**University of Southampton**

Faculty of Medicine

# Lung Surfactant Kinetics for Biomarker Discovery in Pulmonary Disease

By

**Zhang Kewei**

Thesis for the degree of Doctor of Philosophy

Novermber 2019

# University of Southampton

## Abstract

Faculty of Medicine

Doctor of Philosophy

### Lung Surfactant Kinetics for Biomarker Discovery in Pulmonary Disease

by Zhang Kewei

Pulmonary diseases has long suffered from a lack of clinical biomarkers for disease diagnosis, prognostic prediction and treatment response. Pulmonary surfactant is critically important for optimal lung function. While altered surfactant function, concentration and composition have been reported in patients with various lung diseases, these changes have also yet to be translated into clinical practice. An alternative approach to measurement at a single time point is to quantify time course kinetics of substrate fluxes through metabolic pathways. Consequently, in this thesis, incorporation of the stable isotope-labelled substrate deuterated *methyl* choline chloride (*methyl*$D_9$choline) into the major surfactant phospholipid phosphatidylcholine (PC), combined with electrospray ionization tandem mass spectrometry (ESI MS/MS), was employed to monitor the metabolic flux of surfactant phospholipid synthesis *in vivo*. Time series PC data acquired through ESI MS/MS coupled with stable isotope labelling techniques from three studies was analysed using the smoothing spline mixed effect method (SME). This is a pilot study applying time course label enrichment of multiple surfactant PC species as biomarkers to phenotype animal models and patients with various lung diseases.

There is no effective software for data processing mass spectrometry analysis of time series substrate incorporation into multiple molecular species products. Consequently, a bioinformatics platform LipidomeLabelling was developed to facilitate data analysis of dynamic labelling studies and was used for the following biomarker discovery analysis.

Time course *methyl*$D_9$ label enrichment of 16 PC species was modelled using SME method for all the subjects of group pairs in three different studies, which makes the foundation of the multivariate time course biomarker discovery analysis for lung diseases. In a study, a *Ft* statistic was calculated as the scaled difference between the time course label enrichment of two contrasting groups for a single PC species, and the *Ft* statistics for all the species were ranked and according to their significance in distinguishing the two groups, for potential biomarker identification. The surfactant PC kinetic biomarker analysis of a mouse model shows that time course label enrichment of multiple PC species significantly differentiated between wild type mice and granulocyte macrophage colony stimulating factor (GM-CSF) beta chain knock out mice. However, this methodology failed to provide useful biomarkers in the clinical conditions examined, preterm infants at risk of neonatal respiratory distress syndrome and adult patients with acute respiratory distress syndrome compared with healthy volunteers. While differences could be determined between grouped data, subject heterogeneity precluded provision of diagnostic individual for individual patients.

# Table of Contents

Table of Contents

# Table of Tables

Table of Tables

# Table of Figures

Table of Figures

Table of Figures

Table of Figures

# Research Thesis: Declaration of Authorship

| | |
|---|---|
| Print name: | Zhang Kewei |

| | |
|---|---|
| Title of thesis: | Lung surfactant kinetics for biomarker discovery in pulmonary disease |

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. None of this work has been published before submission

| | | | |
|---|---|---|---|
| Signature: | | Date: | |

# Acknowledgements

I would like to express my sincere gratitude to Professor Anthony Postle, Professor Jeremy Frey and Dr Doroteya Staykova for giving me the opportunity to undertake the research pursuing a PhD degree, and for their invaluable supervision, guidance, support, patience and encouragement.

I would also like to thank Dr Grielof Koster, Dr Joost Brandsma, Paul Townsand and the rest of the BRU Mass Spectrometry Laboratory office team, for sharing their knowledge and skills and for their kind support and encouragement. Many thanks to those unnamed who has offered help during my study.

Most importantly, I would like to thank my husband, son and daughter, parents, brother and in-laws, for your infinite love, tremendous support and boundless patience, enabling me to complete my studies. I love you all so much!

# List of abbreviations

| | |
|---|---|
| ABC | ATP-binding cassette |
| APE | Atom percent excess |
| ARDS | Acute respiratory distress syndrome in adults |
| AT II cell | Alveolar type II epithelial cell |
| ATP | Adenosine Triphosphate |
| BALF | Bronchoalveolar lavage fluid |
| CCT | CTP:phosphocholine cytidylyltransferase |
| CDP | Cytidine diphosphate |
| CID | Collision - induced dissociation |
| CK | Choline kinase |
| COPD | Chronic obstructive pulmonary disease |
| CPT | Choline phosphotransferase |
| CWT | Continuous wavelet transform theory |
| DPPC | Dipalmitoylphosphatidylcholine |
| EI | Electron ionization |
| ER | Endoplasmic reticulum |
| ESI | Electrospray ionisation |
| GC | Gas chromatography |
| GC-C-IRMS | GC-combustion isotope ratio mass spectrometry |
| GC-IRMS | Gas chromatography-isotope ratio mass spectrometry |
| GM-CSF | Granulocyte-macrophage colony-stimulating factor |
| ILs | Inflammatory cytokines interleukins |
| LB | Lamellar body |
| LC | Liquid chromatography |
| LPCAT1 | Acyl CoA:lysophosphatidylcholine acyltransferase |
| MS | Mass spectrometry |
| MS/MS | Tandem mass spectrometry |
| m/z | Mass over charge |
| NRDS | Neonatal respiratory distress syndrome |
| PAP | Pulmonary alveolar proteinosis |

| | |
|---|---|
| PC | Phosphatidylcholine |
| PE | Phosphatidylethanolamine |
| PEMT | Phosphatidylethanolamine N-methyltransferase |
| PG | Phosphatidylglycerol |
| PI | Phosphatidylinositol |
| PL | Phospholipids |
| SAM | S-adenosylmethionine |
| SME | Smoothing splined mixed effects |
| SP-A | Surfactant protein A |
| SP-B | Surfactant protein B |
| SP-C | Surfactant protein C |
| SP-D | Surfactant protein D |

# Chapter 1 Introduction

## 1.1 Biomarker discovery in pulmonary diseases

A biomarker is a biological characteristic that can be objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacological responses to a therapeutic intervention (Atkinson *et al.*, 2001). Biomarkers can be characteristic properties and molecules, and biomarkers that are of the most interests include genes, proteins, metabolites, and lipids.

Pulmonary diseases have been suffering from a lack of sensitive and specific biomarkers in clinical practice. The diseases are usually diagnosed based on history of symptoms, physical examinations, pulmonary function tests, radiology, and histopathological tests (Chen *et al.*, 2016; Kim *et al.*, 2017). However, the clinical data is often not sufficient for practitioners to accurately and efficiently identify diseases excluding other diseases with similar characteristics, identify patient subgroups for tailored treatment, monitor and evaluate treatment responses, or predict the disease progression. This is largely due to the complexity and the poor understanding of most of the lung diseases. Therefore, new biomarkers are needed for disclosing the underlying mechanisms of the diseases as well as facilitating effective clinical management.

The lethal acute respiratory distress syndrome in adults (ARDS) has a high mortality rate up to 50% (Máca *et al.*, 2017). Although there are various triggering conditions, ARDS is diagnosed based on the fulfilment of certain clinical diagnostic criteria (Fan *et al.*, 2018), which lack specificity and allow for inclusion of heterogeneous patients groups. The clinical management involves supportive ventilation strategies and treatment for some general conditions, such as infection. There has been no pharmacological treatment, due

to the lack of understanding of disease mechanisms and indicators for guiding customised treatment. Asthma is a highly prevalent pulmonary disease worldwide resulting in 250,000 deaths annually (D'Amato *et al.*, 2016). There has been little improvement in the clinical management of Asthma for decades, despite the tremendous effort. The diagnosis is mainly based on disease history, physical examination and allergy test, which are inadequate to stratify heterogeneous patient groups. Effective markers are not available for identifying patients who may be responsive to the limited treatments options (Miranda *et al.*, 2004; Rosenberg *et al.*, 2007; Shaw *et al.*, 2007; Kim *et al.*, 2017). The debilitating chronic obstructive pulmonary disease (COPD) is one of the leading causes of death worldwide with a rising death rate (Mathers and Loncar, 2006). The natural history of COPD is often marked by periodic exacerbation, in which symptoms of breathlessness and sputum production worsen acutely. Disease progression of COPD is variable, and periodic acute exacerbations only happen in some patients, leading to morbidity and mortality (Lomas *et al.*, 2009). However, there are no reliable indicators that can help assess the state of COPD or predict its progression of each patient for guiding and monitoring treatment intervention (Il Yoon and Sin, 2011; Chen *et al.*, 2016). For other lung diseases including idiopathic pulmonary fibrosis (Ohnishi *et al.*, 2002), pneumonia (Baughman *et al.*, 1993), sarcoidosis (Griese, 1999) and cystic fibrosis (Gray *et al.*, 2017), challenges remain in the diagnosis, prognosis and assessment of disease activity and response to treatment. New biomarkers are urgently required for elucidating the underlying pathophysiological mechanisms of these diseases for improving the clinical outcomes and accelerating drug development.

The ever-growing interests in biomarker discoveries in lung diseases have driven a large amount of investigations on hundreds of potential biomarkers of various forms, such as

proteins, genes, cells, lipids etc. The most intensively studied biological markers are protein indicators, which are involved in various disease mechanisms at different stages of the pathological processes.

Proteins involved in inflammation pathways are of the highest interest in biomarker studies, as inflammation is a common characteristic in most pulmonary diseases. In BALF, plasma, sputum and serum samples, some members of the large family of inflammatory cytokines interleukins (ILs), which are mediators in the inflammatory processes, have been shown to be correlated with pathogenesis of various diseases, like ARDS (Park *et al.*, 2001), neonatal respiratory distress syndrome (NRDS) (Terpstra *et al.*, 2014), asthma (Rosenberg *et al.*, 2007) and cystic fibrosis (Gray *et al.*, 2017). Apart from ILs, other inflammatory related proteins, such as matrix metalloproteinase family (Cederqvist *et al.*, 2001; Sweet *et al.*, 2001) and C-reactive protein (Bajwa *et al.*, 2009) in BALF, tracheal aspirate and plasma samples are also of great interest in the biomarker studies of lung diseases. Potential protein indicators of epithelial and endothelial cell damages are also widely investigated for their suitability of being used as biomarkers. Epithelial cell damage markers including Krebs von den Lungen-6 and soluble receptor for advanced glycation end products in plasma and BALF are elevated with ARDS development (Sato *et al.*, 2004); (Jabaudon *et al.*, 2015), interstitial lung diseases (Ohnishi *et al.*, 2002) and acute lung injury (Uchida *et al.*, 2006). Endothelial related markers for lung injuries include Endothelin 1 (Boscoe *et al.*, 2000) and angiopoietin-2 (Calfee *et al.*, 2012), owing to their pulmonary vasoconstriction and vascular permeability mediating function. Surfactant proteins are important components of pulmonary surfactant system, which are essential for optimal lung function. Biomarker studies of the main surfactant proteins including surfactant protein A (SP-A), surfactant protein B (SP-B), surfactant protein C (SP-C) and

surfactant protein D (SP-D) have also been largely conducted for various lung conditions (Eisner *et al.*, 2003); (Cheng *et al.*, 2003; Lomas *et al.*, 2009). Protein biomarkers involved in other disease mechanisms, such as coagulation, fibrinolysis, fibrosis, extracellular matrix remodelling, are investigated as well.

Cell and genetic biomarkers are also widely analysed for their ability in indicating disease pathogenesis. For example, eosinophils and neutrophils play an important role in the body's response to allergic reactions. They are found to be linked with asthma (Bousquet *et al.*, 1990; Miranda *et al.*, 2004; Shaw *et al.*, 2007) as well as other diseases such as COPD (Stanescu *et al.*, 1996). The genetic biomarkers of interests usually have their products, such as proteins, that have shown significance in indicating abnormalities in certain diseases. Although biomarker studies have been focusing predominantly on targeted individual molecules, omics analysis including proteomics, genomics, transcriptomics, and metabolomics have also been carried out for biomarker analysis (Wheelock *et al.*, 2013). The benefit of omics analysis is the possibility of revealing pathological mechanisms by probing wide range of markers at different functional biological levels at the same time.

Despite the tremendous effort in the last few decades, there are hardly any new biomarkers established in clinical settings for pulmonary diseases with proven benefits, requiring more types of biomarkers to be investigated. Lipid biomarkers have been intensively used in various diseases, such as cancer (Yan *et al.*, 2018), cardiovascular diseases (Li *et al.*, 2017), diabetes (Dorcely *et al.*, 2017) and Alzheimer's disease (Zarrouk *et al.*, 2018) etc. Nevertheless, there are very few lipid biomarker studies for pulmonary diseases apart from some lung cancer studies (Ravipati *et al.*, 2015; Yu *et al.*, 2017). Nearly half decades ago, it was found that the ratio of fetal lung secreted lecithin

(phospholipids) to sphingomyelin presented in amniotic fluid was a significant indicator in determining the fetal lung maturity (Gluck and Kulovich, 1973). However, the studies of surfactant phospholipids (PL), the major component of pulmonary surfactant that keeps normal lung functioning, have not provided any efficient clinical biomarkers for other pulmonary conditions ever since, although their abnormality have been widely found in various lung diseases.

## 1.2 Pulmonary surfactant

Pulmonary surfactant is a surface-active lipid-protein complex secreted by alveolar type II epithelial cells (AT II) through a highly regulated metabolic process (Halliday, 2008). It reduces the high surface tension at the air-liquid interface of the alveoli, which prevents collapse of lung at the end of expiration, and reduces the effort that is needed to expand the lung during inhalation. Pulmonary surfactant is critical for optimal lung function. Deficiency or dysfunction of lung surfactant are associated with various lung diseases and can even lead to severe life-threatening lung failure (Glasser and Mallampalli, 2012; Akella and Deshpande, 2013).

### 1.2.1 Pulmonary surfactant composition

Surfactant lipid is the main component that accounts for around 90% of the lung surfactant (Veldhuizen and Haagsman, 2000), and comprises mainly phospholipids as well as some other lipids such as cholesterol. The identified protein components of surfactant consist of SP-A, SP-B, SP-C and SP-D. SP-B and SP-C are hydrophobic, and their function is to increase the absorption rate of phospholipids (Hawgood, 2004). SP-A and SP-D are hydrophilic, and they contribute to surfactant homeostasis and pulmonary immunity.

Phospholipids are amphiphilic lipids that consist of hydrophobic fatty acids and a phosphate based hydrophilic head, joined together mostly by a glycerol molecule (glycerophospholipids). They make up approximately 80-85% of the lipid component of lung surfactant (Agassandian and Mallampalli, 2013). Among the phospholipid classes, phosphatidylcholine (PC) is the principle constituent and the remainder include phosphotidylglycerol (PG), phosphatidylinositol (PI) and phosphatidylethanolamine (PE). The most abundant PC species present in human lung surfactant is the saturated dipalmitoylphosphatidylcholine (DPPC) (Figure 1)(Postle *et al.*, 2001), which is the major surface-active molecule that reduces the surface tension at the air-liquid surface of alveoli.

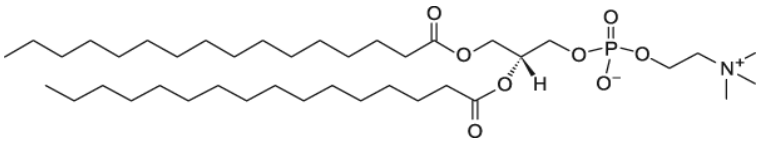


Figure 1: Molecular structure of dipalmitoylphosphatidylcholine (DPPC)

### 1.2.2 Surfactant metabolism

#### 1.2.2.1 Synthesis

Pulmonary surfactant components are synthesised in the endoplasmic reticulum (ER) of AT II cells, where surfactant lipids and proteins are synthesised at smooth ER and rough ER respectively (Olmeda *et al.*, 2017a). *De novo* biosynthesis of the most abundant and functionally important surfactant PC occurs via cytidine diphosphocholine (CDP-choline) pathway, which will be introduced in detail in section 1.3.2. However, CDP-choline pathway is not predominant for surfactant DPPC production. It is shown that in isolated rat lung AT II cells only ~45% of DPPC is *de novo* synthesised, while 55% - 75% comes from remodelling pathway (Batenburg, 1992); (Caesar *et al.*, 1991). The remodelling pathway

of DPPC involves de-acylation of monounsaturated PC by phospholipase (Hook, 1991), which hydrolyses unsaturated fatty acids at the sn-2 position of PC and generates lyso-PC, followed by re-acylation with saturated palmitic acid (16:0) species by acyl CoA:lysophosphatidylcholine acyltransferase (LPCAT1) (Batenburg *et al.*, 1979; Voelker and Snyder, 1979). It is also suggested that LPCAT1 is involved in cross-talk between *de novo* synthesis and remodelling pathways of phospholipid balance (Butler and Mallampalli, 2010). Overexpression of LPCAT1 increased degradation of choline phosphotransferase (CPT) to supresses the *de novo* synthesis. LPCAT1 is also involved in the remodelling pathway of dipalmitoylphosphatidylglycerol.

The biosynthesis of surfactant PC can be affected by various factors. Activities and amount of enzymes catalysing the sequential reactions have direct impacts on the biosynthesis (Goss *et al.*, 2013). Increasing enzyme activity and enzyme mass may result in increased synthesis, vice versa. Certain hormones, such as thyroid hormone, oestrogen and glucocorticoids can also regulate PC biosynthesis by changing the activity and amount of the enzymes (Khosla *et al.*, 1980) or increasing the availability of precursors and substrates involved in the PC synthesis.

#### 1.2.2.2 Lamellar body formation, packaging and transportation

After synthesis, surfactant lipids and hydrophobic proteins SP-B and SP-C are transported to the lamellar bodies (LB) for storage and processing. Phospholipids are likely to be transported to lamellar bodies directly, whereas SP-B and SP-C are initially packed into multivescular bodies and subsequently transported from Golgi apparatus to the lamellar bodies (Weaver *et al.*, 2002). SP-B plays a crucial role in the packaging of surfactant phospholipids and the proper formation of LB (Foster *et al.*, 2003).

Upon arrival at lamellar bodies, surfactant lipids transverse the limiting membrane of LB facilitated by ATP-binding cassette (ABC) transporters. ABCA3 is found in the limiting LB membrane, which is essential for normal mechanism for surfactant synthesis and storage (Ban et al., 2007). It provides energy needed for the translocation of lipid into the lumen of LB by ATP hydrolysis. ABCA3 gene mutation was found to cause fatal surfactant deficiency in newborns (Shulenin et al., 2004).

#### 1.2.2.3 Surfactant secretion

The lamellar bodies are secreted into the extracellular matrix through exocytosis. Mechanical stretching of alveoli during inhalation is the primary physical stimulus for secretion to take place. The distention may lead to elevation of cytoplasmic calcium concentration in alveolar cells (Wirtz and Dobbs, 1990), which triggers surfactant secretion. The elevated calcium leads to fusion of LB with the plasma membrane, fusion pores are then formed, followed by the release of surfactant. Agents that stimulate surfactant secretion includes terbutaline, forskolin, ATP, while SP-A is found to inhibit the secretion (Dobbs *et al.*, 1987) .

#### 1.2.2.4 Surfactant in extracellular space

After secretion, surfactant content is unpacked into the forms of lamellar body like particles, multi-layer packages and tubular myelin. SP-B and SP-C play a role in the LB unpacking process. SP-A and SP-B are suggested to be essential for the formation of tubular myelin (Cabre *et al.*, 2009).

These organised forms of membranous assemblies provide lipid-protein complexes for the formation of the surface-active monolayer at the air-liquid interface of the alveolus. The insertion of surface-active material into the interface is facilitated by SP-B and SP-C (Hobi *et al.*, 2016), and the adsorption rate depends on the concentration of two forms of

the surfactant components in the air space, large aggregates and small aggregates. Large aggregates are the surface active form of surfactant, which can be isolated from tubular myelin, secreted lamellar bodies and large lipid vesicles. Once being used, they are dissociated into small aggregates, the small unilamellar vesicles, which are not surface active.

#### 1.2.2.5 Surfactant degradation and recycling

The surfactant component dissociated from the surfactant monolayer are either internalized by AT-II cells, or cleared by alveolar macrophages and respiratory airways. In the situation of lung injury or other lung conditions, surfactant may leak into the circulation. Internalised surfactant could be remodelled and incorporated into LB for the subsequent secretion or degraded by lysosomes. SP-A has been found to promote the internalisation process, mediated by SP-A receptor and phospholipids were found internalized together with SP-A in rat AT II cells (Wissel *et al.*, 2001).The absence of SP-D are linked with accumulation of DPPC in the alveoli (Postle *et al.*, 2011). SP-C is also speculated as a contributor to extracellular surfactant metabolism, either for degradation or recycling for compositional “re-programming” or both (Olmeda *et al.*, 2017b).

Macrophages are critical for surfactant catabolism and are responsible for about 20% of its clearance. Surfactant saturated PC were observed to have increased four to eight folds in the lungs of granulocyte-macrophage colony-stimulating factor (GM-CSF) deficient mice and GM common receptor β depleted mice, with normal synthesis and secretion (Ikegami, 2006). This is caused by the defect in surfactant catabolism due to impaired alveolar macrophages.

### 1.2.3 Surfactant alteration in pulmonary diseases

Pulmonary surfactant is a complicated and tightly regulated system, and its components are closely coordinated and highly correlated. Dysfunction of any part of the system could result in surfactant alterations, and alterations at any level of metabolism could lead to or be associated with respiratory conditions (Bernhard *et al.*, 1997; Glasser and Mallampalli, 2012; Naroji *et al.*, 2015). Surfactant alteration has been observed in most lung diseases, which is a good indicator of abnormality of surfactant metabolism as well as impaired lung function.

The most aggressive lung condition of respiratory distress syndrome are associated with significant surfactant alteration. Surfactant deficiency is the cause of NRDS due to lung immaturity and the application of surfactant replacement treatment has greatly improved the mortality rate. The ARDS, another fatal secondary lung disease, is also characterised with abnormal surfactant. ARDS patients have decreased fractional concentration of PC, PG and increased concentration of PI, PE and PS, accompanied with alveolar surface tension increase (Gunther *et al.*, 2001). Moreover, Schmidt *et al.* (2007) found that the concentration of major surface-active species DPPC was significantly reduced and the number of unsaturated PC species were increased. Surfactant alterations have also been found in other less aggressive lung conditions. Pulmonary alveolar proteinosis (PAP) is a rare disorder of surfactant catabolism characterised with an accumulation of surfactant and alteration in the constituent lipids. Whitsett et al. (2015) reviewed the genetic disorders of surfactant homeostasis including abnormalities in surfactant packaging, function and degradation caused by mutations in the genes encoding SP-B, SP-C and ABCA3. These mutations are associated with interstitial lung disease and chronic lung diseases. Bacterial and viral infection can also lead to surfactant PC alteration. Wu *et al.*

(2007) have found reduced surfactant DPPC synthesis in mice infected with mucoid isolate of P.aeruginosa bacterial, as a result of reductions in mRNAs and immunoreactivity levels for CTP: phosphocholine cytidylyltransferase. Changes in fatty acyls on phospholipids were also observed in bacterial pneumonia (Baughman et al., 1984). Human adenovirus can alter surfactant phospholipid composition (Miakotina et al., 2007). Apart from the pathogenic germs, surfactant phospholipids also react to the non-pathogen insult such as smoking. Decrease in phosphatidylcholine levels and the increase in phospholipids hydrolysing activity of phospholipase A2 were observed in mice after being exposed to cigarette smoke (Agarwal et al., 2014).

As stated above, alterations of surfactant and especially surfactant PLs are highly associate with various lung conditions. However, the underlying mechanisms of the alterations are still not well understood. Better knowledge of *in vivo* surfactant PL species activities can help gain better understanding of disease mechanisms as well as explore biomarkers that can be indicative of lung abnormalities. The *in vivo* kinetic studies of surfactant phospholipids have been carried out for decades since the last century, benefiting from the ever-developing isotope labelling techniques and label detection techniques, particularly the presently dominated stable isotope labelling coupled with mass spectrometry techniques.

## 1.3 Dynamic surfactant phospholipid analysis with isotope labelling and mass spectrometry techniques

### 1.3.1 Mass spectrometry techniques

History of mass spectrometry (MS) began at the beginning of 19*th* century (Griffiths, 2008), which has experienced huge development ever since and was applied into the

research of physical, chemical and biological sciences. Over the past few decades, MS has become the essential analytical tool in biological research (Finehout and Lee, 2004), which enables scientists to understand the biological systems at the molecular level. Having undergone tremendous technological improvements, with the high sensitivity and high-throughput manner, MS has become an irreplaceable tool in proteomics, metabolomics and the relative young lipidomics (Di Girolamo *et al.*, 2013).

#### 1.3.1.1 Sample introduction approaches: shotgun and chromatography MS analysis

Generally, two sample introduction approaches commonly used in MS based lipidomic analysis are direct infusion, also called “shotgun”, and chromatography (Page *et al.*, 2009). Shotgun approach does not require any sample pre-separation procedures before injecting samples into the MS analyser. This allows MS to analyse all the lipid molecules in a sample simultaneously (Lin *et al.*, 2010). Chromatography approach needs at least one extra step of sample separation before the sample goes into the mass spectrometer (Coskun, 2016). In this way, different molecules in the biological samples are separated in advance for better identification when analysed by the MS. However, separated detection of individual molecules can result in inconsistent ion suppression among the molecules in the same sample, which may lead to inaccurate measurement.

#### 1.3.1.2 Ionisation methods

Ionisation is the process by which molecules acquire charges and turn into molecular ions before being analysed by MS analysers. A variety of ionisation sources have been used for ionising molecules. Before the 1980s, Electron ionization (EI) was the primary ionisation source for mass analysis, and now it is mainly used for gas chromatography (GC) MS analysis in lipidomic research (Gordin *et al.*, 2008). The technique works by bombarding gas phase molecules by a beam of high energetic electrons to produce ions, and due to

the extensive fragmentation power, it is also called hard ionisation technique. However, its usage is limited to analysing smaller molecules with molecular weight under 600, which is deficient in measuring larger biological compounds. This led to the development of electrospray ionisation (ESI) and matrix assisted laser desorption ionisation techniques. ESI is also well known as “soft ionisation”, is the primary ionisation technique used in biological research nowadays, due to its advantage of keeping the entire macromolecules from being dissociated. With ESI, sample solution is sprayed from the tip of a metal nozzle with a potential to form charged droplets, which are subsequently evaporated using dry gas or heat to release ions (Ho *et al.*, 2003). Matrix assisted laser desorption ionisation is another type of ionisation technique that causes transfer of a sample from the condensed phase to the gas phase through laser excitation and ablation of the sample matrix (Karas and Kruger, 2003). It is widely used for imaging mass spectrometry analysis.

#### 1.3.1.3 MS analysers for lipidomic data acquisition

The commonly used MS analysers for lipid data acquisition include quadruple, ion traps, time-of-flight, Orbitrap and ion cyclotron resonance (Wolff and Stephens, 1953; Marshall *et al.*, 1998).These analysers have undergone numerous modifications over the last few decades to be able to interface with various ionisation sources as well as to achieve better performance in accuracy, specification, resolution, scan speed and mass range. Quadrupole analyser is currently the most commonly used mass analyser. It can be interfaced with EI, ESI ionization sources and is capable of analysing molecules with mass over charge (m/z) range between 100 and 4000, which is suitable for macro biomolecules including proteins and lipids that are typically under 3500. Quadrupole mass analysers can be put together to perform tandem mass spectrometry data aquisition (Johnson *et al.*, 1990). Triple quadruple analyser with three quadrupoles placed in series can carry out

collision-induced dissociation on molecules to offer extra component information of molecules for better identification, which is widely used in lipidomic analysis.

Different types of MS analysers have their own advantages and limitations. Quadrupole analysers can be linked together to perform targeted breakdown scans, but it has limited resolution and the peaks may need to be tuned as a function of mass. Time of flight mass analyser scans fast and is able to analyse highest practical mass range, although it is not efficient in breaking down scans. Fourier transform ion cyclotron resonance and Orbitrap use resonance and Fourier transform to provide the highest mass resolution among all mass analysers but have limited dynamic range and sampling speed. With different characteristics of the MS analysers, there is no ideal mass analyser that is good for all applications.

#### 1.3.1.4 Tandem mass spectrometry

The number of steps the mass spectrometry performs for an analysis is dependent upon research requirements. One-step MS is usually for a general detection of molecules in a sample based on their m/z values, sometimes combined with chromatography that offers time information for further recognition of the molecules detected. Tandem mass spectrometry (McLafferty, 1981) is advantageous in targeted MS analysis, which is able to monitor various molecules with certain structure by specifying a certain part of them. A widely used technique is the two-step tandem mass spectrometry (MS/MS) and the instruments can be assembled by combinations of mass analysers (deHoffmann, 1996). In triple quadrupole MS/MS (Hsu and Turk, 2001), the first quadrupole analyser is used to scan entire molecular ions before they are introduced into the second quadrupole, which is called a collision cell. Within the collision cell, ions are collided into fragments using argon or helium gas. In the third quadrupole the generated fragments are analysed. This

method offers structural information of the detected molecules, which has been used to sequence peptides (Shevchenko *et al.*, 2002)and structurally characterise carbohydrates (Carr *et al.*, 1993), small oligo-nucleotides (Ni *et al.*, 1996) and lipids (Shaner *et al.*, 2009) .

#### 1.3.1.5 Direct infusion ESI MS/MS for analysis of dynamic lipidomics with stable isotope labelling

The combination of direct infusion with triple quadruple ESI MS/MS techniques facilitates the rapid lipidomic analysis with high sensitivity, specificity and stability. The ESI technique keeps intact molecules for high throughput lipidomic analysis. Direct infusion approach allows lipid molecules in a sample to be introduced into MS analyser simultaneously. This enables all analytes of interest including labelled and unlabelled lipid species to be analysed under identical ionisation conditions (Pulfer and Murphy, 2003), which is critical to the labelling studies in terms of the measurement accuracy. The tandem MS approach allows monitoring molecules with certain structure by specifying their fragment ions, which is especially useful in investigating certain classes of molecules both in their labelled and unlabelled forms. The instrument can switch quickly between different modes to scan different classes of lipid species almost in parallel. These characteristics of the direct infusion ESI MS/MS makes it a perfect tool for stable isotope labelling analysis of dynamic lipidomics (Shaner *et al.*, 2009; Harkewicz and Dennis, 2011).

### 1.3.2 PC synthesis pathways and *methyl*$D_9$-choline labelling assessment with ESI MS/MS

#### 1.3.2.1 PC synthesis pathways

*De novo* biosynthesis of PC in human relies on two molecular pathways (Gibellini and Smith, 2010). The CDP-choline pathway is the primary PC synthesis pathway that operates in all nucleated mammalian cells (Kennedy and Weiss, 1956). It starts with choline phosphorylation to phosphocholine by choline kinase (CK). In the second reaction,

phosphocholine is converted to CDP-choline by CTP:phosphocholine cytidylyltransferase (CCT). The last step of synthesis is the formation of PC by choline phosphotransferase (CPT), which transfers phosphocholine from CDP-choline to PC. Among the three-step process, the second reaction is considered as being the rate-determining step. Surfactant DPPC is synthesised *de novo* through CDP-choline pathway in AT II lung cells.

The second PC synthesis pathway is the phosphatidylethanolamine N-methyltransferase (PEMT) pathway by hepatocytes (Vance *et al.*, 2007). PEMT enzyme is found in endoplasmic reticulum and mitochondria-associated membranes. It converts PE to PC via three sequential methylation reactions, with S-adenosylmethionine (SAM) as methyl donor in each step. The PEMT pathway contributes approximately 30% of PC biosynthesis in liver, while CDP-choline pathway generates around 70% (Vance, 2014).

#### 1.3.2.2 Deuterium labelled PC by *methyl*$D_9$-choline chloride

The use of *methyl*$D_9$-choline chloride for PC stable isotope labelling, combined with ESI MS/MS determination technique is a novel way for dynamic PC metabolic analysis (Postle *et al.*, 2001; Postle *et al.*, 2004). Through label incorporation, deuterium atoms replace the hydrogen atoms on the methyl groups of the newly synthesised PC species. Consequently the labelled species have a higher mass than the endogenous species, both of which can be identified by MS. ESI MS/MS can perform precursor scans for deuterium labelled PC species detection, which measure the masses of both entire molecular ions and the specific fragments where the methyl groups are positioned. There are three methyl groups on a PC species, and each methyl group has three hydrogen atoms. Therefore, if a PC species are deuterium labelled on all three of its methyl groups, its mass will be nine units higher than the endogenous PC, which happens to the CDP-choline pathway synthesised PC species in the deuterium labelling studies, as well as PEMT

pathway synthesised PC if each of the three sequential methylations involves a *methyl*$D_3$ labelled S-adenosylmethionine. Similarly, if a PC species is deuterium labelled on one or two of its methyl groups, its mass will be three or six units higher than its non-labelled form. This can occur to the PEMT pathway synthesised PC species, which has undergone one or two methylation reactions involving labelled S-adenosylmthionine. The deuterium labelling process on the *methyl* groups of newly synthesised PC through both CDP-choline pathway and PEMT pathway are shown in Figure 2.

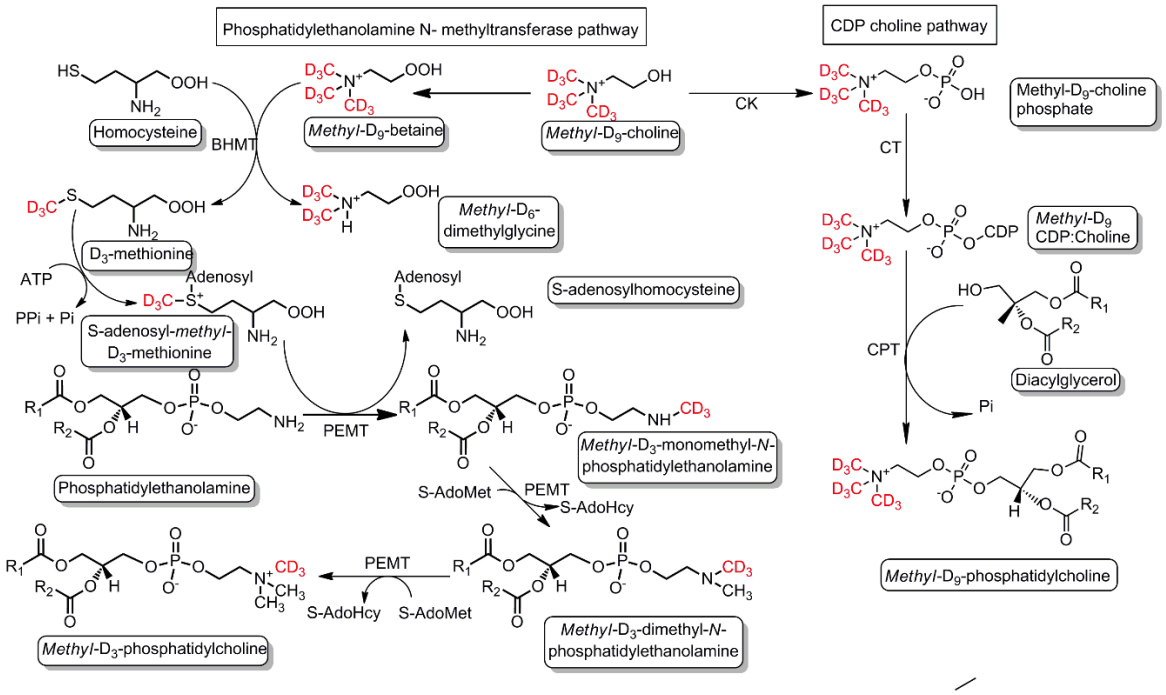



Figure 2: Deuterium labelling on PC methyl group through CDP-choline pathway and PEMT pathway (Dushianthan *et al.*, 2018).

The detection of deuterium labelled PC species is carried out by performing precursor scans using ESI MS/MS. In a typical precursor scan, as illustrated in Figure 3, a lipid sample in solvent is infused into ion source for ionization, and then the sample is sprayed out and desolvated to form gas-phase molecular ions. The first MS analyser MS1 sorts and selects

the molecular ions to be dissociated into fragments in the collision cell. In the second analyser MS2, fragment ions with specified m/z are selected for the subsequent detection and recording by a detector. The intensity of the fragment ions will be taken as the intensity of their precursor molecule ions.

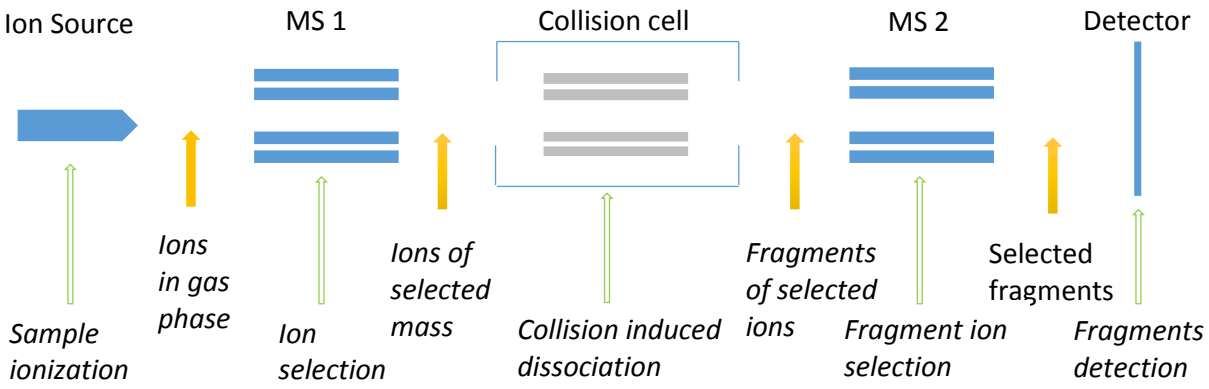



Figure 3: A schematic diagram of precursor scanning process of ESI MS/MS analysis.

The common precursor scan, denoted as P184, for PC detection reports the intensities of all the precursor ions that have a fragment ion with m/z value of 184, which include a phosphate bonded with a choline group making the head group of PC. While for the deuterium labelled forms of PC species, the scan is set to monitor the ions with fragment of a phosphate bonded with choline with deuterium labelled methyl groups. Precursor scan for fragment of m/z 187 (P187), monitors *methyl*$D_3$-PC species with one methyl group labelled with deuterium, which is three mass units higher than the non-labelled PC. P193 scan monitors *methyl*$D_9$-PC with three deuterium labelled methyl groups, nine mass units higher than the non-labelled PC. Figure 4 compares two spectra of ESI MS/MS analysis for PC species and the *methyl*$D_9$-PC species in the same bronchoalveolar lavage fluid (BALF) sample of a mouse model.

(A)

(B)

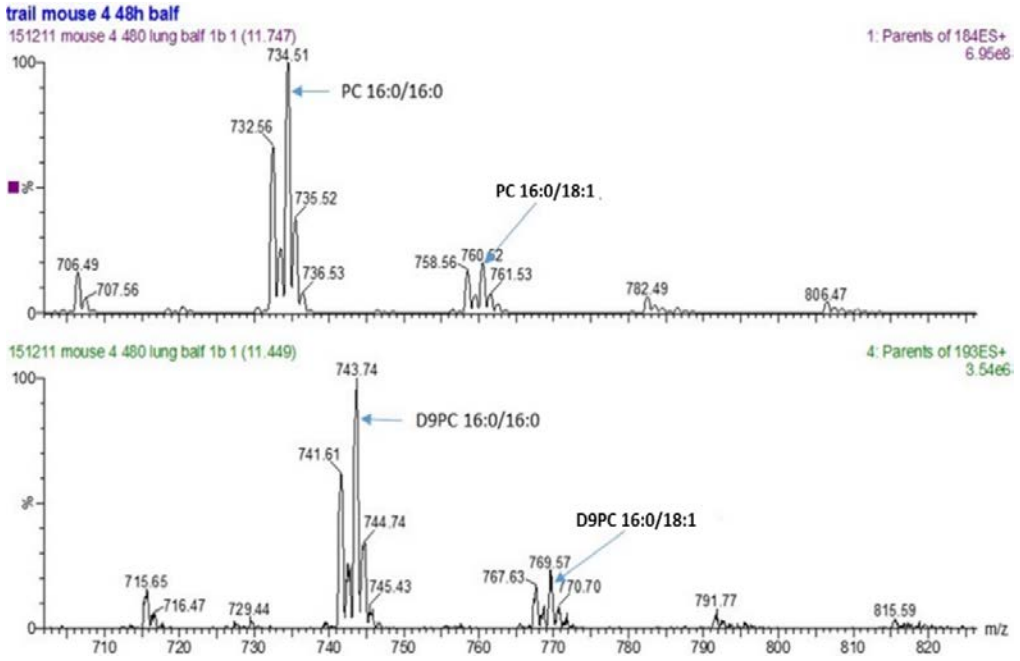



Figure 4: Spectra of P184 and P193 scans for PC and *methyl*$D_9$-PC detection, respectively, in the same BALF sample of a mouse. (A) is the spectrum of P184 scan with detected lipid molecules that have a PC head group. (B) is the spectrum of P193 scan with detected molecules that have a *methyl*$D_9$ labelled PC head group. The two spectra have similar patterns, as species detected in P193 are the labelled version of those in P184. However, there is a 9 mass offset between the two spectra due to deuterium incorporation on the three methyl groups of the PC species, resulting in nine mass higher in the P193 scan. PC16:0/16:0 (m/z 734.51) in the spectrum of P184 corresponds to its deuterium labelled form D9PC16:0/16:0 (m/z 743.74) in the spectrum of P193, and PC16:0/18:1 (m/z 760.62) corresponds to D9PC16:0/18:1 (m/z 769.57).

### 1.3.3 Dynamic surfactant phospholipids analysis with isotope labelling

#### 1.3.3.1 Surfactant phospholipids kinetics

Current knowledge regarding surfactant metabolism has largely been acquired from *in vitro* studies, which may not accurately reflect the real *in vivo* biological process. *In vivo* surfactant metabolic research facilitated by labelling techniques is a better way to understand the underlying mechanism of surfactant system. Using various labelled

materials with corresponding detecting techniques to trace *in vivo* activity of molecules has provided valuable insights of surfactant kinetics in the literature.

In labelling studies, the labelled molecules used for tracing endogenous molecules iare called tracer and the endogenous molecules being traced are called tracee. Early surfactant isotope labelling studies were dominated by using radioactive tracers due to the sufficient supply of radioactive labels, their low costs and the readily available scintillation counters for radioactivity determination. The application of this approach on animals has gained fundamental knowledge of animal pulmonary surfactant metabolism, in terms of surfactant synthesis, secretion, half-life and clearance. The mostly used radioactive isotope labels in surfactant PL kinetic studies include $^{14}C$, $^{3}H$ and $^{32}P$ (Foster and Bloom, 1963). Most of these *in vivo* surfactant investigations were focused on preterm or infant animals. Chida and Adams (1967) found an increased incorporation rate of $^{14}C$ labelled choline, palmitate and glucose into lung PLs, especially PCs, as the lamb fetus matured, which suggested the active synthesis of surfactant in the near term or new born and also implied the possibility of insufficient surfactant production in preterm animals. For the investigation of DPPC turnover, Young and Tierney (1972) used palmitate – 1 – $^{14}C$ as a radioactive label in rats and postulated that not all DPPC is randomly secreted through a single pathway by the same type of cells without recycling into lung tissue. The first study that investigated multiple phospholipid classes (PC, PG, PI and PE) in samples of different lung fractions including alveolar wash, microsome and lamellar bodies was performed by Jobe *et al.* (1978). The authors concluded that PC and PG have shorter half-lives and higher turnover rates than other PL classes. Using similar techniques, Jacobs *et al.* (1982) estimated the flux of surfactant PC from lamellar bodies in rabbits, and observed shorter turnover time and a higher flux rate in adults than baby

rabbits. These early surfactant PL kinetic studies have provided precious knowledge of the *in vivo* activities of lung surfactant from various aspect and laid a good foundation for the later understanding of surfactant metabolism in human.

Despite the success of *in vivo* lipid kinetic analysis, there are limitations of the radioactive isotope labelling methods. Firstly, the health risk radioactive material poses to the experimental subjects makes it not suitable for human studies. Secondly, the method tracks the specificity of the radioactive labels along the entire metabolic process of the radioactive label, rather than certain targeted molecular species of interest. Thirdly, information obtained from animal studies is limited and cannot be directly used in clinical work.

The improvement of mass spectrometry techniques has seen the increased use of the stable isotope labelling approach in dynamic lipid studies. The traditional GC–MS and gas chromatography-isotope ratio mass spectrometry (GC-IRMS) techniques (Torresin *et al.*, 2000; Cogo *et al.*, 2005), have provided powerful tools for precisely measuring stable isotopes and have been largely used in the studies of *in vivo* surfactant PL metabolic kinetics in human. Stable isotope labels are safe for human if used properly. Bunt *et al.* (1998) proved that the novel approach of stable isotope labelling they had used could safely address issues concerning surfactant metabolism. In their study, [U-$^{13}$C] glucose was used for investigating the *in vivo* surfactant synthesis and turnover in preterm infants. They found the labelled PC palmitate reached maximum at 70 ± 18 h, and calculated the secretion time, fractional synthesis rate, absolute production rate and half-life of surfactant PC palmitate. Similar approaches have been used in the studies of NRDS (Torresin *et al.*, 2000; Cavicchioli *et al.*, 2001). Using GC-MS, Cogo *et al.* (2011) have calculated the half-life, pool size and endogenous synthesis of surfactant disaturated PC

isolated from serial tracheal aspirates of preterm infants with moderate to severe NRDS. The authors suggested preterm infants with shorter disaturated PC half-life require additional surfactant administration.

The downside of GC-based techniques is that they generally require derivatisation of non-volatile lipids through saponification to generate methyl fatty acid esters to be analysed. As a result, the derivatised species mixtures, e.g. the overall surfactant disaturated PC rather than individual molecular specie (Facco *et al.*, 2014), are analysed collectively.

