UNIVERSITY OF MANCHESTER

A Systems Approach to Understanding Dupuytren's Disease

A thesis submitted to The University of Manchester for the degree of Doctor of Philosophy in the Faculty of Engineering and Physical Sciences

March 2011

Samrina Rehman

School of Chemical Engineering and Analytical Science

CONTENTS

List of Figures

List of Tables

Abstract

A Systems approach to understanding Dupuytren's Disease, March 2011
Samrina Rehman, PhD, The University of Manchester

Introduction: Dupuytren's disease (DD) is an ill-defined fibroproliferative disorder affecting the palms of the hands of certain patient groups. Whether changes in DD fibroblasts are due to genetic alterations alone or related to metabolic dysregulation has not yet been investigated.

Hypotheses: 1. DD is a disease of several networks rather than of a single gene. 2. DD may be investigated more effectively by employing systems biology. 3. Strict definition of cell passage number is important for the revelation of any DD phenotype. 4. Some of the differences between DD and healthy tissues reside in a difference in their respiratory metabolism. 5. Any such differences are akin the Warburg effect noted for tumour cells in the literature.

Methods: We induced hypoxia in healthy and disease cells to test whether the difference in disease cell types and healthy is the same as the difference in control fibroblasts cultured in normoxia and hypoxia. We investigated both at the metabolic level (intracellular and extracellular) and at the transcript level. This study also employed Fourier transform infrared spectroscopy to permit profiling of cells: (1) DD cords and nodules against the unaffected transverse palmar fascia (internal control), (2) those (1) with carpal ligamentous fascia (external controls) (3) those in (1) against DD fat surrounding the nodule, and skin overlying the nodule. We then compared metabolic profiles of the above to determine the effect of serial passaging by assessment of reproducibility. Subsequently, a novel protocol was employed in carefully controlled culture conditions for the parallel extraction of the metabolome and transcriptome of DD-derived fibroblasts and control at normoxic and hypoxic conditions to investigate this hypothesis. Gas chromatography-mass spectrometry combined with microarrays was employed to identify metabolites and transcript characteristic for DD tissue phenotypes. The extracellular metabolome was also studied for a selected subset. The metabolic and transcriptional changes were then integrated employing a network approach.

Results: Carefully controlled culture conditions combined with multivariate statistical analyses demonstrated metabolic differences in DD and unaffected transverse palmar fascia, in addition to the external control. Differences between profiles of the four DD tissue phenotypes were also demonstrated. In addition early passage (0-3) metabolic differences were observed where a clear separation pattern in clusters was observed. Subsequent passages (4-6) displayed asynchrony, losing distinction between diseased and non-diseased sample phenotypes. A substantial number of dysregulated metabolites involved in amino acid metabolism, carbohydrate metabolism and also metabolism of cofactors and vitamins including downregulated cysteine and aspartic acid have been identified from the integrative analyses. Metabolic and transcriptional differences were revealed between fibroblast cell samples (passage number 3) cultured in 1% and 21% oxygen. The hypothesis that the difference in disease and healthy cells maybe akin to the differences in healthy cells in normoxia and hypoxia was rejected as only a very small number of significant molecules from these studies coincided in perturbed fascia and disease samples. No lactic acid was observed and little difference in the pyruvate concentrations. Yet, upon perturbation several of these transcripts and metabolites involved in the afore-mentioned pathways were significantly dysregulated.

Conclusion: Early, but not late, passage numbers of primary cells provide representative metabolic and transcript fingerprinting for investigating DD. A unique parallel analysis of transcript and metabolic profiles of DD fibroblasts and control, enabled a robust characterization of DD and correlation of parameters across the various levels of systemic description. The tools that should facilitate our understanding of these complex systems are immature, but the pleiotropy of the difference between health and DD tissue suggest the aetiology of a network-based disease.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning;

Copyright Statement

Dedication

I would like to thank my family - Mum and Jalisha who have been very encouraging, understanding and supportive. This thesis is dedicated to them.

Acknowledgements

First, I would like to thank my Director and supervisor Prof. Hans V Westerhoff for his constant support for my research. I learned how to present my research effectively in writing as well as in conferences/talks from him. He also provided a lot of intellectual framework for my thesis. From the beginning, he has been a constant source of inspiration for my research. It was a big adventure altogether.

Secondly, I am thankful to my co-supervisor Dr. Philip Day who was both a frequent source of extremely valuable advice about both research and writing and provided impartial advice through mentorship. I learned how to be clear and precise in conversation and writing from him. Philip has constantly scrutinised my work and this thesis too, encouraging and enabling me to give and perform my best. My interaction with him has been a strong motivating factor for me to continue and conduct my PhD with diligence; a goal to meet high standards of excellence in research every time.

In particular I am thankful to Dr. Ardeshir Bayat (Ardy) for allowing me to access to his precious biopsies and lab consumables to implement & to pursue this exciting systems biology project. In addition I am thankful to Ardy for teaching me relevant medical biology about Dupuytren's disease. Almost all of the DD biology I know is a generous gift from him. He would constantly forward me interesting articles and challenge my understanding of them.

Third, I acknowledge Prof. Royston Goodacre (Roy) for valuable and engaging discussions and in particular for allowing me access to the FT-IR Spectrometer. This association with Roy continually stimulated my analytical thinking in concrete technical and statistical terms and greatly assisted me in developing skills for analysing complex data sets including chemometrics techniques as well as wet-lab experiments.

Fourth, I acknowledge Prof. Magnus Rattray for agreeing to be part of my study and for his continuous involvement and valuable feedback with respect to microarray data analyses. The use of his novel method; puma, has greatly improved my bioinformatics skills and understanding of R-Bioconductor suite.

Throughout my doctoral studies each mentor has contributed whether in critical or supportive form, I admire their high standards for excellence. Combined, this was a unique mentorship par excellence.

I have been especially fortunate to collaborate with not one, two or three, but five independent research groups. I would like to acknowledge the BBSRC/EPSRC (via the DTC), Dr Ardeshir Bayat and Prof Roy Goodacre for supporting me financially to conduct my research. I am grateful for being affiliated with so many groups, in particular the DTC and would like to thank the team members; Dr. Gerold Baier and Dr Maria Nardelli for their continual support to the DTC students and Ms. Lynne Davies, our administrative associate who we all (students and staff) owe a special note of gratitude for providing all administrative support and making our life easier during the doctoral studies. In addition, former members Dr. Anna-Maria Bevilaqua and Dr. Stephen Wilkinson are to be mentioned for their patience and for allowing us to engage and gain experience from their projects in which I was also added as a co-author into book chapter.

I cannot enumerate all individuals but certain names need to be mentioned. I acknowledge key individuals who provided elements to facilitate in decision making including Dr. Yun Xu (chemometrics) and those who helped to generate the data for the studies involved Dr. Warwick Dunn (GC-MS) and Ms. Fiona Salway (Affymetrix microarray facility). In addition, I acknowledge members that provided free training and generous support in the cell culture lab (Ms. Susan Slack and Ms Fiona Flett).

I also acknowledge the extraordinary support from the University of Manchester for my doctoral work. In particular, I would like to thank the DTC ISBML Program from BBSRC/EPSRC for funding my research. The faculties at UoM collectively have been instrumental in shaping my career and analytical thinking through wonderful courses, training workshops, tutorials, seminars, symposiums and conferences. I developed a diverse skill set in the field of systems biology ranging from wet-dry lab including cell culture, qRT-PCR, FT-IR, GC-MS, statistics, microarray technologies, analysis of complex data sets and network analysis through use of a diverse range of SB tools.

I would also like to thank my advisor Prof. Peter Fielden for his help and support during this time. Several discussions with Peter and Prof Nick Goddard not only helped to resolve situations that were mysteriously tangled. I thank them for their help, advice, valuable discussions in meetings, motivation and support and for the fun conversations. Finally, I extend many thanks to all my friends at UoM and outside who made the PhD experience the way it was meant to be…challenging (& seldom cheerful).

Abbreviations

a-SMA - alpha-smooth muscle actin
ANN - Artificial Neural Networks
ANOVA-PCA - Analysis of Variance-principal component analyses

bFGF - basic fibroblast growth factor
BioPAX- Biological Pathway Exchange

C1 – Fibroblasts from Dupuytren cords cultured in 1% Oxygen
C21 - from Dupuytren cords cultured in 21% Oxygen
CAM - cellular adhesion molecules
CE - capillary electrophoresis
CellML- cell markup language
COL1A1- Collagen type I alpha I
COL3A1 - Collagen, type III, alpha 1
COPASI - Complex Pathway Simulator
CREBs - cAMP-response element binding proteins
CT- Computed tomography
CTD-Carpal tunnel decompression
CVA- Canonical Variate Analysis

DAVID - Database for Annotation, Visualization and Integrated Discovery
DC- Dupuytren's Contracture
DD - Dupuytren's disease
DE - Differentially expressed
DFA- Discriminant function analysis
DMEM3- Dulbecco's Modified Eagle's Medium 3
DNA – Deoxyribonucleic acid
DPBS - Dulbecco's Phosphate Buffered Saline
D-PLS - Discriminant-Partial Least Squares

EASE - expression analysis systematic explorer
ECM- Extra cellular matrix
EGF- Epidermal growth factor
EMSC- Extended Multiplicative Scatter Correction

F1 - Fibroblasts from Transverse palmar fascia cultured in 1% Oxygen
F21 - Fibroblasts from Transverse palmar fascia cultured in 21% Oxygen
FDR - False Discovery Rate
FT-IR- Fourier transform infra-red spectroscopy

GC-MS - Gas chromatography-mass spectrometry
GC-RMA - GeneChip Robust Multi Array
GC-TOF-MS - Gas Chromatography Time-of-Flight Mass Spectrometry
GO – Gene Ontology

HCA - Hierarchical Cluster Analysis
HLA-DRB1*15- Major histocompatibility complex, class II, DR beta 1 (aka humanleucocyte antigens)
HMDB - Human metabolome database
H&E - haematoxylin and eosin

IL-1- Interleukin 1
Imaging MS - imaging mass spectrometry
IPA - Ingenuity Pathway Analysis

JWS - Java Web Simulation

KEGG-Kyoto Encyclopedia of Genes and Genomes

LC - liquid chromatography
LC-MS - Liquid chromatography-mass spectrometry
LDA - Linear Discriminant Analysis

Limma – Linear Models for Microarray Data
LOD score - Logarithm of the odds score

MALDI-TOF-MS - Matrix-assisted laser desorption ionization time-of-flight mass spectrometry
MATLAB - Matrix laboratory
MetPA - Metabolomics Pathway Analysis
MMD - Manchester Metabolomics Database
MMP- Metalloproteinases
MMP14 - Matrix metallopeptidase 14 (membrane-inserted)
MafB - Gene- v-maf musculoaponeurotic fibrosarcoma oncogene homolog B
MRI - Magnetic resonance imaging
mRNA- Messenger ribonucleic acid
MS – Mass spectrometer
MVA- Multivariate Data Analysis
MW- Molecular weight

N1 - Dupuytren nodule fibroblasts cultured in 1% Oxygen
N21 - Dupuytren nodule fibroblasts cultured in 21% Oxygen
NCBI- National Centre for Biotechnology Information
NEAA- Non-essential amino acids
NIH - National Institutes of Health
NIPALS- Non-linear Iterative Partial Least Squares
NMR - Nuclear magnetic resonance

PARAFAC - Parallel Factor Analysis
PBS - Phosphate buffered saline
PCs - Principal components
PCA - Principal component analysis
PC-DFA - Principal component-discriminant function analysis
PCR - Polymerase Chain Reaction
PDGF - Platelet derived growth factor
PET- Positron emission tomography
PGA - Program for Genomic Applications
PLS - Partial least squares
pO2 - Oxygen partial pressure
PPLR - probability of positive log-ratio
Puma - Propagating Uncertainty in Microarray Analysis
PyChem - Python and Chemometrics

qRT- PCR- Quantitative Real-time Polymerase Chain Reaction

RMA - Robust Multi Array
RNA- Ribonucleic acid
ROR2 proteins - receptor tyrosine kinase-like orphan receptor 2
ROS - reactive oxygen species
rpm - rotations per minute

S1 - Fibroblasts from skin overlying dupuytren nodules cultured in 1% Oxygen
S21 - Fibroblasts from skin overlying dupuytren nodules cultured in 21% Oxygen
SB – Systems Biology
SBGN- Systems Biology Graphical Notation
SBML - Systems Biology Markup Language
SELDI-TOF-MS- Surface Enhanced Laser Desorption/Ionization Time of Flight Mass Spectrometry
SON- skin overlying nodule

TCA cycle - Tricarboxylic acid cycle
TGF-β- Transforming growth factor beta
TIC - total ion current
TIMP- Tissue Inhibitor of Metalloproteinases
TNC - Tenascin C (hexabrachion)

XTT- tetrazolium salt

Preface / The Author

A bit about me: I obtained MChem (Hons.) Chemistry with Medicinal Chemistry (2004) and MSc. Cheminformatics (2005) from The University of Manchester. Prior to commencing my PhD studies I was employed at Inpharmatica Ltd as a Cheminformatics Associate Scientist. I am an active member of The Royal Society of Chemistry and The Biochemical Society.

Chapter 1

Introduction

This Chapter provides an introduction to the thesis and the motivation to do this research. The study focuses on bringing together two disparate fields; Dupuytren's disease (DD) and Systems Biology (SB). Approximately 180 years have elapsed, since the first literature reference on DD. Conventional research methods have continued to investigate the causes underlying DD formation and progression yet has not identified the precise aetiopathogenesis of this disease. Single gene approaches have not delivered for DD, and DD is a complex disease requiring network analysis. With SB approaches becoming increasingly popular in cancer research, it is timely to implement a novel approach where the two fields may now cross-fertilise. We first review (1.2; Appendix A) how DD may be a network disease, such that a SB approach may help understanding the former. In the latter part of this Chapter research objectives of this study (1.3) and the workflow (1.4) to implement the strategy used to test the hypotheses which follow in later chapters is outlined. Project aims are listed in 1.5.

1.1 About this work

1.1.1 Motivation

Dupuytren's disease (DD) is a benign fibroproliferative tumour of unknown aetiopathogenesis affecting the palm of the hand, which often causes progressive, permanent contracture of the digits [1]. Surgical treatment of DD is symptomatic and associated with a high rate of recurrence which in some cases leads to amputation of the involved digits [2]. Advanced understanding of the mechanisms involved could lead to more-effective and perhaps less invasive therapies. The principal clinical deformity of DD is characterised by a slowly progressive and irreversible flexion of the fingers as a result of decreased distance between the origin and insertion of the palmar fascia. The slowly progressive shortening of the affected digits is termed a contracture [3]. Pathogenesis of DD is characterised by the varying proliferative potential of myofibroblasts, (specialised fibroblasts expressing α-smooth muscle actin (α-SMA)).

The DD process is described by macroscopical investigations of the affected area that demonstrate phenotypic differences between two structurally distinct fibrotic elements (i.e. the nodule and the cord). The nodule is thought to be involved in the most biologically active phase of the disease, characterised by soft-tissue masses containing a dense population of fibroblasts, which are largely myofibroblasts. On the contrary, the cord is relatively avascular and acellular exhibiting a collagen-rich structure that contains a smaller population of myofibroblasts. These abnormal fibroblasts are considered to be responsible for causing the disease [4, 5]. In addition, previous hypotheses yet few, have hinted towards pathological relevance of adipose tissues (the fat cushioning the nodules) [6] and observations from post-dermofasciectomy resulting in lower recurrence rates may indicate a relation to the skin overlying the nodule (SON) [7]. Furthermore, whether any significant changes in DD are due to genetic alterations alone or the consequence of metabolic dysregulation has not to date been demonstrated.

The analysis of transcription has been pivotal in recent analyses aimed at deciphering the control and mechanisms underpinning progression towards a diseased state. In this respect, major advancements have been made. Molecular studies from harvested tissue biopsies and *in-vitro* cultivation of fibroblasts derived from DD tissue have demonstrated presence and dysregulation (over and under expression) of several candidate genes [8-10]. Previous microarray and linkage studies have demonstrated key genes of interest that may potentially be involved in DD pathology. A recent study has investigated the differential gene expression analysis of subcutaneous fat, fascia, and skin overlying the DD nodule in comparison to control tissue [11].

However there still exist huge technical and knowledge limitations that hinder the accumulation of transcriptional data alone to synthesis of information. Further validation is necessary with more sensitive experimental approaches that can be systematically and iteratively investigated and validated with modeling approaches to predict responses with certainty to identify and classify the levels of complexity in disease state of individual DD phenotypes.

SB approaches promise to assist by bridging this deficit, making available novel analytical tools to unravel transcriptional data through the building of models and mathematical predications based on observations. Omics strategies within the framework of a SB context offer platforms such as Fourier Transform-Infrared (FT-IR) spectroscopy (metabolic fingerprints or non-invasive footprint screening) as a potential diagnostic and screening tool, and quantitative real-time polymerase chain reaction (qRT-PCR) for quantifying the transcript levels or making use of microarrays as exploratory tools. Such methods allow cost-effective and rapid methods of detection. SB is one of the most widely discussed fields among the emerging fields of post-genomic disciplines. It applies quantitative, mechanistic modeling to the study of genetic networks, signal transduction pathways and metabolic networks [12-16]; to understand the complexity of biological phenomena at all functional levels in a given cell, organism or tissue. Systems-level approaches are making a definitive pace towards scientific understanding and biotechnological applications [17].

SB for the present largely relates to unicellular organisms, and this study/thesis is one of the first to utilise SB to explore the metabolic characteristics and to advance the understanding of transcript data in a higher organism (i.e. primary cultures from DD (human) biopsies) in the field of medical SB. The data presented in this thesis shall constitute to highly systematic studies for a SB case related to DD.

The purpose of this study is to develop and test our hypothesis that DD is a network disease. This hypothesis has two parts:

(i) The DD and corresponding healthy tissue differ in function through differences between their molecular (and perhaps intercellular) networks, rather than differences in a single molecule, in a plethora of unrelated molecules.

(ii) DD can be caused by any of a variety of perturbations in regulatory networks that lead to the above network differences

Working towards this purpose, I endeavored to develop and engage in a data-driven top-down SB approach, towards the identification of transcriptome and metabolome differences between DD and healthy cells. The aim is to identify some of the mRNAs and some of the metabolites that are different. By projecting the identified molecules onto the existing genome-wide maps, or into functional categories, the objective is to identify pathways that account for differences seen with metabolites and transcripts, separately and in conjunction. In a bottom-up approach, I shall test an additional hypothesis that some of the differences between DD and healthy tissues reside in a difference in their respiratory metabolism. In addition, I shall also examine the hypothesis that any such differences are akin the Warburg effects noted for tumour cells in the literature. Overall the aim is to delineate the above as a SB approach that may also be of use for the analysis of other ill-understood human diseases.

1.1.2 Strategy of the study and its structure in text

This study works towards the above aims in seven more or less sequential steps. Each of these corresponds to a chapter of this thesis. Most steps involve a targeted search for data generating a hypothesis.

The latter part of this Chapter (1.2) reviews how the scientific literature addresses the above aims and to what extent SB approaches have already been implemented. It also develops a perspective on how SB may help DD research and ends in outlining the strategy followed in this study. The philosophy behind this approach and the plan of action is discussed (1.3). Finally, the specific objectives sought in each chapter are outlined followed by a summary of the workflow (1.4). Section 1.2 has been submitted to *Arthritis Research & Therapy* and is currently under review. The background concerning the experimental techniques, design, patient recruitment, sample size and protocols for all materials and methods for Chapters 3-6 are given in Chapter 2.

Because SB attempts to analyse and understand in terms of all the networks and molecules that may affect a system (organism), its approaches tend to be more comprehensive than those of molecular and cell biology in the sense that it accounts for dynamic interactions taking place within a system i.e. the subject in question and its environment. In addition the data generated by SB approaches are likely to be utilised by other groups, either for modeling approaches, or for connecting them to additional experimental data sets (e.g. transcriptome data to proteome data). This makes it important for SB to come to well defined and reproducible experimental (and computational) systems. SB is also more sensitive than molecular biology to differences between the *in vivo* physiological state and experimental *in vitro* states; one is almost forced to work with freshly isolated tissues. However such human tissues are barely accessible, and by virtue of the randomness by which they become available do not lend themselves much to well prepared and reproducible experimentation. An alternative could be the use of cell lines derived from such primary tissues. However, the establishment and maintenance of cell lines inevitably has consequences for the network functioning of the cells. In the case of DD cells and their healthy counterparts for instance, the DD cells may lose their DD characteristics when cultured, or develop differences with normal cell that have little to do with the differences between the diseased and the healthy tissue in the patient and represent positive selection for cells able to survive *in-vitro* culture conditions. In Chapter 3, we examine how DD cells change when taken from tissue and cultured. This will enable definition of which passage of cell culturing would be best to use in systems approaches to DD, as a compromise between retaining *in situ* DD character and having sufficient cell numbers for analysis. We shall investigate this by implementing the concept of a systems signature defined through FT-IR spectroscopy. Using this methodology we shall assess the reproducibility *in vitro* of DD subsets (i.e. nodule, cord, fat & SON) as compared with internal control (transverse palmar fascia) and external controls (carpal ligamentous fascia). To highlight the optimal conditions to further investigate the DD system is the objective of Chapter 3.

SB also tries to integrate possibly all relevant literature data and pre-existing knowledge with new experimental data, all in its attempts to understand network functioning. To this aim, systematic text mining of the literature is important. In Appendix E3(2)(*manuscript in preparation*) we discuss the rationale to develop a novel text mining tool, which may help to depict accurate relationships in bio-networks by using natural language processing approaches. Time constraints have kept us from further implementing this methodology in the present study.

The conclusions from the fingerprint screening study in Chapter 3 are focused upon to investigate whether there are significant differences in functional genomics between DD and healthy cells, and whether a Warburg effect exists in the DD cell. High-throughput studies are used to examine the response to hypoxia at the level of the metabolome and transcriptome. These will be shown to provide a useful - albeit noisy and incomplete - sketch of the cellular network response to hypoxia (or hypoxia reoxygenation) stimulus. Chapters 4 and 5 discuss in detail how this was investigated with the use of omics technologies. In Chapter 4 a metabolomics approach (using gas chromatography-mass spectrometry (GC-MS)) identifies dysregulated metabolites, as well as the pathways in which they occur. A first question we shall address is whether the difference in disease cell types (nodule, cord and SON) and control cell type is the same as the difference in control fibroblasts cultured in normoxia and hypoxia. Second, we shall ask in which specific disease cell type (nodule, cord, and/or SON) is the difference with normal cells in intracellular metabolome the largest? In addition, the extent of the Warburg effect is also tested when hypoxia is induced in the disease cell types. We surmise that not only may DD cells have a Warburg phenotype; they may also have a different propensity to assume such phenotype when exposed to hypoxia. In addition, the external stimuli may induce pathway inhibitors - clues can be used to investigate and demonstrate maximal differential response in a number of DD cell cultures.

In Chapter 5 the disease cell types (nodule) which demonstrated the most significant differences compared with normal and perturbed fascia, were then selected to investigate the Warburg effect in the DD transcriptome. Candidate genes were identified and the question addressed here is whether the identified transcripts/genes and metabolites show congruence across the various levels of systemic description in the context of pathways i.e. the DD metabolome and transcriptome. Integration of these complex data sets is a key challenge and is the objective of the small Chapter 6, which also involves the mapping onto known genome wide pathway maps.

For accurate interpretation of the intracellular dynamics within the DD cell metabolome and transcriptome it is essential to harvest these molecules in identical conditions. Current SOP's do not permit the harvesting and extraction of these moieties simultaneously. This necessitated developing a novel protocol to meet this requirement. The aim here was the simultaneous collections at one time point of the three moieties from all samples i.e. the extracellular metabolome, the intracellular metabolome and RNA. A robust standard operating protocol for three simultaneous sample extractions is given and applied for the first time. This novel protocol is discussed in Chapter 2.3.2 - 2.3.3.

The ultimate goal of this work would be the initiation of a system-wide, data-driven model that includes all genes, enzymes, metabolites, and regulatory proteins that are involved in hypoxia defenses in DD. It is advantageous to decide *a priori* the necessary level of detail and the corresponding modeling strategy. This would then be extended to model metabolism at the level of enzyme fluxes, within the constraint-based framework. The signaling network that modulates these enzymes could then be represented as a map of bidirectional links derived from interaction databases.

Although it is quite a task to build complete representations of these networks from the parts list, factors that define the problem of interest, such as cell type (cord, nodule and control fascia), time window (acute, passage number, metabolome and transcript extraction) and the specific context (hypoxia) help to narrow the number of players. After iterations of experimental validation and refinement, this network model may help to find modulators and targets on which to apply more detailed modeling strategies.

The project encompasses integration of large data-sets from metabolomics and transcriptomics collected from suitable test subjects and matched controls. These data are used as a basis for the construction of probabilistic and statistical models. In each of the studies, the data sets generated have large diversities of characteristics and also have different research objectives including automated high-throughput data analysis, classification and prediction, as well as multivariate pattern comparison. Hence various signal processing and pattern recognition techniques are required to analyse these data sets to reveal their patterns and achieve the objective of research. Several models of the same data are explored including unsupervised methods principal component analysis (PCA), and supervised methods; principal component-discriminant function analyses (PC-DFA) and Analysis of Variance-principal component analyses (ANOVA-PCA). These are described in their respective chapters and background is given is Chapter 2.2. The prediction of each of these models to the same training data set is compared and further experiments conducted on new disease & control cell cultures to predict the response of these cells from their system signatures. Appropriate computational techniques have been used to elucidate models of relationships between gene, protein and metabolite signals and cell functional responses to extracellular cues. Elucidated network relationships from candidate subset of metabolites and candidate transcripts resulting from experiments are given in Chapter 6.

The development of a mathematical model can now be initiated to address clinical implications. This type of approach will ultimately provide a model and rationale for the development of new, dynamic and dual therapeutic strategies and is likely to influence drug design. The implications of the findings of this project are not only important for DD but also for other fibrotic systems onto which these interactions can be projected. A discussion and conclusion of this thesis is given in Chapter 7 followed by references and appendices.

The directions for possible future work are chosen here towards further improvement of fingerprint screening and validating with text mining approaches, increase on accuracy, mapping and integrating pathways and reduction of anomalous results with targeted approaches. This work was supported by the BBSRC and EPSRC via the Doctoral Training Centre at Manchester Centre for Integrative Systems Biology.

1.2 Dupuytren's – A Systems Biology disease?

The following is an extensive review of the scientific literature that addresses the above aims and to what extent SB approaches have already been implemented. It also develops a perspective on how SB may help DD research and ends in outlining the strategy followed in this study. Because Section 1.2 is a manuscript and has been submitted to a peer-reviewed journal, it consists of its own abstract, introduction, results, discussion and references. Please refer to Appendix A.

1.3 Philosophy of Approach

The rest of this chapter summarises the concepts and methods applied in this research thesis from metabolomics and transcriptomics technologies. The practical problems associated with DD research and the rationale behind this approach and statistical data analysis relevant for this thesis are given.

1.3.1 Establishing an optimum cell system to investigate Dupuytren's disease

Life is structured on many levels of biological organisation. Only in the current post-genomic era, after sequencing of many genomes (http://genomesonline.org/), we are now starting to appreciate the complexity of biological organisation at the cellular level. To understand the function and dysfunction of such complex systems requires integrated, systems-level approaches. To understand complex disease systems this rationale is even more significant.

The differentiation of DD subsets (nodule, cord and uninvolved transverse palmar fascia) has classically been performed on the basis of clinical presentation, histopathology and morphological features and more recently with high density microarrays. However, phenotypic variations within the DD tissue have led to plausible questioning with regards to the classification of this disease. The high degree of macroscopical un-relatedness between fibroblasts from the nodule and cord illustrates this. Within the framework of SB, functional analyses of the cell should be investigated at all 'omics levels in order to determine the underlying causes of DD formation. Previous studies have focused on biochemical factors in isolation; whilst mostly concentrating on tissue biopsies harvested from DD subjects, there has been no work to date on the cascades involved in cellular dysfunctioning within a systems context. More recently few studies have focused on the DD myofibroblasts with the use of high throughput technologies.

It is envisaged the fibroblast is the key cell implicated in cell signaling events leading to DD pathogenesis. One study investigated differences in nodule or cord-derived fibroblasts focusing on their cellular metabolic activity [18]. Controls were obtained from carpal ligaments. The fibroblasts derived from cords presented higher metabolic activity compared with fibroblasts derived from nodules or controls (fascial retinaculum) using an XTT assay. Nodular fibroblasts presented lower activity than those of control fascial fibroblast assays.

Despite accumulating evidence that *in vitro* conditions have an impact on gene expression patterns, there are limited studies that have investigated differential gene expression in both tissue culture and biopsies. To address these questions, Shih et al 2009 investigated the gene expression levels of candidate genes differentially expressed in DD tissue phenotypes including not only the cords, nodule and fascia but also the fat and SON to determine whether the observed results from tissue biopsies were comparable to those from cell cultures, in order to identify potential biomarkers [19].

A number of studies have implicated involvement with the TGF-β pathways when compared with tissues from the DD palmar fascia [20]. These distant sites are thought to be uninvolved in the disease process and may serve as internal controls in addition to tissue biopsies harvested from the palmar fascia of patients having Carpal Tunnel Decompression (CTD). While the transverse palmar fascia (from the DD patient) would be thought a more appropriate control and preferred site for study accounting for a homogeneous study set, correlating to the adjacent diseased sites respectively (if excised/harvested on same day), the CTD fascial tissue from individuals unaffected by DD has been documented in many studies as the common choice for control however this attributes to heterogeneity. Cellular profiles in the Dupuytren tissues (nodules and cords) ultrastructural studies display confirmed differences from the transverse palmar fascia, and although it is not completely understood nor proven that these tissue from the Dupuytren patient are confirmed as 100% normal, macroscopical evidence suggests these may be suitable as internal controls for comparison. For this reason, not only is it acceptable but also intriguing in this study to compare the DD tissue (nodule and cords) with those where the site is confirmed to be uninvolved in the diseased palm.

1.3.1.1 Passage Number Effect on Metabolic Fingerprint

Cell line quality is crucial to successful experimentation and an important step to ensure reliable and reproducible results. Continuous cell lines are increasingly being used and may serve as valuable research tools in medicine and biotechnology [21, 22]. However, the ability of continuous cell lines to exist almost indefinitely has opened the possibility of questionable sub-culturing practices [23, 24] and hence, scientific data produced as a result of experiments performed. The issue has been raised in some recent literature that over-sub-culturing may have a profound impact that can lead to changes in cell lines properties over time. Such changes may occur at high passage numbers where cells may experience alterations in cell morphology, response to stimuli, growth rates, protein expression and signaling, compared to lower passage cells. In addition, the *in vitro* culturing conditions may affect gene expression profiles and/or ultimately cell phenotypes [25, 26].

Altered phenotypes in late passage cultures of the DD nodule have been reported, where the authors suggest late passage cultures of DD nodule fibroblasts display phenotypes similar to those of cord-derived fibroblast [26]. It is difficult to make comparisons from previous studies as the reported results are often based on different cellular passages which could have a dramatic effect on their gene expressions. Previous studies on fibroblasts derived from DD tissues have looked at cellular passages without considering factors that may affect not only morphological changes but gene expression and metabolic differences too. Few studies state the passage number used but do not question nor address why those cellular passages were selected. It is possible that a higher or an inconsistent passage number may lead to production of an adequate biomass, but this has not been investigated or reasoned to date.

Moreover, normal cells undergo a finite number of divisions and then cease dividing after some passages (a process known as replicative senescence), whereas tumour cells are able to proliferate indefinitely [27, 28]. The acquisition of an unlimited proliferative potential has been proposed as one of the critical steps in cancer (neoplastic diseases), which arises as a consequence of the accumulation of multiple independent mutations in genes that regulate cell proliferation and survival. Although differing numbers of passages have been reported in the studies investigating the dupuytren tissue, any change in the proliferative potential of the fibroblast has not been stated. DD fibroblasts may possess a higher potential for matrix and collagen production through passages than control fascia cells because the DD nodules and cords result from an uncontrolled proliferative cellular state. Differences in collagen, fibronectin proteins, matrix expression proteins and even proteoglycans could be affected by passaging because it is thought that at earlier passages all cells would mostly be proliferating while at later passages they would tend to become senescent [29].

1.3.1.2 Whole cell fingerprinting to obtain a system signature

Careful consideration should be given to factors which could affect the reproducibility of data or produce instrument drift including sample preparation, instrument contamination and data processing. Time and order of sample analysis could provide significant sources of variability, potentially obscuring the biological variation which we seek to characterise.

FT-IR spectroscopy is a well-established tool for the identification of transition phenomena in systems passing through different phases. Within the biosciences, the applications of FT-IR have been numerous and diverse. FT-IR spectroscopy can detect and identify endogenous (fingerprint) and secreted (footprint) metabolites. It has proven to rapidly and accurately identify bacteria to the sub-species level [30], differentiate between clinically relevant species [31], and has provided a metabolic footprint of tryptophan-metabolism mutants [32] as well definitively discriminating between a range of bacterial genera. It's use as a rapid, yet non-invasive method for embryonic secretome determination from preliminary results has demonstrated its powerful potential as a diagnostic tool [33].

This approach is based on the principle that when a sample is interrogated with an infrared (IR) beam, the functional groups within the sample will absorb the infrared radiation and vibrate in one of a number of ways, either stretching, bending, deformation or combination vibrations [34]. These absorptions/vibrations can then be correlated directly to (bio)chemical species and the resultant infrared absorption spectrum can be described as an infrared 'fingerprint' characteristic of any chemical or biochemical substance.

In this study FT-IR spectroscopy is employed as a metabolic fingerprinting screen with multivariate statistical techniques for cluster analysis to assess reproducibility of cells. In this respect, FT-IR is used to determine the presence of metabolites and their unique fingerprints using a metabolomic analysis of fibroblast cultures derived from different DD tissue phenotypes (nodule, cord, subcutaneous fat and SON) to compare early (primary) cultures to late passages in order to identify the most representative passage for the disease. Controls are both internal (transverse palmar fascia) and external (CTD palmar fascia or transverse carpal ligamentous fascia) fibroblasts grown in the same culture medium and conditions. It is hypothesised that the culturing conditions may have a profound impact on gene expression and metabolic activities as the cells adapt to a new artificial environment and the most suitable passage representative of the disease remains unknown. Such physicochemical spectroscopic methods are increasingly discussed to have a huge potential in disease diagnostics as these platforms offer/facilitate "whole-organism fingerprinting."

The fingerprint study in this project forms the basis and starting point of the subsequent omics approaches investigated in the concept of a systems signature determination. In addition, the fibroblast cultures grown in same culture media are analysed for any change in cellular properties due to the passage effect. Details of materials and methods are given in Chapter 2.3.1.

1.3.2 Metabolism and Warburg effect

Metabolism is a constitutive process within a cell. There is still much to learn about how cell metabolism is regulated during proliferation. In multicellular organisms, most cells are exposed to a constant supply of nutrients. Survival of the organism requires control systems that prevent aberrant individual cell proliferation when nutrient availability exceeds the levels needed to support cell division. Uncontrolled proliferation is prevented because mammalian cells do not normally take up nutrients from their environment unless stimulated to do so by growth factors. Many cancer cells overcome this growth factor dependence by acquiring genetic mutations that functionally alter receptor-initiated signaling pathways [35]. There is growing evidence that some of these pathways constitutively activate the uptake and metabolism of nutrients that both promote cell survival and fuel cell growth [36].

It is known that tumours alter the metabolic profiles of the cells, which display a higher rate of glucose uptake and glycolytic activity when compared to their benign/normal counterparts [37]. These metabolic changes might confer a common advantage on many different types of cancers, which allows the cells to survive and invade. Another important characteristic in many cancer cells is the increase in utilisation of the anaerobic glycolytic pathway, converting pyruvate to lactate. This fermentation phenotype was initially thought to raise as an adaptive response to the hypoxic microenvironment that tumour cells were forced through an inefficient, disordered and commonly insufficient vascular system [38]. However, not all the tumour cells are under hypoxic stress, and the anaerobic glycolytic pathway is used even if the tumour cells are in presence of oxygen [39], displaying a constitutive alteration in carbon metabolism. This phenomenon is termed as the 'Warburg effect' named after Otto Heinrich Warburg [38].

Warburg found that unlike most normal tissues, cancer cells tend to "ferment" glucose into lactate even in the presence of sufficient oxygen to support mitochondrial oxidative phosphorylation and hence their metabolism is often referred to as "aerobic glycolysis [40]." The excess generation of lactate accompanies the Warburg effect. For most proliferating cells, nutrients are not limiting so there is no selective pressure to optimise metabolism for ATP yield. In contrast, a selective pressure for rate of metabolism does exist. For example, immune responses and wound repair depend on the speed of the proliferative expansion of effector cells. To survive, the organism must signal the responding cells to maximize their rate of anabolic growth. Cells that convert glucose and glutamine into biomass most efficiently will proliferate fastest [40]. In addition, there is emerging evidence that cellular metabolism within a tumour can be heterogeneous, with some cells using the excess lactate generated as a fuel for mitochondrial oxidative phosphorylation; a major cellular source of reactive oxygen species (ROS) production [41].

Cells with excess nutrient uptake that have not converted to aerobic glycolysis would be predicted to have increased oxidative phosphorylation and ROS production. Cellular energy supply and demand under hypoxic conditions is regulated by many interacting signaling and transcriptional networks, which complicates studies on individual proteins and pathways. Ambient air has oxygen partial pressure (pO2) = 158 mmHg, tracheal air in the human is pO2 = 149 mmHg, alveolar air; pO2 =100 mmHg; and arterial blood; a pO2 = 95 mmHg [42, 43].

One proposed explanation for Warburg's observation is that tumour hypoxia selects for cells dependent on anaerobic metabolism [44]. However, cancer cells appear to use glycolytic metabolism before exposure to hypoxic conditions. For example, leukemic cells are highly glycolytic [45], yet these cells reside within the bloodstream at higher oxygen tensions than cells in most normal tissues. Similarly, lung tumours arising in the airways exhibit aerobic glycolysis even though these tumour cells are exposed to oxygen during tumorigenesis [46]. Thus, although tumour hypoxia is clearly important for other aspects of cancer biology, the available evidence suggests that it is a late-occurring event that may not be a major contributor in the switch to aerobic glycolysis by cancer cells. Hypoxia is also known as the major cause of necrotic cell death in many diseases.

1.3.2.1 Metabolomics: A component of systems biology

Recent development in 'omics technologies provide an opportunity to study the effects of oxidative stress systematically. Metabolites are potentially a good indication of the state of our health. Therefore measurement and analysis of metabolites can be a precise and potentially valuable source for identifying biomarkers for diagnostics and drug therapy[47]. Metabolomics technologies act as components of systems biology that allow high-throughput analysis of metabolites, allowing determination of their concentrations, in biological samples [48].

While the genome, transcriptome and proteome are estimated to be quite large, by contrast the metabolome is relatively small. It is estimated that there are approximately 7,900 metabolites in humans as has been proposed by the human metabolomics project [49]. This number represents the naturally synthesised biochemicals in humans and now includes some complex oligosaccharides, peptides, etc. The total number of synthesised biochemicals is much smaller; however, the technology to measure all of those small compounds in a single sample is complex and challenging. A major advantage to this smaller number of metabolites is the decreased chance of a random false discovery measurement which often results when large numbers of measurements are made on a small number of samples, such as the small group sizes typically available for most clinical studies – ours is typically one such case.

The primary goal of any metabolomics effort is to extract, identify, and quantify as many as possible small molecular compounds (e.g. metabolites) in a biological sample under set conditions and to correlate the changes observed to changes in environmental condition or perturbation. Yet the size and chemical complexity of the metabolomes itself represent an analytical challenge since metabolites differ significantly in their physical and chemical properties (volatility, solubility); molecules vary from being polar to non-polar. Moreover, these small molecules have a wide range of structures; their molecular weights range from 50Da to over 1500Da and differ in concentrations. As diverse as these compounds may be, the task for metabolomics is to extract, identify and quantify all of the metabolites in a biological sample. To produce an accurate/informative profile of these molecules, not only samples themselves, but every compound present in a sample must be reproducibly extracted and measured, even when the identity is not known or its presence predicted. Sample reproducibly of DD and control cells has been shown and is the objective of Chapter 3.

This complexity necessitates the use of analytical tools and methodologies of high sensitivity, high separation efficiency (specificity) and with the ability to detect metabolites over a wide range of concentrations (typically, mM to sub-pM). Global analysis of the metabolome is achieved with more than one analytical technique and most commonly is achieved using chromatographic separation to fractionate complex samples into simpler components according to their physico-chemical properties and linearly introduced into a mass-spectrometer (MS) or nuclear magnetic resonance (NMR) spectrometer for detection. Gas-chromatography (GC) [50, 51], liquid chromatography (LC) [52, 53] or capillary electrophoresis (CE) are most commonly utilised analytical methods to separate metabolites based on selectivity and specificity.

1.3.2.2 Metabolic profiling of Dupuytren's disease phenotypes in oxidative stress

At present, little is known regarding DD pathogenesis, and even less regarding its cellular functions, i.e. metabolic functions and regulation of intermediates involved in glycolysis, TCA cycle, pentose phosphate shunt and amino acid metabolism. Knowledge of what proliferating cells need in healthy cells and how these cells adapt in hypoxia in terms of energy to generate biomass will help to establish hypotheses if not a direct connection between disease and a healthy cells. This will illuminate signaling pathways that drive cell growth and the regulation of DD cell metabolism.

Here, we test our hypothesis whether DD cells are also under this Warburg effect. This is achieved by inducing a perturbation in healthy cells (fascial cells) cultured in pO2 = 158 mmHg corresponding to a concentration of 21% atmospheric oxygen and compare their metabolic profiles with healthy cells exposed to hypoxia i.e. in pO2 of 8 mm Hg corresponding to concentration of 1% oxygen. Then we examine our hypothesis that any such differences are akin the Warburg effects noted for tumour cells in the literature by comparing the extracts from intracellular (endo-) and extracellular (exo-) metabolomes acquired from DD nodule, cord and SON cultures against those acquired from healthy cells and under hypoxia-induced fascia. The aim here is to identify potential biomarkers of DD by comparing the metabolic profiles obtained from disease and healthy cells. In addition, response to hypoxia is also examined in the intracellular metabolome of disease cells; this may aid in biomarker identification by imposing stress on disease cells. Furthermore the question is; if a Warburg effect exists, in which DD phenotype is this effect greatest? What are these key players (metabolites) and which pathways are these mapped onto? For convenience, fibroblasts from fascia, nodule, cord and SON cultured in pO2 = 158 mmHg (21% oxygen, normoxia) are designated as 'F21', 'N21', 'C21', 'S21' and those cultured in concentrations of 1% oxygen (hypoxia) designated as 'F1', 'N1', 'C1' and 'S1' respectively.

A point to note here, although pO2 =158 mmHg is not physiologically relevant for tissue, the palmar fascia pO2 is expected to be approx 40 mmHg; well above hypoxic conditions. Previous studies reported no culturing conditions other than O_2 at 21% (ambient conditions/setting in standard cell culturing facilities). To make a fair comparison, main consistency with previous reported finding and due to technical limitations we were unable to provide sufficient culture biomass in all three O_2 conditions i.e. 6%, 1% and 21% simultaneously. A comparison between 21% (termed normoxic) and 1% (termed hypoxic) is made to determine the effects of hypoxia upon the respiratory metabolism of all cultures. Metabolic profiling of samples employed GC-MS. Metabolic identification and the effect of perturbation is the objective of Chapter 4. Details of materials and methods used in this study are explained in Chapter 2.3.2.

These biomarkers could open up the possibility of developing new, early or presymptomatic treatments to improve outcomes or even prevent pathology. Furthermore, the validation of biomarkers that can detect early changes specifically correlated to reversal or progression of DD is crucial for intervention. Used as predictors, these biomarkers could help to identify high-risk individuals and disease subgroups potentially useful as targets for chemointervention trials, whilst as surrogate endpoints, biomarkers may be useful for assessing the efficacy and cost effectiveness of preventative interventions at a speed that is not possible when the incidence of manifest DD is used as the endpoint.

1.3.3 Transcriptomics: a component of systems biology

The transcriptome is the complete set of RNA transcripts produced by the genome at any one time. Transcriptomics, the study of the transcriptome involves large-scale analysis of messenger RNAs transcribed from active genes to follow when, where, and under what conditions genes are expressed. Unlike the genome, the transcriptome is extremely dynamic [54]. Our cells contain the same genome regardless of the type of cell, stage of development or environmental conditions. Conversely, the transcriptome varies considerably in these differing circumstances due to different patterns of gene expression. Transcriptomics is therefore a global way of looking at gene expression patterns and in this respect is one such component of systems biology that can aid in discovery experiments.

Microarray technologies allow the examination of gene expression on the scale of a genome when an organism or cells experience a changing status thus providing a practical method for measuring the expression level of thousands of genes simultaneously. Measurement of gene expression with high accuracy and precision is possible with real time PCR [55, 56]. This technique can measure and quantify gene expression one at a time and is used as a standard for validation and measurement of a few selected genes [56]. However, using microarrays it is possible to measure the expression of thousands of genes or more so *'transcripts*' simultaneously [57].

Microarray analysis is a very active area of research now. The fields of systems biology and microarray analysis are both growing exponentially with > 40,000 hits in PubMed using the term 'systems biology' and > 47,000 hits for 'DNA microarray analysis.' In just over a period of six months (July, 2010 to January 2011) 6000+ papers with terms 'microarray' were added to the database. By contrast, only ~2100 publications on DD have appeared since the original publication by Guillaume Dupuytren in 1831, only 56 hits with 'gene expression + dupuytren' and only 11 hits with 'microarray + dupuytren.'

1.3.3.1 Exploring dynamical changes in DD and control transcriptome

In addition to dynamic property changes recurring in DD metabolomes and healthy fascia in response to hypoxia, simultaneous changes in the transcriptome will also inevitably be in progress. A number of key metabolites and pathways that may contribute to DD progression or have been invoked as a consequence of hypoxic stress highlighted from results in Chapter

4 also demand explanation at the transcriptome level. Intermediates involved in amino acid and carbohydrate metabolism have shown significant differences from this analysis. In Chapter 5 Affymetrix microarrays were employed extending this approach to investigate the perturbation effect in their transcriptome. This study examines whether gene expression analysis of such cells could provide a more representative picture of the dynamics involved in DD. It is surmised these transcripts will produce a specific signature for DD complementing the metabolomics study and allow us to look for cell signaling pathways / targets in a controlled systematic manner. The emerging data will form the basis for selecting appropriate models for pathway studies.

In this study, we use 12 x Human Genome U133 Plus 2.0 Arrays (HG U133 Plus 2.0). Each chip contains 54,675 probe sets allowing analysis for relative expression levels of more than 47,000 transcripts and variants, including more than 38,500 well characterised genes and UniGenes [58]. The probe sets represented on these are selected from sequences in GenBank®, dbEST and RefSeq [59]. In addition these chips include a set of constitutively expressed human maintenance genes to facilitate the normalisation of array experiments. This set of normalisation genes has demonstrated consistent levels of expression over a diverse set of tissues and serves as a tool to normalise data prior to performing data analysis. Details of materials and methods used in this study are explained in Chapter 2.3.3.

1.3.4 Discovery of pathway biomarkers through network analysis

The qualitative systematic studies from Chapters 3-6 provide us with robust and reproducible data sets or at least the parts list that can now facilitate the initiation of model construction to study the interplay between theory, experiment and technology and now hypothesis testing by modeling the cellular systems. Key parameters and variables of the DD and control system can be assigned. The parameters of this dynamical system are those properties that are either inherent to the system or whose values can be controlled e.g. to study metabolic networks, these would be delineated as the initial concentrations of enzymes and metabolites, enzyme kinetic properties such as *Km*, *kcat* and *Ki*. The variables by contrast would be those that change during the time evolution of the system e.g. concentrations of metabolites and metabolic fluxes. Since it is the parameters that control the variables, it is more common to measure the variables than the parameters. Methods which start with variables and seek to infer the topology and parameters of the system that generated them are known as 'inverse methods' or 'system identification' methods - considerably more demanding computationally.

At present, knowledge is largely dispersed across various databases ranging from proteome-proteome interaction databases [60], human metabolome database (HMDB) [49], gene ontology (GO)[61] etc. To facilitate in bridging or integrating some of these databases, systems such as Gaggle [62] incorporating geese such as KEGG [63] , Cytoscape [64], STRING [65] and more, Taverna [66] and Ingenuity Pathway Analysis (IPA) [67] are found to be of considerable utility. Such tools consider both the significance of gene expression changes and their topological characteristics in order to better evaluate their impact on the pathways of interest [68].

Functional links between molecules e.g. proteins can often be inferred from genomic associations between the genes that encode them. However some interactions such as protein–protein interactions are not limited to direct physical binding and may also interact indirectly e.g. by sharing a substrate in a metabolic pathway or by regulating each other transcriptionally. For predicting such functional associations (including direct binding), the current growth in completed genomes offers unique opportunities through so-called 'genomic context' or 'non-homology-based' inference methods [69, 70]. The graphical representations of the networks of inferred, weighted entity–entity (e.g. protein-protein, gene-protein or enzyme-substrate) interactions provide a high-level view of functional linkage, facilitating the analysis of modularity in biological processes.

Visualisation of networks is highly important, and for this reason, IPA [67] is employed for topological network analysis. Combining statistically significant filtered gene lists and metabolite lists from the studies in Chapters 4 & 5 to find specific key pathways may be indispensable for understanding the regulation of metabolism in DD. Chapter 6 involves detailed analysis of metabolomic and transcriptomic data using integrative pathway analysis which illuminates molecules some which may be of potential importance in the pathophysiology of DD and may prove to be important as biomarkers. In this experimental model of hypoxia induced in DD cultures and healthy cells, we seek to visualise factors affecting the metabolic profiles resulting from the response to hypoxia which induced discriminating changes in the metabolic pathways. The aim here is to infer metabolic and signaling pathways involved in DD and healthy systems employing both separately and in conjunction the statistically relevant small molecules and transcripts by mapping the molecules identified in key transcriptional pathways (and networks) onto the metabolic pathways (and networks). The molecules highlighted from metabolite mapping and gene mapping will highlight key variables that can then enable construction of model building (e.g. constraint-models, kinetic models or Boolean models).

In addition the highest scoring networks in the analyses both with and without the gene expression data are explored with the aim of illustrating the effects of hypoxia an on inflammation, oxidative stress (production of reactive oxygen species), and metabolism. The networks identify a number of pathways, key molecules which could play a focal role in DD. MetPA (Metabolomics Pathway Analysis) [71]; is also employed for the visualisation and analysis of metabolomic data within the biological context of metabolic pathways. In both these algorithms, topological connections are inferred from fisher's exact ratio. False Discovery Rate (FDR) multiple testing correction (Benjamini and Hochberg analysis) [72] is possible in IPA. Additionally IPA includes modeled relationships between proteins, genes, complexes, cells, tissues, drugs, pathways, and diseases (direct and indirect relationships). It includes information from a broad range of published biomedical literature, internally curated knowledge, and a wide variety of trusted 3rd party sources and databases, so integration from a wide variety of information in one place is possible. All of the content in the Ingenuity Knowledge Base is structured, timely (updates occur weekly), and quality controlled to ensure quality.

1.3.5 Summary

The studies in this thesis form a starting point for metabolic and/or signaling network analyses. There are several benefits to modeling for example testing whether the model can be made to reflect known experimental facts in DD fascia compared to healthy palmar fascia or analysis of the model to understand which parts of the cell system (properties, components) contribute to a certain factor e.g. perturbation effect or so-called sensitivity analysis. Furthermore, hypothesis generation and testing out rapidly the effects of a change or manipulation in the system computationally would be preferred than to perform costly future experiments to determine more 'what if' experiments. Due to cost and time constraints, an integrative metabolomics and transcriptomics approach is applied to detect and identify variables from endogenous (fingerprint) and secreted (footprint) metabolites present in case and controls subjects. Affymetrix microarrays were used to profile a selected subset of samples to both identify and integrate with metabolites data and examine the networks using IPA [67] and MetPA [71] to infer metabolic and transcriptional networks within cell systems (DD, control and perturbed).

1.4 Workflow of the Study

A view of the systems approach (Figure 1), in the framework described involves the following sequence of steps:

1. Gather high-throughput molecular fingerprint data from vibrational spectroscopy using FT-IR to define all the potential components thought to be involved in DD formation and internal & external control.
2. Assess reproducibility of cultures over time (passage number) by determining the hyper spectral signature of each sample with chemometrics & cluster analysis to determine the most suitable *in vitro* representatives for metabolic and transcript level profiling and induction of perturbation effects.
3. Identify key regions of activity from the vibrational bands (i.e. lipids, proteins, sugars, nucleic acids).
4. Examine the extra cellular metabolome with vibrational spectroscopy.
5. Establish a cell culture system to isolate three biomarkers from the same population of cells to enable isolation and measurement of: metabolic footprint and fingerprint, and transcriptome.
6. Simultaneously culture samples in same growth condition and conditioned media in normoxic and hypoxic conditions to examine and monitor the perturbed components.
7. Harvest three cellular components from the same populations of cells in each culture condition 21% and 1% oxygen: metabolic footprint, intracellular metabolites and mRNA.
8. Examine metabolic content harvested from the intracellular and extracellular metabolome using GC-MS.
9. Identify key metabolites in disease and control system.
10. Investigate the effect of hypoxia in the perturbed fascia and compare these resultant molecules with those in disease.
11. Construct key metabolic pathways from this data.

12. Investigate this Warburg effect in the transcriptome with high through-put Affymetrix oligonucleotide microarrays.
13. Identify key dysregulated transcripts/genes
14. Construct key pathways from this data.
15. Map these genes upon metabolic pathways previously identified in step 11 and vice versa.

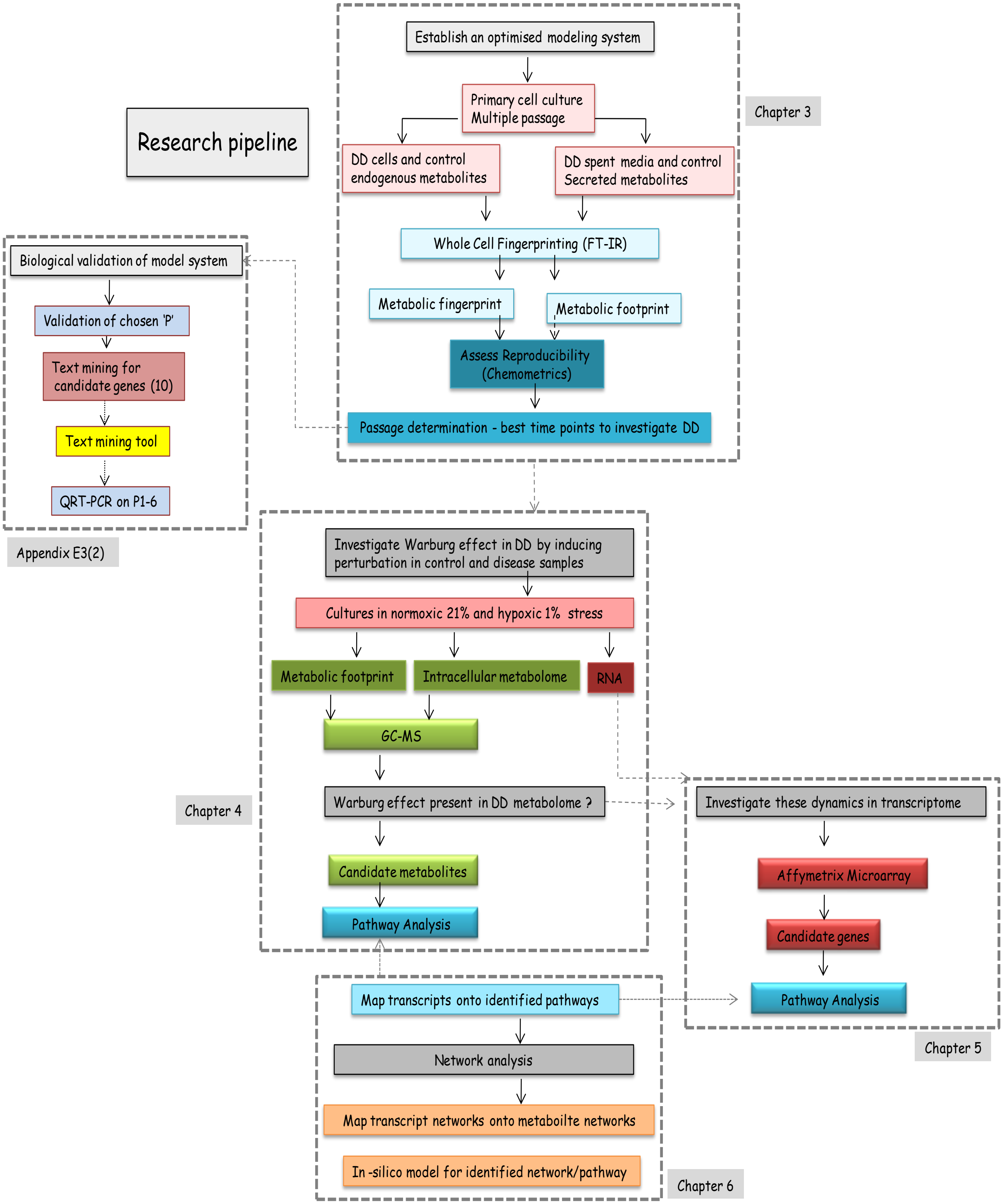

Research pipeline
Establish an optimised modeling system
Primary cell culture
Multiple passage
DD cells and control
endogenous metabolites
DD spent media and control
Secreted metabolites
Whole Cell Fingerprinting (FT-IR)
Metabolic fingerprint
Metabolic footprint
Assess Reproducibility
(Chemometrics)
Passage determination - best time points to investigate DD
Chapter 3
Biological validation of model system
Validation of chosen 'P'
Text mining for
candidate genes (10)
Text mining tool
QRT-PCR on P1-6
Appendix E3(2)
Investigate Warburg effect in DD by inducing
perturbation in control and disease samples
Cultures in normoxic 21% and hypoxic 1% stress
Metabolic footprint
Intracellular metabolome
RNA
GC-MS
Warburg effect present in DD metabolome ?
Candidate metabolites
Pathway Analysis
Chapter 4
Investigate these dynamics in transcriptome
Affymetrix Microarray
Candidate genes
Pathway Analysis
Chapter 5
Map transcripts onto identified pathways
Network analysis
Map transcript networks onto metaboilte networks
In -silico model for identified network/pathway
Chapter 6

Figure 1 Research pipeline to show the steps involved in this highly exploratory systems biology approach to understanding Dupuytren's disease.

1.5 Aims of this project

1. To test the hypothesis that metabolic profiles acquired from DD fibroblast cultures derived from different DD and control tissue phenotypes are unique to their tissue of origin using FT-IR spectroscopy.
2. To compare the metabolic profiles of DD fibroblast cultures derived from different DD and control tissue phenotypes.
3. To determine effect of serial passaging by comparison of early (primary) cultures to late passages in order to identify the most representative passage for the disease using FT-IR spectroscopy by assessment of reproducibility on the metabolic fingerprint/profile). The passage determined shall be used for subsequent studies once established the hypothesis.
4. To identify the metabolites that are different in healthy and DD. This will be investigated using fibroblast cultures selected from the most suitable representative passage number for the disease as determined in steps 1-3. Using a non-targeted approach and a more sensitive analytical technique; GC-MS is employed to profile the pool of metabolites present within the disease and control fibroblasts.
5. To determine whether altered oxygen tension (i.e. hypoxic condition) affects the composition of intracellular and extracellular metabolomes and the transcriptomes (gene expression) of early DD and control fibroblast cultures using GC-MS and Affymetrix microarrays for high throughput profiling respectively. This is to test the hypothesis whether the difference in disease cell types (nodule, cord and SON) and control cell type is the same as the difference in control fibroblasts cultured in normoxia and hypoxia.
6. To test in which specific disease cell type (nodule, cord, and/or SON) is the difference with normal cells in intracellular metabolome the largest?
7. To examine the extent of the Warburg effect when hypoxia is induced within the disease cell types.
8. To integrate transcript and metabolite profiling data through a SB approach in order to determine congruence between the levels of certain metabolites, gene transcripts and their protein product(s) using SB tools to identify metabolic pathways, signaling pathways and key networks (connections and intercellular dynamics) that may attribute to formation of DD.

Chapter 2

Materials and Methods

2.1 Background to Experimental Techniques

This chapter provides the background and theory to the experimental techniques applied in this thesis (2.1). Data analysis methodologies (2.2) and the complete protocols for each of the experiments in Chapters 3-6 (2.3) are provided.

2.1.1 Fourier transform Infra-red Spectroscopy as a metabolic fingerprinting screen

Infrared (IR) spectroscopy is one of the most common spectroscopic techniques used by organic, inorganic chemists and medicinal chemists for the detection of different chemical functional groups in the sample [73, 74]. In addition to its ability to provide information about the structure of a compound it is also used as an analytical tool to assess the purity of a compound [74].

In IR spectroscopy, IR radiation is passed through a sample. Some of the IR radiation is absorbed by the sample and some is transmitted. Different functional groups present in the sample absorb characteristic frequencies of IR radiation [75]. This is recorded in the form of a spectrum which is a plot of intensity vs. frequency with wavelength or wavenumber as the x-axis and absorption intensity or percent transmittance as the y-axis. The resulting spectrum represents the molecular absorption and transmission, creating a molecular fingerprint of the sample. Like a fingerprint no two unique molecular structures produce the same infrared spectrum [76]. In simple terms, IR spectroscopy is the absorption measurement of different IR frequencies by a sample positioned in the path of an IR beam.

The IR region lies between the visible and microwave regions having wavenumbers from approximately 12,500 cm^{-1} to 10 cm^{-1}, or wavelengths from 0.78 to 1000 μm. Frequency ν (nu) is the number of wave cycles that pass through a point in one second. It is measured in Hz, where 1 Hz = 1 cycle/sec. Wavelength, λ (lambda), is the length of one complete wave cycle and is often measured in centimeters (cm).

Wavelength and frequency are inversely related:

$$\nu = c/\lambda \qquad \text{where } \lambda\text{= Wavelength (μm), c = Speed of Light.}$$

and Energy is related to frequency by:

$$E = h\nu \qquad \text{where, } \nu\text{= Frequency (Hz), h = Planck's constant}$$

The IR region is divided into three regions: the near, mid, and far IR (Figure 2). In wavenumbers, the mid IR range is 4000 - 400 $cm^{–1}$. An increase in wavenumber corresponds to an increase in energy. This study is performed in the most frequently used mid IR region, between 4000 - 400 $cm^{–1}$.

Transmittance, *T*, is the ratio of radiant power transmitted by the sample (*I*) to the radiant power incident on the sample (*I*0). Absorbance (*A*) is the logarithm to the base 10 of the reciprocal of the transmittance (*T*).

$$A \log_{10} = (1 / T) = –\log_{10}T = –\log_{10}I / I0$$

The transmittance spectra provide better contrast between intensities of strong and weak bands because transmittance ranges from 0 to 100% *T* whereas absorbance ranges from infinity to zero. Often the same sample will give quite different profiles for the IR spectrum, which is linear in wavenumber, and the IR plot, which is linear in wavelength. This may appear as though some IR bands have been contracted or expanded.

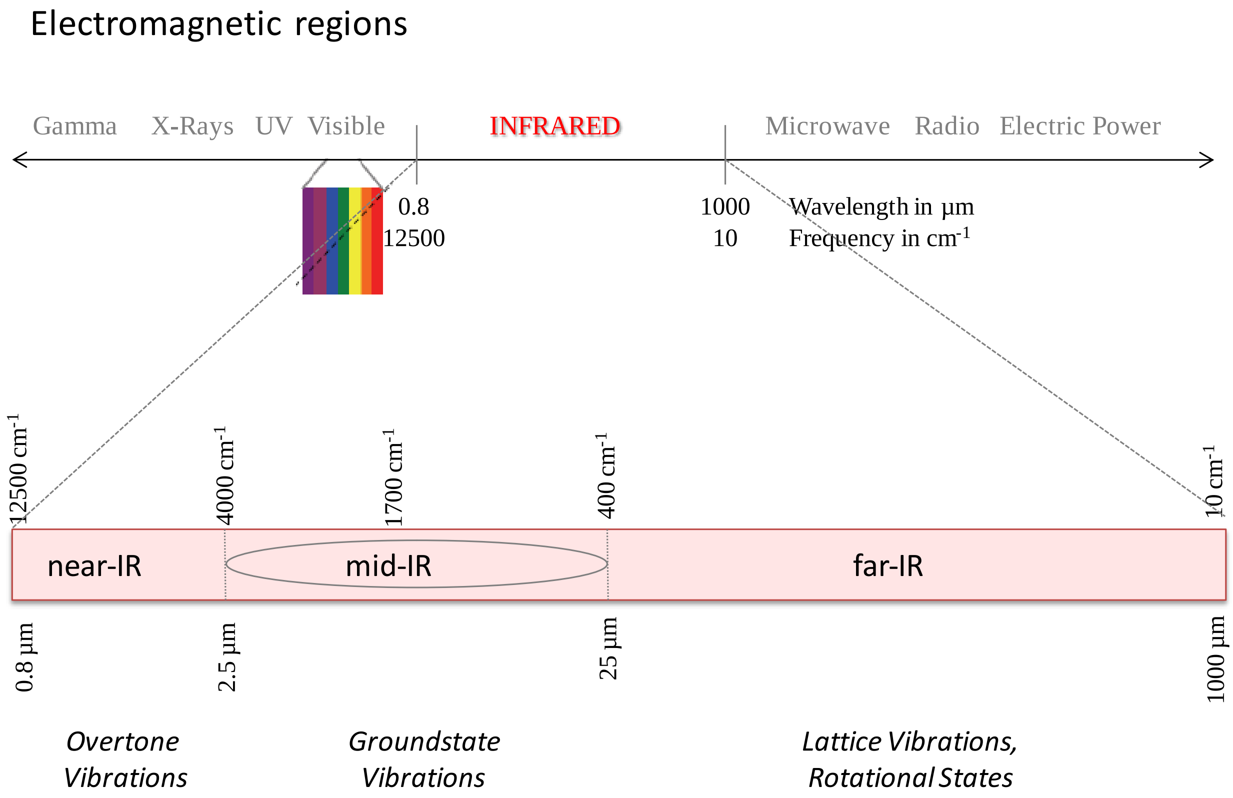


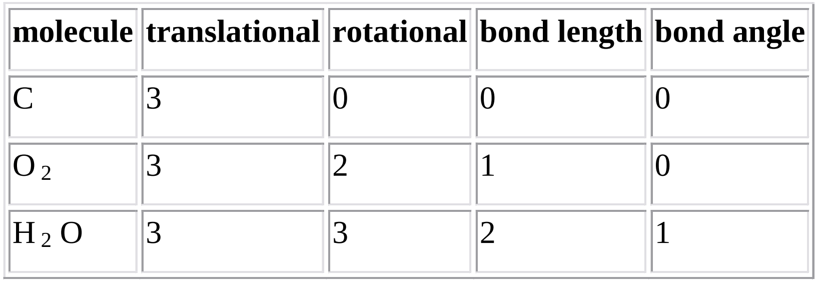
Figure 2 The diagram shows the whole electromagnetic spectrum and the position of IR region divided into three regions; the near, mid, and far IR.

2.1.1.2 Theory of Infrared Absorption

All atoms in a molecule above absolute zero temperature (0 K) are in continuous vibration with respect to each other. A molecule absorbs radiation when the frequency of a specific vibration is equal to the frequency of the IR radiation directed on the molecule. Each atom within the molecule has three degrees of freedom, which corresponds to motions along any of the three Cartesian coordinate axes (x, y, z). A polyatomic molecule of *n* atoms has 3*n* total degrees of freedom. Of these, 3 degrees of freedom are translational, describing the motion of the entire molecule through space. Additionally, there are 3 rotational degrees of freedom for a non-linear molecule and 2 rotational degrees of freedom for a linear molecule which correspond to the rotations of the entire molecule. Therefore, the remaining 3*n* – 6 degrees of freedom are true, fundamental vibrations for nonlinear molecules and for linear molecules 3*n* – 5 fundamental vibrational modes because only 2 degrees of freedom are sufficient to describe rotation [77].

Among the 3*n* – 6 or 3*n* – 5 fundamental vibrations, those that produce a net change in the dipole moment may result in an IR activity and those that give polarisability changes may give rise to Raman activity (complementary technique to FT-IR). Some vibrations can be both IR- and Raman-active. The total number of observed absorption bands can be different and are often reduced from the total number of fundamental vibrations. This is because some modes are not IR active and a single frequency can cause more than one mode of motion to occur. Conversely, additional bands are generated by the appearance of overtones (integral multiples of the fundamental absorption frequencies), combinations of fundamental frequencies, differences of fundamental frequencies, coupling interactions of two fundamental absorption frequencies, and coupling interactions between fundamental vibrations and overtones or combination bands (Fermi resonance). The intensities of overtone, combination, and difference bands are less than those of the fundamental bands. The combination of all factors creates a unique IR spectrum for each compound [78].

Table 1 An example of atoms as point objects with corresponding degrees of freedom.



molecule	translational	rotational	bond length	bond angle
C	3	0	0	0
O_2	3	2	1	0
H_2O	3	3	2	1

Vibrations can be in the form of bonding vibrations (the number of chemical bonds), bending vibrations (change of bonding angles), torsional vibrations, out-of-plane vibrations, wagging scissoring etc. Infrared radiation is absorbed and the associated energy is converted into these types of motions. Table 1 gives an example of atoms, as point objects that have no dimension and hence no moment of inertia and therefore, cannot have rotational degrees of freedom. Besides three translational and one vibrational degrees of freedom, a diatomic molecule can only rotate rigidly about its center of mass, since it has no cross sectional dimension (a bond is idealized as having no thickness). For larger molecules, such as water, there can be an oscillation of the bond angle [78].

Excitations of the vibrational modes of molecular bonds occur when the molecular orbitals absorb photons with infrared wavelengths. A group of these wavelengths are listed in Appendix B; Table 22. For example, carbon dioxide CO_2 is a linear three atomic molecule that has 4 vibrational degrees of freedom. H_2O is non-linear also having three atoms and has 3 degrees of freedom. These degrees of vibrational freedom occur when each bond acts like a spring and carries with it potential and kinetic energy (obeys Hookes law) [79]. Each normal mode has its own quantum number and is independent of the other modes.

2.1.1.3 The interpretation of infrared spectra

Interpretation of vibrations of a polyatomic molecule is complex. The determination of the reduced mass of a specific normal mode for larger molecules is increasingly complex. However, partitioning and identifying functional groups within a molecule gives valuable information and a good approximation about its composition.

The interpretation of infrared spectra can be achieved by correlation of absorption bands in the spectrum of an unknown compound with the known absorption frequencies characteristics for particular types of bonds and functional groups. The most important factors sought for elucidation of structure are Intensity (weak, medium or strong), shape (broad or sharp), and spectral position (cm^{-1}) in the spectrum. For example, X-H stretching vibrations due to the light hydrogen atom are in good approximation independent of any other vibrations and are found in the region of $3400cm^{-1}$, and if an O-H bond is present, a broad peak is observed in this region. A C=O group with two sharp intensity peaks (one longer than the other) would be absorbed around 1715± $100cm^{-1}$ due to asymmetric and symmetric stretching. The four important regions of the IR spectrum are shown Appendix B; Table 23, Figure 69.

For complexed biological systems little can be understood by simply looking at the spectra alone as all spectra will show broad contours in a similar shape due to the typical composition of a biological cell (e.g. cells, nuclei, proteins and nucleic acids are present in all organisms). It is more practical to incorporate multivariate statistical analyses (MVA) methods to help seek trends in the data and to aid interpretation; one such technique is principal component analysis (PCA).

Spectral fingerprints, regardless of the means of acquisition, are highly complex, representing the sum of all compound present in a sample. In general, it is very difficult to identify, let alone quantify, individual compounds. The introduction of group frequencies allows the identification of structural elements of a molecule and makes IR spectroscopy an important tool for the identification of molecular structure and for quantitative analysis http://www.rwc.uc.edu/koehler.

An IR spectrum represents a fingerprint of a sample with absorption peaks which correspond to the frequencies of vibrations between the bonds of the atoms making up the material. However, each different material is a unique combination of atoms and no two compounds produce an identical spectrum. Therefore, IR spectroscopy can result in a positive identification (qualitative analysis) of every different kind of material. In addition, the amplitude of the peaks in the spectrum is a direct indication of the amount of material present. With modern software and signal processing algorithms, IR spectroscopy is an excellent tool for rapid high throughput quantitative chemical analysis [80]. FT-IR Spectroscopy is widely employed in metabolomics for initial finger and footprinting analysis.

2.1.2 Gas chromatography – mass spectrometry in metabolomics

Gas chromatography mass spectrometry (GC/MS) is a hyphenated analytical platform widely employed in the field of analytical science and metabolomic studies for the analysis of volatile organic compounds [81]. The GC is employed as front end separation source where complex chemical samples are fractioned into simpler components via interactions with both the stationary (analytical column) and mobile phases (carrier gas). The type of stationary phase employed is dictated by the chemical nature of the compounds (i.e. polarity, this is different for fat, proteins etc) and its sample matrix. A small volume of liquid sample (1-5 μL) is injected into a high temperature, pressurised injection port assembly causing it to immediately vaporise. The sample is then deposited onto the top of the analytical column under the influence by the carrier gas, it then migrates through the column as it does individual components within the sample will interact with the stationary phase at varying degrees. The separation of each individual component within the sample mixture will be more pronounced and eventually separates into individual discrete bands. These are then eluted from the analytical column. Each band/peak represents an individually resolved component within the sample mixture. The eluents are then ionised using an electron impact source generating a molecular ion and is then introduced into the mass spectrometer. Molecular ions generated are usually unstable and therefore can undergo self fragmentation into smaller sub units. The resulting ion fragments are then detected based on their mass to charge ratio (*m/z*) within the MS producing a mass spectrum of the eluent peak. A fragmentation profile is generated for each eluted peak within the sample; the profile is highly characteristic and indicative of the original parent molecule. Further examination of the isotopic ratios, distribution and composition of these ion fragments, can yield detailed significant chemical information regarding the chemical structure and functional groups present. The fragmentation profiles are then matched against a known mass spectral library such as NIST08 [82] using AMDIS (http://www.amdis.net/) to aid in the identification of the unknown compounds.

2.1.2.1 Chemical derivatisation

Most biofluids and tissue are involatile and this necessitates additional sample processing steps to convert the sample, amenable to GC-MS analysis chemical derivatisation is employed to achieve this. This method transforms molecules to give them a volatile characteristic thus reducing polarity allowing analysis by GC-MS [83]. In smaller molecules low volatility may be the result of strong intermolecular attraction between the polar groups present [84]. Polar groups such N-H, O-H and S-H groups undergo hydrogen bonding and have a significant contribution towards the intermolecular attraction. By replacing the active hydrogen in those groups through alkylation, acylation or silylation will dramatically increase its chemical volatility particularly in compounds with multiple polar groups. Bulky, nonpolar silyl groups such as CH_3 are often used for this purpose. This method also serves to increase the quality of analysis by improving the chromatographic profiles of the chemical compound and overall resolution of the analysis [83].

2.1.3 Microarray: a source of high-throughput biological datasets

Deoxyribonucleic acid (DNA) is the hereditary molecule of all cellular life forms. Found inside the nucleus of a cell, it stores and transmits genetic information. The concept of “gene” is simply a piece of this DNA, yet encoded in so many different ways can depict a phenotype that distinguishes one from another. The complete set of DNA in any cell of an organism is called its genome. However, the genome is only a source of information. In order to function, it must be expressed. Gene expression occurs in two stages. First, DNA (gene sequence) is transcribed to produce an RNA sequence through a process called *transcription*. Next, this messenger RNA, or mRNA travels from the nucleus to the cytoplasm. Each gene sequence in DNA that codes for a protein is expressed as a sequence in mRNA. The phrase "gene expression" in quantitative terms means the amount/abundance of mRNA present inside the cell. The mRNA sequence is then converted to proteins through a process called *translation*.

Microarray technologies allow the examination of gene expression (or study of the transcriptome) on the scale of a genome when an organism or cells experience a changing status thus providing a practical method for measuring the expression level of thousands of genes simultaneously. A microarray is typically a glass slide on to which DNA molecules are fixed in an orderly manner at specific locations called '*features' or 'probes'*. There are two types of microarrays; oligonucleotide microarrays based on Affymetrix technology (single-channel) that uses photolithography-directed combinatorial chemical synthesis to manufacture its GeneChips, the other type is a spotted microarray which is based on cDNA microarray technology [85]. Both these microarrays are based on the mechanisms of DNA hybridisation; a process of combining complementary single stranded nucleic acids into a single molecule. Two perfectly complementary sequences of nucleotides bind to each other under normal condition - a process known as annealing. The microarray contains an array series of thousands of microscopic probes of DNA oligonucleotides attached to a glass or a silicon chip. Each probe contains picoMoles of a specific DNA sequence [86]. This can be a short section of a gene that is expressed in the cell. The oligonucleotide microarrays based on Affymetrix technology synthesize small 20-25mer sequences on the probe. A set of 5-15 probes target a 100 to 200 base pair segments from a known mRNA. These are called '*probe sets*.' Affymetrix GeneChip arrays have multiple probes associated with each target. The probe set can be used to measure the target concentration and this measurement is then used in the downstream analysis to achieve the biological aims of the experiment, e.g. to detect significant differential expression between conditions, or for the visualisation and/or clustering of data.

In single-channel microarrays, the arrays are designed to give estimations of the absolute levels of gene expression. A strength of the single-dye system lies in the fact that an aberrant sample cannot affect the raw data derived from other samples, because each array chip is exposed to only one sample. Another benefit is that data can be compared to arrays from different experiments; the absolute values of gene expression may be compared between studies conducted months or years apart or labs apart around the world.

2.2 Data analysis

2.2.1 An essential tool for Metabolomics Data analysis: Chemometrics

Modern analytical instruments can generate a large amount of data in a short time. Metabolomics techniques enable high-throughput application to separate complex mixtures with several hundreds of chemical compounds, and can provide spectra of the compounds for structure elucidation. These analytical techniques provide as powerful tools to analyse various types of samples and give a very detailed insight into the nature of the sample. However, efficient data analysis methods to extract useful and interpretable information from the huge amount of data are a challenge. Chemometrics is the subject of using statistics, mathematics methods to solve chemical and biological problems [87]. In addition to classical statistical approaches that can be applied to the data sets such as student's t-test, ANOVA chemometrics also employed various multivariate methods for analysing complex data sets which has more advantages in terms of sensitiveness, robustness and wider application prospect. Nevertheless the goal of these statistical or machine learning methods is to identify the biochemicals that best represent the most significant changes between the groups in the study.

The basic principle of chemometrics: given the data under consideration, utilise the methods of mathematics and statistics to help with the aim of extracting and interpreting information for modeling and prediction [87]. Chemometrics deals with experimental design; signal processing; pattern recognition; calibration and evolutionary signal processing and curve resolution [88]. This is a huge field and certain sections are beyond the scope of this study. Experimental design, signal processing and pattern recognition techniques are discussed in this thesis.

2.2.2 Pattern recognition in complex data sets

Pattern recognition techniques can be categorised into two types: unsupervised pattern recognition and supervised pattern recognition [89]. Unsupervised pattern recognition plays the most important role in exploratory data analysis since it does not require much prior information. The information provided by unsupervised pattern recognition is “data driven”, i.e. the pattern revealed by unsupervised techniques purely depends on the data used for modeling/clustering and prediction is often not of concern. There are two main tasks for unsupervised pattern recognition: (1) data visualisation and (2) cluster analysis; allowing examination of the dominated underlying trends within the data set. Therefore, it can be considered as a discovery data analysis tool.

2.2.2.1 Unsupervised pattern recognition: Data visualisation & cluster analysis

Perception of patterns within data can best be captured by visualisation of data points directly through a 2-D or 3-D scatter (score) plot. Since the human eye can only perceive at most in 3 dimensions, it is not easy to directly visualise data with more than 3 dimensions. It is often desirable to summarise such complexity by projecting the data points into a lower dimensional subspaces defined by a few “latent variables”. Such process is under the name of “component analysis” [89]. The most popular component analysis method, is principal component analysis (PCA) [90, 91]. This has become the most commonly used data visualisation and modeling method in Chemometrics. PCA projects the original multivariate data points into a lower dimension space defined by a subset of mutually orthogonal axis (principal components, PC) while still preserving the major variations of the data set. In PCA, the data matrix is decomposed into the product of 2 matrices named scores matrix and loadings matrix[91]. The scores matrix records the relative position of the samples while the loadings matrix records the contribution of each variable to the pattern shown in the scores matrix. The modeling importance of each principal component (PC) is measured by the percentage of total explained variance (TEV) by the PC. If, for example, the sum of TEV of the first two PC is high (e.g. 90% of the total variance) then the scatter plot of PC1 vs. PC2 is a good two-dimensional representation of the original data structure. PCA can also be used for predictive modeling.

Another important area of unsupervised pattern recognition is ‘cluster analysis.’ The objective of cluster analysis is to partition a given data set into a small number of groups (clusters) in terms of similarity as such that data points in the same cluster are more similar to each other relative to than the other data points in different clusters further away. One well known cluster analysis algorithm is Hierarchical Cluster Analysis (HCA) [92]. The clusters identified by such cluster analysis can be further validated by supervised pattern recognition.

2.2.2.2 Supervised pattern recognition

Supervised pattern recognition is used for the purpose of predictive discrimination and calibration purposes, where it aims to correctly associate a known response to the correct output with minimal error [92]. Given a finite number of classes to be separated, a decision rule (classifier) is derived based on a predefined group of samples with known class members (training set) and such a rule has capability to predict the class membership of unknown samples (test set). The accuracy of the prediction model/classifier is assessed by its 'generalisation performance' of the classifier. To successfully solve a supervised pattern recognition problem, two types of problems need to be considered: *underfitting* and *overfitting*. Underfitting refers to the predictive model not having sufficient complexity to fully detect the entire systematic trend in the data set and therefore not being able to give an accurate prediction. In contrast, the problem of overfitting refers to an over-complex predictive model. Not only the systematic trend but also noise has been modelled. Overfitting is particular dangerous in practical applications because it can give a perfect prediction on the training set while giving very high prediction errors on the unknown test set. Linear Discriminant Analysis (LDA) and its variants, Discriminant Function Analysis (DFA) [92, 93]; regression models such as Partial Least Squares for discriminant analysis (PLS-DA) [94] and Artificial Neural Networks (ANNs) [95] are some widely used supervised pattern recognition methods.

2.2.2.2.1 Principal component-discriminate function analysis (PC-DFA)

DFA, also known as canonical variate analysis (CVA) was used in Chapter 3 and 4. Since DFA cannot handle the problem of colinearity well which is common to almost all metabolomics data, it is necessary to "clean" the data beforehand by using PCA. Thus in practical use, PCA and DFA is usually applied in a consecutive manner and such methodology is under the name of PC-DFA. PC-DFA/CVA is a form of supervised pattern recognition as it is applied with a priori knowledge of the class membership of each sample. PC-DFA algorithm is based on the Manly principles [96] and it aims to minimise the within group variance while maximising the between group variance (by maximising its Fisher ratio [97, 98]. The Fisher ratio is defined as the class to class ratio variation divided by the sum of the within class variation:

$$((M_1-M_2)^2)/(V_1+V_2),$$

where M_1 and M_2 are mean of Class 1 and Class 2 and V_1 and V_2 are the variances of the Class 1 and Class 2. This method has been extensively applied in metabolomics for the analysis of spectroscopic data (Raman and FT-IR) [31, 99, 100]. Discrimination of salt stress in tomatoes [101], identification of urinary tract infection bacteria [31] and recently applied to the understanding of plant-pathogen interactions [102].

