level, proteins and their complexes can be separated utilizing their chemical or physical properties. Based on their polarity (Ion-exchange chromatography, IEX) or hydrophobicity (reversed phase or hydrophobic interaction chromatography, HIC), proteins can separate through the interaction of a polar or unpolar solid phase fixed in a chromatographic column. Elution with increasing buffer strength will gradually elute proteins of similar properties which greatly reduces overall sample complexity. Though separation efficiency is high, IEX and HIC require the use of strong organic or inorganic buffers to solubilize proteins from the column. This can interfere with protein interaction or complex formation if the interaction between the proteins is weaker than the interaction of individual proteins with the solid or liquid phase. The required buffering conditions may even have adverse effects on protein structure which would greatly diminish the recovery of intact complexes from the chromatography. Another method of complex separation is to use the physical properties rather than the chemical properties of proteins. Proteins can be separated based on their size through centrifugation sedimentation, native 2D polyacrylamide gel electrophoresis (PAGE) or size exclusion chromatography. Centrifugation sedimentation utilizes the differences in protein density to separate whole protein complexes in, for example, a cellulose or sucrose density gradient. There are no markers required but resolution is rather low and it is more commonly used for whole organelle separation. PAGE-methods are more suitable since they offer a higher resolution but are limited when investigating native complexes. During PAGE, the proteins' movement through the acrylamide molecular sieve is dependent on physical size and intrinsic charge. To drive protein motion, proteins are usually denatured using a strong detergent like sodium dodecyl sulfate (SDS) that anneals to the peptide backbone and masks intrinsic charges so to negate the effects of protein intrinsic charge on electrophoretic mobility and to provide a baseline motive force. Yet, denaturing conditions severely interrupt protein complexes. To preserve protein complexes during native PAGE, none or only mild detergents are utilized. In a first step, native protein complexes are separated utilizing their intrinsic charges and protein complex composition is resolved in a second, denaturing PAGE at a 2D level. The downside is a rather high experimental complexity and sub-optimal performance for proteins with no or low intrinsic charge. During size-exclusion-chromatography (SEC), protein movement is entirely independent of protein charge or ability to interact with the column material. The porous surface structure of the column bed material presents a greater surface area to molecules with a small enough size to enter the pores, which obstructs their movement and delays their

elution time. This is especially advantageous as it allows for native conditions without compromising resolution or increasing analytical complexity and is prime reason why SEC was the method of choice for conducting the research presented in this work.

Yet, several separation methods may be combined to offset each methods drawbacks and paint a more sophisticated picture of protein complex interaction. In *Escherichia coli*, stable protein complexes were characterized using a combination of IEX, HIC and gel filtration as demonstrated by Dong and colleagues in 2008 [112]. Human soluble protein complexes were described by Havugimana and colleagues [113] on a proteomic scale using a combination of IEX and gradient centrifugation in 2012. In *Arabidopsis*, several efforts have been made to characterize protein complexes using SEC as for example for chloroplast stroma complex proteins [114] or whole leaf soluble proteins [115]. The research group around Nils Ruge and Holger Eubel even compared the performance of blue native PAGE and density gradient centrifugation to find the composition of ribosomes in the mitochondria in plants [116]. In their bid to characterize low abundant mitoribosomal complexes, they found that some loosely attached transient interactions were lost during PAGE and that the mitoribosomal subunits showed abnormal migration patterns in the velocity gradient indicating a far higher molecular mass than suspected. The authors attributed this cluster broadening to some variations in molecular mass following a variable subunit composition. Their findings indicate the validity of this research's premise which was to find interacting proteins based on their aberrant chromatographic behaviors.

However, most of available research unsurprisingly focuses on soluble protein complexes. This is also true for SEC since it provides the best conditions for soluble complexes such as aqueous isocratic buffering conditions, no hydrophilic or hydrophobic column material and a low pressure environment. This leaves a scientific gap of knowledge this works strived to fill. Working with membrane protein complexes requires at mild solubilization with adequate detergents to prevent protein aggregates during SEC. In this study, DDM was employed as mild detergent which is known to aid preserve membrane protein structure [108]. Yet, finding optimum solubilization conditions is difficult because of the multitude of complexes in a proteomic context. Different complexes favor different conditions. Mild conditions were chosen to provide the proteins a mostly undisturbed environment to facilitate and preserve interaction. To compensate false positive signals coming from supposedly increased micelle sizes or unspecific protein aggregates, a stringent downstream data processing pipeline was developed with regards to filtering proteins accepted as complex candidates and with regards to assigning relevance through network analysis. Detergent-free techniques to circumvent solubilization related issues exist as for example the Nanodisk platform developed by Denisov in 2016 [117], and improved into the Peptidisc platform by Carlson in 2018 [118] and may be used to further improve co-elution methods.

For the quantification of proteins in mass spectrometry, different options are available. For mass spectrometry analysis, labeled strategies like iTRAQ reagent tracking or SILAC [112, 113] were continuously replaced by the use of label free approaches e.g. [115, 119, 120]. While the ability to pool samples that comes with isobaric labeling techniques is certainly beneficial for comparative studies, they do not scale for more than two conditions. Label-free methods are to be preferred when the datasets are expected to be used in more than one context. In regard to scalability, data-independent acquisition provides the best benefit but comes at increased analysis time and significant computational challenges. A compromise between scalability and accessibility is the use of label-free methods that discern abundance based on the extracted ion-current (XIC) as employed in this work using the MaxQuant LFQ-algorithm [53].

Data processing strategy and data dynamics

For the analysis of co-elution profiles, there are several possible strategies. While the detection or prediction of PPI is not limited to a small set of tagged or expressed proteins, finding pairwise interaction pairs can be challenging and the actual value of information obtained is dependent on the underlying method of data analysis. Published studies use at least three distinct methods of classifying proteins and their respective interaction partners: (i) Correlation strategies such as Pearson correlation strength (e.g. [115]), (ii) clustering metrics such as Euclidean distance (e.g. [121]) or hierarchical clustering (e.g. [114]) and (iii) peak integration to quantify the proteins eluting in the same fraction (e.g. [122]). Recently, machine-learning approaches have been introduced (e.g. [123]). The challenge with a large scale dataset is to separate real data from background noise and each method comes with its own set of advantages and limits. While correlation based methods use the full range of data instead of just detecting if proteins are eluting in the same fraction, an effort to prospect meaningful criteria for accepting correlations has to be conducted as for example by comparison with known interactions. In contrast to clustering, correlation scoring needs context in regard to what score is considered favorable or expected and focusing on co-elution (e.g. peak integration) alone may result in a large number of false-positive results based on random co-elution. This is the reason for combining a Pearson-correlation based metric with a co-elution analysis into a hybrid analysis approach in order to overcome the drawbacks of focusing on co-elution alone and to integrate protein retention behavior for assignment criteria for complex candidates.

The design philosophy for the construction of the data processing pipeline followed the dedicated goal of compensating for potential experimental errors and reducing false positive detection rates for which co-elution approaches are susceptible. For that purpose, the use of a reference of known protein complexes proved to be most beneficial. Data analysis was split into three distinct parts to achieve this goal as detailed in the first attached publication: Finding candidate proteins, correlation scoring and confidence scoring and network analysis for assignment of relevance.

A common source of error is the explosion of combinatorics when working in a proteomic context. By reducing the global scale of identified proteins to that subset which shows the supposed behavior, combinatorics are reduced even more and the focus is sharpened. In the first step of data analysis, a criteria for complex candidate proteins was formulated based on the retention shift. Combined with stringent filtering criteria that aim to filter out proteins with bad sample coverage, this goal was achieved. Compared to other studies, this work does not use the information about co-elution to assign complexes immediately but rather uses co-elution as a categorical property to filter for those proteins that are likely to be in an interaction state. A second benefit of this approach is the reduction of replication errors founded in differing chromatographic performances. As shown in the first included publication, retention shift is a more robust metric than absolute retention time alone and therefore a better metric to use for complex candidate prediction across different replications.

Following the reduction of the dataset to the fraction of complex candidate proteins, correlation scoring was used to assign pairwise interaction. A common pitfall with correlation scoring is random co-elution which results in a high score. To compensate for chance related errors, the data range of every individual pairing was limited to exclude the area where chance-encounters are more likely: that is the area of expected retention time based on the molecular weight of that proteins. While truncating the area for data analysis diminishes a general advantage of correlation profiling – which is that a complete range of data is considered – the elimination of the most pronounced source of error for chance encounters far outweighs the potential drawbacks and gives more relevant meaning to the resulting correlation scores. To further improve scoring relevance, a certain number of data points were required for each pair. This

reduces the dataset by a significant amount of nonsense correlations e.g. correlation of only one or two data points. While in theory, correlation can be done with one or two data points, the information value derived from such measurements is non-informative since such correlations always result in a perfect or inverse perfect score which would inflate the occurrence of well scoring pairings and therefore contradict the use of correlation scoring. This approach of data space truncation and sample size restriction is unique to this study but introduces a filtering requirement that excludes many potential pairings and extends computation time since it is dependent on the pairwise combination. Other studies focus on peak integration to smoothen chromatogram complexity which produces a data space that is arguably better suited for clustering methods, but is more susceptible to outliers, if the dataspace is not checked for data integrity.

Protein dynamics and database considerations

Optimization of the data processing workflow was tightly controlled through comparison to literature confirmed protein-protein interactions. In contrast to other organisms (e.g. CORUM database for mammalia;[124]), at the date this study was performed, there was no unified data source available that collects and curates PPI in Arabidopsis. Therefore, a list of known interactions was compiled using a multitude of available resources as described in the first included publication. Another challenge was the scientific focus of this study which was the plant root membrane environment. Most of available online resources are biased towards interactions of soluble proteins or proteins located in specific cell organelles and limited information is available for membrane proteins in general. To improve complex prediction, genomic data e.g. mRNA co-expression or genetic likelihood. While a great expanse of genomic data is available, it is shown that genomic data introduces a systematic bias and decreases the chance of finding novel interactions. In their effort to map the tissue specific atlas of interactions in mammalia, Skinnider and Foster [125] found that interactome predictions based on gene co-expression underestimate the degree of protein interactions in comparison to their SEC-based approach when compared to experimentally detected interactions. In their findings, interactome predictions based on genetic data are less likely to cover database confirmed interactions and genetic data is less likely to give hints for dynamic protein complex formation. The authors concluded that that dynamic rewiring of interactions is largely uncoupled from genetic effects like co-expression and genetic data alone is insufficient to explain dynamic rewiring at the protein level. This underscores the significance of these studies aims to study protein interactions in a proteomic approach. Using the available information certainly proved to be beneficial in finding the appropriate thresholds or filtering conditions. As a prospect for further considerations, they might be reevaluated and –adjusted as more information is made available. Similarly, information about non-interacting proteins might be incorporated to further improve the filtering conditions.

As reported in the second included publication, it can be observed that, while general variability in identifying protein complex candidates is lower than expected based on technical or biological variation [64], pairwise interactions proved to show a high variance between replications. The fraction of proteins which showed a stable abundance between replications or experiments also tended to be involved with long-term cell homeostasis or housekeeping while the fraction of confirmed interactions increased with application of the high confidence criteria and further reduced the results to reoccurring interactions. Shared interactions found between replicates were also more frequently found than shared interactions between experimental conditions as shown in the second included publication.

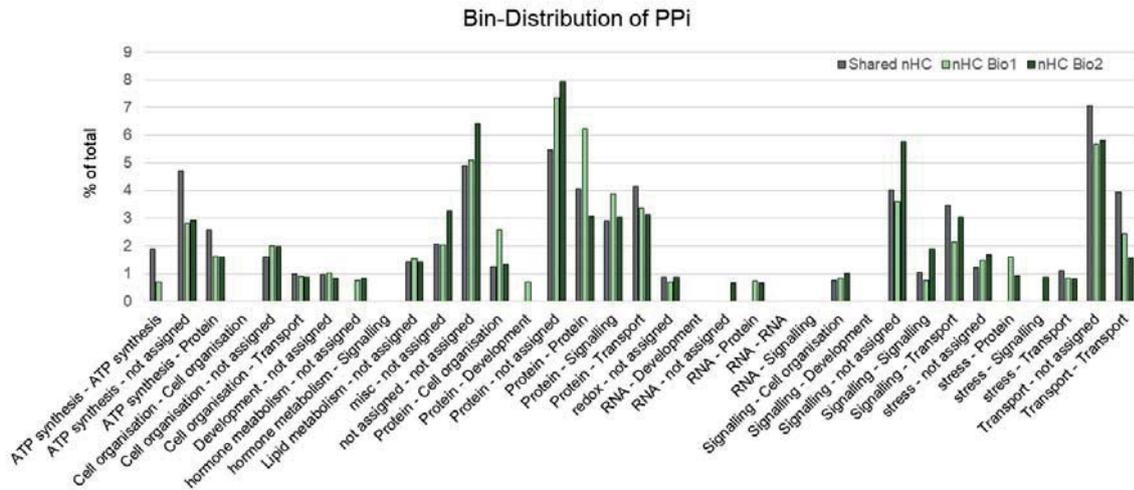


Figure 7: Distribution of protein functions for predicted Interactions. Protein function categories according to the MapMan annotation. Data shown for Full Nutrition condition with not high confidence (nHC) interactions found in replication 1 (light green), replication 2 (dark green) and shared interactions (gray).

Narrow filtering for shared interactions enriched database-confirmed interactions involved in basic cellular processes more than those related to stress or signaling as shown in Figure 7. Based on these phenomena, it can be hypothesized that experimental designs that assign relevance to results based on reliability of repetition introduce systematic biases for constitutive interactions that are conserved independent of the experimental conditions. To overly conform the data processing pipeline to rigid interactions might therefore have an adverse effect on the scope of this study which was to identify dynamic interactions which may be transient or fleeting. These results are in concert with similar studies performed in mice by Skinnider and colleagues [125] which report a very small overlap of interactions predicted by their SEC based approach (around 3%) to curated databases which is comparable but lower than the overlap presented in this study (Full nutrition shared low confidence: ~6%; Full nutrition shared high confidence: ~18% overlap to databases, see Figure 8). They attribute this effect on the fact that literature confirmed interactions are based on a small set of well characterized proteins and show that a proteomic approach for protein interaction prediction is more likely to reveal unknown protein communities and to expand the neighborhood for those curated interactions simply because an unbiased approach includes rarely known proteins or proteins without known function.

Data for Full Nutrition Shared Interactions	Curated Databases only	STRING extended
Interactome Size	23618	73929
Correlation Event Inducing Proteins (FN shared)	899	899
Protein occurrence in database	3743 (~15% of database)	55490 (~73% of database)
Predictable Interactions count	614 (~2% of occurrence)	26188 (~35% of occurrence)
Confirmed Interactions found	65 (~10% of predictable)	1696 (~6% of predictable)
Correlation Events	26491	26491
Correlation Events confirmed	65 (~0.2% of correlation events)	1696 (~6% of correlation events)
High Confidence Correlation Events	1752 (6% of correlation events)	1752 (6% of correlation events)
High Confidence Correlation Events confirmed	22 (~1.2% of High Confidence Correlation Events)	331 (~18% of High Confidence Correlation Events)

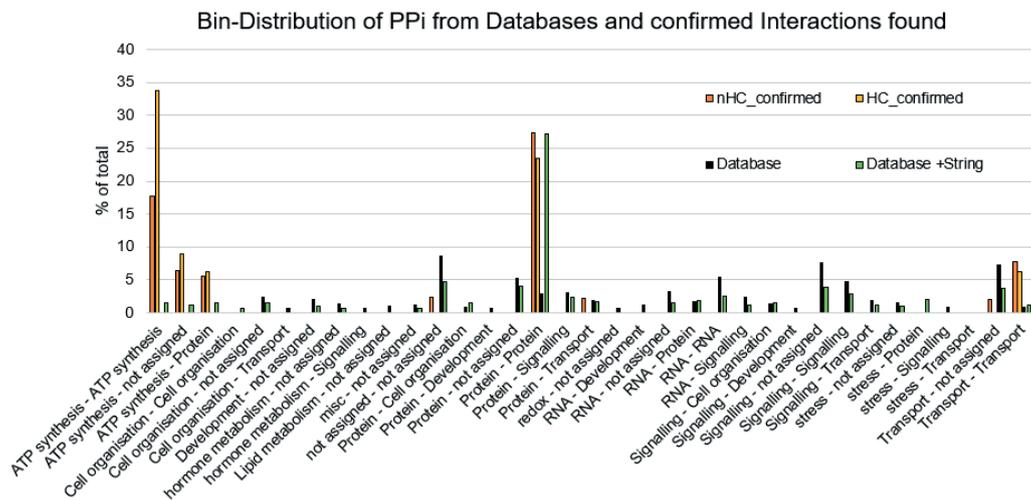


Figure 8: Database coverage and Bin-Distribution. In this study, Confirmed interactions were compiled from different available databases. In addition to curated databases (TAIR [66], MIND [126], BIOGRID [127], AI-1 [128]), PPI information for sample specific proteins were retrieved from the STRING database [129] which includes different source data as for example gene-coexpression. Functional protein annotation according to MapMan [130]. Protein interactions from databases and from the correlation analysis were sorted according to their MapMan-binning and counted. The top 75% frequent interaction types are displayed as bars. Data shown for curated databases (black), curated databases extended with sample specific STRING-database accessions (green), confirmed low confidence interactions in Full Nutrition shared interactions (orange), and confirmed high confidence interactions in Full Nutrition shared interaction (yellow).

