# Quantitative proteomics *on the fly*

ISBN: 978-90-393-4989-2

Printed by PrintPartners Ipskamp, Amsterdam, 2008

The research in this thesis was performed in the Biomolecular Mass Spectrometry and Proteomics Group, Utrecht University, Utrecht, The Netherlands, the Department of Biochemistry, Center for Biomedical Genetics, Erasmus Medical Center, Rotterdam, The Netherlands and the Department of Biochemistry and Molecular Biology, Center of Experimental Bioinformatics, University of Southern Denmark, Odense, Denmark.

**Quantitative proteomics *on the fly***

Kwantitatieve proteomics *on the fly*

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. J.C. Stoof, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op woensdag 14 januari 2009 des middags te 2.30 uur

door

Joost Willem Gouw

geboren op 19 december 1978 te Amersfoort

Promotor: Prof.dr. A.J.R. Heck

Co-promotor: Dr. J. Krijgsveld

# Table of contents

# CHAPTER 1

# Introduction

# I. *Drosophila melanogaster* as a model organism

Over the past 100 years the fruit fly *Drosophila melanogaster* has become a popular model for understanding eukaryotic biology and this invertebrate holds a primary position in the analysis of biological, disease and pharmaceutical processes in modern biology. More than a century ago researchers like Thomas H. Morgan and William E. Castle introduced fruit flies into the field of biological research [1, 2] and used them to establish most of the major principles of classical genetics leading to the Nobel Prize (1933) awarded for the chromosome theory of heredity [1]. Since then, *Drosophila* conquered the field of biological research as a truly remarkable experimental system because of its rapid generation time, high fertility, genetic accessibility and ease of laboratory handling. These characteristic features allow large-scale analysis of mutant phenotypes in a relatively short period at low cost. For instance, large-scale mutagenesis experiments by the mutagenic agent EMS (ethylmethane sulphonate) in the 1960s led to the discovery of 15,000 new mutant alleles [3, 4] and was followed a decade later by the systematic identification of genes involved in embryonic development [5]. The latter Nobel Prize (1995) winning study was one of the first whole-genome screens for mutants with embryonic lethal phenotypes [6-8] and led to the comprehensive understanding of many specific developmental genes which is, compared with any other embryological process, nowhere as clear as in *Drosophila*. Moreover, other genetic technologies [9-11] developed in the pre-genomic era allowed the manipulation of the genome of the fruit fly more easily than any other multicellular organism [12, 13]. With the completion of the sequencing and annotation of the *Drosophila* genome in 1999 [14] fly research entered the genomic era and although only ~15 000 protein-coding genes were estimated (2.5 times less than in the human genome [15]) it was quickly discovered that many of these have clear homologues in higher organisms including humans. Surprisingly, even complete molecular pathways are shown to be evolutionary conserved and in combination with whole-genome gene expression analysis (i.e. microarrays), *Drosophila* serves as a complex system to study these biological processes. For instance, 75% of human disease genes have related sequences in fly and several genetic disorders are currently investigated in flies [16, 17]. More than a century of fly research resulted in a wealth of publicly available material and information [18] and outstanding resources such as FlyBase [19] provide comprehensive knowledge for fly researchers or sometimes called Drosophilists. Today in the post-genomic era, alternative approaches are being developed for the next phase in biology research, understanding the

function of each individual protein and *Drosophila* continues to prove its values as a model organism in the era of large-scale protein research.

## The life cycle and overall development

In FIGURE 1 is shown the life cycle of *Drosophila*. Fertilization takes place inside the mother quickly followed by the process of laying eggs called ovipositing. Initial embryonic development takes place within the 0.5 mm long egg for 12-15 h after which they hatch into first instar larvae. For the first time they can grow by feeding on microorganisms that decompose fruit as well as on the sugar of the fruit itself and during the four day grow period they molt twice. Although larvae do not have wings, legs and other adult organs, these structures are already present as imaginal discs inside the larvae. They grow during the larval life

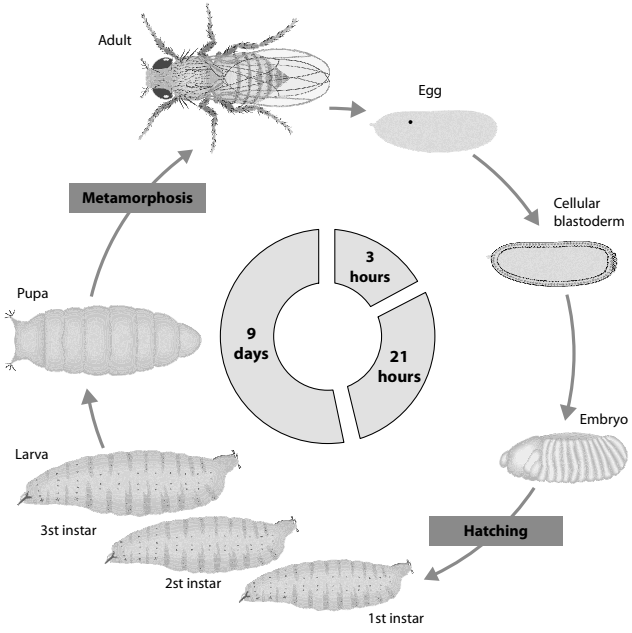


Figure 1. The life cycle of *Drosophila melanogaster*. Shown are the developmental stages as well as critical processes during development. The total development time, from egg to adult, takes more or less 9 days at 25 °C.

and when the third instar larvae form pupa and undergo metamorphosis they develop into legs, wings, antennae, etc. After around four days of transformation adult flies emerge and females become receptive to courting males at about 8-12 h after emergence. As with many ectotherms (cold-blooded species), the developmental period varies with the temperature. At 25 °C the total development time, from egg to adult, takes more or less 8.5 days (FIGURE 1) and can increase to more than 50 days when the temperature falls below 15 °C.

## *Drosophila* embryogenesis

Embryonic development is probably better understood in fly than in any other similar or higher animal. In *Drosophila*, embryonic development can be classified in 17 distinct stages [20] and is initiated when sperm enters the embryo through the micropyle and fuses with the egg nucleus. The first few stages (1-4, 0-130 min) are characterized by series of rapid mitotic divisions (one every 9 minutes) without, however, cytoplasmic cleavage. After around 12 nuclear divisions, 6 000 nuclei share a common cytoplasm, also known as a syncytium, and start to move towards the periphery of the embryo to form the syncytial blastoderm. Shortly after, the cellular blastoderm forms by the invagination of the nuclei by membranes from the surface [21]. It is this blastoderm that is the precursor of all future tissues and cells from this single epithelial layer start to move to their ultimate destination during gastrulation which starts around 3 hours after fertilization. The final destination of these cells depends largely on the body axes, antero-posterior and dorso-ventral of the fly that are already established in the early embryo while still in the syncytial stage. Even well before the oocyte is fertilized, maternal products such as proteins and messenger RNAs are deposited into the egg by the nurse cells that surround the egg [22]. It is these maternal factors that form a spatial pattern and thus define the main body axes and lay down a basic framework of positional information [23, 24]. For instance, the maternal messenger bicoid is localized to the anterior of the oocyte and is, after fertilization, translated into bicoid protein. The protein diffuses towards the posterior of the embryo thereby creating a gradient which in turn activates zygotic genes in the nuclei along the axes [25, 26]. Different zygotic genes are activated depending on the concentration of this so-called morphogen and divide the embryo into a number of regions. Zygotic genes are not transcribed until the gradients are established and therefore development is essentially under maternal control. After two hours of development, transition from maternal to zygotic control (maternal-to-zygotic transition or MZT) is initiated by the start of zygotic transcription and further development is now in the hands of the zygote [27-30].

# II. Mass spectrometry based proteomics

The investigations described in this thesis form part of the work in progress in many places with the aim of improving protein analysis by mass spectrometry. The subject is rapidly gaining coherence and may in fact be called proteomics. A proteome is defined as the set of proteins expressed by a genome, cell, tissue or organism at a given time under defined conditions [31]. After the completion of mapping and sequencing the genome of several organisms (like yeast in 1996 and *Drosophila* in 2000) it was quickly appreciated that a proteome is much more complex than a genome, mostly due to the highly dynamic nature of a proteome. A genome is considered constant, whereas the protein content differs from cell to cell because distinct genes are expressed in distinct cell types. In addition, single genes may possibly encode different proteins due to alternative splicing and post-translational modifications. In the 'post-genomic era' proteins are characterized in terms of expression, localization, post-translational modifications, function and interactions since proteins and not genes determine the complexity of an organism. Mass spectrometry has emerged as an invaluable and indispensible tool in the field of proteomics and is nowadays widely used to study protein interactions [32], protein localization [33], cell signaling [34], functional characterization of protein complexes [35], protein pathways [36], and protein expression analysis [37].

In this thesis several different proteomics strategies have been used to study embryonic development of the fruit fly *Drosophila melanogaster* all of which utilize mass spectrometry. In this part of the thesis several of the most prominent aspects of proteomics-related techniques are discussed starting with the individual components of a mass spectrometer. In general a mass spectrometer consists of three basic parts, a source to ionize molecules, a mass analyzer to separate and/or detect ions based on their mass-to-charge ($m/z$) ratio and a detector to detect and record ions, and all these parts are described and discussed below.

## Ionization techniques

In order to analyze molecules by mass spectrometry, they should be ionized and brought into the gas phase. Two methods that generate ionized gas phase biomolecules are nowadays easily available. Both of these techniques, matrix-assisted laser desorption/ionization (MALDI) and electrospray ionization (ESI), are greatly acknowledged for their ability to ionize biological macromolecules without fragmentation. These techniques paved the

road for comprehensive mass spectrometric analysis of biomolecules and their importance was quickly recognized by the scientific community and ultimately culminated into the award of the Nobel Prize in chemistry to the founders of ESI and MALDI, John Fenn and respectively Koichi Tanaka. Electrospray ionization (ESI) involves the formation of highly charged droplets by nebulization rather than by desorption which takes place in MALDI.

**Matrix-assisted laser desorption/ionization (MALDI)** • Since its introduction in 1987 [38-40], matrix-assisted laser desorption/ionization (MALDI) has been demonstrated to be a powerful technique for the mass analysis of a wide variety of biomolecules. Nowadays MALDI-based mass spectrometry is used to investigate a wide range of compounds and biological systems, including nucleic acids, peptides, proteins and drugs and metabolite systems. In addition, complex synthetic polymers of high molecular weight can also be analyzed by MALDI-MS. The inherent "soft" ionization of MALDI allows promotion of large thermally labile biomolecules into the gas phase without extensive fragmentation. In MALDI, a laser is directed at an analyte imbedded in an excess of a specific matrix material. Crucial to the successful application of this technique is the use of a suitable matrix to promote efficient desorption and ionization of samples [41]. Although a large number of matrix materials have been evaluated, only a few are used routinely and include 2,5-dihydroxybenzoic acid (DHB), sinapinic acid (SA) and α-cyano-4-hydroxycinnamic acid (CHCA) [42]. When the target is irradiated with a laser, the matrix molecules (which are chosen to absorb the wavelength of the laser) are excited and vaporize into the gas phase, taking the analyte molecules with them. During and/or after evaporation, the analyte molecules are ionized. Many ion formation pathways have been suggested, including ionization processes that take place in the gas phase as well as desorption of pre-formed ions [43, 44]. However, the mechanism of ion formation is still poorly understood and no single mechanism can explain the observed phenomena not even in a single MALDI experiment. A better understanding of the ionization process is important for several reasons, such as to improve ion yield and to gain access to new classes of compounds. Generally singly charged ions are formed in MALDI, although higher charged particles have also been reported, resulting in a relatively simple mass spectrum. Other advantages comprise the relative tolerance towards salts and other detergents and, most importantly, small amounts of sample are consumed during analysis. The pulsing laser produces a packet of ions each time it irradiates the target which makes MALDI particularly well-suited to be used in combination with a Time-of-Flight mass analyzer.

**Electrospray ionization (ESI) •** Electrospray ionization (ESI) was described as early as 1968 [45], but gained global interest when it was shown to be capable of ionizing large and fragile biomolecules (up to 130 000 Da) without fragmentation [46, 47]. In addition, it has been shown that ESI excellently interfaces liquid chromatography with mass spectrometry [48]. In ESI, a solution resides in a capillary or nozzle to which a voltage is applied, usually about ± 1-3 kV. The applied voltage (either positive or negative, depending on the analyte) provides the electrical field required to produce charge separation at the surface of the liquid. In the positive mode, negative ions drift away from the surface and positive ions drift towards the meniscus of the liquid and form (after overcoming the surface tension by charge repulsion) a cone, known as the Taylor cone. The emerging liquid is dispersed by Coulomb forces into a fine spray of positively charged droplets. These droplets move through the air towards the counter electrode (i.e. the mass spectrometer) during which solvent evaporates. As a consequence, the charge density on the surface of these droplets increases until it reaches the so-called Rayleigh limit at which the Coulombic repulsion of the surface charge is equal to the surface tension of the solution. Any further evaporation leads to droplet fission, producing charged daughter droplets followed by repeated evaporation and fission, eventually forming gas phase ions. These ions are, in the positive mode, generally formed by the attachment of protons, alkali cations or ammonium ions, whereas in the negative mode hydrogen abstraction or chloride attachment forms negatively charged ions. The exact mechanism of ionization (like in MALDI) is not completely understood and two processes are described. The first mechanism, the charged residue model (CRM) [45, 49], proposes a series of Rayleigh instabilities (droplet fissions including evaporation) producing droplets that consists of a single molecule of solute. That molecule becomes a free gas phase ion by retaining some of the charges in the droplet as the last solvent evaporates. Larger biomolecules appear to be ionized according to this model. The other mechanism, known as the ion evaporation model (IEM) [50, 51], assumes the same sequence as the CRM model, but anticipates that relative small and volatile molecules are able to escape the highly charged droplet as charged species. The differences between both models are subtle but remain subject of debate [52, 53]. In contrast to MALDI, electrospray ionization produces multiply charged species, making it sometimes particularly difficult to interpret a mass spectrum if multiple species (i.e. mixtures) are analyzed. Electrospray ionization is generally conducted under atmospheric pressure, introducing technical difficulties to transfer generated ions into the high vacuum of the mass spectrometer.

## Mass analyzers, hyphenation and ion detectors

The ions that are generated in the source have to be transferred to the mass analyzer so that they can be separated based on their mass-to-charge ($m/z$) ratio. They are usually accelerated by a voltage potential into the high vacuum of the mass spectrometer, eventually reaching the mass analyzer. In a tandem mass spectrometer a collision cell can be placed before or between a mass analyzer and in this configuration structural information can be obtained. Over the years, many different mass analyzers have been developed like the ion trap that has been introduced in the 1950s by Wolfgang Paul. The development of ion trap technology led to the award of the Nobel Prize in physics to Hans G. Dehmelt and Wolfgang Paul for their contributions to ion trap technology. Besides the ion trap also other mass analyzers are discussed below.

**Quadrupole (Q)** • A quadrupole (Q) mass analyzer filters ions based on their (stable) trajectory path through four parallel cylindrical rods, positioned in a radial array. A combination of time-independent (DC) and time-dependent (RF) voltages is applied on opposite, electrically connected rods causing ions to move in a corkscrew kind of motion through the rods. For a given amplitude of the DC and RF voltages, only ions of a given $m/z$ ratio will have a stable trajectory to pass through the quadrupole. In contrast, unstable ions will be neutralized by colliding against the rods. The quadrupole can be used an ion guide by switching off the DC voltage. The scan speed of this device is 5 000 atomic mass units (AMU) per second and has a working mass-to-charge range up to 4 000 Th. The resolution can be as high as 3 000 (FWHM).

**Time-of-Flight (ToF)** • A Time-of-Flight (ToF) mass analyzer [54-56] is probably one of the simplest mass analyzers available. It separates ions in a field-free drift tube based on differences in flight time of ions with different mass-to-charge ratio. A package of ions is injected into the flight tube by a known acceleration potential and since velocity ($v$) is related to mass ($m$), charge ($z$) and acceleration potential ($V_{acc}$):

$$v = \sqrt{\frac{2zeV_{acc}}{m}}$$

ions with different mass-to-charge ratio have different velocities and hence different times of arrival at the detector. An entire mass spectrum can be obtained in a few microseconds for each accelerating pulse making this mass analyzer particularly fast and perfectly suited for MALDI. The initial disadvantage of limited resolution has been overcome by different strategies. Most importantly, the ion mirror or reflectron (FIGURE 2) compensates for the

initial energy spread of the generated ion package by extending the flight time of ions using an electrostatic mirror. Ions with same mass-to-charge ratio but higher velocity penetrate deeper into the Time-of-Flight tube, taking longer to reverse their direction and hence their time of arrival at the detector relative to the slower low-energy ions is delayed. This effect results in improved resolution (20 000 FWHM) in reflector ToF spectra in contrast to linear spectra and a concomitant increase in mass accuracy [57].

**Quadrupole Time-of-Flight (Q-ToF)** • An essential issue in proteomics based mass spectrometry is to sequence low quantities of peptides from complex mixtures. Tandem mass spectrometers (MS/MS) such as triple quadrupole and ion-trap instruments were originally used to deal with these problems, but the lack of sensitivity and/or accuracy opened the door for the design of an instrument with novel geometry, namely the Q-ToF [58]. The configuration, FIGURE 2, can be considered as the replacement of the third quadrupole in a triple quadrupole by a Time-of-Flight mass analyzer (or by the addition of a quadrupole and collision cell to an ESI-ToF). Ions are focused by the first hexapole towards the quadrupole that, depending on the type of analysis (MS or MS/MS) is operated in RF only mode or normal, mass scanning mode, respectively. The collision cell is filled

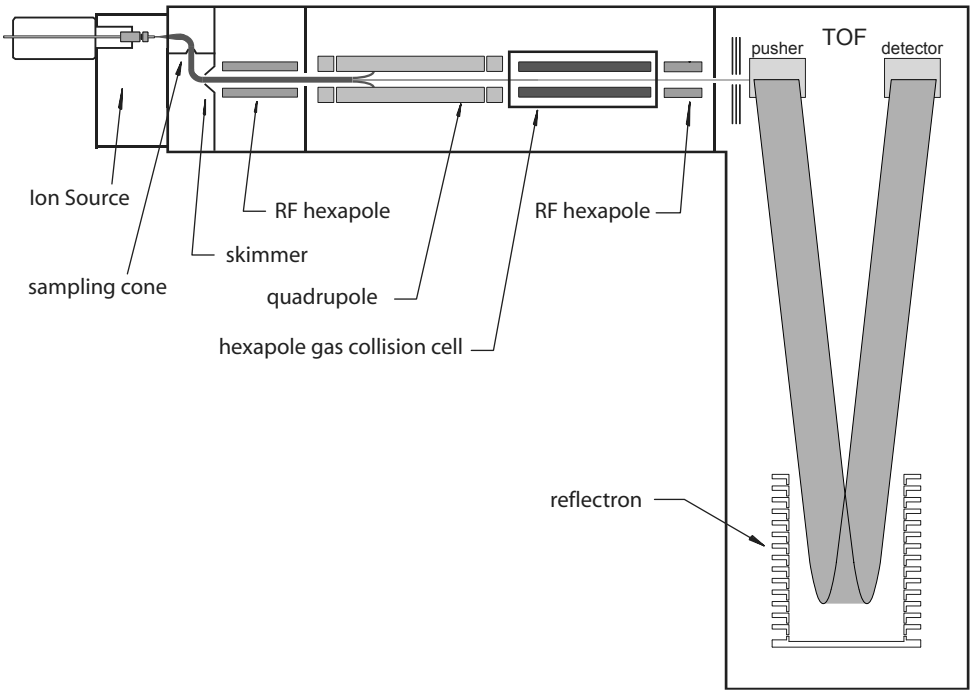


Figure 2. Schematic overview of a quadrupole orthogonal time-of-flight mass spectrometer.

with a collision gas (usually argon or nitrogen) and when operated in MS/MS mode the collision energy is increased to a sufficient level to induce fragmentation. Ions that leave the collision cell are focused by the second hexapole towards the Time-of-Flight mass analyzer and are injected orthogonally, FIGURE 2. Advantages of this configuration include high sensitivity, resolution and accuracy in both MS and MS/MS modes which made this instrument quite popular in the analytical community shortly after its introduction [59].

**Quadrupole (linear) ion trap •** An ion trap mass spectrometer has the ability to capture and store ions by manipulating magnetic or electric fields. This device can be composed of three different traps. The Kingdon [60] and Penning [61] traps are discussed at the respective Orbitrap and FT-ICR sections below. The remaining trap can have a three-dimensional (Paul trap) or linear geometry and is widely used in current mass spectrometers. The 3D-trap is composed of a ring electrode positioned symmetrically between two electrically connected end-cap electrodes [62]. Ions and/or electrons can be gated periodically into the trap by a single small aperture in one of the end-cap electrodes, whereas a small opening in the other end-cap electrode enables ions to be passed towards a detector [63]. Initially, two techniques were used to mass analyze ions with quadrupole ion traps namely mass selective resonance detection [62, 64] and mass selective storage [65]. The latter traps ions of only one $m/z$ value at the time after which they are ejected to a detector. A complete mass spectrum can only be obtained by scanning all possible $m/z$ values which is a relatively long duty cycle (create, store, eject and detect ions). Soon after, it was found that when a light gas (helium, hydrogen, etc.) was used in the trap at relative high pressure, ionic motion was dampened (ions are focused towards the centre of the trap) which increased the resolution, sensitivity and detection limit [66]. In addition, a new mode of operation (mass selective instability mode) was introduced which allowed for the trapping of ions of a complete $m/z$ range. After storing the ions, voltages and frequencies are gradually changed so that ions of consecutive $m/z$ values become unstable and are passed towards the detector [66]. This process, ion excitation, can be applied as radial dipolar and/or axial quadrupolar excitation and can be used to isolate, activate and eject ions. In the case of fragmentation, the kinetic energy of ions is raised after excitation and collisions with buffer gas atoms increases their internal energy which can lead to fragmentation. For a review about 3D ion traps see [63].

By applying stopping potentials to electrodes at the entrance and exit of a quadrupole (Q) mass filter, a linear ion trap is obtained. In this configuration ions are confined radially by RF fields and axially by the stopping potentials. Advantages of a linear trap over

a 3D trap include higher injection efficiencies and higher ion storage capacities resulting in increased sensitivity and improved precision in mass assignment. Two methods exist to eject ions from a linear ion trap in a mass selective manner, radially and axially and form the basis of two recently introduced linear ion trap mass spectrometers [67, 68]. For a review about linear ions traps see [69].

Ion traps have been hyphenated to various kinds of other mass analyzers to improve for instance the duty cycle. To increase the number of ions to be analyzed, a linear trap has been combined with a 3D trap [70]. Ions are accumulated in the linear trap whereas the 3D trap performs other functions such as fragmentation and mass analysis. Also Time-of-Flight mass analyzers have been coupled to both 3D [71] and linear [72] ion traps.

**FT-ICR** • Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS or simply FTMS) is playing an important role in mass spectrometry based proteomics since its supreme resolution, mass accuracy and dynamic range allows for instance for more confident peptide identifications. In FTMS ions are trapped in a Penning trap [61] and under the influence of a strong spatially uniform magnetic field, the ions move in a circular so-called cyclotron motion. The angular frequency ($\omega_c$) of this motion is dependent on the magnetic field ($B_0$) and the mass-to-charge ratio but is completely independent of the velocity of ions:

$$\omega_c = \frac{qB_0}{m}$$

In order to detect (or fragment or eject) ions they are excited by applying for a short period of time an electric field that is oscillating with the same frequency as the cyclotron frequency of ions of a particular *m/z* value. Only ions with the same cyclotron frequency (and therefore the same *m/z* ratio) experience a net continuous outward force and move to a higher orbit (they are ejected when their new radius is bigger than the radius of the trap) where they can be detected. This so-called excitation pulse can also comprise multiple frequencies so that all ions move to the same higher orbit (this excitation is independent of *m/z* ratio as long as the magnitude is constant with the frequency) and can be detected simultaneously. When the new orbit of the ions is close enough to the detector plates, an oscillating differential image can be acquired yielding a time domain signal from all ions passing the detector plates. This transient can be Fourier-transformed into eventually a mass spectrum [73, 74]. Bottom-up proteomics applications using FT-only mass spectrometers were limited since the duty cycle of a tandem FTMS event is relatively long. This problem was solved with the introduction of a hybrid mass spectrometer consisting

of a linear quadrupole ion trap coupled to a FT-ICR mass spectrometer [75], resulting in a highly resourceful instrument that revolutionized global and quantitative proteomics.

**Orbitrap •** Orbital trapping was first described by Kingdon in 1923 where positive ions were trapped along a negatively charged wire that was run axially through a cylindrical anode. These ions described an orbital motion and were lost when they lose sufficient energy by collisions with gas molecules [60]. This principle of electrostatic trapping has been used by Makarov to construct a new type of mass analyzer referred to as the Orbitrap [76]. Ions introduced in the trap describe stable rotational, radial as well as axial oscillating motions along the inner electrode, see FIGURE 3. The latter motion is completely independent of ion energy and position, and the frequency ($\omega_z$) is related to the mass-to-charge ratio of ions:

$$\omega_z = \sqrt{\frac{k}{(m/q)}}$$

Ion oscillation induces a signal voltage detectable by split outer electrodes and an image current is recorded which is amplified and processed exactly in the same way as in FT-ICR [76]. However, the resolving power is proportional to the number of harmonic oscillations of the ions which means that for a fixed acquisition time the resolving power of the FT-ICR declines proportionally to $m/z$ which is more rapidly than the Orbitrap (square root of $m/z$). Other advantages of the Orbitrap over the FT-ICR include the ability to trap more ions and the low cost and complexity results in a powerful new mass spectrometer, especially when coupled to a linear quadrupole ion trap (FIGURE 3) [77, 78]. A similar type, compared to the LTQ-FTMS, hybrid mass spectrometer is obtained with the exception that ions, after accumulation in the linear ion trap, are not directly injected into the Orbitrap but trapped first in a so-called C-Trap, FIG. 3. Orthogonal injection of ions from this curved RF only quadrupole into the Orbitrap ensures fast and uniform extraction for

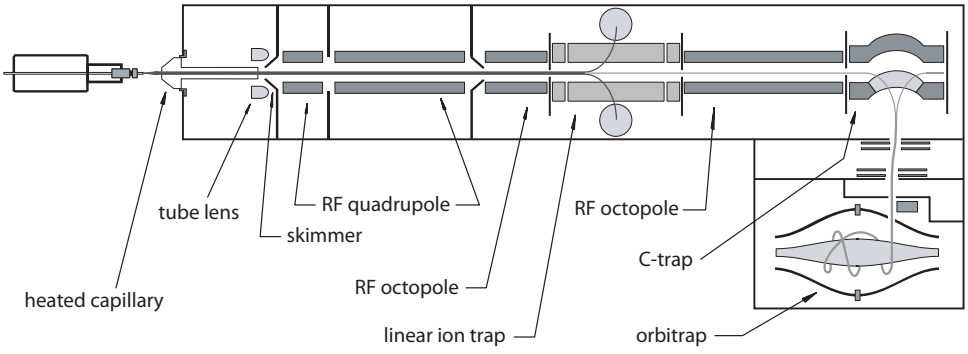


Figure 3. Schematic overview of a LTQ-Orbitrap mass spectrometer.

large ion populations.

## Liquid Chromatography-Mass Spectrometry

Proteomic applications such as described in this thesis form a serious challenge for separation techniques given that protein and/or peptide separation is essential for a successful outcome of a proteomic application. The diversity of protein or peptide samples can be enormous requiring a high-speed, high-resolution and high-sensitivity method to reduce complexity and to prepare such a sample for mass spectrometric analysis. Several decades ago two dimensional gel electrophoresis was the method of choice for protein separation but nowadays this classic technique is simply outperformed by liquid chromatography (LC). One of the main advantages of using LC is that the column effluent can be directly infused into a mass spectrometer using electrospray ionization. The basis that underlies this principle is the compatibility of the mobile phase with electrospray ionization. Hence, revered phase (RP) liquid chromatography is the preferred mode of operation since the mobile phase is usually an acidified (weak acid for the protonation of the peptides) water/acetonitrile mixture which is directly compatible with electrospray ionization.

**Miniaturization and sensitivity •** Although initially flow rates of in the microliter range (1-20 μL/min) [47, 79] were used, the introduction of nanoelectrospray [80] and the miniaturization of analytical columns [81] allowed flow rates of in the nanoliter range (20-200 nL/min) ensuring favorable smaller droplets being emitted from the Taylor cone and high levels of sample entering the mass spectrometer. In addition, reducing the inner diameter (ID) of analytical columns improves sensitivity as well as separation efficiency. Nowadays, fused-silica capillaries with an ID of typically in the range of 25-75 μm are filled with $C_{18}$ reversed phase particles (1-5 μm). These particles can be retained in a capillary by several different methods that include a stainless steel mesh (2 μm pores) [82] or on-column frits made from various sol-gel techniques [83-86] like the polymerization of potassium silicate and formamide [87, 88]. Alternatively, one end of the capillary can be tapered to such a diameter that it functions both as a frit and as an electrospray tip [89, 90]. These fritless columns demonstrate better sensitivity as there is no post-column dead volume that negatively affects chromatographic resolution [84]. An associated problem with the miniaturization of LC systems is the introduction of large amounts (microliters) of sample. Convenient new LC setups have been developed, known as the vented column system, which allows large sample volumes being focused (at a flow rate of 3-5 μL/min) on a trap column which is connected to an analytical column with a zero dead volume connector

thereby minimizing dead volumes. Bound peptides are then eluted (with a split flow at a rate of 20-200 nL/min) from the trap column onto the analytical column by switching a valve and starting the gradient [88, 91]. Sensitivity can even be further improved, down to zeptomole sensitivity, by utilizing columns with an ID of 15 μm and a length of 85 cm operated at a flow rate of ~20 nL/min [82]. Given that LC-MS sensitivity is inversely related to the LC flow rate, reducing the flow rate down to ~10 nL/min improves sensitivity even more and in combination with monolithic stationary phases exceptional results have been obtained [92]. The advantages of these monolithic analytical columns are that they can easily be fabricated, do not need a frit and have no intra-particular void volume [93-95]. Also ultra high performance LC (UPLC) systems have been used to reduce analysis time, improve efficiency and increase peak capacity for columns packed with smaller particles (< 3 μm) [96, 97]. However special pumping equipment capable of creating ultra high pressures (> 1400 bar) is necessary.

**Multidimensional LC** • Despite above delineated improvements in one-dimensional LC-MS, typical proteomics samples are too complex to be comprehensively analyzed in only one dimension. For example, it is estimated that 2 000 – 5 000 proteins are expressed in a typical eukaryotic cell with an unknown number of variants and modifications, producing an order of magnitude more peptides when digested. The complexity of such a (digested) sample is far beyond the resolving capability of one-dimensional liquid chromatography and requires more separation power prior to mass spectrometric detection. To address this, several multidimensional separation methods have been developed which are usually complemented by RP LC-MS in the last dimension due to its compatibility with ESI-MS. Gel electrophoresis (SDS-PAGE) can be used to separate intact proteins in the first dimension after which protein bands or entire gel lanes are cut into smaller pieces and subjected to in-gel proteolytic digestion. Each resulting peptide mixture is then analyzed in the second dimension by RP LC-MS. This approach has been successfully applied to various different types of experiments and large numbers of proteins have been identified [98, 99]. An elegant off-line method to gain separation power is based on the selective modification of peptides between two consecutive reversed phase chromatographic separation steps. Since the modified peptides have different affinity with the stationary phase and hence retention time they can be separated from the non-modified peptides in the second dimension. This procedure, termed combined fractional diagonal chromatography (COFRADIC), is applied to selectively analyze methionine, cysteine and N-terminal peptides [100]. Besides these off-line approaches, also on-line chromatographic and/or elec-

trophoresis procedures have been used like size exclusion chromatography (SEC) [101, 102], affinity chromatography [103, 104] and capillary isoelectric focusing (CIEF) [105, 106]. In addition, capillary zone electrophoresis (CZE) can be used as an alternative second dimension with RP-LC as the primary dimension [107, 108]. Another on-line method to fractionate peptides prior to MS analysis is ion exchange chromatography which can either be strong anion or cation exchange, SAX and SCX respectively. In the so-called multidimensional protein identification technology (MudPIT) a single analytical column is sequentially packed with two different stationary phases, reversed phase complemented by strong cation exchange material. Discrete fractions of peptides can be displaced from the SCX stationary phase to the RP material by injecting salt plugs followed by a regular reversed phase gradient to separate each subset of peptides. This procedure is then repeated for a number of times and this technology has proven to be a powerful tool in proteomics [90, 109]. Although the principle advantage of this on-line technique is the almost complete automation, several disadvantages compared to off-line SCX fractionating have been reported. Most notably, the percentage of acetonitrile that can be added to the buffers to prevent unwanted non-specific hydrophobic interactions with the SCX packing material [110], the number of fractions that can be analyzed with the on-line method and the superiority of a linear (or exponential) gradient compared to a step gradient [111, 112]. For reviews about multidimensional separation see [113, 114].

Taken together, the ability of running samples in an automated fashion and the ease of being hyphenated to mass spectrometers make liquid chromatography superior and a high-throughput separation method for analyzing complex peptide or protein mixtures.

## Methodology to identify peptides and proteins

A fundamental element of proteomic research is the identification of proteins. There are different strategies that lead to the identification of proteins and these can be classified in two complementary approaches, named bottom-up and top-down [115, 116]. Although the top-down method is only introduced several years ago, it is quickly gaining in popularity. Here, intact proteins are introduced and fragmented in the mass spectrometer revealing information about the molecular masses of both the protein and the fragment ions. Moreover, protein modifications (PTMs) can easily be characterized due to the specific nature of this method [115]. The other method and currently the most popular methodology to identify proteins and determining details of their primary sequence and posttranslational modifications (PTMs) is the bottom-up method. This methodology, characterized by very

complex peptide mixtures, has been used in the proteomic experiments described in this thesis and is in more detail discussed below.

**Bottom-up proteomics •** In bottom-up proteomics proteins are, usually before or after a chromatographic separation step, digested into peptides using a proteolytic enzyme. The bottom-up method produces complex peptide mixtures and requires a number of key processes before peptides and subsequently proteins can be identified. Most notably, a mass spectrometer should be able to (*i*) determine the peptide's mass, (*ii*) fragment the peptide and (*iii*) mass analyze the fragments. The lack of time for these events due to the complexity of digests automatically makes separation of the peptides prior to analysis an important feature of bottom-up proteomics. Another important and rapidly evolving aspect of the bottom-up method is processing and interpreting the enormous amounts of data that are typically generated with this approach. Although significant efforts have been made, this aspect is sometimes still underestimated in many places and should be addressed properly before meaningful results can be obtained. Recently, public repositories have been developed with the goal of annotating genomes through the validation of expressed proteins. Proteome information such as peptide and protein identifications can be administered to online databases like PeptideAtlas [117] and PRIDE [118] after which that data can be accessed publicly.

**Proteolytic digestion •** The choice for an enzyme in bottom-up proteomics depends on several factors. Importantly, the method of fragmentation and the position and number of cleavage sites in the protein determine the type of enzyme. An example of such a protease is trypsin which cleaves proteins at the C-terminal side of arginine and lysine residues. This stable and specific enzyme produces peptides that have a favorable size (smaller than 4 kDa) and a basic residue at the C-terminus resulting in usually doubly and triply charged peptides. This not only simplifies the interpretation of collision induced dissociation (CID)mass spectra, but also induces fragmentation in a more predictable manner such that series of y-ions dominate the spectra, see below. Besides trypsin, other enzymes can be used and benefit of combined or sequential digestion for increased sequence coverage. Recently, the properties of the enzyme Lys-N [119] were exploited in a method that allows straightforward ladder sequencing [120]. This enzyme cleaves proteins at the amino side of lysine, resulting in peptides that have a basic residue at the N-terminus and when fragmented using ETD these peptides produce almost complete ladders typically dominated by c-ions. This is particularly interesting for *de novo* sequencing (see below).

**Tandem mass spectrometry and peptide fragmentation •** Primary sequence information can be obtained by fragmenting peptides and mass analyzing the fragments thereby (partly) elucidating the amino acid sequence of the peptides. Several (hybrid) mass spectrometers are capable of performing these experiments and are referred to as tandem mass spectrometers. Examples include ion traps (tandem-in-time) and mass spectrometers that have multiple mass analyzers like a Q-ToF and triple quad (tandem-in-space) [121]. In

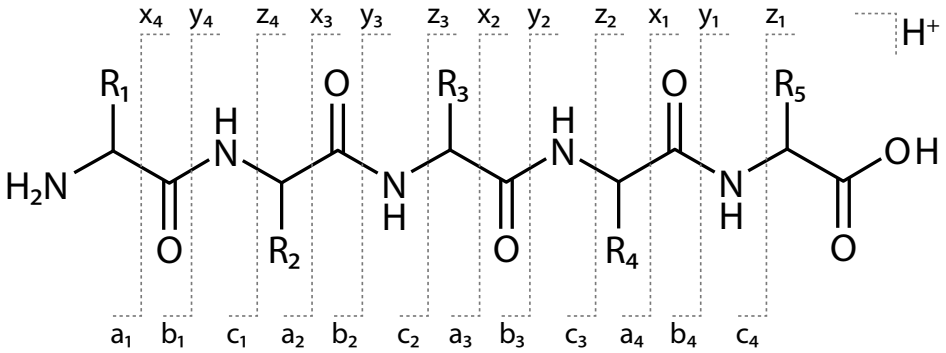


Figure 4. Nomenclature for fragment ions of peptides. N-terminal fragments are of type $a_n$, $b_n$ and $c_n$ whereas C-terminal fragments are of type $x_n$, $y_n$ and $z_n$.

general, the mass spectrometer isolates a precursor ion and induces fragmentation that will result in a second mass spectrum consisting of the product ions. Specific types of ions are obtained depending on the method of fragmentation and these ions are annotated following nomenclature proposed by Roepstorff & Fohlman [122] and subsequently modified by Johnson *et al.* [123, 124], see FIGURE 4.

Two classes of ions can be obtained after fragmentation which depends on the localization of the charge(s). If the charge is retained by the N-terminus, $a_n$, $b_n$ and $c_n$-ions are produced and likewise, $x_n$, $y_n$, $z_n$-ions are obtained when the charge is located to the C-terminus. Furthermore, multiple charged ions can produce favorable complementary ion pairs by separating charge, but also less favorable doubly charged fragments. In addition, the y- and b-type ions can lose water or ammonia, resulting in -18.0106 and -17.0265 Da peaks in the mass spectrum, respectively. For instance, low energy CID of peptides induces cleavage of the peptide amide bonds (NH-CO) thereby producing mainly ions of type $y_n$, $a_n$ and $b_n$, depending on the charge and presence and position of a basic residue in the peptide. Opposed to CID, electron captured dissociation (ECD) [125] and electron transfer dissociation (ETD) [126] produce ions of type $c_n$ and $z_n$ by cleavage of the NH-

CαH amine peptide bond. The major advantage of these elegant fragmentation methods is that labile modifications remain intact during backbone fragmentation which makes them advantageous in characterizing PTMs [127, 128]. However, CID is by far the most used method to fragment peptides since ECD requires an FT-MS which is not readily available in a laboratory and ETD has only recently been introduced to commercially available mass spectrometers, including a hybrid LTQ-Orbitrap MS [129]. CID is available on Q-ToFs, ion traps and triple quadrupole mass spectrometers. For reviews about fragmentation methods and the identification of PTMs see [34, 130, 131].

Regardless of the used dissociation method, a critical issue in bottom-up proteomics remains the fragmentation of peptides originating from complex mixtures. Ideally peptides are separated in such a way that they are presented individually to the mass spectrometer however multiple peptides are usually presented simultaneously. A mass spectrometer has the difficult task to select and fragment as much as possible of these peptides. This can be done in a data-dependent acquisition (DDA) manner where the mass spectrometer first analyzes the intact masses and intensities of all the peptides at that time and, based on user-defined criteria, fragments usually one to ten peptides individually. Alternatively, all peptides can be fragmented concurrently, by alternating between low energy (MS) and elevated energy (MS/MS) which is referred to as $MS^E$ [132] or by isolating and fragmenting precursor windows [133]. The choice for either method depends largely on the available mass spectrometer and database search software.

**Database searching •** The identity of a protein can be unveiled by several different methods, but is simplified if the primary sequence of the protein is known. For instance, the accurately measured mass of peptides (i.e. without fragmenting the peptides) can be compared against theoretical masses of in silico digested proteins from sequence databases. This method, known as peptide mass fingerprinting (PMF) [134], is primarily used for the rapid identification of proteins from simple mixtures, like protein spots from 2D gels. The proteins (usually 1 or 2) in such a spot are in-gel digested and analyzed using a MALDI-ToF mass spectrometer which provides enough accuracy to unambiguously identify the proteins. The peptide masses form the fingerprint of a protein which can be used to reveal the identity of the protein(s) in a database search. A different strategy is necessary when mixtures are more complex and contain proteins beyond the capacity of PMF. Amino acid sequence information obtained by the fragmentation of peptides is, in combination with the accurate peptide mass, used to identify the proteins. Fragmentation information (i.e. intensities and $m/z$ values of all the peaks) from each tandem MS ($MS^2$) spectrum is

submitted to a database search engine which in turn tries to assign a peptide sequence to such a spectrum. Additional sequence information can be gained by performing another round of fragmentation ($MS^3$) where an intense $MS^2$ fragment is selected and fragmented [135]. This gives an in-depth characterization of the peptide at the cost of analysis time and possibly the number of identifications. Nevertheless, large numbers of $MS^n$ spectra are submitted to a database search engine that has the difficult task to identify peptides with small error rates in an automated and timely fashion. Many of such MS/MS database search engines exist [136, 137] and two popular ones, Mascot [138] and Sequest [139], identify peptides using probability-based and autocorrelation algorithms, respectively. These two programs score a peptide fragment ion spectrum against theoretical fragment spectra constructed from candidate peptides from the database. The group of candidate peptides depends on user-defined criteria such as mass tolerances, proteolytic enzyme and possible modifications. Each of these candidates is given a (search engine depend) score and all of them are ranking accordingly. Usually only the top scoring peptide is considered for further processing. Finally, the search engine groups the identified peptides and a list of proteins is reported.

In the absence of protein sequences or even complete sequence databases, experimental MS/MS spectra cannot be compared against theoretical spectra. Instead the amino acid sequence can be determined directly from the ion spectrum. The success of this procedure, called de novo sequencing, is limited due to incomplete backbone fragmentation, overlap of fragment masses and low mass accuracy. However, by using complementary fragmentation techniques and/or favorable enzymes, de novo sequencing is becoming a more powerful and popular tool in proteomics [140].

**Significance of an identification •** The presentation of a list of proteins by the search engine is preceded by two important procedures, separating correct from incorrect peptide identifications and assigning peptide sequences to proteins. Incorrect peptide identifications can be the result of a substantial overlap between the experimental MS/MS ion spectrum and a theoretical fragment spectrum from an unrelated precursor which can result in a random match (by chance) or so-called false-positive. Since there is a substantial overlap between correct and incorrect identifications, it is necessary to find a balance between sensitivity (increasing true-positives) and specificity (minimizing false-positives). The latter can be identified by searching against a so-called 'decoy' or 'reversed' database. The properties of such a database must be similar to the 'target' or 'forward' database in terms of peptide-like sequences that do not occur in the normal database. These decoy databases

can be generated by reversing [141] or shuffling [142] the target sequences or by using a Markov model to create random sequences based on parameters obtained from the target sequences [143]. The number of identifications from the target-decoy search strategy can be used to calculate a false discovery rate (FDR), which is defined as the percentage of accepted identifications that are incorrect and is widely used to describe the statistical significance of the identifications [141, 144-147]. However, the target-decoy strategy is still evolving leading to increased statistical power and increased number of true positives at fixed FDR [148]. Once an optimal peptide FDR is established, peptides have to be clustered into proteins which could potentially be problematic due to splice isoforms for instance. Also homologous proteins, redundant entries in a database and smaller (< 6 amino acids) identified peptides contribute to this issue. As a result peptides may not be unique for a single protein and hence cannot be used to identify a specific protein. This problem is increasing for higher eukaryotic organisms [149, 150]. Again, a statistical measure can be used to determine the plausibility of the protein identification. Also a protein FDR can be determined by using the target-decoy strategy [144, 151], but other means have been reported as well [152, 153].

# III. Quantitative proteomics

The characterization of complex biological processes based only on a qualitative approach (i.e. the mere presence or absence of proteins) may sometimes be fruitful but it ignores the dynamics of protein expression and the fact that biological effects are caused by gradual changes in protein abundance. Therefore, over the past few years extensive efforts have been made to add a quantitative component to comparative proteomics. So far, many different approaches have been published all with the goal of (absolutely) quantifying differences between two or more physiological states of a biological system. These approaches can be classified into different categories (FIGURE 5) based on stable isotope labeling (A-C) and label-free (D) quantitative proteomics [136, 154-156].

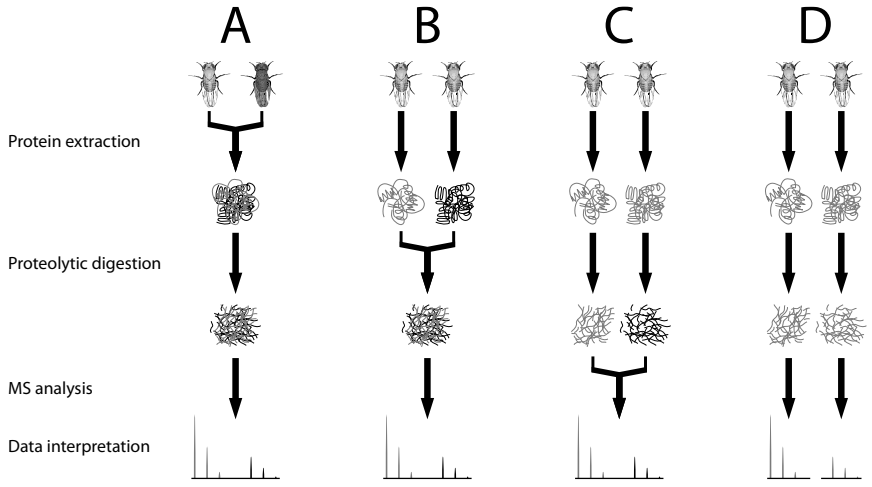


Figure 5. Strategies for quantitative proteomics. Stable isotopes can be incorporated at different stages of the quantitative workflow and are indicated in black. The methods are metabolic labeling (A), protein labeling (B), peptide labeling (C) and label-free (D). Relative expression levels are obtained by mass spectrometry where the signal of the unlabeled peptide is compared to that of the labeled peptide, except for the label-free approach.

## Label-free quantitative proteomics

One of the methods to estimate the abundance of a protein is by looking at the number of sampling events of a peptide from this protein. It has been observed that more MS/MS spectra are assigned to abundant proteins than there are assigned to less abundant proteins [109] and this feature can be used to estimate relative differences in protein abundance [157]. In this approach, termed spectral counting, samples are analyzed and processed

separately and protein lists are compared in terms of sampling events per protein. These sampling events may include multiple identifications of the same peptide, as is usually the case in shotgun experiments like MudPIT, and demonstrates the importance of using the same data acquisition protocol for both samples. In addition, data acquisition features such as dynamic exclusion negatively effects accurate quantification. Although spectral counting has proven to be very accurate at measuring large changes between proteins, accuracy drops dramatically when smaller changes are estimated [158]. The concept of spectral counting has been extended to estimate the absolute protein expression level by computing the protein abundance index (PAI) [159] which was subsequently exponentially modified to emPAI which showed better correlation with known amounts of protein [160]. Another method to estimate absolute protein concentrations is by using absolute protein expression (APEX) [161]. Rather than dividing the number of observed peptides by the number of observable peptides within the mass range of the mass spectrometer (emPAI), APEX corrects each protein's mass spectrometry sampling depth (observed peptide count) by learned probabilities for identifying the peptides.

Whereas protein abundances in spectral counting are estimated from counting and comparing the number of fragment spectra (MS/MS mode), they can also be estimated from measuring and comparing the integrated ion intensity (MS mode) of peptides. Similarly to spectral counting, samples are analyzed and processed separately, but this strategy relies on extracting mass spectrometric peak areas for all peptides that are subsequently integrated for their respective retention times. This approach can be used both for low [162] and high resolution data [163] but the latter is preferred since it increases specificity by reducing the number of similar (hence interfering) but distinct masses. Other important parameters of this strategy include finding the right balance between MS and MS/MS events and optimizing the reproducibility of the chromatographic profile of peptides between different experiments. By using a combination of accurate mass and retention time [164, 165] peptide identifications can be matched to their integrated peak area and hence separate experiments can be conducted with the focus on either qualitative (MS/MS) or quantitative (MS) data acquisition. Alternatively, $LCMS^E$ can be used in combination with a known quantity of a protein spiked into the sample to determine the absolute quantity of proteins [166]. In addition, normalized peak intensities across experiments can be used to correlate profiles of known proteins to identify unknown yet related proteins [167].