A PC-DFA/CVA model is generated by using a portion of the data as a training set to initially construct the PC-DFA model, its generalisation performance is then assessed by the use of a test set, not previously used in the generation of PC-DFA model, this test set is then projected onto the cluster space, by examining the Euclidean distance of projected data point of the test to that of the training test within cluster space it accuracy can be evaluated, typically a third validation set (data points not previously used in either the training and test set) is used to further check the robustness and validity of the PC-DFA model and guard against over fitting of the data set. The PC-DFA model can be optimised by adjusting the number of PC used to increase its generalisation performance. The trends or difference identified in the model are through the loading of DFs and this is followed by validation using other methods or normally confirmed using univariate statistics.

The Euclidean distance between a priori group centers in DFA space using the first two functions (DF1 and DF2), was used to construct a similarity measure, and these distance measures were then processed by an agglomerative clustering algorithm to construct a dendrogram.

2.2.2.2.2 ANOVA-PCA, a semi-supervised approach

ANOVA-PCA can be considered as a method somewhere in the middle between unsupervised and supervised methodologies. A main drawback of PCA is that it is completely variance driven and in this respect, PCA can be disturbed by the variance caused by the variables/metabolites which are not affected by the biological experiment. It is possible that the most significant PCs obtained from PCA are not related to the aim of the study simply because that there are other factors which cause more variances than the factors which the experiment was designed to investigate did. Analysis of variance-principal components analysis (ANOVA-PCA) is developed to cope with such problem by actively incorporating the experimental design into the PCA.

ANOVA-PCA creates a series of matrices which contain the means of the different levels of the factors under investigation and interactions of these factors according to the experimental design, to which are added the residual errors [103]. PCA is then applied to each of these mean plus error matrices in order to evaluate the significance of the effects caused by each factor against the residual error. Construction of submatrices of the data for each factor can be more easily interpreted, visually and statistically, by PCA. Similar to PCA, scores and loadings are obtained, which can be used to study the existence of groupings of individuals and to evaluate the importance of the initial variables in the definition of the effects and the sources of residual variation and to compare it to the different factors in the experimental design. This procedure is not related to the ANOVA-based method that is often used to detect significant variables prior to a multivariate analysis such as PCA. Although this method also uses the labeling information, the key difference between ANOVA-PCA and supervised methods is that ANOVA-PCA "encourages" the PCA to discover the variation of interest by use an ANOVA like data-preprocessing and thus increase the chance of discover the variation of interest in the first or first a few PCs while supervised methods seek an optimal separation boundary between known classes. As a result, the risk of over-fitting of ANOVA-PCA is lower than that of supervised methods because it makes no attempt to *separate* the samples according to their membership information.

The reason ANOVA-PCA is used in this study (2.3.2) is that there are many different types of variations in the data and in many cases only a small subset of variation are of interest. The variance that PCA considered to be important (i.e. the trends shown in the first a few PCs) are not necessarily the ones what we want to see and ANOVA-PCA appeared to be much more effective than PCA in revealing the difference between different groups of samples.

2.2.3 Microarray data analysis

The vast numbers of publications that have reported methods for microarray data analysis follow differing analytical strategies. Microarray data analysis is a complicated process as this technology is associated with many significant sources of experimental uncertainty, which must be considered in order to make confident inferences from the data. However, there is no standardised microarray data analysis pipeline but what is available are opinionated analysis pipelines based on experience, nature of biological question in the study, suitability of method to the experimental design. Following image processing, and transformation using the proprietary Affymetrix image analysis software to generate CEL files (and other files e.g. CHP, DAT), discovery of relationships between genes can be pursued in many ways.

Affymetrix microarrays are designed to give estimations of the absolute levels of gene expression. Relative intensities of each probeset can be used in ratio-based analysis to identify up-regulated and down-regulated genes. Expression ratios are the primary form of comparison. Background correction and quantile normalisation using Robust Multi Array (RMA) [104] and GeneChip Robust Multi Array (GC-RMA) are popular methods used. Exclusion of MM data in RMA reduces noise, but loses information. Inclusion of adjusted MM data in GC-RMA reduces noise, but retains MM data.

Like metabolomics data, analysis of gene expression data can be classified into two different types; unsupervised and supervised learning. In the case of unsupervised learning (exploratory data analysis), the expression data is analysed to identify patterns that can group genes or samples into clusters without the use of any form of *a priori* knowledge. While in supervised learning the use of annotation is incorporated to create genes or sample clusters in order to identify patterns that would be characteristic of respective clusters. Pattern recognition methods have been discussed in 2.2.2.

Among the most common and important analysis tools is data visualisation using heatmaps. Gene expression data is converted to a colour code for visualisation. Common practice uses red for up-regulated genes, green for down-regulated genes, and black for no change. For simultaneous analysis of a set of microarrays, the data is clustered in terms of genes or arrays using different clustering algorithms and the output is visualised by a heatmap. Agglomerative hierarchical clustering is commonly used for the analysis of gene expression data [92]. The representation of this hierarchy is a dendrogram. This can help separate and identify locations of different clusters. Other popular clusters algorithm such as Partitioning Around Medoids (PAM) algorithm [105] and *k*-means [106] are also widely used in gene expression data analysis..

Arbitrary thresholds such as > 2 fold change (FC) have been used to rank/identify genes of interest. Other common methods performed are Pearson’s correlation coefficient; distance metric used in various clustering algorithms including HCA and rank correlation coefficient. Univariate tests can also be applied. Significance analysis of microarray (SAM) performs thousands of T-tests efficiently [107].

A number of methods for finding pairs of co-expressed genes, based on correlation or measures of mutual information have been used to understand gene regulation [108-112]. There are also model based approaches such as Bayesian models which have been very successful and are now used extensively [113-118]. Such models explicitly represent and reason about biological entities in a modular way, as well as capture the mechanistic details of the underlying biological systems. Puma: Propagating Uncertainty in Microarray Analysis [119], one such package incorporated the Gamma Model of Signal [120]. These packages are freely available via the R-Bioconductor software suite [121]. Bioconductor is an open source and open development software project for the analysis and comprehension of genomic data. It is based primarily on the statistical R programming language. Other, emerging new methodologies for gene expression analysis are making use of Boolean logic [122]. Despite this each method has its pros and cons for analysing noisy data.

Due to experiment errors and noise, there will always be some difference in expression between groups. However, it is the size of this difference in comparison to the variance (i.e. the range over which expression values fall) that will tell us if this expression difference is significant or not. Thus, if the difference is large but the variance is also large, then the difference may not be significant. However, a small difference with a very small variance could be significant. For example, T-test and ANOVA test return a p-value that takes into account the mean difference, the variance and the sample size. The p-value is a measure of how likely a particular gene will return if no real difference existed. This is called formulating a null hypothesis. A p-value < 0.05 indicates that the chance of this gene being different is due to random occurrence and there was in fact no real difference is small and therefore the gene could be considered significantly different in the group expression data. However, the p-value itself is only valid to a single test. When the statistics test has been performed on multiple targets, the chance of having false positive discovery increases along with the number of the tests been performed. For example, choosing a threshold of 0.05 means there is a 5% chance (1 in 20) the returned result is false positive in a single test, However, if one perform such test on 2 different genes (assuming the two genes are independent to each other) using the same *p*-value threshold, the chance of at least one of these 2 being false positive becomes 1-0.95×0.95=0.0975. This issue is known as the multiple testing problem [123].

A number of approaches to overcoming this multiple testing have been suggested. These include, assigning an adjusted p-value to each test, or choosing a lower p-value threshold e.g. 0.01 or 0.001. The Bonferroni correction [124] is also a popular method, but often too conservative. While the method reduces the number of false positives, number of true discoveries is also axed. The False Discovery Rate (FDR) approach determines adjusted p-values for each test. However, this method controls the number of false discoveries in only the significant values, hence it is less conservative than Bonferroni approach and is a preferred method to truly identify significant results. An FDR adjusted p-value is now termed the q-value.

When doing lots of tests, as in a microarray experiment, it is more intuitive to interpret p and q values by looking at the entire list of values rather that looking at each one independently. In this way, a threshold of $q < 0.05$ has meaning across the entire experiment indicating how many false positives can be expected by using this cut-off.

2.3 Materials and Methods

2.3.1 Whole-cell fingerprinting using FT-IR (Chapter 3)

2.3.1.1 Experimental design

Patient Recruitment – Study 1

All cases involved in the study were diagnosed to have advanced stage of DD, which was determined by the presence of nodule and cord causing contracture of the metacarpophalangeal joint and the proximal interphalangeal joint in the involved hand. The mean age of the patients participated in Study 1 was 67 ± 10 years. The exact age and demographics for all patients can be seen in Appendix B, Table 24. All patients were male (except DD13; female) caucasians who had not undergone any previous surgical or non-surgical treatments.

Patient Recruitment – Study 2

Four DD cases and five controls subjects (CTD) were included in the study. All recruited DD cases were diagnosed with advanced stage of DD, which was determined clinically by an experienced hand surgeon. All patients presented flexion contracture of the metacarpophalangeal joint and proximal interphalangeal joint as well as presence of nodules. All DD patients in this study were male Caucasians and only one (DD2) had undergone previous surgical treatment. The mean age was 60 ± 12 years. Four of the five control subjects included in the study were Caucasians, and one being Asian Indian. Three of the control subjects were male and two were female. The average age of the control subjects was 57 ± 19 years. The study was approved by the institutional review board for human subjects' research.

Table 2 Demographic details from Dupuytren cases in Study 1 and 2.

Study 1							
Patient ID	**Age**	**Sex**	**Anatomical location**				
DD8	77	M	Nodule	Cord	Fascia		
DD9	58	M	Nodule	Cord	Fascia	Fat	Skin
DD10	76	M	Nodule	Cord	Fascia		
DD11	67	M	Nodule	Cord	Fascia	Fat	Skin
DD12	52	M	Nodule	Cord	Fascia	Fat	Skin
DD13	74	F	Nodule	Cord	Fascia	Fat	Skin

Study 2							
Patient ID	**Age**	**Sex**	**Anatomical location**				
DD16	46	M		Cord	Fascia		
DD17	67	M	Nodule	Cord	Fascia	Fat	Skin
DD2-R			Nodule	Cord	Fascia	Fat	Skin
DD18	68	M	Nodule	Cord		Fat	
CT4	67	M			Fascia*	Fat*	Skin*
CT5	78	M			Fascia*	Fat*	Skin*
CT6	62	F			Fascia*	Fat*	Skin*
CT7	50	F			Fascia*	Fat*	Skin*
CT8	28	M			Fascia*	Fat*	Skin*

*CTD patient, external control

Samples Collections

DD tissue phenotypes (nodule, cord, unaffected transverse palmar fascia, subcutaneous fat superficial to nodule and SON) were carefully dissected using magnifying loupes from each patient at the time of surgery (Figure 3(a), Table 2). Three tissue biopsies, including the skin, subcutaneous fat and palmar fascia (transverse carpal ligament), were obtained from individuals undergoing carpal tunnel release (Figure 3(b). Each biopsy was bisected for either 1) cell culture processing or total ribonucleic acid (RNA) extraction or 2) for cell culture processing and histology tissue processing. The biopsies used for establishing tissue cultures were thoroughly washed for 15 min in 1× Dulbecco's phosphate buffered saline (Lonza, Belgium) and 1% penicillin/streptomycin (Lonza, Belgium), at room temperature. For histological analyses to determine tissue and cell morphology, the biopsies were stored in formalin at 3^{o}C, and processed within 48 hrs. The harvested biopsy samples for total ribonucleic acid (RNA) extraction were kept in RNA*later* (Ambion, UK) at 4^{o}C overnight and stored at -80^{o}C until required for subsequent gene expression analysis.

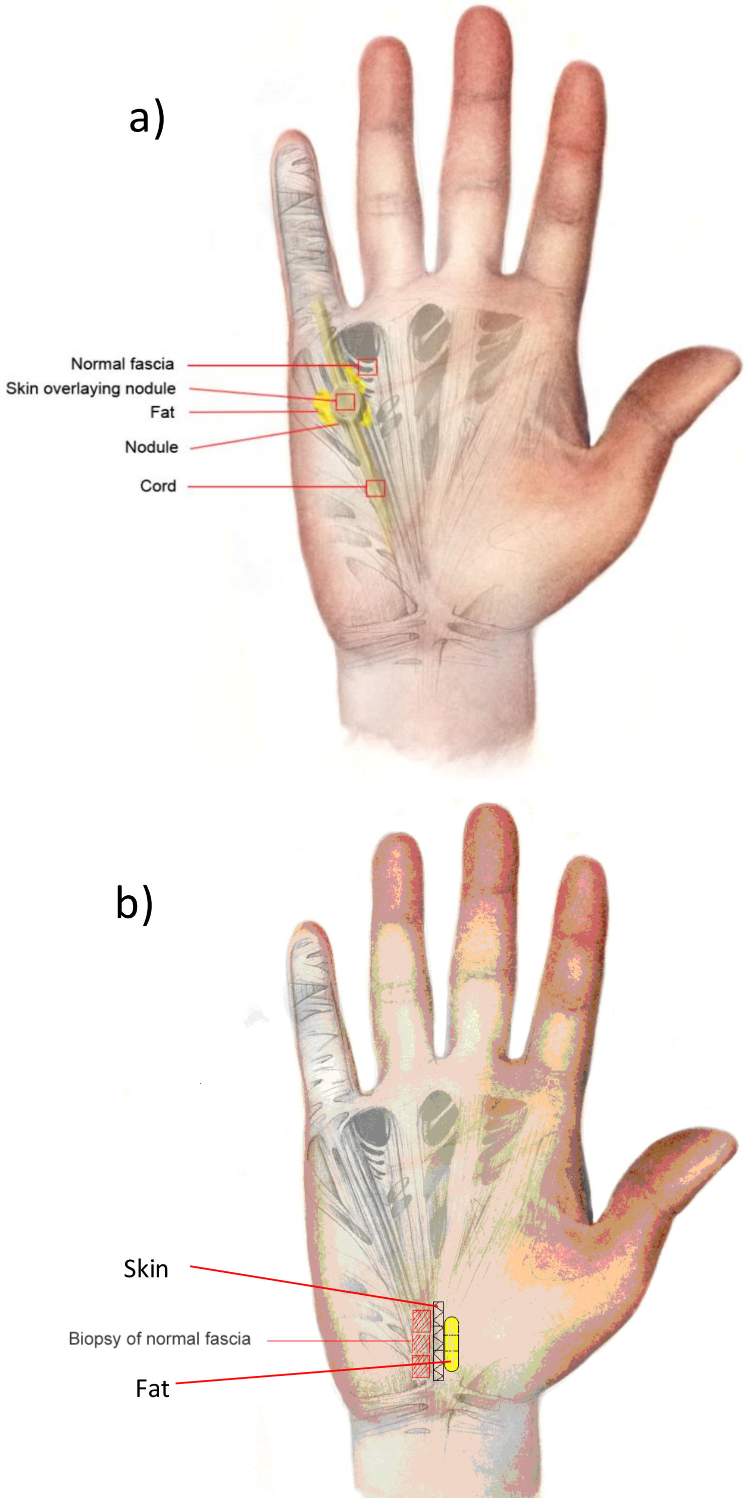


Figure 3 (a) The five sites from the diseased hand. The Dupuytren's disease-associated tissues that are subjected to analysis in this study. Five different Dupuytren's disease-associated tissues in each patient's hand are collected, the normal fascia (unaffected transverse fascia), palmar nodule and cord, skin overlying nodule, and fat. **3 (b) The palm of an unaffected individual used as control.** The palm of the hand of a control subject, where the overlying skin has been removed to demonstrate the position of the palmar fascia harvested. Skin, palmar fascia (transverse carpal ligament) and fat were obtained from control subjects, individuals undergoing carpal tunnel release.

2.3.1.2 Specimen Processing & Tissue/Cell Culture

To establish the tissue cultures, the biopsies were further dissected into small pieces, roughly 1mm^3 in size, with sterile scalpels. The tissue pieces were incubated in 0.25-5% collagenase A solution (Roche Diagnostics, GmBh, Germany) at 37^{o}C for 2.5 to 3 hours. The collagenase activity was inhibited using fibroblast culturing media (Dulbecco's Modified Eagle's Medium 3 (Lonza, Belgium) supplemented with 10% heat-inactivated fetal bovine serum (Lonza, Belgium), 1% penicillin/streptomycin (Lonza, Belgium) and 1% non-essential amino acids (Lonza, Belgium)). The digested samples were centrifuged at 1,500 rpm (approximately 400 ×g) for 5 minutes. Each pellet was re-suspended in 5mL fibroblast culturing media, seeded to 25cm^2 culturing flask (Corning, UK) and incubated at 37^{o}C in 5% CO_2. The culturing media was replaced every 48 hours and cell passages were carried out at approximately 80-90% confluency using trypsin-ethylene diamine tetraacetic acid (200mg/L ethylene diamine tetraacetic acid, 500mg/L trypsin; Lonza, Belgium). The first sub-culturing (passage) was performed on cultures that were grown directly from the biopsies. These were called passage 0 (P0). Following the first centrifugation of cells, ½ of the pellet obtained was seeded onto 1 x 75cm^2 in study 2 (now called P1 cells) and the ½ pellet was frozen down in DMSO containing freezing media. (In Study 1, all cells from the pellet were passaged onto new flasks for the first time (4 x 25cm^2 in equal amounts) and no sample from the previous passage (P0) was retained. From here on further passaging was performed on ¼ of the cells grown from the current passage (called P1) and ¾ was kept in freezing media containing DMSO into 3 separate nunc tubes equally and transferred into Mr Frostie at room temperature. These were then stored at -80ºC until required for FT-IR analysis. During the incubation period the fibroblasts purity was assessed by morphological observation under an inverted phase contrast microscope. The spent culture medium containing all excreted metabolites (footprint) were kept in 15 mL falcon tubes and stored at –80°C until analysis. The approximate % confluency (85-90%) and days each sample was passaged was recorded. All passages of the cell cultures were used in this study. The steps involved in sample preparation to sample interrogation are illustrated in Figure 4. All work was conducted using a single production batch of serum.

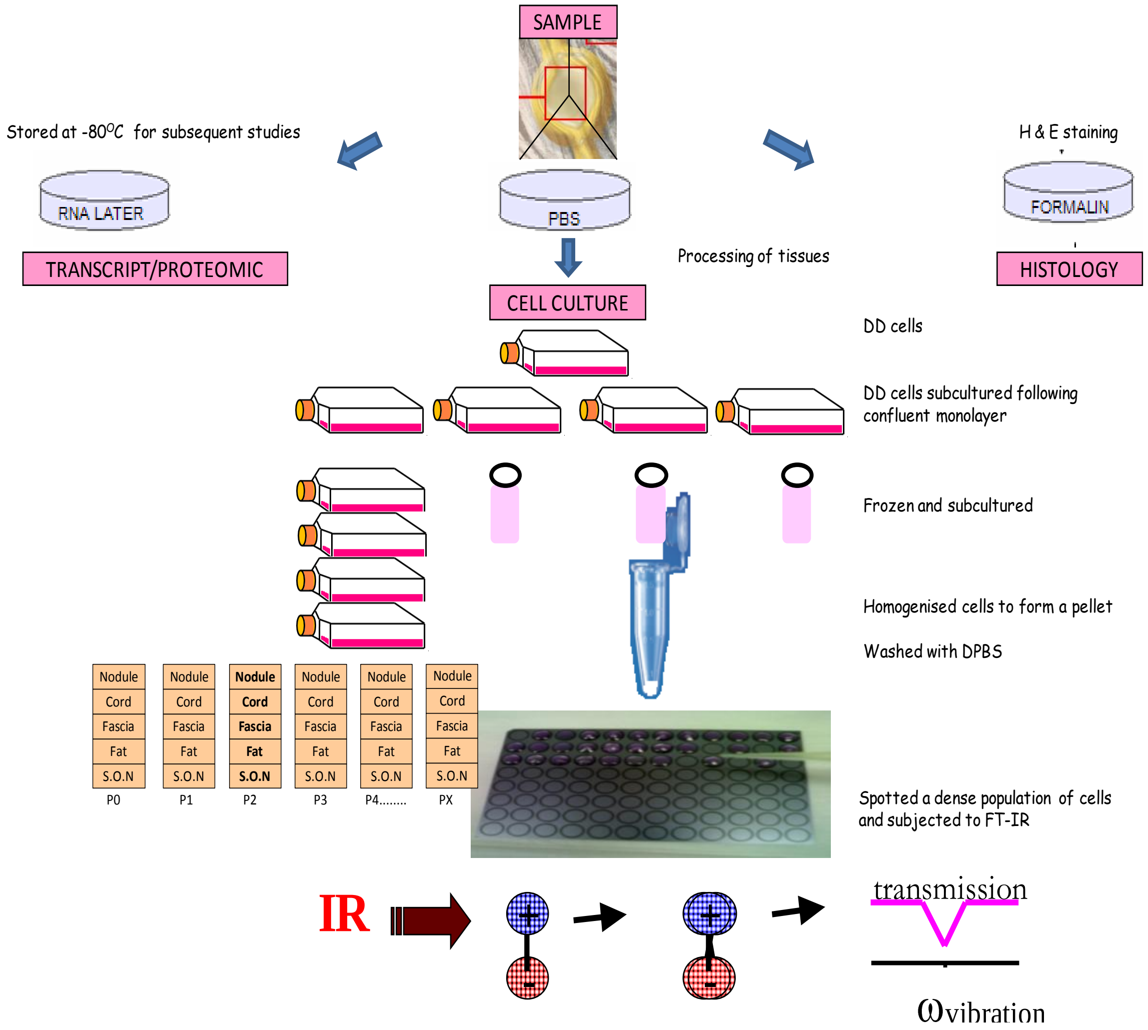


Figure 4 A diagram to show the steps undertaken from tissue processing to sample interrogation by FT-IR.

Morphological Assessment and Haemocytometer Counting

Morphological changes of DD and CTD fibroblasts in all groups were monitored under an inverted phase contrast microscope during the tissue culture experiments. At the end of each passage, cells were washed with phosphate buffered saline (PBS), and detached from the culture dish by 0.25% trypsin. Cell number was then counted through direct visualisation using a haemocytometer. Prior to sub-culturing, 20 μL of cell suspension was then mixed with 20 μL Trypan Blue (Nacalai Tesque, Inc.), a dye exclusion that leaks into cells with damaged plasma membranes. In this way, the dead cells were stained blue, allowing the living and the dead cells to be distinguished. A 10 μL amount of this solution was then placed in a haemocytometer and the number of living cells counted, while being viewed with a light microscope. Figure 70 (i) and (ii) in Appendix B show confluent monolayer images for nodules and cords fibroblast samples. Average no. of cells per T25 cm^2 flask were 600-650,000 for nodules, 400-500,000 for cords, 350-400,000 for the internal control fascia, 400-500,000 for the subcutaneous fat and 600-670,000 for the skin overlying the nodules. The average number of cells per flask and the number of days between each passage for each sample was recorded.

2.3.1.3 FT-IR-Fingerprint Sample Preparation

15μL aliquots of the sample cells dissolved in 50μL DPBS suspensions were evenly applied onto a polished silicon (Bruker, Coventry) microplate containing 96 wells. Each sample was spotted in triplicates in a random arrangement. Prior to analysis the samples were oven dried at 50°C for 30 min. Samples were run in triplicate. The FT-IR instrument used was the Bruker IFS28 FT-IR spectrometer (Bruker Spectrospin Ltd., Coventry, United Kingdom) equipped with a mercurycadmium-telluride detector cooled with liquid nitrogen. The silicon plate was then loaded onto the motorised stage of the FT-IR.

The IBM-compatible personal computer used to control the IFS28 spectrometer was programmed (using Opus, version 2.1, software running under IBM O/S2 Warp provided by the manufacturers) to collect spectra over the wavenumber range of 4,000 cm^{-1} to 600 cm^{-1} at a rate of 20 s^{-1}. A spectral resolution of 4 cm^{-1} was used. To improve the signal-to-noise ratio, the spectra were co-added and averaged (Study 1; 468 and Study 2; 927). Each sample was thus represented by a spectrum containing 1764 points and spectra were displayed in terms of absorbance calculated from the reflectance-absorbance spectra using the Opus software. The combination of both biological and analytical replicates was employed in the FT-IR analysis to measure the biological variance in the data. The employment of ‘analytical or machine replicates’ may be averaged to reduce the heterogeneity of the biological replicates in the analysis of the data. An empty well was used to take a reference background measurement.

2.3.1.4 FT-IR-Footprint Sample Preparation

20uL of the secreted metabolites collected in spent media were spotted directly onto the Silicon microplates, oven dried and subjected to FT-IR as above.

2.3.1.5 FT-IR data pre-processing and multivariate statistical analysis

Preprocessing

All statistical analyses were performed using Matlab R2010a (MathWorks, Inc., MA). The ASCII data were imported into Matlab. To minimise problems arising from baseline shifts, empirical pre-processing techniques to reduce/eliminate light scattering effects were applied. The following procedures in Matlab (methods 1 & 2) were implemented. *Method 1* The spectra were first scaled / normalised so that the smallest absorbance was set to 0 and the highest absorbance was set to +1 for each spectrum. These normalised spectra were then detrended by subtracting a linearly increasing baseline from 4,000 to 600 cm^{-1}; and finally, the smoothed first derivatives of these normalised and detrended spectra were calculated by using the Savitzky-Golay algorithm [125]. *Method 2* – (yielded best separation between tissue phenotypes) To reduce/eliminate light scattering effects an empirical pre-processing technique (extended multiplicative scatter correction (EMSC) [126], order 4) was applied to the spectra. These normalised spectra were then detrended by subtracting a linearly increasing baseline from 4,000 to 600 cm^{-1} where necessary. This method was also tested using PyChem 3.05a [127].

Cluster analyses

Data were grouped into four categories (the samples included are listed in Tables 25-30 in Appendix B)) listed in combinations in Table 3 for comparisons using MVA. The pipeline used for MVA on the hyperspectral FT-IR data is detailed in Figure 5. To reduce the dimensionality of the multivariate data whilst preserving most of the variance, Matlab (and PyChem (3.05a)) [127] was used to perform PCA according to the NIPALS algorithm [91, 94]. For each PCA performed on the chosen combination of samples, of the original 1764 spectral points in the respective study, the maximum percentage of the total variance retained in the first 20 principal components (PCs) was recorded for supervised discriminant function analysis (DFA).

DFA was employed for samples where PCA was insufficient for clear pattern recognition from the scores plot or where the sample and group size was greater than PCA could withstand. DFA; also known as canonical variate analysis [96] then discriminated between groups on the basis of the retained PCs that were used as inputs to the DFA algorithm with the *a priori* knowledge of which spectra were replicates. DFA was programmed to minimise 'within-group' variance and maximise 'between-group' variance, and as this process uses information based on the biological replicates from each sample, it does not bias the analysis, allowing any natural trends or time-dependent trajectories to be observed.

Using PyChem, the Euclidean distance between *a priori* group centers in DFA space using the first two functions (DF1 and DF2), was used to construct a similarity measure, and these distance measures were then processed by an agglomerative clustering algorithm to construct a dendrogram.

2.3.1.6 Haematoxylin and Eosin Staining

Histological examination of palmar aponeurosis tissue specimens stained with haematoxylin eosin (H&E) to confirm macroscopical differences in DD phenotypes and control were performed. The H&E slides from disease and control tissues are shown in Appendix C; Figures 71 (i-viii).

Table 3: List of categories and the sample set chosen for principle component analyses.

CATEGORY	STUDY	SAMPLE SET
1	1A, 2A	Differentiation of diseased subsets (nodule, cord) and fascia (internal control) wrt passage number
2	2C	Differentiation of diseased subsets (nodule, cord) with fascia (internal control) and CTD fascia (external control) wrt passage number
3	2B	Differentiation of diseased subsets (nodule, cord) with fascia (internal control), fat and SON wrt passage number
4	1A, 2A, 2B	Differentiation of individual sites, from all passages wrt to individual patient

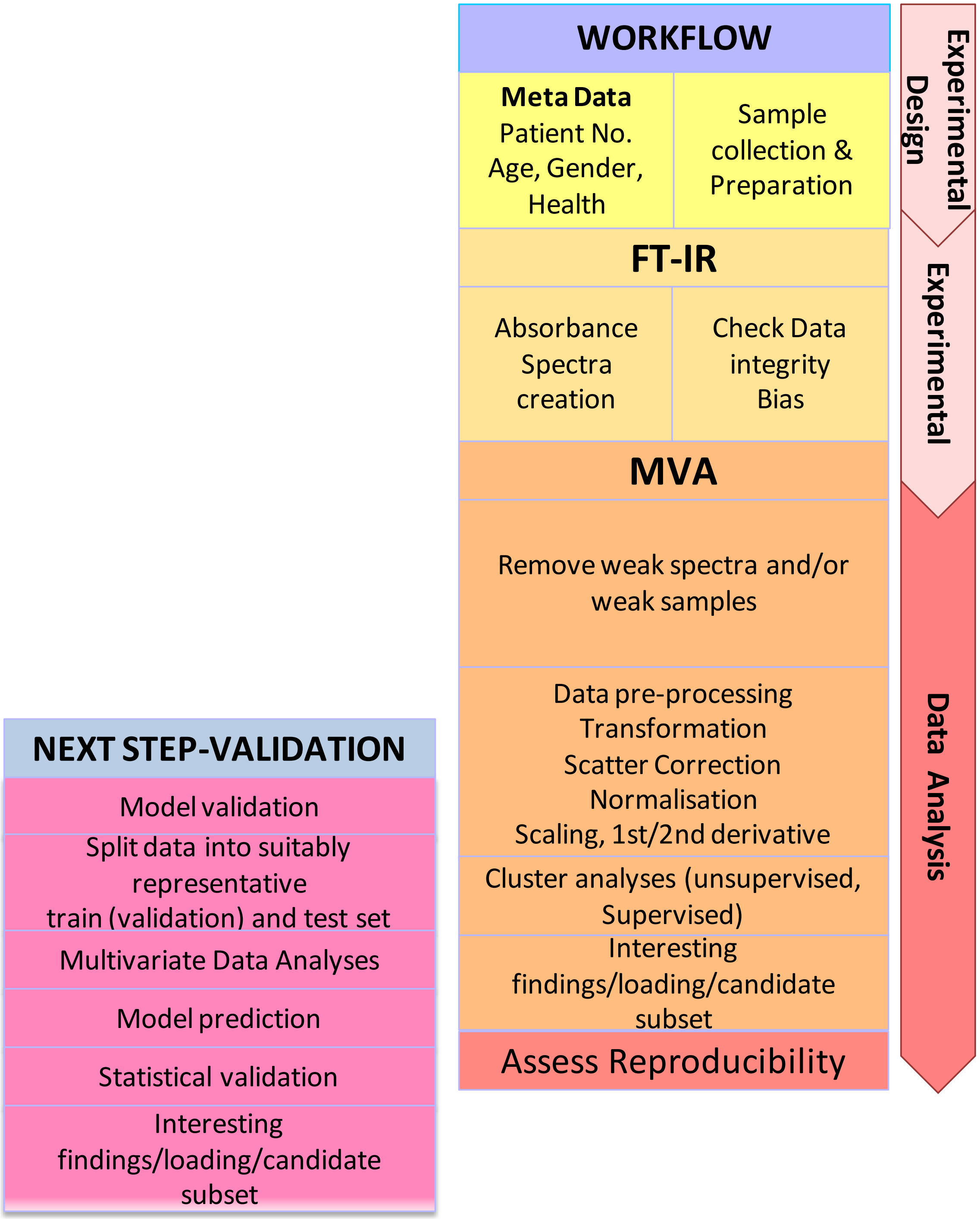

WORKFLOW
Meta Data
Patient No.
Age, Gender,
Health
Sample collection & Preparation
FT-IR
Absorbance Spectra creation
Check Data integrity Bias
MVA
Remove weak spectra and/or weak samples
Data pre-processing
Transformation
Scatter Correction
Normalisation
Scaling, 1st/2nd derivative
Cluster analyses (unsupervised, Supervised)
Interesting findings/loading/candidate subset
Assess Reproducibility
Experimental Design
Experimental
Data Analysis
NEXT STEP-VALIDATION
Model validation
Split data into suitably representative train (validation) and test set
Multivariate Data Analyses
Model prediction
Statistical validation
Interesting findings/loading/candidate subset

Figure 5 A workflow representing the procedure implemented to analyse the samples from Dupuytren's disease and control fibroblast cell cultures.

2.3.2 Metabolic profiling of fibroblasts under oxidative stress using GC-MS (Chapter 4)

2.3.2.1 Experimental design

Patient Recruitment – Study 3 (GC-MS and Microarray Chapter 4 & 5)

DD patients (*n* = 8 men) were used in this study. The age range was between 55 and 70 years with a mean age of 67 years (SD ±7). Patient recruitment procedure was followed as previously stated in Section 2.3.1.1 and demographic information is given in Appendix D. Figure 6 illustrates the experimental design for this study and for part of Chapter 5.

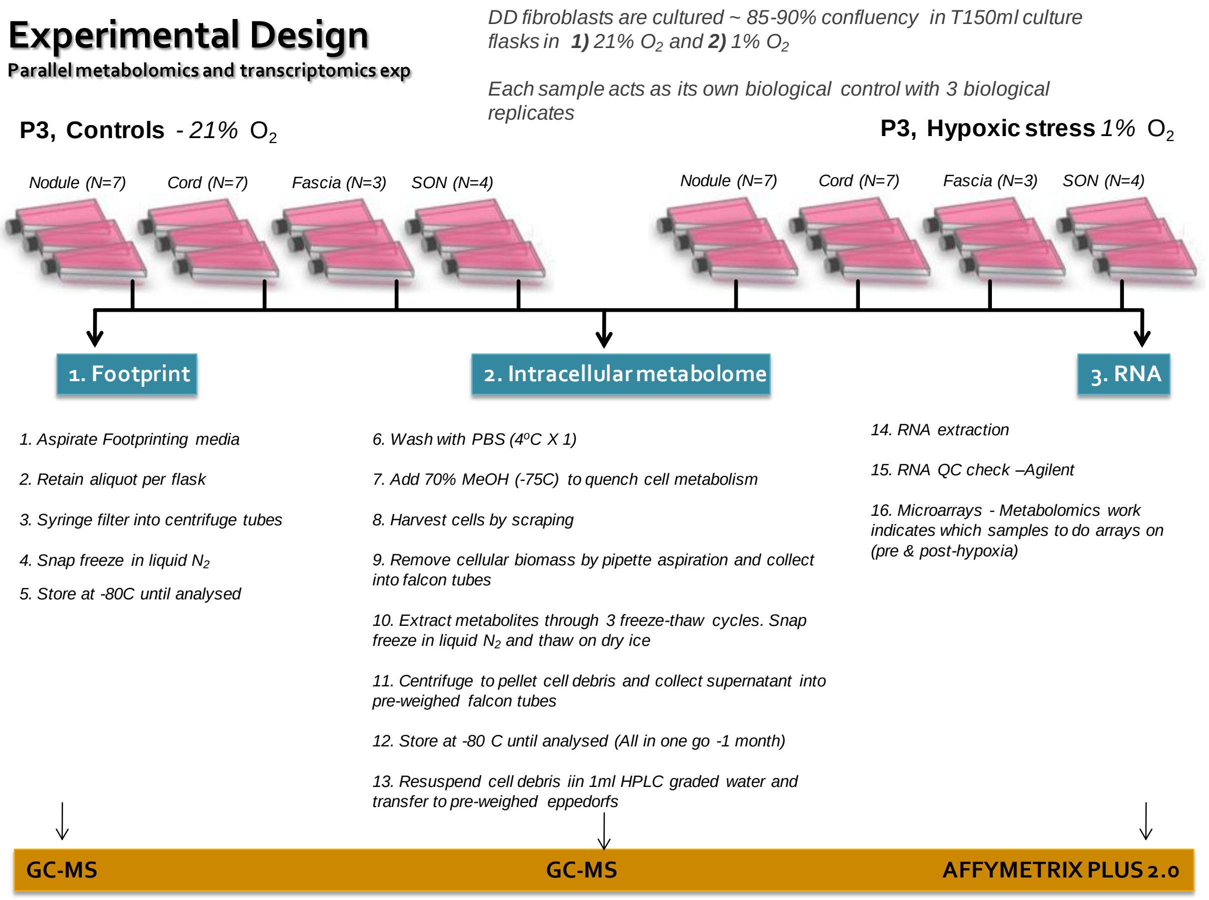
Source of Biopsy Tissue Specimens

Tissue biopsies were obtained from 8 male DD patient having dermofasciectomy: nodules from the palm (*n* = 7), cords from the palm (*n* = 7); of which 6 nodules and cords were obtained from the same individuals. Transverse palmar ligament (internal control; *n* = 3) and SON (*n* = 4) were also obtained where possible. Biopsies were harvested at the time of surgery. Tissue was carefully excised to include merely the diseased or the normal fascia without the adjacent adipose/connective tissue. The tissue was then placed immediately into DPBS and transported to the laboratory within an hour to establish cell cultures.

2.3.2.2 Experimental Protocol

Overview

This protocol describes methodology for quantifying the concentrations of endogenous and secreted metabolites in cultured cells using GC-MS. Cell cultures were established for each sample in 21% and 1% oxygen concentrations. In this way, each sample acts as its own biological control with 3 biological replicates. Many metabolites turn over very rapidly; thus, correctly measuring intracellular metabolite concentrations requires the ability to sample cells quickly. If not, the measured levels will reflect the metabolic state induced by the handling steps leading up to quenching of metabolism, rather than normal cellular physiology. As previously described in Section 2.3.1 DD fibroblast are adherent cells. Metabolism was quenched with minimal perturbation of the culture via quick aspiration of medium, one wash step (4°C) and addition of cold (-75°C) 70% methanol. The solvent addition stopped metabolism (initially due to the temperature drop and subsequently by denaturing enzymes) and simultaneously initiated the extraction process by disrupting the cell membrane. Sample preparation for GC-MS analysis followed as described below.



Experimental Design

Parallel metabolomics and transcriptomics exp

*DD fibroblasts are cultured ~ 85-90% confluency in T150ml culture flasks in **1)** 21% O_2 and **2)** 1% O_2*

Each sample acts as its own biological control with 3 biological replicates

P3, Controls - *21%* O_2

P3, Hypoxic stress *1%* O_2

Nodule (N=7) *Cord (N=7)* *Fascia (N=3)* *SON (N=4)*

Nodule (N=7) *Cord (N=7)* *Fascia (N=3)* *SON (N=4)*

1. Footprint

1. Aspirate Footprinting media

2. Retain aliquot per flask

3. Syringe filter into centrifuge tubes

4. Snap freeze in liquid N_2

5. Store at -80C until analysed

2. Intracellular metabolome

6. Wash with PBS (4°C X 1)

7. Add 70% MeOH (-75C) to quench cell metabolism

8. Harvest cells by scraping

9. Remove cellular biomass by pipette aspiration and collect into falcon tubes

10. Extract metabolites through 3 freeze-thaw cycles. Snap freeze in liquid N_2 and thaw on dry ice

11. Centrifuge to pellet cell debris and collect supernatant into pre-weighed falcon tubes

12. Store at -80 C until analysed (All in one go -1 month)

13. Resuspend cell debris iin 1ml HPLC graded water and transfer to pre-weighed eppedorfs

3. RNA

14. RNA extraction

15. RNA QC check –Agilent

16. Microarrays - Metabolomics work indicates which samples to do arrays on (pre & post-hypoxia)

GC-MS **GC-MS** **AFFYMETRIX PLUS 2.0**

Figure 6 Experimental protocol illustrating steps involved in combined extraction method for metabolomics and transcriptomic level investigation to extract 1. metabolic footprint 2. intracellular metabolome and 3. RNA from same population of cells

Cell culture

A total of 126 samples were processed and stored. Samples from (passage 1) DD nodules (n = 7), cords (n = 7), fascia (n = 3) and (SON (n = 4)) were seeded into T25 cm^2 culture flasks. Upon 90% confluency, samples were passaged into 2 *x* T75cm^2 culture flasks. Approx 1-1.5 million cells were obtained from each T75cm^2 flask. Upon 85-9% confluent, these were subcultured into 6 *x* 150cm^2 flasks; (3 *x* 150cm^2 in normoxic O_2, 3 x 150cm^2 in 1% O_2, 3 replicates for each sample. DD fibroblasts were grown until 85-90% confluent in **a)** 21% O_2, 5 % CO_2 and **b)** 1% O_2, 5% CO_2. A total of 63 flasks were in normoxic culture and 63 in hypoxia. Culture medium for conditions (a) and (b) had same formulation i.e. DMEM

(500mL) supplemented with L-glutamine (1%), NEAA (1%) and FBS Gold (10%) and Pen/strep (1%). Preparation of the 1% oxygen media (termed hypoxic media) is given below. For the purposes of these experiments, 21% O_2 was regarded as normoxic and 1% O_2 as hypoxic conditions for the term DD cultures. After 48-72 h, culture medium was replaced with fresh equilibrated medium, All work reported was conducted using a single production batch of serum.

Preparation of 1 % oxygen (hypoxic) media

Data: 1 mol of gas = 22.4 L at 25^{o}C, hence 23.0 L at 37^{o}C. Therefore the concentration of oxygen (O_2) in pure oxygen is: 1000/22.4 mM at 25^{o}C, 1000/23 at 37^{o}C. Air contains approximately 20.8 % oxygen, hence the oxygen concentration in air is: at 25 ^{o}C: [O_2]air = 0.208*1000/22.4 mM= 208/22.4=9.3 mM and at 37^{o}C: [O_2]air = 0.208*1000/23.0 mM= 208/23.0=9.0 Mm. Oxygen concentration in water at 25^{o}C:- 'Water solubility of oxygen at 25^{o}C and pressure = 1 bar is at 40 mg/L water [128]. In air with a normal composition the oxygen partial pressure is 0.2 atm. This results in dissolution of 40 x 0.2 = 8 mg O_2/L in water that comes in contact with air. A formula that calculates for 760 Torr pressure and 25 °C : 8.3 mg O_2/L is given in [129-131]. In terms of molarity this gives: [O_2] = 8.3/32=0.26 Mm. We need higher temperature however: For 37°C the formula gives: [O_2]= 6.9/32=0.22 Mm (760 mmHg (Torr), 29.92 in Hg, 14.696 PSI, 1013.25 millibars, hence 1 atmosphere is approximately 760 Torr). The ratio in oxygen molarities in gas to water is: At 25°C : 36 and at 37 °C : 41. In what follows we assume that this ratio is 40 at all temperatures.

Making use of the head space

Measure the total volume of the bottle including all the head space. Let us call this volume Vml.

Make the empty bottle anaerobic with N_2 gas. Add a volume of X = x**.**V ml to the bottle. Make sure that the head space is anaerobic by using N_2 gas. Close the bottle with an oxygen impermeable cap. We assume that this is done with medium at 37°C. Now we calculate X:

The total amount of oxygen in the bottle before equilibration is:

$$x \cdot V \cdot 0.22 + (1-x) \cdot V \cdot 0 = x \cdot V \cdot \frac{0.22}{20.8} + (1-x) \cdot V \cdot \frac{0.22}{20.8} \cdot 41$$

$$x \cdot 20.8x \cdot 20.8 = x + (1-x) \cdot 41$$

$60.8 \cdot x = 41$

$x = 0.67$

Two-thirds of the bottle was filled with aerobic medium and the headspace made anaerobic with N_2.

Metabolite quenching & extraction and obtaining RNA for microarrays

Isolation of the exometabolome, intracellular endometabolites and total RNA from the samples was achieved using a novel combined extraction method developed recently and tested on neuroblastoma cell lines [132]. This method permits the isolation of three samples from the same population of cells: metabolic footprint, intracellular metabolites and RNA. Appropriate changes to this protocol were made for optimal extraction of glycolysis intermediates without compromising on RNA integrity or quality.

Isolation of the Metabolic Footprint and Intracellular Metabolites

Cells were grown from each biological sample until 85-90% confluent. Metabolic footprint samples containing exometabolome (secreted metabolites) were obtained from 1mL of used growth media. The media was aspirated from cultures, sampled directly from the flask and passed through a 0.2µm filter to remove any cells and collected in 2mL pre-labeled Eppendorf tubes. The media was immediately snap-frozen in liquid nitrogen, placed in dry ice and then stored at -80^{o}C until ready for GC-TOF-MS analysis.

Following the collection of footprint, the remaining medium was aspirated and the cells were washed once with 4^{o}C PBS (12mL) and quickly aspirated any traces. Immediate addition of 70% methanol (8mL pre-chilled to -75^{o}C) was added to quench metabolic activity. The selection of cold methanol: water 70:30 as an extraction solvent was based on previous systematic studies for extraction of glycolysis intermediates. A large cell scraper was used to harvest cells quickly over ice as addition of the quenching solution increases the temperature. The cellular biomass was removed by pipette aspiration and collected in 15mL centrifuge tubes. Metabolites were immediately snap frozen in liquid nitrogen. Metabolites were then extracted through 3 freeze-thaw cycles (vortexed for 30-40 s each time and thawed on dry ice) in order to permeabilise the cells, resulting in the leakage of the metabolites from the cells. The supernatant (containing the metabolite) was collected by centrifugation (15000 × *g* for 7 min) and transferred into pre-weighed falcon tubes and placed in dry ice. The extracts were stored at -80^{o}C until ready for MS preparation. To one of the three replicates pellets containing the cell debris and the RNA, trizol (1mL) was added and stored at -80^{o}C until ready for RNA extraction post metabolomics data analysis. Two of the three replicates pellets containing cell debris and RNA were dried to remove residual solution and then weighed to determine the mass of dried biomass after extractions to determine volume of metabolites needed for analysis.

Derivatization of metabolites for GC-TOF-MS analysis

Samples were prepared for MS immediately before the analysis was carried out by determining the accurate volume of extracts to dry down. These volumes were then lyophilized for 16 h in a vacuum concentrator (HETO VR MAXI with RVT 4104 refrigerated vapor trap; Thermo Life Sciences, Basingstoke, U.K). For MS analysis endometabolome samples (200 µL) were spiked with 4 µL internal standard solution and lyophilised. Dried extracts were then derivatized as follows; 50 µL of 20 mg mL^{-1} *O*-methoxylamine hydrochloride in pyridine was added, vortexed, and incubated at 80 °C for 15 min in a dri-block heater. A volume of 50 µL of MSTFA was then added and the extracts incubated at 80°C for a further 15 min. On completion, 20 µL of retention index marker solution was added (0.3 mg mL^{-1} docosane, nonadecane, decane, dodecane, and pentadecane in pyridine) prior to centrifugation at 15 800 × *g* for 15 min. The resulting supernatant (90 µL) was transferred to GC-MS vials for analysis.

2.3.2.3 Gas chromatography/Time-of-Flight mass spectrometry analysis

The samples were analyzed in a random order by employing a GC-TOF-MS (Agilent 6890 GC coupled to a LECO Pegasus III TOF mass spectrometer) using a previously described method in [133].

2.3.2.4 Footprint sample analysis

To allow normalisation of response variability, conditioned culture medium was prepared GC-TOF-MS analysis by spiking 200 ml aliquots of cell-free supernatant with 100 ml internal standard solution (0.17 mg/ml succinic d_4 acid).

2.3.2.5 Metabolite identification

The GC-MS data was deconvoluted producing a peak table of the metabolites identified, a three dimensional matrix of information: scan number (related to the time since injection), mass and signal intensity. Raw data were processed using LECO ChromaTof v2.12 and its associated chromatographic deconvolution algorithm, with the baseline set at 1.0, data point averaging of 3 and average peak width of 2.5. A reference database was prepared, incorporating the mass spectrum and retention index of all metabolite peaks detected in a random selection of samples so to allow detection of all metabolites present, whether or not expected from the study of bibliographic data. Each metabolite peak in the reference database was searched for in each sample and if matched (retention index deviation <+/- 10; mass spectral match > 750) the peak area was reported and the response ratio relative to the internal standard (peak area-metabolite/peak area-succinic-d_4 acid internal standard) calculated. These data (matrix of N samples × P metabolite peaks) representing normalised peak lists were exported in ASCII format for further analysis. Metabolites were definitively identified by matching the mass spectrum and retention index of detected peaks to those present in a mass spectral library constructed at the University of Manchester [134]. A match is defined as a match factor greater than 750 and a retention index +/- 10.

Processed data are described by individual metabolite features (chromatographic peaks described by an accurate mass, retention time and peak area). Multiple features can represent a single metabolite. Therefore, raw data are represented by metabolite features and not metabolites. Metabolite features described in the raw dataset were putatively or definitively identified as metabolites in the processed dataset. Putative identification involved the matching of the measured accurate mass to accurate mass(es) present in the Manchester Metabolomics Database (MMD) and as previously described in [134]. Definitive identification involved the matching of accurate mass and retention time of the metabolite to that of authentic chemical standards analysed under identical conditions. The reporting of multiple metabolites for a single feature is a result of several metabolites having the same accurate mass (isomers) which have not been analysed as authentic standards and have identical retention times. Also a metabolite can be detected as different ion types (for example, a protonated species in positive ion mode and a deprotonated ion species in negative ion mode). Only chromatographic peaks assigned a chemical identity as a metabolite are reported, chromatographic peaks not assigned a chemical identity were not reported.

2.3.2.6 Metabolomics Data Analysis - Chemometrics

Analysis of GC-MS raw data and metabolite levels

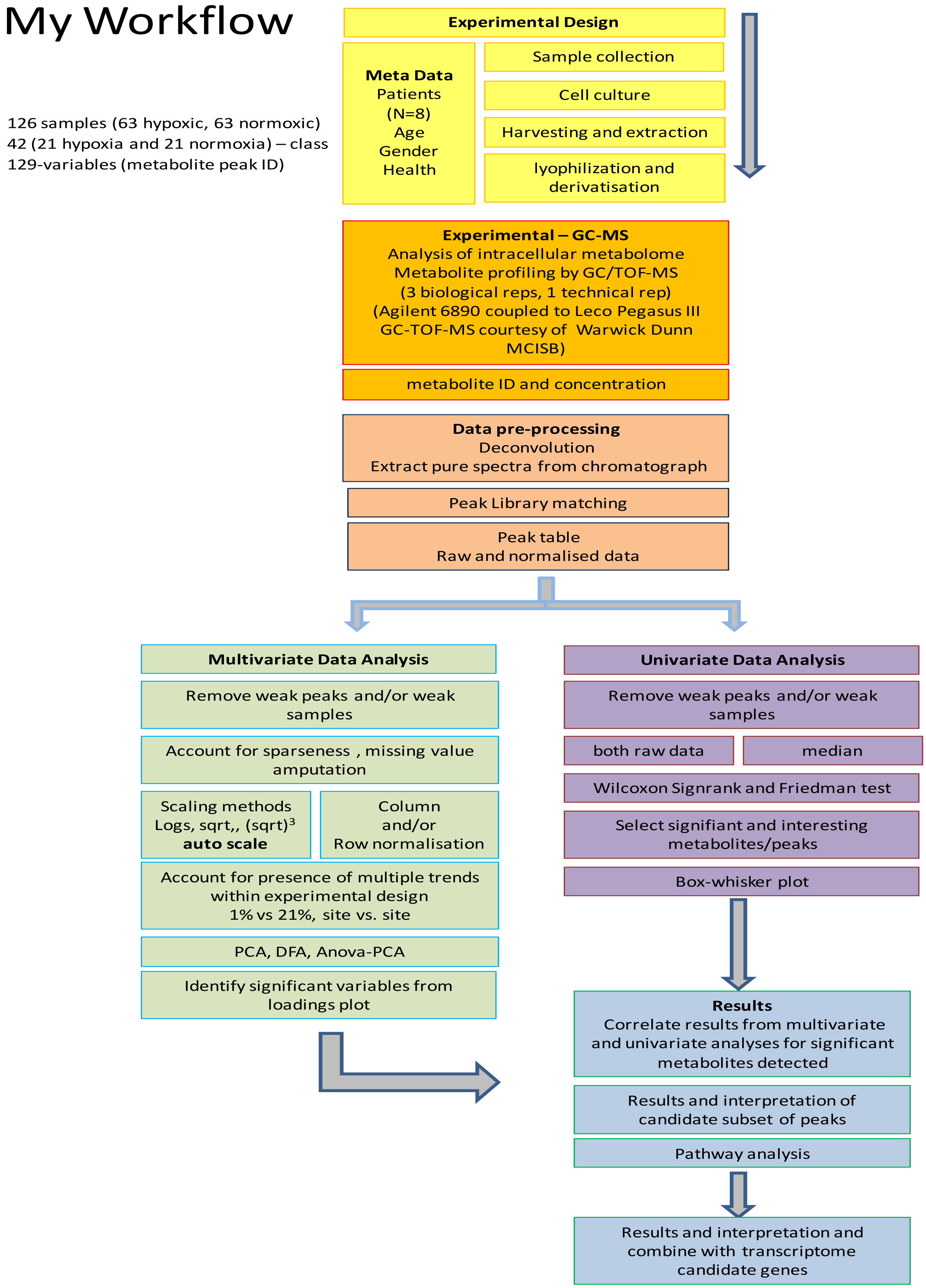
Univariate and multivariate analysis were performed on the ion ratio data sets. All statistical analyses were performed using Matlab R2010a (MathWorks, Inc., MA). Within a GC-MS-based data matrix composed of response ratios (peak area-metabolite/peak area - internal standard), it is possible to obtain zero (or not detected) values for any given metabolite peak caused by either of the following reasons: the metabolite is not present or is present at a concentration below the limit of detection; or the metabolite cannot be resolved from others in the chromatograph by the deconvolution software. In these cases, the following procedure was used to improve data structure for statistical analysis techniques. For univariate analysis, two approaches were applied: 1) All data was accounted and the zero values were replaced with ‘NaN’ (not a number) 2) only the median value of the three replicates was accounted. In any case, all peaks with more than 20% missing values were removed from the analysis. Outliers were suppressed using 95% Winsorisation – eyeballing data from raw PCA scores plots. For multivariate analysis, if two of three replicates were zero values and the third replicate was a non-zero value, the third (non-zero) replicate was replaced with zero. If two of three replicates were non-zero values and the third replicate was a zero value, the zero value for the third replicate was replaced with the mean of the other two replicates.

Prior to multivariate statistical analyses, data was normalised to zero mean, unit variance, (also known as auto-scaling) so that results were not dominated by a small number of high intensity peaks but gave equal weighting to peaks of low intensity. A number of normalisation methods were investigated including taking the $\log_{10}$, square root, cube root and quad root to reduce the dominating effect of higher intensity peaks. MVA were then performed using unsupervised (PCA) and supervised clustering methods using principal components-discriminant function analysis (PC-DFA). Analysis of variance - principal component analysis (ANOVA-PCA) was also used to quantify the relative variance arising from site, oxygen tension, analytical uncertainty and to test the significance of differences in the chemical composition.

The PC-DFA models constructed were cross-validated by iteratively removing one-third of the available data from the training set and using it as a test set, where these data points were projected into the PC space and subsequently the discriminant function space and ANOVA-PCA scores space built by the training data.

Univariate statistical analysis was performed, using the non-parametric Wilcoxon sign rank test and Friedman test to identify metabolites which showed significant difference between two types of samples. The critical *p*-value for rejecting the null hypothesis in a single test is 0.05. However, where many metabolites are tested in parallel, the *p*-value for rejecting the individual hypothesis is typically reduced to lower the probability of type 1 errors (false positives). Therefore, a *p*-value of 0.01 was also used in these experiments and false discovery rate (FDR) controlling test applied. This however, was found too stringent on the data sets returning no metabolites. Boxplots were drawn to show relative concentration distributions for all samples with respect to a given peak. Figure 7 shows the flowchart of steps involved in this experiment, from experimental to data analysis.

Figure 7 Flowchart of metabolomics experiment – from experimental design through to data analysis.


My Workflow
126 samples (63 hypoxic, 63 normoxic)
42 (21 hypoxia and 21 normoxia) – class
129-variables (metabolite peak ID)
Experimental Design
Meta Data
Patients
(N=8)
Age
Gender
Health
Sample collection
Cell culture
Harvesting and extraction
lyophilization and derivatisation
Experimental – GC-MS
Analysis of intracellular metabolome
Metabolite profiling by GC/TOF-MS
(3 biological reps, 1 technical rep)
(Agilent 6890 coupled to Leco Pegasus III GC-TOF-MS courtesy of Warwick Dunn MCISB)
metabolite ID and concentration
Data pre-processing
Deconvolution
Extract pure spectra from chromatograph
Peak Library matching
Peak table
Raw and normalised data
Multivariate Data Analysis
Remove weak peaks and/or weak samples
Account for sparseness , missing value amputation
Scaling methods
Logs, sqrt,, (sqrt)3
auto scale
Column and/or Row normalisation
Account for presence of multiple trends within experimental design
1% vs 21%, site vs. site
PCA, DFA, Anova-PCA
Identify significant variables from loadings plot
Univariate Data Analysis
Remove weak peaks and/or weak samples
both raw data
median
Wilcoxon Signrank and Friedman test
Select signifiant and interesting metabolites/peaks
Box-whisker plot
Results
Correlate results from multivariate and univariate analyses for significant metabolites detected
Results and interpretation of candidate subset of peaks
Pathway analysis
Results and interpretation and combine with transcriptome candidate genes

2.3.3 Dynamic changes in Dupuytren's Disease and control transcriptome in hypoxia (Chapter 5)

2.3.3.1 Experimental design

DD patients 44, 60 & 61 (*n* = 3, men) were entered into the study. These are patients 4, 7 and 8 from the metabolomics study. The 12 samples are listed in Table 4. The following samples were used to perform this study: F1, F21, N1 and N21. Demographic and meta data of these patients can be found in Appendix D.

Table 4 The twelve samples included for transcriptional level analysis.

Sample No.	Patient No.	Patient ID	Site in Hand	Oxygen %
1	4	DD44	Fascia	1
2	4	DD44	Fascia	21
3	4	DD44	Nodule	1
4	4	DD44	Nodule	21
5	7	DD60	Fascia	1
6	7	DD60	Fascia	21
7	7	DD60	Nodule	1
8	7	DD60	Nodule	21
9	8	DD61	Fascia	1
10	8	DD61	Fascia	21
11	8	DD61	Nodule	1
12	8	DD61	Nodule	21

2.3.3.2 Experimental Protocol

RNA extraction

Samples in trizol (collected in Section 2.3.2) were removed from -80°C freezer and allowed to thaw quickly on ice. The contents were transferred to a sterile 1.5mL Eppendorf tube and centrifuged at 13,000 rpm for 10 min at 4°C. The supernatant was then recovered into a new pre-labeled tube and 0.2mL chloroform (Sigma Aldrich) (per mL of trizol used) was added. These were left at room temp for 2 min and then centrifuged at 13,000 rpm for 15 min. The upper aqueous layer was pipetted into a fresh Eppendorf tube and an equal volume of 70% ethanol was added to this and mixed by pipetting up and down. Following this 700µL of the sample, including any precipitate that may have formed was transferred to an RNeasy mini column placed in a 2 mL collection tube. This was centrifuge for 15 sec at 13,000 rpm and the flow through discarded. 700µL buffer RW1 to the RNeasy column was added and then centrifuge for 15 sec at 13,000rpm. The flow through and collection tubes were discarded. The RNeasy column was placed into a new 2 mL collection tube and 500 µL buffer RPE was added (RPE buffer is supplied as a concentrate and 4 vols of 96 – 100% EtOH is added before using). The tubes were centrifuged for 15 sec at 13,000 rpm to wash the column and the flow through was discarded. Another 500 µL buffer RPE was added, centrifuged and the flow through discarded. The RNeasy column was centrifuged for 1 min at 13,000 rpm to dry the RNeasy silica-gel membrane and then placed into a new 1.5mL collection tube, washed with 30 – 50 µL RNase - free water directly on to the RNeasy silica-gel membrane. Tubes were closed and left at room temperature for 1 min and then centrifuged at 10,000 rpm for 1 min. This was followed by sample quantification.

RNA quantification and quality analysis

RNA was quantified and quality checked using a NanoDrop ND-1000 UV-visible spectrophotometer (Labtech International, Ringmer, UK). RNA integrity was assessed using the Agilent 2100 Bioanalyzer (Agilent Technologies UK Limited, Stockport, Cheshire, UK). Only RNA samples that were of sufficient concentration and showed no degradation as evidenced by distinct ribosomal bands at 18S and 28S were used for microarray experiments**.** Ethanol precipitation was also performed on a fractional amount of these samples and these were quantified, quality checked and their integrity assessed as previously. Non- EtOH precipitated samples were used for the cDNA synthesis.

Affymetrix microarray procedure

For each sample, RNA (3µg) was reverse transcribed into cDNA using the 3' IVT Express Kit (Affymetrix) according to the manufacturer’s guidelines [135]. Amplified cDNA was then purified using magnetic beads. Aliquots of labeled cRNA (20 µg) were fragmented and then hybridised to a Human Genome U133 Plus 2.0 GeneChip oligonucleotide array for 16 hours, rotating at 60 rpm at 45°C in a GeneChip Hybridization Oven 640 (Affymetrix). Each chip was washed and stained on a GeneChip Fluidics Station 450 (Affymetrix) and scanned on a GeneChip Scanner 450 (Affymetrix) using manufacturer’s protocol [135].

2.3.3.3 Microarray data analysis

The Affymetrix array data analysis was performed using three methods. The Affymetrix CEL files were uploaded and analysed with the R-Bioconductor tools suite [121] using 1) Limma and 2) puma methods and 3) GeneSifter microarray analysis tool (geopiza, Seattle, WA). The individual methods are described below and Figure 8 illustrates these steps.

Limma: linear models for microarray data

Background correction and quantile normalisation were performed using (1) RMA and (2) GC-RMA in R-Bioconductor. PCA on normalised data was performed to test the quality of the array data in R and MATLAB to confirm correlation between clusters in related gene array chips. Differential expression analysis was performed with Limma using the functions lmFit and eBayes. The analysis was done by creating design and contrast matrices for the following sample replicates: F1*vs.*F21, N21*vs.*F21 and N1*vs.*N21. Gene lists of differentially expressed (DE) genes were created by filtering for probesets with a p-value <0.055 and FC >1.5. These gene expression profiles (gene lists) in terms of p-value and fold change from the mean were compared using the FDR Benjamini and Hochberg multiple test correction [72]. Results with a p-value 0.05 at the 95% confidence level were considered significant. Individual thresholds were further applied (p-value 0.01 to 0.05) to obtain a filtered set of statistically significant genes. Gene ontology (GO) over-representation analysis was performed using the functional annotation tool of the Database for Annotation, Visualization and Integrated Discovery (DAVID) 2.1 program.

PUMA: Propagating Uncertainty in Microarray Analysis

Normalisation and expression analysis was done using multi-mgMOS [136]. Differential expression between the sample groups (F1, F21, N21 & N1) was assessed with puma, a Bayesian method which includes probe-level measurement error when assessing statistical significance. Analysis was performed with the PUMA [119] package in R. Gross differences between arrays were determined using the puma variant of PCA; pumaPCA which can make use of the uncertainty in the expression levels determined by multi-mgMOS. Unlike many other methods, multi-mgMOS provides information about the expected uncertainty in the expression level, as well as a point estimate of the expression level. Standard PCA was applied to multi-mgMOS normalised data on filtered dataset by applying a threshold ≥ 1.5

by comparing variance of the mean expression to mean of standard error for each probe. Differential expression (DE) analysis was performed with the pumaDE function. The results of these commands were ranked gene lists in order of probability of positive log-ratio (PPLR) values, and the FC values. All possible combinations of pair-wise comparisons among experiments were taken to create sets of ratios. A gene list of differentially expressed genes was created by filtering for probe sets with a PPLR values ≤ 0.1 and ≥ 0.9) and further filtered to reduce FDR by increasing to ≤ 0.01 and ≥ 0.99 in any of the comparisons for each patients between F1 *vs.* F2, N2 *vs.* F21 and N1 *vs.* N21 samples.

HCA was applied on the normalised and filtered datasets using Euclidean distance (average linkage). Heatmap was generated based on similarity of expression profiles across the dataset. Pathway analysis of top 1000 and lowest 1000 genes and then on a filtered set of selected genes was carried out with KEGG [63]. Venn diagrams were generated to show the distribution of the probe set for the filtered probesets. Functional and ontology enrichment analysis was performed using the DAVID web-based tool version 2 [137] and the expression analysis systematic explorer (EASE) [138].

GeneSifter microarray analysis tool

Relative changes in gene expression were evaluated by the expression ratio and (FC), as determined from the GC-RMA normalized data reported by GeneSifter. A transcript displaying expression ≥ 1.5 FC from the mean in at least 1 array was used as a cut-off level. The gene expression profile in terms of FC from the mean was compared using the Benjamini and Hochberg multiple test correction. Results with a p-value (adjusted p-value) ≤ 0.05 at the 95% confidence level were considered significant. Individual thresholds were further applied (p-value ≤0.01 to 0.05) to obtain a filtered set of statistically significant genes. Two-way ANOVA test was also applied with above cut off points. Genes and arrays were clustered according to their expression patterns using cluster software, and heat maps were produced. As previously, the subset of filtered genelists were used to study for biological relevance using Gene Ontology, [61] and DAVID.

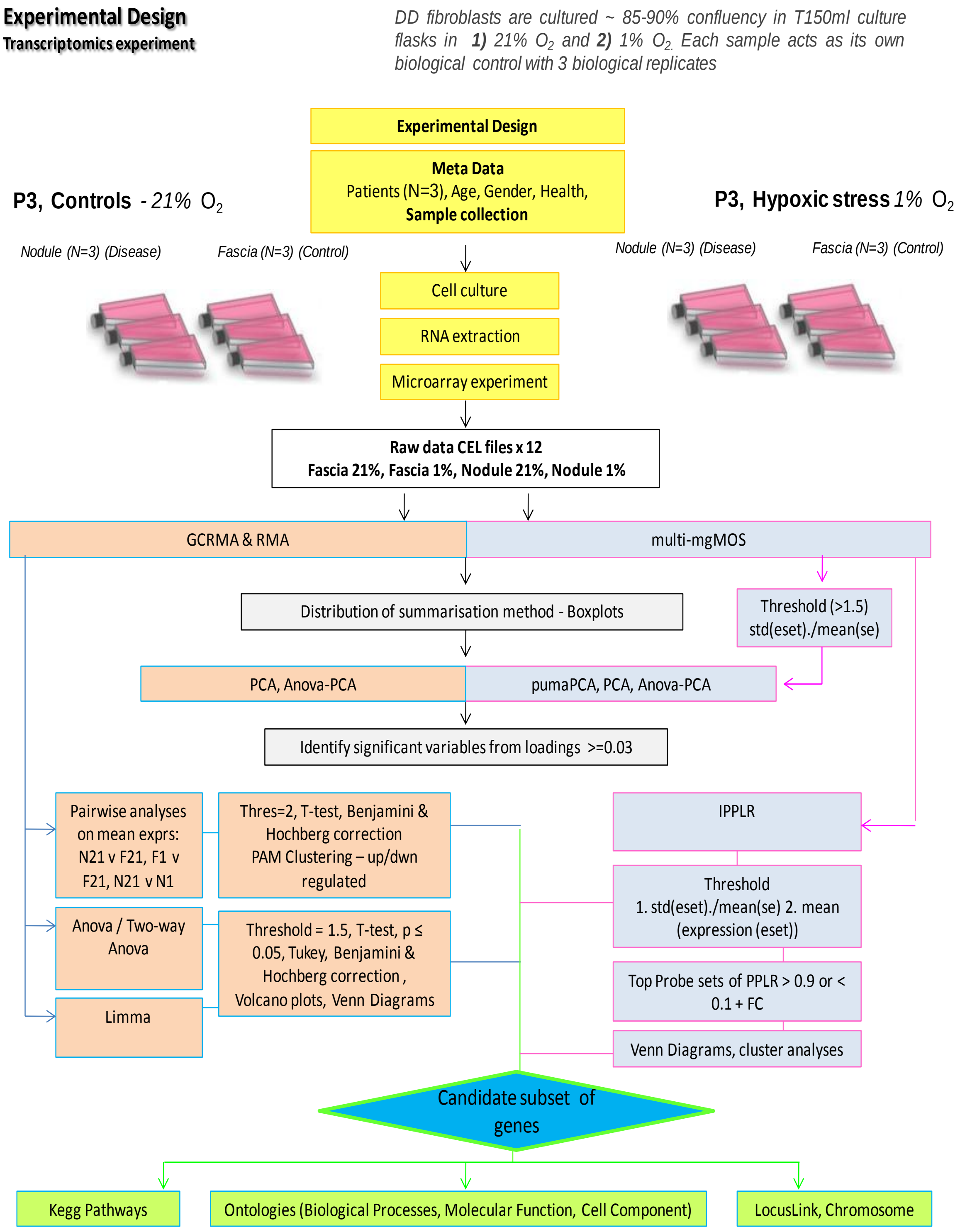

Experimental Design
Transcriptomics experiment
DD fibroblasts are cultured ~ 85-90% confluency in T150ml culture flasks in 1) 21% O_2 and 2) 1% O_2. Each sample acts as its own biological control with 3 biological replicates
Experimental Design
Meta Data
Patients (N=3), Age, Gender, Health,
Sample collection
P3, Controls - 21% O_2
P3, Hypoxic stress 1% O_2
Nodule (N=3) (Disease)
Fascia (N=3) (Control)
Nodule (N=3) (Disease)
Fascia (N=3) (Control)
Cell culture
RNA extraction
Microarray experiment
Raw data CEL files x 12
Fascia 21%, Fascia 1%, Nodule 21%, Nodule 1%
GCRMA & RMA
multi-mgMOS
Distribution of summarisation method - Boxplots
Threshold (>1.5)
std(eset)./mean(se)
PCA, Anova-PCA
pumaPCA, PCA, Anova-PCA
Identify significant variables from loadings >=0.03
Pairwise analyses on mean exprs: N21 v F21, F1 v F21, N21 v N1
Thres=2, T-test, Benjamini & Hochberg correction
PAM Clustering – up/dwn regulated
IPPLR
Threshold
1. std(eset)./mean(se) 2. mean (expression (eset))
Anova / Two-way Anova
Threshold = 1.5, T-test, p ≤ 0.05, Tukey, Benjamini & Hochberg correction, Volcano plots, Venn Diagrams
Top Probe sets of PPLR > 0.9 or < 0.1 + FC
Limma
Venn Diagrams, cluster analyses
Candidate subset of genes
Kegg Pathways
Ontologies (Biological Processes, Molecular Function, Cell Component)
LocusLink, Chromosome

Figure 8 Flowchart illustrating steps involved in microarray study, from experimental design to data analyses.

2.3.4 Inferring the metabolic and transcriptional networks specific to Dupuytren's disease tumours (Chapter 6)

2.3.4.1 Integrated pathway mapping with Ingenuity Pathway Analysis

Data were first analysed using stringent chemometrics and microarray data analysis methods as to generate a list of significantly differentially expressed molecules as described in Section 2.3.3 and 2.3.4. The final set identifiers were combined for each comparison i.e. the HMDB ID's [49] / CAS registry identifiers (for definitive metabolites) and Affymetrix accession Probe ID's were combined into a single data set for each of the pairwise analyses F1 *vs*. F21, N21 *vs*. F21 and N1 *vs*. N21.

IPA: Network generation

A data set containing gene/metabolite identifiers and corresponding expression/mean fold change values was uploaded into the application. Each gene/metabolite identifier was mapped to its corresponding gene/metabolite object in the Ingenuity Pathways knowledge base. Those molecules previously identified as being statistically significant, called focus molecules, were overlaid onto a global molecular network developed from information contained in the Ingenuity Pathways Knowledge Base. Networks of these focus molecules were then algorithmically generated based on their connectivity. The network is a graphical representation of the molecular relationships between the molecules (genes/endogenous chemicals). The molecules are represented as nodes, and the biological relationship between two nodes is represented as an edge (line). All edges are supported by at least one reference from the literature (http://www.ncbi.nlm.nih.gov/pubmed), from a textbook, or from canonical information stored in the Ingenuity Pathways knowledge base. Top scoring networks generated from metabolite data sets alone were mapped onto top scoring transcript data networks and vice versa.

Global functional analysis

The functional analysis identified the biological functions and/or diseases that were most significant to the data set. Molecules from the dataset uploaded that were associated with biological functions and/or diseases in the Ingenuity Pathways Knowledge Base were considered for the analysis. Fisher's exact test was used to calculate a *p*-value determining the probability that each biological function and/ or disease assigned to that data set is due to chance alone.

Global canonical pathways analysis.

Canonical pathways analysis identified the pathways from the IPA library of canonical pathways that were most significant to the data set. Candidates from the data set that were associated with a canonical pathway in the Ingenuity Pathways Knowledge Base were considered for the analysis. The significance of the association between the data set and the canonical pathway was measured in 3 ways: (1) A ratio of the number of genes from the data set that map to the pathway divided by the total number of genes that map to the canonical pathway is displayed. (2) Fischer's exact test was used to calculate a *p*-value determining the probability that the association between the genes in the data set and the canonical pathway is explained by chance alone. (3) Benjamini-Hochberg testing corrected *p*-values were used.

2.3.4.2 Metabolomic pathway analysis with MetPA:

Data Input and Processing

Filtered metabolite lists and their HMDB IDs for each of the three pairwise analyses were uploaded in MetPA [71]. The first step involved standardisation of the compound labels which were subsequently compared with compounds contained in the pathway library. 1 indicated exact match, 2 indicated approximate match, and 0 indicated no match.

Pathway and Over Representation Analysis

The pathway library 'Homo sapiens (human)' containing 80 pathways from KEGG library was selected. The one-tailed Fisher's exact test for pathway enrichment analysis and pathway topology analysis were specified for over-representation analysis. This is to test if a particular group of compounds is represented more than expected by chance within the uploaded compound list. In the context of pathway analysis, we are testing if compounds involved in a particular pathway is enriched compared by random hits.

Pathway Topology Analysis

MetPA [71] uses two well-established node centrality measures to estimate node importance – degree centrality and betweenness centrality. As metabolic networks are directed graphs the relative betweenness centrality algorithm for pathway topology analysis was selected for metabolite importance measure. The betweenness centrality measures focus on global network topology [139].

Chapter 3

Metabolic Fingerprint and Footprint Analysis of Passaged Dupuytren's Disease Cultured Fibroblasts

3.1 Introduction

Previous studies provide compelling evidence that genetic dysregulation plays an important role in DD formation. Macroscopic phenotypic differences between fibrotic elements in DD (i.e. nodule and cord), plus the fat and skin overlying the nodule (SON) also show microscopic differences with variable cellular density. Abnormal fibroblasts are considered to be responsible for causing DD. A recent study investigated phenotypic descriptors of DD tissue (i.e. nodule, cord) and compared with the transverse palmar fascia (internal control) and transverse carpal ligamentous fascia (external control) using the Affymetrix HGU133A GeneChip array reporting several differentially expressed genes thought be involved in the pathogenesis of DD ([10] and Appendix E2(1)). In contrast to our previous study, here, the systemic properties of fibroblast cultures derived from different DD tissue phenotypes including the fat and the SON are investigated. Comparisons across previous studies are impeded because the reported results are often based on different cellular passages which could have a dramatic effect not only for morphological properties but also gene expression and metabolic differences too. Whether any significant changes are due to genetic alterations alone or the consequence of metabolic dysregulation has not to date been demonstrated. In this study, a systemic analysis of Dupuytren fibroblast profiles derived from different DD tissue phenotypes in order to identify the best time to look at the disease pattern is discussed.

The aim of this study is to delineate the morphological components of DD and control by looking at differences in cultures derived from disease tissue phenotypes and comparing different anatomical locations within the DD tissue using an alternative technique to histology. The main advantage is to simultaneously image the quantity and quality of multiple components in DD in relation to normal transverse palmar fascia using a technique with high molecular sensitivity combined with a spatial resolution down to a few micrometres.

A metabolomic analysis of fibroblast cultures derived from different DD tissue phenotypes to compare early (primary) cultures to late passages was employed to identify the most representative passages for the disease. Using FT-IR spectroscopy, a comparison of metabolic profiles of endogenous and secreted metabolites from (1) DD cords and nodules against the unaffected transverse palmar fascia (internal control), (2) DD cords and nodules with fat surrounding the nodule, and SON. Following this metabolic profiles of those in (1) and (2) between different DD patients and (3) those in (1) and were compared between different DD patients and with external controls.

Carefully controlled conditions using multivariate statistical analyses (PCA and DFA) demonstrated early passage (0-3) metabolic differences where a synchronous separation pattern was observed in the PCA scores plots in DD and control fibroblasts. However, higher passages (4-6) demonstrated random asynchronous and overlapping patterns from which it was unclear how to identify and separate the diseased from non-diseased sample phenotypes. The cord and control fibroblast clusters were in closer proximity in the scores plot of PC1 and PC2 across passages 0-3, while the nodule fibroblasts were clearly separable in all PCA plots. The analyses of the PCA scores plots between the five sample types from DD subsets (nodule, cord, fascia, fat and SON) demonstrated overlapping between the clusters in samples across passages 0-3. In the early passages however, the clusters of nodules and SON were in close proximity to each other, while the clusters representing the fibroblasts of cords, fascias and fat were often closer than those of the nodule. The results from FT-IR metabolic fingerprinting combined with chemometrics has demonstrated early passage metabolic differences showing clear separation between the different tissue phenotypes in DD and control fibroblasts, whereas higher passages demonstrated random asynchronous patterns.

3.2 Results

3.2.1 Histological findings

The DD nodule, cord, fat, and the skin overlying the nodule and the internal control – the transverse palmar fascia (*n* = 6) and the carpal ligamentous fascia – the external control from healthy individuals were identified and confirmed by a senior histopathologist. The nodules displayed regions of high cellularity and lots of nuclei were stained/visible, while the cords displayed a tendon like collagen rich structure. Cross sections from the fat and the skin overlying the nodule were also examined and compared with those from normal CTD skin and fat from unaffected patients. All Histological findings can be found in Appendix C.

3.2.2 The raw and preprocessed spectra

Analysis of the metabolic fingerprints

Variations in the passage numbers from cultures were investigated by FT-IR metabolic fingerprinting, using cell samples which had been washed thoroughly to remove any extracellular metabolites medium components. The total number of samples and their combinations selected for MVA can be seen across Tables 25-30 in Appendix B. The results from typical data analysis, preprocessing and MVA techniques are discussed below. The results of the FT-IR spectral data and MVA are presented in the form of PCA, and/or PC-DFA scores plots as discussed under the corresponding headings for each study respectively.

Typical raw and normalised absorbance FT-IR spectra for DD and control fibroblasts are shown in Figures 9 (& Figure 10) and 11 respectively. These are typical vibrational spectra from all samples included in Study 1 and 2. These spectra are the metabolic fingerprints of each sample analysed (metabolic footprint in the case of spent media) on the Si plate respectively; all showing broad and complex contours with relatively little qualitative difference between the spectra visible to the naked eye. Such spectra readily illustrate the need to use multivariate statistical techniques in the analysis of data. The spectra contain information on functional group vibrations resulting in the absorbance of infrared light at specific wavenumbers (1/λ). Some prominent regions (Figure 11) are identified to be arising from vibrational modes of water (O-H stretch centered at 3400cm^{-1}), fatty acids (methyl, methylene and CH_x stretches at 2956-2850 cm^{-1}) proteins (amide I, C=O at 1652-1648 cm^{-1}; amide II N-H, C-N at 1550-1548 cm^{-1}), a mixed region from 1460-110 cm^{-1} which contains information from fatty acids, polysaccharides, nucleic acids, proteins and polysaccharide rings and C-O vibrations at 1085-1052 cm^{-1}). Note here, that large molecules such as proteins as well as small constituents including nucleic acids are detected because the sample preparation used simply DPBS to dissolve the cells and was analysed directly without any interference to the intact cells. FT-IR spectra for various DD phenotype fibroblasts (from all 5 sites), fibroblast growth media and freezing media were also recorded electronically. Again, all spectra showed broad and complex contours, in which there was relatively little qualitative difference between the spectra.

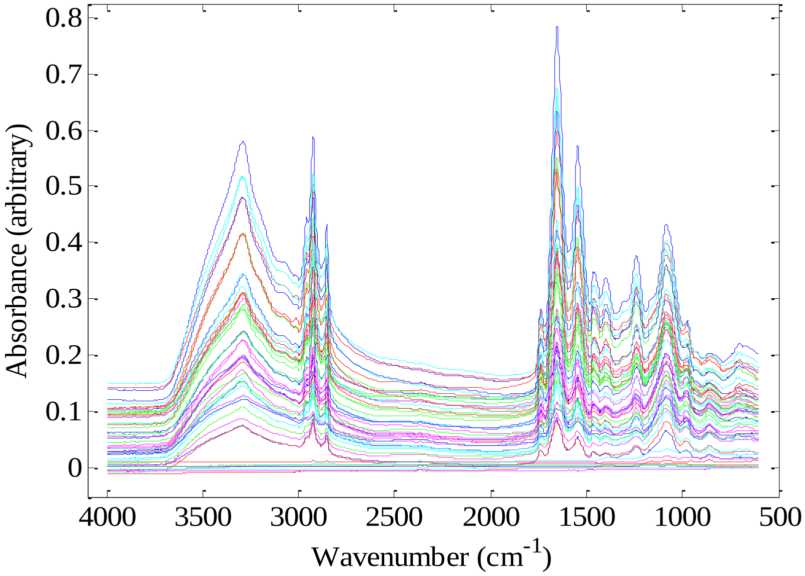


Figure 9 Typical raw FT-IR spectra from metabolic fingerprints of cultured cells. Each sample is represented by an absorbance vs. wavenumber spectrum.

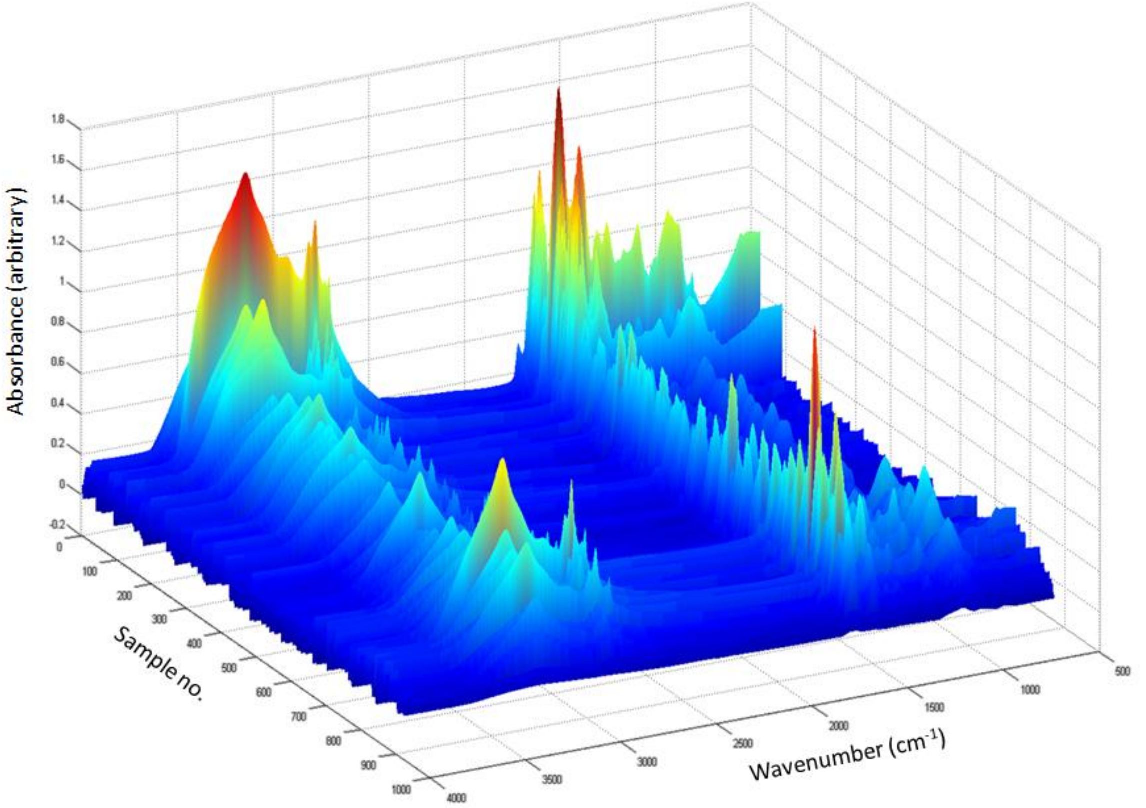


Figure 10 A 3-D raw FT-IR spectra from metabolic fingerprints of cultured cells. Each sample is represented by an absorbance vs. wavenumber spectrum.