As shown in Figure 8, database coverage for proteins found was rather small yet comparable to other studies. The table in Figure 8 gives detailed information about different metrics concerning the databases and correlation-data used in this work with specific regard to database coverage of proteins that are inducing a correlation event in this study. As an example, data for interaction found in both replications in full nutrition condition were shown. In order to increase database coverage, the interactome used in this study was supplemented with sample specific interactions from the STRING-database which encompasses not only strictly curated proteome data but also information about predicted interactions from other contexts such as genetic co-expression or likelihood. However, the integrative approach was additive and duplications were removed to preserve those strictly curated interactions from the more specific databases. This results in a dynamic, sample specific database for each experiment. With this approach, sample-protein coverage among the composite database was improved (Figure 8, table; from 15% to 73%), just as the number of presumably confirmed interactions was (Figure 8, table; from 65 to 1696). Interestingly, the fraction of confirmed high confidence interactions was disproportionately enriched when compared to the approach using curated databases only, while the overall rate of predictable interactions found lowered from 10% to just 6%. Using conventional references, only 1.2% of high confidence interactions could be confirmed, while using the dynamic database, 18% of high confidence interactions were found. In comparison to the conventional database, the dynamic database size tripled, while the fraction of high confidence interactions increased by the factor of 26. In combination, these results suggest, that the formulated criteria for high confidence interactions are more

likely to filter for those correlation events that are also more likely to be occurring in other experimental contexts.

These findings raise the question on what is actually a good representation of database confirmed interactions in the context of complex prediction. As previously mentioned, upon comparison of the protein composition of the curated databases and the protein composition of the experiment, it becomes evident that database coverage of experimentally found proteins is rather small (15%, Figure 8, data for FN shared interactions) and only a small fraction of database confirmed interactions are actually predictable (2%, Figure 8, data for FN shared interactions) with the set of proteins found in the experiment. Still, in the context of full nutrition shared interactions, there are 614 interactions that are actually predictable and of which 65 were found by the correlation analysis (~10%). To answer the question if those findings are just chance related events or taken from a biological context and meaningful, it must be explored what the number of expected findings is if a random distribution is assumed. Assuming a pairwise correlation of 899 proteins (Figure 8, table), there are $c = \binom{n}{k} = \frac{n!}{k!(n-k)!} = \binom{899}{2} = 403\,651$ possible combination of pairings. From that we can construct the binominal distribution function B to describe the probability of finding x true interactions in the experiment:

$$B(x; n, p) = \binom{n}{x} p^x (1 - p)^{n-x},$$

where x is the number of successful predictions, n is the number of tries (number of correlations in the experiment), and p is the probability of a success. For the probability p of a success, different scenarios can be assumed:

- a) A database hit is purely random, therefore $p = \frac{1}{\binom{899}{2}} = 2.47739 \times 10^{-6}$
- b) The complete Arabidopsis interactome encompasses ~290 000 interactions as estimated [128], therefore $p = \frac{290\,000}{\binom{39\,345}{2}} = 0.0003747$, for a Arabidopsis proteome size of 39 345 proteins [68].

With the assumed probabilities of finding a true interactions from a random sampling, the Binominal Probability Distribution can be drawn with n as the number of tries as equal to the number of correlation events observed in the experiment ($n = 23\,618$, FN shared low confidence interactions) as shown in figure 9.

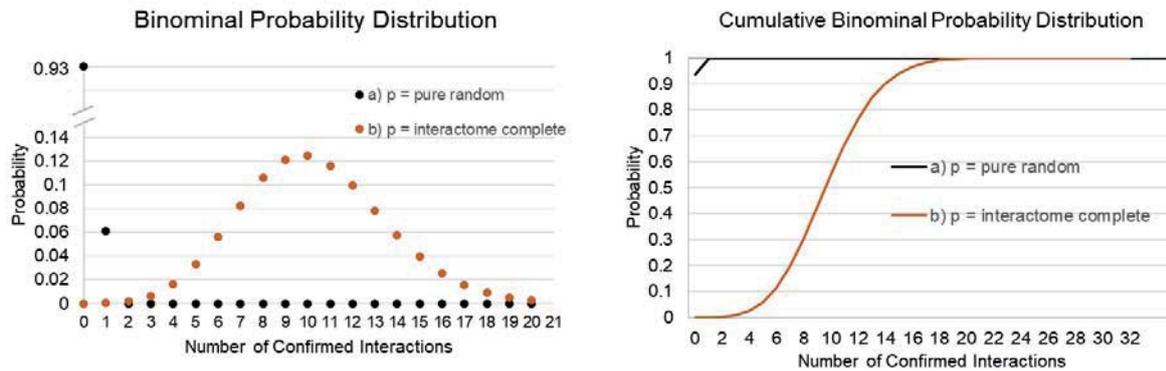


Figure 9: Binominal Probability Distribution of finding true interactions for two different scenarios. Left chart depicts the binominal probability distribution. Right chart depicts the cumulative binominal probability distribution. Scenario a) $p = \text{pure random}$ (black), scenario b) $p = \text{interactome complete}$ (orange). In scenario *a* the probability to find a true interaction is assumed to be based on pure chance. Scenario *b* assumes a p -value based on the predicted interactome size of *Arabidopsis thaliana*.

From Figure 9, the probabilities of finding true interaction partners in the experiment can be read out, if different scenarios for the probability of finding a true interaction are assumed. Assuming scenario a), where pure chance is presumed, it is the most likely outcome to find zero true interactions ($p = 0.936$) and the second most likely outcome is to find either one or zero true interactions ($p=0.997$). This means, the probability to find more than one interaction by chance is just 0.3%, assuming 23 618 correlation events. Considering the experimental results show 65 confirmed curated database hits in the Full Nutrition shared interactions dataset, it can be concluded with confidence, that the correlation results are not the result of pure chance. Assuming scenario b), when perfect knowledge of the complete interactome exists, it is most likely to find 10 confirmed interactions with $p = 0.125$, whereas the cumulative binominal probability to find less than 15 interactions is $p = 0.953$. This means the results from this study produce database-hits at least four times the rate than what would be expected ($\frac{65 \text{ interactions found}}{\text{up to 16 interactions expected}} = 4.06$). Even with perfect knowledge about the probability of finding true interactions from a given pairing – which is of course not realistic – the correlation analysis pipeline as presented in this study is clearly biased towards finding true interactions in a non-random way.

	Confirmed	Unconfirmed	Marginal Row Totals
Experiment	65 (37.53) [20.1]	23553 (23580.47) [0.03]	23618
Theoretical	614 (641.47) [1.18]	403037 (403009.53) [0]	403651
Marginal Column Totals	679	426590	427269 (Grand Total)

Figure 10: Chi-square test statistics. Contingency table for the number of confirmed interactions found in the experiment (Full nutrition, shared interactions) and the maximum number of predictable interactions found in comparison to the all possible correlation events in the same experiment. Expectation values are written in round brackets for each cell. Chi-square values are written in square brackets for each cell. Cells with a chi-square value greater than the critical value are considered to be dependent on their respective categorical variables. The sum of chi-squared values is 21.3109 and the critical value for a chi square distribution with one degree of freedom to reject the independence of variables with a confidence level of $p = 0.05$ is $k = 3.841$.

The correlation results are also enriched for confirmed curated database interactions in comparison to predictable interactions across the whole spectrum of combinations as shown in Figure 10 (Chi square test $c = 21.3109$; $p = 0.00001$, enrichment factor = 1.81).

Adding to those considerations, the confirmed correlations found in course of this study are not showing the same bias in the distribution of protein function as the underlying databases they were matched against. The utilized databases show a distinct preference of interaction typing (Figure 8, lower part) with regards to protein function. The STRING-supplemented database shows as clear bias towards interactions of the “Protein-Protein” MapMan-binning while the curated database predominantly covered “not assigned – not assigned”, RNA-associated interactions and “signaling-not assigned” or “signaling-transporter” type interactions. In contrast, the confirmed interactions found based on this study’s correlation analysis do not adhere to a similar distribution as would be expected if a random association would be assumed. Instead, confirmed interactions found were most likely to be found participating in ATP-related, Protein-related or Transporter-related processes. ATP- and Transporter-related processes were not overly represented by either curated or STRING-supplemented databases. High confidence interaction even showed a trend to be further enriched in those ATP-related processes that mainly take part in cellular housekeeping but slightly depleted for “Protein-” or “Transporter”-related processes. This can be taken as an additional indicator that high-confidence criteria are more likely to describe proteins engaged in not-dynamic processes similar to those that are found consistently between replications or between experiments. In part, this bias can be explained by the experimental conditions which aimed for the enrichment of the microsomal fraction during protein extraction. ATP or transporter-related processes are more prevalent in the microsomal fraction in general.

When investigating protein dynamics, a co-elution approach can also cover the unique aspect of protein association, especially in a plant membrane environment. It is common knowledge that proteins in a membrane environment are not acting in isolation but are surrounded by a variety of molecules that form the natural membrane milieu. The composition of this milieu restricts or enables differential protein movement that does not follow a Brownian diffusion model [131]. Moreover, by regional changing of the lipid composition, membranes are used for compartmentalization and selectively include or exclude proteins. This promotes protein interaction or protein function (Lipid raft model [132]). Thereby, cell function is not only mediated by hard protein interaction, but also through protein association. Investigating those lipid rafts is usually done by cold phase enrichment of the detergent resistant membrane fraction to determine the protein contents. To further the understanding of the significance of protein proximity, several sophisticated methods have been developed such as microscopic methods or TurboID-based proximity labeling as introduced by Mair in 2019 [133], which utilizes enzymes linked to protein of interest that label proximate proteins. Still, these approaches require preparatory

steps which limits the throughput required for a proteomic analysis while those associations will be observed naturally and coincidentally using a co-elution based approach as employed in this study.

In total, implementing the demonstrated data processing pipeline and filtering conditions resulted in a robust identification of complex candidates and their interactions, and this method is very well suited for finding transient and contextual relevant and interactions or associations.

Network analysis

Compared to other methods in PPI research, co-elution strategies are not suited for drawing hard conclusions about pairwise interactions even after pairwise correlation profiling since co-elution is also possible based on indirect interactions or loose association similarly as for example with affinity purification techniques. This may or may not be an advantage or disadvantage dependent on the scientific context. In case of this study, the scientific goal was to observe global and dynamic changes in a membrane environment under different stimuli. A particular emphasis was to find transient interactions or proteins that play a significant role during the treatment. For that purpose, a co-elution strategy as employed in this study is very well suited because proteins were investigated in an unperturbed state. Especially weak or transient interactions might not be detected in heterologous expression systems or after stringent washing steps. On the other hand, co-elution strategies result in a wealth of information and care has to be invested to filter the relevant parts.

A distinct feature of this study is the use of network analysis as tool to assign relevance to the results. While particular efforts are made to identify complexes from correlation or clustering results and specialized algorithms are available for complex identification after network reconstruction (e.g. MCODE, [134]), the information is often lacking context.

Adapted from social science, network analysis makes use of connectivity data to define a proteins relevance for the network [78]. Different centrality indices can be used as described in the second included publication. In contrast to clustering methods, the network is the prerequisite, not the result of the analysis. In this way, another layer of information beyond complex composition may be explored. As shown in this study, network topology and network parameters can even be used to compare and observe the transition of states. This is also true for networks with a general high variability or networks with missing data as commonly found in a biological context where technical limitations can lead to a loss of information. In their study published in 2003, Costenbader and Valente [135] investigated the stability of centrality measures when sampling random sub-samples of a network and thereby creating artificial networks with missing data. Their results indicate a high correlation between the network parameter values in complete networks and those of randomly selected sub-samples. Some network parameters were more robust than others at very high sampling levels. For example, Degree was more robust than betweenness centrality because rather than just being influenced by its direct neighborhood, the betweenness centrality of a node is also affected by indirect missing nodes that are necessary for path accession. This makes it robust at low sampling rates (with a mostly complete network) but inaccurate at very high sampling rates where topology is destroyed because of too many missing nodes or edges. It also demonstrates that betweenness centrality is a strong indicator for changes in network topology. Based on these assumptions, the makeup of the structural organization of a biological system may be explored with confidence even if an incomplete dataset is presumed or errors are suspected within network reconstruction.

Within this studies result, several network parameters showed a change depending on the respective treatment. Specifically, changes are observable and dependent on the nutritional status of the plants. Individual proteins showed reversible perturbations of parameters when subjected to nitrogen starvation or resupply in comparison to the full nutritional status. Thereby, several candidate proteins were identified for which a role in nitrogen-signaling was suggested. Based on these observations, selected pairs of proposed interactions for that proteins were tested and their impact on the phenotypic relevance explored as shown in the second included publication. Therefore, it is suggested, that by adding a network based evaluation approach to the analysis pipeline, valuable information about dynamic changes of networks can be obtained.

Network parameters such as Betweenness Centrality, Stress Centrality, Closeness Centrality and Degree are most suitable to quickly assess the importance of individual proteins in a network and can be used to compare their relevance in different settings. This is especially advantageous over more traditional PPI-research methods which rely on extensive genetic studies or phenotypic experiments to prospect new research perspectives. In theory, a network-based approach can also be implemented in retrospective, as long as the data obtained can be arranged as a network. Thereby, this study's results suggests a high benefit of implementing a network based approach to proteomic interaction studies in general.

However, while studies analyzing the performance of network parameters in simulated environments [136] or random subsamples of real networks [135] showed a general stability of network parameters even in case of missing data, it is up to debate how complete a network has to be to draw a meaningful conclusion. In addition, it is possible that other - non-random factors - might lead to missing values or a loss of information. For example, in gel-based LC-MS/MS proteomics, between 10 to 50% of values have been reported missing [137] and numerous strategies have been proposed for data imputation and normalization to cope with this issue as described for example by Webb-Robertson in 2015 [138] or Lazar and Burger in 2016 [139]. The reasons for missing data in label-free mass spectrometry are either rooted in the experimental, technical, analytical or biological level. Experimental reasons can be small variances in the wet-lab. Technical reasons can be poor chromatography performance or low dynamic range. Analytical reasons can be for example peptide misidentification or false matching. Biological reasons can be misscleavage, poor ionization compatibility, low peptide abundance or peptide ambiguity. While both, random and non-random reasons for missing data are plausible, it is difficult to assess the implications for biological networks because there is no reference as to what a complete or true network is. In contrast to social science (where the completeness of a network can be assessed), confirmed interactions might be biased to only reflect certain types of interactions for which the methods used to construct the database might work best. When trying to draw a conclusion, it might be more beneficial to accept an error that is truly random than to implement a correction mechanism (e.g. integration of confirmed databases or data imputation) that introduces a systematic bias with incalculable effect. A random error affects the entire data uniformly and will not interfere with the conclusion drawn in the context of network analysis. Thereby, the results of this study suggest, that implementing a network based approach to proteomic interaction studies might offer another suitable alternative to cope with general concerns of label-free quantitative proteomics which are missing values. Instead of relying on methods that are clearly introducing bias to the data such as data imputation or replications, this study suggests that researchers could accept a general data variability not only to more likely find transient effects but also to draw a conclusion based on parameters that are shown to be reliable within the scope of random errors.

Dynamic induced changes

In this study, plants were subjected to full nutrition, nitrogen starvation and nitrogen resupply conditions to identify network changes and proteins and their interaction partners that are relevant for the stress response. Several effects on the meta-level and level of individual proteins could be observed.

One advantage of a co-elution approach for PPI research is that besides interaction or association, additional information can be obtained by the interpretation of data, such as protein abundance or protein network connectivity. The more traditional methods only make use of qualitative information which is the presence or absence of an interaction partner. Yet, protein interactions are also regulated via protein abundance and a reaction to different stimuli is often mediated via changing protein abundances. With consideration to the protein abundance analysis, several algorithms or metrics were developed to quantify proteins at least in a relative fashion in mass spectrometry based proteomic experiments as discussed by Anand [140] in 2017 or Ankney in 2018 [141]. For the purpose of comparing cellular steady states, the information about relative changes of abundance is necessary to draw a conclusion. While methods for in- or cross-sample relative quantification improved significantly and were also employed in this study [53], relative protein quantification between different experimental datasets requires additional consideration with regards to protein normalization. In this study, proteins were pooled in different abundance categories dependent on their relative abundance in their respective experimental dataset to make the individual datasets comparable.

As reported in the second included publication, several large scale changes in protein abundance can be observed during nitrogen starvation. Many players showed a perturbation of their network parameters and AT5G49770 was suggested to be involved in nitrogen starvation signaling as novel actor and interactor of QSK1 and BAK1. AT5G49770 (HPCAL1) is the closest homologue of the recently discovered H₂O₂ receptor HPCA1 (AT5G49760) [142].