Application of direct infusion ESI-MS/MS in combination with stable isotope labelling techniques allows simultaneous identification and quantification of multiple lipid molecules, labelled and unlabelled, which has offered deeper insights of surfactant metabolism at species level (Postle and Hunt, 2009). Bernhard *et al.* (2004) applied this technique and investigated the kinetic turnover of surfactant PC species in healthy people. They found an overall incorporation rate of *methyl*D9 choline into sputum PC was linear between 12 and 30 hours. Moreover, the pattern of fractional enrichment of individual PC species over this initial period differed from that of surfactant PC composition. Applying the similar approach, Postle *et al.* (2011) investigated lung PC synthesis and secretion in SP-D null mice, and calculated the rate of accumulation in these mice with impaired surfactant catabolism. Dushianthan *et al.* (2014) analysed PC turnover in BALF samples of ARDS patients and healthy volunteers, both groups being given *methyl*$D_9$-choline infusion. They observed the fraction of deuterated DPPC of total deuterated PC changes along the time. The study found different surfactant PC kinetics between patients and healthy controls. Moreover, it varied considerably between individual patients, suggesting heterogeneity of the patients characterized by different PC metabolism.

#### 1.3.3.2 Limitations of traditional dynamic analysis of surfactant phospholipids

The common practice in surfactant PC turnover analysis facilitated by isotope labelling techniques involves calculating the kinetic parameters to get insights of mechanisms of surfactant regulation. The parameters of interest include fractional synthesis rate, synthesis rate, half-life, decay rates, turnover time, pool size and secretion time, which provide insights of metabolic flux characteristics of PL molecules. However, there are five main areas of concern in the accurate calculation of these parameters.

Firstly, the assumptions for these calculations are not always met. Most of the kinetic parameters formulae assume the surfactant system to be in steady state. However, in many cases we would like to extract the biological activities of compounds in patients with various conditions to understand disease processes. It is not appropriate to assume the metabolism being in steady state under these conditions, therefore the use of these formulae are questionable in such studies (Torresin *et al.*, 2000; Cavicchioli *et al.*, 2001; Cogo *et al.*, 2009)

Secondly, using different labelled precursors or MS instruments may generate different results. Estimated half-life values for a given phospholipid vary depending on the precursor used. For example, a study shows that if labelled palmitate is used, the half-life of PC reported from rabbits range from 17-20 hour. However if labelled choline is used as precursor, the range is 35-45 hours, and rat studies have reported similar trends (Wright and Clements, 1987). This led to the suspicion that labelled parts on PC from different precursors might metabolise differently. Bohlin *et al.* (2005) compared kinetic parameters obtained from several labelling studies on premature infants, which used different tracers, pools (PL or desaturated PL) and analytical methods (GC-MS/ GC-C-IRMS). It was observed

that fractional synthesis rate was specific to labelled precursor used and fractional catabolic rate was independent of tracers but dependent on the analysing methods. GC-IRMS generated slower apparent fractional catabolic rate than GC-MS.

Thirdly, time points chosen for kinetic parameters calculations are crucial. For example, fractional synthesis rate calculation only requires enrichment data at two time points near to the starting point (Foster *et al.*, 1993), which are supposed to be on the linear part of the curve. However, the two data points available for calculation are subject to the experimental design, which may not necessarily fulfil the assumption of being on the linear part or close enough to the starting point. Consequently, the calculated fractional synthesis rate may not be an accurate approximation.

Fourthly, it is not always possible to apply a standard label infusion method due to various reasons. For example, in the study of Dushianthan *et al.* (2014), a short period of intravenous label infusion was applied, rather than a standard bolus infusion or a continuous infusion, due to the need for a continuous labelling period and the unstable and fragile state of patients who may not tolerate longer period of infusion. The time course label enrichment obtained from this experiment may exhibit very different temporal patters from both bolus and continuous infusion, therefore the way of calculating kinetic parameters in normal methods may not directly applicable in this context.

Lastly, different functions or representations of variables are used for kinetic parameter c calculations in different studies. For example, there is a dispute over which function should be used for label enrichment calculation. For synthesis rate calculation, some studies use tracer/tracee ratio as the label enrichment, whereas some use atom percent

excess (APE) (Ramakrishnan, 2007). Different approaches remain in use by different researchers, although it is rarely explicitly stated what assumptions have been made and how the approaches are chosen.

Although kinetic parameter calculation is the main objective of traditional labelling studies, and the only way of getting insights of *in vivo* flux of molecules, the challenges stated above remain for obtaining accurate and consistent results. Therefore, using kinetic information of surfactant PL for biomarker discovery may be hard to achieve by using these kinetic parameters. In biomarker discovery studies, the essence is to extract variables that characterise conditions and are both measurable and comparable between groups for potential marker identification. In labelling studies, apart from kinetic parameters, the time course species label enrichment can also provide dynamic information and is comparable between groups, which is suitable for biomarker analysis. Nevertheless, using time course enrichment as variable adds an extra dimension to the data. Consequently, dynamic lipidomic biomarker discovery analysis will have to take into account multiple species, multiple subjects, different groups and time factors, which requires the aid of appropriate statistical methods.

## 1.4 Statistical methods for multivariate time series data analysis

There are very limited options of statistical methods for multivariate time series biological data analysis, more so when the experiments have limited and irregular time points, small quantities of biological replicates and numerous variables of interest.

Early methods for the analysis of time series biological data simply treat the time course data as a bi-dimensional problem and do not explicitly include the time variable in the model, unable to detect time-related variations. From the last decade, algorithms that are

able extract more information from time series data have been used in genomics and metabolomics studies. Smilde *et al.* (2010) proposed ANOVA-simultaneous component analysis method, combining ANOVA method and principle component analysis for longitudinal metabolic data analysis. The aim of the study was to determine the variation caused by experimental design, such as time, doses or combinations thereof, typically used in nuclear magnetic resonance. The drawback of the method is that it assumes the time related effects to be linear in relation to time, which is rarely a valid assumption (Rantalainen *et al.*, 2008). Breit *et al.* (2015) presented a computational modelling and statistical approach for identifying dynamic metabolic signatures by characterising and categorising kinetic patterns of circulating metabolites during physical exercises. The method utilises maximum fold changes and hypothesis tests to rank the metabolites according to their level of change over time. Putative biomarker candidates exhibiting different temporal patterns can then be selected and classified into several categories, such as early and late markers. This method aims to characterise putative markers time course patterns to be used as standard signatures of certain physical states, and is not for identifying biomarkers from subjects with different conditions. Huang *et al.* (2011) proposed a strategy for time series data biomarker discovery based on dynamic networks. The strategy detects biomarkers by investigating the changes of ratios of metabolites over time, stressing the association between metabolites during disease development. The authors applied the strategy to identify lipid ratio biomarkers of Hepatocellular Carcinoma by observing significant ratio changes at every two adjacent time points during disease progression. This is an attempt of seeking alternative variables by transforming or combining the observation data, which discloses more meaningful information than the observation data itself. Tai and Speed (2006) proposed multivariate empirical Bayes statistical approach for gene ranking in longitudinal replicated developmental microarray

experiments. The method uses Bayes statistics for ranking genes in one sample experiments as well as identifying genes whose temporal patterns differ across two biological conditions. The method has also been applied to proteomic data analysis. Berk *et al.* (2011) introduced a framework for the multivariate time series data analysis based on the method of smoothing splines mixed effects (SME) model. This is a framework designed for modelling high-dimensional time-varying metabolomics profiles and detecting significant differences between two biological groups, being capable of dealing with small sample size, irregular observation time, missing data and extreme outliers. The method was applied to real hydrazine data measured by nuclear magnetic resonance from the COMET longitudinal study, which proved to be able to efficiently identify known metabolites that are differentially regulated in hydrazine Toxicity. According to the author, the framework can be applied to longitudinal data acquired from mass spectrometers.

## 1.5 Bioinformatics tools for mass spectrometry acquired lipidomics data analysis and mass spectrometry data processing techniques

### 1.5.1 Bioinformatics tools for mass spectrometry based lipidomic analysis

Lipidomics is a comparatively new “omic” research field largely driven by the advances in MS techniques, and there are very few bioinformatics tools available for the MS facilitated lipidomic analysis. Around a dozen software applications are currently available offering basic tools for lipidomics studies. These software tools mainly focus on general lipid identification and quantification without comprehensive functionalities. Besides, all of them are designed for certain types of MS instruments and analysing protocols. Nearly half of the tools are developed for liquid chromatography (LC) MS lipid analysis, such as

LOBSTAHS (Collins *et al.*, 2016), LDA (Hartler *et al.*, 2011), MS-DIAL (Tsugawa *et al.*, 2015), Lipid-Pro (Ahmed *et al.*, 2015), LipidSearch (Houjou *et al.*, 2005) and LipidMiner (Meng *et al.*, 2014), leaving few tools for direct infusion MS lipid analysis.

A few software platforms available for direct infusion lipid data analysis are specific for certain instruments and protocols. For example, some tools are specific to time of flight mass spectrometer, which identifies the molecules by combining the fragments spectrum obtained from product ion scans. Such platforms includes LipidView (Ejsing *et al.*, 2006), which exploits multiple precursor scans approach, and LipidInspector, which is designed for the MS analysis in a data dependent acquisition mode. Alex (Husen *et al.*, 2013) is designed only for high resolution Orbitrap MS instrument. LipidQA (Song *et al.*, 2007) and LipidXplorer (Herzog *et al.*, 2012) are able to process data from a wider range of instruments acquired in data dependent acquisition mode or by multiple fragment scans. All of these software tools detect lipid molecules by matching the fragments, which is obtained by simultaneous acquisition of multiple product or neutral loss scans, with the reference spectra from certain databases that are specific to the data acquisition methods. This approach is suitable for global scanning of lipid contents in the samples. For the more focused lipid detection using precursor scanning technique performed by ESI MS/MS, which detect lipid molecules by combining the information of precursor ions and specified fragment ions, there is LIMSA (Haimi *et al.*, 2006) that supports the data processing and lipid identification, although the detected molecules are presented only in the cumulative composition format (e.g. PC 34:0) rather than the detailed form (e.g. PC 16:0/18:0), due to the limited information obtained from the fragment scans. However, just like other software, LIMSA does not support lipid identification for batch samples or labelling studies.

Surfactant lipid kinetic studies investigate the dynamic activity of lipid molecules in the surfactant system. As discussed in the previous section, shotgun ESI MS/MS offers great tool for analysis on focused lipids, which is ideal for labelling lipidomics studies. Despite the useful MS tools, the nature of labelling study requires integrated analysis for labelled and non-labelled lipid species for comparison and calculation, which relies mainly on the dedicated software tools. However, none of the software mentioned above handles lipidomics labelling data or perform kinetic analysis.

The mass spectrometry lipid biology team at Southampton University Hospital has developed a series of applications in Visual Basic for processing lipid data, and has incorporated some functionalities to handle labelling data. However, there are a number of shortcomings with the programme and the approach, which should be addressed further.

- It is cumbersome and not robust. The software crashes very often and running errors occur at times.
- The user interface is unintuitive and not easy to use. It lacks clarity and error preventing means.
- For labelling studies, the programme only allows labelling analysis of one PL class and does not support non-labelling analysis of other PL classes at the same time.
- It does not support labelling analysis of PL labelled with multiple labelling materials.

There is a need for a robust, fast running, user-friendly and more comprehensive software for the PL kinetic labelling studies as well as for the potential usage in clinical work.

### 1.5.2 MS data processing

The mass spectrometry acquired raw data is not directly useable, which requires multiple steps processing for further analysis. There are some common processing procedures needed for the data acquired by most MS methods, including background noise removal, peak identification and isotope correction.

#### 1.5.2.1 Background noise removal

The experimentally measured MS spectrum data is always interfered by background noise. This is mainly electrical noise related to the instrument set-up (Bauer et al. 2011). Consequently, removing background noise becomes a standard procedure of mass spectrum data processing, which aims to separate the analyte signals of interest from the background noise. Different algorithms have been applied for background noise removing on MS data acquired from different instruments. Berndt *et al.* (1999) applied linear fit for sequential mass segments on the mass spectra from MALDI MS and used cubic spline interpolation method to approximate the background. Some researchers estimated the MALDI spectrum background noise using a weighted average of the minimum and maximum peak values within windows of small mass intervals (Samuelsson *et al.*, 2004). Some others use polynomial fittings, such as locally weighted scatter plot smoothing. Wavelet transform theory also became popular for signal processing over the last two decades. Coombes *et al.* (2005) applied the translation-invariant undecimated discrete wavelet transform for noise filtering of SELDI spectra.

#### 1.5.2.2 Peak detection

Peak detection is a key aspect of the MS data processing. A widely used algorithm for peak detection is based on the calculation of signal-to-noise ratio (Petkovic *et al.*, 2001). The algorithm searches for the local maximum intensity among a certain number of

neighbouring data points with local signal-to-noise ratio over a certain threshold for peak detection. Some researchers (Strittmatter *et al.*, 2003) applied mixture models to fit the spectrum data acquired by TOF mass spectrometer for peak identification. Wehofsky and Hoffmann (2002) made use of isotopic patterns of species and fitted the theoretical patterns to the MALDI spectrum data to define peaks. Another set of peak detecting techniques based on the continuous wavelet transform theory (CWT) were developed and widely used in the last decade. Du *et al.* (2006) proposed an algorithm that applies the CWT over spectrum and utilises the information over the 2D CWT coefficients to determine the effective signal-to-noise ratio to identify peaks. The peak detecting techniques available today are often specific to mass spectrometer used, and all of them have their own advantages and drawbacks. Therefore, there is no standard method that out performs others in every aspect (Bauer *et al.*, 2011). As a result, researchers always choose the appropriate methods for peak detection according to specific need of their studies.

#### 1.5.2.3 Isotope distribution calculation

Isotope peak patterns in the mass spectra of compounds are the result of the natural occurrence of polyisotopic elements. In order to remove the isotope effect from the spectra, the isotopic distribution must be calculated first.

The calculation of isotopic distribution has been extensively studied since 1960. The early methods were mainly based on probability theory, and later its combinations with polynomial expansion methods were explored and widely used by researchers (Yergey, 1983). However, these methods require computer intensive calculation and have difficulties in dealing with large molecules, due to the large numbers of possible isotopic combinations that need to be calculated. Consequently, efforts were made to reduce the

amount of computation and increase the computing speed. A core concept of step-wise calculation was applied to tackle this problem by many researchers, using a variety of approaches. Snider (2007) proposed a method that kept track of all isotopes of a built up molecule every time an element is added onto it. This method is based on dynamic programming technique, and has become one of the most widely used algorithm for isotope distribution calculation.

## 1.6 Research questions and structure of the thesis

### 1.6.1 Thesis objectives

Pulmonary surfactant is a critical substance for optimal lung function. The metabolism of the major surfactant component surfactant PLs, as well as the underlying mechanisms of their alteration in lung diseases are still not fully understood. The newly adopted ESI MS/MS coupled with stable isotope labelling technique makes it possible to investigate *in vivo* kinetics of surfactant PL species in human. However, the multi-step processing of raw MS data and the calculation of labelled species enrichment require a high level of expert knowledge, as well as the automatic computational tools for dealing with the high throughput lipidomics data from batches of samples. The first objective of this project is to develop a streamlined software platform for ESI MS/MS based pulmonary surfactant PL labelling studies, which should be fast, robust and user friendly, and could potentially be used for clinical diagnosis.

Despite the endless effort in biomarker discovery in lung diseases, there is hardly any novel biomarkers that are verified and widely adopted in clinical work. There are some limitations with the current studies. Firstly, the predominantly studied molecular biomarkers in the studies of lung disease are proteins, genes and metabolites (Eisner *et*

*al.*, 2003; Terpstra *et al.*, 2014; Ravipati *et al.*, 2015; Chen *et al.*, 2016; Gray *et al.*, 2017; Kim *et al.*, 2017). Lipid biomarkers, especially surfactant PLs, which are found altered in various lung diseases, have not been fully explored as putative biomarkers. Secondly, most of the studies only concentrate on a single species from a certain compartment for biomarker assessment (Eisner *et al.*, 2003; Il Yoon and Sin, 2011; Terpstra *et al.*, 2014; Yu *et al.*, 2017). However, the pulmonary diseases are usually associated with alterations involving multiple interactive mechanisms, resulting in systemic changes of numbers of molecules. Besides, the same molecules from different pathways, compartments and even organs are correlated with each other. It is hard for a single species from a single pool to work individually to indicate changes of the entire metabolic process, recognising a disease or stratifying patients. Furthermore, many lung diseases have similar manifestations, which makes it hard to distinguish them from changes of certain molecules. Collective biomarkers or omics biomarkers from different pathways and compartments could be more helpful in providing sufficient information for disclosing underlying mechanisms from various aspects. Thirdly, the current interests in the molecular biomarker studies are limited in measurable concentration of end product molecules. However, the end product concentration does not always give sufficient information for discriminating biological conditions. Moreover, the snapshot concentration level does not reflect *in vivo* kinetics of molecules while undergoing biological processes. Kinetics are critical for disclosing the underlying mechanism of biological systems, which may contribute to identifying different biological conditions, and become valuable biomarkers. There are a few longitudinal studies that monitor the change of concentration levels of potential markers over a time period. Longitudinal analysis takes time factor into consideration, which provides extra information and offers more complete picture of changing process of molecules. However, time series

concentration of a metabolic product only offers vague and limited information of its metabolism, which does not indicate flux and turnover of molecules that characterise their changing process and causes.

Dynamic surfactant lipidomics is an untapped area for pulmonary disease biomarker discoveries, which can provide rich sources of kinetic knowledge in addition to the current understanding of the pulmonary diseases, paving the way for the discovery of novel individual biomarker or biomarker panels of collective markers. Analysis of lipids from compartments, pathways and organs other than lung surfactant may complement the understanding of surfactant metabolism and provide combined biomarkers for lung disease diagnostic purposes.

Surfactant PLs, especially PC, are the major surface active material maintaining the normal functioning of the lung and PC alteration is found in most lung diseases. Facilitated by the high throughput ESI MS/MS coupled with stable isotope labelling technique, it is possible to calculate the label incorporation of numerous PC molecular species for human *in vivo* studies. Dynamic PC analysis often involves kinetic parameter calculation in the literature. However, there are various issues with the precision and comparability of theses parameters. The label enrichment of PLs over certain time period holds flux information, which reflects the metabolic process of molecules, offering extra layer of information than the end product concentration.

SME method provides a great platform for multivariate time course modelling to characterize temporal features of variables from longitudinal studies, which makes the basis for further kinetic biomarker discovery analysis. The method also benefits from

being able to deal with small sample size and missing data, typical issues in longitudinal biological studies.

The second objective of this project is to investigate the feasibility of using time course enrichment of deuterium labelled surfactant PC species as potential biomarkers for identifying pulmonary diseases, facilitated by SME multivariate time course statistical method. Label enrichment of hepatic PC species will also be evaluated, where data is available, to complement the analysis of surfactant PC.

### 1.6.2 The arrangement of the chapters

Chapter 2 introduces the SME statistical methods and other tools used for PC time course label enrichment biomarker analysis. Chapter 3 introduces the development of a dedicated software platform, LipidomeLabelling, for lipidomics kinetic analysis facilitated by ESI MS/MS coupled with stable isotope labelling techniques. In chapter 4, 5 and 6, time course labelled PC enrichment data from three previous projects is modelled using SME, exploring the possibility of using the time course enrichment as biomarker for lung diseases. In chapter 4, the surfactant PC kinetic biomarker analysis for mice models was performed, aiming to use an established mice model system with a large difference in effect to validate the methodology, which will be applied in the following less clear clinical scenarios. In chapter 5, temporal patterns of surfactant PC label incorporation of NRDS patients were modelled and analysed for investigating their ability in distinguishing patient sub groups in the complex clinical environment. Plasma PC were also analysed in the same way providing information regarding liver function. In chapter 6, time course plasma PC label enrichment of ARDS and healthy people was analysed to check if hepatic

PC are indicative for distinguishing the two groups, which have significant difference in surfactant turnover.

# Chapter 2 Methodologies

## 2.1 ESI MS/MS data acquisition

The PC data from different projects analysed in this thesis were all acquired through ESI MS/MS analysis (Waters, UK). Precursor scans carried out for these projects include P184, P187 and P193 for the detection of endogenous PC, labelled *methyl*$D_3$-PC and *methyl*$D_9$-PC, respectively. The scanned m/z range is 400–900. The lipid sample solvent were infused directly into ion source at a certain rate (5µl/minute or 8µl/minute). Energy for the cone, which is used for extracting charge molecules after ionization and before entering the MS 1, was set to 50V. Collision energy was set to 30V in the collision cell for ion fragmentation. The resolution was set to be the same over the whole mass range with fixed peak width at half height of 0.703 mass unit for the two analysers.

## 2.2 Statistical method

In this thesis, SME statistical framework is used for modelling the time course enrichment of multiple deuterium labelled PC species, as well as group comparison for biomarker discovery. For each investigated species, its temporal label enrichment for every individual subject of the two comparing groups as well as the group mean temporal enrichment were modelled first for subsequent statistical comparisons.

The main functional mixed-effects model of SME method includes a fixed unknown term, a random effect term, and an error term, as shown below

$$y(t_{ij}) = \mu(t_{ij}) + v_i(t_{ij}) + \epsilon_{ij} \quad (1)$$

The observation value of individual $i$ at time $t_{ij}$ is denoted by $y(t_{ij})$, $\mu(t_{ij})$ is the mean value of the whole group at the time point, and $v_i(t_{ij})$ is the individual-specific deviation from the group mean. The additive error term $\epsilon_{ij}$ is assumed to have some group and species-specific variance. Both $\mu$ ( • ) and $v_i$ ( • ) are treated as smooth functions of time. $\mu$ ( • ) is the fixed, unknown population curve, and in the label enrichment analysis it represents the group mean temporal enrichment, whereas $v_i$ ( • ) is treated as a random realisation of an underlying Gaussian process with the mean of zero. $\mu(\cdot) + v_i(\cdot)$ is the smooth underlying PC enrichment function to be estimated. The smoothing parameters of time course models are estimated by optimizing the corrected Akaike Information Criterion (AICc) (Hurvich et al., 1998).

In the biomarker discovery analysis, the mean time course enrichment of certain species for two biological groups are modelled, denoted as $\mu_A$ and $\mu_B$, and SME method offers a moderated functional *t*-type statistic *Ft* for quantifying the difference between them over the entire time course. The formula is shown as

$$Ft = L / (s_e + s_{em}) \quad (2)$$

, where the numerator *L* is the distance between the two mean curves $\mu_C$ ( • ) and $\mu_T$ ( • ).

*L* is the square root of $\int_{t_{min}}^{t_{max}} [\mu_A(t) - \mu_B(t)]^2 \, dt$, indicating the distance between the two mean curves over the period from $t_{min}$ to $t_{max}$. The term $s_e$ is the functional standard error, computed as the square root of $\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}$, is the sum of sample functional variances of the two groups, with $s$ and $n$ being the estimated variance and sample size of each group. Similar to the standard *t*-test, the role of $s_e$ is to scale the *L* distance so that the statistic *Ft* reflects the relative difference between the two mean curves of the groups by

taking into account individual variance, and is comparable across different species observed on different time scales. $s_{em}$ is the term to moderate the *Ft* statistic aiming to correct the effect on the *Ft* statistic caused by inaccurately fitted variances due to small sample size, especially when having large number of variables. However, although the sample sizes are small in all projects analysed in this thesis, the number of variables, i.e. the PC species, is comparatively small, which does not hugely affect the result when comparing *Ft* statistics across all the species. Therefore, in this thesis $s_{em}$ is not adopted, with only the $s_e$ term left as the denominator.

The label enrichment in all PC labelling analysis in the thesis is presented as APE, which is calculated as the ratio of intensities of labelled species over the sum of intensities of labelled and unlabelled species:

$$E_i = \frac{Lint_i}{Lint_i + Uint_i} \quad (3)$$

, where $i$ represents a PC species, $E_i$ is the deuterium label enrichment of species $i$, $Lint_i$ is the intensity of labelled species $i$, and $Uint_i$ denotes the intensity of unlabelled endogenous PC species $i$.

The modelling of time course enrichment of labelled PC species in this thesis has excluded zero time point observations, which is the starting point of the label infusion. This value should always be zero, as there is no immediate label incorporation while label infusion just starts. Moreover, although it is not certain when the incorporation starts, the appearance of the labelled species in the pool is always after zero time point, either sooner or later after infusion. As the SME method is a piecewise smoothing method that fits curves to the available time points, including zero time point, it does not take into

account the time point of species appearance that is not available. Therefore, there is a risk that SME takes zero time points as the emerging time point and links it with the next available time point observation to simulate the species appearance curve, which is incorrect.

The time course data modelling and biomarker discovery analysis in this thesis were carried out using the SME package written in R provided by Berk *et al.* (2011). A VBA programme was written to convert the enrichment data into required format for the SME package.

# Chapter 3 A Software platform for kinetic surfactant phospholipid analysis facilitated by ESI MS/MS coupled with stable isotope labelling techniques

## 3.1 Introduction

The comparatively new approach of stable isotope labelling combined with ESI MS/MS for PL labelling study suffers from a lack of dedicated streamlined software platform that offers efficient and comprehensive functionalities for MS data processing and dynamic lipidomics calculation. Consequently, a platform LipidomeLabelling was developed for the purpose, which is written in Matlab and VBA. The workflow of all functionalities are illustrated in Figure 5.

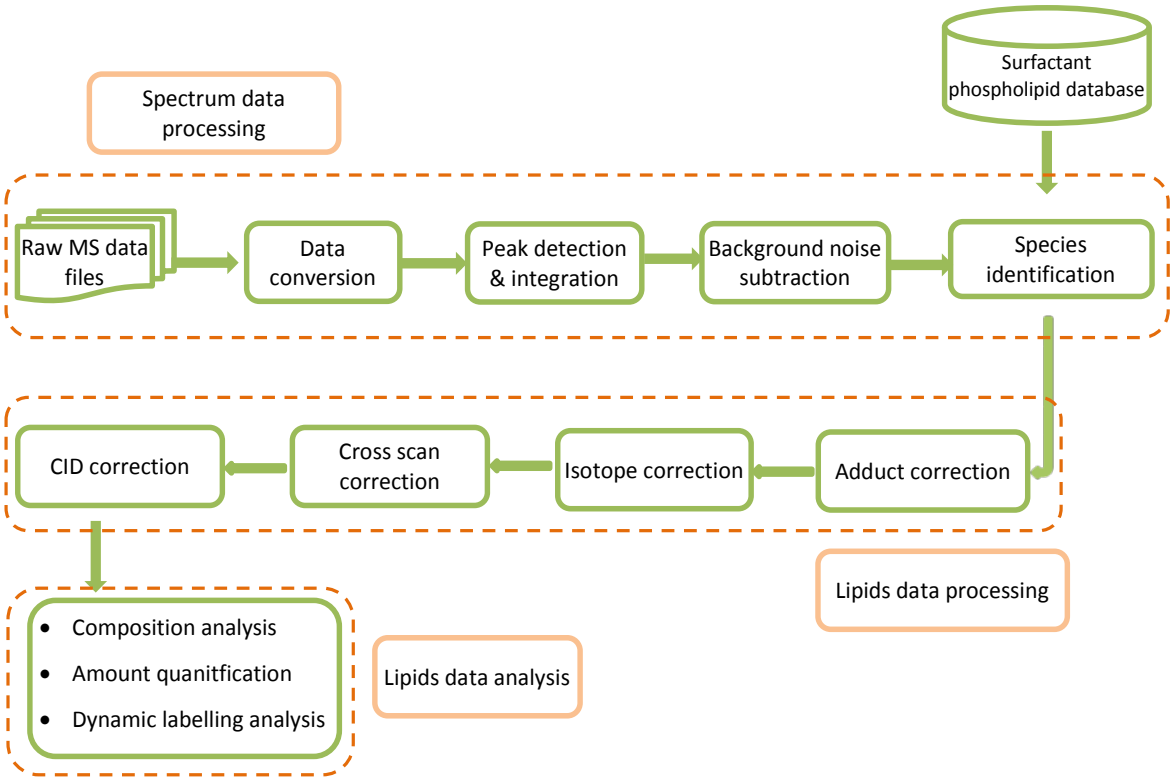



Figure 5: Flow chart of functionalities of LipidomeLabelling.

## 3.2 ESI MS/MS spectrum data processing

### 3.2.1 Data conversion

Spectrum data acquired by ESI MS/MS instrument (Xevo TQD, Waters) is initially analysed in MassLynx (Waters Coporation) and the raw data is stored in a manufacture specific format. As the data is in a format that is not readable or usable, Waters offers a library called Data Access Component that only allows Visual Basic or VC++ programmes to access the raw data. To convert the data into a manageable format, a new version of extracting programme is developed in VBA by modifying the previous in house programme developed by Dr Grielof Koster (Mass spectrometry lipidomics group, University Hospital Southampton). The new extractor programme reads the raw data by calling functions provided in the Data Access Component and exports into excel files. The old programme performs several steps of data processing, so the output is the processed result rather than the original raw spectrum data. In a typical ESI MS/MS analysis, multiple scans are performed on each sample to detect molecules within a certain m/z range. The extractor creates an excel file for every single scan performed for a sample. The acquisition result in each scan excel file is a list of figure pairs with m/z value and its corresponding ion intensity. In a typical scan, the MS instrument is set up to measure the intensities of sixteen evenly scattered m/z points within each mass unit. Table 1 shows a scan result at sixteen points within one mass unit of m/z 401. So if a range of 100 mass unit, say 400 – 499, is scanned, the intensities of 16 *100 mass points will be reported.

| *m/z* | Intensity |
|---|---|
| 401.0337 | 1622 |
| 401.0968 | 1847 |
| 401.1598 | 1701 |
| 401.2228 | 1848 |
| 401.2858 | 13627 |
| 401.3489 | 11004 |
| 401.4119 | 1758 |
| 401.4749 | 2807 |
| 401.5379 | 2279 |
| 401.601 | 1887 |
| 401.664 | 1785 |
| 401.7271 | 23584 |
| 401.7901 | 2686 |
| 401.8531 | 2139 |
| 401.9161 | 1399 |
| 401.9792 | 2627 |

Table 1: An example of raw MS data of *m/z* and intensity pairs measured at sixteen points within the mass unit of 401.

### 3.2.2 Peak detection and integration

Peak detection is the first step for MS data processing. It forms the basis for the following processing procedures, and accurate peak detection is critical for the data analysis. In a spectrum, the detected ions are presented as intensity peaks plotted over their m/z values. A peak is formed by joining up the intensities measured around the m/z value of the ion. A peak detection module is developed to search for the m/z position of the centre of each peak, which is considered as the m/z value of the detected ion, and to integrate the intensities measured at sixteen time points of the peak for this ion. The integrated intensity is taken as the abundance of a measured ion.

Most of the published peak detection algorithms are specific for MALDI-Tof, SELDI-Tof, or LC/MS, GC/MS strategies (Yang *et al.*, 2009), where there is a need to manage the inconsistent peak shape, peak placement (floating mass Tof), curved baseline and peak

width (LC). The ESI MS/MS spectrum benefits from fixed peak width (usually set to be 0.703 mass unit), quasi-normal distributed peak shape and comparatively stable baseline, so it does not suffer from the problems of other MS approaches mentioned above. Nevertheless, the peak centre can be hard to find due to various interference, such as chemical noises.

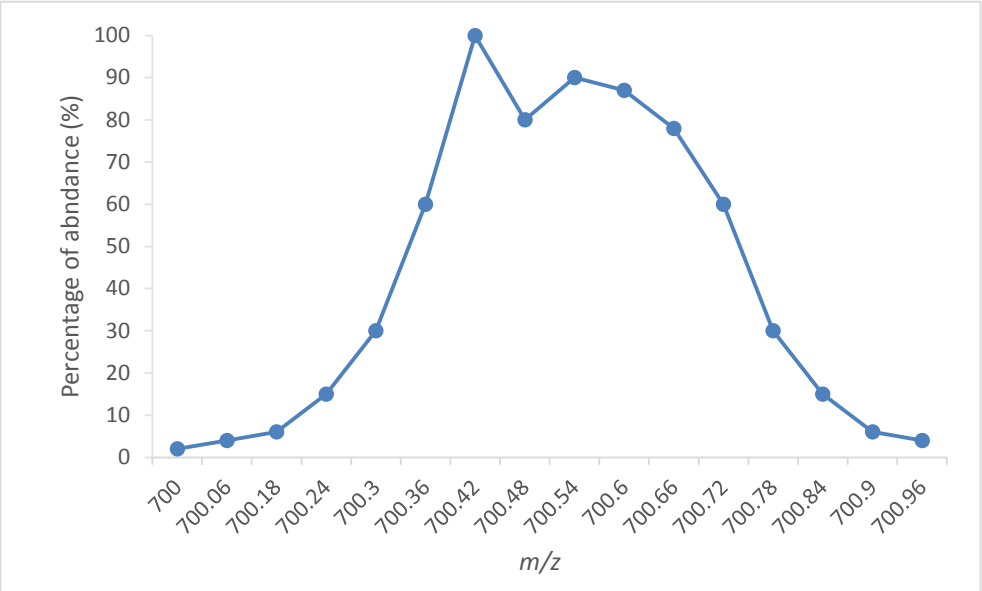



Figure 6: A peak of a molecule with an abnormal shape caused by chemical noise. The highest intensity is observed at 700.42 m/z point, although 700.54 should be the actual centre of the peak.

A perfect peak should have a quasi-normal distribution shape, and the highest intensity within the peak should appear at the centre of the peak, where the m/z value of the molecule locates. However, due to the instrument instability and noise interference, the intensities obtained for a species do not always form a perfect peak, and the detected highest intensity is not necessarily located at the peak centre. In an example shown in Figure 6, the m/z value 700.54 should be the centre of the peak, which is the m/z value of the detected ion, with the highest intensity. However, the highest intensity is observed at

700.42, which is not the real peak centre, due to the noise affecting the accuracy of the measurement.

To remove the noise effect and accurately recognise the peak centre, a weighted moving average algorithm is applied in the peak detection procedure. The programme checks the weighted sum of adjacent intensities of every five mass points through the entire mass unit. The five mass points group that has the largest total of weighted intensities is taken as the top and centre part of the peak, and the m/z of the middle point of the five is taken as the centre mass point of the whole peak. In the spectrum, five points covers about 0.25 mass unit, and a third of a peak width (0.703 mass unit). The weights used for the five adjacent points are 1,2,3,2 and 1 respectively, which gives the middle point the biggest weight. The proposed peak detecting algorithm can be demonstrated by the following model:

$$\text{Maximize} \quad WS_M(j) = Int(j-2) + 2 * Int(j-1) + 3 * Int(j) + 2 * Int(j+1) + Int(j+2)$$

$$\text{Subject to} \quad M \leq j < M+1 \tag{4}$$

, where $WS_M(j)$ denotes the weighted sum of intensities of five adjacent points, the centre of which is mass point $j$, within the mass unit $M$. $Int(j)$ represents the intensity at point $j$ and equation (4) is set up to find the point $j$ that maximize this value.

To demonstrate the effectiveness of the algorithm, it was applied to the example discussed above. Table 2 lists the measured intensities of sixteen m/z points within the 700 as well as the weighted sum of every five adjacent points. Even though the highest intensity (125000) is observed at 700.42 mass point in the original spectrum, the highest weighted sum of intensities (977500) is found at mass point 700.54, so the centre m/z

point of the peak detected by the algorithm is at 700.54. A new spectrum peak defined by the intensities calculated using weighted sum method is shown in Figure 7, which has corrected the effect of chemical noise shown in Figure 6 to present a smooth peak with a quasi-normal shape and find the real peak centre, i.e. the m/z of detected molecule.

| Mass point | Intensity | Weighted Sum |
|---|---|---|
| 700.00 | 2500 | |
| 700.06 | 5000 | |
| 700.18 | 7500 | 110000 |
| 700.24 | 18750 | 226250 |
| 700.3 | 37500 | 432500 |
| 700.36 | 75000 | 668750 |
| 700.42 | 125000 | 875000 |
| 700.48 | 100000 | 958750 |
| 700.54 | 112500 | 977500 |
| 700.6 | 108750 | 921250 |
| 700.66 | 97500 | 810000 |
| 700.72 | 75000 | 622500 |
| 700.78 | 37500 | 405000 |
| 700.84 | 18750 | 226250 |
| 700.9 | 7500 | |
| 700.96 | 5000 | |

Table 2: Intensity and weighted sum calculated for every five adjacent points in the m/z 700 of the peak in Figure 6.

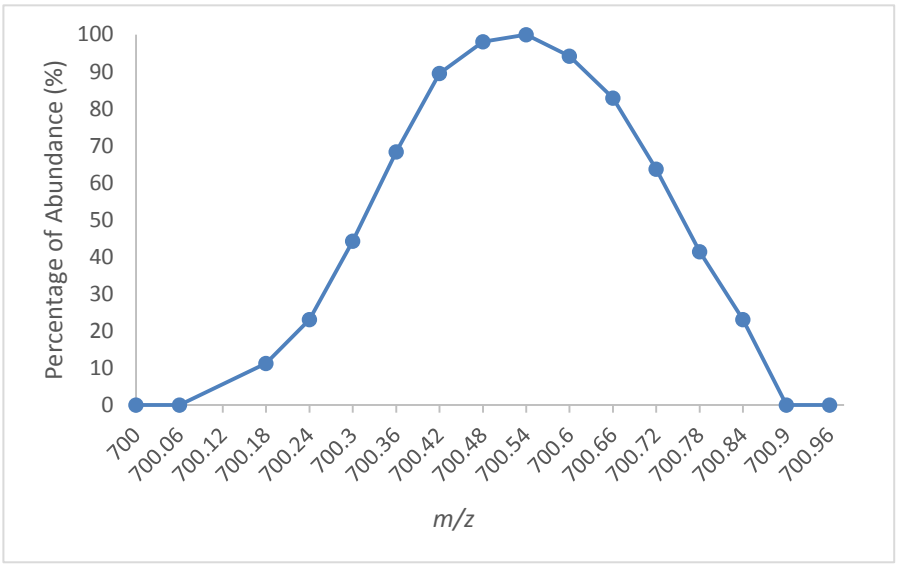



Figure 7: The peak defined by the weighted sum intensities in comparison with the original peak in Figure 6. The peak corrects the effect of chemical noise, and recognise the true peak centre in 700 mass unit, with the highest point at m/z 700.54.

To avoid the situation where a peak with a substantial noise is mistaken as two separate peaks as well as to take into account only one peak in one m/z unit, the following constraint is added to the peak detection procedures.

$$Int_M(m) \; > \; Int_{M+1}(n) \qquad if \quad m + 0.352 < n \qquad (5)$$

, where *m* and *n* are the mass points where the highest intensities are found in two adjacent mass unit *M* and *M*+1, respectively. This constraint states that if the distance between *m* and *n* detected in *M* and *M*+1 is smaller than half of the peak width 0.352, then the intensity of *m* in the mass unit *M* has to be higher than the intensity of mass point *m* in mass unit *M*+1. This is to make sure that the highest point in *M* is not on the slope of the peak in *M*+1. If the distance between *m and n* is no smaller than 0.352, there is no such concern. If the constraint is not satisfied, the peak found in mass *M* will be discarded.

Once the peaks are detected, the intensity of the molecule is approximated by averaging the intensity of points measured around the peak centre. In the ESI MS/MS spectrum, most peaks suffer from unsmooth peak shape, due to noise or measuring errors, and even the rare unaffected peaks are not in perfect shapes, such as Gaussian and Lorentzian, therefore, it is not accurate to fit the peaks using any of the commonly used peak-fitting methods. Consequently, the programme adopts the simple solution of taking the mean values of each peak to be compared with other peaks. It calculates the average of signal intensities of the detected mass points within 0.352 mass distance on both sides of the peak centre, as shown in equation 6. By doing this, the original sixteen data points in each mass unit become one average intensity value, which is used to represent the signal of the molecule.

$$AInt(M) = \frac{1}{N} \times \sum Int_i \qquad j - 0.352 \leq i \leq j + 0.352 \qquad (6)$$

$AInt(M)$ is the average intensity of the molecule detected in the mass unit *M*, $j$ is the mass point of the peak centre, $i$ represents the detected mass points that fall within the range of 0.703 mass unit around the centre point, $Int_i$ is the intensity measured at mass point $i$, and *N* is the number of points included.

Due to the low resolution of the quadrupole mass spectrometer, overlapping peaks can occur. In the peak detection procedure, the peak weighted moving average method with the constraint to drop the peak with lower intensity than a peak closely behind it, attempts to reduce the complexity and only take into account overlapping peaks with reasonable distance of 0.35. The peak integration takes into account the detected points within 0.703 mass range, which is about 0.35 from the peak centre from both sides, to further avoid including intensities of overlapping peak.

Figure 8 shows the integrated peak of the previous example shown in Figure 6 with an average intensity of 52109.38. Figure 9 present a real spectrum and its integrated version. Figure 10 is the flow chart of peak detection and integration procedures.

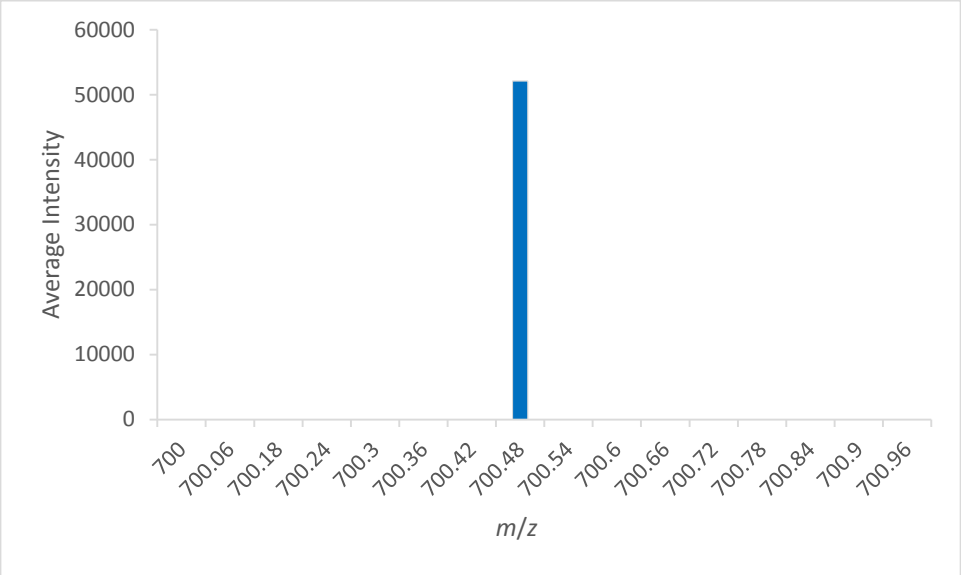



Figure 8: Integrated sample peak in Figure 6. The peak becomes a vertical line after peak integration by taking the average intensity of points around the centre mass point 700.54.

(A)

(B)

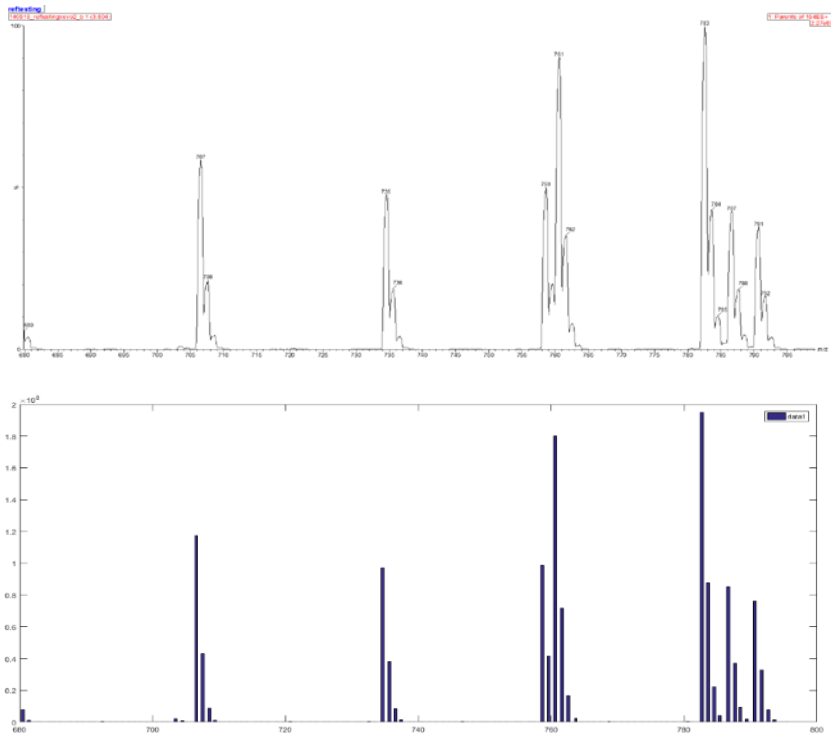


Figure 9: Comparison of spectrum before and after integration. (A) is a part of a real spectrum of P184 scan, and (B) is the result spectrum after peak detection and integration procedures.

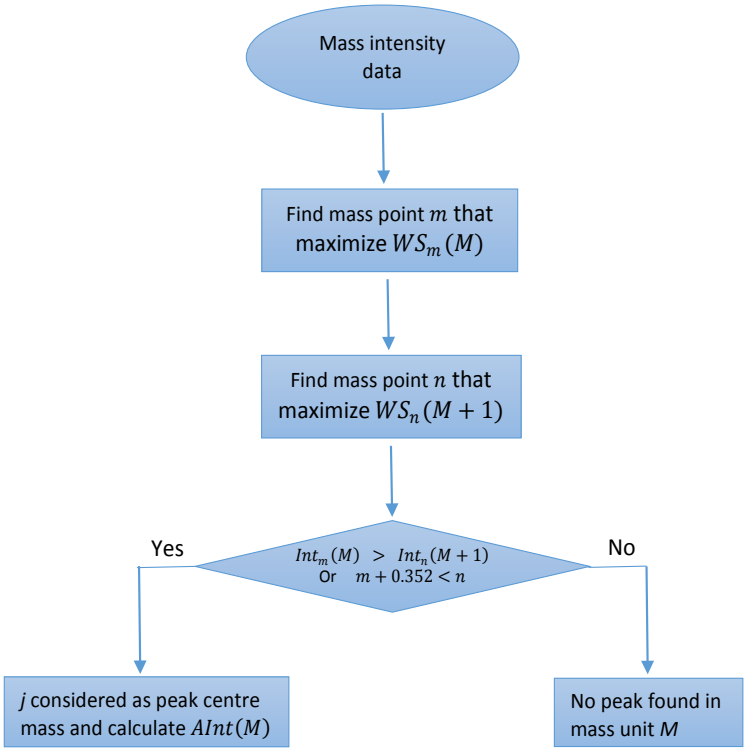



Figure 10: Flow chart of peak detection and integration procedure.