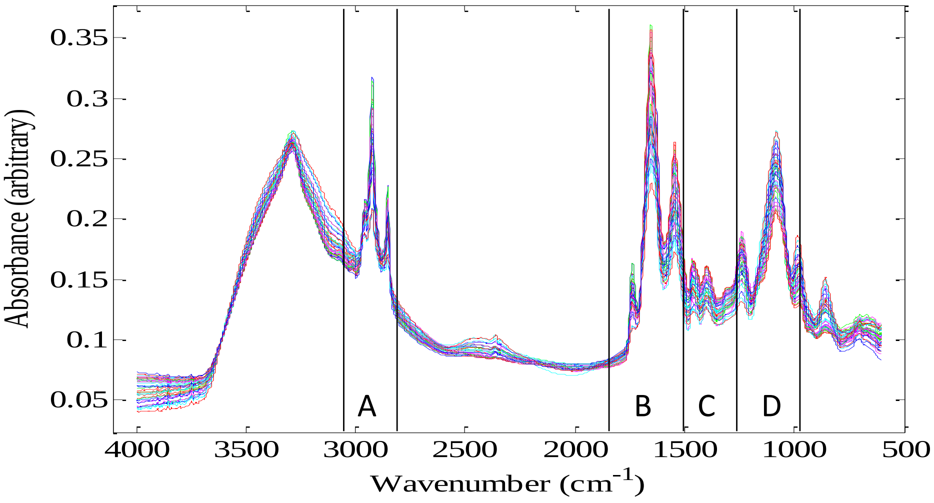


Figure 11 Processed FT-IR spectra of cultured cells with band assignments. Samples are normalised using the EMSC preprocessing method and baseline correction. These spectra have been offset to see the features more readily. Key to vibrational bands: A = fatty acid, B = amide, C = mixed, D = polysaccharide.

3.2.3 Multivariate statistical analyses to determine variability of samples

The next stage was to perform cluster analysis on all preprocessed spectral data. This was done separately for all metabolic fingerprint and footprint data. A total of 309 samples were analysed three times (927) in Study 2 (234 samples analysed twice (468) in Study 1) and PCA was performed in various combinations on the data sets.

Firstly, the relevant passages (e.g. passage 0, 1, 2 etc) from all sites to undergo analysis were grouped by selecting samples from each case and control set. (e.g. all passage (P0) samples from DD nodules, cords and internal fascia, then from another data set all P0 samples from DD nodules, cords and internal fascia and also the external fascia from CTD patient, another data set selecting for P0 cultured cells in all five DD subsets including the fat and SON). This was carried out across all samples selecting on the basis of passage number to determine any similarities and differences across the metabolic profiles (fingerprints in this case) to determine whether the passage had any effect on the individual samples when sub-cultured overtime in same culture medium and conditions. In addition, comparisons of both control fibroblasts were made; i.e. whether internal control demonstrated any similarity to fibroblast from external control and whether the choice of internal control being used for future studies is in fact a more appropriate or suitable representative. The trends sought after were those in which a gradation, similarity and differences within the sample clusters were evident.

Second, PCA was performed across all passages and all samples within the class of dataset in various combinations comparing the sites in individual patients across all passages (e.g. (i) for patient DD2; all passages across nodules, cords, fascias, and then (ii) all passages across (i) plus the fat and SON).

For each study, the two matrices produced from the transformation of the original data matrix, X, in the PCA model, (score and a loading matrix and residual matrix; where the score matrix contains information on how samples are related to each other and the loading matrix shows relationships between variables) were linearly combined with the original data. The determination of the number of significant PCs was solved in three ways. First, the scree plot of eigenvalues, presented in Figure 12 was examined (this is just for demonstration). Its shape suggested selecting between two and three PCs; majority of the variance within the data set are captured by the first few PCs with little importance on the later PCs. Second, the score plots for PC1 & PC2, PC1 & PC3, PC2 & PC3, and subsequent PCs were analysed.

Third, the % cumulative variance plot was examined. It was found that for PCs > 3 the plots became random. From these facts it was decided that 3 PCs would be the best choice. The selection means that most of the variance in absorption of the examined spectra will be further interpreted in terms of the above three factors that will be related to (i) passage number vs. sample type and (ii) individual patients vs. sample type including all passages. Figures 13 and 14 represent the PCA plots from PC1 and PC2 from Study 1 and Figures 16 represent PCA scores plot from DD nodule, cord and internal fascia primary cultures from Study 2 showing the relationships between the three different phenotypes confirming their metabolic differences. Boundaries round the clusters were drawn manually. Each numeric code represents a single biological sample. Observations were made in the relationships between cluster spaces; the closer the samples cluster together the more biochemical similarity they possess.

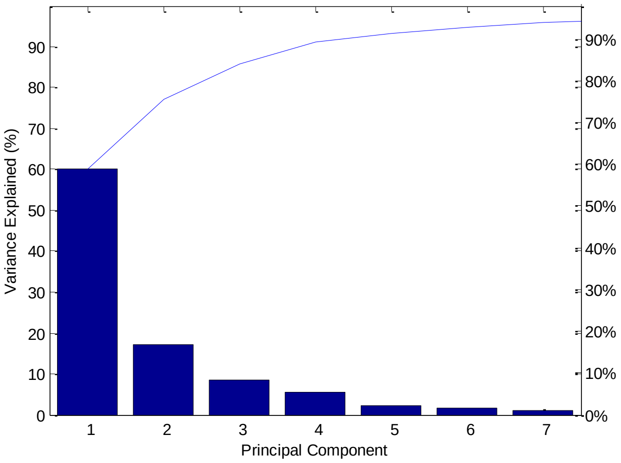


Figure 12 : Eigenvalues scree plot for the FT-IR spectra of passaged cultured fibroblasts subjected to PCA.

3.2.4 Study 1A - DD Nodule and Cord vs. Transverse palmar fascia

In the scores plots, different sites are presented as different colours. The patients are labeled by their numbers e.g. for each patients where three cell types were analysed (e.g. DD nodule, cord and fascia), there are 3 different coloured circles to represent them. For example, in the first plot in Figure 13, patient DD8 has 12 circles, because there were 12 samples from this patient. Among them, 6 are green (nodule), 3 blue (fascia) and 3 red (cord).

PCA on EMSC normalised data was performed to test the quality of spectral data using the NIPALS option in MATLAB to determine covariance between clusters spectra of fibroblast samples. There is a clear separation of control samples and disease tissue. In contrast with normal samples, DD fibroblasts (nodule and cords) also show an additional level of heterogeneity. There were two distinct clusters identified among the DD samples separating it into cord and nodule (circled in green and red) from the transverse palmar fascia (circled in blue) in passage 1. However, as the number of passages increase the samples become increasingly random and show overlapping between those derived from other tissue phenotypes. In Study 1, the separation is minimal between the clusters of individual phenotypes (nodule, cord, fascia) for samples where passage number = 2 and samples from cultures where passage number ≥ 3 do not demonstrate clarity in separation and are overlapping within the boundary of others.

The Fisher's ratio was calculated for each patient vs. passage number to determine separation/variability within class. The relation between Fisher's ratio and passage number is plotted in Figure 15. A linearly decreasing value for ratio as the passage numbers increase is observed. The trend suggests that the larger the ratio, the better the separation between classes (i.e. well separated classes and tight clusters for each class). The decrease in the ratios becoming smaller and smaller when the cells are subcultivated (as passage number increase) correlates with the findings from those of the PCA scores plots that show clear separation between the cultures of different DD phenotypes (nodule, cord and transverse palmar fascia) in the early passages (1 and 2) but as the passage number increases, asynchrony and overlapping in samples are observed. The ratio becoming smaller and smaller as the number of passages increase suggest a similarity (dysregulation) in metabolic profiles of the phenotypes consistent with the findings from the PCA indicating an overlap in the metabolic profiles in higher passage numbers and resembling little differences in their metabolic fingerprint.

3.2.5 Study 1C Patient vs. Passage (Category 4)

The samples in the PCA plots defined by PC1 and PC2 in this study were not clearly separable on the basis of passage number. Though some samples were far apart, the passages were not superimposable on cultures derived from the same phenotype. PCA scores plots from Patients 8-13 are shown below. Patient 9 demonstrating similar trends observed as in Study 1A for the three sites, here, the lower passage number 1, shows clusters from nodule, cord and fascia far apart, as passage number increases, the clusters between classes are now closer and within classes are no longer clustered as tightly. As there is biological variance between patients, this difference is not so greatly observed through PCA scores plot alone, and so PC-DFA is used in Study 2.

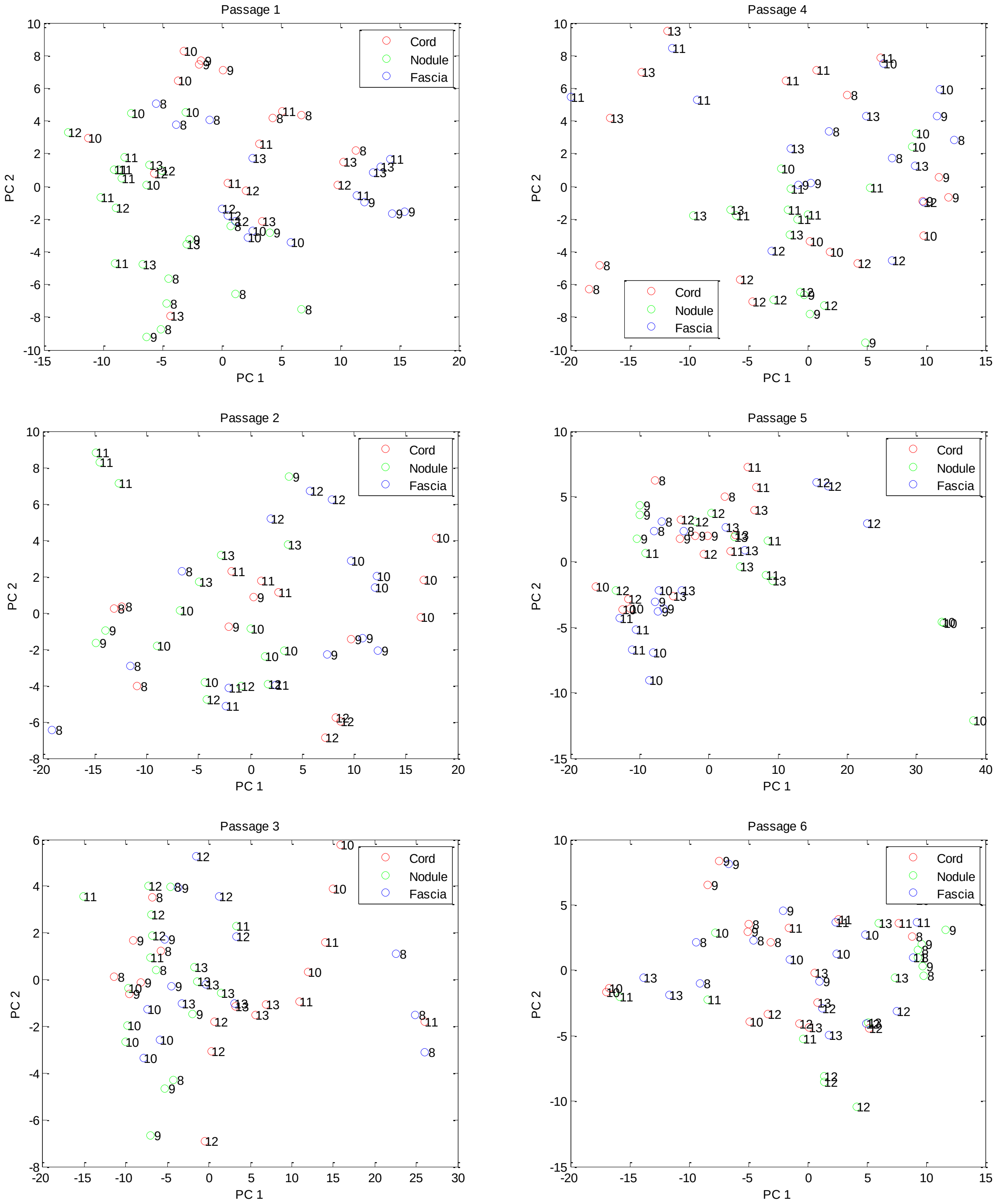


Figure 13 Projection of the FT-IR spectra of DD fibroblasts derived from the nodule, cord and fascia (control) with respect to their passage number onto the plane defined by PC1 and PC2. Patients are labeled by their numbers e.g. for each patients where the three cell types were analysed (e.g. DD nodule, cord and fascia), there are three different coloured circle for it. Red circle=cords, green circle=nodules, blue circle=fascia (internal control).

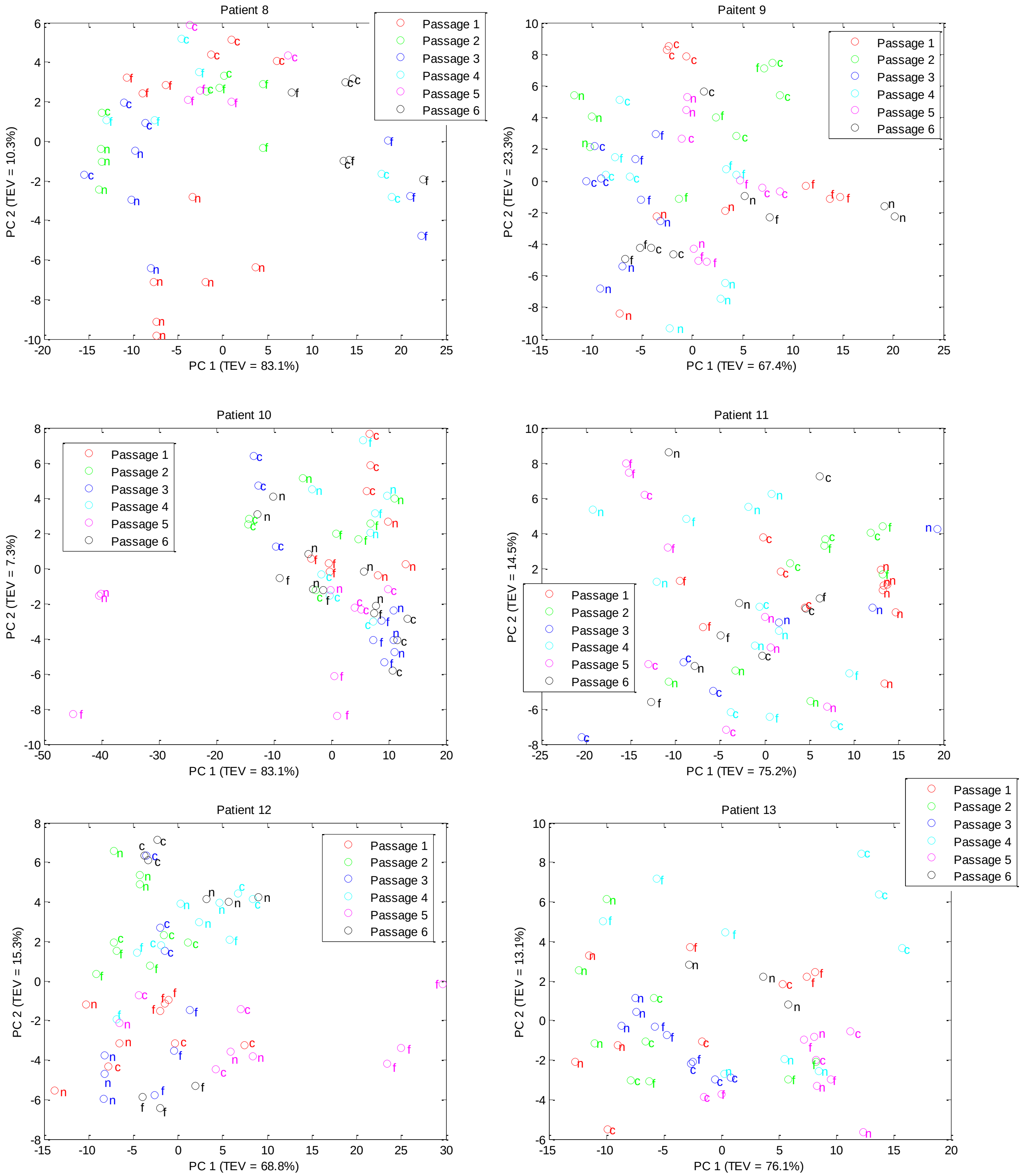


Figure 14 Projection of the FT-IR spectra of DD fibroblasts derived from the nodule, cord and fascia (control) with respect to their passage number onto the plane defined by PC1 and PC2. Samples from variable passage numbers are coloured and DD nodule, cord and fascia are labeled in lower case letters n,c,f respectively.

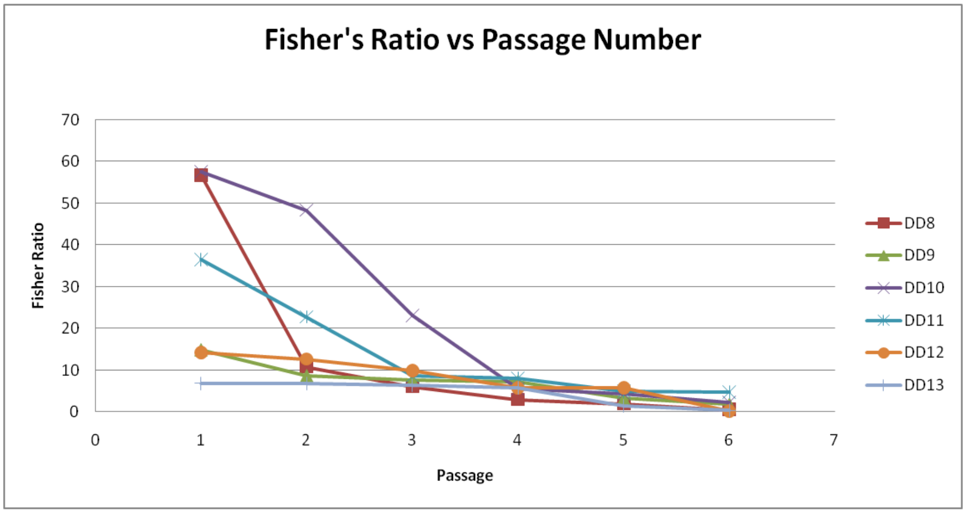


Figure 15 Fisher ratio plot *vs*. passage number

3.2.6 Study 1B - DD Nodule, Cord vs. Transverse Palmar Fascia, Fat, Skin overlying nodule

The cultures from the fat and SON were subjected to FT-IR irradiation at a different time course and not on two consecutive days. This was due to a technical failure that had occurred with the stage of the FT-IR on to which the Si plate is loaded. Therefore, the results of 1B were not analysed with respect to the differentiation between the five DD phenotypes, but only on the basis of the fat and SON passaged cultures. The PCA scores plot defined by PC1 and PC2 showed clear differences between the clusters of two different sites (fat and SON) across all passages. However due to a flaw in the sample arrangement on two plates, (samples not sufficiently randomised); in each passage, all samples from one site were located on one plate while all samples from another site were located on the other plate) one cannot determine whether such separation was caused by biological difference or merely the difference caused by two different plates analysed on two different days or both.

3.2.7 Study 2A - DD Nodule, Cord vs. Transverse Palmar Fascia (Category 1)

The results from Study 1 implied that performing serial passages >3 did not demonstrate clear separation of the metabolic profiles obtained from spectral data of the three sites (nodule, cord and fascia). In Study 2A a rather inconsistent sample size was analysed. This inconsistency in sample no. was due to several factors ranging from a smaller number of DD biopsies, cells that did not grow or survive the tissue processing step, samples arriving on separate occasions; with up to 2-4 weeks gap, due to flaws in cell culture management leading to contamination of some samples, shared incubators, and also cultures grown in two separate buildings.

However, by contrast, cultures from passage 0 were also included in this study. Although there was very little biomass (¾ of the initial monolayer grown from primary fibroblast was subjected to FT-IR spectroscopy), clear separations of the three sample types were observed. For this study cultures from passage 0, 1, 2 and 3 were analysed. As in Study 1A, PCA on EMSC normalised data was performed to test the quality of the spectra using the NIPAL option in MATLAB and results were corroborated with those in PyChem to confirm the covariance between clusters in spectra of fibroblasts samples. Again, there was clear separation of control samples and DD disease tissue. In contrast to normal samples (blue circles), DD fibroblasts (nodule and cords) also show an additional level of heterogeneity. There were 3 distinct clusters between the samples in passage 0, 1, 2 and again little randomisation started to occur in samples of cultures from passage 3. Study 2A confirmed that early passages demonstrated metabolic differences in DD and control fibroblasts, whereas higher passages demonstrated random asynchronous patterns. This study also confirms that early passage numbers are the most suitable representatives for investigating DD.

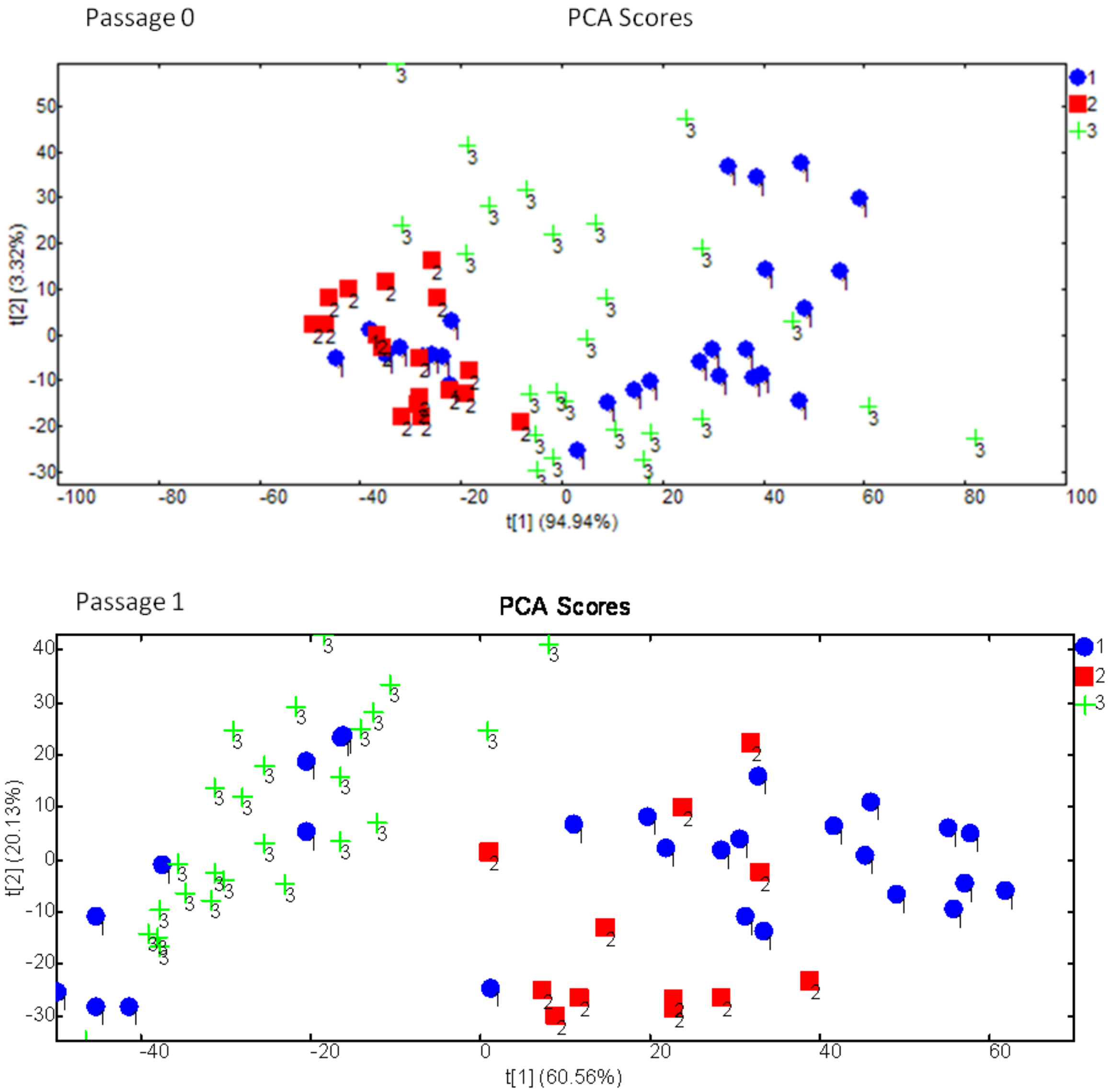


Figure 16 Projection of the FT-IR spectra of DD fibroblasts derived from the nodule, cord and fascia (internal control) with respect to the passage number onto the plane defined by t[1] (PC1) and t[2] (PC2). The samples are labeled by numbers and shapes; 1 blue circle = transverse palmar fascia, 2 red square = nodule, 3 green cross = cord.

.

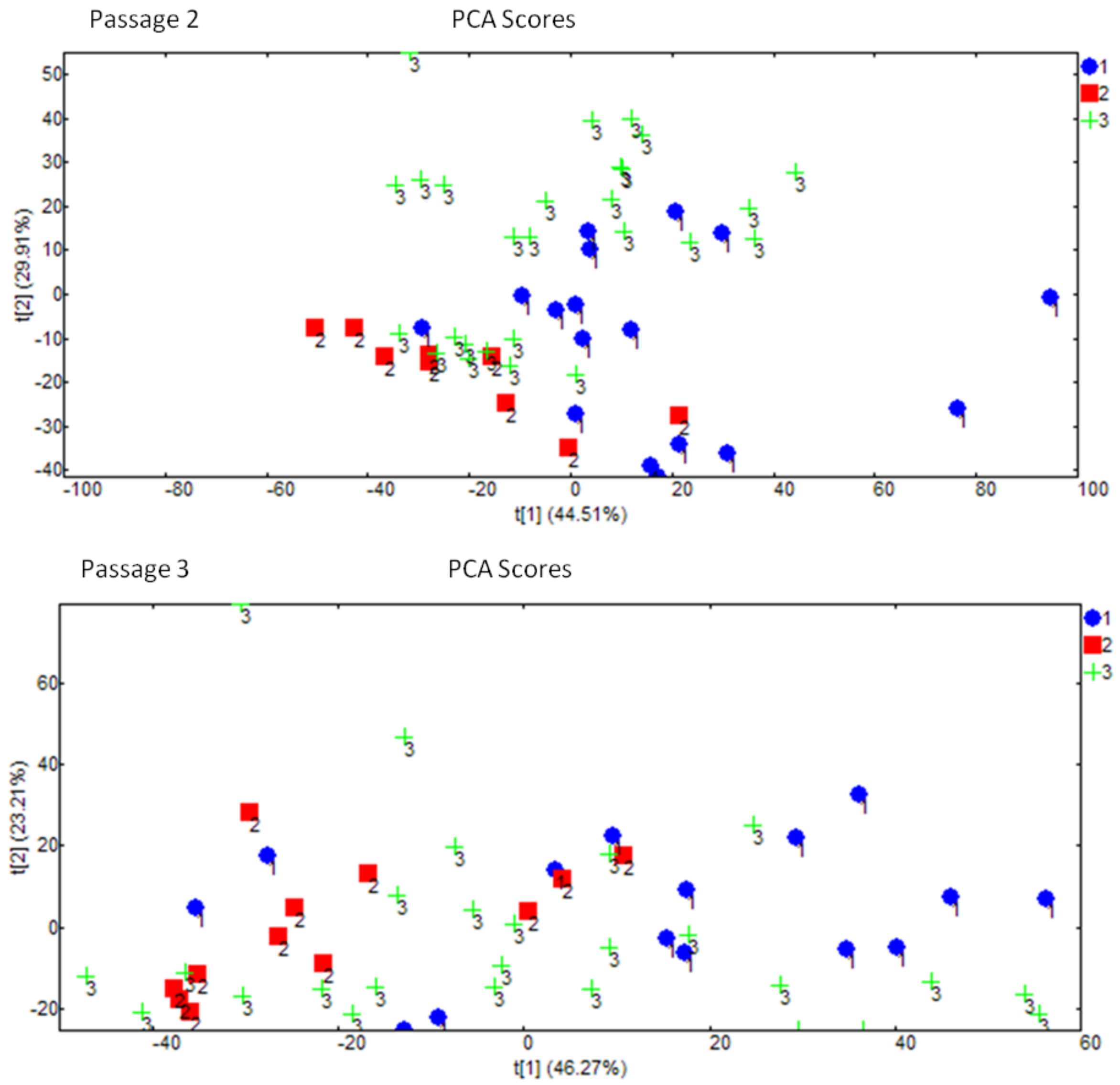


Figure 16 Projection of the FT-IR spectra of DD fibroblasts derived from the nodule, cord and fascia (internal control) with respect to the passage number onto the plane defined by t[1] (PC1) and t[2] (PC2). The samples are labeled by numbers and shapes; 1 blue circle = transverse palmar fascia, 2 red square = nodule, 3 green cross = cord.

3.2.8 Study 2B - DD Nodule, Cord vs., Transverse Palmar Fascia, Fat and Skin over Nodule

The PCA scores plot did not show clear separation between clusters from sample sets from multiple patients combined. Mostly samples from the same phenotype showed asynchronous patterns regardless of the passage number. However, nodule and SON derived fibroblasts were in closer proximity than those cultures derived from cords, fascias and fat. In addition, the samples from the latter three phenotypes demonstrated overlapping in clusters across all passages.

However, clear separation of DD fascial cells nodule, cord and fascia can be observed from fat and SON in Patient 2, passage 0 across PC2, while some fascia can also be separated from all other sample clusters, Separation of fat clusters and SON is observed in PC3 (Figure 17 and 18). Figures 19-21 represent the supervised scores plot from PC-DFA for Patient 2 passage 0-3.

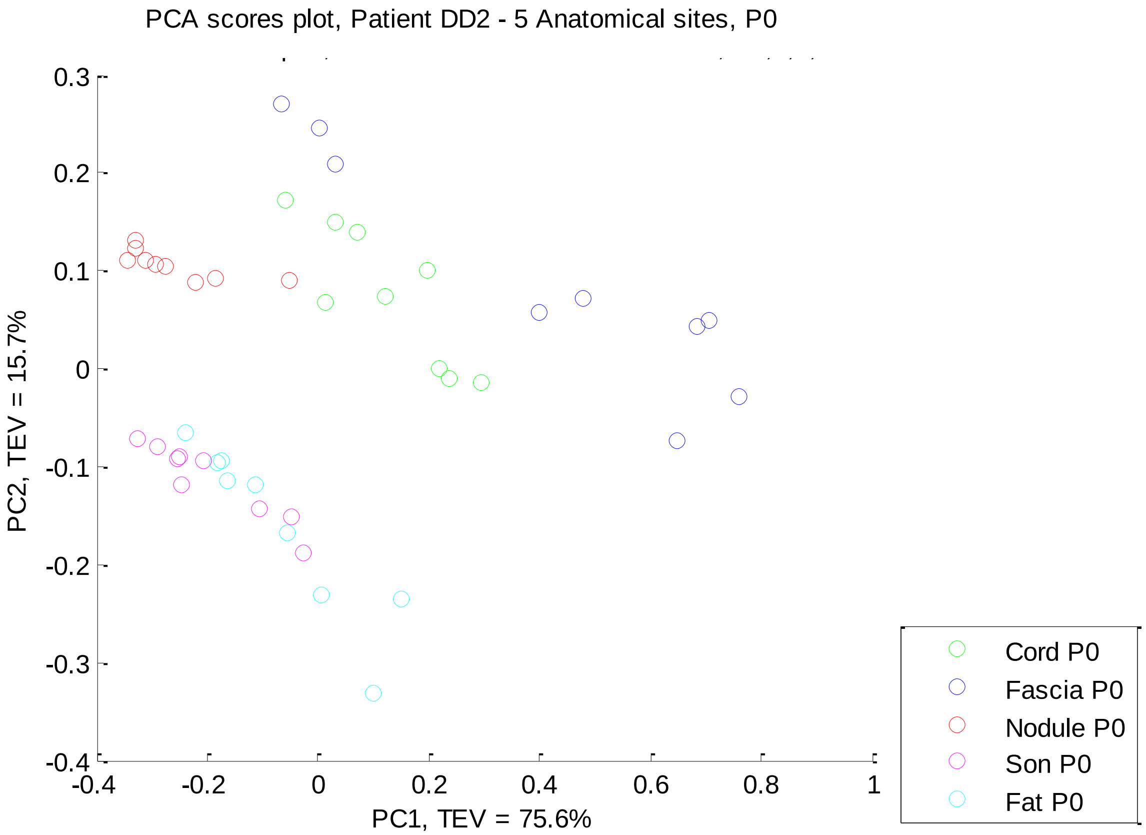


Figure 17 PCA scores plot of the five different sites from a single patient, distinct clusters are observed, the spectral profile of SON and fat are highly similar to each other with the cord and fascia showing a similar relationship, the nodules cluster within its own space and can be separated from the rest using PC1 and PC2.

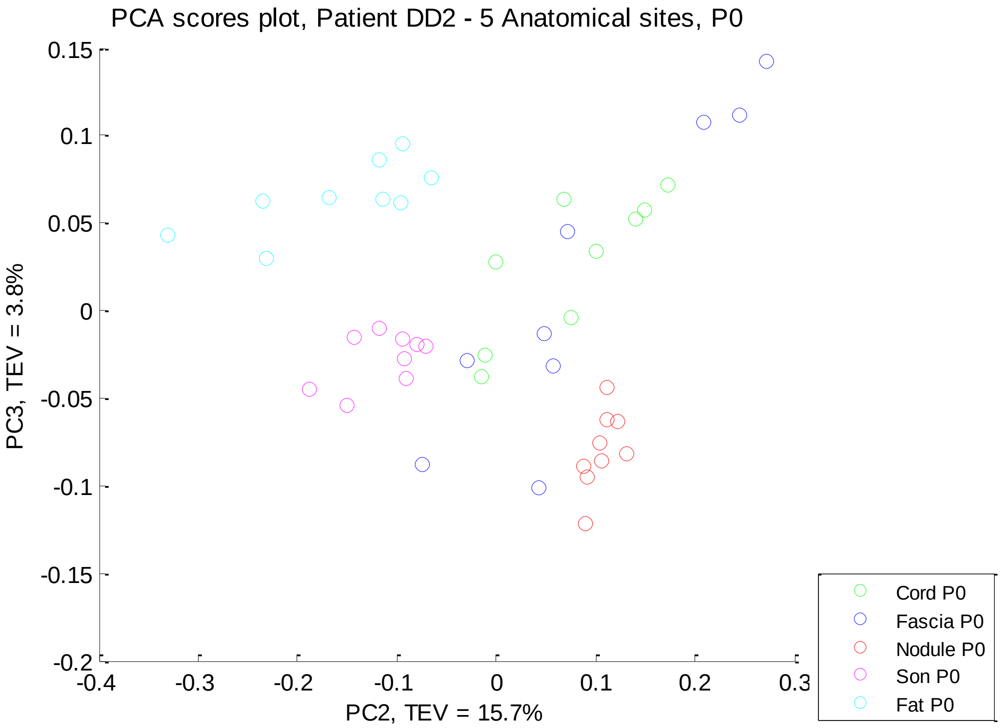


Figure 18 PCA score plot of PC2 *vs.* PC3 from patient DD2. The Fat and SON displays greater separation compare to PC1 *vs.* PC2. The overlap between Cord and Fascia is still observed.

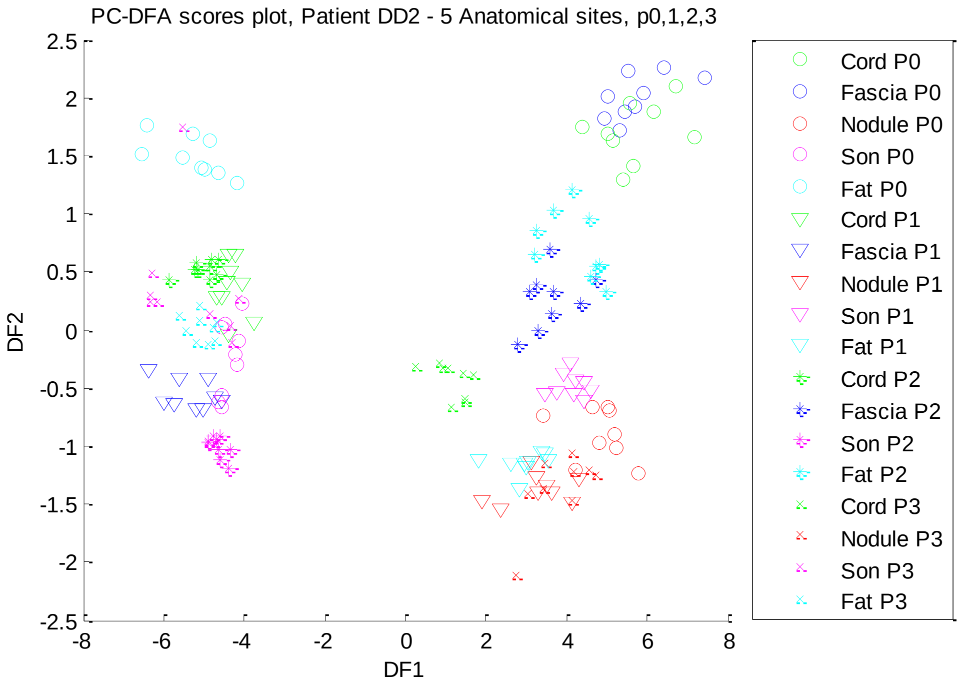


Figure 19 DFA score plot of all site and passage (0-3) from patient DD2. The profile for each site with respect to increasing passage number is observed, where discriminate from the same site is possible based upon passage number.

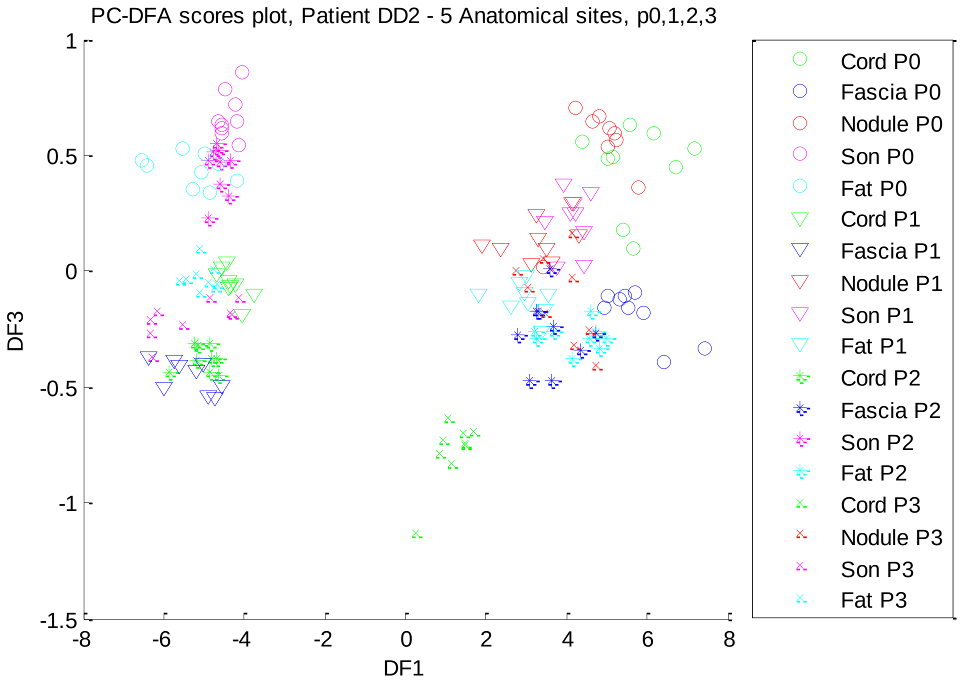


Figure 20 DFA score of patient DD2 for DF1 *vs.* DF3, the effect of different passage number within each respective site is still observed.

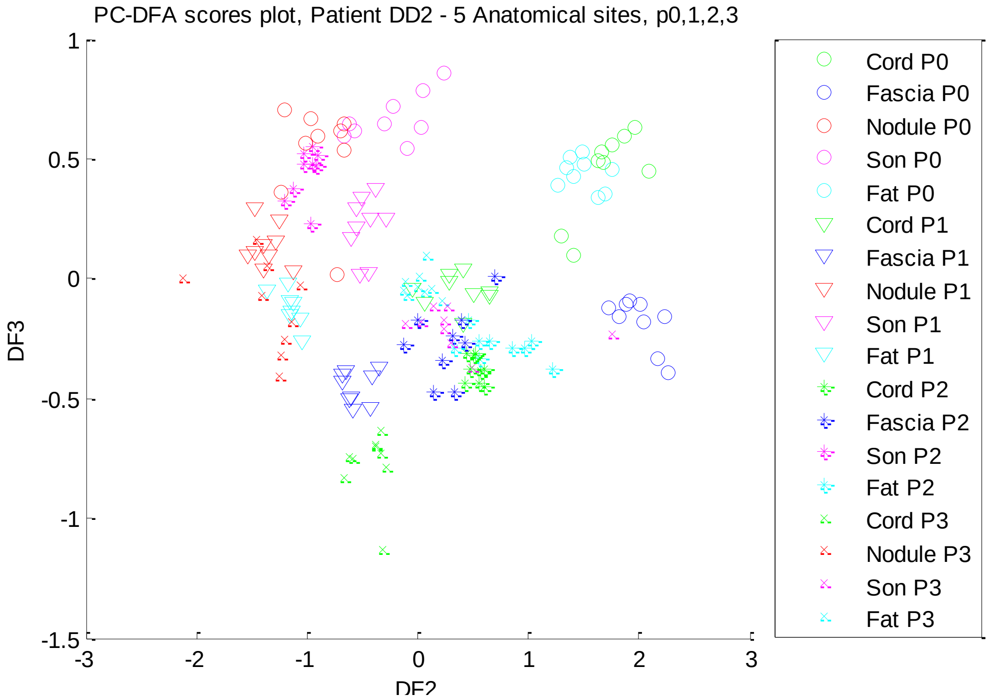


Figure 21 DFA score plot of DF2 *vs.* DF3 for patient DD2. The trends are less clear compared to previous two plots.

3.2.9 Study 2C - DD Nodule, Cord *vs.* Transverse Palmar Fascia and Unaffected (CTD) normal palmar fascia

Figure 22 shows the DF1 *vs.* DF2 score plot from the DD subsets compared to internal and external control fascia. Its analysis along the DF1 axis shows that this component divides the whole set of spectra into four groups; a cluster of red circles (nodules), majority of blue circles, (DD fascia-internal control) on the right half of the plot, majority of turquoise/cyan circles (CTD fascia-external control) on the left half of the plot, and green circles (cords) closer to the centre (zero on *x*-axis).

There was clear separation between the DD fascia, nodule, cord and also external CTD control. Clusters of samples from DD cells always clustered apart (while showing separation between the individual samples) from those of CTD. This separation suggests that internal fascia is an appropriate control and can be distinguished from diseased fibroblasts using chemometrics techniques. The use of internal fascia as the control will attribute to homogeneity in future studies.

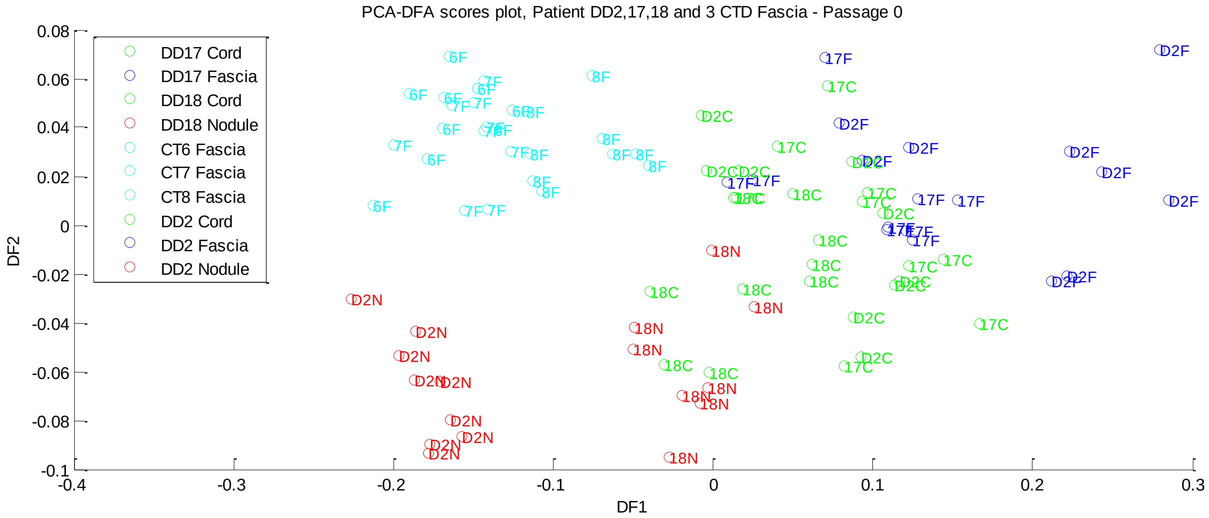


Figure 22 Projection of the FT-IR spectra of DD fibroblasts derived from the nodule, cord, fascia (internal control) and CTD fascia (external control) from passage 0 onto the plane defined by DF1 and DF2. Patients are labeled by their numbers e.g. for each patient where the four cell types were analysed (e.g. DD nodule, cord, fascia and CT external fascia), there are four different colours to represent them; blue circle = fascia (internal control); red circle = nodules; green circle = cords; and turquoise circles = CTD fascia (external control).

3.2.10 Study 2D- CTD Fascia, Fat, Skin

Clear separation was observed in the fibroblasts derived from CTD fat, fascia and skin.

3.2.11 PCA of Footprint Spectra

No separation was observed using PCA on footprint data as the culture medium they were collected in was nutrient rich and undefined. An example of the processed FT-IR spectra from DD footprint and the PCA plot from this data are shown in Appendix C; Figures 72 and 73.

3.3 Discussion

3.3.1 Principal Findings

The goal of the present work was the analysis of Dupuytren fibroblast cultures derived from different DD tissue phenotypes & control (CTD) fibroblast cultures comparing early (primary) cultures to late passages in order to identify the most representative passage for the disease.

Using FT-IR spectroscopy, the metabolic profiles of endogenous and secreted metabolites were acquired from (1) DD cords and nodules from the palm against the unaffected transverse palmar fascia (internal control), (2) DD cords and nodules with the cushioning fat surrounding the nodule, and with the skin overlying the nodule and (3) those in (1) and were compared between different DD patients and with external controls.

Observational changes in cell growth were recorded to determine whether cultures derived from DD tissue phenotypes and control changed with cell passage. The haemocytometer cell counts were constantly high for fibroblasts derived from DD nodule, cord, fat and SON across passage 1 to passage 6, while the growth of fascial cells considered as normal (the internal and external fascia controls) slowed down at higher passage numbers. The difference in the cell numbers through variable passages suggest that cell senescence may have been achieved by some of the control fibroblasts whereas the proliferative potential displayed by fibroblasts derived from the diseased nodule, cord, fat and SON consistently provided a high cell count. The evaluation of growth in cultivated fibroblasts from different samples has not only shown significant differences in their morphological appearance but also differences in generating the number of myofibroblasts. These numbers can be correlated with tissue type as well as their unique metabolic profiles which may be of relevance to disease stage in some cases.

The FT-IR spectra showed broad and complex contours, in which there was relatively little visible qualitative difference between the spectra. In lesser complicated molecules such as chemical or medicinal compounds such as drugs whose MW $\leq$500, the locations of transition points are based on an arbitrary selection of the frequencies that are more pronounced for the compound being analysed as the functional groups present are few. Such an approach assumes that information about the structural changes leading to the phase transition is known a priori. However, for large molecules such as the proteins, biomolecular compounds, cells (and samples in this study), the conformational changes cannot be easily anticipated because the bands suitable for monitoring complex molecular processes cannot be chosen without any *a priori* assumption about the structure and its infrared intensity relationship. This is due to the fact that essential information on any dynamic process being analysed by FT-IR spectroscopy does not lie at any individual wavenumber, but across many wavenumbers, mostly correlated to each other. These contain many overlapping bands and so data interpretation cannot be made by simple visual inspection and therefore, to uncover the structure–infrared intensities relationship, unrestricted by any presumptions, suitable multivariate methods such as PCA and hierarchical cluster analysis (HCA) are used to look at differences and similarities between these spectra.

Visual comparisons of the general contours of the spectra were made with bands assignments according to peak wavenumber regions and the shape of the normalised spectra were compared with metabolic fingerprints of sera, urine and synovial fluids. While the spectral contours were similar in shape to all three, they shared spectral contours closer to those in synovial fluid. As the spectral data illustrated the need to use multivariate statistical techniques in the analysis, PCA and PC-DFA were applied to discriminate between the data sets.

Metabolic differences were identified between DD fibroblasts and in control (internal and external) fibroblasts through variable passages from the different sites. Analyses of these profiles lead to identification of metabolic differences where a synchronous separation pattern in early passages was observed in Study 1A & 2A). PCA demonstrated clear separation of the three sites from early cell passage cultures (0, 1 & 2). Samples from Study 2C also displayed this trend and showed separation of the respective sites from early cell passage cultures (0, 1 & 2). The cord and control fibroblast clusters were in closer proximity in the scores plot of PC1 and PC2 across passages 0-3, while the nodules fibroblast were clearly separable in all PCA plots and were confirmed by Euclidean distances calculated in hierarchical cluster analysis using PyChem.

However, in higher passage numbers (4-6) a random asynchronous pattern in the scores plot was observed across all samples. The analyses from both studies demonstrate that cultures of different tissue phenotypes from passage 3 show little separation across the cell types, (while not being too far apart), and from thereon, subsequent passage numbers demonstrated a random asynchronous pattern in metabolic profiles that appear the PCA scores plots.

The analyses of the PCA scores plots between the five sample types from DD subsets in Study 2B demonstrated overlapping between the clusters in samples across passages 0-3. In the early passages however, the clusters of nodules and SON were in close proximity to each other, while the clusters representing the fibroblasts of cords, fascias and fat were often closer than those of the nodule. While PCA showed some differences, an alternative MVA method could to be applied to discriminate between these individual cell types.

3.3.2 Strengths and weaknesses of the study

The FT-IR data together with PCA provide good classification of spectra of DD subsets in Study 1 and 2 and also distinguishes spectra from external controls in study 2C with PC-DFA. Classification of FT-IR data may be explored further using first and second derivatives of spectra or by exploring alternative preprocessing methods. However, the classification achieved so far is encouraging, as this technique not only separates DD nodule and cord fibroblasts from the two controls, but also between the two fibrotic elements i.e. nodules from cords. In addition the analyses across all passages suggest that early passage cultures are close representatives of the metabolic fingerprints of those in disease state *in vivo* and may be more appropriate for further DD studies. In addition data from the late passages support the hypothesis that the cell culture monolayer environment may alter the functional characteristics of the samples, possibly by selecting against a subpopulation of cells which survived the *in vitro* conditions. Differences in FT-IR spectral profiles in DD and in controls can be exploited further to test hypotheses with the use of formalin-fixed samples for discrimination of normal and diseased tissue. To discriminate between footprint spectral data, supervised chemometrics analysis such as artificial neural networks and genetic algorithms may be applied.

In order to validate PCA and PC-DFA analyses independently, the data can be projected into the PCA or PC-DFA ordinate space. This involves each class within the data set being randomly split, and forming a test sample set and a training set. Generate PCA and/or PC-DFA models independently for the training set, and the test set then to be subsequently projected into the same ordinate space. A close alignment between training and test clusters would indicate the validity of the PCA and PC-DFA model. This is particularly necessary when the number of variables is considerably larger than the number of samples and when one wants to ensure that the hypotheses generated are robust [140].

We have shown that FT-IR spectroscopy is a powerful fingerprinting tool when combined with multivariate analyses. It has enabled the detection and acquisition of several endogenous metabolic fingerprints from cultures derived from different tissue phenotypes. However, due to the complicated nutrient rich medium, this technique was not appropriate for detecting patterns of secreted (footprint) metabolites. Nevertheless, it has proven to rapidly and accurately identify and distinguished on the basis of separation of the unique fingerprints of intact fibroblast cells from diseased and non-diseased regions. Preliminary results obtained from Study 1 and validation with those of Study 2 has shown FT-IR to be powerful as a diagnostic tool.

While it has been recognised that FT-IR is not as specific and sensitive as some techniques such as GC-TOF-MS, the high throughput, rapidity and reproducibility of FT-IR is demonstrable through the large body of research published using this technology. FT-IR is recognised as a valuable tool for metabolic fingerprinting as is able to analyse cellular constituents composed of carbohydrates, fatty acids, amino acids, protein, nucleic acids, and polysaccharides rapidly and simultaneously with a minimum amount of sample preparation.

One of the potential limitations of FT-IR is that the absorption of water is very intense. An attempt to overcome this problem was the dehydration of samples in 50^{O}C incubator prior to loading onto the stage of the spectrometer. In future analysis, the water signal could be subtracted. Alternatively, a related vibrational technique, Raman spectroscopy may be used as a complementary tool to confirm the results of FT-IR , yet this would be more expensive, low through put, and require much longer analysis time.

This study has enabled characterisation and classification of the various sites thought to be involved in DD. In addition, chemometrics has strongly facilitated pattern recognition of disease state when compared with normal palmar fascia from several cultures (early and late) allowing us make an informed choice of a suitable passage to perform subsequent studies in this project. Thus, the present study is not only taken up as a more complicated case to discriminate from macroscopical and genetic studies, but also to verify the feasibility of spectroscopic methods in predicting the DD tumour response to *in vitro* cultivation.

3.3.3 Trypsin Effect

The differences between metabolic profiles achieved from those of early and late passages determined by FT-IR and multivariate analyses are clear from the statistical analyses. The relation between the Fisher ratio and passage number showing a linearly decreasing value for ratio as the passage number increases is also suggestive of a decrease in the ratios due to cell subcultivation with the use of trypsin (passage increase). This correlates with findings from those of the PCA plots that show separation between DD disease phenotypes in the passage 1 only (Study 1A) and little separation of samples in subsequent passage numbers. However, why this change occurs at such an early onset is not understood. One possible reason for this could be the treatment of trypsin to collect these anchorage-dependent cells that were cultivated on a solid culture substrate (polystyrene flasks). Trypsinisation of cells may have had detrimental effects such as the proteolysis of the cell membrane proteins. Furthermore subcultivation and collection of the cells performed in this way could have had detrimental effects on the cell number count as well as morphological and conformational changes of the cell. This type of methodology and its suitability to retrieve cells could be discussed with a cytogeneticist who would have a better perspective. The analysis to locate this change in cell cultures (or even cell line) is required early on in establishing a cell line. Little is published about the effects of trypsinisation and this method is considered as standard protocol to collect anchorage-dependent cells. It is also possible that no study has tried to see what changes are/could be taking place at this early stage and may put this down to cells undergoing an immortalisation transformation process. A few studies have raised concerns about trypsinisation [141, 142]. The synchrony seen between a decrease in the Fishers ratio and passage number could be due to a trypsin treatment effect as all passages were produced at a convenient point in time in connection to the cells approaching confluence, where many cells had already experienced confluence. With the view that the trypsin had no effect, the 'passage effect' on metabolite fingerprint is simply because the metabolites were grouped into the same culture media representing each passage. Or by contrast, if the process of the passage had no effect at all, it would be possible to measure (metabolite) flux as a function of passage as the flux would be a continuum. To solve this problem mechanically scraping cells (which although may break the intact cell) may avoid the problem of conformational changes during sub-cultivation using trypsin (this method is used for harvesting cells from cultures for subsequent studies). Additionally, the average concentration of metabolites over each passage could be measured. Close examination of the plot at each passage suggests there as though there is an inflection (in the plot) at each passage when in fact the process is one of gradual change which should be independent of trypsin. Furthermore, this could be also be investigated by frequent changes of media to increase the resolution of metabolite secretion (on the metabolic footprint).

3.3.4 Conclusion

The application of FT-IR spectroscopy conducted under carefully controlled conditions with appropriate chemometric techniques to differentiate between DD and control fibroblasts has been demonstrated to be a powerful tool for discriminating between these cell types in individual DD patients, as well as samples from controls. No study to date has attempted to demonstrate metabolic analytical differences of dupuytren fibroblast cultures derived from different DD tissue phenotypes in early (primary) cultures compared to late passages. This study implies that early passage numbers are suitable representatives for DD studies. In addition based on their separation, these data demonstrate a gradation in profiles of certain metabolites (obtained from their unique fingerprints) in fibroblasts of DD phenotypes compared with control. Metabolic fingerprinting is predictive not only of disease but also of disease phenotype. The technique has major advantages of speed, sensitivity, and the ability to analyse many hundreds of samples simultaneously.

Chapter 4

Metabolic profiling of Dupuytren's disease fibroblasts under oxidative stress

4.1 Introduction

4.1.1 Metabolism and Warburg effect in DD

At present, little is known regarding DD pathogenesis, and even less regarding its cellular functions, i.e. metabolic functions and regulation of intermediates involved in glycolysis, TCA cycle, pentose phosphate shunt and amino acid metabolism. Studies of genetic abnormalities and environmental factors have provided evidence for a multifactorial nature of DD. The disease progresses with the growth of abnormal fibroblasts and localised ischemia/hypoxia. It is speculated that microvessel narrowing preceding localised hypoxia may be one possible cause of DD, where fibroblast proliferation ensues during perivascular connective tissue damage.

In this Chapter we test our hypothesis whether DD cells are under the Warburg effect. This was achieved by inducing a perturbation in healthy cells (fascial cells) cultured in pO2 = 158 mmHg corresponding to a concentration of 21% atmospheric oxygen and compare their metabolic profiles with healthy cells exposed to hypoxia i.e. in pO2 = 8 mmHg corresponding to concentration of 1% oxygen. Then we examine the hypothesis that any such differences are akin the Warburg effects noted for tumour cells in the literature by comparing the extracts from intracellular (endo-) metabolomes acquired from DD nodule, cord and SON cultures against those acquired from healthy cells and under hypoxia-induced fascia. The extracellular (exo-) metabolomes acquired from DD nodule and healthy fascia in normal and hypoxia-induced cultures. In addition, response to hypoxia is also examined in the intracellular metabolomes of disease cells; this may aid in biomarker identification by imposing stress on the disease cells. Furthermore the question is; if a Warburg effect exists, in which DD phenotype is this effect greatest? What are these key players (metabolites) and which pathways are these mapped onto? In addition, we test another hypothesis; whether changes occurring in abnormal fibroblasts are due to gene expression alone or dictated by metabolism or a combination of both? (see Chapter 5). The objectives were to investigate 1) perturbation effect on healthy fascia cells to investigate whether healthy cells in hypoxia mimic disease state scenario 2) progression of disease compared to healthy (nodule, cord and SON *vs.* fascia) 3) the effect of 1% oxygen tension on these samples (1% *vs.* 21%). The aim is to get a deeper understanding of the dynamics involved in the DD cellular system. Sample reproducibly of DD and control cells has been shown in Chapter 3. This study employed the use of GC-MS to determine metabolic profiles.

Metabolic differences between fibroblast cell samples (passage number 3) cultured in normoxic and hypoxic conditions have identified a number of significantly dysregulated metabolites involved in both amino acid metabolism and carbohydrate metabolism pathways in nodule, cord and SON pairwise analyses. The comparisons between disease and control fascia reveal that cysteine, aspartic acid and a sugar molecule were significantly down-regulated in disease. The perturbation effect in control fascia resulted in the identification of a number of significant metabolites involved in carbohydrate metabolism and amino acid metabolism. While relatively fewer metabolites have been identified in disease when compared with control, the perturbations effects have revealed a relatively larger number of significant metabolites. It has been demonstrated that GC-MS is a highly sensitive and high throughput analytical technique for biomarker screening and identification in DD samples, able to discriminate between different oxygen tension and tissues types (cultures), also enabling detailed profiling of the induced perturbation of metabolome. This is the first time that GC-MS and metabolomics analysis methodology been applied to the characterisation of DD samples.

4.2 Results

4.2.1 MVA on GC-MS data acquired from all samples in hypoxia and normoxia

The intracellular distribution of metabolites and metabolic changes induced by these abiotic perturbations were investigated by metabolomics. GC-TOF-MS was employed, providing in this study the detection of 129 different metabolite peaks from the intracellular metabolome (in each of the 126 samples) and 79 from the exometabolome (36 samples). A typical total ion current (TIC) chromatogram from mammalian endometabolome is shown in Figure 23.

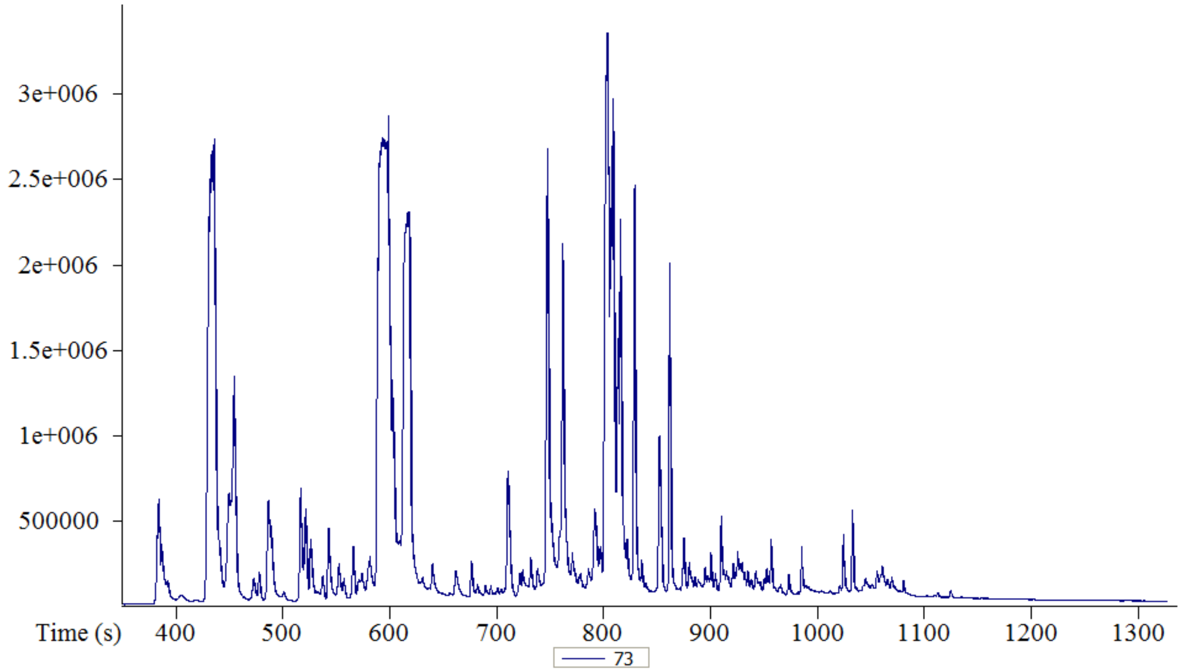


Figure 23 Typical TIC for mammalian cells.

To assess sample variability, GC-TOF-MS analysis was performed on 126 samples, constructed from 4 different tissue phenotypes (nodule, cord, skin overlying nodule and transverse palmar fascia – i.e. control) cultured in triplicates for each O_2 tension. In this study there are three factors of interest; 1) progression of disease compared to healthy (nodule, cord and SON *vs*. fascia) 2) the effect of 1% oxygen tension on these samples (1% vs. 21%), 3) perturbation effect on healthy fascia cells to investigate whether healthy cells in hypoxia mimic disease state scenario. To identify the source of greatest variation within the combined and individual groups of data for all samples MVA was employed. The initial stage of the data analysis strategy was to use unsupervised exploratory data analysis; PCA was employed to discover any natural groups within the data and also used for discovering any outliers before pre-processing. Other methods for outlier detection such as Winsorisation may not have provided suitable results as this was an unbalanced data set with respect to sample type (phenotypic differences e.g. nodule, *n* = 7, fascia *n* = 3). A number of known and unknown metabolites were identified. 69 of the 129 metabolites were known metabolites or indicatives of sugar. Table 5 shows the identified metabolites and the super families of pathways in which they contribute. The results of the PCA for all 8 samples types cultured in 21% O_2 and hypoxia 1% O_2 showed separation in clusters from fascia cells to those from all skin samples in PC1. There was no obvious separation between these two samples types when compared with cords and nodule samples. There was overlap in these clusters in PC1 and PC2. The results of the PCA are shown in Figure 24(a). Most N21 samples also clustered together and separated from N1 in PC1. No trends were clearly observed in cords. The 4 samples in Figure 24(b) cultured in 21% O_2, plus F1 shows S21 samples separate in PC1 from fascia and cord (except patient 5) in PC1. This confirms that SON derived fibroblasts were a different sample type (from epidermis) compared with fibroblasts derived from nodule, cord and fascia (fascial cells beneath dermis and fat).

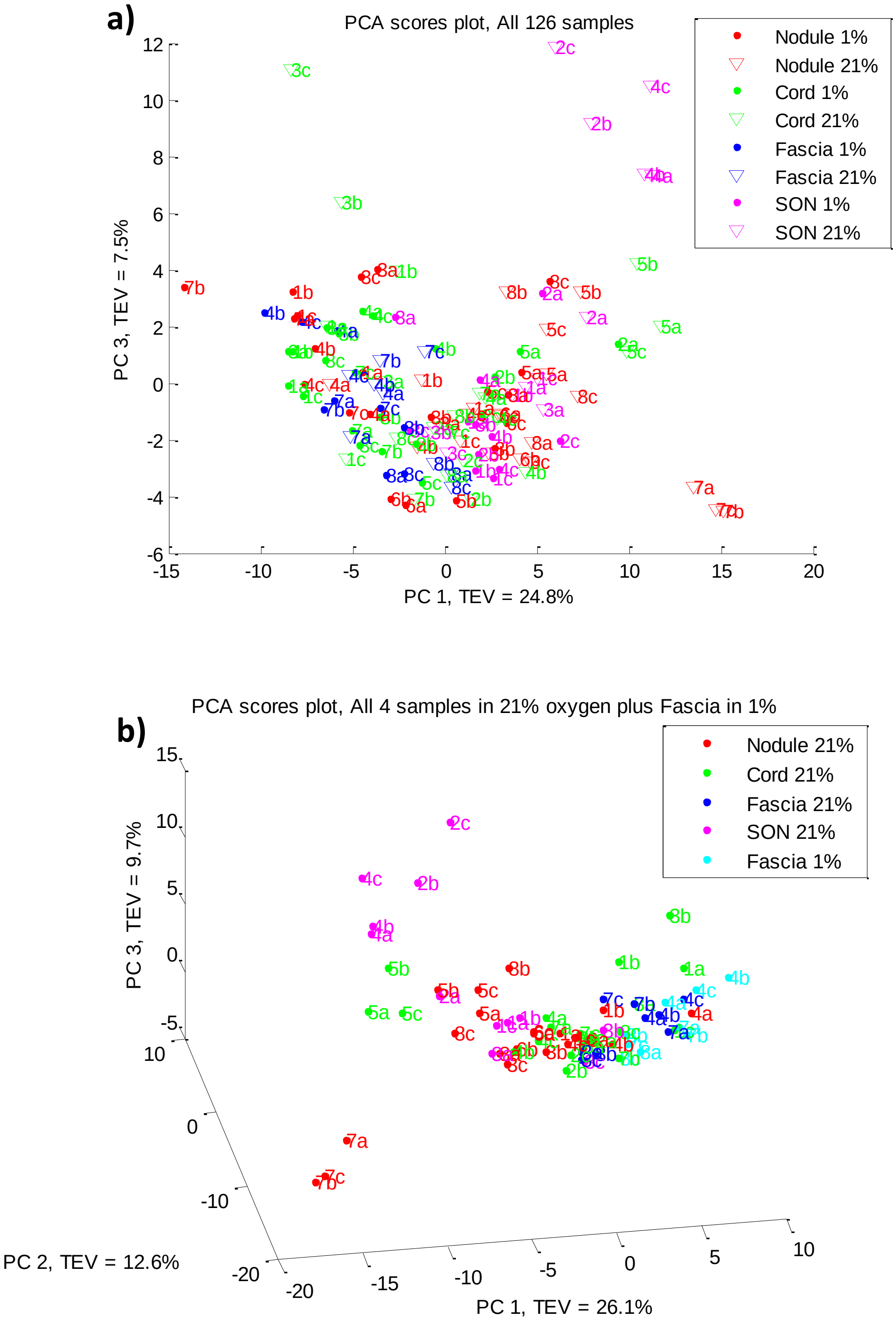


Figure 24 a) PCA score plot from GC-MS data acquired from 126 samples; nodule, cord, fascia and SON fibroblasts cultured in 1% and 21% oxygen. Red dot = N1, red triangle = N21, green dot = C1, green triangle = C21, blue dot = F1, blue triangle = F21, magenta dot = S1, magenta triangle = S21. **b)** PCA scores plot from nodule, cord, fascia and SON fibroblasts cultured in 21% oxygen and fascia cultured in 1% oxygen. In each plot the numbers represent patients (each sample with three biological replicates. The numbers represent samples (each patient with three biological replicates. Red round dot = N21, green dot green dot = C21, blue dot = F21, magenta dot = S21, cyan dot = F1.

Table 5 List of 69 known detected metabolite features (peaks) and their involvement in corresponding pathways. 'Yes' to Definitive ID = the metabolite in the sample has matched (by retention index and mass spectrum) to an authentic chemical standard present in MMD's EI-MS mass spectral library. 'No' = the mass spectrum only matches to a metabolite in other mass spectral libraries (i.e. not the Manchester library). Definitive ID = Due to the nature of the analysis, not all metabolites detected can be accurately identified and some have similar chemical structure and were assigned by best possible match score.

Metabolite No	HMDB Accession ID	Metabolite Identification	Definitive	Pathway
1		trimethylamine-N-oxide	no	Amino Acid Metabolism
2	HMDB00883	valine	yes	
3	HMDB00192	cystine	yes	
4	HMDB00929	tryptophan	yes	
5	HMDB00687	leucine	yes	
6	HMDB00050	adenosine	yes	
7	HMDB00172	isoleucine	yes	
8	HMDB00687	leucine	yes	
9	HMDB00172	isoleucine	yes	
10	HMDB00123	glycine	yes	
11	HMDB00162	proline	yes	
12	HMDB00187	serine	yes	
13	HMDB00161	alanine	yes	
14	HMDB00167	threonine	yes	
15	HMDB00191	aspartic acid	yes	
16	HMDB00191	aspartic acid	yes	
17	HMDB00574	cysteine	yes	
18	HMDB00696	methionine	yes	
19	HMDB00574	cysteine	yes	
20	HMDB00696	methionine	yes	
21	HMDB00159	phenylalanine	yes	
22	HMDB00159	phenylalanine	yes	
23	HMDB00159	phenylalanine	yes	
24	HMDB00123	glycine	yes	
25	HMDB00182	lysine	yes	
26	HMDB00182	lysine	yes	
27	HMDB00182	lysine	yes	
28	HMDB11733	glycylglycine	no	
29	HMDB00158	tyrosine	yes	
30	HMDB00158	tyrosine	yes	
31	HMDB00883	valine	yes	
32	HMDB00167	threonine	yes	
33	HMDB00191	aspartic acid	yes	
34	HMDB00143	galactose	yes	Carbohydrate Metabolism
35		sugar	no	
36		inositol	no	
37		sugar	no	
38		sugar	no	
39		sugar	no	
40		sugar	no	
41		sugar	no	
42		sugar	no	
49		sugar	no	
43	HMDB00827	octadecanoic acid	yes	Fatty Acid Metabolism
44	HMDB10368	cholesterol	yes	
45	HMDB00482	octanoic acid	yes	
46	HMDB00220	hexadecanoic acid	yes	
47	HMDB00131	glycerol	yes	Glycerolipid Metabolism
48	HMDB00131	glycerol	yes	
50	HMDB00126	glycerol-3-phosphate	yes	Glycolysis pathway
51	HMDB01401	glucose-6-phosphate	yes	
52	HMDB01401	glucose-6-phosphate	yes	
53	HMDB02730	nicotinamide	yes	Metabolism of Cofactors and Vitamins
54	HMDB00210	Pantothenic acid	yes	
55		3-ureidopropionic acid and/or beta-alanine	no	Metabolism of Other Amino Acids
56	HMDB00641	glutamine	yes	
57	HMDB00641	glutamine	yes	
58	HMDB00641	glutamine	yes	
59	HMDB00300	uracil	yes	
60	HMDB00300	uracil	yes	
61		sucrose	no	Starch and sucrose metabolism
62		sucrose	no	
63	HMDB00254	succinic acid	yes	Tricarboxylic acid cycle
64	HMDB00094	citric acid	yes	
65	HMDB00243	pyruvic acid	yes	
66		thiourea	no	Urea cycle
67		thiourea	no	
68		thiourea	no	
69		à-D-Galactopyranosiduronic acid	no	Uronic acid pathway

4.2.2 Effect of O_2 tension in intracellular metabolomes of healthy cells (F1 *vs.* F21)

The results from PCA for healthy fascial cells and those under hypoxic insult (F21 and F1) are shown in Figure 25(a). The PCA scores plots show within patient's signature was stronger than between patients. To analyse the variance a semi-supervised method, ANOVA-PCA was used which removes uninteresting sources of variance and the results are more interpretable. Most of the variance (TEV *vs.* # PCs) was captured in the first three components. The relationship between the data is very well elucidated in Figures 25(c) that display PC1 *vs*. PC2 scores plot. Analysis along the PC1 axis shows that this component divides the data into two groups (albeit broad clusters) i.e. F1 clusters and F21 clusters. To identify which chemical compounds might be responsible for such separation the corresponding loading plot (Figure 25(b)) was assessed. Because the separation between healthy and hypoxic fascia appeared in PC1, the variables that show a large diversity in PC1 are more likely to be the chemicals that differentiate these classes. The numbers represent the unique metabolite ID given to the 129 variables, each number corresponds to one peak observed in chromatograms.

From the loadings plot, the variable peaks on the extremes were mainly responsible for the separation exhibited in the scores plot whereas those close to the origin had little or no contribution to such separation. Examination of the loading plot revealed peaks with the highest variability. This was used to compare relative metabolite concentrations in samples and whether peaks (metabolite ID's) from loading plots correspond with those from univariate analysis. Many of these factors were induced by the hypoxic insult. The results were mapped onto pathways such as glycolysis, citric acid cycle, amino acid metabolism to see which metabolites were up/down regulated as a consequence of hypoxic insult.

A number of known and unknown metabolites were identified. Variables identified from both loadings and univariate analysis (Wilcoxon - sign rank test [143]) in Table 6 and 7 were considered as significant and box-whisker plots were drawn to display relative concentration distributions of metabolites with respect to sample and perturbation effect. Using an appropriate level of statistical significance ($p \leq 0.05$) major differences in 23 metabolites were identified (also from PCA) as significantly dysregulated in the endometabolomes. Of these 8 were known metabolites; pantothenic acid, a sugar and cystine elevated significantly while citric acid, cysteine (identified twice metabolite 40 and 44), aspartic acid, and 3-ureidopropionic acid and/or beta-alanine levels decreased. A further 15 metabolite features (peaks) were significant but remain unidentified. The metabolites ID (e.g. MET 100) are given in Table 7.

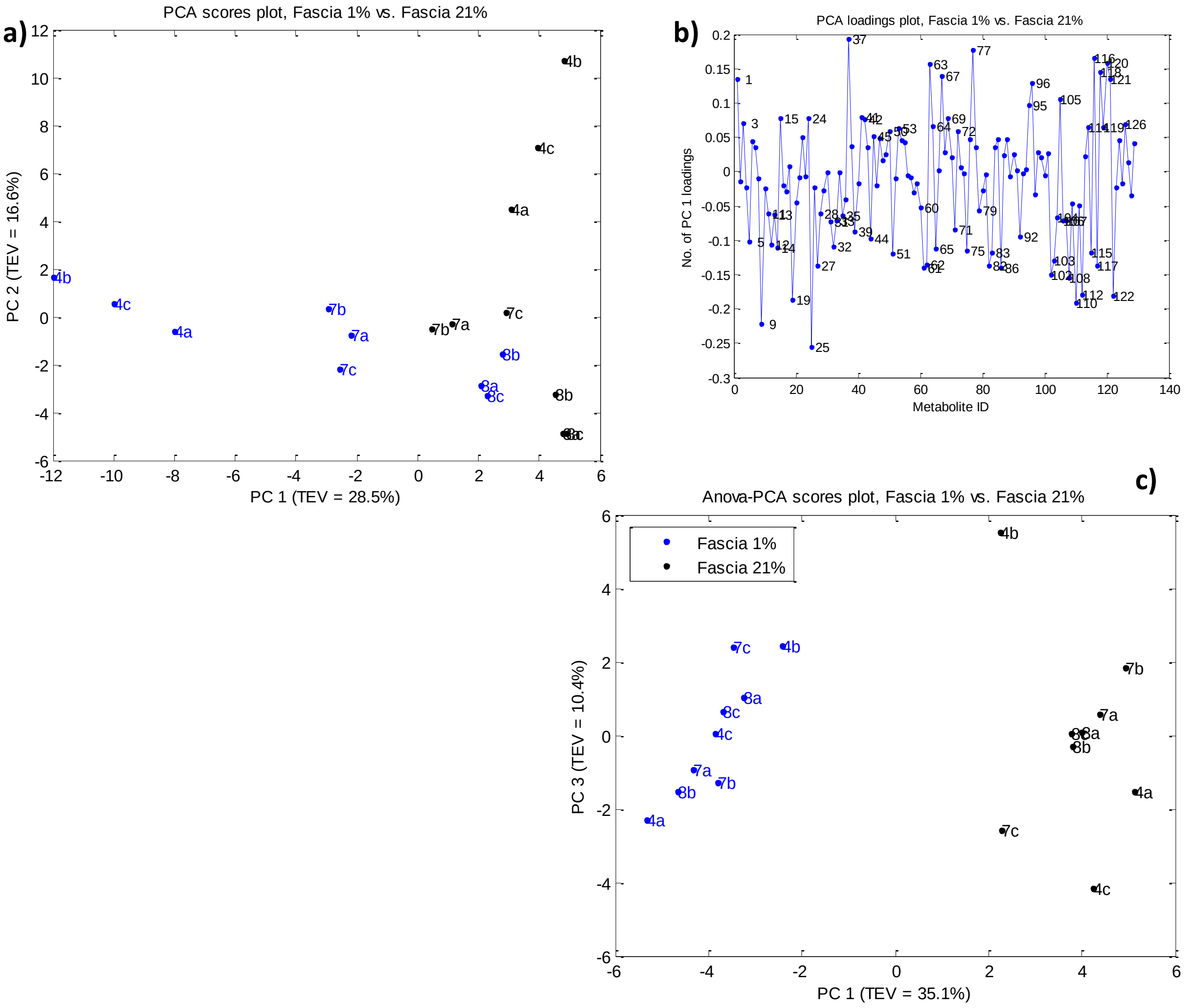


Figure 25 (a) PCA scores plot showing the separation between hypoxic fascia cells and healthy cells. The numbers represent patient the three patients that were entered into the study; patient 4, 7 & 8; each with three biological replicates Black denotes control F21 and blue is perturbed fascia; F1. **(b)** The corresponding PCA loadings plot from the first principal component illustrating which features are important for separation in PC1; the numbers correspond to the metabolite peaks. These data points were used for univariate analysis. Data points that lie close to origin (zero), have little or no contributions toward separations, whereas points that are further away from the origin (zero) have more significant contributions toward the separations. **(c)** PCA scores plot from ANOVA-PCA

4.2.3 Disease *vs.* hypoxia induced and healthy fascia

PCA was then performed on the two disease fibroblasts, with F21 and F1, separately and in conjunction. DFA was applied on the PCs (from PCA) to compare more closely the clustering in C21 with hypoxia induced and healthy fascia and the same was done for N21. Figures 26(a) and (b) represent the PC-DFA scores plot. In Figure 26(a) DF1 separates C21 from F21 & F1. In Figure 26(b) DF2 separates N21 from F21, while DF1 separates hypoxic F1 from normoxic; F21 & N21 clusters. Figure 27(a) shows the scores plot following ANOVA-PCA on N21, C21 and F21 and then including F1 respectively in Figure 27(b). It can be observed from both scores plots the separation between N21 & F21 appears in PC1 while separation between C21 & F1 appears in PC2. In addition, PC2 also separates F1 from F21 and demonstrates these clusters to be closer to those in C21. As previously, loadings were plotted and the variables with data points further away from the origin (zero) that had more significant contributions toward the separations were examined.

The results from Wilcoxon-rank test [143] are as follows. For nodule samples compared with normal fascia (N21 *vs.* F21), 11 metabolites were identified as significantly different. Of these 5 known metabolites; cysteine (identified twice - due to the nature of the analysis, not all metabolites detected can be accurately identified and some having similar chemical structure and were assigned by best possible match score hence all metabolites are assigned a Definitive ID), a sugar, phenylalanine, leucine and aspartic acid a decrease was observed in nodules. A further 6 metabolite features (peaks) were significant but remain unidentified (Table 7). For cords compared with normal fascia (C21 *vs.* F21), 9 metabolites were identified as significantly different. Of these 4 known metabolites; levels of glycerol-3-phosphate increased while leucine, a sugar and pantothenic acid decreased. Some of these trends have been shown in the box-whisker plots in Figures 32(a-j). A further 5 unidentified metabolite features were also significant.

The majority of known metabolites displayed disconcordant changes when disease (N21 & C21) and F1were compared with normal fascia F21. Sugar and pantothenic acid displayed this trend; up in F1 and down in C21. However, cysteine levels were downregulated in both N21 and F1.