## Label-based quantitative proteomics

Accurate quantitation requires a reference point or internal standard that accounts for analyte losses, if any. This internal standard should have physicochemical properties similar to the analyte so that it behaves identical during chromatographic and/or mass spectrometric analysis. Internal standards that have these properties are stable isotope labeled variants of the analyte that increase the mass by several Daltons. The best isotopes for labeling are $^{15}N$, $^{13}C$ and $^{18}O$ in contrast to $^{2}H$ which may produce a significant isotope effect. Deuterated internal standards elute ahead of their unlabeled counterparts which makes data interpretation more challenging. However, this effect seems to be less pronounced when deuterated methyl groups are used such as in the stable isotope dimethyl labeling procedure, see below. Many different labeling strategies exist and internal standards can be introduced in virtually every step of the quantitative workflow (FIGURE 5A-C) but depends mostly on the type of sample. Clearly, the best position to introduce the internal standard is before any sample processing, hence, the first step in the quantitative workflow, FIG. 5A. This can be accomplished via for instance stable isotope labeling by amino acids in cell culture (SILAC) [168] where specific labeled amino acids are in vivo incorporated into growing cells. The success of SILAC depends on the metabolic incorporation of preferably but not necessarily essential amino acids like leucine (either deuterated [168] or $^{13}C_6$-leucine [98, 169]), methionine [170] and lysine [171] to ensure that the supplemented amino acid is the only source in the medium. Non-essential amino acids like tyrosine [172] and arginine [173, 174] can be used but negative side effects such as the arginine-to-proline conversion [175] can occur. Although initially only one labeled amino acid was used, also combinations of differently labeled amino acids can be used to label cells like triple arginine labeling [176, 177] or labeled arginine and lysine amino acids [178, 179]. For a review and protocol for SILAC see [180]. Another way to metabolically incorporate stable isotopes into organisms is by growing these organisms on isotope enriched media, such as $^{15}N$-labeled media. An extensive review about metabolic labeling of model organisms for quantitative proteomics can be found in chapter two of this thesis.

In the second step of the quantitative workflow, proteins can be chemically modified to function as internal standards thereby facilitating relative quantitation (FIG. 5B). Cysteine residues of intact and reduced proteins, for instance, can react with a reagent that besides the reactive group also contains an isotope labeled linker and an affinity tag. In this procedure, termed isotope-coded affinity tags (ICAT) [181], proteins from two samples are treated with either labeled or unlabeled ICAT after which they are combined and digested. The tagged peptides are then purified, analyzed and quantitated. Second genera-

tion ICAT reagents [182, 183] overcome several drawbacks of the first generation ICAT, such as size and isotope effect.

During enzymatic digestion, proteins can be $^{18}O$-labeled by performing the digestion with trypsin in either unlabeled ($H_2{}^{16}O$) or labeled ($H_2{}^{18}O$) water (FIG. 5C) [184, 185]. Here the dual function of trypsin (proteolysis and oxygen exchange) is exploited to incorporate two heavy oxygen atoms into the C-terminus of peptides. The mass difference of 4 Da per peptide is sufficient to mass spectrometric distinguish labeled from unlabeled peptides. However, optimal experimental conditions are necessary to avoid inefficient labeling and/or back exchange of the labeled oxygen [186]. For a review about $^{18}O$-labeling see [187].

A procedure related to ICAT, referred to as isotope tags for relative and absolute quantification (iTRAQ) [188], is based on the reaction of one of four isobaric tags with N-termini and lysine residues of peptides (FIG. 5C). The four isobaric tags facilitate the relative comparison of four different samples per analysis and since these tags are indistinguishable in MS mode, quantitation is accomplished in MS/MS mode. Due to unique stoichiometry of reporter and balance groups in the tags fragmentation reveals their identity by different reporter ions (*m/z* 114-117). The concept of 4-plex labeling was extended to an 8-plex procedure to simultaneously quantify eight biological samples [189]. Many more different chemical tagging concepts on the peptide level have been described and these generally target functional groups in peptides, such as the N-terminus and lysine groups but also carboxyl groups on C-termini or glutamic and aspartic acid residues albeit to a lesser extend [190]. For instance, stable isotope dimethyl labeling is based on the reductive amination of primary amine moieties by using formaldehyde and cyanoborohydride. This reaction introduces two methyl groups on the e-amine group of lysine residues and on the N-terminus. The mass increase depends on the type of formaldehyde that is used in the reaction and can be 28 Da or 32 Da per derivatized site with native or deuterated formaldehyde, respectively [191].

## Processing quantitative mass spectrometric data

A critical issue in the successful application of a quantitative proteomics experiment depends on the possibility to deduce peptide and/or protein ratios from mass spectrometric data. Many software solutions capable of quantitating proteomics data exist but currently the type of labeling and data format dictate the kind of software that can be used [192]. Raw data from mass spectrometer vendors is not interchangeable and therefore analysis

packages have to be tailored to specific experimental setups resulting in a wide variety of (commercial) software packages. Due to the different labeling strategies and data formats, no single software package exists to perform universal quantitation. For instance, quantitation based on stable isotope labeling requires the software to find and integrate peak pairs, except in iTRAQ labeling where the reporter ions always have the same mass. The mass difference between these peaks could depend on (*i*) the number of modified amino acids in SILAC, (*ii*) the total number of nitrogen atoms per peptide in $^{15}N$ labeling, (*iii*) the N-terminus and number of lysine amino acids in dimethyl labeling, (*iv*) the C-terminus in $^{18}O$ labeling, etc. In addition, the software must be compatible of reading the data format of the vendor. The latter issue is subject of debate in many places and the need for common data formats is increasing [193]. Thus far, some XML-based standards, like mzData and mzXML, have been introduced [194-196].

# IV. Outline of this thesis

Development of multicellular organisms is a fascinating process that has been the subject of research for more than a century. Complex structures such as eyes, arms and brain arise from a single cell, the fertilized egg, and require millions of cell divisions and highly specialized mechanisms before these structures are formed. Although (embryonic) development has been studied for a long time, the molecular mechanisms that underlie the principles of development are only beginning to be unraveled. Over the years, the fruit fly *Drosophila melanogaster* has become a popular model organism and provides a wealth of information about fundamental biological processes that are shown to be conserved during evolution. Nowadays *Drosophila* is one of the most effective tools to not only study development but also to analyze the function of for instance human disease genes. Development in *Drosophila* has been studied extensively and in the post-genomic era primarily genomic techniques were used to uncover large-scale gene expression and regulation. However, mRNA abundances do not necessarily correlate with protein expression levels making it necessary to develop alternative approaches for the ultimate phase in genomic research, understanding the function of each individual protein. The maturation of proteomics-based mass spectrometric techniques allows, together with the sequencing and annotation of the *Drosophila* genome, to capture the identity of many proteins in a single experiment. In the research described in this thesis, we extend such a qualitative approach by adding a quantitative component to comparative proteomics to investigate early embryonic development in *Drosophila melanogaster*. In addition, we developed tools that facilitate the analysis of data from large-scale quantitative proteomic studies.

Metabolic labeling of model organisms is increasingly used in comparative proteomics to accurately study many different aspects of biology. In chapter 2 is given an overview of model organisms that have been metabolically labeled for quantitative proteomic applications. Stable isotopes continue moving up the evolutionary ladder witnessed by the recent metabolic labeling of mice and rats.

The maternal-to-zygote transition in *Drosophila* is described in chapter 3. We demonstrate that by combining stable-isotope labeling of fruit flies in vivo with high accuracy quantitative mass spectrometry changes in protein expression levels of more than 2 200 proteins can be captured in a single experiment. Extensive analysis of the dynamic profiles of known and novel proteins provided distinction between maternal proteins and proteins as a result of embryonic translation. We show that novel control mechanisms in genome

activation can be determined by directly comparing transcription profiles and protein dynamics. The method described in this chapter proved highly accurate and reproducible, and has revealed detailed information giving insight that cannot be obtained from genomic approaches.

Chapter 4 provides insights into the process of protein identification and quantitation by investigating the influence of incomplete metabolic heavy nitrogen labeling on the number of identifications as well as on the error in protein quantitation. Two datasets with different suboptimal enrichments in heavy nitrogen were systematically explored and are used to illustrate that specifically larger labeled peptides are underrepresented among all identifications and that peptide ratios lack accuracy and precision due to incomplete enrichment. We propose correction methods that can be used to improve both qualitative and quantitative data obtained by heavy nitrogen labeling with enrichment less than 100%. Although we have applied this to $^{15}$N-labeled proteins, this could as well be used for other types of labeling such as $^{13}$C-labeling.

The development of MSQuant, a tool for interpreting quantitative proteomics data, is described in chapter 5. We show that MSQuant is a multifunctional algorithm that is capable of reading raw data from different mass spectrometer vendors as well as visualization and validation of peptide identification results directly on the raw mass spectrometric data. We demonstrate that MSQuant increases the confidence of peptide identifications by using an incorporated scoring scheme for $MS^3$ spectra and by computing a posttranslational modification score. In addition, it is illustrated that MSQuant supports many different quantitation modes including the heavy nitrogen labeling approach and allows quantitation in an automated fashion.

In chapter 6 we explore the possibility of using a label-free approach in comparative proteomics. Although stable isotope (metabolic) labeling has proven to be an invaluable strategy to determine protein expression levels, we demonstrate that label-free quantitation is a good alternative. By directly comparing the protein expression levels obtained by label-free quantitation based on spectral peak areas of peptides and metabolic labeling based on stable isotope $^{15}$N-labeling we show that there is a weak positive correlation between differential protein abundance levels. The preliminary results from the investigations described here are part of a larger experiment, and we anticipate that the correlation between these methods will improve with the analysis of more samples.

# References


[1] Castle W.E., *Inbreeding, Cross-Breeding and Sterility in Drosophila*. Science, **1906**, *23*, 153.

[2] Bridges C.B., *Direct Proof through Non-Disjunction That the Sex-Linked Genes of Drosophila Are Borne by the X-Chromosome*. Science, **1914**, *40*, 107-109.

[3] Alderson T., *Chemically induced delayed germinal mutation in Drosophila*. Nature, **1965**, *207*, 164-167.

[4] Lewis E.B. and Bacher F., *Methods of feeding ethylmethane sulfonate (EMS) to Drosophila males*. Dros. Inf. Serv., **1968**, 193.

[5] Nusslein-Volhard C. and Wieschaus E., *Mutations affecting segment number and polarity in Drosophila*. Nature, **1980**, *287*, 795-801.

[6] Jürgens G., Wieschaus E., Nüsslein-Volhard C. and Kluding H., *Mutations affecting the pattern of the larval cuticle in Drosophila melanogaster*. Development Genes and Evolution, **1984**, *193*, 283-295.

[7] Nüsslein-Volhard C., Wieschaus E. and Kluding H., *Mutations affecting the pattern of the larval cuticle in Drosophila melanogaster*. Development Genes and Evolution, **1984**, *193*, 267-282.

[8] Wieschaus E., Nüsslein-Volhard C. and Jürgens G., *Mutations affecting the pattern of the larval cuticle in Drosophila melanogaster*. Development Genes and Evolution, **1984**, *193*, 296-307.

[9] Fischer J.A., Giniger E., Maniatis T. and Ptashne M., *GAL4 activates transcription in Drosophila*. Nature, **1988**, *332*, 853-856.

[10] Rubin G.M. and Spradling A.C., *Genetic transformation of Drosophila with transposable element vectors*. Science, **1982**, *218*, 348-353.

[11] Golic K.G. and Golic M.M., *Engineering the Drosophila genome: chromosome rearrangements by design*. Genetics, **1996**, *144*, 1693-1711.

[12] Venken K.J. and Bellen H.J., *Emerging technologies for gene manipulation in Drosophila melanogaster*. Nat Rev Genet, **2005**, *6*, 167-178.

[13] St Johnston D., *The art and design of genetic screens: Drosophila melanogaster*. Nat Rev Genet, **2002**, *3*, 176-188.

[14] Adams M.D., Celniker S.E., Holt R.A., Evans C.A., Gocayne J.D., Amanatides P.G., Scherer S.E., Li P.W., Hoskins R.A., Galle R.F., et al., *The genome sequence of Drosophila melanogaster*. Science, **2000**, *287*, 2185-2195.

[15] Venter J.C., Adams M.D., Myers E.W., Li P.W., Mural R.J., Sutton G.G., Smith H.O., Yandell M., Evans C.A., Holt R.A., et al., *The sequence of the human genome*. Science, **2001**, *291*, 1304-1351.

[16] Bier E., *Drosophila, the golden bug, emerges as a tool for human genetics*. Nat Rev Genet, **2005**, *6*, 9-23.

[17] Schneider D., *Using Drosophila as a model insect*. Nat Rev Genet, **2000**, *1*, 218-226.

[18] Matthews K.A., Kaufman T.C. and Gelbart W.M., *Research resources for Drosophila: the expanding universe*. Nat Rev Genet, **2005**, *6*, 179-193.

[19] Flybase Consortium, *The FlyBase database of the Drosophila genome projects and community literature*. Nucleic Acids Res, **2002**, *30*, 106-108.

[20] Campos-Ortega J.A. and Hartenstein V. *The embryonic development of Drosophila melanogaster.*; Springer-Verlag: Berlin, 1985.

[21] Sokac A.M. and Wieschaus E., *Local actin-dependent endocytosis is zygotically controlled to initiate*

*Drosophila cellularization*. Dev Cell, **2008**, *14*, 775-786.

[22] Mische S., Li M., Serr M. and Hays T.S., *Direct observation of regulated ribonucleoprotein transport across the nurse cell/oocyte boundary*. Mol Biol Cell, **2007**, *18*, 2254-2263.

[23] Semotok J.L. and Lipshitz H.D., *Regulation and function of maternal mRNA destabilization during early Drosophila development*. Differentiation, **2007**, *75*, 482-506.

[24] Steinhauer J. and Kalderon D., *Microtubule polarity and axis formation in the Drosophila oocyte*. Dev Dyn, **2006**, *235*, 1455-1468.

[25] Bergmann S., Sandler O., Sberro H., Shnider S., Schejter E., Shilo B.Z. and Barkai N., *Pre-steady-state decoding of the Bicoid morphogen gradient*. PLoS Biol, **2007**, *5*, e46.

[26] Ephrussi A. and Johnston D.S., *Seeing Is Believing: The Bicoid Morphogen Gradient Matures*. Cell, **2004**, *116*, 143-152.

[27] Bashirullah A., Halsell S.R., Cooperstock R.L., Kloc M., Karaiskakis A., Fisher W.W., Fu W., Hamilton J.K., Etkin L.D. and Lipshitz H.D., *Joint action of two RNA degradation pathways controls the timing of maternal transcript elimination at the midblastula transition in Drosophila melanogaster*. Embo J, **1999**, *18*, 2610-2620.

[28] De Renzis S., Elemento O., Tavazoie S. and Wieschaus E.F., *Unmasking activation of the zygotic genome using chromosomal deletions in the Drosophila embryo*. PLoS Biol, **2007**, *5*, e117.

[29] Stitzel M.L. and Seydoux G., *Regulation of the oocyte-to-zygote transition*. Science, **2007**, *316*, 407-408.

[30] Tadros W., Westwood J.T. and Lipshitz H.D., *The mother-to-child transition*. Dev Cell, **2007**, *12*, 847-849.

[31] Wilkins M.R., Pasquali C., Appel R.D., Ou K., Golaz O., Sanchez J.C., Yan J.X., Gooley A.A., Hughes G., Humphery-Smith I., et al., *From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis*. Biotechnology (N Y), **1996**, *14*, 61-65.

[32] Bader S., Kuhner S. and Gavin A.C., *Interaction networks for systems biology*. FEBS Lett, **2008**, *582*, 1220-1224.

[33] Cornett D.S., Reyzer M.L., Chaurand P. and Caprioli R.M., *MALDI imaging mass spectrometry: molecular snapshots of biochemical systems*. Nat Methods, **2007**, *4*, 828-833.

[34] Jensen O.N., *Interpreting the protein language using proteomics*. Nat Rev Mol Cell Biol, **2006**, *7*, 391-403.

[35] Gavin A.C., Bosche M., Krause R., Grandi P., Marzioch M., Bauer A., Schultz J., Rick J.M., Michon A.M., Cruciat C.M., et al., *Functional organization of the yeast proteome by systematic analysis of protein complexes*. Nature, **2002**, *415*, 141-147.

[36] Cravatt B.F., Simon G.M. and Yates J.R., 3rd, *The biological impact of mass-spectrometry-based proteomics*. Nature, **2007**, *450*, 991-1000.

[37] Cox J. and Mann M., *Is proteomics the new genomics?* Cell, **2007**, *130*, 395-398.

[38] Karas M., Bachmann D., Bahr U. and Hillenkamp F., *Matrix-assisted ultraviolet laser desorption of non-volatile compounds*. Int. J. Mass Spectrom. Ion Processes, **1987**, *78*, 53-68.

[39] Karas M. and Hillenkamp F., *Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons*. Anal Chem, **1988**, *60*, 2299-2301.

[40] Tanaka K., Waki H., Ido Y., Akita S., Yoshida Y. and Yoshida T., *Protein and polymer analyses of up to m/z 100,000 by laser ionization time-of-flight mass spectrometry*. Rapid Commun. Mass Spectrom.,

**1988**, *2*, 151-153.

[41] Cohen S.L. and Chait B.T., *Influence of matrix solution conditions on the MALDI-MS analysis of peptides and proteins*. Anal Chem, **1996**, *68*, 31-37.

[42] Pan C., Xu S., Zhou H., Fu Y., Ye M. and Zou H., *Recent developments in methods and technology for analysis of biological samples by MALDI-TOF-MS*. Anal Bioanal Chem, **2007**, *387*, 193-204.

[43] Chang W.C., Huang L.C., Wang Y.S., Peng W.P., Chang H.C., Hsu N.Y., Yang W.B. and Chen C.H., *Matrix-assisted laser desorption/ionization (MALDI) mechanism revisited*. Anal Chim Acta, **2007**, *582*, 1-9.

[44] Zenobi R. and Knochenmuss R., *Ion formation in MALDI mass spectrometry*. Mass Spectrom. Rev., **1998**, *17*, 337-366.

[45] Dole M., Mack L.L., Hines R.L., Mobley R.C., Ferguson L.D. and Alice M.B., *Molecular Beams of Macroions*. J. Chem. Phys., **1968**, *49*, 2240-2249.

[46] Yamashita M. and Fenn J.B., *Electrospray ion source. Another variation on the free-jet theme*. J. Phys. Chem., **1984**, *88*, 4451-4459.

[47] Fenn J.B., Mann M., Meng C.K., Wong S.F. and Whitehouse C.M., *Electrospray ionization for mass spectrometry of large biomolecules*. Science, **1989**, *246*, 64-71.

[48] Whitehouse C.M., Dreyer R.N., Yamashita M. and Fenn J.B., *Electrospray interface for liquid chromatographs and mass spectrometers*. Anal Chem, **1985**, *57*, 675-679.

[49] Fernandez de la Mora J., *Electrospray ionization of large multiply charged species proceeds via Dole's charged residue mechanism*. Analytica Chimica Acta, **2000**, *406*, 93-104.

[50] Iribarne J.V. and Thomson B.A., *On the evaporation of small ions from charged droplets*. The Journal of Chemical Physics, **1976**, *64*, 2287-2294.

[51] Thomson B.A. and Iribarne J.V., *Field induced ion evaporation from liquid surfaces at atmospheric pressure*. The Journal of Chemical Physics, **1979**, *71*, 4451-4463.

[52] Kebarle P. and Peschke M., *On the mechanisms by which the charged droplets produced by electrospray lead to gas phase ions*. Analytica Chimica Acta, **2000**, *406*, 11-35.

[53] Cech N.B. and Enke C.G., *Practical implications of some recent studies in electrospray ionization fundamentals*. Mass Spectrom Rev, **2001**, *20*, 362-387.

[54] Cameron A.E. and Eggers J.D.F., *An Ion "Velocitron"*. Rev. Sci. Instrum., **1948**, *19*, 605-607.

[55] Stephens W.E., *A pulsed mass spectrometer with time dispersion*. Phys. Rev., **1946**, *69*, 691.

[56] Wiley W.C. and McLaren I.H., *Time-of-Flight Mass Spectrometer with Improved Resolution*. Rev. Sci. Instrum., **1955**, *26*, 1150-1157.

[57] Grix R., Kutscher R., Li G., Grüner U., Wollnik H. and Matsuda H., *A time-of-flight mass analyzer with high resolving power*. Rapid Commun. Mass Spectrom., **1988**, *2*, 83-85.

[58] Morris H.R., Paxton T., Dell A., Langhorne J., Berg M., Bordoli R.S., Hoyes J. and Bateman R.H., *High Sensitivity Collisionally-activated Decomposition Tandem Mass Spectrometry on a Novel Quadrupole/Orthogonal-acceleration Time-of-flight Mass Spectrometer*. Rapid Commun. Mass Spectrom., **1996**, *10*, 889-896.

[59] Chernushevich I.V., Loboda A.V. and Thomson B.A., *An introduction to quadrupole-time-of-flight mass spectrometry*. J. Mass Spectrom., **2001**, *36*, 849-865.

[60] Kingdon K.H., *A Method for the Neutralization of Electron Space Charge by Positive Ionization at Very Low Gas Pressures*. Physical Review, **1923**, *21*, 408.

[61] Dehmelt H.G. and Walls F.L., *"Bolometric" Technique for the rf Spectroscopy of Stored Ions*. Phys. Rev. Lett., **1968**, *21*, 127-131.

[62] Paul W. and Steinwedel H., *Ein neues Massenspektrometer ohne Magnetfeld*. Z. f. Naturforschung A, **1953**, *8*, 448-450.

[63] March R.E., *An Introduction to Quadrupole Ion Trap Mass Spectrometry*. J. Mass Spectrom., **1997**, *32*, 351-369.

[64] Fischer E., *Die dreidimensionale Stabilisierung von Ladungsträgern in einem Vierpolfeld*. Z. Physik, **1959**, *156*, 1-26.

[65] Dawson P.H., Hedman J.W. and Whetten N.R., *A Simple Mass Spectrometer*. Review of Scientific Instruments, **1969**, *40*, 1444-1450.

[66] Stafford G.C., Kelley P.E., Syka J.E.P., Reynolds W.E. and Todd J.F.J., *Recent improvements in and analytical applications of advanced ion trap technology*. International Journal of Mass Spectrometry and Ion Processes, **1984**, *60*, 85-98.

[67] Hager J.W., *A new linear ion trap mass spectrometer*. Rapid Communications in Mass Spectrometry, **2002**, *16*, 512-526.

[68] Schwartz J.C., Senko M.W. and Syka J.E., *A two-dimensional quadrupole ion trap mass spectrometer*. J Am Soc Mass Spectrom, **2002**, *13*, 659-669.

[69] Douglas D.J., Frank A.J. and Mao D., *Linear ion traps in mass spectrometry*. Mass Spectrom Rev, **2005**, *24*, 1-29.

[70] Voyksner R.D. and Lee H., *Investigating the use of an octupole ion guide for ion storage and high-pass mass filtering to improve the quantitative performance of electrospray ion trap mass spectrometry*. Rapid Communications in Mass Spectrometry, **1999**, *13*, 1427-1437.

[71] Michael S.M., Chien M. and Lubman D.M., *An ion trap storage/time-of-flight mass spectrometer*. Review of Scientific Instruments, **1992**, *63*, 4277-4284.

[72] Campbell J.M., Collings B.A. and Douglas D.J., *A new linear ion trap time-of-flight system with tandem mass spectrometry capabilities*. Rapid Communications in Mass Spectrometry, **1998**, *12*, 1463-1474.

[73] Heeren R.M.A., Kleinnijenhuis A.J., McDonnell L.A. and Mize T.H., *A mini-review of mass spectrometry using high-performance FTICR-MS methods*. Analytical and Bioanalytical Chemistry, **2004**, *378*, 1048-1058.

[74] Marshall A.G., Hendrickson C.L. and Jackson G.S., *Fourier transform ion cyclotron resonance mass spectrometry: A primer*. Mass Spectrometry Reviews, **1998**, *17*, 1-35.

[75] Syka J.E.P., Marto J.A., Bai D.L., Horning S., Senko M.W., Schwartz J.C., Ueberheide B., Garcia B., Busby S., Muratore T., et al., *Novel Linear Quadrupole Ion Trap/FT Mass Spectrometer: Performance Characterization and Use in the Comparative Analysis of Histone H3 Post-translational Modifications*. J. Proteome Res., **2004**, *3*, 621-626.

[76] Makarov A., *Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis*. Anal. Chem., **2000**, *72*, 1156-1162.

[77] Hu Q., Noll R.J., Li H., Makarov A., Hardman M. and Cooks R.G., *The Orbitrap: a new mass spectrometer*. Journal of Mass Spectrometry, **2005**, *40*, 430-443.

[78] Makarov A., Denisov E., Kholomeev A., Balschun W., Lange O., Strupat K. and Horning S., *Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer*. Anal Chem, **2006**, *78*, 2113-2120.

[79] Smith R.D., Loo J.A., Edmonds C.G., Barinaga C.J. and Udseth H.R., *New developments in biochemical mass spectrometry: electrospray ionization*. Anal. Chem., **1990**, *62*, 882-899.

[80] Wilm M. and Mann M., *Analytical properties of the nanoelectrospray ion source*. Anal Chem, **1996**, *68*, 1-8.

[81] Davis M.T., Stahl D.C., Hefta S.A. and Lee T.D., *A Microscale Electrospray Interface for Online, Capillary Liquid Chromatography/Tandem Mass Spectrometry of Complex Peptide Mixtures*. Anal. Chem., **1995**, *67*, 4549-4556.

[82] Shen Y., Tolic N., Masselon C., Pasa-Tolic L., Camp D.G., Hixson K.K., Zhao R., Anderson G.A. and Smith R.D., *Ultrasensitive Proteomics Using High-Efficiency On-Line Micro-SPE-NanoLC-NanoESI MS and MS/MS*. Anal. Chem., **2004**, *76*, 144-154.

[83] Cortes H.J., Pfeiffer C.D., Richter B.E. and Stevens T.S., *Porous ceramic bed supports for fused silica packed capillary columns used in liquid chromatography*. Journal of High Resolution Chromatography, **1987**, *10*, 446-448.

[84] Maiolica A., Borsotti D. and Rappsilber J., *Self-made frits for nanoscale columns in proteomics*. Proteomics, **2005**, *5*, 3847-3850.

[85] Schmid M., Bäuml F., Köhne A.P. and Welsch T., *Preparation of On-Column Frits in Packed Fused Silica Capillaries by Sol-Gel Technology*. Journal of High Resolution Chromatography, **1999**, *22*, 438-442.

[86] Wang L.C., Okitsu C.Y., Kochounian H., Rodriguez A., Hsieh C.L. and Zandi E., *A simple and inexpensive on-column frit fabrication method for fused-silica capillaries for increased capacity and versatility in LC-MS/MS applications*. Proteomics, **2008**, *8*, 1758-1761.

[87] Behnke B., Grom E. and Bayer E., *Evaluation of the parameters determining the performance of electrochromatography in packed capillary columns*. Journal of Chromatography A, **1995**, *716*, 207-213.

[88] Meiring H.D., van der Heeft E., ten Hove G.J. and de Jong A.P.J.M., *Nanoscale LC-MS$^{(n)}$: technical design and applications to peptide and protein analysis*. J. Sep. Sci., **2002**, *25*, 557-568.

[89] Gatlin C.L., Kleemann G.R., Hays L.G., Link A.J. and Yates J.R., *Protein Identification at the Low Femtomole Level from Silver-Stained Gels Using a New Fritless Electrospray Interface for Liquid Chromatography-Microspray and Nanospray Mass Spectrometry*. Analytical Biochemistry, **1998**, *263*, 93-101.

[90] Link A.J., Eng J., Schieltz D.M., Carmack E., Mize G.J., Morris D.R., Garvik B.M. and Yates J.R., 3rd, *Direct analysis of protein complexes using mass spectrometry*. Nat Biotechnol, **1999**, *17*, 676-682.

[91] Licklider L.J., Thoreen C.C., Peng J. and Gygi S.P., *Automation of nanoscale microcapillary liquid chromatography-tandem mass spectrometry with a vented column*. Anal Chem, **2002**, *74*, 3076-3083.

[92] Luo Q., Page J.S., Tang K. and Smith R.D., *MicroSPE-nanoLC-ESI-MS/MS Using 10-μm-i.d. Silica-Based Monolithic Columns for Proteomics*. Anal. Chem., **2007**, *79*, 540-545.

[93] Gusev I., Huang X. and Horváth C., *Capillary columns with in situ formed porous monolithic packing for micro high-performance liquid chromatography and capillary electrochromatography*. Journal of Chromatography A, **1999**, *855*, 273-290.

[94] Nunez O., Nakanishi K. and Tanaka N., *Preparation of monolithic silica columns for high-performance liquid chromatography*. J Chromatogr A, **2008**, *1191*, 231-252.

[95] Unger K.K., Skudas R. and Schulte M.M., *Particle packed columns and monolithic columns in high-performance liquid chromatography-comparison and critical appraisal*. J Chromatogr A, **2008**, *1184*, 393-415.

[96] MacNair J.E., Lewis K.C. and Jorgenson J.W., *Ultrahigh-Pressure Reversed-Phase Liquid Chromatogra-*

*phy in Packed Capillary Columns*. Anal. Chem., **1997**, *69*, 983-989.

[97] Shen Y., Zhang R., Moore R.J., Kim J., Metz T.O., Hixson K.K., Zhao R., Livesay E.A., Udseth H.R. and Smith R.D., *Automated 20 kpsi RPLC-MS and MS/MS with Chromatographic Peak Capacities of 1000-1500 and Capabilities in Proteomics and Metabolomics*. Anal. Chem., **2005**, *77*, 3090-3100.

[98] Romijn E.P., Christis C., Wieffer M., Gouw J.W., Fullaondo A., van der Sluijs P., Braakman I. and Heck A.J., *Expression clustering reveals detailed co-expression patterns of functionally related proteins during B cell differentiation: a proteomic study using a combination of one-dimensional gel electrophoresis, LC-MS/MS, and stable isotope labeling by amino acids in cell culture (SILAC)*. Mol Cell Proteomics, **2005**, *4*, 1297-1310.

[99] Lasonder E., Ishihama Y., Andersen J.S., Vermunt A.M.W., Pain A., Sauerwein R.W., Eling W.M.C., Hall N., Waters A.P., Stunnenberg H.G., et al., *Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry*. Nature, **2002**, *419*, 537-542.

[100] Gevaert K., Van Damme P., Ghesquière B., Impens F., Martens L., Helsens K. and Vandekerckhove J., *A la carte proteomics with an emphasis on gel-free techniques*. Proteomics, **2007**, *7*, 2698-2718.

[101] Moore A.W. and Jorgenson J.W., *Comprehensive three-dimensional separation of peptides using size exclusion chromatography/reversed phase liquid chromatography/optically gated capillary zone electrophoresis*. Anal. Chem., **1995**, *67*, 3456-3463.

[102] Opiteck G.J., Jorgenson J.W. and Anderegg R.J., *Two-Dimensional SEC/RPLC Coupled to Mass Spectrometry for the Analysis of Peptides*. Anal. Chem., **1997**, *69*, 2283-2291.

[103] Geng M., Ji J. and Regnier F.E., *Signature-peptide approach to detecting proteins in complex mixtures*. Journal of Chromatography A, **2000**, *870*, 295-313.

[104] Pinkse M.W., Mohammed S., Gouw J.W., van Breukelen B., Vos H.R. and Heck A.J., *Highly robust, automated, and sensitive online TiO2-based phosphoproteomics applied to study endogenous phosphorylation in Drosophila melanogaster*. J Proteome Res, **2008**, *7*, 687-697.

[105] Chen J., Balgley B.M., DeVoe D.L. and Lee C.S., *Capillary Isoelectric Focusing-Based Multidimensional Concentration/Separation Platform for Proteome Analysis*. Anal. Chem., **2003**, *75*, 3145-3152.

[106] Chen J., Lee C.S., Shen Y., Smith R.D. and Baehrecke E.H., *Integration of capillary isoelectric focusing with capillary reversed-phase liquid chromatography for two-dimensional proteomics separation*. Electrophoresis, **2002**, *23*, 3143-3148.

[107] Moore A.W. and Jorgenson J.W., *Rapid comprehensive two-dimensional separations of peptides via RPLC-optically gated capillary zone electrophoresis*. Anal. Chem., **1995**, *67*, 3448-3455.

[108] Issaq H.J., Chan K.C., Janini G.M. and Muschik G.M., *A simple two-dimensional high performance liquid chromatography/high performance capillary electrophoresis set-up for the separation of complex mixtures*. Electrophoresis, **1999**, *20*, 1533-1537.

[109] Washburn M.P., Wolters D. and Yates J.R., *Large-scale analysis of the yeast proteome by multidimensional protein identification technology*. Nat Biotech, **2001**, *19*, 242-247.

[110] Lorne Burke T.W., Mant C.T., Black J.A. and Hodges R.S., *Strong cation-exchange high-performance liquid chromatography of peptides : Effect of non-specific hydrophobic interactions and linearization of peptide retention behaviour*. Journal of Chromatography A, **1989**, *476*, 377-389.

[111] Peng J., Elias J.E., Thoreen C.C., Licklider L.J. and Gygi S.P., *Evaluation of Multidimensional Chromatography Coupled with Tandem Mass Spectrometry (LC/LC-MS/MS) for Large-Scale Protein Analysis: The Yeast Proteome*. J. Proteome Res., **2003**, *2*, 43-50.

[112] Davis M.T., Beierle J., Bures E.T., McGinley M.D., Mort J., Robinson J.H., Spahr C.S., Yu W., Luethy R. and Patterson S.D., *Automated LC-LC-MS-MS platform using binary ion-exchange and gradient*

*reversed-phase chromatography for improved proteomic analyses*. Journal of Chromatography B: Biomedical Sciences and Applications, **2001**, *752*, 281-291.

[113] Issaq H.J., Chan K.C., Janini G.M., Conrads T.P. and Veenstra T.D., *Multidimensional separation of peptides for effective proteomic analysis*. Journal of Chromatography B, **2005**, *817*, 35-47.

[114] Nice E.C., Rothacker J., Weinstock J., Lim L. and Catimel B., *Use of multidimensional separation protocols for the purification of trace components in complex biological samples for proteomics analysis*. Journal of Chromatography A, **2007**, *1168*, 190-210.

[115] McLafferty F.W., Breuker K., Jin M., Han X., Infusini G., Jiang H., Kong X. and Begley T.P., *Top-down MS, a powerful complement to the high capabilities of proteolysis proteomics*. FEBS J, **2007**, *274*, 6256-6268.

[116] Chait B.T., *Chemistry. Mass spectrometry: bottom-up or top-down?* Science, **2006**, *314*, 65-66.

[117] Desiere F., Deutsch E.W., Nesvizhskii A.I., Mallick P., King N.L., Eng J.K., Aderem A., Boyle R., Brunner E., Donohoe S., et al., *Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry*. Genome Biol, **2005**, *6*, R9.

[118] Martens L., Hermjakob H., Jones P., Adamski M., Taylor C., States D., Gevaert K., Vandekerckhove J. and Apweiler R., *PRIDE: the proteomics identifications database*. Proteomics, **2005**, *5*, 3537-3545.

[119] Nonaka T., Hashimoto Y. and Takio K., *Kinetic characterization of lysine-specific metalloendopeptidases from Grifola frondosa and Pleurotus ostreatus fruiting bodies*. J Biochem, **1998**, *124*, 157-162.

[120] Taouatas N., Drugan M.M., Heck A.J. and Mohammed S., *Straightforward ladder sequencing of peptides using a Lys-N metalloendopeptidase*. Nat Methods, **2008**, *5*, 405-407.

[121] Hopley C., Bristow T., Lubben A., Simpson A., Bull E., Klagkou K., Herniman J. and Langley J., *Towards a universal product ion mass spectral library - reproducibility of product ion spectra across eleven different mass spectrometers*. Rapid Commun Mass Spectrom, **2008**, *22*, 1779-1786.

[122] Roepstorff P. and Fohlman J., *Proposal for a common nomenclature for sequence ions in mass spectra of peptides*. Biomed Mass Spectrom, **1984**, *11*, 601.

[123] Johnson R.S., Martin S.A., Biemann K., Stults J.T. and Watson J.T., *Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass spectrometer: differentiation of leucine and isoleucine*. Anal. Chem., **1987**, *59*, 2621-2625.

[124] Johnson R.S., Martin S.A. and Biemann K., *Collision-induced fragmentation of (M + H)+ ions of peptides. Side chain specific sequence ions*. International Journal of Mass Spectrometry and Ion Processes, **1988**, *86*, 137-154.

[125] Zubarev R.A., Kelleher N.L. and McLafferty F.W., *Electron Capture Dissociation of Multiply Charged Protein Cations. A Nonergodic Process*. J. Am. Chem. Soc., **1998**, *120*, 3265-3266.

[126] Syka J.E., Coon J.J., Schroeder M.J., Shabanowitz J. and Hunt D.F., *Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry*. Proc Natl Acad Sci U S A, **2004**, *101*, 9528-9533.

[127] Bakhtiar R. and Guan Z., *Electron capture dissociation mass spectrometry in characterization of peptides and proteins*. Biotechnol Lett, **2006**, *28*, 1047-1059.

[128] Mikesh L.M., Ueberheide B., Chi A., Coon J.J., Syka J.E., Shabanowitz J. and Hunt D.F., *The utility of ETD mass spectrometry in proteomic analysis*. Biochim Biophys Acta, **2006**, *1764*, 1811-1822.

[129] McAlister G.C., Berggren W.T., Griep-Raming J., Horning S., Makarov A., Phanstiel D., Stafford G., Swaney D.L., Syka J.E., Zabrouskov V., et al., *A Proteomics Grade Electron Transfer Dissociation-Enabled Hybrid Linear Ion Trap-Orbitrap Mass Spectrometer*. J Proteome Res, **2008**.

[130] Wysocki V.H., Resing K.A., Zhang Q. and Cheng G., *Mass spectrometry of peptides and proteins*. Methods, **2005**, *35*, 211-222.

[131] Steen H. and Mann M., *The ABC's (and XYZ's) of peptide sequencing*. Nat Rev Mol Cell Biol, **2004**, *5*, 699-711.

[132] Bateman R.H., Carruthers R., Hoyes J.B., Jones C., Langridge J.I., Millar A. and Vissers J.P., *A novel precursor ion discovery method on a hybrid quadrupole orthogonal acceleration time-of-flight (Q-TOF) mass spectrometer for studying protein phosphorylation*. J Am Soc Mass Spectrom, **2002**, *13*, 792-803.

[133] Venable J.D., Dong M.Q., Wohlschlegel J., Dillin A. and Yates J.R., *Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra*. Nat Methods, **2004**, *1*, 39-45.

[134] Henzel W.J., Watanabe C. and Stults J.T., *Protein identification: the origins of peptide mass fingerprinting*. Journal of the American Society for Mass Spectrometry, **2003**, *14*, 931-942.

[135] Olsen J.V. and Mann M., *Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation*. Proc Natl Acad Sci U S A, **2004**, *101*, 13417-13422.

[136] Nesvizhskii A.I., Vitek O. and Aebersold R., *Analysis and validation of proteomic data generated by tandem mass spectrometry*. Nat Meth, **2007**, *4*, 787-797.

[137] Sadygov R.G., Cociorva D. and Yates J.R., 3rd, *Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book*. Nat Methods, **2004**, *1*, 195-202.

[138] Perkins D.N., Pappin D.J., Creasy D.M. and Cottrell J.S., *Probability-based protein identification by searching sequence databases using mass spectrometry data*. Electrophoresis, **1999**, *20*, 3551-3567.

[139] Eng J.K., Mccormack A.L. and Yates J.R., *An Approach to Correlate Tandem Mass-Spectral Data of Peptides with Amino-Acid-Sequences in a Protein Database*. Journal of the American Society for Mass Spectrometry, **1994**, *5*, 976-989.

[140] Savitski M.M., Nielsen M.L., Kjeldsen F. and Zubarev R.A., *Proteomics-grade de novo sequencing approach*. J Proteome Res, **2005**, *4*, 2348-2354.

[141] Moore R.E., Young M.K. and Lee T.D., *Qscore: an algorithm for evaluating SEQUEST database search results*. J Am Soc Mass Spectrom, **2002**, *13*, 378-386.

[142] Klammer A.A. and MacCoss M.J., *Effects of modified digestion schemes on the identification of proteins from complex mixtures*. J Proteome Res, **2006**, *5*, 695-700.

[143] Colinge J., Masselot A., Giron M., Dessingy T. and Magnin J., *OLAV: towards high-throughput tandem mass spectrometry data identification*. Proteomics, **2003**, *3*, 1454-1463.

[144] Weatherly D.B., Atwood J.A., 3rd, Minning T.A., Cavola C., Tarleton R.L. and Orlando R., *A Heuristic method for assigning a false-discovery rate for protein identifications from Mascot database search results*. Mol Cell Proteomics, **2005**, *4*, 762-772.

[145] Nielsen M.L., Savitski M.M. and Zubarev R.A., *Improving protein identification using complementary fragmentation techniques in fourier transform mass spectrometry*. Mol Cell Proteomics, **2005**, *4*, 835-845.

[146] Higgs R.E., Knierman M.D., Freeman A.B., Gelbert L.M., Patil S.T. and Hale J.E., *Estimating the statistical significance of peptide identifications from shotgun proteomics experiments*. J Proteome Res, **2007**, *6*, 1758-1767.

[147] Gouw J.W., Pinkse M.W., Vos H.R., Moshkin Y.M., Verrijzer C.P., Heck A.J.R. and Krijgsveld J., *In vivo stable isotope labeling of fruit flies reveals post-transcriptional regulation in the maternal-to-zygotic transition*. Unpublished work, **2008**.

[148] Kall L., Storey J.D., MacCoss M.J. and Noble W.S., *Assigning significance to peptides identified by tandem mass spectrometry using decoy databases*. J Proteome Res, **2008**, *7*, 29-34.

[149] Nesvizhskii A.I. and Aebersold R., *Interpretation of shotgun proteomic data: the protein inference problem*. Mol Cell Proteomics, **2005**, *4*, 1419-1440.

[150] Rappsilber J. and Mann M., *What does it mean to identify a protein in proteomics?* Trends Biochem Sci, **2002**, *27*, 74-78.

[151] Elias J.E. and Gygi S.P., *Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry*. Nat Methods, **2007**, *4*, 207-214.

[152] Nesvizhskii A.I., Keller A., Kolker E. and Aebersold R., *A statistical model for identifying proteins by tandem mass spectrometry*. Anal Chem, **2003**, *75*, 4646-4658.

[153] States D.J., Omenn G.S., Blackwell T.W., Fermin D., Eng J., Speicher D.W. and Hanash S.M., *Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study*. Nat Biotechnol, **2006**, *24*, 333-338.

[154] Bantscheff M., Schirle M., Sweetman G., Rick J. and Kuster B., *Quantitative mass spectrometry in proteomics: a critical review*. Anal Bioanal Chem, **2007**, *389*, 1017-1031.

[155] Ong S.E. and Mann M., *Mass spectrometry-based proteomics turns quantitative*. Nat Chem Biol, **2005**, *1*, 252-262.

[156] MacCoss M.J. and Matthews D.E., *Quantitative MS for proteomics: teaching a new dog old tricks*. Anal Chem, **2005**, *77*, 294A-302A.

[157] Liu H., Sadygov R.G. and Yates J.R., 3rd, *A model for random sampling and estimation of relative protein abundance in shotgun proteomics*. Anal. Chem., **2004**, *76*, 4193-4201.

[158] Old W.M., Meyer-Arendt K., Aveline-Wolf L., Pierce K.G., Mendoza A., Sevinsky J.R., Resing K.A. and Ahn N.G., *Comparison of label-free methods for quantifying human proteins by shotgun proteomics*. Mol Cell Proteomics, **2005**, *4*, 1487-1502.

[159] Rappsilber J., Ryder U., Lamond A.I. and Mann M., *Large-scale proteomic analysis of the human spliceosome*. Genome Res, **2002**, *12*, 1231-1245.

[160] Ishihama Y., Oda Y., Tabata T., Sato T., Nagasu T., Rappsilber J. and Mann M., *Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein*. Mol Cell Proteomics, **2005**, *4*, 1265-1272.

[161] Lu P., Vogel C., Wang R., Yao X. and Marcotte E.M., *Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation*. Nat Biotechnol, **2007**, *25*, 117-124.

[162] Bondarenko P.V., Chelius D. and Shaler T.A., *Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry*. Anal. Chem., **2002**, *74*, 4741-4749.

[163] Ono M., Shitashige M., Honda K., Isobe T., Kuwabara H., Matsuzuki H., Hirohashi S. and Yamada T., *Label-free quantitative proteomics using large peptide data sets generated by nanoflow liquid chromatography and mass spectrometry*. Mol Cell Proteomics, **2006**, *5*, 1338-1347.

[164] Silva J.C., Denny R., Dorschel C.A., Gorenstein M., Kass I.J., Li G.Z., McKenna T., Nold M.J., Richardson K., Young P., et al., *Quantitative proteomic analysis by accurate mass retention time pairs*. Anal Chem, **2005**, *77*, 2187-2200.

[165] Smith R.D., Anderson G.A., Lipton M.S., Pasa-Tolic L., Shen Y., Conrads T.P., Veenstra T.D. and Udseth H.R., *An accurate mass tag strategy for quantitative and high-throughput proteome measure-*

*ments*. Proteomics, **2002**, *2*, 513-523.

[166] Silva J.C., Gorenstein M.V., Li G.Z., Vissers J.P. and Geromanos S.J., *Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition*. Mol Cell Proteomics, **2006**, *5*, 144-156.

[167] Andersen J.S., Wilkinson C.J., Mayor T., Mortensen P., Nigg E.A. and Mann M., *Proteomic characterization of the human centrosome by protein correlation profiling*. Nature, **2003**, *426*, 570-574.

[168] Ong S.E., Blagoev B., Kratchmarova I., Kristensen D.B., Steen H., Pandey A. and Mann M., *Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics*. Mol. Cell. Proteomics, **2002**, *1*, 376-386.

[169] Ishihama Y., Sato T., Tabata T., Miyamoto N., Sagane K., Nagasu T. and Oda Y., *Quantitative mouse brain proteomics using culture-derived isotope tags as internal standards*. Nat. Biotechnol., **2005**, *23*, 617-621.

[170] Ong S.E., Mittler G. and Mann M., *Identifying and quantifying in vivo methylation sites by heavy methyl SILAC*. Nat Methods, **2004**, *1*, 119-126.

[171] Gu S., Pan S., Bradbury E.M. and Chen X., *Precise peptide sequencing and protein quantification in the human proteome through in vivo lysine-specific mass tagging*. Journal of the American Society for Mass Spectrometry, **2003**, *14*, 1-7.

[172] Ibarrola N., Molina H., Iwahori A. and Pandey A., *A novel proteomic approach for specific identification of tyrosine kinase substrates using [13C]tyrosine*. J Biol Chem, **2004**, *279*, 15805-15813.

[173] Blagoev B., Kratchmarova I., Ong S.E., Nielsen M., Foster L.J. and Mann M., *A proteomics strategy to elucidate functional protein-protein interactions applied to EGF signaling*. Nat Biotechnol, **2003**, *21*, 315-318.

[174] Ong S.E., Kratchmarova I. and Mann M., *Properties of 13C-substituted arginine in stable isotope labeling by amino acids in cell culture (SILAC)*. J Proteome Res, **2003**, *2*, 173-181.

[175] Van Hoof D., Pinkse M.W., Oostwaard D.W., Mummery C.L., Heck A.J. and Krijgsveld J., *An experimental correction for arginine-to-proline conversion artifacts in SILAC-based quantitative proteomics*. Nat Methods, **2007**, *4*, 677-678.

[176] Andersen J.S., Lam Y.W., Leung A.K., Ong S.E., Lyon C.E., Lamond A.I. and Mann M., *Nucleolar proteome dynamics*. Nature, **2005**, *433*, 77-83.

[177] Blagoev B., Ong S.E., Kratchmarova I. and Mann M., *Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics*. Nat Biotechnol, **2004**, *22*, 1139-1145.

[178] Mousson F., Kolkman A., Pijnappel W.W., Timmers H.T. and Heck A.J., *Quantitative proteomics reveals regulation of dynamic components within TATA-binding protein (TBP) transcription complexes*. Mol Cell Proteomics, **2008**, *7*, 845-852.