### 3.2.3 Background noise removal

Background noise exists in every spectrum and LipidomeLabelling provides an optional function for removing it. As MS data generated from ESI MS/MS usually has stable and low baseline noises, a uniform background approximation for the whole spectrum is estimated by taking the average of intensities of a user specified percent of the lowest peaks in the spectrum. The estimated background is then subtracted from all detected peaks over the whole spectrum. Figure 11 displays the plots of a spectrum before and after background is removed.

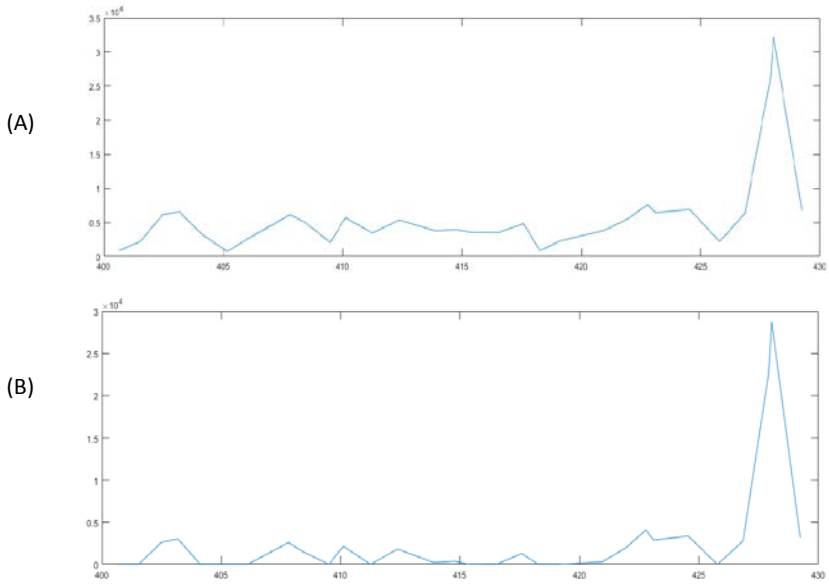



Figure 11: Comparison of spectrum before and after background removing procedure. (A) is the original spectrum and (B) is the result spectrum after an estimated background noise is removed.

## 3.3 Knowledge-based phospholipid annotation

### 3.3.1 Phospholipid database for stable isotope labelling studies

Once the ion peaks are detected, the species annotation are performed to give biological meanings to the peak data, getting ready for the following molecule-specific processing steps and analysis.

Species annotation for m/z value - intensity pair data acquired from ESI MS/MS relies on ·referencing the appropriate lipid databases which support ion mass searching approach. There are very few lipidomics databases currently available for lipid research, and LIPID MAPS (LIPID MAPS, 2018) and LipidBank (LipidBank, 2007) are the most popular ones. However, these databases are not suitable for MS/MS acquired lipid data identification. They are mainly for global search of lipids, and a confident identification of a species requires more information, such as mass, chemical formula, lipid category, lipid

class and molecular structure. As stated above, the ESI MS/MS acquired data only provides *m/z* values of ions and their fragments detected by break down scans, but these databases do not support lipid searching based on these pieces of information. Moreover, none of the available databases support labelled species identification, which is critical for the stable isotope labelling dynamic lipidomic analysis.

A dedicated database for identification of phospholipid and stable isotope labelled phospholipid based on spectrum data acquired through ESI MS/MS is created by Dr Grielof Koster. This database comprises of numerous species lists of certain classes of phospholipids detected by numerous MS scans. Each list is for a particular scan type (e.g. P184 for PC species) and a certain sample type, including BALF, lung tissue, cell suspension of BALF and plasma. The species lists are created based on the empirical knowledge of phospholipids that can be found in each sample type as well as theoretically possible lipid species ,which is for preliminary species identification, and more precise distributions of positions of acyl groups of the observed molecules can be verified by doing further break down scans.

There are hundreds of empirical and theoretical phospholipid species in each species list in ascending order of their m/z values. The lists are made to be complete to cover as many masses as possible. Table 3 displays part of the species list of a normal P184 scan. The species are annotated in the commonly accepted shorthand form. The annotation describes the species chemical structure to the fatty acid level and number of double bonds. For example, “PC 18:0/18:2” is the notation of a PC species that has two fatty acyl chains of 18 carbons, with no double bonds on sn-1 position and two double bonds on sn-2 position.

| m/z | Species list for P184 scan |
|---|---|
| 580 | PC12:0a/10:0 |
| 582 | LPC24:6a |
| 590 | PC12:1/10:1 |
| 592 | PC12:0/10:1 |
| 594 | PC12:0/10:0 |
| 596 | LPC24:6 |
| 603 | SM10:1a |
| 604 | PC10:1a/14:1 |
| 605 | SM10:0a |
| 606 | PC14:1a/10:0 |
| 607 | H2SM10:0a |
| 608 | PC10:0a/14:0 |
| 610 | ox2PC16:0/5:1 |
| 616 | LPC26:3a |
| 617 | SM10:1 |
| 618 | PC14:1/10:1 |
| 619 | SM10:0 |
| 620 | PC14:1/10:0 |
| 621 | H2SM10:0 |
| 622 | PC10:0/14:0 |

Table 3: Part of the species list of a normal P184 scan. It displays 20 species (out of 349 species in the full list) arranged in the ascending order of their m/z value.

The problem with the database is that the annotations are more of indicative rather than absolutely accurate, which assigns the detected molecules only to the most likely occurred species, because the precursor scan does not provide enough information for a confident identification in terms of the length of fatty acyl chains as well as the position of double bond. The database makes this compromises to generate a speedy assignment for the subsequent data processing. Further fragmentation scans will be needed for a more accurate identification.

### 3.3.2 Species assignment

LipidomeLabelling adopts the database created by Dr Grielof Koster for species identification and annotation. The PL species assignment is carried out by mutually

matching the m/z values of the detected peaks with the m/z values of species in the species list of the corresponding scan type. If there is a match, the m/z -intensity pair data will be assigned with the matching species, which is how the original spectrum data in the form of m/z intensity pair becomes real species with intensities. The detected peaks of which the m/z value could not find a match in the species list will be removed. Therefore, the species lists have to be as complete as possible to avoid removing of real species. Table 4 is an example of part of the result of an annotation. Unlike the old programme that does not allow amendment to the database, LipidomeLabelling allows users to access and edit the library for specific use.

| m/z | Combined | Detail | Intensity |
|---|---|---|---|
| 580 | PC22:0a | PC12:0a/10:0 | 38355 |
| 582 | LPC24:6a | LPC24:6a | 1524 |
| 590 | PC22:2 | PC12:1/10:1 | 113077 |
| 592 | PC22:1 | PC12:0/10:1 | 25324 |
| 594 | PC22:0 | PC12:0/10:0 | 25110 |
| 596 | LPC24:6 | LPC24:6 | 10914 |
| 603 | SM10:1a | SM10:1a | 1216 |
| 604 | PC24:2a | PC10:1a/14:1 | 6227 |
| 605 | SM10:0a | SM10:0a | 32691 |
| 606 | PC24:1a | PC14:1a/10:0 | 74621 |
| 607 | H2SM10:0a | H2SM10:0a | 23715 |
| 608 | PC24:0a | PC10:0a/14:0 | 59230 |
| 610 | ox2PC21:1 | ox2PC16:0/5:1 | 9686 |
| 616 | LPC26:3a | LPC26:3a | 4096 |
| 617 | SM10:1 | SM10:1 | 11967 |
| 618 | PC24:2 | PC14:1/10:1 | 15054 |
| 619 | SM10:0 | SM10:0 | 48147 |
| 620 | PC24:1 | PC14:1/10:0 | 25519 |
| 621 | H2SM10:0 | H2SM10:0 | 9830 |
| 622 | PC24:0 | PC10:0/14:0 | 57862 |

Table 4: Example of species annotation. This is part of a species annotation result of a P184 scan. The identified species of certain m/z values together with their intensities are assigned with the corresponding species name in detail form as well as in combined form.

### 3.3.3 Combination form annotation

In order to tackle the issue that the database only assign the detected ions to the most likely species, which may not be the real molecules, LipidomeLabelling offers the combined form of a species alongside its detailed form from the database in the final result. For example, the annotation of PC16:0/16:0 from the database for the detected species will be accompanied by another annotation of PC32:0. This combined form is offered to cover all possibilities of isobars and isomers of detected ions with the same m/z values.

## 3.4 Lipid data processing

Once species are annotated, the subsequent procedures are specific to the identified molecular species. LipidomeLabelling offers functional tools for species intensity data processing and analysis, including adduct correction, isotope effect correction, CID correction, cross scan isotope correction, composition analysis, quantification and label enrichment calculation.

### 3.4.1 Adduct effect correction

Adduct formation takes place sometimes and adducts should be considered as another presentation of the original molecule, rather than a distinguished molecule. Therefore, the intensity of an adduct should be counted as the intensity of the original molecule. For example, a sodiated DPPC as a result of sodiation during experiment, is an adduct of DPPC, and the intensity of it should be added to that of DPPC.

The platform offers an optional function for adduct correction. For each scan type, the user can define the m/z values for an ion and its adduct, then the programme will search

for the intensities of them. The adduct-ion ratio of the specified adduct over the ion will be calculated and used for calculating the adduct intensities of all other species within the scan. The intensity of an adduct will be removed from the intensity of the species that has the same m/z with the adducts. The function is as follows:

$$I(m) = I_u(m) - I\ (m - M_a) \times \frac{S_a}{S_i} \qquad m\ \in D \qquad (7)$$

, where $I(m)$ is the adduct corrected intensity of a species with the m/z value of $m$. $I_u(m)$ is the uncorrected intensity. $M_a$ denotes the user specified m/z difference between the adduct and its original ion, and $I\ (m - M_a)$ represents the corrected intensity of a species, the adduct of which has m/z of $m$. $\frac{S_a}{S_i}$ is the ratio of intensities of the specified adduct and its original ion, and $I\ (m - M_a) \times \frac{S_a}{S_i}$ is the intensity of the adduct of the species with m/z value of $m - M_a$, which is removed from the intensity of $I(m)$. $D$ is the set of all detected m/z values. This procedure only removes adduct intensity from the species that have the same m/z with the adduct. It does not add it to the intensity of original species, although it can be done by dividing the ratio after the correction has been performed.

### 3.4.2 Isotope effect correction

Isotope correction is an essential procedure in MS data processing, due to the naturally occurring isotope effect of elements, which result in multiple isotopologues of the same species. Isotopologues of a species have different masses and are identified as different molecules by MS, which should be corrected and their intensities should be added up together as the intensity of one species. Besides, an isotopologue could be mistaken for another species having the same mass, so the correction can also correct the intensity

result for this distinct species. Isotope correction is critical important for stable isotope labelling studies, which require complete removal of naturally occurred isotope of a species in order to get the accurate abundance of artificially labelled molecules for kinetic analysis.

#### 3.4.2.1 Isotope distribution calculation

To correct isotope effect of a detected molecule, the isotope distribution of the species has to be calculated first. The programme adopts a Markov chain stepwise addition dynamic programming algorithm (Snider, 2007) for the isotope distribution calculation. Instead of calculating the probability of each isotope variant, the problem is seen as a dynamic process. Every isotope variant is considered as a state within a Markov model and every atom added onto the molecule is considered as a state transition step. The programme calculates the probabilities of isotope variants every time an atom is added and replace the probabilities calculated in the previous step before the atom is added. By doing this, the probabilities of the isotope variants at each step are only related to the calculations in the previous step and the isotope probabilities of the newly added element.

The starting point of the isotope distribution algorithm is no atom status (Atom 0) with a number of nodes of all possible isotope variants of a molecule. It then starts to add one single atom of an element of the molecule at a time and calculates the probabilities of all the isotope variants. Then the next atom is added and the probabilities of the variants are calculated again by computing the transition from the previous step. This procedure keeps on until all the atoms of the element are added, and then it starts to add the atoms of another element of the molecule until all the elements are calculated. Figure 12 is a simple example of this dynamic process of adding two atoms of the only element of a

molecule to calculate its isotope variants distribution. The element has three stable isotopes.

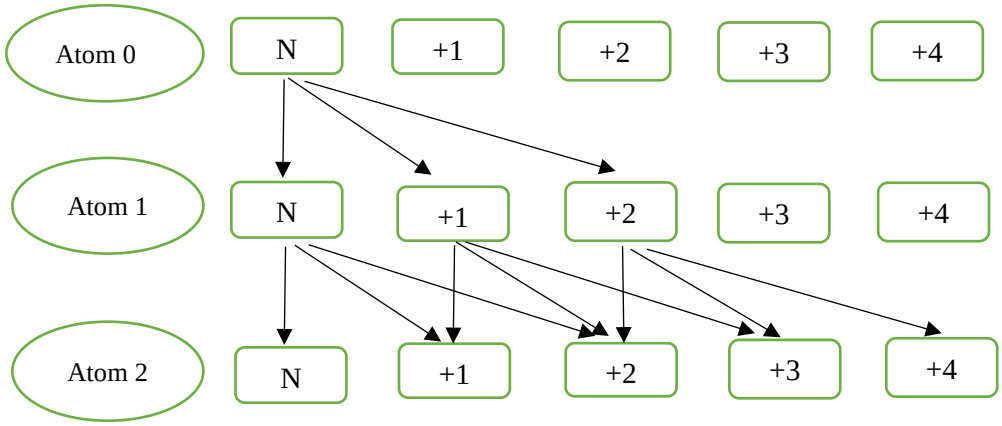



Figure 12: Schematic diagram of dynamic calculation of isotope distribution for a molecules with two atoms of the same element. The atom has three isotopes and therefore the molecule can have five isotope variants denoted as “N“, “+1”, “+2”, “+3”, “+4”, which represent the principle ion state with monoisotopic mass, and isotope states that are one, two, three and four mass higher. Three atom states include no adding atom (Atom 0), adding the first atom (Atom 1) and adding the second atom (Atom 2). The arrows represent the possible transitions from one isotope state to another after adding an atom.

A molecule that has one element with two atoms, each having three isotopes, can have five possible isotope states in total, including the no isotope effect or principle ion state “N” having monoisotopic mass, isotope state with one mass higher than the principle ion state “+1”, two mass higher state “+2”, three mass higher state “+3” and four mass higher state “+4”. At initial status “Atom 0”, no atom is added, the no isotope effect state “N” is given the probability of one. The next step is to add “Atom 1”, and as this element has three stable isotopes, it can only transit from the previous state “N” to “N”, “+1” or “+2” states with the transition probabilities equal to the natural abundance of the three stable

isotopes of atom added. Now there are three isotope states at this stage “N”, “+1” and “+2” as the starting points for the following transitions when “Atom 2” is added, where the transitions are more complicated as each of the three starting states has three possible transitions. For starting point “N”, the result state after adding the second atom could be “N”, “+1” and “+2”. If starts from “+1”, the transit isotope states could be “+1”, “+2”, and “+3”. From starting point “+2”, it can get to the isotope states of “+2”,”+3” and “+4”. The probability of an isotope variant can be obtained by summing up all the possible transitions to this state, each obtained by multiplying the probability of starting isotope state at “Atom 1” with the one step transition probabilities. For example, as shown in Figure 12, the “+1” isotope state at “Atom 2” can be from either “N” or “+1” at “Atom 1”, and its probability should be the sum of the two transition probabilities from these two starting points. The first part is the transition from “N” at “Atom 1” to the current state, which needs an atom with its isotope of one mass higher than the most abundant isotope to get “+1” state. The probability can be obtained by multiplying the probability of “N” at “Atom 1” with the transition probability, which is the natural abundance of the isotope with one mass higher than the most abundant isotope of the added atom. The second part is obtained by multiplying the probability of “+1” at “Atom 1” with the natural abundance of the most abundant isotope of the added atom. The algorithm is shown below:

$$Iso_N(i) = \sum_{j=1}^{K} Iso_{N-1}(j) * P_{ji,N} \qquad \forall\, i,j = 1,\dots,K\ ;\ N \geq 1 \tag{8}$$

, and the starting point is set as

$$\begin{cases} Iso_0(i) = 1 & i = 1 \\ Iso_0(i) = 0 & 1 \leq i \leq K \end{cases} \tag{9}$$

, where $Iso_N(i)$ denotes the isotope state $i$ when adding the *N*th atom to the molecule, and $K$ is the number of isotope variants. $P_{ji,N}$ is the probability of the one step transition from isotope state *j* to the isotope state $i$, which is related to the natural abundance of the isotope of the *N*th atom being added, and it satisfies $P_{ji,N} \geq 0$ and $\sum_{j=1}^{k} P_{ji,N} = 1$. Formula (8) describes that the probability of a molecule at a certain isotope state $i$ after adding the *N*th new atom on it, equals to the sum of the probabilities of all possible isotope state transitions from the previous status. Formula (9) gives the probabilities of the initial isotope states at "Atom 0", and the probability of the no isotope effect is given the value of 1 and other isotope states 0.

| **m/z of Isotope variants** | **Probabilities** |
|---|---|
| 734 | 0.672590888 |
| 735 | 0.262635973 |
| 736 | 0.055411745 |
| 737 | 0.008296727 |
| 738 | 0.000976817 |
| 739 | 0.000095285 |
| 740 | 0.000007945 |
| 741 | 0.000000578 |
| 742 | 0.000000037 |
| 743 | 0.000000002 |

Table 5: Probabilities of 10 isotope variants of PC16:0/16:0 ion. The ion has a probability of 0.672590888 to be in the monoisotopic form with the mass of 734, and a probability of 0.262635973 to be in the state with one mass higher of 735.

The isotope distribution function is set to calculate the probabilities of the first 10 isotopic variants of a species, including the molecules of monoisotopic mass with only the principle isotopes, as well as the ones with other isotopes up to 10 masses higher than

the monoisotopic mass. This is reasonable for a typical PL analysis, as for a molecule with a mass over 1700, the probability of it having more than 10 isotopologues could be well under $10^{-6}$, which is negligible. The PL molecules mass are usually under 1700 and the possibility of these specie having more than 10 isotopologues is even less. Table 5 is an example of the calculated isotope distribution of PC16:0/16:0 ion using the programme.

#### 3.4.2.2 Isotope effect correction

To correct the isotope effect shown in the MS analysis, LipidomeLabelling utilizes a formula (Eibl *et al.*, 2008) commonly used in isotope effect removal, shown as below

$$I(m) = (I_u(m) - \sum_{n=1}^{9} I(m-n) \times Iso_n(m-n))/Iso_0(m) \quad m \in D \quad (10)$$

, where $I(m)$ is the corrected intensity of a species with m/z value $m$, and $I_u(m)$ is the observed intensity of it. $I(m-n)$ is the corrected intensity of species with m/z value of $m-n$, and so $Iso_n(m-n)$ represents the natural abundance of "*+n*" isotopologues of this species. The product of these two items $I(m-n) \times Iso_n(m-n)$ gives the intensity of the "*+n*" isotoplogues of the species. The programme corrects the isotope effect on a species from 9 other species, with m/z values from *m-1* to *m-9*, the isotoplogues of which may have the same mass with *m*. $I_u(m) - \sum_{n=1}^{9} I(m-n) \times Iso_n(m-n)$ is the remaining of the intensity after removing the isotope effect. $Iso_0(m)$ denotes the percentage abundance of principle ion of species *m*. Dividing the remaining intensity by this term generates the real abundance of species *m*. Figure 13 shows spectrum of the previous sample shown in Figure 9 after the isotope effect correction.

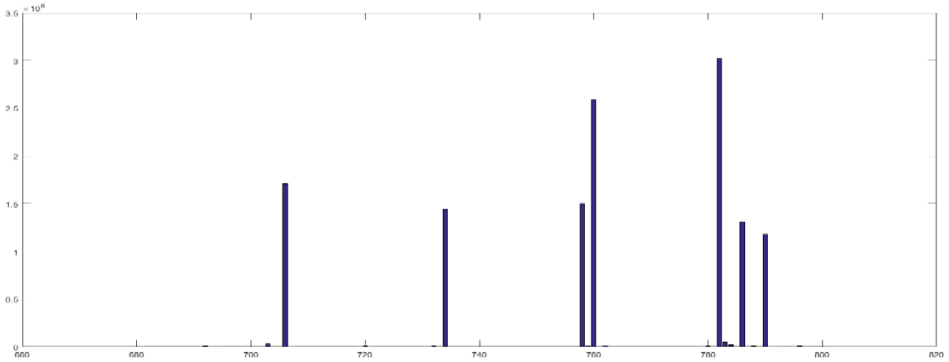


Figure 13: Spectrum of the previous example in Figure 9 after isotope correction.

#### 3.4.2.3 Cross scan isotope effect correction

The cross scan isotope effect correction step is a unique feature of the LipidomeLabelling and the in house programme, which aims to correct the isotope effect occurring on the fragment that is monitored in a precursor scan. In a precursor scan, the isotope effect of a detected species which occurs on the non-monitored part of the molecule can be corrected by calculating and removing the effect from other species detected in the same scan using the algorithm discussed in the previous section. However, the species may also suffer from the isotope effect from other different species that have the same m/z with the detected species, but isotope effect occurs on their fragment making it has the same mass with the fragment being monitored. For example, the monoisotopic mass of ion A is 100 and its fragment of 50 is monitored. The monoisotopic mass of ion B is 99, but isotope effect occurs on its fragment with m/z of 49, making the fragment of 50 and the m/z of the isotopologue of B 100. This variant of ion B will be detected by the MS and taken as ion A, as it has the same m/z for the monitored fragment and the whole ion of A. However, this is a totally different species, which should be removed if possible. Correcting this requires the intensity of the source species, the isotope effect of which affects the result of the species of interest, but the intensity is not derivable from the

scan itself. Nevertheless, multiple scans are often carried out for an ESI MS/MS facilitated PL labelling analysis, which makes it possible to partly correct this isotope effect on a species by working out the isotope effects from other species detected from other scans. This is particularly important in labelling studies that require removing the naturally occurred isotope effect in order to get an accurate abundance of artificially labelled species for enrichment calculation.

For example, in a PC labelling study, DPPC (m/z 734) and *methyl*$D_3$ labelled DPPC (m/z 737) can be detected by P184 and P187 scans, respectively. However, the detected D3DPPC molecules by P187 is the combination of artificially labelled D3DPPC and the naturally occurred "+3" isotopologues of DPPC. With the information obtained from P184 scan, this problem can be solved by calculating the abundance of "+3" isotopologues of DPPC detected by P184 scan, and removing it from the D3DPPC measurement by P187. The correction formula is as follows

$$I_f(m) = I_{uf}(m) - \sum_{f-7 \le n < f} I_n\big(m - (f-n)\big) \times Iso_{f-n}(n) \qquad (11)$$

, where $I_f(m)$ denotes the corrected intensity of species with m/z of $m$, the monitored fragment of which has a mass of $f$. $I_{uf}(m)$ is the uncorrected intensity of the species. $n$ is the mass of the fragment monitored by another scan, which satisfies $f - 7 \le n < f$, meaning the programme only corrects the effect of the species, the fragments of which are smaller than that of the species of interest by maximum of 7 mass units. $I_n\big(m - (f-n)\big)$ is the corrected intensity of a species in the scan of fragment $n$, and the mass difference between this species and species *m* is the same as that between the two scanned fragments $f - n$. Taking the example above, if we are correcting an isotope effect on species D3DPPC with m/z of 737 in P187 scan from a isotope variant of a species

detected in P184, making "*m*" being 737 and "*f*" being 187, we need to consider the "+3" (187-184) isotopologues of a species from P184 having a mass of 737 – (187-184), which is 734. The intensity of the species from scan *n* is used to calculate the isotopologues that affect the species *m* in scan *f*, by multiplying their percentage abundance of the $Iso_{f-n}(n)$ . All the scans in the same mode will be treated in the same way as long as their mass differences between the scanning fragments are smaller than 7.

### 3.4.3 CID correction

A tandem mass spectrometer with a collision cell can perform collision - induced dissociation (CID), through which molecular ions are fragmented. However, the bigger a molecule is, the more vibration modes it has, so it is less likely that a given collision energy can induce fragmentation. Besides, the more energy put in for dissociating the big ion, the greater the speed of the fragments, and the more difficult it is for the containment field to keep them in, which leads to scatter loss. The effect of above is that the detection rates of bigger precursor molecules tend to be lower than smaller precursor molecules. LipidomeLabelling provides an optional CID correction function to correct the effect occurred due to the first issue mentioned above, with an empirical formula as follows

$$I(m) = I_u(m) \div (2 \times 10^{13} \times m^{-3.9873}) \times 100 \qquad (12)$$

, where $I(m)$ and $I_u(m)$ are the corrected and uncorrected intensities of species with m/z value of $m$.

## 3.5 Phospholipids data analysis

### 3.5.1 Composition analysis

LipidomeLabelling offers a functionality for PL composition analysis. It calculates the percentage of all the species detected in a scan. This calculation excludes the internal standards, which are exogenous molecules artificially added into samples for species quantification. The internal standards are specified by users in the setting up for analysis by LipidomeLabelling.

| Ion Mass | Species | Intensities | Quantification (nmol) |
|---|---|---|---|
| 678 | PC14:0/14:0 | 161446616.5 | 10 |
| 780 | PC16:0/20:5 | 882214.6 | 0.0546 |
| 781 | SM22:3 | 0 | 0 |
| 782 | PC16:0/20:4 | 302661855 | 18.7469 |
| 783 | SM22:2 | 5145607.8 | 0 |
| 784 | PC18:1/18:2 | 1769086.9 | 0.1096 |
| 785 | SM22:1 | 59115.9 | 0 |
| 786 | PC18:0/18:2 | 131416916 | 8.13997 |
| 787 | SM22:0 | 341345.5 | 0 |
| 788 | PC18:0/18:1 | 1004836.6 | 0.06224 |
| 789 | SM24:6a | 0 | 0 |
| 790 | PC18:0/18:0 | 118075112 | 7.3134 |
| 792 | PC18:2a/20:4 | 0 | 0 |
| 794 | PC18:1a/20:4 | 56540 | 0.0035 |
| 796 | PC18:0a/20:4 | 759896 | 0.0471 |
| 797 | SM24:2a | 60661 | 0 |
| 798 | PC18:0a/20:3 | 503165 | 0.03117 |
| 799 | SM24:1a | 109304 | 0 |
| 800 | PC18:0a/20:2 | 139302 | 0.008628 |

Table 6: Part of the result from a P184 scan of a sample. 10nmol of internal standard PC14:0/14:0 (first row) are used specifically for PC species quantification, so only the amounts of PC species are calculated.

### 3.5.2 Amount quantification

Quantifying the amount of species in a sample is one of the main purposes of MS analysis. For lipid quantification, internal standards are added into samples before MS monitoring. Internal standards are some exogenous compounds that have similar properties with the endogenous species to be quantified when analysed by MS. It is possible to determine the quantity of endogenous lipid by multiplying the amount of internal standard with the intensity ratio of endogenous species over the internal standard. The LipidomeLabelling supports three internal standards for multiple lipid class quantification. Table 6 is part of the lipid quantification result of an ESI MS/MS analysis calculated by the programme.

### 3.5.3 Dynamic stable isotope labelling analysis

The isotopic enrichment of a stable isotope labelled material is calculated based on the intensities of the detected labelled and unlabelled molecules. In the literature of tracer kinetics study, there are two commonly used forms of enrichment, tracer/tracee ratio (Cobelli *et al.*, 1987) and APE. APE is the ratio of tracer over the sum of tracer and tracee (Garlick *et al.*, 1994). There are no clear explanations about which form should be used in what situation. In the surfactant phospholipid stable isotope labelling studies, APE has been widely used (Cogo *et al.*, 1999; Torresin *et al.*, 2000). The APE is described by the following formula

$$APE = \frac{Tracer}{Tracer+Tracee} \qquad (13)$$

LipidomeLabelling supports both forms of enrichment calculation to be chosen by users for lipid labelling analysis. The platform support studies with multiple labelling materials. It can also analyse labelling and non-labelling PL studies at the same time. Table 7 lists the

enrichment of some labelled PC species in a mouse sample measured by ESI MS/MS through P184, P187, P190 and P193 scans.

| Endogenous | m/z | Enri | D3 labelled | m/z | Enri | D6 labelled | m/z | Enri | D9 labelled | m/z | Enri |
|---|---|---|---|---|---|---|---|---|---|---|---|
| PC16:0/14:0 | 706 | 0.9973 | D3PC16:0/14:0 | 709 | 0.001 | D6PC16:0/14:0 | 712 | 0.0001 | D9PC16:0/14:0 | 715 | 0.0016 |
| PC22:6a/10:0 | 708 | 0.9671 | D3PC22:6a/10:0 | 711 | 0.0096 | D6PC22:6a/10:0 | 714 | 0.0128 | D9PC22:6a/10:0 | 717 | 0.0105 |
| PC22:5a/10:0 | 710 | 0.9581 | D3PC22:5a/10:0 | 713 | 0.0198 | D6PC22:5a/10:0 | 716 | 0.0098 | D9PC22:5a/10:0 | 719 | 0.0123 |
| PC14:1a/18:3 | 712 | 0.9604 | D3PC14:1a/18:3 | 715 | 0.0109 | D6PC14:1a/18:3 | 718 | 0.0174 | D9PC14:1a/18:3 | 721 | 0.0113 |
| PC14:1a/18:2 | 714 | 0.9536 | D3PC14:1a/18:2 | 717 | 0.0064 | D6PC14:1a/18:2 | 720 | 0.0354 | D9PC14:1a/18:2 | 723 | 0.0046 |
| PC14:0a/18:2 | 716 | 0.9824 | D3PC14:0a/18:2 | 719 | 0.0082 | D6PC14:0a/18:2 | 722 | 0.0031 | D9PC14:0a/18:2 | 725 | 0.0063 |
| PC18:0a/14:1 | 718 | 0.9935 | D3PC18:0a/14:1 | 721 | 0.002 | D6PC18:0a/14:1 | 724 | 0.0003 | D9PC18:0a/14:1 | 727 | 0.0042 |
| PC16:0a/16:0 | 720 | 0.9966 | D3PC16:0a/16:0 | 723 | 0.0014 | D6PC16:0a/16:0 | 726 | 0.0001 | D9PC16:0a/16:0 | 729 | 0.0019 |
| PC22:6/10:0 | 722 | 0.9816 | D3PC22:6/10:0 | 725 | 0.0094 | D6PC22:6/10:0 | 728 | 0.0032 | D9PC22:6/10:0 | 731 | 0.0059 |
| PC22:5/10:0 | 724 | 0.9242 | D3PC22:5/10:0 | 727 | 0.0558 | D6PC22:5/10:0 | 730 | 0.0152 | D9PC22:5/10:0 | 733 | 0.0048 |
| PC14:1/18:3 | 726 | 0.9762 | D3PC14:1/18:3 | 729 | 0.0159 | D6PC14:1/18:3 | 732 | 0.0069 | D9PC14:1/18:3 | 735 | 0.001 |
| PC14:1/18:2 | 728 | 0.9732 | D3PC14:1/18:2 | 731 | 0.0059 | D6PC14:1/18:2 | 734 | 0.0018 | D9PC14:1/18:2 | 737 | 0.0191 |
| PC14:0/18:2 | 730 | 0.9945 | D3PC14:0/18:2 | 733 | 0.0019 | D6PC14:0/18:2 | 736 | 0.0003 | D9PC14:0/18:2 | 739 | 0.0034 |
| PC16:0/16:1 | 732 | 0.9961 | D3PC16:0/16:1 | 735 | 0.0008 | D6PC16:0/16:1 | 738 | 0.0000 | D9PC16:0/16:1 | 741 | 0.0031 |
| PC16:0/16:0 | 734 | 0.9958 | D3PC16:0/16:0 | 737 | 0.001 | D6PC16:0/16:0 | 740 | 0.0000 | D9PC16:0/16:0 | 743 | 0.0032 |

Table 7: Enrichment of PC species in a mouse sample analysed by P184, P187, P190 and P193 scans measuring the endogenous, *methyl*$D_3$ labelled, *methyl*$D_6$ labelled and *methyl*$D_9$ labelled PC species, respectively. The isotopologues of a species detected in the four scans are put in the same row. *m/z* values are listed behind species names. APE is applied for enrichment calculation, so that the enrichment of a species in either labelled or unlabelled form is the ratio of its intensity over the sum of the intensities of all isotopologues detected. Therefore, the enrichment shown on each row add up to 1.

## 3.6 Graphical user interface

LipidomeLabelling offers a graphical user interface (GUI) (Figure 14) for users of all levels with no programming skills requirement. The GUI is stored in GUI.mat file. To run a labelling analysis, users need to make a few steps of setting up, including file selection, species list selection, baseline noise removal specification, internal standards specification, and criteria setting for species selection. The programme allows modifying the specification of internal standards and selection criteria after analysis has been done and only adjust the previous results, which does not require the recalculation of all the results. This feature makes it easier for users to correct input errors made by mistake

when setting up the programme, which saves the time of recalculation of the unaffected parts.

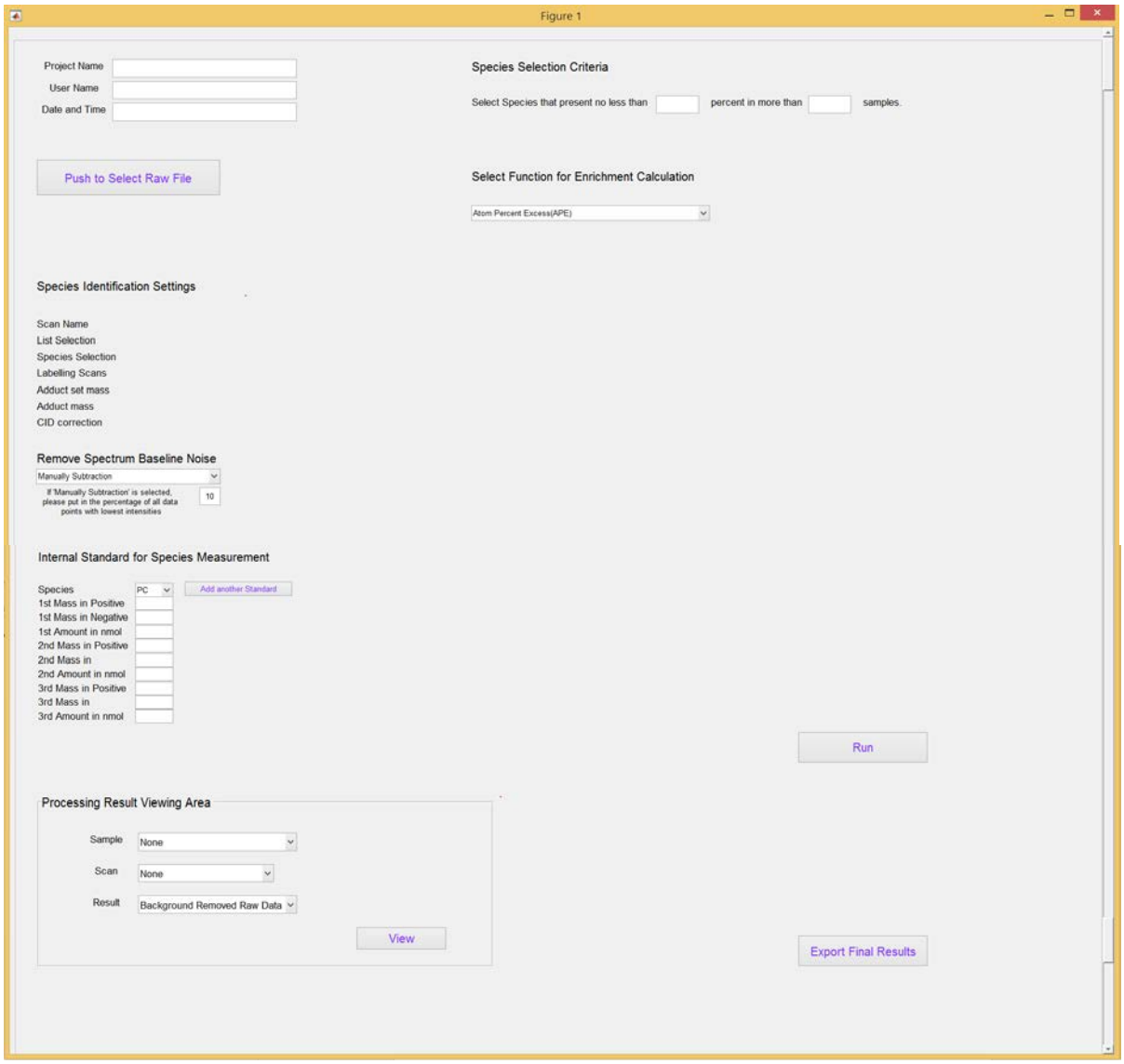



Figure 14: Graphical user interface of LipidomeLabelling.

## 3.7 Validation

Validation of LipidomeLabelling is performed by comparing it with the old in house software on the results of the common functionalities they have. The comparison has

found small differences in the composition and enrichment, which mainly stems from the different background subtraction algorithms applied, resulting in the difference in the calculated molecule intensities. A comparison of molecule intensities of a mouse sample calculated by the two programmes are listed in Table 8.

| Species | LPC10:1 | LPC10:0 | LPC12:1 | PC16:0/16:1 | PC16:0/18:2 | PC16:0/18:1 | PC18:1/20:4 | PC18:0/20:4 |
|---|---|---|---|---|---|---|---|---|
| m/z | 410 | 412 | 438 | 732 | 758 | 760 | 808 | 810 |
| Old analyser | 41855 | 132945 | 76469 | 425395368 | 7467709331 | 13079711579 | 1383499560 | 3361057215 |
| Lipidomelabelling | 44344 | 138086 | 81315 | 425399571 | 7467714456 | 13079715941 | 1383506189 | 3361064179 |

Table 8: Comparisons of intensities of selected species processed by the old analyser and Lipidomelabelling. The intensities calculated for species from different mass ranges are generally higher in the Lipidomelabelling, which approximates a smaller background than the old analyser. The first three species within small mass range are affected more due to their relatively small intensities compared to the background noise subtracted from all species.

The old programme is more aggressive on the background removal, which generates slightly higher background estimation. This mainly affects molecules with lower abundance, because the procedure removes a fixed background noise from all spectrum and so the smaller the intensity is, the more effect the removed background can make. Although the mostly affected species are usually not of interest, as these species are with an intensity level close to noises, it is still worth to evaluate further for a more appropriate background algorithm for a better approximation of the molecule intensities.

It was also shown that LipidomeLabelling is much faster than the in house software. In a test of analysing a typical PC labelling analysis with 4 scan types, the new platform runs 6 times faster than the old programme.

## 3.8 Conclusions

LipidomeLabelling is a platform for ESI MS/MS lipidomics data processing and kinetic analysis. It consists of a data converting programme and streamlined analysing modules, offering functionalities for background noise removal, peak detection and integration, adduct correction, isotope effect correction, CID correction, as well as composition, quantification and kinetic analysis. It facilitates fast, comprehensive, automatic and user friendly analysis for stable isotope labelling as well as non-labelling PL studies.

The algorithm used in LipidomeLabelling for peak integration takes the average of the intensities measured at several points around a detected peak centre, due to difficulty in modelling the peak based on the information available. Better algorithms should be explored for integrating the whole peak area for better abundance approximation. Besides, background subtraction algorithm of LipidomeLabelling could be tested and examined in different studies for better parameter setting and background estimation.

# Chapter 4 Surfactant phosphatidylcholine dynamic biomarker discovery analysis for animal models

## 4.1 Introduction

In this chapter, the pulmonary surfactant PC labelling data of mouse models from a previous surfactant turnover study is used for dynamic biomarker discovery analysis. The methodology for time course PC label enrichment modelling and biomarker analysis is established for the following application to human clinical data. The mouse model is ideal for investigating the possibility of using temporal PC label incorporation as a biomarker, because the transgenic mouse in the study is an artificially established model with single defined condition. It benefits from having a genetically homogeneous background and common nutrition, conditions that do not apply in clinical studies. The SME statistical method is used to characterise the time course enrichment of *methyl*$D_9$-PC, an indication of PC turnover, to assess if it can potentially provide improved diagnostic insights over static measurements.

### 4.1.1 Project description

The mice data analysed in this chapter is from a previous study carried out in Southampton funded by Medimmune LLC. Medimmune has developed a therapeutic antibody against the GM-CSF receptor, which they are using a biological drug to inhibit macrophage-driven inflammation and so reduce painful symptoms in patients with rheumatoid arthritis. As reduced GM-CSF stimulation of alveolar macrophages development is a major cause of PAP disease, Medimmune was concerned that the antibody for treatment of rheumatoid arthritis might have a harmful side effect of accumulation of excess surfactant in the lungs of their patients. Therefore, Medimmune

carried out a two-step study aiming to evaluate how their drug would affect the turnover of the PC species in the respiratory system. The first part of the study sought to compare the PC turnover of mice with and without PAP. They use GM-CSF receptor beta chain depleted mice, also called KO mice in this chapter, as diseased mice to establish the effect of blocking alveolar catabolism of surfactant PL due to the absence of macrophages. Then they compared the turnover of PC species in BALF and lung tissue samples of these gene depleted mice models with that of normal mice. There are three age groups of normal mice including 169 days, 113 days and 84 days groups. The second part of the study deployed another two groups of normal mice treated with low dose and high dose of the developed antibody drug to see if excessive antibody caused any harm, where the low dose is equivalent to the amounts given to rheumatic arthritis patients. Low dose and high dose of drug were given to both 113 days and 169 days mice groups, to assess the dose effect across two different age groups. The whole study found that the total PC turnover of the KO mice was significantly lower than that of the normal mice at the designed time points. Besides, PC turnover of low dose drug treated group was not different from the normal mice. However, although the high dose treatment did not have an effect on physiological functioning of the mice, it did have an impact on the PC label incorporation pattern, which is more significant in 169 days high dose treated mice group.

In order to set up and validate the methodologies for time course PC label incorporation modelling as well as biomarker discovery analysis, data of the mice groups that were found with significant differences in the study were chosen to be analysed in this chapter. Thus, the three age groups of normal mice, KO mice and 169 days high dose treated mice were analysed to investigate if there is any difference between groups that can be

recognized by temporal enrichment of labelled PC and which species are the most significant in marking the difference.

### 4.1.2 Data description and method

The mice surfactant PC label incorporation data analysed in this chapter was provided in the form of calculated *methyl*$D_9$-PC enrichment of sixteen PC species. The enrichment data of BALF and lung tissue samples of three age groups of normal mice, KO group mice and 169 days high dose treated mice, observed at four time points are included in the analysis. The normal mice are considered as control mice, therefore the control group has subsequent 169, 113 and 84 days age groups.

All of the mice received intraperitoneal injections of 0.1mg *methyl*D9-choline chloride for three hours. The mice were sacrificed 5, 10, 24 and 48 hours after label infusion, and BALF and lung tissue samples were collected immediately. Time series data for five mice are kept in each group, although the samples were actually from 20 mice. As mice could not be resampled, the samples collected at the 4 time points were actually from 4 different mice, rather than a single mouse resampled four times.

As the time series samples are not from the same mice, a random combination of observations from 20 mice at the four time points was carried out to be considered as repeatedly sampled time series data for five mice subjects. This is necessary as the nature of the longitudinal study is to feature the time course characteristics of PC species, and in biological experiments it is common that resampling is impossible. The made-up mice subjects are denoted as Mouse 1 to Mouse 5 in 169 days group, Mouse 5 to Mouse 10 in 113 days group, and Mouse 11 to Mouse 15 in 84 days group for control group, and KO1 to KO5 for KO group. Due to unknown error, the labelled material in some samples are at

an undetectable level, including Mouse 1 of 169 days group and KO5 of KO group at 5 hour. PC enrichment data of these samples were excluded from the analysis.

SME method was applied for model fitting for time course enrichment of sixteen *methyl*$D_9$-PC species. *Ft* statistic was used for comparisons between groups and the SME R package was used for the analysis.

### 4.1.3 Chapter objects and structure

This chapter aims to use SME method to model time course pattern of *methyl*D9 labelled surfactant PC enrichment of mice samples and compare between groups for biomarker discovery analysis, as well as to validate the methodologies for the analysis of clinical data in the following chapters. The objects are as follows.

- Compare the D9DPPC temporal enrichment between control age groups, to explore the appropriate application of SME framework to the data for best performance, as well as to assess if age factor has any effects on the PC turnover. The variation detected between the close assembled sub control groups can be used as a reference for comparisons between other more diverse groups.
- Compare label incorporation patterns of sixteen PC species between control and KO mice to evaluate the difference between the theoretically contrast groups for biomarker analysis. The result could be used as reference of difference of contrast groups, to be compared with that of the control age group, both of which could be referred to when comparing other groups.
- Compare time course PC label incorporation between control and high dose antibody treated groups, which should have mild difference comparing to that

between control and KO mice, to check how well the established methodology can identify and scale differences between groups.

- Analyse PC label incorporation in both BALF and lung tissue samples, for a comprehensive longitudinal investigation of surfactant PC metabolism across different compartments of the respiratory system.

## 4.2 Characterisation of time course *methyl*$D_9$ labelled DPPC enrichment in BALF and group comparisons

In this section, the time course modelling and group comparison analysis were first applied to a single DPPC species in BALF samples in mice groups, which aims to establish the methodology as well as to demonstrate the use of it, before applying to multiple PC species for biomarker discovery. SME method was used to model time course enrichment of *methyl*$D_9$ labelled DPPC species (D9DPPC or D9PC16:0/16:0) for three age groups of control mice, combined control mice as well as KO group mice. The comparisons of the D9DPPC temporal enrichment were then carried out between three control age groups, between age groups and KO group, as well as between combined control group and KO group. Comparisons between control groups would indicate how much the comparatively resembled age groups of control mice are different from each other in DPPC turnover. The comparisons between control groups with KO group are to assess how much the mice in the two biological states are different from each other in DPPC turnover. Furthermore, the necessity of comparing combined control group with KO group instead of individual sub age groups for the following analysis was discussed.

### 4.2.1 Analysis of time course D9DPPC enrichment and group comparison

#### 4.2.1.1 Dynamic modelling for control mice

DPPC is the most abundant PC species in pulmonary surfactant. The 169 days, 113 days and 84 days age groups of control mice are all normal mice with genetically homogeneous background. DPPC temporal label incorporation of the three age groups were firstly modelled individually as independent groups to examine the age effect on the species turnover, and then the DPPC label incorporation was modelled for the combined age groups as a whole control group.