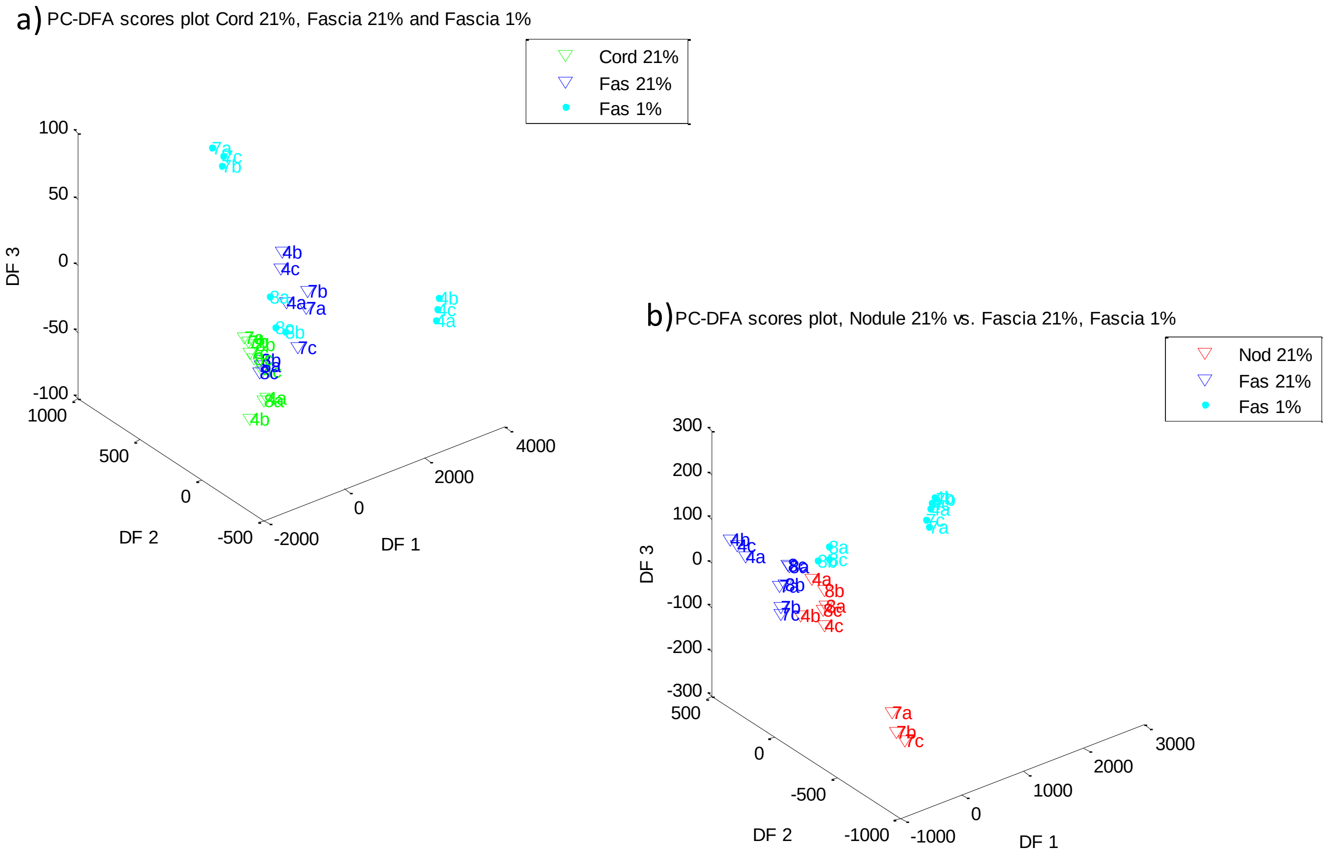


Figure 26 3D DFA plot based on GC-MS analysis showing the relationship between the disease, healthy and hypoxia induced fascia cells (**a)** DD cords, hypoxia induced fascia cell and normal (healthy) subjects, (**b)** DD nodules, hypoxia induced fascia and healthy subjects. Each symbol represents an individual subject (green triangles = C21, red triangles = N21, blue triangles = F21, and cyan dots = F1.

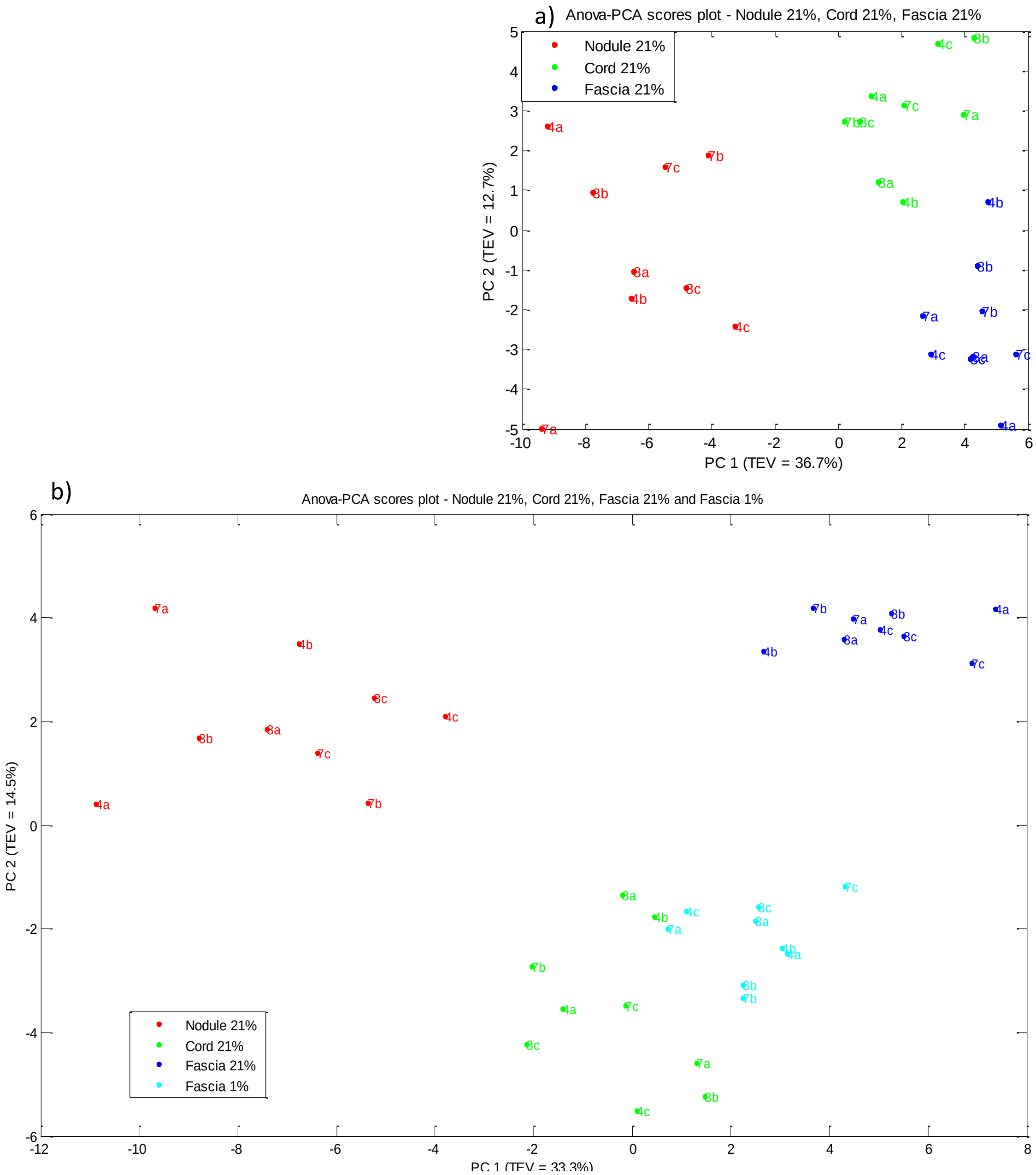


Figure 27 a) ANOVA-PCA scores plot based on autocsaled GC-MS data showing the relationship between ANOVA-PCA scores plot based on autocsaled GC-MS data showing the relationship between two disease and control; N21,C21,F21. The numbers represent the patients. b) ANOVA-PCA scores plot based on autocsaled GC-MS data showing the relationship between the four samples; N21, C21, F21 and F1. The numbers represent three patients. Each coloured dot/point represents a sample type. Number refers to Patients 4, 7 and 8. Letters refer to replicates.

Table 6 Metabolite peaks observed to be statistically different when comparing the intracellular extracts detected for O_2 tensions of 1 and 21%. The legend denotes super families of pathways corresponding to the metabolite detected. The red arrow denotes an increase in metabolite concentration. The green arrow denotes a decrease in metabolite concentration.

No.	Metabolite ID	F1 *vs* F21	N1 *vs* N21	C1 *vs* C21	S1 *vs* S21	N21 *vs* F21	C21 *vs* F21	C21 *vs* N21
		p -value 🟦	*p* -value 🟧	*p* -value 🟩	*p* -value 🟪	*p* -value 🟥	*p* -value 🟩	*p* -value 🟩
1	leucine	0.3828	0.0526 ↑	0.0011 ↑	0.0098 ↑	0.0313 ↓	0.0078 ↓	0.4887
2	sugar	0.0469 ↑	0.2146	0.3981	0.5781	0.6250	0.0156 ↓	0.4631
3	glycerol-3-phosphate	1.0000	0.0024 ↓	0.7344	0.4961	0.2500	0.0156 ↑	0.0391 ↑
4	Pantothenic acid	0.0078 ↑	0.0218 ↑	0.0080 ↑	0.4961	0.6875	0.0547 ↓	0.4887
5	glucose-6-phosphate	0.3828	0.1075	0.5755	0.0098 ↓	0.6875	0.0781	0.4887
6	sucrose	0.9375	0.0413 ↓	0.9479	0.1094	0.5625	0.0781	0.6257
7	isoleucine	0.2188	0.0040 ↑	0.3259	0.7002	1.0000	0.1094	0.3394
8	glucose-6-phosphate	0.5469	0.0854	0.5503	0.0322 ↓	1.0000	0.1094	0.4543
9	citric acid	0.0156 ↓	0.0025 ↓	0.0043 ↓	0.0098 ↓	0.3125	0.1484	0.4212
10	tyrosine	0.4609	0.9359	0.3317	0.0244 ↓	0.3125	0.1484	0.0353 ↑
11	sugar	0.3125	0.1337	0.0152 ↑	0.1309	0.3125	0.1953	0.8904
12	succinic acid	0.1953	0.0486 ↑	0.0522 ↑	0.7646	0.4375	0.1953	0.6387
13	sugar	0.1953	0.0176 ↑	0.0169 ↑	0.2324	0.4375	0.2500	0.3028
14	cystine	0.0469 ↑	0.0029 ↑	0.0353 ↑	0.0313 ↑	0.5000	0.2500	0.9375
15	sugar	0.4375	0.0067 ↑	0.0586	0.4316	0.0313 ↓	0.2500	0.2412
16	glutamine	0.3125	0.0353 ↓	0.1627	0.2754	0.4375	0.2500	0.3804
17	valine	0.3125	0.0395 ↑	0.0836	0.6875	0.8125	0.2969	0.1688
18	tryptophan	0.1953	0.0072 ↑	0.0031 ↑	0.0391 ↑	0.0625	0.3125	0.2061
19	glycylglycine	0.1094	0.0023 ↑	0.0771	0.1250	0.1250	0.3125	0.1484
20	glycerol	1.0000	0.0582	0.0005 ↑	0.8457	0.8438	0.3828	0.5830
21	phenylalanine	0.8438	0.0141 ↑	0.0304 ↑	0.0244 ↑	0.0313 ↓	0.3828	0.0302 ↑
22	methionine	0.9453	0.0129 ↑	0.0479 ↑	0.1094	0.0625	0.3828	0.4263
23	lysine	0.5469	0.1119	0.2322	0.2783	0.2188	0.3828	0.0353 ↑
24	cholesterol	0.5469	0.9359	0.3703	0.0098 ↓	0.5625	0.3828	0.1514
25	cysteine	0.0547 ↓	0.0218 ↓	0.2471	1.0000	0.0313 ↓	0.4609	0.2524
26	glycine	0.3828	0.1365	0.2959	0.0244 ↓	0.2188	0.4609	0.3591
27	uracil	0.2500	0.4997	0.0141 ↓	0.2783	0.5625	0.5469	0.8904
28	lysine	0.3750	0.9359	0.0187 ↓	0.3203	0.6875	0.5469	0.0353 ↑
29	nicotinamide	0.7422	0.6529	0.0366 ↑	0.3594	0.8750	0.5469	0.2439
30	leucine	0.6406	0.0017 ↑	0.1790	0.4131	0.8438	0.5469	0.4212
31	3-ureidopropionic acid and/or beta-alanine	0.0469 ↓	0.2485	0.5862	0.0273 ↓	1.0000	0.5469	1.0000
32	octadecanoic acid	0.4688	0.0148 ↓	0.8789	0.0098 ↓	0.3125	0.5469	0.2163
33	adenosine	0.6406	0.0055 ↑	0.0910	0.5195	0.2188	0.5781	0.2293
34	sugar	0.3125	0.2311	0.0366 ↑	0.4961	1.0000	0.6406	0.9341
35	aspartic acid	0.0391 ↓	0.3547	0.5016	0.2402	0.0625	0.6406	0.5614
36	isoleucine	0.3125	0.0005 ↑	0.0218 ↑	0.0420 ↑	0.0625	0.6875	0.1514
37	glycerol	0.6406	0.0168 ↑	0.0206 ↑	0.0840	0.6875	0.7422	0.9460
38	aspartic acid	0.9453	0.2273	0.1084	0.1934	0.0313 ↓	0.8438	0.0637
39	methionine	0.8438	0.3271	0.3317	0.0420 ↑	0.0625	0.8438	0.2676
40	valine	0.9453	0.0056 ↑	0.3958	0.6377	0.0938	0.8438	0.2166
41	sucrose	0.8438	0.0386 ↓	0.3981	0.8125	0.8125	0.8438	0.3054
42	cysteine	0.0156 ↓	0.4445	0.2180	0.7002	0.0313 ↓	0.9453	0.0637
43	octanoic acid	0.1094	0.0129 ↑	0.2432	0.2324	1.0000	0.9453	0.8552
44	sugar	0.8438	0.0218 ↓	0.8228	0.9658	0.0625	0.9453	0.0043 ↓
45	thiourea	0.1250	0.0078 ↓	0.0039 ↓	0.3750	0.5000	1.0000	0.8125
46	thiourea	0.1094	0.0049 ↓	0.0161 ↓	0.0156 ↓	0.7500	1.0000	0.5703
47	lysine	0.0938	0.0026 ↑	0.2273	0.4961	0.0625	1.0000	0.0023 ↑
48	serine	0.2500	0.0012 ↑	0.6788	0.5703	0.1250	1.0000	0.1934

Pathway 🟧 Amino acid metabolism 🟦 Carbohydrate metabolism 🟨 Fatty acid metabolism ▨ Metabolism of Cofactors and Vitamins 🟫 Urea cycle

Metabolite level ↑ up regulated ↓ down regulated

Table 7 Metabolite peaks observed to be statistically different when comparing the intracellular extracts detected for O2 tensions of 1 and 20%. These metabolites are unknown and hence MET ID in order detected is given. The colours indicate the significant metabolites.

No.	Metabolite ID	F1 *vs* F21	N1 *vs* N21	C1 *vs* C21	S1 *vs* S21	N21 *vs* F21	C21 *vs* F21	N21 *vs* C21
		p -value	*p* -value	*p* -value	*p* -value	*p* -value	*p* -value	*p* -value
1	MET 31	0.0078	0.0836	0.0304	0.2061	0.5625	0.0781	0.3591
2	MET 54	0.0078	0.0003	0.0080	0.0371	0.4375	1.0000	0.7197
3	MET 55	0.0078	0.0766	0.3760	0.1748	0.4375	0.0781	0.8904
4	MET 83	0.0078	0.0702	0.0442	0.6875	0.3125	0.4609	0.5416
5	MET 107	0.0078	0.0910	0.0620	0.0273	0.5625	0.3828	0.2293
6	MET 115	0.0078	0.0641	0.0569	0.0830	0.8438	0.1094	0.7615
7	MET 14	0.0156	0.0641	0.0930	0.1475	0.0938	0.2500	0.7615
8	MET 27	0.0156	0.1712	0.8107	0.3652	0.8438	0.5469	0.0580
9	MET 45	0.0234	0.4688	0.1454	0.2324	0.8438	0.3828	0.6387
10	MET 71	0.0234	0.6009	0.0479	0.5195	0.6875	0.0781	0.1876
11	MET 9	0.0313	1.0000	1.0000	0.5000	1.0000	1.0000	1.0000
12	MET 19	0.0313	0.0437	0.0002	0.1250	0.8750	0.2188	0.7002
13	MET 122	0.0313	1.0000	0.0005	1.0000	1.0000	1.0000	1.0000
14	MET 32	0.0391	0.0836	0.3317	0.1230	0.5625	0.0781	0.6788
15	MET 114	0.0391	0.0329	0.9405	0.8311	1.0000	0.3828	0.0554
16	MET 100	0.8438	0.0009	0.5862	0.0273	0.0313	0.4609	0.1272
17	MET 13	0.6875	0.0070	0.1221	0.7695	0.6250	0.2188	0.3910
18	MET 29	0.7422	0.0079	0.0400	0.0186	0.0313	0.3125	0.7615
19	MET 59	0.1094	0.0100	0.0072	0.7344	0.0625	0.2500	0.3303
20	MET 58	0.5469	0.0100	0.2790	0.3203	0.0313	0.2500	0.3894
21	MET 21	0.8438	0.0112	0.0364	0.0137	0.0313	0.4609	0.7148
22	MET 3	1.0000	0.0158	0.5257	0.5771	0.4375	0.3828	0.4212
23	MET 77	0.1875	0.0181	1.0000	0.0313	0.0625	0.0156	0.1475
24	MET 62	0.3828	0.0329	0.6813	0.7646	1.0000	0.6406	0.9780
25	MET 96	1.0000	0.0340	0.3271	0.8438	0.0313	0.6875	0.3894
26	MET 99	0.3125	0.0442	0.8228	0.7646	1.0000	0.8438	0.1688
27	MET 126	0.8438	0.0479	0.5228	0.7695	0.6250	0.3750	0.0186
28	MET 28	0.3828	0.0534	0.8721	0.1016	0.6875	0.9453	0.8904
29	MET 1	0.4375	0.9697	0.0002	0.4688	0.2500	1.0000	0.7344
30	MET 102	0.5000	1.0000	0.0020	1.0000	0.2500	0.3750	0.0039
31	MET 82	0.1484	0.2954	0.0022	0.2402	0.0938	0.3828	0.4543
32	MET 121	1.0000	1.0000	0.0078	1.0000	0.0625	0.0313	1.0000
33	MET 112	0.2500	0.0625	0.0117	0.5000	1.0000	0.3125	0.4375
34	MET 61	0.1094	0.3144	0.0152	0.0322	0.5625	0.0547	0.9780
35	MET 4	0.2500	0.0645	0.0156	0.3750	1.0000	0.2500	0.8125
36	MET 124	0.8750	0.7002	0.0156	0.7500	0.5000	0.2500	0.6875
37	MET 60	0.4609	0.7782	0.0169	0.1016	0.3125	0.0391	0.4212
38	MET 106	0.3125	0.0582	0.0171	1.0000	0.4375	0.8438	0.9658
39	MET 17	0.4609	0.3981	0.0401	0.9658	0.1563	0.3828	0.0256
40	MET 25	0.1484	0.0836	0.0522	0.1016	0.8438	0.2500	0.8904
41	MET 123	0.5469	0.7475	0.1560	0.0049	0.2188	0.1484	0.7615
42	MET 120	0.7422	0.9679	0.8519	0.0049	1.0000	0.6406	0.3894
43	MET 125	1.0000	0.1701	0.2180	0.0244	0.6875	0.1484	0.6788
44	MET 119	0.3828	0.1590	0.8228	0.0322	0.4375	0.2500	0.5614
45	MET 74	0.7422	0.8092	0.8789	0.8125	0.0313	0.9453	0.7609
46	MET 128	0.0547	0.4939	0.1169	0.2783	0.5625	0.0234	0.0730
47	MET 94	1.0000	0.1250	1.0000	1.0000	1.0000	0.1250	0.0010
48	MET 16	0.4609	0.2273	1.0000	0.3203	0.1563	0.8438	0.0084
49	MET 7	0.7422	0.8405	0.4209	0.3750	1.0000	1.0000	0.0215

4.2.4 Disease *vs.* Disease – Nodule *vs.* Cord

MVA was also performed on the DD cords against DD nodules i.e. C21 *vs*. N21. Little could be depicted from the PCA scores plots (Figure 28). From the Wilcoxon-rank test [143], 5 metabolites showed a significant increase in cords compared with nodules. These are glycerol-3-phosphate, tyrosine, phenylalanine and lysine (identified three times). A sugar was markedly down regulated in nodules. In addition 6 unknown metabolites were also shown to have significant differences (Table 7).

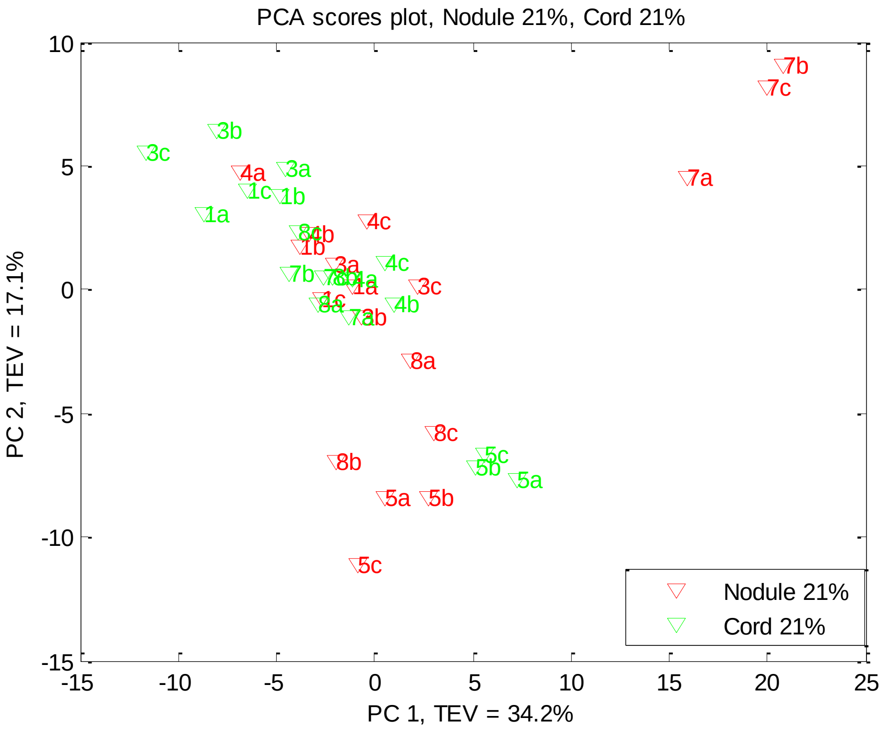


Figure 28 PCA scores plot comparing two disease samples.

4.2.5 Nodule, Cord and Skin overlying nodule under hypoxic stress

MVA was then performed on the DD nodule, cord and SON data to compare against those cultured in hypoxic stresses i.e. N1 *vs.* N21, C1 *vs.* C21 and S1 *vs.* S21. Figures 29 (a-c) shows the PCA scores plot following ANOVA-PCA in each case. It can be observed from all scores plots the separation between samples cultured in normoxic state compared with hypoxic conditions separate in PC1. This is clearer in the case of nodules and SON. The cords (C21) however show overlap in clusters across PC1 and in PC2, while the hypoxic cords (C1) are mostly observed in PC2. The loadings plot and univariate analyses results are as follows.

For nodules in hypoxia 1% compared with cultures in 21% (N1 *vs.* N21) 46 metabolites were identified as significantly different. Of these 30 known metabolites; 20 increased in hypoxic cultures: leucine (identified twice), pantothenic acid, isoleucine (identified twice), succinic acid, a sugar, cystine, another sugar, valine (twice), tryptophan, glycylglycine, phenylalanine, methionine, adenosine, glycerol, octanoic acid, lysine and serine. A decrease in the relative concentrations for the following 10 metabolites was observed: glycerol-3-phosphate, sucrose (twice), citric acid, glutamine, cysteine, octadecanoic acid, a sugar and thiourea (twice). The 16 unknown metabolites that displayed significant differences are listed in Table 7.

For cords in hypoxia 1% compared with cultures in 21% (C1 *vs.* C21) 40 metabolites were identified as significantly different. Of these 19 known metabolites; 14 increased in their concentration in hypoxic cultures: leucine, pantothenic acid, sugars (three), succinic acid, cystine, tryptophan, glycerol (twice), phenylalanine, methionine, nicotinamide and isoleucine. A decrease in the relative concentrations for the following 5 metabolites was observed: citric acid, uracil, lysine and thiourea (twice). The 21 unknown metabolites that displayed significant differences are listed in Table 7.

For SON in hypoxia 1% compared with cultures in 21% (S1 *vs.* S21) 26 metabolites were identified as significantly different. Of these 15 known metabolites; 6 increased in their concentration in hypoxic cultures: leucine, cystine, tryptophan, phenylalanine, isoleucine and methionine. A decrease in the relative concentrations for the following 9 metabolites was observed: glucose-6-phosphate (twice), citric acid, tyrosine, cholesterol, glycine, 3-ureidopropionic acid and/or beta-alanine, octadecanoic acid and thiourea. The 11 unknown metabolites that displayed significant differences are listed in Table 7.

The majority of metabolites displayed concordant changes when cord and nodules were perturbed in hypoxia (i.e. increased in both or decreased in both). Citric acid was downregulated in hypoxic samples including F1. Pantothenic acid levels increased in both disease hypoxic cultures (N1, C1) and F1. Leucine is elevated in N1, C1 and S1 but decreases in N21 and C21 compared with controls F21. Figure 30 displays the general trend observed in DD nodules, cords and fascia from 3 patients cultured in two oxygen tensions. Hypoxia has an influence on disease and healthy cells. Figure 31 shows the separation between DD nodules and fascia and their respective hypoxic counterparts in scores plot from PCA and ANOVA-PCA.

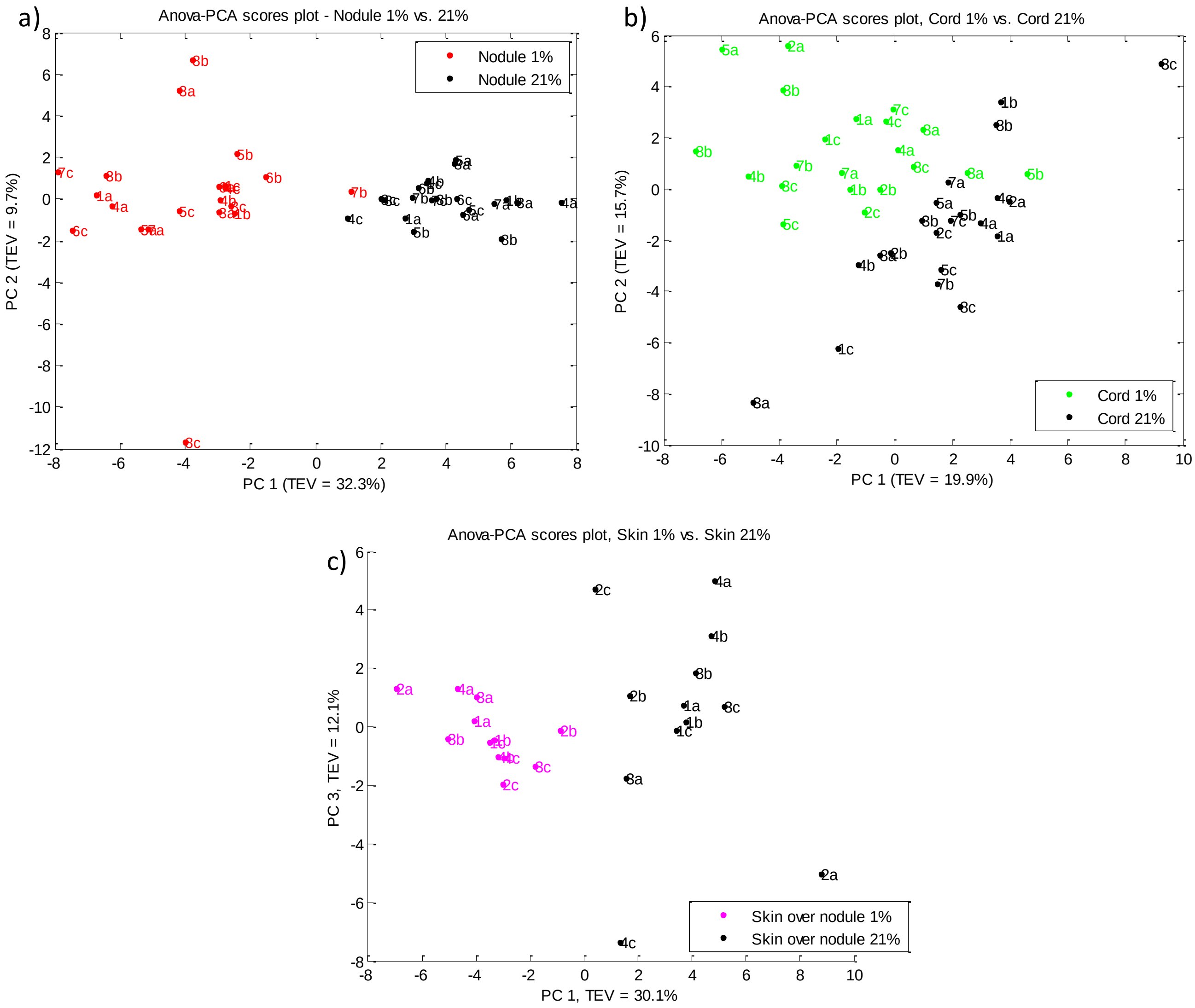


Figure 29 PCA scores plot demonstrating changes due to hypoxic effect in a) nodules, b) cords and c) skin overlying nodules.

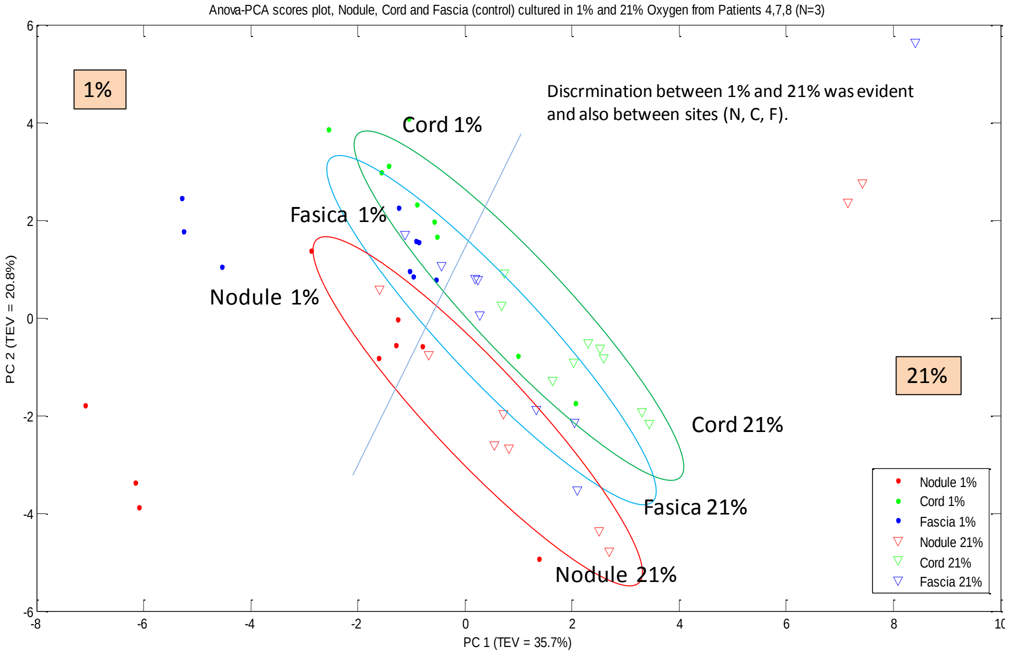


Figure 30 PCA scores plot demonstrating changes due to hypoxic effect in nodules, cords and control samples from three patients. The ellipses do not have statistical significance and are for illustrative purpose only.

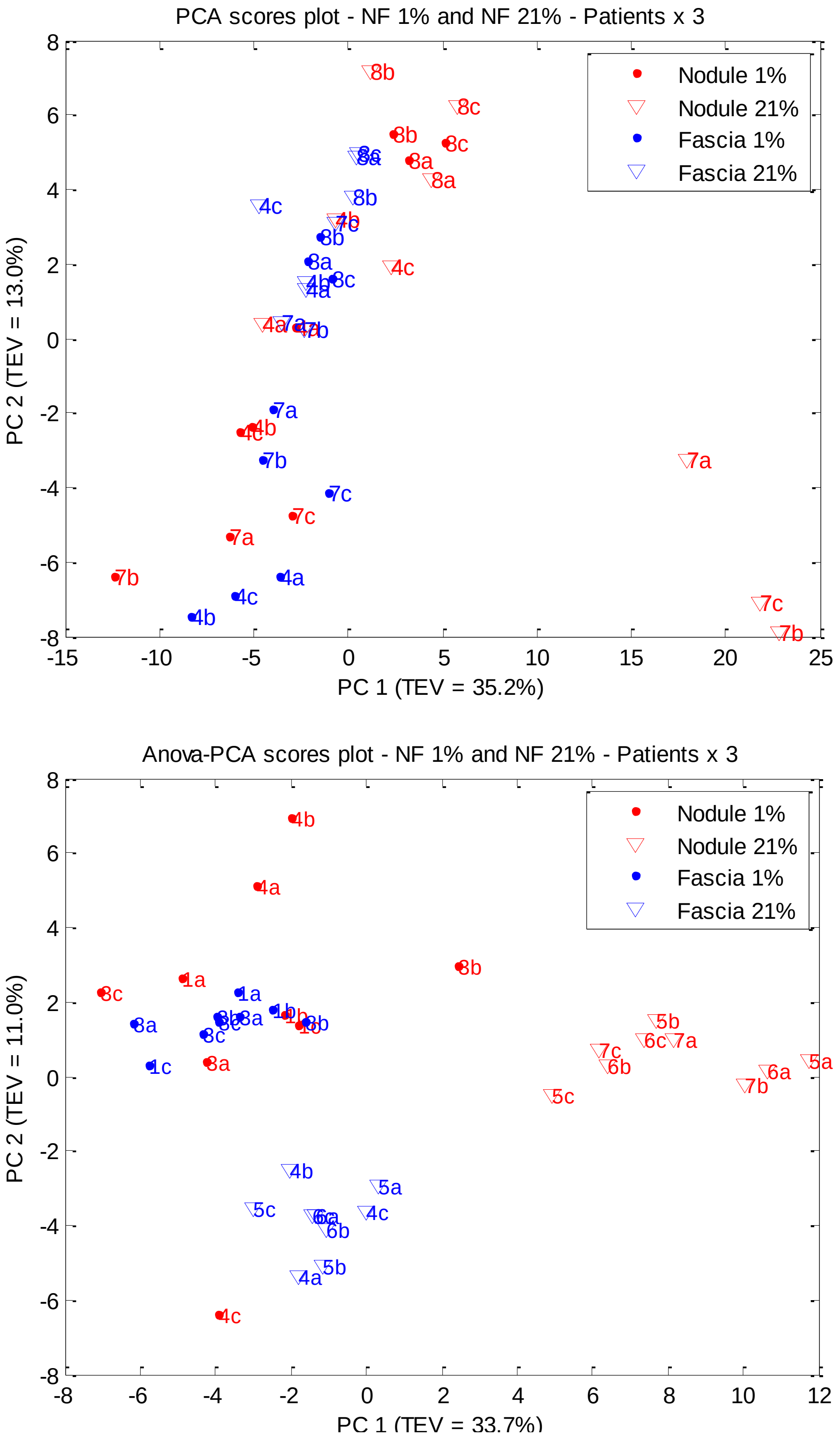
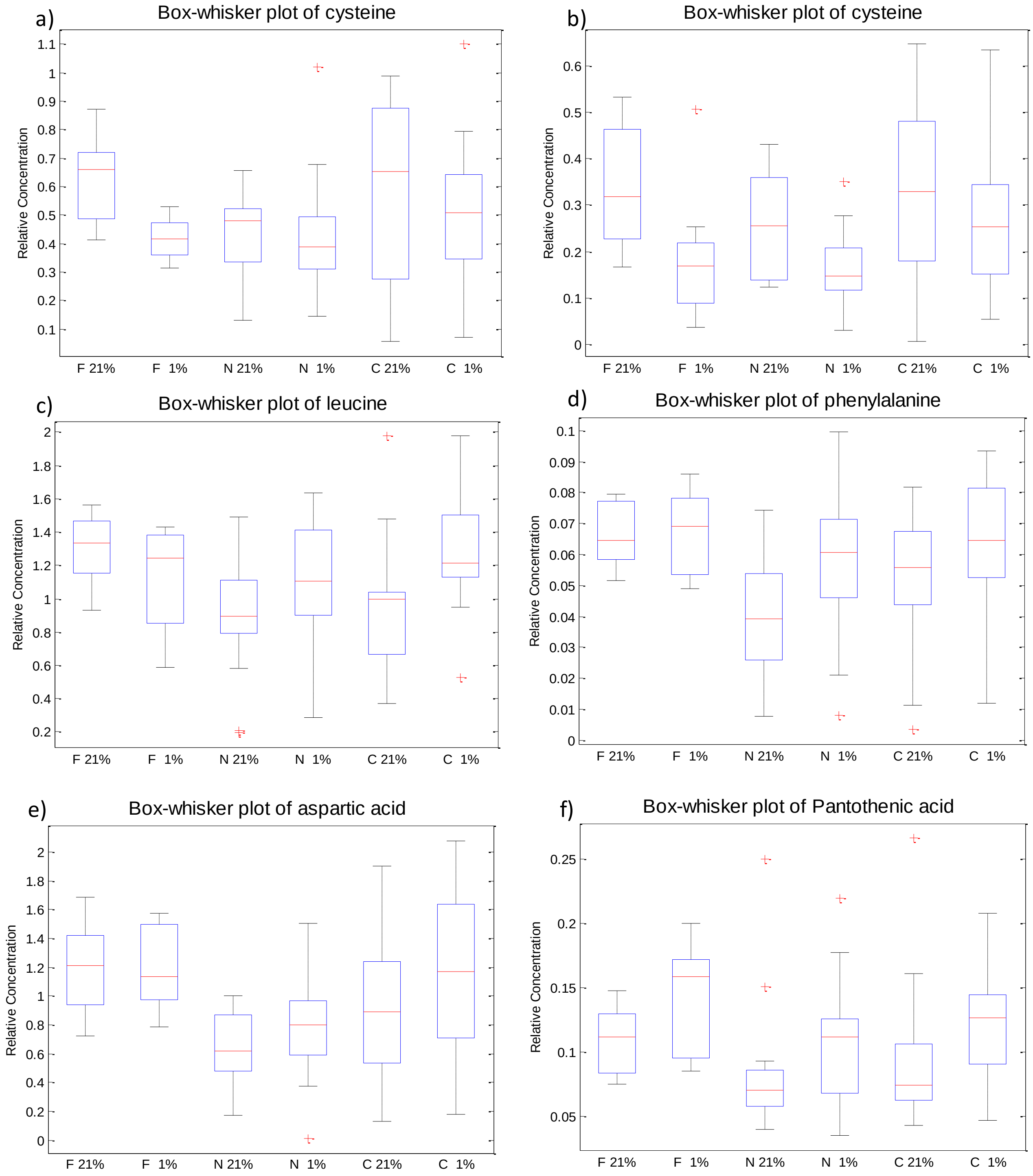


Figure 31 PCA and ANOVA-PCA scores plot demonstrating changes due to hypoxic effect in both nodules and fascia.


a)
Box-whisker plot of cysteine
Relative Concentration
1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
F 21%
F 1%
N 21%
N 1%
C 21%
C 1%
b)
Box-whisker plot of cysteine
Relative Concentration
0.6
0.5
0.4
0.3
0.2
0.1
0
F 21%
F 1%
N 21%
N 1%
C 21%
C 1%
c)
Box-whisker plot of leucine
Relative Concentration
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
F 21%
F 1%
N 21%
N 1%
C 21%
C 1%
d)
Box-whisker plot of phenylalanine
Relative Concentration
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
F 21%
F 1%
N 21%
N 1%
C 21%
C 1%
e)
Box-whisker plot of aspartic acid
Relative Concentration
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
F 21%
F 1%
N 21%
N 1%
C 21%
C 1%
f)
Box-whisker plot of Pantothenic acid
Relative Concentration
0.25
0.2
0.15
0.1
0.05
F 21%
F 1%
N 21%
N 1%
C 21%
C 1%

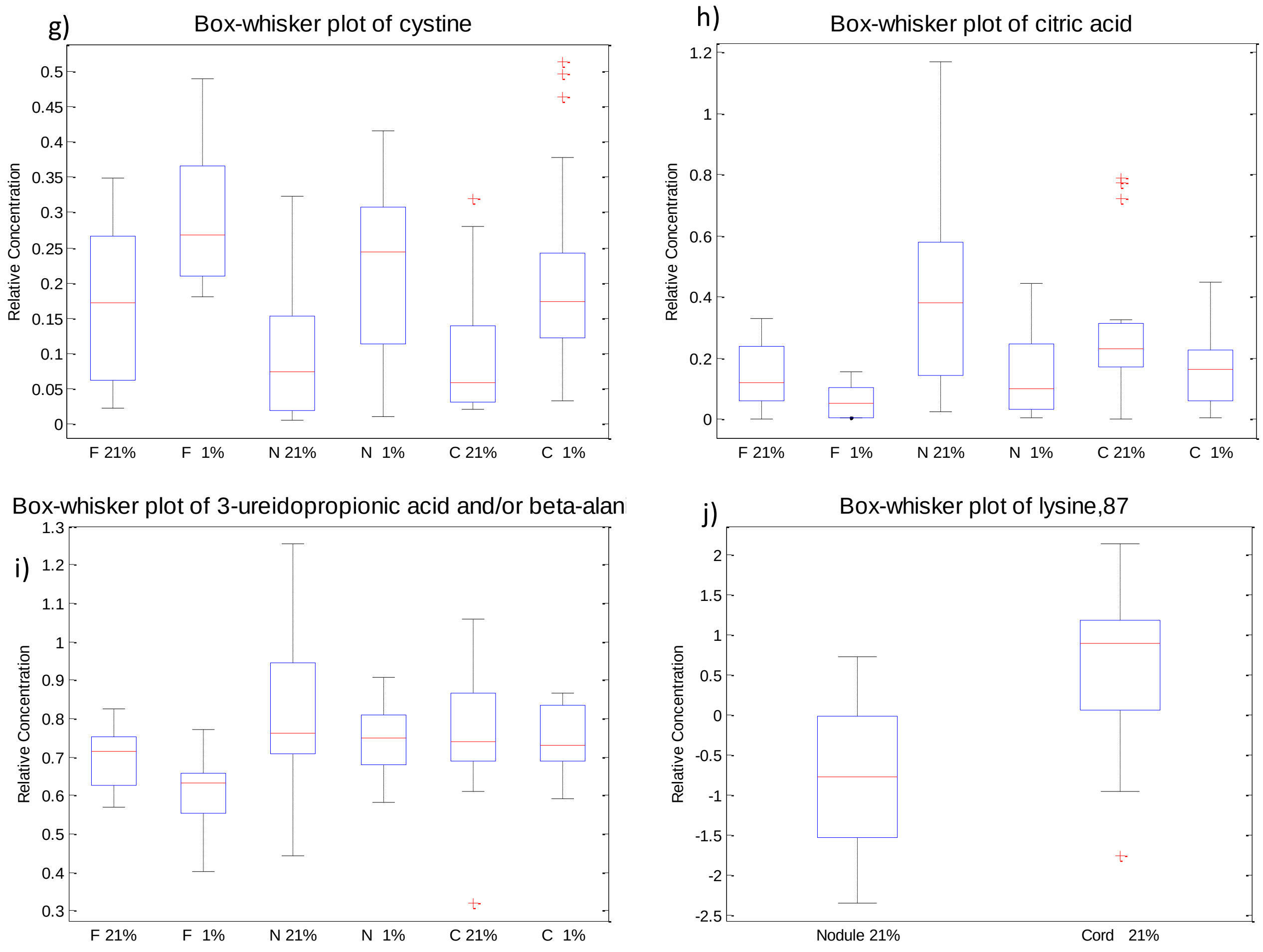


Figure 32 Box-whisker plots demonstrating altered expression of metabolites in intracellular metabolomes of nodules, cords and Fascia in 1% and 21%. a) cysteine, b)cysteine, c) leucine, d) phenylalanine, e) aspartic acid, f) pantothenic acid, g) cystine, h) citric acid, i) 3-ureidopropionic acid &/or beta-alanine and j) lysine. The red line in the box represents the median change in the peak value; the lower and upper boundaries of the box represent the 25th and 75th percentiles, respectively; the lower and upper whiskers represent the 5th and 95th percentiles, respectively; and crosses represent the outliers.**p ≤ 0.05 Wilcoxon sign rank test.

4.2.6 Effect of O2 tension on the metabolic footprint (exometabolome)

Two known (ornithine and leucine) and one unknown metabolites (um01) were identified as significantly different between samples in N21 *vs.* F21 (*n = 3* nodule, *n = 3* fascia; in triplicates). None of these metabolites were found to be significantly different between N1 and N21. Um01 and ornithine were significantly elevated in samples from N21 compared to F21 (p = 0.007 and 0.03 respectively). Leucine levels were found to be significantly decreasing in samples from N21 (p = 0.03) and F1 (p = 0.05) compared to F21 (Figure 33(a-c)).

4.2.7 Pathway analysis

Analysis of all known metabolites in the dataset using KEGG database revealed that members of the amino acid metabolism pathway were significantly overrepresented in the list of metabolites that changed specifically in the setting of hypoxic ischemia. In the case of DD nodule and cord samples, the results indicate, that dysregulation in intermediates involved in carbohydrate and amino acid metabolism may attribute to DD formation. Table 6 illustrates the directional change with red arrow (upregulated) and green arrow (downregulated).

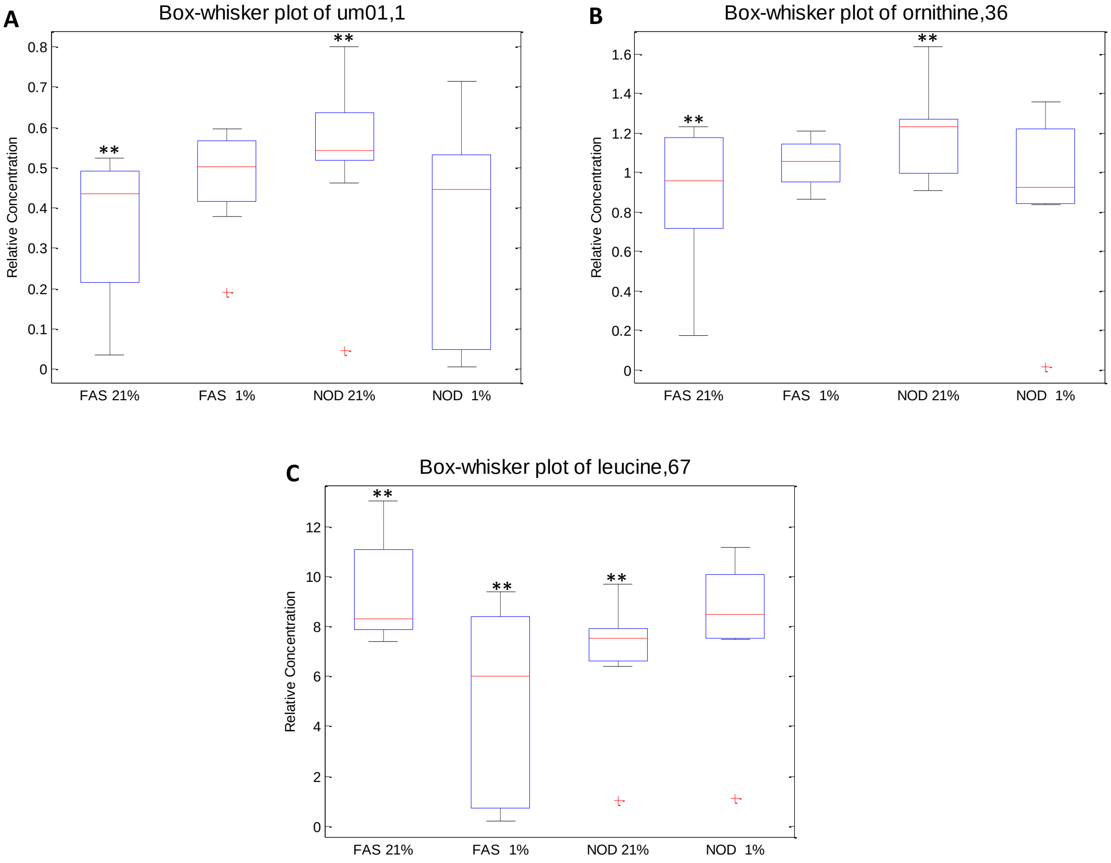


Figure 33 Box and whisker plots demonstrating altered expression of metabolites in conditioned cultured medium in DD Footprints in response to different atmospheric O2 tensions. (a) UM01, (b) ornithine, (c) leucine, **p ≤ 0.05 Wilcoxon signed rank test.

4.3 Discussion

4.3.1 Principal Findings

Metabolomics has expanded rapidly into new scientific fields in recent years. Despite this expansion, there are no reports of metabolomic (metabolic fingerprint or footprint) investigations in DD or any studies of DD cells following exposure to altered oxygenation. Variations in the intracellular metabolomes represent biochemical reactions in cells and therefore can aid in hypothesis generation regarding the activities within DD compared to control cells. From these variations, the pathways affected can be deduced, and the mode of action of disease pathogenesis better understood.

In these preliminary studies despite a small clinical cohort of patients the aim was to identify novel biomarkers and determine whether metabolomic strategies were appropriate for the investigation of DD, including the assessment of comparison of technical and biological variability. In Chapter 3, FT-IR spectroscopy was used to produce unique fingerprints of DD and control fibroblasts based on biological provenance of each sample. This method represents a novel approach for sample characterisation and has allowed for classification of disease samples based on their genotype-phenotype differences and identified a suitable passage representative allowing further investigation of this disease in a controlled manner through construction of a reproducible cell culture model. Although this method proved to be both effective and rapid, it did not yield specific metabolite information. To elucidate key metabolites (and their levels in terms of relative abundance), an untargeted metabolic profiling approach using GC-MS was applied on selected test subjects to identify groups of metabolites associated with a specific pathways thought to be involved in the pathogenesis of DD. A targeted approach could then be applied to focus on specific groups of metabolites (*e.g.* lipids, carbohydrates, amino acids). The results from metabolomics analyses combined with transcriptomic data are confronted with many systems biology tools that facilitate investigation in a more controlled manner in order to deduce key pathways thought to be involved in disease. This will be discussed in Chapter 6.

Using GC-MS, the metabolic profiles of endogenous and secreted metabolites were acquired from (1) DD cords and nodules from the palm against the unaffected transverse palmar fascia (internal control), (2) DD the skin overlying the nodule and (3) those in (1) and

(2) were compared between their counterparts under hypoxic stresses. A total of 129 metabolites were detected in the analysis of the intracellular metabolome and 79 metabolites detected from the footprint. While transport of metabolites between sub-cellular compartments may play a key role in regulation of metabolism, the current quenching, extraction and analytical approach provides only average metabolite levels throughout the full cellular volume.

The first aim was to determine whether we could detect metabolites (endogenous) in healthy cells under hypoxic stress (F1) and compared these with the metabolic profiles acquired from healthy transverse palmar fascia (F21). Next, these metabolic profiles were compared with DD nodule, cord and SON. This was done to investigate whether DD tumour cells were akin to the tumours in which exists the Warburg effect. Furthermore, these metabolic profiles were compared with their counterparts under hypoxic stress to test the response to hypoxia in the metabolomes of nodules, cords and SON. The key players from this study were then used to investigate the cellular dynamics in the transcriptomes (Chapter 5). For these samples differences between metabolic footprints in two different O_2 tensions were also investigated. This advanced sensitivity in detecting changes in DD cultures.

Multivariate analyses demonstrated within-class inter-patient/site variation was less than between-class inter-patient/site variation. This shows that there are greater differences between classes (between patients and their respective samples cultured in different O_2 tensions) than within classes (patients/samples cultured at the same O_2 tension), which are necessary to deduce metabolic changes between the experimental classes. These results show that metabolomic methodologies are suitable for distinguishing metabolic variations in cultures of DD at different O_2 tensions. They also indicate that when appropriate experimental design is applied, observed differences are likely to originate from true biological variation rather than technical inaccuracies. The groups of metabolites that were identified as significantly different in the endometabolomes were amino acids, sugars or intermediates in carbohydrate metabolism as explained below.

Cysteine, decreased in N21, N1 and F1 compared to control samples. This sulfur-containing amino acid is a building block to most proteins. It is unique among the twenty common amino acids because it contains a thiol group with can undergo oxidation of a pair of cysteine residues producing cystine, a disulfide-containing derivative. This reaction is reversible. The disulphide bonds of cystine are crucial to defining the structures of many proteins. Cystine levels however notably increase in all hypoxic samples. This disulphide molecule is generally found in high concentrations in the cells of the immune system, skeletal and connective tissues and skin. Hair and skin are 10-14% cystine [49]. A decrease in sugars levels in disease cells and increase in hypoxic cells indicate altered glucose metabolism in fibroblasts in response to changes in O_2 tension. Leucine decreases in N21 and C21. By contrast it increases in N1, C1 and S1. This essential amino acid is a group of three amino acids known as branched-chain amino acids – BCAA (valine and isoleucine being the other two) whose carbon structure is marked by a branch point. These three amino acids are critical to human life and are particularly involved in stress, energy and muscle metabolism. Stress state e.g. surgery, trauma, cirrhosis, infections, fever and starvation require proportionately more BCAA than other amino acids and probably proportionately more leucine than either valine or isoleucine [49]. Phenylalanine, an essential amino acid is observed as downregulated in N21 but upregulated in C21 compared with N21. Previous literature reports some tumours use more phenylalanine, particularly in melanoma. One strategy suggested to deal with this is to exclude this amino acid from the diet, the other is to increase phenylalanine's competing amino acids, i.e., tryptophan, valine, isoleucine and leucine, but not tyrosine [49]. Aspartic acid levels decrease in N21, This non-essential amino acid is synthesised from glutamic acid by enzymes using vitamin B6. The amino acid has important roles in the urea cycle and DNA metabolism. Aspartic acid is a major excitatory neurotransmitter which is sometimes found to be increased in epileptic and stroke patients. Pantothenic acid, also called vitamin B5, was elevated in all hypoxic samples. By contrast this water-soluble vitamin decreased in C21. Pantothenic acid is required to sustain life; is needed to form coenzyme-A (CoA), and is thus critical in the metabolism and synthesis of carbohydrates, proteins, and fats. It is found almost in all foods [49].

The final aim of the metabolomics experiments was to determine whether GC-MS could define differences in exometabolome of cells exposed to different atmospheric O_2 tensions. Using an appropriate level of statistical significance ($p \leq 0.05$) only three metabolites were identified as significantly different in the exometabolomes; ornithine, leucine and the third is unknown. Ornithine; an amino acid not encoded by DNA, i.e. it is not involved directly in protein synthesis. In mammalian non-hepatic tissues, the main use of the urea cycle is in arginine biosynthesis, ornithine is an intermediate, in this metabolic process, and thus is quite important [144]. It is believed not to be a part of genetic code because polypeptides containing unprotected ornithines undergo spontaneous lactamization. Overall, no major differences were demonstrated in the exometabolomes. This may suggest that cultures in O_2 levels exceeding those estimated as physiological for term human cell cultures (6-7%) may not significantly alter the secreted metabolic content.

4.3.2 Strengths and weaknesses of the study

For mammalian cells in culture, characterisation of metabolic footprint and within the cells (intracellular metabolites; "fingerprinting") both have validity for assessment of cellular function [145, 146]. Metabolic footprinting offers technical simplicity, high-throughput, and automation as samples are simply centrifuged to separate medium and cells prior to analysis and is an appropriate tool for screening of large sample libraries. None of the difficulties associated with sampling intracellular metabolomes, including cell leakage during quenching, are present. It provides a picture of metabolism over a period of time instead of a snapshot, as is the case for intracellular metabolism [145]. However, the intracellular metabolite profile most accurately defines the metabolic status of the cell, and one of the most crucial technical features for characterisation of intracellular metabolites is the quenching and extraction process. Many metabolites are extremely labile and, for example, ATP and glucose 6-phosphate have turnover rates of less than 1-2s and hence cellular metabolism must be stopped immediately (quenched) upon sampling of the cells to prevent / minimise metabolite turnover. Here a metabolomics approach within an integrated systems biology framework is applied to understand the metabolic and functional response to hypoxia in DD cells compared to DD and healthy cells. This high-throughput approach has facilitated the identification of several biomarkers (small molecules) thought to be involved in DD pathogenesis. In addition, simultaneous measurements of two factors in 126 samples i.e. difference in cell type and effect of hypoxia has been possible. Oxidative stress has been linked to many diseases [147-149]. Increased levels of oxidative biomarkers can be detected in aging population; oxidants are produced by endogenous sources, such as mitochondrial respiration, which generates reactive oxygen species (ROS) as by products. Increased production of oxidants results from saturation of antioxidant defenses [150], such as inactivation of antioxidant enzymes and depletion of sulfhydryls due to nutritional imbalance or exposure to xenobiotics [151].

In order to improve the statistical validity of the acquired data, and to facilitate epidemiological studies, it is highly desirable to employ data set sizes composed of hundreds if not thousands of samples. This requires reproducible analyses over time scales ranging from several months to years. The Human Serum Metabolome project [133, 152], conducted by the University of Manchester, AstraZeneca, and GlaxoSmithKline, is a study which presents this particular challenge. Raw and processed analytical data, together with clinical and physiological metadata for the subjects, is a highly useful resource. However, such sample collection regimes of many diseases such as DD are not always possible for metabolomics experiment, e.g. the fact that patient samples are not always available at the time that is optimal for the research project and so in this experiment a small clinical number *n*=8 was only possible.

The findings from this experiment can now lead to a more targeted profiling approach using LC-MS/MS or NMR spectroscopy. The principal carbohydrate product of glycolysis is pyruvate, which was not measured as significant here. A key metabolite of pyruvate is alanine, which is produced via pyruvate transamination. Levels of alanine were also detected. It is clear that several amino acid, sugars and fatty acids altered in DD. This investigation provides evidence of this and future studies should be targeted at the amino acid utilisation patterns.

An apparent issue with sample collection/preparation was observed in TIC by the high chromatogram baselines. This suggested a high concentration of phosphate and glucose in the intra-cellular extracts not due to high concentrations of these metabolites in the cell but the issue of all the footprint and wash step involved potentially leaving minute traces before the quenching solution was added. Rapid quenching following footprint collection was performed instantly to prevent changes in metabolism. This may have been the case in few samples due to the large volume of media being aspirated (35mL) and one wash (10mL) step which then was also aspirated quickly. The high phosphate concentration interfered in accurate determination of the internal standard (succinic acid-d_4) as both elute close to each other. To overcome this problem all data was normalised according to total peak area for all detected peaks for a single sample, i.e. (peak area-metabolite/total peak area for all metabolites) x 100.

Both the strength and weakness of PCA is that it is a non-parametric analysis. One only needs to make the assumptions outlined in Chapter 2.2.2 and then calculate the corresponding answer. There are no parameters to tweak and no coefficients to adjust based on user experience - the answer is unique and independent of the user [153]. This same strength can also be viewed as a weakness. If one knows *a priori* some features of the structure of a system, then it makes sense to incorporate these assumptions into a parametric algorithm - or an algorithm with selected parameters. PCA showed little separation of different samples types in Figure 24(a-b), therefore, a supervised method, ANOVA-PCA was used. This method removed uninteresting sources of variance and the results were more interpretable than without ANOVA. This can be seen in Figures 27(a-b). PCA is most commonly used to gain an intuitive view of the multivariate data, however, when there are 2 or more underlying influential factors PCA is not always the best method to reveal the influence of these factors. Methods such as PARAFAC [154] and PLS [94] may be more appropriate. Again, due to small sample number of patients and only 3 biological replicates such methods were not applicable here.

Supervised classification attempts to build a predictive model based on a subset of samples with known origin (training set). If there are sufficient chemical differences between the samples that are detected by GC-MS the model should be able to predict the class membership of unknown samples. The accuracy of such prediction can then be assessed by using an independent data set (test set) not used during the training stage. DFA was applied and the scores plots can be seen in Figures 26(a-b). Despite supervised learning methods, the variability between patients is observed in this analysis. This could be due to the patients recruited not displaying a similar proliferative disease stage (e.g. early *vs.* late stage, contracted vs. non-contracted hand) and could mostly be at interface or in transformation from one state to another. DD cords do however have the least significant no. of differences compared with normal fascia, while nodules show 6 significant defected metabolites.

The study has potential limitations that should be considered. First, although sampling in patients who served as their own biological controls (1% and 21%) helped diminish interindividual variability and signal-to-noise problems, the study population was nevertheless small. Thus, it is important to note that changes in metabolites that failed to reach nominal significance in this study still may be scientifically important and should be further investigated. For this reason, biological pathway trend analysis offered increased advantage to detect subtle but significant differences. Further testing in larger cohorts will provide the opportunity for both confirmation and exploration of subgroups of interest, including those based on varying proliferative stages, which this study was underpowered to do. Moreover, larger datasets will provide sufficient precision in the estimates of the utility of each marker to allow for appropriate relative weighting of each component. Another weakness is that independent measurements of lactate and ATP have not been made and therefore it cannot be confirmed whether the Warburg is being used or not. The study does highlight many trends and these shall be sought in the transcriptome profile in Chapter 5 and 6.

4.3.3 Conclusions

Currently, metabolomics has been applied as a hypothesis generation strategy, as there is little known of expected metabolic differences in DD. This study is novel in that it also is testing for the first time analytical tools and various DD passaged cells. The difficulties faced and successfully overcoming challenges associated with primary cultures (e.g. high variability and low *n* number) should not be understated. Metabolic profiling is used to detect a wide range of metabolites covering a number of different metabolic classes to provide as large an overview of metabolism as achievable. The intracellular and selective extracellular metabolome have been studied, giving clues to metabolic pathways that may be utilised within the DD cell environment and the effects these cells may be having on their environment through released products. Furthermore this approach reveals the effect these cells may exert upon induction of a hypoxic insult to disease and healthy cells. However, from Table 6 and 7, it can be observed that the metabolites found significantly dysregulated in F1 *vs.* F21 analysis were not akin the N21 *vs.* F21 metabolites. From this we may be tempted to falsify the first hypothesis that difference in disease and healthy cells maybe akin to the differences in healthy cells in normoxia and hypoxia as the number of identified metabolites do not coincide in the cases (except cysteine). Because such few dysregulated metabolites have been identified in disease it is more appropriate to test this hypothesis in the transcriptome for a more clear understanding as it is expected that changes occurring in the cells metabolome would also be observed at the transcript level.

What can be concluded is that these data suggest that hypoxia possesses a role in DD investigations as a number of metabolites were significantly upregulated in disease upon perturbations which were identified previously as downregulated in disease. Two areas of metabolism were highlighted for systems biology and transcriptomics investigation; amino acid (leucine cysteine) metabolism and carbohydrate metabolism. Fatty acid and metabolism of cofactors and vitamins should be investigated. These findings will be correlated with those in Chapter 5 as a better understanding of the mechanistic links between cellular metabolism and transcriptome profiles may ultimately lead to better treatments for DD and other disease.

Chapter 5

Dynamic changes in Dupuytren's Disease and control transcriptome in hypoxia

5.1 Introduction

5.1.1 Exploring dynamical changes in DD and control transcriptome

In Chapter 4 we investigate the dynamic changes occurring in DD phenotypes and healthy fascia in response to hypoxia. A number of key metabolites and pathways that may contribute to DD progression or have been invoked as a consequence of hypoxic stress are highlighted. Intermediates involved in amino acid and carbohydrate metabolism have shown significant differences from this analysis. In this Chapter we extend this approach to investigate this effect in their transcriptome. This study examines whether gene expression analysis of such cells could provide a more representative picture of the dynamics involved in DD. It is surmised these transcripts will produce a specific signature for DD complementing the metabolomics study and allow us to look for metabolic pathways and cell signaling pathways / targets in a controlled systematic manner. The emerging data will form the basis for selecting appropriate models for pathway studies.

In this study, we test the experimental model of hypoxia induced in DD and healthy cells, the effect on its transcriptome profile and seek to understand factors that induce discriminating changes in both the metabolic and signaling pathways. The aim is to identify potential biomarkers of DD and characterise a possible altered biochemical profile of the DD cells compared to healthy fascia in patients with and to determine the metabolic impact of hypoxia in both. From results in Chapter 4, the nodule displays the maximum differential response (compared with cord and SON is a different cell type) compared with healthy fascia.

In this study, we set out to (1) understand progression of disease compared to healthy (nodule *vs.* fascia) (2) profile DD nodules to investigate the effect of 1% oxygen tension on these samples (1% *vs.* 21%), and 3) investigate the perturbation effect on healthy fascia cells to examine whether healthy cells in hypoxia mimic disease state scenario. The study employed Affymetrix Human Genome U133 Plus 2.0 GeneChip oligonucleotide arrays [58] to determine transcript profiles of fibroblasts cultured in normoxic and hypoxic conditions.

The study revealed a small number of DE transcripts that were common in N21 *vs.* F21 and F1 *vs.* F21 analysis. These transcripts were involved in the following pathways: - MAPK signaling pathway, ECM-receptor interaction, p53 signals pathway, tyrosine metabolism, nicotinate and nicotinamide metabolism, phenylalanine metabolism and vitamin B6 metabolism.

Landmark genes previously identified to be associated with DD were also confirmed and key collagens, collagenases, metalloproteinases, and inhibitors, cell adhesion molecules and integrins were identified. Genes involved in focal adhesion, regulation of actin cytoskeleton, ECM-receptor interaction, vascular smooth muscle contraction pathways, cysteine and methionine metabolism were markedly dysregulated. From the perturbation effect in N1 *vs.* N21 fatty acid metabolism, toll-like receptor signaling pathway, biosynthesis of unsaturated fatty acids, PPAR signaling pathway, citrate cycle (TCA cycle), glycine, serine and threonine metabolism pathways were enriched. These results suggested correlation with the molecules dysregulated in the metabolomes. In addition strengthened our hypothesis that DD is a disease of networks, where molecules are interconnected and a number of amino acid metabolism molecules are actively DE in DD nodules and this perturbation effect is observed at both the transcriptome and metabolome levels within a cell. These results indicate a number of important candidate genes associated with DD formation, which may provide clues for molecular mechanisms involved in DD pathogenesis.

5.2 Results

5.2.1 Gene expression Analysis

Affymetrix microarray technology was adopted to address whether the metabolites found to be significantly differentially expressed in Chapter 4 between F1 *vs.* F21, N21 *vs.* F21 and N1 *vs.* N21 correlate with differences at the mRNA level. A total of 12 samples were analysed using three analysis methods (*n = 3* transverse palmar ligaments from DD cases acting as control, *n* = 3 transverse palmar fascia cultured in hypoxia, *n = 3* DD nodule samples and *n = 3* DD nodules cultured in hypoxia). The boxplots shown in Figure 34 illustrate the distribution for expression values created by each of the normalisation methods RMA, GC-RMA, and multi-mgMOS. The PCA scores plot from the multi-mgMOS normalised and GC-RMA normalised expression set are given in Figures 35(a) and 35(b) respectively.

It can be seen from Figure 35(a) PC1 appears to be separating the arrays by patient, whereas PC2 appears to be separating the arrays by oxygen tension. The grouping of N1 & N21 clusters and F1 & F21 clusters is much tighter in pumaPCA using multi-mgMOS normalised data. Figure 35(b) using GC-RMA normalised data and standard PCA also shows a similar trend across PC1 where separation is observed with respect to patients and PC2 separates for disease or healthy cells. This is not the case with patient 60 because PC2 separates the samples based on oxygen tension not disease. But separation can still be observed between disease and healthy samples (in all conditions) in patient 60. One reason the difference observed here between pumaPCA and standard PCA is because expression levels have been normalised using GC-RMA in Figure 3b; the quantile normalisation used in GC-RMA removes such differences. On the whole, both methods show similar observations – patients separated along PC1 and oxygen tension (1 always higher than 21 for the same patient/cell type) along PC2. Samples of same type from each patient are clustered but there is no strong systematic trend for the disease vs. healthy samples that can be seen in PC2.

HCA using Euclidean distance (average linking) matrix was applied to GC-RMA normalised data following a Two-way ANOVA, cut off value $p \leq 0.05$ and fold change > 2 resulting in 461 significant genes. Genes were clustered by applying two-way clustering where on one axis (horizontal) array samples (F21 (control), F1, N21 and N1) and on the other axis (vertical) are the genes. The heat map and clustering algorithms identified subsets of genes that were co-expressed similarly. An expanded region (cluster of genes) resulting from agglomerative HCA of all samples is shown in Figure 36. Presence of large contiguous patches of colour represent groups of genes that share similar expression patterns over multiple conditions. Branch lengths represent the degree of similarity between the genes. Co-regulated and functionally related genes were statistically grouped into clusters. Larger groups of clustered genes were examined where we observed a strong tendency for these genes to share common roles in cellular processes. Genes illustrated in Cluster A consist of co-expressed genes which appear to separate disease from fascia irrespective of oxygen concentration. Cluster B groups co-expressed genes impacted by oxygen tension.

In order to identify genes unique to each of these conditions differential gene expression was analysed for F1 *vs.* F21, N21 *vs.* F21 and N1 *vs.* N21 for each of the three patients using 1) Limma that produced *p*-values and *q*-values (adjusted *p* values), and $\log_{10}$ fold change in genes (up, down, or unchanged) between the conditions and these probesets were compared with 2) Puma that generated the probability of PPLR statistical scores for each probe set by two-way comparisons between pairs of these sample groups. PPLR values range from 0 to 1, with values closest to 0 representing the most significantly down-regulated probe sets and values closest to 1, representing the most significantly up-regulated probe sets. Values of 0.5 represent no significant change.

Probe sets ≥ 1.5 threshold value and with a PPLR value > 0.9 (significantly up-regulated) and < 0.1 in at least one of the comparisons were compared for their presence across the statistically and biologically significant transcript lists retrieved from these individual analyses (i.e., transcripts over- and under expressed in gene lists generated from pairwise analyses of F1 *vs.* F21 compared with those in N21 *vs.* F21. The same was done for other N1 *vs.* N21 gene lists. Statistical false-positives in this data have been minimized by using a high cut-off PPLR value for puma methods (0.9 and 0.1) and a FDR (Benjamini and Hochberg) for Limma p-values to locate more disease specific and hypoxia-responsive genes. A correction method for PPLR values does not exist at present and filtering in this way was deemed appropriate.

Venn diagrams were generated to illustrate the distribution of the probe set for the Limma probe sets with FDR applied. Over representation analysis of probes falling in top 1000 PPLRs values and weakest 1000 PPLR values was followed by ontological mapping to investigate significant overrepresentations and associations between genes belonging to specific patterns and Gene Ontology categories. The same gene lists were also used for KEGG pathway analysis (refer to Tables 31-33 in Appendix D). To test whether the genes could be mapped onto the pathways detected by metabolomics and also show changes at mRNA level by any chance, a less stringent selection criteria was applied by including PPLR values >0.8 and <0.2 counting for the genes showing 1.5-fold or more changes in two out of three independent experiments e.g. patient to patient comparison. Table 8-9 show significant DE genes identified from each of the pairwise analyses that were common across three patients from F1 *vs.* F21 analysis and N21 *vs.* F21 analysis. Figure 42 show GO analysis from top over expressed genes in N21 *vs.* F21. Table 10-12 show the top scoring pathways from the overlapping (common) DE genes identified from pairwise analyses in clusters from heatmaps (Figure 43-44, 48). Tables 13 and 14 show a section of DE from N1 *vs.* N21 analysis.

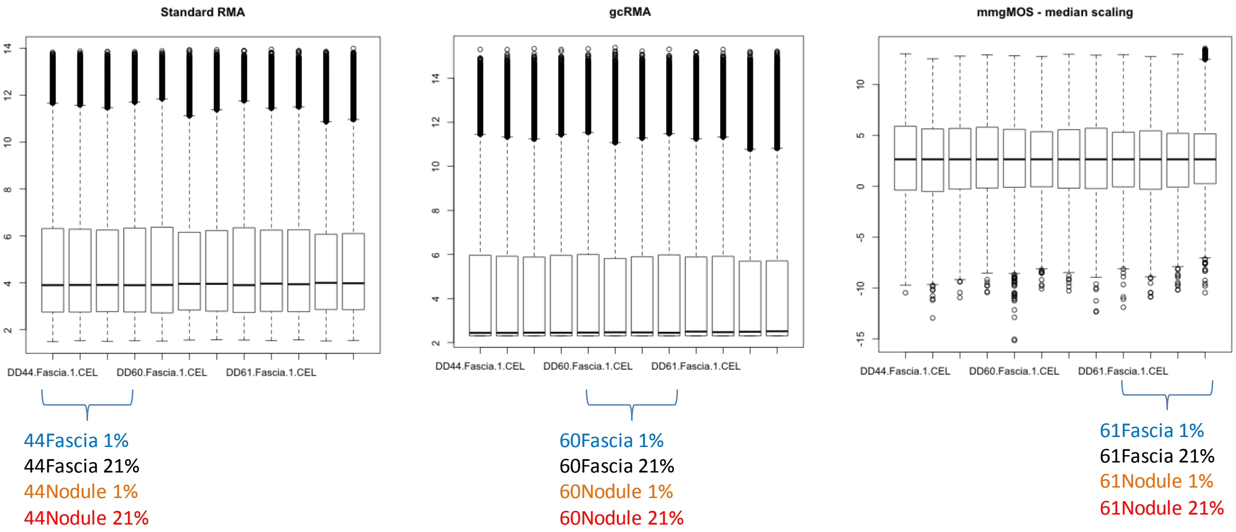


Figure 34 Boxplots to show distributions of expression values created by each of the summarisation methods. Data set processed by RMA, GCRMA and multi-mgMOS (global median scaling) respectively.

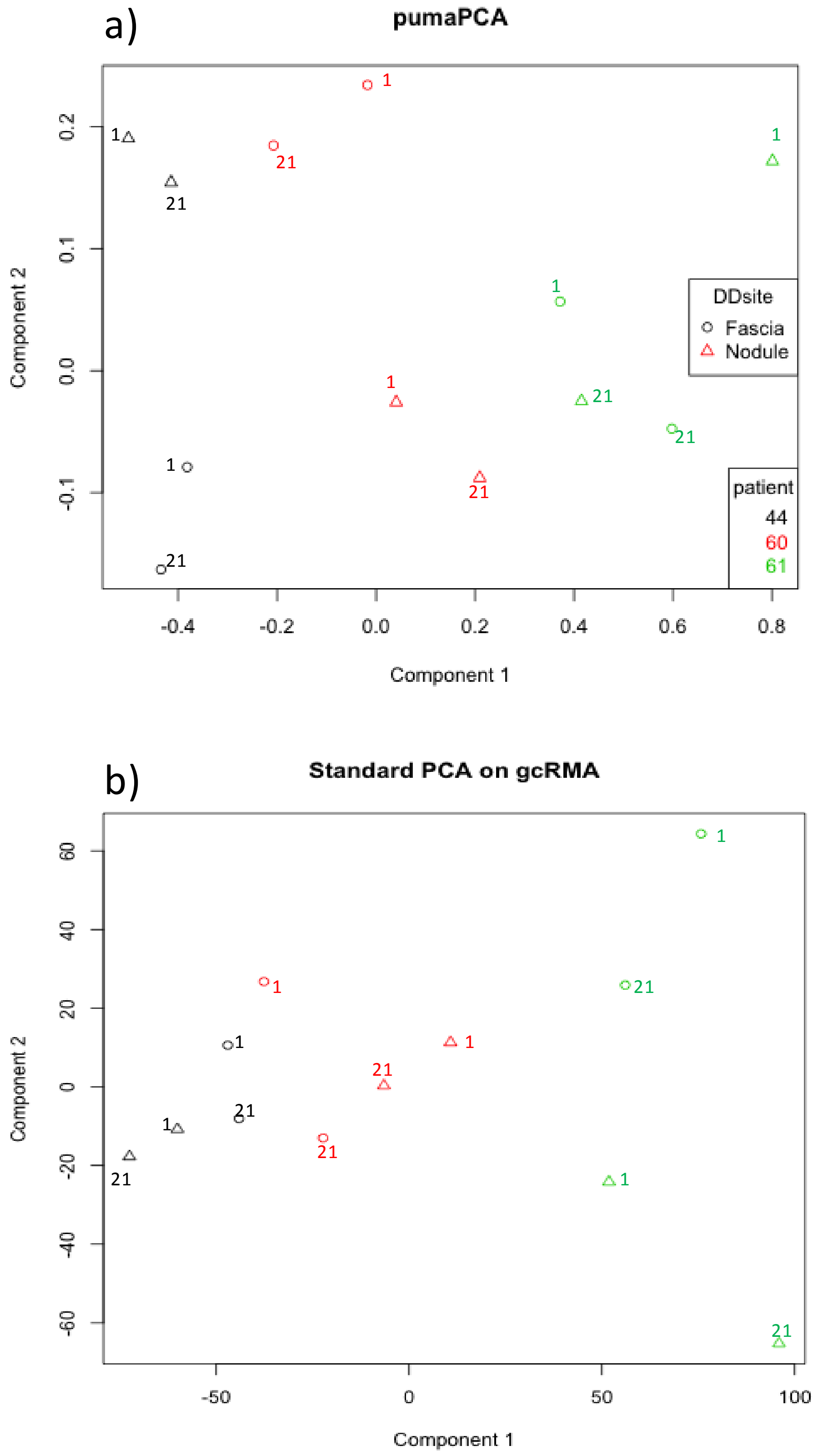


Figure 35 First two components after applying a) pumapca and b) prcomp to the 12 samples.

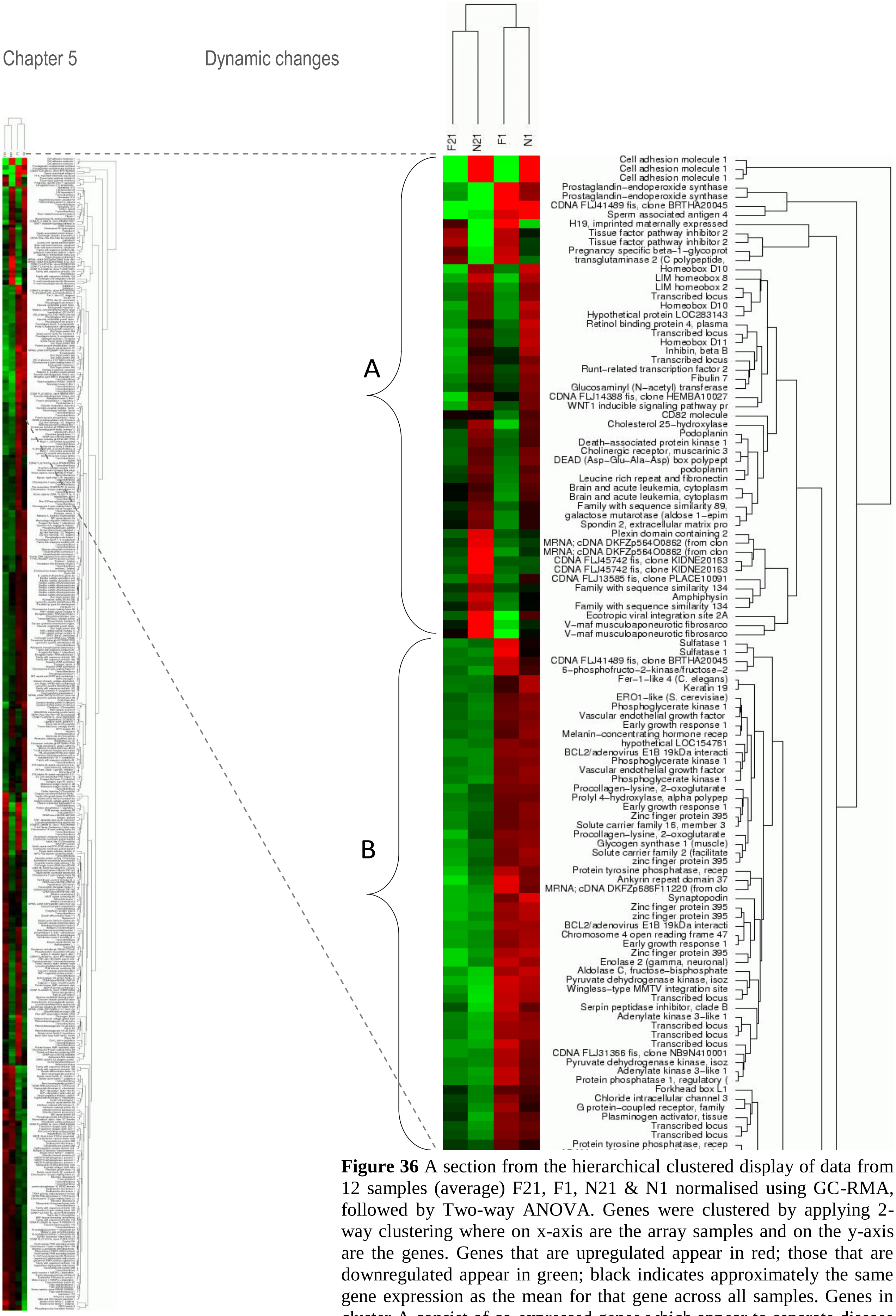


Figure 36 A section from the hierarchical clustered display of data from 12 samples (average) F21, F1, N21 & N1 normalised using GC-RMA, followed by Two-way ANOVA. Genes were clustered by applying 2-way clustering where on x-axis are the array samples and on the y-axis are the genes. Genes that are upregulated appear in red; those that are downregulated appear in green; black indicates approximately the same gene expression as the mean for that gene across all samples. Genes in cluster A consist of co-expressed genes which appear to separate disease from fascia. Cluster B groups co-expressed genes impacted by oxygen tension.

Probe sets with a probability of positive log-ratio (PPLR) value greater than 0.9 (significantly up-regulated) in at least one of the comparisons generated a list of 669 probe sets. 47 probe sets are uniquely up-regulated when comparing the gene expression profile of F1 *vs.* F21 and N21 *vs.* F21. When comparing N21 *vs.* F21, 130 probe sets are uniquely upregulated, interestingly however, 4 probe sets are upregulated in all comparisons. The number of probe sets associated with a significantly down-regulated PPLR value (0.1) for N1 *vs.*N21was higher than the number for up-regulated probe sets. 256 probe sets alone were unique and significantly down regulated in N1 *vs.*N21 (Figure 37).