[179] Muñoz J., van Hoof D., Pinkse M.W.H., Braam S.R., Linding R., Heck A.J.R., Mummery C.L. and Krijgsveld J., *Phosphorylation dynamics during early differentiation of human embryonic stem cells*. Unpublished work, **2008**.

[180] Ong S.E. and Mann M., *A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC)*. Nat Protoc, **2006**, *1*, 2650-2660.

[181] Gygi S.P., Rist B., Gerber S.A., Turecek F., Gelb M.H. and Aebersold R., *Quantitative analysis of complex protein mixtures using isotope-coded affinity tags*. Nat. Biotechnol., **1999**, *17*, 994-999.

[182] Hansen K.C., Schmitt-Ulms G., Chalkley R.J., Hirsch J., Baldwin M.A. and Burlingame A.L., *Mass spectrometric analysis of protein mixtures at low levels using cleavable 13C-isotope-coded affinity tag and multidimensional chromatography*. Mol Cell Proteomics, **2003**, *2*, 299-314.

[183] Zhou H., Ranish J.A., Watts J.D. and Aebersold R., *Quantitative proteome analysis by solid-phase isotope tagging and mass spectrometry*. Nat Biotechnol, **2002**, *20*, 512-515.

[184] Mirgorodskaya O.A., Kozmin Y.P., Titov M.I., Korner R., Sonksen C.P. and Roepstorff P., *Quantitation of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry using (18) O-labeled internal standards*. Rapid Commun. Mass Spectrom., **2000**, *14*, 1226-1232.

[185] Schnolzer M., Jedrzejewski P. and Lehmann W.D., *Protease-catalyzed incorporation of 18O into peptide fragments and its application for protein sequencing by electrospray and matrix-assisted laser desorption/ionization mass spectrometry*. Electrophoresis, **1996**, *17*, 945-953.

[186] Reubsaet L., Kool J., Wesseldijk F., Gouw J.W., Mohammed S., Jansen J.W.A., Maravilha R.T., Zijlstra F.J., Heck A.J.R. and van Hilten J., *Quantitative proteome analysis of human blister fluids using 18O labeling for the identification of potential biomarkers of inflammation in Complex Regional Pain Syndrome*. Unpublished work, **2008**.

[187] Miyagi M. and Rao K.C., *Proteolytic 18O-labeling strategies for quantitative proteomics*. Mass Spectrom. Rev., **2007**, *26*, 121-136.

[188] Ross P.L., Huang Y.N., Marchese J.N., Williamson B., Parker K., Hattan S., Khainovski N., Pillai S., Dey S., Daniels S., et al., *Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents*. Mol. Cell. Proteomics, **2004**, *3*, 1154-1169.

[189] Choe L., D'Ascenzo M., Relkin N.R., Pappin D., Ross P., Williamson B., Guertin S., Pribil P. and Lee K.H., *8-plex quantitation of changes in cerebrospinal fluid protein expression in subjects undergoing intravenous immunoglobulin treatment for Alzheimer's disease*. Proteomics, **2007**, *7*, 3651-3660.

[190] Leitner A. and Lindner W., *Chemistry meets proteomics: the use of chemical tagging reactions for MS-based proteomics*. Proteomics, **2006**, *6*, 5418-5434.

[191] Hsu J.L., Huang S.Y., Chow N.H. and Chen S.H., *Stable-isotope dimethyl labeling for quantitative proteomics*. Anal Chem, **2003**, *75*, 6843-6852.

[192] Mueller L.N., Brusniak M.Y., Mani D.R. and Aebersold R., *An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data*. J Proteome Res, **2008**, *7*, 51-61.

[193] Orchard S., Jones P., Taylor C., Zhu W., Julian R.K., Jr., Hermjakob H. and Apweiler R., *Proteomic data exchange and storage: the need for common standards and public repositories*. Methods Mol Biol, **2007**, *367*, 261-270.

[194] Stromback L., Hall D. and Lambrix P., *A review of standards for data exchange within systems biology*. Proteomics, **2007**, *7*, 857-867.

[195] Orchard S., Taylor C., Hermjakob H., Zhu W., Julian R. and Apweiler R., *Current status of proteomic standards development*. Expert Rev Proteomics, **2004**, *1*, 179-183.

[196] Pedrioli P.G., Eng J.K., Hubley R., Vogelzang M., Deutsch E.W., Raught B., Pratt B., Nilsson E., Angeletti R.H., Apweiler R., et al., *A common open representation of mass spectrometry data and its application to proteomics research*. Nat Biotechnol, **2004**, *22*, 1459-1466.

CHAPTER 2

# Quantitative proteomics in model organisms

Joost W. Gouw, Albert J.R. Heck and Jeroen Krijgsveld

Biomolecular Mass Spectrometry and Proteomics Group, Bijvoet Center for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands.

# I. Abstract

Mass spectrometry-based quantitative proteomics is increasingly becoming the standard to comprehensively analyze proteomes. Especially the labeling with stable isotopes has proved an effective means to accurately determine differential expression levels of proteins. Model organisms have been used for many decades, and they have produced a large proportion of our present-day knowledge of basic biological processes and their derailments in human diseases. While the focus has primarily been on genetic and genomic approaches, it is important that methods become available to extend this to quantitative proteomics. As a result, a range of methodologies have been explored to stable isotope-label model organisms in vitro and in vivo, ranging from relatively simple organisms as bacteria and yeast, via *C. elegans*, *Drosophila*, *Arabidopsis* up to mammals such as rats and mice. This review summarizes the status of quantitative proteomics as applied to various important model organisms, and describes how this has opened up ways to investigate biological processes at the protein level in health and disease, across the evolutionary tree.

# II. Introduction

Over the past few decades, proteomic research experienced significant advances and evolved into an indispensible technology to investigate the proteomic composition of biological samples. Proteomics shifted from the analysis of small sets of proteins towards the comprehensive investigation of a much larger number of proteins expressed in a cell, tissue or organisms. Moreover, the addition of a quantitative component to comparative proteomics has allowed the detection of differences in protein expression levels in two or more biological samples. Classically, this has been approached by two-dimensional gel electrophoresis technologies where spot densities represent changes in protein expression levels. However, due to increasingly powerful mass spectrometry-based technologies 2D gel electrophoresis is largely replaced by LC-based approaches [1]. Nowadays, a typical proteomics experiment is peptide centric and starts with the enzymatic digestion of a protein mixture after which one or more chromatographic steps are used to reduce sample complexity followed by mass spectrometric analysis. Importantly, by the addition of one or more internal standards this can be extended with a quantitative component. Because mass spectrometry is not inherently quantitative, it is necessary to compare the mass spectrometric response of a peptide to the internal standard. Often, each peptide carries a mass-label where the intensity ratio between this peptide and the unlabeled peptide accurately reflects the change in expression level. In addition, the internal standard accounts for variation such as sample losses during preparation.

It is crucial to consider where in the experimental workflow the internal standard is introduced to the sample. In FIGURE 1 is given an overview of the different positions in the

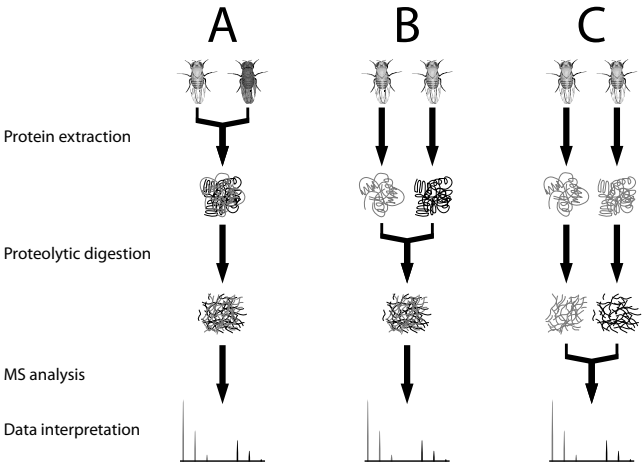


Figure 1. Strategies for quantitative proteomics. Stable isotopes can be incorporated at different stages of the quantitative workflow and are indicated in black. The methods are metabolic labeling (A), protein labeling (B) and peptide labeling (C). Relative expression levels are obtained by mass spectrometry where the signal of the unlabeled peptide is compared to that of the labeled peptide, except for the label-free approach.

quantitative workflow where the internal standard can be introduced, which can roughly be divided into four strategies. These mass labels can be metabolically incorporated into intact organisms or cells (FIG. 1A), after protein extraction by derivatizing proteins (e.g. with ICAT [2]) (FIG. 1B), during proteolytic digestion in $^{18}O$-labeled water [3, 4] FIG. 1C), or by chemically labeling peptides like the iTRAQ and stable isotope dimethyl labeling procedures [5-7] (FIG. 1C). Importantly, these methods introduce the mass-label at different positions in the workflow and therefore determine the moment of sample mixing. When the internal standard is introduced further downstream of the workflow higher levels of variation can be expected due to parallel sample processing. Clearly, the best place to introduce an internal standard is by metabolically incorporating the stable isotope into living organisms or cells, thereby producing the lowest variation before any sample processing occurs. In this type of labeling, stable isotope-labeled atoms such as $^{13}C$, $^{15}N$ or $^{18}O$ either in salts or amino acids, are metabolically incorporated into organisms or cells. Intact labeled and unlabeled cells or organisms can be combined where further processing is performed in the combined sample. Any effect due to sample handling occurs to the same extent to both the labeled and unlabeled sample (FIG. 1A), which is in contrast to parallel sample preparation at the protein level or at the peptide level (FIGURE 1B and 1C). Metabolic labeling requires the addition of an isotopically enriched element (e.g. as a salt or amino acid) to the growth media, in a form that makes it available for incorporation in the entire organism. Given this restriction, this method has so far been applied to a limited number of organisms. Nevertheless, the procedure of $^{15}N$ substitution in living organisms is very attractive since there are little to no side effects (cytological or morphological) reported [8, 9]. The alternative offered by label-free quantitation suffers from lower sensitivity because of parallel sample handling and therefore only extreme differences can be detected with enough accuracy. However, label-free quantitation can be applied to any sample since this technique does not require the incorporation of any internal standard.

In the pre-genomic era, so-called model organisms were primarily selected based on their size, ease of laboratory handling, life span and genetic accessibility. In fact, most of the current knowledge about development, evolution and genetics originated from such organisms including bacteria, yeast, fruit flies and mice. With increasing numbers of completed genome annotations, however, the focus shifts towards organisms that have unique genetic properties, are economically interesting or are related to human disease such as puffer fish, rice and plasmodium, respectively. Consequently, the definition model organism has broadened over the past decade and today model organisms are found in nearly

all branches of the 'tree of life' and provide extensive means to further investigate conservation or diversification of biological principles through evolution [10]. In recent years, increasing numbers of model organisms are being used for comparative proteomics and various methodologies have been developed to metabolically label them. In FIGURE 2 is

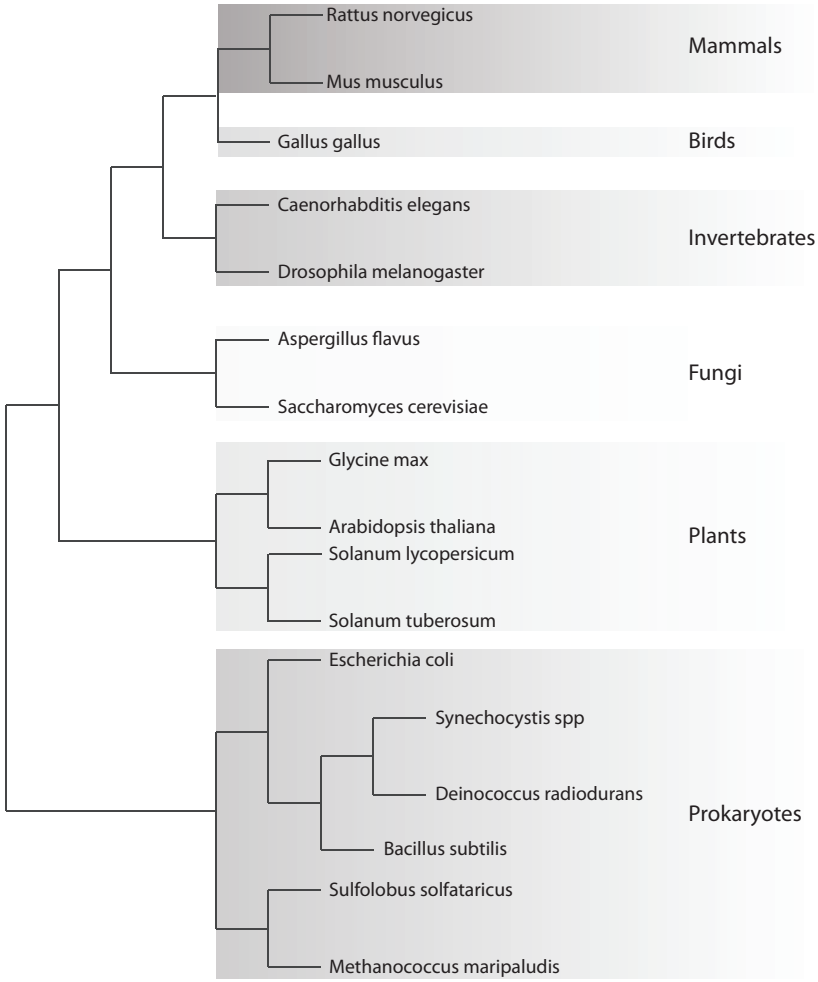


Figure 2. The tree of metabolically labeled life. Branch lengths are not proportional to time.

shown a compendium of the current 'tree of metabolically labeled life' and it can be seen that species in almost any branch can be labeled.

Recently, the development of strategies to label cells in culture with stable isotopes (SILAC) has found broad application and continues to expand to a wider range of cell lines. A number of excellent studies have been published [11, 12], reviewed elsewhere

[13, 14] showing the power of metabolic labeling with stable isotopes. Most of these studies focus on the effect of an isolated treatment (e.g. stress condition, growth factor) of cells grown in a petri dish, which is a somewhat artificial system since cells have often been maintained for thousands of cell doublings over many years. Moreover, any cell system neglects the differences in responses that would normally occur in an organism, because of the nature of the cells or organs, or communication between them. If the aim is to understand cells in their natural environment one should do so in the intact organism. Therefore, here we present an overview of intact model organisms that have been metabolically labeled for the purpose of quantitative proteomics. We provide per class of species the current status of the methodology as well as examples of their application to biological problems.

# III. Overview of metabolically labeled species

## Prokaryotes

The prokaryotes *Escherichia coli*, *Deinococcus radiodurans* and *Synechocystis spp.* were used to establish different $^{15}$N-metabolic labeling methodologies. In the first study to accomplish this, unstressed and $Cd^{2+}$ stressed *E. coli* were grown on normal and heavy isotope ($^{13}$C, $^{15}$N and $^{2}$H) depleted media, respectively. After mixing aliquots from both cultures, relative expression levels of intact proteins were determined by fitting the experimental results with the predicted shapes of the calculated isotopic envelopes [15]. Although this procedure readily lends itself for the determination of relative expression levels, identifying these proteins is somewhat more complicated and requires sophisticated fragmentation methods such as ECD and ETD [16]. To overcome this identification issue and instead of obtaining expression levels directly from intact proteins, proteins can be digested into peptides for both quantitative and qualitative information. This procedure was developed by growing *D. radiodurans* on fully $^{15}$N-enriched (> 98%) media after which proteins were digested into peptides to facilitate protein identification and quantitation [17]. A third method to determine protein expression levels employs the labeling of proteins with a subtle enrichment (~1-2%) of a stable isotope instead of using fully enriched (> 98%) labeled media. This approach, termed 'subtle modification of isotope ratio proteomics' or SMIRP, has been used to label *Synechocystis spp.* with various $^{13}$C/$^{12}$C isotope ratios to determine the feasibility of this approach. Optimal results were obtained with enrichments as low as ~1-2% of $^{13}$C which have no effect on either data-dependent acquisition or database searching algorithms but yields a measurable effect on peptide isotopic distribution of which an isotope ratio can be inferred [18]. A method similar to the labeling procedure of *D. radiodurans* (i.e. using fully enriched heavy nitrogen) included the incorporation of stable isotope labeled $^{13}$C atoms in the hyperthermophilic crenarchaeon *Sulfolobus solfataricus*. In this approach three different versions of the same protein are produced (unlabeled, $^{15}$N- and $^{13}$C-labeled) and allows for the analysis of three different samples in a single experiment [19]. The same group also investigated the applications and limitations of $^{15}$N-labeling and showed that stable isotope labeling with heavy nitrogen in combination with mass spectrometric detection provides an excellent tool for dynamic proteomic studies [20, 21]. This was further corroborated in another study where the archaeon *Methanococcus maripaludis* was $^{15}$N-labeled and protein expression levels were validated using real

time PCR [22]. An additional benefit of $^{15}$N-labeling is its use to increase the number and confidence of protein identifications. In $^{15}$N-labeling the number of nitrogen atoms in a peptide can easily be determined from the mass difference between labeled and unlabeled peptide pairs, which can be used as an additional criterion to accept or refute peptide identification [23]. One of the first applications utilizing metabolic labeling of a prokaryote included a combination of $^{15}$N-labeling and stable isotope labeling of amino acids in cell culture (SILAC). Changes in the membrane proteome during stationary phase adaptation of *Bacillus subtilis* were monitored and both techniques showed similar valuable data for quantification of bacterial membrane proteins [24].

## Plants

Metabolic labeling of plants dates back to the mid-sixties where labeled nitrogen (in the form of ammonium sulfate) was added in tracer amounts to fertilizer and used to evaluate the efficiency of fertilizer applications in rice (*Oryza sativa*) [25]. Metabolic incorporation of labeled nitrogen was accomplished by growing rice hydroponically (i.e. using mineral nutrient solutions instead of soil) and the nitrogen accumulated to specific parts of the plant. Although only ~30% enriched labeled nitrogen was used ($(^{15}NH_4)_2SO_4$ and $K^{15}NO_3$) to investigate the uptake and distribution of nitrogen, it was sufficient to spectroscopically determine the amount of heavy nitrogen in different organs of the rice plant [26]. The same method was used to label the potato plant (*Solanum tuberosum*) [27]. The first report of a highly enriched, uniformly labeled plant was published in 1994 where soybean plants (*Glycine max*) were labeled with ~98% $^{15}$N to investigate nutrient absorption and metabolism in human or animal studies [28]. Next, potato plants were uniformly $^{15}$N-labeled (> 98%) for the purpose of structural proteomics [29]. Intact plants were obtained by growing aseptic seed potato tubers using a hydroponics set-up for a period of 93 days in a climate controlled greenhouse on a nutrient solution containing 99% potassium nitrate ($K^{15}NO_3$) as the sole nitrogen source. Although *Solanum tuberosum* was the first plant to be fully $^{15}$N-labeled, *Arabidopsis thaliana* is the prime model species in plant biology, and the genome of this plant was the first to be sequenced [30]. Complemented with methods to incorporate stable isotopes *in vivo* this has facilitated mass spectrometry-based quantitative proteomics applied to plant biology.

***Arabidopsis thaliana*** • The genome of the flowering plant *Arabidopsis* is one of the smallest among plants, and was sequenced in 2000. It was concluded that this plant has 11 000 – 15 000 protein families, a number that is similar to other multicellular eukaryotes like *C.*

*elegans* and *Drosophila* [30]. To date, around 35 000 proteins are encoded by approximately 27 000 genes but these numbers change due to combined efforts by The Arabidopsis Information Resource (TAIR) that updates the Arabidopsis genome annotation annually [31]. *Arabidopsis* is an autotrophic species synthesizing all amino acids from inorganic nitrogen. This makes it particularly difficult to metabolically incorporate labeled amino acids into the proteome of the plant at high efficiency. However, labeling enrichments of 70-80% have been reported when suspension cells were grown in the presence of exogenously supplied heavy arginine for 7 days [32]. In this experiment, SILAC was used to identify regulated glutathione S-transferases and 14-3-3 proteins in response to treatment with salicylic acid. Disadvantages of this approach include incomplete labeling, the use of only one labeled amino acid reducing the number of quantifiable peptides and expensive labeled amino acids that are needed to reach efficient enrichments. This method is therefore limited to plant cell culture. Alternatively, these drawbacks can be circumvented by growing cell cultures on media containing highly enriched $^{15}N$ as the sole nitrogen source [33-36]. Typically, cells were grown for around 21 days in modified liquid media containing > 98% $^{15}N$ and complete uniform incorporation was achieved. No detrimental isotopic effects were observed since morphology and growth rate of the $^{15}N$-labeled cells were indistinguishable from their $^{14}N$ counterparts [33]. In some cases inverse labeling was performed to confirm results [35, 36]. Besides proteomics, also metabolomics can be studied using this type of labeling, but this application is limited to metabolites that contain nitrogen [33, 34]. Labeling with heavy carbon ($^{13}C$) is considered to be more efficient for this purpose.

Like for the labeling of cell cultures, heavy nitrogen has been used as the prime method to metabolically label intact plants *in vivo* with high levels of enrichment. The first publication of labeling of *Arabidopsis* for proteomic experiments was in 2007 [37]. Since then, several different $^{15}N$-labeling techniques have been investigated and include the comparison of partial versus full labeling [38], the automated analysis of uniformly labeled proteins using Mascot peptide identification in conjunction with the trans-proteomic pipeline [39] and a procedure referred to as HILEP (hydrophonic isotope labeling of entire plants) [40]. A notable aspect in the method to metabolically incorporate heavy nitrogen in intact plants is that almost all proteomics studies on *Arabidopsis* grow plants in liquid media. It can be argued that biologically meaningful results can only be obtained when plants are grown on solid medium thereby simulating growth conditions as natural as possible. Leaf senescence was investigated in Arabidopsis by growing the plants on solid labeled and unlabeled media [41]. Recently, the approach SILIP (Stable Isotope Labeling *In Planta*)

was published where the tomato plant *Solanum lycopersicum* was $^{15}$N-labeled by growing them for two months on solid media [42]. Besides metabolic labeling, most other quantitative proteomics methods such as iTRAQ, ICAT and $^{18}$O-labeling have been established in plants as well. For an overview see [43].

## Fungi

***Saccharomyces cerevisiae*** • The yeast *Saccharomyces cerevisiae* fulfills an important dual role being an industrially applied organism in itself while serving as a model organism for higher eukaryotes. First, this microorganism is extensively used in industry as it has the ability to produce alcohol and carbon dioxide. In addition, it can be used as a host for the production of proteins and small molecules exploited by pharmaceutical companies like for the production insulin [44]. Second, yeast has been used for a long time as a model organism in the field of molecular biology. Properties such as the rapid and easy growth under numerous conditions, its lack of pathogenicity and the availability of powerful genetic manipulation techniques make this organism perfectly suited for laboratory conditions. In addition, it was the first eukaryote to have its genome sequenced in 1996 which resulted in the prediction of approximately 6 000 protein coding genes [45]. Since then, large amounts of data based on genome, transcriptome, proteome and metabolome studies were generated. A great deal of this information is curated, stored and managed in the online database Saccharomyces Genome Database (SGD) [46]. In the post-genomic era, information from studies at different genomic levels is combined in a systems biology approach, resulting in a comprehensive understanding of eukaryotic cell biology. In this context, it is expected that yeast continues to provide new knowledge and insights [47, 48].

Metabolic labeling for quantitative proteomics is extensively applied in yeast and can be accomplished by two different approaches. Stable isotope labeled atoms, such as $^{15}$N, can be incorporated by growing yeast in media containing labeled ammonium sulphate as the sole nitrogen source [49]. The other method involves the incorporation of one or more stable isotope-labeled amino acids into the proteome of yeast. The former methodology was used in 1999 by Oda *et al.* to uniformly $^{15}$N-label yeast which they used to investigate protein expression as well as site-specific changes in phosphorylation levels. They reported that relatively small changes in phosphorylation (> 20%) can be reliably detected from small amounts of gel-separated proteins [50]. The group of Yates extensively used $^{15}$N-labeled yeast to establish methodologies for the analysis of quantitative proteomic samples as well as to address biological questions. For instance, by combining meta-

bolic labeling with multidimensional protein identification technology (MudPIT) they described a system useful for detailed quantitative proteomic analysis [51] which they used to investigate the correlation between mRNA and protein expression levels [52]. This approach was complemented with a correlation algorithm called RelEx for the automated analysis of quantitative data [53] and an algorithm called The Atomizer to determine isotope enrichment levels [54]. Both approaches were validated using complex mixtures with known enrichment levels of $^{15}$N-labeled yeast and, moreover, the RelEx algorithm was applied to the study of NaCl osmotic stress on protein level in yeast. An alternative data-acquisition method that was developed using $^{15}$N-labeled yeast in the Yates lab relies on data-independent analysis where quantitative results are determined directly from tandem mass spectra ($MS^2$) using a modified version of RelEx [55]. In addition, this group also evaluated the use of a LTQ-Orbitrap hybrid mass spectrometer for quantitative analysis using labeled yeast [56]. Other groups that focused on more fundamental research using $^{15}$N-labeled yeast investigated top-down quantitative proteomics approaches [57] and the correlation between spectral counting and metabolic labeling [58, 59]. Besides this important basic research, biological effects like response to nutrient limitations [60], adaptation to anaerobiosis [61] or phosphorylation level changes related to glucose activation [62] were studied using $^{15}$N-labeled yeast.

The other metabolic labeling strategy established in yeast involves the incorporation of one or more labeled amino acids. The first two papers describing such an approach both used deuterium-labeled leucine, but differed in the number of deuterium atoms in the amino acid. In one case triple deuterated leucine (Leu-$d_3$) was used [63] while the other used deca deuterated leucine (Leu-$d_{10}$) [64]. In both cases a yeast strain was used that is auxotrophic for leucine, ensuring that all leucine was replaced by labeled leucine. Once established, the Leu-$d_{10}$ metabolic labeling approach was used to investigate the response of the yeast proteome to $H_2O_2$ [65]. Besides labeled leucine, also stable isotope labeled arginine and lysine were used to label yeast. A double auxotroph strain was used to uniformly label yeast to study the changes in pheromone-induced phosphorylation [66]. However, a disadvantage of using labeled arginine in eukaryotes is that the accuracy of peptide ratio calculation is compromised by the metabolic conversion of arginine to proline [67, 68]. Although several solutions have been published [69, 70], prevention of this effect is advantageous to ensure accurate quantitation. A study that used labeled lysine only focussed on determining factors that prevent complete proteome analysis [71]. The sequencing speed of the mass spectrometer was found as one of these factors which was determined by com-

paring the number of identified 'light' and 'heavy' peptides. If the sequencing speed is sufficiently high then both forms of a SILAC pair should be sequenced and identified, but this study showed that this was only true for abundant peptides. The other factor that limited complete proteome coverage was the restricted dynamic range of the mass spectrometer compared to the dynamic range of the yeast proteome (i.e. 100 versus 10 000) [72]. The dynamic range of the mass spectrometer was determined by comparing the intensity of the most abundant SILAC pair to the least abundant pair in the same mass spectrum [71].

***Aspergillus flavus*** • Recently, two papers were published that described the SILAC labeling of the fungus *Aspergillus flavus* by growing this multicellular prototroph on media containing heavy arginine. This amino acid was chosen over lysine since it occurs at a frequency of about 6% in the proteome [73]. By determining the incorporation efficiency of 50 proteins they found an average enrichment level of 78% which is similar to the enrichment level of 81% reported by the SILAC labeling of *Arabidopsis* suspension cells [32] (see above). This suboptimal enrichment level was sufficient to quantify protein changes in response to environmental stimuli regulating biosynthesis of the carcinogen aflatoxin [73]. Another study by the same group, however, showed that this suboptimal enrichment hampered the quantification of intact proteins that have a large number of arginine residues. Here, they establish a quantitative top-down proteomics approach using this labeled fungus and showed that proteins with few arginines can readily be identified and quantified [74]. However, they suggest to reduce endogenous amino acid incorporation by using an arginine auxotroph strain thereby facilitating straightforward quantification of proteins that contain larger number of arginines [74].

## *Drosophila melanogaster* and *Caenorhabditis elegans*

Since the introduction of the nematode *Caenorhabditis elegans* and the fruit fly *Drosophila melanogaster* as model organisms in biological research [75-77], many fundamental biological principles were disclosed using these species. They were the first multicellular organisms to have their genome sequenced in respectively 1998 and 1999 [78, 79] and nowadays truly outstanding resources are available for both species [80, 81] such as the online databases Wormbase [82] and Flybase [83].The cellular complexity and the conservation of biological pathways between these invertebrates and higher organisms, including humans opened the door for using them as tools to study human genetics [84, 85]. Indeed, many biomedical discoveries were fuelled by *C. elegans* and *D. melanogaster* research.

Although quantitative genetics and functional genomics have been established in

these model organisms for some time now, quantitative proteomics based on metabolic labeling is only beginning to be introduced. Flies and worms can be metabolically $^{15}$N-labeled by feeding them on uniformly $^{15}$N-labeled *S. cerevisiae* and *E. coli*, respectively [86]. Whereas flies are completely labeled after one generation, worms need to be grown on labeled media for at least one more generation after which second-generation worms are completely labeled. However, due to different protein turnover rates in different tissues, it is necessary to analyze the enrichment of different proteins to ensure complete labeling. Labeled worms have been used to identify targets of insulin signaling [87] and labeled flies have been used to identify novel seminal fluid proteins [88] and proteins involved in the maternal-to-zygote transition [89]. In addition, both species have been used to optimize the identification and quantitation of $^{15}$N-labeled proteins in comparative proteomics [90]. The method for metabolic labeling of both *Drosophila* and *C. elegans* should be easily adoptable to many fly and worm labs since it requires only minor adaptations compared to routine protocols for growing flies and worms. Therefore, we expect this approach to find broad application in fly and worm developmental biology and beyond.



## Birds and mammals

One of the first reports where stable isotopes were enclosed in birds involved the partial SILAC labeling of the chicken *Gallus gallus*. A diet that consisted of 50% labeled valine was fed to six days old chickens to monitor the incorporation of the heavy amino acid and thereby facilitating the determination of protein turnover rates. Although these chickens are not fully labeled, such an approach allows for the accurate detection of protein turnover rates *in vivo* [91]. Wu and co-workers even extended the metabolic $^{15}$N-labeling approach to mammals by labeling rats using a diet source that was supplemented with $^{15}$N-enriched (> 99%) algal cells. After feeding the rats for 44 days, enrichment levels ranged from 74% to 92% depending on the type of tissue [92]. Using an improved labeling strategy they increased the enrichment to 94% throughout all tissues of the rat and these tissues can serve as internal standards to facilitate the quantitative proteomic analyses of complex mammalian tissue samples [93]. For instance, labeled rat brain was used to investigate the synaptosomal proteome of the rat cerebellum during post-natal development [94]. In addition to labeled rats, other mammals such as mice can be metabolically labeled by using a SILAC diet and second generation mice showed complete labeling whereas no obvious effects on growth, behavior or fertility were observed [95]. In these experiments $^{13}$C-labeled lysine was used, which is an essential amino acid thus preventing synthesis

from other (unlabeled) sources of amino acids. As a result, it is not straightforward to use trypsin as the preferred protease since this enzyme produces both C-terminal arginine and lysine peptides of which only the lysine-containing peptides can be used for quantitation. The endoprotease Lys-C, in contrast, produces only peptides that have a C-terminal lysine which makes Lys-C the preferred choice as the proteolytic enzyme.

A limitation to the use of labeled rats or mice could be the investment that is required to produce isotope-enriched offspring. However, the amount of protein that can be collected from basically every tissue, even from a single animal, is sufficient for hundreds of proteomic experiments. Furthermore, if this material is produced from a wild type, commonly-used strain it could be distributed and used as an internal standard even in facilities without resources to label these animals themselves.

## Outlook

Although stable isotope labeling has been achieved across many species in almost all branches of life (FIGURE 2), there are still some model organisms where the introduction of a labeling strategy would be very useful. For example, *Danio rerio* (zebrafish) has been established as an important model organism to study vertebrate biology [96, 97] especially in immunology [98], cancer [99], neurological disorders [100] and toxicology [101]. This research is facilitated by the development of excellent genetic techniques in zebrafish such as the targeted knock-down of genes by morpholino antisense technology [102]. There are several reasons accounting for the fact that this organism has not been metabolically labeled yet. Notably, the extensive developmental period and the absence of a labeled food source limit straightforward metabolic labeling. Yet, metabolic labeling of fish is not unprecedented with the incorporation of $^{15}N$ into Rainbow trout to study protein turnover [103].

The last but certainly not least important species that has not been metabolically labeled so far are humans. Up to now, SILAC approaches to metabolic label human cell lines have been firmly established [14], but the next logical step would be to extend this to human tissue. However, the holy grail of quantitative proteomics is to metabolically incorporate stable isotopes into intact humans. Heavy nitrogen would be an excellent candidate since there are delightful products for all kinds of people including vegetarians. The menu consists of plants, rice, potatoes, insects and rodents, but would we find volunteers?

# IV. Conclusions

A significant increase in both biological knowledge and technical expertise has been gained with the metabolic incorporation of stable isotopes in model organisms. On one hand this labeling procedure produces the most accurate quantitation and lowest variation of all labeling approaches. On the other hand, with the availability of methods for labeling many model organisms virtually unlimited biological problems can be investigated at the protein level. Most of the studies presented so far have focused on methodological development to metabolically incorporate stable isotopes into model organisms. In addition, mass spectrometry has been established as a powerful and indispensible tool for quantitative proteomics. Therefore the time is right to take advantage of these advancements and integrate them into the continued efforts in using model organisms for the understanding of fundamental biology, as well as for applied research.

# References


[1] Aebersold R. and Mann M., *Mass spectrometry-based proteomics*. Nature, **2003**, *422*, 198-207.

[2] Gygi S.P., Rist B., Gerber S.A., Turecek F., Gelb M.H. and Aebersold R., *Quantitative analysis of complex protein mixtures using isotope-coded affinity tags*. Nat Biotechnol, **1999**, *17*, 994-999.

[3] Mirgorodskaya O.A., Kozmin Y.P., Titov M.I., Korner R., Sonksen C.P. and Roepstorff P., *Quantitation of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry using (18) O-labeled internal standards*. Rapid Commun. Mass Spectrom., **2000**, *14*, 1226-1232.

[4] Schnolzer M., Jedrzejewski P. and Lehmann W.D., *Protease-catalyzed incorporation of $^{18}$O into peptide fragments and its application for protein sequencing by electrospray and matrix-assisted laser desorption/ionization mass spectrometry*. Electrophoresis, **1996**, *17*, 945-953.

[5] Ross P.L., Huang Y.N., Marchese J.N., Williamson B., Parker K., Hattan S., Khainovski N., Pillai S., Dey S., Daniels S., et al., *Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents*. Mol. Cell. Proteomics, **2004**, *3*, 1154-1169.

[6] Hsu J.L., Huang S.Y., Chow N.H. and Chen S.H., *Stable-isotope dimethyl labeling for quantitative proteomics*. Anal Chem, **2003**, *75*, 6843-6852.

[7] Lemeer S., Jopling C., Gouw J.W., Mohammed S., Heck A.J., Slijper M. and den Hertog J., *Comparative phosphoproteomics of zebrafish Fyn/Yes morpholino knockdown embryos*. Mol Cell Proteomics, **2008**, *7*, 2176-2187.

[8] Meselson M. and Stahl F.W., *The Replication of DNA in Escherichia Coli*. Proc Natl Acad Sci U S A, **1958**, *44*, 671-682.

[9] Uphaus R.A., Flaumenhaft E. and Katz J.J., *A living organism of unusual isotopic composition. Sequential and cumulative replacement of stable isotopes in Chlorella vulgaris*. Biochim Biophys Acta, **1967**, *141*, 625-632.

[10] Hedges S.B., *The origin and evolution of model organisms*. Nat Rev Genet, **2002**, *3*, 838-849.

[11] Andersen J.S., Lam Y.W., Leung A.K., Ong S.E., Lyon C.E., Lamond A.I. and Mann M., *Nucleolar proteome dynamics*. Nature, **2005**, *433*, 77-83.

[12] Olsen J.V., Blagoev B., Gnad F., Macek B., Kumar C., Mortensen P. and Mann M., *Global, in vivo, and site-specific phosphorylation dynamics in signaling networks*. Cell, **2006**, *127*, 635-648.

[13] Ong S.E. and Mann M., *Mass spectrometry-based proteomics turns quantitative*. Nat Chem Biol, **2005**, *1*, 252-262.

[14] Ong S.E. and Mann M., *A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC)*. Nat Protoc, **2006**, *1*, 2650-2660.

[15] Pasa-Tolic L., Jensen P.K., Anderson G.A., Lipton M.S., Peden K.K., Martinovic S., Tolic N., Bruce J.E. and Smith R.D., *High Throughput Proteome-Wide Precision Measurements of Protein Expression Using Mass Spectrometry*. J. Am. Chem. Soc., **1999**, *121*, 7949-7950.

[16] McLafferty F.W., Breuker K., Jin M., Han X., Infusini G., Jiang H., Kong X. and Begley T.P., *Top-down MS, a powerful complement to the high capabilities of proteolysis proteomics*. FEBS J, **2007**, *274*, 6256-6268.

[17] Conrads T.P., Alving K., Veenstra T.D., Belov M.E., Anderson G.A., Anderson D.J., Lipton M.S., Pasa-Tolic L., Udseth H.R., Chrisler W.B., et al., *Quantitative analysis of bacterial and mammalian proteomes using a combination of cysteine affinity tags and $^{15}$N-metabolic labeling*. Anal. Chem., **2001**, *73*, 2132-2139.

[18] Whitelegge J.P., Katz J.E., Pihakari K.A., Hale R., Aguilera R., Gomez S.M., Faull K.F., Vavilin D. and Vermaas W., *Subtle modification of isotope ratio proteomics; an integrated strategy for expression proteomics*. Phytochemistry, **2004**, *65*, 1507-1515.

[19] Snijders A.P., de Vos M.G. and Wright P.C., *Novel approach for peptide quantitation and sequencing based on $^{15}N$ and $^{13}C$ metabolic labeling*. J Proteome Res, **2005**, *4*, 578-585.

[20] Snijders A.P., de Koning B. and Wright P.C., *Perturbation and interpretation of nitrogen isotope distribution patterns in proteomics*. J Proteome Res, **2005**, *4*, 2185-2191.

[21] Snijders A.P., de Vos M.G., de Koning B. and Wright P.C., *A fast method for quantitative proteomics based on a combination between two-dimensional electrophoresis and $^{15}N$-metabolic labelling*. Electrophoresis, **2005**, *26*, 3191-3199.

[22] Xia Q., Hendrickson E.L., Zhang Y., Wang T., Taub F., Moore B.C., Porat I., Whitman W.B., Hackett M. and Leigh J.A., *Quantitative proteomics of the archaeon Methanococcus maripaludis validated by microarray analysis and real time PCR*. Mol Cell Proteomics, **2006**, *5*, 868-881.

[23] Zhong H., Marcus S.L. and Li L., *Two-dimensional mass spectra generated from the analysis of $^{15}N$-labeled and unlabeled peptides for efficient protein identification and de novo peptide sequencing*. J Proteome Res, **2004**, *3*, 1155-1163.

[24] Dreisbach A., Otto A., Becher D., Hammer E., Teumer A., Gouw J.W., Hecker M. and Volker U., *Monitoring of changes in the membrane proteome during stationary phase adaptation of Bacillus subtilis using in vivo labeling techniques*. Proteomics, **2008**, *8*, 2062-2076.

[25] Patnaik S. and Broadbent F.E., *Utilization of tracer nitrogen by rice in relation to time of application*. Agron. J., **1967**, *59*, 287-288.

[26] Muhammad S. and Kumazawa K., *The absorption, distribution, and redistribution of $^{15}N$-labelled ammonium and nitrate nitrogen administered at different growth stages of rice*. Soil Sci. Plant Nutr., **1974**, *20*, 47-55.

[27] Osaki M., Shirai J., Shinano T. and Tadano T., *$^{15}N$-Allocation of $^{15}NH_4$-N and $^{15}NO_3$-N to nitrogenous compounds at the vegetative growth stage of potato plants*. Soil Sci. Plant Nutr., **1995**, *41*, 699-708.

[28] Grusak M. and Pezeshgi S., *Uniformly $^{15}N$-labeled soybean seeds produced for use in human and animal nutrition studies: Description of a recirculating hydroponic growth system and whole plant nutrient and environmental requirements*. J. Sci. Food Agric., **1994**, *64*, 223-230.

[29] Ippel J.H., Pouvreau L., Kroef T., Gruppen H., Versteeg G., van den Putten P., Struik P.C. and van Mierlo C.P., *In vivo uniform (15)N-isotope labelling of plants: using the greenhouse for structural proteomics*. Proteomics, **2004**, *4*, 226-234.

[30] The Arabidopsis Genome Initiative, *Analysis of the genome sequence of the flowering plant Arabidopsis thaliana*. Nature, **2000**, *408*, 796-815.

[31] Swarbreck D., Wilks C., Lamesch P., Berardini T.Z., Garcia-Hernandez M., Foerster H., Li D., Meyer T., Muller R., Ploetz L., et al., *The Arabidopsis Information Resource (TAIR): gene structure and function annotation*. Nucleic Acids Res, **2008**, *36*, D1009-1014.

[32] Gruhler A., Schulze W.X., Matthiesen R., Mann M. and Jensen O.N., *Stable isotope labeling of Arabidopsis thaliana cells and quantitative proteomics by mass spectrometry*. Mol Cell Proteomics, **2005**, *4*, 1697-1709.

[33] Kim J.K., Harada K., Bamba T., Fukusaki E.-i. and Kobayashi A., *Stable Isotope Dilution-Based Accurate Comparative Quantification of Nitrogen-Containing Metabolites in Arabidopsis thaliana T87 Cells Using in Vivo $^{15}N$-Isotope Enrichment*. Biosci. Biotechnol. Biochem., **2005**, *69*, 1331-1340.

[34] Engelsberger W.R., Erban A., Kopka J. and Schulze W.X., *Metabolic labeling of plant cell cultures with*

*$K^{15}NO_3$ as a tool for quantitative analysis of proteins and metabolites*. Plant Methods, **2006**, *2*, 14.

[35] Benschop J.J., Mohammed S., O'Flaherty M., Heck A.J., Slijper M. and Menke F.L., *Quantitative phosphoproteomics of early elicitor signaling in Arabidopsis*. Mol. Cell. Proteomics, **2007**, *6*, 1198-1214.

[36] Lanquar V., Kuhn L., Lelievre F., Khafif M., Espagne C., Bruley C., Barbier-Brygoo H., Garin J. and Thomine S., *$^{15}$N-metabolic labeling for comparative plasma membrane proteomics in Arabidopsis cells*. Proteomics, **2007**, *7*, 750-754.

[37] Nelson C.J., Huttlin E.L., Hegeman A.D., Harms A.C. and Sussman M.R., *Implications of $^{15}$N-metabolic labeling for automated peptide identification in Arabidopsis thaliana*. Proteomics, **2007**, *7*, 1279-1292.

[38] Huttlin E.L., Hegeman A.D., Harms A.C. and Sussman M.R., *Comparison of full versus partial metabolic labeling for quantitative proteomics analysis in Arabidopsis thaliana*. Mol Cell Proteomics, **2007**, *6*, 860-881.

[39] Palmblad M., Bindschedler L.V. and Cramer R., *Quantitative proteomics using uniform $^{15}$N-labeling, MASCOT, and the trans-proteomic pipeline*. Proteomics, **2007**, *7*, 3462-3469.

[40] Bindschedler L.V., Palmblad M. and Cramer R., *Hydroponic isotope labelling of entire plants (HILEP) for quantitative plant proteomics; an oxidative stress case study*. Phytochemistry, **2008**.

[41] Hebeler R., Oeljeklaus S., Reidegeld K.A., Eisenacher M., Stephan C., Sitek B., Stuhler K., Meyer H.E., Sturre M.J., Dijkwel P.P., et al., *Study of early leaf senescence in Arabidopsis thaliana by quantitative proteomics using reciprocal $^{14}$N/$^{15}$N labeling and difference gel electrophoresis*. Mol Cell Proteomics, **2008**, *7*, 108-120.

[42] Schaff J.E., Mbeunkui F., Blackburn K., Bird D.M. and Goshe M.B., *SILIP: A novel stable isotope labeling method for in planta quantitative proteomic analysis*. Plant J, **2008**.

[43] Thelen J.J. and Peck S.C., *Quantitative proteomics in plants: choices in abundance*. Plant Cell, **2007**, *19*, 3339-3346.

[44] Kjeldsen T., *Yeast secretory expression of insulin precursors*. Appl Microbiol Biotechnol, **2000**, *54*, 277-286.

[45] Goffeau A., Barrell B.G., Bussey H., Davis R.W., Dujon B., Feldmann H., Galibert F., Hoheisel J.D., Jacq C., Johnston M., et al., *Life with 6000 genes*. Science, **1996**, *274*, 546, 563-547.

[46] Cherry J.M., Adler C., Ball C., Chervitz S.A., Dwight S.S., Hester E.T., Jia Y., Juvik G., Roe T., Schroeder M., et al., *SGD: Saccharomyces Genome Database*. Nucleic Acids Res, **1998**, *26*, 73-79.

[47] Castrillo J.I. and Oliver S.G., *Yeast as a touchstone in post-genomic research: strategies for integrative analysis in functional genomics*. J Biochem Mol Biol, **2004**, *37*, 93-106.

[48] Kumar A. and Snyder M., *Emerging technologies in yeast genomics*. Nat Rev Genet, **2001**, *2*, 302-312.

[49] Gao H., Shen Y., Veenstra T.D., Harkewicz R., Anderson G.A., Bruce J.E., Pasa-Tolic L. and Smith R.D., *Two-dimensional electrophoretic/chromatographic separations combined with electrospray ionization FTICR mass spectrometry for high throughput proteome analysis*. Journal of Microcolumn Separations, **2000**, *12*, 383-390.

[50] Oda Y., Huang K., Cross F.R., Cowburn D. and Chait B.T., *Accurate quantitation of protein expression and site-specific phosphorylation*. Proc. Natl. Acad. Sci. U.S.A., **1999**, *96*, 6591-6596.

[51] Washburn M.P., Ulaszek R., Deciu C., Schieltz D.M. and Yates J.R., 3rd, *Analysis of quantitative proteomic data generated via multidimensional protein identification technology*. Anal Chem, **2002**, *74*,

1650-1657.

[52] Washburn M.P., Koller A., Oshiro G., Ulaszek R.R., Plouffe D., Deciu C., Winzeler E. and Yates J.R., 3rd, *Protein pathway and complex clustering of correlated mRNA and protein expression analyses in Saccharomyces cerevisiae*. Proc Natl Acad Sci U S A, **2003**, *100*, 3107-3112.

[53] MacCoss M.J., Wu C.C., Liu H., Sadygov R. and Yates J.R., 3rd, *A correlation algorithm for the automated quantitative analysis of shotgun proteomics data*. Anal. Chem., **2003**, *75*, 6912-6921.

[54] MacCoss M.J., Wu C.C., Matthews D.E. and Yates J.R., 3rd, *Measurement of the isotope enrichment of stable isotope-labeled proteins using high-resolution mass spectra of peptides*. Anal. Chem., **2005**, *77*, 7646-7653.

[55] Venable J.D., Dong M.Q., Wohlschlegel J., Dillin A. and Yates J.R., *Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra*. Nat Methods, **2004**, *1*, 39-45.

[56] Venable J.D., Wohlschlegel J., McClatchy D.B., Park S.K. and Yates J.R., 3rd, *Relative quantification of stable isotope labeled peptides using a linear ion trap-Orbitrap hybrid mass spectrometer*. Anal Chem, **2007**, *79*, 3056-3064.

[57] Du Y., Parks B.A., Sohn S., Kwast K.E. and Kelleher N.L., *Top-down approaches for measuring expression ratios of intact yeast proteins using Fourier transform mass spectrometry*. Anal Chem, **2006**, *78*, 686-694.

[58] Zybailov B., Coleman M.K., Florens L. and Washburn M.P., *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*, 6218-6224.

[59] Zybailov B., Mosley A.L., Sardiu M.E., Coleman M.K., Florens L. and Washburn M.P., *Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae*. J Proteome Res, **2006**, *5*, 2339-2347.

[60] Kolkman A., Daran-Lapujade P., Fullaondo A., Olsthoorn M.M., Pronk J.T., Slijper M. and Heck A.J., *Proteome analysis of yeast response to various nutrient limitations*. Mol. Syst. Biol., **2006**, *2*, 2006 0026.

[61] de Groot M.J., Daran-Lapujade P., van Breukelen B., Knijnenburg T.A., de Hulster E.A., Reinders M.J., Pronk J.T., Heck A.J. and Slijper M., *Quantitative proteomics and transcriptomics of anaerobic and aerobic yeast cultures reveals post-transcriptional regulation of key cellular processes*. Microbiology, **2007**, *153*, 3864-3878.