##### 4.2.1.1.1 Modelling for 169 days control mice

Based on the observations at different time points, the time course D9DPPC enrichment was modelled for each of the 5 individual mouse of the 169 days age group, as well as the mean enrichment of the whole group, which is taken as the underlying biological process of the whole group. Table 9 gives the estimated D9DPPC enrichment values at the 4 observation time points for individuals and the whole group. It also lists out the variation of fitted individuals values from the fitted group mean, as well as the original observations.

The fitted group mean model gives an estimated D9DPPC enrichment value of 0.555% at 5 hour time point, which goes up to 1.158% at 10 hour. It then gets to a much higher value of 1.699% at 24 hour, and decreases to 1.097% at 48 hour. The temporal trends for all the individuals have the similar pattern. The variation of individual fitted values from group mean values are generally small, ranges from -0.020% to 0.039% at 5 hour, -0.057% to 0.109% at 10 hour, -0.046% to 0.084% at 24 hour, and -0.26% to 0.136% at 48 hour. The fitted values among the individuals are very close suggesting similar time course characteristics of the mice in 169 age group.

| Subjects | | 5 hour | 10 hour | 24 hour | 48 hour |
|---|---|---|---|---|---|
| Fitted mean enrichment of the group | | 0.555 | 1.158 | 1.699 | 1.096 |
| Mouse 1 | Original enrichment | | 1.084 | 1.613 | 1.229 |
| | Fitted enrichment | 0.535 | 1.100 | 1.654 | 1.233 |
| | Variation from group mean | -0.020 | -0.057 | -0.046 | 0.136 |
| Mouse 2 | Original enrichment | 0.538 | 1.258 | 1.747 | 1.060 |
| | Fitted enrichment | 0.564 | 1.186 | 1.722 | 1.033 |
| | Variation from group mean | 0.009 | 0.028 | 0.023 | -0.063 |
| Mouse 3 | Original enrichment | 0.493 | 1.222 | 1.726 | 1.089 |
| | Fitted enrichment | 0.557 | 1.1685 | 1.709 | 1.074 |
| | Variation from group mean | 0.002 | 0.011 | 0.010 | -0.022 |
| Mouse 4 | Original enrichment | 0.592 | 1.196 | 1.763 | 0.944 |
| | Fitted enrichment | 0.576 | 1.214 | 1.744 | 0.959 |
| | Variation from group mean | 0.021 | 0.057 | 0.044 | -0.138 |
| Mouse 5 | Original enrichment | 0.550 | 1.325 | 1.737 | 0.806 |
| | Fitted enrichment | 0.594 | 1.266 | 1.783 | 0.834 |
| | Variation from group mean | 0.039 | 0.109 | 0.084 | -0.263 |

Table 9: Fitted enrichment values (%) of D9DPPC in BALF samples of 169 days control mice at four observation time points. In this table, the fitted group mean enrichment values are listed in the first row. For each individual mouse, the original enrichment observation value, model fitted value as well as the variation of the fitted value from the group mean are all listed at each observation time point. Rows highlighted in blue are fitted values through modelling.

A variation value as listed in the table is the difference between a fitted individual value and the fitted group mean value at a certain time point, rather than that between the fitted value and the original observation of an individual. For example, the variation of -0.02% of Mouse 3 at 5 hour is the difference between the group mean value of 0.555% and the fitted value 0.557%. But the difference between the original observational value of 0.493% and the fitted value is -0.56%, which is bigger than the variation term. The fitted values of individuals reflect the underlying biological feature of each individual subjects, whereas the originally measured values are real observations, so the difference between them reflects random variations of an individual model, while the variation of fitted individual values from group means could be considered as variations within group.

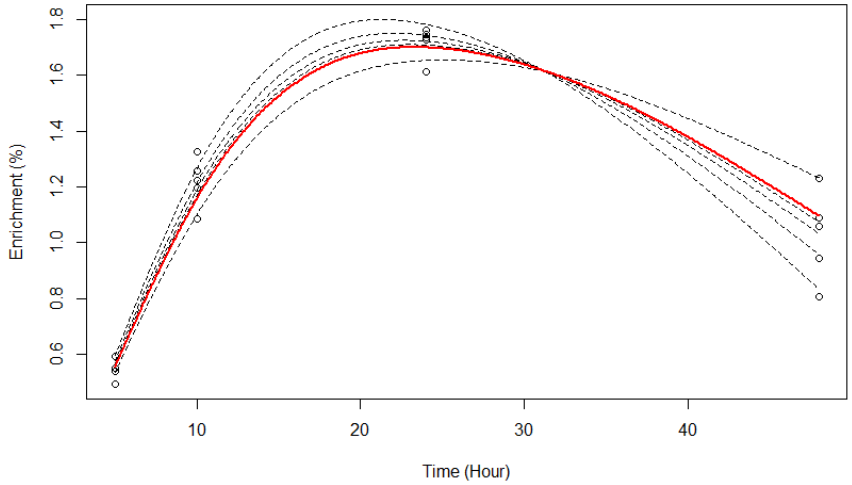



Figure 15: Time course enrichment model of D9DPPC in BALF sample of 169 day age group control mice. The empty circles are originally observed data points. The grey dashed curves are enrichment models of individual mice, and the red solid curve is that of the whole group.

Figure 15 shows the fitted curves of temporal enrichment of D9DPPC for each individual mouse and the group mean of 169 days mice. The enrichment of D9DPPC increases with a big gradient at the beginning and gets to its highest level of 1.65% at 22 hour post labelling. After reaching the highest point, the enrichment goes down slowly. The maximum value as well as the time trend are approximated through modelling, which cannot be obtained directly from the observed values at separated time points or by simply lining them up to simulate the process.

The curves reflect the underlying biological activities of PC species by characterising the change of their label incorporation over time. It indicates that D9DPPC species come into the surfactant system and accumulate, and in the meanwhile they leave the pool at a certain rate. In the steady state, when the amount of the entering D9DPPC is higher than that of the leaving DPPC, it accumulates in the pool, and the enrichment curve shows an increasing trend. If the two equal to each other, the curve reaches a plateau. However,

when the amount of incoming labelled species is less than that of the outgoing, the enrichment values get smaller, therefore the curve drops down.

It can be seen from the plot that the group mean curve is not obtained by calculating the mean values at each time point and joining them up. In fact, all the fitted values are slightly deviated from the mean values calculated from observations at single time points. For example, the fitted enrichment value of 1.096% at 48 hour on the group mean curve is higher than the mean value 1.06% calculated directly from the individual mice at this time point. This is because the SME method assumes a smoothing biological process of the variables of interest, where the time effect is taken into account. As a result, critical points can be found at unobserved time points, such as maximal and minimal points. Simply taking means of observations at individual time points and lining them up would lose the time information and leave discontinuity of variables at time knots. Moreover, it is not able to estimate critical points other than observation time points.

Surfactant PC are synthesised by AT II cells and are secreted into alveolar space, where BALF samples are taken from, through exocytosis. If control mice are assumed in a steady state, the swinging of the enrichment curves should be associated with the relationship between the enrichment of labelled PC species before entering the pool, and the current enrichment of labelled PC species in the pool. This is because in the steady state, the pool size is constant, and the amount of entering species matches the leaving, which means the rate and amount of new PC species coming into the pool should equal to those of PC species leaving the pool. When it comes to the labelled species, under the constant secretion, degradation and exchange rate, the rate of newly appeared labelled PC species is related to its enrichment before coming into the pool, whereas the rate of disappearance of the labelled species is related to their enrichment of the current pool.

Therefore, as the enrichment curve reflects the relationship between the entering and leaving labelled product PC species, it is actually down to the relation of the enrichment of labelled precursor species and the enrichment of labelled species in the current pool. Therefore, the increasing part in the curve means the label enrichment before entering the pool is higher than the current pool enrichment, and the decreasing part of curves indicates the label enrichment of species ready to get into the pool has already gone lower than the current pool enrichment. As the label enrichment in the current pool is mainly determined by the enrichment in the precursor pool steady state, the swing of an enrichment curve is primarily due to the change of precursor pool label enrichment.

#### 4.2.1.1.2 Modelling for 113 days age group mice

The fitted model of D9DPPC enrichment of 113 days group is very close to that of 169 days group. The approximated group mean values are 0.505%, 1.212%, 1.707% and 1.08% at the four sample taking time points (Table 10). The enrichment curves increase from the beginning and go down after reaching the maximum point (Figure 16). The maximum enrichment is estimated to be 1.65% reached at about 21 hour after label infusion, which is similar to that of 169 days group. The fitted individual curves and the group mean curve are very close to each other, with a few observation points dispersed away from the fitted curves. This indicates the small differences between fitted individual values and the group means, but comparatively bigger differences between fitted individual values and their observation, which implies a high group consistency and some variation among the individual. The observed values deviating from modelled curves also show that the data has not been over fitted by the model, in which case the fitted curves would go through every single points on the plot.

| Subjects | | 5 hour | 10 hour | 24 hour | 48 hour |
|---|---|---|---|---|---|
| Fitted mean enrichment of the group | | 0.506 | 1.212 | 1.707 | 1.080 |
| Mouse 6 | Original enrichment | 0.416 | 1.1348 | 1.540 | 1.065 |
| | Fitted enrichment | 0.470 | 1.173 | 1.669 | 1.080 |
| | Variation from group mean | -0.036 | -0.039 | -0.038 | -0.001 |
| Mouse 7 | Original enrichment | 0.434 | 1.431 | 1.710 | 0.973 |
| | Fitted enrichment | 0.529 | 1.234 | 1.722 | 1.068 |
| | Variation from group mean | 0.023 | 0.022 | 0.0146 | -0.0124 |
| Mouse 8 | Original enrichment | 0.508 | 1.411 | 1.756 | 1.071 |
| | Fitted enrichment | 0.536 | 1.244 | 1.734 | 1.076 |
| | Variation from group mean | 0.031 | 0.032 | 0.0274 | -0.004 |
| Mouse 9 | Original enrichment | 0.392 | 1.213 | 1.829 | 1.183 |
| | Fitted enrichment | 0.497 | 1.208 | 1.715 | 1.098 |
| | Variation from group mean | -0.009 | -0.004 | 0.008 | 0.018 |
| Mouse 10 | Original enrichment | 0.542 | 1.192 | 1.636 | 1.1034 |
| | Fitted enrichment | 0.501 | 1.206 | 1.700 | 1.080 |
| | Variation from group mean | -0.005 | -0.006 | -0.007 | -0.001 |

Table 10: Fitted enrichment values (%) of D9DPPC in BALF sample of 113 day age group control mice at four observing time points.

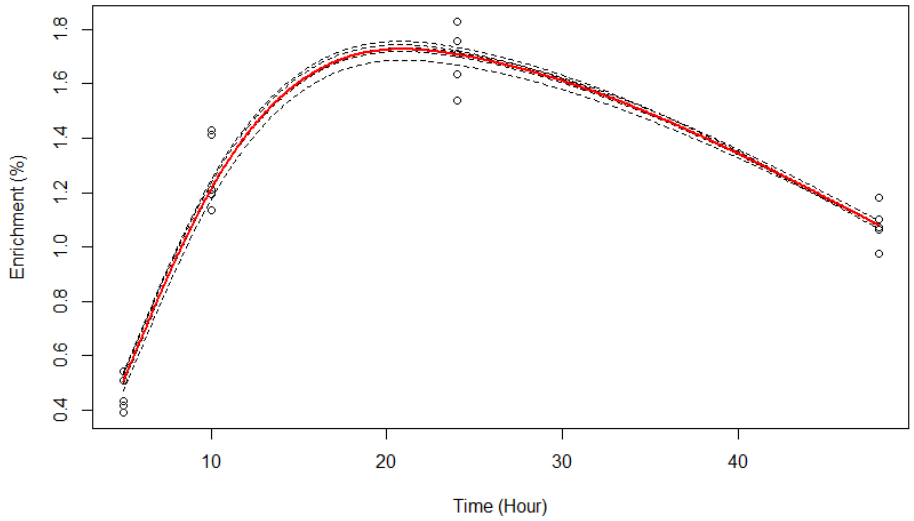



Figure 16: Time course enrichment model of D9DPPC in BALF sample of 113 day age group control mice. The empty circles are originally observed data points. The grey dashed curves are enrichment models of individual mice, and the red solid curve is that of the whole group.

#### 4.2.1.1.3 Modelling for 84 days age group mice

PC label enrichment data of 84 days control mice group is more spread out than 169 and 113 days groups, so the value range at each time point is wider (Table 11, Figure 17). The enrichment of D9DPPC is between 0.986% and 1.902% at 24 hour, much wider than that for 113 days group which is between 1.636% and 1.829%. The fitted curves have similar temporal trend with those in 169 and 113 groups, which increase from the beginning to the highest point and then decay towards 48 hour. However, the approximated values of individuals and group mean are slightly different from 169 and 113 age groups, with values at 5, 10 and 48 hour higher and values at 24 hour lower. The estimated maximum value is about 16.7% reached at around 18 hour time point post labelling, which is earlier than those of 113 and 169 days groups. The reason of more variation between individuals in this group is unknown. But with such small sample size, the spread of these observations, especially with the extreme values in them, has heavily affected the curve shape. For example, it can been seen from the plot that the highest observation value at 12 hour and the lowest at 24 hour make a huge influence on the fitted curves. The high value at 12 hour largely lifts up the mean value of all 5 observations at the time point and the one 24 hour drags down the mean a lot. It can be imagined that without these extreme values, the mean curve could shift to the right and reach the maximum at a later time point, which may overlap with the curves of 113 and 169 days groups. However, with data from only 5 mice, time course observation of each even from different mice, it is hard to treat the extreme values properly or investigate further by looking into other information. Besides, the plot also shows that although the observed data spreads out far from the group mean curve, the fitted individual curves gather closely suggesting small within group difference. However, this could potentially be an indication of slightly over-

smoothing of the model. If this is the case, one possible reason for overfitting in this example can be the effect of the AICc fitting criteria, which can be further investigated when necessary.

| Subjects | | 5 hour | 10 hour | 24 hour | 48 hour |
|---|---|---|---|---|---|
| Fitted mean enrichment of the group | | 0.670 | 1.425 | 1.573 | 1.1557 |
| Mouse 11 | Original enrichment | 0.617 | 1.974 | 1.783 | 1.372 |
| | Fitted enrichment | 0.690 | 1.445 | 1.592 | 1.170 |
| | Variation from group mean | 0.020 | 0.020 | 0.019 | 0.0144 |
| Mouse 12 | Original enrichment | 0.615 | 1.184 | 1.838 | 0.963 |
| | Fitted enrichment | 0.666 | 1.421 | 1.569 | 1.152 |
| | Variation from group mean | -0.004 | -0.004 | -0.004 | -0.004 |
| Mouse 13 | Original enrichment | 0.781 | 1.129 | 1.306 | 1.233 |
| | Fitted enrichment | 0.660 | 1.415 | 1.565 | 1.151 |
| | Variation from group mean | -0.010 | -0.010 | -0.009 | -0.005 |
| Mouse 14 | Original enrichment | 0.607 | 1.521 | 1.902 | 0.853 |
| | Fitted enrichment | 0.675 | 1.430 | 1.576 | 1.154 |
| | Variation from group mean | 0.005 | 0.005 | 0.003 | -0.002 |
| Mouse 15 | Original enrichment | 0.607 | 1.488 | 0.986 | 1.360 |
| | Fitted enrichment | 0.659 | 1.414 | 1.564 | 1.151 |
| | Variation from group mean | -0.011 | -0.011 | -0.009 | -0.004 |

Table 11: Fitted enrichment values (%) of D9DPPC in BALF of 84 day age group control mice at four time points.

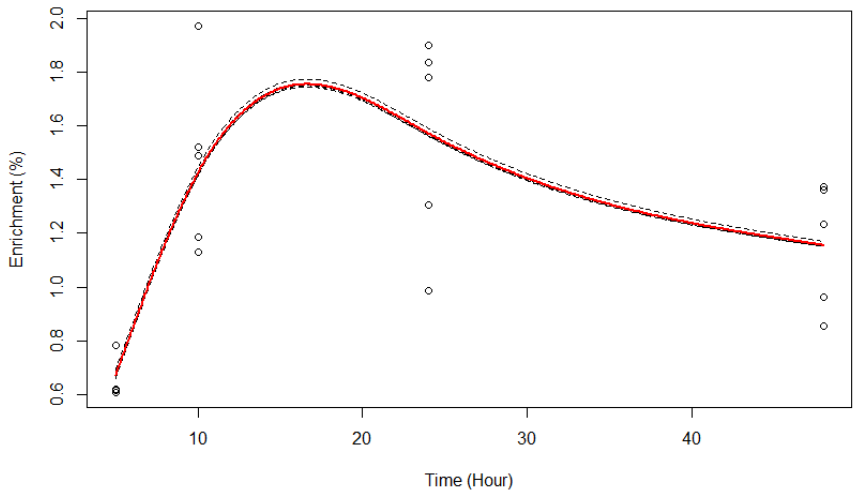



Figure 17: Time course enrichment model of D9DPPC in BALF sample of 84 day age group control mice. The empty circles are originally observed data points. The grey dashed curves are enrichment models of individual mice, and the red solid curve is that of the whole group.

#### 4.2.1.1.4 Analysis of combined control group

##### 4.2.1.1.4.1 Time course DPPC label enrichment of combined control group

The analysis of the three age group mice show a small difference of *methyl*$D_9$ label incorporation of DPPC species, although 84 days group mice have more scattered observations. All of the three groups have a very small sample sizes, with just 5 mice in each group. In this section, the three age groups were combined to form a larger control group with observations of 15 mice to characterise the time course D9DPPC enrichment of control mice population. Figure 18 shows a time trend of the combined group similar with the three age groups modelled in previous sections. The maximum enrichment is approximated at around 19 hour after labelling.

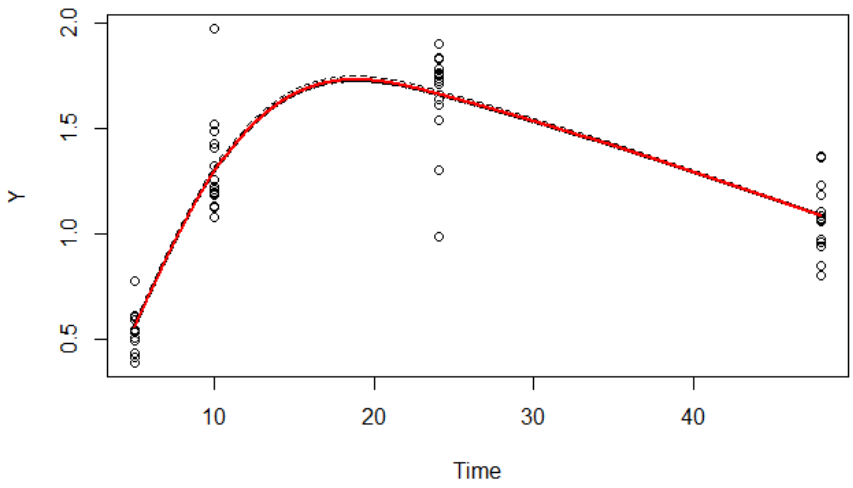



Figure 18: Time course enrichment model of D9DPPC in BALF of combined control group mice.

Table 12 lists the fitted values of the combined group. Comparing the results with previous analysis, the fitted values for individual Mouse 1 to 5 of 169 days group analysed

Table 12: Fitted enrichment values (%) of D9DPPC in BALF of combined control group mice at four time points.

| Subjects | | 5 hour | 10 hour | 24 hour | 48 hour |
|---|---|---|---|---|---|
| Group Mean Fitted enrichment Value | | 0.564 | 1.298 | 1.663 | 1.087 |
| Mouse 1 | Original enrichment | 0.000 | 1.084 | 1.613 | 1.229 |
| | Fitted enrichment | 0.558 | 1.293 | 1.659 | 1.088 |
| | Variation from group mean | -0.006 | -0.006 | -0.003 | 0.001 |
| Mouse 2 | Original enrichment | 0.538 | 1.258 | 1.747 | 1.060 |
| | Fitted enrichment | 0.564 | 1.298 | 1.663 | 1.087 |
| | Variation from group mean | 0.000 | 0.000 | 0.000 | 0.000 |
| Mouse 3 | Original enrichment | 0.493 | 1.222 | 1.726 | 1.089 |
| | Fitted enrichment | 0.562 | 1.296 | 1.661 | 1.087 |
| | Variation from group mean | -0.003 | -0.002 | -0.002 | 0.000 |
| Mouse 4 | Original enrichment | 0.592 | 1.196 | 1.763 | 0.944 |
| | Fitted enrichment | 0.564 | 1.298 | 1.662 | 1.085 |
| | Variation from group mean | 0.000 | 0.000 | -0.001 | -0.002 |
| Mouse 5 | Original enrichment | 0.550 | 1.325 | 1.737 | 0.806 |
| | Fitted enrichment | 0.565 | 1.299 | 1.662 | 1.083 |
| | Variation from group mean | 0.001 | 0.001 | -0.001 | -0.004 |
| Mouse 6 | Original enrichment | 0.416 | 1.135 | 1.540 | 1.065 |
| | Fitted enrichment | 0.554 | 1.288 | 1.655 | 1.085 |
| | Variation from group mean | -0.010 | -0.010 | -0.007 | -0.002 |
| Mouse 7 | Original enrichment | 0.434 | 1.431 | 1.710 | 0.973 |
| | Fitted enrichment | 0.565 | 1.298 | 1.663 | 1.086 |
| | Variation from group mean | 0.000 | 0.000 | 0.000 | -0.001 |
| Mouse 8 | Original enrichment | 0.508 | 1.411 | 1.756 | 1.071 |
| | Fitted enrichment | 0.567 | 1.301 | 1.665 | 1.088 |
| | Variation from group mean | 0.003 | 0.003 | 0.002 | 0.001 |
| Mouse 9 | Original enrichment | 0.392 | 1.213 | 1.829 | 1.183 |
| | Fitted enrichment | 0.561 | 1.295 | 1.662 | 1.089 |
| | Variation from group mean | -0.004 | -0.003 | -0.001 | 0.002 |
| Mouse 10 | Original enrichment | 0.542 | 1.192 | 1.636 | 1.103 |
| | Fitted enrichment | 0.561 | 1.295 | 1.660 | 1.087 |
| | Variation from group mean | -0.004 | -0.003 | -0.003 | 0.000 |
| Mouse 11 | Original enrichment | 0.617 | 1.974 | 1.783 | 1.372 |
| | Fitted enrichment | 0.586 | 1.318 | 1.680 | 1.095 |
| | Variation from group mean | 0.021 | 0.020 | 0.017 | 0.008 |
| Mouse 12 | Original enrichment | 0.615 | 1.184 | 1.838 | 0.963 |
| | Fitted enrichment | 0.566 | 1.299 | 1.664 | 1.086 |
| | Variation from group mean | 0.001 | 0.001 | 0.001 | -0.001 |
| Mouse 13 | Original enrichment | 0.781 | 1.129 | 1.306 | 1.233 |
| | Fitted enrichment | 0.560 | 1.294 | 1.659 | 1.086 |
| | Variation from group mean | -0.004 | -0.004 | -0.003 | -0.001 |
| Mouse 14 | Original enrichment | 0.607 | 1.521 | 1.902 | 0.853 |
| | Fitted enrichment | 0.575 | 1.308 | 1.669 | 1.086 |
| | Variation from group mean | 0.010 | 0.010 | 0.006 | -0.001 |
| Mouse 15 | Original enrichment | 0.607 | 1.488 | 0.986 | 1.360 |
| | Fitted enrichment | 0.559 | 1.293 | 1.659 | 1.087 |
| | Variation from group mean | -0.005 | -0.005 | -0.004 | 0.000 |

in combined group are close to those analysed as an independent group, although at 10 hour the values are higher. However, for Mouse 6 to 10 of 113 days group, the fitted values in combined group analysis are greater than individual group analysis at 5, 10 and 48 hours, and smaller at 24 hour. On the contrary, for individual Mouse 11 to 15 of 84 days group, the values in combined group analysis are lower at 5, 10, 48 hour, and higher at 24 hour. The analysis of 84 days group mice (Figure 17) suggests the outlier with a high value at 10 hour lifts up the fitted values at 5 and 10 hour, and the outlier with a small value at 24 hour pull down the fitted values at these time points. The combined group analysis has reduced the effect of these outliers, and the opposite way of handling 113 and 84 days mice indicates the SME takes all the individual subjects as a whole group for modelling, which neutralises the variations among individual subjects.

#### 4.2.1.1.4.2 Investigation of recombined time course data of control group

As stated before, the time series data for each mouse subject in this study is actually from different mice rather than repeated samples from a single mouse. In order to check if the combination of time points for each mouse subject affects the statistical analysis results, the time point data is recombined to form new time series data of mice in each age group, and the age groups are put together as a new control group for analysis. Two versions of recombination of mice data were constructed for comparisons with the original combination in the last section. The two plots are shown in Figure 19 and Figure 20.

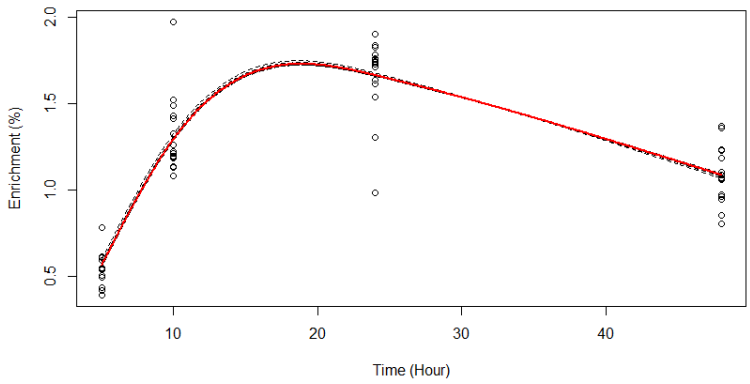



Figure 19: Modelling of D9DPPC enrichment of the first recombination of control group.

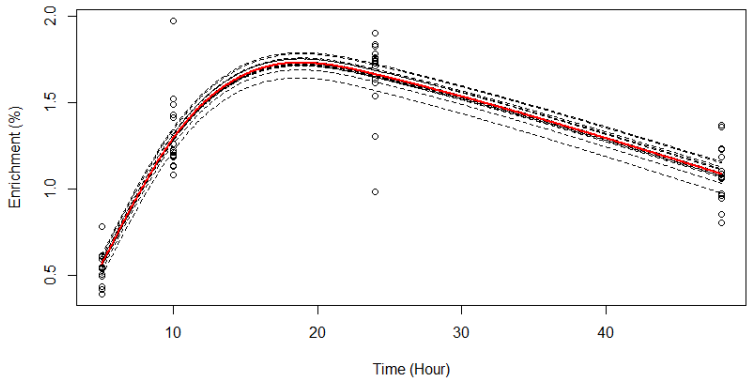



Figure 20: Modelling of D9DPPC enrichment of the second recombination of control group.

Figure 19 shows that the modelling of the first recombination data of control mice is very close to the original previously analysed. However, Figure 20 shows the individual curves of the second recombination version are more scattered, although the group mean curve is still similar to the other versions. The fitted group means of the two recombination versions of time series data are both similar to the original combination. Table 13 lists the fitted group mean values at the four observation time points for the three versions of combinations, and the difference is subtle. This indicates that the way of combination of

time series data has slightly affected the variation among individual models but not the underlying group incorporation pattern.

| Version of combination | 5 hour | 10 hour | 24 hour | 48 hour |
|---|---|---|---|---|
| Original combination | 0.564378638 | 1.298180555 | 1.662924 | 1.087009 |
| First recombination | 0.566041783 | 1.296100238 | 1.663433 | 1.087032 |
| Second recombination | 0.565236142 | 1.297654328 | 1.663010 | 1.086973 |

Table 13: The group mean values of D9DPPC enrichment in BALF fitted for three combination versions of time course data of 169 days control mice.

#### 4.2.1.2 Model fitting for KO mice group

The KO mice group consists of longitudinal enrichment data of five mice. Table 14 lists the PC enrichment approximation results of KO mice at four observation time points. The enrichment levels of KO mice are very low, and the fitted mean values of the whole group are 0.018%, 0.067%, 0.130% and 0.186% at 5, 10, 24 and 48 hour, respectively. The variations of individual fitted values are also small from the group means, ranging from -0.013% to 0.007% at 5 hour, -0.016% to 0.035% at 10 hour, -0.026% to 0.018% at 24 hour, and -0.044% to 0.025% at 48 hour. The differences between fitted values and observations are generally small as well. D9DPPC incorporation patterns are very close among the individual mouse in this group, and the consistency is an ideal group characteristic for biomarker discovery analysis. As mentioned before, KO5 mouse has an undetectable level of labelled PC species at 5 hour, which has been excluded from the analysis, but the SME method has made an estimation of the value of 0.005% without the observation, showing its ability in dealing with missing values.

| Subjects | | 5 hour | 10 hour | 24 hour | 48 hour |
|---|---|---|---|---|---|
| Fitted mean enrichment of the group | | 0.018 | 0.067 | 0.130 | 0.186 |
| KO1 | Original enrichment | 0.021 | 0.056 | 0.104 | 0.212 |
| | Fitted enrichment | 0.021 | 0.056 | 0.104 | 0.212 |
| | Variation from group mean | 0.003 | -0.011 | -0.026 | 0.026 |
| KO2 | Original enrichment | 0.026 | 0.054 | 0.137 | 0.200 |
| | Fitted enrichment | 0.026 | 0.054 | 0.137 | 0.200 |
| | Variation from group mean | 0.008 | -0.012 | 0.006 | 0.014 |
| KO3 | Original enrichment | 0.019 | 0.051 | 0.129 | 0.187 |
| | Fitted enrichment | 0.019 | 0.051 | 0.129 | 0.187 |
| | Variation from group mean | 0.001 | -0.016 | -0.002 | 0.001 |
| KO4 | Original enrichment | 0.018 | 0.075 | 0.133 | 0.185 |
| | Fitted enrichment | 0.018 | 0.075 | 0.133 | 0.185 |
| | Variation from group mean | 0.000 | 0.008 | 0.002 | -0.001 |
| KO5 | Original enrichment | | 0.102 | 0.149 | 0.142 |
| | Fitted enrichment | 0.005 | 0.102 | 0.149 | 0.142 |
| | Variation from group mean | -0.013 | 0.035 | 0.018 | -0.044 |

Table 14: Fitted enrichment values (%) of D9DPPC in BALF of KO mice at four time points.

The model plot (Figure 21) shows that the enrichment of D9DPPC in KO mice group generally increases along the time course. In previous discussion for control mice, the assumption was made as the mice being in a steady state with stable surfactant turnover. However, for the KO mice, it is not appropriate to make such assumption, as these mice models were PAP positive, with impaired macrophage degradation. With PAP the amount of entering surfactant species outweighs that of the outgoing, which results in the accumulation of the PC species as well as their labelled forms in the pool.

It is also noted that the curve for KO4 mouse, which has higher enrichment values at 10 and 24 hour and slightly lower value at 48 hour, exhibits a different temporal trend comparing to others. However, it does not affect the overall increasing pattern of the whole group. Besides, as described before, the time course data for each mouse subject is actually from different mice, so it is possible that some of the observations are from the

mice that have general higher incorporations and some are from the mice with low incorporations, and the combination can make fluctuating time trends.

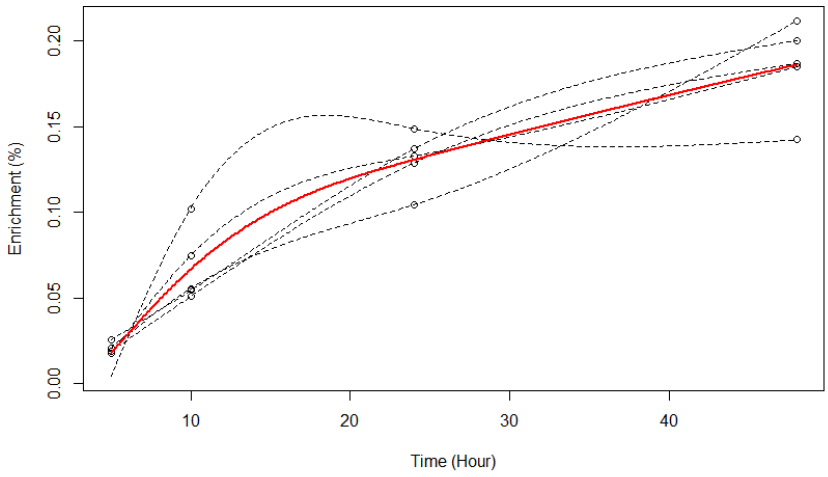



Figure 21: Time course enrichment model of D9DPPC in BALF of KO mice group.

#### 4.2.1.3 Comparisons between control and KO groups

The modelling of time course D9DPPC enrichment for control and KO groups has provided a wellspring of information regarding the species turnover. The curves do not only exhibit the activities of D9DPPC in DPPC pool during the time period, but also contain the information of activities of their precursors before turning up in the pool. The time of appearance, enrichment level and the rate of change reflect the metabolic kinetics of DPPC species indirectly, such as its synthesis, secretion, degradation and pool size. Moreover, the experimental information are also implied in the curve, e.g. the bolus infusion experiment has an enrichment curve with decreasing part after certain time point. All these pieces of information can be integrally manifested in the time course model, although they are not explicitly calculable as variables. Nevertheless, the temporal patterns are comparable between subjects for metabolic insights, especially when certain

conditions are identical to the subjects. In this study, general experimental procedures were applied to all of the mice and mice in the same group are with identical biological conditions, which allows the comparisons focus only on the different PC label incorporation between groups without being interfered by other factors.

The following comparison analysis of time course enrichment of D9DPPC is performed between every two age groups of control mice, as well as control with KO group, to get a general idea about how much difference there is between control mice and between control and KO groups. Table 15 shows the *Ft* statistics calculated for comparing the time course D9DPPC enrichment between control groups as well as control group with KO group.

(A): *Ft* statistics for control age groups.

| Within Control groups | G169 vs G113 | G169 vs G84 | G113 vs G84 |
|---|---|---|---|
| *Ft* | 1.340691 | 5.808352 | 15.03815 |
| *L* | 0.279276 | 1.174293 | 0.9763674 |
| *se* | 0.2083076 | 0.2021732 | 0.06492603 |

(B): *Ft* statistics for control and KO groups.

| Control with KO | G169 vs KO | G113 vs KO | G84 vs KO | All control vs KO |
|---|---|---|---|---|
| *Ft* | 41.87362 | 111.4986 | 142.3241 | 162.8645 |
| *L* | 8.657869 | 8.702522 | 8.50837 | 8.621543 |
| *se* | 0.2067619 | 0.07805052 | 0.05978164 | 0.05293691 |

Table 15: *Ft* statistics calculated for comparing every two age groups of control mice (A) and control groups vs KO group (B). *L* denotes the distance between the two group mean curves over the time period, and *se* is the functional standard error.

As shown in Table 15 (A), the calculated *Ft*s are 1.34 for the D9DPPC temporal enrichment of 169 days vs 113 days group mice, 5.81 for 169 vs 84, and 15.04 for 113 vs 84, respectively. In the table, along with the *Ft* are the two statistics denoted as *L* and *se*. *Ft* is calculated as the ratio of *L* over *se*. The numerator *L* is the distance between the two

group means. It is worth to stress here that the difference is not calculated from independent time point enrichment, but from the curves over the entire period. The denominator *se* is the term of functional standard error. The ratio of *L* over *se*, namely *Ft*, represents the scaled distance, which not only considers the difference between group means over time, but also take into account how spread out the data is. So the bigger the distance *L* between two groups is, the bigger the *Ft* will be. But the more scattered the data is, meaning big *se*, the smaller the *Ft* will be. This is reasonable because when the individuals in a group scatter far away from the group mean, the fitted group mean is less effective in representing the group and explaining variations. Therefore, the individual variations are taken into account by using *se* statistics. This is well demonstrated from the *Ft* values computed between the control groups. The *Ft* for 169 vs 113 group is very small with the value of 1.34, due to the small *L* of 0.279, indicating small difference between the two groups. Although the 169 vs 84 has a similar *se* with 169 vs 113, it has a much bigger *L*, resulting in a higher *Ft* of 5.81. While for the comparison between 113 vs 84, although the *L* is similar with that of 169 vs 84, the *se* is much smaller with the value of 0.0649, making the *Ft* much higher at 15.04, suggesting a greater scaled difference than other group pairs.

The *Ft* values between control groups are generally small, whereas the *Ft* statistics calculated for control group vs KO group are considerably bigger. Table 15(B) shows that the *Ls* of control vs KO group pairs are all over 8, much higher than those of control groups which are around 1. The *se* values are comparatively low, leading to much higher *Ft* values. The much higher *Ft* values of control vs KO comparing to those of control vs control suggest that there is significant difference in D9DPPC turnover between control and KO mice.

The *Ft* of combined control vs KO is 162.86, which is highly significant comparing with those of the individual age control groups vs KO. This can be explained by checking the *L* and *se*. Although the calculated *L* between combined control and KO groups is 8.62, not much different from the *L*s between individual age groups vs KO group, the *se* with the value 0.0529 is smaller than the others, which results in a bigger *Ft* ratio. This suggests the combination of all age groups not only keeps the similar distance with individual age groups to the KO group, but also reduces the variations of models of individual mice, which makes a good representation of control mice population and gives better results when comparing with KO mice model.

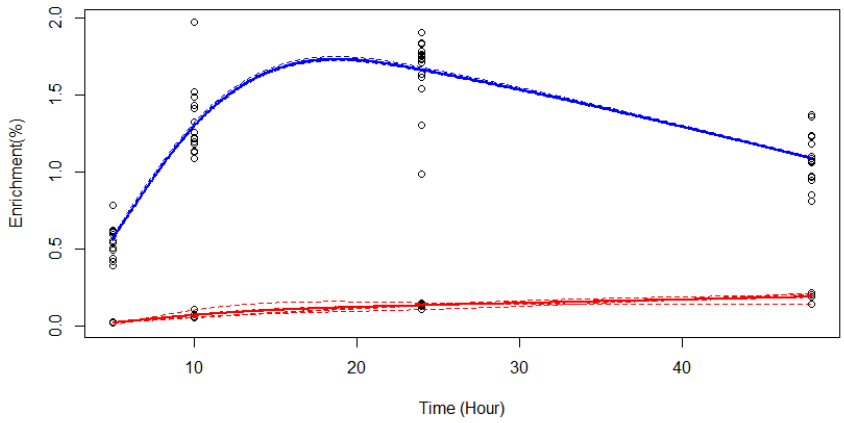



Figure 22: Comparison between time model of D9DPPC in BALF between combined control group and KO group

In Figure 22 the D9DPPC incorporation model of combined control group and KO group are plotted together for comparison. The control group curve has a trend of increasing at the beginning and decreasing after getting the highest level along the time axis. In contrast, the KO group has a trend of increasing slowly along the time with no sign of decreasing within the time window. The value ranges of the two groups are very different.

The lowest enrichment level of control group is shown around 5 hour with the fitted group mean value is 0.564%. While for the KO group, the highest enrichment values appears around 48 hour with the group mean of 0.18%. The enrichment level in control group, including group mean, individual mean and real observations, are all substantially higher than the highest point in KO group. The distinctive time trend and value ranges show the control and KO mice are significantly different.

As discussed before, a temporal enrichment model does not explicitly deliver any specific parameters regarding the metabolic process of labelled or unlabelled species, but rather it manifests the information integrally. Therefore, even though it is not possible to quantify the factors that affect the timing of appearance in the pool, the enrichment level, or the curve shapes, temporal models themselves are informative, and the difference between them is meaningful and quantifiable for group comparison.

It is worth to mention that, although *Ft* is a *t*-style statistic for assessing the difference between groups, there is no established distribution for the statistic under the null hypothesis of no difference between the two mean curves. Thus, it is not possible to make a standard statistical test to evaluate the significance of difference identified by the variables directly. However, by checking the *Ft* values and observing the graph based on the understanding of the mice model, it is possible to get a general idea about how far the two groups are from each other and how significant it is.

#### 4.2.1.4 Combining age groups as control group for the biomarker discovery analysis

In the last section, both individual age groups and combined group of control mice have been compared with KO group to assess the capability of D9DPPC time course incorporation in distinguishing control and KO mice. However, there are some downsides of using individual age groups for the analysis, which suggests that it may be better to

group them together as a whole group to be compared with KO group. The sample size is very small in each group, with data for only 5 subjects, some of which are even excluded due to measuring errors. It is hard for an individual group with such few subjects to cover the underlying population. Besides, it is found that some extreme outliers exist in the already scarce dataset and heavily influence the modelling of the time series data, however, removing the outliers would result in even less observations. Figure 23 shows the fitted temporal enrichment models of species D9PC16:0a/16:1 for the three age groups. It is shown that under the effect of an extreme outlier at 5 hour, the mean value of 169 days group mice has been lifted to be around 1.3%, which could have been well under 1.0% and had a curve shape close to the other two groups. Note that there are only 4 observations at this time point for 169 mice, due to the exclusion of one data point of measuring error, which results in the outlier asserting greater impact to the overall group mean value, and removing the outlier would worsen the data insufficiency.

Another issue with the mice model data is that the time series are actually a random combination of different mice at different time points rather than repeated sampling from a single subject over a time period. The individual curve obtained by putting together the time point data for a made-up mouse can be different from a real mouse from the underlying population, which can affect the estimation of the mean curve of the whole population, especially when the sample size is small.

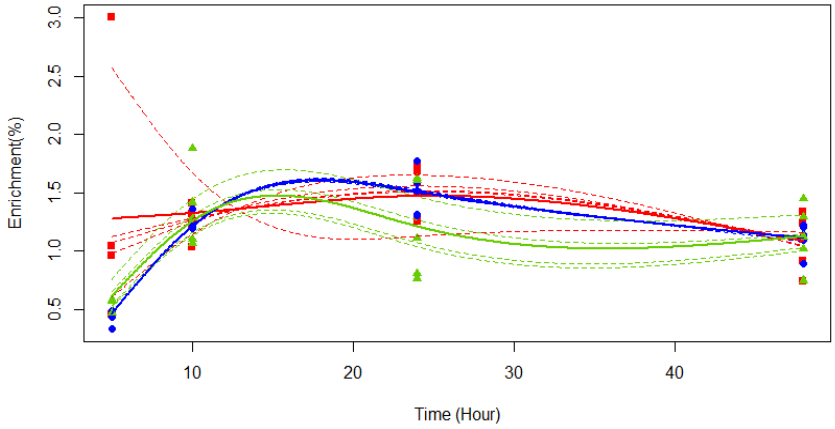



Figure 23: Time course enrichment model of species D9PC16:0a/16:1 for the three age groups of control mice. The lines in red, blue, and green are the enrichment curves of 169, 113 and 84 days mice groups. The extreme outlier of a mouse in 169 day age group at five hour time point, with the value of 3%, hugely affects the mean value of the whole group.

In order to tackle these issues, it is appropriate and more sensible to combine all the three age groups as one control group for the comparison with KO group, given that the three groups resemble each other quite well (Figure 24). The benefit of combining has already been demonstrated in the last section when comparing with KO group, which has generated better results than using individual age groups. Besides, the rare extreme outlier should be removed, as they may significantly twist the shape of the underlying group mean curve. The impact can be even worse when inappropriately combining the time point observations from significantly different mice, which could result in great fluctuations in the fitted curves for made-up mice. Therefore, in the following group comparisons for biomarker discoveries, the combined control group will be used and extreme outliers will be removed.

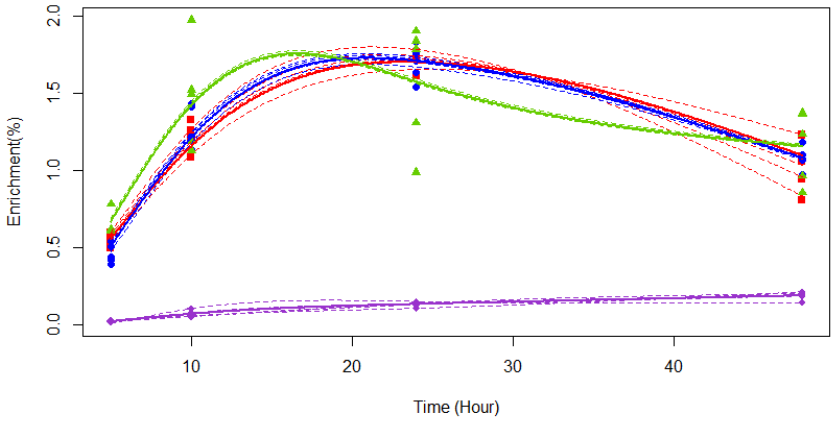



Figure 24: Integrated graph of D9DPPC time course enrichment modelled for three individual age control groups and KO group. The curves for the three age groups of 169 days, 113 days and 84 days, highlighted in red, blue and green respectively, have similar time trend and mean curve values. 84 days group has wider spread of values and outliers, which makes the whole curve shape lean forward slightly comparing to the other two groups. Generally, the time series models of the three control groups are very close to each other, and all of them are far above the KO group highlighted in purple, suggesting the substantial difference between control and KO groups.

## 4.3 Characterisation of time course methylD9 -PC enrichment in BALF of control vs KO groups and biomarker analysis

Based on the methodology established through analysing label incorporation of the single species DPPC, the time course enrichment of sixteen labelled PC species in BALF were modelled and analysed for both control and KO mice groups in this section. Comparisons were carried out between the two groups for multivariate time course biomarker analysis.