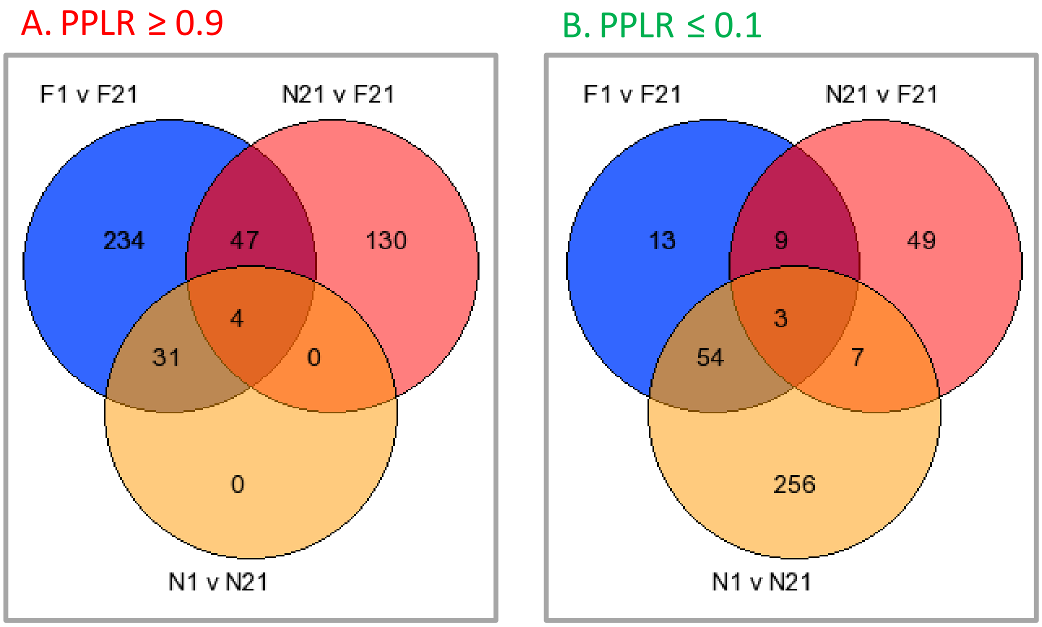
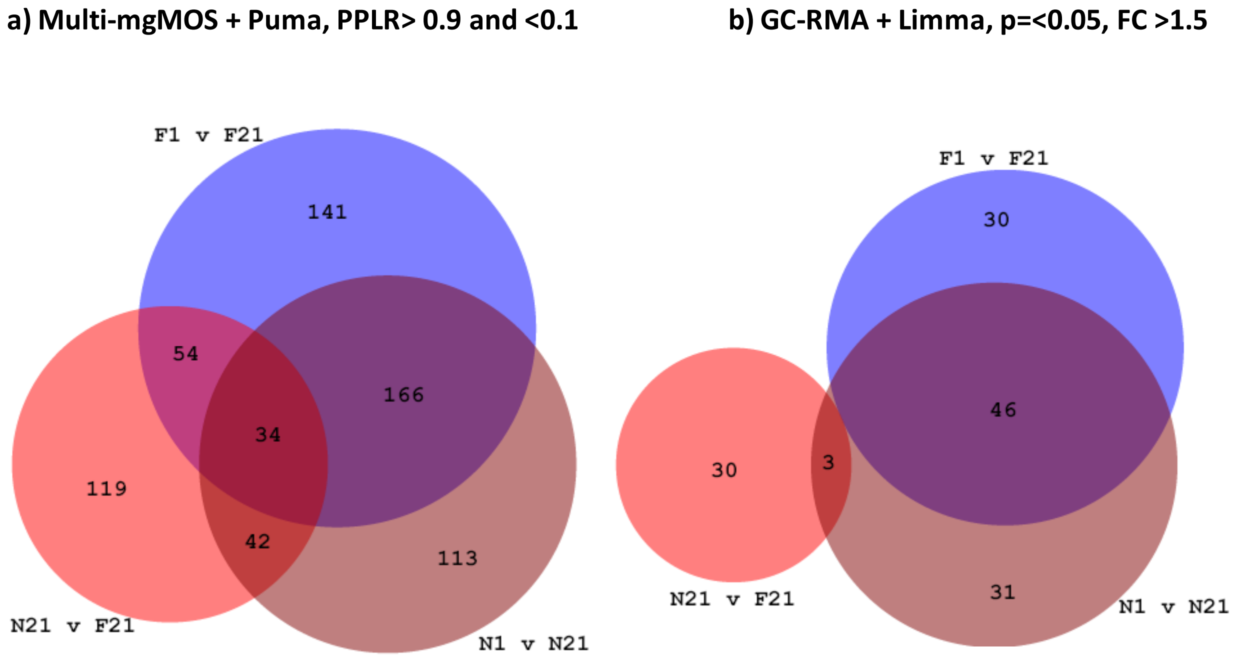


Figure 37 Distribution of probe sets with a PPLR value greater than 0.9 (A), PPLR less than 0.1 (B) in any one of the three comparisons (F1 *vs.* F21, N21 *vs.* F21 or N1 *vs.* N21 are shown in Venn diagrams.

Combining probe sets that were either significantly upregulated (0.9) or significantly down-regulated (0.1) in at least one of the comparisons (F1 *vs.* F21, N21 *vs.* F21 and N1 *vs.* N21) generated a list of 669 probe sets (Figure 38(a)). 141 of these probe sets were unique to F1 *vs.* F21, 119 of these were unique to N21 *vs.* F21 and 113 unique to N1 *vs.* N21. A total of 54 genes were significantly up-regulated or down-regulated and overlapping in F1 *vs.* F21 and N21 *vs.* F21, 166 overlapped in F1 *vs.* F21 and N1 *vs.* N21). 34 genes were common in all three comparisons (F1 *vs.* F21, N21 *vs.* F21 and N1 *vs.* N21). By contrast GC-RMA normalisation followed by Limma models revealed only 30 probe sets unique to N21 *vs.* F21 of which none overlapped with F1 *vs.* F21 combination, and relatively lower number of probesets unique to F1 *vs.* F21 (30 probe sets), and unique to N1 *vs.* N21(31) can be seen in Figure 38(b). The 669 probe sets with a PPLR value greater than 0.9 or less than 0.1 in at least one of the comparisons (F1 *vs.* F21, N21 *vs.* F21 and N1 *vs.* N21) underwent clustering analysis. A number of distinct clusters were generated and probe sets from each cluster were subjected to expression analysis systematic explorer (EASE) online tool and DAVID [63]. For each cluster, overrepresented GO groups were identified. For each individually patient, significant DE was examined and Figure 39 shows the extent of patient variability. Patient 61 demonstrated the maximum number of DE genes for N21 *vs.* F21 analysis and N1 *vs.* N21 (514 and 752 respectively).



a) Multi-mgMOS + Puma, PPLR> 0.9 and <0.1

b) GC-RMA + Limma, p=<0.05, FC >1.5

Figure 38 Analysis of microarray data of DD samples using a) mmgMOS and IPPLR for differential expression PPLR values were calculated for fascia in normal and hypoxia (F1 *vs.* F21), nodules in normal and hypoxia (N1 *vs.* N21) and nodules compared to healthy fascia (N21 *vs.* F21). Distribution of probe sets with a PPLR value greater than 0.9 and less than 0.1 in any one of the three comparisons (F1 *vs.* F21, N21 *vs.* F21 and N1 *vs.* N21) are shown in Venn diagram b) GC-RMA and Limma to identify differential expression, probe sets where p-value ≤ 0.05 and fold change > 1.5 are shown Venn diagram.

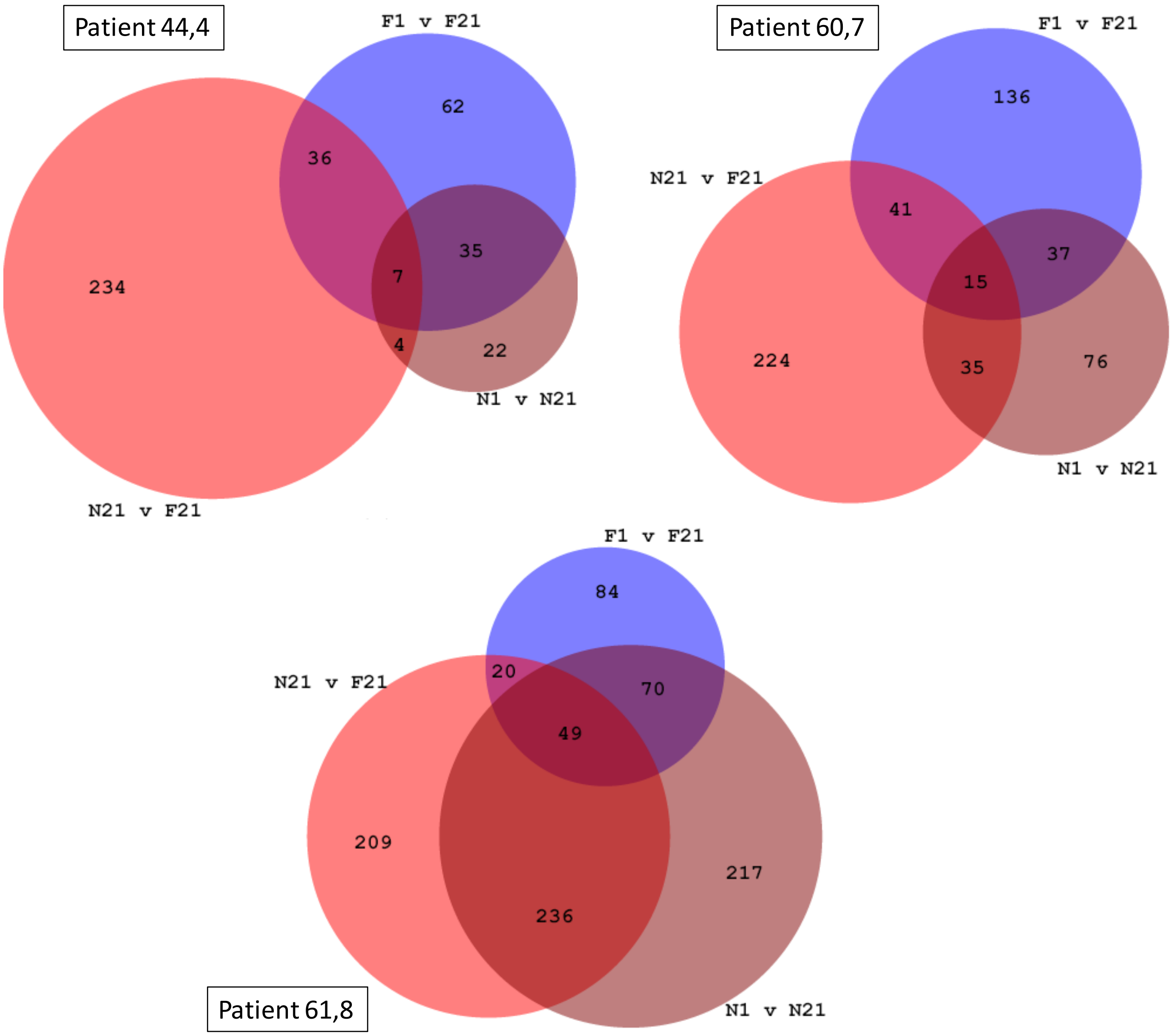


Figure 39 Venn diagrams to show filtered probelists following PPLR thresholds applied to data sets from individual patients to determine extent of patient variability. Patient 44 (4 from metabolomics study), patient 60 (7 from metabolomics study) and patient 61(8 from metabolomics study). PPLR values were calculated for Fascia in normal and hypoxia (F1 *vs.* F21), nodules in normal and hypoxia (N1 *vs.* N21) and nodules compared to healthy fascia (N21 *vs.* F21). Distribution of probe sets with a PPLR value greater than 0.9 and less than 0.1 in any one of the three comparisons (F1 *vs.* F21, N21 *vs.* F21 and N1 *vs.* N21) are shown.

5.2.2 Effect of O_2 tension in intracellular transcriptomes of healthy cells

The Venn diagrams from puma analyses in Figure 39 show the number of transcripts that were significantly differentially expressed. In hypoxic conditions F1 cells show upregulation in a number of intermediates involved in glycolysis and carbohydrate metabolism. Some hypothetical proteins are present. This was observed using both methods (puma + limma). Figure 40 demonstrates the individual filtered probelists from each patient which allows determination of the variability between patients upon hypoxia. Patient 60 consists of 136 uniquely significant DE genes in F1 *vs.* F21, while patient 44 demonstrates the smallest change with 62 differentially expressed genes. The 25 gene names relevant to the 36 probe IDs common across patients are given in Table 8.

A number of transcripts are consistent across all patients; some showing greater/lower PPLR values, however the trends observed are mostly similar across the three patients with varying levels of significant expression. In patients 61, 70 probe sets from N1 *vs.* N21 list are overlapping with many of the differentially expressed transcripts in F1 are also observed in the N1. While no transcripts were identified common from Limma analysis in this list for N21 *vs.* F21**.**

Effect of Hypoxia in healthy cells across patients F1 *vs.* F21

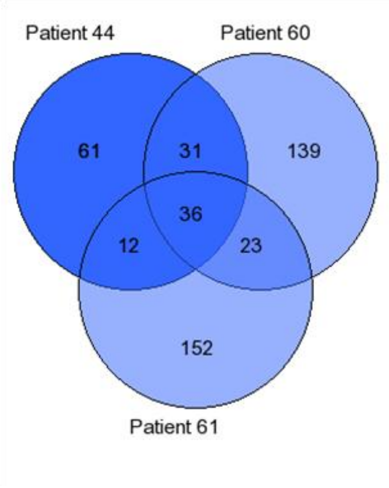


Figure 40 Venn diagram showing comparison across patient samples where healthy fascia was compared to perturbed fascia.

Table 8 The 25 genes (36 probes) significant common in all 3 samples where healthy fascia was induced with hypoxic environment.

Affymetrix Probe ID	Gene Name
202464_s_at	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 3
204347_at	adenylate kinase 3-like 2; adenylate kinase 3-like 1
207692_s_at	aggrecan
227337_at	ankyrin repeat domain 37
201623_s_at	aspartyl-tRNA synthetase
201848_s_at	BCL2/adenovirus E1B 19kDa interacting protein 3
222646_s_at	ERO1-like (S. cerevisiae)
217967_s_at	family with sequence similarity 129, member A
217871_s_at	hypothetical protein LOC284889
201650_at	junction plakoglobin
212689_s_at	lysine (K)-specific demethylase 3A
200738_s_at	phosphoglycerate kinase 1
202619_s_at	procollagen-lysine, 2-oxoglutarate 5-dioxygenase 2
207543_s_at	prolyl 4-hydroxylase, alpha polypeptide I
1554997_a_at	prostaglandin-endoperoxide synthase 2
226452_at	pyruvate dehydrogenase kinase, isozyme 1
226682_at	RAR-related orphan receptor A
219888_at	sperm associated antigen 4
212353_at	sulfatase 1
202796_at	synaptopodin
228483_s_at	TAF9B RNA polymerase II, TATA box binding protein (TBP)-associated factor, 31kDa
201010_s_at	thioredoxin interacting protein
200822_x_at	TPI1 pseudogene; triosephosphate isomerase 1
219410_at	transmembrane protein 45A
218149_s_at	zinc finger protein 395

5.2.3 Disease vs. healthy fascia

GO biological process and molecular function overrepresentations are illustrated in Figure 42. The genes enriched in biological processes range from cellular processes involving biological regulation. More than half of 1000 gene lists are involved in binding or protein binding. Other molecular functions involve catalytic activity and regulator activities.

From PPLR analyses, 119 transcripts were significantly differentially expressed unique to N21 *vs.* F21 analysis combining data from all patients. The GO overrepresentations from these gene lists show upregulation in a number of intermediates involved in oxidative phosphorylation, ribosomal subunits, albeit direct glycolysis and carbohydrate metabolism intermediates are not clear from this list. A large number of downregulated genes have been enriched in pathways associated in Focal cell adhesion pathway, ECM-receptor (Figure 46) and TGF beta pathway. A small number was also linked with pentose phosphate metabolism. The PPLR method generated a greater number of significantly expressed gene lists.

The PPLR values compare N21 from F21 samples in individual patients which allow determination of variability between patients. 36 probe sets are consistent across all patients. The gene names associated with the 36 Probe IDs common across the three patients are shown in the Venn diagram (Figure 41) and the 25 corresponding genes are listed in Table 9. The trends observed are mostly similar across the three patients with varying level of significant expression. Many of the differentially expressed transcripts in N21 *vs.* F21 are also observed in N1 *vs.* N21 but these show an opposite effect (i.e. downregulation upon perturbation).

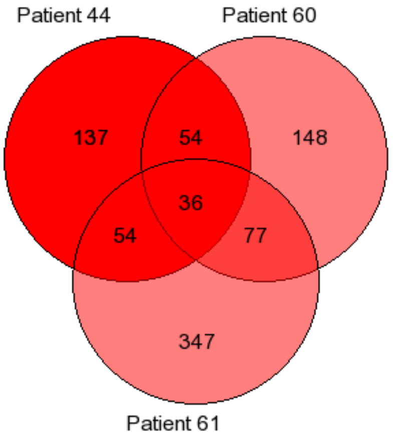
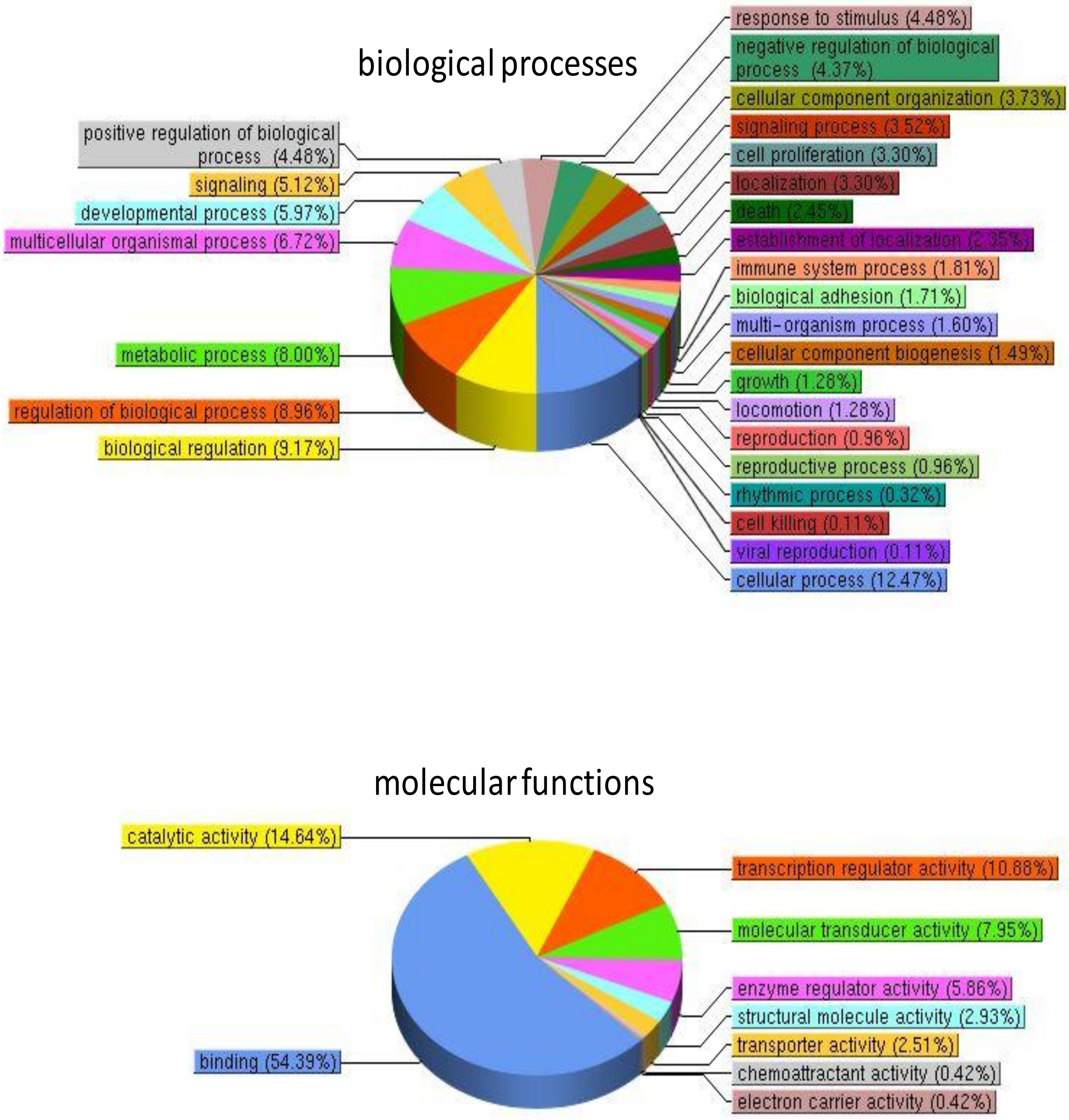


Figure 41 Venn diagram showing comparison across patient samples where DD nodule was compared to healthy fascia.

Table 9 The 25 genes (36 probes from Venn diagram) significant common in all 3 samples where DD nodule was compare with fascia.

Affymetrix Probe ID	Gene Name
209555_s_at	CD36 molecule (thrombospondin receptor)
225496_s_at	synaptotagmin-like 2
219148_at	PDZ binding kinase
1568618_a_at	UDP-N-acetyl-alpha-D-galactosamine
209396_s_at	chitinase 3-like 1 (cartilage glycoprotein-39)
203764_at	discs, large (Drosophila) homolog-associated protein 5
219685_at	transmembrane protein 35
209047_at	aquaporin 1 (Colton blood group)
201843_s_at	EGF-containing fibulin-like extracellular matrix protein 1
201291_s_at	topoisomerase (DNA) II alpha 170kDa
228407_at	signal peptide, CUB domain, EGF-like 3
209806_at	histone cluster 1, H2bk
211959_at	insulin-like growth factor binding protein 5
224941_at	PAPPA antisense RNA (non-protein coding)
222608_s_at	anillin, actin binding protein
201539_s_at	four and a half LIM domains 1
221593_s_at	ribosomal protein L31 pseudogene 49
225647_s_at	cathepsin C
205923_at	reelin
227345_at	tumor necrosis factor receptor superfamily, member 10d, decoy with truncated death domain
209120_at	nuclear receptor subfamily 2, group F, member 2
202503_s_at	KIAA0101
208792_s_at	clusterin
229400_at	homeobox D10
218009_s_at	protein regulator of cytokinesis 1



biological processes

molecular functions

Figure 42 Pie charts showing gene ontology overrepresentations from 1000 up and 1000 downregulated genes illustrated as a percentage of the 2000 probesets involved in nodules 21 % compared with fascia 21%.

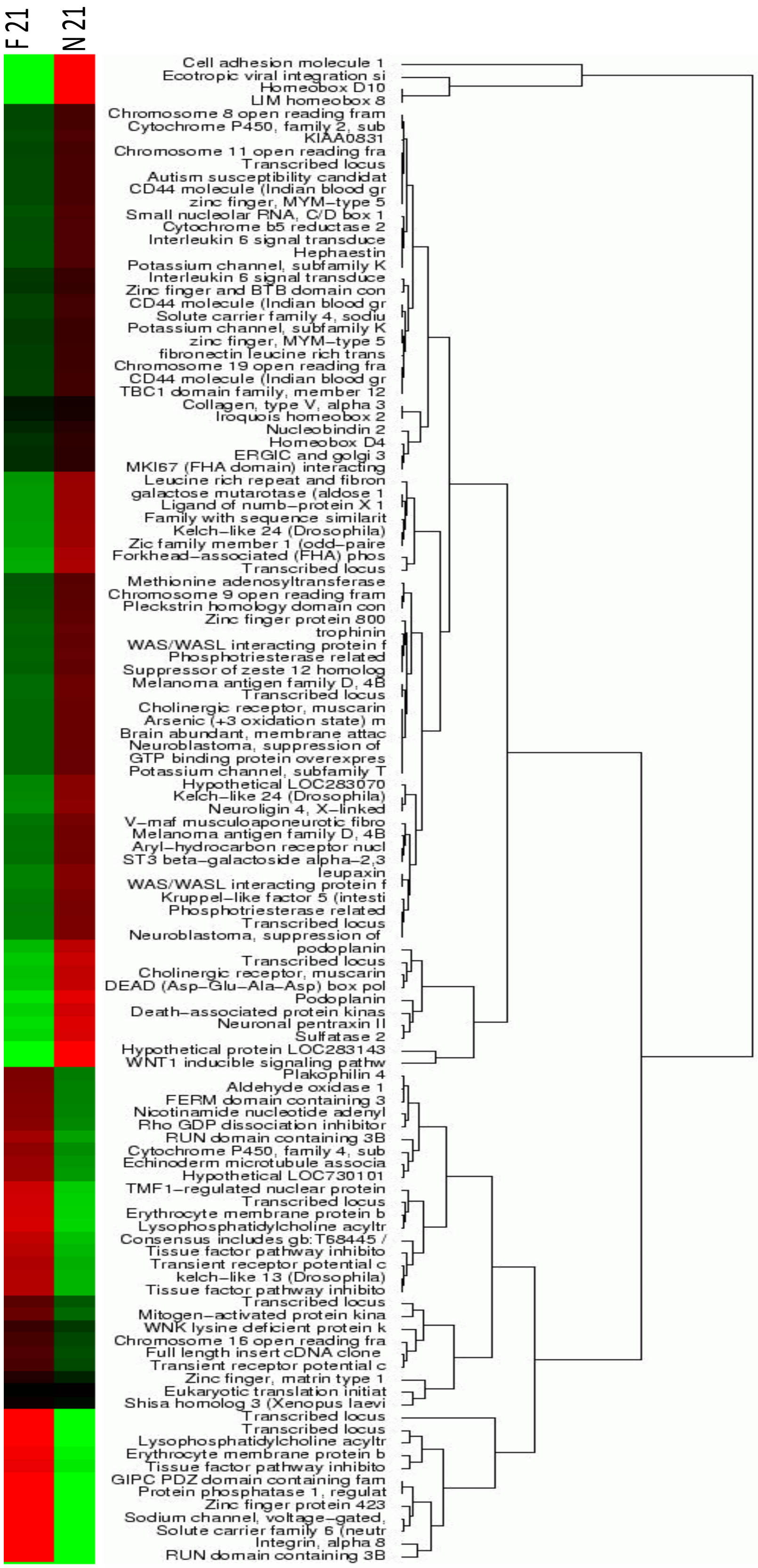


Figure 43 Probelist (119) unique to N21 *vs.* F21 where PPLR values ≥0.9 and ≤0.1

Table 10 KEGG pathways retrieved from 119 probe lists unique to N21 *vs*. F21.

KEGG Pathway	**Gene Name**
Cell adhesion molecules (CAMs)	Cell adhesion molecule 1 Integrin, alpha 8 Neuroligin 4, X-linked
ECM-receptor interaction	CD44 molecule Collagen, type V, alpha 3 Integrin, alpha 8
Neurotrophin signaling pathway	Mitogen-activated protein kinase 13 Rho GDP dissociation inhibitor (GDI) beta
Nicotinate and nicotinamide metabolism	Aldehyde oxidase 1 Nicotinamide nucleotide adenylyltransferase 2
Shigellosis	CD44 molecule Mitogen-activated protein kinase 13
Bladder cancer	Death-associated protein kinase 1
Cysteine and methionine metabolism	Methionine adenosyltransferase II, alpha
Ether lipid metabolism	Lysophosphatidylcholine acyltransferase 2
Glycosphingolipid biosynthesis - ganglio series	ST3 beta-galactoside alpha-2,3-sialyltransferase 5
Selenoamino acid metabolism	Methionine adenosyltransferase II, alpha
Tryptophan metabolism	Aldehyde oxidase 1
Tyrosine metabolism	Aldehyde oxidase 1
Valine, leucine and isoleucine degradation	Aldehyde oxidase 1
Vitamin B6 metabolism	Aldehyde oxidase 1

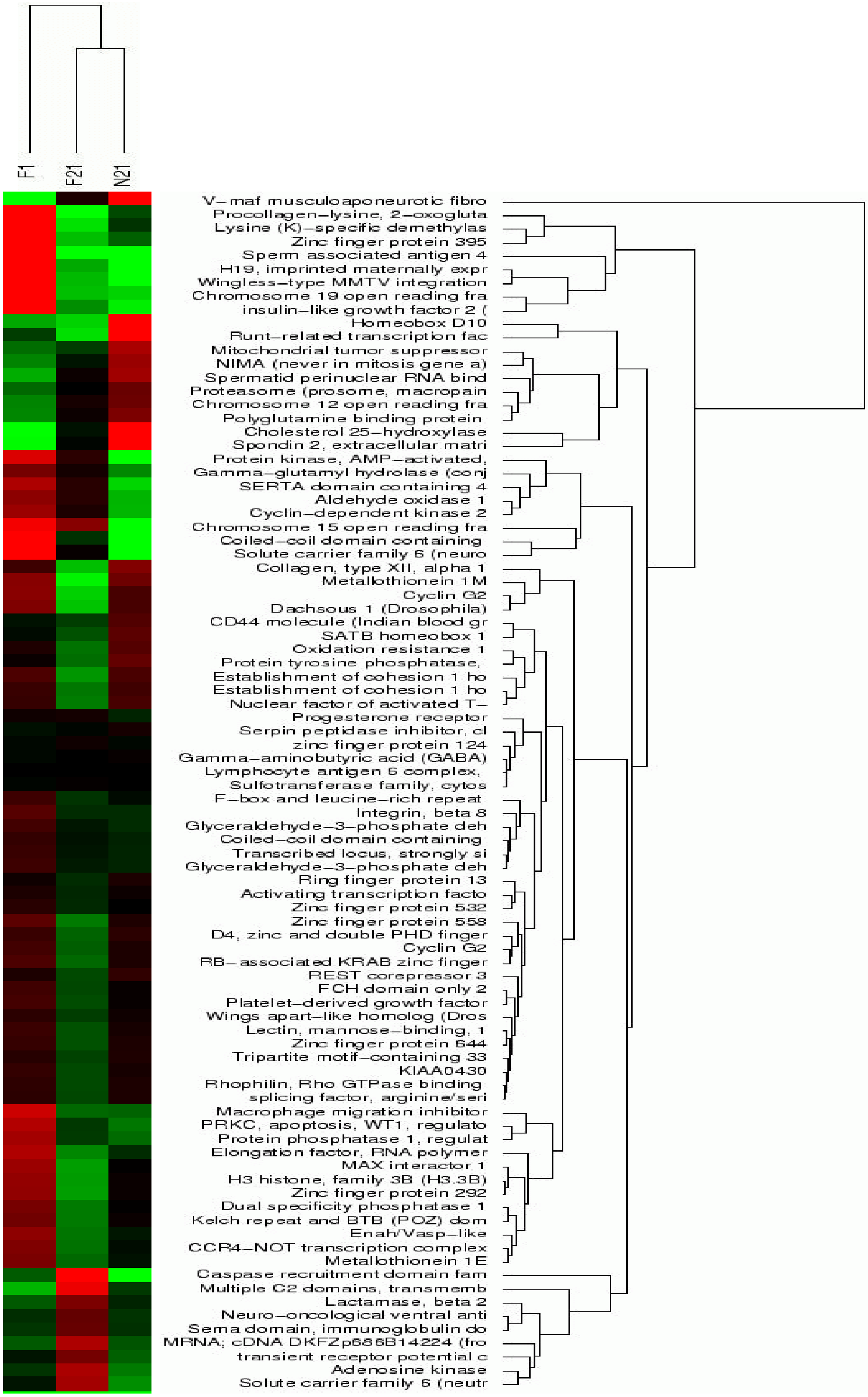


Figure 44 Heatmap displaying expression ratios from 88 probelist overlapping from the pairwise analysis on N21 *vs.* F1 and F1 *vs.* F21. Few F1 and N21 genes are correlated.

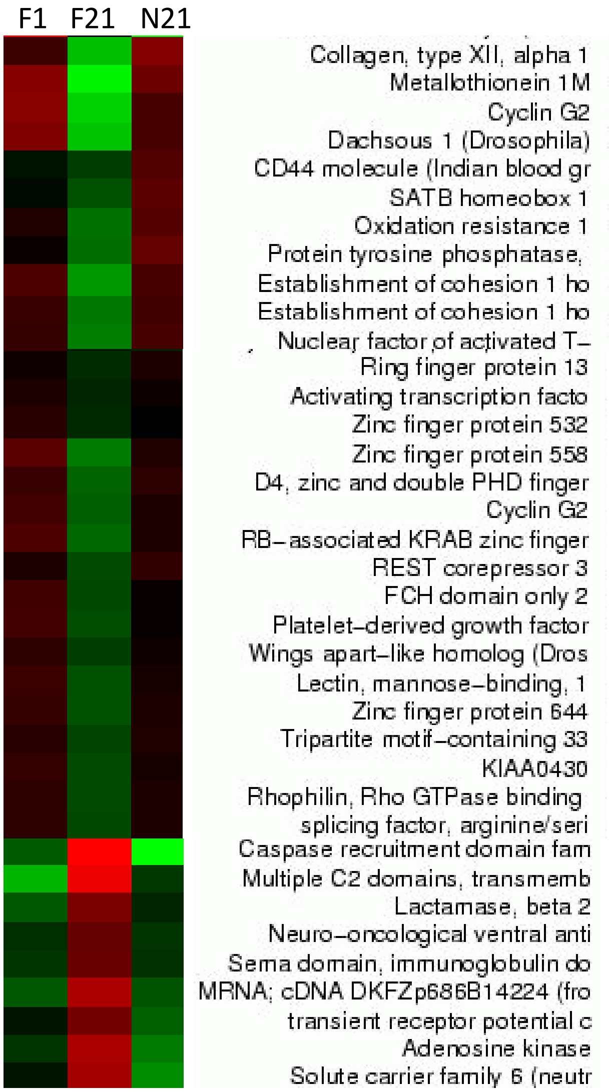


Figure 45 Heatmap displaying expression ratios from the most significant transcripts overlapping from the pairwise analyses in N21 *vs.* F1 and F1 *vs.* F21. Red = upregulated and green = downregulated.

Table 11 KEGG pathways retrieved from 88 overlapping probe lists common to N21 *vs.* F21 and F1 *vs.* F21.

KEGG Pathway	**Gene Name**
MAPK signaling pathway	Activating transcription factor 4 Dual specificity phosphatase 1 Nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 4 Platelet-derived growth factor receptor, alpha polypeptide
Prostate cancer	Activating transcription factor 4 Cyclin-dependent kinase 2 Platelet-derived growth factor receptor, alpha polypeptide
ECM-receptor interaction	CD44 molecule (Indian blood group) Integrin, beta 8
Hypertrophic cardiomyopathy (HCM)	Integrin, beta 8 Protein kinase, AMP-activated, alpha 2 catalytic subunit
p53 signaling pathway	Cyclin G2 Cyclin-dependent kinase 2
Tyrosine metabolism	Aldehyde oxidase 1 Macrophage migration inhibitory factor (glycosylation-inhibiting factor)
Folate biosynthesis	Gamma-glutamyl hydrolase
Nicotinate and nicotinamide metabolism	Aldehyde oxidase 1
Phenylalanine metabolism	Macrophage migration inhibitory factor (glycosylation-inhibiting factor)
Primary bile acid biosynthesis	Cholesterol 25-hydroxylase
Regulation of autophagy	Protein kinase, AMP-activated, alpha 2 catalytic subunit
Vitamin B6 metabolism	Aldehyde oxidase 1

Table 12 KEGG pathways retrieved from 113 problelists unique to N1 *vs.* N21.

KEGG Pathway	**Gene Name**
Lysosome	Adaptor-related protein complex 1, sigma 2 subunit Arylsulfatase G Cathepsin K Tripeptidyl peptidase I
Toll-like receptor signaling pathway	Cathepsin K Mitogen-activated protein kinase kinase 2 Toll-like receptor 1
Biosynthesis of unsaturated fatty acids	Fatty acid desaturase 1 Stearoyl-CoA desaturase (delta-9-desaturase)
Fatty acid metabolism	Acyl-CoA synthetase long-chain family member 3 Dodecenoyl-Coenzyme A delta isomerase
Peroxisome	Acyl-CoA synthetase long-chain family member 3 Isocitrate dehydrogenase 1 (NADP+), soluble
PPAR signaling pathway	Acyl-CoA synthetase long-chain family member 3 Stearoyl-CoA desaturase (delta-9-desaturase)
Terpenoid backbone biosynthesis	Farnesyl diphosphate synthase Isopentenyl-diphosphate delta isomerase 1
Citrate cycle (TCA cycle)	Isocitrate dehydrogenase 1 (NADP+), soluble
Glycine, serine and threonine metabolism	Serine racemase
Porphyrin and chlorophyll metabolism	Biliverdin reductase A
RNA polymerase	Polymerase (RNA) III (DNA directed) polypeptide B
Thyroid cancer	Mitogen-activated protein kinase kinase 2

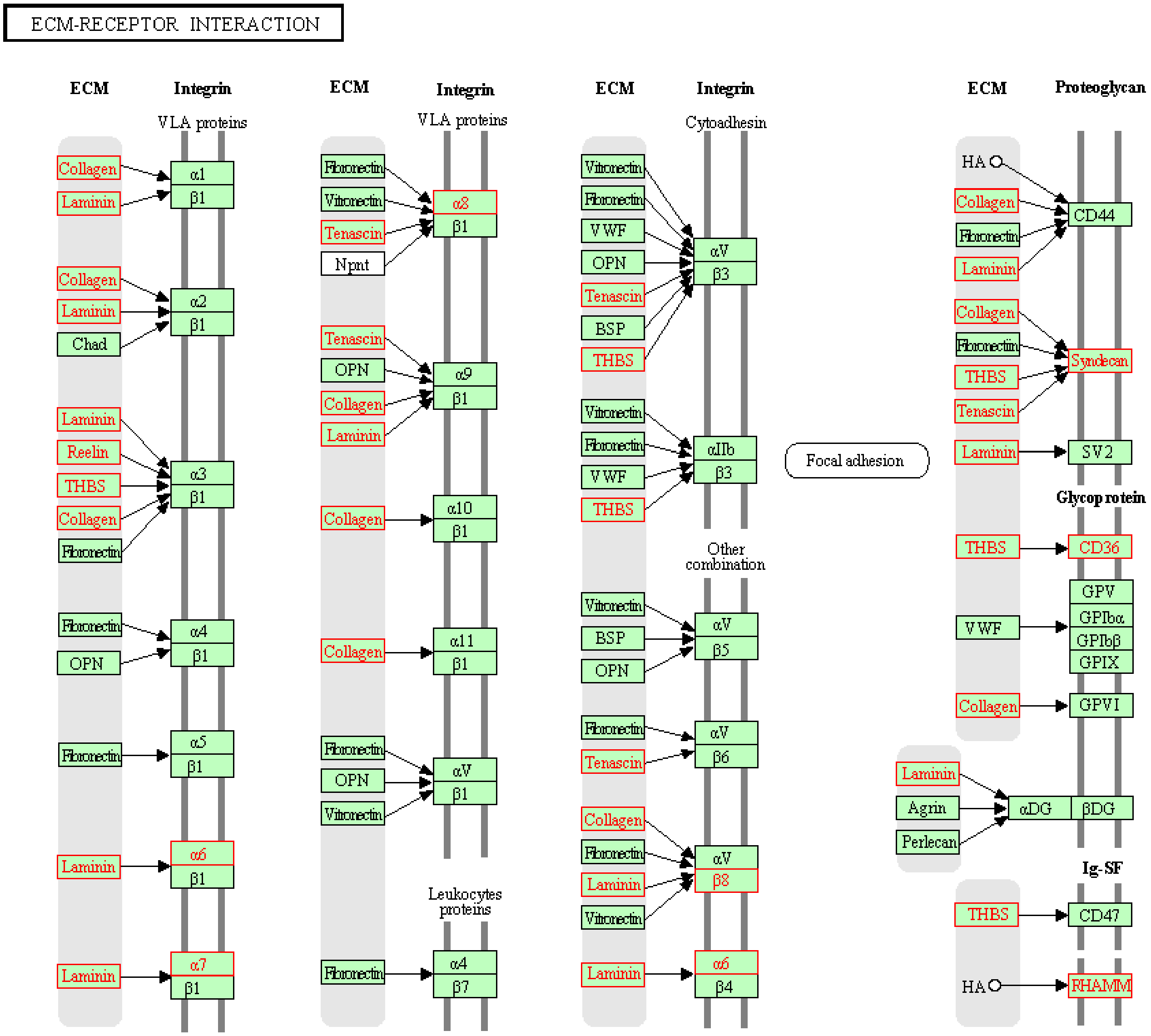
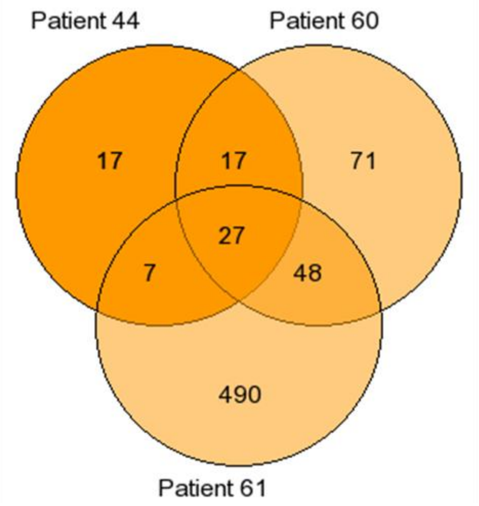


Figure 46 ECM-receptor interaction diagram schema obtained from KEGG database. These genes are significantly differentially expressed in N21 *vs.*F21.

5.2.4 Disease in hypoxic stress

From PPLR analyses, 113 transcripts were significantly differentially expressed unique to N1 *vs.* N21 analysis combining data from all patients. The GO overrepresentations from these genelists show downregulation in a number of intermediates involved in oxidative phosphorylation, ribosomal subunits, glycolysis, citric acid cycle, carbohydrate metabolism and fatty acid metabolism intermediates. A large number of upregulated genes have been enriched in pathways associated in Focal cell adhesion pathway, ECM-receptor and TGF beta pathway. A small number was also linked with pentose phosphate metabolism. The PPLR method generated a greater number of significantly expressed gene lists. Table 13 and 14 show a section of the clusters where inducing hypoxia in nodules has up regulating effect, particularly in patient 61 and down regulating effect from the three patients combined. The Venn diagram in Figure 47 shows a comparison across patient where perturbed nodule was compared to unperturbed nodule (N1 *vs.* N21). Patient 61's nodules demonstrated the maximum number of DE genes.

Figure 47 Venn diagram showing comparison across patient samples where perturbed nodules N1 was compared to unperturbed disease nodules N21.



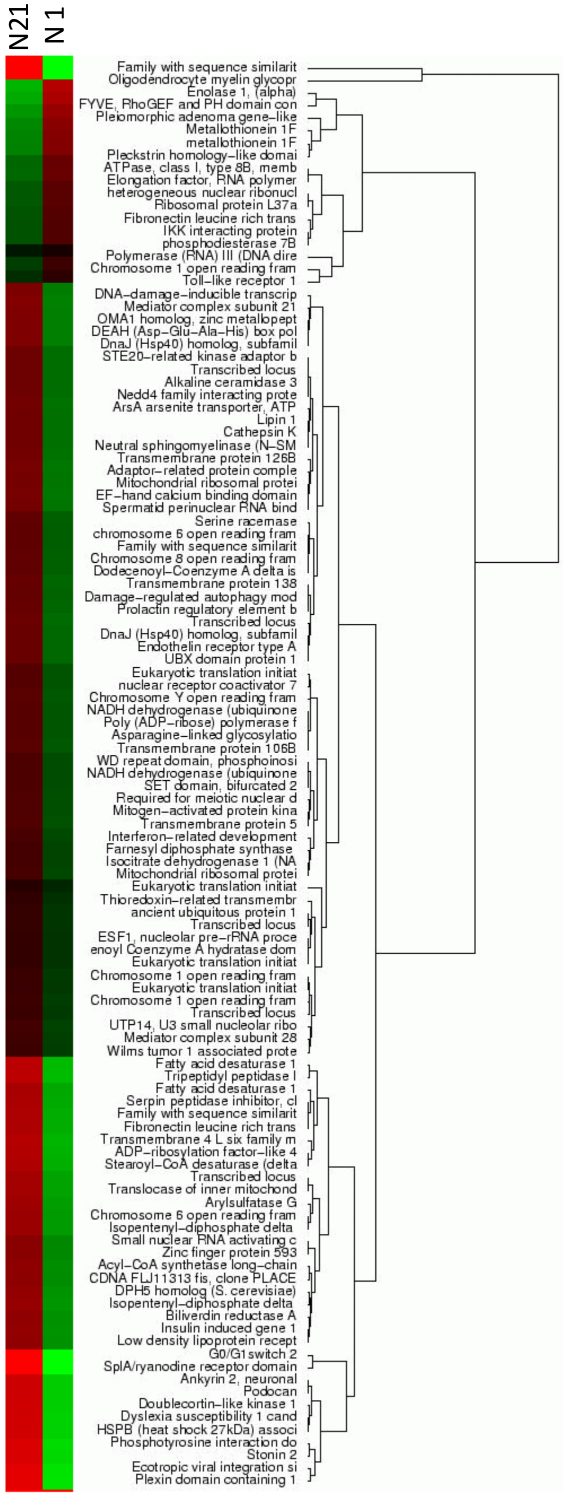


Figure 48 Heatmap displaying 113 DE genes unique to N1 *vs.* N21**.**

Table 13 A cluster from the most significantly upregulated genes in nodules in 1% in descending order for patient 61. The PPLR values for each pairwise analysis and each patient is given. Coloured PPLRs values were filtered through applied thresholds >0.9 and <0.1 and cut off at 1.5 threshold calculated by variance /mean of standard error. Upregulated genes in nodule in 21% were downregulated upon hypoxic induction. Blue PPLR value = F1 *vs.*F21, red PPLR value = N21 *vs.*F21 and orange PPLR value = N1 *vs.* N21.

Fas 1- Fas 21			Nod 21 - Fas 21			Nod 1 - Nod 21			Gene Title
44	60	61	44	60	61	44	60	61	
0.718574	0.887815	0.637286	0.00822	0.9871	2.256E-07	0.709087	0.645745	1	anillin, actin binding protein
0.969455	0.99376	0.892637	0.99301	0.854193	0.0057309	0.321391	0.00579	1	cartilage oligomeric matrix protein
0.968656	0.99793	0.98547	0.06015	0.99202	0.54178045	0.99317	0.99851	0.99999	keratin 19
0.702811	0.538372	0.847788	0.827663	0.781848	0.0301109	0.504834	0.894362	0.99999	keratin associated protein 1-5
0.716923	0.861019	0.579127	0.114924	0.99327	2.589E-05	0.486039	0.603754	0.99997	ribonucleotide reductase M2
0.988427	0.99531	0.955704	0.476238	0.694773	0.17141395	0.97107	0.97624	0.99996	phosphoglycerate kinase 1
0.933711	0.97894	0.795848	0.544678	0.94342	0.0976347	0.804438	0.819971	0.99995	enolase 1, (alpha)
0.483066	0.97582	0.202194	0.829364	0.892457	0.21871269	0.544958	0.602791	0.99995	microtubule-associated protein 7
0.766935	0.98115	0.243967	0.600533	0.89107	0.19367931	0.790868	0.251839	0.99984	tenascin C
0.841894	0.99288	0.765051	0.875038	0.740007	0.0083822	0.762495	0.704613	0.99971	ADAM metallopeptidase with thrombospondin type 1 motif, 1
0.740832	0.93558	0.370647	0.03545	0.96865	0.0007852	0.586164	0.537268	0.99966	topoisomerase (DNA) II alpha 170kDa
0.735155	0.889697	0.447871	0.06701	0.97229	0.0014163	0.514132	0.528171	0.9993	protein regulator of cytokinesis 1
0.920004	0.86878	0.963535	0.826189	0.0153	0.22709253	0.854154	0.98938	0.99885	Hyaluronan synthase 2
0.961601	0.892888	0.221875	0.874232	0.22141	0.0037772	0.492348	0.55322	0.99875	TIMP metallopeptidase inhibitor 3
0.9797	0.98946	0.727395	0.676385	0.813496	0.12474992	0.85177	0.840167	0.99863	glucose-6-phosphate isomerase
0.998009	1	0.99986	0.89911	0.98869	0.67676494	0.96854	0.99255	0.99843	procollagen-lysine, 2-oxoglutarate 5-dioxygenase 2
0.649619	0.582144	0.425659	0.06597	0.464525	0.54276319	0.799399	0.796252	0.99838	lymphocyte antigen 6 complex, locus K
0.634552	0.562206	0.815759	0.02089	0.95202	0.0001784	0.473758	0.588634	0.998	PDZ binding kinase
0.953371	0.99866	0.668381	0.97358	0.857032	0.0585368	0.514532	0.65614	0.99777	collagen, type V, alpha 1
0.93749	0.97806	0.235076	0.206388	0.93781	0.0471915	0.294236	0.456295	0.99698	follistatin
0.687632	0.894347	0.039551	0.828983	0.747221	0.0750765	0.435088	0.587703	0.9966	protein disulfide isomerase family A, member 6
0.734488	0.99987	0.084193	0.96719	0.99904	0.20530817	0.336546	0.117982	0.99644	collagen, type XII, alpha 1
0.873885	0.838389	0.127601	0.011	0.99767	0.0051529	0.866715	0.521115	0.99615	four and a half LIM domains 1
0.742307	0.97551	0.796964	0.861947	0.770534	0.18283947	0.522767	0.706115	0.99592	lysyl oxidase-like 2
0.71332	0.96421	0.90394	0.475086	0.87881	0.091007	0.676519	0.708455	0.99568	cysteine-rich, angiogenic inducer, 61
0.753956	0.93508	0.069555	0.595971	0.871871	0.17986166	0.50156	0.462219	0.99563	prolyl 4-hydroxylase, beta polypeptide
0.810472	0.96983	0.913577	0.572733	0.90831	0.11940655	0.636146	0.709332	0.99546	cysteine-rich, angiogenic inducer, 61
0.736274	0.97777	0.428559	0.9254	0.629151	0.0039913	0.487503	0.508989	0.9954	lysyl oxidase
0.996448	0.99519	0.995476	0.398058	0.489721	0.39197187	0.99765	0.99508	0.99515	pyruvate dehydrogenase kinase, isozyme 1
0.986508	0.94687	0.043057	0.99959	0.887674	0.0134854	0.809172	0.04709	0.99426	aggrecan
0.892211	0.91928	0.886742	0.511751	0.675846	0.26925903	0.808951	0.755733	0.99399	phosphofructokinase, platelet
0.983147	0.97257	0.974628	0.627699	0.67941	0.46854296	0.9077	0.96069	0.9927	triosephosphate isomerase 1
0.835124	0.96378	0.147373	0.700977	0.833064	0.16435574	0.579176	0.478601	0.9927	prolyl 4-hydroxylase, beta polypeptide
0.713036	0.479589	0.859703	0.133181	0.776016	0.0269158	0.485595	0.79558	0.9925	pituitary tumor-transforming 1
0.999576	0.99456	0.997872	0.97057	0.150293	0.10488757	0.93717	0.99418	0.99248	adenylate kinase 3-like 1
0.597714	0.836998	0.516529	0.54815	0.90954	0.0972981	0.289431	0.321367	0.99243	solute carrier family 7
0.999333	0.99709	0.999809	0.419378	0.662824	0.7528722	0.99841	0.99941	0.9924	BCL2/adenovirus E1B 19kDa interacting protein 3

Table 14 A cluster from the most significantly upregulated genes in nodules in 1% in descending order for patient 61. The PPLR values for each pairwise analysis and each patient is given. Coloured PPLRs values were filtered through applied thresholds >0.9 and <0.1 and cut off at 1.5 threshold calculated by variance /mean of standard error. Upregulated genes in nodule in 21% were downregulated upon hypoxic induction. Blue PPLR value for F1 *vs.* F21, red PPLR value for N21 *vs.*F21 and orange PPLR value N1 *vs.* N21. A cluster of differentially expressed genes. Blue PPLR value = F1 *vs.* F21, red PPLR value = N21 *vs.* F21 and orange PPLR value = N1 *vs.*N21.

F1 v F21	N21 v F21	N1 v N21	Gene Title
0.214384	0.99999	8.1E-06	stonin 2
2.01E-05	0.99999	1.3E-11	chitinase 3-like 2
0.551767	0.99998	1.9E-07	lysophospholipase-like 1
0.500286	0.99997	0.115965	WNT1 inducible signaling pathway protein 2
0.948089	0.999948	6.7E-05	ribonuclease P RNA component H1
0.911325	0.99992	0.00015	ribosomal protein L31
0.924233	0.99983	2.7E-06	cytochrome c oxidase subunit VIIc
0.742703	0.99982	0.00108	cytochrome P450, family 24, subfamily A, polypeptide 1
0.017217	0.99981	0.00088	tumor necrosis factor, alpha-induced protein 6
0.086406	0.9998	0.01343	GTP binding protein overexpressed in skeletal muscle
0.587924	0.99969	7.3E-07	lymphocyte antigen 96
0.907669	0.99956	0.01174	microRNA 21
0.868802	0.99954	5E-05	prefoldin subunit 4
0.763007	0.99947	0.468237	runt-related transcription factor 2
0.585672	0.99947	0.00593	cystatin A (stefin A)
0.685392	0.99933	0.160908	neuritin 1
0.027009	0.99916	1E-06	interleukin 1 receptor, type II
0.643232	0.99913	0.00286	cartilage intermediate layer protein
0.411317	0.99892	0.00016	ATP-binding cassette, sub-family B (MDR/TAP), member 5
0.811695	0.99876	7.6E-05	hypothetical LOC388789
0.311654	0.9986	0.00315	basic helix-loop-helix family, member e22
0.619723	0.9985	0.401579	glucosaminyl (N-acetyl) transferase 1, core 2
0.527067	0.99842	0.0001	hypothetical protein MGC5566
0.093916	0.99834	0.07707	dipeptidyl-peptidase 4
0.724518	0.99813	0.00022	TatD DNase domain containing 1
0.462208	0.99802	0.0014	glycoprotein, alpha-galactosyltransferase 1 pseudogene
0.645576	0.99796	0.00041	alkB, alkylation repair homolog 7 (E. coli)
0.727138	0.99794	2.4E-05	growth arrest-specific 5 (non-protein coding)
0.626711	0.99752	0.00605	parathyroid hormone-like hormone
0.467238	0.99602	0.00671	chitinase 3-like 1 (cartilage glycoprotein-39)
0.965408	0.99589	0.00098	integrin beta 3 binding protein (beta3-endonexin)
0.013607	0.99563	4E-05	plexin domain containing 1
0.777867	0.99493	0.00041	methylmalonyl CoA epimerase
0.736525	0.99491	0.00175	synaptonemal complex central element protein 1-like
0.902797	0.99437	0.00055	LSM domain containing 1
0.847522	0.99423	0.01113	ribosomal L24 domain containing 1
0.000479	0.99413	0.101812	clusterin

5.3 Discussion

5.3.1 Principal Findings

Transcriptome profiling has enabled the investigation of the expression levels of thousands of genes associated with DD formation simultaneously. A quantitative comparison was made between 1) mRNA expression levels in normal oxygenated and hypoxic transverse palmar fascia in search for novel hypoxia-induced genes to investigate whether this change was mimicked in disease state scenario. 2) The transcriptional response to hypoxia was also investigated in nodules from the same patients. 3) mRNA expression levels in normal oxygenated disease with healthy fascia. The study and funds permitted for 12 Affymetrix microarrays, 4 samples cultured at two oxygen levels from 3 patients.

Landmark genes identified in previous studies [9, 19, 155] to be involved in DD were also validated in this study. In addition, this study has made possible identification of genes not only displaying unique characteristics of disease cells compared with normal palmar fascia but also transcriptionally hypoxia-activated genes in both healthy and disease cells. Several common effects of hypoxia were seen in the DD nodules and fascia, such as an increase in glycolytic metabolism; however, the response to hypoxia varies greatly between the individual patients. We discuss these key findings below.

A meta-analysis of top1000 genes with PPLR values > 0.8 and lowest 1000 genes with PPLR value < 0.2 for each pair in a given analysis were entered into KEGG pathway database to identify molecular pathways (metabolic and signaling) that were statistically significantly enriched (from uploaded gene lists). Significantly up regulated gene lists produced enrichment in 20 molecular pathways enriched for F1 *vs.* F21. Significantly down regulated gene lists produced enrichment in 33 molecular pathways enriched for F1 *vs.* F21 (Appendix; Table J). Gene enrichment analysis was then performed using top 1000 and bottom 1000 genes significantly dysregulated in gene lists from pairwise comparison of N21 *vs*. F21 and then repeated with gene lists for the combination pair N1 *vs.* N21. An attempt to connect the common pathways into a hypothetical common molecular network for each combination/pair and compare with pathways identified in Chapter 4 from metabolomics data is given Chapter 6 currently in progress.

Differential alterations in gene expression in N21 (compared with F21) were observed in pathways associated with focal cell adhesion, apoptosis, and inflammation.

Among the dysregulated transcripts, marked enrichment was observed in those directly involved in developmental processes including cell growth, proliferation, differentiation, regulation of cell death, biological cell adhesion, localisation, extracellular matrix-receptor interaction, and cell communication. Neuritin and Amphiphysin were upregulated in N21. Neuritin is a growth-promoting protein and may be involved in tumorigenesis [156]. Cell adhesion molecules were up regulated in N21. Major down regulated groups of genes were involved in focal cell adhesion pathway. Focal adhesions connect the ECM to actin filaments of the cell. This cell-to-ECM adhesion is regulated by specific cell surface cellular adhesion molecules (CAM) known as integrins. Integrins are cell surface proteins that bind cells to ECM structures, such as fibronectin and laminin, and also to integrin proteins on the surface of other cells [157]. Fibronectins bind to ECM macromolecules and facilitate their binding to transmembrane integrins. The attachment of fibronectin to the extracellular domain initiates intracellular signaling pathways as well as association with the cellular cytoskeleton via a set of adaptor molecules such as actin [157]. CD36 was significantly down regulated in N21. Collagen, type VI, alpha 6, was found upregulated in N21, but downregulated in N1. No change was observed in F1. Vascular cell adhesion molecule 1, collagen, type XII, alpha 1 and collagen, type XIV, alpha 1 were upregulated in disease but downregulated upon perturbation in N1.

However, from Table 11 and Figure 44(a) and 44(b), it can be observed that relatively few transcripts identified as significantly dysregulated in F1 *vs.* F21 analysis were present in N21 *vs.* F21 analysis. From these analyses it is tempting to falsify the first hypothesis that difference in disease and healthy cells maybe akin to the differences in healthy cells in normoxia and hypoxia as only a very small number of significant DE transcripts coincide in F1 and N21 candidate lists. In Chapter 6 these significant candidates (i.e. metabolites and transcripts) from Chapter 4 and 5 are integrated to visualize relationships using network analysis (e.g. candidate metabolites and transcripts in F1 *vs.*F21 using IPA and these networks are mapped on the networks from significant molecules in N21 *vs.* F21 analysis.

These results imply that DD is associated with a stimulation of collagen gene expression at the transcriptional and translational levels together with an increase in the rate of collagenolytic activity in up regulated inhibitors such as ADAM metallopeptidase domains (ADAM12 and ADAM19). A disintegrin-like and metalloproteinase with thrombospondin type 1 motif, 2 (ADAMTS2) and TIMP were notably differentially expressed in DD nodule tissues compared with external controls only.

A list of few DE genes unique to disease

CD44 molecule, encodes a cell-surface glycoprotein involved in cell-cell interactions, cell adhesion and migration. It is a receptor for hyaluronic acid (HA) and can also interact with other ligands, such as osteopontin, collagens, and matrix metalloproteinases (MMPs) [158]. This protein participates in a wide variety of cellular functions including lymphocyte activation, recirculation and homing, hematopoiesis, and tumour metastasis. Alternative splicing is the basis for the structural and functional diversity of this protein, and may be related to tumour metastasis. Collagen, type V, alpha 3, encodes an alpha chain for one of the low abundance fibrillar collagens. Fibrillar collagen molecules are trimers that can be composed of one or more types of alpha chains. Type V collagen is found in tissues containing type I collagen and appears to regulate the assembly of heterotypic fibers composed of both type I and type V collagen [159]. This gene product is closely related to type XI collagen and it is possible that the collagen chains of types V and XI constitute a single collagen type with tissue-specific chain combinations. Mutations in this gene are thought to be responsible for the symptoms of a subset of patients with Ehlers-Danlos syndrome type III [160].

Mitogen-activated protein kinase 13; encodes proteins which are members of the MAP kinase family. MAP kinases act as an integration point for multiple biochemical signals, and are involved in a wide variety of cellular processes such as proliferation, differentiation, transcription regulation and development [161]. This kinase is closely related to p38 MAP kinase, both of which can be activated by proinflammatory cytokines and cellular stress. MAP kinases 3, and 6 can phosphorylate and activate this kinase. Nicotinamide nucleotide adenylyltransferase 2; this gene product belongs to the nicotinamide mononucleotide adenylyltransferase (NMNAT) enzyme family, members of which catalyse an essential step in NAD (NADP) biosynthetic pathway. Ganglioside GM3 is known to participate in the induction of cell differentiation, modulation of cell proliferation, maintenance of fibroblast morphology, signal transduction, and integrin-mediated cell adhesion. Aldehyde oxidase produces hydrogen peroxide and, under certain conditions, can catalyze the formation of superoxide. Aldehyde oxidase is a candidate gene for amyotrophic lateral sclerosis.

Some DE genes common to N21 and F1 albeit not always in the same direction.
Activating transcription factor 4; The protein encoded by this gene belongs to a family of DNA-binding proteins that includes the AP-1 family of transcription factors, cAMP-response element binding proteins (CREBs) and CREB-like proteins. These transcription factors share a leucine zipper region that is involved in protein-protein interactions, located C-terminal to a stretch of basic amino acids that functions as a DNA binding domain. Dual specificity phosphatase 1, The expression of DUSP1 gene is induced in human skin fibroblasts by oxidative/heat stress and growth factors. It specifies a protein with structural features similar to members of the non-receptor-type protein-tyrosine phosphatase family, and which has significant amino-acid sequence similarity to a Tyr/Ser-protein phosphatase. DUSP1 may play an important role in the human cellular response to environmental stress as well as in the negative regulation of cellular proliferation.

Nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 4; the product of this gene is a member of the nuclear factors of activated T cells DNA-binding transcription complex. The product of this gene plays a role in the inducible expression of cytokine genes in T cells, especially in the induction of the IL-2 and IL-4. Cyclin-dependent kinase 2; the protein encoded by this gene is a member of the Ser/Thr protein kinase family. It is a catalytic subunit of the cyclin-dependent protein kinase complex, whose activity is restricted to the G1-S phase, and essential for cell cycle G1/S phase transition. This protein associates with and regulated by the regulatory subunits of the complex including cyclin A or E, CDK inhibitor p21Cip1 (CDKN1A) and p27Kip1 (CDKN1B). Integrin, beta 8; in general, integrin complexes mediate cell-cell and cell-extracellular matrix interactions and this complex plays a role in human airway epithelial proliferation. Protein kinase, AMP-activated, alpha 2 catalytic subunit; encodes a protein that is a catalytic subunit of the AMP-activated protein kinase (AMPK). AMPK is an important energy-sensing enzyme that monitors cellular energy status. In response to cellular metabolic stresses, AMPK is activated, and thus phosphorylates and inactivates acetyl-CoA carboxylase (ACC) and beta-hydroxy beta-methylglutaryl-CoA reductase (HMGCR), key enzymes involved in regulating de novo biosynthesis of fatty acid and cholesterol. Cholesterol 25-hydroxylase; an intronless gene that is involved in cholesterol and lipid metabolism.

Some DE genes unique to N1 list

Arylsulfatase G; Sulfatases, such as ARSG, hydrolyze sulfate esters from sulfated steroids, carbohydrates, proteoglycans, and glycolipids. They are involved in hormone biosynthesis, modulation of cell signaling, and degradation of macromolecules. Cathepsin K; The protein encoded by this gene is a lysosomal cysteine proteinase involved in bone remodeling and resorption. Toll-like receptor 1; This family of protein plays a fundamental role in pathogen recognition and activation of innate immunity. TLRs are highly conserved from Drosophila to humans and share structural and functional similarities. They recognize pathogen-associated molecular patterns (PAMPs) that are expressed on infectious agents, and mediate the production of cytokines necessary for the development of effective immunity. Stearoyl-CoA desaturase (delta-9-desaturase); an iron-containing enzyme that catalyzes a rate-limiting step in the synthesis of unsaturated fatty acids. The principal product of SCD is oleic acid, which is formed by desaturation of stearic acid. The ratio of stearic acid to oleic acid has been implicated in the regulation of cell growth and differentiation through effects on cell membrane fluidity and signal transduction.

Within our candidate gene lists across the multiple comparisons a group of hypoxia-responsive genes have been identified. Hypoxia affected the expression of multiple genes in F21, and N21 which were either induced or repressed under these conditions. Some of these lists are novel candidates for hypoxia-driven angiogenesis including vascular endothelial growth factor (VEGF) and matrix metalloproteinases MMPs [162]. The KEGG pathways enriched by these genes include glycolysis and gluconeogenesis, hypertrophy model, wnt signaling pathway, smooth muscle contraction and pentose phosphate pathway.

Interestingly a cluster of these genes were also dysregulated in N1. Tables 10, 11 and 12 display several concordant differentially expressed genes in these two samples i.e. up in fascia 1% and up in nodule 1% following the perturbation effect. In addition to the KEGG pathways identified by genelists F1, gene enrichment was observed in the following pathways: eicosanoid synthesis, prostaglandin synthesis regulation, fatty acid degradation, glycogen metabolism, blood clotting cascade, GPCRDB class A rhodopsin-like pathway.

VEGFA was upregulated in N21 and also in F1. VEGF has been reported to be a major contributor to angiogenesis, increasing the number of capillaries in a given network [163]. Previous in vitro studies clearly demonstrate that VEGF is a potent stimulator of angiogenesis because, in the presence of this growth factor, plated endothelial cells will proliferate and migrate, eventually forming tube structures resembling capillaries.

Up-regulated genes included those for glycolysis, the tricarboxylic acid (TCA) cycle. Ethanol and lactate production was not observed under hypoxic conditions indicating that glucose was not fermented to these compounds via the glycolytic pathway. Hypoxia down-regulated some genes involved in transcription initiation by RNA polymerase II, and also some hypothetical protein coding genes that were previously upregulated in N21.

5.3.2 Strengths and weakness of study

For any system under study, we need to understand the magnitude and diversity of gene expression in the unperturbed state and normal state. It is accepted that, for any particular parameter, physiological "normalcy" is not a strict value but is rather a range of values presented by healthy individuals. It is perhaps due to this reason "normal" fascia gene expression displays similar variability, while this not the case for disease samples. This study has used microarray technology to simultaneously profile healthy cells in normal and perturbed state to understand the gene expression profiles of these healthy cells prior to comparison with disease. Additionally, perturbation effects in disease cultures have implicated several genes to be downregulated which were previously found to be upregulated. The profiling of healthy palmar fascia and in perturbed conditions is a key strength of this study. A key strength of this study is that samples were obtained from previous optimised methods using FT-IR in order to confirm reproducibility. Optimised modeling cell culture system and sample harvesting methodology from same population of cells allowed for comparison with the intracellular metabolome.

Another key strength of this study was the use of Affymetrix oligonucleotide arrays that allows the examination of 47,000 transcripts simultaneously. Previous studies report the use of microarrays with a significantly lower number and we used a chip with 22,500 previously. This same strength can be considered as a weakness as increase in chip probe size also increases noise in the data generated and this may result in an increase in false discoveries as well as true discoveries in subsequent data analysis. For this reason, robust data analysis methods using 3 independent methods with three normalisation methods have been employed to infer unique discoveries. The method also allowed comparisons across fascia and nodule within and across individual patients by generating a range of contrasts matrices from the p-data uploaded (readaffy files). Also contrasts between individual probe sets of uncertainty using PPLR methods were used. The limma methods accounted for false discoveries, while puma methods accounted for probe level uncertainties and yielded a much greater number of DE genes. Regardless of the different methods employed, high noise to signal ratio is still a significant factor in the data set. FDR removed all such noise and false genes resulting in only handful i.e. 20 genes. Occasional occurrence where the adjusted p-values are equal to 1 suggests this is not an error situation but rather an indication that there is no evidence of DE in the data after adjusting for multiple testing. To make objective judgments about the most promising candidates for follow-up studies, a trading off of both p-values and log odd values could be a good criteria but selection of genes solely on multiple correction or one method would a be a criteria too stringent.

While an attempt to profile normal here has been made, $n = 3$ is an insufficient number and only acceptable for a pilot study. To formulate new hypotheses, as previously stated this sample number must increase as the variability within these 3 patients, (different growth stage of disease, individual patient signature, and other traits) may have a significant impact on results generated from this study. Such variability between individuals emphasises the need to pool samples, and to perform additional biological and analytical replicates. The current study could not afford replicates, as the cost for 12 arrays was already £6000 alone. Three patients too small a sample number, large noise to signal ratio within the dataset and high patient to patient variation is evident from PCA. The individual patient signature was greater in some case. e.g. collagen V was up regulated in patients 44 and 60 but not in 61. This biased analysis does not give rise to correct pathways prediction.

The statistical analysis of microarray data represents a significant challenge as the aim is to apply standard statistical approaches to determine gene expression and gene expression alteration significance, thus enabling the extraction of significant biological information from a morass of noise and variability. However, present data analysis methodologies do not adequately deal well with the number of possible combinations. Statisticians are experienced with handling data involving a limited number of variables, but a large number of samples (e.g., the average weight of persons in UK is a problem of a single variable and 62 million samples). Microarrays turn this problem on its head, producing thousands of variables from a small number of samples; the number of samples is significantly less compare to the number of observation (e.g. 1:10000). Hence the reason a number of different methods are employed in this study. In particular the puma method employs the multi-chip modified gamma Model Of Signal and noise-propagation in principal component analysis (NPPCA) method [164] which propagates the expression level uncertainty to improve the results of PCA. Puma-PCA scores plot show clarity in observations than standard PCA due to the scale difference and the purity and compactness of the clusters observed. The large scale on PCA plots generally implies incorrect or unsuitable data preprocessing (large scale does not reveal differences within the data set as high intensity variables could be dominating low intensity variables).

Because of the statistical issues raised by microarray technology, it is necessary that findings be confirmed using independent methodological criteria, preferably with samples from same RNA used to derive the original targets and a larger cohort followed. A limitation of this study is that these alterations have not been confirmed or validated due to time constraints. An ideal, rapid, high through-put, method for confirmation of microarray data would be quantitative (real time) RT-PCR using the TaqMan LightCycler (Roche Diagnostics). Alternatively, Northern blots may provide the benefit of direct quantification. This is important, as significant gene expression changes detected on a microarray may be related to a small fraction of the cells.

Because a microarray experiment may reveal putative changes in the expression of hundreds of genes, it is practically impossible to confirm all of the data. However, it is important to evaluate a reasonable number of genes. That said, confirmational studies may raise other issues such that although microarray experiments might indicate an increase or decrease in the expression of a gene, an independent method might reveal a greater or a lesser change which could then lead to further new plausible questions about the validity of the microarray data.

The hypothesis behind using an exploratory tool of this kind is that clustered genes may be coregulated and therefore may be involved in similar functions. To make sense of these data, and depict gene function the hypotheses that emerge from analysis of systemic expression information must be tested empirically. This requires integration of transcriptomic knowledge with metabolomics and other omics to test hypotheses within the physiological integrity of the DD cell.

5.3.3 Conclusion

This study is the first to show that hypoxia elicits systematic transcriptional responses in DD and healthy palmar fascia cells. Global transcriptome profiling of the cellular response to hypoxia has revealed a multitude of novel mechanisms and functions which may be affected by hypoxia in DD and merit further study. The study has demonstrated a unique approach to the analysis of DD.

Chapter 6

Inferring the metabolic and transcriptional networks specific to Dupuytren’s disease tumours

6.1 Introduction

6.1.1 Metabolic Pathway analysis and Integration with transcriptomics data to aid biomarker discovery in Dupuytren’s Disease

In Chapters 4 and 5 systematic studies investigating the changes the in metabolomes and transcriptomes of DD phenotypes and healthy fascia have been performed. Methods utilising cluster analyses or correlation of gene-expression profiles are common approaches to predict function based on the assumption that genes with similar functions are likely to be co-expressed. This however cannot advance our understanding of the more complex system, consisting of many yet to be identified (unravel) network relationships between these molecules which in turn shall lead to better understanding the functionality of these cells and monitor situations to predict DD progression. For this reason it is important to study the disease as a system and the many interacting components at the global level and also at the subcellular organisational levels.

A full understanding of metabolic networks requires quantitative data about transcript levels, protein levels or enzyme activities, and metabolite levels. This information will then allow construction of probabilistic and mechanistic models to investigate effects of perturbations imposed on the system. For example, investigating any changes in metabolite levels that would contribute to the regulation of gene expression and any change in the levels of transcripts would then have an impact upon the levels of the encoded enzymes and the levels of subsequent metabolites and most likely on the DD metabolic phenotype. The proportion of metabolite-transcript correlations identified from the transcript-metabolic profiling of DD and control fibroblast will provide clues and give possible direction in understanding whether these correlations (between metabolites and transcripts) are due to regulation of gene expression by metabolites, or the metabolites being changed as a consequence of a change in gene expression.

Because our metabolic and transcript profile data originates from the same biological samples in steady state growth rates and these data have been normalised and statistically analysed using stringent methods to select for statistically significant DE transcripts and dysregulated metabolites, it is anticipated that now a realistic determination of the congruence between the levels of certain metabolites, gene transcripts and their cognate protein product(s) can be inferred. However, as noted from the previous chapter the abundance of significantly DE genes alone (and now in combination with metabolites) makes their study a formidable yet challenging task. There is a clear need for methods that would allow sorting these molecules and their families and selecting the most important ones, i.e. prioritising the targets for experimental studies.

In this study, we utilise the statistically significant filtered data sets generated from studies in Chapters 4 and 5 from DD and healthy cells in normal and perturbed conditions to construct 1) metabolic networks i.e. metabolic pathways based on topological analysis using MetPA 2) metabolic and transcriptional networks employing IPA to integrate molecules and construct networks from the focus molecules (network eligible molecules in Ingenuity knowledge base). Construction of networks in MetPA enabled identification of metabolic pathways from overrepresentation analysis, The construction of networks in IPA from the most statistically significant molecules facilitated data visualisation and identified key interactions between the molecules e.g. direct or indirect binding between the molecules residing in the cell including subcelluar locations i.e. nucleus, cytoplasm, plasma membrane and extracellular space (confirmed from literature; PubMed). Any molecules not specific to a subcelluar location for example, some endogenous chemicals (metabolites) were assigned to an unknown location.

Inferences are made from constructed networks where transcriptomics data has been superimposed on metabolite networks as well as mapping of perturbed state onto disease specific networks; e.g. trends of transcription factor binding site enrichment in the promoters of these gene groups led to the identification of regulatory metabolic and signaling pathways that implicate discrete metabolic-transcriptional networks associated with specific molecular subtypes of DD (i.e. nodule).

This analysis indicates that approaches of this type can provide unique insights into the differential regulatory molecular studies associated with DD and connective tissue disorders and will aid in discovery / identification of specific transcriptional networks and pathways as potential targets for tumour subtype-specific therapeutic intervention. The detailed analysis of metabolomic and transcriptomic data using integrative pathway analysis in this study has identified a number of candidate metabolites and pathways that may be of potential importance in the pathophysiology of DD and with further studies may prove to be important as biomarkers of DD. The mapping of perturbed networks (from F1 *vs.* F21 analysis) upon unperturbed networks (from N21 *vs.* F21 analysis) is suggestive that molecules (found from this study) under hypoxic stresses in transverse palmar fascia do not correlated with the statistically significant molecules in disease. Some correlation is present, but not prominent. A major observation from mapping perturbed networks (from N1 *vs.* N21 analysis) upon disease networks (N21 *vs.* F21) was that upon hypoxic perturbation most if not all disease molecules under hypoxia became downregulated. The integration of metabolomics and transcriptomics data in the current study via integrative pathway analysis has also facilitated the contextualization of probable biomarkers related to DD. Future validation work is needed to assess the utility of potential novel biomarkers of DD and to determine their ability to characterise this disease.

6.2 Results

6.2.1 Metabolic Pathway enrichment and topological analysis using MetPA

Pathway analyses in MetPA were conducted to construct metabolic pathways based on topological analysis for each of the candidate metabolite lists from the pairwise analyses generated in Table 6 in Chapter 4. The analyses were conducted through two routes: 1) pathway enrichment analysis supporting both over-representation analysis as well as gene-set enrichment analysis based approaches. The Fishers' exact test was applied to generate a list of significant matched pathways arranged by p-values from pathway enrichment analysis. 2) Pathway topological analysis was performed based on the centrality measures of a metabolite in a given metabolic network. Centrality is a local quantitative measure of the position of a node relative to the other nodes, and is used in analysis to estimate a node's relative importance or role in network organisation [139]. Since metabolic networks are directed graphs, relative betweenness centrality measures were selected to calculate compound importance. The pathway impact is calculated here as the sum of the importance measures of the matched metabolites normalised by the sum of the importance measures of all metabolites in each pathway.

The results from pathway analysis are presented graphically as well as in a detailed table. The graphical output contains three levels of information to view; metabolome view (Figure 49), pathway view (Figure 50), and compound view. The first two were employed in this study, the latter is not as the analyses for this dataset have been performed and shown in Chapters 2 and 4. The metabolome view contains all the matched pathways (the metabolome) arranged by p-values (from pathway enrichment analysis) on *y*-axis, and pathway impact values (from pathway topology analysis) on *x*-axis. The node color is based on its p-value and the node radius is determined based on their pathway impact values. The pathway name is given in the table. From these nodes the metabolic pathway corresponding to the node on the metabolome view is selected. This pathway is essentially a simplified KEGG pathway map showing only chemical compounds. The default node color within the reference metabolome is light blue. The matched nodes display varied heat map colors based on their p-values. The common compound names (KEGG IDs) can be obtained from any node which reveals more detailed information as well as links to curated database. The results are presented below. Selected figures are only provided here.

For each analysis, a table containing detailed results from the pathway analysis was generated. The list of matched compounds from the F1 *vs*. F21 is given below. Since multiple pathways are tested at the same time, the statistical p-values from enrichment analysis are further adjusted for multiple testings. In particular, the **Total** is the total number of compounds in the pathway; the **Hits** is the actually matched number from the user uploaded data; the **Raw p** is the original p value calculated from the enrichment analysis; the

Holm p is the p-value adjusted by Holm-Bonferroni method; the **FDR p** is the p-value adjusted using False Discovery Rate; the **Impact** is the pathway impact value calculated from pathway topology analysis.

Table 15 lists the significant pathways derived from the metabolome pathways summary for F1 *vs*. F21 analysis in Figure 48 where each node represents a pathway. Pantothenate and CoA biosynthesis and beta-Alanine metabolism are the top two pathways identified from the pathway topological analysis (impact 0.29 and 0.30 respectively) and are also significant in the pathway enrichment analysis (4.52E-3 after adjustment of multiple testing). Further checking of the metabolite concentrations by plotting the boxplot and investigating any significant downstream effect on metabolites/compounds in the pathways was explored. In the pathways below (Figure 49) the enriched endogenous compounds are Pantothenic acid (KEGG ID: C00864), Ureidopropionic acid (KEGG ID: C02642 1) and Beta-Alanine (KEGG ID: C00099).

Table 16 lists the significant pathways derived from N1 *vs.* N21 analysis. Alanine, aspartate and glutamate metabolism, glycerolipid metabolism and pantothenate and CoA biosynthesis are the top three pathways from the pathway topological analysis (impact 0.21, 0.20 and 0.18 respectively). Although the pathway enrichment analysis p-values after adjustment for multiple testing are not small, these are more likely to be significant than the Aminoacyl-tRNA biosynthesis pathway which has a relatively small p-value after multiple testing, but lower impact (0.11).

Pantothenate and CoA biosynthesis was the only significant pathway derived from C1 *vs.* C21 analysis with impact 0.18 and FDR corrected p-value = 2.63E-01. For the S1 *vs.* S21 analysis beta-Alanine metabolism (impact 0.30, FDR p-value = 1.62E-01) and Glycine, serine and threonine metabolism pathways (impact 0.20, FDR p-value = 2.85E-01) were significant. For disease N21 *vs.* F21 analysis relatively fewer metabolites were identified and so the phenylalanine metabolism pathways was identified with a relatively low impact (0.12) but p-value = 1.07E-01. Aminoacyl-tRNA biosynthesis pathway had a raw p-value = 1.32E-02 and 0 impact. For the C21 *vs.* F21 analysis Pantothenate and CoA biosynthesis impact 0.18 and raw p-value = 4.41E-02 and FDR 7.59E-01.

Table 15 Result from Pathway Analysis with MetPA for F1 *vs.* F21 matched list of identifiers.

	Total	Expected	Hits	Raw p	Holm p	FDR	Impact
Pantothenate and CoA biosynthesis	27	0.10	3	1.01E-04	8.09E-03	4.52E-03	0.29
beta-Alanine metabolism	28	0.10	3	1.13E-04	8.94E-03	4.52E-03	0.30
Pyrimidine metabolism	60	0.22	2	1.97E-02	1.00E+00	5.24E-01	0.01
Citrate cycle (TCA cycle)	20	0.07	1	7.25E-02	1.00E+00	1.00E+00	0.06
Alanine, aspartate and glutamate metabolism	24	0.09	1	8.64E-02	1.00E+00	1.00E+00	0.02
Propanoate metabolism	35	0.13	1	1.24E-01	1.00E+00	1.00E+00	0.09
Galactose metabolism	41	0.15	1	1.43E-01	1.00E+00	1.00E+00	0.02
Glyoxylate and dicarboxylate metabolism	50	0.19	1	1.72E-01	1.00E+00	1.00E+00	0.00
Starch and sucrose metabolism	50	0.19	1	1.72E-01	1.00E+00	1.00E+00	0.06
Cysteine and methionine metabolism	56	0.21	1	1.91E-01	1.00E+00	1.00E+00	0.01

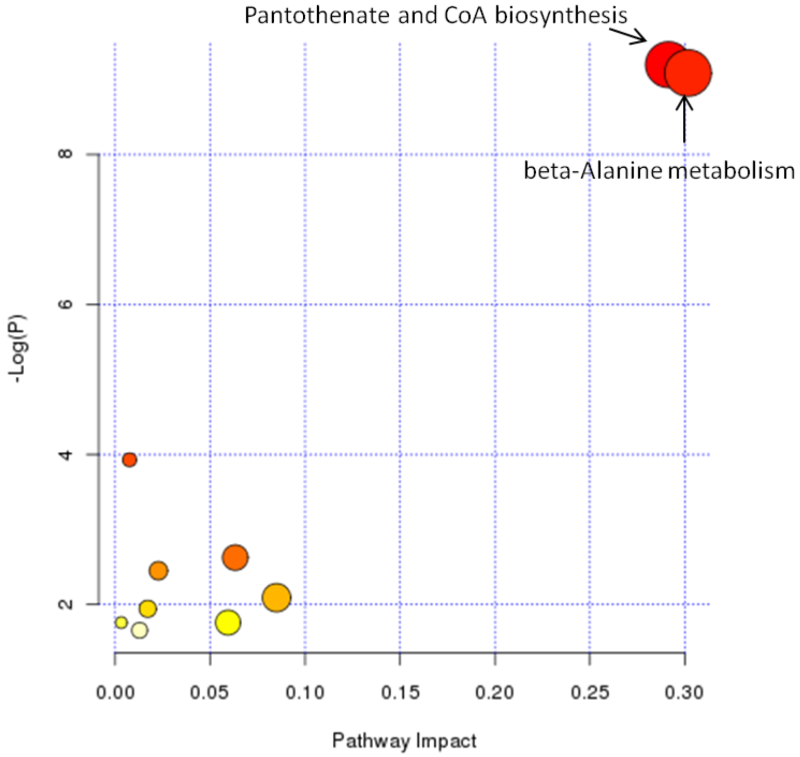


Figure 49 Summary of Pathway Analysis from F1 *vs.* F21 metabolite list with MetPA. Topological pathway analysis in metabolome view.