[62] Lecchi S., Nelson C.J., Allen K.E., Swaney D.L., Thompson K.L., Coon J.J., Sussman M.R. and Slayman C.W., *Tandem phosphorylation of Ser-911 and Thr-912 at the C terminus of yeast plasma membrane H+-ATPase leads to glucose-dependent activation*. J Biol Chem, **2007**, *282*, 35471-35481.

[63] Zhu H., Pan S., Gu S., Bradbury E.M. and Chen X., *Amino acid residue specific stable isotope labeling for quantitative proteomics*. Rapid Commun Mass Spectrom, **2002**, *16*, 2115-2123.

[64] Jiang H. and English A.M., *Quantitative analysis of the yeast proteome by incorporation of isotopically labeled leucine*. J Proteome Res, **2002**, *1*, 345-350.

[65] Jiang H. and English A.M., *Evaluation of D10-Leu metabolic labeling coupled with MALDI-MS analysis in studying the response of the yeast proteome to H2O2 challenge*. J Proteome Res, **2006**, *5*, 2539-2546.

[66] Gruhler A., Olsen J.V., Mohammed S., Mortensen P., Faergeman N.J., Mann M. and Jensen O.N., *Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway*. Mol Cell Proteomics, **2005**, *4*, 310-327.

[67] Everley P.A., Bakalarski C.E., Elias J.E., Waghorne C.G., Beausoleil S.A., Gerber S.A., Faherty B.K.,

Zetter B.R. and Gygi S.P., *Enhanced analysis of metastatic prostate cancer using stable isotopes and high mass accuracy instrumentation*. J Proteome Res, **2006**, *5*, 1224-1231.

[68] Ong S.E., Kratchmarova I. and Mann M., *Properties of 13C-substituted arginine in stable isotope labeling by amino acids in cell culture (SILAC)*. J Proteome Res, **2003**, *2*, 173-181.

[69] Van Hoof D., Pinkse M.W., Oostwaard D.W., Mummery C.L., Heck A.J. and Krijgsveld J., *An experimental correction for arginine-to-proline conversion artifacts in SILAC-based quantitative proteomics*. Nat Methods, **2007**, *4*, 677-678.

[70] Bendall S.C., Hughes C., Stewart M.H., Doble B., Bhatia M. and Lajoie G.A., *Prevention of Amino Acid Conversion in SILAC Experiments with Embryonic Stem Cells*. Mol Cell Proteomics, **2008**, *7*, 1587-1597.

[71] de Godoy L.M., Olsen J.V., de Souza G.A., Li G., Mortensen P. and Mann M., *Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system*. Genome Biol, **2006**, *7*, R50.

[72] Ghaemmaghami S., Huh W.K., Bower K., Howson R.W., Belle A., Dephoure N., O'Shea E.K. and Weissman J.S., *Global analysis of protein expression in yeast*. Nature, **2003**, *425*, 737-741.

[73] Georgianna D.R., Hawkridge A.M., Muddiman D.C. and Payne G.A., *Temperature-dependent regulation of proteins in Aspergillus flavus: whole organism stable isotope labeling by amino acids*. J Proteome Res, **2008**, *7*, 2973-2979.

[74] Collier T.S., Hawkridge A.M., Georgianna D.R., Payne G.A. and Muddiman D.C., *Top-down identification and quantification of stable isotope labeled proteins from Aspergillus flavus using online nano-flow reversed-phase liquid chromatography coupled to a LTQ-FTICR mass spectrometer*. Anal Chem, **2008**, *80*, 4994-5001.

[75] Brenner S., *The genetics of Caenorhabditis elegans*. Genetics, **1974**, *77*, 71-94.

[76] Castle W.E., *Inbreeding, Cross-Breeding and Sterility in Drosophila*. Science, **1906**, *23*, 153.

[77] Bridges C.B., *Direct Proof through Non-Disjunction That the Sex-Linked Genes of Drosophila Are Borne by the X-Chromosome*. Science, **1914**, *40*, 107-109.

[78] Adams M.D., Celniker S.E., Holt R.A., Evans C.A., Gocayne J.D., Amanatides P.G., Scherer S.E., Li P.W., Hoskins R.A., Galle R.F., et al., *The genome sequence of Drosophila melanogaster*. Science, **2000**, *287*, 2185-2195.

[79] *C. elegans* Sequencing Consortium, *Genome sequence of the nematode C. elegans: a platform for investigating biology*. Science, **1998**, *282*, 2012-2018.

[80] Antoshechkin I. and Sternberg P.W., *The versatile worm: genetic and genomic resources for Caenorhabditis elegans research*. Nat Rev Genet, **2007**, *8*, 518-532.

[81] Matthews K.A., Kaufman T.C. and Gelbart W.M., *Research resources for Drosophila: the expanding universe*. Nat Rev Genet, **2005**, *6*, 179-193.

[82] Rogers A., Antoshechkin I., Bieri T., Blasiar D., Bastiani C., Canaran P., Chan J., Chen W.J., Davis P., Fernandes J., et al., *WormBase 2007*. Nucleic Acids Res, **2008**, *36*, D612-617.

[83] Wilson R.J., Goodman J.L. and Strelets V.B., *FlyBase: integration and improvements to query tools*. Nucleic Acids Res, **2008**, *36*, D588-593.

[84] Bier E., *Drosophila, the golden bug, emerges as a tool for human genetics*. Nat Rev Genet, **2005**, *6*, 9-23.

[85] Kaletta T. and Hengartner M.O., *Finding function in novel targets: C. elegans as a model organism*. Nat Rev Drug Discov, **2006**, *5*, 387-398.

[86] Krijgsveld J., Ketting R.F., Mahmoudi T., Johansen J., Artal-Sanz M., Verrijzer C.P., Plasterk R.H. and Heck A.J., *Metabolic labeling of C. elegans and D. melanogaster for quantitative proteomics*. Nat Biotechnol, **2003**, *21*, 927-931.

[87] Dong M.Q., Venable J.D., Au N., Xu T., Park S.K., Cociorva D., Johnson J.R., Dillin A. and Yates J.R., 3rd, *Quantitative mass spectrometry identifies insulin signaling targets in C. elegans*. Science, **2007**, *317*, 660-663.

[88] Findlay G.D., Yi X., Maccoss M.J. and Swanson W.J., *Proteomics reveals novel Drosophila seminal fluid proteins transferred at mating*. PLoS Biol, **2008**, *6*, e178.

[89] Gouw J.W., Pinkse M.W., Vos H.R., Moshkin Y.M., Verrijzer C.P., Heck A.J.R. and Krijgsveld J., *In vivo stable isotope labeling of fruit flies reveals post-transcriptional regulation in the maternal-to-zygotic transition*. Unpublished work, **2008**.

[90] Gouw J.W., Tops B.B.J., Mortensen P., Heck A.J.R. and Krijgsveld J., *Optimizing identification and quantitation of $^{15}$N-labeled proteins in comparative proteomics*. Anal. Chem., **2008**, *80*, 7796-7803.

[91] Doherty M.K., Whitehead C., McCormack H., Gaskell S.J. and Beynon R.J., *Proteome dynamics in complex organisms: using stable isotopes to monitor individual protein turnover rates*. Proteomics, **2005**, *5*, 522-533.

[92] Wu C.C., MacCoss M.J., Howell K.E., Matthews D.E. and Yates J.R., 3rd, *Metabolic labeling of mammalian organisms with stable isotopes for quantitative proteomic analysis*. Anal. Chem., **2004**, *76*, 4951-4959.

[93] McClatchy D.B., Dong M.Q., Wu C.C., Venable J.D. and Yates J.R., 3rd, *$^{15}$N metabolic labeling of mammalian tissue with slow protein turnover*. J. Proteome Res., **2007**, *6*, 2005-2010.

[94] McClatchy D.B., Liao L., Park S.K., Venable J.D. and Yates J.R., *Quantification of the synaptosomal proteome of the rat cerebellum during post-natal development*. Genome Res, **2007**, *17*, 1378-1388.

[95] Kruger M., Moser M., Ussar S., Thievessen I., Luber C.A., Forner F., Schmidt S., Zanivan S., Fassler R. and Mann M., *SILAC mouse for quantitative proteomics uncovers kindlin-3 as an essential factor for red blood cell function*. Cell, **2008**, *134*, 353-364.

[96] Grunwald D.J. and Eisen J.S., *Headwaters of the zebrafish -- emergence of a new model vertebrate*. Nat Rev Genet, **2002**, *3*, 717-724.

[97] Lieschke G.J. and Currie P.D., *Animal models of human disease: zebrafish swim into view*. Nat Rev Genet, **2007**, *8*, 353-367.

[98] Langenau D.M. and Zon L.I., *The zebrafish: a new model of T-cell and thymic development.* Nat Rev Immunol, **2005**, *5*, 307-317.

[99] Feitsma H. and Cuppen E., *Zebrafish as a cancer model*. Mol Cancer Res, **2008**, *6*, 685-694.

[100] Flinn L., Bretaud S., Lo C., Ingham P.W. and Bandmann O., *Zebrafish as a new animal model for movement disorders.* J Neurochem, **2008**, *106*, 1991-1997.

[101] McGrath P. and Li C.Q., *Zebrafish: a predictive model for assessing drug-induced toxicity.* Drug Discov Today, **2008**, *13*, 394-401.

[102] Eisen J.S. and Smith J.C., *Controlling morpholino experiments: don't stop making antisense*. Development, **2008**, *135*, 1735-1743.

[103] Carter C., Owen S., He Z., Watt P., Scrimgeour C., Houlihan D. and Rennie M., *Determination of Protein Synthesis in Rainbow Trout, Oncorhynchus Mykiss, Using a Stable Isotope*. J Exp Biol, **1994**, *189*, 279-284.

# CHAPTER 3

# *In vivo* stable isotope labeling of fruit flies reveals post-transcriptional regulation in the maternal-to-zygotic transition

Joost W. Gouw[†], Martijn W.H. Pinkse[†,‡], Harmjan R. Vos[†,§], Yuri Moshkin[ϡ], C. Peter Verrijzer[ϡ], Albert J.R. Heck[†] and Jeroen Krijgsveld[†]

[†]Biomolecular Mass Spectrometry and Proteomics Group, Bijvoet Center for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands. [ϡ]Department of Biochemistry, Center for Biomedical Genetics, Erasmus University Medical Center, Rotterdam, The Netherlands.

[‡]Current address: Analytical Biotechnology, Delft University of Technology, Julianalaan 67, 2628 BC Delft, The Netherlands.

[§]Current address: Laboratory of Physiological Chemistry, University Medical Center Utrecht, Universiteitsweg 100, 3584 CG Utrecht, The Netherlands.

# I. Abstract

An important hallmark in embryonic development is characterized by the maternal-to-zygotic transition (MZT), where zygotic transcription is activated by a maternally controlled environment. Post-transcriptional and translational regulation is critical for this transition and has been investigated in considerable detail at the gene level. We have used a proteomic approach employing metabolic labeling of Drosophila to quantitatively assess changes in protein expression levels before and after the MZT. By combining stable isotope labeling of fruit flies in vivo with high accuracy quantitative mass spectrometry we could quantify 2,232 proteins, of which about half changed in abundance during this process. We show that ~500 proteins increased in abundance, providing direct evidence of the identity of proteins as a product of embryonic translation. The group of down-regulated proteins is dominated by maternal factors involved in translational control of maternal and zygotic transcripts. Surprisingly, a direct comparison of transcript and protein levels showed that the mRNA levels of down-regulated proteins remained relatively constant, indicating a translational control mechanism specifically targeting these proteins. In addition, the finding of specific processing of a number of proteins provides evidence for post-translational regulation. Altogether, this quantitative proteomic study provides a dynamic profile of known and novel proteins of maternal as well as embryonic origin. This provides insight into the production, stability and modification of individual proteins, while discrepancies between transcriptional profiles and protein dynamics indicate novel control mechanisms in genome activation during early fly development.

# II. Introduction

In many organisms, the first few hours of development are controlled by maternal proteins and mRNAs, which are deposited into the egg during oogenesis. After fertilization, the primary roles of these factors are to facilitate zygotic transcription and to establish the initial body framework. In Drosophila, zygotic transcription is initiated after approximately two hours of development, when the first 13 synchronous nuclear divisions give rise to the formation of the syncytial blastoderm. This is also referred to as the maternal-to-zygotic transition, or MZT. The first developmental processes controlled by zygotic factors are mitotic cycle 14 and the cellularization of the blastoderm thereby hallmarking the midblastula transition (MBT). Although the existence of this process has been known for a long time [1-5], the molecular mechanisms regulating the transition from mother to zygote are only beginning to be unraveled.



The mother transfers a large number of mRNAs to the oocyte, estimated at ~7 000 transcripts [6, 7]. The bulk of these are degraded [8], while a selected set needs to be stabilized allowing translation to sustain development of the embryo. This is achieved by the combined effects of mRNA localization, (de)stabilization by maternal and zygotic proteins, and translational repression and activation. Over the years it has become clear that in Drosophila multiple mechanisms act simultaneously to achieve protein expression at the right dose, at the right time and at the right location. One way of localizing a particular protein is to stabilize and localize its mRNA transcript prior to translation which ensures high levels of protein to restricted well-defined cytoplasmic positions [9, 10]. This complements mechanisms suppressing activation of untranslated transcripts, which have been shown to aggregate in specific cytoplasmic granules, known as P bodies [11, 12].

One of the important questions in the activation of the zygotic genome relates to the origin of proteins, either by deposition in the oocyte by the mother or by transcriptional and translational activity in the embryo. Although recent proteomics studies aimed to define the *Drosophila* proteome [13-15], they investigated a different developmental event or they did not specifically focus on fly development. In a number of recent studies genomic techniques were used to distinguish maternal from zygotic gene expression. Lécuyer et al. used high-resolution fluorescent in situ hybridization assuming maternal and zygotic transcripts localize in the cytoplasm and nucleus, respectively. De Renzis et al. addressed a similar question by investigating chromosomal ablated mutants to discriminate between transcriptional and post-transcriptional regulation of gene expression, and it

was estimated that approximately 20% of the transcripts at cycle 14 were of zygotic origin [6]. Although the presence and precise localization of transcripts are crucial to understand developmental activation of the embryo, they do not necessarily allow extrapolation to protein expression. Notably, multiple mechanisms shown to determine mRNA stability and translational activity, (e.g. dependent on or independent of deadenylation, targets of RNA silencing or transacting factors) provide an additional level of regulation [16]. The result of the combined effect of these post-transcriptional processes can only be captured by determining expression levels of individual proteins before and after MZT.

Therefore, we have used a proteomic approach quantifying the relative protein expression levels before (1.5 hrs after oviposition, embryonic stages 1-3) and after MZT (4.5 hrs after oviposition, embryonic stages 6-9). By applying a combined approach using in vivo labeling of fruit flies by the incorporation of stable isotope-coded nitrogen ($^{15}N$) [17] combined with LC-MS/MS, more than 2 200 proteins could be quantitated. About half of these changed in abundance, of which ~500 proteins increased, providing for the first time direct evidence of the identity of proteins as a product of embryonic translation in a large-scale approach. While these up-regulated proteins represent a wide variety of functional classes, maternal proteins were among the most dramatically down-regulated proteins including transacting factors involved in regulation of mRNA stability (including ME31B, SMG and a number of proteins interacting with these). Moreover, specific down-regulation of these proteins appears to be governed by a post-transcriptional mechanism, as evidenced by direct comparison of protein and transcript levels in the same samples. In addition, evidence was found that a limited number of proteins, including PABP and CP1, were subject to post-translational processing leading to truncation, possibly resulting in an altered function of these proteins. Altogether, this study provides a dynamic profile of known and novel proteins of maternal as well as embryonic origin associated with embryonic development in *Drosophila*.

# III. Materials and Methods

**Flies stock, labeling and embryo collection •** Wild-type OregonR flies were maintained by standard methods at 25 °C and were labeled as described [17]. Briefly, larvae were grown in boxes containing $^{15}$N-labeled or unlabeled yeast and kept at 25 °C throughout larval and pupal developmental stages. Hatched flies were transferred to fly cages and kept on $^{15}$N-labeled or unlabeled yeast. Embryos were collected on agarose-agar plates completed with a small amount of $^{15}$N-labeled or unlabeled yeast which were removed from the fly cage after 90 minutes. Unlabeled stage 6-9 embryos were obtained by aging the 0-90 minutes embryos at standard methods for another 180 minutes whereas $^{15}$N-labeled embryos were processed immediately. Embryos were washed in water and dechorionated by incubation in 2.5% sodium hypochlorite for 90 s followed by another wash and kept at -20 °C. A biological duplicate, independent experiment was performed with swapped labels (i.e. 0-90 min. unlabeled, 180-270 min. $^{15}$N-labeled).

**Sample preparation •** Equal amounts of labeled and unlabeled embryos were combined and lysed in 8 M urea and 50 mM ammonium bicarbonate. Cellular debris was pelleted by centrifugation at 20 000 *g* for 20 minutes. Prior to digestion, proteins were reduced with 1 mM DTT and alkylated with 2 mM iodoacetamide. The mixture was diluted 4-fold to 2 M urea using 250 µL of 50 mM ammonium bicarbonate and 50 µL of trypsin solution, 0.1 mg/mL, and incubated overnight at 37 °C.

**Strong cation exchange •** Strong cation exchange was performed using a Zorbax BioSCX-Series II column (0.8 mm i.d. × 50 mm length, 3.5 µm), a FAMOS autosampler (LC-packing, Amsterdam, The Netherlands), a Shimadzu LC-9A binary pump and a SPD-6A UV-detector (Shimadzu, Tokyo, Japan). Prior to SCX chromatography, protein digests were desalted using a small plug of $C_{18}$ material (3 M Empore $C_{18}$ extraction disk) packed into a GELoader tip (Eppendorf) similar to what has been previously described [18],onto which ~10 µL of Aqua $C_{18}$ (5 µm, 200 Å) material was placed. The eluate was dried completely and subsequently reconstituted in 20% acetonitrile and 0.05% formic acid. After injection, a linear gradient of 1% $min^{-1}$ solvent B (500 mM KCl in 20% acetonitrile and 0.05% formic acid, pH 3.0) was performed. A total of 28 SCX fractions (1 min each, i.e., 50 µL elution volume) were manually collected and dried in a vacuum centrifuge.

**Nanoflow-HPLC-MS •** Dried residues were reconstituted in 50 µL of 0.1 M acetic acid and were analyzed by nanoflow liquid chromatography using an Agilent 1100 HPLC sys-

tem (Agilent Technologies) coupled on-line to a 7-Tesla LTQ-FT-ICR mass spectrometer (Thermo Electron). The liquid chromatography part of the system was operated in a setup essentially as described previously [19]. Aqua $C_{18}$, 5 µm, (Phenomenex) resin was used for the trap column, and ReproSil-Pur $C_{18}$-AQ, 3 µm, (Dr. Maisch GmbH) resin was used for the analytical column. Peptides were trapped at 5 µL/min in 100% solvent A (0.1 M acetic acid in water) on a 2 cm trap column (100 µm i.d., packed in-house) and eluted to a 40 cm analytical column (50 µm i.d., packed in-house) at ~100 nL/min in a 150-min gradient from 10 to 40% solvent B (0.1 M acetic acid in 8/2 (v/v) acetonitrile/water). The eluent was sprayed via standard coated emitter tips (New Objective), butt-connected to the analytical column. The mass spectrometer was operated in data-dependent mode, automatically switching between MS and MS/MS. Full scan MS spectra (from *m/z* 300 to 1 500) were acquired in the FT-ICR with a resolution of 100 000 at *m/z* 400 after accumulation to target value of 500 000. The three most intense ions at a threshold above 5 000 were selected for collision-induced fragmentation in the linear ion trap at normalized collision energy of 35% after accumulation to a target value of 15 000.

**Peptide identification** • All $MS^2$ spectra were converted to single DTA files using Bioworks 3.1 (Thermo) with default parameters and merged into a Mascot generic format file which was searched twice to identify both unlabeled and $^{15}N$-labeled peptides using an in-house licensed Mascot v2.1.0 search engine (Matrix Science) against a concatenated database containing both forward and reversed entries from an Integr8 *D. melanogaster* database (http://www.ebi.ac.uk/integr8, version 20060806) consisting of 32 508 sequences. Carbamidomethyl cysteine was set as a fixed modification; oxidized methionine, protein N-acetylation and N-terminal pyroglutamate were set as variable modifications. Trypsin was specified as the proteolytic enzyme, and up to two missed cleavages were allowed. The mass tolerance of the precursor ion was set to 15 ppm, and that of fragment ions was set to 0.8 Da. Both search results ($^{14}N$ and $^{15}N$ identifications) were merged into one HTML result page using an in-house developed Perl script for qualitative and quantitative analysis. A false-positive discovery rate (FDR) of < 1% was estimated [20] and accomplished by using a peptide cut-off score of 20 (horizontal line in SUPPLEMENTARY FIG. 1), a minimum of two peptides per protein and a protein cut-off score of 60.

**Protein quantitation** • An in-house developed $^{15}N$-version of MSQuant [21] was used to quantify relative protein levels. MSQuant was modified is such a way that the position of the partner peptide can be detected in the case of metabolic $^{15}N$-labeling (i.e., the position

of the $^{15}$N-labeled peptide in case when the unlabeled ($^{14}$N) peptide was identified and *vice versa*) and no modifications were made to the quantitation algorithm. Briefly, extracted ion chromatograms (XIC) for both unlabeled and labeled peptides are calculated and summed over consecutive MS cycles for the duration of their respective LC-MS peaks using monoisotopic peaks only. All full-MS scans were manually verified for sufficient signal and absence of interference by other signals. A minimum XIC threshold of at least 3 400 was used to exclude low intensity peptides. The total XIC intensity of labeled peptides was corrected for incomplete enriched nitrogen as described before [22]. Briefly, by using the average $^{15}$N enrichment level of 98.2% and the peptide's chemical formula the theoretical isotope pattern using natural abundances of elements as well as the pattern based on the $^{15}$N enrichment level was calculated and used to sum the intensities of the '-1' and '-2' isotopes. Relative peptide levels were computed by dividing the labeled (corrected) total XIC by the unlabeled total XIC. Relative protein expression levels ($\log_2$) with standard deviation were obtained by averaging individual peptide $\log_2$ ratios that identified the same protein with a minimum of 2 quantitated peptides per protein. Protein ratios were normalized to the average ratio of all proteins. Finally, the protein ratios from the two independent experiments were averaged and extreme protein ratios (i.e. more than a 2-fold change) were used for further data analysis.

**Supporting Information •** The peptide identifications have been made publicly available on the proteomics identifications database PRIDE (http://www.ebi.ac.uk/pride) and can be found under experiment accession numbers 8170 and 8171 in the project 'In vivo stable isotope labeling of fruit flies reveals post-transcriptional regulation in the maternal-to-zygotic transition'. Since PRIDE is not setup to handle quantitative data, these have been parsed into a database along with annotated spectra of all identified peptides. This can be contacted via https://bioinformatics.chem.uu.nl/drosophila. In addition, quantitative and qualitative results of all 2 232 quantitated proteins, including individual peptide ratios, sequence coverage and protein information, can be found in SUPPLEMENTAL TABLE 7. All the supplemental files can be found at https://bioinformatics.chem.uu.nl/thesis/joost.

**RNA Extraction and Affymetrix Microarray Hybridization •** RNA was extracted using the SV Total RNA Isolation System (Promega) and tested on an Agilent BioAnalyzer (Agilent). Samples with RNA integrity numbers > 8 were selected. Labeling, hybridization, washes and staining of microarrays were performed according to Affymetrix specifications.

**Microarray Statistical Analysis •** Statistical analysis of the microarray data was performed using R and Bioconductor free software as described previously [23]. Gene expression indexes were calculated using Robust Multichip Average (RMA) algorithm implemented in Bioconductor *affy* package. Distribution of the expression indexes is bimodal therefore we applied Multiple Covariance Determinant (MCD) algorithm implemented in *rrcov* R package to filter non-expressing genes. In total 4 657 genes have been selected for further analysis. Details of the statistical analysis and R scripts will be provided upon request.

# IV. Results

**Strategy for High-Throughput Proteomics** • The strategy used to identify and differentially quantify proteins in embryos in Browne's stages 1-3 and 6-9 is shown schematically in FIGURE 1. We started our approach with approximately 10 mL of $^{15}N$-labeled and unlabeled flies and collected 'heavy' and 'light' embryos for 90 minutes. Light embryos were allowed to develop for another 180 minutes to reach stage 6-9. The reverse ('label swap') experiment was conducted as well. Embryos were visually inspected after harvesting to confirm their desired developmental stage (SUPPLEMENTAL FIGURE 2A-C). Labeled and unlabeled samples were combined (FIGURE 1A), proteins were extracted and digested using trypsin and subjected to strong cation exchange (SCX) as the first separation step (FIG. 1B). One-minute SCX fractions were collected with peptides eluting in fractions 5 to 33. Each of these 28 fractions was subjected to the second dimension, nano-LC MS/MS. A typical chromatogram of one fraction is shown in FIGURE 1C. Extended column length (40

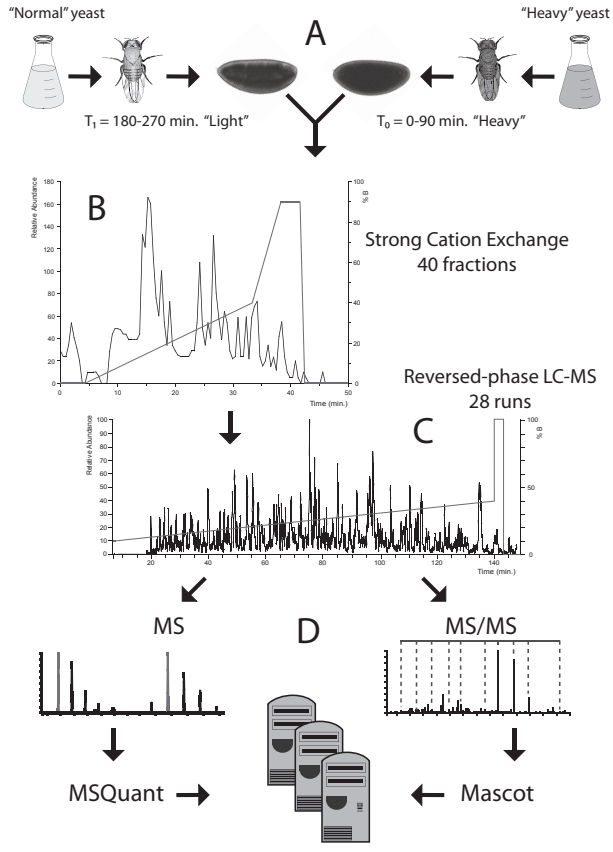


Figure 1. Experimental strategy used to analyze the maternal-to-zygotic transition (MZT). (A) Labeled and unlabeled yeast was used to grow embryos that were harvested before (right) and after (left) the MZT. Embryos were combined, lysed, digested and subjected to (B) SCX fractionation. Each of the 28 SCX fractions were analyzed by (C) reversed-phase LC-MS and the resulting (D) MS and MS/MS spectra were used to respectively quantify and identify peptides and stored in a PostgreSQL database.

cm) and gradient times (2.5 hrs) were used to obtain optimal peptide separation over more than 2 hours (FIGURE 1C). Analysis of these fractions led to an average of 10 000 fragment spectra per SCX fraction. Altogether, more than 250 000 spectra were searched against a *Drosophila* protein database consisting of 'forward' and 'reversed' protein sequences for protein identifications and an estimation of the false-positive (FP) discovery rate. To determine an optimal false-positive rate, peptide scores from both parts of the database were plotted (SUPPLEMENTAL FIGURE 1). With an arbitrary peptide cut-off score of 20 (horizontal line in SUPPLEMENTAL FIG. 1), a peptide FP rate of approximately 4% resulted in 79 196 peptide identifications. In addition to the peptide cut-off score, we used a minimum of two peptides per protein and a protein cut-off score of 60. This led to 2 736 protein identifications with a false-positive discovery rate of < 1%. The complete list of all proteins identified in stages 1-3 as well as stages 6-9 is given in SUPPLEMENTAL TABLE 1. Subsequently, proteins were quantified by integrating all MS peak areas of identified peptides using an in-house adapted version of MSQuant [21] (FIG. 1C). Using a minimum of two quantifiable peptides per protein, a total of 2 232 proteins were quantified (FIG. 1D). Peptides that could not be quantified represent mainly low abundant peptides that disappear in the noise or produce an insufficient number of data points for proper integration (each peptide was manually verified). Qualitative as well as the quantitative data was parsed into a PostGreSQL database (FIGURE 1D), accessible and searchable via https://bioinformatics.chem.uu.nl/drosophila.

**Protein Identification and Quantification •** Protein identifications were obtained by stringent filtering criteria, and therefore the compendium of the proteins found in stages 1-3 and 6-9 (SUPPLEMENTAL TABLE 1) provides a valuable resource for further exploration. To discover the quantitative difference between these datasets, thereby uncovering proteins associated with the maternal-to-zygotic transition, an overview of all relative expression patterns of the proteins that were quantified is shown in FIGURE 2. Extreme protein ratios (i.e. more than a 2-fold change) were validated in a duplicate, independent experiment where labels were swapped (i.e. 0-90 min. unlabeled, 180-270 min. $^{15}$N-labeled). SUPPLEMENTAL FIGURE 3 shows a scatter plot of the protein ratios of both experiments where a correlation coefficient of 0.871 indicates that protein ratios are in good agreement.

Protein quantitation was of high quality, reflected by the overall average coefficient of variation (CV) of 18%. The accuracy of protein quantitation in our approach was evidenced further by looking at ribosomal proteins only. Among the expression ratios of the 127 subunits identified in this study, a remarkable observation can be made indicating a

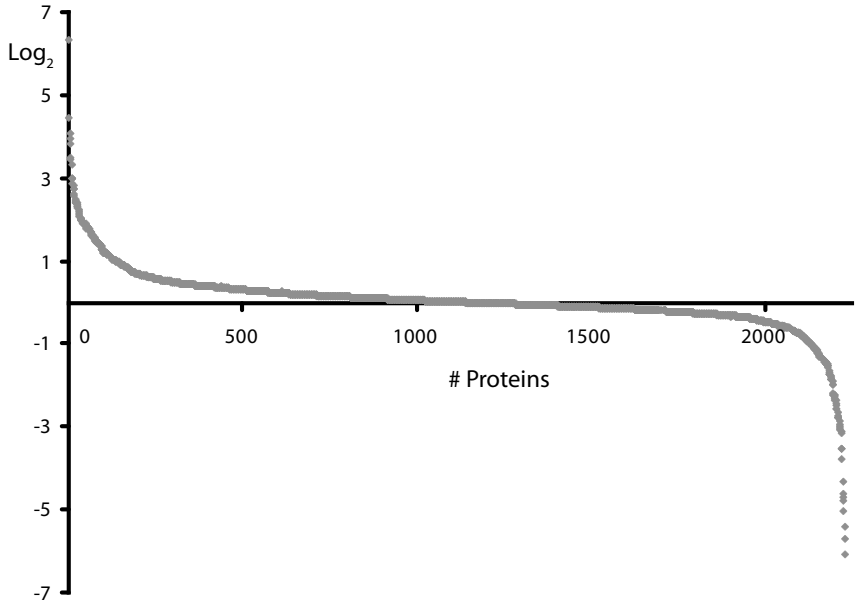


Figure 2. Expression profiles of all quantified proteins. $Log_2$ ratios of all 2 232 quantified proteins, where up-regulated and down-regulated proteins indicate zygotic and maternal expression, respectively.

subtle but clear distinction between cytoplasmic and mitochondrial ribosomal proteins (SUPPLEMENTAL TABLE 2). 90% of the cytoplasmic ribosomal subunits (67) were down-regulated whereas 70% of the mitochondrial subunits (42) were up-regulated. However, the absolute difference in expression ratios of both classes is very small (average $log_2$ ratio cytoplasmic -0.09, mitochondrial 0.03). We think that our method for protein quantitation is highly accurate, permitting the distinction of small differences. We next tested which proteins in the entire dataset significantly changed in abundance before and after MZT. Therefore, we employed a two-tailed *t*-test and calculated the probability if there was a significant ($p \leq 0.05$) difference between the ratio of individual proteins and the arithmetic mean of all protein ratios. This turned out to be the case for more than half of all quantified proteins (~1 300). These proteins were clustered into two distinct groups (i.e. up- and down regulated) for further functional analysis.

The first group of proteins consists of proteins that were significantly up-regulated with a ratio of at least one time the average coefficient of variation (i.e. > $log_2$ 0.238). Out of the 490 proteins falling in this category (SUPPLEMENTAL TABLE 3), 8 were found exclusively in the later time point and not in the 'maternal' sample, permitting the conclusion that these are purely zygotic proteins. Three of these eight proteins (Hiiragi, Neurotactin and CG13427) were previously annotated by De Renzis as part of a set of 59 genes that are

expressed as early as mitotic cycle 12 [6]. The only interpretation of the observed expression profile is that these are the product of zygotic transcription and translational activity that starts after approximately two hours of development. The up-regulation of the other 482 proteins indicates that these are present at some level at stages 1-3 and that their expression level is increased due to zygotic translation. Alternatively, these expression profiles could be the result of specific timed translation of maternal mRNA.

Using the same criteria as for the up-regulated proteins, the second group includes 121 proteins that were down-regulated (SUPPLEMENTAL TABLE 4). These can be interpreted as maternal proteins which are degraded over time, although it cannot be excluded that the expression ratio is a combined effect of zygotic translation and (stronger) degradation of maternal products. Yet, for simplicity we will refer to this set as 'maternal proteins'.

**Comparing Identity of Zygotic Proteins with Previous Datasets •** Previously, genomic techniques were used to infer maternal or zygotic origin by interpreting localization of transcripts (cytoplasmic or nuclear, resp.) [24] and by chromosomal ablation to discriminate between transcriptional and post-transcriptional regulation of gene expression [6].

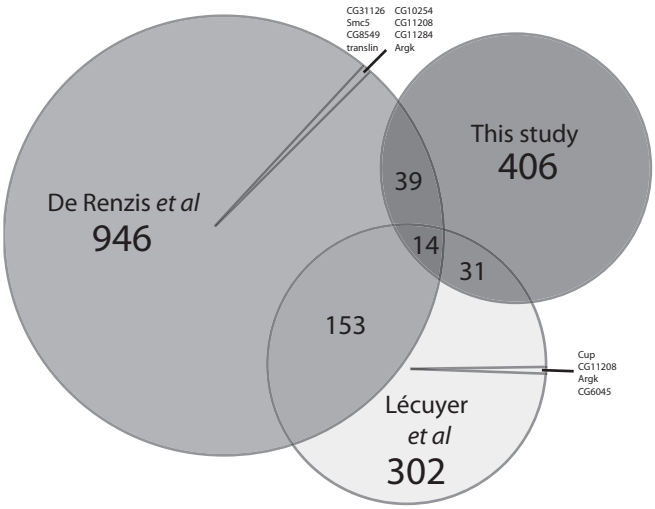


Figure 3. Comparison of datasets annotating zygotic products. Venn diagram showing the overlap between our dataset and published data to define zygotic transcripts and proteins. De Renzis et al [6] investigated chromosomal ablated mutants to discriminate between transcriptional and post-transcriptional regulation of gene expression, Lécuyer et al [24] used high-resolution fluorescent in situ hybridization to assess differential localization of maternal and zygotic transcripts. The gene symbols indicating 8 and 4 genes in the De Renzis [6] and Lécuyer [24] data, respectively, were claimed to be zygotic but proteins were found down-regulated in our work, suggesting maternal expression.

We believe the best proof for a protein to be zygotically expressed is its higher abundance in late compared to early embryos. We therefore propose to define the proteins up-regulated in our study as the first proteins to be expressed by the embryo. By comparing these proteins to the transcripts described by Lécuyer *et al.* and De Renzis *et al.* as being zygotic, we found that the agreement between the datasets is rather low. In FIGURE 3, a Venn diagram shows the identified zygotic products from the three datasets. Only 14 proteins were found to be present in all three experiments whereas 39 and 31 proteins were found overlapping with the De Renzis and the Lécuyer dataset, respectively. Interestingly, 8 proteins in the De Renzis dataset and 4 in the Lécuyer dataset defined as zygotic (indicated in FIGURE 3) were found to be significantly down-regulated in our dataset, suggesting that classifying both proteins and transcripts as being zygotic products is in some cases highly ambiguous.



**Functional classification of zygotic and maternal proteins •** Both zygotic and maternal proteins were clustered in functional groups based on gene ontology (GO) classification [25]. The list of all enriched GO terms (EASE score below 0.05) can be found in SUPPLEMENTAL TABLE 5 whereas the most prominent GO classes for both zygotic and maternal proteins are shown in TABLE 1. Classification of zygotic proteins by biological process led to a great number of enriched GO-terms, as might be expected from the dramatic events that occur in the embryo during this period of development. Further interpretation of these results

**Table 1. Selected set of significant gene ontology categories for both zygotic and maternal proteins that were functionally annotated by biological process using David (http://david.abcc.ncifcrf.gov/).**

| | Category[a] | EASE-score[b] |
|---|---|---|
| Zygotic | Cell organization and biogenesis | $1.84 \cdot 10^{-30}$ |
| | RNA metabolism | $1.35 \cdot 10^{-20}$ |
| | Establishment of cellular localization | $5.15 \cdot 10^{-20}$ |
| | Morphogenesis of an epithelium | $1.74 \cdot 10^{-08}$ |
| | Regulation of signal transduction | $6.63 \cdot 10^{-08}$ |
| | Morphogenesis of embryonic epithelium | $7.12 \cdot 10^{-08}$ |
| | Cell differentiation | $3.75 \cdot 10^{-06}$ |
| Maternal | Macromolecule biosynthesis | $6.63 \cdot 10^{-05}$ |
| | Oogenesis | $1.31 \cdot 10^{-04}$ |
| | Female gamete generation | 2.04 • 10-04 |
| | Organelle organization and biogenesis | 3.25 • 10-04 |
| | Cellular biosynthesis | 6.58 • 10-04 |
| | Regulation of translation | 8.10 • 10-04 |

[a] Biological process category [b] EASE-score, a modified Fisher Exact P-Value, for protein-enrichment analysis

is not straightforward given the large number of proteins in some classes and the widely differing abundance ratios of individual proteins in each class.

Some of the most dramatically down-regulated (maternal) proteins are those involved in translational control of RNAs and are listed in TABLE 2. These include Oskar (OSK), a key player in germline as well as abdominal development, found to be more than 9-fold down-regulated. Other examples include Bicaudal C, Vasa and Cup which are involved in translational repression of *oskar* mRNA and were down-regulated in a similar fashion as OSK itself ($\log_2$ ratios of -4.8, -1.6 and -5.4, respectively). In addition, Cup can be recruited by Smaug (SMG, $\log_2$ ratio of -2.8) to translationally repress *nanos* mRNA.

**Table 2. List of maternal proteins that are involved in translational regulation. Subset of down-regulated (maternal) proteins that are involved in mediating Oskar translational regulation.**

| Protein | ID[a] | Protein ratio[b] ($\log_2$ ± SD) |
|---|---|---|
| Oskar | OSKA_DROME | -3.62 ± 0.35 |
| Bicaudal-C | BICC_DROME | -4.77 ± 0.09 |
| Cup | CUP_DROME | -5.31 ± 0.21 |
| Smaug | SMG_DROME | -2.79 ± 0.31 |
| Eukaryotic initiation factor 4E | IF4E_DROME | -1.41 ± 0.18 |
| Eukaryotic initiation factor 4G | 061380_DROME | -0.29 ± 0.01 |
| Trailer hitch | Q9VTZ0_DROME | -5.08 ± 0.29 |
| Tudor | TUD_DROME | -2.56 ± 0.25 |
| Vasa | VASA_DROME | -1.63 ± 0.19 |
| Maternal expression at 31B | ME31_DROME | -3.66 ± 0.18 |
| Ypsilon schachtel | Q95RE4_DROME | -2.35 ± 0.67 |

[a] UniprotKB/Swiss-Prot entry name [b] Average ratio from two biological independent experiments

**Post-transcriptional regulation •** From our findings above it is evident that a large proportion of the proteins identified here change in abundance, due to both translation and degradation. Some of these proteins are involved in translationally repressing, localizing and stabilizing mRNAs. Our quantitative proteomic approach provides an opportunity to link expression profiles of proteins to their cognate transcript levels, which could reveal mechanisms of post-transcriptional regulation. To directly correlate protein and transcript expression levels, microarray experiments were conducted on identical samples (including isotope labels) used for proteomic analyses. We could obtain mRNA expression levels for 4 657 transcripts that could be correlated to 1 548 entries in the quantitative proteomic dataset. Direct comparison of mRNA and protein expression levels is shown in FIGURE 4. The Spearman's rank correlation coefficient between these datasets is weakly positive (0.39),

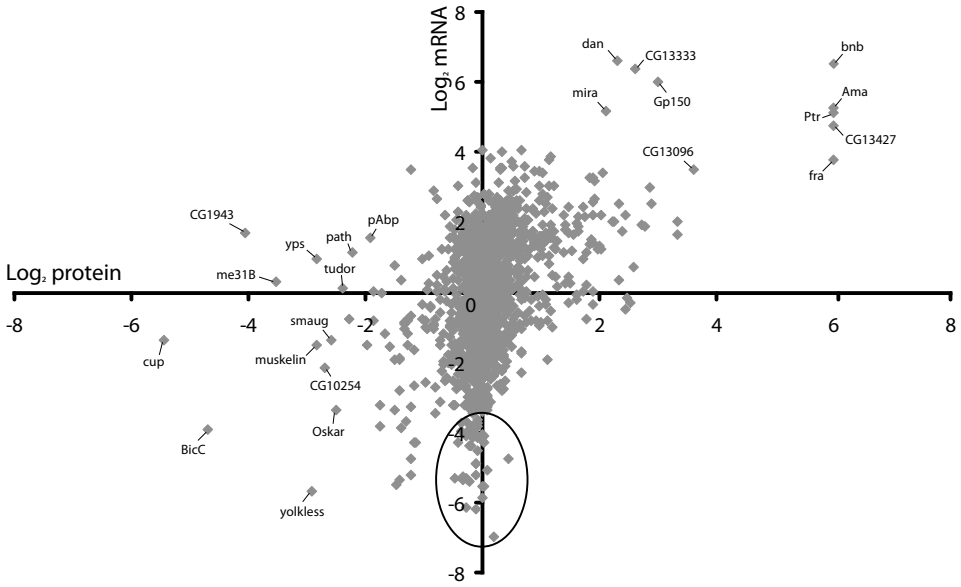


Figure 4. Correlation of protein and transcript expression. Scatter plot representing changes in expression levels of proteins and transcripts during the MZT. The encircled region contains data points where transcript levels show decreased abundance, while protein levels do not change significantly. Of these 41 proteins, 37 have 'catalytic activity', and are listed in Supplemental Table 9.

indicating that mRNA levels are poor indicators for protein expression levels. While the majority of the data points are clustered around the centre of the plot, a number of more extreme values permit some interesting observations. For instance, a cluster of data points can be found on the y-axis (encircled in FIGURE 4) in the region where transcript levels show decreased abundance, while protein levels do not change appreciably. Apparently, transcripts are degraded while proteins remain stable even though transcription (and translation) ceases. Strikingly, 37 out of the 41 proteins fall into the GO biological process category 'catalytic activity' (SUPPLEMENTAL TABLE 6). Apparently, these are mainly enzymes whose expression is regulated by a common mechanism. A possible explanation could be that these proteins are stabilized by an unknown mechanism to prevent degradation. Alternatively, this could be supplemented by a mechanism regulating enzyme activity. It is conceivable that this would provide a mechanism to engage or fine-tune enzymatic activity in metabolic processes without the need to newly synthesize proteins.

Another region in FIGURE 4 comprises the part where protein levels are down-regulated while mRNA levels do not change appreciably, or are even up-regulated. Strikingly, several prominent maternal proteins fall in this group, such as Smaug and maternal expres-

sion at 31B (ME31B). Moreover, a number of proteins known to interact with these proteins are observed here (Cup, Ypsilon Schachtel, Poly(A) Binding Protein, Bicaudal C and Trailer Hitch) as well as Tudor and Pathetic, indicated in FIGURE 4. This pattern could be the result of post-transcriptional or transcriptional silencing, where translation is blocked and protein levels drop, while transcript levels remain constant or even increase. Since this specific correlation of transcript and protein levels seem to be restricted to this functional class of proteins, it is tempting to speculate that a dedicated, functionally relevant, but as yet unknown mechanism is involved. Along with above described proteins, another protein (CG1943) is located in the same region. Although this protein has weak homology to HN1-like proteins, no function is known for this protein, but its expression is likely to be regulated post-transcriptionally.

**Post-translational protein processing •** During the quantitiation and validation process some proteins attracted attention because of their high coefficient of variation (CV) compared to the average CV of 18% for all proteins. A high deviation is indicative of large differences in individual peptide ratios within a single protein. We examined quantitation details of proteins where the CV was higher than 40%, establishing whether the source of this error was in the quantitation process itself or caused by biological phenomena, like post-translational processing.

Examples of proteins with high variation included the eight proteins that were off-on regulated (BNB, SCA, PTR, HRG, AMA, FRA, CG13427 and NRT). The intensity of the labeled peptides of these proteins was at noise level, resulting in variation of abundance ratios and thus a high standard deviation (SD) for the protein.

A more interesting example is the protein Cysteine Proteinase-1 (CP1), a member of the papain family of cysteine proteinases which is involved in the process of intra- and extra-cellular protein degradation and turnover ($\log_2$ ratio -0.84, SD 1.28). It is known that these proteases are activated by removal of the N-terminal propeptide region [26]. Indeed, two N-terminal peptides from CP1 displayed a more dramatic down-regulated pattern ($\log_2$ ratio -3.01) compared to the other peptides (-0.22). This result was validated by the swap experiment where the same peptides showed the same regulation. It can therefore be concluded that CP1 is activated during the MZT.

Another example is the Polyadenylate Binding Protein PABP ($\log_2$ ratio -1.93, SD 2.05). A closer look to the sequence of this protein and the individual peptides that were quantified (FIGURE 5) revealed that specifically C-terminal peptides exhibited a ratio deviating strongly from the other peptides. The average $\log_2$ ratio of 13 identified peptides (indi-

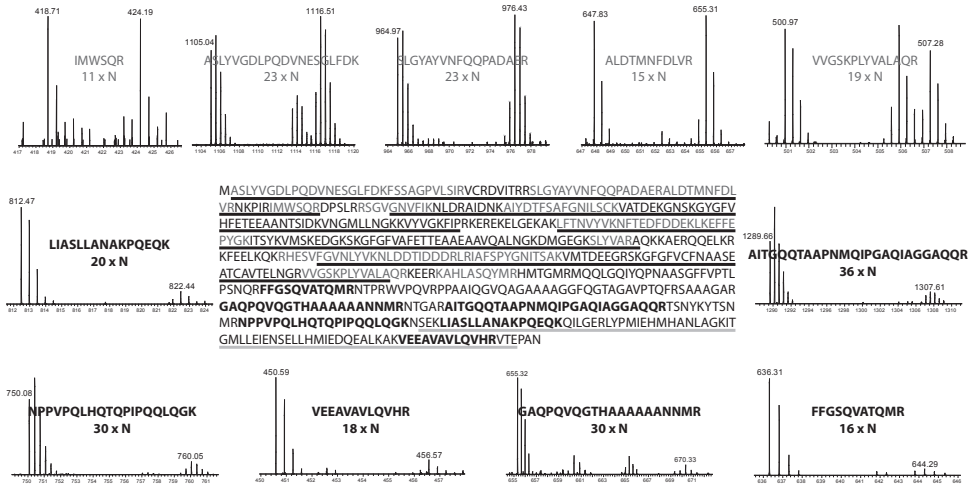


Figure 5. Post-translational processing of Poly(A) Binding Protein (PABP). Identified peptides of PABP along with spectra indicating their abundance before and after MZT. Peptides originating from the N-terminal part of the protein (shown in red) show no change, apparent from equal intensities of the unlabeled and labeled peptides. This is in contrast to peptides originating from the C-terminal part of the protein (blue). The intensity of the unlabeled peptides is dramatically less compared to that of the labeled peptides, indicating truncation of this part of the protein during MZT.

cated grey in FIG. 5) until amino acid residue 410 is -0.94 whereas the C-terminal peptides (four in total, bold black in FIG. 5) showed a $\log_2$ ratio of -5.17. This indicates that PABP is C-terminally processed. This is unlikely to be the result of aspecific proteolytic cleavage during sample handling since (*i*) this was the only protein in the entire dataset for which we observed this phenomenon and (*ii*) ratios were confirmed in the label-swap experiment ($\log_2$ ratio -0.82 for the N-term and -1.94 for the C-term peptides). Moreover, C-terminal truncation of PABP may serve a functional role. PABP contains four RNA recognition motifs located on the N-terminal part of the protein with the last domain ending at amino acid 362, black underlined in FIGURE 5. In addition, there is a polyadenylate motif at the C-terminal end of this protein (ranging from amino acid 552 to 629, grey underlined in FIG. 5). Through these motifs, PABP facilitates the formation of the 'closed loop' structure of the mRNA-protein complex, which is essential to stimulate translation. More specifically, the RNA recognition motifs of PABP interact with poly(A) tails of transcripts while the C-terminal domain is used to interact with factors regulating polyadenylation, deadenylation and translational activities [27]. It is this latter domain that is more down-regulated than the other part of the protein containing the RNA recognition motifs, indicating that the polyadenylate domain is eliminated by post-translational processing.