However, the results could also give a perspective on other general processes. Several clusterings of receptors and co-receptors were found in the network analysis with previously unsuspected relations. Most notable examples are the predicted relations of Co-Receptor BAK1 [143] with the AHA2 [144] inhibitor FERONIA [145], and AT5G49770 with a cluster of Ca²⁺ transporter of the ACA family [146] and BAK1.

BAK1 has been identified as co-receptor of BRI and FLS2 in a ligand dependent manner. Brassinosteroid (BR) induces BRI1 binding to BAK1 and negatively regulates FLS2 signaling. The interaction between FLS2 and BAK1 is induced by the Flagellin peptide flg22, which induces a plant defense reaction [143, 147, 148]. BAK1 also interacts with multiple LRR-receptors, such as PSKRs, RGFs, HAE/HSLs and EMS1/EXS and is involved in cell division, cell expansion, root meristem development, cell separation, stomata development and others, balancing plant immunity and development [149, 150]. The RALF-FERONIA pathway is also shown to be involved in regulating immune response and cell growth [151]. In their paper, Stegman and colleagues describe that RALF-FER functions as a scaffold and modulates the FLS2-BAK1 and ERF-BAK1 heterodimer complex formation [152]. Further, BAK1 is involved in RALF1-induced inhibition of root cell expansion [153]. In a recent preprint-paper, Gronnier and colleagues propose the idea, that FERONIA regulates FLS2-signaling through affecting dynamics and size of FLS2 nanodomains [154]. Based on these findings and the results from the network analysis as presented in this study, it is reasonable to assume, that BAK1 is indeed interacting with FERONIA in a condition dependent manner. Further, there is supporting evidence, that FERONIA serves another purpose as sensor for cell-wall

related damages and salinity integrating its signaling pathway via Ca^{2+} [155]. While the involvement is clear, the linkage to Ca^{2+} downstream signaling is still missing.

On the other hand, HPCA1 was found as a cell-surface sensor for H_2O_2 which induces Ca^{2+} signaling [142]. As result of the network analysis, the interaction of AT5G49770, BAK1 and various ACA Ca^{2+} transporters were suggested in full nutrition conditions which could provide an explanation of the aforementioned relations. Moreover, HPCA1 was also found to be involved as quinone receptor [156] and therefore, AT5G49770, which is the closest homologue to HPCA1, might possibly be linked to plants immune response as well, especially in conjunction with BAK1 [156]. Similarly, these interactions were lost during nitrogen starvation while other interactions e.g. the interaction between the NRT2.1 interacting kinase AT5G49770 and BAK1 occurred. Since AT5G49770 and HPCA1 are members of the same leucine-rich-repeat-receptor-like-kinases protein family, the involvement of AT5G49770 in these processes as predicted in this study can be hypothesized with confidence. AHA2 ATPase-activity is also shown to be enhanced after Ca^{2+} triggered dissociation from the SCABP3-calcium sensor [157], indicating another avenue of linkage.

H_2O_2 is known to be universal messenger molecules for signal integration and plays a role manifold processes as summarized in the recent review article by Noctor and Rechheld in 2018 [158]. Likewise, Calcium signaling regulates many physiological processes through complex regulation of calcium dependent kinases. The possibility of a connection between Calcium-, ROS- and nitrogen-signaling is suggested based on the predicted interactions in this study. In full nutrition conditions, the immediate cellular environment would be constantly monitored for cell wall integrity and environmental effects by FERONIA and AT5G49770 utilizing BAK1 as co-receptor to stop root growth in case of any interference. During nitrogen starvation, root growth is inhibited and the primary signal input comes via NRT2.1 while BAK1 coordinates a response with AT5G49770 as co-receptor. AT5G49770 was also shown to interact with receptor kinase QSK1 which is a co receptor of the Sucrose Induced Receptor Kinase 1 deeply involved in monitoring the carbon status [159]. Both, Carbon and Nitrogen supply are important checkpoints for plant growth promotion or inhibition. It can be further hypothesized that AT5G49770 is involved in Carbon and Nitrogen signaling crosstalk. This is also supported by the findings Chaput and colleagues [160]. While this study only draws a definitive conclusion for the receptor kinase AT5G49770, the produced data offers more than enough material for further considerations.

As demonstrated, many observed changes under nitrogen starvation are reverted within a short 15 minute Nitrogen resupply. This phenomenon is an indicator for network resilience and promotes the idea of a signal network that can dynamically shift and create a response without any underlying genetic changes. Protein abundance analysis showed that proteins with a higher variability between replicates also showed a greater response to and involvement in the stress condition response. As shown in the second included publication, this inter-replication variability is not the underlying reason for the changes observed in stress response since different groups of proteins showed a different type of variability. Thereby, proteins related to cell homeostasis like mitochondrial electron transport or ribosomal protein metabolism showed neither response to the stress nor abundance-variability within samples or replications. At the same time, proteins associated to signaling, transporters and stress showed a great response to a change in nutritional status but also a trend to more variability both in their abundance and correlation events. It can therefore be hypothesized that the stress response is regulated by proteins that form dynamic complexes and are variable in their protein abundance and potential partners. Thereby, this study concludes, that to exclude results because of too stringent filtering or a focus on replication stability could potentially introduce a bias for highly stable proteins and complexes unrelated to stress response.

Data processing automatization

All presented results were acquired using the MaxQuant [39] software as data processing tool to analyze mass spectrometric raw data of SEC fractions. The MaxQuant result tables were further processed using self-made Microsoft VBA [161] scripts for the Microsoft Excel spreadsheet calculation software. Microsoft Excel was chosen as a platform because its broad functionality, widespread popularity and use familiarity. Experienced Excel users can easily integrate the produced scripts and prepare worksheets for processing. The scripts were designed to be independent of the size of the underlying dataset and to produce results in a non-destructive way so that the original data is not altered or discarded. Results are transferred to a new worksheet and can be used further. The commented source code of the scripts is attached in the Addendum together with a short explanation.

The scripts were designed to be used consecutively so that after each step, the intermediate results can be used for additional independent analysis.

Following flow-chart demonstrates script usage (Figure 11).

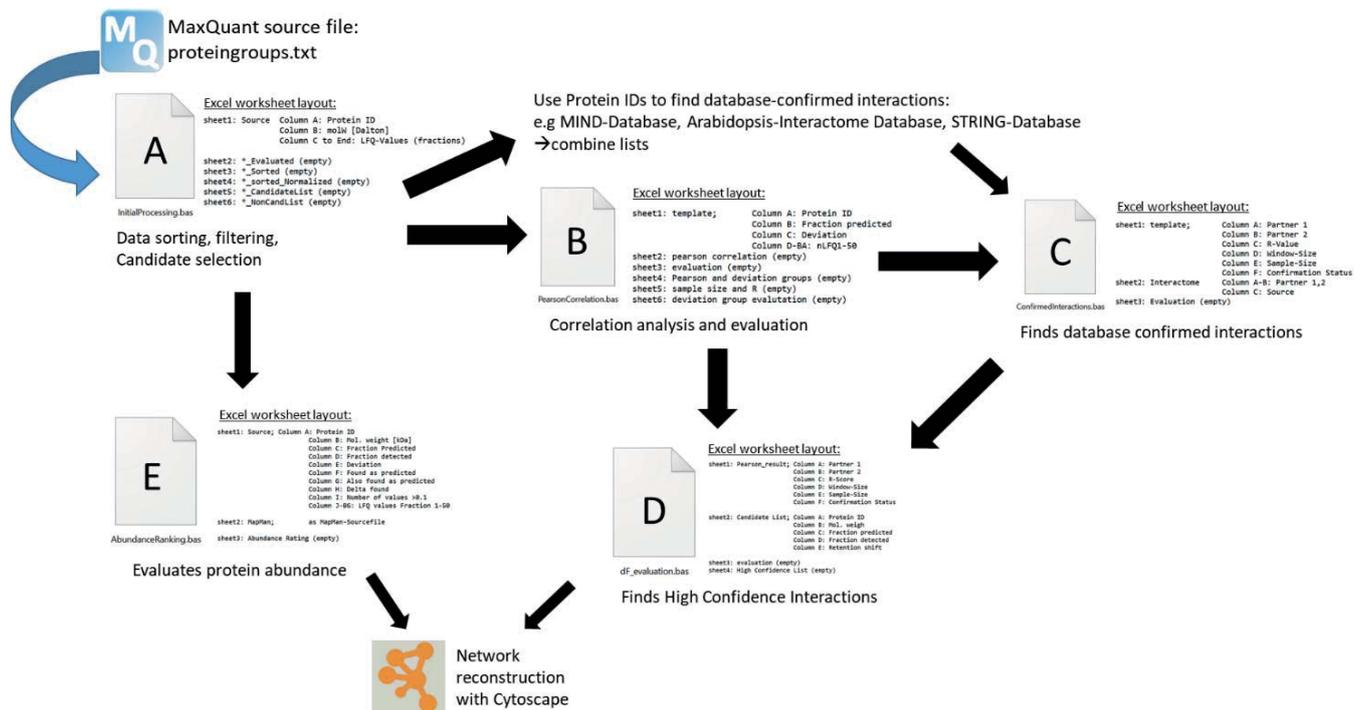


Figure 11: Flowchart of data processing using VBA scripts for Microsoft Excel. Initial data from mass spectrometry is analyzed with the MaxQuant software for protein identification and quantification. Scripts require a specific input-layout of columns as indicated.

Closing remarks

In course of this study, using a co-elution strategy, mass spectrometry and a sophisticated data processing pipeline, a new perspective on protein interaction dynamic analysis in *Arabidopsis* root was explored. It was demonstrated, that protein interactions or associations can be predicted based on this approach, even in the membrane environment. Different criteria for describing the confidence of prediction were found and it was shown that the high confidence criteria selectively favor known interactions from different published databases. Integrating a network analysis to the results uncovered several proteins of relevance to the *Arabidopsis* nitrogen response and the dynamic adjustment of protein interaction and protein abundance could be retraced not only on a large scale, but also for individual proteins. While the results of this study show that large scale protein interaction prediction is certainly possible and produces results that are relevant, on first glance, it is hard to evaluate what niche such a prediction model could occupy in the context of general research. The lack of finality and the requirement for the exact knowledge of the conceptual limitations of the results seems too prohibitive for a wider application in contrast to methods that deliver apparently more direct or clearer result as for example Yeast-Hybrid systems or FRET-methods. Yet, those methods are mostly used to confirm a hypothesis that has been formulated based on previous consideration and are suitable for that task. Although true interactions can be found, the author of the study suggests that the primary purpose of utilizing a co-elution based approach for the prediction of protein interaction is not the confirmation of suspected relations. Instead, on a conceptual basis, it is distinct because it produces results that can be used to formulate a hypothesis. Thus, it is highly complementary to other methods and offers a different approach to systems biology in general: Instead of reconstructing the big picture from small experiments, the results are presented as a big picture with the option to go into detail. This is also true, independent of the scientific context. Nitrogen availability was the primary example for the proof-of-principle, but the methodology can be used in mostly any context without the need for adaptation of the data analysis pipeline, especially in the context of plant nutritional supply using a hydroponic system and *Arabidopsis thaliana*. Although no results were presented in this work, this system was also used in the context of Flagellin-, or Brassinosteroid treatment to demonstrate, that the methodology can be seamlessly applied to different scenarios. In the long term, different complementary datasets could be produced and combined to reconstruct an atlas of the dynamic interactome in plants.

Literature

1. Tirard, S., *Origin of life and definition of life, from Buffon to Oparin*. Orig Life Evol Biosph, 2010. **40**(2): p. 215-20.
2. Napoli, C., C. Lemieux, and R. Jorgensen, *Introduction of a Chimeric Chalcone Synthase Gene into Petunia Results in Reversible Co-Suppression of Homologous Genes in trans*. Plant Cell, 1990. **2**(4): p. 279-289.
3. Agrawal, N., et al., *RNA interference: biology, mechanism, and applications*. Microbiol Mol Biol Rev, 2003. **67**(4): p. 657-85.
4. Muhammad, T., et al., *RNA Interference: A Natural Immune System of Plants to Counteract Biotic Stressors*. Cells, 2019. **8**(1).
5. Matthews, B.W., et al., *Three-dimensional Structure of Tosyl- α -chymotrypsin*. Nature, 1967. **214**(5089): p. 652-656.
6. Anfinsen, C.B., Jr., *Structural basis of ribonuclease activity*. Fed Proc, 1957. **16**(3): p. 783-91.
7. Millar, A.H., et al., *The Scope, Functions, and Dynamics of Posttranslational Protein Modifications*. Annu Rev Plant Biol, 2019. **70**: p. 119-151.
8. Caffarri, S., et al., *A comparison between plant photosystem I and photosystem II architecture and functioning*. Current protein & peptide science, 2014. **15**(4): p. 296-331.
9. Tabita, F.R., et al., *Function, structure, and evolution of the RubisCO-like proteins and their RubisCO homologs*. Microbiol Mol Biol Rev, 2007. **71**(4): p. 576-99.
10. Braun, P. and A.C. Gingras, *History of protein-protein interactions: From egg-white to complex networks*. Proteomics, 2012. **12**(10): p. 1478-1498.
11. Altmann, R., *Über Nucleinsäuren*. Arch. f. Anatomie u. Physiol, 1889: p. 524–536.
12. de Vries, H., *Das Spaltungsgesetz der Bastarde*. Berichte der Deutschen Botanischen Gessellschaft, 1900. **18**: p. 83-90.
13. Hedin, S.G., *Trypsin and Antitrypsin*. Biochem. J., 1906. **1**: p. 474-483.
14. Chick, H., *On the 'heat coagulation' of proteins*. The Journal of physiology, 1910. **40**(5): p. 404-430.
15. Morgan, T.H., *SEX LIMITED INHERITANCE IN DROSOPHILA*. Science, 1910. **32**(812): p. 120-2.
16. d'Herelle, F., *Sur une épizootie de nature bactérienne sévissant sur les sauterelles au Mexique*. CR Acad. Sci, 1911. **152**: p. 1413-1415.
17. d'Herelle, F., *Technique de la recherche du microbe filtrant bactériophage (Bacteriophagum intestinale)*. CR Soc. Biol, 1918. **81**: p. 1160-1162.
18. Sumner, J.B. and V.A. Graham, *The nature of insoluble urease*. Proceedings of the Society for Experimental Biology and Medicine, 1925. **22**(8): p. 504-506.
19. Svedberg, T. and E. Chirnoaga, *THE MOLECULAR WEIGHT OF HEMOCYANIN*. Journal of the American Chemical Society, 1928. **50**(5): p. 1399-1411.
20. Levene, P., *PROPERTIES OF THE NUCLEOTIDES OBTAINED FROM YEAST NUCLEIC ACID*. Studies from the Rockefeller Institute for Medical Research: Reprints, 1921. **36**: p. 189.
21. Levene, P., *THE STRUCTURE OF YEAST NUCLEIC ACID: V. AMMONIA HYDROLYSIS*. Journal of Biological Chemistry, 1920. **41**(1): p. 19-23.
22. Koltzoff, N.K., *Physikalisch-chemische grundlage der morphologie*. Biol. Zbl, 1928. **48**(6): p. 345.
23. Mirsky, A.E. and L. Pauling, *On the Structure of Native, Denatured, and Coagulated Proteins*. Proceedings of the National Academy of Sciences of the United States of America, 1936. **22**(7): p. 439-447.
24. Weber, H.H., *Der Feinbau und die mechanischen Eigenschaften des Myosinfadens*. Pflüger's Archiv für die gesamte Physiologie des Menschen und der Tiere, 1935. **235**(1): p. 205-233.