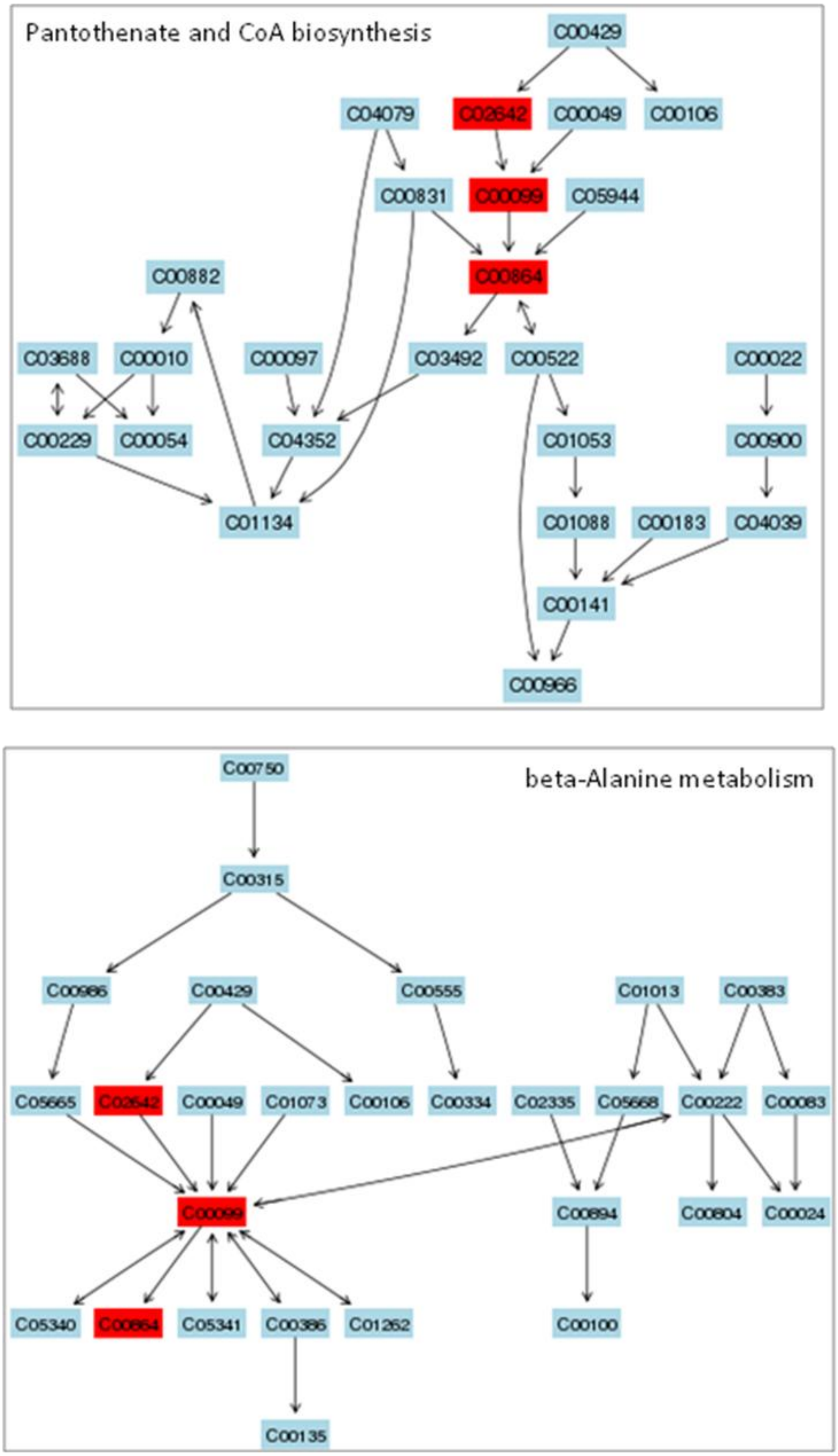
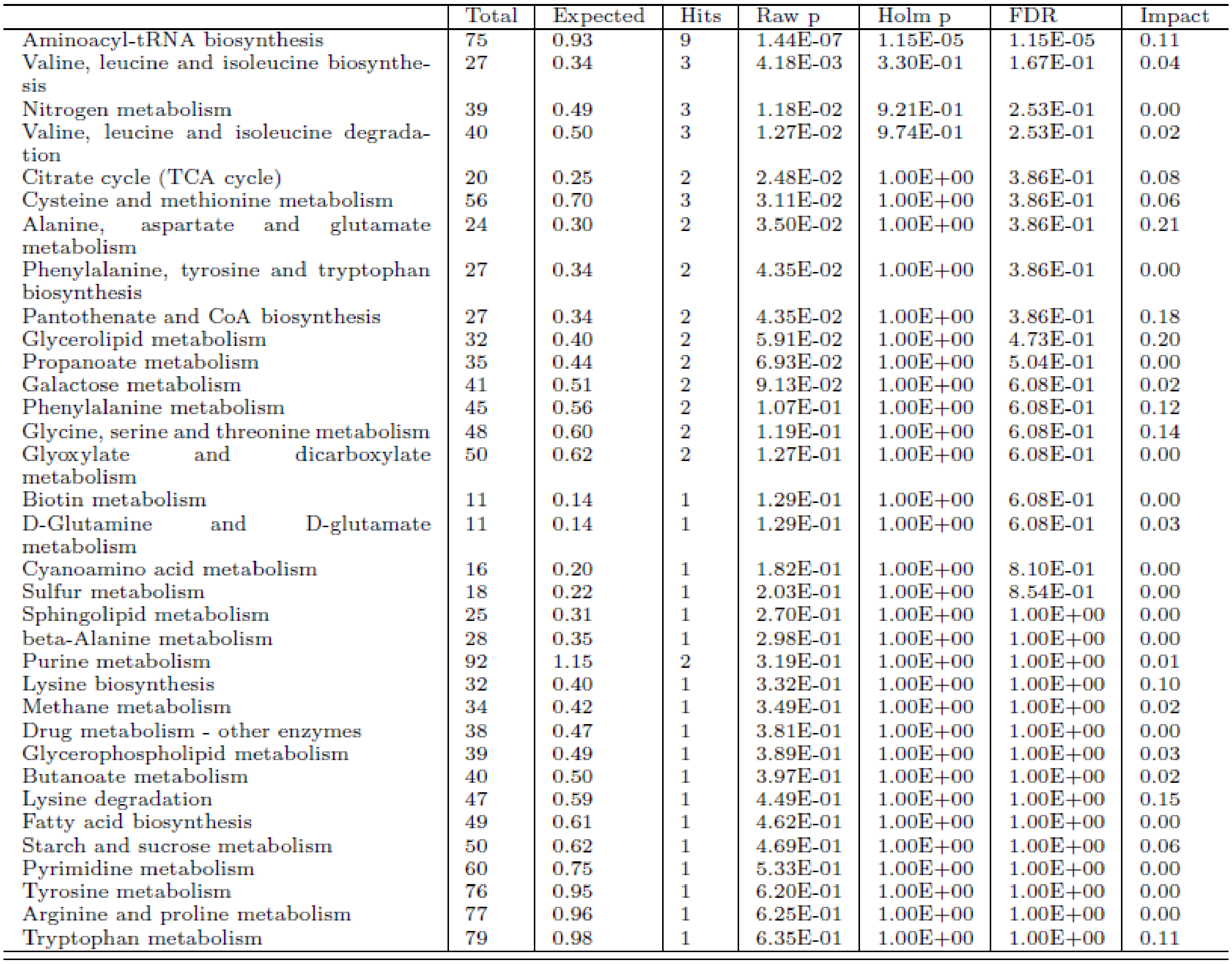


Figure 50 Pathway view of the top two highest scoring pathways (Pantothenate and CoA biosynthesis and beta-Alanine metabolism).

Table 16 Result from Pathway Analysis with MetPA for N1 *vs.* N21 matched list of identifiers.



	Total	Expected	Hits	Raw p	Holm p	FDR	Impact
Aminoacyl-tRNA biosynthesis	75	0.93	9	1.44E-07	1.15E-05	1.15E-05	0.11
Valine, leucine and isoleucine biosynthesis	27	0.34	3	4.18E-03	3.30E-01	1.67E-01	0.04
Nitrogen metabolism	39	0.49	3	1.18E-02	9.21E-01	2.53E-01	0.00
Valine, leucine and isoleucine degradation	40	0.50	3	1.27E-02	9.74E-01	2.53E-01	0.02
Citrate cycle (TCA cycle)	20	0.25	2	2.48E-02	1.00E+00	3.86E-01	0.08
Cysteine and methionine metabolism	56	0.70	3	3.11E-02	1.00E+00	3.86E-01	0.06
Alanine, aspartate and glutamate metabolism	24	0.30	2	3.50E-02	1.00E+00	3.86E-01	0.21
Phenylalanine, tyrosine and tryptophan biosynthesis	27	0.34	2	4.35E-02	1.00E+00	3.86E-01	0.00
Pantothenate and CoA biosynthesis	27	0.34	2	4.35E-02	1.00E+00	3.86E-01	0.18
Glycerolipid metabolism	32	0.40	2	5.91E-02	1.00E+00	4.73E-01	0.20
Propanoate metabolism	35	0.44	2	6.93E-02	1.00E+00	5.04E-01	0.00
Galactose metabolism	41	0.51	2	9.13E-02	1.00E+00	6.08E-01	0.02
Phenylalanine metabolism	45	0.56	2	1.07E-01	1.00E+00	6.08E-01	0.12
Glycine, serine and threonine metabolism	48	0.60	2	1.19E-01	1.00E+00	6.08E-01	0.14
Glyoxylate and dicarboxylate metabolism	50	0.62	2	1.27E-01	1.00E+00	6.08E-01	0.00
Biotin metabolism	11	0.14	1	1.29E-01	1.00E+00	6.08E-01	0.00
D-Glutamine and D-glutamate metabolism	11	0.14	1	1.29E-01	1.00E+00	6.08E-01	0.03
Cyanoamino acid metabolism	16	0.20	1	1.82E-01	1.00E+00	8.10E-01	0.00
Sulfur metabolism	18	0.22	1	2.03E-01	1.00E+00	8.54E-01	0.00
Sphingolipid metabolism	25	0.31	1	2.70E-01	1.00E+00	1.00E+00	0.00
beta-Alanine metabolism	28	0.35	1	2.98E-01	1.00E+00	1.00E+00	0.00
Purine metabolism	92	1.15	2	3.19E-01	1.00E+00	1.00E+00	0.01
Lysine biosynthesis	32	0.40	1	3.32E-01	1.00E+00	1.00E+00	0.10
Methane metabolism	34	0.42	1	3.49E-01	1.00E+00	1.00E+00	0.02
Drug metabolism - other enzymes	38	0.47	1	3.81E-01	1.00E+00	1.00E+00	0.00
Glycerophospholipid metabolism	39	0.49	1	3.89E-01	1.00E+00	1.00E+00	0.03
Butanoate metabolism	40	0.50	1	3.97E-01	1.00E+00	1.00E+00	0.02
Lysine degradation	47	0.59	1	4.49E-01	1.00E+00	1.00E+00	0.15
Fatty acid biosynthesis	49	0.61	1	4.62E-01	1.00E+00	1.00E+00	0.00
Starch and sucrose metabolism	50	0.62	1	4.69E-01	1.00E+00	1.00E+00	0.06
Pyrimidine metabolism	60	0.75	1	5.33E-01	1.00E+00	1.00E+00	0.00
Tyrosine metabolism	76	0.95	1	6.20E-01	1.00E+00	1.00E+00	0.00
Arginine and proline metabolism	77	0.96	1	6.25E-01	1.00E+00	1.00E+00	0.00
Tryptophan metabolism	79	0.98	1	6.35E-01	1.00E+00	1.00E+00	0.11

6.2.2 Metabolic and Signaling pathway analysis using IPA

Integration of transcript and metabolite candidates

'Core' and 'Metabolomics Analyses' were performed in the IPA system using candidate gene and metabolites lists produced from data analyses from the pairwise analyses (F1 *vs*. F21, N21 *vs*. F21 and N1 *vs*. N21) in Chapters 4 and 5. Initial analysis using integrative pathway mapping (metabolite with transcript data) has shown several observed markers in DD (nodules) involved in key networks, biological functions and signaling and metabolic pathways are involved in amino acid metabolism and metabolism of cofactors and vitamins.

The most significant biological functions and canonical pathways across multiple datasets (i.e. comparing results between gene/compound list from all three pairwise analyses; F1 *vs*. F21, N21 *vs*. F21 and N1 *vs*. N21), are shown in (Figures 51-54). The canonical pathways are displayed as bar charts across one axis. The other axis displays the ratio which is calculated as follows: the number of genes in a given pathway that meet cut-off criteria, divided by total number of genes that make up that pathway. Taller bars have more genes associated with the canonical pathway than shorter bars. For the single analysis (e.g. gene/compound list from the N21 *vs*. F21 analysis), a graph displaying the various pathways is presented from largest ratio to smallest ratio. For comparison analyses (i.e. gene/compound list from all three pairwise analyses F1 *vs*. F21, N21 *vs*. F21 and N1 *vs*. N21), various pathways are presented from largest ratio to smallest ratio according to the first experimental group**.** Canonical pathways that are associated with a particular network can be superimposed on molecules within the respective network.

Based on Fishers exact test for top canonical pathways associated in disease (N21) resulted in nicotinate and nicotinamide metabolism (p-value = 7.98E-03), p38 MAPK signaling (p-value = 1.03E-02), ATM signaling (p-value = 2.88E-02), antiproliferative role of TOB in T cell signaling (p-value = 4.64E-02) and ERK5 signaling (p-value = 4.95E-02). The top molecules identified from this analysis include upregulated: amphiphysin (AMPH), ankyrin-3 (ANK3), ankyrin repeat domain 28 (ANKRD28), adaptor-related protein complex 1, sigma 2 subunit (AP1S2), aryl hydrocarbon receptor nuclear translocator 2 (ARNT2), arylsulfatase D (ARSD), arsenite methyltransferase (AS3MT), activating transcription factor 4 (ATF4), autism susceptibility candidate 2 (AUTS2), cell adhesion molecule 1 (CADM1) and downregulated: zinc finger protein 423 (ZNF423), zinc finger protein 124 (ZNF124), wingless-type MMTV integration site family member 2 (WNT2), WNK lysine deficient protein kinase 4 (WNK4), transient receptor potential cation channel, subfamily C, member 6 (TRPC6), tissue factor pathway inhibitor 2 (TFPI2), tissue factor pathway inhibitor (TFPI), sulfotransferase family, cytosolic, 1C, member 4 (SULT1C4), solute carrier family 6 (neurotransmitter transporter, taurine), member 6 (SLC6A6), and sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3E (SEMA3E).

The top canonical pathways associated perturbation in healthy fascia (F1) analysis resulted in inositol metabolism (p-value = 7.41E-05), fructose and mannose metabolism (p-value = 1.85E-04), glycolysis/gluconeogenesis (p-value = 2.48E-04), AMPK signaling (p-value = 3.7E-03) and pentose phosphate pathway (p-value = 3.88E-03). The top molecules identified from this analysis include upregulated: ATP-binding cassette, sub-family B (MDR/TAP), member 10 (ABCB10), abl-interactor 2 (ABI2), A disintegrin and metalloproteinase with thrombospondin motifs 6 (ADAMTS6), adenylosuccinate synthase like 1 (ADSSL1), 1-acylglycerol-3-phosphate O-acyltransferase 5 (AGPAT5), adenylate kinase 2 (AK2), adenylate kinase 4 (AK4), v-akt murine thymoma viral oncogene homolog 3 (AKT3), aldolase A, fructose-bisphosphate (ALDOA), aldolase C, fructose-bisphosphate (ALDOC) and downregulated: zinc finger protein 124 (ZNF124), WD repeat domain 78 (WDR78), tRNA splicing endonuclease 2 homolog (TSEN2), tumor protein D52-like 1 (TPD52L1), tumor necrosis factor receptor superfamily, member 21 (TNFRSF21) transmembrane protein 97 (TMEM97), toll-like receptor 3 (TLR3), sulfotransferase family, cytosolic, 1C, member 4 (SULT1C4), spermatid perinuclear RNA binding protein (STRBP), sulfide quinone reductase-like (SQRDL).

The top canonical pathways associated with perturbation in disease (N1) analysis identified were glycolysis/gluconeogenesis (p-value = 1.99E-06), inositol metabolism (p-value = 5.18E-05), pentose phosphate pathway (p-value = 2.52E-03), PI3K/AKT signaling (p-value = 4.5E-03) and fructose and mannose metabolism (p-value = 6.65E-03). The top biofunctions associated were cell death, carbohydrate metabolism, small molecule biochemistry, cell morphology and cellular development with top molecules identified upregulated: aldolase C, fructose-bisphosphate (ALDOC), apelin (APLN), BCL2/adenovirus E1B 19kDa interacting protein 3 (BNIP3), aspartyl-tRNA synthetase (DARS), egl nine homolog 1 (EGLN1), F-box protein (FBXO16), fibroblast growth factor 1 (FGF11), gamma-glutamyl hydrolase (GGH), macrophage migration inhibitory factor (MIF), prolyl 4-hydroxylase, alpha polypeptide I (P4HA1) and downregulated: zinc finger protein 654

(ZNF654), zinc finger protein 593 (ZNF593), Wilms tumor 1 associated protein (WTAP), WD repeat domain, phosphoinositide interacting 1 (WIPI1), WD repeat domain 78 (WDR78), vascular endothelial growth factor A (VEGFA), tripeptidyl peptidase I (TPP1), triosephosphate isomerase 1 (TPI1), tumor protein D52-like 1 (TPD52L1) and thioredoxin-related transmembrane protein 4 (TMX4).

Figures 51 and 52 display the comparison analyses for the top canonical pathways. It can be seen that the perturbation effect has revealed other pathways than those identified for disease. The perturbation effect has many molecules associated with the glycolysis pathway or sugar metabolism, DE molecules in F1 score highest in cancer associated pathways, while top disease molecules (from N21) score the highest in pathways involved in vitamin B5 and signaling ways.

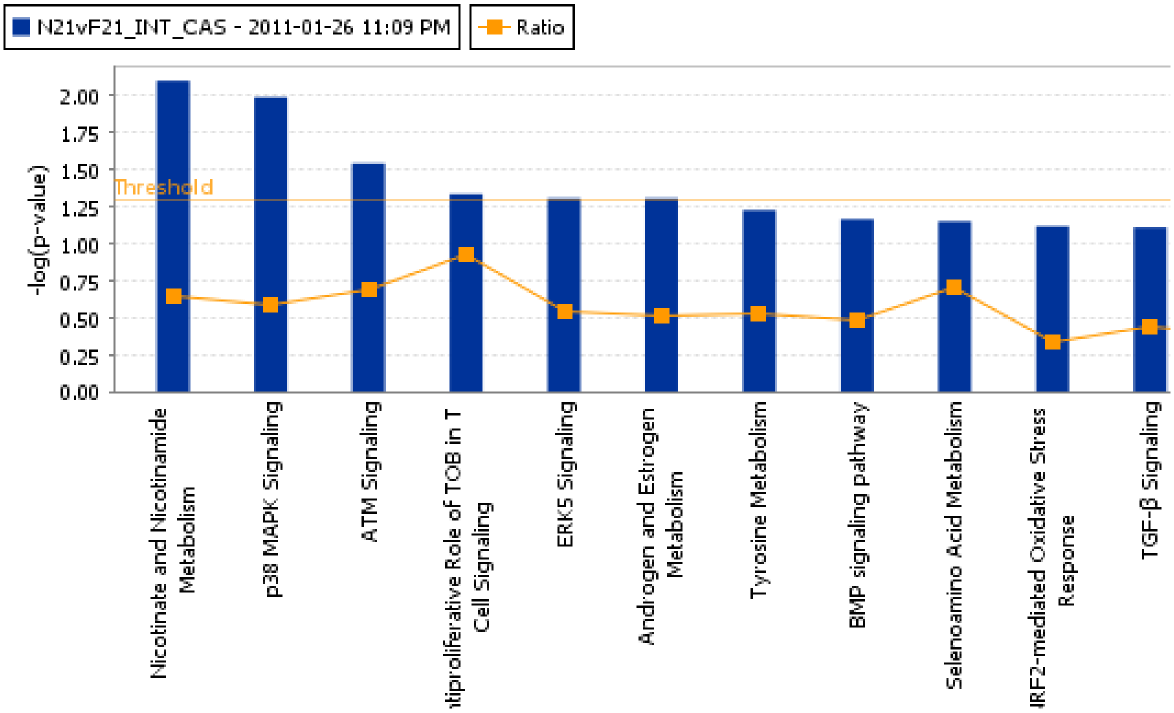


Figure 51 A summary of the distribution of metabolite markers across top scoring pathways from N21 *vs.* F21 analysis. This indicates that metabolite markers reflect CoA enzyme metabolism, key signaling pathways nicotinate and nicotinamide metabolism, p38 MAPK Signaling, antiproliferative Role of TOB in T Cell Signaling, ERK5 Signaling.

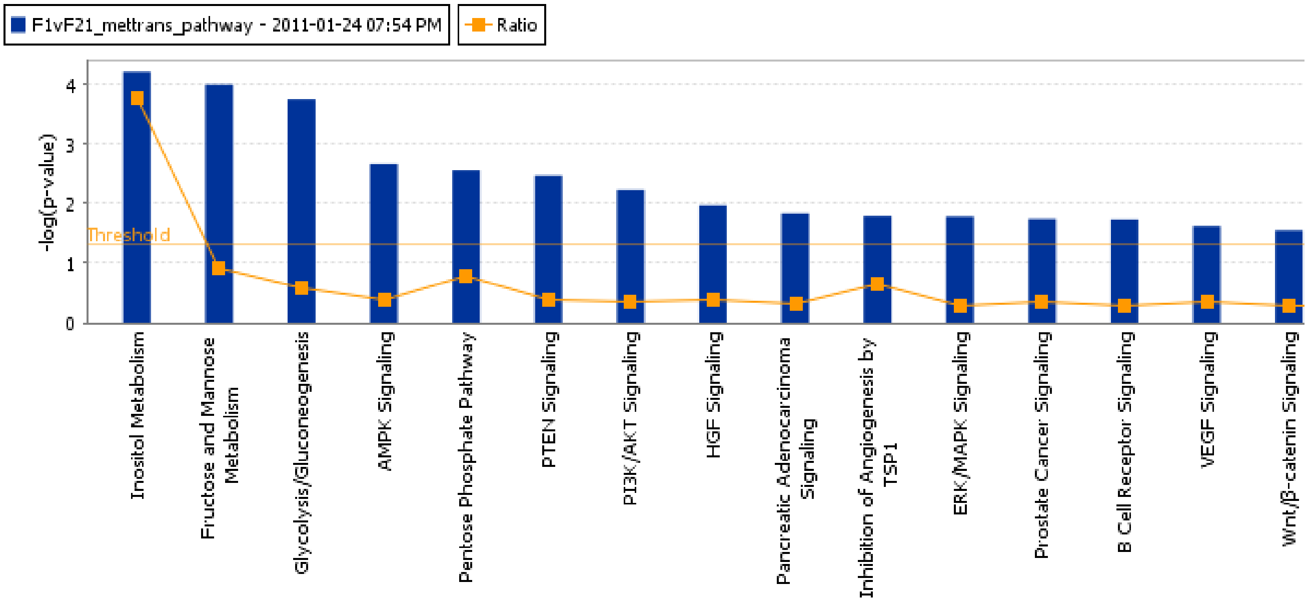


Figure 52 A summary of the distribution of metabolite markers across top scoring pathways from F1 *vs.* F21 analysis.

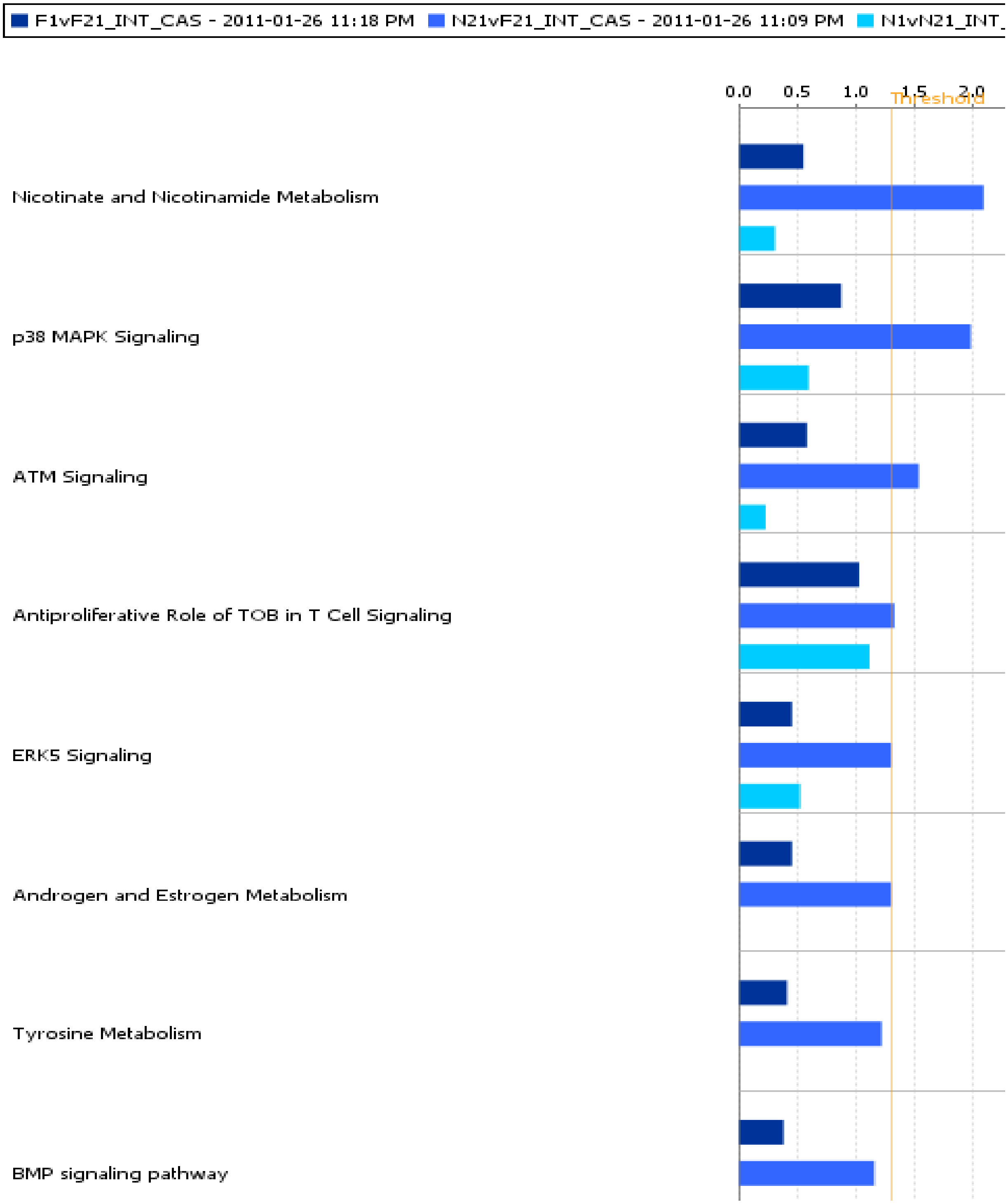


Figure 53 A comparison analysis of top scoring pathways displaying a summary of the distribution of metabolite markers across top scoring pathways.

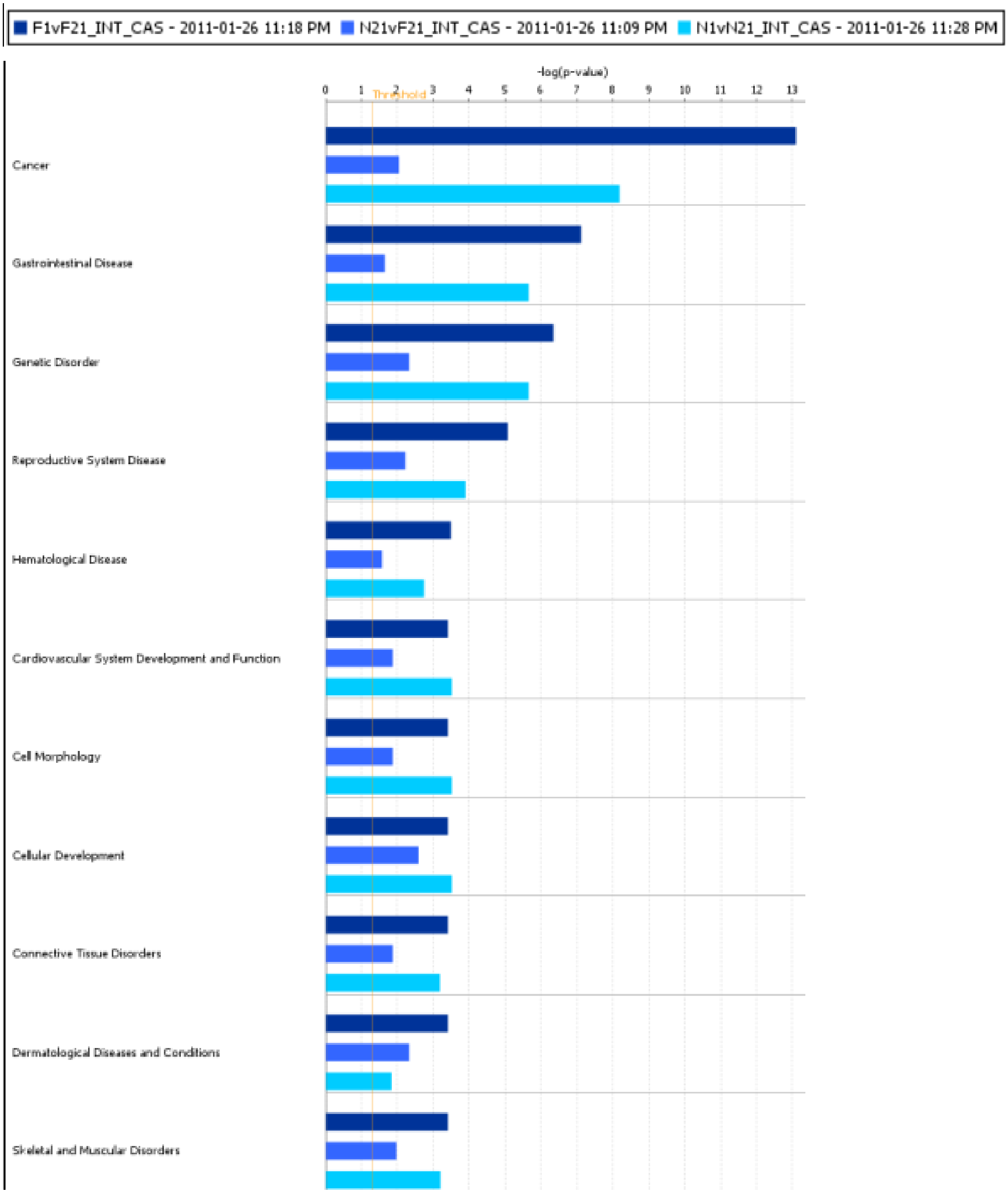


Figure 54 A comparison analysis across the top scoring biological functions in the three pairwise analyses.

Network analysis integrating candidate transcripts with candidate metabolites

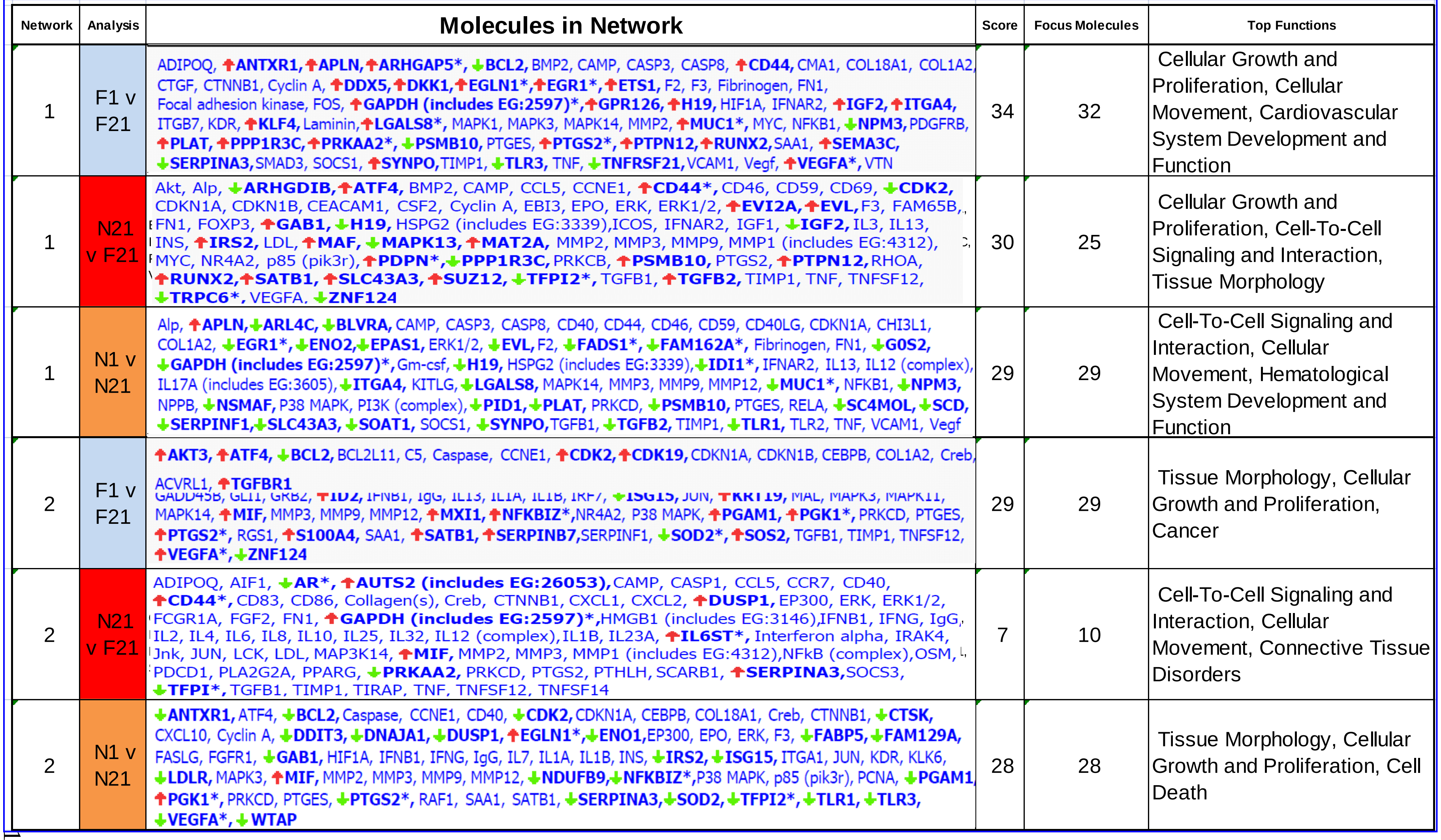
Following transcript analysis a list of networks were generated *de novo* based on the input data. These networks do not have directionality and contain molecules involved in several pathways. Table 17 shows the two most significant networks resulting from each analysis. Molecules of interest which interacted with other molecules in the Ingenuity Knowledge Base were identified as Network Eligible Molecules which served as "seeds" for generating networks. More details on The Network Generation Algorithm can be found in [165]. All of the molecules that compose each network are listed. Network eligible molecules appear in bold, other molecules are in regular font. An asterisk appears next to any gene for which the input file contained more than one identifier. A score based on a p-value calculation is determined. This calculates the likelihood that the Network Eligible Molecules that are part of a network are found therein by random chance alone. Mathematically, the score is simply the negative exponent of the right-tailed Fisher's exact test result. For example, if the score is 3, then there is a 1 in 1000 chance that the Network Eligible Molecules found in that network appeared there just by chance. The score is a measure of the number of Network Eligible Molecules in a network; the greater the number of Network Eligible Molecules in a network, the higher the score (lower the p-value) will be. The number of focus molecules indicates the number of Network Eligible Molecules per network. The maximum number of molecules per network is currently limited to 35. The three most significant functions for each network are listed under the column Top Functions. Further analysis into the high-level functional categories for individual networks was performed. For the analysis N21 *vs.* F21, Network 1 revealed the three functions with 24 molecules associated with immunological disorder, 21 molecules associated with autoimmune disease and 15 molecules associated with rheumatoid arthritis.

Network Explorer tool in IPA was used to visualise molecular relationships representing inter-relationships between molecules. Genes, proteins, and metabolites (endogenous chemicals) were displayed as various shapes. The shapes are indicative of the molecular class (protein family, or metabolite). Colouring is based on the expression values that were uploaded with the dataset. Red indicates up-regulation (positive values), green indicates down-regulation (negative values), gray indicates the molecule was part of the dataset but did not meet the specified cut-off value, and white indicates the molecule was added from the Ingenuity Knowledge Base. Lines connecting molecules indicate molecular relationships. Dashed lines indicate indirect interactions; solid lines indicate direct interactions. The style of the arrow indicates specific molecular relationships and the directionality of the interaction (A acts on B). A legend is provided in Figure 55. Canonical Pathways that are associated with a particular network were then overlaid highlighting individual molecules that are involved in a specific canonical pathway. As visualising these networks becomes increasingly difficult, the list of canonical pathways most significant for disease (N21 *vs*. F21 analysis) is shown only; Figure 62.

The networks generated displayed relevant relationships as specified by the filters applied. This may have excluded some of the relationships for every network eligible molecule. Also because some interactions present in the Ingenuity Knowledge Base were not used in the network generation process the molecules of interest and their relationships were reinserted as could be biologically important. For this reason the full complement of direct and indirect interactions for a molecule within the network are highlighted. Each network illustrates the molecules within their subcellular compartments i.e. nucleus, cytoplasm, plasma membrane and extracellular space. For molecules where no subcellular localisation information was currently available in the IPA knowledge database were categorized into an 'unknown space.'

Figures 56-60 illustrate the networks listed in Table 17. These are the top networks resulting from the combined candidate gene/metabolite identifiers. Table 18 displays an enhanced view of the top function and molecules from Network 1 that are generated from DD nodule candidate list of metabolites/transcripts.

Table 17 Top scoring networks from Core analysis (significant transcriptomic and metabolite data sets).



Network	Analysis	Molecules in Network	Score	Focus Molecules	Top Functions
1	F1 v F21	ADIPOQ, ↑**ANTXR1,**↑**APLN,**↑**ARHGAP5*,** ↓**BCL2,** BMP2, CAMP, CASP3, CASP8, ↑**CD44,** CMA1, COL18A1, COL1A2, CTGF, CTNNB1, Cyclin A, ↑**DDX5,**↑**DKK1,**↑**EGLN1*,**↑**EGR1*,**↑**ETS1,** F2, F3, Fibrinogen, FN1, Focal adhesion kinase, FOS, ↑**GAPDH (includes EG:2597)*,**↑**GPR126,**↑**H19,** HIF1A, IFNAR2, ↑**IGF2,**↑**ITGA4,** ITGB7, KDR, ↑**KLF4,** Laminin,↑**LGALS8*,** MAPK1, MAPK3, MAPK14, MMP2, ↑**MUC1*,** MYC, NFKB1, ↓**NPM3,** PDGFRB, ↑**PLAT,** ↑**PPP1R3C,**↑**PRKAA2*,** ↓**PSMB10,** PTGES, ↑**PTGS2*,** ↑**PTPN12,**↑**RUNX2,** SAA1, ↑**SEMA3C,** ↓**SERPINA3,** SMAD3, SOCS1, ↑**SYNPO,** TIMP1, ↓**TLR3,** TNF, ↓**TNFRSF21,** VCAM1, Vegf, ↑**VEGFA*,** VTN	34	32	Cellular Growth and Proliferation, Cellular Movement, Cardiovascular System Development and Function
1	N21 v F21	Akt, Alp, ↓**ARHGDIB,**↑**ATF4,** BMP2, CAMP, CCL5, CCNE1, ↑**CD44*,** CD46, CD59, CD69, ↓**CDK2,** CDKN1A, CDKN1B, CEACAM1, CSF2, Cyclin A, EBI3, EPO, ERK, ERK1/2, ↑**EVI2A,**↑**EVL,** F3, FAM65B, FN1, FOXP3, ↑**GAB1,** ↓**H19,** HSPG2 (includes EG:3339),ICOS, IFNAR2, IGF1, ↓**IGF2,** IL3, IL13, INS, ↑**IRS2,** LDL, ↑**MAF,** ↓**MAPK13,** ↑**MAT2A,** MMP2, MMP3, MMP9, MMP1 (includes EG:4312), MYC, NR4A2, p85 (pik3r),↑**PDPN*,**↓**PPP1R3C,** PRKCB, ↑**PSMB10,** PTGS2, ↑**PTPN12,** RHOA, ↑**RUNX2,**↑**SATB1,** ↑**SLC43A3,** ↑**SUZ12,** ↓**TFPI2*,** TGFB1, ↑**TGFB2,** TIMP1, TNF, TNFSF12, ↓**TRPC6*,** VEGFA, ↓**ZNF124**	30	25	Cellular Growth and Proliferation, Cell-To-Cell Signaling and Interaction, Tissue Morphology
1	N1 v N21	Alp, ↑**APLN,**↓**ARL4C,** ↓**BLVRA,** CAMP, CASP3, CASP8, CD40, CD44, CD46, CD59, CD40LG, CDKN1A, CHI3L1, COL1A2, ↓**EGR1*,**↓**ENO2,**↓**EPAS1,** ERK1/2, ↓**EVL,** F2, ↓**FADS1*,** ↓**FAM162A*,** Fibrinogen, FN1, ↓**G0S2,** ↓**GAPDH (includes EG:2597)*,** Gm-csf, ↓**H19,** HSPG2 (includes EG:3339),↓**IDI1*,** IFNAR2, IL13, IL12 (complex), IL17A (includes EG:3605),↓**ITGA4,** KITLG,↓**LGALS8,** MAPK14, MMP3, MMP9, MMP12, ↓**MUC1*,** NFKB1, ↓**NPM3,** NPPB, ↓**NSMAF,** P38 MAPK, PI3K (complex),↓**PID1,**↓**PLAT,** PRKCD, ↓**PSMB10,** PTGES, RELA, ↓**SC4MOL,**↓**SCD,** ↓**SERPINF1,**↓**SLC43A3,** ↓**SOAT1,** SOCS1, ↓**SYNPO,** TGFB1, ↓**TGFB2,** TIMP1, ↓**TLR1,** TLR2, TNF, VCAM1, Vegf	29	29	Cell-To-Cell Signaling and Interaction, Cellular Movement, Hematological System Development and Function
2	F1 v F21	↑**AKT3,** ↑**ATF4,** ↓**BCL2,** BCL2L11, C5, Caspase, CCNE1, ↑**CDK2,**↑**CDK19,** CDKN1A, CDKN1B, CEBPB, COL1A2, Creb, ACVRL1, ↑**TGFBR1** GADD45B, GLI1, GRB2, ↑**ID2,** IFNB1, IgG, IL13, IL1A, IL1B, IRF7, ↓**ISG15,** JUN, ↑**KRT19,** MAL, MAPK3, MAPK11, MAPK14, ↑**MIF,** MMP3, MMP9, MMP12, ↑**MXI1,** ↑**NFKBIZ*,** NR4A2, P38 MAPK, ↑**PGAM1,** ↑**PGK1*,** PRKCD, PTGES, ↑**PTGS2*,** RGS1, ↑**S100A4,** SAA1, ↑**SATB1,** ↑**SERPINB7,** SERPINF1, ↓**SOD2*,** ↑**SOS2,** TGFB1, TIMP1, TNFSF12, ↑**VEGFA*,**↓**ZNF124**	29	29	Tissue Morphology, Cellular Growth and Proliferation, Cancer
2	N21 v F21	ADIPOQ, AIF1, ↓**AR*,** ↑**AUTS2 (includes EG:26053),** CAMP, CASP1, CCL5, CCR7, CD40, ↑**CD44*,** CD83, CD86, Collagen(s), Creb, CTNNB1, CXCL1, CXCL2, ↑**DUSP1,** EP300, ERK, ERK1/2, FCGR1A, FGF2, FN1, ↑**GAPDH (includes EG:2597)*,** HMGB1 (includes EG:3146),IFNB1, IFNG, IgG, IL2, IL4, IL6, IL8, IL10, IL25, IL32, IL12 (complex),IL1B, IL23A, ↑**IL6ST*,** Interferon alpha, IRAK4, Jnk, JUN, LCK, LDL, MAP3K14, ↑**MIF,** MMP2, MMP3, MMP1 (includes EG:4312),NFkB (complex),OSM, PDCD1, PLA2G2A, PPARG, ↓**PRKAA2,** PRKCD, PTGS2, PTHLH, SCARB1, ↑**SERPINA3,** SOCS3, ↓**TFPI*,** TGFB1, TIMP1, TIRAP, TNF, TNFSF12, TNFSF14	7	10	Cell-To-Cell Signaling and Interaction, Cellular Movement, Connective Tissue Disorders
2	N1 v N21	↓**ANTXR1,** ATF4, ↓**BCL2,** Caspase, CCNE1, CD40, ↓**CDK2,** CDKN1A, CEBPB, COL18A1, Creb, CTNNB1, ↓**CTSK,** CXCL10, Cyclin A, ↓**DDIT3,**↓**DNAJA1,**↓**DUSP1,**↑**EGLN1*,**↓**ENO1,** EP300, EPO, ERK, F3, ↓**FABP5,**↓**FAM129A,** FASLG, FGFR1, ↓**GAB1,** HIF1A, IFNB1, IFNG, IgG, IL7, IL1A, IL1B, INS, ↓**IRS2,**↓**ISG15,** ITGA1, JUN, KDR, KLK6, ↓**LDLR,** MAPK3, ↑**MIF,** MMP2, MMP3, MMP9, MMP12, ↓**NDUFB9,**↓**NFKBIZ*,** P38 MAPK, p85 (pik3r), PCNA, ↓**PGAM1,** ↑**PGK1*,** PRKCD, PTGES, ↓**PTGS2*,** RAF1, SAA1, SATB1, ↓**SERPINA3,**↓**SOD2,**↓**TFPI2*,**↓**TLR1,** ↓**TLR3,** ↓**VEGFA*,** ↓**WTAP**	28	28	Tissue Morphology, Cellular Growth and Proliferation, Cell Death

Table 18 Enhanced view of function and molecules from Network 1 generated from disease (N21 *vs.* F21 analysis) candidate list of metabolites/transcripts.

Function Annotation	Molecules	# Molecules
immunological disorder	ADK, AMPH, ANK3, ANKRD28, AR, ARHGDIB, ARNT2, AUTS2 (includes EG:26053), CADM1, CAMK1D, CD44, CDK2, CHRM3, DAPK1, DCHS1, DPF3, DUSP1, EIF1, ELL2, ERP29, ESD, FCHO2, FRMD3, FTH1	24
autoimmune disease	ADK, AMPH, ANK3, ANKRD28, ARHGDIB, AUTS2 (includes EG:26053), CADM1, CAMK1D, CD44, CDK2, CHRM3, DAPK1, DCHS1, DPF3, DUSP1, ELL2, ERP29, ESD, FCHO2, FRMD3, FTH1	21
rheumatoid arthritis	ADK, AMPH, ANKRD28, ARHGDIB, AUTS2 (includes EG:26053), CADM1, CAMK1D, CD44, DAPK1, DCHS1, DUSP1, ESD, FCHO2, FRMD3, FTH1	15

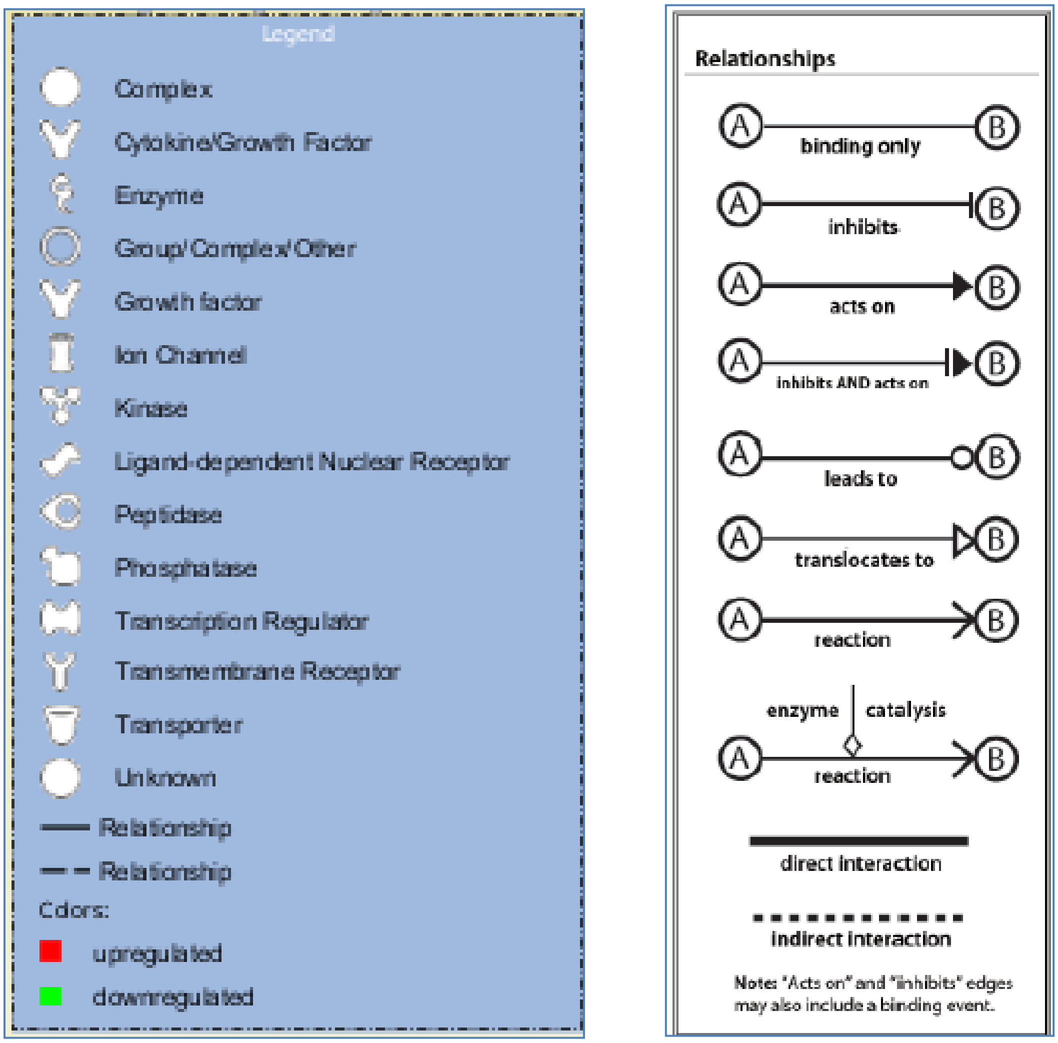


Figure 55 a) Key for all networks. b) Relationship legend.

Network 1 – Effect of hypoxia on healthy fascia

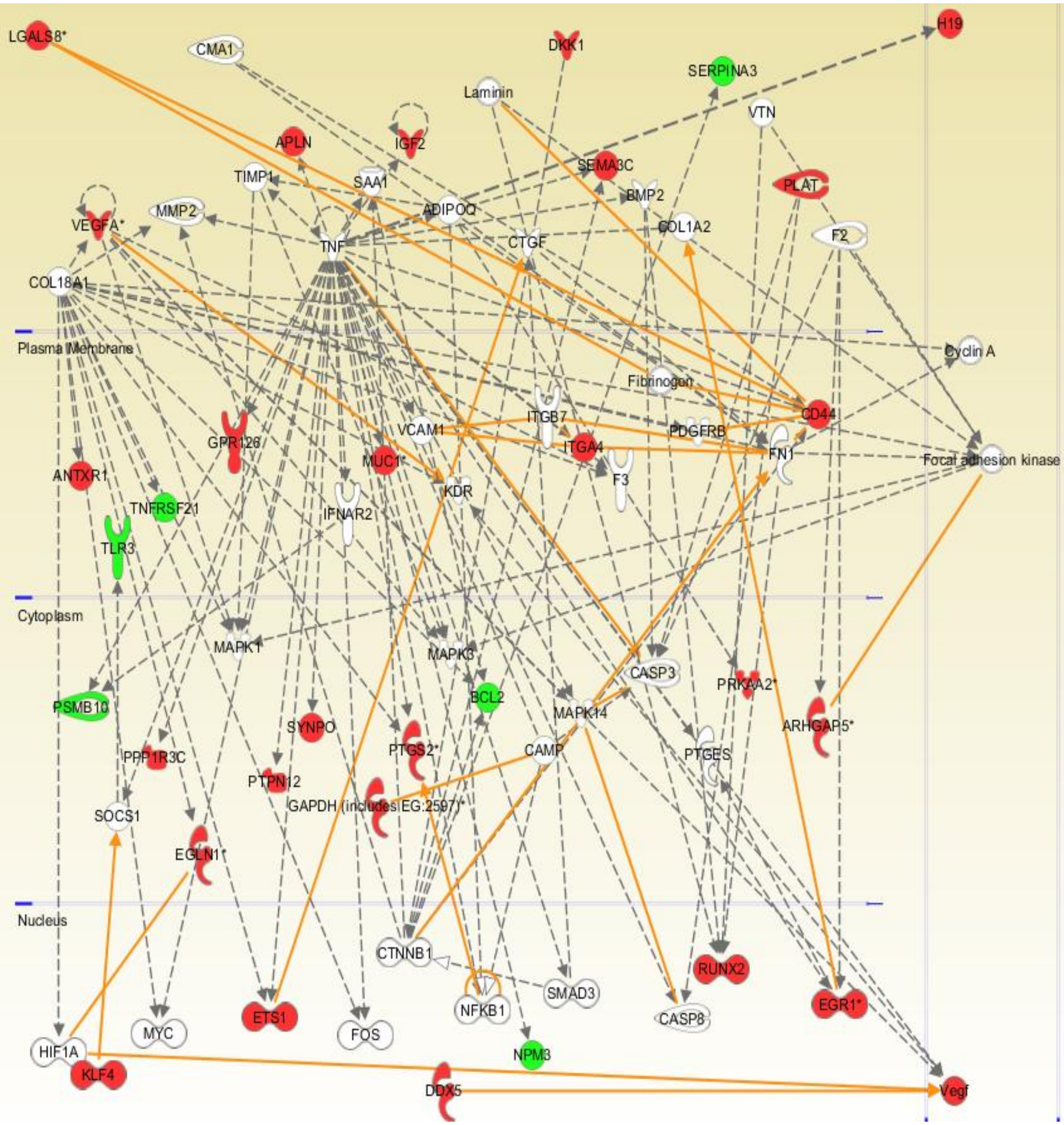


Figure 56 Top scoring Network 1 based on the significant differentially expression transcripts and metabolites from **F1** ***vs.*** **F21** gene/metabolite lists. Colour coding of the nodes corresponds to the direction of responses (up-regulation shown in red and down-regulation shown in green). Lines that connect two molecules represent relationships. Any two molecules that bind, act upon one another are orange.

Network 2 - F1 *vs.* F21 analysis

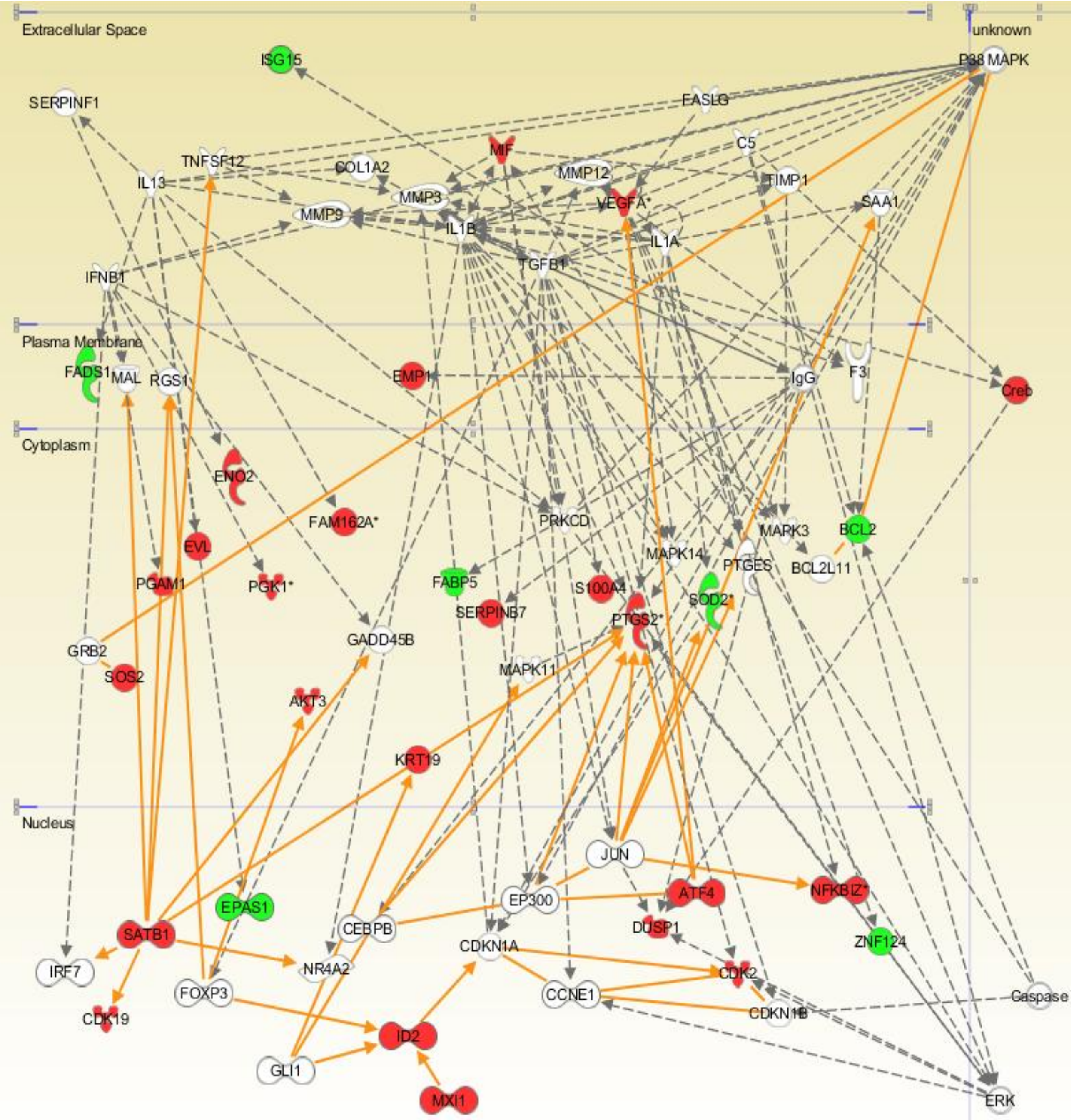


Figure 57 Top scoring Network 2 based on the significant differentially expression transcripts and metabolites from **F1 *vs.* F21** gene/metabolite lists. Color coding of the nodes corresponds to the direction of responses (up-regulation shown in red and down-regulation shown in green. Any two molecules that bind, act upon one another, are shown in orange.

Network 1 – Top molecules highlighted from DD Nodules *vs.* healthy fascia analysis

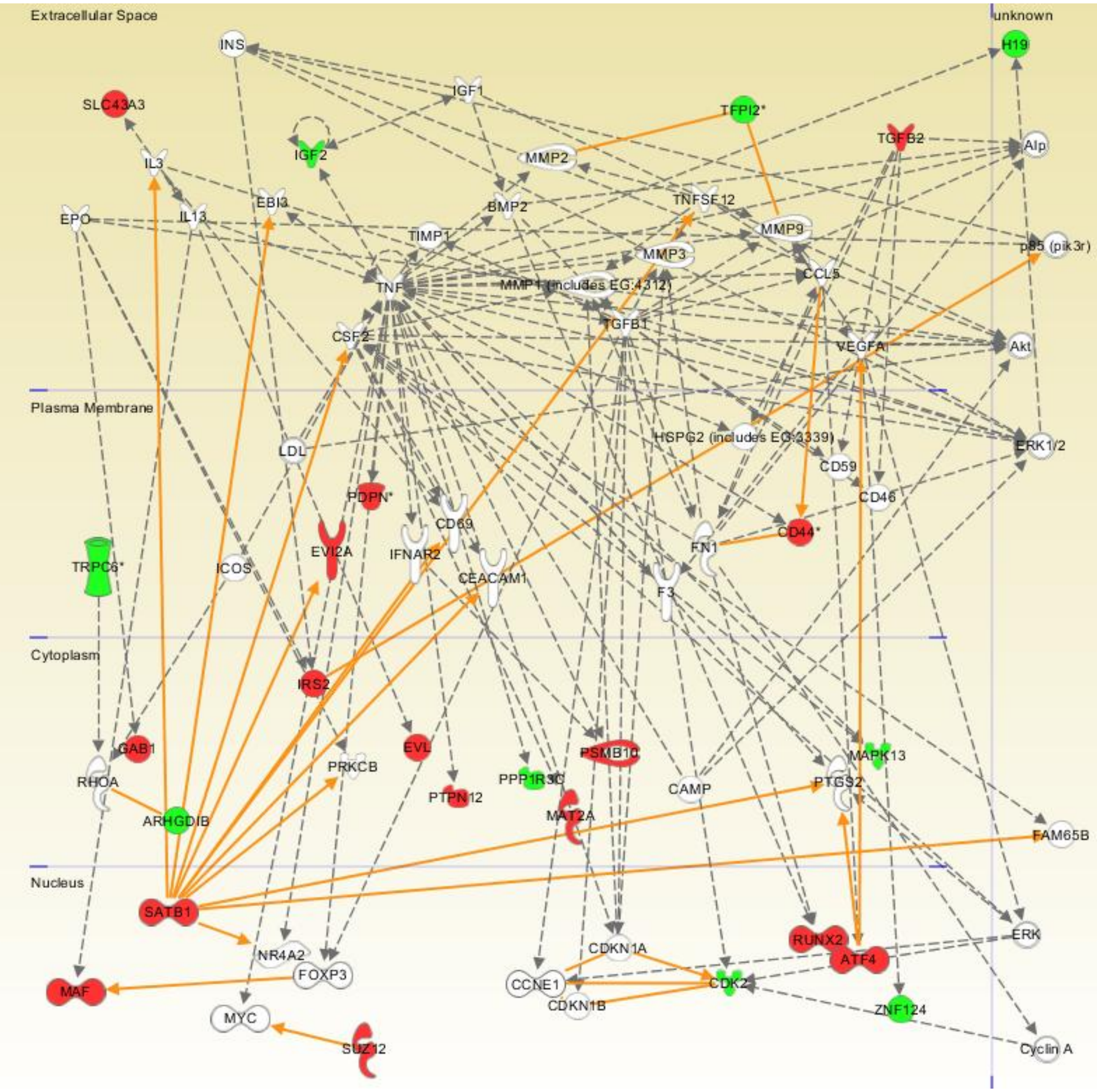


Figure 58 Top scoring network based on the significant differentially expression transcripts and metabolites From **N21 *vs.* F21** gene/metabolite lists. Color coding of the nodes corresponds to the direction of responses (up-regulation shown in red and down-regulation shown in green). Lines that connect two molecules represent relationships. Any two molecules that bind, act upon one another, or that are involved with each other in any other manner would be considered to possess a relationship between them (orange). Each relationship between molecules is created using scientific information contained in the Ingenuity Knowledge Base.

Network 1 from N1 *vs.* N21 analysis

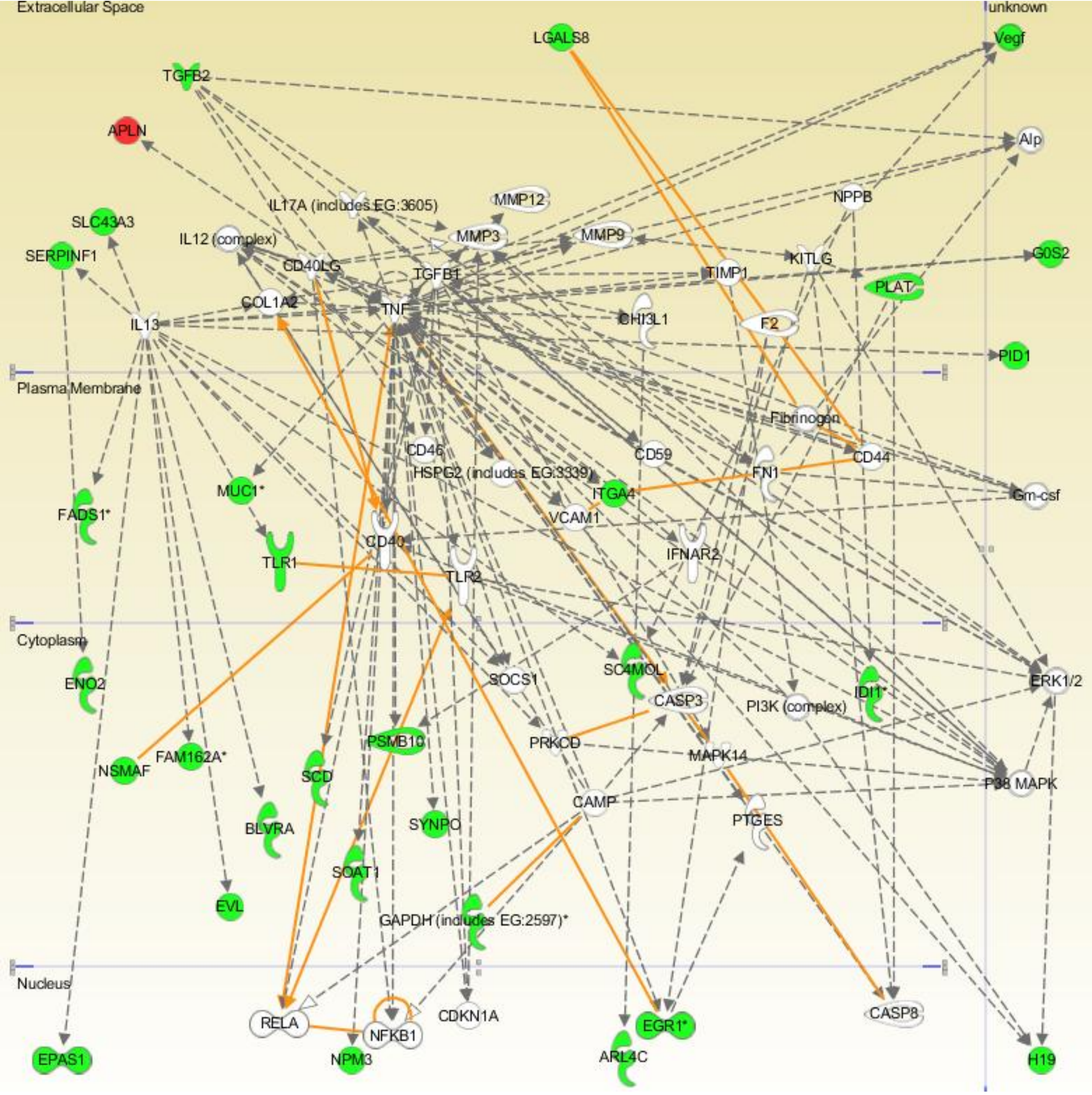


Figure 59 Top scoring network based on the significant differentially expression transcripts and metabolites From **N1 *vs.* N21** gene/metabolite lists. Colour coding of the nodes corresponds to the direction of responses (up-regulation shown in red and down-regulation shown in green). Lines that connect two molecules represent relationships. Any two molecules that bind, act upon one another are orange.

Network 2 from N1 *vs.* N21 analysis

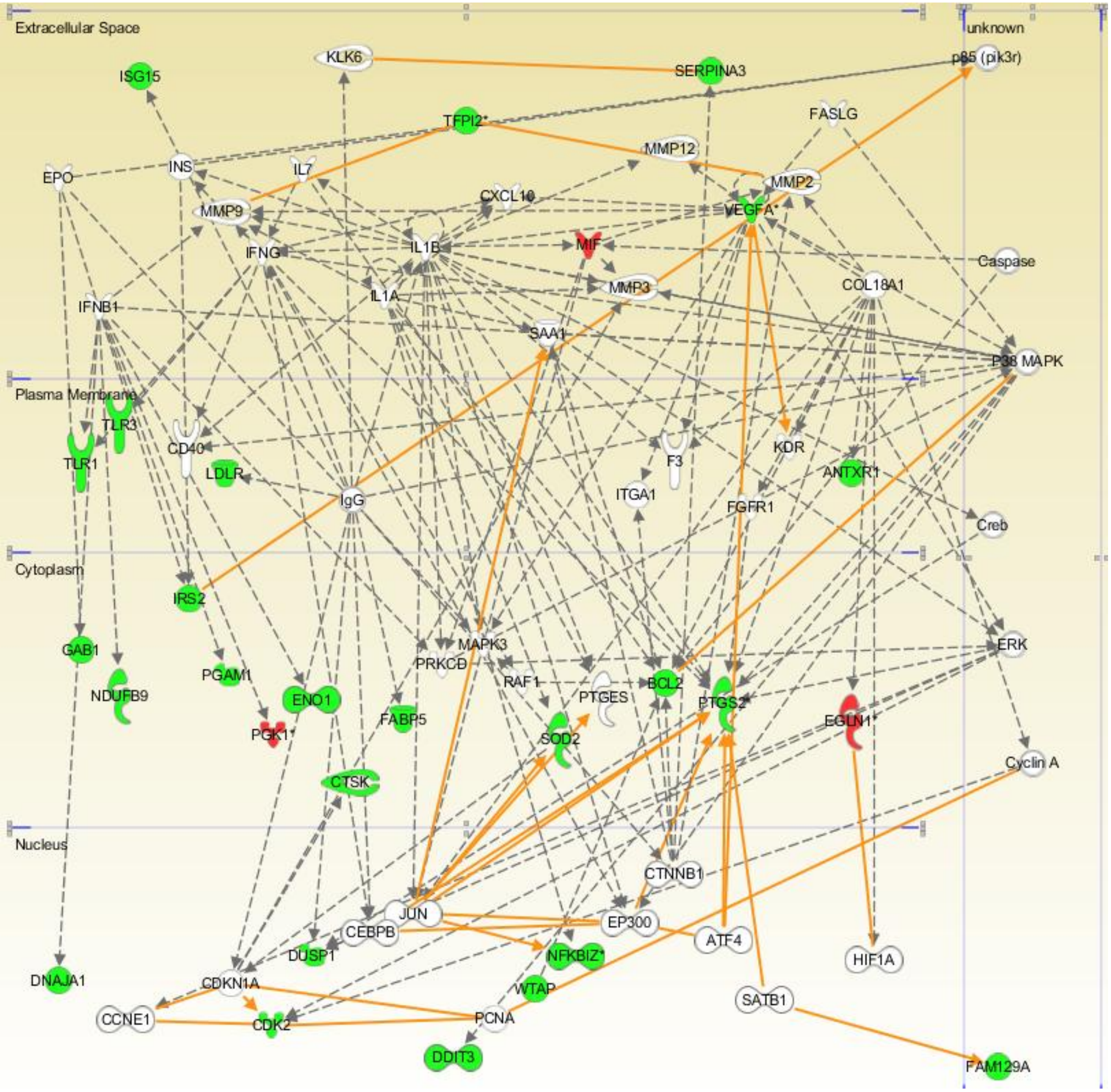


Figure 60 Top scoring Network 2 based on the significant differentially expression transcripts and metabolites From **N1 *vs.* N21** gene/metabolite lists. Colour coding of the nodes corresponds to the direction of responses (up-regulation shown in red and down-regulation shown in green). Lines that connect two molecules represent relationships. Any two molecules that bind, act upon one another are shown in orange. Dashed lines are indicative of indirect binding and/or relationships with confirmed literature references.

6.2.3 Metabolite and gene mapping

One of the aims of the DD metabolomics work was to provide putative biomarkers of disease phenotype that can be monitored to determine the effect of hypoxia and discover novel metabolite entities and possibly translate a metabolic network for *in vitro* studies. To gain additional confidence that these metabolite changes are associated with transcriptional changes, rather than just a product of altered phenotype based upon induced perturbation transcriptomics data were overlaid on the metabolite-centric networks and pathways. Table 19 shows the two most significant networks resulting from metabolomics analysis. Top transcriptomics information from Network 1 was overlaid on the metabolic network shown in Figures 61 - 65 to illustrate the analysis of disease (N21) and for effect of hypoxia on networks molecules upon perturbation in disease cells (N1). This now provides smaller networks from which key molecules can be used to depict our components which may be used for modeling; *in silico* studies for further experimentation and validation. The networks were then overlaid with the most significant canonical pathways specific for molecules with individual network (e.g. Figure 62). Figures 66 and 67 mapped the top scoring molecules from F1 and N1 upon N21 Network 1. A notable switching in direction (up or downregulated) of some molecules is observed.

Table 19 Top scoring networks from metabolomics analysis (significant metabolite data sets).

Network ID	Analysis	Molecules in Metabolite Network	Score	Focus Molecules	Top Functions
1	N1vN21	↓**BLVRA,** CDKN1A, COL1A2, ↓**CTSK,** EDN1, ↓**EGR1*,**↓**ENO2,**↓**EPAS1,** EPO, ↓**EVL,** ↓**FADS1*,** ↓**FAM162A*,** ↓**G0S2,** ↓**GAB1,** ↓**H19,** HSPG2 (includes EG:3339),↓**IDI1*,** IFNG, IL13, ↓**IRS2,** ↓**ITGA4,** NPPB, ↓**PID1,** PTGES, ↓**SC4MOL,**↓**SERPINF1,**↓**SLC43A3,** SOCS1, ↓**SOD2,**↓**TGFB2,**↓**TLR1,**↓**TLR3,** TNF, TP53, VCAM1	27	22	Cardiovascular System Development and Function, Organismal Development, Hematological System Development and Function
1	F1vF21	APP, ↑**ATF4,** ↓**BCL2,** C5, CCL5, CCNE1, ↑**CDK2,**↑**CDK19,** CEBPB, CHI3L1, CSF1, CXCL12, ↑**DUSP1,**↑**EMP1,** EP300, ERK, ↓**FABP5,** FASLG, GRB2, IgG, IL17A (includes EG:3605), ↓**ISG15,** JUN, P38 MAPK, PRKCD, ↑**PTGS2*,** SAA1, ↑**SATB1,** ↑**SERPINB7,**SHC1, ↑**SOS2** TERT, TNFSF12, ↑**VEGFA*,**↓**ZNF124**	13	14	Tissue Morphology, Cellular Growth and Proliferation, Hematological System Development and Function
1	N21vF21	Alp, ↑**APLN,**↓**ARL4C,** CAMP, CASP3, CD40, CD44, CD40LG, CDKN1A, CHI3L1, COL1A2, IgG, IL2, IL4, IL6, IL8, IL10, IL13, IL12 (complex),IL1B, IRAK3, ↑**MAF,** MYC, NID1, ↑**PDPN*,**PI3K (complex),↓**PPP1R3C,**↑**PSMB10,**↑**PTPN12,**RFTN1, RHOA, ↑**SLC43A3,** ↑**SUZ12,** TNF, ↓**TRPC6***	13	12	Cell-mediated Immune Response, Cellular Development, Cellular Function and Maintenance
2	N1vN21	↓**ANTXR1,**↓**BCL2,** CASP8, CD40LG, COL18A1, ↓**DNAJA1,** ↑**EGLN1*,**↓**ENO1,**F3, FOS, HIF1A, IFNB1, IL1A, IL1B, Jnk, JUN, KITLG,↑**MIF,** MMP2, MMP3, MMP9, MMP1 (includes EG:4312),↓**NDUFB9,**↓**NFKBIZ*,**↓**PGAM1,** ↑**PGK1*,**↓**PLAT,** PRKCD, PTGES, ↓**PTGS2*,** ↓**SOD2,** STAT1, TERT, ↓**TFPI2*,**↓**VEGFA***	15	15	Cardiovascular System Development and Function, Organismal Development, Cellular Growth and Proliferation
2	F1vF21	↑**ANTXR1,**↑**ARHGAP5*,** ↓**BCL2,** CCL5, ↑**CD44,** CMA1, COL18A1, CTGF, CTNNB1, Cyclin A, ↑**DKK1,**↑**EGLN1*,**↑**EGR1*,**↑**ETS1,** F2, Fibrinogen, FN1, Focal adhesion kinase, HIF1A, IL12 (complex),↑**ITGA4,** ITGB1, ITGB7, Laminin,↑**LGALS8*,** MAPK1, MMP2, MMP1 (includes EG:4312),PDGFRB, ↑**RUNX2,**↑**SEMA3C,**↓**SERPINA3,**SPP1, ↓**TNFRSF21,** VTN	13	14	Cellular Movement, Cellular Development, Cardiovascular System Development and Function
2	N21vF21	Alp, ↓**AR*,** CCNB1, CCNE1, CD46, CD59, CD69, ↓**CDK2,** CDKN1B, CEACAM1, CSF2, Cyclin A, ↑**DUSP1,**↑**EVI2A,** FAM65B, FN1, IL6, ↑**IL6ST*,** Jnk, LDL, ↓**MAPK13,**↑**MAT2A,** ↑**MIF,** MMP2, MMP3, MMP1 (includes EG:4312),NFkB (complex),OSM, ↑**RUNX2,**↑**SATB1,** ↑**SERPINA3,**TGFB1, ↑**TGFB2,** TIMP1, TNFSF12	13	12	Cellular Growth and Proliferation, Cell Death, Skeletal and Muscular Disorders

Transcript data superimposed on metabolite Network 1 - N21 (disease molecules)

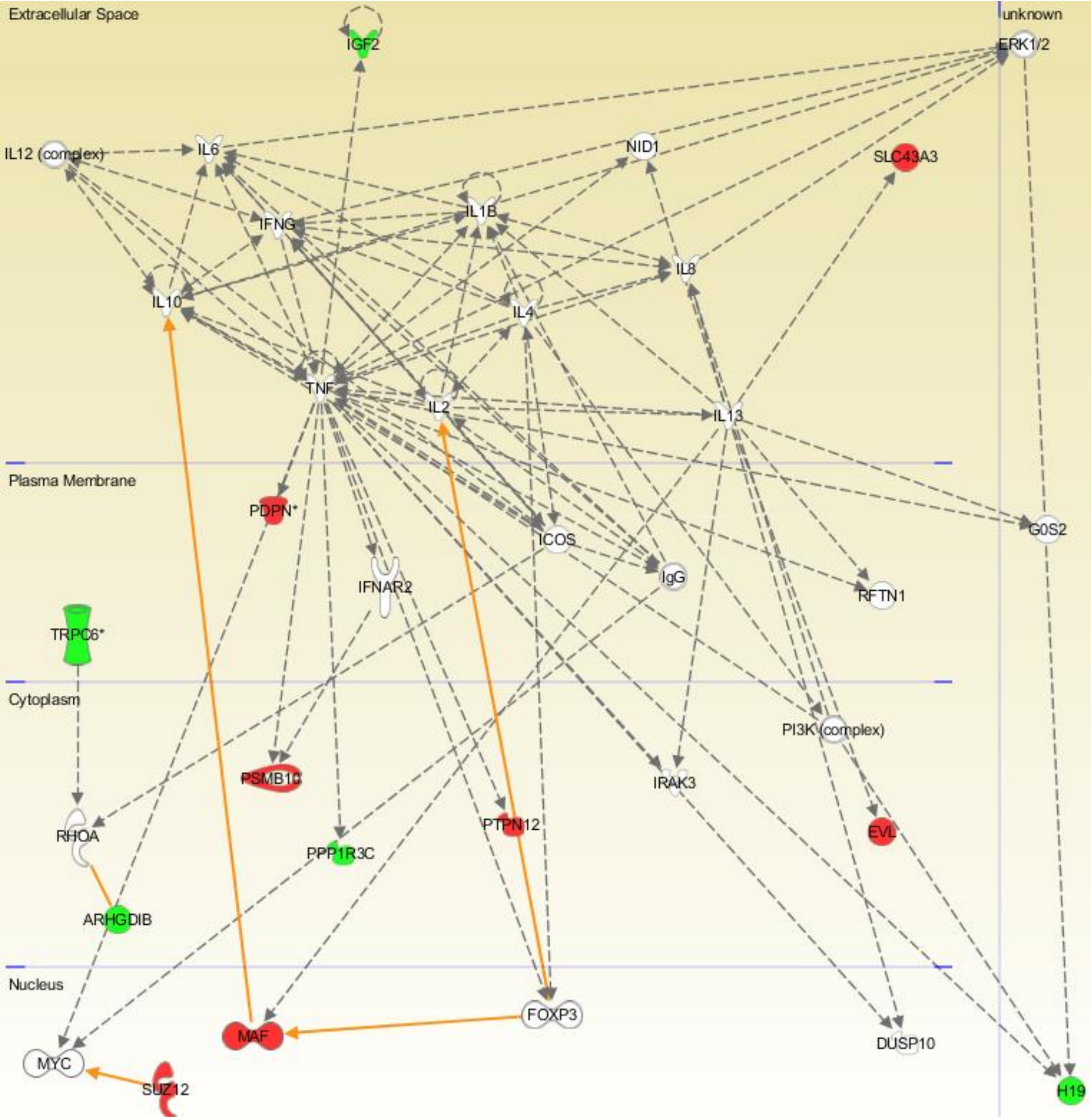


Figure 61 Transcriptomics data superimposed on the metabolite network 1 generated from molecules in the analysis of N21 *vs.* F21**.**

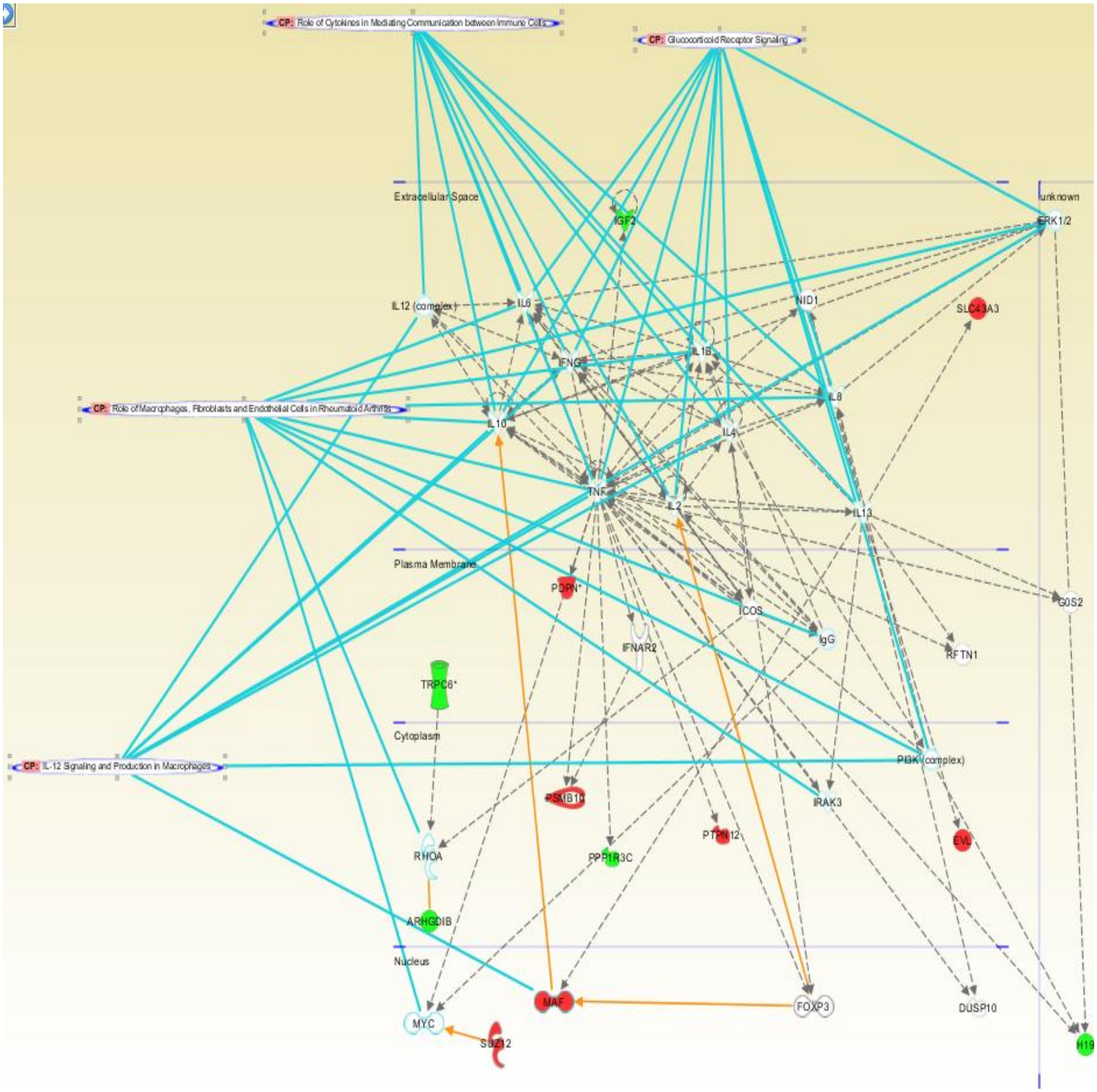

CP: Role of Cytokines in Mediating Communication between Immune Cells
CP: Glucocorticoid Receptor Signaling
Extracellular Space
unknown
IGF2
ERK1/2
IL12 (complex)
IL6
NID1
SLC43A3
IL1B
IFNG
CP: Role of Macrophages, Fibroblasts and Endothelial Cells in Rheumatoid Arthritis
IL8
IL10
IL4
TNF
IL2
IL13
Plasma Membrane
PDPN*
ICOS
G0S2
IgG
RFTN1
IFNAR2
TRPC6*
Cytoplasm
CP: IL-12 Signaling and Production in Macrophages
PI3K (complex)
PSMB10
IRAK3
PTPN12
EVL
RHOA
PPP1R3C
ARHGDIB
Nucleus
MAF
FOXP3
DUSP10
MYC
SUZ12
H19

Figure 62 Transcriptomics data superimposed on the metabolite network (1) generated with analysis of N21 v F21 with top scoring canonical pathways overlaid. These are: role of cytokines in mediating communication between immune cells, glucocorticoid receptor signaling, role of macrophages, fibroblasts and endothelial cells in rheumatoid arthritis and signaling and production in macrophages.