# V. Discussion

Since the notion that early embryonic development is determined to a large extent by maternal mRNA deposited in the oocyte, there has been great interest to investigate the relationship between gene expression and establishment of embryonic organization. In the first place this concerns the identity of the gene products involved in governing these processes, but extends to the mechanism whereby gene expression is controlled in space and time. It has been known for a long time that embryos contain a large number of maternally-derived mRNAs during the initial stages of development [28], while early proteomic studies showed that not all transcripts were translated, indicating that they must be controlled at the translational level [29]. Since the introduction of genomic and imaging techniques it has become possible to study this in far greater detail and now, for the first time, we complement these studies by a large-scale quantitative proteomic approach.

To successfully investigate the maternal-to-zygotic transition (MZT) in *Drosophila* embryos at the proteome level, we applied an approach that allows the identification and accurate quantitation of multiple proteins in a single experiment. This proteomics approach includes the stable-isotope labeling of all proteins in fruit flies *in vivo* by the incorporation of heavy nitrogen ($^{15}N$) [17]. Two-dimensional peptide separation, combined with high-accuracy mass spectrometry, resulted in the identification of 2 736 proteins with a false-positive rate of < 1%. Out of these proteins, 2 232 proteins were quantified. These data now allows to distinguish and to relatively quantify proteins that are expressed before and after the MZT. This thereby defines which proteins are of maternal origin and which ones are (also) zygotic. Especially the latter class of proteins is of primary interest since they define the earliest products of zygotic translation, initiating processes further downstream. In our dataset, these proteins are defined by their elevated expression after the MZT.

We classified a total of 490 proteins as being zygotic, of which 482 had maternal contribution. It cannot be excluded, however, that this list of zygotic proteins contains maternally derived proteins that are silenced and stabilized during the first few hours of development and are expressed after the MZT. The other 8 proteins were identified exclusively after the MZT, indicating that these are purely zygotic and have no maternal contribution. Previous genomic studies revealed a large number of zygotic genes that were either classified based on cellular localization (nuclear or cytoplasmic) [24] or based on the location on chromosome-ablated mutants compared to wild-type embryos [6]. Combining these studies, only 14 products were classified as zygotic in all three experiments and 39 respec-

tively 31 in the De Renzis and Lécuyer dataset (FIGURE 3). Several reasons could account for the small overlap related primarily to the fact that widely different techniques were used. Proteomic studies, as presented here, tend to be biased towards abundant proteins, unlike genomic experiments such as microarrays. Some proteins classified as zygotic by De Renzis (8) and Lécuyer (4) were found to be down-regulated in our study, suggesting maternal expression. Alternatively, the decrease in protein level could be the combined effect of zygotic translation and (stronger) degradation. Finally, it is important to note that no direct relation may be expected from the genomic studies [24, 6] and ours, since transcript and protein levels do not necessarily correlate (see FIG. 3).

Functional classification of both maternal and zygotic proteins led to a wide variety of functional categories (Table 1). A group of proteins enriched in early embryos and found in the category 'regulation of translation' were RNA binding proteins. An interesting example is Oskar, responsible for setting up body axes and abdominal development together with mediators that are involved in regulating silenced localization and expression of *Osk* mRNA [30]. These factors include Cup, Ypsilon Schachtel (YPS) and Maternal Expression at 31B (ME31B) all of which were identified in this study (TABLE 2). Once localized to and translated at the posterior pole, OSK has the ability to interact with the RNA binding domain of Smaug [31, 32] and this interaction prevents Smaug from binding *nanos* (*nos*) mRNA and subsequently deadenylation. This results in *nos* stability and translation at the posterior pole [33], whereas unlocalized and silenced *nos* in the bulk cytoplasm of the embryo is subjected to degradation [8, 34].

The translational regulators discussed above (YPS, ME31B, TRAL, CUP, EIF4E, PABP, and SMG) are among the most severely down-regulated proteins in our study, suggesting shut-down of translational silencing mechanism. Remarkably, their respective transcript levels all remain relatively constant (FIGURE 4). This observation may be explained by a post-transcriptional regulatory mechanism that stabilizes transcripts while inhibiting translation. This could be mediated by miRNAs, which bind to the 3' untranslated regions of specific mRNAs and inhibit their translation. Gene silencing by miRNAs is active in *Drosophila*, and is indicated in this study by the distinct up-regulation of Argonaute-1, which is critically involved in miRNA-mediated cleavage of target mRNAs. Nakahara *et al.* searched for miRNA targets in *Drosophila* oocytes by comparing wild-type oocytes to null *dicer-1* oocytes, in which the biogenesis of miRNAs is blocked [35]. Forty-one proteins were found (of which 22 identified) and 18 candidates were presented whose expression was elevated in the *dicer* mutant, suggesting they are targets of miRNAs. Of these targets, 15

were identified in our study as well, of which only ME31B was down-regulated. Although it cannot be excluded that apart from ME31B also other translational regulators are miRNA targets, there is no direct evidence that this is indeed the case.

In stress granules, Poly(A)-binding protein (PABP) is critically involved in mRNA translation and stabilization. This is achieved through domains in its N-terminal part that bind poly(A) tails of target transcripts, and a C-terminal motif that can interact with factors regulating translation initiation and termination, polyadenylation and deadenylation [27, 36]. By these specific binding properties, PABP bridges 3' poly(A) tails of target mRNAs with components of the translational initiation machinery (e.g. EIF4G) at the 5' end thereby forms so-called 'closed loop' structures that stimulate translation [27, 37]. It is interesting that we observed C-terminal cleavage forms of PABP, which eliminates the conserved C-terminal protein-interaction domain. Elimination of this domain would abolish interaction with, for instance, EIF4G and hence the bridge function, while it could still bind to poly(A) tails of mRNA. This would prevent translation, but also inhibit mRNA degradation by preventing interaction with the CCR4-POP2-NOT deadenylase complex and hence shortening of the poly(A) tail.

Apart from a functional role for truncated PABP, other questions that remain are how this product is formed, and whether cleavage is specific. Remarkably, proteolytic cleavage of PABP is not unprecedented since a number of viral proteases (poliovirus 2A and 3C proteases and coxsackievirus B3 2A protease) have been shown to specifically cleave PABP in the C-terminal domain [38, 39]. Interestingly, proteolytic cleavage was accompanied by loss of translational activity [38] suggesting that translation could be inhibited in *Drosophila* as well if a similar endogenous mechanism indeed exists. Although we cannot exactly pinpoint the site of cleavage, it is restricted to a well-defined region (between residues Arg398 and Phe411) based on quantitative data of individual peptides in PABP. It is striking to note that the viral proteases cleave exactly in this region [39], suggesting a large degree of conservation. We have searched for proteins homologous to the viral proteases in the *Drosophila* genome, but obvious candidates have not been found. All together, it remains speculative whether cleavage of PABP has a specific role in the expression profiles of SMG, ME31B, CUP, YPS, TUD and PATH (both at the level of transcript and protein, FIG. 4) or in the context of disintegration of granules or even developmental progression in general.

# VI. Conclusion

Our quantitative proteomic approach using *in vivo* labeling of fruit flies with stable isotopes combined with extensive analysis by LC-MS/MS has permitted the relative quantitation of thousands of proteins during early embryonic development. The method proved highly accurate and reproducible, and has revealed detailed information giving insight that cannot be obtained from genomic approaches. This clearly demonstrates the contribution of quantitative proteomics for our understanding of early fly development. Although we have focused our study on early fly development, the quantitative proteomic technique used is applicable to any other process where changes in protein expression are to be studied (development, environmental condition, mutation). Furthermore, the method for metabolic labeling should be easily adoptable to many fly labs since it requires only minor adaptations compared to routine protocols for growing flies. Therefore, we expect this approach to find broad application in fly developmental biology and beyond.



## Acknowledgements

We like to thank Bas van Breukelen (Utrecht University) for statistical analyses. This work was supported by the Netherlands Proteomics Centre (www.netherlandsproteomicscentre.nl).

## Supplemental data

The 3 supplemental figures and 7 supplemental tables can be found online as pdf files at https://bioinformatics.chem.uu.nl/thesis/joost.

# References


[1] Wieschaus E. and Gehring W., *Clonal analysis of primordial disc cells in the early embryo of Drosophila melanogaster*. Dev Biol, **1976**, *50*, 249-263.

[2] Zalokar M., *Autoradiographic study of protein and RNA formation during early development of Drosophila eggs*. Dev Biol, **1976**, *49*, 425-437.

[3] Anderson K.V. and Lengyel J.A., *Rates of synthesis of major classes of RNA in Drosophila embryos*. Dev Biol, **1979**, *70*, 217-231.

[4] Newport J. and Kirschner M., *A major developmental transition in early Xenopus embryos: I. characterization and timing of cellular changes at the midblastula stage*. Cell, **1982**, *30*, 675-686.

[5] Edgar B.A. and Schubiger G., *Parameters controlling transcriptional activation during early Drosophila development*. Cell, **1986**, *44*, 871-877.

[6] De Renzis S., Elemento O., Tavazoie S. and Wieschaus E.F., *Unmasking activation of the zygotic genome using chromosomal deletions in the Drosophila embryo*. PLoS Biol, **2007**, *5*, e117.

[7] Tadros W., Goldman A.L., Babak T., Menzies F., Vardy L., Orr-Weaver T., Hughes T.R., Westwood J.T., Smibert C.A. and Lipshitz H.D., *SMAUG is a major regulator of maternal mRNA destabilization in Drosophila and its translation is activated by the PAN GU kinase*. Dev Cell, **2007**, *12*, 143-155.

[8] Bashirullah A., Halsell S.R., Cooperstock R.L., Kloc M., Karaiskakis A., Fisher W.W., Fu W., Hamilton J.K., Etkin L.D. and Lipshitz H.D., *Joint action of two RNA degradation pathways controls the timing of maternal transcript elimination at the midblastula transition in Drosophila melanogaster*. Embo J, **1999**, *18*, 2610-2620.

[9] Kloc M. and Etkin L.D., *RNA localization mechanisms in oocytes*. J Cell Sci, **2005**, *118*, 269-282.

[10] St Johnston D., *Moving messages: the intracellular localization of mRNAs*. Nat Rev Mol Cell Biol, **2005**, *6*, 363-375.

[11] Eulalio A., Behm-Ansmant I. and Izaurralde E., *P bodies: at the crossroads of post-transcriptional pathways*. Nat Rev Mol Cell Biol, **2007**, *8*, 9-22.

[12] Parker R. and Sheth U., *P bodies and the control of mRNA translation and degradation*. Mol Cell, **2007**, *25*, 635-646.

[13] Gong L., Puri M., Unlu M., Young M., Robertson K., Viswanathan S., Krishnaswamy A., Dowd S.R. and Minden J.S., *Drosophila ventral furrow morphogenesis: a proteomic analysis*. Development, **2004**, *131*, 643-656.

[14] Brunner E., Ahrens C.H., Mohanty S., Baetschmann H., Loevenich S., Potthast F., Deutsch E.W., Panse C., de Lichtenberg U., Rinner O., et al., *A high-quality catalog of the Drosophila melanogaster proteome*. Nat Biotechnol, **2007**, *25*, 576-583.

[15] Zhai B., Villen J., Beausoleil S.A., Mintseris J. and Gygi S.P., *Phosphoproteome Analysis of Drosophila melanogaster Embryos*. J Proteome Res, **2008**, *7*, 1675-1682.

[16] Vardy L. and Orr-Weaver T.L., *Regulating translation of maternal messages: multiple repression mechanisms*. Trends Cell Biol, **2007**.

[17] Krijgsveld J., Ketting R.F., Mahmoudi T., Johansen J., Artal-Sanz M., Verrijzer C.P., Plasterk R.H. and Heck A.J., *Metabolic labeling of C. elegans and D. melanogaster for quantitative proteomics*. Nat Biotechnol, **2003**, *21*, 927-931.

[18] Rappsilber J., Ishihama Y. and Mann M., *Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics*. Anal Chem, **2003**, *75*, 663-670.

[19] Meiring H.D., van der Heeft E., ten Hove G.J. and de Jong A.P.J.M., *Nanoscale LC-MS$^{(n)}$: technical design and applications to peptide and protein analysis*. J Sep Sci, **2002**, *25*, 557-568.

[20] Zhang B., Chambers M.C. and Tabb D.L., *Proteomic Parsimony through Bipartite Graph Analysis Improves Accuracy and Transparency*. J. Proteome Res., **2007**, *6*, 3549-3557.

[21] Mortensen P., The MSQuant home page. http://msquant.sourceforge.net/, 2006.

[22] Gouw J.W., Tops B.B.J., Mortensen P., Heck A.J.R. and Krijgsveld J., *Optimizing identification and quantitation of $^{15}$N-labeled proteins in comparative proteomics*. Anal. Chem., **2008**, *80*, 7796-7803.

[23] Moshkin Y.M., Mohrmann L., van Ijcken W.F. and Verrijzer C.P., *Functional differentiation of SWI/SNF remodelers in transcription and cell cycle control*. Mol Cell Biol, **2007**, *27*, 651-661.

[24] Lecuyer E., Yoshida H., Parthasarathy N., Alm C., Babak T., Cerovina T., Hughes T.R., Tomancak P. and Krause H.M., *Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function*. Cell, **2007**, *131*, 174-187.

[25] Hosack D.A., Dennis G., Jr., Sherman B.T., Lane H.C. and Lempicki R.A., *Identifying biological themes within lists of genes with EASE*. Genome Biol, **2003**, *4*, R70.

[26] Yamamoto Y., Kurata M., Watabe S., Murakami R. and Takahashi S.Y., *Novel cysteine proteinase inhibitors homologous to the proregions of cysteine proteinases*. Curr Protein Pept Sci, **2002**, *3*, 231-238.

[27] Mangus D.A., Evans M.C. and Jacobson A., *Poly(A)-binding proteins: multifunctional scaffolds for the post-transcriptional control of gene expression*. Genome Biol, **2003**, *4*, 223.

[28] Arthur C.G., Weide C.M., Vincent W.S., 3rd and Goldstein E.S., *mRNA sequence diversity during early embryogenesis in Drosophila melanogaster*. Exp Cell Res, **1979**, *121*, 87-94.

[29] Trumbly R.J. and Jarry B., *Stage-specific protein synthesis during early embryogenesis in Drosophila melanogaster*. Embo J, **1983**, *2*, 1281-1290.

[30] Nakamura A., Sato K. and Hanyu-Nakamura K., *Drosophila cup is an eIF4E binding protein that associates with Bruno and regulates oskar mRNA translation in oogenesis*. Dev Cell, **2004**, *6*, 69-78.

[31] Dahanukar A., Walker J.A. and Wharton R.P., *Smaug, a novel RNA-binding protein that operates a translational switch in Drosophila*. Mol Cell, **1999**, *4*, 209-218.

[32] Smibert C.A., Lie Y.S., Shillinglaw W., Henzel W.J. and Macdonald P.M., *Smaug, a novel and conserved protein, contributes to repression of nanos mRNA translation in vitro*. Rna, **1999**, *5*, 1535-1547.

[33] Zaessinger S., Busseau I. and Simonelig M., *Oskar allows nanos mRNA translation in Drosophila embryos by preventing its deadenylation by Smaug/CCR4*. Development, **2006**, *133*, 4573-4583.

[34] Semotok J.L., Cooperstock R.L., Pinder B.D., Vari H.K., Lipshitz H.D. and Smibert C.A., *Smaug recruits the CCR4/POP2/NOT deadenylase complex to trigger maternal transcript localization in the early Drosophila embryo*. Curr Biol, **2005**, *15*, 284-294.

[35] Nakahara K., Kim K., Sciulli C., Dowd S.R., Minden J.S. and Carthew R.W., *Targets of microRNA regulation in the Drosophila oocyte proteome*. Proc Natl Acad Sci U S A, **2005**, *102*, 12023-12028.

[36] Jacobson A. and Favreau M., *Possible involvement of poly(A) in protein synthesis*. Nucleic Acids Res, **1983**, *11*, 6353-6368.

[37] de Moor C.H., Meijer H. and Lissenden S., *Mechanisms of translational control by the 3' UTR in development and differentiation*. Semin Cell Dev Biol, **2005**, *16*, 49-58.

[38] Joachims M., Van Breugel P.C. and Lloyd R.E., *Cleavage of poly(A)-binding protein by enterovirus proteases concurrent with inhibition of translation in vitro*. J Virol, **1999**, *73*, 718-727.

[39] Kuyumcu-Martinez N.M., Joachims M. and Lloyd R.E., *Efficient cleavage of ribosome-associated poly(A)-binding protein by enterovirus 3C protease*. J Virol, **2002**, *76*, 2062-2074.

# CHAPTER 4

# Optimizing identification and quantitation of $^{15}$N-labeled proteins in comparative proteomics

Joost W. Gouw[†], Bastiaan B.J. Tops[†], Peter Mortensen[‡], Albert J.R. Heck[†] and Jeroen Krijgsveld[†]

[†]Biomolecular Mass Spectrometry and Proteomics Group, Bijvoet Center for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands, [‡]Center for Experimental Bioinformatics, University of Southern Denmark, Odense, Denmark

# I. Abstract

Comparative proteomics has emerged as a powerful approach to determine differences in protein abundance between biological samples. The introduction of stable-isotopes as internal standards especially paved the road for quantitative proteomics for comprehensive approaches to accurately determine protein dynamics. Metabolic labeling with $^{15}N$ isotopes is applied to an increasing number of organisms, including *Drosophila*, *C. elegans,* and rats. However, $^{15}N$-enrichment is often suboptimal (< 98%) which may hamper identification and quantitation of proteins. Here, we systematically investigated two independent $^{15}N$-labeled data sets to explore the influence of heavy nitrogen enrichment on the number of identifications as well as on the error in protein quantitation. We show that specifically larger $^{15}N$-labeled peptides are under-represented when compared to their $^{14}N$ counterparts and propose a correction method, which significantly increases the number of identifications. In addition, we developed a method that corrects for inaccurate peptide ratios introduced by incomplete $^{15}N$ enrichment. This results in improved accuracy and precision of protein quantitation. Altogether, this study provides insight into the process of protein identification and quantitation, and the methods described here can be used to improve both qualitative and quantitative data obtained by labeling with heavy nitrogen with enrichment less than 100%.

# II. Introduction

The identification of peptides and proteins is an important first step in proteomic experiments. Typically, the proteomic composition of biological samples is determined by mass spectrometry preceded by one or more chromatographic steps depending on sample complexity. In a more extended setup, samples with different origins or treatments can be analyzed in parallel to get insight into the identity of the proteins that may underlay the difference between them, such as in biomarker studies. Although in some cases such a qualitative approach may be fruitful, it usually ignores the dynamics of protein expression and the fact that biological effects are caused by gradual changes in protein abundance rather than their mere presence or absence. Therefore, extensive efforts have been made over the past few years to add a quantitative component to comparative proteomics. Although protein abundance can be estimated from spectral intensities [1, 2] or peptide counts [3] under some experimental conditions, in general, approaches make use of stable isotope labels that are incorporated in proteins of one sample, which can then be mixed with an unlabeled sample to serve as an internal control. The use of internal standards in mass spectrometry-based quantitative proteomics is nowadays common practice for global relative quantitation of proteins. Since the ionization efficiency of the analyte and internal standard are identical, their spectral intensities accurately reflect the relative differences in protein amount.



Different methods of introducing these internal standards exist, ranging from chemical tagging of proteins or peptides using isotope-containing reagents to the enrichment of stable isotope labeled atoms into proteins *in vivo*. Stable isotope labeling that is achieved *in vitro* such as the chemical labeling procedures isotope coded affinity tags (ICAT) [4] and isobaric tags for relative and absolute quantitation (iTRAQ) [5] as well as proteolytic $^{18}O$ labeling [6] can be applied to virtually any type of sample. However, these types of labeling suffer from higher levels of variation compared to *in vivo* labeling since the samples can only be mixed after labeling which can be even after digestion as is the case for $^{18}O$ labeling. Nevertheless, reasonably low levels of variation (median coefficient of variation of 18.6%) are reported by employing ICAT on *Escherichia coli* cells [7]. A major drawback of the ICAT strategy is that not all peptides can be used for quantitation due to the cysteine-specific reagent that is used to label proteins.

An alternative approach which does not have this disadvantage is labeling *in vivo*. Several types of this labeling procedure have been developed of which stable isotope label-

ing by amino acids in cell culture (SILAC) is probably the most well-known [8]. Although SILAC labels proteins uniformly with a preferred labeled amino acid (usually Lys and/or Arg), its application is limited to cells that can be grown in isotopically enriched media. Alternatives that circumvent such drawbacks have been explored using stable isotope-labeled atoms such as nitrogen and carbon ($^{15}N$ and $^{13}C$) that can be metabolically incorporated *in vivo.* Global coding of proteomes using $^{15}N$ has been reported for cell cultures [9, 10], plants [11, 12], microorganisms [10, 13, 14], fruit flies [15, 16], and *C. elegans* [15, 17]. Wu and co-workers even extended this labeling approach to mammals by labeling rats with enrichments ranging from 74% to 92% depending of the type of tissue [18]. Using an improved labeling strategy, they increased the enrichment to 94% throughout all tissues of the rat [19]. Previously, we metabolically labeled fruit flies with an average enrichment level of 98.2% heavy nitrogen [16]. At this point we would like to distinguish two separate, yet equally important, phenomena that contribute independently to the final percentage of the stable isotope in proteins. First, the purity of the stable isotope that is obtained from the supplier (e.g., 98% $^{15}N$) and, second, the degree of incorporation of that stable isotope into proteins. Throughout this article we use the term 'enrichment' to indicate the final percentage of $^{15}N$ in peptides and proteins.

Although labeling with $^{15}N$ can be very efficient, even enrichment levels as high as 98% cause some challenges for proper peak selection, peptide identification, and quantitation. This is caused primarily by isotope peaks that appear in front of the monoisotopic peak, as is illustrated in FIGURE 1. These are isotopologues containing a progressive num-

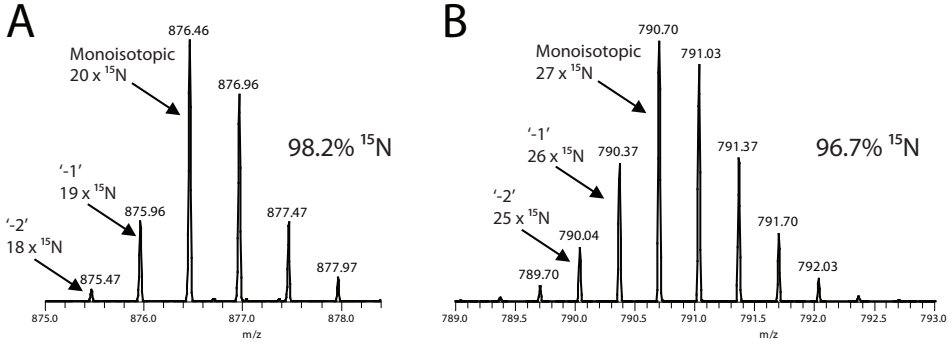


Figure 1. Isotope distributions of two typical $^{15}N$-labeled peptides. Peptide (A) contains 20 heavy nitrogen atoms and has a $^{15}N$ enrichment of 98.2% whereas peptide (B) contains 27 heavy nitrogen ($^{15}N$) atoms and is 96.7% enriched in $^{15}N$. The isotopes '-1' and '-2' originate from the incomplete enriched nitrogen and contain 1 and 2 unlabeled nitrogen ($^{14}N$) atoms, respectively and are already present prominently, even for these average-sized peptides and high enrichment levels.

ber of $^{14}N$ atoms. Even for average-sized peptides and high enrichment levels, the phenomenon is prominent, as shown in FIG. 1A (20 N atoms, 98.2% $^{15}N$). The intensities of these peaks increase with lower enrichment levels and with larger numbers of nitrogen atoms in the peptide. In FIGURE 1B is shown a peptide that contains more nitrogen atoms and also exhibits a lower enrichment level (96.7%) compared to the peptide in FIGURE 1A, resulting in not only more intense but also more satellite peaks.

The appearance of additional isotopologues potentially affects both peptide identification and quantitation. If such a satellite peak is mistakenly considered as the monoisotopic peak, proper mass assignment will fail resulting in a false identification or no identification altogether. The quantitation process may be compromised if only the monoisotopic peak is used to derive the ratio between labeled and unlabeled peptides. Neglecting the intensity of the satellite peaks will affect the accuracy of the quantitation. Several ways of eliminating this effect have been described. For instance, by using the same internal standard in two samples, any systematic error (including the error from incomplete labeling) can be canceled since both samples contain the same systematic errors [9, 18, 20]. Another way is to integrate the areas under the ion chromatograms of the labeled and unlabeled isotopes *m/z* range (i.e., the complete isotope distribution), like the correlation algorithm RelEx [20] does. Nonetheless, the chance of isobaric interference increases dramatically, reducing the number of usable peptides for quantitation. In addition, RelEx can also quantitate on the most intense isotope *m/z* only and computes automatically an individual correction factor using the base isotope peak in the entire distribution.

Here we have investigated the influence of $^{15}N$ enrichment on the number of identifications as well as on the error in protein quantitation. We show that upon a simple correction of the precursor mass, the number of identifications can be increased dramatically. In addition, we designed a method that improves the quality of protein quantitation. This results in more accurate quantitative data among peptides as well as more precise protein ratios. The methods described here can be applied to any type of labeling experiment with varying degrees of nitrogen enrichment. These methods significantly improve quantitative results of highly enriched samples (> 98%), and are necessary to obtain any meaningful results if enrichment is lower.

## III. Materials and methods

**Sample Preparation** • *D. melanogaster*. Labeling, embryo collection, and sample preparation was performed as described previously [16]. Briefly, equal amounts of labeled and unlabeled embryos were mixed and lysed in 8 M urea and 50 mM ammonium bicarbonate. Cellular debris was pelleted by centrifugation at 20 000 g for 20 min. Prior to digestion, proteins were reduced with 1 mM DTT and alkylated with 2 mM iodoacetamide. The mixture was diluted 4-fold to 2 M urea using 250 µL of 50 mM ammonium bicarbonate and 50 µL of trypsin solution, 0.1 mg/mL, and incubated overnight at 37 °C.

*C. elegans*. Strains used are the canonical Bristol N2 strain and VC576 (*mir-1-(gk276)*, obtained from the *Caenorhabditis* Genetics Center). Strains were cultured for either $^{14}N$ or $^{15}N$ enrichment. Unlabeled worms were cultured on NGM plates, containing 1% agarose instead of agar, seeded with OP50 bacteria grown in regular LB. Strains grown for $^{15}N$ enrichment were grown on plates, containing 1% agarose, 0.3% (w/v) NaCl and 25% (v/v) Spectra9-N (Spectra Stable Isotopes), seeded with OP50 bacteria grown in Spectra9-N medium. Extracts were prepared as follows. Synchronized L3/L4 animals were harvested in M9 buffer on ice. The resulting pellet was subsequently dried by vacuum centrifugation and extracted in 8 M urea. After determining the protein concentrations, extracts were mixed in a 1:1 ratio (N2 vs $^{15}N$ VC576).

**1D-SDS-PAGE** • Worm protein mixture was resolved by an SDS-PAGE gel (10%) and Coomassie-stained. The gel lane was cut into 24 pieces and subjected to in-gel reduction, alkylation, and tryptic digestion.

**Strong Cation Exchange** • Strong cation exchange of the fly sample was performed using a Zorbax BioSCX-Series II column (0.8 mm i.d. × 50 mm length, 3.5 µm), a FAMOS autosampler (LC-packing, Amsterdam, The Netherlands), a LC-9A binary pump, and a SPD-6A UV-detector equipped with a micro UV-cell (V = 0.6 µL) (Shimadzu, Tokyo, Japan). Prior to SCX chromatography, protein digests were desalted using a small plug of $C_{18}$ material (3 M Empore $C_{18}$ extraction disk) packed into a GELoader tip (Eppendorf) similar to what has been previously described [21], onto which ~10 µL of Aqua $C_{18}$ (5 µm, 200 Å) material was placed. The eluate was dried completely and subsequently reconstituted in 20% acetonitrile and 0.05% formic acid. After injection, a linear gradient of 1% $min^{-1}$ solvent B (500 mM KCl in 20% acetonitrile and 0.05% formic acid, pH 3.0) at a flow rate of 50 µL/min was started. A total of 24 SCX fractions (1 min each, i.e., 50 µL elution volume)

were manually collected and dried in a vacuum centrifuge.

**Nanoflow-HPLC-MS** • Dried SCX residues were reconstituted in 50 µL of 0.1 M acetic acid of which 5 µL was used for analysis. All samples were analyzed by nanoflow liquid chromatography using an Agilent 1100 HPLC system (Agilent Technologies) coupled on-line to a 7-Tesla LTQ-FTICR mass spectrometer (Thermo Electron). The liquid chromatography part of the system was operated in a setup essentially as described previously [22]. Aqua $C_{18}$, 5 µm, (Phenomenex) resin was used for the trap column, and ReproSil-Pur $C_{18}$-AQ, 3 µm, (Dr. Maisch GmbH) resin was used for the analytical column. Peptides were trapped at 5 µL/min in 100% solvent A (0.1 M acetic acid in water) on a 2 cm trap column (100 µm i.d., packed in-house) and eluted to a 40 cm analytical column (50 µm i.d., packed in-house) at ~100 nL/min in a 150 min gradient from 10 to 40% solvent B (0.1 M acetic acid in 8/2 (v/v) acetonitrile/water). The eluent was sprayed via standard coated emitter tips (New Objective), butt-connected to the analytical column. The mass spectrometer was operated in the data-dependent mode, automatically switching between MS and MS/MS. Full scan MS spectra (from *m/z* 300 to 1 500) were acquired in the FTICR with a resolution of 100 000 at *m/z* 400 after accumulation to a target value of 500 000. The three most intense ions at a threshold above 5000 were selected for collision-induced fragmentation in the linear ion trap at a normalized collision energy of 35% after accumulation to a target value of 15 000.

**Peptide Identification and Quantitation** • All $MS^2$ spectra were converted to single dta files using Bioworks Browser 3.1 (Thermo) and merged into a Mascot generic format file which was searched using an in-house licensed Mascot v2.2.04 search engine (Matrix Science) against a *D. melanogaster* database (downloaded 2008-3-15; containing 22 785 sequences) or *C. elegans* database (downloaded 2007-4-9; containing 22 622 sequences). In both cases carbamidomethyl cysteine was set as a fixed modification and oxidized methionine as a variable modification. The quantitation field was set to '15N Metabolic' to identify both labeled and unlabeled peptides. Trypsin was specified as the proteolytic enzyme, and one missed cleavage was allowed. The mass tolerance of the precursor ion was set to 15 ppm and that of fragment ions was set to 0.6 Da. Peptide and protein quantitation was performed using the open source program MSQuant [23]. Briefly, peptide ratios were obtained by calculating the extracted ion chromatograms (XIC) of the unlabeled and labeled forms of the peptide using the monoisotopic peaks only. The total XIC for each of the peptide forms was obtained by summing the XIC in consecutive MS cycles for the duration

of their respective LC-MS peaks in the total ion chromatogram using FTMS scans. These total XICs were then used to compute the peptide ratio.

**Determination of the $^{15}$N enrichment level** • All MS scans of a $^{15}$N-labeled peptide were combined, and the resulting isotope distribution was manually correlated against theoretical isotope patterns varying in degree of $^{15}$N enrichment. In this way, the enrichment of 20 randomly selected peptides of each data set were determined. The isotope enrichment level of each data set was determined by averaging the 20 experimental enrichments (data not shown). This procedure can also be automated by using a software package called The Atomizer [24].

**Correction of precursor mass and quantitative results** • All queries in the Mascot generic format file were precursor mass corrected (+1 Da) in case if the wrong isotope was selected (i.e. '-1' isotope in FIGURE 1) using an in-house written Perl script and searched again using similar Mascot settings (see above). Correction of quantitative data was accomplished by summing the intensities of the isotopes originating from the incomplete $^{15}$N-enrichment (i.e., '-1' and '-2' isotopes in FIGURE 1). This was done by calculating the theoretical isotope pattern using natural abundances of elements as well as calculating the pattern based on the $^{15}$N enrichment level. Both theoretical isotope patterns as well as the monoisotopic intensity obtained from MSQuant were used to calculate intensities of the isotopologues originating from the impurity and their $^{13}$C isotopologues. These intensities were then used to correct the monoisotopic intensity for the labeled peptide. An in-house written Perl script was used to correct quantitative data and is available from the authors upon request.

# IV. Results and discussion

**Origin of samples •** To establish the enrichment level of $^{15}$N-labeled proteomes, we systematically investigated qualitative and quantitative data that was obtained from the analyses of two independent experiments. In the first experiment, two different time points in fly development were investigated [16]. Fruit flies were metabolically labeled by feeding them with $^{15}$N-labeled yeast leading to an average enrichment level of 98.2 ± 0.2% $^{15}$N. In the second labeling experiment, the nematode *C. elegans* was labeled by feeding them with $^{15}$N-labeled bacteria resulting in an average enrichment level of 96.7 ± 0.3%. In both experiments, 20 randomly selected peptides have been used to determine the average enrichment level. Several methods exist to determine the enrichment level [24, 25], and we used a similar method as reported by MacCoss *et al.* [24]. Although there is a small variation in the average enrichments, we expect the enrichments to be homogeneous since the *Drosophila* embryos are essentially single cells with little compartments and organelles (excluding tissue-specific protein turnover), and moreover, the flies have been kept on $^{15}$N for an extended period (at least two generations). Furthermore, the *C. elegans* nematodes used in these experiments were grown on labeled medium for five generations ensuring complete labeling. These samples were used to investigate and compare various aspects in qualitative and quantitative analysis of samples varying in the degree of $^{15}$N enrichment.

**Qualitative data •** To follow the process from detection in the mass spectrometer to identification in a search engine we started to investigate raw data obtained from the *Drosophila* experiment. Ten labeled peptides increasing in mass (1 500-3 000 Da), were randomly selected (TABLE 1). The first step in the identification process of a data-dependent analysis is the selection of a precursor for fragmentation. Because of the additional isotopologues in $^{15}$N-labeled peptides (FIGURE 1), this could potentially be problematic. If these isotopologues increase in intensity, the chance that they are mistaken for the monoisotopic peak and are selected for fragmentation increases. The masses of fragment ions, however, are not dramatically affected since the isolation window is usually large enough (two mass-to-charge units) to also include the monoisotopic peak for fragmentation. In this experiment, it turned out that in six cases the mass-to-charge ratio ($m/z$) selected for fragmentation did not correspond to the peptide's monoisotopic $m/z$ (TABLE 1). In fact, the $m/z$ selected for fragmentation for the first four peptides (1-4) was too high and corresponded to the $m/z$ of one of the $^{13}$C-isotopologues of the peptide. For peptides 7 and 8, the $m/z$ was lower than expected corresponding to isotopologues originating from not fully enriched heavy

nitrogen (i.e., '-1' and '-2' isotopes containing 1 and 2 $^{14}N$, respectively in FIGURE 1). As a result, peptides 1-4, 7 and 8 were not identified in a database search due to their incorrect precursor *m/z*.

**Table 1. Monoisotopic, Fragmented and dta Mass-to-Charge Ratio for 10 $^{15}N$-Labeled Peptides Originating from SCX Fraction 16**

| peptide no. | sequence[a] | monoisotopic *m/z*[b] | fragmented *m/z*[c] | Δ*m/z*[d] | *m/z* in dta file[e] | Δ*m/z*[d] |
|---|---|---|---|---|---|---|
| 1 | K.ILDEEVASLHLEK.L | 504.59 | 504.95 | 0.36 | 504.59 | 0 |
| 2 | R.LLQDLFNGKELNK.S | 517.27 | 517.94 | 0.67 | 517.27 | 0 |
| 3 | R.ELISNASDALDKIR.Y | 521.93 | 522.26 | 0.33 | 521.93 | 0 |
| 4 | K.DPNTHLIFLPTLLR.W | 557.30 | 557.63 | 0.33 | 557.30 | 0 |
| 5 | R.VTGGLHLINPSELTEK.A | 576.63 | 576.63 | 0 | 576.63 | 0 |
| 6 | K.GPVLPIKNEQPAVVEFR.E | 639.33 | 639.33 | 0 | 639.33 | 0 |
| 7 | K.ISADYDVLLDKEGISLR.E* | 643.32 | 642.99 | -0.33 | 642.99 | -0.33 |
| 8 | R.LILEFQPESHESNVLVR.S* | 678.67 | 678.34 | -0.33 | 678.34 | -0.33 |
| 9 | K.SSRPALNGVPAQEGEPFGDEALTFTR.E* | 927.43 | 927.43 | 0 | 927.09 | -0.34 |
| 10 | K.VLWDSEGSAAPGEIEIFPQYHASPFSR.L* | 1008.79 | 1008.79 | 0 | 1008.46 | -0.33 |

[a] Peptide sequence obtained from Mascot. Peptides denoted with * were identified after manually correcting the monoisotopic mass. [b] Monoisotopic mass-to-charge ratio, the charge was 3 for all the peptides in this table. [c] Mass-to-charge ratio that was selected and fragmented. [d] Mass-to-charge ratio difference compared to the monoisotopic mass-to-charge ratio. [e] Mass-to-charge ratio that was written to the output file (.dta).

The origin of the mass error could be related to the fact that the mass spectrometer used in these experiments (LTQ-FTICR) uses a fast, low-resolution scan to determine which precursors should be fragmented. Moreover, a precursor is already detected in the beginning of its elution profile where the signal-to-noise ratio is relatively low making it particularly difficult to determine the correct *m/z*. However, in the next step of the identification process, relevant information for each fragmentation event, including precursor *m/z*, is exported to data files (.dta) to facilitate database searching. After investigating each of the files belonging to the 10 prototype peptides, the precursor *m/z* was different from the fragmented *m/z* for peptides 1-4 and 9-10. The first four peptide masses, which were initially one or two *m/z* units too high, corresponded after .dta creation to the peptide's monoisotopic *m/z*, but the last four peptides were one mass unit too low (TABLE 1). This leads to two observations. First, it seems that the postanalysis software (Bioworks Browser from Thermo) reassesses the *m/z* selected for fragmentation but apparently not always correctly. It should be noted that this could be a specific problem related to the mass spectrometer and/or the Bioworks Browser software used in these experiments. However, to

date, the majority of metabolic labeling experiments requiring high resolution is conducted using either LTQ-Orbitrap or LTQ-FTICR mass spectrometers. Interestingly, Mayampurath *et al.* reported similar problems for (unlabeled) qualitative data obtained from similar instruments and similar software and introduced the program DeconMSn to determine more accurately the monoisotopic mass and charge [26]. Second, it seems that often the *m/z* of large labeled peptides does not correspond to the monoisotopic *m/z*. Searching the data from the .dta files resulted in six peptide identifications (1-6) and only after manually changing the precursor *m/z* to the monoisotopic *m/z* of the last four peptides (7-10), all ten peptides could be identified (TABLE 1). This is important, since the quantitation process would depend completely on the identification of the unlabeled form of the peptide if the labeled form of a peptide is not identified. Most quantification programs require only one of the forms of the peptide to be identified and calculate the position of the other form, but it will nevertheless be difficult to identify and quantitate proteins that are highly regulated in favor of the labeled condition.



Next, we wanted to get a more global picture of the peptide identifications in large data sets and investigated the total number of peptides that were identified in both the *Drosophila* and *C. elegans* labeling experiments. Although both experiments presented similar results we used the results from the *Drosophila* experiment to illustrate our findings. In FIGURE 2A is shown in dark grey the precursor mass distribution of all the 253 486 queries that were searched in the search engine Mascot. In light grey is shown the distribution of the 69 465 identified peptides, both unlabeled and labeled, that have a Mascot ion score higher than 20. Only a little more than 27% of the queries resulted in peptide identifications (31% in the *C. elegans* experiment), and although an equally shaped distribution is expected, it is clear that between 1 700 and 3 500 Da fewer peptides were identified. To see if there was a preference for either unlabeled (i.e., $^{14}N$) or labeled ($^{15}N$) identifications, all peptides were sorted accordingly. Surprisingly, there were almost 3 times more $^{14}N$ identifications than $^{15}N$ identifications observed in the *Drosophila* and 2 times more in the *C. elegans* data (TABLE 2). More strikingly, FIGURE 2B shows the resulting individual mass distributions, and not only is the number of $^{14}N$ identifications considerably larger than the number of $^{15}N$ identifications, but no or little labeled peptides were identified in the high mass area (1 500 Da and higher). In contrast, unlabeled peptides are identified over the entire mass range, comparable to the distribution of all the precursor masses. The region without labeled peptide identifications corresponded to the results presented in TABLE 1 where it became clear that the precursor mass of large (> 1 900 Da) labeled peptides did not match

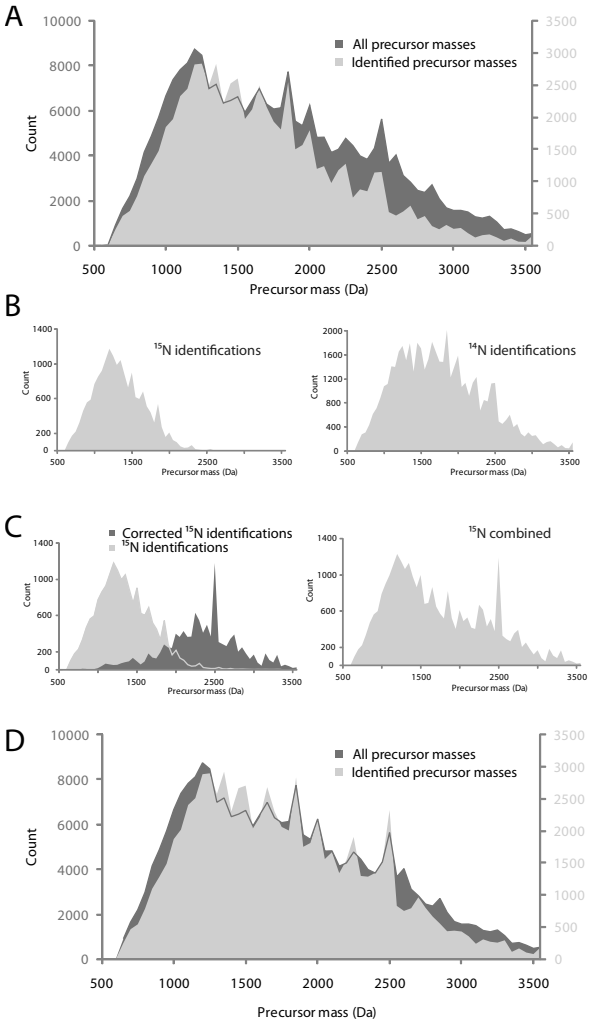


Figure 2. Distribution (A) of precursor masses (dark grey) and identified (light grey) precursor masses. A total of 27% of the precursor masses resulted in a peptide identification. The identified peptides can be divided into $^{15}N$ and $^{14}N$ identifications (B) and it is clear that fewer labeled peptides were identified compared to the unlabeled peptide distribution. After correction, a substantial number of extra identifications (dark grey distribution in part C) is obtained and when combined with the original $^{15}N$ identifications (C), the $^{14}N$ distribution is more closely resembled although the total number of labeled identifications is still lower than the unlabeled identifications. After combining all identifications (light grey distribution in part D) the precursor distribution (dark grey in parts A and D) is more closely resembled. The bin size of all distributions was set to 50 Da.

the monoisotopic mass and was one *m/z* unit too low.

In an attempt to identify more labeled peptides and to correct for the selection of a $^{14}N$ isotopologue as the monoisotopic peak, we changed the precursor mass of all the queries by adding 1 Dalton (Da) and searched again. In this way 11 762 additional labeled peptides with a Mascot score of at least 20 were indentified in the *Drosophila* and 2 826 in the *C. elegans* experiment (TABLE 2). Of these additional identified peptides, around 11% (1 297 and 329) is identified exclusively after precursor mass correction (i.e., these peptides have no $^{14}N$ identification) and this fraction represents the number of extra peptides that can be used for quantitative analysis. Also a small number of unlabeled peptides (618 and 389) were identified in both experiments after mass correction, indicating that also normal

peptides are affected, although this phenomenon is less prominent. The mass distribution, of the additional 11 762 labeled peptides (FIGURE 2C, shown in dark grey) indicates that particularly larger peptides are identified and when combined with the original identifications (second distribution, FIGURE 2C) the distribution resembles the distribution of the $^{14}N$ identifications more closely. Although the new total number of labeled identifications increased (29 532 and 17 485, TABLE 2), this is still not equal to the total number of unlabeled identifications (51 695 and 25 665). A possible explanation could be the fact that, in the case of a 1:1 mixture, the intensity of the labeled form is always less than the intensity of the unlabeled form due to the satellite isotopologues. The mass spectrometer detects the unlabeled peptide before the labeled peptide is detected, resulting in more unlabeled peptide identifications. We also explored the possibility that another isotopologue ('-2' in FIGURE 1) is selected and changed the precursor mass by adding 2 Da, but this resulted in no significant extra identifications (data not shown). Nevertheless, the distribution of the new total number of identifications (81 227) shows a remarkable resemblance to the distribution of the queries, FIGURE 2D. This indicates that an equal percentage of peptides have been identified over a mass range of 3 000 Da (500 – 3 500). Moreover, the percentage of identified queries was, with the extra $^{15}N$ identifications, increased from 27% to 32% and from 31% to 33% for the two experiments.



**Table 2. Number of Peptide Identifications before and after Precursor Mass Correction Resulting in an Additional Number of Labeled Peptide Identifications**

| identifications | *Drosophila* peptides[a] (%[b]) | *C. elegans* peptides[a] (%[b]) |
|---|---|---|
| $^{14}N$ and $^{15}N$ | 69 465 (100) | 40 324 (100) |
| $^{14}N$ | 51 695 (74.4) | 25 665 (63.6) |
| $^{15}N$ | 17 770 (25.6) | 14 659 (36.4) |
| $^{15}N$ corrected | 11 762 (16.9) | 2 826 (7.0) |
| exclusive $^{15}N$[c] | 1 297 (1.9) | 329 (0.8) |

[a] Number of redundant peptide identifications that have a Mascot score of at least 20. [b] Percentage of identifications relative to the initial ($^{14}N$ and $^{15}N$) number of identifications. [c] The number of peptides that are identified exclusively after precursor mass correction.