### 4.3.1 Time course enrichment of labelled PC species

The SME method has worked effectively in modelling the time course enrichment of D9DPPC for control and KO groups, as well as comparing groups to investigate if and how well the DPPC label incorporation can distinguish between different groups. In this section the temporal enrichment of sixteen *methyl*$D_9$ labelled PC species in BALF are

modelled and compared between combined control and KO groups, for understanding the species metabolism as well as looking for potential biomarkers.

| *m/z* | Species | Group | Enrichment (%) | | | |
|---|---|---|---|---|---|---|
| | | | 5 hour | 10 hour | 24 hour | 48 hour |
| 715 | D9PC16:0/14:0 | **Control** | **0.466** | **1.104** | **1.512** | **1.020** |
| | | KO | 0.018 | 0.051 | 0.152 | 0.213 |
| 727 | D9PC16:0a/16:1 | **Control** | **0.505** | **1.227** | **1.409** | **1.102** |
| | | KO | 0.036 | 0.057 | 0.116 | 0.216 |
| 729 | D9PC16:0a/16:0 | **Control** | **0.563** | **1.188** | **1.521** | **1.016** |
| | | KO | 0.025 | 0.044 | 0.143 | 0.167 |
| 741 | D9PC16:0/16:1 | **Control** | **0.643** | **1.436** | **1.724** | **1.041** |
| | | KO | 0.029 | 0.086 | 0.186 | 0.248 |
| 743 | D9PC16:0/16:0 | **Control** | **0.564** | **1.298** | **1.663** | **1.087** |
| | | KO | 0.018 | 0.067 | 0.130 | 0.186 |
| 767 | D9PC16:0/18:2 | **Control** | **1.042** | **1.686** | **1.716** | **1.051** |
| | | KO | 0.032 | 0.092 | 0.162 | 0.228 |
| 769 | D9PC16:0/18:1 | **Control** | **1.016** | **1.813** | **1.898** | **1.107** |
| | | KO | 0.056 | 0.113 | 0.184 | 0.233 |
| 771 | D9PC16:0/18:0 | **Control** | **0.074** | **0.252** | **1.169** | **0.751** |
| | | KO | 0.005 | 0.031 | 0.098 | 0.137 |
| 791 | D9PC16:0/20:4 | **Control** | **0.911** | **1.560** | **1.773** | **0.981** |
| | | KO | 0.043 | 0.136 | 0.230 | 0.297 |
| 793 | D9PC18:1/18:2 | **Control** | **0.756** | **1.273** | **1.556** | **0.964** |
| | | KO | 0.048 | 0.100 | 0.248 | 0.169 |
| 795 | D9PC18:0/18:2 | **Control** | **0.409** | **0.599** | **1.129** | **0.841** |
| | | KO | 0.058 | 0.073 | 0.110 | 0.145 |
| 797 | D9PC18:0/18:1 | **Control** | **0.501** | **0.680** | **1.081** | **1.203** |
| | | KO | 0.091 | 0.075 | 0.170 | 0.167 |
| 815 | D9PC16:0/22:6 | **Control** | **1.003** | **1.435** | **1.857** | **0.998** |
| | | KO | 0.047 | 0.112 | 0.199 | 0.231 |
| 817 | D9PC18:1/20:4 | **Control** | **0.878** | **1.308** | **1.981** | **1.042** |
| | | KO | 0.112 | 0.165 | 0.176 | 0.223 |
| 819 | D9PC18:0/20:4 | **Control** | **0.439** | **0.582** | **0.942** | **0.894** |
| | | KO | 0.044 | 0.050 | 0.078 | 0.172 |
| 843 | D9PC18:0/22:6 | **Control** | **0.365** | **0.522** | **1.080** | **0.978** |
| | | KO | 0.015 | 0.043 | 0.105 | 0.117 |

Table 16: Fitted group mean values of time course enrichment of sixteen *methyl*$D_9$ labelled PC species in BALF samples of control and KO mice groups at four sampling time points. The species are classified into three categories in terms of temporal enrichment levels of control group based on the result of hierarchical clustering analysis. The species with comparatively higher time course enrichment are in the high category highlighted in pink, and the species with comparatively lower time course enrichment values are in low category highlighted in blue, with the rest of species in the middle category.

The modelling results are listed in Table 16, which gives the fitted values of group means of sixteen labelled PC species at four sampling time points, for both control and KO groups. The species in control group are classified into three categories based on their time course enrichment level using hierarchical clustering analysis, and the three groups can be generally recognised as high, middle and low label incorporation groups. The clustering result is shown in Figure 25, and the species in the three categories are highlighted Table 16.

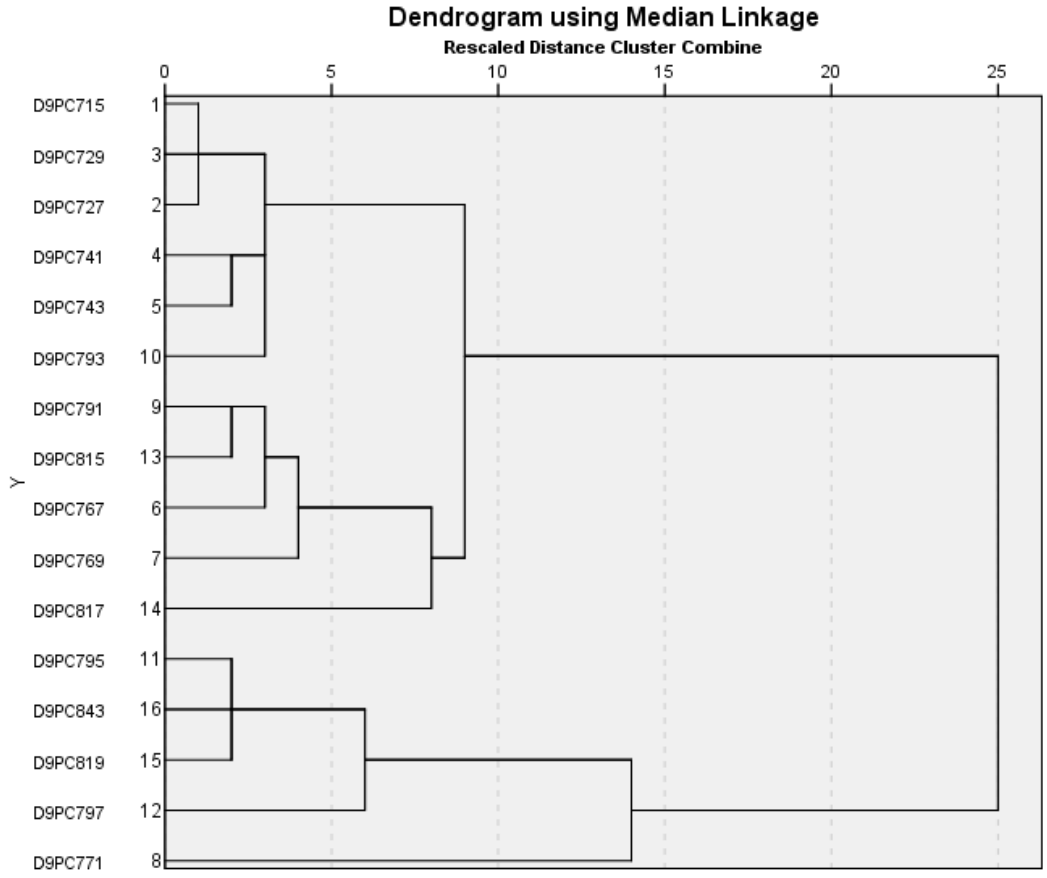



Figure 25: Dendrogram of hierarchical clustering analysis for time course enrichment of sixteen *methyl*$D_9$ labelled PC species in BALF samples of control mice. Species is denoted as "D9PC" with their m/z value in the dendrogram.

The species that have similar level of time course enrichment with D9PC16:0/16:0 (D9DPPC) are in the middle category, including the characteristic surfactant PC species

D9PC16:0/16:0, D9PC16:0/14:0, D9PC16:0a/16:1, D9PC16:0a/16:0, D9PC16:0/16:1, and D9PC18:1/18:2. In the high category, most of the species with higher enrichment are with a fatty acyl chain of 16:0 and another unsaturated longer chain, including D9PC 16:0/18:2, D9PC 16:0/18:1, D9PC 16:0/20:4 and D9PC 16:0/22:6, with one exemption of D9PC 18:1/20:4 that does not have a 16:0 chain. Species in the low category include D9PC16:0/18:0, D9PC18:0/18:1, D9PC18:0/18:2, D9PC18:0/20:4, and D9PC18:0/22:6, all of which are with a saturated 18:0 fatty acyl chain. Moreover, D9PC16:0/18:0 is the only species with a 16:0 acyl that falls in this category. As 16:0 is the component of major surfactant PC with high synthesis, the low level of its combination with 18:0 into D9PC16:0/18:0 is a strong evidence that the species with 18:0 have lower incorporation than other species.

As discussed before, the control mice are assumed to be in a steady biological state. In steady state, the fractional rate of appearance of species is directly proportional with the initial change rate of labelled PC species enrichment upon arrival at the pool (at time zero), and inversely proportional with the initial enrichment of labelled species in the source pool. It is known that surfactant PC species in BALF is synthesised and remodelled in AT II cells (Chen *et al.*, 2006), which are wrapped in lamella bodies and secreted into the BALF through exocytosis. In the steady state where secretion rate is constant, it is reasonable to assume all surfactant PC species at alveolar surface turn up in the pool at the same time and have the same fractional appearance/disappearance rate.

$$FAR = \frac{Z'_B(0)}{Z_A(0)} \qquad (14)$$

In equation (14) $FAR$ is the fractional appearance rate of a species secreted from pool *A* into pool *B.* $Z'_B(0)$ is the initial change of enrichment of the labelled form of the species

appeared in pool *B* at time 0, and $Z_A(0)$ is the enrichment of the labelled species in pool *A* before secreted into pool *B*.

Although we do not have the enrichment values of labelled PC species in the pools at time zero to calculate the parameters, based on the assumption of same turning up time in the surfactant pool of all PC species, a higher enrichment level of a species at very early time point indicates a higher initial change rate of incorporation, hence the greater $Z'_B(0)$ the species has. Under the assumption of same fractional appearance rate, the bigger $Z'_B(0)$ is, the higher $Z_A(0)$ will be. Therefore, the species in BALF samples with high early time enrichment should be down to their initial high enrichment level in the source pools in the lung cells.

The analysis above categorises the species in BALF into three categories. The species in the high category with higher start of enrichment level should also have higher enrichment level in their source pool in AT II cells than those in the middle category. While within the category the species having similar early time enrichment level may probably have similar source pool enrichment as well. In particular, the characteristic lung surfactant species PC16:0/16:0, PC16:0/14:1 and PC16:0/16:1 in the middle category have a similar time course enrichment level at early time as well as during the whole time period, therefore, they may have similar $Z_A(0)$ in the source pool before secretion. As to the lower category species with 18:0 acyl, the low early time point enrichment indicates low initial enrichment from the source pool.

From the modelled time course enrichment of KO group, it was found that the PC species in KO group have the similarities in that, species in the higher/lower enrichment level category of control group can generally fall in the higher/lower category in KO group. This

may imply that despite the impaired degradation of surfactant PC, their secretion from lung cells into the surfactant pool may not be affected much.

### 4.3.2 Temporal trend of labelled PC species

The time course enrichment of each labelled PC species in control groups all follows an increasing - decreasing trend. The curve shapes of their models can be grouped into two categories. Figure 26 displays the enrichment curves of some species that represent the two main types of curve shapes among the sixteen analysed species. The curves of species without 18:0 have larger gradient at the increasing part and smaller slope at the decreasing part. The curves rise up to reach their maximum level at around 18 hours after labelling and then gradually drop down. The species with 18:0 acyl have a very different time trend from other species. They all start out at lower enrichment but keep increasing slowly to get the highest level at around 30 hour post labelling, much later than other species, and then go down. Another finding is that, the enrichment of all the species get close to 1% at 48 hour time point post labelling, no matter if they start high or low, increase or decrease. However, as the observations do not cover the full decreasing part, it is not possible to further examine the rate of disappearance of the species. The temporal enrichment of most of the labelled PC species of KO group has an increasing trend during the 48 hour time window, except D9DPPC16:0a/16:1, D9DPPC16:0a/16:0, D9DPPC18:1/18:2 and D9DPPC18:0/18:1, which decreases a bit at the end of the time period. Overall, the curves of control and KO group are distanced far away from each other, making significant gaps for all the PC species.

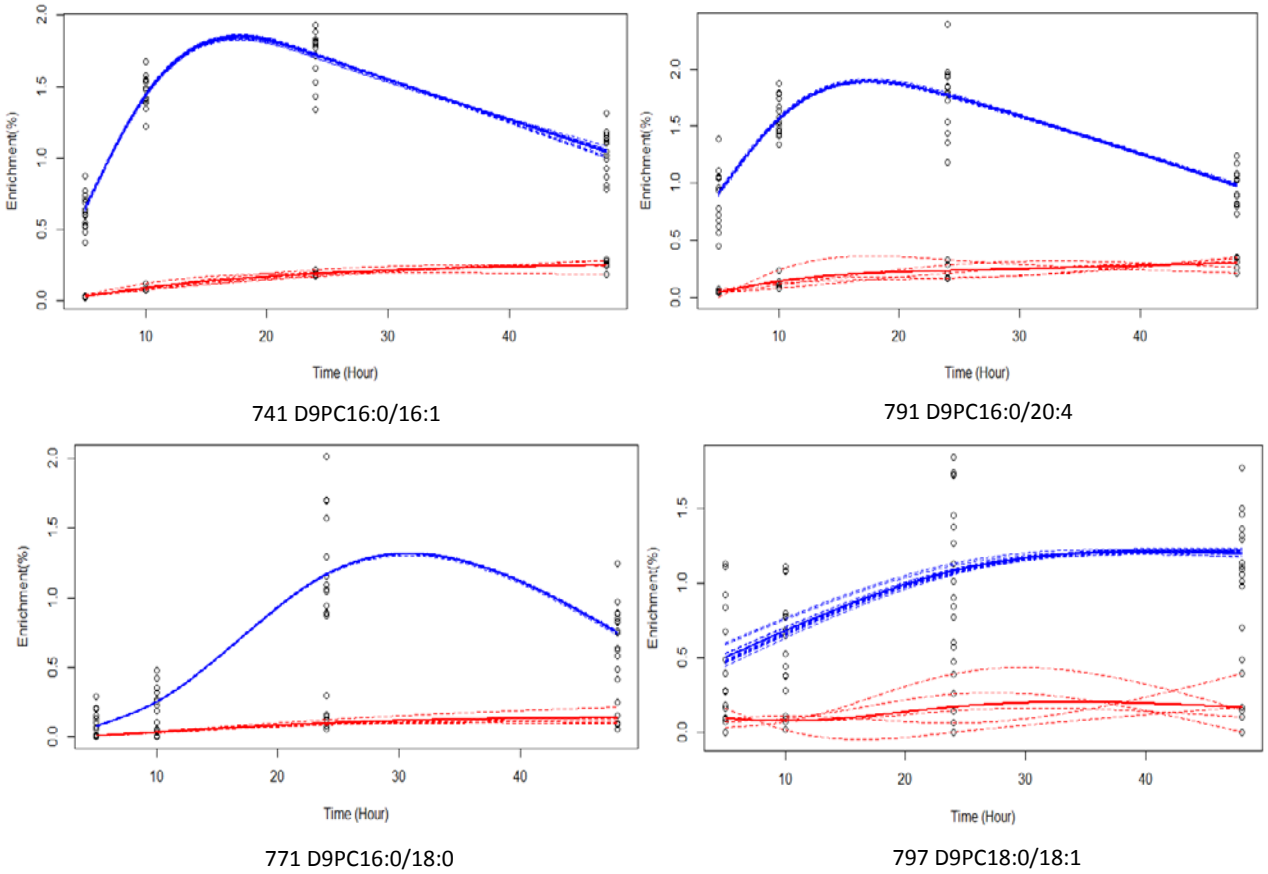



Figure 26: Time course enrichment models of four species in BALF of control and KO mice groups. There are two types of curve shapes for control mice (in blue). Species D9PC16:0/16:1 and D9PC16:0/20:4 have the typical curve shape of the species without 18:0 acyl chain, which increases fast at the beginning, and decrease slowly after reaching the highest enrichment level at around 18 hour. Species D9PC16:0/18:0 and D9PC18:0/22:6 represent the species with 18:0 acyl chain and have a curve shape of slower increase to the highest enrichment level at around 30 hour. The models of KO mice (in red) generally show an increasing trend with very small gradient.

It has been discussed in the previous sections that the swing of time course enrichment curves of species is directly related to the label enrichment in the source pool before secretion, and the enrichment level can imply species turnover characteristics. The curve shape can also provide valuable information. Benefiting from being able to analyse multiple species, it is possible to get extra information by comparing the curve shapes of these species. As shown in the plots in Figure 26, the PC species without 18:0 enrichment

have a similar temporal pattern, indicating the similar time patterns of their enrichment in the source pool, which means the flood of these labelled species change in the same way in the AT II cells before secretion into surfactant pool. Similarly, all species with an 18:0 fatty acyl exhibit a similar pattern, suggesting similar flood change of labelled molecules in lung cells. However, the pattern of these species is very different from that of the species without 18:0, which may suggest a different metabolism of the two group of species.

### 4.3.3 Group comparison and biomarker analysis

The control and KO groups are distinguished in their PC label incorporation, in terms of enrichment values and model curve shapes of species. Yet for multivariate biomarker discovery analysis, it is needed to quantify and compare the group difference recognised by each species to identify the most significant species that can distinguish the two groups. This is done by calculating and comparing the *Ft* statistics. The *Ft*, *L,* and *se* for the sixteen species are listed in Table 17, and the species are sorted in descending order of *Ft*.

The result shows that all the species with high *L* values above 9, which features larger distance between the two groups, have higher time course enrichment and falling in the high category in model fitting. These species include D9PC16:0/18:2, D9PC16:0/18:1, D9PC16:0/22:6 and D9PC18:1/20:4. Other species with high *L* values over 8 include D9PC16:0/16:0, D9PC16:0/16:1 and D9PC16:0/24:0. The species with 18:0 fatty acyl chain have the lowest *L* values about 5. This is due to the low incorporation of these species especially between the starting time point and the maximum time point, resulting in the bigger gaps between the enrichment of the control and KO comparing to those of the species without 18:0 acyl chain.

| *m/z* | Species | *L* | *se* | *Ft* |
|---|---|---|---|---|
| 743 | D9PC16:0/16:0 | 8.621543 | 0.0529369 | 162.864500 |
| 727 | D9PC16:0a/16:1 | 7.572277 | 0.0469461 | 161.297200 |
| 715 | D9PC16:0/14:0 | 7.572791 | 0.0474698 | 159.528600 |
| 741 | D9PC16:0/16:1 | 8.660240 | 0.0627928 | 137.917767 |
| 769 | D9PC16:0/18:1 | 9.936843 | 0.0723025 | 137.434293 |
| 767 | D9PC16:0/18:2 | 9.137802 | 0.0967007 | 94.495746 |
| 771 | D9PC16:0/18:0 | 5.746057 | 0.0647328 | 88.765731 |
| 815 | D9PC16:0/22:6 | 9.148791 | 0.1154591 | 79.238371 |
| 729 | D9PC16:0a/16:0 | 7.844602 | 0.1103914 | 71.061713 |
| 791 | D9PC16:0/20:4 | 8.693447 | 0.1558689 | 55.774096 |
| 819 | D9PC18:0/20:4 | 5.079186 | 0.1082473 | 46.922057 |
| 795 | D9PC18:0/18:2 | 5.685439 | 0.1466385 | 38.771790 |
| 843 | D9PC18:0/22:6 | 5.699652 | 0.1503399 | 37.911780 |
| 817 | D9PC18:1/20:4 | 9.630963 | 0.4165925 | 23.118426 |
| 793 | D9PC18:1/18:2 | 7.564625 | 0.3800108 | 19.906342 |
| 797 | D9PC18:0/18:1 | 5.814019 | 0.3003244 | 19.359130 |

Table 17: *Ft* statistics for sixteen *methyl*$D_9$ labelled species in BALF of control and KO mice. The species are sorted in descending order of their *Ft* values. D9PC16:0/16:0 has the highest *Ft* value and is the most significant species in distinguishing control and KO groups.

The D9PC16:0/16:0, D9PC16:0a/16:0, D9PC16:0/14:0 and D9PC16:0/16:1 are not the species with the highest *L* values, however, they have the highest *Ft* values due to small low *se* values. Species D9PC16:0/18:1 has the highest *L*, but its *Ft* is ranked behind the species mentioned above due to a moderate *se* value. Species D9PC18:1/20:4, the only one in the high category without 16:0 fatty acyl, is ranked very low due to a bigger *se*. Another interesting species D9PC16:0/18:0, with the lowest enrichment over the time period, has a small *se* and a comparatively high *Ft.* The results demonstrate that *Ft* statistic not only reflects the group distance *L,* but also takes into account the how scattered, represented by *se*, the individuals are from their group mean curve.

The analysis in previous sections compared time course incorporation of D9DPPC of individual age groups of control mice separately, and it also compared the control with

KO mice. The comparisons between the three age groups get small *Ft* ranging from 5 to 15, due to the small differences between the groups, whereas the comparison between control and KO mice obtained considerably big *Ft* value. The comparisons between KO mice with both 113 days group and 84 days group have generated high *Ft* values over 100, and even the smallest *Ft* for 169 days group vs KO group is over 40, due to the larger variations of individual mice in 169 group. The range of *Ft* values obtained from these comparisons make a good reference for other group comparisons. By referring to these figures, the high *Ft* values for the sixteen BALF species listed in Table 16 suggest the significance of most labelled PC species in identifying the difference between control and KO groups.

Sorting the species by descending order of *Ft* means sorting them according to their significance in distinguishing the two groups. Hence, the higher a species is ranked, the better it can identify the difference between the two groups. It can be seen from the list that the species with a 16:0 or 16:0a acyl are all ranked higher than those without them, which suggests higher significance of these species in differentiating control and KO groups. Therefore, in BALF sample, these species are more likely to be taken as biomarker candidates for discriminating the two conditions.

The analysis shows that control mice are significantly different from KO mice, in terms of time pattern of enrichment as well as the value range of analysed labelled PC species in BALF. This result is consistent with the current knowledge about the pulmonary surfactant metabolism. The macrophages are known to play an important role in surfactant catabolism, which degrade around 20% surfactant from the alveoli. GM-CSF receptor beta chain knock out mice have impaired macrophages degradation of pulmonary surfactant, leading to its abnormal accumulation and the development of PAP.

The accumulation of both labelled and unlabelled PC leads to the largely increased value of the denominator term in the enrichment function. Despite the increased labelled species in the numerator, the sum of these labelled species and huge proportion of unlabelled species in the denominator terms makes the overall enrichment much smaller than that of control mice having a stable pool size with normal catabolism.

It can be hard to elucidate what affects the temporal trend of each individual PC species of the mice in each group, which may involve various factors. These could be the physical status of the mice, age factor, strain of the mice, label infusion approach, sample processing procedures and many other. These factors can have an impact either directly or indirectly, identifiable or unidentifiable with current knowledge. Although details of these factors are yet to be found out, luckily, their influences manifest integrally in the species time course enrichment. The SME method has successfully characterized the temporal enrichment patterns and facilitated the biomarker discovery analysis by quantifying and comparing the group differences identified by the individual species, which gives reliable results that are consist with the current knowledge. Therefore, it is viable to use the surfactant *methyl*$D_9$-PC time course enrichment as potential biomarkers for identifying PAP mice from healthy mice.

## 4.4 Characterisation of time course *methyl*$D_9$ -PC enrichment in lung tissue of control vs KO groups and biomarker analysis

### 4.4.1 Time course enrichment of labelled PC species

To further investigate surfactant PC species metabolism across different compartments in the respiratory system, the temporal enrichment of labelled PC species in lung tissue

| *m/z* | Species | Group | Enrichment (%) | | | |
|---|---|---|---|---|---|---|
| | | | 5 hour | 10 hour | 24 hour | 48 hour |
| 715 | D9PC 16:0/14:0 | **Control** | **1.571** | **1.318** | **1.244** | **0.942** |
| | | KO | 0.253 | 0.243 | 0.278 | 0.278 |
| 727 | D9PC16:0a/16:1 | **Control** | **0.575** | **0.531** | **0.633** | **0.698** |
| | | KO | 0.247 | 0.244 | 0.297 | 0.348 |
| 729 | D9PC16:0a/16:0 | **Control** | **0.982** | **0.968** | **0.933** | **0.892** |
| | | KO | 0.217 | 0.218 | 0.225 | 0.276 |
| 741 | D9PC16:0/16:1 | **Control** | **2.355** | **1.869** | **1.493** | **0.968** |
| | | KO | 0.362 | 0.357 | 0.344 | 0.330 |
| 743 | D9PC16:0/16:0 | **Control** | **1.636** | **1.373** | **1.222** | **1.144** |
| | | KO | 0.306 | 0.295 | 0.269 | 0.262 |
| 767 | D9PC16:0/18:2 | **Control** | **2.261** | **1.749** | **1.332** | **0.894** |
| | | KO | 0.775 | 0.703 | 0.531 | 0.373 |
| 769 | D9PC16:0/18:1 | **Control** | **1.997** | **1.619** | **1.352** | **0.994** |
| | | KO | 0.911 | 0.818 | 0.600 | 0.458 |
| 771 | D9PC16:0/18:0 | **Control** | **0.721** | **0.765** | **0.919** | **0.815** |
| | | KO | 0.252 | 0.257 | 0.271 | 0.297 |
| 791 | D9PC16:0/20:4 | **Control** | **1.837** | **1.524** | **1.332** | **0.909** |
| | | KO | 1.292 | 1.104 | 0.681 | 0.547 |
| 793 | D9PC18:1/18:2 | **Control** | **1.957** | **1.477** | **1.304** | **0.880** |
| | | KO | 1.440 | 1.333 | 1.037 | 0.557 |
| 795 | D9PC18:0/18:2 | **Control** | **1.499** | **1.437** | **1.261** | **0.960** |
| | | KO | 1.080 | 1.021 | 0.854 | 0.567 |
| 797 | D9PC18:0/18:1 | **Control** | **1.187** | **1.147** | **1.334** | **1.100** |
| | | KO | 1.127 | 1.089 | 0.980 | 0.788 |
| 815 | D9PC16:0/22:6 | **Control** | **2.010** | **1.721** | **1.206** | **0.871** |
| | | KO | 0.937 | 0.689 | 0.472 | 0.383 |
| 817 | D9PC18:1/20:4 | **Control** | **2.075** | **1.616** | **1.276** | **0.871** |
| | | KO | 1.478 | 1.105 | 0.754 | 0.400 |
| 819 | D9PC18:0/20:4 | **Control** | **1.386** | **1.334** | **1.234** | **0.901** |
| | | KO | 1.244 | 1.150 | 0.899 | 0.536 |
| 843 | D9PC18:0/22:6 | **Control** | **1.202** | **1.160** | **1.041** | **0.833** |
| | | KO | 0.705 | 0.666 | 0.558 | 0.372 |

Table 18: Fitted group mean values of time course enrichment of sixteen *methyl*$D_9$ labelled PC species in lung tissue samples of control and KO mice groups at four sampling time points. The species are classified into three categories in terms of temporal enrichment levels of control group based on the result of hierarchical clustering analysis. The species with comparatively higher time course enrichment are in the high category highlighted in pink, and the species with comparatively lower time course enrichment values are in low category highlighted in blue, with the rest of species in the middle category.

were analysed in addition to those in BALF samples. The same methodology was applied to lung tissue samples and the results are shown in Table 18. By applying the hierarchical clustering analysis to the control group (Figure 27), the species can be categorized into three groups based on their time course enrichment levels. The species in the high enrichment level group include D9PC16:0/16:1, D9PC16:0/18:2, D9PC16:0/18:1, D9PC16:0/20:4, D9PC18:1/18:2, D9PC16:0/22:6, and D9PC18:1/20:4. D9PC16:0/16:1 has the highest time course enrichment. The D9PC16:0a/16:1 and D9PC16:0/18:0 are in the low category with extremely low incorporation comparing to other species. The remained species are in the middle category. Generally, among all the sixteen species, the enrichment levels of those with 18:0 are lower than other species, except D9PC16:0a/16:1, which has the lowest enrichment of all.

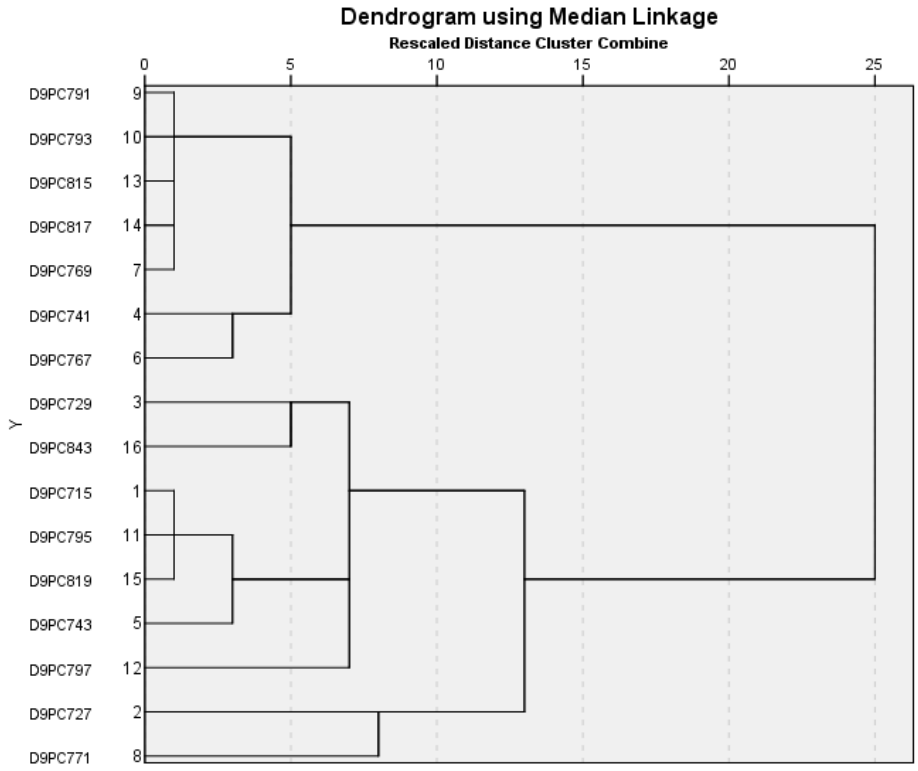



Figure 27: Dendrogram of hierarchical clustering analysis for time course enrichment of the sixteen deuterium labelled PC species in lung tissue samples of control mice. Species is denoted as “D9PC” with their m/z value.

Although species of KO group mice are not categorised into groups, the characteristic surfactant species and some species, including D9PC16:0/14:0, D9PC16:0a/16:1, D9PC16:0a/16:0, D9PC16:0/16:1, D9PC16:0/16:0 and D9PC16:0/18:0, have much lower enrichment levels than the other species. In particular, the time course enrichment of some unsaturated species with longer fatty acyl chains are very high, which are close to those of control group, such as D9PC18:0/20:4. Generally, the PC label enrichment of KO group is much higher in lung tissue samples than in BALF samples, and the gaps between control and KO group are smaller in lung tissue samples.

The PC species detected in lung tissue samples can be from several pools, such as synthesised surfactant, cell membrane, and blood. Therefore, the label incorporation pattern of a PC species observed in the sample actually reflects the collective activities of it from multiple pools. This may partly explain the much lower enrichment of characteristic surfactant species in KO groups than other species. The surfactant accumulation of PAP mice mainly affects the surfactant species pool, and other pools like membrane or blood PC are not affected. Consequently, the characteristic surfactant species, which only come from the surfactant pool, suffer from the impact of accumulation more significantly, but the same effect on the non-characteristic surfactant species has been reduced by the normal activities of the same species from other pools, as the calculation of enrichment takes all of them into account. Similarly, the multiple pools in the lung tissue sample may also be the reason of smaller gap of label enrichment between control and KO groups in lung tissue than in BALF samples. This is because PC species in BALF are mainly surfactant PC, while in lung tissue samples, additional pools of the same PC species with higher enrichment and fractional appearance rate can raise up the overall enrichment level.

In the analysis of PC species in BALF, fractional appearance rate was discussed to characterise the surfactant secretion process, and same fractional appearance rate was assumed to investigate the label enrichment of PC before and after secretion. Providing the multiple pools of the same species in lung tissue, it is not appropriate to assume the same turnover as well as the fractional appearance rate of the integrated PC pools. However, if the whole labelling process, from label infusion to the detection of a labelled PC species, is to be seen as an entire synthesis process regardless the intermediate phases, then all product PC species can be assumed to have the same initial precursors, the *methyl*$D_9$ substrate, provided they are synthesised in the same mouse under the same conditions. The same label precursor means the label enrichment of precursor pool is the same for all. The method of evaluating the fractional appearance rate can also be applied to characterise the fractional synthesis rate. Therefore, according to equation (14), if the initial enrichment of precursor pool is the same to all species, the initial change of label enrichment in product pool should be directly proportional to the fractional appearance rate. Using early time point enrichment level to represent the initial change of product pool enrichment, then the higher the early time point label enrichment is, the higher the fractional appearance rate will be. Consequently, the PC species in the high category in control group would probably have higher collective fractional rates of appearance/synthesis and the low category would have the lower ones.

Table 18 shows that the species with 18:0 or 16:0a chain have general lower rate of appearance in the pool than other species, indicated by the low enrichment level early time point. The low fractional appearance rate of species with 18:0 acyl is consistent with the result of the same species in BALF. As the surfactant PC from lung cells are subsequently secreted into the surfactant pool in the air space, which is collected as BALF

sample, the lower enrichment level in the lung cells may result in the lower enrichment in the product pool. Contrarily, the species with 16:0a alkyl, especially D9PC16:0a/16:1 has much lower rank of early time enrichment and fractional appearance rate in the lung tissue in contrast to its high rank in the BALF, which may indicate different pool sources or even different metabolism of the species from others.

### 4.4.2 Temporal trend of labelled PC species

In lung tissue samples, the time course enrichment of the species without 18:0 and 16:0a chains have a decreasing trend in both control and KO group, which indicates fast label incorporation of these species in the lung cells after label infusion, as there must be an increasing part starting from zero enrichment to reach certain enrichment level before going down, which may have happened before the first observation. The enrichment decreases as the labelled species leaving the lung tissues without enough newly synthesised labelled species coming in the pool. This enrichment pattern of labelled PC species in lung tissue is different from that in BALF, which slowly increases to the maximum and decreases gradually. Combining the two patterns may present a process that the newly labelled surfactant PC are firstly synthesised and remodelled in the lung cells, and then secreted into the alveolar space with time delay. The continuous supply of labelled species from lung cells results in the accumulation in surfactant monolayer where the species enrichment increases, and as the labelled species leaves the pool without new ones coming in, the enrichment goes down.

A feature in the time course curves of labelled PC without 18:0 and 16:0a chain is that the enrichment decreases rapidly between 5 to around 15 hour (Figure 28). Then the decrease slows down and the enrichment level stabilizes for a period of time until around

24 hour, followed by another decreasing trend. The stable state period from 15 to 24 hour should be paid more attention as it should not be the case in a bolus infusion experiment if the labelled species only leave the system without coming back. Therefore, there may probably be a new flood of labelled PC species coming into the pool during or even before this period. It is also possible to see from the figure that the label incorporation of characteristic lung surfactant species in KO group are more affected than other species, where the curves of characteristic surfactant species in KO group lie far away from the control group, but those of non-characteristic species are closer to control group. This may be due to the multiple pools combined in lung tissue samples as explained before.

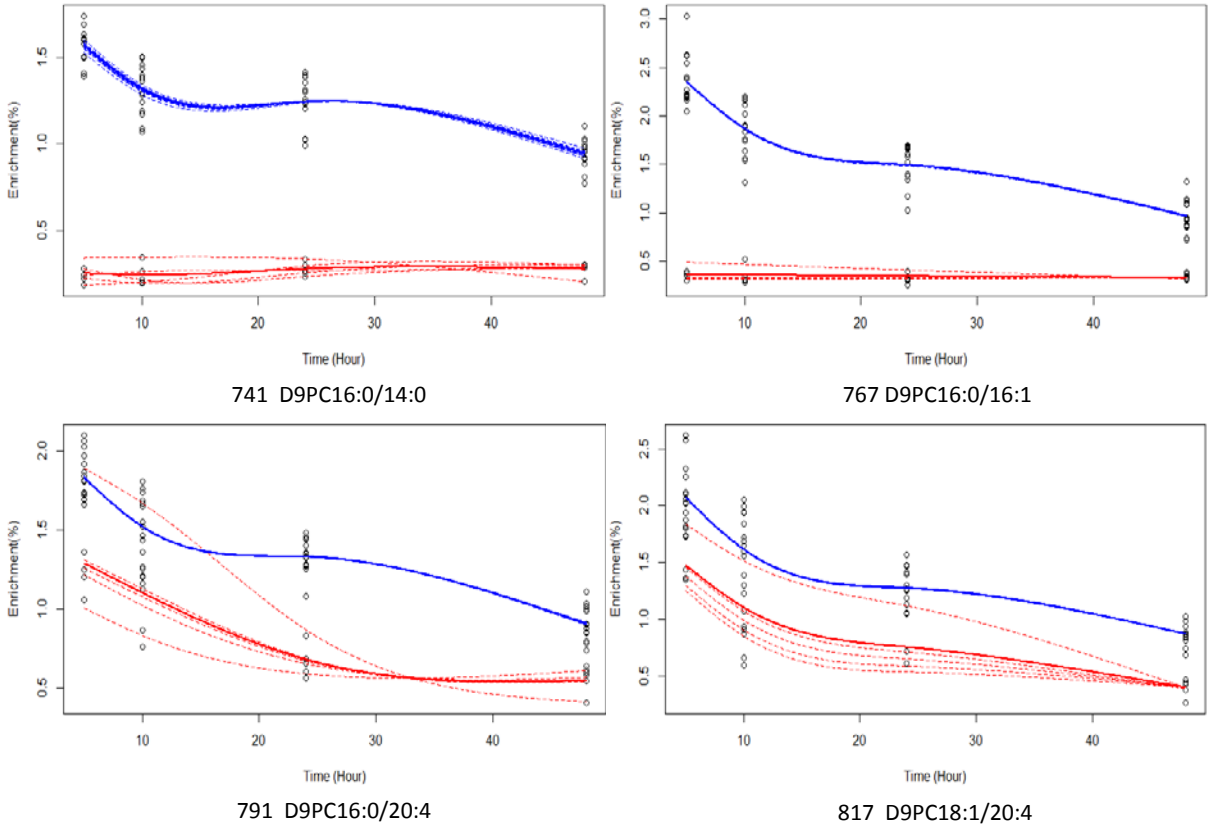



Figure 28: Time course enrichment curves of four *methyl*$D_9$-PC species in lung tissue samples of control (in blue) and KO (in red) mice models. The species have similar time trend over the 48 hour period. The label incorporation of characteristic surfactant species D9PC16:0/14:0 and D9PC16:0/16:1 in KO group are further away from the control group than that of the other two non-characteristic surfactant species.

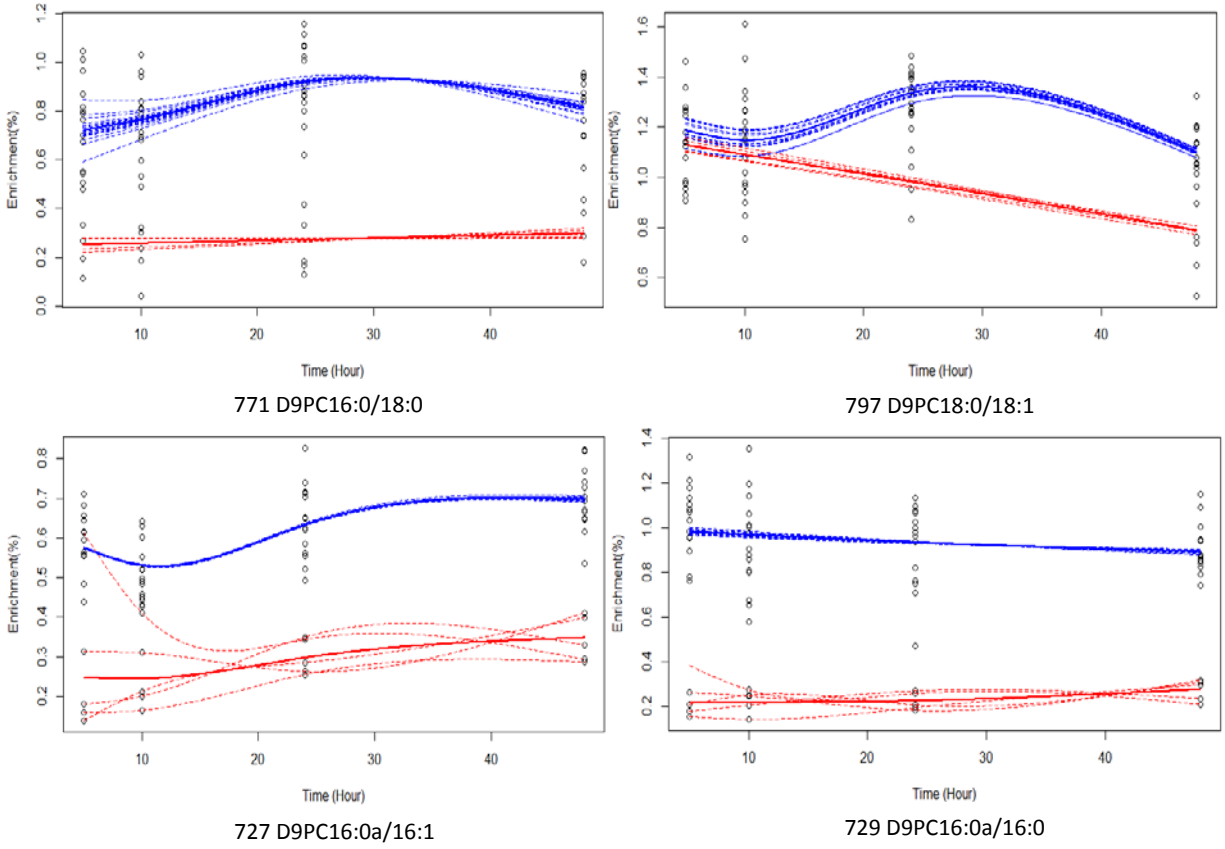



Figure 29: Time course model of D9PC16:0/18:0, D9PC18:0/18:1, D9PC16:0a/16:1 and D9PC16:0a/16:0, which have different enrichment patterns from the other species shown in Figure 28.

There are some species with different curve shapes to the others (Figure 29). Species PC16:0/18:0 has a low incorporation and the enrichment increases towards 30 hour time point, which is consist with that in BALF. Another species with the 18:0 acyl chain, PC18:0/18:1 exhibits a different time trend, but it also has a significant increasing part towards 30 hour. Within the same experiments, different shapes of time trend of species enrichment to other species in the same compartment could be due to that the species have different source pools in lung tissue, or that they metabolise differently. Interestingly, apart from these two species, others with 18:0 do not have a significant difference from the species without it, suggesting that species with 18:0 may not be from the same source or metabolise differently. In addition, species with 16:0a have low

incorporation and different time trend from other species. They do not have a decreasing trend over the time period, and species PC16:0a/16:1 even has an increasing part from 10 hour towards 40 hour, which may indicate different metabolism from other species.

### 4.4.3 Group comparison and biomarker analysis

Comparisons of time course enrichment of each labelled species between control and KO groups were carried out using SME method. The *L, se* and *Ft* values are listed in Table 19, and the species are sorted in descending order of their *Ft* values. The time course enrichment of D9PC16:0/16:1 marks the largest distance between the two group means, with an *L* value of 7.670, and is ranked at the second place with a high *Ft* value of 71.2798. Some other surfactant characteristic species also have comparatively higher *L* and *Ft* values, including D9PC16:0/14:0, D9PC16:0a/16:0 and D9PC16:0/16:0.

| *m/z* | Species | *L* | *se* | *Ft* |
|---|---|---|---|---|
| 815 | D9PC16:0/18:0 | 3.884709 | 0.04981282 | 77.98612 |
| 741 | D9PC16:0/16:1 | 7.670280 | 0.107608 | 71.27983 |
| 727 | D9PC16:0/14:0 | 6.182472 | 0.1151509 | 53.69019 |
| 743 | D9PC16:0a/16:0 | 4.576028 | 0.1065365 | 42.95267 |
| 769 | D9PC16:0/16:0 | 6.430485 | 0.2182161 | 29.46843 |
| 795 | D9PC18:0/18:1 | 2.152155 | 0.09407033 | 22.87814 |
| 767 | D9PC16:0/18:2 | 5.470157 | 0.2395183 | 22.83816 |
| 843 | D9PC16:0/22:6 | 4.999620 | 0.3224717 | 15.50406 |
| 771 | D9PC16:0/18:1 | 4.836043 | 0.3460205 | 13.97618 |
| 715 | D9PC16:0a/16:1 | 2.193747 | 0.1783389 | 12.30100 |
| 797 | D9PC18:0/22:6 | 3.149484 | 0.2615119 | 12.04337 |
| 729 | D9PC16:0/20:4 | 3.677153 | 0.4655047 | 7.899281 |
| 817 | D9PC18:1/20:4 | 3.372458 | 0.5335027 | 6.321352 |
| 819 | D9PC18:0/18:2 | 2.784551 | 0.4544229 | 6.127664 |
| 793 | D9PC18:0/20:4 | 2.102729 | 0.3528222 | 5.959741 |
| 791 | D9PC18:1/18:2 | 2.033169 | 0.5399264 | 3.765641 |

Table 19: *Ft* statistics for lung tissue PC specie of control and KO groups.

Unexpectedly the species in the first place is D9PC16:0/18:0, which has very low label incorporation over the period. Although the difference between the two groups is small, with the *L* value of just 3.8847, the species has an extremely small *se* value of 0.0498, suggesting very small variations among individual mice in the two groups. This is illustrated in the plots of the species in Figure 29, where the fitted curves of individual mouse are more closely around the group mean especially in the KO group, comparing with other species.

It is obvious that the *Ft* calculated between control and KO groups in lung tissue are not as high as those in BALF, and many of them are under 15, indicating the temporal enrichment of these species are less significant in lung tissue samples than in BALF. The previous comparison analysis of BALF D9DPPC between age groups of control mice generated comparatively small *Ft* values ranging between 5 and 15, and the *Ft* values calculated for control and KO groups have values from 40 to over 160. Using these ranges as standards, the fact that the *Ft* values calculated for many PC species in lung tissues being under 15, suggests the low viability of their temporal enrichment in differentiating KO mice from control. The small number of species ranking at the top of the list, with higher *Ft* values over 40, can be considered as potential markers for distinguishing the two groups in future studies.