Transcript data superimposed on metabolite Network 2 - N21 vs. F21

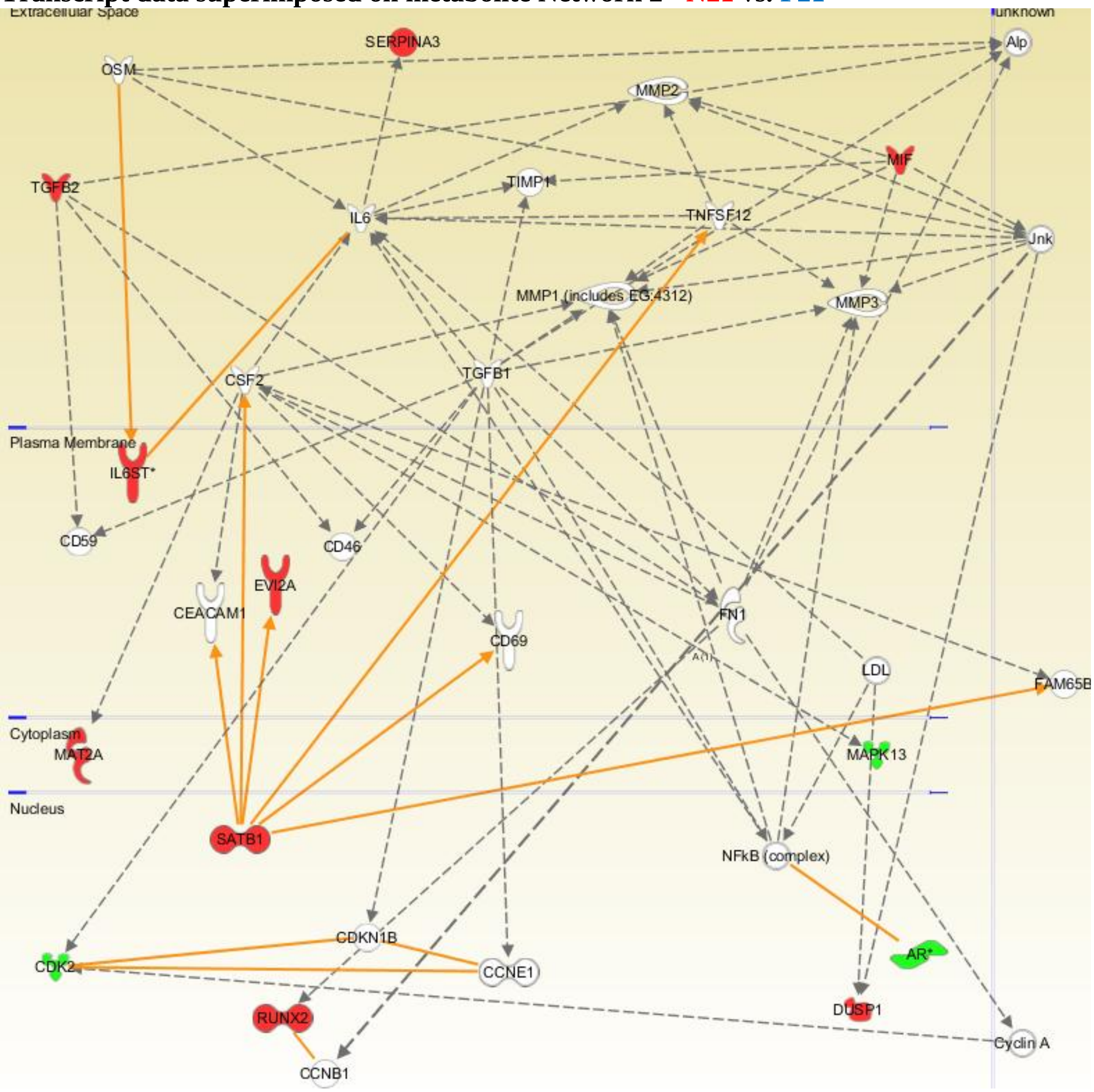


Figure 63 Transcriptomics data superimposed on the metabolite network (2) generated from molecules in the analysis of N21 *vs.* F21.

Transcript data superimposed on metabolite-centric Network 1 - N1 *vs.* N21

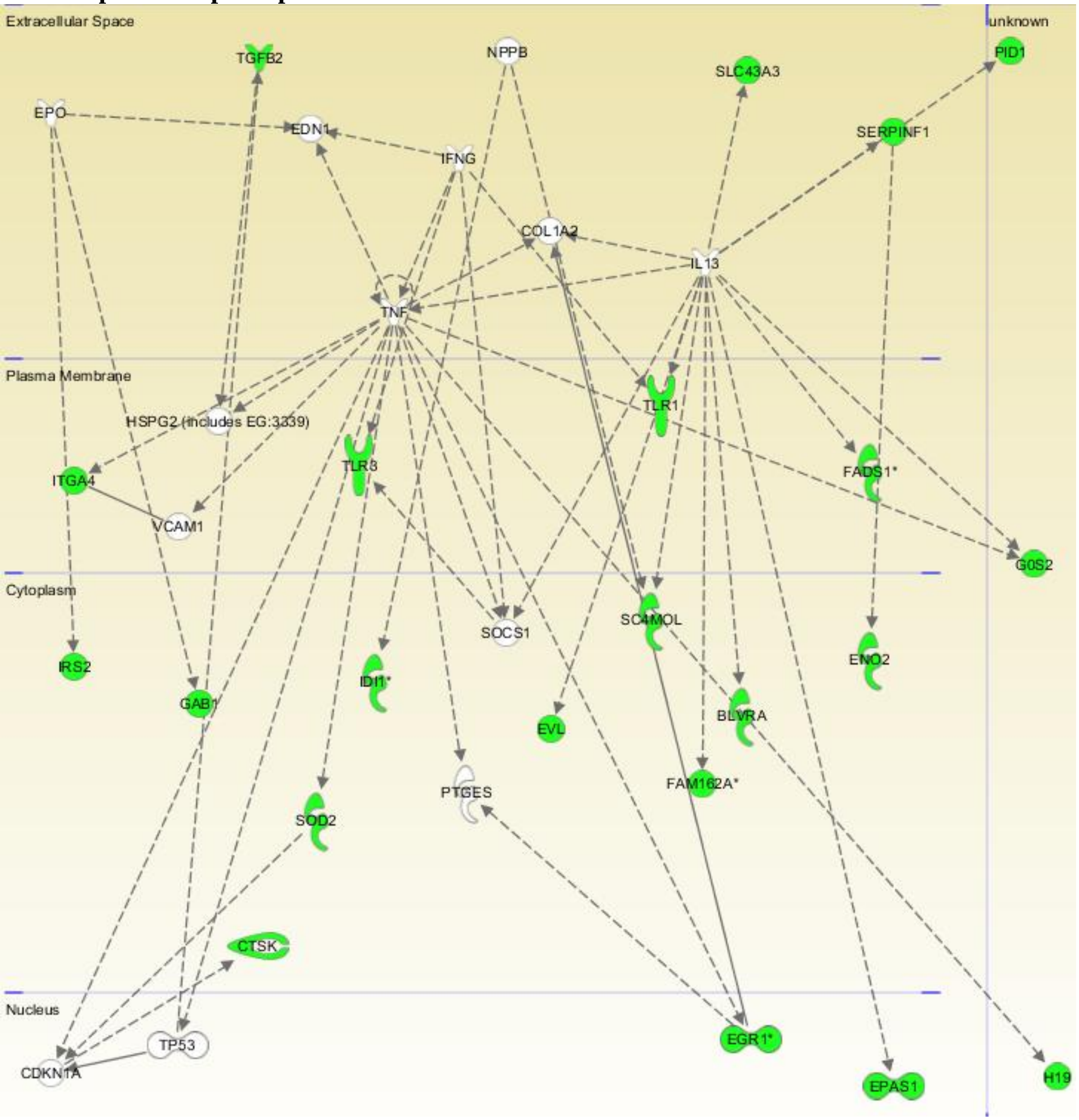


Figure 64 Transcriptomics data superimposed on the metabolite network (1) generated from molecules in N1 *vs.* N21 analysis.

Transcript data superimposed on metabolite-centric Network 2 - N1 *vs.* N21

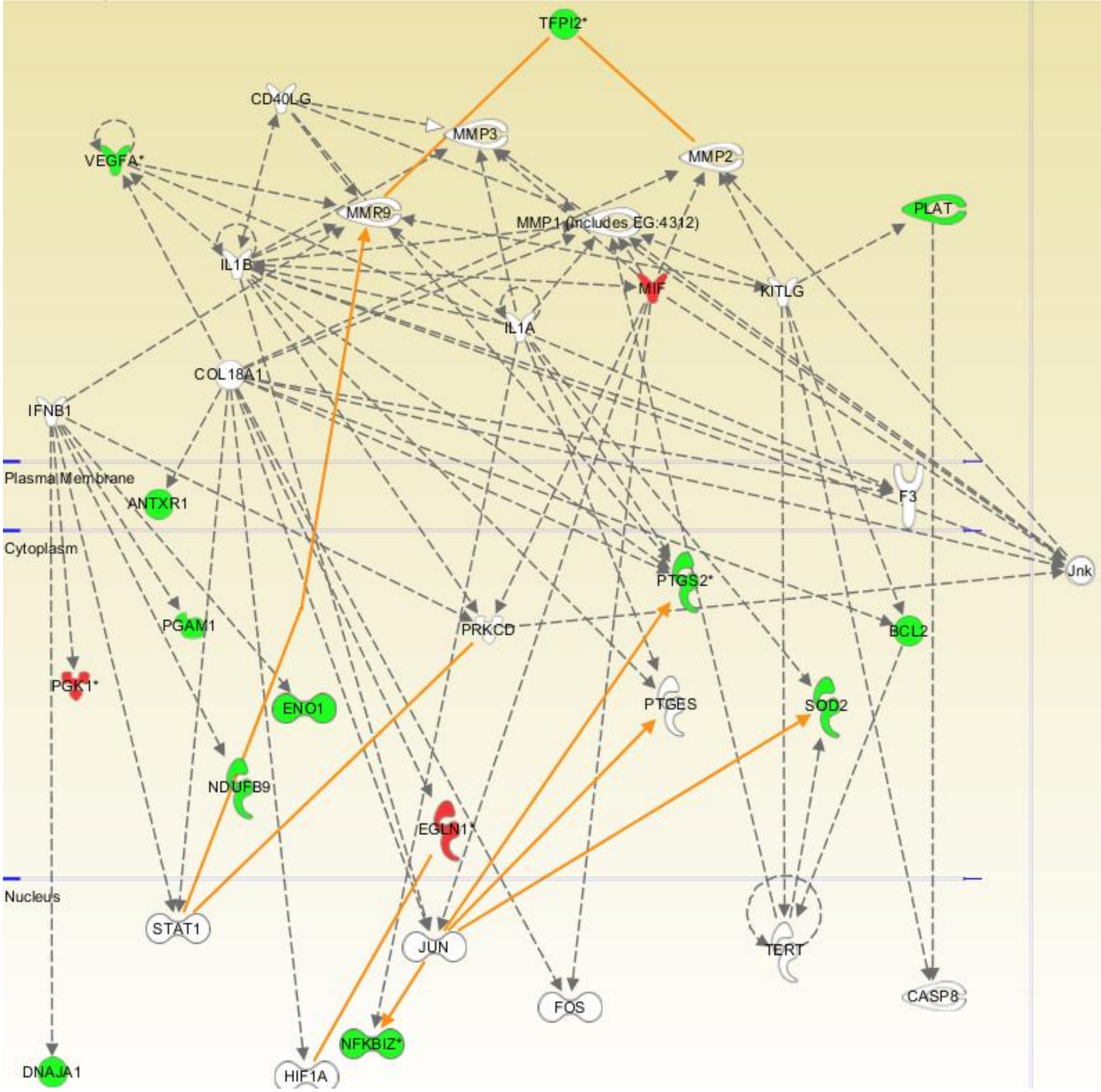


Figure 65 Transcriptomics data superimposed on the metabolite network (2) generated from molecules in N1 *vs.* N21 analysis.

F1 molecules mapped onto N21 network (superimposed on Figure 58)

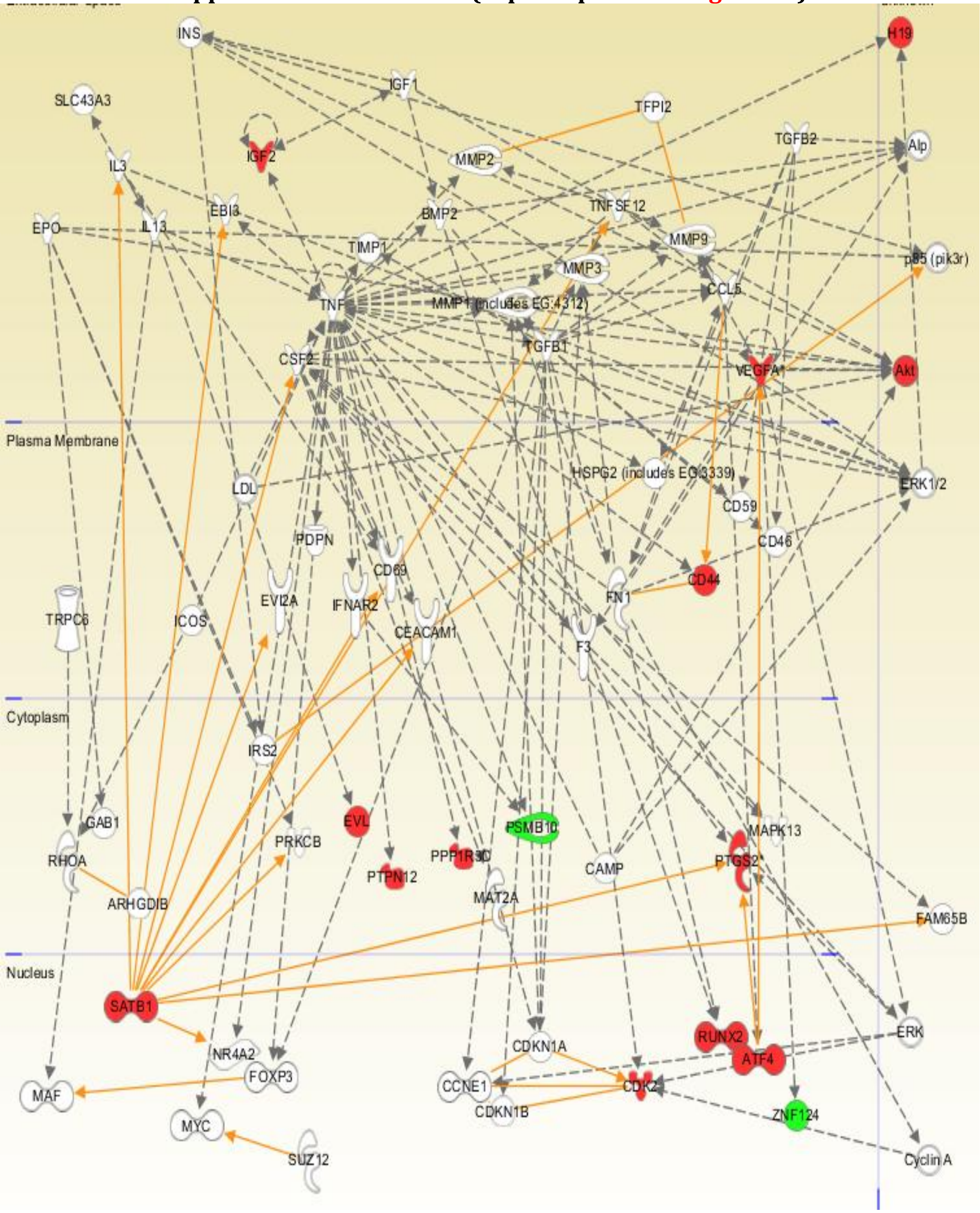


Figure 66 Molecules from F1 transcript-metabolite network superimposed on the N21 network (1).

N1 molecules mapped onto N21 (superimposed on Figure 58)

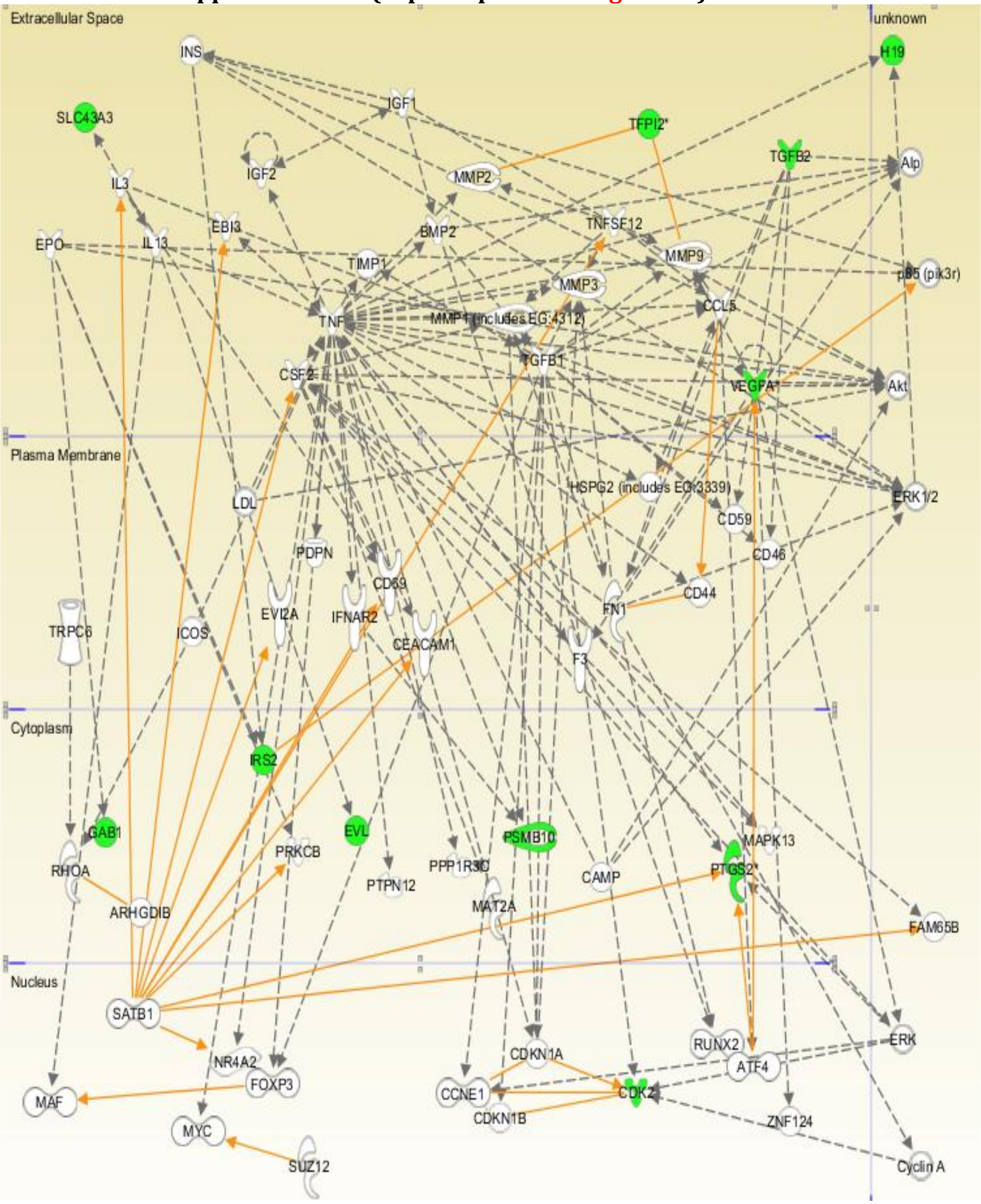


Figure 67 Molecules from N1 transcript-metabolite network superimposed on the N21 network (1).

6.3 Discussion

6.3.1 Principal Findings

In this study we analysed the transcriptome and metabolome profiles representing specific signatures of DD that have been shown in studies Chapter 4 and 5. The analysis identified several transcriptional pathways and implicated multiple regulatory networks that characterise and classify the different molecular subtypes.

Among the networks identified for the disease *vs.* healthy analysis (N21 *vs.* F21), two notable ones were found to be associated with molecular subtypes of autoimmune disease, connective tissue disorder and rheumatic disease. Network 1 revealed the three biological functions with 24 molecules associated with immunological disorder, 21 molecules associated with autoimmune disease and 15 molecules associated with rheumatoid arthritis. The first network, characterised by the SATB1, TNF, CSF2, TGFβ2, MMP1, TGFβ2 and molecules dominated by molecular signaling and possible cross-talking interactions (direct and indirect relationships) between transcriptional pathways. Landmark genes identified in previous studies associated or thought to be involved in DD and were also validated in this study (Table 20). In addition, other genes not previously described in the study following perturbation in disease are reported in Table 21.

The first goal of this study was to determine key metabolic pathways based on topological pathway analysis. The result from this analysis is a list of canonical pathways and gene sets relevant to progression in DD. In addition, the analysis indicates in which stage of progression a pathway is relevant: i.e. cord development pathways, response to hypoxia, and nodule development and also its response to hypoxia and pathways that are activated when normal cells are induced with hypoxia.

Mapping of the molecules associated with perturbation effect in healthy fascia (F1) and disease (N1) on to the network constructed for N21 *vs.* F21 analysis (disease) have highlighted some key switching patterns i.e. changes in molecules. Notably, upon perturbation of disease, key molecules that were identified as upregulated were either down regulated or no change in those molecules was observed. These opposing trends suggest (perhaps an ambitious statement) that careful and effective decrease of oxygen in nodules, may be indicative that some of these molecule can be restored to normal activity. Mapping of transcript data on metabolite networks has shown some good correlation at the two subcelluar levels depicting a change in transcriptome has a downstream effect on metabolome networks (and perhaps vice versa too).

Table 20 18 key molecules identified in this study previously confirmed in literature to affect connective tissue disorders including rheumatic disease.

Function Annotation	Molecules in disease (N21) networks	# Molecules
affects connective tissue disorder (18/18)	ADK, AMPH, ANKRD28, ARHGDIB, AUTS2 (includes EG:26053), CADM1, CAMK1D, CD44, CDK2, COL12A1, COL5A3, DAPK1, DCHS1, DUSP1, ESD, FCHO2, FRMD3, FTH1	18
affects rheumatic disease (16/16)	ADK, AMPH, ANKRD28, ARHGDIB, AUTS2 (includes EG:26053), CADM1, CAMK1D, CD44, CDK2, DAPK1, DCHS1, DUSP1, ESD, FCHO2, FRMD3, FTH1	16

Table 21 Previously reported functions of molecules highlighted to be of significantly DE upon perturbation in nodule.

Function Annotation	Molecules perturbed in disease network - **N1**	# Molecules
affects apoptosis of connective tissue cells (18/31)	BCL2, CASP3, CASP8, CDKN1A, CEBPB, DDIT3, DUSP1, FASLG, FN1, GAB1, JUN, MAPK14, MMP9, NSMAF, RAF1, RELA, TGFB1, TNF	18
decreases apoptosis of connective tissue cells (15/31)	BCL2, CD44, DUSP1, EP300, FN1, IL13, IL1A, IL1B, INS, RELA, SAA1, TGFB1, TGFB2, TIMP1, TNF	15
increases apoptosis of connective tissue cells (12/31)	CD44, CD40LG, EP300, FASLG, IFNG, JUN, LGALS8, MMP9, RAF1, TGFB1, TNF, VEGFA	12

6.3.2 Strengths and weaknesses of the study

The metabolites are interconnected through metabolic reactions, generally grouped into metabolic pathways [166]. Classical metabolic maps provide a relational context to interpret metabolomics experiments and a wide range of tools have been developed to locate metabolites within metabolic pathways. However, the representation of metabolites within separate disconnected pathways overlooks most of the connectivity of the metabolome [166]. By definition, the reference pathways cannot integrate novel pathways nor show relationships between metabolites that may be linked by common neighbors without being considered as joint members of a classical biochemical pathway [166]. The IPA is a proof of knowledge based comprehensive software of data analysis that can help researchers model, analyse, and understand the complex biological and chemical systems at the core of life science research [67]. By metabolomics analysis in the IPA, the phenotypic data of the metabolites can be validated and correlated with the targeted metabolites with the potential metabolic pathways, and related to biochemical functions.

An importantly successful application for metabolomics has been in diagnosing drugs-induced toxicities in kidney, liver and heart. Here we demonstrated the use of metabolomics screening combined with transcriptomics and IPA analysis as a novel usable strategy to characterise the biochemical perturbation induced by hypoxia, revealing several biomarkers resulting from this perturbation effect that could be suggestive of the potential metabolic pathways that can be targeted for future studies in higher, more complex systems i.e. primary cells from human samples with high biodiversity. This strategy confirms and strengthens the applicability of metabolomics analysis methodology to investigate the perturbation effect in DD phenotypes and it highlights the strong potential of the network analysis made possible by IPA.

The strength of being able to visualise possible interactions between molecules at the cellular level can be seen as a step closer to building constraint based models, or kinetic models or even abstract models to generate novel yet directed hypotheses. Once assembled, the model provides a means to organise the data about the DD system, to rapidly and inexpensively test hypotheses through *in silico* experimentation and to generate new hypotheses which can be tested in the laboratory. Not only can this kind of modeling approach transform raw data into actionable insights, but it can be an invaluable addition to an experimental program, that will allow researchers to ask more pertinent questions and to plan and design more focused experiments that have a much higher chance of discovering meaningful knowledge adding to existing knowledge and may translate into directed research; a step closer to understanding and targeting mechanisms in DD. The ultimate goal would be to integrate and process all these measurements to formulate mathematical models that recapitulate all previous observations and predict new behaviour together with environmental perturbations. Unfortunately time constraints have prevented the implementation of this.

Chapters 2, 4 and 5 utilised numerous computational methods to identify trends and patterns of gene expression and small endogenous and exogenous chemicals (from metabolome) specific to different DD cells, and healthy cells. These have led to the discovery of several genetic patterns or molecular signatures that aid in distinguishing biologically relevant aspects of tumour behaviour, their identity and some functional knowledge. It should be stressed that data integration from different data sources imposes major tasks such as including careful assembly of similar and complementary information from heterogeneous data sources and deletion of duplicated data. Such requirements demand considerable hand coded programming efforts, as different data formats have to be combined into a common schema.

6.3.3 Conclusions

In a systems approach, the various cellular networks are perturbed by changing environmental conditions of disease and healthy cells. The impact of the perturbation is assessed in by constructing key networks from the most significantly dysregulated molecules to evaluate any changes as measured and the consequences of these changes as they present themselves throughout the cellular networks. Key molecules in DD nodules network have been identified. However, the data still requires the development and assistance of new software tools and algorithms that can extract meaningful biological insights which yield in systems level understanding of cellular responses. Furthermore, models (e.g. kinetic models) from these data should be initiated. The application of high-throughput gene expression and metabolic profiling to the study of DD will now broaden our thinking about the biology of this quasi–neoplasia by providing deeper insights into the mechanisms underlying tumour promotion and progression.

Chapter 7

Conclusion and outlook

7.1 Thesis Summary

The sequencing of multiple genomes and the concomitant parallel development of computational SB tools start to realise the promise of functional genomics has created an opportunity for investigating DD in a way that was hardly imaginable 10 years ago. Today SB studies together with advances in molecular biological techniques and instrumentation allow for a directed, systematic effort aimed at producing a complete catalogue of biochemical activities, biological functions and their interactions, at least for simple unicellular life forms (e.g. bacteria and yeast).

In this thesis a major advancement to the application of functional genomics and metabolomics methodologies has been made to investigate human disease and this study for a higher organism is highly challenging, novel and original. The aim of this research thesis was to develop and test the hypothesis that DD is a network disease. This hypothesis was tested in two parts with the following questions in mind:

(i) The DD and corresponding healthy tissue differ in function through differences between their molecular and intercellular networks, rather than differences in a single molecule, or in a plethora of unrelated molecular species.

(ii) DD can be caused by any of a variety of perturbations in the regulatory networks that lead to the network differences.

With this in mind systematic studies were conducted employing a SB approach which involved making use of metabolomics and transcriptomics methodologies to enable the profiling and characterisation of signatures from DD phenotypes. The major innovation presented in this thesis is the use of metabolites and gene (transcript) sets in a data-centric mechanistic modeling approach for the construction of potential networks that may be associated or involved in tumour progression rather than single genes. Previous research in the field of DD where single genes were studied have provided a wealth of important information and these molecules may now be regarded as markers of certain stages of disease progression. While it is important to identify these individual genes as biomarker landmarks, a broader understanding of the functional biological processes occurring during disease progression has been missing.

Before we can interpret any data related to physiological and pathological transitions in DD, a question one should answer is, “What is normal?” as for any system under study, there exists a requirement to understand the magnitude and diversity of its typical activity whether in metabolism, gene expression or influence of its environment in the unperturbed state. Recent research in DD progression has taken a step closer to gaining insight into biologically related gene sets from the transverse palmar fascia, but none to date have explored the profound detrimental effects of subcultivation on their unique molecular (biochemical) signatures.

In Chapter 1 an introduction to this thesis is given presenting the approach and hypotheses tested in this study. Section 1.2; Appendix A reviews scientific literature available on DD and discusses the rationale to undertake a SB approach to investigate DD as it proposes it to be a disease of networks. The techniques, instruments and algorithms used to test these hypotheses are given in Chapter 2.

As it is important for SB to be employed on defined and reproducible experimental (and computational) systems, Chapter 3 examined how DD and healthy cells changed following excision and cultured over a period of time. This highlighted the optimal conditions to further investigating the DD system, as a compromise between retaining *in situ* DD character and having sufficient cell numbers for analysis; the concept of a systems signature defined through FT-IR spectroscopy was applied. Using this procedure, the reproducibility *in vitro* of DD subsets (i.e. nodule, cord, fat & SON) was compared with internal control (transverse palmar fascia) and external controls (carpal ligamentous fascia) based on their unique biochemical signatures. The results indicated that different DD phenotypes exhibit marked variability in the overall pattern of metabolic fingerprints (nodule, cord, skin and fat). The results from the metabolic fingerprinting based on PCA of DD nodule, cord and fascia has implied an early passage number (0-3) would faithfully represent the test subject as its phenotype would be closest to the *in vivo* state. Due to uneven sample size (and cellular content), certain supervised methods such as ANN could not be applied to the model to assess validity of the model, but PCA and PC-DFA indicated good clustering and separation between different sample types (from disease and healthy and also between samples from highly biodiverse individuals i.e. patients). The application of FT-IR spectroscopy conducted under carefully controlled conditions with appropriate chemometric techniques to differentiate phenotype between DD and control fibroblasts has been demonstrated to be a powerful tool for the rapid screening and discrimination between the anatomical cell types within individual DD patients as well as samples from controls. This study highlights early passage cultures are close representatives of the metabolic fingerprints of those in disease phenotype state *in vivo* and are more appropriate for further DD studies. However, PCA of FT-IR footprint spectra did not yield any statistically significant results. This may be because the culture medium in which secreted metabolites were collected was nutrient rich and contained undefined reagents including FBS. The study supports the hypothesis that the cell culture monolayer environment may alter the functional characteristics of the DD samples, possibly by abiotically selecting against a subpopulation of cells which survived the in *vitro* conditions, and result in degradation of the sample's phenotypic identity. 3D cultures may improve over cell monolayer culturing and should be considered in future studies.

The classification achieved is encouraging, as FT-IR spectroscopy not only discriminates DD phenotypes from the two controls, but also between the two fibrotic elements i.e. nodules from cords, fat and SON. Despite high biodiversity in patients, this trend could be observed in most patients when examined alone, if not altogether in the PCA scores plots when combined. In addition this technique demonstrates internal fascia is an appropriate control and can be distinguished from diseased fibroblasts using chemometrics techniques. The use of internal transverse palmar fascia as the control will attribute to the homogeneity and consistency in future studies.

One problem with the current implementation of the PCA model is its unsuitability to assess multiple influential trends (too many variables - as this could lead to poor generalisation performance of the model due to under fitting). Given the clinical nature of this project, samples sets are often too few. Any experimental error reducing the already few numbers interferes with the chosen models applied and cluster analyses, making this a biased data set depicting an unbalanced sample number and experimental design. Alternatively, it may be necessary to apply other supervised methods that complement the existing *in silico* model if it is to be used on larger data sets. This may involve a slightly different optimisation technique.

Future studies should consider the following to improve and optimise the current cell culture methodology: culture and harvest a greater total amount and full sets of DD samples e.g. $n = 20$ nodules, $n = 20$ cords, $n = 20$ transverse palmar fascia. Initial confirmation of phenotypic identity, morphological assessment and fibroblast purity, should be obtained using immunocytochemistry and/or immunohistochemistry. Furthermore, alternative sample processing and extraction methods may be compared for future metabolomics studies e.g. snap freeze biopsies in methanol or subject samples to FT-IR spectroscopy immediately from culture flask and not thaw from frozen (but this would not be valid for different time points (days) as technical variation could be introduced) to make a comparison of current methods using chemical information from formalin fixed tissues.

To validate the current results, that a 'passage effect' is one possible effect on the fibroblast cultures, an independent validation of the trypsin and passage effect should be applied where cells should be isolated from the test tissue from a minimum of three patients and passaged a total of six times (Passage 0 to Passage 6). Using qRT-PCR, expression of all genes of interest (approx 10-12 initially) selected from bioinformatics and text mining analysis should be quantitatively assessed (e.g. (collagen type I ligament/tendon's main matrix constituent), collagen type III, fibronectin, metalloprotease-13 [MMP-13], and tissue inhibitor of metallopreotease-1[TIMP-1]). An alternative approach could be to determine the passage effect by simultaneous growth of fibroblasts from two sets of biopsies cut in half arriving on the same day and grown in different sized culture flasks and passaged upon confluence. For example (one piece of nodule tissue, cut into two and processed for fibroblast cell culture. one pellet transferred into a T25cm^2 flask and the other into a T75cm^2.

Here the use of natural language processing (NLP), text mining and machine learning methodologies would be of considerable utility if one could train these to represent more quantitative, effective and efficient ways to capture data generated from SB experiments. Due to advances in high-throughput -omics technologies, e.g. gene expression data from microarray experiments and protein interaction databases, together with large volume of digital textural documents such as PubMed, an unprecedented opportunity is created to apply computational techniques for a comprehensive study of the structure and dynamics of the systems components, and thus provide a robust foundation to systems biologists. It must not be understated that despite the high quality obtained through manual curation, NLP and machine learning methodologies represent a more effective and efficient way to capture and curate large data sets resulting from iterative SB experiments. Though efficiencies can be built in to the manual curation systems, it cannot be compared to the speeds achieved by NLP tools. Automated extraction of data using NLP technologies is fast but the accuracy of the data captured and the data points that are omitted comprise major areas that need to be improved. In principle, the requirement is the discovery of new knowledge from hidden texts that will help with recognition of extracted **relations** from **entities** that are shown to be truly interesting and not merely erroneous trends that could occur from the processed data.

In Appendix E2(3), a novel method (application) under construction which may later be incorporated into an existing text mining tool (FACTA) [167] is proposed. This addition to the existing tool would play a crucial role in SB research and should allow the end user to make decisions and perform text processing and clustering of selected categories of interest. Various discussions with Dr Philip J Day arising from this study, have led to the initial design and set-up of an application which may facilitate in quantitative and qualitative selection of candidate genes (following reproducibility study in Chapter 3) to validate from the literature. The project involves chunking/parsing/zoning of data sets from DD specific methodologies from which entities (proteins, genes, and chemicals) are retracted. The strategy involved retrieval of these entities from the literature based on an ordered ranking system that would specifically identify the technique (e.g. genes discovered through a microarray study or immunohistochemistry) from which the molecule was discovered. Furthermore, which group (authors, laboratories) report these findings and to gain a deeper understanding the sub sections in a given article e.g. abstract, method or results section of an article would enable the users to select and filter a sub group of entities based on their individual experimental predictions/queries as identification of entities from irrelevant sections in a article could be avoided e.g., molecules identified in the results section of a peer-reviewed journal than found by chance say in the methods which could lead to ambiguity in novel findings.

The workflow is briefly shown in Appendix E3 where I started to build the reference collections (libraries) from existing and known information retrieval sources such as PubMed, Scopus, Science direct etc. This consists of full articles on 1) DD scientific and literature, and 2) entities (i.e. techniques from which these proteins, genes, metabolites were discovered to obtain information directly from domain experts or from biological annotation databases. Here one of the challenges was to determine the optimal queries for selecting all relevant annotations from various biological resources, to cope with the different granularities of heterogeneous information types and the incompleteness of manually curated information. The idea is once the entities in sentences were parsed/chunked/curated etc the completed ‘mini tool’ would then be connected with the larger tool, FACTA, to enable prediction and inference of entities relevant to DD biology, gene expression and specific techniques the genes were inferred from using this test model (in this case the idea was selection of candidate genes from text mining to validate results from Chapter 3).

In Chapter 4 the above two hypotheses were tested with DD and corresponding healthy and normal metabolomes. One hypothesis was, whether the difference in disease cell types (nodule, cord and SON) and control cell type is the same as the difference in control fibroblasts (transverse palmer fascia) cultured in normoxia and hypoxia. Secondly, in which specific disease cell type (i.e. nodule or cord) is the difference with normal cells in intracellular metabolome the largest? Thirdly, the Warburg effect was tested by inducing hypoxia in the disease cell types. The study employed GC-MS for metabolic profiling and identification revealed a few yet important metabolites that were significantly dysregulated in disease compared with fascia. Supervised analysis methodologies such as DFA and ANOVA-PCA were employed in addition to unsupervised methods (PCA) to make inferences from the mass spectral data. Metabolites involved in amino acid metabolism were significantly down-regulated in nodules including leucine, phenylalanine, cysteine, aspartic acid and a sugar. Leucine and a sugar were also significantly down-regulated in cords, in addition to these, metabolites from carbohydrate metabolism, cofactors and vitamin metabolism pathways i.e. glycerol-3-phosphate (up-regulated) and pantothenic acid (down-regulated) respectively were identified.

The question whether the metabolites dysregulated in fascia upon perturbation are the same as those dysregulated in nodule and cord is difficult to address from this analysis alone, as only cysteine and aspartic acid (both down-regulated) were common to both. Whether this was a true correlation or by chance alone has been investigated in the transcriptomes of these subjects addressed in Chapter 5. Furthermore from Table 6, the perturbation effect caused by inducing hypoxia to disease cells revealed a large numbers of metabolites that were significantly up/down-regulated. These were mostly involved in amino acid metabolism and also carbohydrate metabolism. Whether this correlation could also be observed at the transcriptome level was then further investigated in Chapter 5. Total RNA from DD nodules and fascia from 1% and 21% (from three patients) were used as the maximum difference in metabolites numbers was observed in nodules than cords. Since the SON samples were a different cell type from all others (fascia *vs.* epidermis in PCA analysis), it was excluded from the transcriptome analysis.

One of the major strength of this study was the novel approach to harvest, extract both metabolites and RNA without affecting the RNA integrity. This method was previously attempted in neuroblastoma cells lines [132] but an improved method had been implemented here. The extraction steps were in total three (not four as previously described) at 4^{o}C and the RNA quality was preserved. The co-extraction of transcripts and metabolites is believed to be important because the amount and integrity of both mRNA and metabolites are unaffected from the dual analyte extraction procedure.

In Chapter 5 the DD nodules were compared with normal and perturbed fascia, to investigate the Warburg effect in the DD transcriptome. The study employed Affymetrix Human Genome U133 Plus 2.0 GeneChip oligonucleotide arrays to determine transcript profiles of fibroblasts cultured in normoxic and hypoxic conditions. The study revealed a small number of DE transcripts that were common in N21 *vs.* F21 and F1 *vs.* F21 analysis. These transcripts were involved in the following pathways: - MAPK signaling pathway, ECM-receptor interaction, p53 signals pathway, tyrosine metabolism, nicotinate and nicotinamide metabolism, phenylalanine metabolism and vitamin B6 metabolism. Since so many DE transcripts were identified from microarray analysis it is difficult to assess whether these correlations were true or by chance. This was further explored through network analysis using IPA in Chapter 6. The common DE genes can be seen in Figure 44b.

The perturbation effect in DD nodules caused by hypoxia was also examined to address whether the dysregulated metabolites also can be measured by this study to correlate at the transcriptome level. Further we would like to address which of the identified candidate DE transcripts/genes and metabolites from Chapter 4 show congruence across the various levels of systemic description in the context of pathways i.e. the DD metabolome and transcriptome. Integration of these complex data sets has been a key challenge and is the objective of Chapter 6, which involved mapping molecules from one analysis onto disease networks of known genome wide pathway maps. Several methods for microarray data analysis and metabolite datasets were explored. As the vast number of results from the microarray study resulted in an abundance of DE transcripts, three normalization methods complemented by two Bayesian methods; limma and puma were used to make statistically significant inferences from the data. The limma method was found too stringent while the puma algorithm allowed for maximal inference from the data. Based on the puma method, little synchrony in DE transcripts were observed from N21 *vs.* F21 and F1 *vs.* F21 analysis gene lists. From the unique 119 DE probe lists identified in N21 *vs.* F21 analysis, the following pathways were highlighted:- cell adhesion molecules, ECM-receptor interaction, neurotrophin signaling pathway, nicotinate and nicotinamide metabolism, cysteine and methionine metabolism, ether lipid metabolism, selenoamino acid metabolism, tryptophan metabolism, tyrosine metabolism, valine, leucine and isoleucine degradation and vitamin B6 metabolism. A number of these pathways are amino acid metabolism pathways, and so we can confirm there is a strong correlation in DD at both the transcriptome and downstream in the intracellular metabolome. From the perturbation effect in N1 *vs.* N21 fatty acid metabolism, toll-like receptor signaling pathway, biosynthesis of unsaturated fatty acids, PPAR signaling pathway, citrate cycle (TCA cycle), glycine, serine and threonine metabolism pathways were enriched. Again this strengthened our hypothesis that DD is a disease of networks, where molecules are interconnected and a number of amino acid metabolism molecules are actively DE in DD nodules and this perturbation effect is observed at both the transcriptome and metabolome levels within a cell.

The same analysis methods were not applicable to metabolomic data sets as by contrast the variables (metabolites) are far fewer and appropriate chemometrics approaches were employed. If both ‘omic data sets could be normalized using the same methods then statistical significance and congruence between the identified transcript and metabolites would be more powerful to facilitate interpretation in terms biological relationships. At present the two ‘omics studies must be analysed separately using different yet most suitable methods.

New data analysis methods need to be derived to integrate these heterogeneous data sets. A correlation between protein and metabolites is being made [168]; these molecules can be identified through mass spectrometry. This is not possible for transcripts. To enable data integration Ingenuity Systems now allow incorporation of small molecules (HMDB ID’s and CAS registry numbers) in addition to gene/transcript/enzyme data for network analysis of network eligible molecules present in the Ingenuity Pathways Knowledge Base. In Chapter 6 we first integrate the candidate metabolite with the candidate transcripts for each pairwise analysis, and then map these networks onto networks generated for N21 *vs.* F21 analysis.

Using the currently available technologies, we have intended to address the overlap between transcripts, and metabolic activity altered by oxidative stress using DD fibroblasts as an experimental system. As environmental factors are considered to play a role in the development of DD, data from proteome and epigenomes should also be studied within a SB framework.

A major conclusion from this correlation analysis is that (from Tables 6,7, 11 and Figures 44, 58, 66 and 67), it relatively few metabolites/transcripts identified as significantly dysregulated in F1 *vs.* F21 analysis were present in N21 *vs.* F21 analysis. From these systematic analyses it is now appropriate to challenge the first hypothesis that difference in disease and healthy cells maybe akin to the differences in healthy cells in normoxia and hypoxia. As only a very small number of significant molecules coincide in F1 and N21 candidate lists. In Chapter 6 these significant candidates (i.e. metabolites and transcripts) from Chapters 4 and 5 were integrated to visualize relationships using network analysis (e.g. candidate metabolites and candidate (prominent) transcripts in F1 *vs*. F21 using IPA and these networks are mapped on the networks from significant molecules in N21 *vs.* F21 analysis. From Figures 66 (F1 candidates) and 67 (N1 candidates) superimposed onto Figure 58 (N21 *vs.* F21 candidate molecules), this can be observed and that hypoxia has an opposite effect, few molecules were in common and SATB1 being one of them, was upregulated in F1 as well as N21.

It should be emphasized, here that the above conclusion is based on a correlation analysis only and not considering the biological relationships between transcripts, proteins and their downstream metabolites. In depth knowledge of cell biology would greatly facilitate further examination of these molecules in the respective networks and also knowledge of which metabolites are related to certain proteins.

In this study we have also validated the hypothesis that changes of metabolite levels upon perturbation would reveal changes at the transcript levels (or perhaps vice versa). This has been observed both in healthy and disease phenotypes. Large and quantitative transcript and metabolite data was acquired from techniques currently allowed. However until the relevant biology is uncovered there are concerns regarding the interpretation expressed above. Following-up with future experimental work will ratify these conclusions/hypotheses.

Through this thesis some of the technical challenges associated with working with complex organisms (i.e. human cells) using a SB approach are addressed. These successfully addressed aspects are:- (i) system wide component identification and quantification ("omics" data) at the level of mRNA, and small molecular weight metabolites; (ii) experimental identification of physical component interactions, for information processing networks; (iii) and integration of heterogeneous data sets. The next major challenges are to address which computational inferences ought to be made from the structure, type, and quantity of component interactions from these data. This step is essentially required to achieve a holistic, quantitative, and predictive understanding through mathematical models that will enable an iterative cycle between prediction and experiments, the hallmark of systems biology. In addition a major challenge is to understand heterogeneity both at the biological and cellular levels (e.g. patient, technical, biological variability).

Finally future work that could be pursued from this thesis would involve the following as illustrated in Figure 68 (i.e. the next steps from page 42):-

1. Reconstruct integrated cellular networks into an *in silico* model.
2. Reconcile the experimentally observed responses with those predicted by the model.
3. Design and perform new perturbation experiments to distinguish between competing model hypotheses.

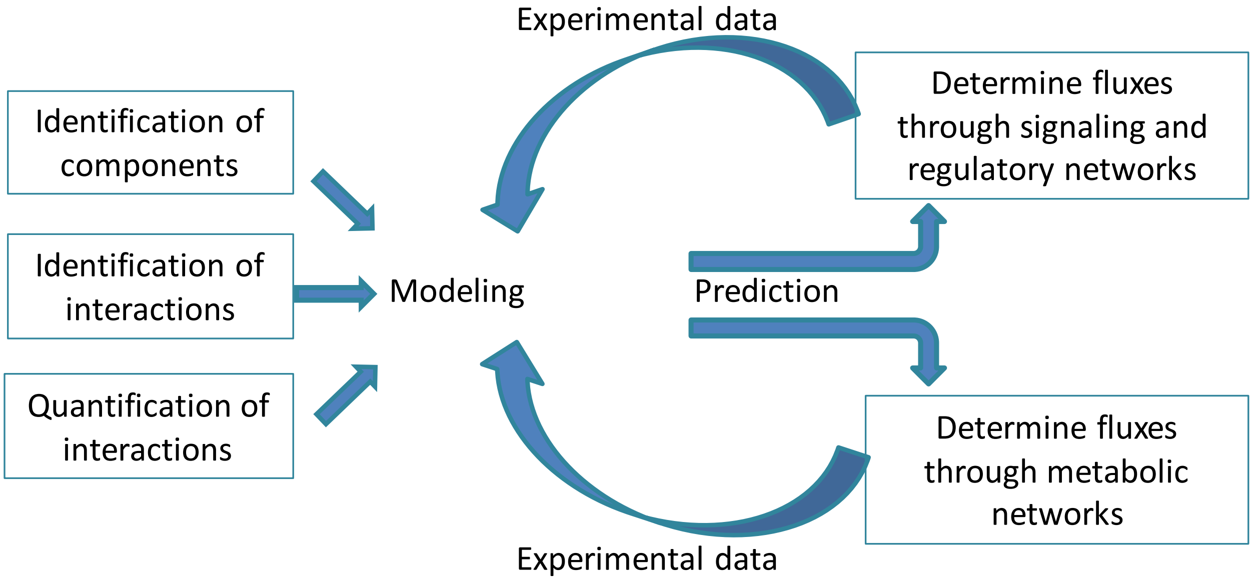

Experimental data
Identification of components
Identification of interactions
Quantification of interactions
Modeling
Prediction
Determine fluxes through signaling and regulatory networks
Determine fluxes through metabolic networks
Experimental data

Figure 68 Future outlook for a DD systems roadmap. The comprehensive component concentrations identified in this thesis through omics studies provide input data for inferring component interactions using computational methods. The challenge is for computational modeling methods yet to be developed to enable prediction of the functional network state from the concentrations and to infer the information processing network that controls the functional state.

In conclusion, it should be recognised that increasing sample replications is not always possible, particularly for clinical samples where patients are under treatment and any new diagnostic test is a secondary goal. Therefore, there needs to be greater emphasis on more robust experimental design. For example, to avoid sample bias, adequate matching of patients with disease to those who are healthy, in terms of gender, age, BMI, and ethnicity, as well as other extrinsic factors such as metabolic rates, pharmaceutical use and diet. In addition, if future studies include females or patients with pre-existing medical history of other disease e.g. diabetics then additional factors such as diurnal effects should be considered. All of these factors will significantly influence the subject’s physiological profile, and these need to be carefully considered when generating a robust experimental design.

This novel study has demonstrated a major advancement in understanding DD. Systematic hypothesis driven studies have been performed to investigate the DD state dynamics. A number of dysregulated metabolites and transcripts involved in amino acid metabolism, carbohydrate metabolism and also metabolism of cofactors and vitamins have been identified from these integrative analyses. Upon perturbation several of these transcripts and metabolites involved in the mentioned pathways were significantly dysregulated. For the first time, early passage numbers are shown to provide representative metabolic and transcript fingerprinting for investigating DD. A parallel analysis of transcript and metabolic profiles of DD fibroblasts is performed and the parameters are correlated across the various levels of systemic description. This will now enable us to examine the extent to which systems biology helps investigate pathological mechanisms in DD other related connective tissue disorders. This thesis should provide a fundamental change in DD research.

APPENDIX

Appendix A

1.2 Literature Review

Dupuytren’s – a Systems biology disease?

Dupuytren's: a Systems Biology Disease?

Samrina Rehman[a], Royston Goodacre[b], Philip J. Day[c] Ardeshir Bayat[d] and Hans V. Westerhoff[a,e*]

[a]Manchester Centre for Integrative Systems Biology, [b]School of Chemistry, [c]Quantitative Molecular Medicine Research, CIGMR, [d]Plastic and Reconstructive Surgery Research, Manchester Interdisciplinary Biocentre, University of Manchester, Manchester, UK, EU

[e]Netherlands Institute for Systems Biology, VU University Amsterdam, EU

*Hans.Westerhoff@manchester.ac.uk.

Contents

Abstract

Dupuytren’s disease (DD) is an ill-defined fibroproliferative disorder affecting the palms of the hands of some Northern Europeans. Cellular components and processes associated with DD pathogenesis include cytokines, growth factors, adhesion molecules, and extracellular matrix components. Free radicals and localised ischemia may trigger the proliferation of DD tissue. Histology has confirmed the presence of collagens, myofibroblast and myoglobin proteins in DD, but at widely varying abundances. Several genes are over- or under-expressed in DD tissue. Although the existing data sets contain potentially valuable (though largely un-interpreted) information, the precise aetiology of DD remains unknown. We question whether seeking further information about DD in ways that assume the disease to be due to a single faulty gene, is likely to lead to a breakthrough. Thus we propose that DD is in fact a disease of several networks rather than of a single molecule per se, and show that this accounts for the experimental observations. We outline how DD may be investigated more effectively by employing systems biology which considers the disease process as a whole rather than focusing on specific molecules.

Key words

Dupuytren’s disease/contracture, systems biology, transcriptomics, gene expression profiling, metabolomics, proteomics, modelling, free radicals, myofibroblast, multifactorial disease.

Introduction

Although Dupuytren's disease (DD) is commonplace in General Practice and despite decades of both experimental and clinical investigations into the disease, its precise aetiology remains unclear. DD resides within the poorly understood, yet important category of superficial quasi-neoplastic proliferative fibromatoses [1]. Phenotypically, it is a nodular palmar fibromatosis. It often causes permanent flexion contracture of the metacarpophalangeal (MCP) and proximal interphalangeal joints (PIPJ) of the digits [2] (Figure 1) leading to loss of function, discomfort and deformity of the hand. Ultimately, this then leads to a permanent contracture of the involved digits [2-6].

DD is a non malignant disease as it does not metastasize [7]: it may invade locally within the palmar aponeurosis of the hand (sparingly supplied with blood vessels as compared to other parts of the body) but it does not disseminate to other tissues. DD behaves as a benign neoplastic disorder: DD is progressive and irreversible with a high rate of recurrence after its main treatment, i.e. surgical excision [8]. The increasing severity and recurrence can lead to amputation of the affected digit [9-11]. A thorough understanding of the diagnostic, prognostic and theragnostic indicators affecting the disease is therefore important.

The three stages of DD growth (proliferative, involutional, and residual) involve myofibroblasts [12-15]. DD is associated with abundance of collagen, fibronectin, integrins, cytokines and many other growth factors [2, 9, 16-18]. Differential gene expression microarray studies and biomarker identification strategies found the expression of several genes correlate with DD [19-28]. Nevertheless, none of these studies has established the molecular mechanism of DD formation and hence the molecular basis of its diagnosis and therapy is not in sight. This is unlike diseases such as cancer, where at least the oncogenes have been identified, albeit not as single causes [29].

Systems biology is one of the most widely discussed fields among the emerging post-genomic disciplines [30]. Its bottom-up variant combines mechanistic modelling with quantitative experimentation in studies of genetic networks, signal transduction pathways, metabolic networks and their integrations [30-37]. It aims at understanding how the interaction of multiple components within a cell, tissue, organ or indeed individual leads to much of the biological function. The mathematics and the quantitative experimentation are not aims in themselves but merely serve to address the full complexity and emergence of biological phenomena at all functional levels within the system being studied (from cell to ecosystem) [30].

Systems-level approaches are beginning to make some pace towards scientific comprehension of pathway control, regulation, and function [38-41]. This has improved the understanding of some diseases [42], and has provided new rationales for drug discovery [43-45]. More than 1800 publications on DD have appeared since the original publication by Guillaume Dupuytren in 1831. The field of Systems Biology is growing exponentially with more than 8500 papers to date. Many characteristics and the complex biological behaviour of DD fibroblasts may constitute an invitation to a systems level approach to DD at both the cellular and the molecular level. In this review, we outline such an approach.

Review criteria

PubMed and Scopus searches were conducted to identify all relevant scientific literature evaluating the pathogenesis of DD, published in English to in or before June 2010 (with the addition of early literature in French). Studies were confined to cytogenetic studies, histological findings, genomic approaches and biomarker-related discoveries of biomolecular components involved in the pathogenesis of DD. The keywords used in various combinations included (* used as wildcard truncation): Dupuytren* disease, contracture, gene expression, microarray*, chromosome, cytogenetic*, *array, treat*, therap*, manag*, surg*, excis* fibroblast*, tissue* biopsie*, cel* plus character searches for animal or *in vitro* studies and studies evaluating prophylactic treatment. Review articles were used as additional sources for primary papers by cross-checking the reference sections with the master list of compiled articles. Full text articles were retrieved for those studies in which subsequent findings had been reported from the same research group. The methodological quality of each paper was examined (for example, whether statistical data analysis was sufficiently detailed to allow for reproducibility). Each study was assessed in terms of the direction of the conclusions vis-à-vis the phenomenon investigated, whether positive, negative or inconclusive.

Dupuytren's disease and its many faces

Connective tissue fibrosis

Human skin exhibits a remarkable diversity in structure and function across anatomic sites [46]. Fibroblasts are the principal cells of this stromal tissue. They are responsible for extracellular matrix (ECM) synthesis in connective tissues and thereby play a vital role in tissue repair and wound healing [7, 47]. In the various connective tissues where they occur, they constitute a heterogeneous population of cells [47-49]. Many diseases are associated with fibroblasts, either through fibroblasts being implicated in their aetiology or because of the fibrosis consequent to damage to other cell types. Fibrosis is a consequence of excessive synthesis and deposition of collagen by abnormal fibroblasts [27]. The ensuing excessive ECM accumulation is a common feature of many connective tissue diseases. Fibrosis can affect not only the skin but also internal organs such as lungs and kidneys, leading to organ dysfunction and failure. Clinical examples include renal interstitial fibrosis [50], scleroderma, sarcoidosis [51], idiopathic pulmonary fibrosis [52], retroperitoneal fibrosis [53] and DD [54]. A complex network of intercellular interactions involving a diverse range of molecules, including growth factors, cytokines, chemokines and endothelin may drive the pathological events that ultimately lead to uncontrolled connective tissue fibrosis.

Pathophysiology

The pathophysiology of DD is thought to arise either from a defect in the wound repair process or from an abnormal response to wounding. These hypotheses are based upon biochemical characterisation of affected tissues showing increased deposition of collagen III relative to collagen I and increased levels of collagen hydroxylation and glycosylation [55]. We note that these hypotheses are non-molecular and may associate the disease with effects, rather than with a single cause. The following sections will reflect upon the current understanding of DD pathophysiology.

The myofibroblast

An early attempt at a functional interpretation of the histopathological changes observed in DD settled on the assumption that the cellularity (quantified as the cellular density) of the DD nodules (see below) was indicative of the activity of the disease [6]. Later, the disease was classified into three stages: proliferative, involutional and residual. The diseased tissue was further subdivided into the essentially fibrous nodules, reactive tissue, and residual tissue.

Normal palmar skin is similar to the skin of the sole, and differs from skin covering the rest of the body. Essential to the dermis are the fibroblasts with spindle-shaped cell bodies and nuclei that produce fibres, which can be seen by light microscopy. Collagen is the most abundant constituent of these filamentous components. Further investigations into the ultrastructure of DD tissue revealed the presence of myofibroblasts at various abundances. These specialised mesenchymal cells express smooth muscle α-actin, in which an acquisition of a smooth muscle like function may explain the contractility observed in DD [13].

Cell heterogeneity

Fibroblasts are identified by their spindle-shaped morphology. They possess the ability to adhere to plastic culture vessels, and are thus identifiable upon the absence of markers for other cell lineages. The presence and role of myofibroblasts in DD have been studied by many independent authors, mostly revolving to similar conclusions, which are summarized below.

Ultrastructural studies confirm that DD tissues can be classified into several stages according to their cellularity [6]. These different stages may co-exist in the same specimen, with different cell populations at maximum density in each stage. DD contains two structurally distinct fibrotic elements; the nodule and the cord. The nodule is described as a highly vascularised tissue containing a large number of fibroblasts, with a high percentage being recognised as myofibroblasts due to their expression of the α-smooth muscle actin. The cord is relatively avascular and acellular; collagen-rich with few myofibroblasts. The haematoxylin and eosin stained DD sections shown in Figures 2(A) and 2(B) demonstrate large numbers of relatively cellular (nodules) and relatively a-cellular, tendon-like regions (cords). There are different opinions regarding the origin and development of this aspect of the DD phenotype, viewing either the nodule as developing into the cord as the disease progresses over time or, the two structures representing independent stages of the disease.

The presence of myofibroblasts in DD has inspired the histopathological literature [56-58]: Their morphologic characteristics of being both a fibroblast and a smooth muscle cell might be related to the contraction involved in DD. In terms of their ultrastructure, DD myofibroblasts resemble myofibroblasts of granulation tissue thought to be responsible for contraction during wound healing. The Dupuytren myofibroblast synthesizes fibronectin, an extracellular glycoprotein that connects myofibroblasts and connects them to the extracellular stromal matrix through an integrin. In a study of 43 cases of DD tissue, histopathological changes suggested that certain growth factors may induce proliferation of such genetically abnormal myofibroblasts [15].

Not all palmodigital aponeurotic structures are affected by DD. The areas that are affected macroscopically, do so at an irregular depth and distribution, with the more superficial layers and ulnar side of the palm being affected most. Macroscopically, neither the deep retinacular tissue that includes the transverse palmar ligament or fascia also known as “Skoog’s fibres” nor the fibrous flexor tendon sheaths, appear to be involved in DD.

Recent advances in microarray technology and bioinformatics have revealed an appreciable cell-to-cell heterogeneity: According to genome-wide gene expression profiles, fibroblasts come in various subtypes [48]. In one study, fibroblast samples were clustered on the basis of the expression levels, using the Partitioning Around Medoids algorithm [59]. This identified diverse sets of genes being expressed in the different subtypes. The authors proposed that different anatomic sites have characteristic, distinct phenotypes, which persisted *in vitro* even when fibroblasts were isolated from the influence of other cell types. They termed this phenomenon ‘topographic differentiation’. We note that expression differences at the level of mRNA need not necessarily lead to functional differences, as control and regulation of cell function also involves other levels of cellular organisation such as translation and posttranslational modification [40]. Further substantiation that the differences in mRNA expression correlate with differences at the proteomics, metabolomics, or morphological level should be welcome [60]. Chang *et al.* [48] did not find such differences when they evaluated cultured fibroblasts from diverse DD sites but all with the morphology of elongated, spindle-shaped cells. Immunofluorescence microscopy showed that the fibroblast cultures were uniformly positive for a mesenchymal marker, but negative for markers of epithelial, smooth muscle, endothelial, perineural, and histiocytic cells. The study revealed that the different passages of the same fibroblast culture clustered with each other, indicating that their in vitro phenotypes were stable. One aspect of the topographic genomic program in fibroblasts may be the coordinate regulation and synthesis of the ECM proteins in a site-specific manner. Taken together, the information suggests that fibroblasts with the morphology of elongated, spindle-shaped cells may themselves be more homogeneous and that the heterogeneity observed by Kaufman *et al.* [59] concerned fibroblasts differentiated to morphologically different subtypes.

Growth factors

Table 1 lists the components that have been implicated as modulators of the DD fibroblast transdifferentiation into myofibroblasts. Possible roles of these growth factors in DD pathogenesis have been discussed in previous reviews [61] and we summarise just a few [62-70]. Among the cytokines, TGF-β is thought to be a significant inducer of myofibroblast transdifferentiation because of its ability to up-regulate α-smooth muscle actin and collagen in fibroblasts, both *in vivo* and in vitro [66].

Linkage analysis

A study performed in a five generation Swedish family suggested that DD was inherited in an autosomal dominant pattern [71]. Mitochondrial and X-linked inheritance of this dominant factor in this family were ruled-out because of the male to female and male to male transmissions of DD. Linkage analysis implicated a single region of approximately 30cM on chromosome 16 bounded by microsatellite markers D16S3131 and D16S514, and produced a logarithm of the odds (LOD) score >1.5. Genotyping of individuals made up of four siblings affected by the disease but from another branch of the family together with the use of additional microsatellite markers supported linkage to that region and produced a maximal LOD score of 3.2 for D16S415, with four other markers producing LODs of >1.5. Linkage was further restricted to a single 6cM region between markers D16S419 and D16S3032 on chromosome 16.

When a disease is dominant, it is likely to be caused by a single allele of a single gene, and by the molecule it encodes. From this perspective, the above findings would suggest that DD is a single gene disease. However, to date, the linkage to a single gene has not been reported up to an LOD that is much more significant than the marginal value of 3 in this Swedish study and the penetrance in this study was incomplete. In addition, the disease develops at an advanced age, there are many more sporadic cases of DD, and there are few such families for which the genetic analysis has been performed.

Other studies have shown assocation of the disease with other loci, including a positive association with HLA-DRB1*15 on chromosome 6 in Caucasians [72]. A study of 20 British DD patients with a maternally transmitted inheritance pattern demonstrated a mutation within the mitochondrial genome (mitochondrial 16 S ribosomal RNA region) in 90% of patients [73]. The defective mitochondria generated abnormally high levels of free radicals and induced defects in apoptotic mechanisms, and might hence directly participate in the pathogenesis of the disease.

Free radicals

Oxygen free radical production has been proposed to be one of the many factors contributing to tissue damage in DD [74]. A relation between localised ischaemia, superoxide free radicals, hydrogen peroxide, hydroxyl radicals and DD was projected from this study in which palmar fascia from 10 individuals who had DD were subjected to 0-60 minutes of tourniquet ischaemia. Palmar fascia obtained from 10 suitable control patients (having carpal tunnel decompression) were subjected to the same insult. The concentration of hypoxanthine was 6-fold higher in Dupuytren's palmar fascia as compared to the control palmar fascia. The suggestion was that before the ischaemia and in DD patient tissue more than in non-DD patient tissue, xanthine oxidase present in the endothelial cells around the small vessels was converting hypoxanthine into xanthine and perhaps further into uric acid, and this implies that measuring metabolites directly in tissue could be an important step to understanding the network of events involved in DD. Both these steps are catalyzed by xanthine oxidase, with super oxide free radicals and hydrogen peroxide as by-products. This mechanism is illustrated in Figure 3. These free radicals would damage the perivascular connective tissue, with fibroblasts attempting to repair the damage. The free radicals might directly stimulate proliferation of fibroblasts, as upon addition of free radicals to fibroblast cultures from DD palmar fascia, higher concentrations of free radicals led to toxicity, but lower concentrations stimulated fibroblast proliferation. This group also suggested that the observed increase in collagen type III might be a result of fibroblast proliferation [74].

Hypoxanthine was more abundant in nodular areas than in the tight fibrous cords. Based on these studies it is speculated that microvessel narrowing, leading to localised hypoxic conditions may be one cause of DD, secondary to age, smoking and other environmental factors. Although this is a crisp hypothesis, most of the extensive histologic and biochemical studies on DD continue to be controversial. Examples include studies on the types of collagen present in DD cell or tissue, the presence of the myofibroblast and factors such as vascularity including microvascular and macrovascular contributions.

Animal models

No animal model exists for the study of DD fibromatoses. Yet, results of animal studies of possibly related diseases may be informative. An attempt to explore the level of basic fibroblast growth factor (bFGF), a known angiogenic factor in dermal fibrosarcoma in transgenic mice, has revealed three stages of that disease, i.e. mild fibromatosis, aggressive fibromatosis, and fibrosarcoma. The latter two stages were highly vascularised when compared with both the normal dermis and the initial mild lesion. Analysis of cell cultures derived from biopsies of these lesions, revealed that bFGF synthesis occurred in all three stages as well as in normal dermal fibroblasts derived from the same mice. However, the location of bFGF changed from its normal cell-associated state in the fibromatotic to extracellular in the subsequent two stages. In this multistep tumorigenesis pathway a discrete switch to the angiogenic phenotype may correlate with bFGF export [75].

Another study discussed the effects of electrical stimulation on joint contracture in rats [76, 77], while others have monitored loss of motion with time as well as myofibroblast numbers in a rabbit knee model of post-traumatic joint contractures [78]. The latter study revealed that myofibroblasts in the posterior joint capsules were elevated 4-5 times in the knees with contractures when compared with the contralateral knees. The initial decrease in severity was followed by stabilisation of motion loss. The association of motion loss with myofibroblast abundance mimics the human scenario of permanent post-traumatic joint contractures [78].

As a substitute for animal models for DD, in silico models using an integrated systems approach could ultimately investigate the effects caused by amplification or modification of various factors, such as determining change upon tensile force application. The creation of a virtual hand using in-silico modeling [79] would be of interest and may elucidate functional outcomes/behaviours of the proliferation of diseased cells. Hereto it should be useful to map expression patterns onto such an anatomical model and to hook it up to in silico models of metabolism and gene expression [80]. The changes in diseased palmar fascia could be compared to the normal fascial state in the hand and allow detection of abnormal fibroblasts early on. Such an approach could overcome the ethical issues in animal research studies [45].

Hunting for the candidate gene by transcriptomics

DNA microarray studies allow differential analysis of the expression of multiple genes [81]. This should permit correlating gene-expression variations with DD. Of course both effects and causes of DD would show up. However, such microarray studies would test the proposal that (see above) DD is nothing more complictated than an (autosomal, dominant) single-gene disease. In the simplest case, the DNA micro arrays should then show a strong correlation with a single such gene, which should moreover localize to a single 6cM region between markers D16S419 and D16S3032. The genes localised on chromosome 16 listed in our previous study [21] are hemoglobin, alpha 2, cadherin 11, type 2, OB-cadherin (osteoblast), matrix metallopeptidase 2, hemoglobin, alpha 1, periplakin, tryptase alpha/beta

1, tryptase beta 2. However, the issue of bulk measurement caused by sample heterogeneity serves to raise the baseline signal of all transcripts and reduce the contribution of signal from the specifically targeted cells.

Alterations of gene expression in Dupuytren's nodules [28], Peyronie's plaques [82], and cultured fibroblasts have been reported. Subsequently, we optimized on unbiased experimental design, sample size, and on data sets to be large enough to elucidate the mechanisms underlying the disease process. The macroscopically distinct, fibrotic elements of the DD tissue, i.e. the nodule and the cord, were considered as two separate entities, as if arisen from separate precursor cells. Gene-expression profiles were compared between diseased Dupuytren tissue biopsies (both nodules and cords) and corresponding healthy tissue (the transverse palmar fascia adjacent to the diseased site) from the same patients. We also compared these gene-expression profiles with those from the palmar fascia of individuals not affected by DD [21]. This study confirmed the DD specific expression of many genes, some of which had been documented as such previously [16, 61]. Using a pathway orientated approach, several additional genes were found to be of statistical significance for DD. The genes established as deregulated in DD belonged to a wide variety of categories, including immune response, angiogenesis, apoptosis, cell adhesion and cell-matrix adhesion, cell cycle and proliferation, cell differentiation, transcription, development, signaling and signal transduction, protein synthesis and folding, oxygen transport, and carbohydrate metabolism (Figure 4).

More recently, a study compared the gene expression profiles of fibroblasts isolated from Dupuytren patients and controls. Here, the authors used two microarray platforms (Illumina™ and GE CodeLink™ Bioarray Systems) initiating a quantitative and comparable approach [83]. They again found tens of genes to be altered in mRNA expression level, which differed between the two platforms. They confirmed the down regulation of three of the genes by QPCR, which included those encoding a proteoglycan, a fibulin and type XV collagen alpha 1 chain [83]. Again using PCR, Ulrich and colleagues found that DD tissue amplified mRNA encoding one metalloprotease and two tissue inhibitors of metalloproteases [84].

The results are not particularly supportive of the hypothesis that altered expression of a single autosomal gene is solely responsible for DD. Many more genes seem to correlate with the disease. The possibility remains however that all these genes are downstream a single key gene, the expression of which might have changed significantly but currently unnoticeably. The observation that genes with fairly obvious functional connotations to DD, such as metalloproteases, proteoglycans and collagen components, have altered expression, brings home the message however that even if the disease were to have a single-gene origin, its aetiology is likely to involve multiple regulatory pathways and genes downstream of such a key gene, the diversity of which in the human population would then cause appreciable dispersion in its pathology. Up to now the hunt for the single DD gene has not only failed but also weakened its own motivation; there may be many genes involved. The more definitive evidence of the effect of targeted knock outs or anti sense RNA is still lacking, in part because of the absence of a clear hypothesis on which gene would cause the disease, and the lack of animal models.

Hunting for the candidate protein by proteomics

Proteomics is the systematic, genome-wide analysis of protein identity, quantity, and function. Proteins are closer to function than transcripts are, therefore proteomics has been thought to have more potential than genomics and transcriptomics for deriving clinically useful applications, in diagnostics, prognostics and therapeutics. The DD proteome analysis venture began in 2006 with the study of protein expression profiles in an attempt to identify potential disease protein biomarkers [85, 86]. In one study, 2-D gel electrophoresis was performed to extract proteins from diseased tissue (nodule and cord), the Skoogs fibres, and normal control tissues. MALDI-TOF-MS (matrix-assisted laser desorption ionization time-of-flight mass spectrometry) was used to generate a peptide mass fingerprint that was used to search protein databases. However, the authors did not report names of identified proteomic changes in their abstract.

Another study employed Surface Enhanced Laser Desorption/Ionization Time-of-Flight Mass Spectrometry (SELDI-TOF-MS) to analyse normal and disease palmar fascia from DD patients using Ciphergen's SELDI-TOF-MS Protein Biological System II (PBSII) ProteinChip reader [86]. The study revealed several differentially expressed low molecular weight (<20 kDa) tissue proteins and identified three disease-associated protein features (4600.8 Da, 10254.5 Da, and 11405.1 Da) that were elevated (5, 12, and 4-fold respectively).

Three potential low molecular weight protein markers (p4.6DC, p1ODC, p11.7DC) for DD were identified. However, there has been some debate as to the reproducibility of SELDI-TOF-MS [87], and this and perhaps the lack of overall coverage of the proteome means that the two studies are not comparable.

More recently, a study using an integrative proteomic-interactomic approach [88] proposed several molecular processes (see below) to be involved in DD progression. It coupled 2-D gel electrophoresis with MS and compared the proteomic profile of DD tissue with that of unaffected patient-matched palmar fasciae tissue. Several proteins correlated with DD. The findings were used to create a protein-protein interaction network (interactome) map on the basis of the proposed interactions in the Human Interactome Map (HiMAP) [89] and the Search Tool for the Retrieval of Interacting Proteins (STRING) [90]: Because it was not possible to design a complete interactome from the experimental data, several proteins were added to the experimental set to fill gaps so as to yield a complete network. This integrated approach suggested several different pathways to be involved such as extra- and intra-cellular signalling, oxidative stress, cytoskeletal changes, and alterations in cellular metabolism. In particular, ERBB-2 and IGF-1R receptors and Akt signalling pathway emerged as novel components of pro-survival signalling in Dupuytren's fibroblasts. One should exercise care however not to over-interpret these results, as they are partly based more on inference involving other protein interaction data obtained in different contexts. In addition, increased activity of pathways need not involve increased protein levels [91] and increased pathway expression may be a homeostatic rather than a primary aetiological event.

The dilemma: more or less data; less or more understanding

With every paper about the genomics of DD, the enigma of the disease seems to increase. As more and more aspects associating with DD are spotted, we see less and less understanding of the disease for its many molecules [92]. The hypothesis that in DD a single gene is at fault, is likely to be false. Even if the disease were set in motion by a single genetic factor, then it encounters so many diverse processes during its aetiology, that it will be co-determined by the many factors that regulate those processes. Indeed, it is more likely that the networks governing differentiation of normal fibrocytes of the palm of the hand are perturbed irreversibly such that they differentiate into muscle tissue without the proper controllers of contraction and relaxation.   It is also more likely then that there are a number of different sets of genetic perturbations that could cause the same type of perturbation that then leads to DD. Here DD is much like cancer [42].

The dilemma is that although we now avail of an unprecedented set of methodologies for the identification and analysis of all the molecules in living cells, that methodology alone is not enough. We need something substantially more to understand how all those molecules interact to create functional networks. Seeing more molecules may not help our understanding. Seeing the connections between them and more mechanism might.

Dupuytren's disease and Systems Biology

Where Systems Biology might come in

Abnormal development of tissue such as in DD and cancer, stresses the surrounding tissue. Hence an association of these diseases with normal tissue repair processes should be expected. Such an assocation does not imply that repair processes are causally involved in the development of disease. The same is true for other homeostatic processes. Because persistence of the disease may depend on the success of the homeostatic response of the surrounding tissue, it is difficult to distinguish between the genes involved causally in the development of such a disease and the genes involved in the homeostatic response.

In addition, a characteristic of tissue repair processes mediated by growth factors is that once the original tissue state has been re-achieved, the autocrine and/or paracrine mechanisms return to normal; the repair process tends to be transient and experiments with insufficient temporal resolution may miss it. A classic example is wound healing [17]. In the temporarily differentiated state, some genes are active and others are repressed. Identification of mechanisms of differentiation then requires the identification of the temporal patterns of gene activity that are causally involved rather than just consequences without causal significance. These patterns may depend on DNA and chromatin modification [93], as well as on the activity of multiple regulatory proteins [94]. In turn these modifications and activities are functions of the state of the cells and of their growth-factor enriched environment [95]. Their regulation may occur at different levels (Table 1) and thus, cell lineage and cell types such as fat, skin, epidermal, dermal fibroblasts and the roles played by them in, for instance, differentiation, should be another aspect of the exploration of genes involved in DD pathogenesis.

More generally, there may be no process in any living organism that stands alone [96]. Processes are linked extensively, if not through metabolic or gene expression networks, then through RNAs (sense, anti-sense or micro [97]), or though dynamic ultrastructure [93, 98, 99]. To understand how living cells function, one needs to have a way to look at the operation and integration of several simultaneous processes, as a function of time. Since the sum of a negative and a positive effect is important, yet uncertain in the absence of precise assessment of magnitudes, the required approach needs to be precise experimentally and quantitative in the analysis. Since virtually all molecules of a living cell are connected, the approach needs to relate to molecular biology as well as functional genomics [30]. It is in appreciating these concertations of the various levels of cell functioning, that systems biology might come to the rescue.

This systems biology activity should differ from the traditional molecular biological approach, where reductionist strategies assist in the characterization of the molecular components of the much larger cellular system [100]. In molecular biology a ‘favourite’ candidate gene or gene product may be investigated on its own (e.g. by immunohistochemistry) using some pre-existing knowledge of its function but without simultaneously looking at the pathway in which it is active. Whilst this approach is useful when detailing the molecular mechanism of action, it does not automatically lead to understanding of how the protein or gene acts responsively so as to make the whole cell and ultimately the whole organism, function. Organisms depend on many more than a single gene function [101, 102]. Unless one gene or its product were the single rate limiting step for the organism or cell function (e.g. survival, proliferation rate) of interest, more than one gene product must be considered.

The attractive and thereby rather persistent concept of ‘the’ rate limiting step in a network process has been invalidated for metabolic pathways [103], for gene expression circuits [104-106], and for signal transduction [107, 108]. Already in Escherichia coli gene expression is under tight and subtle homeostatic control, involving gene-expression and metabolic regulation through topoisomerase I, DNA gyrase, the supercoiling state of the DNA itself, and ATP simultaneously [104]. For signal transduction pathways the subtleties in control are not much less [107, 109-111]. Much the same applies to regulation [40, 112]. Accordingly, the default way of examining how the functioning and malfunctioning of network processes is controlled and regulated in DD should reckon with the likelihood that many more than one molecular process is involved. The systems biology developed and tested in simpler organisms and systems offers various methodologies to do so [41, 98, 100].

An immense number of studies have been conducted at the genome and transcriptome (set of all mRNA molecules, or ‘transcripts’, produced in a cell or a tissue) levels. More are now following at the levels of the proteome, (set of entire complement of proteins expressed by a genome, cell, tissue or organism), interactome (interactions between all molecules within a cell), and metabolome (complete set of small-molecule metabolites). A priori this functional genomics may have seemed a promising approach. It should show most if not all the factors involved in a disease such as DD. However, it has not yet delivered on its promises, neither in any disease, nor (yet) in DD (see the above discussion). This might be attributed to the sheer number of molecules involved, but then bioinformatics should know how to deal with those numbers.

What is the key issue here? Why might the direct accumulation of knowledge of ultimately all components of a living organism, or of a disease, without much further processing, not lead to understanding of function or malfunction, of physiology and pathology? The reason may be: most functions depend on the concerted action of a number of processes the rates and efficiencies of which influence each other. For instance the synthesis of a given protein (e.g. myosin) will depend on the synthesis of 20 aminoacyl tRNAs which again depends on the synthesis of the corresponding amino acids. A stark example is that of the molecule cyclin D which is involved in the mechanism of the cell cycle by oscillating in terms of its activity [113]. Alone in a test tube, cyclin D activity would not oscillate and it is only because of its dynamic interaction with a dynamic network of other proteins that cyclin D cycles and the cell with it [114]. Moreover these oscillations are generally found at the single cell level and as such analyses should be aimed at these entities rather than averages of the population [115].

Systems Biology specializes in understanding the roles such interactions play in bringing about biological function [116]. It thereby aims to complete the large part of the job molecular biology and genomics fail to accomplish, i.e. the job of understanding physiology and pathology. This is more easily said than done however. It is the activities of all the components of living systems and their relationships to each other that multiply or integrate rather than add up to the living organism. In recognition of this complexity, systems biology attempts to harness the power of mathematics, engineering, and computer science to analyse and integrate data from all the ‘omics’ [35], ultimately through the construction of experiment-based, in silico models [117].

Molecular biological components and the systems in which they function (e.g. substrates, enzymes, metabolites, genes in a cell, tissue or organism) and pathways mediating their functional outcome, are many times more complex than networks and circuits such as the London Underground. Yet, the London Underground is already complex: making sure that one particular station functions efficiently such that each minute one train could depart, does not guarantee that indeed one train will depart per minute. Mostly likely, many fewer trains will depart because further down the tube there are other stations that are less efficient, or because at some stations up the tube excessive numbers of people wish to board or leave the train, or because the train driver overslept. Network studies reside at the crossroads of disciplines, from mathematics (graph theory, combinatorics, probability theory) to physics (statistical thermodynamics, macromolecular crowding), and from computer science (network generating algorithms, combinatorial optimisation) to the life sciences (metabolic and regulatory networks between proteins and nucleic acids). The impact of network theory on understanding is strong in all natural sciences [118], especially in systems biology with gene networks [119], metabolic networks [41, 120, 121], plant systems biology [122], and even food webs [123]. Yet, biological systems will not be understood by existing network theory alone. Their properties are much more complex than the properties of standard networks, for instance in that their networks adapt and change temporary [124], are hierarchical in terms of space, time and organization [98], and have been optimized through evolution for multiple properties that we do not yet understand [41, 100, 125]. New network theories are needed and will have to be more targeted towards understanding biological systems functionally [126]. These will have to integrate strongly with genomics and molecular data, because different biological networks may need somewhat different theories, if only because their objective (evolutionary purpose) is different. The Flux Balance Analysis objective function of maximum growth yield for instance [127], is irrelevant for the human erythrocyte and muscle cell, which do not grow [100]. For the former, the objective functions of oxygen carrying capacity plus binding capacity in the lungs and delivery capacity in the tissues, without consuming the oxygen needed by those tissues, will make more sense; for the latter, performance independent of oxygen supply will do. It remains to be seen whether in the development of DD the flux pattern adjusts to a new criterion of optimality, such as in some cases of tumorigenesis [128].

System diseases versus molecular diseases

Noting that molecular biology and genomics do not suffice for understanding living systems is one thing, but actually making a difference by applying systems approaches, may be quite another. We shall here sketch how a systems approach to DD might make a difference. The first role systems biology may have is that of clarifying that DD is a type of disease that, although common, is not engrailed in scientific tradition, and should not be approached by traditional methodologies alone. For this we first need to clarify what the difference is between a systems-biology disease and a single-molecule disease. Figure 5 illustrates this difference. A single-gene disease depends in principle on a single gene only, or at least that is how it is often approached. But of course, a disease cannot depend on a gene (if defined as the corresponding DNA sequence) alone: it will depend on its gene product (F in Figure 5(A)), and in fact on the molecular function of the latter. For instance a muscular dystrophy could result from a mutation in the gene encoding myosin, the molecular function of which is muscle contraction. If that muscular dystrophy would only be found when the myosin gene has been mutated and if the severity of the disease would not be influenced by other factors, then that muscular dystrophy would be a single-gene disease. In actuality there are many different genetic lesions that lead to similar muscular dystrophies, including lesions in mitochondrially encoded genes [129].