**Quantitative data •** Besides the identification process that is compromised by the enrichment level, the quantitation process is also affected. When the signal of the monoisotopic peak is considered for peptide ratio calculations, a substantial amount of signal that resides in the satellite peaks (i.e., '-1' and '-2' peaks in FIGURE 1) is in this way neglected. Therefore, quantitation can be improved by summation of the intensities of the monoisotopic peak and the first $^{14}N$ isotopologue. The effect of this procedure can be exemplified with the ra-

tios of 32 peptides that identified the *Drosophila* protein Q9VVA4_DROME. The average ratio of this protein before correction was 0.86 with a standard deviation of 0.13, resulting in a coefficient of variation of 16%. The ratios of the peptides before and after correction are plotted against their mass indicated by grey circles and black squares respectively in FIGURE 3. These peptides varied widely in mass (700-2 900 Da) and although the enrichment level is assumed identical for all these peptides, their difference in chemical composition (i.e., nitrogen and carbon content) necessitates the calculation of individual correction factors instead of applying a general factor to all peptides. Although we used an average enrichment level in these calculations (see also 'Origin of samples'), ideally, the individual peptide enrichment level should be determined and used to derive the correction factor, especially when the labeling period is short or with heterogeneous tissue samples. The need for individual correction factors is reflected in FIGURE 3 where the first 17 peptides (700 - 1 700 Da) show a positive correction, followed by 6 peptides (1 700 - 2 100 Da) that show almost no change in ratio, and by the last 9 peptides that are negatively corrected. Taking together all corrected peptide ratios, a new protein average ratio of 0.87 with a standard deviation of 0.09 was calculated. The coefficient of variation decreased from 16 to 10%, indicating a significant improvement in precision. The most important contributors to the

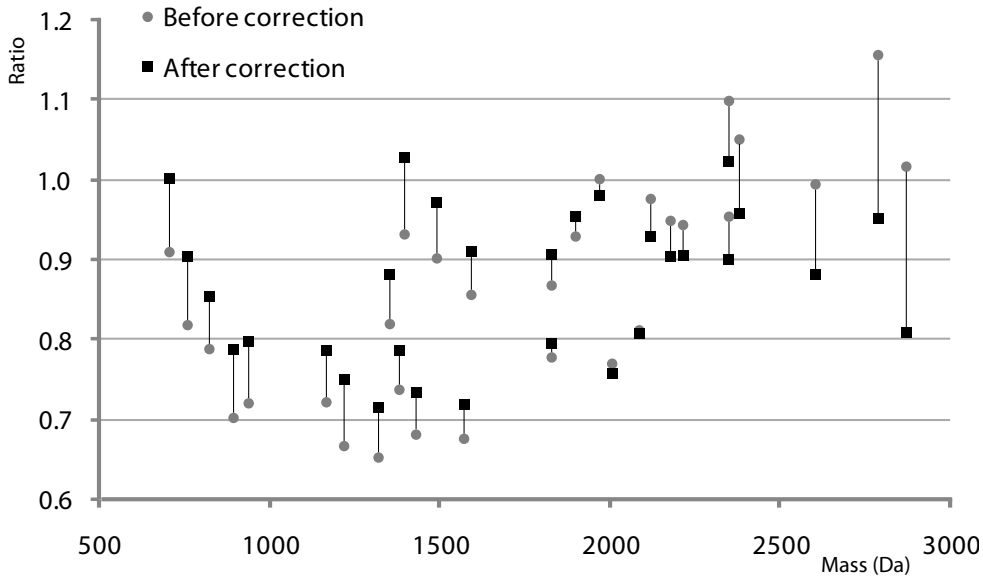


Figure 3. Peptide ratios before (grey circles) and after (black squares) correction of the protein Q9VVA4_DROME plotted against their mass. The smallest 15 peptides are positively corrected, the next 6 are minimally corrected, and the 9 biggest peptides are negatively corrected, resulting in more accurate peptide ratios and a more precise protein ratio.

correction factor, besides the enrichment level, are the carbon and nitrogen content of a peptide. For relatively small peptides, the carbon content is low resulting in a minor contribution of the $^{13}C$-isotopologue of the first satellite isotope to the monoisotopic peak, indicative for a positive correction factor. This changes when the carbon content of the peptide increases up to a point where the carbon-13 contribution equals the intensity of the satellite peaks, resulting in a correction factor of 1. For large peptides, the correction factor becomes negative because carbon-13 is now the most important contributor to the intensity of the monoisotopic peak.

This effect is explored systematically for the two data sets, 96.7 and 98.2% enriched in $^{15}N$ (FIGURE 4A). The correction factor is plotted against the numbers of carbon atoms in peptides containing 7 up to 19 nitrogen atoms. For this subset of quantified peptides from the *Drosophila* experiment, the correction factor changes significantly depending on the number of nitrogen and carbon. It can be seen that the correction factor for peptides containing a constant number of nitrogen atoms decreases linearly with an increasing number of carbon atoms. For example, the correction factor for peptides containing 13 nitrogen atoms varies from 1.15 (33 carbon atoms) to 1.05 (69 carbon atoms). Besides the carbon and nitrogen contribution, the enrichment level has a major effect on the correction factor. A drop of 1.5% in enrichment level (98.2 - 96.7%) has a significant effect on the correction factor, shown in FIGURE 4B. Average correction factors are calculated and plotted for peptides quantified from both the *Drosophila* and *C. elegans* experiments against the number of nitrogen atoms. Although this average correction factor cannot be applied to all peptides

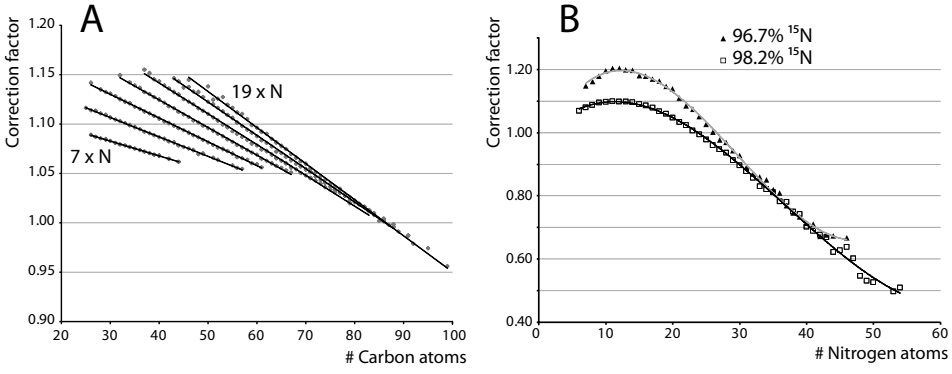


Figure 4. Quantitative correction factor calculated for peptides containing 7 - 19 nitrogen atoms plotted against the number of carbon atoms (A). The correction factor linearly decreases as the number of carbon atoms increases ($R^2$ was better than 0.998 in all cases). Average correction factors plotted against the number of nitrogen atoms in peptides enriched in 96.7% (solid triangles) or 98.2% (open squares) heavy nitrogen (B).

containing a fixed number of nitrogen atoms, it is illustrative that the profile changes depending on the enrichment level. The average correction factor for the *C. elegans* experiment (96.7%) is much higher for peptides containing up to 32 nitrogen atoms than for the *Drosophila* experiment (98.2%). In addition, the average correction factor for the 96.7% enrichment levels off faster (41 or more nitrogen atoms) than for the 98.2% enrichment (FIGURE 4B). With the use of this method, all quantified peptides in both experiments were corrected. The average coefficient of variation in the *Drosophila* experiment was 22% before correction and improved to 18% after correction. Likewise, the average coefficient of variation decreased by 3% - 15% in the *C. elegans* experiment. This indicates that a significant improvement in precision can be obtained using this correction method, meaning that more subtle changes can be measured. Although we have applied this to $^{15}$N-labeled proteins, this could as well be used for $^{13}$C-labeling. This method could then be used to include signal that resides in the $^{13}$C peaks of the unlabeled peptides and any residual intensity of the $^{12}$C peak of labeled peptides (in case of less than 100% labeling).

## V. Conclusion

Metabolic enrichment of heavy stable isotopes is nowadays routinely used in quantitative proteomic applications. Enrichments of stable isotope labeled atoms is never complete (100%), making it necessary to compensate for the unfavorable effects in protein identification and quantitation. Using postanalysis protocols, we showed that large $^{15}N$-labeled peptides are under-represented in the total number of identified peptides, and we have described a method that increases the number of identifications specifically of these peptides. In addition, mass spectrometry-derived quantitative data based on monoisotopic peptide intensities lacks accuracy due to incomplete enrichment. This can be corrected using the enrichment level and the peptide chemical formula resulting in more accurate peptide ratios as well as more precise protein ratios. Our approach was tested and applied on two independent experiments differing in enrichment levels and clearly show that a significant improvement of both qualitative and quantitative data can be obtained. Furthermore, with the increased number of organisms that can be labeled with $^{15}N$, we expect that this approach will become part of standard procedures to correct for the suboptimal enrichment levels that are often observed.



## Acknowledgement

This work was supported by The Netherlands Proteomics Centre (www.netherlandsproteomicscenter.nl).

# References


[1] Bondarenko P.V., Chelius D. and Shaler T.A., *Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry*. Anal. Chem., **2002**, *74*, 4741-4749.

[2] Chelius D. and Bondarenko P.V., *Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry*. J. Proteome Res., **2002**, *1*, 317-323.

[3] Liu H., Sadygov R.G. and Yates J.R., 3rd, *A model for random sampling and estimation of relative protein abundance in shotgun proteomics*. Anal. Chem., **2004**, *76*, 4193-4201.

[4] Gygi S.P., Rist B., Gerber S.A., Turecek F., Gelb M.H. and Aebersold R., *Quantitative analysis of complex protein mixtures using isotope-coded affinity tags*. Nat. Biotechnol., **1999**, *17*, 994-999.

[5] Ross P.L., Huang Y.N., Marchese J.N., Williamson B., Parker K., Hattan S., Khainovski N., Pillai S., Dey S., Daniels S., et al., *Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents*. Mol. Cell. Proteomics, **2004**, *3*, 1154-1169.

[6] Mirgorodskaya O.A., Kozmin Y.P., Titov M.I., Korner R., Sonksen C.P. and Roepstorff P., *Quantitation of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry using (18) O-labeled internal standards*. Rapid. Commun. Mass Spectrom., **2000**, *14*, 1226-1232.

[7] Molloy M.P., Donohoe S., Brzezinski E.E., Kilby G.W., Stevenson T.I., Baker J.D., Goodlett D.R. and Gage D.A., *Large-scale evaluation of quantitative reproducibility and proteome coverage using acid cleavable isotope coded affinity tag mass spectrometry for proteomic profiling*. Proteomics, **2005**, *5*, 1204-1208.

[8] Ong S.E., Blagoev B., Kratchmarova I., Kristensen D.B., Steen H., Pandey A. and Mann M., *Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics*. Mol. Cell. Proteomics, **2002**, *1*, 376-386.

[9] Ishihama Y., Sato T., Tabata T., Miyamoto N., Sagane K., Nagasu T. and Oda Y., *Quantitative mouse brain proteomics using culture-derived isotope tags as internal standards*. Nat. Biotechnol., **2005**, *23*, 617-621.

[10] Conrads T.P., Alving K., Veenstra T.D., Belov M.E., Anderson G.A., Anderson D.J., Lipton M.S., Pasa-Tolic L., Udseth H.R., Chrisler W.B., et al., *Quantitative analysis of bacterial and mammalian proteomes using a combination of cysteine affinity tags and 15N-metabolic labeling*. Anal. Chem., **2001**, *73*, 2132-2139.

[11] Bindschedler L.V., Palmblad M. and Cramer R., *Hydroponic isotope labelling of entire plants (HILEP) for quantitative plant proteomics; an oxidative stress case study*. Phytochemistry, **2008**, *69*, 1962-1972.

[12] Nelson C.J., Huttlin E.L., Hegeman A.D., Harms A.C. and Sussman M.R., *Implications of $^{15}$N-metabolic labeling for automated peptide identification in Arabidopsis thaliana*. Proteomics, **2007**, *7*, 1279-1292.

[13] Oda Y., Huang K., Cross F.R., Cowburn D. and Chait B.T., *Accurate quantitation of protein expression and site-specific phosphorylation*. Proc. Natl. Acad. Sci. U.S.A., **1999**, *96*, 6591-6596.

[14] Kolkman A., Daran-Lapujade P., Fullaondo A., Olsthoorn M.M.A., Pronk J.T., Slijper M. and Heck A.J.R., *Proteome analysis of yeast response to various nutrient limitations*. Mol Syst Biol, **2006**, *2*.

[15] Krijgsveld J., Ketting R.F., Mahmoudi T., Johansen J., Artal-Sanz M., Verrijzer C.P., Plasterk R.H. and Heck A.J., *Metabolic labeling of C. elegans and D. melanogaster for quantitative proteomics*. Nat. Biotechnol., **2003**, *21*, 927-931.

[16] Gouw J.W., Pinkse M.W., Vos H.R., Moshkin Y.M., Verrijzer C.P., Heck A.J.R. and Krijgsveld J., *In vivo stable isotope labeling of fruit flies reveals post-transcriptional regulation in the maternal-to-zygotic transition*. Unpublished work, **2008**.

[17] Dong M.Q., Venable J.D., Au N., Xu T., Park S.K., Cociorva D., Johnson J.R., Dillin A. and Yates J.R., 3rd, *Quantitative mass spectrometry identifies insulin signaling targets in C. elegans*. Science, **2007**, *317*, 660-663.

[18] Wu C.C., MacCoss M.J., Howell K.E., Matthews D.E. and Yates J.R., 3rd, *Metabolic labeling of mammalian organisms with stable isotopes for quantitative proteomic analysis*. Anal. Chem., **2004**, *76*, 4951-4959.

[19] McClatchy D.B., Dong M.Q., Wu C.C., Venable J.D. and Yates J.R., 3rd, *15N metabolic labeling of mammalian tissue with slow protein turnover*. J. Proteome Res., **2007**, *6*, 2005-2010.

[20] MacCoss M.J., Wu C.C., Liu H., Sadygov R. and Yates J.R., 3rd, *A correlation algorithm for the automated quantitative analysis of shotgun proteomics data*. Anal. Chem., **2003**, *75*, 6912-6921.

[21] Rappsilber J., Ishihama Y. and Mann M., *Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics*. Anal. Chem., **2003**, *75*, 663-670.

[22] Meiring H.D., van der Heeft E., ten Hove G.J. and de Jong A.P.J.M., *Nanoscale LC-MS$^{(n)}$: technical design and applications to peptide and protein analysis*. J. Sep. Sci., **2002**, *25*, 557-568.

[23] Schulze W.X. and Mann M., *A novel proteomic screen for peptide-protein interactions*. J. Biol. Chem., **2004**, *279*, 10756-10764.

[24] MacCoss M.J., Wu C.C., Matthews D.E. and Yates J.R., 3rd, *Measurement of the isotope enrichment of stable isotope-labeled proteins using high-resolution mass spectra of peptides*. Anal. Chem., **2005**, *77*, 7646-7653.

[25] Jennings M.E., 2nd and Matthews D.E., *Determination of complex isotopomer patterns in isotopically labeled compounds by mass spectrometry*. Anal. Chem., **2005**, *77*, 6435-6444.

[26] Mayampurath A.M., Jaitly N., Purvine S.O., Monroe M.E., Auberry K.J., Adkins J.N. and Smith R.D., *DeconMSn: a software tool for accurate parent ion monoisotopic mass determination for tandem mass spectra*. Bioinformatics, **2008**, *24*, 1021-1023.

# CHAPTER 5

# MSQuant, an open source platform for the interpretation of tandem mass spectra and for quantitative proteomics

Peter Mortensen[†], Joost W. Gouw[‡], Jesper V. Olsen[§], Jens S. Andersen[†], Leonard J. Foster[ǯ], Blagoy Blagoev[†], Shao-En Ong[ǀ], Albert J.R. Heck[‡], Goiqing Li[§], Yong Zhang[§], Jürgen Cox[§] and Matthias Mann[†,§]

[†]Center for Experimental Bioinformatics, University of Southern Denmark, Odense, Denmark , [‡]Biomolecular Mass Spectrometry and Proteomics Group, Bijvoet Center for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands, [§]Department for Proteomics and Signal Transduction, Max-Planck Institute for Biochemistry, Martinsried, Germany, [ǯ]Department of Biochemistry and Molecular Biology, University of British Columbia, Vancouver, Canada and [ǀ]Broad Institute of MIT and Harvard, Cambridge, United States

# I. Abstract

Mass spectrometry based proteomics critically depends on algorithms for data interpretation. A current bottleneck in the rapid advance of proteomics technology is the closed nature and slow development cycle of vendor-supplied software solutions. We have created an open source software environment, called MSQuant, which allows visualization and validation of peptide identification results directly on the raw mass spectrometric data. MSQuant iteratively recalibrates MS data thereby significantly increasing mass accuracy. A peptide identification score using $MS^3$ spectra is incorporated, as well as a post-translational modification (PTM) score. This PTM score determines the probability that a modification such as a phosphorylation is placed at a specific residue in an identified peptide. MSQuant supports relative protein quantitation including the SILAC, ICAT and $^{15}N$-labeling approaches, as well as protein correlation profiling. MSQuant is freely available including an installer and supporting scripts at http://msquant.sourceforge.org.

# II. Introduction

Even though protein science and 2D gel electrophoresis have been around for decades, proteomics – the large-scale study of proteins – has only recently proven its value as a post-genomic tool. Several areas of proteomics use different technological platforms [1-3] and mass spectrometry (MS) is a particularly powerful method to characterize endogenous proteins including their modifications. MS-based proteomics [4] produces data which is very different from that familiar to most bioinformaticians and as a result, only certain aspects of the field have attracted their attention and input. To gain biologically or clinically relevant insights from proteomic data, however, robust and effective software solutions are required [5]. On one level, vast quantities of data need to be stored and managed. On another level, data needs to be reduced and information needs to be extracted. In MS-based proteomics, two main areas for software algorithms have emerged. Firstly, proteins need to be identified from the peptide masses and fragmentation data. Early breakthroughs in this area include the development of efficient search tools for peptide mass fingerprinting and for algorithms that matched tandem mass spectra to their cognate peptide sequences in databases (reviewed in [6]). The second area of algorithmic development concerns quantitative measurements, either by extracting peptide ratios from isotopically labeled peptide pairs or by the determination of peptide ion currents (reviewed in [7]). Despite significant progress, both areas still require much research. As an example, statistical validation of the peptide hits generated by various search engines, as well as what constitutes acceptable peptide hits, remain active areas of research and controversy [8].

Here we describe an open-source program called MSQuant, which implements a novel design strategy with significant advantages for proteomics projects. Firstly, being open source it can be modified and extended rapidly and independently of the MS manufacturer. Secondly, we use building blocks of the manufacturer's software to gain direct access to the raw data without any loss of information and without having to re-create basic data manipulation and visualization tools. As a result, we have been able to quickly implement a range of analysis tools, which are not available in any commercial software. These analysis tools have been crucial in enabling a number of large-scale proteomics projects in our laboratories ([9-17]). Currently, MSQuant parses the output of Mascot search files and allows interaction with Thermo Electron, ABI-Sciex and Micromass/Waters data files.

As we hope that this paper will be of interest to both the bioinformatics and the pro-

teomics communities, we begin with a short overview of a typical proteomics experiment and the informatics tasks relevant to MS-based proteomics.

**The Experimental Process in MS-Based Proteomics** • FIGURE 1 shows a flow diagram of the currently most popular form of MS-based proteomics (see ref. [4] for an introduction). Briefly, proteins are first extracted from the source of interest, for instance protein complexes, organelles or organisms and contain anywhere between a few proteins to thousands of proteins. In case of an isotope labeling experiment, the two or more populations to be compared are mixed as is shown in FIGURE 1A. Proteins are enzymatically digested to peptides, often after separation by one-dimensional gel electrophoresis, the resulting complex mixture of peptides is applied to a chromatographic column and the peptides are analyzed by on-line chromatography. As they elute from the column they are vaporized and ionized by electrospray. The mass spectrometer measures the masses and signal heights of the eluting peptides and selects peptides in turn for fragmentation, giving rise to MS/MS spectra (also called $MS^2$ or tandem mass spectra). These fragmentation spectra contain in-

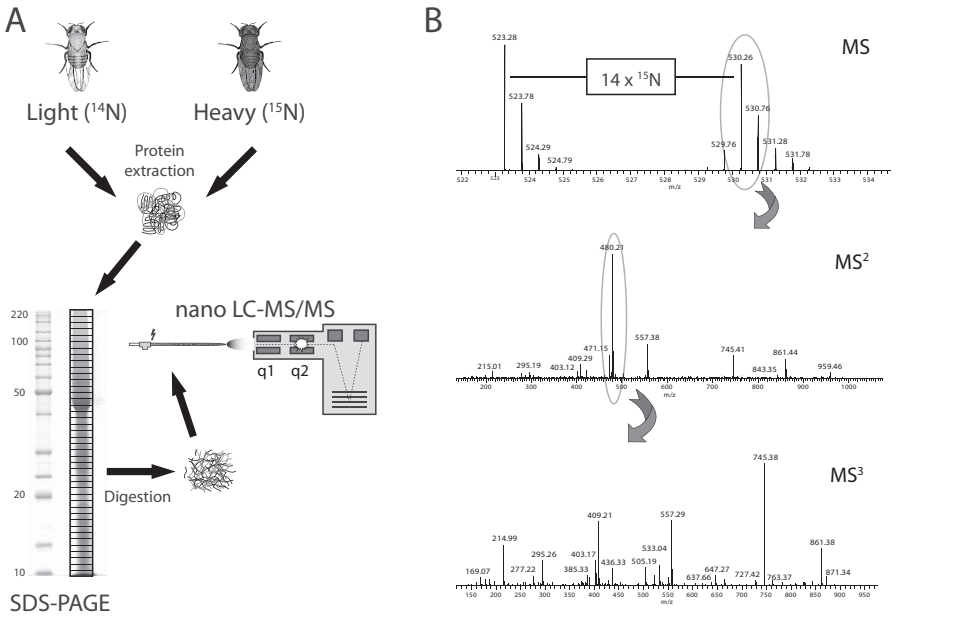


Figure 1. Data generation in MS-based quantitative proteomics. (A) Two populations of flies are metabolically labeled with a 'light' ($^{14}N$) and a 'heavy' ($^{15}N$) isotope. After optional fractionation and protein separation, proteins are digested to peptides and analysed by electrospray on-line coupled to a tandem mass spectrometer. (B) The upper panel shows the extracted ion current of a peptide pair. The middle panel depicts the $MS^2$ fragmentation spectrum of the 'heavy' peptide and the bottom panel shows the $MS^3$ spectrum of the most intense $MS^2$ signal.

formation related to the amino acid sequence of the peptide and are used to determine the presence of the peptide in a protein database. Fragmenting peaks in the MS/MS spectrum is called $MS^3$ and adds further specificity to the database search, FIGURE 1B.

In a qualitative experiment, the objective is to identify as many of the proteins as possible in a given sample whereas in a quantitative experiment, the relative or absolute amount of proteins is also determined [7]. The quantitative information is derived from the intensity of the MS peak of the peptide signal as it elutes from the column, the so-called 'extracted ion current' (XIC). In isotope based quantitative proteomics, such as in the ICAT [18] or metabolic $^{15}N$-labeling [19] technologies, this XIC is compared between two isotopically different forms of the same peptide. Depending on the experiment, the relative amount of protein modifications such as phosphorylations may also be quantified.

**The Computational Process in MS-Based Proteomics** • The LC-MS/MS experiment described above is controlled by software provided by the manufacturer of the mass spectrometer. The resulting data resides in raw files (layer 1 in FIGURE 2) in a proprietary format and can be visualized by the same software. However, this software is normally only used during acquisition of the data and not in the subsequent processing steps. The route from raw data to relevant proteomic results involves four levels of data processing, (*i*) feature extraction, (*ii*) database searching, (*iii*) quantitation and (*iv*) bioinformatics analysis (FIGURE 2).

Mass spectra initially are just mass and intensity pairs whereas we are interested in peptide or fragment peaks. Thus, processing begins with the generation of peak lists from the fragmentation spectra (layer 2 in FIGURE 2). Signal processing techniques can be applied to remove chemical and electronic noise (reviewed in [20]). Peak picking algorithms then attempt to find peaks, isotope clusters and charge states. These algorithms are either built into the vendor's software or as a 'pre-processing' step into the database search software. In our experience, current peak picking algorithms have a significant error rate, especially in the case of metabolic labeling with suboptimal $^{15}N$-enrichments [21]. For example, for peptides larger than 1 000 Da, the $^{13}C$ peak is often mistaken for the monoisotopic peak and the charge state is also sometimes incorrect. Therefore we have developed a program, called DTAsuperCharge, which attempts to assign the most probable monoisotopic peak and charge state but gives the expert user the option to correct the software assignment (see below). Since there can be thousands of potential fragments signals in a spectrum, generally only a few are relevant for a database search which makes the selection

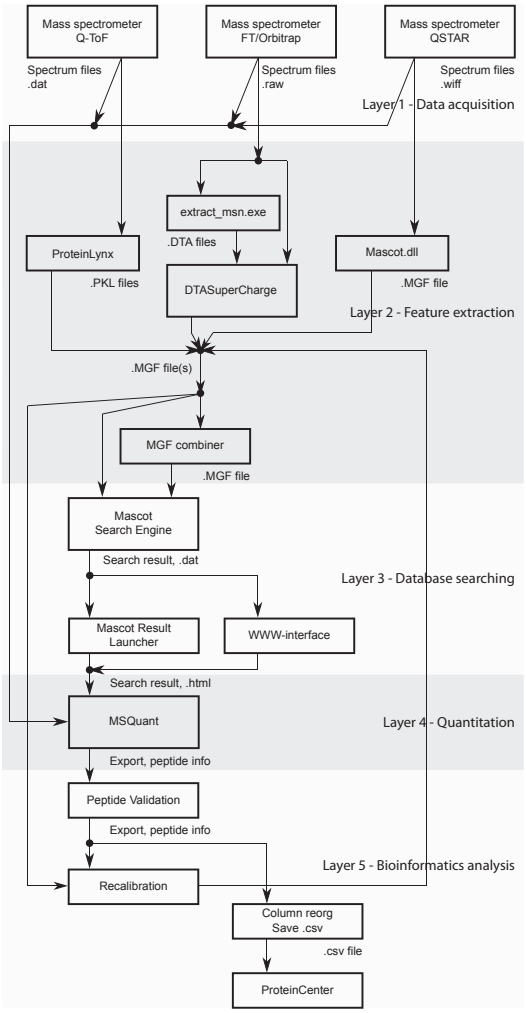


Figure 2. Data generation in MS-based quantitative proteomics. (A) Two populations of flies are metabolically labeled with a 'light' ($^{14}N$) and a 'heavy' ($^{15}N$) isotope. After optional fractionation and protein separation, proteins are digested to peptides and analysed by electrospray online coupled to a tandem mass spectrometer. (B) The upper panel shows the extracted ion current of a peptide pair. The middle panel depicts the $MS^2$ fragmentation spectrum of the 'heavy' peptide and the bottom panel shows the $MS^3$ spectrum of the most intense $MS^2$ signal.

of the peaks most likely to be informative an important step in the generation of the peak list.

The peak lists, consisting of the *m/z* value and charge state of the precursor and the *m/z* value and intensity of the fragments are then submitted to a database search program (layer 3 in FIGURE 2). The results of this search are a list of identified peptides and their scores. Typically zero to ten sequences are returned for each MS/MS event. Note that current proteomics experiments often involve dozens of raw data files, each containing of the order of 10 000 MS/MS spectra. The database engine must be sufficiently robust to handle and report these large amounts of data. These peptide lists are then combined and clustered into lists of proteins identified with one or more peptides. Although often

treated in a simple way this 'protein inference problem' actually poses challenging statistical and conceptual problems for 'bottom up' proteomics [22]. Users apply certain criteria to the output of the search engine. For example, valid proteins may be required to have at least two tryptic peptides with a certain minimum peptide score and the protein may be required to have a certain minimum protein score. All proteins fulfilling these criteria are then reported as the outcome of the qualitative proteomic experiment.

In a quantitative proteomics experiment a further processing step is performed (before or after the identification of peptides). In isotope labeling experiments the XIC of both isotope forms of the peptide is determined. This requires finding the 'isotope partner' of an identified peptide, even when that partner was not sequenced. The quantitative ratios of the different peptides identifying each protein are averaged to produce a protein ratio and identified and quantified proteins can be ordered by fold-change between the two or more experimental conditions, layer 4 in FIGURE 2.

FIGURE 2 also lists 'bioinformatics analysis' as a step in the analysis pipeline (layer 5). This bioinformatics analysis is not yet always part of the routine processing of proteomic data. However, in our opinion, it is a necessary part. For example, identified proteins need to be checked for redundancy using homology tools. It is also necessary to pinpoint peptides distinguishing between isoforms of proteins, if they are present. More broadly, depending on the projects, proteomics results should be analyzed for novel proteins, proteins with certain domains, proteins with certain disease associations and so on. Such an analysis should be performed for all proteins in a qualitative experiment and for the proteins with a significant fold-change in a quantitative experiment. Currently, there are few tools to perform such analyses, with ProteinCenter (www.proxeon.com) and Scaffold (www.proteomesoftware.com) as prominent examples.

**Problems with the Current Processing Model** • There are a number of bioinformatics challenges associated with the proteomic workflow as outlined in FIG. 2. Today, mass spectrometers have fast scan speeds and can have high resolution resulting in large amounts of data that need to be processed and stored (Gbytes per machine per day). Automation is a necessity because manual interpretation of hundreds of thousands of fragmentation spectra is not feasible. However, the proteomics community has begun to realize that such automated workflows and large data amounts can lead to significant errors [8], with published data sets sometimes containing large fractions of false identifications. This problem did not arise to the same extent in traditional MS because data was usually obtained on high resolution instruments by expert users and was interpreted by those users. Such user

intervention is not possible any more in the 'pipeline' workflow as depicted in FIG. 1. In fact, in all proteomics software solutions that we are aware of, the identification data generated by search engines has 'lost connection' to the raw data and the user can only validate congruence of the search results with a small amount of the actual information.

A second problem with the pipeline model arises from the fact that the software used to acquire and display the raw MS data is of necessity proprietary. Thus tools to analyze the raw data have to be implemented by the manufacturer, who may not be interested in specialized questions and usually observe a conservative software release policy. MS vendors offer various proprietary software suites for the analysis of proteomic data. However, these often lag behind when new technologies allow new types of data to be acquired. Due to their closed nature, they typically do not allow much interaction or experimentation with the data. A related complication is that the database search program is often developed by a different party than the mass spectrometer vendor. Thus data types that can be acquired by the instrument may not be 'searchable' later.

One solution to this dilemma is the introduction of open standards for mass spectrometric data [23]. This seems to be a good and necessary strategy for the future but for now, the resulting large data files and the inability of the standards to represent fully the data that can be generated with sophisticated scan functions have so far prevented us and others to adopt such open data standards.

## III. Materials and Methods

**MSQuant Development Environment •** Open source bioinformatics projects are usually implemented in an environment such as Java. MSQuant, however, is written in the Microsoft .NET integrated development environment. The main reason for this choice was the fact that MS vendor's software is implemented on the Windows platform and programs created in the .NET environment easily interface to those programs, partially solving the problem of proprietary software. The Microsoft .NET environment has some similarities to the Java environment and contains more than 6 000 classes providing low and high level functionality. Like other modern development environments, .NET enables object oriented architecture, exception handling, rapid development of graphical user interfaces as well as a host of other powerful features that enable efficient software construction. A limitation of the .NET environment in our experience is that it often interfaces poorly to cross platform tools used in bioinformatics. As an example, the R and Bioconductor environment [24, 25] is used extensively in our department but we currently interface data between MSQuant and these platforms via a spreadsheet format.

MSQuant interfaces with MS vendor's software through standard Microsoft binary component standards (COM objects, for Common Object Model). The spectra displayed in MSQuant are the native display components of the manufacturer's software (ABI-Sciex Analyst and Thermo Electron Xcalibur), thus the user familiar with these programs does not have to learn a new spectra display and navigation scheme. The COM model is also used for exporting the data directly into Microsoft Office software such as Excel.

**DTAsuperCharge •** As mentioned above, vendor software sometimes miss-assigns the charge state or monoisotopic mass of the precursor ion or not even a single precursor mass is assigned at all. As part of the MSQuant environment, we also developed DTAsuperCharge, a program which attempts to assign the correct charge state and isotope state for each precursor. The algorithm generates an average isotope distribution and determines the square deviation to the observed isotope cluster for each possible charge state and isotope position. The square deviation serves as a score to determine the most likely result. These *m/z* value and charge state are then exported into the file used for database searching. Furthermore, DTAsuperCharge allows the user to decide to directly export the MS/MS peak list as determined by the vendor software or to add an additional preprocessing step. Such a step can be necessary to limit the number of peaks for database searching. This assures that the database search engine will accept the data and speeds up the search. In

an attempt to retain the most informative peaks DTAsuperCharge will recursively find the largest peaks in the fragment spectrum. First the largest data point is found and the corresponding peak is determined by peak recognition. This defines two new mass ranges to the left and to the right of the peak in which the procedure is repeated until the user defined number of iterations has been achieved. Alternatively, the user can also specify that this algorithm is performed independently in smaller mass ranges, such as 300 *m/z* windows.

Other optional pre-processing steps included in DTAsuperCharge include the proper assignment of the precursor mass in $MS^3$ spectra of phosphopeptides. In these experiments, the $MS^2$ spectrum serves to recognize phosphopeptide candidates and to generate a neutral loss precursor, which is then fragmented in an additional step to identify it and to locate the phosphorylation site. In this case, to satisfy database engine requirements, the precursor mass of the $MS^3$ spectrum is specified as the peak in the original spectrum rather than the neutral loss peak of the MS/MS spectrum.

The improvement of identification results obtained with this preprocessing of the MS/MS spectra depends on the sophistication of the vendor program and ranges from no difference to being an essential step. It is in any case essential for transforming Thermo Electron ".dta" files into a format that MSQuant recognizes. The DTAsuperCharge program is open source and available via the MSQuant home page.

**Handling of Multiple Files •** In some situations, such as protein correlation profiling, MSQuant needs to keep track of the raw file that any peptide was identified from. To accomplish this, a script with a GUI interface writes the file information into the header of each $MS^2$or $MS^3$ spectrum.

**Getting Started as an MSQuant User •** At the moment, MSQuant supports the ABI-Sciex and the Thermo Electron data files directly. The Micromass/Waters format is also supported but since there are no COM components for visualization in MassLynx, we employ the Xcalibur (Thermo Electron) data visualization components instead. Therefore, Micromass/Waters users need to have both data systems installed in order to use MSQuant. On the MSQuant homepage detailed instructions for the entire process of MSQuant installation can be found, in particular regarding dependencies on other required software. An installer for both MSQuant and DTAsuperCharge can be downloaded from http://msquant.sourceforge.net. Before executing the installer, we recommend to study the installation notes and examples in detail. Note that MSQuant is freely available for academic and commercial users. MSQuant and associated programs are supplied 'as is', without any

warranty; without even the implied warranty of merchantability or fitness for a particular purpose. The authors, CEBI and the University of Southern Denmark assume no liability nor do they guarantee that all parts of MSQuant are free of intellectual property.

**The User Interface •** MSQuant has four main windows: The start screen associates Mascot result files with the corresponding raw data files. It also allows the user to specify the quantitation mode, various filters used in parsing the Mascot file and a number of other parameters. The protein list window is the main document window and contains a list of identified proteins with the number of identified peptides as well as the protein score. Double clicking a protein brings up the protein validation window (FIGURE 3). The user can manually verify or reject peptides and proteins. The fourth window is used for quantitation and contains peptides identifying the protein, a visual display of the LC profile, a list of the SILAC ratios for each MS spectrum and a panel for a selected MS spectrum (FIGURE 4). This window allows manual check of quantitation. For example, if the SILAC ratio of one peptide is different from the others identifying the same protein, stepping through the zoomed part of the MS spectrum will quickly reveal if there is a problem in quantitation. Saving of results is carried out from the protein list window and acts on the user selected proteins. Data can be saved to a tab separated text file or directly into Microsoft Excel. The user interface also allows specification of the quantitation mode, including a large range of possible labeling conditions. Moreover, arbitrary labels can be specified in an XML file.

**Batch Processing •** MSQuant allows batch processing of an arbitrary subset of identified proteins. Due to limitations in the .NET environment, MSQuant can experience memory problems with large data sets, necessitating saving the current state and manual continuation of the quantitation operation after restart of the program. Quantitation, $MS^3$ scoring (see below) and post-translational modification (PTM) scoring can all be performed in batch mode.

**Getting Started as an MSQuant Developer •** The full source code for MSQuant can be obtained via standard CVS (Concurrent Versions System) procedures from the MSQuant homepage. MSQuant is written as four projects in one 'solution' in VB.NET. These projects can also be downloaded and information about installation is contained in a separate text file.

# IV. Results and Discussion

The MSQuant Program. In 2002 we started to develop a program we called Mascot Parser. This program was first developed out of necessity, as a platform for the interpretation of Mascot database search results against raw MS/MS spectra. Furthermore, in 2001 we had begun to develop the SILAC method but could not find any software to extract quantitative information from the mass spectra. Manual quantitation of even a simple experiment could take several months. We then added a quantitation module to the program, which was then accordingly renamed to MSQuant. In this way, MSQuant had a significant part in the rapid success of SILAC. We have since used MSQuant as a platform for quickly implementing algorithmic solutions. In early 2004, we have posted MSQuant and its source code at the sourceforge website, and it has since been in use in a number of proteomics laboratories around the world.

**Modules and Functionality of MSQuant** • The program DTAsuperCharge or similar programs are first used to preprocess the raw data to prepare it for database searching. This step also incorporates information required by MSQuant to locate the correct raw data during processing. Then a database search program such as Mascot is used to identify peptide hits corresponding to the MS/MS data. At the start of a session, MSQuant interprets the output of the database search and identified peptides and proteins are parsed into an internal data structure. Information for each peptide includes its charge state, observed mass, position of the MS/MS spectrum in the raw data file, score, sequence, modification state, theoretical mass and start of the peptide sequence in the protein. The second best matching sequence is also included into the internal data structure.

**Micromass Support** • MSQuant aims to support access to raw data formats from a range of vendors. Initially designed for ABI-Sciex data format (.wiff), MSQuant has now been expanded to support data from Thermo Electron (.raw) and Micromass/Waters as well. This is facilitated by the design of MSQuant, to which modules can be added to incorporate support for specific data files, provided that access to this kind of data via standard Microsoft binary component standards (COM) is possible. MSQuant is programmed in such a way that data from the individual vendor specific modules is generalized immediately after extraction, making it relatively easy to add other modules. Interaction with Waters/Micromass data is made available via the MassLynx Datafile Access Component (DAC), which allows simple access to the MassLynx raw data files and is included in the installa-

tion of MassLynx. DAC is used by MSQuant, for example, to extract data points from the raw data files to display fragmentation spectra and full-scan mass spectra. Moreover, data points from these mass spectra are used to derive ratios between unlabeled and labeled peptides.

**Recalibration •** The first functionality implemented in MSQuant is an iterative recalibration [9]. Two sources determine the error in mass determination, the systematic and random error. Systematic error is caused by a variety of factors, for example drift of calibration constants with time or temperature. This kind of systematic error can be minimized by frequent calibration or by using internal standards. However, in practice, neither of these is always performed or practical. As a result researchers often search proteome data with very wide windows, needlessly degrading the mass accuracy of the instrument.

MSQuant sidesteps this problem by using several hundred high scoring peptides as internal calibrants. Optimal linear calibration constants are calculated from these peptides using their observed versus calculated masses and all measured masses are corrected accordingly. MSQuant reports the improvement in average mass accuracy due to this procedure. On a hybrid quadrupole TOF instrument (ABI-Sciex QSTAR), average absolute mass accuracy improves from 50-200 ppm (depending on the calibration state of the instrument) to typically around 10 ppm. A script changes the precursor masses in the peak list file after which a second search can be performed using the improved mass tolerance. In practice, this simple algorithm improves the mass accuracy of the instrument several fold, leading to much more specific search results. Elimination of this type of systematic mass error would have been straightforward to implement either in MS vendor or database search software but is still not widely employed. This illustrates the value of user access to the data flow as is possible in MSQuant.

For FTICR data using SIM scans as well as LTQ-Orbitrap data with lock mass enabled [26] mass accuracy does not improve very much by the procedure outlined above. In full scan FTICR data, a systematic error is introduced by the space charge effects. In this case MSQuant employs a script to recalibrate masses in the frequency domain [27].

**Validation of Protein and Peptide Hits •** As mentioned above, the typical pipeline model in proteomics does not allow evaluation of raw data. Coupled with the large amount of data generated in proteomics experiments, users are 'educated' to rely on purely statistical measures (i.e. the Mascot score), or at best, to view a static picture of the spectrum at very low resolution. While this latter option is better than no evaluation, it does not allow

appreciation of high resolution data or focusing on details in the spectra. Thus, validation using raw data has become rare in proteomics experiments.

MSQuant retains the location of the spectrum identifying a peptide in its internal data structure. When evaluating a protein, its peptides are displayed in a list. Double clicking a peptide brings up the raw fragment spectrum, see FIGURE 3. This is accomplished using COM interfaces to the vendor's acquisition and visualization software (Xcalibur in the case of Thermo Electron, Analyst for ABI-Sciex and MassLynx for Micromass/Waters). In this way the functionality of the vendor's software is retained in MSQuant and the data can be zoomed at high resolution. Importantly, the high resolution raw data is overlaid with the information calculated from the peptide identification. The b- and y-ion series are automatically marked on the respective peaks and peaks matching a theoretical fragment mass are especially highlighted. When validating spectra, expert mass spectrometrists look for certain telltale features, such as the presence of an intense fragment N-terminal to proline residues in the sequence or the characteristic $a_2$-$b_2$ pair. Database search software ignores these critical features. MSQuant automatically annotates N-terminal proline breaks, the

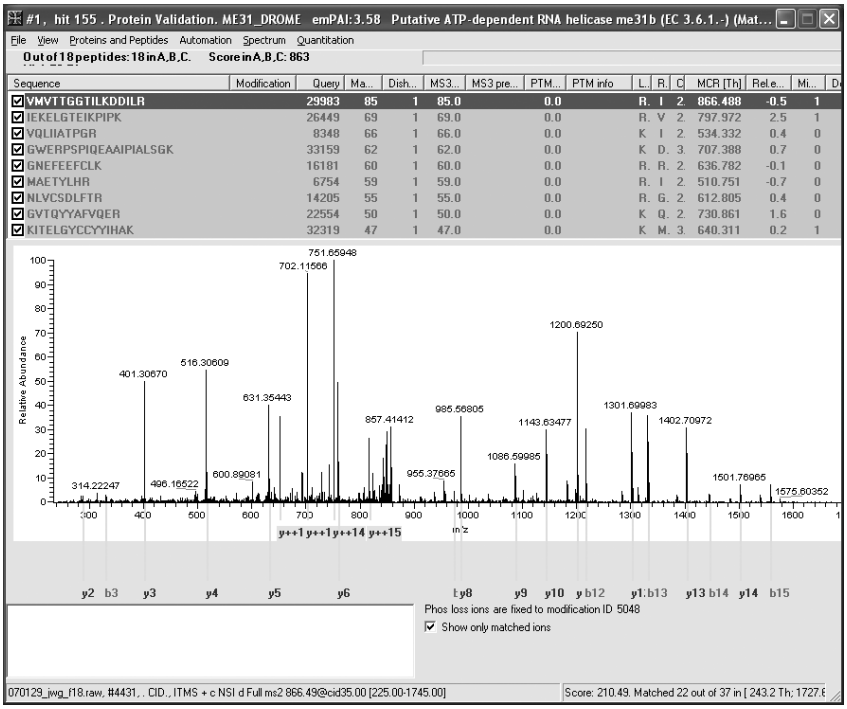


Figure 3. Screenshot of the Validation window in MSQuant. Shown is the raw fragment spectrum of a peptide from the *Drosophila* protein ME31. Note that matched fragments are indicated with cyan colored lines.

$a_2$-$b_2$ pair and complementary ions, facilitating general use of these features. MSQuant also allows annotation with the second best matching peptide sequence for the MS/MS spectrum. Toggling between the two possibilities gives a measure of how much better the top score is compared to the second best score (i.e. the difference may be a single, noisy fragment or a number of highly significant fragments). Furthermore, it is easy to see if the calculated fragments for the two sequences distribute to different peaks, in which case it is likely that both peptides have been fragmented at once.

Importantly, validation in MSQuant is highly visual and streamlined, requiring only a few seconds per spectrum. In a typical proteomics experiment, hundreds of proteins may have been identified but only a small number are biologically interesting (those which change quantitatively in the experiment, for example). Furthermore, many proteins do not need manual validation since they are identified with many peptides and high scores. In our experience, validation in MSQuant of a few hundred MS/MS spectra is almost always sufficient, does not take much time and greatly adds to the overall data quality of the experiment. We highly recommend verification with raw data – as exemplified here in MSQuant – for all cases where extensive biological follow-up is dependent on correct interpretation of an MS/MS spectrum.

**$MS^3$ Scoring •** Proteomic strategies normally involve a single peptide fragmentation step. The introduction of instruments such as the linear ion trap has recently made it technically feasible to perform two stages of fragmentation on a chromatographic time scale and with high sensitivity ($MS^3$). Such a second step of fragmentation should dramatically increase the information of a peptide and greatly increase identification confidence. However, no software was available to use the $MS^3$ fragment information for peptide identification and consequently $MS^3$ was only performed in special cases, such as phosphopeptide identification where standard database search algorithms could be used [28].

Using the MSQuant framework, we quickly implemented an $MS^3$ score as follows [29]. In a first step the Mascot identification is used to determine the putative $MS^3$ spectrum precursor and the theoretical $MS^3$ fragments. The peak list of the $MS^3$ spectrum is then reduced to four peaks per 100 *m/z* and the overlap of calculated and observed fragments is counted. Finally, the probability for observing these fragments by chance is calculated using a binominal distribution. If there are several $MS^3$ spectra for one $MS^2$ spectrum, MSQuant calculates all scores but retains only the highest scoring spectrum. Likewise, if there are several possible precursors (y-ion or b-ion, for example), MSQuant will consider both possibilities and retain the one with the best score. For compatibility with Mascot, we

report 10 times the negative logarithm of this probability. This $MS^3$ score has proven to be very useful and allows identification of proteins on the basis of a single peptide [30-32].

**Post-Translational Modification (PTM) Score •** For modified peptides both the identity of the peptide and the placement of the modification needs to be determined. Sometimes there is only one possibility for the location of the modification. However, often the nature of the modification and identity of the primary peptide sequence can be clearly established but the exact site of the modification may be less clear. For example, in the analysis of phosphorylation, the phosphogroup could be located on any serine, threonine or tyrosine residue in the primary sequence. The site of modification must be within the sequence stretch corresponding to the peptide, but can only be localized precisely if the corresponding, distinguishing fragments are present. Search engines such as Mascot in principle generate scores for each of the different phosphorylation sites but they do not specially score for fragments distinguishing between these possibilities.

We developed a probability score (PTM-score) based on assigning a probability that the observed fragments match the fragments calculated for a given sequence by chance. In the MSQuant framework, we first applied this score for $MS^3$ spectra as described above [29]. We then further developed the algorithm for phosphorylation matching. It iterates through all the possible modification sites and generates a score based on the number of matching fragment masses. The score also handles the placement of several phosphorylation sites in a sequence, each of which may have different probabilities and is described in [33]. We applied this score to a large scale quantitative study of protein phosphorylation. In that study more than 6 000 phosphorylation sites were identified and classified according to probability of phosphorylation site placement [33]. While developed for phosphorylation, the PTM-score can be used for any modification.

**Quantitation •** As mentioned above, one of the primary purposes of MSQuant was quantitation of SILAC pairs. In SILAC, stable isotope labeled amino acids such as $^{13}C_6$-Arg and $^{13}C_6{}^{15}N_2$-Lys are metabolically incorporated into the proteome. Two populations, the light and heavy SILAC populations, are treated differentially and mixed to compare their proteomes. Peptides occur in pairs, and the quantitative ratios between the two forms of the peptide accurately reflect the abundance of the proteins in the two proteomes.

MSQuant quantifies SILAC pairs on the basis of peptide identifications (rather than directly from the data). However, only one of the members of a SILAC pair needs to be identified and the position of the other SILAC partner is then calculated. A somewhat

complicating issue in stable-isotope coding with heavy nitrogen ($^{15}N$) is that the mass increase introduced by the label is not a fixed value (such as in SILAC), but varies depending on the total number of nitrogen atoms in the peptide. Finding peak pairs and quantifying their intensities therefore requires a different approach compared to SILAC-labels with fixed masses. To this end, MSQuant iterates through the sequence of the identified peptide (i.e. residue by residue) and computes the mass increase for each amino acid. Conversely, a mass decrease will be determined if the identified peptide is labeled. As an example, in FIGURE 1B is shown a mass spectrum of a doubly charged unlabeled and $^{15}N$-labeled peptide from *Drosophila* which contains 14 nitrogen atoms reflected by the mass difference of 7 Th. Subsequently, the monoisotopic signal of the light and heavy forms is integrated scan by scan and MSQuant displays the results in the protein quantitation window, FIGURE 4. Users can click on the result for any MS scan and inspect the corresponding raw data. Often, single scans are unreliable due to interference from co-eluting peaks, for instance. These

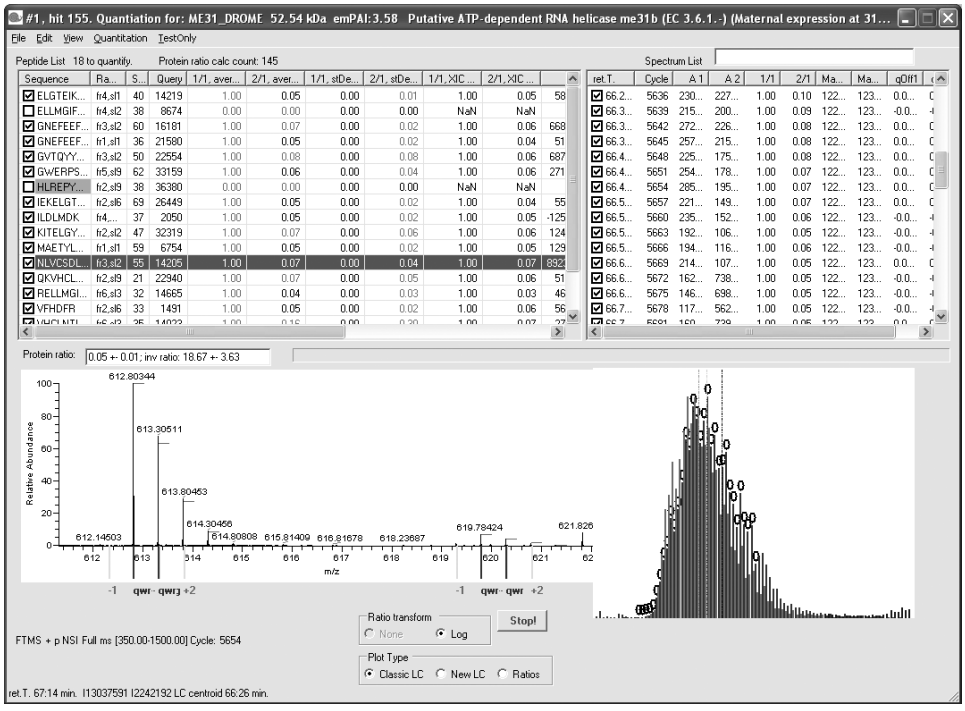


Figure 4. Screenshot of the Quantitation window in MSQuant. This window can be divided into four panels. (A) The list of identified peptides of a protein that can be quantified. (B) A list of MS scans with the retention time, scan number, intensities and ratio indicated. The raw mass spectrum of each scan can be visualized in (C) by activating (e.g. double clicking) a scan. And (D) a graph where the peak area of the peptides is plotted against the retention time.

scans can be removed from consideration under user-control.