## 4.5 Characterisation of time course *methyl*$D_9$ - PC enrichment in BALF of control and high dose antibody treated mice groups and biomarker analysis

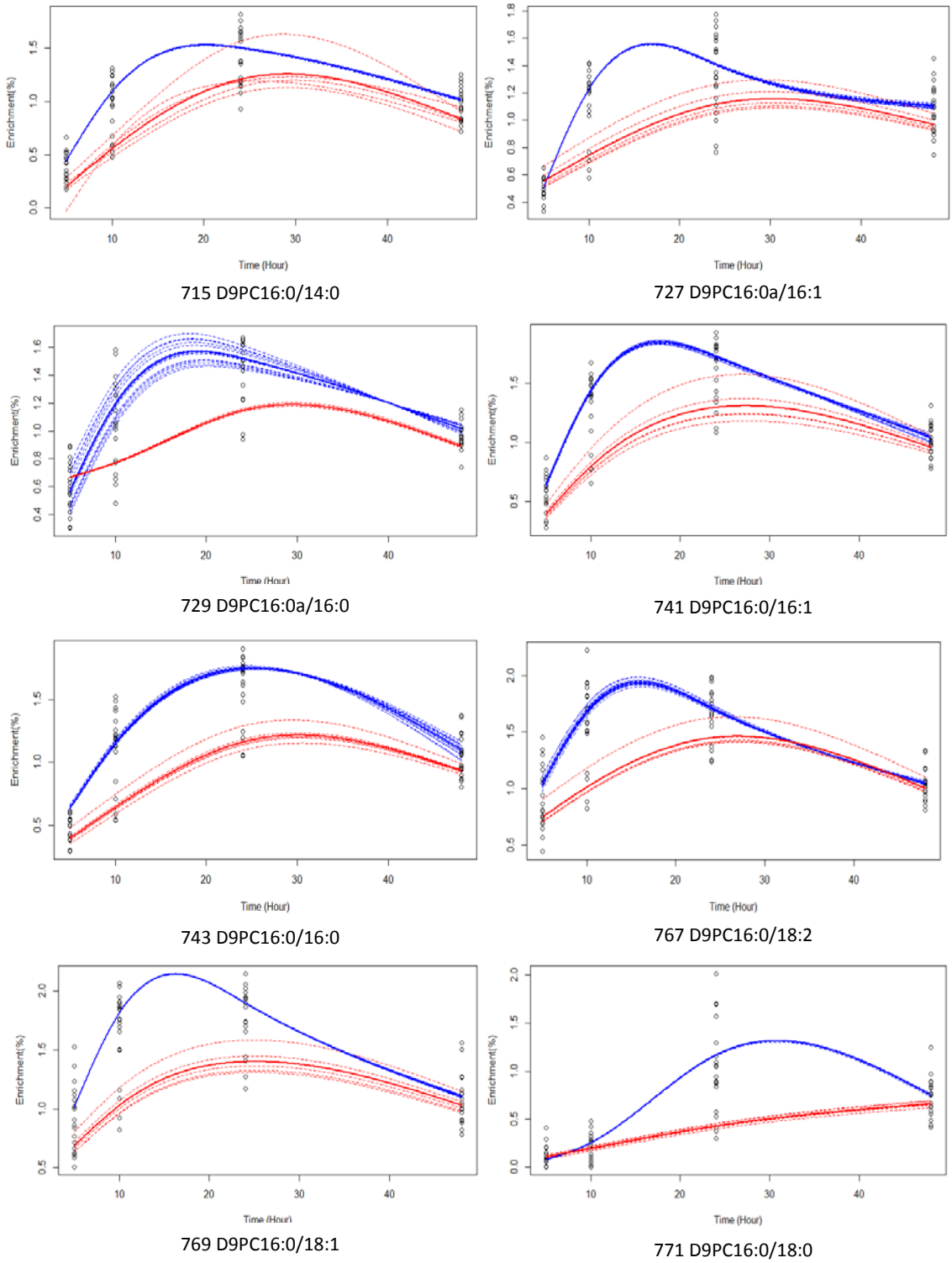
The last two sections compared time course enrichment of labelled PC species in BALF and lung tissue between control and KO groups facilitated by SME method. To further assess and validate the methodology, in this section, the time course PC enrichment of a

group of 169 days mice, which were treated with high dose antibody drug, were modelled and compared with the control group. The study conducted in Southampton has shown changes in PC turnover of this group of mice. This analysis will further verify the result, and explore possible dynamic biomarkers. Moreover, as the antibody drug has a milder effect of accumulation of surfactant PC species than that of PAP disease, the difference between control and high dose groups should be theoretically smaller than that between control and KO groups analysed in the last section. Therefore, performing SME analysis on them is helpful to test if the statistical method is sensitive in identifying different degrees of group variation.

The SME modelling shows that, in this high dose group, the time course enrichment of labelled PC in BALF has some similar features with the control group. The temporal trend is similar, and the ranking and categorization of the species are similar as well. Species with a 16:0 acyl chain and an unsaturated longer acyl chain, including D9PC16:0/18:1, D9PC16:0/18:2, D9PC16:0/22:6 and D9PC16:0/20:4 have the highest enrichment in the high dose group. Species with 18:0 acyl are still the ones with the lowest enrichment levels.

The plots of time models (Figure 30) show that the label incorporation of most of the species follow an increasing – decreasing trend during the 48 hours after label infusion. High dose drug treated mice group have generally lower enrichment levels over the time comparing with control group. This is consistent with what is expected that the antibody treated mice may have impaired degradation of surfactant PC species, which increases the surfactant pool, and results in a different temporal enrichment trend comparing with the control mice. Nonetheless, the difference between the groups are much smaller than

that between control and KO groups, suggesting a smaller impact from the antibody drug than the PAP on the surfactant PC turnover of mice.




715 D9PC16:0/14:0

727 D9PC16:0a/16:1

729 D9PC16:0a/16:0

741 D9PC16:0/16:1

743 D9PC16:0/16:0

767 D9PC16:0/18:2

769 D9PC16:0/18:1

771 D9PC16:0/18:0

Figure 30: Time course enrichment curves of *methyl*$D_9$-PC species for both control group (in blue) and high dose treated mice group (in red) in BALF.

It appears that the highest enrichment levels of many species are reached at later time points in high dose than in control mice. Especially for D9PC16:0/14:0, D9PC16:0a/16:1, D9PC16:0a/16:0, D9PC16:0/16:1, D9PC16:0/16:2, D9PC16:0/18:2 and D9PC16:0/18:1, the group difference at the initial increasing part makes the major contribution to the total difference between the two groups over time. As stated above, the species with 18:0 acyl in the high dose treated mice have lower label incorporation than other species, which is the same with control mice. Species D9PC16:0/18:0 in high dose group has an enrichment trend that increases slowly in a linear way, which is contrast to that of control group. Moreover, the species has the lowest time course enrichment among all the species analysed.

*Ft* values are calculated for evaluating the differences between control and high dose group recognised by the temporal enrichment of labelled PC. Table 20 lists the sorted species on descending order of their *Ft* values. Again, D9PC16:0/18:0 turns out unexpectedly to be the most significant species with the highest *Ft*, owing to its greater *L* and smaller s*e* values. As mentioned above, D9PC16:0/18:0 has the lowest temporal enrichment both in control and high dose mice in BALF sample. The difference is mainly owing to the distinguished temporal trend of the two groups.

The *Ft* for control vs high dose group are generally much smaller than those calculated for control vs KO group. This suggests that the differences of time course label incorporation of the *methyl*$D_9$-PC species between control and high dose antibody treated mice are much smaller than those between control and KO group, which is consistent with the current understanding that the antibody has milder impact on surfactant turnover than the PAP disease. This also demonstrates the SME method is effective in quantifying and scaling the difference between groups. Apart from D9PC16:0/18:0, the *Ft* values of the

other species are not very high. Nevertheless, it is still worth to choose the species ranked higher in the list for further investigation on potential markers.

| *m/z* | Species | *L* | *se* | *Ft* |
|---|---|---|---|---|
| 771 | D9PC16:0/18:0 | 3.53755 | 0.06748898 | 52.4167 |
| 729 | D9PC16:0a/16:0 | 2.252287 | 0.1030562 | 21.85494 |
| 795 | D9PC18:0/18:2 | 0.8717616 | 0.0406007 | 21.47159 |
| 791 | D9PC16:0/20:4 | 2.366708 | 0.124248 | 19.04825 |
| 743 | D9PC16:0/16:0 | 3.141018 | 0.1653191 | 18.99973 |
| 769 | D9PC16:0/18:1 | 3.375497 | 0.2612953 | 12.91833 |
| 843 | D9PC18:0/22:6 | 1.990326 | 0.1691978 | 11.76331 |
| 767 | D9PC16:0/18:2 | 2.547251 | 0.230024 | 11.07385 |
| 727 | D9PC16:0a/16:1 | 2.175126 | 0.2106802 | 10.3243 |
| 819 | D9PC18:0/20:4 | 1.599397 | 0.1781391 | 8.978359 |
| 741 | D9PC16:0/16:1 | 2.824162 | 0.3271821 | 8.631773 |
| 817 | D9PC18:1/20:4 | 3.141313 | 0.4067864 | 7.722268 |
| 815 | D9PC16:0/22:6 | 2.422177 | 0.3809698 | 6.357925 |
| 715 | D9PC16:0/14:0 | 2.251787 | 0.4029425 | 5.588357 |
| 793 | D9PC18:1/18:2 | 3.594759 | 0.6737384 | 5.335541 |
| 797 | D9PC18:0/18:1 | 1.604938 | 0.5262081 | 3.050007 |

Table 20: *Ft* statistics for BALF PC of control and high dose drug treated mice groups.

## 4.6 Characterisation of time course *methyl*$D_9$ - PC enrichment in lung tissue of control and high dose treatment mice groups and biomarker analysis

PC label incorporation in lung tissue samples are analysed for the high dose group mice in addition to that in BALF for a deeper insight of surfactant PC turnover. The SME analysis suggests the labelled PC time course enrichment of high dose treated mice are very similar to the control mice, and the categories of the species are similar too. D9PC16:0/16:1, D9PC16:0/18:2 and D9PC16:0/22:6 are the top three species that have the highest time course label enrichment level. D9PDC16:0a/16:1, D9PC16:0/18:0, and D9PC16:0a/16:1 are the three species with the lowest enrichment in the lung samples

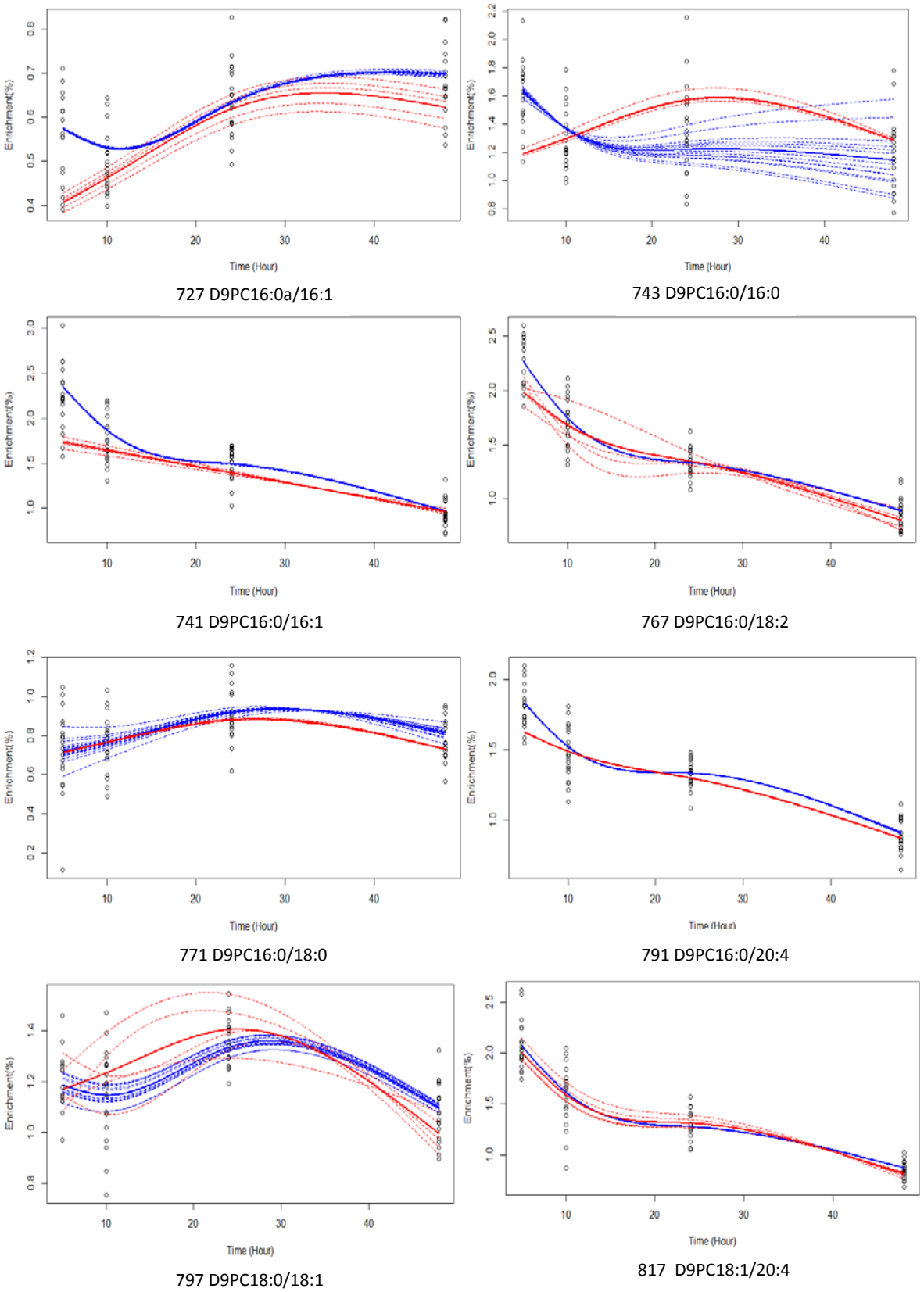



Figure 31: Time course enrichment curves for *methyl*$D_9$-PC species in lung tissue of control (in blue) and high dose antibody drug treated mice (in red).

of high dose group.

The enrichment curves of most species have the similar temporal enrichment patterns except a few species with 18:0 and 16:0a acyl. Besides, the species of the high dose group are very similar to the control group in terms of the curve shapes and enrichment levels, although there are still some exceptions shown in Figure 31. The fitted model of D9PC16:0/16:0 of the control mice group has a decreasing trend over the whole period, whereas that of the high dose group displays an increasing trend at the beginning and decreasing after reaching the maximum at around 27 hour. The curve of species D9PC16:0a/16:1 of control group decreases at the early time point and later lifts up to reach the highest point after 30 hour, whereas the curve of the high dose group increases all the way and reaches the maximum at around 30. This feature also happens to the time model of species D9PC18:0/18:1. The reason of these abnormal features are not clear and should be further investigated.

Comparison of the time models between control and high dose group were performed by calculating the *L*, *Se* and *Ft*, listed in Table 21. The top three species with high *L* values greater than 1 include D9PC16:0/16:0, D9PC16:0/14:0 and D9PC16:0/16:1. However, they are only ranked at the $6^{th}$, $2^{nd}$, and $4^{th}$ places due to the comparatively big *se* values.

D9PC16:0/24:0 has a small *L* value, indicating small difference between the two groups, but due to the extremely small *se*, it gets the highest *Ft* and ranks in the first place. However, the plot (Figure 31) has shown that models for this species in both groups are very close to each other, therefore the result of highest *Ft*, which indicates high capability in differentiating the two groups does not look straightforward. This should be

investigated further as to if the modelling of the species is appropriate, such as if the model is over smoothed, which can results in extremely small variance term.

Overall, the comparison between control and high dose groups has not identified as many viable species for discriminating groups as the comparisons between control and KO groups, with very small *Ft* values of most species. Nevertheless, the species with highest *Ft* values, including D9PC16:0/20:4, D9PC16:0/14:0 and D9PC16:0a/16:0 should be checked further as potential biomarker candidates.

| *m/z* | Species | *L* | *se* | *Ft* |
|---|---|---|---|---|
| 791 | D9PC16:0/20:4 | 0.4347934 | 0.00645087 | 67.40078 |
| 715 | D9PC16:0/14:0 | 1.144752 | 0.02560512 | 44.70796 |
| 729 | D9PC16:0a/16:0 | 0.7866485 | 0.02102845 | 37.40878 |
| 741 | D9PC16:0/16:1 | 1.127695 | 0.06496118 | 17.35952 |
| 793 | D9PC18:1/18:2 | 0.5634926 | 0.03941952 | 14.29476 |
| 743 | D9PC16:0/16:0 | 1.830349 | 0.1917868 | 9.543662 |
| 795 | D9PC18:0/18:2 | 0.3605815 | 0.03900632 | 9.244183 |
| 771 | D9PC18:0/16:0 | 0.3388316 | 0.03861647 | 8.774277 |
| 843 | D9PC18:0/22:6 | 0.438177 | 0.05632649 | 7.779236 |
| 815 | D9PC16:0/22:6 | 0.8427376 | 0.1279054 | 6.588759 |
| 819 | D9PC18:0/20:4 | 0.2415565 | 0.04138286 | 5.837116 |
| 727 | D9PC16:0a/16:1 | 0.3704806 | 0.07703749 | 4.809095 |
| 769 | D9PC16:0/18:1 | 0.4831807 | 0.1686932 | 2.864258 |
| 797 | D9PC18:0/18:1 | 0.4918722 | 0.252976 | 1.944344 |
| 767 | D9PC16:0/18:2 | 0.4982655 | 0.2843049 | 1.752574 |
| 817 | D9PC18:1/20:4 | 0.2012468 | 0.1480195 | 1.359596 |

Table 21: *Ft* statistics for lung tissue PC of control and high dose drug treated mice.

## 4.7 Conclusions

In this chapter, time course *methyl*$D_9$-PC label enrichment modelling and biomarker discovery analysis were carried out for control mice, GMCSF beta receptor knock out mice

and high dose antibody drug treated mice. The comparison was performed between individual age groups of control mice, between control and KO groups, as well as between control and high dose group. Sixteen *methyl*$D_9$ labelled PC species in both BALF and lung tissue samples were analysed, and the SME method is used for the time series data modelling, group comparison, as well as identification of species that are most significant in distinguishing the groups.

Time course enrichment of D9DPPC in BALF of mice models was analysed in detail to set up the methodology for the following analysis. The temporal models of multiple species have revealed different incorporation patterns of species with 18:0 both in BALF and lung tissue pools. Many of the BALF PC species are significant in distinguishing the control and KO groups, and species with 16:0 acyl in BALF are generally more significant than other species. PC label incorporation in BALF sample is more capable than that in lung tissue in identifying two compared groups as BALF PC species have generally higher *Ft* values than those lung tissue. Temporal enrichment of the species are not as significant in distinguishing control and high dose drug treated mice in both BALF and lung tissue, and only a few species have the potential to be used as biomarkers. Besides, the difference between the PC label incorporation of control group and KO group is greater than that between control group and high dose group, suggesting the high dose antibody drug does induce some side effects on the healthy mice in terms of surfactant PC turnover, but the effect is smaller than that of PAP disease. All these results are consistent with the current understanding as well as assumptions of surfactant metabolism. The analysis demonstrates it is viable to use time course enrichment of *methyl*$D_9$ labelled surfactant PC species, which reflects the PC species metabolism, as potential biological markers for identifying mice with PAP diseases from healthy ones.

SME method works well in modelling time course multivariate data. It is good at dealing with data with small sample size and missing values, which are common problems in biological studies. The method is sensitive and effective in identifying group difference as well as quantifying and scaling the differences, which is critical for comparisons among large numbers of variables for biomarker discovery analysis. An issue with the method is that there are a couple of models that might be possibly over smoothed, which needs to be investigated further as to if the results are reliable. Another issue with the SME method is that, as there is no theoretical probability distribution for the *Ft* statistic under the null hypothesis of no difference, it is not possible to do a standard statistical test to evaluate its significance. However, in this chapter, the comparison of D9DPPC between three different age groups, as well as between control and KO groups provided a general standard of the *Ft* ranges which were referred by the following analysis. Therefore, despite the lack of theoretical distribution for a significance test, this chapter sets good reference ranges for *Ft* for the analysis in the following chapters.

# Chapter 5 Surfactant and hepatic phosphatidylcholine dynamic biomarker discovery analysis for NRDS patients

## 5.1 Introduction

In the last chapter, time course enrichment modelling of surfactant PC and biomarker analysis were carried out for mice models, which has established the methodology for dynamic PC biomarker analysis. As ESI MS/MS coupled with stable isotope labelling technique is capable of facilitating investigation of *in vivo* kinetics of molecules in human, in this and the following chapters, *methyl*$D_9$ labelled PC species are analysed for human subjects, for characterising human surfactant and hepatic PC turnover as well as investigating the feasibility of using them as biomarkers in less clear clinical scenarios.

NRDS is a primary lung disease caused by lung immaturity, where the AT II cells are underdeveloped and unable to synthesise pulmonary surfactant. All NRDS patients are given surfactant supplement at birth, which has largely improved the mortality rate. However, the survived babies have high probability of getting long term lung conditions in later life. It will be beneficial to be able to identify patient groups in terms of disease severity, tendency of developing lung diseases in the long term, and response to the treatment, for the application of early interventions to prevent bad outcomes as well as for developing new treatment. All of these require effective biomarkers, which will also be helpful in understanding the underlying mechanisms of different biological conditions.

The NRDS infants in the study are categorised into two groups based on disease severity, indicated by on or off ventilation five days after label infusion. The grouping aims to check if PC *in vivo* activity can indicate any difference between the two groups of patients with

different disease severity, and predict outcomes as biomarkers. Time course enrichment of labelled surfactant PC species as well as hepatic PC species were modelled and analysed. Assessing hepatic PC species can provide additional view to the surfactant PC analysis regarding liver function, as both of them are integrated into the whole body PC metabolism and are interacted with each other.

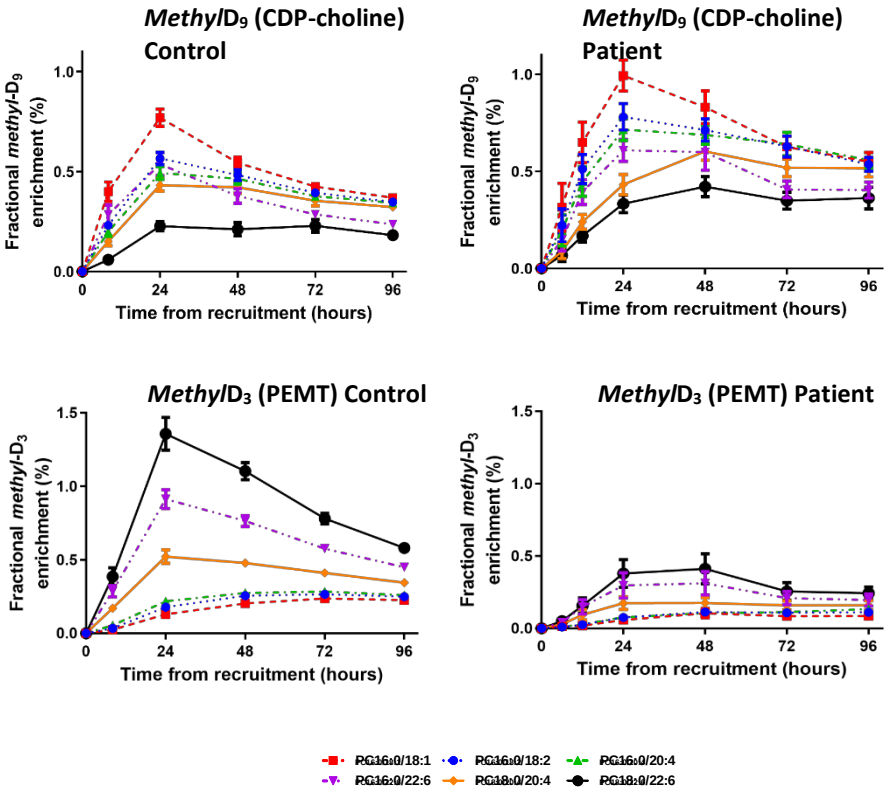



Figure 32: Enrichment levels of liver synthesised *methyl*$D_9$-PC through CDP-choline pathway and *methyl*$D_3$-PC through PEMT pathway in plasma samples of healthy people and ARDS patients at six time points. In each graph, an enrichment line is obtained by connecting the mean values of observations of a group at six time points. The enrichment levels of the *methyl*$D_9$-PC species synthesised through CDP-choline pathway are slightly lower in control group than in ARDS patients group. Whereas *methyl*$D_3$-PC synthesised through PEMT pathway are significantly higher in control group (Dushianthan *et al.*, 2018).

In liver, PC species can be synthesised through both CDP-choline pathway and PEMT pathway. Dushianthan *et al.* (2018) has found that label incorporation of PC synthesised through CDP-choline pathway in liver are similar between healthy people and ARDS patient, but that of PC synthesised through PEMT pathway are significantly different between the two groups (Figure 32), suggesting that the label incorporation of hepatic PC via PEMT pathway are more indicative in ARDS disease than those via CDP-choline pathway. Therefore, in this chapter, the PEMT pathway PC were assessed along with those of pulmonary surfactant PC synthesised through CDP-choline pathway to examine if they can characterise different stages of NRDS disease and predict outcomes.

### 5.1.1 Project description

The data analysed in this chapter is obtained in the form of raw MS datasets from a previous study carried out in Southampton, named as "The Surfactant, Nutrition and Microorganism Interaction study in babies at risk of developing neonatal chronic lung disease (TsuNaMI)". The study was designed to identify and correlate risk factors for development of neonatal chronic lung disease (nCLD) in preterm infants. The preterm babies are recruited from Princess Anne Hospital, Southampton. They all received infusion of *methyl*$D_9$ choline chloride in dextrose for three hours and had sequential blood and endotracheal aspirate (ETA) samples collected. Those infants still ventilated five days after label infusion received a second infusion of *methyl*$D_9$ choline chloride.

The samples are processed following the TsuNaMI Sample Processing Protocol. As described in their report, blood samples were centrifuged and the upper plasma layer was aspirated. Lipid extraction was performed using Bligh and Dyer method (Bligh and Dyer, 1959). Plasma and ETA were added with methanol and chloroform, and then

centrifuged. The supernatant was carefully drawn off and added with distilled water and chloroform to be centrifuged again, after which the lower layer was carefully aspirated and dried for the subsequent ESI MS/MS analysis. Precursor scans P184, P187 and P193 were performed to monitor endogenous surfactant PC molecules, *methyl*$D_3$ labelled PC molecules and *methyl*$D_9$ labelled PC molecules.

### 5.1.2 Data description and methods

In the clinical study, five days after first label infusion was used as a cut-off time point, and the infants were categorized into two groups based on their status at the time point, i.e. on or off ventilation. The labelled PC enrichment data of the two groups before the time points are analysed in the following analysis.

For ETA samples, the time course enrichment of sixteen *methyl*$D_9$ labelled PC molecules were modelled for probing the lung surfactant PC kinetics. These species were selected based on the confidence of existence of the endogenous species and deuterium labelled species based on the MS analysis result. Restricted by the availability of samples, there are PC observations of twelve NRDS patients in off ventilation group and eleven patients in on ventilation group.

The time course enrichment of *methyl*$D_3$ labelled form of the same sixteen PC molecules in plasma samples were also analysed for assessing the hepatic PC turnover through PEMT pathway in liver. There were observations of seventeen patients in off ventilation group and thirteen patients in on ventilation group. Observations at seven time points were included for the time course PC kinetic study, including 6, 12, 24, 48, 72, 96 and 120 hour after initial label infusion.

MS data processing and calculation in this chapter were all performed by the newly developed LipidomeLabelling platform. The procedures conducted include raw data extraction, background removal, peak identification, peak integration, species identification, isotope correction, between scan isotope correction, and enrichment calculation. The multivariate time course modelling and biomarker discovery analysis were performed using SME method facilitated by the SME R package.

## 5.2 Characterisation of time course *methyl*$D_9$ - PC enrichment in ETA and comparison between off and on ventilation NRDS patients

The model fitting for the *methyl*$D_9$ labelled PC time course enrichment in ETA samples was conducted using SME method and the fitted values at seven time points are listed in Table 22. The result shows that, the off ventilation group tend to have higher group means than the off ventilation group for most PC species at the early time of the whole period, whereas the situation changes after certain time point and the mean values of on ventilation group become higher towards the end of the period. The enrichment levels of the labelled PC species of the off ventilation group are not much different, although the ones with the highest masses, especially those with 20:4 and 20:6 acyl chains are comparatively higher. The D9PC16:0a/16:0 has an exceptional high enrichment at the end of the period, while other characteristic surfactant species are comparatively low. In the on ventilation group, D9PC18:0/22:6 has the highest enrichment level over time.

Table 22: Fitted group mean values of time course enrichment of *methyl*$D_9$-PC in ETA samples of the off and on ventilation NRDS patient groups.

| Mass | Species Assignment | Group | Enrichment (%) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | 6h | 12h | 24h | 48h | 72h | 96h | 120h |
| 715 | D9PC 16:0/14:0 | **Off Venti** | **0.006** | **0.015** | **0.058** | **0.157** | **0.209** | **0.236** | **0.239** |
| | | On Venti | 0.000 | 0.022 | 0.069 | 0.207 | 0.373 | 0.393 | 0.392 |
| 727 | D9PC16:0a/16:1 | **Off Venti** | **0.114** | **0.126** | **0.152** | **0.203** | **0.254** | **0.304** | **0.355** |
| | | On Venti | 0.031 | 0.064 | 0.131 | 0.257 | 0.351 | 0.407 | 0.449 |
| 729 | D9PC16:0a/16:0 | **Off Venti** | **0.011** | **0.043** | **0.108** | **0.235** | **0.358** | **0.477** | **0.592** |
| | | On Venti | 0.007 | 0.024 | 0.060 | 0.179 | 0.336 | 0.356 | 0.367 |
| 741 | D9PC16:0/16:1 | **Off Venti** | **0.012** | **0.039** | **0.093** | **0.188** | **0.249** | **0.286** | **0.304** |
| | | On Venti | 0.003 | 0.043 | 0.104 | 0.276 | 0.525 | 0.466 | 0.474 |
| 743 | D9PC16:0/16:0 | **Off Venti** | **0.005** | **0.022** | **0.055** | **0.118** | **0.169** | **0.209** | **0.242** |
| | | On Venti | 0.000 | 0.021 | 0.049 | 0.190 | 0.389 | 0.430 | 0.414 |
| 767 | D9PC16:0/18:2 | **Off Venti** | **0.012** | **0.042** | **0.101** | **0.191** | **0.238** | **0.278** | **0.316** |
| | | On Venti | 0.008 | 0.037 | 0.098 | 0.265 | 0.453 | 0.464 | 0.444 |
| 769 | D9PC16:0/18:1 | **Off Venti** | **0.022** | **0.051** | **0.107** | **0.203** | **0.258** | **0.281** | **0.287** |
| | | On Venti | 0.019 | 0.050 | 0.114 | 0.305 | 0.525 | 0.510 | 0.496 |
| 771 | D9PC16:0/18:0 | **Off Venti** | **0.010** | **0.026** | **0.060** | **0.128** | **0.177** | **0.207** | **0.228** |
| | | On Venti | 0.002 | 0.029 | 0.066 | 0.181 | 0.319 | 0.401 | 0.397 |
| 791 | D9PC16:0/20:4 | **Off Venti** | **0.019** | **0.046** | **0.098** | **0.192** | **0.254** | **0.287** | **0.301** |
| | | On Venti | 0.010 | 0.049 | 0.128 | 0.347 | 0.588 | 0.560 | 0.537 |
| 793 | D9PC18:1/18:2 | **Off Venti** | **0.015** | **0.044** | **0.102** | **0.207** | **0.269** | **0.283** | **0.269** |
| | | On Venti | 0.017 | 0.051 | 0.124 | 0.342 | 0.560 | 0.546 | 0.525 |
| 795 | D9PC18:0/18:2 | **Off Venti** | **0.031** | **0.071** | **0.148** | **0.261** | **0.308** | **0.317** | **0.308** |
| | | On Venti | 0.013 | 0.072 | 0.134 | 0.326 | 0.597 | 0.523 | 0.532 |
| 797 | D9PC18:0/18:1 | **Off Venti** | **0.025** | **0.056** | **0.116** | **0.216** | **0.264** | **0.277** | **0.275** |
| | | On Venti | 0.006 | 0.058 | 0.095 | 0.231 | 0.490 | 0.425 | 0.427 |
| 815 | D9PC16:0/22:6 | **Off Venti** | **0.037** | **0.061** | **0.153** | **0.271** | **0.300** | **0.352** | **0.396** |
| | | On Venti | 0.025 | 0.079 | 0.192 | 0.471 | 0.696 | 0.600 | 0.578 |
| 817 | D9PC18:1/20:4 | **Off Venti** | **0.043** | **0.097** | **0.201** | **0.342** | **0.413** | **0.435** | **0.417** |
| | | On Venti | 0.007 | 0.066 | 0.188 | 0.444 | 0.634 | 0.648 | 0.592 |
| 819 | D9PC18:0/20:4 | **Off Venti** | **0.026** | **0.054** | **0.110** | **0.204** | **0.260** | **0.290** | **0.306** |
| | | On Venti | 0.014 | 0.054 | 0.134 | 0.291 | 0.469 | 0.490 | 0.494 |
| 843 | D9PC18:0/22:6 | **Off Venti** | **0.071** | **0.103** | **0.162** | **0.246** | **0.291** | **0.329** | **0.367** |
| | | On Venti | 0.046 | 0.098 | 0.182 | 0.482 | 0.838 | 0.568 | 0.724 |

Figure 33 for the time course enrichment models of labelled PC shows that for most species, the on ventilation group in red lines has wider range of enrichment levels and more diverse individual curves than the off ventilation group. The substantially high enrichment levels of some ventilated patients account for the overall greater mean curve of this group than the off ventilation group.

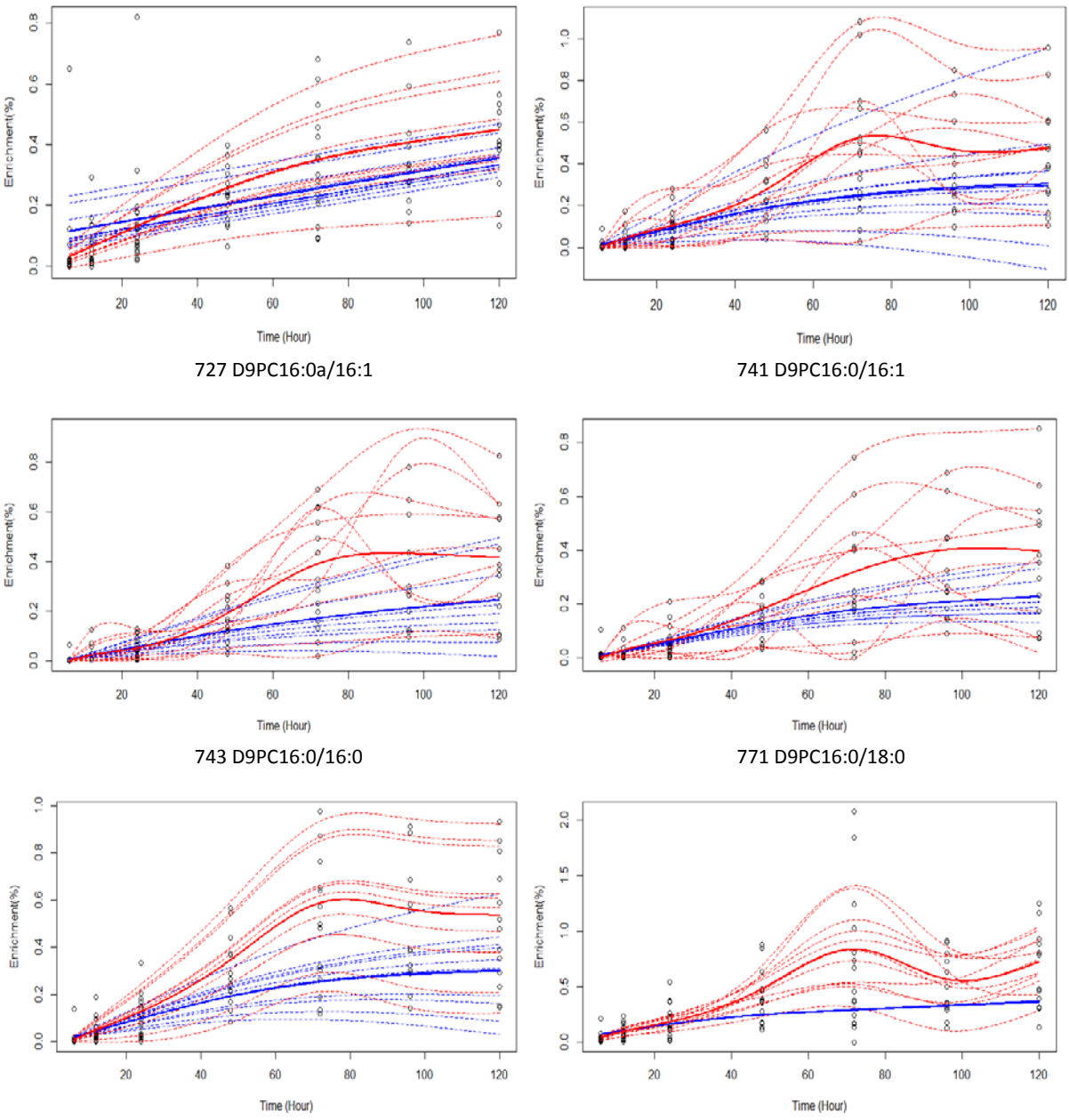



Figure 33: Time course enrichment of some *methyl*$D_9$-PC in ETA samples of the off ventilation (in blue) and on ventilation (in red) NRDS patients.

The temporal curves of species of on ventilation group slowly increase from the starting point and reach the highest enrichment level at around 72 hour, and then gradually decrease towards the end of the period, except for some species picking up slightly at the

end. The maximum enrichment of D9PC16:0/18:0 is estimated to be reached at around 96 hour, which is different from most of other species. Another exception is D9DPPC16:0a/16:1, which has an increasing trend throughout the period. The most varied individual curve shapes are found in the model of the species D9PC16:0/16:1, D9PC16:0/16:0 and D9PC16:0/18:0, suggesting different turnover patterns among these ventilated babies. Moreover, in these species, the curves of many individuals have more than one peak within the time frame, which can imply either more waves of labelled species coming into the system, or the change of the pool size.

The PC species of the off ventilation babies exhibit flat and lower temporal enrichment trend with small gradient and there are hardly any stationary point on the curves. It is important to note that the data points of the off ventilation group suffers from a lack of observations after 72 hour, with only three or four observations for each time point. As the babies recovered and went off ventilation, ETA samples were not collected, so there was no observation available after the time point. Consequently, it is possible that the insufficient data may only represent part of the population.

The group comparison and biomarker analysis were conducted using SME method and the *Ft*, *L* and *se* values are listed in Table 23. The distance between the two groups is quite small, indicated by the *L* values. Among all the species, D9PC18:0/22:6 has the highest *L* value of 3.11, followed by D9PC16:0/22:6 of 2.42, and D9PC16:0/20:4 and 2.24, suggesting the differences between group means of these species are higher than others with *L* values range between 1 and 2. The *se* values are comparatively large with values between 0.4 and 0.9. The individual models show diverse incorporation trends among individuals within each group, particularly the on ventilation group with largely varied individual curves, which contributes to the variance term *se*. Due to the generally small *L*

values and large *se* values, the *Ft* values are generally small. The highest *Ft* with the value of 7.28 belongs to D9PC18:0/22:6, owing to its biggest *L* and smallest *se* value.

| *m/z* | Species | *L* | *se* | *Ft* |
|---|---|---|---|---|
| 843 | D9PC18:0/22:6 | 3.112788 | 0.4276522 | 7.278784 |
| 729 | D9PC16:0a/16:0 | 1.696044 | 0.4765379 | 3.559095 |
| 767 | D9PC16:0/18:2 | 1.682209 | 0.5436523 | 3.094274 |
| 815 | D9PC16:0/22:6 | 2.419692 | 0.8090925 | 2.990625 |
| 791 | D9PC16:0/20:4 | 2.241368 | 0.7919274 | 2.83027 |
| 727 | D9PC16:0a/16:1 | 1.725831 | 0.6294649 | 2.741742 |
| 793 | D9PC18:1/18:2 | 1.650062 | 0.6069545 | 2.718592 |
| 769 | D9PC16:0/18:1 | 1.724096 | 0.6612318 | 2.6074 |
| 743 | D9PC16:0/16:0 | 1.836189 | 0.7086587 | 2.591076 |
| 715 | D9PC16:0/14:0 | 1.601133 | 0.6524921 | 2.453873 |
| 741 | D9PC16:0/16:1 | 1.867165 | 0.7873556 | 2.371437 |
| 817 | D9PC18:1/20:4 | 1.960294 | 0.8445815 | 2.321024 |
| 797 | D9PC18:0/18:1 | 1.339566 | 0.6187157 | 2.165075 |
| 795 | D9PC18:0/18:2 | 1.435755 | 0.7166049 | 2.003552 |
| 819 | D9PC18:0/20:4 | 1.2818 | 0.6556711 | 1.954944 |
| 771 | D9PC18:0/16:0 | 1.176324 | 0.6910945 | 1.702117 |

Table 23: *Ft* statistics for ETA PC species of the off and on ventilation NRDS patients.

The small *Ft* values of these species in the ETA samples suggest that the temporal enrichment of these labelled PC species are of little significance in distinguishing the two groups. Despite the different curve shapes for the two groups, close group mean and especially the large variances between individuals weaken the ability of PC turnover in discriminating the two groups. The large differences among individuals in ventilated group, leading to small *Ft* statistics, should also be considered as to if the grouping is appropriate. It is worth to investigate what causes the big differences in the time patterns and enrichment levels between individual patients, and if there are subgroups within this group. These facts indicate that the classification of groups directly affect the result of group comparison, and biomarker discovery requires reasonable grouping of subjects.

Besides, it should be noted that the observations of the off ventilation group is scarce, which could possibly fail to cover some part of the underlying population and should be taken into consideration when analysed further.

## 5.3 Characterisation of time course PEMT pathway *methyl*$D_3$ - PC enrichment in plasma and comparison between off and on ventilation NRDS patients

In order to analyse the turnover of liver synthesised PC through PEMT pathway, the time course enrichment of *methyl*$D_3$ labelled PC in plasma samples were modelled using SME method and the fitted values of the four designed time points are listed in Table 24.

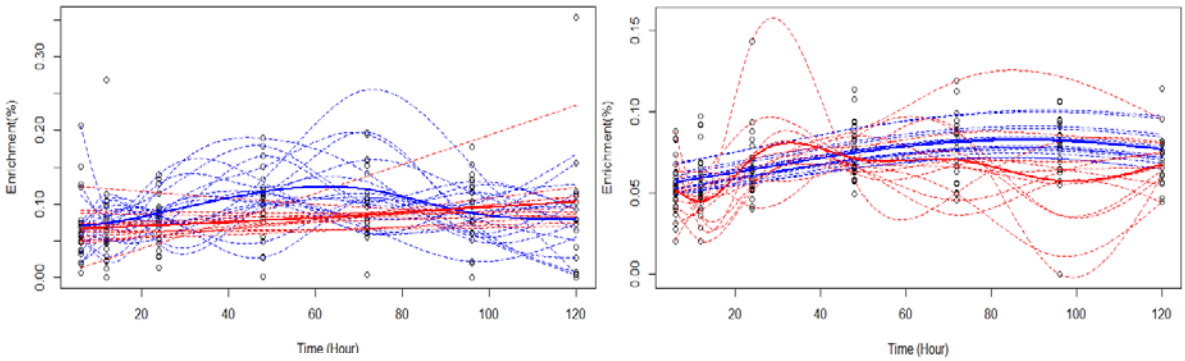
The figures in Table 24 show that the PEMT pathway PC enrichment values in plasma samples are much lower than those in ETA samples, suggesting their lower incorporation levels. It shows that the group mean values of most PC species are not significantly different between the on ventilation and off ventilation groups, but many of them are slightly higher in the off ventilation group, which is opposite to the results of ETA PC species analysis, where the mean values of on ventilation group are notably higher. The species with the highest enrichment are D3PC18:0/22:6, D3PC16:0/22:6, D3PC18:1/20:4 and D3PC18:0/20:4 in the off ventilation group. In the on ventilation group, the enrichment values are also high in these species. Moreover, there are some species in this group that have slightly different trends, e.g., D3PC16:0a/16:1 and D3PC16:0a/16:0 that have the highest enrichment only at the end of the investigated time period, and D3PC16:0/18:0 has the highest enrichment value just at 72 hour.