25. Beadle, G.W. and E.L. Tatum, *Genetic Control of Biochemical Reactions in Neurospora*. Proc Natl Acad Sci U S A, 1941. **27**(11): p. 499-506.
26. Avery, O.T., C.M. Macleod, and M. McCarty, *STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III*. J Exp Med, 1944. **79**(2): p. 137-58.
27. Sanger, F., *The free amino groups of insulin*. Biochemical Journal, 1945. **39**(5): p. 507.
28. Sanger, F. and E. Thompson, *The amino-acid sequence in the glycol chain of insulin. 1. The identification of lower peptides from partial hydrolysates*. Biochemical Journal, 1953. **53**(3): p. 353-366.
29. Fischer, E.H. and E.G. Krebs, *Conversion of phosphorylase b to phosphorylase a in muscle extracts*. J Biol Chem, 1955. **216**(1): p. 121-32.
30. Walsh, D.A., J.P. Perkins, and E.G. Krebs, *An adenosine 3',5'-monophosphate-dependant protein kinase from rabbit skeletal muscle*. J Biol Chem, 1968. **243**(13): p. 3763-5.
31. Clegg, G.A. and M. Dole, *Molecular beams of macroions. 3. Zein and polyvinylpyrrolidone*. Biopolymers, 1971. **10**(5): p. 821-6.
32. Fenn, J.B., et al., *Electrospray ionization for mass spectrometry of large biomolecules*. Science, 1989. **246**(4926): p. 64-71.
33. Fenn, J.B., *Ion formation from charged droplets: roles of geometry, energy, and time*. Journal of the American Society for Mass Spectrometry, 1993. **4**(7): p. 524-535.
34. Holley, K., Nirenberg, *The Nobel Prize in Physiology or Medicine 1968*. 1968.
35. Fields, S. and O.K. Song, *A novel genetic system to detect protein-protein interactions*. Nature, 1989. **340**(6230): p. 245-246.
36. Wolters, D.A., M.P. Washburn, and J.R. Yates, 3rd, *An automated multidimensional protein identification technology for shotgun proteomics*. Anal Chem, 2001. **73**(23): p. 5683-90.
37. Gavin, A.C., et al., *Functional organization of the yeast proteome by systematic analysis of protein complexes*. Nature, 2002. **415**(6868): p. 141-147.
38. Gavin, A.C., et al., *Proteome survey reveals modularity of the yeast cell machinery*. Nature, 2006. **440**(7084): p. 631-6.
39. Cox, J. and M. Mann, *MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification*. Nat Biotechnol, 2008. **26**(12): p. 1367-72.
40. Watson, J.D. and F.H. Crick, *Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid*. Nature, 1953. **171**(4356): p. 737-8.
41. Krogan, N.J., et al., *Global landscape of protein complexes in the yeast Saccharomyces cerevisiae*. Nature, 2006. **440**(7084): p. 637-43.
42. Wolfgang, P. and S. Helmut, *Apparatus for separating charged particles of different specific charges*. 1960, Google Patents.
43. Yost, R.A. and C.G. Enke, *Selected ion fragmentation with a tandem quadrupole mass spectrometer*. Journal of the American Chemical Society, 1978. **100**(7): p. 2274-2275.
44. Makarov, A., *Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis*. Analytical Chemistry, 2000. **72**(6): p. 1156-1162.
45. Olsen, J.V., et al., *Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap*. Mol Cell Proteomics, 2005. **4**(12): p. 2010-21.
46. Michalski, A., et al., *Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer*. Mol Cell Proteomics, 2011. **10**(9): p. M111011015.
47. Wells, J.M. and S.A. McLuckey, *Collision-induced dissociation (CID) of peptides and proteins*. Methods Enzymol, 2005. **402**: p. 148-85.

48. Olsen, J.V., et al., *Higher-energy C-trap dissociation for peptide modification analysis*. Nature Methods, 2007. **4**(9): p. 709-712.
49. Zubarev, R.A., N.L. Kelleher, and F.W. McLafferty, *Electron capture dissociation of multiply charged protein cations. A nonergodic process*. Journal of the American Chemical Society, 1998. **120**(13): p. 3265-3266.
50. Boyd, R. and A. Somogyi, *The Mobile Proton Hypothesis in Fragmentation of Protonated Peptides: A Perspective*. Journal of the American Society for Mass Spectrometry, 2010. **21**: p. 1275-8.
51. Zubarev, R.A., et al., *Towards an understanding of the mechanism of electron-capture dissociation: a historical perspective and modern ideas*. European Journal of Mass Spectrometry, 2002. **8**(5): p. 337-349.
52. Roepstorff, P. and J. Fohlman, *Proposal for a common nomenclature for sequence ions in mass spectra of peptides*. Biomed Mass Spectrom, 1984. **11**(11): p. 601.
53. Cox, J., et al., *Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ*. Molecular and Cellular Proteomics, 2014. **13**(9): p. 2513-2526.
54. Tyanova, S., T. Temu, and J. Cox, *The MaxQuant computational platform for mass spectrometry-based shotgun proteomics*. Nat Protoc, 2016. **11**(12): p. 2301-2319.
55. Baptista, A.M., et al., *The origin of trypsin: evidence for multiple gene duplications in trypsins*. J Mol Evol, 1998. **47**(3): p. 353-62.
56. Nesvizhskii, A.I., O. Vitek, and R. Aebersold, *Analysis and validation of proteomic data generated by tandem mass spectrometry*. Nat Methods, 2007. **4**(10): p. 787-97.
57. Elias, J.E. and S.P. Gygi, *Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry*. Nat Methods, 2007. **4**(3): p. 207-14.
58. Cox, J. and M. Mann, *MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification*. Nature Biotechnology, 2008. **26**(12): p. 1367-1372.
59. Cox, J., et al., *Andromeda: A peptide search engine integrated into the MaxQuant environment*. Journal of Proteome Research, 2011. **10**(4): p. 1794-1805.
60. Reiter, L., et al., *Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry*. Molecular & cellular proteomics : MCP, 2009. **8**(11): p. 2405-2417.
61. Nesvizhskii, A.I., et al., *A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry*. Analytical Chemistry, 2003. **75**(17): p. 4646-4658.
62. Schaffer, J., *What Not to Multiply Without Necessity*. Australasian Journal of Philosophy, 2015. **93**(4): p. 644-664.
63. Zybailov, B., et al., *Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling*. Anal Chem, 2005. **77**(19): p. 6218-24.
64. Al Shweiki, M.R., et al., *Assessment of Label-Free Quantification in Discovery Proteomics and Impact of Technological Factors and Natural Variability of Protein Abundance*. J Proteome Res, 2017. **16**(4): p. 1410-1424.
65. Bensimon, A., A.J. Heck, and R. Aebersold, *Mass spectrometry-based proteomics and network biology*. Annu Rev Biochem, 2012. **81**: p. 379-405.
66. Huala, E., et al., *The Arabidopsis Information Resource (TAIR): A comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant*. Nucleic Acids Research, 2001. **29**(1): p. 102-105.
67. Cheng, C.-Y., et al., *Araport11: a complete reannotation of the Arabidopsis thaliana reference genome*. The Plant Journal, 2017. **89**(4): p. 789-804.

68. Consortium, T.U., *UniProt: a worldwide hub of protein knowledge*. Nucleic Acids Research, 2018. **47**(D1): p. D506-D515.
69. Mann, M., et al., *The coming age of complete, accurate, and ubiquitous proteomes*. Mol Cell, 2013. **49**(4): p. 583-90.
70. McClung, C.R., P.A. Salomé, and T.P. Michael, *The Arabidopsis circadian system*. The arabidopsis book, 2002. **1**: p. e0044-e0044.
71. Erdős, P. and A. Rényi, *On Random Graphs I*. Publicationes Mathematicae Debrecen, 1959. **6**: p. 290--297.
72. Han, J.D., et al., *Evidence for dynamically organized modularity in the yeast protein-protein interaction network*. Nature, 2004. **430**(6995): p. 88-93.
73. Tanay, A., et al., *Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data*. Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(9): p. 2981-2986.
74. Barabasi, A.L. and R. Albert, *Emergence of scaling in random networks*. Science, 1999. **286**(5439): p. 509-12.
75. Evlampiev, K. and H. Isambert, *Conservation and topology of protein interaction networks under duplication-divergence evolution*. Proc Natl Acad Sci U S A, 2008. **105**(29): p. 9863-8.
76. Mosca, R., R.A. Pache, and P. Aloy, *The role of structural disorder in the rewiring of protein interactions through evolution*. Mol Cell Proteomics, 2012. **11**(7): p. M111 014969.
77. Artico, I., et al., *How rare are power-law networks really?* Proc Math Phys Eng Sci, 2020. **476**(2241): p. 20190742.
78. Jalili, M., et al., *Evolution of Centrality Measurements for the Detection of Essential Proteins in Biological Networks*. Frontiers in Physiology, 2016. **7**: p. 375.
79. Uetz, P., et al., *A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae*. Nature, 2000. **403**(6770): p. 623-7.
80. Thamiy, S., J. Miller, and I. Stagljar, *The split-ubiquitin membrane-based yeast two-hybrid system*. Methods Mol Biol, 2004. **261**: p. 297-312.
81. Bush, S.M., S. Folta, and D.A. Lannigan, *Use of the yeast one-hybrid system to screen for mutations in the ligand-binding domain of the estrogen receptor*. Steroids, 1996. **61**(3): p. 102-109.
82. Padlan, E.A., *Anatomy of the antibody molecule*. Molecular Immunology, 1994. **31**(3): p. 169-217.
83. von Mering, C., et al., *Comparative assessment of large-scale data sets of protein-protein interactions*. Nature, 2002. **417**(6887): p. 399-403.
84. Bork, P., et al., *Protein interaction networks from yeast to human*. Curr Opin Struct Biol, 2004. **14**(3): p. 292-9.
85. Hart, G.T., A.K. Ramani, and E.M. Marcotte, *How complete are current yeast and human protein-interaction networks?* Genome Biol, 2006. **7**(11): p. 120.
86. Forde, B.G., *Nitrogen signalling pathways shaping root system architecture: an update*. Curr Opin Plant Biol, 2014. **21**: p. 30-36.
87. Jinek, M., et al., *A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity*. Science, 2012. **337**(6096): p. 816-21.
88. Woodward, A.W. and B. Bartel, *Biology in Bloom: A Primer on the Arabidopsis thaliana Model System*. Genetics, 2018. **208**(4): p. 1337-1349.
89. Arabidopsis Genome, I., *Analysis of the genome sequence of the flowering plant Arabidopsis thaliana*. Nature, 2000. **408**(6814): p. 796-815.
90. Cheng, C.Y., et al., *Araport11: a complete reannotation of the Arabidopsis thaliana reference genome*. Plant J, 2017. **89**(4): p. 789-804.
91. Alonso, J.M., et al., *Genome-wide insertional mutagenesis of Arabidopsis thaliana*. Science, 2003. **301**(5633): p. 653-7.

92. Yamada, K., et al., *Empirical analysis of transcriptional activity in the Arabidopsis genome*. Science, 2003. **302**(5646): p. 842-6.
93. Earley, K.W., et al., *Gateway-compatible vectors for plant functional genomics and proteomics*. Plant J, 2006. **45**(4): p. 616-29.
94. Trigg, S.A., et al., *Cry2H-seq: a massively multiplexed assay for deep-coverage interactome mapping*. Nature methods, 2017. **14**(8): p. 819-825.
95. Klepikova, A.V., et al., *A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling*. Plant J, 2016. **88**(6): p. 1058-1070.
96. Hirakawa, Y. and S. Sawa, *Diverse function of plant peptide hormones in local signaling and development*. Curr Opin Plant Biol, 2019. **51**: p. 81-87.
97. Arsova, B., M. Watt, and B. Usadel, *Monitoring of Plant Protein Post-translational Modifications Using Targeted Proteomics*. Frontiers in Plant Science, 2018. **9**(1168).
98. Vidal, E.A., et al., *Nitrogen control of developmental phase transitions in Arabidopsis thaliana*. Journal of Experimental Botany, 2014. **65**(19): p. 5611-5618.
99. Lin, Y.L. and Y.F. Tsay, *Influence of differing nitrate and nitrogen availability on flowering control in Arabidopsis*. J Exp Bot, 2017. **68**(10): p. 2603-2609.
100. Weber, K. and M. Burow, *Nitrogen – essential macronutrient and signal controlling flowering time*. Physiologia Plantarum, 2018. **162**(2): p. 251-260.
101. Xu, G., X. Fan, and A.J. Miller, *Plant Nitrogen Assimilation and Use Efficiency*. Annual Review of Plant Biology, 2012. **63**(1): p. 153-182.
102. Krapp, A., *Plant nitrogen assimilation and its regulation: a complex puzzle with missing pieces*. Current Opinion in Plant Biology, 2015. **25**: p. 115-122.
103. Krapp, A., et al., *Nitrate transport and signalling in Arabidopsis*. Journal of Experimental Botany, 2014. **65**(3): p. 789-798.
104. Jacquot, A., et al., *Post-translational regulation of nitrogen transporters in plants and microorganisms*. J Exp Bot, 2017. **68**(10): p. 2567-2580.
105. Ho, C.H. and Y.F. Tsay, *Nitrate, ammonium, and potassium sensing and signaling*. Curr Opin Plant Biol, 2010. **13**(5): p. 604-10.
106. Kotur, Z., et al., *Nitrate transport capacity of the Arabidopsis thaliana NRT2 family members and their interactions with AtNAR2.1*. New Phytologist, 2012. **194**(3): p. 724-731.
107. Laurent, T.C. and J. Killander, *A theory of gel filtration and its experimental verification*. Journal of Chromatography A, 1964. **14**: p. 317-330.
108. Arnold, T. and D. Linke, *The use of detergents to purify membrane proteins*. Curr Protoc Protein Sci, 2008. **Chapter 4**: p. Unit 4 8 1-4 8 30.
109. Pearson, K. and O.M.F.E. Henrici, *VII. Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia*. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 1896. **187**: p. 253-318.
110. Gilbert, M. and W.X. Schulze, *Global Identification of Protein Complexes within the Membrane Proteome of Arabidopsis Roots Using a SEC-MS Approach*. J Proteome Res, 2019. **18**(1): p. 107-119.
111. Gilbert, M., et al., *Comparison of path-based centrality measures in protein-protein interaction networks revealed proteins with phenotypic relevance during adaptation to changing nitrogen environments*. Journal of Proteomics, 2021. **235**: p. 104114.
112. Dong, M., et al., *A "tagless" strategy for identification of stable protein complexes genome-wide by multidimensional orthogonal chromatographic separation and iTRAQ reagent tracking*. J Proteome Res, 2008. **7**(5): p. 1836-49.
113. Havugimana, P.C., et al., *A census of human soluble protein complexes*. Cell, 2012. **150**(5): p. 1068-1081.

114. Olinares, P.D., L. Ponnala, and K.J. van Wijk, *Megadalton complexes in the chloroplast stroma of Arabidopsis thaliana characterized by size exclusion chromatography, mass spectrometry, and hierarchical clustering*. Mol Cell Proteomics, 2010. **9**(7): p. 1594-615.
115. Aryal, U.K., et al., *Analysis of protein complexes in Arabidopsis leaves using size exclusion chromatography and label-free protein correlation profiling*. Journal of Proteomics, 2017. **166**: p. 8-18.
116. Rugen, N., et al., *Complexome Profiling Reveals Association of PPR Proteins with Ribosomes in the Mitochondria of Plants*. Mol Cell Proteomics, 2019. **18**(7): p. 1345-1362.
117. Denisov, I.G. and S.G. Sligar, *Nanodiscs for structural and functional studies of membrane proteins*. Nat Struct Mol Biol, 2016. **23**(6): p. 481-6.
118. Carlson, M.L., et al., *The Peptidisc, a simple method for stabilizing membrane proteins in detergent-free solution*. Elife, 2018. **7**.
119. McBride, Z., et al., *A Label-free Mass Spectrometry Method to Predict Endogenous Protein Complex Composition*. Mol Cell Proteomics, 2019. **18**(8): p. 1588-1606.
120. Heusel, M., et al., *Complex-centric proteome profiling by SEC-SWATH-MS*. Mol Syst Biol, 2019. **15**(1): p. e8438.
121. Wan, C., et al., *Panorama of ancient metazoan macromolecular complexes*. Nature, 2015. **525**(7569): p. 339-44.
122. Goebels, F., et al., *Automated Computational Inference of Multi-protein Assemblies from Biochemical Co-purification Data*. Methods Mol Biol, 2018. **1764**: p. 391-399.
123. Hawe, J.S., F.J. Theis, and M. Heinig, *Inferring Interaction Networks From Multi-Omics Data*. Front Genet, 2019. **10**: p. 535.
124. Giurgiu, M., et al., *CORUM: the comprehensive resource of mammalian protein complexes-2019*. Nucleic Acids Res, 2019. **47**(D1): p. D559-D563.
125. Skinnider, M.A., et al., *An atlas of protein-protein interactions across mammalian tissues*. bioRxiv, 2018: p. 351247.
126. Jones, A.M., et al., *Border control - A membrane-linked interactome of Arabidopsis*. Science, 2014. **344**(6185): p. 711-716.
127. Oughtred, R., et al., *The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions*. Protein Sci, 2021. **30**(1): p. 187-200.
128. Arabidopsis Interactome Mapping, C., *Evidence for network evolution in an Arabidopsis interactome map*. Science (New York, N.Y.), 2011. **333**(6042): p. 601-607.
129. Franceschini, A., et al., *STRING v9.1: Protein-protein interaction networks, with increased coverage and integration*. Nucleic Acids Research, 2013. **41**(D1): p. D808-D815.
130. Thimm, O., et al., *MAPMAN: A user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes*. Plant Journal, 2004. **37**(6): p. 914-939.
131. Metzler, R., J.H. Jeon, and A.G. Cherstvy, *Non-Brownian diffusion in lipid membranes: Experiments and simulations*. Biochimica et Biophysica Acta (BBA) - Biomembranes, 2016. **1858**(10): p. 2451-2467.
132. Simons, K. and D. Toomre, *Lipid rafts and signal transduction*. Nature Reviews Molecular Cell Biology, 2000. **1**(1): p. 31-39.
133. Mair, A., et al., *Proximity labeling of protein complexes and cell-type-specific organellar proteomes in Arabidopsis enabled by TurboID*. Elife, 2019. **8**.
134. Bader, G.D. and C.W.V. Hogue, *An automated method for finding molecular complexes in large protein interaction networks*. BMC Bioinformatics, 2003. **4**(1): p. 2.
135. Costenbader, E. and T.W. Valente, *The stability of centrality measures when networks are sampled*. Social Networks, 2003. **25**(4): p. 283-307.
136. Bolland, J.M., *Sorting out centrality: An analysis of the performance of four centrality models in real and simulated networks*. Social Networks, 1988. **10**(3): p. 233-253.