SILAC can be performed on single or several amino acids and more than one state can be encoded. For example, in 'double-triple' SILAC labeling, both arginine and lysine are present in three different states. In 'Heavy Methyl' labeling, methionine serves as a labeled methyl group donor, which allows the quantitation of protein methylation in a site specific manner. All these instances of SILAC labeling are supported by MSQuant. In addition to SILAC and $^{15}$N-labeling, MSQuant supports quantitation of HysTag [29], $^{18}$O [34] and dimethyl [35] labeled peptides.

**Protein Correlation Profiling •** While mass spectrometry is not in itself quantitative, signals for the same peptides can be compared between runs ('label-free relative quantitation'). In a technique called 'protein correlation profiling' we made use of this feature in distinguishing background proteins from genuine organellar proteins [11]. The centrosome, an important organelle involved in organizing the microtubule network and cell division, was partially separated by centrifugation. Fractions surrounding and including the peak centrosomal fraction were digested and analyzed by LC-MS/MS. Quantitation of the XIC for each peptide through the gradients gave a characteristic profile for genuine centrosomal proteins but not for co-purifying proteins [11]. While straightforward in principle, this algorithm required matching thousands of peptides in adjacent gradient fractions on the basis of mass and retention time. To align retention times, MSQuant uses a simple linear fit of the high scoring peptides common to each run. In this way, peaks that are not sequenced in every run can nevertheless be quantified. The ability to accurately assign proteins to either centrosome or background and the observed consistency of centrosomal profiles of usually better than 30% implies excellent matching of peptides between runs. It also implies good reproducibility of the peptide signal across fractions in very complex mixtures. We have since used MSQuant to perform protein correlation profiling of all membrane bound compartments of the cell [14] and – in a two compartment comparison – to the changes in the membrane proteome of stem cells during differentiation [36].

## V. Conclusions and Perspectives

In this paper we have shown several advantages of an open and extensible software environment. Data quality and quality of manual validation is tremendously enhanced through convenient access to the raw data. We were able to quickly implement a number of novel data manipulation ideas due to access to all parts of the data processing scheme. A strong point of MSQuant is that it allows optimal use of the raw data which results in very high quality identification and quantitation. A current limitation of our software environment is that it still requires much user intervention. We are currently working on a more comprehensive 'quantitation engine', which will eliminate tedious, user-directed quantiation steps. In our opinion, technological progress in proteomics will depend just as much on algorithmic advances as on improvements in sample preparation strategies and mass spectrometric hardware. We therefore hope that MSQuant can make a contribution to the development of these crucial data analysis methods.

## Acknowledgements

Work at the Center for Experimental BioInformatics is supported by a generous grant of the Danish National Research Foundation. We thank all the members of CEBI for their contributions to the development of MSQuant. We also thank all users of MSQuant elsewhere for their active support of this project.

# References


[1] Tyers M. and Mann M., *From genomics to proteomics*. Nature, **2003**, *422*, 193-197.

[2] de Hoog C.L. and Mann M., *Proteomics*. Annu Rev Genomics Hum Genet, **2004**, *5*, 267-293.

[3] Zhu H. and Snyder M., *Protein chip technology*. Curr Opin Chem Biol, **2003**, *7*, 55-63.

[4] Aebersold R. and Mann M., *Mass spectrometry-based proteomics*. Nature, **2003**, *422*, 198-207.

[5] Patterson S.D., *Data analysis--the Achilles heel of proteomics*. Nat Biotechnol, **2003**, *21*, 221-222.

[6] Sadygov R.G., Cociorva D. and Yates J.R., 3rd, *Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book*. Nat Methods, **2004**, *1*, 195-202.

[7] Ong S.E. and Mann M., *Mass spectrometry-based proteomics turns quantitative*. Nat Chem Biol, **2005**, *1*, 252-262.

[8] Steen H. and Mann M., *The ABC's (and XYZ's) of peptide sequencing*. Nat Rev Mol Cell Biol, **2004**, *5*, 699-711.

[9] Lasonder E., Ishihama Y., Andersen J.S., Vermunt A.M.W., Pain A., Sauerwein R.W., Eling W.M.C., Hall N., Waters A.P., Stunnenberg H.G., et al., *Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry*. Nature, **2002**, *419*, 537-542.

[10] Andersen J.S., Lam Y.W., Leung A.K., Ong S.E., Lyon C.E., Lamond A.I. and Mann M., *Nucleolar proteome dynamics*. Nature, **2005**, *433*, 77-83.

[11] Andersen J.S., Wilkinson C.J., Mayor T., Mortensen P., Nigg E.A. and Mann M., *Proteomic characterization of the human centrosome by protein correlation profiling*. Nature, **2003**, *426*, 570-574.

[12] Kratchmarova I., Blagoev B., Haack-Sorensen M., Kassem M. and Mann M., *Mechanism of divergent growth factor effects in mesenchymal stem cell differentiation*. Science, **2005**, *308*, 1472-1477.

[13] Kerner M.J., Naylor D.J., Ishihama Y., Maier T., Chang H.C., Stines A.P., Georgopoulos C., Frishman D., Hayer-Hartl M., Mann M., et al., *Proteome-wide analysis of chaperonin-dependent protein folding in Escherichia coli*. Cell, **2005**, *122*, 209-220.

[14] Foster L.J., de Hoog C.L., Zhang Y., Xie X., Mootha V.K. and Mann M., *A mammalian organelle map by protein correlation profiling*. Cell, **2006**, *125*, 187-199.

[15] Gouw J.W., Pinkse M.W., Vos H.R., Moshkin Y.M., Verrijzer C.P., Heck A.J.R. and Krijgsveld J., *In vivo stable isotope labeling of fruit flies reveals post-transcriptional regulation in the maternal-to-zygotic transition*. Unpublished work, **2008**.

[16] Vermeulen M., Mulder K.W., Denissov S., Pijnappel W.W., van Schaik F.M., Varier R.A., Baltissen M.P., Stunnenberg H.G., Mann M. and Timmers H.T., *Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4*. Cell, **2007**, *131*, 58-69.

[17] Romijn E.P., Christis C., Wieffer M., Gouw J.W., Fullaondo A., van der Sluijs P., Braakman I. and Heck A.J., *Expression clustering reveals detailed co-expression patterns of functionally related proteins during B cell differentiation: a proteomic study using a combination of one-dimensional gel electrophoresis, LC-MS/MS, and stable isotope labeling by amino acids in cell culture (SILAC)*. Mol Cell Proteomics, **2005**, *4*, 1297-1310.

[18] Gygi S.P., Rist B., Gerber S.A., Turecek F., Gelb M.H. and Aebersold R., *Quantitative analysis of complex protein mixtures using isotope-coded affinity tags*. Nat Biotechnol, **1999**, *17*, 994-999.

[19] Ong S.E., Blagoev B., Kratchmarova I., Kristensen D.B., Steen H., Pandey A. and Mann M., *Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics*. Mol Cell Proteomics, **2002**, *1*, 376-386.

[20] Listgarten J. and Emili A., *Statistical and computational methods for comparative proteomic profiling*

*using liquid chromatography-tandem mass spectrometry*. Mol Cell Proteomics, **2005**, *4*, 419-434.

[21] Gouw J.W., Tops B.B.J., Mortensen P., Heck A.J.R. and Krijgsveld J., *Optimizing identification and quantitation of $^{15}$N-labeled proteins in comparative proteomics*. Anal. Chem., **2008**, *80*, 7796-7803.

[22] Nesvizhskii A.I. and Aebersold R., *Interpretation of shotgun proteomic data: the protein inference problem*. Mol Cell Proteomics, **2005**, *4*, 1419-1440.

[23] Pedrioli P.G., Eng J.K., Hubley R., Vogelzang M., Deutsch E.W., Raught B., Pratt B., Nilsson E., Angeletti R.H., Apweiler R., et al., *A common open representation of mass spectrometry data and its application to proteomics research*. Nat Biotechnol, **2004**, *22*, 1459-1466.

[24] Gentleman R.C., Carey V.J., Bates D.M., Bolstad B., Dettling M., Dudoit S., Ellis B., Gautier L., Ge Y., Gentry J., et al., *Bioconductor: open software development for computational biology and bioinformatics*. Genome Biol, **2004**, *5*, R80.

[25] R Development Core Team *R: A Language and Environment for Statistical Computing*: Vienna, Austria, 2007.

[26] Olsen J.V., de Godoy L.M., Li G., Macek B., Mortensen P., Pesch R., Makarov A., Lange O., Horning S. and Mann M., *Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap*. Mol Cell Proteomics, **2005**, *4*, 2010-2021.

[27] de Godoy L.M., Olsen J.V., de Souza G.A., Li G., Mortensen P. and Mann M., *Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system*. Genome Biol, **2006**, *7*, R50.

[28] Beausoleil S.A., Jedrychowski M., Schwartz D., Elias J.E., Villen J., Li J., Cohn M.A., Cantley L.C. and Gygi S.P., *Large-scale characterization of HeLa cell nuclear phosphoproteins*. Proc Natl Acad Sci U S A, **2004**, *101*, 12130-12135.

[29] Olsen J.V. and Mann M., *Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation*. Proc Natl Acad Sci U S A, **2004**, *101*, 13417-13422.

[30] Adachi J., Kumar C., Zhang Y., Olsen J.V. and Mann M., *The human urinary proteome contains more than 1500 proteins, including a large proportion of membrane proteins*. Genome Biol, **2006**, *7*, R80.

[31] Pilch B. and Mann M., *Large-scale and high-confidence proteomic analysis of human seminal plasma*. Genome Biol, **2006**, *7*, R40.

[32] de Souza G.A., Godoy L.M. and Mann M., *Identification of 491 proteins in the tear fluid proteome reveals a large number of proteases and protease inhibitors*. Genome Biol, **2006**, *7*, R72.

[33] Olsen J.V., Blagoev B., Gnad F., Macek B., Kumar C., Mortensen P. and Mann M., *Global, in vivo, and site-specific phosphorylation dynamics in signaling networks*. Cell, **2006**, *127*, 635-648.

[34] Reubsaet L., Kool J., Wesseldijk F., Gouw J.W., Mohammed S., Jansen J.W.A., Maravilha R.T., Zijlstra F.J., Heck A.J.R. and van Hilten J., *Quantitative proteome analysis of human blister fluids using $^{18}$O labeling for the identification of potential biomarkers of inflammation in Complex Regional Pain Syndrome*. Unpublished work, **2008**.

[35] Lemeer S., Jopling C., Gouw J.W., Mohammed S., Heck A.J., Slijper M. and den Hertog J., *Comparative phosphoproteomics of zebrafish Fyn/Yes morpholino knockdown embryos*. Mol Cell Proteomics, **2008**.

[36] Foster L.J., Zeemann P.A., Li C., Mann M., Jensen O.N. and Kassem M., *Differential expression profiling of membrane proteins by quantitative proteomics in a human mesenchymal stem cell line undergoing osteoblast differentiation*. Stem Cells, **2005**, *23*, 1367-1377.

CHAPTER 6

# Comparison of label-free and metabolic stable isotope labeling

Joost W. Gouw[†], Johannes P.C. Vissers[‡], Scott J. Geromanos[§], Albert J.R. Heck[†], James I. Langridge[‡] and Jeroen Krijgsveld[†]

[†]Biomolecular Mass Spectrometry and Proteomics Group, Bijvoet Center for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands, [‡]Waters Corporation, Atlas Park, Manchester, Great Britain and [§]Waters Corporation, Milford, Massachusetts, United States.

# I. Abstract

Comparative proteomics has emerged as a powerful approach to determine differences in protein abundance between biological samples. Stable-isotopes are commonly used as internal standards to facilitate quantitative proteomics, but quantitation based on label-free approaches are gaining in popularity. Here, we have investigated the correlation between metabolic labeling based on stable isotope $^{15}$N-labeling and label-free quantitation based on spectral peak areas. We show that the label-free approach identified more peptides, but similar numbers of proteins were identified. Even though completely different systems were used we show that peptide separation is of high quality and similar in both situations. However, although there was some overlap between the identified peptides and proteins, sample complexity is too high for either approach to detect and/or identify all peptides from the investigated (sub) proteome. By directly comparing the protein expression levels obtained by both methods, we show that there is a weak positive correlation between differential protein abundance levels. It is anticipated that the correlation will improve with the analysis of more samples.

# II. Introduction

Comparative proteomics aims at the global analysis of protein expression levels between two or more different proteomes. Until recently, the method of choice for expression proteomics was two-dimensional gel electrophoresis (2D-PAGE) but this has shifted in the last few years towards liquid chromatography mass spectrometry (LC-MS) based approaches [1]. Whereas 2D-PAGE relies on comparing spot densities from separated intact proteins, in a typical LC-MS experiment usually peptide abundances are compared to obtain relative protein expression levels. Many different approaches have been developed over the past decade to compare peptide abundances and differential stable isotope labeling is widely used to determine relative abundance ratios of peptides. In effect, the stable isotope is introduced in one sample which then serves as an internal standard in another sample after they are mixed in usually a 1:1 ratio. After digestion and mass spectrometric analysis, the intensity ratio between the unlabeled peptide and the unlabeled peptide accurately reflects the change in expression level. In addition, the internal standard accounts for experimental and technical variation.

Stable isotopes can be metabolically introduced in tissue culture using labeled amino acids [2-4] or in organisms by growing them on minimal media in which a labeled atom is enriched. The latter method has successfully been used to $^{15}$N-label various simple organisms like yeast, *E. coli* and algal cells [5, 6] which in turn can be used to label higher organisms such as fruit flies, worms and rats, respectively [7-9]. Both these methods have proven to be accurate and reproducible for quantitative proteomic analysis. However, it is not always feasible to metabolically incorporate such a label, especially in the case of human proteomics. In these situations alternative labeling strategies are used that introduce an isotopic label at different positions in the quantitative workflow. For instance, ICAT can be used to label at the protein level [10] , proteins can be digested in $^{18}$O-labeled water to label peptides [11, 12], or iTRAQ, TMT and dimethyl labeling can be used to label at the peptide level [13-15].

An alternative method that can be used to accomplish quantitative proteomics without the need to use stable isotope labeling is label-free quantitation. Essentially two complementary methods exist that facilitate label-free quantitation of which one relies on counting and comparing the number of MS/MS events per protein. The number of tandem mass spectra that identify a protein increases with the abundance of that protein and therefore can be used as a measure of the protein's abundance [16, 17]. Several different

methods have been developed to determine protein expression levels based on spectral counting and include the APEX [18] and emPAI [19] methods. Although spectral counting has proven to be very accurate at measuring large changes between proteins, accuracy drops dramatically when smaller changes are estimated [20]. The other approach depends on extracting and integrating ion chromatograms from full MS scans for relative abundance measurements. Spectral peak intensities of peptides show good correlation with protein abundances in complex mixtures [21-23]. This approach can be used both for low and high resolution data [24] but the latter is preferred since it increases specificity by reducing the number of similar (hence interfering) but distinct masses. The advantages of these two label-free methods are that relatively simple analytical protocols are needed and that they can be applied to virtually any sample.

The performance of these different methods for quantitative proteomics has been evaluated in a number of studies where changes in protein abundances were directly compared between different methods. For instance, Oda *et al.* compared spectral counting and label-free ion intensity quantitation and showed that ratios obtained from both methods agreed well with each other. Nevertheless they suggested that these label-free methods are less sensitive than isotopic labeling [20].This is further investigated by Hendrickson *et al.* who compared spectral counting with metabolic stable isotope labeling and showed that metabolic labeling is certainly more sensitive than spectral counting. In addition, they also concluded that spectral counting performs poorly when counts are low, but performs quite well when counts and signal-to-noise are high [25]. These results are corroborated by other groups who compared similar methods [26, 27]. Nevertheless, label-free quantitation remains attractive and subject of active research [28].

In this investigation, we evaluated the correlation between label-free quantitation based on spectral peak areas of peptides and metabolic labeling based on stable isotope $^{15}$N-labeling by directly comparing protein expression levels obtained by both methods. A complex mixture consisting of unlabeled and $^{15}$N-labeled peptides from *Drosophila* embryos was prepared and analyzed in triplicate using Orbitrap and Q-ToF mass spectrometers for the metabolic labeling and label-free quantitation, respectively. We demonstrate a weak positive correlation between protein abundance ratios derived from the label-free and metabolic labeling quantitation approaches. The preliminary results from the investigations described here are part of a larger experiment, and we expect that the correlation between these methods will improve when more samples will be analyzed.

# III. Materials and methods

**Sample Preparation** • Labeling, embryo collection and sample preparation was performed as described previously [29]. Briefly, equal amounts of labeled and unlabeled embryos were mixed and lysed in 8 M urea and 50 mM ammonium bicarbonate. Cellular debris was pelleted by centrifugation at 20 000 *g* for 20 min. Prior to digestion, proteins were reduced with 1 mM DTT and alkylated with 2 mM iodoacetamide. The mixture was diluted 4-fold to 2 M urea using 250 µL of 50 mM ammonium bicarbonate and 50 µL of trypsin solution, 0.1 mg/mL, and incubated overnight at 37 °C.

**Strong Cation Exchange** • Strong cation exchange was performed using a Zorbax Bio-SCX-Series II column (0.8 mm i.d. × 50 mm length, 3.5 µm), a FAMOS autosampler (LC-Packings, Amsterdam, The Netherlands), an LC-9A binary pump and an SPD-6A UV-detector equipped with a micro UV-cell (V = 0.6 µL) (Shimadzu, Tokyo, Japan). Prior to SCX chromatography, protein digests were desalted using a small plug of $C_{18}$ material (3 M Empore $C_{18}$ extraction disk) packed into a GELoader tip (Eppendorf) similar to what has been previously described [30], onto which ~10 µL of Aqua $C_{18}$ (5 µm, 200 Å) material was placed. The eluate was dried and subsequently reconstituted in 20% acetonitrile and 0.05% formic acid. After injection, a linear gradient of 1% $min^{-1}$ solvent B (500 mM KCl in 20% acetonitrile and 0.05% formic acid, pH 3.0) at a flow rate of 50 µL/min. was started. A total of 24 SCX fractions (1 min each, i.e., 50 µL elution volume) were manually collected and dried in a vacuum centrifuge.

**Nanoflow-HPLC-Orbitrap-MS** • Dried SCX fraction 16 was reconstituted in 50 µL of 0.1 M acetic acid of which 3 µL was used for analysis. The three replicates of the sample were analyzed by nanoflow liquid chromatography using an Agilent 1100 HPLC system (Agilent Technologies) coupled on-line to a LTQ-Orbitrap mass spectrometer (Thermo Scientific, San Jose, CA). The liquid chromatography part of the system was operated in a setup essentially as described previously [31]. Aqua $C_{18}$, 5 µm, (Phenomenex, Torrance, CA) resin was used for the trap column, and ReproSil-Pur $C_{18}$-AQ, 3 µm, (Dr. Maisch GmbH, Ammerbuch-Entringen, Germany) resin was used for the analytical column. Peptides were trapped at 5 µL/min. in 100% solvent A (0.1 M acetic acid in water) on a 25 mm trap column (100 µm i.d., packed in-house) and eluted to a 25 cm analytical column (50 µm i.d., packed in-house) at ~150 nL/min. in a 120-min gradient from 10 to 40% solvent B (0.1 M acetic acid in 8/2 (v/v) acetonitrile/water). The eluent was sprayed via standard coated

emitter tips (New Objective, Woburn, MA), butt-connected to the analytical column. All analyses were performed in positive mode ESI. The mass spectrometer was operated in data-dependent mode, automatically switching between MS and MS/MS. Full scan MS spectra (from $m/z$ 300 to 1500) were acquired in the Orbitrap with a resolution of 60 000 at $m/z$ 400 after accumulation to target value of 500 000. The three most intense ions at a threshold above 5000 were selected for collision-induced fragmentation in the linear ion trap at normalized collision energy of 35% after accumulation to a target value of 10 000.

**Nanoflow-HPLC-Q-ToF-MS** • Nanoscale LC separation of tryptic peptides from SCX fraction 16 was performed with a nanoACQUITY system (Waters Corporation, Milford, MA), equipped with a Symmetry $C_{18}$ 5 µm, 5 mm x 300 µm precolumn and an Atlantis $C_{18}$ 1.7 µm, 25 cm x 50 µm analytical reversed phase column (Waters Corporation). The samples, 5 µL full loop injection, were initially transferred with an aqueous 0.1% formic acid solution to the precolumn at a flow rate of 15 µL/min for 1 min. Mobile phase A was water with 0.1% formic acid whilst mobile phase B was 0.1% formic acid in acetonitrile. After desalting and preconcentration, the peptides were eluted from the precolumn to the analytical column and separated with a gradient of 3% to 40% mobile phase B over 100 minutes at a flow rate of 150 nL/min, followed by a 5 minute rinse with 95% of mobile phase B. The column was re-equilibrated at initial conditions for 20 minutes. The column temperature was maintained at 35 °C. The lock mass compound, [$Glu_1$]-Fibrinopeptide B, was delivered by the auxiliary pump of the LC system at 250 nL/min at a concentration of 100 fmol/µL to the reference sprayer of the NanoLockSpray source of the mass spectrometer. SCX fraction 16 was analyzed in triplicate.

Mass spectrometric analysis of tryptic peptides was performed using a Q-Tof Premier mass spectrometer (Waters Corporation, Manchester, UK). For all measurements, the mass spectrometer was operated in v-mode with a typical resolution of at least 10 000 FWHM. All analyses were performed in positive mode ESI. The time-of-flight analyzer of the mass spectrometer was externally calibrated with a NaI mixture from $m/z$ 50 to 1990. The data were post-acquisition lock mass corrected using the doubly charged monoisotopic ion of [$Glu_1$]-Fibrinopeptide B. The reference sprayer was sampled with a frequency of 60 s. Accurate mass LC-MS data was collected in an alternating, low energy and elevated energy mode of acquisition. The spectral acquisition time in each mode was 1.0 s with an 0.1 s interscan delay. In low energy MS mode, data was collected at constant collision energy of 4 eV. In elevated energy MS mode, the collision energy was ramped from 12 eV to 35 eV during each 1.0 s integration. One cycle of low and elevated energy data was acquired

every 2.2 s. The RF amplitude applied to the quadrupole mass analyzer was adjusted such that ions from *m/z* 300 to 2000 were efficiently transmitted; ensuring that any ions observed in the LC-MS data less than *m/z* 300 were known to arise from dissociations in the collision cell.

**Sequence Database and Common Search Parameters** • Protein identifications were obtained by searching a *D. melanogaster* species-specific database (downloaded 2007-11-20; containing 22 785 sequences) to which a randomized version of the database was appended including human keratins and trypsin. In both database search engines the maximum number of missed cleavage sites was set to 1, carbamidomethyl cysteine was set as a fixed modification and oxidized methionine as a variable modification. Trypsin was specified as the proteolytic enzyme. Moreover, identification of a protein had to occur in at least 2 out of 3 injections of the same condition.

**Orbitrap Peptide Identification and Quantitation** • All $MS^2$ spectra were converted to single DTA files using Bioworks Browser 3.2 (Thermo) and merged into a Mascot generic format file which was searched using an in-house licensed Mascot v2.2.03 search engine (Matrix Science) with the quantitation field set to 15N Metabolic to identify both labeled and unlabeled peptides. The mass tolerance of the precursor ion was set to 5 ppm and that of fragment ions was set to 0.6 Da. Peptide and protein quantitation was performed using the open source program MSQuant [32]. Briefly, peptide ratios were obtained by calculating the extracted ion chromatograms (XIC) of the unlabeled and labeled forms of the peptide using the monoisotopic peaks only. The total XIC for each of the peptide forms was obtained by summing the XIC in consecutive MS cycles for the duration of their respective LC-MS peaks in the total ion chromatogram using FT-MS scans. These total XICs were then used to compute the peptide ratio.

**Q-ToF Peptide Identification and Quantitation** • Data independent, alternate scanning LC-MS data were processed and searched using ProteinLynx GlobalServer v2.3. The ion detection, data clustering and normalization of the multiplexed, data independent LC-MS data has been explained in previous reports [33]. Briefly, the lock mass corrected spectra are first centroided, deisotoped, and charge-state-reduced to produce a single accurately mass measured monoisotopic mass for each peptide and the associated fragment ions. The correlation of a precursor and a potential fragment ion is initially achieved by means of time alignment, followed by a further correlation process during the database searches that is based on the physicochemical properties of peptides when they undergo

collision induced fragmentation.

Q-ToF search specific criteria (besides the common criteria, see above) included that peptide and fragment ion tolerances were determined automatically by the software (typically 10 ppm precursor and 20 ppm fragment ion). The protein identifications were based on the detection of at least 3 fragment ions per peptide, with a minimum of 7 fragment ions per protein and initially 1 peptide identified to a protein.

Quantitative results for the label-free analysis approach were obtained by comparing the peak area/intensity of each peptide within the two investigated fractions. Automatic-normalization was applied to the complete data set. Briefly, for complex samples, it is often possible to measure and correct for systematic errors taking into account slight differences in protein loading amount without using an internal standard. The assumption is that changes in protein expression occur against a dominant background of proteins which are unaffected by the perturbation being studied. Each peptide or cluster is initially treated as an internal standard by the quantification algorithm. During this step, peptides showing real changes are naturally suppressed as it occurs for inappropriate assignments or interferences in normal quantification. After this procedure, the entire data set is corrected and quantified.

# IV. Results and Discussion

**Experimental Design** • The strategy used to compare protein expression ratios obtained from metabolic labeling and label-free quantitation is illustrated in FIGURE 1. We started out with combining unlabeled and $^{15}$N-labeled *Drosophila* embryos that differed in developmental age. Unlabeled ('light') embryos were collected for a period of 90 minutes (0-90 min.) whereas $^{15}$N-labeled ('heavy') embryos were allowed to develop for 60 minutes after a similar collection span (60-150 min.). Although there is a small overlap between the two stages (i.e. 30 minutes), visual verification of the collected embryos showed that 90% of the light embryos were not older than 70 minutes (< stage 3) and 95% of the heavy embryos were older than 90 minutes (> stage 4). This indicates that there is in fact little overlap between the samples (data not shown). Intact labeled and unlabeled embryos were combined (FIG. 1A), proteins extracted and proteolytically digested using trypsin followed by strong cation exchange (SCX) chromatography. In total, 37 one-minute SCX fractions were collected but 4 fractions that contained the highest number of peptides were selected for further analysis (FIGURE 1B, grey bars). Each of these four fractions was analyzed in triplicate by subjecting them to reversed-phase nano LC-MS/MS using an LTQ-Orbitrap mass spectrometer operated in data-depended acquisition (DDA) mode (left track FIGURE 1C). The same SCX fraction was also analyzed in sextuple using a similar reversed-phase nano LC setup but different mass spectrometer, namely a Q-ToF Premier operated in MS$^E$ mode (right track FIGURE 1C). In the metabolic labeling approach, proteins were identified and quantified using the database search engine Mascot and quantitation software MSQuant, respectively. Protein ratios were determined in triplicate by comparing the mass spectrometric response of the light peptide to that of the heavy peptide in the same mass spectrum, left track in FIGURE 1C. Alternatively, in the case of label-free quantitation proteins were identified and quantified by using the software package ProteinLynx GlobalServer. Here, protein ratios were obtained by comparing the intensity of the unlabeled peptide from one LC-MS run to that of the labeled peptide in the consecutive LC-MS run, right track in FIGURE 1C. By using six LC-MS analyses of the same sample, protein ratios can be determined in triplicate.

Since our goal is to compare protein expression ratios, we reduced experimental variation by using the same methods wherever possible, or kept equipment and methods similar. For instance, by analyzing the same SCX fractions in both strategies any possible sample variation was eliminated. However, some experimental settings were inevitable dif-

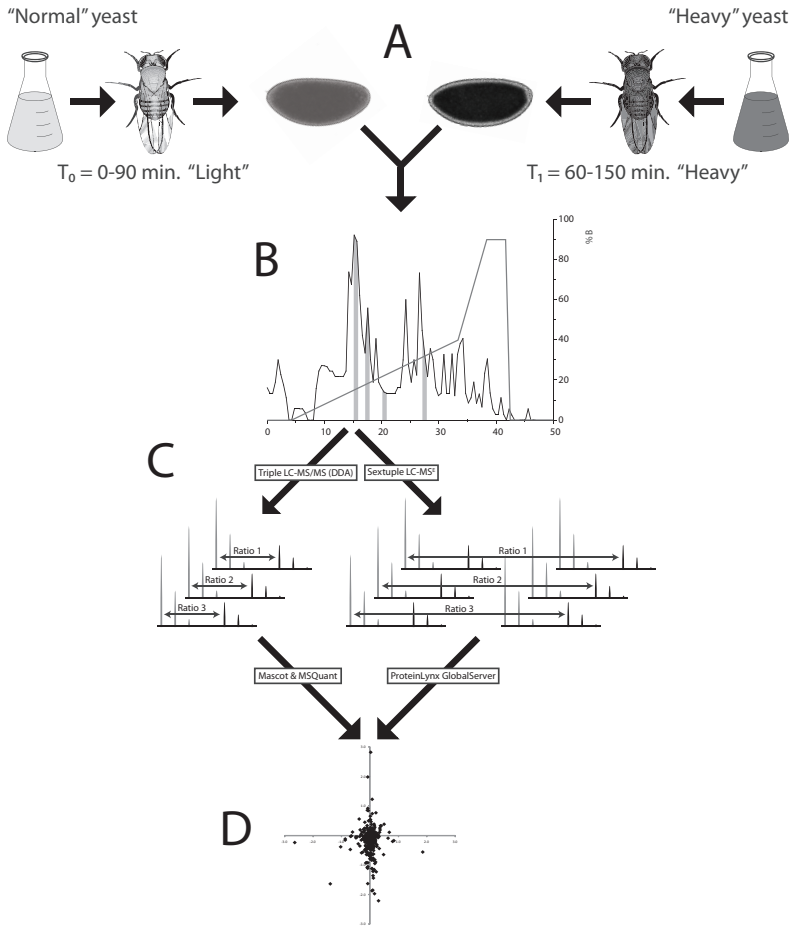


Figure 1. Experimental design used to compare label-free and metabolic labeling. (A) Labeled and unlabeled yeast was used to grow embryos that were mixed, lysed and digested. The resulting peptide mixture was subjected to (B) strong cation exchange chromatography. One fraction was (C) analyzed in triplicate (red) or sextuple (green) in case of metabolic labeling quantitation or label-free quantitation, respectively. Protein ratios obtained with both methods (D) were used for further analysis.

ferent. Notably, (*i*) two different liquid chromatography systems were used to separate and deliver peptides to the mass spectrometer, but variation was minimized by using similar analytical columns, gradients and flow rates, see below. Fundamental differences between the approaches dictated the use of (*ii*) different mass spectrometers and subsequently (*iii*) different post-acquisition processing software such as database search engines. Data obtained by both mass spectrometers is not interchangeable and therefore $MS^E$ data, obtained by a Q-ToF [33], can only be processed by using ProteinLynx GlobalServer (PLGS). Vice versa, the processing of data obtained by an Orbitrap operated in DDA mode cannot be

done by PLGS and is restricted to different post-acquisition processing software such as Mascot. At this moment we have to stress that we are not investigating differences between mass spectrometers and/or database search engines but our primary goal is to compare protein expression levels as determined by label-free and metabolic labeling as is depicted in the final part (D) of FIGURE 1.

Although four SCX fractions were analyzed in total, the results described below are preliminary and based on the results from SCX fraction 16 only. Restricted time availability hampered the analyses of the remaining fractions. The combined results from all four SCX fractions will be processed and published at a later stage.

**Peptide and Protein Identifications •** Despite the fact that different systems were used in both approaches, an interesting observation appears from the number of detected precursors. Regardless of the used strategy (Orbitrap DDA or Q-ToF $MS^E$) similar numbers of precursors per injection (~20 000, TABLE 1) were submitted to the respective database search engines. For the Orbitrap DDA strategy this means that on average 20 000 precursors per injection (number of queries in TABLE 1) were selected for fragmentation and searched in the database search engine Mascot. Similarly, PLGS detected equal numbers of non-redundant precursors in the low energy mass spectra that were searched. Although there seems to be similarity between these numbers, peptide and protein identifications show that there are differences.

**Table 1. Number of identified peptides and proteins from SCX fraction 16**

| | Orbitrap DDA | | | Q-ToF $MS^E$ | | |
|---|---|---|---|---|---|---|
| Injection | # Queries | # Unique peptides[a] | # Proteins | # Clusters[b] | # Unique peptides[a] | # Proteins |
| 1 | 20 787 | 1 387 | 548 | 20 721 | 2 702 | 566 |
| 2 | 21 028 | 1 446 | 548 | 21 245 | 2 244 | 556 |
| 3 | 20 789 | 1 436 | 547 | 19 998 | 2 430 | 521 |
| Total | 62 604 | 1 412[c] | 552[c] | 61 694 | 2 498[c] | 512[c] |

[a] Uniquely identified peptides with a minimum probability of 95%. [b] De-isotoped and charge stated reduced feature (accurate mass – retention time pair). [c] Identified in at least 2 out of 3 injections

The first observed difference between the two approaches comes from the number of identified peptides. Searching the Orbitrap data resulted in on average 7% unique peptide identifications per injection (TABLE 1) with a minimum peptide probability of 95%. In total 1412 unique peptides were identified in at least 2 out of 3 injections. Using the same criteria, the Q-ToF $MS^E$ strategy yielded more peptide identifications (~12%) with

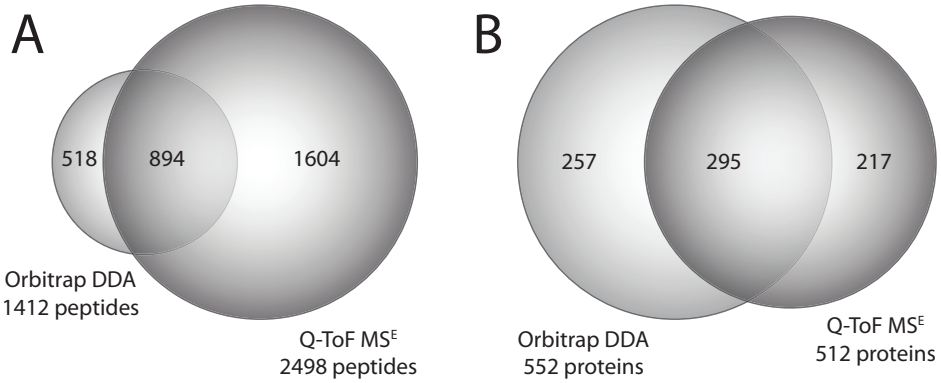


Figure 2. Venn diagrams of the identified peptides and proteins. (A) Of the 1 412 Orbitrap identified peptides, 894 were also identified by the Q-ToF MSE approach. These peptides clustered into (B) 552 and 512 proteins, respectively, of which 295 were identified by both methods.

a total of 2498 uniquely identified peptides in at least two injections. These results could be partially explained by the more independent nature of the applied label-free scanning method. Here, all co-eluting precursors at any given time are fragmented simultaneously, whereas the Orbitrap selects precursors for fragmentation based on user defined criteria. Due to the complex nature of these samples, more precursors co-elute than can be selected and ultimately sequenced by the DDA method, resulting in a reduced set of identified peptides. Alternatively, these differences could be the result of the different database search engines that were used. Nevertheless, the overlap between the uniquely identified peptides, shown in FIGURE 2A, revealed that a large proportion of the peptides were identified by both methods. The fact that there are 518 peptides identified by the Orbitrap and 1604 by the Q-ToF could be the result of several factors, including the above mentioned discrepancy between data independent and dependent acquisition and the database search engines. On the other hand, the different mass spectrometers also have impact on the observed differences. Although all these experimental factors influence the number of identifications to some extent, the most likely factor that introduces most differences is the complexity of the sample. These results indicate that neither approach is capable of detecting and/or identifying all precursors that are co-eluting. The overlap will probably increase when the separation power increases or when a less complex sample is analyzed and compared.

Peptide identifications were generally of high quality, reflected by the large overlap between the two strategies (FIGURE 2A). This is evidenced further by looking at the retention times of the 894 overlapping peptides. When these retention times are plotted against each other, a good linear relationship between them is observed ($R^2$ of 0.86), shown in FIG-

URE 3. It can be noted, however, that the retention time of the peptides is not exactly similar (the slope of the trend is not equal to 1, but 0.6 instead), but nevertheless a linear relationship indicates that these small differences are negligible and were probably introduced by the different liquid chromatography systems. Moreover, the fact that these peptides elute at comparable times also indicates that the gradient is analogous in both approaches even though completely different LC systems were used in both strategies.

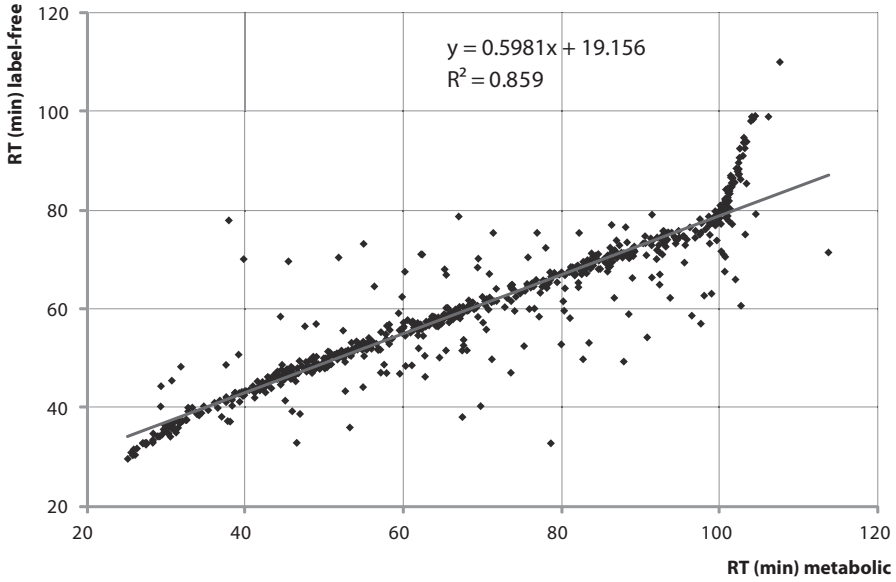


Figure 3. Scatter plot of the retention times of 894 commonly identified peptides. A linear relationship ($R^2$ of 0.86) indicates that peptide retention times are in good agreement.

The uniquely identified peptides clustered into 552 and 512 unique proteins that were identified in at least 2 injections by the DDA and $MS^E$ method, respectively (TABLE 1). This indicates that although more peptides were identified in the label-free approach, similar numbers of proteins were identified. The ratio of identified peptides per protein is higher in the $MS^E$ approach (4.7) than in the DDA method (2.5), which is again expected from the data independent nature of the label-free approach. However, the overlap of identified proteins between the two methods, shown in FIGURE 2B, indicates that these methods are in fact complementary as only ~55% of the proteins (295) were identified by both methods. Although these differences are interesting, the numbers are likely going to change as the results from the other three SCX fractions are included. Therefore, we will not focus on the differences between protein identifications in more detail at this stage. Moreover, our

aim of this project was to compare protein expression levels. The 295 overlapping proteins are therefore used for further quantitative analysis and the other proteins disregarded from any additional processing.

**Peptide and Protein Quantitation** • Interestingly, out of the 295 identified proteins, 290 could be quantitated by the metabolic labeling approach in at least 2 injections. The remaining 5 proteins could be quantitated but in only 1 injection, therefore these were not further used as we considered a minimum of two quantitation events per protein. The quantified 290 proteins were of high quality evidenced by the average coefficient of variation (CV) of 3%. The highest CV observed was 23% and the median was 2% which reflects nicely the accuracy and sensitivity of this approach. The label-free method, using the same 2-out-of-3 criteria, could quantitate 220 out of the 295 proteins with an average CV of 17%. The higher average CV in the label-free method is somewhat expected as the low flow rate (~150 nL/min.) applied in the experiments challenges spray stability, i.e. experimental variation, which is important if there is no internal standard present to account for these effects. Although almost 6 times higher than the metabolic labeling method, this average CV is still lower than most of the stable isotope labeling studies report [23, 34] and favorable compared to other label-free studies [17, 35].

The protein ratios obtained by both methods are plotted against each other and shown in FIGURE 4. While the majority of the data points are clustered around the centre of the plot, a number of more extreme values permit some interesting observations. For instance, most of the highly regulated proteins are only found to be regulated in one dimension (i.e. only one method found these proteins to be highly regulated) with the exception of a few proteins like Defective chorion-1 protein, CG9796 and C-terminal binding protein. These last three proteins were found highly regulated in both approaches. One reason for the observed differences between protein ratios could originate from the fact that only a few peptides per protein were used for quantitation. It is known that proteins can be post-translationally processed resulting in different ratios for peptides from different parts of the protein [29]. In addition, since it is unknown what the true protein ratios are, we cannot pinpoint causes to either method. However, the protein ratios can be compared to previous obtained ratios [29]. While in this experiment older embryos were compared to the same age embryos used in this study, the previous obtained protein ratios can be used as rough indicators. This protein ratio is given in between brackets in FIGURE 4 and it can be seen that the metabolic labeling ratios agree better with these ratios. That is, if a protein was found down- or up-regulated in the previous experiment it was similarly regulated in

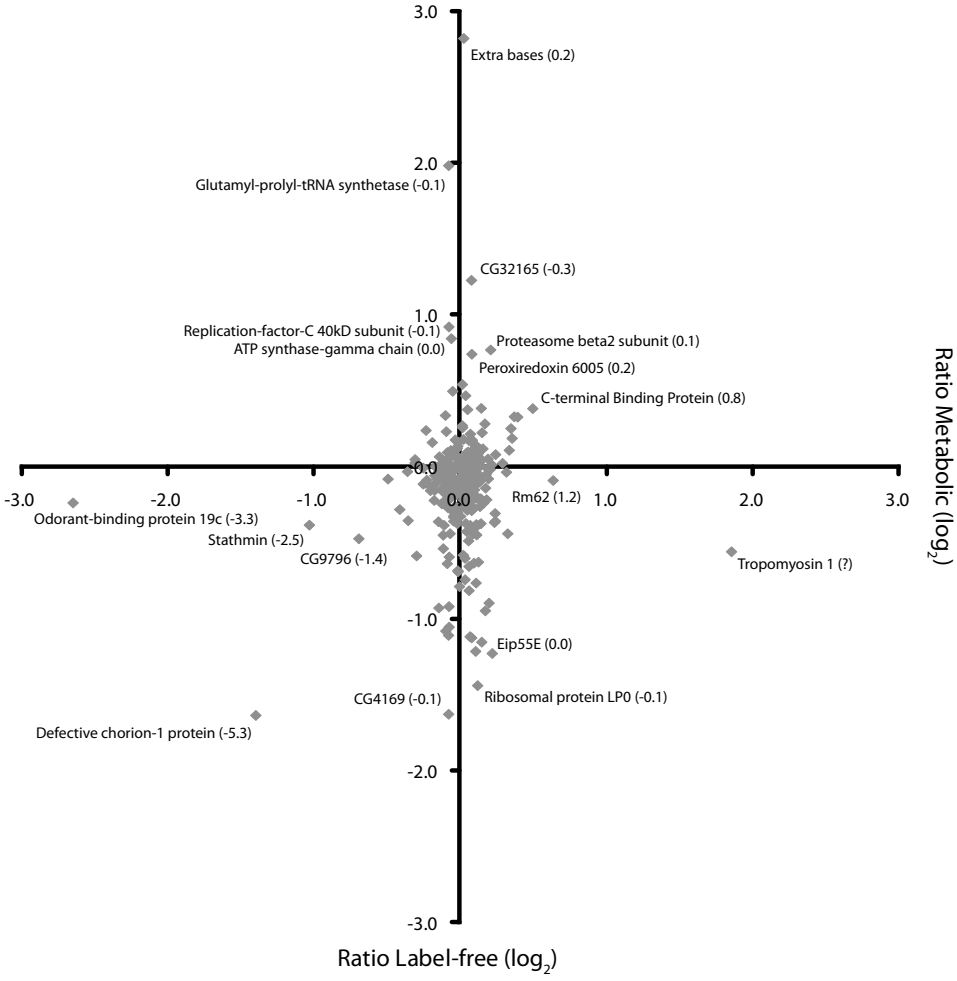


Figure 4. Correlation of protein expression levels obtained with label-free and metabolic labeling quantitation. Extreme data points are indicated with the protein name together with in brackets the expression level obtained in a previous experiment [29]. While in this experiment older embryos were compared to the same age embryos used in this study, the previous obtained protein ratios can be used as rough indicators.

this study, albeit to a lesser extent. This is in contrast to some protein ratios obtained with the label-free approach like Extra bases, Glutamyl-prolyl-tRNA synthase, CG4169 and Ribosomal protein LP0. These could be examples of the lower afforded accuracy in label-free quantitation. However, the number of peptides per protein are expected to increase with the analysis of the other SCX fractions and therefore we anticipate that the scatter plot as presented in FIGURE 4 will improve, providing more consistent intra-method protein ratios.

## V. Conclusion

Metabolic labeling of heavy stable isotopes is nowadays routinely used in quantitative proteomic applications. In addition, label-free based approaches are becoming increasingly popular to determine relative protein expression levels. By analyzing a complex peptide mixture consisting of unlabeled and $^{15}N$-labeled peptides we investigated the correlation between label-free quantitation based on spectral peak areas of peptides and metabolic labeling based on stable isotope $^{15}N$-labeling. We showed that similar numbers of proteins can be identified by both methods. By directly comparing protein expression levels obtained by both methods we showed that there is a weak positive correlation between these two approaches. However, since the results presented here are preliminary, we expect that the correlation will improve with the analysis of additional samples.

## Acknowledgement

This work was supported by The Netherlands Proteomics Centre (www.netherlandsproteomicscenter.nl).

# References


[1] Aebersold R. and Mann M., *Mass spectrometry-based proteomics*. Nature, **2003**, *422*, 198-207.

[2] Ong S.E., Blagoev B., Kratchmarova I., Kristensen D.B., Steen H., Pandey A. and Mann M., *Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics*. Mol Cell Proteomics, **2002**, *1*, 376-386.

[3] Zhu H., Pan S., Gu S., Bradbury E.M. and Chen X., *Amino acid residue specific stable isotope labeling for quantitative proteomics*. Rapid Commun Mass Spectrom, **2002**, *16*, 2115-2123.

[4] Jiang H. and English A.M., *Quantitative analysis of the yeast proteome by incorporation of isotopically labeled leucine*. J Proteome Res, **2002**, *1*, 345-350.

[5] Oda Y., Huang K., Cross F.R., Cowburn D. and Chait B.T., *Accurate quantitation of protein expression and site-specific phosphorylation*. Proc. Natl. Acad. Sci. U.S.A., **1999**, *96*, 6591-6596.

[6] Zhong H., Marcus S.L. and Li L., *Two-dimensional mass spectra generated from the analysis of $^{15}N$-labeled and unlabeled peptides for efficient protein identification and de novo peptide sequencing*. J Proteome Res, **2004**, *3*, 1155-1163.