Table 24: Fitted group mean values of time course enrichment *methyl*$D_3$-PC synthesised by liver through PEMT pathway in plasma samples of the off and on ventilation NRDS patient.

| *m/z* | Species | Group | Enrichment (%) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | 6h | 12h | 24h | 48h | 72h | 96h | 120h |
| 709 | D3PC 16:0/14:0 | **Off Venti** | **0.070** | **0.074** | **0.087** | **0.116** | **0.119** | **0.087** | **0.079** |
| | | On Venti | 0.066 | 0.068 | 0.070 | 0.077 | 0.085 | 0.094 | 0.103 |
| 721 | D3PC16:0a/16:1 | **Off Venti** | **0.124** | **0.123** | **0.123** | **0.126** | **0.135** | **0.149** | **0.166** |
| | | On Venti | 0.092 | 0.089 | 0.086 | 0.084 | 0.146 | 0.123 | 0.133 |
| 723 | D3PC16:0a/16:0 | **Off Venti** | **0.090** | **0.091** | **0.094** | **0.100** | **0.105** | **0.111** | **0.116** |
| | | On Venti | 0.090 | 0.092 | 0.097 | 0.105 | 0.114 | 0.124 | 0.134 |
| 735 | D3PC16:0/16:1 | **Off Venti** | **0.058** | **0.071** | **0.080** | **0.092** | **0.096** | **0.094** | **0.080** |
| | | On Venti | 0.042 | 0.054 | 0.079 | 0.085 | 0.071 | 0.090 | 0.076 |
| 737 | D3PC16:0/16:0 | **Off Venti** | **0.056** | **0.059** | **0.064** | **0.074** | **0.081** | **0.082** | **0.077** |
| | | On Venti | 0.052 | 0.045 | 0.071 | 0.071 | 0.070 | 0.057 | 0.067 |
| 761 | D3PC16:0/18:2 | **Off Venti** | **0.056** | **0.066** | **0.082** | **0.100** | **0.099** | **0.094** | **0.081** |
| | | On Venti | 0.048 | 0.056 | 0.069 | 0.079 | 0.081 | 0.078 | 0.074 |
| 763 | D3PC16:0/18:1 | **Off Venti** | **0.056** | **0.063** | **0.074** | **0.087** | **0.090** | **0.084** | **0.075** |
| | | On Venti | 0.049 | 0.053 | 0.065 | 0.073 | 0.075 | 0.066 | 0.074 |
| 765 | D3PC16:0/18:0 | **Off Venti** | **0.062** | **0.065** | **0.070** | **0.082** | **0.094** | **0.107** | **0.119** |
| | | On Venti | 0.080 | 0.044 | 0.099 | 0.028 | 0.168 | 0.135 | 0.047 |
| 785 | D3PC16:0/20:4 | **Off Venti** | **0.061** | **0.082** | **0.096** | **0.120** | **0.110** | **0.109** | **0.098** |
| | | On Venti | 0.055 | 0.065 | 0.082 | 0.096 | 0.095 | 0.094 | 0.090 |
| 787 | D3PC18:1/18:2 | **Off Venti** | **0.064** | **0.076** | **0.098** | **0.122** | **0.119** | **0.108** | **0.094** |
| | | On Venti | 0.062 | 0.067 | 0.076 | 0.088 | 0.090 | 0.088 | 0.087 |
| 789 | D3PC18:0/18:2 | **Off Venti** | **0.059** | **0.084** | **0.094** | **0.111** | **0.105** | **0.097** | **0.085** |
| | | On Venti | 0.051 | 0.061 | 0.083 | 0.086 | 0.086 | 0.072 | 0.075 |
| 791 | D3PC18:0/18:1 | **Off Venti** | **0.052** | **0.067** | **0.071** | **0.084** | **0.080** | **0.074** | **0.071** |
| | | On Venti | 0.049 | 0.053 | 0.060 | 0.065 | 0.064 | 0.063 | 0.068 |
| 809 | D3PC16:0/22:6 | **Off Venti** | **0.111** | **0.128** | **0.156** | **0.174** | **0.159** | **0.140** | **0.120** |
| | | On Venti | 0.067 | 0.104 | 0.145 | 0.126 | 0.109 | 0.094 | 0.095 |
| 811 | D3PC18:1/20:4 | **Off Venti** | **0.083** | **0.129** | **0.155** | **0.150** | **0.143** | **0.129** | **0.117** |
| | | On Venti | 0.078 | 0.091 | 0.114 | 0.127 | 0.121 | 0.111 | 0.101 |
| 813 | D3PC18:0/20:4 | **Off Venti** | **0.077** | **0.104** | **0.134** | **0.146** | **0.134** | **0.119** | **0.100** |
| | | On Venti | 0.058 | 0.083 | 0.115 | 0.118 | 0.116 | 0.099 | 0.091 |
| 837 | D3PC18:0/22:6 | **Off Venti** | **0.112** | **0.129** | **0.159** | **0.187** | **0.177** | **0.151** | **0.125** |
| | | On Venti | 0.097 | 0.103 | 0.114 | 0.121 | 0.115 | 0.110 | 0.106 |

It can be seen from the plots (Figure 34) that the plasma PC species of PEMT pathway have lower incorporation and smaller gradients along the temporal curves than the ETA samples. Moreover, the shapes of the mean curves are different among species. Many of the individual curves fluctuate with multiple local maximum and minimum points, which indicates multiple waves of flooding of remodelled or resynthesised species entering the

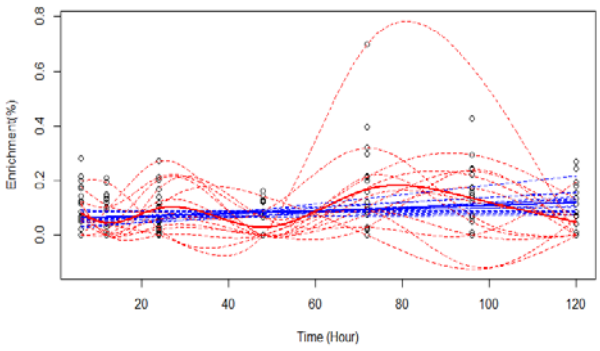
pool during the time period. As shown in the figure, the enrichment curve shapes of most PC species are very different among individuals, due to the diverse incorporation patters, which indicates heterogeneous turnover patterns of different PC species through PEMT pathway. There are a few exceptions where the fitted individual curve shapes are consistent, such as D3PC16:0/18:2, D3PC18:1/18:2 and D3PC18:0/22:6. This may suggest more consistent label incorporation patterns among the individuals, but it is also possible that the SME takes heterogeneous individuals as being in the same category and imposes a consistent shape for them. For example, on the plot of D3PC18:0/22:6, the individual curves have a similar time pattern, although there are many points scatter randomly away from the curves. It should be noted that the SME method uses the same smoothing parameter to all individuals when fitting each curve of them to reduce the computational burden, which may have contributed to the consistent species pattern among individuals in a group, such as D3PC18:0/22:6. However, there is another important step of the curve fitting that determines the final curve shape, the model selection, for which AICc criteria is adopted in the SME method. The smoothing parameters are selected to achieve the best fit, and it may result in individual curves with disparate shapes within the same group, which is seen in most of the species in the figure. It can be helpful to further evaluated these factors that affect the model shapes as to if the AICc criteria is appropriate in all cases.



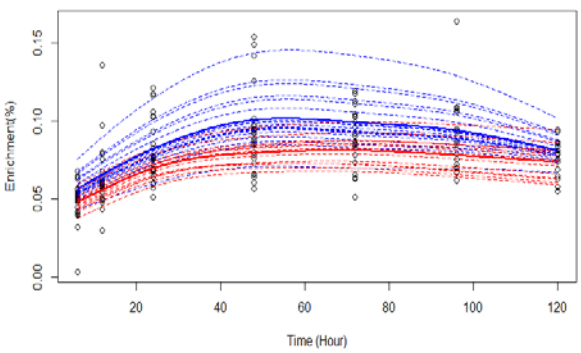

709 D3PC16:0/14:0



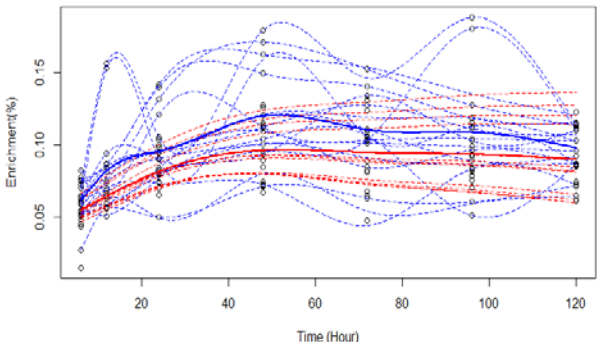

737 D3PC16:0/16:0



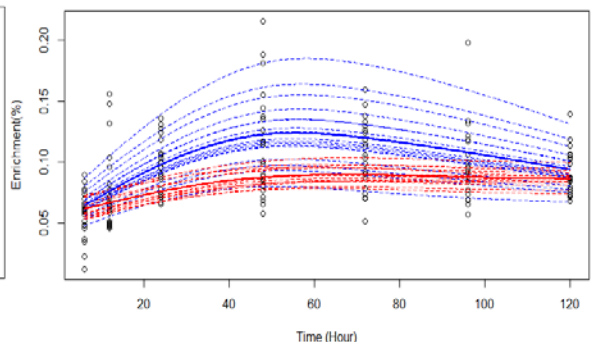

765 D3PC16:0/18:0



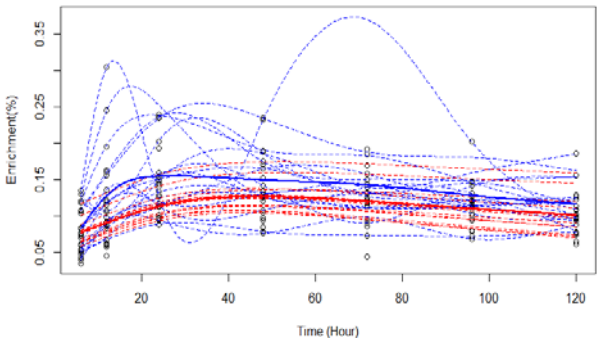

761 D3PC16:0/18:2



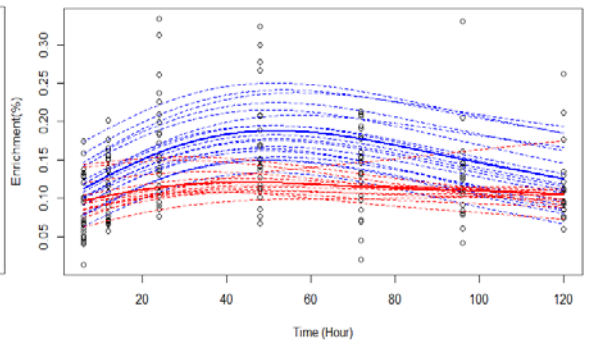

785 D3PC16:0/20:4




787 D3PC18:1/18:2




811 D3PC18:1/20:4




837 D3PC18:0/22:6

Figure 34: Time course enrichment of *methyl*$D_3$-PC species synthesised by liver through PEMT pathway for the off and on ventilation NRDS patients.

| *m/z* | Species | *L* | *se* | *Ft* |
|---|---|---|---|---|
| 837 | D3PC18:0/22:6 | 0.5428842 | 0.1010202 | 5.374019 |
| 787 | D3PC18:1/18:2 | 0.2681372 | 0.0567689 | 4.723312 |
| 761 | D3PC16:0/18:2 | 0.1725278 | 0.04394618 | 3.925889 |
| 721 | D3PC16:0a/16:1 | 0.3282075 | 0.08770852 | 3.742025 |
| 809 | D3PC16:0/22:6 | 0.4292514 | 0.1302551 | 3.295468 |
| 789 | D3PC18:0/18:2 | 0.2143859 | 0.0756707 | 2.833142 |
| 791 | D3PC18:0/18:1 | 0.1460655 | 0.05667729 | 2.577143 |
| 763 | D3PC16:0/18:1 | 0.1430773 | 0.05688063 | 2.515395 |
| 709 | D3PC 16:0/14:0 | 0.2802902 | 0.1163214 | 2.409618 |
| 737 | D3PC16:0/16:0 | 0.1534859 | 0.06562033 | 2.338999 |
| 811 | D3PC18:1/20:4 | 0.2767172 | 0.1372274 | 2.016486 |
| 813 | D3PC18:0/20:4 | 0.2202025 | 0.1113 | 1.97846 |
| 785 | D3PC16:0/20:4 | 0.1788182 | 0.0916905 | 1.950237 |
| 765 | D3PC16:0/18:0 | 0.4672063 | 0.3760173 | 1.242513 |
| 723 | D3PC16:0a/16:0 | 0.1027319 | 0.09802702 | 1.047995 |
| 735 | D3PC16:0/16:1 | 0.1381838 | 0.1640108 | 0.8425282 |

Table 25: *Ft* statistics for plasma *methyl*$D_3$-PC for the off and on ventilation NRDS patients.

Table 25 lists the *L*, *se* and *Ft* values calculated between the off ventilation and on ventilation groups. The *L* values range from 0.13 to 0.54, which are much smaller than those for ETA samples. The *se* values are also quite small, ranging from 0.004 to 0.16. Due to the extremely small *L* values, indicating small gaps between time models of the two groups, the *Ft* values are generally small with values between 0.84 and 5.37. D3PC18:0/22:6 is the species that has the biggest *L* and the highest *Ft*, just as it is in the ETA sample. D3PC16:0/22:6 has the second highest *L* value of 0.43 and its *Ft* value is 3.3, ranking at the fifth place. The overall small *Ft* values suggest that the time course enrichment of PEMT pathway PC species may be not significantly indicative in discriminating between the two groups. Similar to the result of ETA PC species analysis, the PEMT pathway species have heterogeneous enrichment patterns among individuals. This may suggest that the current grouping is not appropriate, which can lead to the

failure of using PC label enrichment as biomarkers for identifying group differences, and it underlines the critical importance of appropriate grouping in biomarker discovery analysis.

## 5.4 Conclusions

In this chapter, the time course enrichment of *methyl*$D_9$-PC species in ETA samples and *methyl*$D_3$-PC of PEMT pathway in plasma samples were modelled for both off ventilation and on ventilation NRDS patients. PC label incorporation level of the off ventilation group is lower than that of the on ventilation group over time, and also have wider range. The individual patients in this group have shown heterogeneous patterns. The time course enrichment of labelled PC in ETA samples are not significant in distinguishing the on and off ventilation NRDS patient groups.

PEMT pathway PC species have not shown significance as biomarkers in discriminating the two groups in the biomarker analysis, however, it provided extra knowledge for a more comprehensive understanding of PC turnover in pulmonary disease. The PEMT PC of neither the off ventilation nor the on ventilation NRDS patients have consistent incorporation patterns. Multiple peaks are shown on the fitted curves, indicating multiple waves of labelled species coming into the circulation. Different to surfactant PC in ETA samples, the enrichment levels of the off ventilation group are higher than those of on ventilation group.

The individual NRDS patients show varied labelled PC incorporation patterns in both groups, which may suggest they may have heterogeneous nature, and the categorising of the two groups may not be appropriate to be distinguished by PC kinetics. This also demonstrates group categorisation is critical important in biomarker discovery. The

modelling has shown consistent incorporation patterns in a couple of species with large amount of unfitted points dispersed far away from the fitted curves, which may be an indication of over-smoothing and should be carefully examined.

# Chapter 6 Hepatic phosphatidylcholine dynamic biomarker discovery for ARDS patients

## 6.1 Introduction

ARDS is a secondary lung disease triggered by various insults causing damage to the lung. Clinical management of ARDS includes ventilation strategies as well as therapies for the known underlying conditions, and there has been no newly developed treatment with proven benefit for improving the mortality rate. There is a need for a better understanding of the disease mechanism to guide treatment development. It would be useful to characterise the *in vivo* kinetics of critical materials for pulmonary function of ARDS patients and assess the implications from it for potential effective clinical management of the disease.

The PC label enrichment data analysed in this chapter is from a study investigating surfactant PC kinetics in ARDS patients and healthy people. Longitudinal observations of label incorporation of PC in BALF and plasma samples were made for both healthy control and ARDS patients. The study found the altered PC composition and label incorporation in BALF as well as plasma in patients comparing to healthy control. The mean enrichment of a few *methyl*$D_9$-PC species at different time points were calculated for both patient and control groups. As the BALF samples were only collected at 24 and 48 after label infusion for control group, the comparison between the two groups were carried out only at the two time points, where higher mean enrichment values of total surfactant PC as well as D9PC16:0/16:0 were found in ARDS patients (Dushianthan *et al.*, 2014). In plasma samples, increased flux of PC through the CDP-choline pathway and reduced flux through the PEMT pathway were found in ARDS patients comparing to healthy people. Based on

the findings of the study, it would be helpful to characterise PC kinetics for the two groups of individual subjects, offering better insights of time patterns of PC label incorporation than simply evaluating the mean values at each time points. Modelling the time course PC label incorporation also facilitates biomarker analysis through group comparison. However, BALF samples were only collected at 24 and 48 hour time points for healthy control group, which is not enough for time course modelling and biomarker analysis by SME method. As the study found that, in plasma samples hepatic PC synthesised through PEMT pathway is more indicative than those through CDP-choline pathway while comparing ARDS with control groups (Figure 32), therefore in this chapter the kinetics of PEMT pathway PC will be analysed for biomarkers identification for the control and ARDS patient groups.

## 6.2 Data description and methods

The data analysed in this chapter is obtained in the form of raw MS data sets from two previous studies carried out by Dushianthan *et al.* (2018). The first study "Pulmonary surfactant in adult patients with acute respiratory distress syndrome" was conducted in the General Intensive Care Unit at University Hospital Southampton, where the ARDS patients were recruited. The second study "Comparison of bronchoalveolar lavage and induced sputum surfactant phospholipid kinetics in healthy adult volunteers" was conducted in the Wellcome Trust Clinical Research facility at University Hospital Southampton. Detailed experimental methods can be found in the study of Dushianthan *et al.* (2018).

In the study, ARDS patients' blood sample were collected at 6, 12, 24, 48, 72 and 96 hours after label infusion, and healthy volunteers blood samples were taken at 8, 12, 24, 48, 72

and 96 hours post labelling. There were 10 people in both patient and healthy groups, although not all the subjects had complete samples collected at each time points. In this chapter sixteen *methyl*$D_3$ PC, selected based on the confidence of existence, in these plasma samples from patient and healthy groups were analysed.

MS data processing and kinetic calculations were carried out using LipidomeLabelling. The procedures conducted for MS data processing include raw data extraction, background subtraction, peak identification, peak integration, species identification, isotope correction, between scan isotope correction, and enrichment calculation. The multivariate time course modelling and biomarker discovery analysis were performed using SME method facilitated by the SME R package.

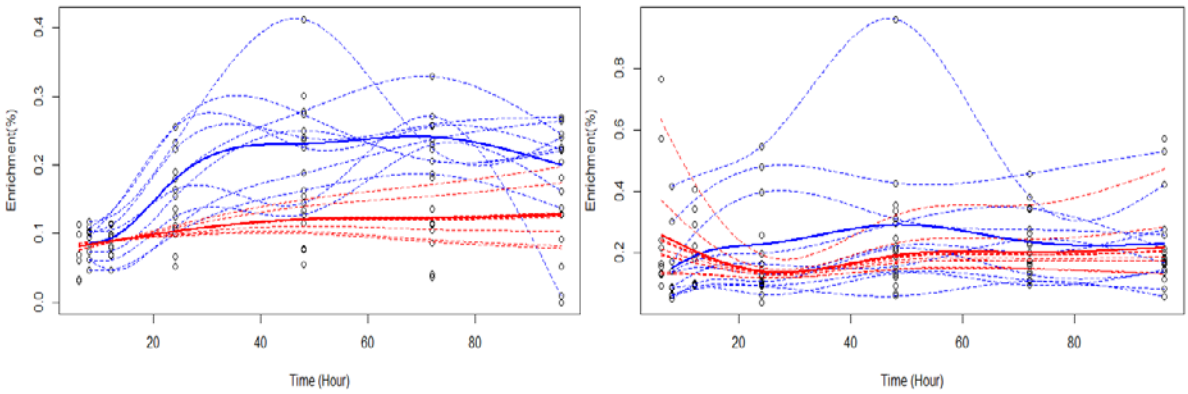
## 6.3 Characterisation of time course PEMT pathway *methyl*$D_3$ - PC enrichment in plasma and comparison between control and ARDS patients

In this analysis, time course enrichment of *methyle*$D_3$-PC in plasma samples are modelled using SME method, for healthy control group for the time period from 8 to 92 hour post label infusion, and for ARDS patient group from 6 to 92 hour. The fitted mean values at the four sample collecting time points are listed in Table 26. As shown in the table, the healthy control group has generally higher enrichment level than ARDS patients over the time period, which is opposite with that of surfactant PC observed at 24 and 48 hour points described in the introduction section. The species with the highest temporal enrichment level in healthy control group is D3PC18:0/22:6. The other species with comparative high enrichment include D3PC16:0/22:6, D3PC18:0/20:4, D3PC18:1/20:4, and D3PC16:0a/16:1. D3PC16:0/16:0, D3PC16:0/18:0 and D3PC16:0a/16:0 are the three species with the lowest enrichment level over time. For ARDS patients, the species with

the highest time course enrichment are the same with those of the healthy volunteers, and the species with the lowest enrichment level is D3PC16:0/16:0, which is also low in healthy people. But species D3PC16:0/18:0, which is the lowest in healthy group, is not as low in patient plasma. It exhibits a fluctuating pattern and gets quite high at some time points, which can be seen from Figure 35.

Table 26: Fitted group mean values for the time course enrichment of PEMT pathway *methyl*$D_3$-PC in plasma samples of healthy control and ARDS patients.

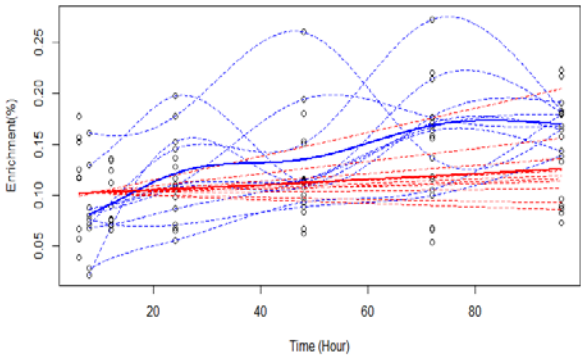
| ***m/z*** | **Species** | **Group** | **Enrichment (%)** | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | **6h** | **8h** | **12h** | **24h** | **48h** | **72h** | **96h** |
| 709 | D3PC 16:0/14:0 | **Control** | | **0.086** | **0.093** | **0.180** | **0.230** | **0.240** | **0.200** |
| | | Patient | 0.081 | | 0.089 | 0.103 | 0.121 | 0.123 | 0.128 |
| 721 | D3PC16:0a/16:1 | **Control** | | **0.151** | **0.189** | **0.227** | **0.290** | **0.234** | **0.228** |
| | | Patient | 0.258 | | 0.205 | 0.139 | 0.191 | 0.200 | 0.216 |
| 723 | D3PC16:0a/16:0 | **Control** | | **0.081** | **0.091** | **0.121** | **0.135** | **0.168** | **0.169** |
| | | Patient | 0.102 | | 0.103 | 0.106 | 0.112 | 0.119 | 0.126 |
| 735 | D3PC16:0/16:1 | **Control** | | **0.096** | **0.084** | **0.204** | **0.253** | **0.283** | **0.249** |
| | | Patient | 0.060 | | 0.075 | 0.119 | 0.133 | 0.135 | 0.155 |
| 737 | D3PC16:0/16:0 | **Control** | | **0.057** | **0.063** | **0.101** | **0.166** | **0.191** | **0.195** |
| | | Patient | 0.050 | | 0.061 | 0.081 | 0.093 | 0.112 | 0.115 |
| 761 | D3PC16:0/18:2 | **Control** | | **0.083** | **0.070** | **0.227** | **0.303** | **0.314** | **0.298** |
| | | Patient | 0.059 | | 0.078 | 0.122 | 0.160 | 0.157 | 0.160 |
| 763 | D3PC16:0/18:1 | **Control** | | **0.072** | **0.067** | **0.178** | **0.251** | **0.285** | **0.274** |
| | | Patient | 0.060 | | 0.068 | 0.103 | 0.149 | 0.131 | 0.133 |
| 765 | D3PC16:0/18:0 | **Control** | | **0.054** | **0.062** | **0.085** | **0.131** | **0.177** | **0.223** |
| | | Patient | 0.132 | | 0.162 | 0.073 | 0.225 | 0.139 | 0.069 |
| 785 | D3PC16:0/20:4 | **Control** | | **0.102** | **0.077** | **0.266** | **0.323** | **0.332** | **0.308** |
| | | Patient | 0.059 | | 0.081 | 0.120 | 0.148 | 0.156 | 0.176 |
| 787 | D3PC18:1/18:2 | **Control** | | **0.102** | **0.084** | **0.276** | **0.330** | **0.317** | **0.309** |
| | | Patient | 0.062 | | 0.092 | 0.139 | 0.148 | 0.163 | 0.179 |
| 789 | D3PC18:0/18:2 | **Control** | | **0.116** | **0.091** | **0.352** | **0.395** | **0.363** | **0.326** |
| | | Patient | 0.061 | | 0.095 | 0.157 | 0.200 | 0.191 | 0.178 |
| 791 | D3PC18:0/18:1 | **Control** | | **0.095** | **0.079** | **0.229** | **0.290** | **0.292** | **0.289** |
| | | Patient | 0.057 | | 0.070 | 0.107 | 0.152 | 0.164 | 0.140 |
| 809 | D3PC16:0/22:6 | **Control** | | **0.324** | **0.280** | **0.961** | **0.817** | **0.626** | **0.497** |
| | | Patient | 0.090 | | 0.211 | 0.313 | 0.334 | 0.246 | 0.234 |
| 811 | D3PC18:1/20:4 | **Control** | | **0.210** | **0.095** | **0.528** | **0.501** | **0.436** | **0.373** |
| | | Patient | 0.079 | | 0.145 | 0.228 | 0.219 | 0.199 | 0.186 |
| 813 | D3PC18:0/20:4 | **Control** | | **0.216** | **0.119** | **0.566** | **0.524** | **0.451** | **0.388** |
| | | Patient | 0.077 | | 0.157 | 0.217 | 0.221 | 0.206 | 0.206 |
| 837 | D3PC18:0/22:6 | **Control** | | **0.395** | **0.635** | **1.211** | **1.154** | **0.813** | **0.623** |
| | | Patient | 0.090 | | 0.214 | 0.406 | 0.440 | 0.292 | 0.284 |



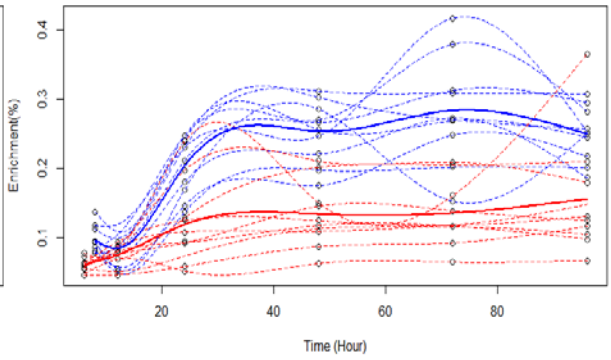

709 D3PC 16:0/14:0



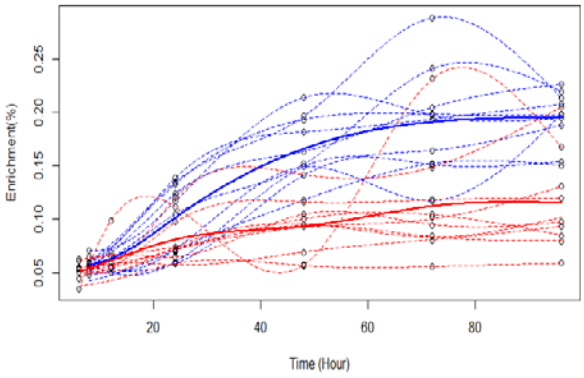

721 D3PC16:0a/16:1



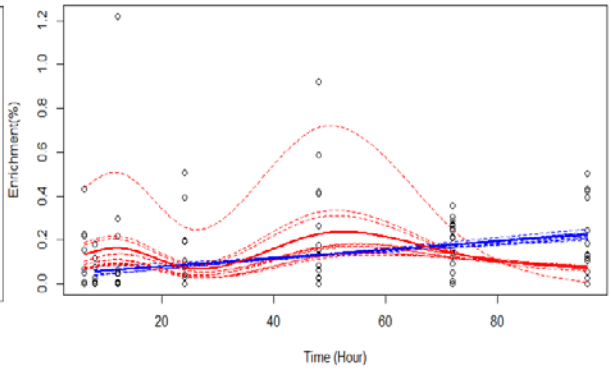

723 D3PC16:0a/16:0




735 D3PC16:0/16:1

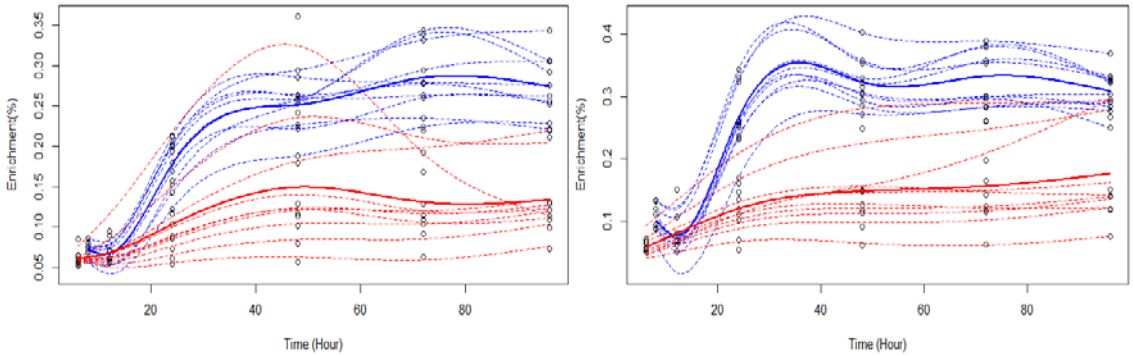



737 D3PC16:0/16:0




765 D3PC16:0/18:0

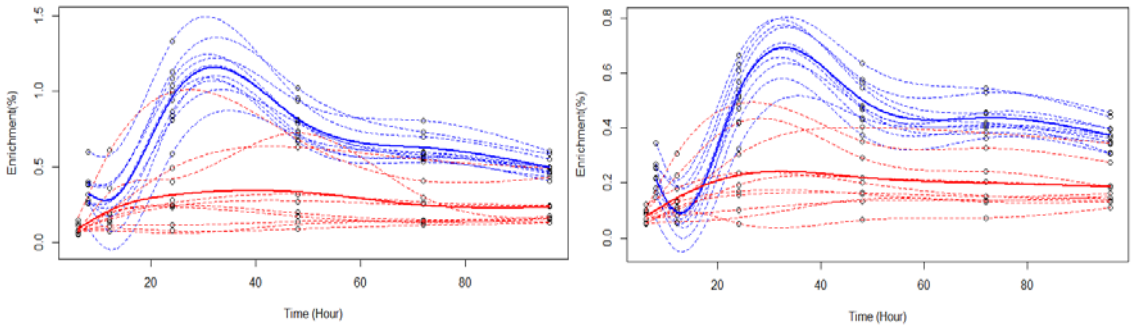
Figure 35: Time course enrichment of *methyl*$D_3$-PC for healthy control (in blue) and ARDS groups (in red). (To be continued)




763 D3PC16:0/18:1



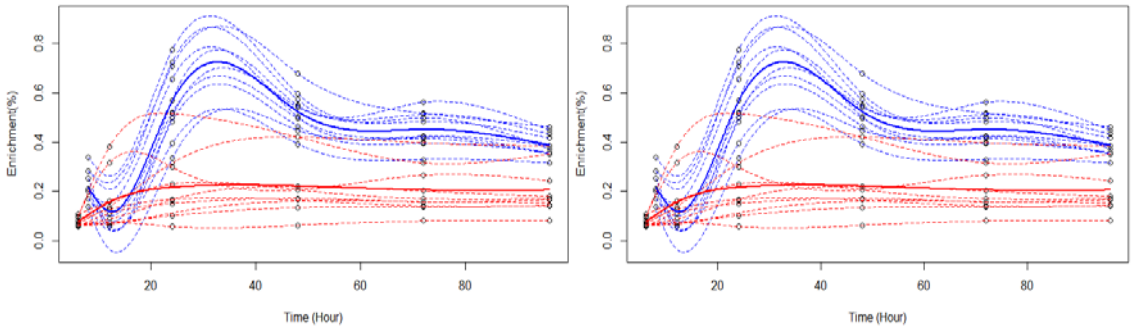

785 D3PC16:0/20:4




809 D3PC16:0/22:6




811 D3PC18:1/20:4




813 D3PC18:0/20:4




837 D3PC18:0/22:6

(Figure 35 continued)

Figure 35 shows the plots of fitted enrichment curves of plasma *methyl*$D_3$-PC species for the two groups. The healthy group has apparent higher enrichment levels than ARDS patients over the time period. It can be found that the incorporation of most species

exhibit tidal style temporal patterns. This is more obvious in the control group, where two peaks can be found in each curve along the time curves.

In control group, the first peak of the labelled PC species appears at around 30 hour, and the second appears at around 72 hour, which can be either lower or higher than the previous peak in different species. Another feature of the control group is that, for most of the species, there is a short term of dropping between 6 and 12 hours, and after this the enrichment goes up dramatically, implying decreasing label incorporation at the early time period before picking up. As healthy control people are assumed to be in steady state with constant PC turnover, the multiple peaks on the curves indicates there are multiple waves of labelled PC species coming into circulation. The increasing part is the period that the amount of newly entered labelled species overweighs those leaving the pool, and the decreasing part means the opposite. Based on the signs of decreasing trend at early time period on the curves, it is appropriate to assume that the experiment may have missed the early waves of incoming labelled species after label infusion, but captures two later ones that peak at around 33 hour and 72 hour. The four species having the fastest increasing rates also have the highest temporal enrichment levels over time, including D3PC16:0/22:6, D3PC18:0/22:6, D3PC18:0/20:4 and D3PC18:1/20:4. The decreasing rates of them are also fast, which can be seen from the steep slopes at the beginning of the time period and after the first peak, suggesting higher fractional appearance and disappearance rate of these species than others in the blood of healthy people. This further implies higher fractional synthesis rate of these species through PEMT pathway by liver.

A species that should be paid more attention is D3PC16:0/16:0, the labelled form of PC16:0/16:0, which is the most abundant species in pulmonary surfactant. However, this

species turns out to have low temporal enrichment level, suggesting its low fractional turnover in blood. Besides, the group mean curve does not show an obvious two-peak pattern, but keeps increasing over the time period. However, the individual curves on the plot show that many of them have more than one peak, although they do not change at the same pace, and even go in the opposite direction especially at the 48 hour, which cancels out the fluctuation and results in an increasing mean curve. The species that act most diversely among individuals, in terms of enrichment level as well as temporal pattern, in the healthy group, include D3PC16:0/14:0, D3PC16:0a/16:1, D3PC16:0a/16:0, D3PC16:0/16:1 and PC16:0/16:0. This shows that these species, the counterpart of which are characteristic species in pulmonary surfactant, are not dominant in the blood and the turnover vary a lot among individuals.

The ARDS patient group has generally low enrichment level and small change over time for all species. Besides, although the multi-peak curve shape appears in some species in some individual patients, the multiple peaks are not obvious in the group mean curves. But D3PC16:0/18:0 is an exception, which increases at the beginning and reaches a local maximum value at around 10 hour, and then decreases to a minimum at around 24 hour followed by another increase to reach the local maximum at around 50 hour. Moreover, its enrichment level is not lower than the control group all the time, but their relationship changes with the fluctuation of the curves. However, a close examination of the individual patients shows that some extreme values of a few patients have largely affected the curve shape of the whole group.

The incorporation curves of most individuals as well as the whole ARDS group are much lower and flatter comparing to the healthy group, but there are three patients with different shapes and much higher enrichment levels than the other patients. On the plot

of D3PC16:0/18:1, the three fitted individual curves with red dotted lines that are above the group mean curve are for patient No.4, 3 and 1, from high to low. Furthermore, these three patients are also found to have different curve shapes and higher enrichment levels than other patients in most of other species, especially the longer fatty chain species. Taking species D3PC18:1/20:4 as an example, the curves for the three patients are much higher than other individuals and largely lift up the group mean values, which results in the overall group difference between ARDS and control groups much smaller. Without these three patients, the gap between the two groups would have been much greater. However, there are only three patients in question, which is not enough to be analysed as a group by the SME for further examination. Nevertheless, other analysis could be done by looking into other clinical information of these patients to investigate the reason behind it.

Group comparisons were carried out for identifying potential markers. The observation time points of the two groups are not completely the same, with the first time point of control group at 8 hour after label infusion, and 6 hour for ARDS group. In spite of this, the SME can model the enrichment over the whole time period for both groups and compare them over the overlapping period from 8 hour to 120 hour after label injection.

The *Ft* values are generally very small for all the species (Table 27), suggesting their limited ability of distinguishing the two groups by the time course label enrichment. The species with the highest *L* values include D3PC18:0/22:6, D3PC16:0/22:6, D3PC18:0/20:4 and D3PC18:1/20:4, exactly the ones with the highest time course enrichment in both groups, with the value of 5.928, 4.501, 2.745 and 2.534, respectively. Despite the comparative high *L* values, these specie do not get big *Ft*. This is due to the big *se* values owing to the large variances among individual patients. Species with longer acyl chains

are ranked higher, due to their dominancy and high turnover in blood. Characteristic lung surfactant species are not as active in plasma as in lung surfactant, and they are ranked lower in the list with the smallest *L* values. Species D3PC16:0/16:0, D3PC16:0/18:0, D3PC16:0a/16:0 and D3PC16:0a/16:0 have the smallest *Ft* values among all the species. This is reasonable as these species are not the major PC species synthesised through PEMT pathway.

| *m/z* | Species | *L* | *se* | *Ft* |
|---|---|---|---|---|
| 837 | D3PC18:0/22:6 | 5.927751 | 0.8062072 | 7.35264 |
| 785 | D3PC16:0/20:4 | 1.514275 | 0.209139 | 7.240519 |
| 811 | D3PC18:1/20:4 | 2.53435 | 0.3737603 | 6.780682 |
| 813 | D3PC18:0/20:4 | 2.745033 | 0.4144251 | 6.623713 |
| 809 | D3PC16:0/22:6 | 4.501139 | 0.7016266 | 6.415292 |
| 761 | D3PC16:0/18:2 | 1.281495 | 0.2024036 | 6.331383 |
| 787 | D3PC18:1/18:2 | 1.453808 | 0.2353299 | 6.177744 |
| 763 | D3PC16:0/18:1 | 1.103629 | 0.1871716 | 5.896346 |
| 789 | D3PC18:0/18:2 | 1.709253 | 0.2922198 | 5.849202 |
| 709 | D3PC 16:0/14:0 | 0.9187931 | 0.1744889 | 5.265626 |
| 735 | D3PC16:0/16:1 | 1.091156 | 0.208337 | 5.237459 |
| 791 | D3PC18:0/18:1 | 1.218338 | 0.2473827 | 4.92491 |
| 737 | D3PC16:0/16:0 | 0.6014878 | 0.1383454 | 4.347724 |
| 765 | D3PC16:0/18:0 | 0.7256573 | 0.3026004 | 2.398071 |
| 723 | D3PC16:0a/16:0 | 0.3236736 | 0.1370244 | 2.36216 |
| 721 | D3PC16:0a/16:1 | 0.6685675 | 0.5415068 | 1.234643 |

Table 27: *Ft* statistics for plasma PC species of control and ARDS patient groups.

As mentioned before, the individual fitted curves of ARDS patient group indicate that the big variances are mainly from the three individual patients who have substantial higher incorporation than other patients. The temporal PC enrichment of these three patients hugely affect the patient group mean, and largely narrow the distance between ARDS and control groups. This effect brings down the *L* values, and enlarges the variance within the group, which increases the *se* values. Taking species D3PC16:0/20:4 as an example, it has the second largest *Ft*, although it is not one of the species with highest enrichment level

or *L* value, but it has a comparatively smaller variance. By checking the individual models of the group, it was found that only two of the three patients mentioned above, rather than all three, have higher enrichment levels of this species than the group mean. This generates much smaller variance and reduces the *se* value, resulting in a big *Ft* value for the species. Therefore, these patients should be further investigated as to 1) what causes the significant difference of these patients comparing with others, 2) if they are appropriate to be put in the same group with other patients.

According to the analysis, due to the small *Ft* values of all the species, they are not significant as potential biomarkers. However, if the grouping of the heterogeneous individual patients is properly treated, the result will be hugely improved, and the label incorporation of PEMT pathway PC species can be promising biomarkers for ARDS. This analysis and the original study integrally suggest that, the PC turnover of the ARDS patients are altered comparing with healthy people, but the alteration is different between surfactant PC and PEMT pathway PC, where the label incorporation of surfactant PC of the patients are higher than healthy people, but that of PEMT PC are lower.

## 6.4 Conclusions

In this chapter, the time course enrichment of plasma *methyl*$D_3$-PC synthesised through PEMT pathway for both healthy people and ARDS patients were characterised, and the group comparisons were carried out for identifying potential biomarkers for distinguish the group difference.

In plasma, the incorporation level of PEMT pathway PC species at 24 and 48 hour are higher in control healthy group than the ARDS patient group, which is different from that

of surfactant PC which is higher in ARDS group. The species with longer unsaturated fatty acyl chains are more active than the species that are highly active in lung surfactant. The group comparisons generated small *Ft* values, indicating poor capability of PEMT pathway PC species in distinguishing healthy people and ARDS patients. Nevertheless, the analysis has recognised comparatively large difference between group mean enrichment levels, denoted by *L* values, and found that the small *Ft* values are due to three ARDS patients with significantly high PC label incorporation. Further investigation was suggested to identify the cause of the heterogeneity and improve the grouping, which may achieve high significance of the PEMT PC time course incorporation as potential biomarkers.

The SME method has successfully modelled the temporal PC label enrichment for both individual subjects and groups, estimated the group differences for all the species, as well as identified the unusual high PC label enrichment of three ARDS patients. The results are reliable for biomarker discovery analysis.

# Chapter 7 General discussion

The clinical management of various pulmonary diseases suffers from a lack of specific and sensitive biomarkers. Despite the endless effort put into the biomarker discovery studies, there have been hardly any established new biomarkers proven to be sensitive and effective enough to replace the traditional measures for diagnosing and predicting in clinical settings. Pulmonary surfactant is critical for normal lung function, and PC species are the major component and surface active material of lung surfactant. The PL alterations in surfactant are associated with malfunctions of lungs and pulmonary diseases, and *in vivo* PL turnover studies have provided valuable insights of surfactant PL metabolism in healthy and disease states. The traditional techniques for analysing species *in vivo* kinetics have their drawbacks. The radioactive labels do not trace target molecules and they are harmful to human health. Lipid detection by GC techniques are not informative regarding molecular species. The relatively new ESI MS/MS coupled with stable isotope labelling techniques offer good tools for analysing molecular species *in vivo* kinetic activities in human. This methodology not only identifies intact molecular species, but can also trace multiple targeted species at the same time, which has provided new insights of pulmonary surfactant system.

Due to the comparatively new ESI MS/MS coupled with stable isotope labelling techniques applied in lilpidmoics analysis, there is a lack of efficient and comprehensive software for dealing the generated data for dynamic PC analysis. In this study a new comprehensive, robust, user friendly, and fast running platform, LipidomeLabelling was developed for the special needs. It allows PL species of multiple lipid categories and

classes as well as their labelled forms to be analysed together in the dynamic analysis. It also supports data analysis of labelling studies with more than one type of label used for different researching purposes. The software facilitates large-scale high-dimensional data processing and dynamic pulmonary surfactant lipidomic analysis. In spite of the comprehensive functionalities, the platform can be further improved to serve wider use in the future. Firstly, although the platform is able to manage various labelling techniques, the current library only offers reference lists of species labelled with deuterium and $^{13}C$. Therefore, a more comprehensive database will be needed to facilitate studies using different labelling materials. Secondly, as the platform offers batch analysis for omics data analysis that generates large datasets in parallel, which is not easy to be transferred into common statistical software for statistical analysis. Therefore, it would be useful to add some statistical functions into the platform or to add some features that can make the results compatible with other statistical packages. Thirdly, the platform can also be improved in user convenience. For instance, there is only one output format of the PL analysis, and more choices of output format can better satisfy the various needs of different users.

This is the first study to characterise the time course enrichment of multiple *methyl*$D_9$ labelled PC species, and to use them as variables for biomarker discovery analysis for pulmonary diseases. As the major lipid class in surfactant PL, the abnormal turnover of surfactant PC have been found in various lung diseases, so the surfactant PC kinetics has the potential to indicate lung disease. The traditional metabolic flux analysis tries to calculate metabolic kinetic parameters, such as synthesis rate, pool size, degradation rate, and half-life. However, it is not always possible to calculate these parameters accurately due to various limitations. Nevertheless, time course label incorporation of a species in a

certain pool contains the integrated information of the species metabolism, including that the kinetic parameters can offer, although inexplicitly. This study modelled and used time course label enrichment of PC species, from three studies facilitated by ESI MS/MS analysis, as variables to compare between different groups, and evaluated their significance in distinguishing the two groups for biomarkers discovery analysis.

This study is the first to model the time course PC label incorporation in bolus label infusion experiment, facilitated by SME method. SME has its roots in functional data analysis that sees an observed time series as the realisation of the underlying stochastic process, i.e. a smooth curve (Berk *et al.*, 2011). The fitted continuous curves can capture the time factor of the process and estimate turning points and inflection points that are not observed, such as maximum and maximum points. This is a big step forward for the current longitudinal PC labelling analysis, which normally characterise the time trend by joining up the observations at separate time points. It may lose the time effect, and is not possible to approximate any unobserved turning points or inflection points within the observation time period. This study has characterised temporal incorporation trend of sixteen PC species for group of subjects from different studies, for better understanding of PC metabolism in various subjects, whereas previous studies only concentrate on a few major species, such as PC16:0/16:0, and the total PC species.

The modelling of the time course enrichment of multiples species in various sample types and different subject groups of the three studies has provided massive new insights of *in vivo* kinetics of surfactant PC species. By comparing the early time point enrichment level of multiple species, it is possible to compare relative fractional appearance, synthesis and

secretion rate among these species. In addition, by comparing the shapes of time curves of multiple species, it allows us to recognise species with different metabolisms. In the mice study, BALF and lung PC with 16:0 acyl chain present much higher fractional appearance rate than those with 18:0 indicated by the high enrichment level at early time points. The delayed increasing and low level of enrichment of the species with 18:0 suggest a different metabolism comparing with the other species. In the ARDS study, the PC species synthesised through PEMT pathway in healthy people have shown multiple waves of incoming labelled PC species into the plasma pool within 96 hour time frame after label infusion. The ETA PC enrichment curves of NRDS patients illustrate low incorporation and late arrival of labelled species into the pool. The enrichment curves for ARDS and NRDS patients vary largely among individuals, indicating heterogeneous flux kinetics of PC species between them. The modelling of time course PC label incorporation takes into account the time factor and simulates the dynamic biological process, which lays the basis for the biomarker discovery analysis.

The biomarker analysis for the GM-CSF beta chain depleted mice and normal mice were conducted first for validating the methodology for the following application to the more complicated clinical data. The mice models have clear and consistent incorporation patterns among individuals within each group. The most significant differences were found between control and KO mice groups with large *Ft* values, and small difference were between control age groups with very small *Ft* values. The characteristic lung surfactant PC species and those with 16:0 acyls are more significant than others in discriminating the control and KO groups. Most of the PC species in BALF are highly significant and are promising potential biomarkers, and in lung tissue there are also many significant species, although fewer than those in BALF. The comparisons between control

and high dose antibody treated mice showed lower enrichment at the increasing part of the curves of the high dose group than the control, indicating the difference in PC turnover, and a few species are significant in identifying the differences. However, the *Ft* values suggest that the difference between control and high dose group recognised by the species are much smaller than that between the control and KO groups. The *Ft* range obtained from the mice analysis is a good reference to be compared to the analysis of less clear clinical scenarios, in which the group difference are less significant, to evaluate to what extent patient groups are different from healthy groups. The analysis between the off ventilation and on ventilation NRDS patients suggests that the time course enrichment of surfactant PC is higher in the weaker ventilated infants, but there is no significant species in distinguishing the two groups, largely due to the small group difference and large variances among individual patients. Besides, ETA sample is not as effective as BALF in recovering pulmonary surfactant, which can be the reason that the surfactant PC dynamic biomarker analysis for the NRDS patient ETA samples are not as significant as that of the BALF in the mice study.