137. Albrecht, D., et al., *Missing values in gel-based proteomics*. PROTEOMICS, 2010. **10**(6): p. 1202-1211.
138. Webb-Robertson, B.J., et al., *Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics*. J Proteome Res, 2015. **14**(5): p. 1993-2001.
139. Lazar, C., et al., *Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies*. J Proteome Res, 2016. **15**(4): p. 1116-25.
140. Anand, S., et al., *Label-Based and Label-Free Strategies for Protein Quantitation*. Methods Mol Biol, 2017. **1549**: p. 31-43.
141. Ankney, J.A., A. Muneer, and X. Chen, *Relative and Absolute Quantitation in Mass Spectrometry-Based Proteomics*. Annu Rev Anal Chem (Palo Alto Calif), 2018. **11**(1): p. 49-77.
142. Wu, F., et al., *Hydrogen peroxide sensor HPCA1 is an LRR receptor kinase in Arabidopsis*. Nature, 2020. **578**(7796): p. 577-581.
143. Li, J., et al., *BAK1, an Arabidopsis LRR receptor-like protein kinase, interacts with BRI1 and modulates brassinosteroid signaling*. Cell, 2002. **110**(2): p. 213-22.
144. Palmgren, M.G., *PLANT PLASMA MEMBRANE H⁺-ATPases: Powerhouses for Nutrient Uptake*. Annu Rev Plant Physiol Plant Mol Biol, 2001. **52**: p. 817-845.
145. Haruta, M., et al., *A peptide hormone and its receptor protein kinase regulate plant cell expansion*. Science, 2014. **343**(6169): p. 408-411.
146. Yu, H., et al., *Overlapping and differential roles of plasma membrane calcium ATPases in Arabidopsis growth and environmental responses*. Journal of experimental botany, 2018. **69**(10): p. 2693-2703.
147. Chinchilla, D., et al., *A flagellin-induced complex of the receptor FLS2 and BAK1 initiates plant defence*. Nature, 2007. **448**(7152): p. 497-500.
148. Wang, Z.-Y., *Brassinosteroids modulate plant immunity at multiple levels*. Proceedings of the National Academy of Sciences, 2012. **109**(1): p. 7.
149. Gou, X. and J. Li, *Paired Receptor and Coreceptor Kinases Perceive Extracellular Signals to Control Plant Development*. Plant Physiology, 2020. **182**(4): p. 1667.
150. Xi, L., et al., *Classification and Interactions of LRR Receptors and Co-receptors Within the Arabidopsis Plasma Membrane – An Overview*. Frontiers in Plant Science, 2019. **10**(472).
151. Zhang, X., et al., *RALF–FERONIA Signaling: Linking Plant Immune Response with Cell Growth*. Plant Communications, 2020. **1**(4): p. 100084.
152. Stegmann, M., et al., *The receptor kinase FER is a RALF-regulated scaffold controlling plant immune signaling*. Science, 2017. **355**(6322): p. 287.
153. Dressano, K., et al., *BAK1 is involved in AtRALF1-induced inhibition of root cell expansion*. PLOS Genetics, 2017. **13**(10): p. e1007053.
154. Gronnier, J., et al., *FERONIA regulates FLS2 plasma membrane nanoscale dynamics to modulate plant immune signaling*. bioRxiv, 2020: p. 2020.07.20.212233.
155. Feng, W., et al., *The FERONIA Receptor Kinase Maintains Cell-Wall Integrity during Salt Stress through Ca(2+) Signaling*. Curr Biol, 2018. **28**(5): p. 666-675 e5.
156. Laohavisit, A., et al., *Quinone perception in plants via leucine-rich-repeat receptor-like kinases*. Nature, 2020. **587**(7832): p. 92-97.
157. Yang, Y., et al., *The Ca²⁺ Sensor SCaBP3/CBL7 Modulates Plasma Membrane H⁺-ATPase Activity and Promotes Alkali Tolerance in Arabidopsis*. The Plant Cell, 2019. **31**(6): p. 1367-1384.
158. Noctor, G., J.P. Reichheld, and C.H. Foyer, *ROS-related redox regulation and signaling in plants*. Semin Cell Dev Biol, 2018. **80**: p. 3-12.
159. Wu, X.N., et al., *Sucrose-induced receptor kinase SIRK1 regulates a plasma membrane aquaporin in Arabidopsis*. Molecular and Cellular Proteomics, 2013. **12**(10): p. 2856-2873.

160. Chaput, V., A. Martin, and L. Lejay, *Redox metabolism: the hidden player in carbon and nitrogen signaling?* J Exp Bot, 2020. **71**(13): p. 3816-3826.
161. Microsoft *Visual Basic for Applications*. 2013.

Addendum

Script for initial processing

Following script is used for data sorting, filtering and candidate protein selection. It requires six worksheets whereas a collection of data from the MaxQuant proteingroups.txt file is to be copied into the first worksheet in the sequential order as indicated in the flowchart step A

```
1
2 Attribute VB_Name = "modProcessing"
3 Option Explicit
4
5 Sub DataEvaluation()
6
7 Dim i As Long
8 Dim k As Long
9
10 Dim intTemp As Integer
11 Dim intFractionPredicted As Integer
12 Dim dbMAXLFQ As Double
13 Dim dbIntermediatePrediction As Double
14 Dim lgnLFQCounter As Long
15 Dim dbMolW As Double
16 Dim intFractionDetected As Integer
17
18 Dim counter1 As Long
19 Dim counter2 As Long
20 Dim counter3 As Long
21
22
23 Dim adject4 As Boolean
24 Dim nLFQ3 As Boolean
25
26
27 Dim lgLFQLBorder As Long
28 Dim lgLFQRBorder As Long
29
30 Dim lgTempLastRow As Long
31 Dim lgTempLastColumn As Long
32
33 Dim wsSource As Worksheet
34 Dim wsEvaluated As Worksheet
35 Dim wsTIRMSorted As Worksheet
36 Dim wsTIRMSortedNorm As Worksheet
37 Dim wstarget As Worksheet
38 Dim wsCandidateList As Worksheet
39 Dim wsNonCandidateList As Worksheet
40
41
42 Dim rgTemporary As Range
43
44 Dim strTempString As String
45
46 Dim dtStart As Date
47 Dim dtEnd As Date
48
49
50 Set wsSource = Sheet1
51 Set wsEvaluated = Sheet2
52 Set wsTIRMSorted = Sheet3
53 Set wsTIRMSortedNorm = Sheet4
54 Set wsCandidateList = Sheet5
55 Set wsNonCandidateList = Sheet6
56
57 dtStart = Now
58
59
```

```

60 '-----Initialisieren-----
61 With wsEvaluated
62
63 .Cells(1, 3) = "Fraction Predicted (Rounded)"
64 .Cells(1, 4) = "Fraction detected"
65 .Cells(1, 5) = "Found as predicted"
66 .Cells(1, 6) = "Also found as predicted"
67 .Cells(1, 7) = "Retention shift"
68
69 End With
70
71 Application.StatusBar = "Progress: Initializing..."
72
73 With wsSource
74 lgTempLastRow = .Cells(.Rows.Count, 1).End(xlUp).Row
75 lgTempLastColumn = .Cells(1, .Columns.Count).End(xlToLeft).Column
76
77 Set rgTemporary = .Range(.Cells(1, 1), .Cells(lgTempLastRow, 2))
78
79 rgTemporary.Copy wsEvaluated.Cells(1, 1)
80
81 Set rgTemporary = .Range(.Cells(1, 3), .Cells(lgTempLastRow, lgTempLastColumn))
82 lgTempLastColumn = wsEvaluated.Cells(1, wsEvaluated.Columns.Count).End(xlToLeft).Column
83
84 rgTemporary.Copy wsEvaluated.Cells(1, lgTempLastColumn + 1)
85 lgTempLastColumn = wsEvaluated.Cells(1, wsEvaluated.Columns.Count).End(xlToLeft).Column
86 End With
87
88 lgTempLastColumn = wsEvaluated.Cells(1, wsEvaluated.Columns.Count).End(xlToLeft).Column '58
89 Application.StatusBar = "Progress: Finding LFQ-values..."
90 For i = 2 To lgTempLastColumn 'Bereich der LFQs in der Tabelle finden
91     strTempString = CStr(wsEvaluated.Cells(1, i).Value)
92     intTemp = InStr(strTempString, "LFQ")
93     If intTemp > 0 Then
94         lgLFQLBorder = i
95         Exit For
96     Else
97         End If
98 Next i
99 intTemp = 0
100 Application.StatusBar = "Progress: Finding LFQ-values: 50%"
101 For i = lgLFQLBorder To lgTempLastColumn + 1
102     strTempString = (wsEvaluated.Cells(1, i).Value)
103     intTemp = InStr(strTempString, "LFQ")
104     If intTemp > 0 Then
105         Else
106             lgLFQRBorder = i - 1
107             Exit For
108         End If
109 Next i
110 Application.StatusBar = "Progress: Finding LFQ-values: 100%"
111
112 '-----Start auswertung-----
113 counter1 = 2 'counter für die row-nr der neuen worksheets
114 counter2 = 2
115 counter3 = 2
116
117
118 For i = 2 To lgTempLastRow 'lgTempLastRow 'fraction predicted, retention shift
119
120 Application.StatusBar = "Progress: Evaluation and Sorting: " & ((Round(i / lgTempLastRow,
0)) * 100) & " %"
121
122 lgnLFQCounter = 0
123 adjunct4 = False
124
125 With wsEvaluated
126
127 dbMolW = .Cells(i, 2).Value
128 If dbMolW = 0 Then
129 Else

```

```

130 dbIntermediatePrediction = ((-3.464) * Application.WorksheetFunction.Ln(dbMolW)) +
59.067
131 intFractionPredicted = Application.WorksheetFunction.Round(dbIntermediatePrediction, 0)
'vorhersage der fraktion der elution
132 .Cells(i, 3) = intFractionPredicted
133 Set rgTemporary = .Range(.Cells(i, lgLFQLBorder), .Cells(i, lgLFQRBOrder))
134 dbMAXLFQ = Application.WorksheetFunction.Max(rgTemporary) 'maxLFQ
135 End If
136
137 If dbMAXLFQ = 0 Then
138 .Cells(i, 6) = "#NA"
139 .Cells(i, 4) = "#NA"
140 .Cells(i, 5) = "#NA"
141 .Cells(i, 7) = "#NA"
142 Else
143
144 For k = lgLFQLBorder To lgLFQRBOrder
145
146 If (intFractionPredicted = k - (lgLFQLBorder - 1)) And (.Cells(i, k) > 0) Then
'also found as predicted
147 .Cells(i, 6) = "TRUE"
148 Else
149
150 End If
151
152 If .Cells(i, k) = dbMAXLFQ Then
153 .Cells(i, 4) = k - (lgLFQLBorder - 1) 'fraction detected
154 intFractionDetected = .Cells(i, 4)
155 Else
156 End If
157
158 If .Cells(i, k) = dbMAXLFQ And (intFractionPredicted = k - (lgLFQLBorder - 1))
Then
159 .Cells(i, 5) = "TRUE" 'found as predicted
160 Else
161
162 End If
163
164 If (.Cells(i, k) > 0) And (.Cells(i, k + 1) > 0) And (.Cells(i, k + 2) > 0) And
(.Cells(i, k + 3) > 0) Then
165 adjunct4 = True 'ueberprueft ob 4 werte nebeneinander liegen
166 Else
167 End If
168
169 If dbMAXLFQ = 0 Then
170 Else
171 If .Cells(i, k) / dbMAXLFQ > 0.1 Then 'zählt LFQ-Werte über 0.1 per
protein
172 lgnLFQCounter = lgnLFQCounter + 1
173 Else
174 End If
175 End If
176 Next k
177
178
179
180 .Cells(i, 7) = intFractionDetected - intFractionPredicted
181 End If
182 If adjunct4 = True And lgnLFQCounter > 3 Then 'wenn die bedingungen zutreffen
183
184 .Range(.Cells(i, 1), .Cells(i, lgLFQRBOrder)).Copy wsT1RMSorted.Cells(counter1,
1)
185 .Range(.Cells(i, 1), .Cells(i, 7)).Copy wsT1RMSortedNorm.Cells(counter1, 1)
186
187 For k = lgLFQLBorder To lgLFQRBOrder
188 wsT1RMSortedNorm.Cells(counter1, k) = .Cells(i, k) / dbMAXLFQ 'normalisiert
die LFQ-Werte
189 Next k
190
191 If .Cells(i, 7) < (-5) Then 'kopiert gleich die kandidaten und nicht-
kandidaten in die entsprechenden worksheets

```

```

192 wsTlRMSortedNorm.Range(wsTlRMSortedNorm.Cells(counter1, 1),
wsTlRMSortedNorm.Cells(counter1, lgLFQRBorder)).Copy wsCandidateList.Cells(counter2, 1)
193 counter2 = counter2 + 1
194 Else
195 wsTlRMSortedNorm.Range(wsTlRMSortedNorm.Cells(counter3, 1),
wsTlRMSortedNorm.Cells(counter3, lgLFQRBorder)).Copy wsNonCandidateList.Cells(counter3, 1)
196 counter3 = counter3 + 1
197 End If
198 counter1 = counter1 + 1
199 Else
200 End If
201
202 End With
203 Next i
204 Application.StatusBar = "Progress: Finalizing: " & ((Round(i / lgTempLastRow, 0)) * 100) &
" %"
205
206 For i = 2 To lgTempLastRow 'filling empty cells (finalizing)
207
208 If wsEvaluated.Cells(i, 4) > 0 And Not wsEvaluated.Cells(i, 5) = True And Not
wsEvaluated.Cells(i, 5) = "#NA" Then 'finalize list
209 wsEvaluated.Cells(i, 5) = "FALSE"
210 Else
211 End If
212
213 If wsEvaluated.Cells(i, 4) > 0 And Not wsEvaluated.Cells(i, 6) = True And Not
wsEvaluated.Cells(i, 6) = "#NA" Then
214 wsEvaluated.Cells(i, 6) = "FALSE"
215 Else
216 End If
217
218 If wsTlRMSorted.Cells(i, 4) > 0 And Not wsTlRMSorted.Cells(i, 5) = True And Not
wsTlRMSorted.Cells(i, 5) = "#NA" Then 'finalize sorted List
219 wsTlRMSorted.Cells(i, 5) = "FALSE"
220 Else
221 End If
222
223 If wsTlRMSorted.Cells(i, 4) > 0 And Not wsTlRMSorted.Cells(i, 6) = True And Not
wsTlRMSorted.Cells(i, 6) = "#NA" Then
224 wsTlRMSorted.Cells(i, 6) = "FALSE"
225 Else
226 End If
227
228 If wsTlRMSortedNorm.Cells(i, 4) > 0 And Not wsTlRMSortedNorm.Cells(i, 5) = True And
Not wsTlRMSortedNorm.Cells(i, 5) = "#NA" Then 'finalized sorted list norm
229 wsTlRMSortedNorm.Cells(i, 5) = "FALSE"
230 Else
231 End If
232
233 If wsTlRMSortedNorm.Cells(i, 4) > 0 And Not wsTlRMSortedNorm.Cells(i, 6) = True And
Not wsTlRMSortedNorm.Cells(i, 6) = "#NA" Then
234 wsTlRMSortedNorm.Cells(i, 6) = "FALSE"
235 Else
236 End If
237
238 If wsCandidateList.Cells(i, 4) > 0 And Not wsCandidateList.Cells(i, 5) = True And
Not wsCandidateList.Cells(i, 5) = "#NA" Then 'finalize candidates
239 wsCandidateList.Cells(i, 5) = "FALSE"
240 Else
241 End If
242
243
244 If wsCandidateList.Cells(i, 4) > 0 And Not wsCandidateList.Cells(i, 6) = True And
Not wsCandidateList.Cells(i, 6) = "#NA" Then
245 wsCandidateList.Cells(i, 6) = "FALSE"
246 Else
247 End If
248
249 If wsNonCandidateList.Cells(i, 4) > 0 And Not wsNonCandidateList.Cells(i, 5) = True
And Not wsNonCandidateList.Cells(i, 5) = "#NA" Then 'finalize non candidates
250 wsNonCandidateList.Cells(i, 6) = "FALSE"
251 Else