[7] Krijgsveld J., Ketting R.F., Mahmoudi T., Johansen J., Artal-Sanz M., Verrijzer C.P., Plasterk R.H. and Heck A.J., *Metabolic labeling of C. elegans and D. melanogaster for quantitative proteomics*. Nat Biotechnol, **2003**, *21*, 927-931.

[8] Wu C.C., MacCoss M.J., Howell K.E., Matthews D.E. and Yates J.R., 3rd, *Metabolic labeling of mammalian organisms with stable isotopes for quantitative proteomic analysis*. Anal. Chem., **2004**, *76*, 4951-4959.

[9] McClatchy D.B., Dong M.Q., Wu C.C., Venable J.D. and Yates J.R., 3rd, *$^{15}N$ metabolic labeling of mammalian tissue with slow protein turnover*. J. Proteome Res., **2007**, *6*, 2005-2010.

[10] Gygi S.P., Rist B., Gerber S.A., Turecek F., Gelb M.H. and Aebersold R., *Quantitative analysis of complex protein mixtures using isotope-coded affinity tags*. Nat Biotechnol, **1999**, *17*, 994-999.

[11] Mirgorodskaya O.A., Kozmin Y.P., Titov M.I., Korner R., Sonksen C.P. and Roepstorff P., *Quantitation of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry using (18) O-labeled internal standards*. Rapid Commun. Mass Spectrom., **2000**, *14*, 1226-1232.

[12] Schnolzer M., Jedrzejewski P. and Lehmann W.D., *Protease-catalyzed incorporation of $^{18}O$ into peptide fragments and its application for protein sequencing by electrospray and matrix-assisted laser desorption/ionization mass spectrometry*. Electrophoresis, **1996**, *17*, 945-953.

[13] Ross P.L., Huang Y.N., Marchese J.N., Williamson B., Parker K., Hattan S., Khainovski N., Pillai S., Dey S., Daniels S., et al., *Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents*. Mol. Cell. Proteomics, **2004**, *3*, 1154-1169.

[14] Hsu J.L., Huang S.Y., Chow N.H. and Chen S.H., *Stable-isotope dimethyl labeling for quantitative proteomics*. Anal Chem, **2003**, *75*, 6843-6852.

[15] Thompson A., Schafer J., Kuhn K., Kienle S., Schwarz J., Schmidt G., Neumann T., Johnstone R., Mohammed A.K. and Hamon C., *Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS*. Anal Chem, **2003**, *75*, 1895-1904.

[16] Washburn M.P., Wolters D. and Yates J.R., *Large-scale analysis of the yeast proteome by multidimensional protein identification technology*. Nat Biotech, **2001**, *19*, 242-247.

[17] Pang J.X., Ginanni N., Dongre A.R., Hefta S.A. and Opiteck G.J., *Biomarker Discovery in Urine by Proteomics*. J. Proteome Res., **2002**, *1*, 161-169.

[18] Lu P., Vogel C., Wang R., Yao X. and Marcotte E.M., *Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation*. Nat Biotechnol, **2007**, *25*, 117-

124.

[19] Ishihama Y., Oda Y., Tabata T., Sato T., Nagasu T., Rappsilber J. and Mann M., *Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein*. Mol Cell Proteomics, **2005**, *4*, 1265-1272.

[20] Old W.M., Meyer-Arendt K., Aveline-Wolf L., Pierce K.G., Mendoza A., Sevinsky J.R., Resing K.A. and Ahn N.G., *Comparison of label-free methods for quantifying human proteins by shotgun proteomics*. Mol Cell Proteomics, **2005**, *4*, 1487-1502.

[21] Bondarenko P.V., Chelius D. and Shaler T.A., *Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry*. Anal. Chem., **2002**, *74*, 4741-4749.

[22] Chelius D. and Bondarenko P.V., *Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry*. J. Proteome Res., **2002**, *1*, 317-323.

[23] Wang W., Zhou H., Lin H., Roy S., Shaler T.A., Hill L.R., Norton S., Kumar P., Anderle M. and Becker C.H., *Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards*. Anal Chem, **2003**, *75*, 4818-4826.

[24] Ono M., Shitashige M., Honda K., Isobe T., Kuwabara H., Matsuzuki H., Hirohashi S. and Yamada T., *Label-free quantitative proteomics using large peptide data sets generated by nanoflow liquid chromatography and mass spectrometry*. Mol Cell Proteomics, **2006**, *5*, 1338-1347.

[25] Hendrickson E.L., Xia Q., Wang T., Leigh J.A. and Hackett M., *Comparison of spectral counting and metabolic stable isotope labeling for use with quantitative microbial proteomics*. Analyst, **2006**, *131*, 1335-1341.

[26] Zybailov B., Coleman M.K., Florens L. and Washburn M.P., *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*, 6218-6224.

[27] Usaite R., Wohlschlegel J., Venable J.D., Park S.K., Nielsen J., Olsson L. and Yates Iii J.R., *Characterization of global yeast quantitative proteome data generated from the wild-type and glucose repression saccharomyces cerevisiae strains: the comparison of two quantitative methods*. J Proteome Res, **2008**, *7*, 266-275.

[28] America A.H. and Cordewener J.H., *Comparative LC-MS: a landscape of peaks and valleys*. Proteomics, **2008**, *8*, 731-749.

[29] Gouw J.W., Pinkse M.W., Vos H.R., Moshkin Y.M., Verrijzer C.P., Heck A.J.R. and Krijgsveld J., *In vivo stable isotope labeling of fruit flies reveals post-transcriptional regulation in the maternal-to-zygotic transition*. Unpublished work, **2008**.

[30] Rappsilber J., Ishihama Y. and Mann M., *Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics*. Anal. Chem., **2003**, *75*, 663-670.

[31] Meiring H.D., van der Heeft E., ten Hove G.J. and de Jong A.P.J.M., *Nanoscale LC-MS$^{(n)}$: technical design and applications to peptide and protein analysis*. J. Sep. Sci., **2002**, *25*, 557-568.

[32] Schulze W.X. and Mann M., *A novel proteomic screen for peptide-protein interactions*. J. Biol. Chem., **2004**, *279*, 10756-10764.

[33] Silva J.C., Denny R., Dorschel C.A., Gorenstein M., Kass I.J., Li G.Z., McKenna T., Nold M.J., Richardson K., Young P., et al., *Quantitative proteomic analysis by accurate mass retention time pairs*. Anal Chem, **2005**, *77*, 2187-2200.

[34] Radulovic D., Jelveh S., Ryu S., Hamilton T.G., Foss E., Mao Y. and Emili A., *Informatics platform for global proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry*. Mol Cell Proteomics, **2004**, *3*, 984-997.

[35] Liu H., Sadygov R.G. and Yates J.R., 3rd, *A model for random sampling and estimation of relative protein abundance in shotgun proteomics*. Anal. Chem., **2004**, *76*, 4193-4201.

# CHAPTER 7

## Summary

# I. Introduction

The development of multicellular organisms is characterized by complex processes that progressively transform essentially a single cell into a creature with complicated structures and highly specialized functions. The spatio-temporal regulation of this developmental process is tightly regulated and is already laid down in the early embryo. In fact, well before the fertilization of an oocyte, the body framework is deployed by multiple mechanisms that act simultaneously to achieve protein expression at the right time, at the right dose and at the right position. This blueprint made of proteins is later, after fertilization, interpreted by mechanisms initiated by the activation of the zygotic genome. These mechanisms ensure specialization of newly formed cells that in turn move to their final destination by once again interpreting the body plan of the embryo. After reaching their ultimate location they start, when the time is right, to divide and give rise to future structures. The fruit fly *Drosophila melanogaster* provides an excellent model system to investigate these principles of life in great detail, especially after recognizing that many of the mechanisms are evolutionary conserved. Over the past few decades, *Drosophila* has been used to investigate and elucidate fundamental aspects that underlie the mechanisms described above. Although this research mainly involved large-scale analysis of genes, many principles are now understood at the gene level. However, it is nowadays clear that transcript abundances no not necessarily correlate with protein expression levels and since the latter determine the complexity of cells it is the ultimate goal to investigate and understand these processes directly at the protein level.

# II. Summary of Results

The primary goal of the work described in this thesis focused on gaining insights into early embryonic development of the fruit fly *Drosophila melanogaster* at the protein level by combining stable isotope labeling high-accuracy mass spectrometry. In **chapter 1** is given an introduction about fruit flies as model organisms, their overall development and a more focused view on embryogenesis. Moreover, an overview of presently available mass spectrometric techniques to analyze proteins is outlined which includes various liquid chromatographic approaches as well as methodology to identify peptides and proteins. Finally, different labeling strategies that facilitate the quantitative analysis of proteins are discussed. A crucial step in the accurate analysis of protein expression levels is the ability to introduce an internal standard to account for sample losses and to facilitate quantitation and the most accurate quantitative results are obtained when this label is metabolically incorporated into living cells or organisms. Therefore **chapter 2** gives an overview of the current status of the metabolic labeling of model organisms for quantitative proteomics. It is shown that stable isotopes are increasingly used to metabolic label model organisms and different methodologies to incorporate stable isotopes are presented and discussed. As a consequence, species in almost every branch of the 'tree of life' can be metabolically labeled for the purpose of quantitative proteomics. It is shown that the focus is shifting from fundamental research (i.e. developing and validating methodology) to applied research where metabolic labeled model organisms are now used to investigate various biological processes. In **chapter 3** we described the application of a proteomic approach to investigate the well defined mother-to-child transition in *Drosophila melanogaster* at the protein level. By combining stable isotope labeling *in vivo* with high accuracy quantitative mass spectrometry we have quantified 2,232 proteins with at least two peptides per protein. The expression levels of around 500 proteins increased during this transition indicating that these proteins are a product of embryonic translation. In contrast, ~120 proteins were down-regulated and this group of proteins is dominated by maternal factors involved in translational control of maternal and zygotic transcripts. To uncover post-transcriptional regulation we have directly compared protein levels to transcript abundances by performing microarray experiments on identical samples and we have shown that mRNA levels are poor indicators for protein expression levels. In addition, the mRNA levels of down-regulated proteins remained relatively constant, indicating a translational control mechanism specifically targeting these (mainly maternal) proteins. We have shown that post-transla-

tional regulation can be uncovered by interrogating individual peptide ratios of a protein which resulted in the finding of specific processing of a number of proteins. Taken together, our quantitative proteomic approach using *in vivo* labeling of fruit flies with stable isotopes combined with extensive analysis by LC-MS/MS has permitted the relative quantitation of thousands of proteins during early embryonic development. This provided insight into the production, stability and modification of individual proteins, while discrepancies between transcriptional profiles and protein dynamics indicated novel control mechanisms in genome activation during early fly development.

Alternative objectives that were addressed in the work described in this thesis aimed at developing methods to optimize, improve or facilitate the quantitative analysis of multiple proteins in a single experiment. In **chapter 4** is shown that metabolic labeling with suboptimal (< 98%) $^{15}$N-enrichments negatively affects protein identification and quantitation which was discovered by the systematic investigation of two independent $^{15}$N-labeled datasets. We have shown that the number of specifically larger identified $^{15}$N-labeled peptides is underrepresented compared to their identified $^{14}$N-counterparts. Moreover, accuracy of peptide ratios and precision of protein ratios are revealed to be compromised by incomplete or suboptimal labeling. To overcome or compensate these shortcomings we have developed methods that can be applied to qualitative and quantitative $^{15}$N-labeled data. We have shown that the precursor mass of specifically larger $^{15}$N-labeled peptides is not corresponding to the precursor's monoisotopic mass which is a crucial factor for correct peptide identification. Upon a simple correction of altering the precursor's mass into the monoisotopic mass, the number of identifications of specifically larger $^{15}$N-labeled peptides can be increased dramatically. This leads to more confident protein identifications as well as more data points that can be used for quantitation. Additionally, when the signal of the monoisotopic peak is considered for peptide ratio calculations, a substantial amount of signal is not accounted for due to suboptimal labeling. We have developed a method that uses the enrichment level and the peptide chemical formula to calculate an individual peptide correction factor that rectifies the peptide ratio. This results in more accurate quantitative data among peptides as well as more precise protein ratios. The methods described here can be applied to any type of labeling experiment with varying degrees of nitrogen enrichment.

In **chapter 5** the open-source program MSQuant is presented which allows the visualization, quantitation and validation of quantitative mass spectrometric data directly from raw MS data. Besides the validation of peptide identifications also scoring algorithms

are implemented that allow the interpretation of second stage fragment spectra ($MS^3$) and the localization of post-translational modifications. This versatile software package currently interacts with data files from Thermo Electron, ABI-Sciex and Micromass/Waters and allows the extraction of data points from these data files without losing any information. MSQuant supports relative protein quantitation including the SILAC, ICAT and $^{15}N$-labeling approaches, as well as protein correlation profiling. Due to its open-source nature, MSQuant can be modified and extended rapidly and independently of the MS manufacturer.

**Chapter 6** shows the preliminary results of a comparative study between label-free and metabolic labeling approaches for quantitative proteomics. Although metabolic labeling produces most accurate quantitative data, it is not always feasible to incorporate an internal standard metabolically into an organism. So-called label-free techniques allow the quantitation of proteins in two or more samples without the need of using an internal standard. By directly comparing protein expression levels obtained by metabolic labeling based on stable isotope $^{15}N$-labeling and label-free quantitation based on spectral peak areas we have correlated both approaches. Although differences in the numbers of peptide identifications we show that similar numbers of proteins can be identified with both methods. Interestingly, the overlap of identified proteins between the two methods is only ~55% indicating that neither approach is capable of identifying all proteins in a complex sample. We directly compared protein expression levels obtained by both methods and showed that there is a weak positive correlation between these two approaches. However, since the results presented here are preliminary we anticipate that the correlation will improve with the analysis of more samples.

# III. Conclusion

In the research described in this thesis early embryonic development of the fruit fly *Drosophila melanogaster* was studied by investigating different stages of the developing fly. We combined stable isotope labeling *in vivo* with high accuracy quantitative proteomics thereby identifying many proteins involved in early fly development. Differential regulation of numerous proteins was observed and allowed the classification of these proteins into specialized classes. Although this approach shows its strength in identifying and quantitating proteins, some drawbacks including suboptimal labeling and software processing were identified, and methods were developed to overcome these. Taken together, we have presented a powerful comprehensive proteomics approach that we have applied to early embryonic developed of the fruit fly which provided insights into the production, stability and modification of individual proteins during the developing fly.

CHAPTER 8

# Nederlandse samenvatting
# Curriculum vitae
# List of publications
# Dankwoord

# Nederlandse samenvatting

In dit hoofdstuk beschrijf ik voor niet-ingewijden het doel en de resultaten van mijn promotieonderzoek. Tevens leg ik uit wat de perspectieven en implicaties van het onderzoek zijn.

Embryonale ontwikkeling is een fascinerend proces waarbij één enkele cel uitgroeit tot een organisme met complexe structuren en gespecialiseerde functies. De embryonale ontwikkeling is een uitstekend voorbeeld van wat voor processen er zich allemaal in een cel kunnen afspelen. Als eerste wordt een eicel gemaakt en voorbereid op bevruchting, waarna de cel in leven moet blijven om te kunnen groeien (door middel van celdeling) en tijdens dat proces krijgen bepaalde cellen specifieke functies. Daarna groeien die gespecialiseerde cellen uit tot complete nieuwe structuren, zoals het hart en brein, om uiteindelijk een nieuw organisme te vormen. Al deze processen (en nog vele anderen) worden uitgevoerd door de verschillende eiwitten die zich in alle cellen bevinden. Zonder eiwitten kan een cel, en dus ook een organisme niet leven. Eiwitten vormen bijvoorbeeld het skelet van de cel, zijn verantwoordelijk voor de communicatie van zowel binnen als buiten de cel en zorgen ervoor dat een cel zich kan delen. Elk eiwit is opgebouwd uit bouwstenen, de zogenaamde aminozuren, en dit kunnen ketens zijn van wel honderden aminozuren. Er zijn in totaal 20 verschillende aminozuren dus er kunnen talloze combinaties gemaakt worden wat leidt tot een enorme diversiteit in eiwitten. Dit is ook noodzakelijk, want meestal heeft elk eiwit een specifieke functie. De volgorde van deze bouwstenen bepalen de structuur en de functie van het eiwit en dit is vastgelegd in het DNA van het organisme. Als een eiwit met een specifieke functie nodig is, wordt een stukje DNA (een gen) 'gelezen' en 'vertaald' in aminozuren. De aminozuren worden één voor één aan elkaar vastgemaakt om uiteindelijk een eiwit te vormen. Het proces van lezen en vertalen heet translatie en wordt uiteraard ook door eiwitten uitgevoerd. In een cel kunnen wel meer dan 5 000 verschillende eiwitten aanwezig zijn en al deze eiwitten hebben een bepaalde taak.

Een belangrijk proces tijdens de vroege embryonale ontwikkeling en wat eigenlijk direct na de bevruchting begint is de initiatie van extreem snelle celdeling. Zo zijn er binnen een paar uur al meer dan duizenden cellen en dit vindt plaats onder toezicht van eiwitten die verantwoordelijk zijn voor de celdeling. Enerzijds controleren deze eiwitten het celdelingproces en corrigeren waar nodig terwijl anderzijds andere eiwitten garant staan voor de aanvoer van bouwstoffen. Een misschien wel net zo belangrijk proces wat tegelijk met de celdeling plaats vindt, is de vorming van de definitieve lichaamsassen. Dit zijn de

assen, van rug naar buik bijvoorbeeld, die grotendeels bepalen waar structuren ontwikkeld moeten worden. Al tijdens de ontwikkeling van de eicel worden door eiwitten bepaald welke structuren waar komen en nog voordat een eicel bevrucht wordt staat het al vast waar bijvoorbeeld het toekomstige hoofd zich zal ontwikkelen. Het zijn in dit geval de eiwitten van de moeder die een soort van blauwprint klaar leggen zodat de eiwitten van het embryo dit kunnen interpreteren. Ze weten dan waar ze naar toe moeten gaan en alleen daar hun werk uitvoeren. De vraag is echter welke eiwitten een rol van betekenis spelen tijdens deze belangrijke processen en wat voor functie ze hebben.

Om dit in detail te bestuderen, is er een flink aantal embryo's nodig. Er is daarom gekozen voor de fruitvlieg *Drosophila melanogaster* als modelorganisme te gebruiken om deze processen te bestuderen. De fruitvlieg wordt al meer dan 100 jaar gebruikt als model voor het onderzoeken van onder andere biologische, biochemische en genetische processen. Van oudsher werd de fruitvlieg in het laboratorium gebruikt vanwege zijn hoge voortplantingssnelheid, korte ontwikkelingsperiode, eenvoudige kweekbaarheid en de mannetjes kunnen gemakkelijk van de vrouwtjes onderscheiden worden. Zo is er ontzettend veel onderzoek gedaan in de fruitvlieg naar het vinden van genen die essentieel zijn voor de (embryonale)ontwikkeling, wat zelfs geleid heeft tot het uitreiken van de Nobelprijs voor Fysiologie of Geneeskunde in 1995 voor deze studie. Het belang van de fruitvlieg als modelorganisme werd opnieuw benadrukt nadat in 1999 het DNA van de fruitvlieg in kaart gebracht werd. Zo werd al snel duidelijk dat de fruitvlieg evolutionair gezien dicht bij de mens staat. Hoewel de mens ongeveer twee keer zoveel genen heeft en daardoor uiteraard ook meer eiwitten, blijken de meeste eiwitten van de fruitvlieg ook vergelijkbare partners te hebben in de mens. De mens heeft alleen meer verschillende eiwitten die dezelfde functie hebben wat te vergelijken is met een familie. Een eiwitfamilie (verschillende eiwitten met dezelfde functie) in de fruitvlieg bestaat uit 1 of 2 eiwitten terwijl dezelfde familie in de mens uit 4 of 5 eiwitten bestaat. Interessant genoeg blijken 75% van de menselijke ziektegenen een vergelijkbaar gen te hebben in de fruitvlieg wat ertoe geleid heeft dat de fruitvlieg nu gebruikt wordt voor genetisch onderzoek naar ziektes zoals Parkinson, Huntington, Alzheimer, etc. Met dit gegeven kunnen gekarakteriseerde eiwitten in de fruitvlieg geëxtrapoleerd worden naar eiwitten in de mens. Dan blijft echter staan hoe de eiwitten in een embryo, cel of weefsel geïdentificeerd kunnen worden.

Aan de hand van de techniek massaspectrometrie kan bepaald worden welke eiwitten zich precies in een cel bevinden. Allereerst worden de eiwitten geïsoleerd en in kleinere stukjes 'geknipt' met behulp van een enzym. Dit soort enzymen (ook wel proteasen

genoemd) verbreken de binding tussen twee specifieke aminozuren waardoor een eiwit in tweeën valt. De protease blijft net zolang knippen totdat alleen kleine stukjes van het eiwit overblijven en deze stukjes noemen we peptiden. De massaspectrometer meet de massa van de peptiden ontzettend nauwkeurig en bepaalt de volgorde van de aminozuren in de peptiden. Alleen met deze twee gegevens kan het oorspronkelijke eiwit geïdentificeerd worden. Het identificeren en kwantificeren (zie hieronder) van eiwitten vallen onder de techniek proteomics. Dankzij belangrijke innovaties in de afgelopen jaren kunnen er op dit moment binnen korte tijd duizenden eiwitten tegelijk geïdentificeerd worden. Onder deze innovaties valt bijvoorbeeld de manier hoe peptiden in de massaspectrometer gebracht worden en het belang hiervan werd een paar jaar geleden benadrukt met het uitreiken van de Nobelprijs voor scheikunde aan de onderzoekers die deze technieken ontwikkelden. Dat het hele proces van identificeren nogal complex is wordt snel duidelijk wanneer bedacht wordt dat één enkel eiwit meerdere peptiden geeft, soms wel meer dan 100, en dat er duizenden eiwitten in cel zitten. Daarbuiten is de concentratie van elk eiwit ook nog eens verschillend. Zo kunnen er van een eiwit misschien 100 kopieën zijn, terwijl er van een ander eiwit miljarden kopieën zijn. Omdat het dynamisch bereik van de massaspectrometer vele malen lager is dan het dynamisch bereik van de eiwitten in een cel is het uitermate lastig om het eerstgenoemde eiwit te identificeren. Gelukkig gaan de ontwikkelingen in instrumentatie redelijk snel en is de verwachting dat binnen niet al te lange tijd ook laag abundante eiwitten gedetecteerd en geïdentificeerd kunnen worden. Dit geeft overigens wel aan dat er nog volop ontwikkelingen gaande zijn op het gebied van proteomics. Zo is het bijvoorbeeld ontzettend belangrijk dat een eiwit geïdentificeerd kan worden, maar misschien nog wel belangrijker is de mogelijkheid om te bepalen in welke mate een eiwit aanwezig is. Als er twee verschillende ontwikkelingsstadia vergeleken worden is een eiwitidentificatie op zich niet voldoende, maar wordt het pas echt interessant als het dynamisch profiel, of de relatieve expressie van datzelfde eiwit vastgesteld kan worden. Dit wordt kwantificeren genoemd en vertelt hoe het concentratieverloop van een eiwit is, neemt de hoeveelheid toe, blijft die gelijk of neemt de hoeveelheid juist af. Er is een aantal manieren ontwikkeld die het mogelijk maakt om een eiwit te kwantificeren. Zo is het mogelijk om alle eiwitten in de fruitvlieg van een tag te voorzien door middel van de fruitvliegen te laten groeien op speciaal voer. Op deze manier krijgen de vliegen op natuurlijke manier de tag binnen en wordt deze automatisch door natuurlijke processen ingebouwd in alle eiwitten. Dit wordt in vaktermen metabool labelen genoemd. De tag is een zwaardere versie van het natuurlijk element stikstof, een stabiele isotoop. Elk aminozuur bevat minstens 1

stikstofatoom dus de gelabelde peptiden zijn, afhankelijk van het aantal stikstofatomen, zwaarder dan de niet gelabelde peptiden. Als de gelabelde peptiden met ongelabelde peptiden gemixt worden, kan de massaspectrometer gemakkelijk onderscheid maken tussen beide vormen, de massa van de peptiden wordt immers gemeten, en dankzij deze truc kan achterhaald worden wat de herkomst van een peptide en dus een eiwit was. Door de intensiteit van beide peptiden met elkaar te vergelijken kan de relatieve expressie van een eiwit bepaald worden. Deze techniek wordt kwantitatieve proteomics genoemd en ik heb dit toegepast in de meeste experimenten die beschreven staan in dit proefschrift.

Het metabool labelen van complete organismen is echter nog geen standaardtechniek en daarom heb ik in hoofdstuk 2 een overzicht gemaakt van de beschikbare methoden om intacte organismen metabool te labelen. Tevens heb ik elke methode in detail besproken evenals de organismen op welke de methoden toegepast zijn. Zo laat FIGUUR 2 in Hoofdstuk 2 zien dat organismen in bijna elke tak van de 'boom des levens' metabool gelabeld kunnen worden voor kwantitatieve proteomics. Hoewel het meeste onderzoek zich initieel richtte op fundamenteel werk, zoals het ontwikkelen en valideren van nieuwe methoden, begint die focus zich te verplaatsen richting de toegepaste wetenschap. De conclusie van Hoofdstuk 2 is dat de tijd rijp is om de beschikbare methoden toe te passen op biologische systemen om meer inzicht te krijgen in de biologische processen.

Dit is dan ook precies wat in Hoofdstuk 3 beschreven staat. Ik heb fruitvliegen metabool gelabeld met zwaar stikstof en de embryo's gebruikt om een belangrijke overgang in de vroege embryonale ontwikkeling te karakteriseren op eiwitniveau. Deze overgang, genaamd maternal-to-zygotic transition (MZT), is het moment dat de moederlijke eiwitten afgebroken worden en de taken overgenomen worden door eiwitten van het embryo zelf. De MZT vindt plaats ongeveer 2 uur na bevruchting en de eiwitten van het embryo zelf worden niet eerder dan de MZT geproduceerd, wat automatisch betekent dat de moederlijk eiwitten de eerste 2 uur van de embryonale ontwikkeling controleren. Ik heb in totaal meer dan 2 000 eiwitten kunnen kwantificeren en daarvan nemen er ongeveer 500 toe in hoeveelheid. Dit zijn eiwitten die door het embryo zelf gemaakt worden. Daar tegenover staan ruim 120 eiwitten die in hoeveelheid afnemen en deze groep kan gezien worden als moederlijke eiwitten die na de MZT afgebroken worden. Ook heb ik naar de expressieniveaus van de corresponderende genen gekeken. Hieruit blijkt dat het expressieniveau van een gen niet altijd direct correleert met het expressieniveau van het eiwit. Dit laat niet alleen zien dat complexe processen verantwoordelijk zijn voor het transleren van een gen naar een eiwit, maar ook dat proteomics onmisbaar is om een gedetailleerd beeld van wat

er zich afspeelt in een cel te krijgen. Ook heb ik laten zien dat niet alle peptiden van een eiwit hetzelfde gedrag vertonen. Dit kan verklaard worden door het feit dat sommige eiwitten geactiveerd worden door een stukje van het eiwit af te halen. Door de peptiden afkomstig van dat stukje eiwit te vergelijken met de andere peptiden van hetzelfde eiwit, kan geconcludeerd worden dat sommige eiwitten geactiveerd worden tijdens de MZT. Deze studie geeft inzicht in de stabiliteit, productie en modificatie van individuele eiwitten terwijl verschillen in gen- en eiwitexpressie laten zien dat nieuwe mechanismen een belangrijke rol spelen in het activeren van het genoom van het embryo.

Zoals meestel het geval is bij het ontwikkelen van een nieuwe methode, komen er tijdens het opzetten en/of verwerken van de resultaten bepaalde kleine problemen aan het licht. Zo ook met de methode zoals beschreven in Hoofdstuk 3. Tijdens het uitwerken van de resultaten heb ik gemerkt dat voornamelijk de data afkomstig van de zware peptiden (die dus zware stikstofatomen bevatten) niet altijd dezelfde accuraatheid en juistheid heeft als de data van de niet gelabelde peptiden. Door in detail naar deze data te kijken, ben ik tot de conclusie gekomen dat de massaspectrometer de gelabelde peptiden niet precies hetzelfde behandeld als de ongelabelde peptiden. Door dit probleem te identificeren heb ik procedures ontwikkeld die hiervoor corrigeren. Dit alles staat beschreven in Hoofdstuk 4. Dankzij deze procedures wordt de kwaliteit van data een stuk beter waardoor er minder variatie is tussen de gelabelde en ongelabelde peptiden van hetzelfde eiwit. Dit heeft vooral effect op de nauwkeurigheid waarop de hoeveelheid van een eiwit bepaald wordt en na correctie kunnen we dus kleinere verschillen waarnemen.

In Hoofdstuk 5 heb ik het software pakket MSQuant beschreven. Met dit programma is het mogelijk om de relatieve hoeveelheid van eiwitten te bepalen en hoewel dit programma oorspronkelijk niet specifiek ontwikkeld was voor zwaar stikstof bevattende data, heb ik het zo aangepast dat dit nu wel mogelijk is. Ik heb dit programma dan ook veel gebruikt voor de data-analyse in Hoofdstukken 3 en 6. Naast het bepalen van de hoeveelheid van een eiwit heeft dit programma nog veel meer toepassingen. Zo kan bepaalde data gevalideerd worden en is het mogelijk om gemodificeerde peptiden te karakteriseren. Ook kan dit programma meetresultaten op een overzichtelijke manier sorteren en gemakkelijk presenteren.

Naast het labelen met zwaar stikstof zijn er nog andere methoden om de relatieve hoeveelheid van een eiwit te bepalen. In Hoofdstuk 6 heb ik een zogenaamde label-free methode vergeleken met de stikstof labelingsprocedure. Omdat niet alle organismen met zwaar stikstof (of welke andere methode dan ook) gelabeld kunnen worden, denk bijvoor-

beeld aan de mens, kan de label-free methode uitkomst bieden. Zoals de naam al suggereert, wordt er in deze methode geen gebruik gemaakt van een label, wat betekent dat de relatieve hoeveelheid van een eiwit bepaald wordt aan de hand van andere fysische eigenschappen van peptiden. Het is dan wel belangrijk te weten in welke mate deze methode presteert omdat metabool labelen over het algemeen de meest accurate data genereert. Hoewel de data zoals die gepresenteerd is in Hoofdstuk 6 slechts deel uitmaakt van een grotere set, blijkt echter wel dat de overlap tussen de geïdentificeerde eiwitten niet heel groot is. Dit geeft aan dat geen van beide methoden in staat is om alle eiwitten die aanwezig zijn in een complexe oplossing te identificeren. Misschien wel interessanter is de vergelijking tussen de relatieve hoeveelheid van de eiwitten die door beide methoden bepaald zijn. Hieruit blijkt dat er maar een kleine positieve correlatie tussen deze twee datasets bestaat. Desalniettemin neem ik aan dat deze correlatie een stuk beter wordt zodra meer data geanalyseerd wordt.

Het onderzoek zoals beschreven in dit proefschrift richt zich op de vroege embryonale ontwikkeling van de fruitvlieg *Drosophila melanogaster* door verschillende tijdspunten van de ontwikkelend e vlieg te vergelijken en bestuderen. Door het in vivo labelen met stabiele isotopen te combineren met kwantitatieve proteomics, heb ik vele eiwitten kunnen identificeren tijdens deze fase van ontwikkeling. Aan de hand van de expressieniveaus van de eiwitten heb ik deze kunnen classificeren in gespecialiseerde categorieën. Hoewel de kracht van deze techniek duidelijk is, zijn er enkele problemen geassocieerd met deze methode. Ik heb deze problemen kunnen identificeren en ik heb methoden ontwikkeld die hiervoor corrigeren. Alles samengenomen, ik heb een krachtige proteomics techniek toegepast en gebruikt om de vroege embryonale ontwikkeling van de fruitvlieg te onderzoeken welke ons inzicht geeft in de productie, stabiliteit en modificatie van individuele eiwitten tijdens deze fase van ontwikkeling.

# Curriculum vitae

De schrijver van dit proefschrift werd geboren op 19 december 1978 te Amersfoort. Na het behalen van het MAVO diploma aan het Vallei college in dezelfde stad begon hij met de Middelbare Laboratorium Opleiding (MLO) aan het ROC De Amerlanden in Amersfoort met als richting analytische chemie. Tussentijds maakte hij de overstap naar de Hogere Laboratorium Opleiding (HLO) aan de Hogeschool van Utrecht met als afstudeerrichting analytische chemie. Als onderdeel van zijn afstuderen deed hij 10 maanden onderzoek bij Hercules European Research Center te Barneveld onder leiding van Dr. Peter Burgers. Na het behalen van het diploma in januari 2002 begon hij in dezelfde maand met de Master Biomolecular Sciences aan de Universiteit van Utrecht. De minor van de Master voerde hij deels uit bij het Center for Experimental Bioinformatics (CEBI) in Odense, Denemarken en de major van de Master werd uitgevoerd in de vakgroep Biomoleculaire Massa Spectrometrie. In juni 2004 werd het Master of Science diploma gehaald waarna hij in september van datzelfde jaar starte als onderzoeker in opleiding bij de sectie Biomoleculaire Massa Spectrometrie van Prof. Dr. Albert Heck onder leiding van Dr. Jeroen Krijgsveld. De resultaten van dat onderzoek staan beschreven in dit proefschrift. Onderzoeksresultaten zijn onder meer gepresenteerd op verschillende conferenties waaronder die van de American Society for Mass Spectrometry in 2006, te Seattle in de Verenigde Staten. Sinds 15 januari 2009 is Joost werkzaam als postdoc in het 'Netherlands Proteomics Centre' op de Universiteit van Utrecht.

# List of publications

**Gouw, J.W.**, Pinkse, M.W., Vos, H.R., Moshkin, Y.M., Verrijzer, C.P., Heck, A.J.R. and Krijgsveld, J. *In vivo stable isotope labeling of fruit flies reveals post-transcriptional regulation in the maternal-to-zygotic transition.* **2008** Submitted.

**Gouw, J.W.**, Tops, B.B.J., Mortensen, P., Heck, A.J.R. and Krijgsveld, J. *Optimizing identification and quantitation of 15N-labeled proteins in comparative proteomics.* Anal Chem **2008**, *80*, 7796-7803.

**Gouw, J.W.**, Heck, A.J.R. and Krijgsveld, J. *Quantitative proteomics in model organisms.* **2008** Manuscipt in preparation.

**Gouw, J.W.**, Vissers, J.P.C., Geromanos, S.J., Langridge, J.I., Heck, A.J.R. and Krijgsveld, J. *Comparison of label-free and metabolic stable isotope labeling.* **2008** Manuscipt in preparation.

Lemeer, S., Jopling, C., **Gouw, J.W.**, Mohammed, S., Heck, A.J., Slijper, M. and den Hertog, J. *Comparative phosphoproteomics of zebrafish Fyn/Yes morpholino knockdown embryos.* Mol Cell Proteomics **2008**, *7*, 2176-2187.

Pinkse, M.W., Mohammed, S., **Gouw, J.W.**, van Breukelen, B., Vos, H.R. and Heck, A.J. *Highly robust, automated, and sensitive online TiO2-based phosphoproteomics applied to study endogenous phosphorylation in Drosophila melanogaster.* J Proteome Res **2008**, *7*, 687-697.

van den Toorn, H.W.P., Mohammed, S., **Gouw, J.W.**, van Breukelen, B. and Heck, A.J.R. *Targeted SCX based peptide fractionation for optimal sequencing by collision induced, and electron transfer dissociation.* J Proteomics Bioinform **2008**, *1*, 374-386.

Dreisbach, A., Otto, A., Becher, D., Hammer, E., Teumer, A., **Gouw, J.W.**, Hecker, M. and Volker, U. *Monitoring of changes in the membrane proteome during stationary phase adaptation of Bacillus subtilis using in vivo labeling techniques.* Proteomics **2008**, *8*, 2062-2076.

Romijn, E.P., Christis, C., Wieffer, M., **Gouw, J.W.**, Fullaondo, A., van der Sluijs, P., Braakman, I. and Heck, A.J. *Expression clustering reveals detailed co-expression patterns of functionally related proteins during B cell differentiation: a proteomic study using a combination of one-dimensional gel electrophoresis, LC-MS/MS, and stable isotope labeling by amino acids in cell culture (SILAC).* Mol Cell Proteomics **2005**, *4*, 1297-1310.

**Gouw, J.W.**, Burgers, P.C., Trikoupis, M.A. and Terlouw, J.K. *Derivatization of small oligosaccharides prior to analysis by matrix-assisted laser desorption/ionization using glycidyltrimethylammonium chloride and Girard's reagent T.* Rapid Commun Mass Spectrom **2002**, *16*, 905-912.

# Dankwoord

Het is af! Na 4 jaar met veel plezier, maar natuurlijk ook met de nodige frustraties, aan dit onderwerp gewerkt te hebben is het zover. Het voelt goed om 'nog even' het dankwoord te schrijven om daarna alles terug te zien in een mooi boekje. De 4 jaar zijn omgevlogen, het gaat trouwens helemaal snel als je bepaalde experimenten twee keer doet, en dankzij de hulp en gezelligheid van vele waren het vier leuke jaren. Daarom wil ik een aantal hier bedanken want zonder jullie was dit niet gelukt.

Allereerst wil ik mijn co-promotor Jeroen Krijgsveld en promoter Albert Heck bedanken. Jeroen, dankzij jou nooit aflatende steun, ideeën, waardevolle discussies en het snelle correctiewerk is het een heel mooi boekje geworden. In die 4 jaar zijn er, zoals die eerste keer toen we tijdens jullie verhuizing de wasmachine aan het verslepen waren, heel wat kritische blikken uitgewisseld. Gelukkig waren we het altijd snel met elkaar eens wat ertoe geleid heeft dat vooral de laatste fase erg vlot verlopen is. Het was wel even een schok toen je vertelde dat je ons ging verlaten, maar dat heeft geen effect gehad op het afronden van mijn boekje. Inmiddels werk je in Duitsland bij het EMBL waar het je ongetwijfeld gaat lukken om iets moois op te bouwen, succes daar en bedankt voor alles. Albert, jij stond iets verder van het onderzoek af, maar dankzij jouw kennis van zaken, inzicht en enthousiasme heeft dat een prima effect gehad. Bedankt dat je mij de mogelijkheid gegeven hebt om te promoveren in je groep en om wat langer te mogen blijven (nog steeds boos?). Overigens heb ik mee mogen maken hoe jouw groep, zelf stevig aan het NPC-roer, uitgegroeid is tot één van de grootste proteomics groepen van de wereld.

Verder wil ik Peter Verrijzer bedanken voor de vrijheid die ik in Rotterdam kreeg om met de vliegen en embryo's te werken. Hans Vissers wil ik bedanken voor de prettige samenwerking en alle hulp bij de experimenten bij Waters in Manchester. Hoe vaak zijn wij ook alweer tegen de Wet van Murphy aangelopen? I would like to thank Matthias Mann and Peter Mortensen from CEBI. Matthias, thanks to for giving me the opportunity to work in your group. Peter, thank you for having the patience to teach me the insights of MSQuant and keeping me involved in the project. I had a great time in Odense and I will never forget the Christmas dinner and the Julebryg. Juliana, my only student, thanks for all the work you did on the phosphorylation project and too bad this didn't result in a publication. Good luck with your new job.

Na meer dan 5 jaar op de vakgroep te hebben rondgelopen, zijn er heel wat collega's, kamergenoten en studenten voorbijgekomen. Arjen, kamergenoot van het eerste uur, jij

kwam net terug uit Vermont toen ik als AiO begon en samen hebben we Albert's oude kamer overgenomen. Bedankt voor de gezellige momenten. Bas van Balkom, één van de vele Bassen, wil ik bedanken voor de uitermate amuserende momenten toen jij weer eens aan het discussiëren was met één van de energiebedrijven; heb je inmiddels al een directe lijn met de helpdesk? Bas van Breukelen, bedankt voor alle hulp bij het programmeren en de computerproblemen. Ook de kamergenoten van korte duur, Janssûh (het waren twee gezellige weken), Asier en Reinout wil ik bedanken. Asier, eskerrik asko! Een kamergenoot bij uitstek is Annemieke. Bedankt voor het bijbrengen van de beginselen der proteomics toen ik mijn hoofdvak bij je deed, de gezellige tijd in de 'terminale' AiO kamer en natuurlijk het uitstapje naar San Diego. SD stond garant voor even bubbelen, let's get the f@#* out of here en natuurlijk dig your own grave. Natuurlijk wil ik Edwin en Kirsten ook bedanken voor het goede onderkomen tijdens dit gezellige weekje. En Kirsten, nogmaals sorry dat we jullie die eerste avond wakker hielden toen we nogal melig waren, terwijl jullie toch echt wilden slapen. Ook wil ik jullie en Ewald, Bjørn en Jeroen bedanken voor de gezellige Katan-avondjes, welke ook wel gebruikt werden om alle (oude maar nog steeds aanwezige) frustraties te uiten. Jeroen, bedankt voor alle gezelligheid en hulp in Rotterdam en de … die stuk voor stuk eindigden met gele rakkers en bitterballen in Wester Paviljoen of Coenen. Daarbij hoort natuurlijk ook Harmjan ('ik kan geen spuugje meer maken'), steun, toeverlaat en mede-alcoholist drinkbroeder in Rotterdam. Zonder jou was het mij nooit gelukt om die vliegen te labelen, bedankt voor al je hulp hierbij. Ik denk dat niet veel mensen kunnen zeggen dat een restaurant direct gesloten wordt na een bezoekje van ons.

Ontspanning is ontzettend belangrijk tijdens het promoveren. Het is aan te raden een paar dagen in Gulpen door te brengen met Martijn, Paul, Jeroen K, Jeroen J, Harmjan, Simone, Eef, Annemieke en Léon. Of om een weekje te gaan skiën in Frankrijk met Jeroen, Simone, Paul en Javier. Paul, je weet je nu prima te gedragen als coole boarder en misschien kun je dit combineren met een optreden in de skihut. Jouw Seal performance doet namelijk echt niet onder voor het zingende duo die we in de kroeg in Frankrijk zagen. Jeroen, mede NEC'er, jammer dat je verhuisd bent naar Diergeneeskunde maar gelukkig zit je dichtbij zodat we gemakkelijk Heroes en 24 kunnen blijven kijken. Bedankt voor de squash en bieravondjes (zelfs in Wijchen) en succes met je nieuwe baan. Simone, wat dacht je van een spoortje in Whistler nabij Vancouver? Bedankt voor de gezelligheid hulp in het lab en succes in Duitsland. Wanneer gaan we nou die vissen labelen? Ik kijk ontzettend uit naar het volgende board-tripje in januari na mijn promotie. Possibly the best way to become completely relaxed is to go to the festival of San Fermín (aka running of the bulls) in

Pamplona. I have to thank Javier for inviting me to this fantastic, crazy, shocking and insane event. Javi, since we became roommates we could get along perfectly, became good friends and now we are sharing the best office (couch, fridge, entertainment system and some nice decoration) of the University. Thank you for learning me the essentials of Spanish and all the things I can do with 'my beautiful eyes'. It is great that you want to be my Paranimf!

If you want to go to a conference and have a good night sleep, than be sure to not share a room with Shabaz and Manuel and book the room yourself. Although we ended up sleeping in the same bed we had a really good time in Seattle. Manuel ('let's order room service and have a couple of Coronas and burgers') thank you for the good time, playing tennis and the nice discussions. Shabaz, you have been a good friend since we met in Odense some 5 years ago. Thanks for all the help in and around the lab. Onno, ondanks je twijfelachtige muzieksmaak, ik noem een BJ of BA, vond ik het ontzettend leuk om samen naar concerten/festivals te gaan. Succes met je vervolg in de groep en waarschijnlijk kun je jezelf over een tijdje ook quantitation-guru noemen. De buurvrouw bij uitstek, Renkse, bedankt voor al je gezellige bezoekjes, lunches bij Tricolore en commentaar op de muziek (het is bewonderenswaardig dat je het hebt uitgehouden naast ons met die muziek). Sharon, zet 'em op, je bent er bijna. Nadia, your Gajol and Spunk are without doubt the best self made liquor I ever tasted (even with technical ethanol). Esther, na een jaartje weg geweest te zijn weer terug op het oude nest, bedankt voor de gezelligheid. Martina thanks for your help in the lab and all your stories. Kees, een bezoekje van jou duurt altijd iets langer dan gedacht, maar is altijd wel leuk en gezellig. Bedankt voor al je hulp en adviezen. Nicolai, thanks for your help with the students, have fun with them. Tieneke, nog maar een paar maanden en dan ben je er ook, succes. Waneer gaan we weer naar een concertje? Ook natuurlijk alle andere (ex-)collega's ontzettend bedankt. De gasten van het HLO Arjan (aka dregt, ar-bar), Wouter en Erik mogen natuurlijk niet ontbreken. Arjan, als jij er bent is het nooit saai en gebeurt er altijd wel wat. We hebben het maar getroffen met de wetenschap op de eerste verdieping. Wouter, wat dacht je van een weekendje Vancouver? Wie van jullie belt Cees trouwens?

Verder wil ik Peter Burgers bedanken voor de leuke en vooral leerzame tijd tijdens mijn stageperiode bij Hercules. Peter, ik heb ontzettend veel geleerd van je en het is bewonderenswaardig hoe gemakkelijk jij de moeilijkste onderwerpen kunt uitleggen.

Uiteraard zijn er naast het werk ook de nodige mensen die ik wil bedanken voor de ontspanning. Voornamelijk mijn ex-teamgenoten van de verschillende teams bij Keistad. Jammer genoeg kon ik de afgelopen 2 jaar niet meer mee doen, maar in ieder geval bedankt

voor de leuke wedstrijden en vooral de gezellige tijd na de wedstrijden. Natuurlijk wil ik ook alle mensen die de afgelopen 3 jaar mee geweest zijn op wintersport bedanken. Bart ('Je m'appelle un pain'), ik heb me prima vermaakt tijdens die paar extra dagen in Gargellen in de quaterpipe, wanneer gaan je ouders weer? Emile zorg er als opponent wel voor dat je het over fruitvliegen hebt en niet over vuur- of strontvliegen.

De familie Vergeer, Jo, Ida en René, ik heb me vanaf het begin al direct bij jullie thuis gevoeld. Ontzettend bedankt voor de goede zorgen, leuke uitstapjes en gezelligheid. Ida, jij weet als geen ander hoe je iemand in de watten kan leggen en mede daarom kom ik heel graag bij jullie. Je bent zonder twijfel mijn liefste schoonmoeder. Jo, ik vond het ontzettend leuk dat je me meegenomen hebt naar de leukste club van Nederland (maar wanneer gaan ze nou eens goed spelen) en ik kan gelukkig nu weer wat vaker mee. Ook vond ik het leuk om samen met jou, Leo en later René op de beurs rond te scharrelen. Jammer genoeg zal het wel even duren voordat ik mijn eigen baan heb. René, ik heb nu gelukkig meer tijd om wat vaker te komen kijken bij een optreden en om samen met jou en Jo naar de beurs te gaan.

Het leven is een stuk aangenamer met een broer als Matthijs. Samen zijn wij twee handen op een buik en hebben altijd ontzettend veel lol, hoewel er momenten zijn geweest dat niet iedereen een nachtelijke sessie Mario Kart kon waarderen ('als jullie niet ophouden gaat de stop eruit'). Ik heb enorm veel van je geleerd en vind het altijd super om even een klein biertje met je te gaan drinken. Ik vind het een eer dat je mijn Paranimf wilt zijn en hopelijk gaat het je lukken om ons in Vancouver te bezoeken. Ik kijk nu al uit naar een klein Canadees biertje.

Lieve Pa en Ma, zonder jullie onvoorwaardelijke steun, interesse en vertrouwen zou dit allemaal niet gelukt zijn. Ik kon gelukkig altijd bij jullie terecht en even alles vergeten. Jullie hebben mij altijd mijn eigen keuzes laten maken (maar je gaat wel eerst naar Volleybal) en dat heb ik ontzettend gewaardeerd. Het is altijd (soms te) gezellig als jullie er zijn en daarnaast staan jullie altijd klaar voor ons. Het weekje Griekenland heeft mij prima gedaan (en al helemaal met al die kilo's rosé) en hopelijk kunnen we dat nog vaker doen, misschien zelfs in Canada. Bedankt voor alles!

Natuurlijk zijn de laatste woorden voor mijn lieve Sandra. Al die jaren heb je mij ontzettend gesteund en heb je heel veel begrip getoond als ik weer eens in het weekend moest werken. San, je bent buitengewoon zorgzaam, lief en biedt altijd een luisterend oor en daar heb ik zeer veel respect voor. Ik heb enorm veel zin in ons Canada avontuur en vind het super dat we daar samen gaan wonen. Ik hou van je!