The dynamic biomarker analysis of hepatic PC synthesised through PEMT pathway for NRDS patients groups as well as ARDS patients and healthy people have provided valuable insights of surfactant PC turnover. By integrating the result with the analysis of surfactant PC, it was found that the physically better group have lower label incorporation of surfactant PC than the physically worse group, whereas the situation is opposite for the PEMT pathway PC, where the incorporation of the physically better group is higher than the physically worse group. The off ventilation group of NRDS patients have lower PC incorporation in ETA samples but higher PEMT PC label incorporation in plasma than the

ventilated group. The healthy group has lower surfactant PC incorporation at certain time points (Dushianthan *et al.*, 2014), but higher PEMT PC label incorporation than ARDS group.

Although the analysis of PEMT PC species have not been recognised as significant in distinguishing groups in the two analysis, large gaps between group means of healthy and ARDS, i.e. *L* values, were recognised in many species. Due to three heterogeneous patients in the ARDS group, the distance between the two groups has been largely narrowed and the variance of ARDS group increased, resulting in low *Ft* values. An improved grouping and proper management of the heterogeneous patients may generate much better results and find promising biomarkers.

SME method worked well in modelling the time course label incorporation, and was able to manage the issues of small sample sizes and missing values. It provides efficient tools for group comparison for biomarker discovery analysis. It can tackle the issues of different sample collecting times between groups by modelling over the whole time period and comparing two groups over the overlapping time periods. The only issue with the method is that a couple of models might be slightly over smoothed in the study, and if that is the case, it could somehow affect the *se* values and hence the *Ft* statistics. This issue should be further investigated in terms of what cause this issue and if it is appropriate to use AICc criteria in all situations.

The time course modelling and biomarker discovery analysis carried out in the three studies have demonstrated the importance of the following points. Firstly, the more experimental subjects there are, the better approximation of the underlying population we can get. The scarce enrichment data of ETA PC of the off ventilation NRDS patients has

shown weak representation of the underlying group. Although the SME method can deal with small sample size with its correcting features, it is always the best solution to have more subjects when possible.

Secondly, the more time points within a time frame are observed, the more reliable the model fitting will be. The SME method models the time series data based on the values from the observed time points, and is able to approximate turning points and inflection points within the observation period. However, it is still possible that some important features may be lost due to missing observations at some critical time points, and even SME method would not be able to capture it, resulting in incorrect curve shapes. An example is shown in Figure 36, where the plot of D9PC16:0/16:0 in BALF of mice sample is compared with the plot of the same data but with an extra fake time point observation between the 24 and 48 hour. The time trend has been dramatically changed by adding the fake observation. This indicates that sample collecting time should be carefully designed in an experimental study to cover the critical part of changing as much as possible, ideally at least two observations for each curvature between two stationary points for the whole biological process, although this design will need expertise theoretical and empirical knowledge with appropriate assumptions.

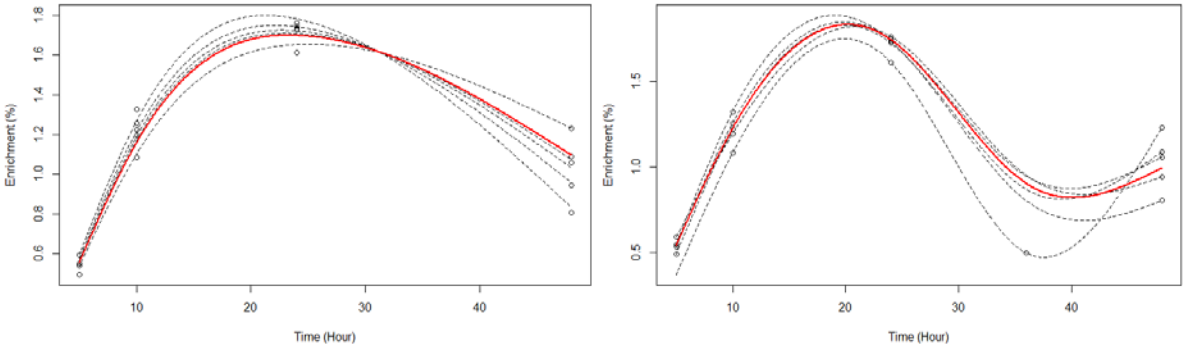



Figure 36: Time course enrichment of D9PC16:0/16:0 enrichment in BALF of 169 days control mice group (left) and the same data with an extra fake observation of with a small value of 0.5% at 36 hour (right).

Thirdly, grouping is critical important for biomarker discovery. The groups in the three studies are classified based on the current knowledge and assumptions, some of which may not be appropriate for PC label incorporation biomarker analysis. This can be due to the heterogeneous subjects within groups as well as wrong assumptions. The three patients in ARDS group with extremely high PEMT pathway PC incorporation largely lift the group mean and narrows the gap between the groups, which is an example of grouping issue due to heterogeneity within a group. The *Ft* analysis found no difference between the two groups of NRDS patients, and the great variances occurred in each groups, which is another example of heterogeneity individuals as well as the inappropriate grouping criteria. An ideal grouping should generate large group distance and small individual variation.

This study concentrates on the feasibility of using time course PC incorporation as biomarkers for pulmonary diseases. There are still a lot to be done in future studies. Firstly, the modelling of the PC species has provided large amount of information on their metabolism, which has not been fully explored. More investigation could be done on the

turnover patterns of individual PC species in different compartments, hydrolysed species turnover patterns, investigation of surfactant secretion by comparing species between lung and BALF, and transporting information by comparing those between plasma and ETA samples. These would give wider views and deeper insights of PC species metabolism. Secondly, the time course enrichment modelling in this study does not report confidence interval in the results. This is because the main purpose is to investigate possible biomarkers by comparing between species, and confidence interval of each group mean curve is not much of a concern in this context. The confidence interval can be further investigated in future studies, especially when biomarkers are identified and the value range should be defined in both groups for diagnostic or prognostic purposes. Thirdly, in this study the time course PC enrichment of the entire observing time frame was modelled and used as a variable for biomarker analysis. In future studies, it can also be considered to use enrichment within certain time window, in which the two groups have the largest gap and smallest variation, or even as precise as enrichment at certain time points, as biomarker variables, which may be more significant, efficient and applicable. Fourthly, SME can cope with large numbers of variables for biomarker discovery. Berk *et al.* (2011) has evaluated around 2000 variables in a biomarker analysis. A typical biomarker analysis by SME method involves calculating moderated *Ft* values of each variables, simulating the *Ft* distribution and selecting the most significant species as potential biomarkers. However, the number of species assessed in this pilot study is limited due to the detectability of labelled species as well as their capability in indicating PC metabolism. Therefore, the *Ft* values were not moderated and the distributions were not determined. In future studies, it can be considered to add more relevant and

detectable species into the analysis, for distribution determination and biomarker selection. Lastly, in this study, only PC species were analysed for dynamic biomarker discovery for lung diseases. However, kinetics of other major surfactant phospholipids like PG, PI and PE may also be indicative, which is worth to be assessed, although the labelling techniques for these species might be different from those for PC species.

# Bibliography

Agarwal, A.R., Yin, F. and Cadenas, E. (2014) Short-term cigarette smoke exposure leads to metabolic alterations in lung alveolar cells. *American Journal of Respiratory Cell and Molecular Biology,* 51 (2), 284-293.

Agassandian, M. and Mallampalli, R.K. (2013) Surfactant phospholipid metabolism. *Biochim Biophys Acta,* 1831 (3), 612-625.

Ahmed, Z., Mayr, M., Zeeshan, S., Dandekar, T., Mueller, M.J. and Fekete, A. (2015) Lipid-Pro: a computational lipid identification solution for untargeted lipidomics on data-independent acquisition tandem mass spectrometry platforms. *Bioinformatics,* 31 (7), 1150-1153.

Akella, A. and Deshpande, S.B. (2013) Pulmonary surfactants and their role in pathophysiology of lung disorders. *Indian Journal of Experimental Biology,* 51 (1), 5-22.

Atkinson, A.J., Colburn, W.A., Degruttola, V.G., Demets, D.L., Downing, G.J., Hoth, D.F., Oates, J.A., Peck, C.C., Schooley, R.T., Spilker, B.A., Woodcock, J., Zeger, S.L. and Biomarkers Definitions Working, G. (2001) Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clinical Pharmacology & Therapeutics,* 69 (3), 89-95.

Bajwa, E.K., Khan, U.A., Januzzi, J.L., Gong, M.N., Thompson, T. and Christiani, D.C. (2009) Plasma C-reactive protein levels are associated with improved outcome in ARDS. *Chest,* 136 (2), 471-480.

Ban, N., Matsumura, Y., Sakai, H., Takanezawa, Y., Sasaki, M., Arai, H. and Inagaki, N. (2007) ABCA3 as a lipid transporter in pulmonary surfactant biogenesis. *J Biol Chem,* 282 (13), 9628-9634.

Batenburg, J.J. (1992) Surfactant phospholipids - synthesis and storage. *American Journal of Physiology,* 262 (4), L367-L385.

Batenburg, J.J., Longmore, W.J., Klazinga, W. and Van Golde, L.M. (1979) Lysolecithin acyltransferase and lysolecithin: lysolecithin acyltransferase in adult rat lung alveolar type II epithelial cells. *Biochim Biophys Acta,* 573 (1), 136-144.

Bauer, C., Cramer, R. and Schuchhardt, J. (2011) Evaluation of peak-picking algorithms for protein mass spectrometry. *Methods Mol Biol,* 696, 341-352.

Baughman, R.P., Stein, E., Macgee, J., Rashkin, M. and Sahebjami, H. (1984) Changes in fatty-acids in phospholipids of the bronchoalveolar fluid in bacterial pneumonia and in adult respiratory-distress syndrome. *Clinical Chemistry,* 30 (4), 521-523.

Baughman, R.P., Sternberg, R.I., Hull, W., Buchsbaum, J.A. and Whitsett, J. (1993) Decreased surfactant protein-a in patients with bacterial pneumonia. *American Review of Respiratory Disease,* 147 (3), 653-657.

Berk, M., Ebbels, T. and Montana, G. (2011) A statistical framework for biomarker discovery in metabolomic time course data. *Bioinformatics,* 27 (14), 1979-1985.

Berndt, P., Hobohm, U. and Langen, H. (1999) Reliable automatic protein identification from matrix-assisted laser desorption/ionization mass spectrometric peptide fingerprints. *Electrophoresis,* 20 (18), 3521-3526.

Bernhard, W., Pynn, C.J., Jaworski, A., Rau, G.A., Hohlfeld, L.M., Freihorst, J., Poets, C.F., Stoll, D. and Postle, A.D. (2004) Mass spectrometric analysis of surfactant metabolism in human volunteers using deuteriated choline. *American Journal of Respiratory and Critical Care Medicine,* 170 (1), 54-58.

Bernhard, W., Wang, J.Y., Tschernig, T., Tummler, B., Hedrich, H.J. and Vonderhardt, H. (1997) Lung surfactant in a cystic fibrosis animal model: increased alveolar phospholipid pool size without altered composition and surface tension function in cftr(m1HGU/m1HGU) mice. *Thorax,* 52 (8), 723-730.

Bligh, E.G. and Dyer, W.J. (1959) A rapid method of total lipid extraction and purification. *Can J Biochem Physiol,* 37 (8), 911-917.

Bohlin, K., Patterson, B.W., Spence, K.L., Merchak, A., Zozobrado, J.C.G., Zimmermann, L.J.I., Carnielli, V.P. and Hamvas, A. (2005) Metabolic kinetics of pulmonary surfactant in newborn infants using endogenous stable isotope techniques. *Journal of Lipid Research,* 46 (6), 1257-1265.

Boscoe, M.J., Goodwin, A.T., Amrani, M. and Yacoub, M.H. (2000) Endothelins and the lung. *International Journal of Biochemistry & Cell Biology,* 32 (1), 41-62.

Bousquet, J., Chanez, P., Lacoste, J.Y., Barneon, G., Ghavanian, N., Enander, I., Venge, P., Ahlstedt, S., Simonylafontaine, J., Godard, P. and Michel, F.B. (1990) Eosinophilic inflammation in asthma. *New England Journal of Medicine,* 323 (15), 1033-1039.

Breit, M., Netzer, M., Weinberger, K.M. and Baumgartner, C. (2015) Modeling and classification of kinetic patterns of dynamic metabolic biomarkers in physical activity. *Plos Computational Biology,* 11 (8), 30.

Bunt, J.E.H., Zimmermann, L.J., Wattimena, J.L.D., Van Beek, R.H., Sauer, P.J. and Carnielli, V.P. (1998) Endogenous surfactant turnover in preterm infants measured with stable isotopes. *American Journal of Respiratory and Critical Care Medicine,* 157 (3), 810-814.

Butler, P.L. and Mallampalli, R.K. (2010) Cross-talk between remodeling and de novo pathways maintains phospholipid balance through ubiquitination. *Journal of Biological Chemistry,* 285 (9), 6246-6258.

Cabre, E.J., Malmstrom, J., Sutherland, D., Perez-Gil, J. and Otzen, D.E. (2009) Surfactant protein SP-B strongly modifies surface collapse of phospholipid vesicles: insights from a quartz crystal microbalance with dissipation. *Biophys J,* 97 (3), 768-776.

Caesar, P.A., Mcelroy, M.C., Kelly, F.J., Normand, I.C.S. and Postle, A.D. (1991) Mechanisms of phosphatidylcholine acyl remodeling by human fetal lung. *American Journal of Respiratory Cell and Molecular Biology,* 5 (4), 363-370.

Calfee, C.S., Gallagher, D., Abbott, J., Thompson, B.T., Matthay, M.A. and Network, N.A. (2012) Plasma angiopoietin-2 in clinical acute lung injury: Prognostic and pathogenetic significance. *Critical Care Medicine,* 40 (6), 1731-1737.

Carr, S.A., Huddleston, M.J. and Bean, M.F. (1993) Selective identification and differentiation of n-linked and o-linked oligosaccharides in glycoproteins by liquid-chromatography mass-spectrometry. *Protein Science,* 2 (2), 183-196.

## Bibliography


Cavicchioli, P., Zimmermann, L.J.I., Cogo, P.E., Badon, T., Giordano, G., Torresin, M., Zacchello, F. and Carnielli, V.P. (2001) Endogenous surfactant turnover in preterm infants with respiratory distress syndrome studied with stable isotope lipids. *American Journal of Respiratory and Critical Care Medicine,* 163 (1), 55-60.

Cederqvist, K., Sorsa, T., Tervahartiala, T., Maisi, P., Reunanen, K., Lassus, P. and Andersson, S. (2001) Matrix metalloproteinases-2,-8, and-9 and TIMP-2 in tracheal aspirates from preterm infants with respiratory distress. *Pediatrics,* 108 (3), 686-692.

Chen, X., Hyatt, B.A., Mucenski, M.L., Mason, R.J. and Shannon, J.M. (2006) Identification and characterization of a lysophosphatidylcholine acyltransferase in alveolar type II cells. *Proceedings of the National Academy of Sciences of the United States of America,* 103 (31), 11724-11729.

Chen, Y.W.R., Leung, J.M. and Sin, D.D. (2016) A systematic review of diagnostic biomarkers of copd exacerbation. *Plos One,* 11 (7).

Cheng, I.W., Ware, L.B., Greene, K.E., Nuckton, T.J., Eisner, M.D. and Matthay, M.A. (2003) Prognostic value of surfactant proteins A and D in patients with acute lung injury. *Critical Care Medicine,* 31 (1), 20-27.

Chida, N. and Adams, F.H. (1967) Incorporation of palmitate, glucose and choline into lecithin by fetal and newborn lamb lung. *Pediat Res,* 1 ((5)), 364-371.

Cobelli, C., Toffolo, G., Bier, D.M. and Nosadini, R. (1987) Models to interpret kinetic data in stable isotope tracer studies. *American Journal of Physiology,* 253 (5), E551-E564.

Cogo, P.E., Carnielli, V.P., Bunt, J.E.H., Badon, T., Giordano, G., Zacchello, F., Sauer, P.J.J. and Zimmermann, L.J.I. (1999) Endogenous surfactant metabolism in critically ill infants measured with stable isotope labeled fatty acids. *Pediatric Research,* 45 (2), 242-246.

Cogo, P.E., Facco, M., Simonato, M., De Luca, D., De Terlizi, F., Rizzotti, U., Verlato, G., Bellagamba, M.P. and Carnielli, V.P. (2011) Pharmacokinetics and clinical predictors of surfactant redosing in respiratory distress syndrome. *Intensive Care Medicine,* 37 (3), 510-517.

Cogo, P.E., Gucciardi, A., Traldi, U., Hilkert, A.W., Verlato, G. and Carnielli, V. (2005) Measurement of pulmonary surfactant disaturated-phosphatidylcholine synthesis in human infants using deuterium incorporation from body water. *Journal of Mass Spectrometry,* 40 (7), 876-881.

Cogo, P.E., Ori, C., Simonato, M., Verlato, G., Isak, I., Hamvas, A. and Carnielli, V.P. (2009) Metabolic precursors of surfactant disaturated-phosphatidylcholine in preterms with respiratory distress. *J Lipid Res,* 50 (11), 2324-2331.

Collins, J.R., Edwards, B.R., Fredricks, H.F. and Van Mooy, B.A. (2016) LOBSTAHS: An adduct-based lipidomics strategy for discovery and identification of oxidative stress biomarkers. *Anal Chem,* 88 (14), 7154-7162.

Coombes, K.R., Tsavachidis, S., Morris, J.S., Baggerly, K.A., Hung, M.C. and Kuerer, H.M. (2005) Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics,* 5 (16), 4107-4117.

Coskun, O. (2016) Separation techniques: Chromatography. *Northern clinics of Istanbul,* 3 (2), 156-160.

D'amato, G., Vitale, C., Molino, A., Stanziola, A., Sanduzzi, A., Vatrella, A., Mormile, M., Lanza, M., Calabrese, G., Antonicelli, L. and D'amato, M. (2016) Asthma-related deaths. *Multidisciplinary Respiratory Medicine,* 11, 5.

Dehoffmann, E. (1996) Tandem mass spectrometry: A primer. *Journal of Mass Spectrometry,* 31 (2), 129-137.

Di Girolamo, F., Lante, I., Muraca, M. and Putignani, L. (2013) The role of mass spectrometry in the "omics" era. *Current organic chemistry,* 17 (23), 2891-2905.

Dobbs, L.G., Wright, J.R., Hawgood, S., Gonzalez, R., Venstrom, K. and Nellenbogen, J. (1987) Pulmonary surfactant and its components inhibit secretion of phosphatidylcholine from cultured rat alveolar type II cells. *Proc Natl Acad Sci U S A,* 84 (4), 1010-1014.

Dorcely, B., Katz, K., Jagannathan, R., Chiang, S.S., Oluwadare, B., Goldberg, I.J. and Bergman, M. (2017) Novel biomarkers for prediabetes, diabetes, and associated complications. *Diabetes, metabolic syndrome and obesity : targets and therapy,* 10, 345-361.

Du, P., Kibbe, W.A. and Lin, S.M. (2006) Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics,* 22 (17), 2059-2065.

Dushianthan, A., Cusack, R., Grocott, M.P.W. and Postle, A.D. (2018) Abnormal liver phosphatidylcholine synthesis revealed in patients with acute respiratory distress syndrome. *Journal of Lipid Research,* 59 (6), 1034-1045.

Dushianthan, A., Goss, V., Cusack, R., Grocott, M.P.W. and Postle, A.D. (2014) Altered molecular specificity of surfactant phosphatidycholine synthesis in patients with acute respiratory distress syndrome. *Respiratory Research,* 15.

Eibl, G., Bernardo, K., Koal, T., Ramsay, S.L., Weinberger, K.M. and Graber, A. (2008) Isotope correction of mass spectrometry profiles. *Rapid Commun Mass Spectrom,* 22 (14), 2248-2252.

Eisner, M.D., Parsons, P., Matthay, M.A., Ware, L., Greene, K. and Acute Resp Distress Syndrome, N. (2003) Plasma surfactant protein levels and clinical outcomes in patients with acute lung injury. *Thorax,* 58 (11), 983-988.

Ejsing, C.S., Duchoslav, E., Sampaio, J., Simons, K., Bonner, R., Thiele, C., Ekroos, K. and Shevchenko, A. (2006) Automated identification and quantification of glycerophospholipid molecular species by multiple precursor ion scanning. *Anal Chem,* 78 (17), 6202-6214.

Facco, M., Nespeca, M., Simonato, M., Isak, I., Verlato, G., Ciambra, G., Giorgetti, C., Carnielli, V.P. and Cogo, P.E. (2014) In Vivo effect of pneumonia on surfactant disaturated-phosphatidylcholine kinetics in newborn infants. *Plos One,* 9 (12).

Fan, E., Brodie, D. and Slutsky, A.S. (2018) Acute respiratory distress syndrome advances in diagnosis and treatment. *Jama-Journal of the American Medical Association,* 319 (7), 698-710.

Finehout, E.J. and Lee, K.H. (2004) An introduction to mass spectrometry applications in biological research. *Biochemistry and Molecular Biology Education,* 32 (2), 93-100.

Foster, C.D., Zhang, P.X., Gonzales, L.W. and Guttentag, S.H. (2003) In vitro surfactant protein B deficiency inhibits lamellar body formation. *Am J Respir Cell Mol Biol,* 29 (2), 259-266.

Foster, D.M., Barrett, P.H., Toffolo, G., Beltz, W.F. and Cobelli, C. (1993) Estimating the fractional synthetic rate of plasma apolipoproteins and lipids from stable isotope data. *J Lipid Res,* 34 (12), 2193-2205.

Foster, D.W. and Bloom, B. (1963) The synthesis of fatty acids by rat liver slices in tritiated water. *The Journal of biological chemistry,* 238, 888-892.

Garlick, P.J., Mcnurlan, M.A., Essen, P. and Wernerman, J. (1994) Measurement of tissue protein-synthesis rates in-vivo - a critical analysis of contrasting methods. *American Journal of Physiology,* 266 (3), E287-E297.

Gibellini, F. and Smith, T.K. (2010) The Kennedy pathway—De novo synthesis of phosphatidylethanolamine and phosphatidylcholine. *IUBMB Life,* 62 (6), 414-428.

Glasser, J.R. and Mallampalli, R.K. (2012) Surfactant and its role in the pathobiology of pulmonary infection. *Microbes Infect,* 14 (1), 17-25.

Gluck, L. and Kulovich, M.V. (1973) Lecithin/sphingomyelin ratios in amniotic-fluid in normal and abnormal pregnancy. *American Journal of Obstetrics and Gynecology,* 115 (4), 539-546.

Gordin, A., Fialkov, A.B. and Amirav, A. (2008) Classical electron ionization mass spectra in gas chromatography/mass spectrometry with supersonic molecular beams. *Rapid Commun Mass Spectrom,* 22 (17), 2660-2666.

Goss, V., Hunt, A.N. and Postle, A.D. (2013) Regulation of lung surfactant phospholipid synthesis and metabolism. *Biochim Biophys Acta,* 1831 (2), 448-458.

Gray, R.D., Downey, D. and Taggart, C.C. (2017) Biomarkers to monitor exacerbations in cystic fibrosis. *Expert Rev Respir Med,* 11 (4), 255-257.

Griese, M. (1999) Pulmonary surfactant in health and human lung diseases: state of the art. *Eur Respir J,* 13 (6), 1455-1476.

Griffiths, J. (2008) A brief history of mass spectrometry. *Analytical Chemistry,* 80 (15), 5678-5683.

Gunther, A., Ruppert, C., Schmidt, R., Markart, P., Grimminger, F., Walmrath, D. and Seeger, W. (2001) Surfactant alteration and replacement in acute respiratory distress syndrome. *Respiratory Research,* 2 (6), 353-U352.

Haimi, P., Uphoff, A., Hermansson, M. and Somerharju, P. (2006) Software tools for analysis of mass spectrometric lipidome data. *Anal Chem,* 78 (24), 8324-8331.

Halliday, H.L. (2008) Surfactants: past, present and future. *J Perinatol,* 28 Suppl 1, S47-56.

Harkewicz, R. and Dennis, E.A. (2011) Applications of mass spectrometry to lipids and membranes. *Annual review of biochemistry,* 80, 301-325.

Hartler, J., Trotzmuller, M., Chitraju, C., Spener, F., Kofeler, H.C. and Thallinger, G.G. (2011) Lipid Data Analyzer: unattended identification and quantitation of lipids in LC-MS data. *Bioinformatics,* 27 (4), 572-577.

Hawgood, S. (2004) Surfactant protein B: Structure and function. *Biology of the Neonate,* 85 (4), 285-289.

Herzog, R., Schuhmann, K., Schwudke, D., Sampaio, J.L., Bornstein, S.R., Schroeder, M. and Shevchenko, A. (2012) LipidXplorer: a software for consensual cross-platform lipidomics. *PLoS One,* 7 (1), e29851.

Ho, C.S., Lam, C.W.K., Chan, M.H.M., Cheung, R.C.K., Law, L.K., Lit, L.C.W., Ng, K.F., Suen, M.W.M. and Tai, H.L. (2003) Electrospray ionisation mass spectrometry: principles and clinical applications. *The Clinical biochemist. Reviews,* 24 (1), 3-12.

Hobi, N., Giolai, M., Olmeda, B., Miklavc, P., Felder, E., Walther, P., Dietl, P., Frick, M., Perez-Gil, J. and Haller, T. (2016) A small key unlocks a heavy door: The essential function of the small hydrophobic proteins SP-B and SP-C to trigger adsorption of pulmonary surfactant lamellar bodies. *Biochimica Et Biophysica Acta-Molecular Cell Research,* 1863 (8), 2124-2134.

Hook, G.E.R. (1991) Alveolar proteinosis and phospholipidoses of the lungs. *Toxicologic Pathology,* 19 (4), 482-513.

Houjou, T., Yamatani, K., Imagawa, M., Shimizu, T. and Taguchi, R. (2005) A shotgun tandem mass spectrometric analysis of phospholipids with normal-phase and/or reverse-phase liquid chromatography/electrospray ionization mass spectrometry. *Rapid Commun Mass Spectrom,* 19 (5), 654-666.

Hsu, F.-F. and Turk, J. (2001) Studies on phosphatidylglycerol with triple quadrupole tandem mass spectrometry with electrospray ionization: Fragmentation processes and structural characterization. *Journal of the American Society for Mass Spectrometry,* 12 (9), 1036-1043.

Huang, Y., Zaas, A.K., Rao, A., Dobigeon, N., Woolf, P.J., Veldman, T., Oien, N.C., Mcclain, M.T., Varkey, J.B., Nicholson, B., Carin, L., Kingsmore, S., Woods, C.W., Ginsburg, G.S. and Hero, A.O., 3rd (2011) Temporal dynamics of host molecular responses differentiate symptomatic and asymptomatic influenza a infection. *PLoS Genet,* 7 (8), e1002234.

Husen, P., Tarasov, K., Katafiasz, M., Sokol, E., Vogt, J., Baumgart, J., Nitsch, R., Ekroos, K. and Ejsing, C.S. (2013) Analysis of lipid experiments (ALEX): a software framework for analysis of high-resolution shotgun lipidomics data. *PLoS One,* 8 (11), e79736.

Ikegami, M. (2006) Surfactant catabolism. *Respirology,* 11, S24-S27.

Il Yoon, H. and Sin, D.D. (2011) Biomarkers of therapeutic response in patients with chronic obstructive pulmonary disease a critical review of the literature. *Drugs,* 71 (14), 1821-1837.

Jabaudon, M., Blondonnet, R., Roszyk, L., Bouvier, D., Audard, J., Clairefond, G., Fournier, M., Marceau, G., Dechelotte, P., Pereira, B., Sapin, V. and Constantin, J.M. (2015) Soluble receptor for advanced glycation end-products predicts impaired alveolar fluid clearance in acute respiratory distress syndrome. *American Journal of Respiratory and Critical Care Medicine,* 192 (2), 191-199.

Jacobs, H., Jobe, A., Ikegami, M. and Jones, S. (1982) Surfactant phosphatidylcholine source, fluxes, and turnover times in 3-day-old, 10-day-old, and adult-rabbits. *Journal of Biological Chemistry,* 257 (4), 1805-1810.

Jobe, A., Kirkpatrick, E. and Gluck, L. (1978) Labeling of phospholipids in surfactant and subcellular-fractions of rabbit lung. *Journal of Biological Chemistry,* 253 (11), 3810-3816.

Johnson, J.V., Yost, R.A., Kelley, P.E. and Bradford, D.C. (1990) Tandem-in-space and tandem-in-time mass-spectrometry - triple quadrupoles and quadrupole ion traps. *Analytical Chemistry,* 62 (20), 2162-2172.

Karas, M. and Kruger, R. (2003) Ion formation in MALDI: The cluster ionization mechanism. *Chemical Reviews,* 103 (2), 427-439.

Kennedy, E.P. and Weiss, S.B. (1956) The function of cytidine coenzymes in the biosynthesis of phospholipides. *J Biol Chem,* 222 (1), 193-214.

Khosla, S.S., Gobran, L.I. and Rooney, S.A. (1980) Stimulation of phosphatidylcholine synthesis by 17beta-estradiol in fetal rabbit lung. *Biochimica Et Biophysica Acta,* 617 (2), 282-290.

Kim, H., Ellis, A.K., Fischer, D., Noseworthy, M., Olivenstein, R., Chapman, K.R. and Lee, J. (2017) Asthma biomarkers in the age of biologics. *Allergy, Asthma & Clinical Immunology,* 13 (1), 48.

Li, S., Guo, Y.L., Zhao, X., Zhang, Y., Zhu, C.G., Wu, N.Q., Xu, R.X., Qing, P., Gao, Y., Li, X.L., Sun, J., Liu, G., Dong, Q. and Li, J.J. (2017) Novel and traditional lipid-related biomarkers and their combinations in predicting coronary severity. *Scientific Reports,* 7.

Lin, L., Yu, Q., Yan, X., Hang, W., Zheng, J., Xing, J. and Huang, B. (2010) Direct infusion mass spectrometry or liquid chromatography mass spectrometry for human metabonomics? A serum metabonomic study of kidney cancer. *Analyst,* 135 (11), 2970-2978.

Lipid Maps (2018) *LIPID MAPS Databases*. Available from: http://www.lipidmaps.org/data/databases.html.

Lipidbank (2007) LipidBank. Available from: http://lipidbank.jp.

Lomas, D.A., Silverman, E.K., Edwards, L.D., Locantore, N.W., Miller, B.E., Horstman, D.H., Tal-Singer, R. and Evaluation, C.L.I. (2009) Serum surfactant protein D is steroid sensitive and associated with exacerbations of COPD. *European Respiratory Journal,* 34 (1), 95-102.

Máca, J., Jor, O., Holub, M., Sklienka, P., Burša, F., Burda, M., Janout, V. and Ševčík, P. (2017) Past and Present ARDS Mortality Rates: A Systematic Review. *Respiratory Care,* 62 (1), 113-122.

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

Mathers, C.D. and Loncar, D. (2006) Projections of global mortality and burden of disease from 2002 to 2030. *Plos Medicine,* 3 (11).

Mclafferty, F. (1981) Tandem mass spectrometry. *Science,* 214 (4518), 280-287.

Meng, D., Zhang, Q., Gao, X., Wu, S. and Lin, G. (2014) LipidMiner: a software for automated identification and quantification of lipids from multiple liquid chromatography/mass spectrometry data files. *Rapid Commun Mass Spectrom,* 28 (8), 981-985.

Miakotina, O.L., Mccoy, D.M., Shi, L., Look, D.C. and Mallampalli, R.K. (2007) Human adenovirus modulates surfactant phospholipid trafficking. *Traffic,* 8 (12), 1765-1777.

Miranda, C., Busacker, A., Balzar, S., Trudeau, J. and Wenzel, S.E. (2004) Distinguishing severe asthma phenotypes: Role of age at onset and eosinophilic inflammation. *Journal of Allergy and Clinical Immunology,* 113 (1), 101-108.

Naroji, S., Aramini, B., Kim, C., Chan, R., Zhou, B., Raza, K., Robbins, H., Shah, L., Lederer, D., Arcasoy, S., Sonett, J., Di Paolo, G. and D'ovidio, F. (2015) Surfactant phospholipids as a marker of chronic lung allograft dysfunction: A targeted lipidomics approach. *The Journal of Heart and Lung Transplantation,* 34 (4), S255-S256.

Ni, J.S., Pomerantz, S.C., Rozenski, J., Zhang, Y.H. and Mccloskey, J.A. (1996) Interpretation of oligonucleotide mass spectra for determination of sequence using electrospray ionization and tandem mass spectrometry. *Analytical Chemistry,* 68 (13), 1989-1999.

Ohnishi, H., Yokoyama, A., Kondo, K., Hamada, H., Abe, M., Nishimura, K., Hiwada, K. and Kohno, N. (2002) Comparative study of KL-6, surfactant protein-A, surfactant protein-D, and monocyte chemoattractant protein-1 as serum markers for interstitial lung diseases. *American Journal of Respiratory and Critical Care Medicine,* 165 (3), 378-381.

Olmeda, B., Martinez-Calle, M. and Perez-Gil, J. (2017a) Pulmonary surfactant metabolism in the alveolar airspace: Biogenesis, extracellular conversions, recycling. *Ann Anat,* 209, 78-92.

Olmeda, B., Martinez-Calle, M. and Perez-Gil, J. (2017b) Pulmonary surfactant metabolism in the alveolar airspace: Biogenesis, extracellular conversions, recycling. *Annals of Anatomy-Anatomischer Anzeiger,* 209, 78-92.

Page, J.S., Marginean, I., Baker, E.S., Kelly, R.T., Tang, K. and Smith, R.D. (2009) Biases in ion transmission through an electrospray ionization-mass spectrometry capillary inlet. *Journal of the American Society for Mass Spectrometry,* 20 (12), 2265-2272.

Park, W.Y., Goodman, R.B., Steinberg, K.P., Ruzinski, J.T., Radella, F., Park, D.R., Pugin, J., Skerrett, S.J., Hudson, L.D. and Martin, T.R. (2001) Cytokine balance in the lungs of patients with acute respiratory distress syndrome. *American Journal of Respiratory and Critical Care Medicine,* 164 (10), 1896-1903.

Petkovic, M., Schiller, J., Muller, J., Muller, M., Arnold, K. and Arnhold, J. (2001) The signal-to-noise ratio as the measure for the quantification of lysophospholipids by matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry. *Analyst,* 126 (7), 1042-1050.

Postle, A.D., Heeley, E.L. and Wilton, D.C. (2001) A comparison of the molecular species compositions of mammalian lung surfactant phospholipids. *Comparative Biochemistry and Physiology a-Molecular and Integrative Physiology,* 129 (1), 65-73.

Postle, A.D., Henderson, N.G., Koster, G., Clark, H.W. and Hunt, A.N. (2011) Analysis of lung surfactant phosphatidylcholine metabolism in transgenic mice using stable isotopes. *Chem Phys Lipids,* 164 (6), 549-555.

Postle, A.D. and Hunt, A.N. (2009) Dynamic lipidomics with stable isotope labelling. *J Chromatogr B Analyt Technol Biomed Life Sci,* 877 (26), 2716-2721.

Postle, A.D., Madden, J., Clark, G.T. and Wright, S.M. (2004) Electrospray ionisation mass spectrometry analysis of differential turnover of phosphatidylcholine by human blood leukocytes. *Phys. Chem. Chem. Phys.,* 6 (5), 1018-1021.

Pulfer, M. and Murphy, R.C. (2003) Electrospray mass spectrometry of phospholipids. *Mass Spectrometry Reviews,* 22 (5), 332-364.

Ramakrishnan, R. (2007) Alternative equations for whole-body protein synthesis and for fractional synthetic rates of proteins. *Metabolism-Clinical and Experimental,* 56 (11), 1550-1560.

Rantalainen, M., Cloarec, O., Ebbels, T.M., Lundstedt, T., Nicholson, J.K., Holmes, E. and Trygg, J. (2008) Piecewise multivariate modelling of sequential metabolic profiling data. *BMC Bioinformatics,* 9, 105.

Ravipati, S., Baldwin, D.R., Barr, H.L., Fogarty, A.W. and Barrett, D.A. (2015) Plasma lipid biomarker signatures in squamous carcinoma and adenocarcinoma lung cancer patients. *Metabolomics,* 11 (6), 1600-1611.

Rosenberg, H.F., Phipps, S. and Foster, P.S. (2007) Eosinophil trafficking in allergy and asthma. *Journal of Allergy and Clinical Immunology,* 119 (6), 1303-1310.

Samuelsson, J., Dalevi, D., Levander, F. and Rognvaldsson, T. (2004) Modular, scriptable and automated analysis tools for high-throughput peptide mass fingerprinting. *Bioinformatics,* 20 (18), 3628-3635.

Sato, H., Callister, M.E.J., Mumby, S., Quinlan, G.J., Welsh, K.I., Dubois, R.M. and Evans, T.W. (2004) KL-6 levels are elevated in plasma from patients with acute respiratory distress syndrome. *European Respiratory Journal,* 23 (1), 142-145.

Schmidt, R., Markart, P., Ruppert, C., Wygrecka, M., Kuchenbuch, T., Walmrath, D., Seeger, W. and Guenther, A. (2007) Time-dependent changes in pulmonary surfactant function and composition in acute respiratory distress syndrome due to pneumonia or aspiration. *Respiratory Research,* 8, 11.

Shaner, R.L., Allegood, J.C., Park, H., Wang, E., Kelly, S., Haynes, C.A., Sullards, M.C. and Merrill, A.H., Jr. (2009) Quantitative analysis of sphingolipids for lipidomics using triple quadrupole and quadrupole linear ion trap mass spectrometers. *Journal of lipid research,* 50 (8), 1692-1707.

Shaw, D.E., Berry, M.A., Hargadon, B., Mckenna, S., Shelley, M.J., Green, R.H., Brightling, C.E., Wardlaw, A.J. and Pavord, I.D. (2007) Association between neutrophilic airway inflammation and airflow limitation in adults with asthma. *Chest,* 132 (6), 1871-1875.

Shevchenko, A., Chernushevich, I., Shevchenko, A., Wilm, M. and Mann, M. (2002) "De novo" sequencing of peptides recovered from in-gel digested proteins by nanoelectrospray tandem mass spectrometry. *Molecular Biotechnology,* 20 (1), 107-118.

Shulenin, S., Nogee, L.M., Annilo, T., Wert, S.E., Whitsett, J.A. and Dean, M. (2004) ABCA3 gene mutations in newborns with fatal surfactant deficiency. *New England Journal of Medicine,* 350 (13), 1296-1303.

Smilde, A.K., Westerhuis, J.A., Hoefsloot, H.C.J., Bijlsma, S., Rubingh, C.M., Vis, D.J., Jellema, R.H., Pijl, H., Roelfsema, F. and Van Der Greef, J. (2010) Dynamic metabolomic data analysis: a tutorial review. *Metabolomics,* 6 (1), 3-17.

Snider, R.K. (2007) Efficient calculation of exact mass isotopic distributions. *J Am Soc Mass Spectrom,* 18 (8), 1511-1515.

Song, H., Hsu, F.F., Ladenson, J. and Turk, J. (2007) Algorithm for processing raw mass spectrometric data to identify and quantitate complex lipid molecular species in mixtures

by data-dependent scanning and fragment ion database searching. *J Am Soc Mass Spectrom,* 18 (10), 1848-1858.

Stanescu, D., Sanna, A., Veriter, C., Kostianev, S., Calcagni, P.G., Fabbri, L.M. and Maestrelli, P. (1996) Airways obstruction, chronic expectoration, and rapid decline of FEV(1) in smokers are associated with increased levels of sputum neutrophils. *Thorax,* 51 (3), 267-271.

Strittmatter, E.F., Rodriguez, N. and Smith, R.D. (2003) High mass measurement accuracy determination for proteomics using multivariate regression fitting: Application to electrospray ionization time-of-flight mass spectrometry. *Analytical Chemistry,* 75 (3), 460-468.

Sweet, D.G., Mcmahon, K.J., Curley, A.E., O'connor, C.M. and Halliday, H.L. (2001) Type I collagenases in bronchoalveolar lavage fluid from preterm babies at risk of developing chronic lung disease. *Archives of Disease in Childhood,* 84 (3), F168-F171.

Tai, Y.C. and Speed, T.P. (2006) A multivariate empirical Bayes statistic for replicated microarray time course data. *Annals of Statistics,* 34 (5), 2387-2412.

Terpstra, M.L., Aman, J., Amerongen, G.P.V. and Groeneveld, A.B.J. (2014) Plasma biomarkers for acute respiratory distress syndrome: a systematic review and meta-analysis. *Critical Care Medicine,* 42 (3), 691-700.

Torresin, M., Zimmermann, L.J.I., Cogo, P.E., Cavicchioli, P., Badon, T., Giordano, G., Zacchello, F., Sauer, P.J.J. and Carnielli, V.P. (2000) Exogenous surfactant kinetics in infant respiratory distress syndrome: A novel method with stable isotopes. *American Journal of Respiratory and Critical Care Medicine,* 161 (5), 1584-1589.

Tsugawa, H., Cajka, T., Kind, T., Ma, Y., Higgins, B., Ikeda, K., Kanazawa, M., Vandergheynst, J., Fiehn, O. and Arita, M. (2015) MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat Methods,* 12 (6), 523-526.

Uchida, T., Shirasawa, M., Ware, L.B., Kojima, K., Hata, Y., Makita, K., Mednick, G., Matthay, Z.A. and Matthay, M.A. (2006) Receptor for advanced glycation end-products is a marker of type I cell injury in acute lung injury. *American Journal of Respiratory and Critical Care Medicine,* 173 (9), 1008-1015.

Vance, D.E. (2014) Phospholipid methylation in mammals: from biochemistry to physiological function. *Biochimica Et Biophysica Acta-Biomembranes,* 1838 (6), 1477-1487.

Vance, D.E., Li, Z.Y. and Jacobs, R.L. (2007) Hepatic phosphatidylethanolamine N-methyltransferase, unexpected roles in animal biochemistry and physiology. *Journal of Biological Chemistry,* 282 (46), 33237-33241.

Veldhuizen, E.J.A. and Haagsman, H.P. (2000) Role of pulmonary surfactant components in surface film formation and dynamics. *Biochimica et Biophysica Acta (BBA) - Biomembranes,* 1467 (2), 255-270.

Voelker, D.R. and Snyder, F. (1979) Subcellular site and mechanism of synthesis of disaturated phosphatidylcholine in alveolar type-ii cell adenomas. *Journal of Biological Chemistry,* 254 (17), 8628-8633.

Weaver, T.E., Na, C.L. and Stahlman, M. (2002) Biogenesis of lamellar bodies, lysosome-related organelles involved in storage and secretion of pulmonary surfactant. *Semin Cell Dev Biol,* 13 (4), 263-270.

Wehofsky, M. and Hoffmann, R. (2002) Automated deconvolution and deisotoping of electrospray mass spectra. *J Mass Spectrom,* 37 (2), 223-229.

Wheelock, C.E., Goss, V.M., Balgoma, D., Nicholas, B., Brandsma, J., Skipp, P.J., Snowden, S., Burg, D., D'amico, A., Horvath, I., Chaiboonchoe, A., Ahmed, H., Ballereau, S., Rossios, C., Chung, K.F., Montuschi, P., Fowler, S.J., Adcock, I.M., Postle, A.D., Dahlen, S.E., Rowe, A., Sterk, P.J., Auffray, C., Djukanovic, R. and Grp, U.B.S. (2013) Application of 'omics technologies to biomarker discovery in inflammatory lung diseases. *European Respiratory Journal,* 42 (3), 802-825.

Whitsett, J.A., Wert, S.E. and Weaver, T.E. (2015) Diseases of pulmonary surfactant homeostasis IN: Abbas, A.K., Galli, S.J. and Howley, P.M. (eds.) *Annual Review of Pathology: Mechanisms of Disease, Vol 10*. 371-393.

Wirtz, H.R. and Dobbs, L.G. (1990) Calcium mobilization and exocytosis after one mechanical stretch of lung epithelial cells. *Science,* 250 (4985), 1266-1269.

Wissel, H., Lehfeldt, A., Klein, P., Muller, T. and Stevens, P.A. (2001) Endocytosed SP-A and surfactant lipids are sorted to different organelles in rat type II pneumocytes. *American Journal of Physiology-Lung Cellular and Molecular Physiology,* 281 (2), L345-L360.

Wolff, M.M. and Stephens, W.E. (1953) A pulsed mass spectrometer with time dispersion. *Review of Scientific Instruments,* 24 (8), 616-617.

Wright, J.R. and Clements, J.A. (1987) Metabolism and turnover of lung surfactant. *American Review of Respiratory Disease,* 136 (2), 426-444.

Wu, Y.H., Xu, Z.W., Henderson, F.C., Ryan, A.J., Yahr, T.L. and Mallampalli, R.K. (2007) Chronic Pseudomonas aeruginosa infection reduces surfactant levels by inhibiting its biosynthesis. *Cellular Microbiology,* 9 (4), 1062-1072.

Yan, F.R., Zhao, H. and Zeng, Y.M. (2018) Lipidomics: a promising cancer biomarker. *Clinical and Translational Medicine,* 7, 3.

Yang, C., He, Z. and Yu, W. (2009) Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis. *BMC Bioinformatics,* 10, 4.

Yergey, J.A. (1983) A general-approach to calculating isotopic distributions for mass-spectrometry. *International Journal of Mass Spectrometry,* 52 (2-3), 337-349.

Young, S.L. and Tierney, D.F. (1972) Dipalmitoyl lecithin secretion and metabolism by rat lung. *American Journal of Physiology,* 222 (6), 1539-&.

Yu, Z., Chen, H., Ai, J., Zhu, Y., Li, Y., Borgia, J.A., Yang, J.-S., Zhang, J., Jiang, B., Gu, W. and Deng, Y. (2017) Global lipidomics identified plasma lipids as novel biomarkers for early detection of lung cancer. *Oncotarget,* 8 (64), 107899-107906.

Zarrouk, A., Debbabi, M., Bezine, M., Karym, E.M., Badreddine, A., Rouaud, O., Moreau, T., Cherkaoui-Malki, M., El Ayeb, M., Nasser, B., Hammami, M. and Lizard, G. (2018) Lipid biomarkers in Alzheimer's disease. *Curr Alzheimer Res,* 15 (4), 303-312.