```

```

252     End If
253
254
255     If wsNonCandidateList.Cells(i, 4) > 0 And Not wsNonCandidateList.Cells(i, 6) = True
And Not wsNonCandidateList.Cells(i, 6) = "#NA" Then
256         wsNonCandidateList.Cells(i, 5) = "FALSE"
257     Else
258     End If
259 Next i
260
261 Set rgTemporary = wsEvaluated.Range(wsEvaluated.Cells(1, 1), wsEvaluated.Cells(1,
lgLQRBorder))
262
263 rgTemporary.Copy wsTlRMSorted.Cells(1, 1)
264 rgTemporary.Copy wsTlRMSortedNorm.Cells(1, 1)
265 rgTemporary.Copy wsCandidateList.Cells(1, 1)
266 rgTemporary.Copy wsNonCandidateList.Cells(1, 1)
267
268 dtEnd = Now
269 MsgBox "Script for processing data and sorting finished!" & vbCrLf & "Start: " & dtStart &
vbCrLf & "End: " & dtEnd & vbCrLf & "Duration: " & Format(dtEnd - dtStart, "hh:mm:ss") & "
hh:mm:ss", , "Script for Data Processing by MAX"
270 Application.StatusBar = False
271 End Sub
272

```

Script Pearson Correlation

Following script is performing the Pearson correlation analysis to predict pairwise interaction partners after evaluation of the respective correlation window and sample sizes. It requires a preset of six Excel worksheets with data arranged in worksheet 1 as indicated in the flowchart step B

```
1 Attribute VB_Name = "modPearsonCalculation"
2 Option Explicit
3
4 Sub PearsonCalculation()
5
6
7 Dim i As Long
8 Dim k As Long
9 Dim l As Integer
10 Dim m As Long
11 Dim n As Long
12 Dim o As Long
13 Dim p As Long
14 Dim q As Long
15 Dim r As Long
16 Dim s As Long
17 Dim t As Long
18 Dim u As Long
19 Dim v As Long
20 Dim w As Long
21 Dim x As Long
22 Dim y As Long
23 Dim z As Long
24
25 Dim dStart As Date
26 Dim dEnd As Date
27
28
29 Dim intMinimumSampleSize As Integer
30 Dim intLength As Integer
31
32 Dim intWindowSize1 As Integer
33 Dim intWindowSize2 As Integer
34 Dim intWindowSizeShared As Integer
35 Dim intWindowBorder As Integer
36
37 Dim intSampleSizeCounter As Integer
38 Dim varInhalt As Variant
39 Dim varInhalt1 As Variant
40 Dim varInhalt2 As Variant
41 Dim strInhalt1a As String
42 Dim strInhalt2a As String
43 Dim strInhaltL As String
44
45
46 Dim wsTemplate As Worksheet
47 Dim wstarget As Worksheet
48 Dim wsresultsev As Worksheet
49 Dim wsPearsonDev As Worksheet
50 Dim wsPearsonDevEva As Worksheet
51 Dim wsDatapairs As Worksheet
52
53
54 Dim rgPartner1 As Range
55 Dim rgPartner2 As Range
56 Dim rgWindow As Range
57 Dim rgSampleSize As Range
58 Dim rgDeviation As Range
59
60
61 Dim dbPearsonScore As Double
```

```

62
63 Dim lgtemplatelastrow As Long
64 Dim lgTargetLastRow As Long
65 Dim lgEvaluationLastRow As Long
66 Dim lgWindowsLastRow As Long
67 Dim lgSampleSizeLastRow As Long
68 Dim lgDeviationLastRow As Long
69 Dim lgDeviationLastColumn As Long
70 Dim lgTemporaryLastRow As Long
71 Dim lgDatapairsLastColumn As Long
72
73 Dim lgBereichAnfang As Long
74 Dim lgBereichEnde As Long
75
76 Dim Array1() As Variant
77 Dim Array2() As Variant
78 Dim Array3() As Variant
79 Dim Array4() As Variant
80
81 ' _____
82
83
84
85 Set wsTemplate = Sheet1           'anpassen
86 Set wstarget = Sheet2           'anpassen
87 Set wsresultsev = Sheet3       'anpassen
88 Set wsPearsonDev = Sheet4      'anpassen
89 Set wsPearsonDevEva = Sheet6   'anpassen
90 Set wsDatapairs = Sheet5       'anpassen
91
92
93
94 intMinimumSampleSize = 5       'anpassen der Bedingung
95
96
97 lgtemplatelastrow = wsTemplate.Cells(wsTemplate.Rows.Count, "A").End(xlUp).Row
98 dStart = Now
99
100 wstarget.Cells.Clear
101 wsresultsev.Cells.Clear
102 wsPearsonDev.Cells.Clear
103 wsPearsonDevEva.Cells.Clear
104 wsDatapairs.Cells.Clear
105
106 ' _____Convert 0 to #NA_____
107
108 For i = 2 To lgtemplatelastrow
109
110 Application.StatusBar = "Progress: Converting Zero to #NA - " & n / lgtemplatelastrow * 100
& " %"
111
112     For k = 4 To 53
113
114         varInhalt = wsTemplate.Cells(i, k)
115         intLength = Len(varInhalt)
116
117         If intLength = 1 And Not wsTemplate.Cells(i, k) = 1 Then
118             wsTemplate.Cells(i, k) = "#NA"
119         Else
120             End If
121
122     Next k
123 Next i
124
125
126
127
128
129 ' _____START PEARSON CALCULATION_____
130
131
132 q = 2 'startreihe des Ziel-Sheets (Results_Pearson) (fix)

```

```

133
134 For i = 2 To lgtemplatelastrow
135 wsTemplate.Cells(i, 54) = "DONE" 'verhindert doppelte kalkulationen
136
137 Application.StatusBar = "Progress: Calculating Pearson-Correlation - " & i /
lgtemplatelastrow * 100 & " %"
138 intWindowSize1 = wsTemplate.Cells(i, 2).Value
139
140 For k = 2 To lgtemplatelastrow
141
142 intSampleSizeCounter = 0
143
144 If wsTemplate.Cells(k, 54) = "DONE" Then
145 Else
146 If wsTemplate.Cells(i, 1) = wsTemplate.Cells(k, 1) Then 'eigentlich obsolet
147 GoTo HierHer
148 Else
149 intWindowSize2 = wsTemplate.Cells(k, 2).Value
150 End If
151
152 If intWindowSize1 > intWindowSize2 Then
153 intWindowSizeShared = intWindowSize2
154 Else
155 intWindowSizeShared = intWindowSize1
156 End If
157
158 intWindowBorder = 3 + intWindowSizeShared
159
160
161 For l = 4 To intWindowBorder
162
163 varInhalt1 = wsTemplate.Cells(i, l)
164 varInhalt2 = wsTemplate.Cells(k, l)
165
166 If IsNumeric(varInhalt1) And IsNumeric(varInhalt2) Then
167 intSampleSizeCounter = intSampleSizeCounter + 1
168 Else
169 End If
170
171 Next l
172
173 If intSampleSizeCounter >= intMinimumSampleSize Then 'zweite Bedingung
(SampleSize)
174
175 Set rgPartner1 = Range(wsTemplate.Cells(i, 4), wsTemplate.Cells(i, 53))
'columns anpassen wenn sich die template liste verschiebt
176 Set rgPartner2 = Range(wsTemplate.Cells(k, 4), wsTemplate.Cells(k, 53))
177
178 dbPearsonScore = Application.WorksheetFunction.Pearson(rgPartner1, rgPartner2)
179
180 If dbPearsonScore >= 0.7 Then 'dritte Bedingung (Pearson score mindestens gr als
0.7)
181
182 wstarget.Cells(q, 1) = wsTemplate.Cells(i, 1)
183 wstarget.Cells(q, 2) = wsTemplate.Cells(k, 1)
184 wstarget.Cells(q, 3) = dbPearsonScore
185 wstarget.Cells(q, 4) = intWindowSizeShared
186 wstarget.Cells(q, 5) = intSampleSizeCounter
187 q = q + 1
188
189 Else
190 End If
191
192 Else
193 End If
194 End If
195
196 HierHer: Next k
197
198 Next i
199
200 wstarget.Cells(1, 1) = "Partner 1"

```

```

201 wstarget.Cells(1, 2) = "Partner 2"
202 wstarget.Cells(1, 3) = "R"
203 wstarget.Cells(1, 4) = "Window Size"
204 wstarget.Cells(1, 5) = "Sample Size"
205 wsTemplate.Columns(54).Delete
206
207 'Pearson Liste alphabetisch sortieren (klein=Partner1)
208
209 lgTargetLastRow = wstarget.Cells(wstarget.Rows.Count, "A").End(xlUp).Row
210
211 For m = 2 To lgTargetLastRow
212
213     strInhalt1a = wstarget.Cells(m, 1)
214     strInhalt2a = wstarget.Cells(m, 2)
215
216     If strInhalt1a > strInhalt2a Then
217         wstarget.Cells(m, 1) = strInhalt2a
218         wstarget.Cells(m, 2) = strInhalt1a
219
220     Else
221     End If
222
223 Next m
224
225
226
227 'Liste nach Partner A sortieren
228
229 wstarget.Columns("A:D").Sort key1:=wstarget.Range("A1"), order1:=xlAscending, Header:=xlYes
230
231
232
233 'Result Evaluation
234
235 '1) Kopieren des Templates auf ein neues Blatt + Anzahl der Pearson-Werte pro protein
236
237
238
239 Application.StatusBar = "Progress: Evaluation"
240
241 wsresultsev.Cells(1, 1) = "Name"
242 wsresultsev.Cells(1, 2) = "Fraction predicted"
243 wsresultsev.Cells(1, 3) = "Count of R"
244 wsresultsev.Cells(1, 4) = "Average of R"
245 wsresultsev.Cells(1, 5) = "Median of R"
246
247
248 wsresultsev.Cells(1, 18) = "Total Count of R"
249 wsresultsev.Cells(1, 19) = lgTargetLastRow - 1
250 wsresultsev.Cells(2, 18) = "Possible Correlation Events"
251 wsresultsev.Cells(2, 19) = (lgtemplatelastrow * lgtemplatelastrow) / 2
252 wsresultsev.Cells(3, 18) = "Ratio"
253 wsresultsev.Cells(3, 19) = wsresultsev.Cells(1, 19) / wsresultsev.Cells(2, 19)
254
255
256 For s = 2 To lgtemplatelastrow
257
258     Application.StatusBar = "Progress: Evaluation, Count of R, average and median - " & s /
lgtemplatelastrow * 100 & " %"
259
260     r = 0
261     v = 0
262     ReDim Array1(lgTargetLastRow)
263
264     wsresultsev.Cells(s, 1) = wsTemplate.Cells(s, 1) 'namen
265     wsresultsev.Cells(s, 2) = wsTemplate.Cells(s, 2) 'Fraction predicted
266
267     For q = 2 To lgTargetLastRow
268         If wstarget.Cells(q, 1) = wsresultsev.Cells(s, 1) Or wstarget.Cells(q, 2) =
wsresultsev.Cells(s, 1) Then
269             r = r + 1
270             Array1(v) = wstarget.Cells(q, 3).Value

```

```

271         v = v + 1
272     Else
273     End If
274
275     Next q
276
277     wsresultsev.Cells(s, 3) = r
278     ReDim Preserve Array1(r + 1)
279
280     If Not r = 0 Then
281         wsresultsev.Cells(s, 5) = Application.WorksheetFunction.Median(Array1)
282         'Median (per protein)
283         wsresultsev.Cells(s, 4) = Application.WorksheetFunction.Average(Array1)
284         'Average (per protein)
285     Else
286         wsresultsev.Cells(s, 4) = "#NA"
287         wsresultsev.Cells(s, 5) = "#NA"
288     End If
289
290     Next s
291
292
293
294 '2) Average und Median schreiben fr die Proteine
295
296
297
298
299 '3) Windowsize Auswertung
300
301 Set rgWindow = wstarget.Range(wstarget.Cells(1, 4), wstarget.Cells(lgTargetLastRow, 4))
302
303 rgWindow.Copy wsresultsev.Cells(1, 8) 'kopiert die window sizes nach spalte H (alle)
304
305 wsresultsev.Columns("H").Sort key1:=wsresultsev.Range("H1"), order1:=xlAscending,
306 Header:=xlYes 'sortieren ascending
307
308 Set rgWindow = wsresultsev.Range(wsresultsev.Cells(1, 8),
309 wsresultsev.Cells(lgTargetLastRow, 8))
310
311 rgWindow.RemoveDuplicates Array(1), xlYes 'duplikate entfernen (elegant)
312
313 lgWindowsLastRow = wsresultsev.Cells(wsresultsev.Rows.Count, "H").End(xlUp).Row
314
315 For w = 2 To lgWindowsLastRow 'jedes window wird durchsucht_____
316
317     x = 0
318     v = 0
319     ReDim Array2(lgTargetLastRow)
320
321     For u = 2 To lgTargetLastRow
322         Application.StatusBar = "Progress: Window Size Evaluation " & w & " of " &
323         lgWindowsLastRow & " - " & u / lgTargetLastRow * 100 & " %"
324
325         If wsresultsev.Cells(w, 8) = wstarget.Cells(u, 4) Then
326             Array2(v) = wstarget.Cells(u, 3)
327             v = v + 1
328             x = x + 1
329         Else
330         End If
331
332     Next u
333
334
335     ReDim Preserve Array2(x + 1)
336

```

```

337     wsresultsev.Cells(w, 9) = Application.WorksheetFunction.Average(Array2)
'Average (per window)
338     wsresultsev.Cells(w, 10) = Application.WorksheetFunction.Median(Array2)
'Median (per window)
339     wsresultsev.Cells(w, 11) = x
'Count (per window)
340
341
342     Next w
343
344     wsresultsev.Cells(1, 9) = "Average R"
345     wsresultsev.Cells(1, 10) = "Median of R"
346     wsresultsev.Cells(1, 11) = "Count of R"
347
348
349     '4) Sample size Evaluation (to do: window size evaluation kopieren)
350
351     Set rgSampleSize = wstarget.Range(wstarget.Cells(1, 5), wstarget.Cells(lgTargetLastRow,
5))
352
353     rgSampleSize.Copy wsresultsev.Cells(1, 13)     'kopiert die SampleSizes nach Spalte M
354
355     wsresultsev.Columns("M").Sort key1:=wsresultsev.Range("M1"), order1:=xlAscending,
Header:=xlYes     'sortieren ascending
356
357     Set rgSampleSize = wsresultsev.Range(wsresultsev.Cells(1, 13),
wsresultsev.Cells(lgTargetLastRow, 13))
358
359     rgSampleSize.RemoveDuplicates Array(1), xlYes     'duplikate entfernen (elegant)
360
361
362     lgSampleSizeLastRow = wsresultsev.Cells(wsresultsev.Rows.Count, "M").End(xlUp).Row
363
364     Set rgSampleSize = wsresultsev.Range(wsresultsev.Cells(2, "M"),
wsresultsev.Cells(lgSampleSizeLastRow, "M"))
365
366     rgSampleSize.Copy
367     wsDatapairs.Range("A1").PasteSpecial , , , True
368
369
370     For w = 2 To lgSampleSizeLastRow 'jede SampleSize wird durchsucht_____
371
372
373         x = 0
374         v = 0
375         ReDim Array3(lgTargetLastRow)
376
377         For u = 2 To lgTargetLastRow
378
379             Application.StatusBar = "Progress: Sample Size Evaluation " & w & " of " &
lgSampleSizeLastRow & " - " & u / lgTargetLastRow * 100 & " %"
380
381             If wsresultsev.Cells(w, 13) = wstarget.Cells(u, 5) Then
382                 lgTemporaryLastRow = wsDatapairs.Cells(wsDatapairs.Columns.Count, w -
1).End(xlUp).Row 'test
383                 Array3(v) = wstarget.Cells(u, 3)
384                 v = v + 1
385                 x = x + 1
386                 wsDatapairs.Cells(lgTemporaryLastRow + 1, w - 1) = wstarget.Cells(u, 3)
'test
387             Else
388                 End If
389
390         Next u
391
392         ReDim Preserve Array3(x + 1)
393
394         wsresultsev.Cells(w, 14) = Application.WorksheetFunction.Average(Array3)
'Average (per SampleSize)
395         wsresultsev.Cells(w, 15) = Application.WorksheetFunction.Median(Array3)
'Median (per SampleSize)

```

```

396         wsresultsev.Cells(w, 16) = x
397     'Count (per SampleSize)
398
399     Next w
400
401
402
403     wsresultsev.Cells(1, 14) = "Average R"
404     wsresultsev.Cells(1, 15) = "Median of R"
405     wsresultsev.Cells(1, 16) = "Count of R"
406
407
408
409 ' 5) Deviation Groups schreiben
410
411 Application.StatusBar = "Progress: Finding Deviation groups"
412
413
414 Set rgDeviation = wsTemplate.Range(wsTemplate.Cells(1, 3),
wsTemplate.Cells(lgtemplatelastrow, 3))
415
416 rgDeviation.Copy wsPearsonDev.Cells(1, 1) 'kopiert die Deviation Groups nach Spalte A
417
418 wsPearsonDev.Columns("A").Sort key1:=wsPearsonDev.Range("A1"), order1:=xlAscending,
Header:=xlYes 'sortieren ascending
419
420 Set rgDeviation = wsPearsonDev.Range(wsPearsonDev.Cells(1, 1),
wsPearsonDev.Cells(lgtemplatelastrow, 1))
421
422 rgDeviation.RemoveDuplicates Array(1), xlYes 'duplikate entfernen (elegant)
423
424
425 lgDeviationLastRow = wsPearsonDev.Cells(wsPearsonDev.Rows.Count, "A").End(xlUp).Row
426
427 Set rgDeviation = wsPearsonDev.Range(wsPearsonDev.Cells(2, 1),
wsPearsonDev.Cells(lgDeviationLastRow, 1))
428
429 rgDeviation.Copy wsPearsonDevEva.Cells(2, 1)
430 wsPearsonDevEva.Cells(1, 1) = "Deviation Group"
431 wsPearsonDevEva.Cells(1, 2) = "Average R"
432 wsPearsonDevEva.Cells(1, 3) = "Median of R"
433 wsPearsonDevEva.Cells(1, 4) = "Count of R"
434 rgDeviation.Copy
435 wsPearsonDev.Range(wsPearsonDev.Cells(1, 1), wsPearsonDev.Cells(1,
1)).PasteSpecial , , , True
436 rgDeviation.Delete
437
438 ' 5.1) Deviation groups fuellen
439
440 lgDeviationLastColumn = wsPearsonDev.Cells(1,
wsPearsonDev.Columns.Count).End(xlToLeft).Column
441
442
443 For i = 1 To lgDeviationLastColumn
444
445
446     x = 0
447     v = 0
448     ReDim Array4(lgTargetLastRow)
449
450     For m = 2 To lgtemplatelastrow
451         Application.StatusBar = "Progress: Fill deviation group " & i & " of " &
lgDeviationLastColumn & " - " & m / lgtemplatelastrow * 100 & " %"
452         If wsTemplate.Cells(m, 3) = wsPearsonDev.Cells(1, i) Then
453             For n = 2 To lgTargetLastRow
454                 If (wsTemplate.Cells(m, 1) = wstarget.Cells(n, 1)) Or
(wsTemplate.Cells(m, 1) = wstarget.Cells(n, 2)) Then
455                     lgTemporaryLastRow = wsPearsonDev.Cells(wsPearsonDev.Rows.Count,
i).End(xlUp).Row
456                     o = lgTemporaryLastRow + 1
457                     wsPearsonDev.Cells(o, i) = wstarget.Cells(n, 3)

```

```

458         Array4(v) = wstarget.Cells(n, 3)
459         v = v + 1
460         x = x + 1
461     Else
462         End If
463     Next n
464 Else
465     End If
466 Next m
467
468 ReDim Preserve Array4(x + 1)
469
470
471 If x > 0 Then
472     wsPearsonDevEva.Cells(i + 1, 4) = x
473     'Count (per Deviation Group)
474     wsPearsonDevEva.Cells(i + 1, 2) = Application.WorksheetFunction.Average(Array4)
475     'Average (per Deviation Group)
476     wsPearsonDevEva.Cells(i + 1, 3) = Application.WorksheetFunction.Median(Array4)
477     'Median (per Deviation Group)
478 Else
479     wsPearsonDevEva.Cells(i + 1, 4) = 0
480     'Count (per Deviation Group)
481     wsPearsonDevEva.Cells(i + 1, 2) = "#NA"
482     'Average (per Deviation Group)
483     wsPearsonDevEva.Cells(i + 1, 3) = "#NA"
484     'Median (per Deviation Group)
485 End If
486 Next i
487
488 '6) Pearson and Datapair-Count
489
490 lgDatapairsLastColumn = wsDatapairs.Cells(1,
wsDatapairs.Columns.Count).End(xlToLeft).Column
491
492 For i = 1 To lgDatapairsLastColumn
493     q = 2
494     For k = 2 To lgTargetLastRow
495         If wstarget.Cells(k, 5) = wsDatapairs.Cells(1, i) Then
496             wsDatapairs.Cells(q, i) = wstarget.Cells(k, 3)
497             q = q + 1
498         Else
499             End If
500     Next k
501 Next i
502
503
504
505
506
507
508
509 Application.StatusBar = "Progress: Complete"
510 dEnd = Now
511 MsgBox "Script for calculation of Pearson-Correlation finished!" & vbCrLf & "Start: " &
dtStart & vbCrLf & "End: " & dtEnd & vbCrLf & "Duration: " & Format(dtEnd - dtStart,
"hh:mm:ss") & " hh:mm:ss", , "Script by MAX"
512
513 Debug.Print "Ende"
514 Application.StatusBar = False
515
516
517 End Sub
518
519

```

Script “Confirmed Interactions”

This script is used to compare the predictions of the pearson correlation analysis with database confirmed protein-protein interactions. Three worksheets need to be provided whereas sheet 1 and 2 should list the results from the pearson correlation analysis (sheet 1) or database confirmed interactions (sheet 2). Columns should be arranged as indicated in the flowchart step C.

```
1 Attribute VB_Name = "Module1"
2 Option Explicit
3
4 Sub deltaFEvaluation()
5
6 Application.ScreenUpdating = False
7
8 Dim i As Long
9 Dim k As Long
10 Dim m As Long
11 Dim q As Long
12
13 Dim intCounter As Integer
14
15 Dim wsPearson As Worksheet
16 Dim wsCandidates As Worksheet
17 Dim wsEvaluation As Worksheet
18
19 Dim lgPearsonLastRow As Long
20 Dim lgCandidatesLastRow As Long
21 Dim lgTempLastRow As Long
22 Dim lgTemp As Long
23
24 Dim dtStart As Date
25 Dim dtEnd As Date
26
27 dtStart = Now
28
29
30 Dim lgTempLastColumn As Long
31
32 Dim boo1 As Boolean
33 Dim boo2 As Boolean
34
35 Set wsPearson = Sheet1
36 Set wsCandidates = Sheet2
37 Set wsEvaluation = Sheet3
38
39 intCounter = 20
40 q = 90
41
42 lgPearsonLastRow = LastRow(1, wsPearson)
43 lgCandidatesLastRow = LastRow(1, wsCandidates)
44 lgTempLastColumn = LastColumn(1, wsCandidates)
45
46 '_____Initialize_____'
47
48
49 With wsPearson
50 .Cells(1, 7) = "Partner 1 Fraction detected"
51 .Cells(1, 8) = "Partner 2 Fraction detected"
52 .Cells(1, 9) = "dFraction"
53 End With
54
55 With wsEvaluation
56 .Cells.Delete
57 lgTempLastRow = LastRow(1, wsEvaluation)
```

```

58 .Cells(lgTempLastRow, 1) = "Total"
59 .Cells(lgTempLastRow + 1, 1) = "dF"
60 .Cells(lgTempLastRow + 1, 2) = "Count"
61 .Cells(lgTempLastRow + 1, 3) = "% of total"
62
63 .Cells(1, lgTempLastColumn + 5) = "Total"
64 lgTempLastColumn = LastColumn(1, wsEvaluation)
65
66 For m = 0 To intCounter
67 lgTempLastRow = LastRow(1, wsEvaluation)
68 .Cells(lgTempLastRow + 1, 1) = m
69 .Cells(2, lgTempLastColumn) = m
70 lgTempLastColumn = lgTempLastColumn + 1
71 Next m
72
73 lgTempLastRow = LastRow(1, wsEvaluation)
74 .Cells(lgTempLastRow + 2, 1) = "Confirmed"
75 .Cells(lgTempLastRow + 3, 1) = "dF"
76 .Cells(lgTempLastRow + 3, 2) = "Count"
77 .Cells(lgTempLastRow + 3, 3) = "% of total"
78
79 lgTempLastColumn = lgTempLastColumn + 2
80 .Cells(1, lgTempLastColumn) = "Confirmed"
81
82 For m = 0 To intCounter
83 lgTempLastRow = LastRow(1, wsEvaluation)
84 .Cells(lgTempLastRow + 1, 1) = m
85 .Cells(2, lgTempLastColumn) = m
86 lgTempLastColumn = lgTempLastColumn + 1
87 Next m
88
89 lgTempLastRow = LastRow(1, wsEvaluation)
90 .Cells(lgTempLastRow + 2, 1) = "Unconfirmed"
91 .Cells(lgTempLastRow + 3, 1) = "dF"
92 .Cells(lgTempLastRow + 3, 2) = "Count"
93 .Cells(lgTempLastRow + 3, 3) = "% of total"
94
95 lgTempLastColumn = lgTempLastColumn + 2
96 .Cells(1, lgTempLastColumn) = "Unconfirmed"
97
98 For m = 0 To intCounter
99 lgTempLastRow = LastRow(1, wsEvaluation)
100 .Cells(lgTempLastRow + 1, 1) = m
101 .Cells(2, lgTempLastColumn) = m
102 lgTempLastColumn = lgTempLastColumn + 1
103 Next m
104
105 End With
106
107 ''_____Add F and dF to Pearson List_____''
108
109
110
111 With wsPearson
112 lgTempLastColumn = LastColumn(1, wsPearson)
113
114 For i = 2 To lgPearsonLastRow
115
116
117 If (i / lgPearsonLastRow) * 100 = 100 - q Then 'Status bar
118 Application.ScreenUpdating = True
119 Application.StatusBar = "Progress: Evaluation - " & (100 - q) & "%"
120 q = q - 10
121 Application.ScreenUpdating = False
122 Else
123 End If
124
125
126 bool = False
127 boo2 = False
128 For k = 2 To lgCandidatesLastRow
129

```

```

130     If .Cells(i, 1) = wsCandidates.Cells(k, 1) Then
131         .Cells(i, 7) = wsCandidates.Cells(k, 4)
132         bool = True
133     Else
134     End If
135
136     If .Cells(i, 2) = wsCandidates.Cells(k, 1) Then
137         .Cells(i, 8) = wsCandidates.Cells(k, 4)
138         boo2 = True
139     Else
140     End If
141
142     If bool And boo2 Then
143         .Cells(i, 9) = Abs(.Cells(i, 7).Value - .Cells(i, 8).Value)
144
145         For m = 0 To intCounter
146             If .Cells(i, 9).Value = m Then
147
148                 wsEvaluation.Cells(m + 3, 2) = wsEvaluation.Cells(m + 3, 2) + 1 'total
149                 lgTempLastRow = LastRow(m + 10, wsEvaluation) + 1
150                 wsEvaluation.Cells(lgTempLastRow, m + 10) = .Cells(i, 3)
151
152                 If .Cells(i, 6) = "Confirmed" Then
153                     wsEvaluation.Cells(m + 27, 2) = wsEvaluation.Cells(m + 27, 2) + 1
154                     'confirmed
155                     lgTempLastRow = LastRow(m + 33, wsEvaluation) + 1
156                     wsEvaluation.Cells(lgTempLastRow, m + 33) = .Cells(i, 3)
157                 Else
158                     wsEvaluation.Cells(m + 51, 2) = wsEvaluation.Cells(m + 51, 2) + 1
159                     'unconfirmed
160                     lgTempLastRow = LastRow(m + 56, wsEvaluation) + 1
161                     wsEvaluation.Cells(lgTempLastRow, m + 56) = .Cells(i, 3)
162                 End If
163             Exit For
164         Else
165         End If
166     Next m
167
168     Exit For
169 Else
170 End If
171
172
173 Next k
174 Next i
175 Next i
176
177 End With
178
179
180
181 ''_____Finalize_____
182
183 With wsEvaluation
184     lgTemp = 0 '% total
185     For i = 3 To 23
186         If .Cells(i, 2) = "" Then
187             .Cells(i, 2) = 0
188         Else
189         End If
190     lgTemp = lgTemp + .Cells(i, 2)
191 Next i
192
193 For i = 3 To 23
194     .Cells(i, 3) = Round((.Cells(i, 2) / lgTemp) * 100, 5)
195 Next i
196
197
198 lgTemp = 0 '% Confirmed
199 For i = 27 To 47

```

```

200 If .Cells(i, 2) = "" Then
201 .Cells(i, 2) = 0
202 Else
203 End If
204 lgTemp = lgTemp + .Cells(i, 2)
205 Next i
206
207 For i = 27 To 47
208 If .Cells(i, 2) = "" Then
209 .Cells(i, 2) = 0
210 Else
211 End If
212 .Cells(i, 3) = Round((.Cells(i, 2) / lgTemp) * 100, 5)
213 Next i
214
215 lgTemp = 0 '% Unconfirmed
216 For i = 51 To 71
217 If .Cells(i, 2) = "" Then
218 .Cells(i, 2) = 0
219 Else
220 End If
221 lgTemp = lgTemp + .Cells(i, 2)
222 Next i
223
224 For i = 51 To 71
225 .Cells(i, 3) = Round((.Cells(i, 2) / lgTemp) * 100, 5)
226 Next i
227
228 End With
229
230
231
232 '_____END_____''
233 Application.ScreenUpdating = True
234
235 ThisWorkbook.Save
236
237 dtEnd = Now
238 MsgBox "Script finished!" & vbCrLf & "Start: " & dtStart & vbCrLf & "End: " & dtEnd &
vbCrLf & "Duration: " & Format(dtEnd - dtStart, "hh:mm:ss") & " hh:mm:ss", , "Script by MAX"
239
240
241 End Sub
242
243
244 Sub FindHC()
245
246 Dim i As Long
247
248 Dim wsPearson As Worksheet
249 Dim wsHC As Worksheet
250
251 Dim lgPearsonLastRow As Long
252 Dim lgHCLastRow As Long
253
254 Dim value1 As Double 'pearson
255 Dim value2 As Double 'dF
256
257 Dim rgTemp As Range
258
259 Set wsPearson = Sheet1
260 Set wsHC = Sheet4
261
262
263 lgPearsonLastRow = LastRow(1, wsPearson)
264
265
266 With wsPearson
267
268 Set rgTemp = .Range(.Cells(1, 1), .Cells(1, 9))
269
270 rgTemp.Copy wsHC.Cells(1, 1)

```

```

271
272
273 For i = 2 To lgPearsonLastRow
274
275     value1 = .Cells(i, 3)
276     value2 = .Cells(i, 9)
277
278     If value1 >= 0.89 And value2 <= 1 Then
279         Set rgTemp = .Range(.Cells(i, 1), .Cells(i, 9))
280         lgHCLastRow = LastRow(1, wsHC) + 1
281         rgTemp.Copy wsHC.Cells(lgHCLastRow, 1)
282     Else
283     End If
284
285 Next i
286
287 End With
288
289 ThisWorkbook.Save
290
291
292 End Sub
293
294
295
296
297 Function LastRow(lgColumn As Long, wsSheet As Worksheet) As Long
298     With wsSheet
299         LastRow = .Cells(.Rows.Count, lgColumn).End(xlUp).Row
300     End With
301 End Function
302
303
304 Function LastColumn(lgRow As Long, wsSheet As Worksheet) As Long
305     With wsSheet
306         LastColumn = .Cells(lgRow, .Columns.Count).End(xlToLeft).Column
307     End With
308 End Function
309

```

Script “dF Evaluation”

This script is used to analyze the pairwise interaction predictions from the results of the Pearson correlation analysis. It assigns a co-elution value (dF; the difference of elution fraction between two partners) and decides if a correlation event is to be considered to be a high confidence interaction as introduced in the first included publication. For this script, data needs to be imported from the previous scripts B and C as indicated in the flowchart step D. The script is split into two subroutines that need to be activated sequentially.

```
1 Attribute VB_Name = "Module1"
2 Option Explicit
3
4 Sub deltaFEvaluation()
5
6 Application.ScreenUpdating = False
7
8 Dim i As Long
9 Dim k As Long
10 Dim m As Long
11 Dim q As Long
12
13 Dim intCounter As Integer
14
15 Dim wsPearson As Worksheet
16 Dim wsCandidates As Worksheet
17 Dim wsEvaluation As Worksheet
18
19 Dim lgPearsonLastRow As Long
20 Dim lgCandidatesLastRow As Long
21 Dim lgTempLastRow As Long
22 Dim lgTemp As Long
23
24 Dim dtStart As Date
25 Dim dtEnd As Date
26
27 dtStart = Now
28
29
30 Dim lgTempLastColumn As Long
31
32 Dim boo1 As Boolean
33 Dim boo2 As Boolean
34
35 Set wsPearson = Sheet1
36 Set wsCandidates = Sheet2
37 Set wsEvaluation = Sheet3
38
39 intCounter = 20
40 q = 90
41
42 lgPearsonLastRow = LastRow(1, wsPearson)
43 lgCandidatesLastRow = LastRow(1, wsCandidates)
44 lgTempLastColumn = LastColumn(1, wsCandidates)
45
46 '''_____Initialize_____'''
47
48
49 With wsPearson
50 .Cells(1, 7) = "Partner 1 Fraction detected"
51 .Cells(1, 8) = "Partner 2 Fraction detected"
52 .Cells(1, 9) = "dFraction"
53 End With
54
55 With wsEvaluation
56 .Cells.Delete
57 lgTempLastRow = LastRow(1, wsEvaluation)
```

```

58 .Cells(lgTempLastRow, 1) = "Total"
59 .Cells(lgTempLastRow + 1, 1) = "dF"
60 .Cells(lgTempLastRow + 1, 2) = "Count"
61 .Cells(lgTempLastRow + 1, 3) = "% of total"
62
63 .Cells(1, lgTempLastColumn + 5) = "Total"
64 lgTempLastColumn = LastColumn(1, wsEvaluation)
65
66 For m = 0 To intCounter
67 lgTempLastRow = LastRow(1, wsEvaluation)
68 .Cells(lgTempLastRow + 1, 1) = m
69 .Cells(2, lgTempLastColumn) = m
70 lgTempLastColumn = lgTempLastColumn + 1
71 Next m
72
73 lgTempLastRow = LastRow(1, wsEvaluation)
74 .Cells(lgTempLastRow + 2, 1) = "Confirmed"
75 .Cells(lgTempLastRow + 3, 1) = "dF"
76 .Cells(lgTempLastRow + 3, 2) = "Count"
77 .Cells(lgTempLastRow + 3, 3) = "% of total"
78
79 lgTempLastColumn = lgTempLastColumn + 2
80 .Cells(1, lgTempLastColumn) = "Confirmed"
81
82 For m = 0 To intCounter
83 lgTempLastRow = LastRow(1, wsEvaluation)
84 .Cells(lgTempLastRow + 1, 1) = m
85 .Cells(2, lgTempLastColumn) = m
86 lgTempLastColumn = lgTempLastColumn + 1
87 Next m
88
89 lgTempLastRow = LastRow(1, wsEvaluation)
90 .Cells(lgTempLastRow + 2, 1) = "Unconfirmed"
91 .Cells(lgTempLastRow + 3, 1) = "dF"
92 .Cells(lgTempLastRow + 3, 2) = "Count"
93 .Cells(lgTempLastRow + 3, 3) = "% of total"
94
95 lgTempLastColumn = lgTempLastColumn + 2
96 .Cells(1, lgTempLastColumn) = "Unconfirmed"
97
98 For m = 0 To intCounter
99 lgTempLastRow = LastRow(1, wsEvaluation)
100 .Cells(lgTempLastRow + 1, 1) = m
101 .Cells(2, lgTempLastColumn) = m
102 lgTempLastColumn = lgTempLastColumn + 1
103 Next m
104
105 End With
106
107 ''_____Add F and dF to Pearson List_____''
108
109
110
111 With wsPearson
112 lgTempLastColumn = LastColumn(1, wsPearson)
113
114 For i = 2 To lgPearsonLastRow
115
116
117 If (i / lgPearsonLastRow) * 100 = 100 - q Then 'Status bar
118 Application.ScreenUpdating = True
119 Application.StatusBar = "Progress: Evaluation - " & (100 - q) & "%"
120 q = q - 10
121 Application.ScreenUpdating = False
122 Else
123 End If
124
125
126 bool = False
127 boo2 = False
128 For k = 2 To lgCandidatesLastRow
129

```

```

130     If .Cells(i, 1) = wsCandidates.Cells(k, 1) Then
131         .Cells(i, 7) = wsCandidates.Cells(k, 4)
132         bool = True
133     Else
134     End If
135
136     If .Cells(i, 2) = wsCandidates.Cells(k, 1) Then
137         .Cells(i, 8) = wsCandidates.Cells(k, 4)
138         boo2 = True
139     Else
140     End If
141
142     If bool And boo2 Then
143         .Cells(i, 9) = Abs(.Cells(i, 7).Value - .Cells(i, 8).Value)
144
145         For m = 0 To intCounter
146             If .Cells(i, 9).Value = m Then
147
148                 wsEvaluation.Cells(m + 3, 2) = wsEvaluation.Cells(m + 3, 2) + 1 'total
149                 lgTempLastRow = LastRow(m + 10, wsEvaluation) + 1
150                 wsEvaluation.Cells(lgTempLastRow, m + 10) = .Cells(i, 3)
151
152                 If .Cells(i, 6) = "Confirmed" Then
153                     wsEvaluation.Cells(m + 27, 2) = wsEvaluation.Cells(m + 27, 2) + 1
154                     'confirmed
155                     lgTempLastRow = LastRow(m + 33, wsEvaluation) + 1
156                     wsEvaluation.Cells(lgTempLastRow, m + 33) = .Cells(i, 3)
157                 Else
158                     wsEvaluation.Cells(m + 51, 2) = wsEvaluation.Cells(m + 51, 2) + 1
159                     'unconfirmed
160                     lgTempLastRow = LastRow(m + 56, wsEvaluation) + 1
161                     wsEvaluation.Cells(lgTempLastRow, m + 56) = .Cells(i, 3)
162                 End If
163             Exit For
164         Else
165         End If
166     Next m
167
168     Exit For
169 Else
170 End If
171
172 Next k
173
174 Next i
175 Next i
176
177 End With
178
179
180
181 ''_____Finalize_____
182
183 With wsEvaluation
184     lgTemp = 0 '% total
185     For i = 3 To 23
186         If .Cells(i, 2) = "" Then
187             .Cells(i, 2) = 0
188         Else
189         End If
190     lgTemp = lgTemp + .Cells(i, 2)
191 Next i
192
193 For i = 3 To 23
194     .Cells(i, 3) = Round((.Cells(i, 2) / lgTemp) * 100, 5)
195 Next i
196
197
198 lgTemp = 0 '% Confirmed
199 For i = 27 To 47

```

```

200 If .Cells(i, 2) = "" Then
201 .Cells(i, 2) = 0
202 Else
203 End If
204 lgTemp = lgTemp + .Cells(i, 2)
205 Next i
206
207 For i = 27 To 47
208 If .Cells(i, 2) = "" Then
209 .Cells(i, 2) = 0
210 Else
211 End If
212 .Cells(i, 3) = Round((.Cells(i, 2) / lgTemp) * 100, 5)
213 Next i
214
215 lgTemp = 0 '% Unconfirmed
216 For i = 51 To 71
217 If .Cells(i, 2) = "" Then
218 .Cells(i, 2) = 0
219 Else
220 End If
221 lgTemp = lgTemp + .Cells(i, 2)
222 Next i
223
224 For i = 51 To 71
225 .Cells(i, 3) = Round((.Cells(i, 2) / lgTemp) * 100, 5)
226 Next i
227
228 End With
229
230
231
232 ' _____END_____ '
233 Application.ScreenUpdating = True
234
235 ThisWorkbook.Save
236
237 dtEnd = Now
238 MsgBox "Script finished!" & vbCrLf & "Start: " & dtStart & vbCrLf & "End: " & dtEnd &
vbCrLf & "Duration: " & Format(dtEnd - dtStart, "hh:mm:ss") & " hh:mm:ss", , "Script by MAX"
239
240
241 End Sub
242
243
244 Sub FindHC()
245
246 Dim i As Long
247
248 Dim wsPearson As Worksheet
249 Dim wsHC As Worksheet
250
251 Dim lgPearsonLastRow As Long
252 Dim lgHCLastRow As Long
253
254 Dim value1 As Double 'pearson
255 Dim value2 As Double 'dF
256
257 Dim rgTemp As Range
258
259 Set wsPearson = Sheet1
260 Set wsHC = Sheet4
261
262
263 lgPearsonLastRow = LastRow(1, wsPearson)
264
265
266 With wsPearson
267
268 Set rgTemp = .Range(.Cells(1, 1), .Cells(1, 9))
269
270 rgTemp.Copy wsHC.Cells(1, 1)

```

```

271
272
273 For i = 2 To lgPearsonLastRow
274
275     value1 = .Cells(i, 3)
276     value2 = .Cells(i, 9)
277
278     If value1 >= 0.89 And value2 <= 1 Then
279         Set rgTemp = .Range(.Cells(i, 1), .Cells(i, 9))
280         lgHCLastRow = LastRow(1, wsHC) + 1
281         rgTemp.Copy wsHC.Cells(lgHCLastRow, 1)
282     Else
283     End If
284
285 Next i
286
287 End With
288
289 ThisWorkbook.Save
290
291
292 End Sub
293
294
295
296
297 Function LastRow(lgColumn As Long, wsSheet As Worksheet) As Long
298     With wsSheet
299         LastRow = .Cells(.Rows.Count, lgColumn).End(xlUp).Row
300     End With
301 End Function
302
303
304 Function LastColumn(lgRow As Long, wsSheet As Worksheet) As Long
305     With wsSheet
306         LastColumn = .Cells(lgRow, .Columns.Count).End(xlToLeft).Column
307     End With
308 End Function
309

```

Script "Abundance Rating"

This script is used to normalize protein abundance to enable comparisons between different experiments. Protein LFQ-Values are normalized, sorted and ranked as described in the second included publication. Layout and contents of the required worksheets as shown in the Flowchart step E. The second worksheet (e.g. "MapMan") optional and may be left empty. Protein annotation data can be included there and will be cross-referenced to the proteins of interest if provided.

```
1 Attribute VB_Name = "modAbundanceRanking"
2 Option Explicit
3
4 Sub AbundancyRanking() 'ist abundancy ein korrektes Wort??
5
6 Dim i As Integer
7 Dim k As Double
8
9 Dim lgLastRow As Long
10 Dim lgMapManLastRow As Long
11 Dim lgTargetLastRow As Long
12 Dim dbLFQValue As Double
13 Dim dbLFQTotalSum As Double
14 Dim dbLFQProteinSum As Double
15 Dim dbMaxLFQ As Double
16 Dim dbMolW As Double
17 Dim lgMaxRank As Long
18 Dim lgLeftBorder As Long
19 Dim lgRightBorder As Long
20 Dim lgCategorySize As Long
21 Dim dbCategoryStep As Double
22 Dim rngLFQ As Range
23 Dim rngAbundanceCo As Range
24 Dim rngAbundancRanking As Range
25 Dim wsSource As Worksheet
26 Dim wsTarget As Worksheet
27 Dim wsMapMan As Worksheet
28
29 Dim dtStart As Date
30 Dim dtEnd As Date
31
32 Set wsSource = Sheet1
33 Set wsTarget = Sheet3
34 Set wsMapMan = Sheet2
35
36
37 dtStart = Now
38
39
40 '_____1) Adding all LFQ-Values to variable
dbLFQTotalSum_____
41
42 lgLastRow = wsSource.Cells(wsSource.Rows.Count, 1).End(xlUp).Row
43 dbLFQTotalSum = 0
44 For i = 2 To lgLastRow
45 Application.StatusBar = "Progress: Adding LFQ-Values of each Protein - " & i & " of " &
lgLastRow
46 For k = 10 To 59
47 dbLFQValue = wsSource.Cells(i, k).Value
48 dbLFQTotalSum = dbLFQTotalSum + dbLFQValue
49 Next k
50 Next i
51
52
53 '_____1.) Writing Abundance-Score and
Ranking_____
54
55
56 For i = 2 To lgLastRow ' erste i schleife schreibt den Abundance coefficient
((ProteinSum/TotalSum)/MW)
57 With wsSource
```

```

58 dbLFQProteinSum = 0
59 dbMolW = .Cells(i, 2).Value
60 Application.StatusBar = "Progress: Writing Abundance coefficient - " & i & " of " &
lgLastRow
61
62 For k = 10 To 59 'k-schleife addiert alle LFQ-Werte des Proteins zur Variablen dbLFQValue
63 dbLFQValue = .Cells(i, k).Value
64 dbLFQProteinSum = dbLFQProteinSum + dbLFQValue
65 Next k
66
67 wsTarget.Cells(i, 9) = (dbLFQProteinSum / dbLFQTotalSum) / dbMolW 'coeffizient aus dem
quotienten des normalisierten LFQ-Wertes und dem Mol-Gewicht[kDa]
68
69 .Range(.Cells(i, 1), .Cells(i, 5)).Copy wsTarget.Cells(i, 1)
70
71 End With
72
73
74 Next i
75
76 Set rngAbundanceCo = wsTarget.Range(wsTarget.Cells(1, 9), wsTarget.Cells(lgLastRow, 9)) '
definiert die Gesamtheit der Koeffizienten als range lngAbundanceCo
77
78 For i = 2 To lgLastRow 'zweite i-Schleife berprft, ob der jeweilige Koeffizient in der I
der k-grist oder nicht
79 Application.StatusBar = "Progress: Obtaining Abundance Rank - " & i & " of " & lgLastRow
80
81 For k = 1 To lgLastRow 'variable k lndet die folgende large-funktion nach dem k-grm wert
suchen von 1 zu lgLastRow (jedes Protein ist im ranking vertreten)
82
83 If wsTarget.Cells(i, 9) = WorksheetFunction.Large(rngAbundanceCo, k) Then
'vergleicht den Koeffizienten der Reihe i mit dem k-grm Wert
84 wsTarget.Cells(i, 10) = k
85 Exit For
86 Else
87 End If
88
89 Next k
90 Next i
91
92
93 '_____2) Writing Abundance-
Category_____
94
95
96 With wsTarget
97 lgTargetLastRow = .Cells(.Rows.Count, "A").End(xlUp).Row
98 Set rngAbundancRanking = .Range(.Cells(2, 10), .Cells(lgTargetLastRow, 10)) 'spalte in der
die rankings stehen
99 End With
100
101 lgMaxRank = Application.WorksheetFunction.Max(rngAbundancRanking)
102
103 lgCategorySize = 100
104 'dbCategoryStep = Application.WorksheetFunction.Round(lgMaxRank / lgCategorySize, 0)
105 dbCategoryStep = (lgMaxRank / lgCategorySize)
106
107
108 For k = 2 To lgTargetLastRow 'k-schleife rastert jede zeile ab
109 Application.StatusBar = "Progress: Summarizing to Categories - " & k & " of " & lgLastRow
110
111 For i = 1 To lgCategorySize 'i-schleife berprft
112
113 lgLeftBorder = ((i * dbCategoryStep) - dbCategoryStep) 'kalkuliert Grenzen fr
jede Kategorie
114 lgRightBorder = i * dbCategoryStep
115
116
117 If lgLeftBorder < wsTarget.Cells(k, 10) And wsTarget.Cells(k, 10) <=
lgRightBorder Then
118 wsTarget.Cells(k, 11) = i
119 Exit For

```

```

120         Else
121             End If
122     Next i
123 Next k
124
125 Next k
126
127
128 ' _____2) Writing MapMan-
Attributes_____
129
130
131 lgMapManLastRow = wsMapMan.Cells(wsMapMan.Rows.Count, 1).End(xlUp).Row
132
133
134
135 For i = 2 To lgTargetLastRow
136 Application.StatusBar = "Progress: Writing MapMan-Attributes - " & i & " of " & lgLastRow
137
138     For k = 2 To lgMapManLastRow
139
140         With wsTarget
141
142             If LCase(.Cells(i, 1)) = LCase(wsMapMan.Cells(k, 1)) Then
143
144                 .Cells(i, 12) = wsMapMan.Cells(k, 3) ' Schreibt den Name
145                 .Cells(i, 13) = wsMapMan.Cells(k, 4) ' Schreibt Description
146                 .Cells(i, 14) = wsMapMan.Cells(k, 12) ' Schreibt Bin l(short)
147                 .Cells(i, 15) = wsMapMan.Cells(k, 8) ' Schreibt Suba
148
149             Else
150
151             End If
152
153         End With
154     Next k
155
156
157 Next i
158
159
160 With wsTarget
161
162 .Cells(1, 1) = "Protein ID"
163 .Cells(1, 2) = "Mol. weight [kDa]"
164 .Cells(1, 3) = "Fraction Predicted (Rounded)"
165 .Cells(1, 4) = "Fraction detected"
166 .Cells(1, 5) = "Deviation"
167 .Cells(1, 6) = "Found as predicted"
168 .Cells(1, 7) = "Also found as predicted"
169 .Cells(1, 8) = "Delta Found (Found-Predicted)"
170 .Cells(1, 9) = "Abundance coefficient ((ProteinSum/TotalSum)/MW)"
171 .Cells(1, 10) = "Abundance Rank"
172 .Cells(1, 11) = "Abundance Category"
173 .Cells(1, 12) = "Name"
174 .Cells(1, 13) = "Description"
175 .Cells(1, 14) = "Bin"
176 .Cells(1, 15) = "Suba"
177
178
179 End With
180
181 dtEnd = Now
182 MsgBox "Script finished the calculation and ranking of the abundance coefficient!" & vbCrLf
& "Proteins were rated according to the rank of their respective coefficient." & vbCrLf &
vbCrLf & "Start: " & dtStart & vbCrLf & "End: " & dtEnd & vbCrLf & "Duration: " & Format(dtEnd
- dtStart, "hh:mm:ss") & " hh:mm:ss", , "Script for Rating Relative Protein Abundance by MAX"
183 Debug.Print "Ende"
184 Application.StatusBar = False
185
186
187 End Sub

```