A better candidate for a mono-gene disease may be phenylketonuria, an inherited (autosomal recessive) metabolic disease that is largely due to mutations in the phenylalanine hydroxylase (PAH) gene [130]. However, its therapy (dietary restriction) shows that the disease can be influenced by external factors. Moreover, mutations in genes involved in the synthesis of a cofactor of the phenylalanine hydroxylation reaction also lead to the disease, and there are multiple alleles of the PAH gene with different implications for the severity of the disease. Hence even this disease exhibits characteristics of systems biology diseases, and therapy could well benefit from a corresponding approach.

The dubious concept that diseases are essentially due to the perturbation of a single gene is pervasive, be it mostly in an implicit sense. The concept is recognized in the gene hunt approaches to diseases where a transcriptome analysis is aimed at the identification of a so-called ‘candidate gene’, where the word ‘candidate’ refers to the implicit expectation that a single molecular culprit may be found for a disease or other phenotypic feature. The early molecular biology of cancer was an example, with a hunt for the oncogene and before that, for the single molecular processes (membrane fluidity for instance) that could be held responsible for the disease. Of course for cancer this concept has long faded, thanks to the demonstration that cancer development requires cooperativity between at least three oncogenes [29, 131]. Indeed, many more than 70 oncogenes and tumor suppressor genes have been identified, which are not all involved in all cancers [29].

Figure 5(A) may be judged a caricature of single-gene diseases then, as most diseases have multiple genes associated with them. However, such diseases could be considered to be a group of single-molecule diseases; i.e. many diseases each caused by a different single gene lesion, but all with similar phenotypes [132]. This would explain the association of multiple genes with the disease and still essentially reduce it to single-gene diseases (Figure 5(A)). The difference between a disease being a group of single-gene diseases and being a systems biology disease is that the former should still only depend on the single molecule that is faulty. In the case of a group of single gene diseases, there should be no other faulty molecules important for that individual disease and there should be no influences on the disease severity of other gene changes (e.g. polymorphisms) or conditions (e.g. diet) on that disease. Notably, a single patient's transcriptome should then show only changes in the single molecular culprit and not in many factors controlling the network leading to the disease. And in the transcriptomes of different patients suffering from the same disease group, that single molecular culprit should be different.

The suspicion that single-gene diseases may be rare, mirrors the point made above that there are few completely rate-limiting steps in cell biology. Indeed, dominance should be common and loss of function in heterozygous deletions rare: The very fact that most pathways consist of many steps, strongly reduces the average effect a heterozygous deletion should have on their flux [133, 134]. The phenomenon of parallel pathways further reduces the effect of deletions [135]. Single gene perturbations would thereby rarely lead to disease. For disease to occur multiple genetic lesions are likely to be required.

What then is a systems biology disease? As illustrated by Figure 5(B), in a systems biology disease the function that is compromised depends cooperatively on a number of pathways, the functioning of each of which again depends on many cooperating molecular factors. In systems biology diseases one would typically find multiple changes in the transcriptome of each patient, differing between individual patients but such that all have a very similar disabling effect on network function.

It might seem that the identification of the variety of transcriptome changes that lead to the same network change and hence to the same effect on function could be achieved by top-down systems biology, i.e. by measuring variations of transcriptomes and function and by then designing the linear combination of transcriptome changes that would computationally lead to the most extreme effect on function and malfunction [136]. However, functional dependencies are frequently nonlinear in biological systems and therefore more mechanism, kinetics and topology of the network needs to be known to make this strategy effective. Ultimately a dynamic model that is true-to-reality should be able to do this job, but that of course requires an enormous experimental effort before such a model would be reliable enough [137, 138]. Until that time, it would seem that an experimental diagnosis to see whether the observed changes in gene expression do cause the changes in function that are essential for the disease to occur, remains an essential components of systems biology.

Identifying a disease as a systems biology disease does not dispel molecules from its pathology: molecules are always involved. The issue is whether the change in networking of the molecules is crucial for the disease, i.e. whether the disease is more a consequence of faulty networking than of malfunctioning of individual molecules. If the disease is due to molecules defaulting at their own molecular function only because of changes inherent to those molecules (such as through mutations), it should perhaps be called a molecular disease. For every disease it may be useful to decide whether it is more molecular or more systems biological in the above sense.

Once a disease has been recognized as a systems biology disease, then what difference should this make for research, diagnosis and therapy? The answer is straightforward: When dealing with a network disease, one should deal with the network; when dealing with a molecular disease one should concentrate on the molecule. For systems biology diseases, transcriptome patterns should be mapped onto the known cellular pathways. One should not try to establish correlations between individual mRNAs and disease, but rather between the effects of mRNA changes on a pathway and the disease. The concept ‘candidate pathway’ or even ‘candidate network’ should be substituted for ‘candidate gene’. In addition, cell function is only partly controlled and only partly regulated at the level of transcription [98]. Hence one should also involve the levels of the proteome, of the metabolome, and of function, and not each independently but all together;

then, one should ensure that the changes observed make collective sense. For if some factors in a pathway change up and others down, this may not enhance function; the changes should be related to pathway kinetics or control and it should be checked whether together they affect the function or the malfunction that is of interest for the disease. Figure 6 shows that the issue may be complicated further because a disease may readily involve more than one function. Persistent malignant cancer for instance may involve proliferation, lack of apoptosis, metastasis and multiple drug resistance.

Is DD a molecular or a systems biological disease?

After defining the differences between molecular and systems diseases, we should now establish which of the two DD is. One aspect comes to mind immediately: DD has been identified as a disease inherited in an autosomal dominant pattern [71]. It was linked to a single 6cM region on chromosome 16. This would suggest that all DD patients should have a mutation in this part of their genome, and that transcriptomes of DD patients should be altered in terms of the level of the transcripts encoded by this part of the genome, or in terms of the coding sequence of one of those transcripts. However, the dominance was incomplete and has only been observed in a single Swedish family. This suggests that the genes on chromosome 16 are only dominant when other genes in the genome are of certain allelic forms (i.e. the ones that happen to dominate in that particular Swedish family). Moreover, in many other cases many other mRNAs were changed in expression levels, although it remains to be analyzed whether in those papers there was always a change in an mRNAs from the 6 cM region on chromosome 16. In our own studies, the DD-nodule transcriptomes of individual patients all exhibited multiple changes in mRNA levels and these changes overlapped, but were not identical between individuals. The proteome did not point at a single protein either. The functional studies pointed at myofibroblast enrichment, but not clearly as the sole cause and neither was a causal relationship between a gene on chromosome 16 in the 6 cM region and differentiation of myofibroblasts established. This all shows that DD is not a single-gene disease and suggests that it is not just a group of pure single-gene diseases either.

For further understanding of DD then, a hypothesis-driven systems biology approach should allow for selection of candidate pathways. This could be based on a priori observations in human *in vitro* or *in vivo* (linkage and expression studies, for example), or on the basis of knowledge in related diseases (such as plantar fibromatosis, peyronies, musculo-aponeurotic fibromatosis and even keloid disease). Inter-relationships may be sought between hypothesized underlying mechanisms governing these fibrotic disorders and physiological changes predicted for changes in molecular and environmental changes impacting on those mechanisms. This could then be extended to understand inter- vs. intra-individual variability. Such analysis should lead to hypotheses about mechanisms by which network changes would cause the DD phenotype. By bringing about those network changes by multiple molecular interventions in a tissue culture model system for DD, the hypotheses could then be tested. Such an approach should also help put into perspective existing inconclusive discoveries. It would thereby maximize the utilization of data obtained in molecular approaches from molecular biology, which provide an extensive database. Then systems biology would reduce this to a smaller but more implicative knowledge base, perhaps in the sense of a live model repository [117].

Systems Biology advances

Already, systems-level approaches are making a pace towards scientific understanding and biotechnological applications [30]. Recently, the National Heart, Lung and Blood Institute (NHLBI) of the US' NIH initiated the Program for Genomic Application (PGA) of advancing functional genomic research. One such PGA is known as CardioGenomics in which the primary aim is to study how the transcriptional network of the cardiovascular system responds to genetic and environmental stresses and how the network is altered in disease conditions. Research in CardioGenomics focuses on transcriptional profiling of murine models of cardiomyopathy, on the diseased *versus* normal human myocardium, and on identification of the gene mutations that cause familial hypertrophic cardiomyopathy and left-heart obstructive lesions. Meanwhile studies using a CardioChip (a custom built cDNA microarray), have identified differentially expressed genes in diseased conditions as compared to the normal heart [139]. With this approach the limitations of molecular, genetic and functional genomics research in gaining a complete understanding of a complex disease such as heart failure are again being highlighted. Many recent articles illustrate the use of proteomics technology to investigate the disease-state proteome using a multitude of experimental techniques and elucidate biomarkers that may be indicative of pathological network states [140, 141], and the same process is happening within the area of metabolomics [142]. However, the subject is in urgent need of a systems biology approach. There is an extensive and successful systems biology activity that focuses on the heart [143]. The focus of that activity has however been on the electrophysiology and the relationships of its models with metabolism and gene expression is still in the making, e.g. in an important project, the virtual physiological human [144, 145].

Cancer cannot be considered as a single molecule disease as so many oncogenes have been identified. Because there is no cancer that is caused by a single mutation (even retinoblastoma is focal, i.e. does not occur in all cells of the individual with the responsible mutation), cancer is not just a collection of single-gene diseases either [29, 131]: Complex networks of relationships between genes, gene products and/or proteins govern neoplastic processes [29, 42]. Cancer requires the simultaneous deregulation of a set of genes [29, 131], which act in a cooperative mode. In other words, cancer requires the deregulation of a network of genes or even of various networks of genes. It is a systems biology disease therefore [42], or even a collection of systems biology diseases [132]. Whole-genome association scans and mutational screens of cancer genomes identify gene interactions that associate with cancer [146]. It is now time to project these discoveries onto cancer pathways, or onto normal pathways that are essential for tumors to develop into malignancy [147].

Systems Biology as integrator of heterogeneous ¬omic datasets

The ¬omics experiments generate heterogeneous data and meta data torrents. Not only should this data be put to best use when studying the complex system of interest through systems biology, the modelling by systems biology would serve as a way to organize the data rationally. These heterogeneous datasets are normally deposited in a broad spectrum of public and commercial databases and for them to be put into context it is important that both their own format and their annotation are standardized. Emerging standard languages such as the systems biology Markup Language (SBML) [148], Cell Markup Language (CellML) [149], Cell System Markup Language (CSML) [150], Biological Pathway Exchange (BioPAX) [151] and Systems Biology Graphical Notation (SBGN) [152], and modelling tools such as COmplex PAthway SImulator (COPASI) [153], Cytoscape [154] and Pathway databases (e.g. Ingenuity pathway analysis software) facilitate data representation and inter-operability from leading multidisciplinary research groups [155]. Important also is the Java Web Simulation (JWS) facility which quality controls kinetic models and puts them into a live repository, enabling through-web experimentation in silico for scientists naive with respect to modeling but experiment prone [156]. BioModels is a parallel model repository not built for in silico experimentation but focusing on annotation [157].

Systems biology is gaining increasing support from a wide variety of sources. To name but a few, UK's BBSRC and EPSRC have funded six Systems Biology research centres and three doctoral training centres dedicated to Systems Biology, the Wellcome Trust supports the Heart Physiome Project and the Integrative Animal and Human Physiology Initiative [158, 159]. Evidence from medical research charities [160] such as the Arthritis Research Campaign, British Heart Foundation and Cancer Research UK is indicative of systems biology becoming an increasingly important component of their research programmes. The FP7 Marie Curie Training Network NucSys presents a nutrigenomic approach to aging-related diseases *vis-a-vis* the systems biology of nuclear receptors [161]. German BMBF's HepatoSys is focusing on quantitative understanding of complex and dynamic cellular processes in detoxification, endocytosis, iron regulation and regeneration in mammalian hepatocytes [162]. Switzerland and Luxembourg are the most active countries pro capita. Other emerging Systems Biology includes the development of personalised medicine [163] and of the systems biology of ageing [164, 165] as well as integrative pharmacogenomics, biomarker discovery for disease prognoses, and therapy monitoring [166-171]. The field applies quantitative, mechanistic approaches to understand disease, cell, tissue or organism holistically from bottom up, top down and/or middle out [172] approaches. There is great interest in biomarker discovery, and as many diseases are a result of genetic and/or metabolic disorders [173], it makes a great deal of sense to measure gene expression (transcriptome) and metabolites (metabolome) directly [142] at these functional levels with systems biology enabling correct interpretation and prediction of functional outcomes [174].

Imaging

Finally, it would be incomplete to mention systems biology in medicine and not mention the advantages technological imaging has given us to capture disease [175]. Perhaps the biggest growth area in imaging technologies is fluorescence imaging, with various technologies being adapted for *in vivo* analysis; although recent advances in label-free methods such as Raman microspectrometry are also showing promise [176]. The development of imaging techniques, will allow researchers to address some of the questions in molecular oncology such as how the components of intracellular signalling pathways interact in real time (e.g. [177, 178]). The current and developing instrument-based technological platforms (computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), confocal microscopy, and Raman microspectroscopy) already provide a platform for initiating a knowledgebase relationship between the clinician, patient and scientist. However, a systematic analysis would fill missing gaps in this knowledge base, as illustrated in Figure 7. Integration of the observation platform with dynamic-spatial models [79] has not yet been achieved much but should now raise the utility of the imaging approaches to new levels. For DD the potential is to image the metabolic state of the various cells in the nodule and the cord as a function of contractile activity, in order to examine how that activity could be modulated. At the cellular level the de- and re-differentiation of fibroblasts and myofibroblasts might be monitored and modelled.

Discussion

DD research has so far implicated a re-differentiation of fibroblasts into myofibroblasts. Functional genomics studies have highlighted multiple mRNAs of which the expression level differs between diseased and healthy tissue. Extracellular matrix enzymes are involved in the disease. High throughput and candidate-gene association studies have suggested multiple biomarkers for the disease, but these would require invasive sampling. Association of the disease with a single gene has been inconclusive. Thus in the absence of radically new findings, the present mainstream research paradigm is unlikely to lead to a full understanding of the disease. Progress is impressive in terms of collecting more data, but less so in terms of understanding these data and hence the disease.

What is known about DD suggests that DD is neither a discrete single-gene disease nor a group of single-gene diseases. The experimental data are much more consistent with DD being a systems biology disease. Such diseases are determined much more by the malfunctioning of the network of the macromolecules rather than by the malfunctioning of the individual macromolecules themselves. At present this is a suggestion only, because the data are inconclusive. However, we have indicated how a progressive systems approach could test this hypothesis. It would identify the network states that correlate with DD and then bring about changes in these network states with predicted outcomes in terms of disease progression, aggressiveness and the ability to reoccur after therapeutic intervention. Comparison with actual outcomes would serve as the test of the suggestion. Experimentally the approach would not be much different from approaches that are being developed anyway, but the experimental design would be different, involving systems biology, multiple modulations and monitoring of time dependence.

If indeed DD turns out to be a systems biology disease, we propose that the approach to the disease should change drastically: analysis, diagnosis and therapy should target pathways rather than genes or their products. The concept of a 'candidate gene' should be replaced with that of 'candidate pathway(s)'. Studies should be aimed at elucidating cause-effect chains, rather than correlating and thereby not being able to distinguish consequences of the diseased state from its causes. To be more concrete, from the experimental data, alterations in pathways should be inferred. Using transgenic and antisense approaches in cell lines, these pathway alterations should then be induced and the predicted development into a DD cellular phenotype tested. The pathways are expected to be integrals of gene expression, signalling and metabolic networks and so should be the approach and data analysis. At present DD research is not conducted in the required integrative manner. Consequently, what we propose has substantial ramifications for the organization of DD research.

If we are right, the rewards would be substantial: No longer will the data collected in this field disappear into the diasporas of the experimental literature; they will be analyzed in terms of network models and when informative connected to them with proper data annotation [179]. Once the network hypotheses turn out to be successful, they will underpin the development of new rational biomarkers strategies, and become starting points for therapeutic intervention and prophylaxis.

We suggest that adopting a systems biology approach to analyze DD has the potential to improve the knowledge gained in both *in vivo* and *in vitro* DD research. This may be in the form of hypotheses generation in a carefully controlled experiment, or in the form of making experimental data address issues that limit DD understanding.

Even though myofibroblasts obtained from different stages of DD may exhibit features that could trigger contraction and uncontrollable growth, neither the diversity of these cells nor the extent or nature of local specificity in situ in their differentiation has been examined systematically. Remodeling of vascular connective tissue should be of fundamental importance as DD progresses over time and the ability of that tissue to be remodelled could be a gratuitous factor in the development of the disease. Matrix remodeling and matrix turnover is controlled by a complex network of cell-cell and cell-matrix interactions [180]. Mathematical modelling in a systems biology approach could help deduce the net outcome where the product would be a robust balance between proliferative and degradative processes. This could extend the work from previous findings where the expressions of the family of enzymes and inhibitors are directly associated with matrix turnover (the matrix metalloproteinases (MMPs) and their natural inhibitors (TIMPs) [23]). Furthermore, the constraints imposed by non-functional protein–protein interactions on gene expression and proteome size could be studied.

Metabolomics alone can help identify metabolite markers and has many exciting clinical applications [163, 173, 181]. As well as genomics, transcriptomics and proteomics, the fluxomics, lipidomics, interactomics, glycomics and secretomics studies of biofluids within the Dupuytren’s system together have the potential to improve our understanding immeasurably, especially if they are integrated, and full integration is only plausible through systems approaches. Investigating this deforming complex fibromatosis as part of a systems biology approach (Figure 8) will benefit not only the understanding of the diseased sites, but will also address the effects on the ECM surroundings and excreted by-products, and could offer suggestions for early diagnosis. This could then be extended to explore relations between DD, bone formation and mineralization, as many deregulated genes that co-exist may lead to further complications. A novel systems approach along with existing knowledge will give a comprehensive view of the network of gene interactions played in gene regulation and their functional properties involved in patterning DD from normal state. Systems biology approaches may provide clues for new diagnostic and prognostic markers for DD and as well as design of innovative novel therapeutic tools.

Conclusions

DD is likely to be a systems biology disease. The corresponding mechanistic strategy offered by systems biology may lead to important new insights into DD tumorigenesis, based on the analysis of molecular interactions that become deregulated in the DD tumour phenotypes (nodule, cord). This approach will not only extend but also integrate and complement existing methods and information. The existing genomic data will serve as metadata for a systems oriented knowledgebase, partly in the form of experimentally tested mathematical models, which then lead to breakthroughs in prophylactic and therapeutic measures in DD.

List of abbreviations

bFGF: basic fibroblast growth factor;

BioPAX: Biological Pathway Exchange;

CellML: Cell Markup Language;

COPASI: COmplex PAthway SImulator;

CSML: Cell System Markup Language;

CT: computed tomography;

DD: Dupuytren's disease;

ECM: extracellular matrix ECM;

HiMAP: Human Interactome Map;

HLA-DRB1: major histocompatibility complex, class II, DR beta 1;

JWS: Java Web Simulation;

LOD: logarithm of the odds;

MCP: metacarpophalangeal;

MMPs: matrix metalloproteinases;

MRI: magnetic resonance imaging;

NHLBI: National Heart, Lung and Blood Institute;

PAH: phenylalanine hydroxylase;

PBSII: Protein Biological System II;

PET: positron emission tomography;

PGA: Program for Genomic Application;

PIPJ: proximal interphalangeal joints;

QPCR: Quantitative polymerase chain reaction;

SBGN: Systems Biology Graphical Notation;

SBML: systems biology Markup Language;

SELDI-TOF-MS: Surface Enhanced Laser Desorption/Ionization Time-of-Flight Mass Spectrometry;

STRING: Search Tool for the Retrieval of Interacting Proteins;

TGF-β: Transforming growth factor-beta;

Competing interests

The author(s) declare that they have no competing interests.

Acknowledgements and Funding

SR thanks BBSRC and EPSRC for her studentship via the Doctoral Training Centre, at The Manchester Centre for Integrative Systems Biology, UK. HW and RG thank various funders [http://www.systembiology.net/support/] but in particular the EPSRC and BBSRC for their support of the above two centres.

References for Appendix A; Section 1.2

1. McFarlane RM MD, Flint MH, editors: *Dupuytren’s disease: biology and treatment.* New York: Churchill Livingstone; 1990.
2. Rayan GM: **Dupuytren Disease: Anatomy, Pathology, Presentation, and Treatment.** *J Bone Joint Surg Am* 2007, **89:**189-198.
3. Horner RL, Bralliar F: **Dupuytren's contracture. Analysis of 100 consecutive surgical cases.** *Rocky Mountain medical journal* 1971, **68:**49-52.
4. Schroter G: **The recognition of Dupuytren's contracture as an occupational disease.** *Beitr Orthop Traumatol* 1971, **18:**78-80.
5. Tubiana R: **Dupuytren's disease. Present status of the treatment.** *Cah Med* 1971, **12:**1305-1309.
6. Meyerding HW BJ, Broders AC: **The Etiology and Pathology of Dupuytren's Contracture.** *Surgery, Gynecology and Obstetrics* 1941.
7. Seemayer TA LR, Schürch W, Thelmo WL: **The myofibroblast: biologic, pathologic, and theoretical considerations.** *Pathology Annual* 1980, **15 (Pt 1):**443-470.
8. Bayat A, McGrouther DA: **Management of Dupuytren's disease - Clear advice for an elusive condition.** *Ann R Coll Surg Engl* 2006, **88:**3-8.
9. Shaw RB, Chong AKS, Zhang A, Hentz VR, Chang J: **Dupuytren's disease: History, diagnosis, and treatment.** *Plast Reconstr Surg* 2007, **120**.
10. Hindocha S, Stanley JK, Watson S, Bayat A: **Dupuytren's Diathesis Revisited: Evaluation of Prognostic Indicators for Risk of Disease Recurrence.** *J Hand Surg Am* 2006, **31:**1626-1634.
11. Hayton MJ, Gray ICM: **Dupuytren's contracture: A review.** *Current Orthopaedics* 2003, **17:**1-7.
12. Tse R, Howard J, Wu Y, Gan B: **Enhanced Dupuytren's disease fibroblast populated collagen lattice contraction is independent of endogenous active TGF-beta2.** *BMC Musculoskelet Disord* 2004, **5:**41.
13. Drake MJ, Hedlund P, Andersson KE, Brading AF, Hussain I, Fowler C, Landon DN: **Morphology, phenotype and ultrastructure of fibroblastic cells from normal and neuropathic human detrusor: Absence of myofibroblast characteristics.** *J Urol* 2003, **169:**1573-1576.
14. Moyer KE, Banducci DR, Graham WP, Ehrlich HP: **Dupuytren's disease: physiologic changes in nodule and cord fibroblasts through aging in vitro.** *Plast Reconstr Surg* 2002, **110:**187 - 193.
15. Iwasaki H MH, Stutte HJ, Brennscheidt V: **Palmar fibromatosis (Dupuytren’s contracture). Ultrastructural and enzyme histochemical studies of 43 cases.** *Virchows Arch A [pathol Anat] Histopathol* 1984, **405:**41-53.
16. Lubahn JD, Pollard M, Cooney T: **Immunohistochemical Evidence of Nerve Growth Factor in Dupuytren's Diseased Palmar Fascia.** *J Hand Surg Am* 2007, **32:**337-342.
17. Clozel M SH: **Role of endothelin in fibrosis and antifibrotic potential of bosentan.** *Ann Med* 2005, **37:**2-12.
18. Cordova A, Tripoli M, Corradino B, Napoli P, Moschella F: **Dupuytren's contracture: An update of biomolecular aspects and therapeutic perspectives.** *J Hand Surg Br* 2005, **30:**557-562.
19. Hnanicek J, Cimburova M, Putova I, Svoboda S, Stritesky J, Kratka K, Sosna B, Horak J: **Lack of association of iron metabolism and Dupuytren's disease.** *J Eur Acad Dermatol Venereol* 2008, **22:**476-480.
20. Kaur S, Forsman M, Ryhänen J, Knuutila S, Larramendy ML: **No gene copy number changes in Dupuytren's contracture by array comparative genomic hybridization.** *Cancer Genet Cytogenet* 2008, **183:**6-8.
21. Rehman S, Salway F, Stanley JK, Ollier WER, Day P, Bayat A: **Molecular Phenotypic Descriptors of Dupuytren's Disease Defined Using Informatics Analysis of the Transcriptome.** *J Hand Surg Am* 2008, **33:**359-372.
22. Forsman M, Paakkonen V, Tjaderhane L, Vuoristo J, Kallioinen L, Salo T, Kallioinen M, Ryhanen J: **The Expression of Myoglobin and ROR2 Protein in Dupuytren's Disease.** *J Surg Res* 2008, **146:**271-275.
23. Johnston P, Chojnowski AJ, Davidson RK, Riley GP, Donell ST, Clark IM: **A Complete Expression Profile of Matrix-Degrading Metalloproteinases in Dupuytren's Disease.** *J Hand Surg Am* 2007, **32:**343-351.

24. Lee LC, Zhang AY, Chong AK, Pham H, Longaker MT, Chang J: **Expression of a Novel Gene, MafB, in Dupuytren's Disease.** *J Hand Surg Am* 2006, **31:**211-218.

25. O'Gorman D, Wu Y, Seney S, Zhu R, Gan B: **Wnt expression is not correlated with beta-catenin dysregulation in Dupuytren's Disease.** *J Negat Results Biomed* 2006, **5:**13.

26. Samuel CS, Hewitson TD: **Relaxin in cardiovascular and renal disease.** *Kidney Int* 2006, **69:**1498-1502.

27. Maureen DM: **Endothelin and endothelin receptor antagonists in systemic rheumatic disease.** *Arthritis Rheum* 2003, **48:**1190-1199.

28. Pan D, Watson HK, Swigart C, Thomson JG, Honig SC, D N: **Microarray gene analysis and expression profiles of Dupuytren's contracture.** *Ann Plast Surg* 2003, **50:**618-622.

29. Hanahan D, Weinberg RA: **The Hallmarks of Cancer.** *Cell* 2000, **100:**57-70.

30. Westerhoff HV, Palsson BO: **The evolution of molecular biology into systems biology.** *Nat Biotechnol* 2004, **22:**1249-1252.

31. Kholodenko BN, Kiyatkin A, Bruggeman FJ, Sontag E, Westerhoff HV, Hoek JB: **Untangling the wires: A strategy to trace functional interactions in signaling and gene networks.** *Proc Natl Acad Sci* 2002, **99:**12841-12846.

32. An G, Hunt CA, Clermont G, Neugebauer E, Vodovotz Y: **Challenges and rewards on the road to translational systems biology in acute illness: four case reports from interdisciplinary teams.** *J Crit Care* 2007, **22:**169-175.

33. Ahn AC, Tewari M, Poon C-S, Phillips RS: **The Limits of Reductionism in Medicine: Could Systems Biology Offer an Alternative?** *PLoS Medicine* 2006, **3:**e208.

34. Katagiri F: **Attacking Complex Problems with the Power of Systems Biology.** *Plant Physiol* 2003, **132:**417-419.

35. Spivey A: **Systems Biology: The Big Picture.** *Environ Health Perspect* 2004, **112:** A938–A943.

36. Jensen PR, Gugten AAvd, Bier M, Heeswijk WCv, Rohwer JM, Molenaar D, Workum Mv, Richard P, Teusink B, Bakker BM, et al: **Hierarchies in control.** *Journal of Biological Systems* 1995, **3:**139-144.

37. Kahn D, Westerhoff HV: **Control theory of regulatory cascades.** *J Theor Biol* 1991, **153:**255-285.

38. Steven Wiley H, Shvartsman SY, Lauffenburger DA: **Computational modeling of the EGF-receptor system: a paradigm for systems biology.** *Trends Cell Biol* 2003, **13:**43-50.

39. Schoeberl B, Gaudet S, Albeck JG, Janes K, Sorger PK, DA. L: **The apoptotic decision process in TNF alpha-stimulated HT-29 cells: a combined computational and experimental approach.** *Mol Biol Cell* 2002, **13:**11A-A.

40. Daran-Lapujade P, Rossell S, van Gulik WM, Luttik MAH, de Groot MJL, Slijper M, Heck AJR, Daran J-M, de Winde JH, Westerhoff HV, et al: **The fluxes through glycolytic enzymes in Saccharomyces cerevisiae are predominantly regulated at posttranscriptional levels.** *Proc Natl Acad Sci* 2007, **104:**15753-15758.

41. Westerhoff HV: **Signalling control strength.** *J Theor Biol* 2008, **252:**555-567.

42. Hornberg JJ, Bruggeman FJ, Westerhoff HV, Lankelma J: **Cancer: A Systems Biology disease.** *Biosystems* 2006, **83:**81-90.

43. Boshoff HI, Barry CE, Hornberg J, Bruggeman F, Bakker B, Westerhoff H: **Metabolic control analysis to identify optimal drug targets.** In *Systems Biological Approaches in Infectious Diseases. Volume* 64: Birkhäuser Basel; 2007: 171-189: *Progress in Drug Research*].

44. Bakker BM, Assmus HE, Bruggeman F, Haanstra JR, Klipp E, Westerhoff H: **Network-Based Selectivity of Antiparasitic Inhibitors.** *Mol Biol Rep* 2002, **29:**1-5.

45. Westerhoff HV, Mosekilde E, Noe CR, Clemensen AM: **Integrating systems approaches into pharmaceutical sciences** *Eur J Pharm Sci* 2008, **35:**1-4.

46. Rinn JL, Wang JK, Liu H, Montgomery K, van de Rijn M, Chang HY: **A Systems Biology Approach to Anatomic Diversity of Skin.** *J Invest Dermatol* 2008, **128:**776-782.

47. Hinz B: **Formation and Function of the Myofibroblast during Tissue Repair.** *J Invest Dermatol* 2007, **127:**526-537.

48. Chang HY, Chi JT, Dudoit S, Bondre C, Van De Rijn M, Botstein D, Brown PO: **Diversity, topographic differentiation, and positional memory in human fibroblasts.** *Proc Natl Acad Sci* 2002, **99:**12877-12882.

49. Darby IA, Hewitson TD, Kwang WJ: **Fibroblast Differentiation in Wound Healing and Fibrosis.** In *Int Rev Cytol. Volume* Volume 257: Academic Press; 2007: 143-179

50. Eddy A: **Interstitial macrophages as mediators of renal fibrosis.** *Exp Nephrol* 1995**:**76-79.

51. Hunninghake G: **Sarcoidosis: linking inflammation and fibrosis.** *Am J Med Sci* 1995**:**124-133.

52. Selman M, eacute, Pardo A, Kaminski N: **Idiopathic Pulmonary Fibrosis: Aberrant Recapitulation of Developmental Programs?** *PLoS Medicine* 2008, **5:**e62.

53. Vaglio A, Salvarani C, Buzio C: **Retroperitoneal fibrosis.** *The Lancet* 2006, **367:**241-251.

54. Hindocha S, John S, Stanley JK, Watson SJ, Bayat A: **The heritability of Dupuytren's disease: Familial aggregation and its clinical significance.** *J Hand Surg Am* 2006, **31:**204-210.

55. Fitzgerald AMP, Kirkpatrick JJR, Naylor IL: **Dupuytren's disease - the way forward?** *J Hand Surg Br* 1999, **24:**395-399.

56. Gabbiani G, Majno G: **Dupuytren's contracture: fibroblast contraction? An ultrastructural study.** *Am J Pathol* 1972, **66:**131-146.

57. Tomasek JJ, Gabbiani G, Hinz B, Chaponnier C, RA B: **Myofibroblasts and mechano-regulation of connective tissue remodelling.** *Nat Rev Mol Cell Biol* 2002**:**349-363.

58. Tomasek JJ, Vaughan MB, CJ H: **Cellular structure and biology of Dupuytren's disease.** *Hand Clin* 1999**:**21-34.

59. Kaufman L, Rousseeuw P: *Finding Groups in Data: An Introduction to Cluster Analysis (Wiley Series in Probability and Statistics).* Wiley-Interscience; 2005.

60. Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, Kell DB: **Metabolomics by numbers: acquiring and understanding global metabolite data.** *Trends Biotechnol* 2004, **22:**245-252.

61. Kloen P: **New insights in the development of Dupuytren's contracture: a review.** *British Journal of Plastic Surgery* 1999, **52:**629-635.

62. Alioto RJ, Rosier RN, Burton RI, Edward Puzas J: **Comparative effects of growth factors on fibroblasts of Dupuytren's tissue and normal palmar fascia.** *J Hand Surg Am* 1994, **19:**442-452.

63. Badalamente MA, Hurst LC, Grandia SK, Sampson SP: **Platelet-derived growth factor in Dupuytren's disease.** *J Hand Surg Am* 1992, **17:**317-323.

64. Badalamente MA, Sampson SP, Hurst LC, Dowd A, Miyasaka K: **The role of transforming growth factor beta in Dupuytren's disease.** *The Journal of Hand Surgery* 1996, **21:**210-215.

65. Bayat A, Alansar A, Hajeer AH, Shah M, Watson JS, Stanley JK, Ferguson MWJ, Ollier WER: **Genetic susceptibility in Dupuytren's disease: lack of association of a novel transforming growth factor beta(2) polymorphism in Dupuytren's disease.** *J Hand Surg Br* 2002, **27 B:**47-49.

66. Vaughan MB, Howard EW, Tomasek JJ: **Transforming Growth Factor-[beta]1 Promotes the Morphological and Functional Differentiation of the Myofibroblast.** *Experimental Cell Research* 2000, **257:**180-189.

67. Berndt A, Kosmehl H, Mandel U, Gabler U, Luo X, Celeda D, Zardi L, Katenkamp D: **TGF? and bFGF synthesis and localization in Dupuytren's disease (nodular palmar fibromatosis) relative to cellular activity, myofibroblast phenotype and oncofetal variants of fibronectin.** *Histochemical Journal* 1995, **27:**1014-1020.

68. Lappi DA, Martineau D, Maher PA, Florkiewicz RZ, Buscaglia M, Gonzalez AM, Farris J, Hamer M, Fox R, Baird A: **Basic fibroblast growth factor in cells derived from Dupuytren's contracture: Synthesis, presence, and implications for treatment of the disease.** *J Hand Surg Am* 1992, **17:**324-332.

69. Gonzalez AM, Buscaglia M, Fox R, Isacchi A, Sarmientos P, Farris J, Ong M, Martineau D, Lappi DA, Baird A: **Basic fibroblast growth factor in Dupuytren's contracture.** *Am J Pathol* 1992, **141:**661-671.

70. Serini G, Gabbiani G: **Mechanisms of Myofibroblast Activity and Phenotypic Modulation.** *Experimental Cell Research* 1999, **250:**273-283.

71. Hu FZ, Nystrom A, Ahmed A, Palmquist M, Dopico R, Mossberg I, Gladitz J, Rayner M, Post JC, Ehrlich GD, Preston RA: **Mapping of an autosomal dominant gene for Dupuytren's contracture to chromosome 16q in a Swedish family.** *Clinical Genetics* 2005, **68:**424-429.

72. Brown JJ, Ollier W, Thomson W, Bayat A: **Positive association of HLA-DRB1*15 with Dupuytren's disease in Caucasians.** *Tissue Antigens* 2008, **72:**166-170.

73. Bayat A, Walter J, Lambe H, Watson JS, Stanley JK, Marino M, Ferguson MWJ, Ollier WER: **Identification of a novel mitochondrial mutation in Dupuytren's disease using multiplex DHPLC.** *Plastic and Reconstructive Surgery* 2005, **115:**134-141.

74. Murrell GA FM, Bromley L: **Free radicals and Dupuytren's contracture.** *Br Med J (Clin Res Ed)* 1987, **28:**1373-1375.

75. Kandel J, Bossy-Wetzel E, Radvanyi F, Klagsbrun M, Folkman J, Hanahan D: **Neovascularization is associated with a switch to the export of bFGF in the multistep development of fibrosarcoma.** *Cell* 1991, **66:**1095-1104.

76. Akai M SY, Tateishi T.: **Electrical stimulation on joint contracture: an experiment in rat model with direct current.** *Arch Phys Med Rehabil* 1997 Apr;78(4):405-9.

77. Tart RP, Dahners LE: **Effects of electrical stimulation on joint contracture in a rat model.** *J Orthop Res* 1989, **7:**538-542.

78. Hildebrand KA, Sutherland C, M. Z: **Rabbit knee model of post-traumatic joint contractures: the long-term natural history of motion loss and myofibroblasts.** *J Orthop Res* 2004 Mar;22(2):313-20.

79. Soh J, Turinsky AL, Trinh QM, Chang J, Sabhaney A, Dong X, Gordon PM, Janzen RP, Hau D, Xia J, et al: **Spatiotemporal integration of molecular and anatomical data in virtual reality using semantic mapping.** *Int J Nanomedicine* 2009, **4:**79-89.

80. Westerhoff HV: **The Silicon Cell, Not Dead but Live!** *Metabolic Engineering* 2001, **3:**207-210.

81. Forster T, Roy D, Ghazal P: **Experiments using microarray technology: limitations and standard operating procedures.** *J Endocrinol* 2003, **178:**195-204.

82. Qian A, Meals RA, Rajfer J, Gonzalez-Cadavid NF: **Comparison of gene expression profiles between Peyronie's disease and Dupuytren's contracture.** *Urology* 2004, **64:**399-404.

83. Satish L, Laframboise WA, O'Gorman DB: **Identification of differentially expressed genes in fibroblasts derived from patients with Dupuytren's Contracture.** *BMC Med Genomics* 2008, **1:**1-10.

84. Ulrich D, Ulrich F, Piatkowski A, Pallua N: **Expression of matrix metalloproteinases and their inhibitors in cords and nodules of patients with Dupuytren's disease.** *Arch Orthop Trauma Surg* 2009, **129:**1453-1459.

85. Bayat A, Winder C, Stanley J, Day P, Goodacre R: **Proteome analysis of dupuytren's disease differentiating between disease tissue phenotypes (nodule, cord and transverse palmar fascia) and control palmar fascia.** *J Hand Surg* 2006, **31:**5.

86. O'Gorman D, Howard JC, Varallo VM, Cadieux P, Bowley E, McLean K, Pak BJ, Gan BS: **Identification of protein biomarkers in Dupuytren's contracture using Surface Enhanced Laser Desorption Ionization Time-of-Flight Mass Spectrometry (SELDI-TOF-MS).** *Clin Invest Med* 2006, **29:**136-145.

87. Baggerly KA, Morris JS, Coombes KR: **Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments.** *Bioinformatics* 2004, **20:**777-785.

88. Kraljevic Pavelic S, Sedic M, Hock K, Vucinic S, Jurisic D, Gehrig P, Scott M, Schlapbach R, Cacev T, Kapitanovic S, Pavelic K: **An integrated proteomics approach for studying the molecular pathogenesis of Dupuytren's disease.** *J Pathol* 2009, **217:**524-533.

89. Rhodes DR, Tomlins SA, Varambally S, Mahavisno V, Barrette T, Kalyana-Sundaram S, Ghosh D, Pandey A, AM C: **Probabilistic model of the human protein-protein interaction network.** *Nat Biotechnol* 2005, **23:**951 - 959

90. von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Krüger B, Snel B, P B: **STRING 7 — recent developments in the integration and prediction of protein interactions.** *Nucleic Acids* 2007, **35:**D358–D362.

91. ter Kuile BH, Westerhoff HV: **Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway.** *FEBS Lett* 2001, **500:**169-171.

92. Lazebnik Y: **Can a biologist fix a radio?Or, what I learned while studying apoptosis.** *Cancer Cell* 2002, **2:**179-182.

93. van Driel R, Fransz PF, Verschure PJ: **The eukaryotic genome: a system regulated at different hierarchical levels.** *J Cell Sci* 2003, **116:**4067-4075.

94. de-Leon SB-T, H. E: **Gene Regulation: Gene Control Network in Development.** *Annu Rev Biophys Biomol Struct* 2007, **36:**191-212.

95. Westerhoff HV, Koster JG, van Workum M, Rudd KE: **On the control of gene expression. .** In *Control of metabolic processes.* Edited by Cornish-Bowden A C, M.L., editor. New York: Plenum Press; 1989: 399-413

96. Wagner A, Fell D: **The small world inside large metabolic networks.** *Proc Biol Sci* 2001 **7:**1803–1810.

97. Hendrickson DG, Hogan DJ, McCullough HL, Myers JW, Herschlag D, Ferrell JE, Brown PO: **Concordant Regulation of Translation and mRNA Abundance for Hundreds of Targets of a Human microRNA.** *PLoS Biol* 2009, **7:**e1000238.

98. Westerhoff H, Kolodkin A, Conradie R, Wilkinson S, Bruggeman F, Krab K, van Schuppen J, Hardin H, Bakker B, Moné M, et al: **Systems biology towards life in silico: mathematics of the control of living cells.** *J Math Biol* 2009, **58:**7-34.

99. Westerhoff HV, Aon MA, Vandam K, Cortassa S, Kahn D, M. V: **Dynamic and hierarchical coupling** *Biochim Biophys Acta* 1990, **1018:**142-146.

100. Westerhoff HV, Winder C, Messiha H, Simeonidis E, Adamczyk M, Verma M, Bruggeman FJ, Dunn W: **Systems Biology: The elements and principles of Life.** *FEBS Lett* 2009, **583:**3882-3890.

101. Glass JI, Assad-Garcia N, Alperovich N, Yooseph S, Lewis MR, Maruf M, Hutchison CA, Smith HO, Venter JC: **Essential genes of a minimal bacterium.** *Proc Natl Acad Sci* 2006, **103:**425-430.

102. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, et al: **The minimal gene complement of Mycoplasma genitalium.** *Science* 1995, **270:**397-394.

103. Groen AK, Wanders RJA, Westerhoff HV, Vandermeer R, JM. T: **Quantification of the contribution of various steps to the control of mitochondrial respiration.** *J Biol Chem* 1982, **257:**2754-2757.

104. Stuger R, Woldringh CL, van der Weijden CC, Vischer NOE, Bakker BM, van Spanning RJM, Snoep JL, Weterhoff HV: **DNA Supercoiling by Gyrase is Linked to Nucleoid Compaction.** *Mol Biol Rep* 2002, **29:**79-82.

105. Koster JG, Destrée OHJ, Westerhoff HV: **Kinetics of histone gene expression during early development of Xenopus laevis.** *J Theor Biol* 1988, **135:**139-167.

106. Snoep JL, Van Der Weijden CC, Andersen HW, Westerhoff HV, Jensen PR: **DNA supercoiling in Escherichia coli is under tight and subtle homeostatic control, involving gene-expression and metabolic regulation of both topoisomerase I and DNA gyrase.** *Eur J Biochem* 2002, **269:**1662-1669.

107. Hornberg JJ, Bruggeman FJ, Binder B, Geest CR, De Vaate AJMB, Lankelma J, Heinrich R, Westerhoff HV: **Principles behind the multifarious control of signal transduction.** *FEBS J* 2005, **272:**244-258.

108. Hornberg JJ, Binder B, Bruggeman FJ, Schoeberl B, Heinrich R, Westerhoff HV: **Control of MAPK signalling: from complexity to what really matters.** *Oncogene* 2005, **24:**5533-5542.

109. Heinrich R, Neel BG, TA. R: **Mathematical models of protein kinase signal transduction.** *Mol Cell* 2002, **9:**957-970.

110. Kholodenko BN: **Cell-signalling dynamics in time and space.** *Nat Rev Mol Cell Biol* 2006, **7:**165-176.

111. Bastiaens P: **Systems biology: When it is time to die.** *Nature* 2009, **459:**334-335.

112. Rossell S, van der Weijden CC, Lindenbergh A, van Tuijl A, Francke C, Bakker BM, Westerhoff HV: **Unraveling the complexity of flux regulation: A new method demonstrated for nutrient starvation in Saccharomyces cerevisiae.** *Proc Natl Acad Sci* 2006, **103:**2166-2171.

113. Gardner TS, Dolnik M, Collins JJ: **A theory for controlling cell cycle dynamics using a reversibly binding inhibitor.** *Proc Natl Acad Sci* 1998, **95:**14190-14195.

114. Alberghina L, Westerhoff HV, Novák B, Chen K, Tyson J: **Systems biology of the yeast cell cycle engine.** In *Systems Biology. Volume* 13: Springer Berlin / Heidelberg; 2005: 305-324: *Topics in Current Genetics*].

115. Nelson DE, Ihekwaba AEC, Elliott M, Johnson JR, Gibney CA, Foreman BE, Nelson G, See V, Horton CA, Spiller DG, et al: **Oscillations in NF-{kappa}B Signaling Control the Dynamics of Gene Expression.** *Science* 2004, **306:**704-708.

116. Alberghina L, HV. W: *Systems Biology: Definitions and Perspectives (Topics in Current Genetics). .* Springer-Verlag Berlin and Heidelberg GmbH & Co. K; 2005.

117. Snoep JL, Bruggeman F, Olivier BG, Westerhoff HV: **Towards building the silicon cell: A modular approach.** *Biosystems* 2006, **83:**207-216.

118. Barab, aacute, si A-L, aacute, szl, oacute, Albert R, eacute, ka: **Emergence of Scaling in Random Networks.** *Science* 1999, **286:**509-512.

119. Alon U: **Network motifs: theory and experimental approaches.** *Nat Rev Genet* 2007, **8:**450-461.

120. Schuster S, Hilgetag C, Woods JH, Fell DA: **Reaction routes in biochemical reaction systems: Algebraic properties, validated calculation procedure and example from nucleotide metabolism.** *J Math Biol* 2002, **45:**153-181.

121. Reed J, Palsson B: **Genome-scale in silico models of E. coli have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states.** *Genome Res* 2004, **14:**1797-1805.

122. Marshall A, Gollapudi S, de Silva J, Hodgman C: **A systems biology approach to modelling tea (Camellia sinensis).** *BMC Syst Biol* 2007, **1:**P13.

123. Getz WM, Westerhoff HV, Hofmeyr J-HS, Snoep JL: **Control analysis of trophic chains.** *Ecological Modelling* 2003, **168:**153-171.

124. Reijenga KA, Bakker BM, Van Der Weijden CC, Westerhoff HV: **Training of yeast cell dynamics.** *FEBS J* 2005, **272:**1616-1624.

125. Schuetz R, Kuepfer L, Sauer U: **Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli.** *Mol Syst Biol* 2007, **3**.

126. Molenaar D, van Berlo R, de Ridder D, Teusink B: **Shifts in growth strategies reflect tradeoffs in cellular economics.** *Mol Syst Biol* 2009, **5**.

127. Edwards JS, Covert M, Palsson B: **Metabolic modelling of microbes: the flux-balance approach.** *Environmental Microbiology* 2002, **4:**133-140.

128. Rodríguez-Enríquez S, Marín-Hernández A, Gallardo-Pérez JC, Carreño-Fuentes L, Moreno-Sánchez R: **Targeting of cancer energy metabolism.** *Mol Nutr Food Res* 2009, **53:**29-48.

129. Emery AE: **The muscular dystrophies.** *The Lancet* 2002, **359:**687-695.

130. Williams R, Mamotte C, Burnett J: **Phenylketonuria: an inborn error of 1413 phenylalanine metabolism.** *Clin Biochem Rev* 2008, **29:**31-41.

131. Robert A: **The genetic origins of human cancer.** *Cancer* 1988, **61:**1963-1968.

132. Weigelt B, Reis-Filho JS: **Histological and molecular types of breast cancer: is there a unifying taxonomy?** *Nat Rev Clin Oncol* 2009, **6:**718-730.

133. Kacser H, Burns J: **The molecular basis of dominance.** *Genetics* 1981, **97:**639-666.

134. Kacser H: **Dominance not inevitable but very likely** *J Theor Biol* 1987, **126:**505-506.

135. Price ND, Papin JA, Palsson BO: **Determination of redundancy and systems properties of the metabolic network of Helicobacter pylori using genome-scale extreme pathway analysis.** *Genome Res* 2002, **12:**760-769.

136. Janes KA, Albeck JG, Gaudet S, Sorger PK, Lauffenburger DA, Yaffe MB: **A Systems Model of Signaling Identifies a Molecular Basis Set for Cytokine-Induced Apoptosis.** *Science* 2005, **310:**1646-1653.

137. Bakker BM, Michels PAM, Opperdoes FR, Westerhoff HV: **Glycolysis in Bloodstream Form Trypanosoma brucei Can Be Understood in Terms of the Kinetics of the Glycolytic Enzymes.** *J Biol Chem* 1997, **272:**3207-3215.

138. Teusink B, Passarge J, Reijenga CA, Esgalhado E, Van Der Weijden CC, Schepper M, Walsh MC, Bakker BM, Van Dam K, Westerhoff HV, Snoep JL: **Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry.** *Eur J Biochem* 2000, **267:**5313-5329.

139. Sihag S, Cresci S, Li AY, Sucharov CC, Lehman JJ: **PGC-1[alpha] and ERR[alpha] target gene downregulation is a signature of the failing human heart.** *J Mol Cell Cardiol* 2009, **46:**201-212.

140. McGregor E, Dunn MJ: **Proteomics of the Heart: Unraveling Disease.** *Circ Res* 2006, **98:**309-321.

141. Mayr M, Yusuf S, Weir G, Chung Y-L, Mayr U, Yin X, Ladroue C, Madhu B, Roberts N, De Souza A, et al: **Combined Metabolomic and Proteomic Analysis of Human Atrial Fibrillation.** *J Am Coll Cardiol* 2008, **51:**585-594.

142. Dunn WB, Broadhurst DI, Atherton HJ, Goodacre R, Griffin JL: **Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy.** *Chem Soc Rev* 2010. DOI: 10.1039/B906712B

143. Noble D: **Computational models of the heart and their use in assessing the actions of drugs.** *J Pharmacol Sci* 2008, **107:**107-117.

144. Kohl P, Noble D: **Systems biology and the virtual physiological human.** *Mol Syst Biol* 2009, **5**.

145. Turinsky AL, Fanea E, Trinh Q, Wat S, Hallgrímsson B, Dong X, Shu X, Stromer JN, Hill JW, Edwards C, et al: **CAVEman: Standardized anatomical context for biomedical data mapping.** *Anat Sci Educ* 2008, **1:**10-18.

146. Maxwell C, Moreno V, Sole X, Gomez L, Hernandez P, Urruticoechea A, Pujana M: **Genetic interactions: the missing links for a better understanding of cancer susceptibility, progression and treatment.** *Molecular Cancer* 2008, **7:**4.

147. Franovic A, Holterman CE, Payette J, Lee S: **Human cancers converge at the HIF-2Î± oncogenic axis.** *Proc Natl Acad Sci* 2009, **106:**21306-21311.

148. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, and the rest of the SF, Arkin AP, Bornstein BJ, Bray D, et al: **The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models.** *Bioinformatics* 2003, **19:**524-531.

149. Cuellar AA, Lloyd CM, Nielsen PF, Bullivant DP, Nickerson DP, Hunter PJ: **An Overview of CellML 1.1, a Biological Model Description Language.** *Simulation* 2003, **79:**740-747.

150. Jeong E, Nagasaki M, Saito A, Miyano S: **Cell System Ontology: Representation for Modeling, Visualizing, and Simulating Biological Pathways.** *In Silico Biol* 2007, **7:**623-638.

151. Demir E, Cary MP, Paley S, Fukuda K, Lemer C, Vastrik I, Wu G, D'Eustachio P, Schaefer C, Luciano J, et al: **The BioPAX community standard for pathway data sharing.** *Nat Biotechnol*, **28:**935-942.

152. Le Novère N, Hucka M, Mi H, Moodie S, Schreiber F, Sorokin A, Demir E WK, Aladjem MI, Wimalaratne SM, Bergman FT, Gauges R, Ghazal P, Kawaji H, Li L, Matsuoka Y, Villéger A, Boyd SE, Calzone L, Courtot M, Dogrusoz U, Freeman TC, Funahashi A, Ghosh S, Jouraku A, Kim S, Kolpakov F, Luna A, Sahle S, Schmidt E, Watterson S, Wu G, Goryanin I, Kell DB, Sander C, Sauro H, Snoep JL, Kohn K, Kitano H: **The Systems Biology Graphical Notation.** *Nat Biotechnol* 2009, **27:**735-741.

153. Hoops S, Sahle S, Gauges R, Lee C, Pahle Jr, Simus N, Singhal M, Xu L, Mendes P, Kummer U: **COPASI - a COmplex PAthway SImulator.** *Bioinformatics* 2006, **22:**3067-3074.

154. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks.** *Genome Res* 2003, **13:**2498-2504.

155. Calvano SE, Xiao W, Richards DR, Felciano RM, Baker HV, Cho RJ, Chen RO, Brownstein BH, Cobb JP, Tschoeke SK, et al: **A network-based analysis of systemic inflammation in humans.** *Nature* 2005, **437:**1032-1037.

156. Olivier BG, Snoep JL: **Web-based kinetic modelling using JWS Online** *Bioinformatics* 2004, **20:**2143-2144.

157. Le Novere N, Bornstein B, Broicher A, Courtot Ml, Donizelli M, Dharuri H, Li L, Sauro H, Schilstra M, Shapiro B, et al: **BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems.** *Nucl Acids Res* 2005, **34:**D689-D691.

158. Trew ML, Smaill BH, Bullivant DP, Hunter PJ, Pullan AJ: **A generalized finite difference method for modeling cardiac electrical activation on arbitrary, irregular computational meshes.** *Math Biosci* 2005, **198:**169-189.

159. [http://www.physiome.ox.ac.uk]

160. [http://www.mrc.ac.uk]

161. http://www.uku.fi/nucsys.

162. [http://www.systembiologie.de]

163. van der Greef J, Hankemeier T, McBurney RN: **Metabolomics-based systems biology and personalized medicine: moving towards n = 1 clinical trials?** *Pharmacogenomics* 2006, **7:**1087-1094.

164. Kirkwood TBL: **A systematic look at an old problem.** *Nature* 2008, **451:**644-647.

165. João F. Passos, Zglinicki TV, Kirkwood TBL: **Mitochondria and ageing: winning and losing in the numbers game.** *Bioessays* 2007, **29:**908-917.

166. Shreenivasaiah PK, Rho S-H, Kim T, Kim DH: **An overview of cardiac systems biology.** *J Mol Cell Cardiol* 2008, **44:**460-469.

167. Lemberger T: **Systems biology in human health and disease.** *Mol Syst Biol* 2007, **3**.

168. Bader S, Kühner S, Gavin A-C: **Interaction networks for systems biology.** *FEBS Lett* 2008, **582:**1220-1224.

169. Mani KM, Lefebvre C, Wang K, Lim WK, Basso K, Dalla-Favera R, Califano A: **A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas.** *Mol Syst Biol* 2008, **4**.

170. Miller JA, Oldham MC, Geschwind DH: **A Systems Level Analysis of Transcriptional Changes in Alzheimer's Disease and Normal Aging.** *J Neurosci* 2008, **28:**1410-1420.

171. Krishna R, Schaefer HG, Bjerrum OJ: **Effective integration of systems biology, biomarkers, biosimulation and modelling in streamlining drug development.** *Eur J Pharm Sci* 2007, **31:**62-67.

172. Noble D: *The Music of Life: biology beyond genes.* Oxford Oxford University Press; 2006.

173. Hollywood K, Brison DR, Goodacre R: **Metabolomics: Current technologies and future trends.** *Proteomics* 2006, **6:**4716-4723.

174. Heazell AEP, Brown M, Dunn WB, Worton SA, Crocker IP, Baker PN, Kell DB: **Analysis of the Metabolic Footprint and Tissue Metabolome of Placental Villous Explants Cultured at Different Oxygen Tensions Reveals Novel Redox Biomarkers.** *Placenta* 2008, **29:**691-698.

175. Megason SG, Fraser SE: **Imaging in Systems Biology.** *Cell* 2007, **130:**784-795.

176. Ellis DI, Goodacre R: **Metabolic fingerprinting in disease diagnosis: biomedical applications of infrared and Raman spectroscopy.** *Analyst* 2006, **131:**875-885.

177. Maeder CI, Hink MA, Kinkhabwala A, Mayr R, Bastiaens PIH, Knop M: **Spatial regulation of Fus3 MAP kinase activity through a reaction-diffusion mechanism in yeast pheromone signalling.** *Nat Cell Biol* 2007, **9:**1319-1326.

178. Chen Y-CM, Kappel C, Beaudouin J, Eils R, Spector DL: **Live Cell Dynamics of Promyelocytic Leukemia Nuclear Bodies upon Entry into and Exit from Mitosis.** *Mol Biol Cell* 2008, **19:**3147-3162.

179. Novere NL, Finney A, Hucka M, Bhalla US, Campagne F, Collado-Vides J, Crampin EJ, Halstead M, Klipp E, Mendes P, et al: **Minimum information requested in the annotation of biochemical models (MIRIAM).** *Nat Biotechnol* 2005, **23:**1509-1515.

180. Humphries M, Reynolds A: **Cell-to-cell contact and extracellular matrix.** *Curr Opin Cell Biol* 2009, **21:**613-615.

181. Kim YS, Maruvada P, Milner JA: **Metabolomics in biomarker discovery: future uses for cancer prevention.** *Future Oncology* 2008, **4:**93-102.

Figure Legends

Figure 1: Different stages of disease progression. Stage A generally starts as a small lump in the palm of the hand often just under the digit on the palmar crease. In Stage B the disease spreads up the fascia and into the fingers leading to the development of a cord. Following this, Stage C demonstrates that as the disease spreads up the fingers eventually creating a tight cord, the fingers are forced to progressively bend, and are unable to straighten, effecting an irreversible contracture.

Figure 2: Haematoxylin and Eosin staining of DD tissue from (A) the nodule and (B) cord. The nodule is a focus of proliferating young fibroblasts having no particular orientation and associated with minimal collagen deposition. Nodules slowly disappear leaving a dense contracted fibrous cord which becomes increasingly acellular and tendon-like.

Figure 3: Hypoxia-induced mechanisms leading to production of reactive oxygen species triggering tissue damage at the cellular level. A mechanistic drawing illustrating free radical production leading to tissue damage. Reproduced from reference 73.

Figure 4: The cellular components, molecular functions and biological processes derived from genes in expression studies using Gene Ontology.

Figure 5: Molecular (A) versus Systems Biological (B) disease. A: In the molecular, or single-gene disease, a mutation in or around a piece of DNA causes a change in function of the gene product F. F is solely responsible (or the rate limiting step) for the physiological function that is impaired in the disease, or for the pathology itself. B: In the Systems Biological or network disease, the biological function that is impaired in the disease, or the new pathological function, depends on many factors (called Z here) at the same time. Factors Z themselves depend on many other factors, on genes and environmental (e.g. nutritional, hormonal, age) factors, and ultimately even on the development of the pathology itself. In terms of transcriptomics, changes in any factors shown could correlate somewhat with the disease, in either type of disease. In the molecular disease (A) however, the correlation between the disease and changes in the single causative disease gene should be 100 %. When, as in Systems Biology the cause-effect relationships are investigated, the correlations should be time and perturbation dependent as consistent with the network drawn (e.g. a deletion of Y might not affect the disease totally, but should destroy the causal correlation between gene 2 and disease). The Systems Biology paradigm is not soft however, as in that case the correlation between disease and network state should be 100 %.

Figure 6: The complex reality of most diseases, as proposed here for DD. DD depends of the simultaneous occurrence of a number of malfunctions, each of which is controlled by a network of internal and environmental factors. These networks also have other effects (Z) than DD. In this case of a systems biology disease, only a careful dissection of the network changes on the basis of accurate experiments that involve (i) different points in time/progression of the disease and different genetic and environmental backgrounds, (ii) quantitative experimentation at the transcriptomic, proteomic, metabolomic and functional level, and (iii) computation assisted analysis and experimental design, can lead to understanding of the disease and rational and optimally effective therapies.

Figure 7: A proposed information flow of DD research versus normal fibroblast biology research. In the top-down branch of the systems biology effort, data maps generated by large scale experiments first need to be annotated and subjected to statistical analysis in order to extract biologically relevant information. That information should then generate hypotheses concerning patterns of molecular behaviour or dynamic parameters of the networks. Phenomenological or partly mechanistic mathematical modelling can already help here to weed the impossible from the possible and to enable one to put multiple complex

interactions into single testable hypotheses. Then, predictions can be made and tested. This may spiral through iterations of top-down systems biology into an ever improving set of hypotheses which may become more and more mechanistic. A bottom-up Systems Biology branch of the research may begin with proposed mechanisms (such as stimulation of fibroblast growth because of enhanced ROS production) and make mathematical models of this in order to assist with experimental design. By a spiral of testing and adjusting the hypothesis this will ultimately lead to a hypothesis that is better and better tested and involves more and more of the network. At each step, data will be churned or sublimed into information with a reduction in the amount of unnecessary bits but an increase in accuracy, quality and usefulness to improve and generate stronger models of the DD cell. A metabolic or signaling network can then be represented in silico and its properties studied using computer-simulated perturbations. For instance, the flux balance model could be applied to predict the behaviour of metabolic networks upon perturbation of the optimised metabolites within a metabolic pathway.

Figure 8: A Systems Biology approach to understanding DD. An overview of the proposed integrative network-based analysis characterising DD tumorigenic versus oncogenic and normal mechanisms and guiding therapeutic interventions. Biological organisation has a nested, hierarchical structure. Each hierarchy, however, is bounded to some epistemological degree due to the principles of biocomplexity. Systems Biology takes an integrative approach and tries to synthesize the biological knowledge to understand how the molecules act together within the network of interaction that makes up Life (e.g. the living cell). Classical systems biology approaches will focus on the characterisation and description of mechanisms of cellular control with an emphasis on genetic regulation and intracellular signaling through inferred information from omics experiments. Translational systems biology attempts to extrapolate the knowledge generated at the subcellular level to the type of systemic behavior seen in the clinical setting e.g. imaging and histopathology. Exploding amounts of unravelled biological data cannot be understood by simply drawing lines between interacting molecules, but requires a more organic approach where model building and simulations advance the understanding of DD.

Table Legends

Table 1 Components implicated as modulators of the DD fibroblast transdifferentiation into myofibroblasts.

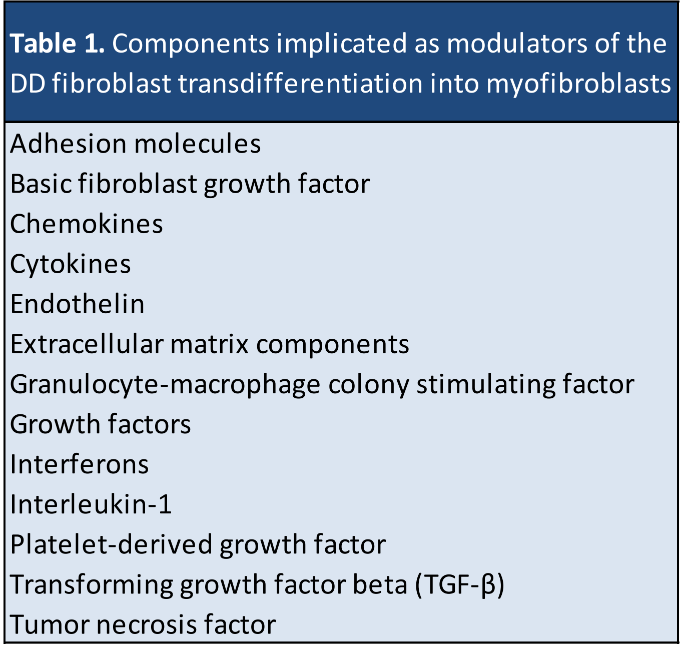
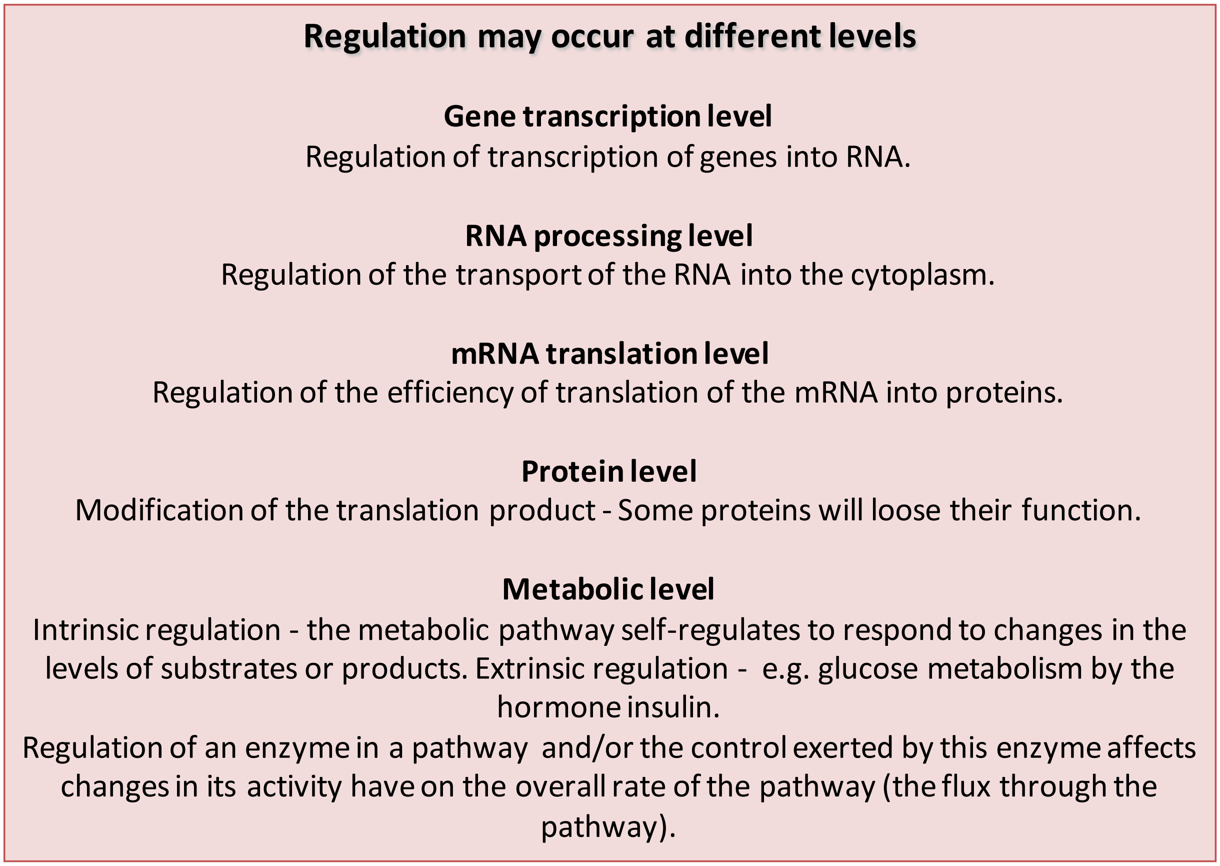


Table 1. Components implicated as modulators of the DD fibroblast transdifferentiation into myofibroblasts
Adhesion molecules
Basic fibroblast growth factor
Chemokines
Cytokines
Endothelin
Extracellular matrix components
Granulocyte-macrophage colony stimulating factor
Growth factors
Interferons
Interleukin-1
Platelet-derived growth factor
Transforming growth factor beta (TGF-β)
Tumor necrosis factor

Table 2 Multiple levels of regulation in the biocomplexed cellular system.



Regulation may occur at different levels

Gene transcription level
Regulation of transcription of genes into RNA.

RNA processing level
Regulation of the transport of the RNA into the cytoplasm.

mRNA translation level
Regulation of the efficiency of translation of the mRNA into proteins.

Protein level
Modification of the translation product - Some proteins will loose their function.

Metabolic level
Intrinsic regulation - the metabolic pathway self-regulates to respond to changes in the levels of substrates or products. Extrinsic regulation - e.g. glucose metabolism by the hormone insulin.
Regulation of an enzyme in a pathway and/or the control exerted by this enzyme affects changes in its activity have on the overall rate of the pathway (the flux through the pathway).

Appendix

Figure 1

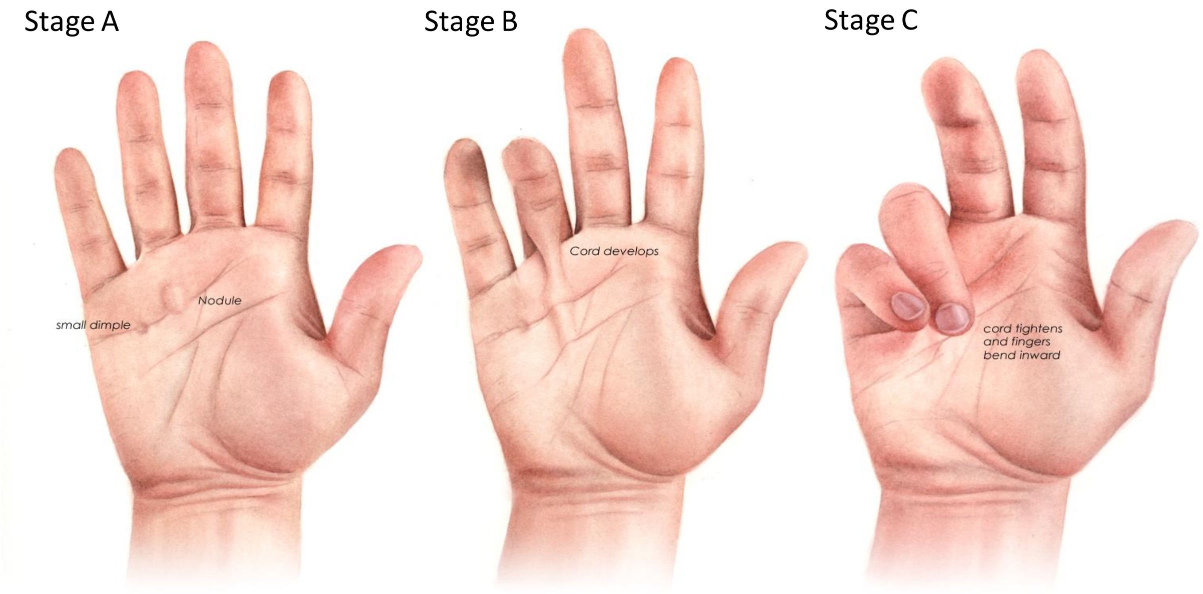
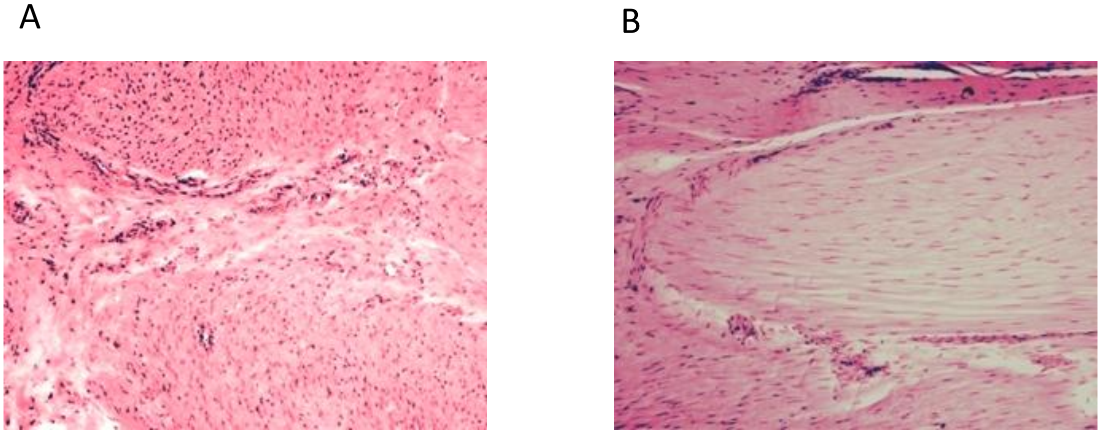


Figure 2



Appendix

Figure 3

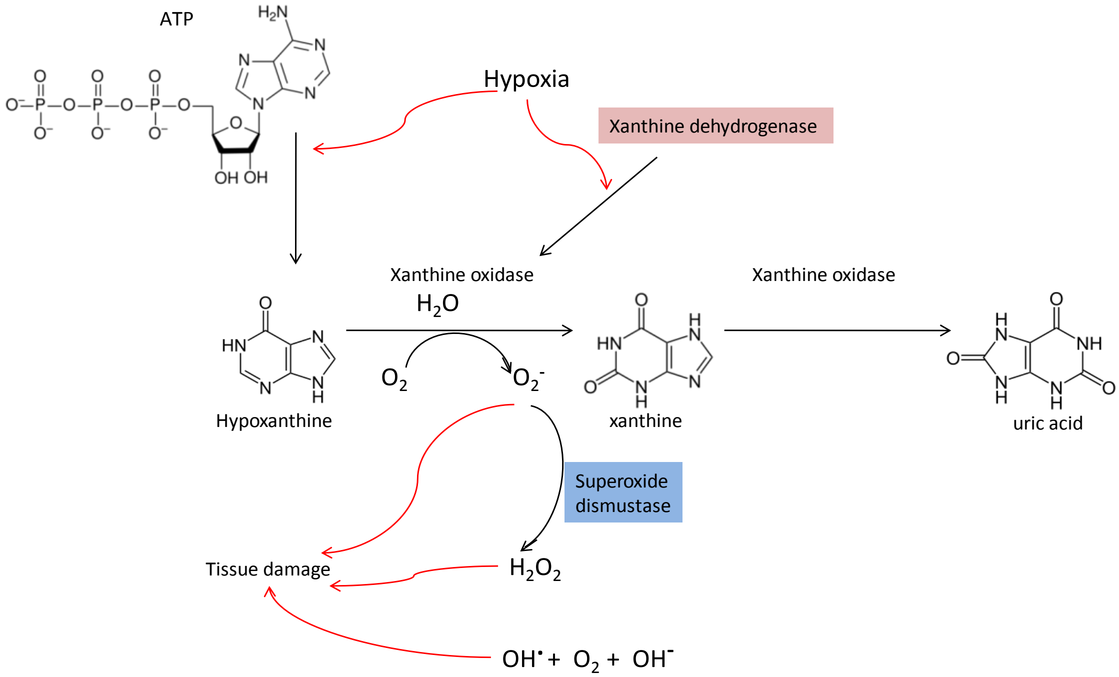
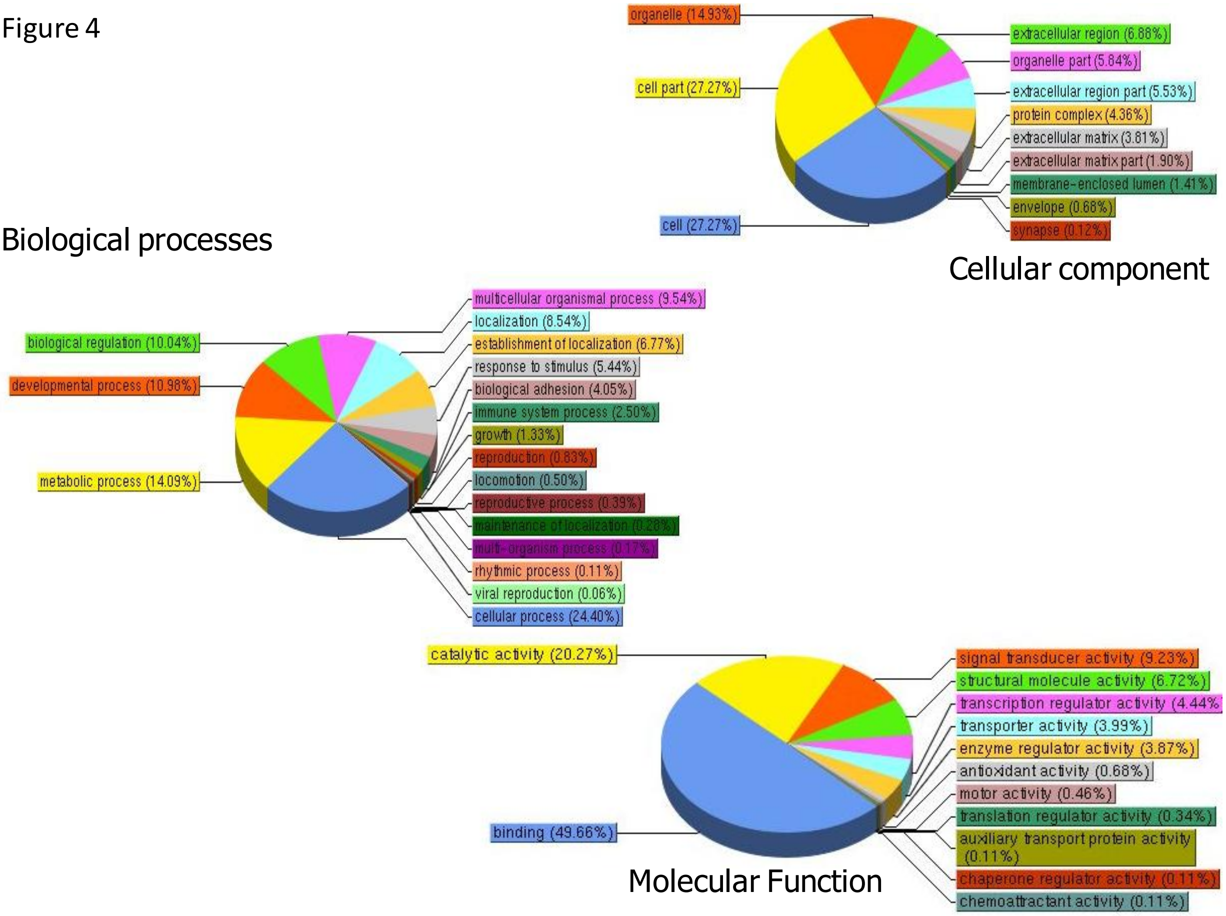


Figure 4



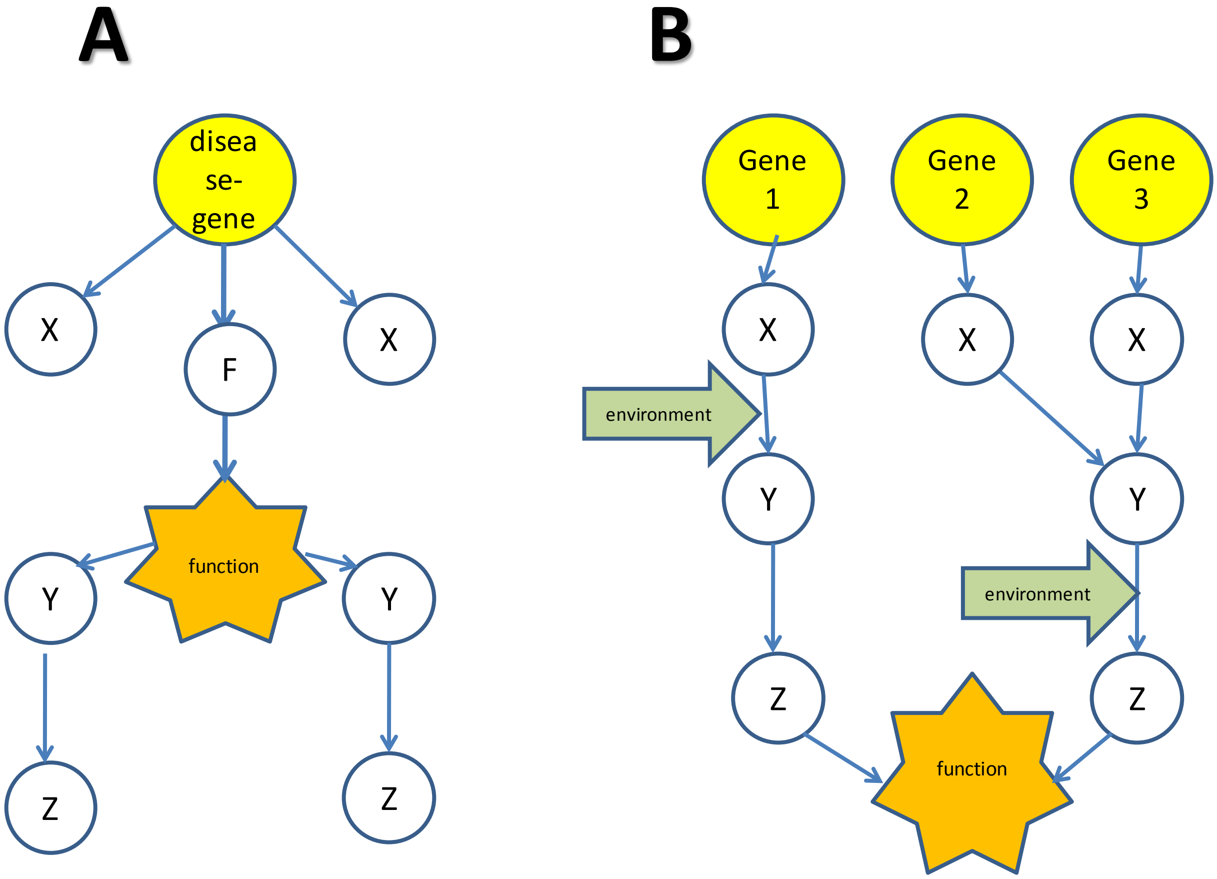
Biological processes

Cellular component

Molecular Function

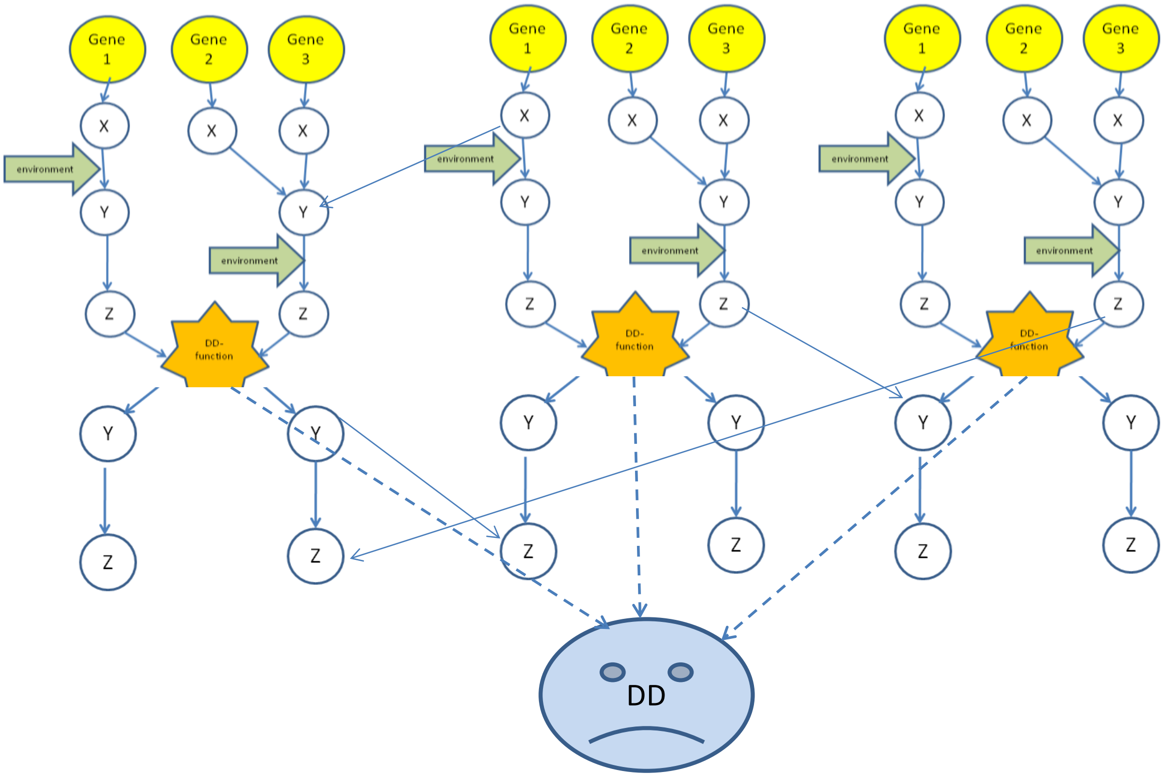
Appendix

Figure 5



Appendix

Figure 6



Appendix

Figure 7

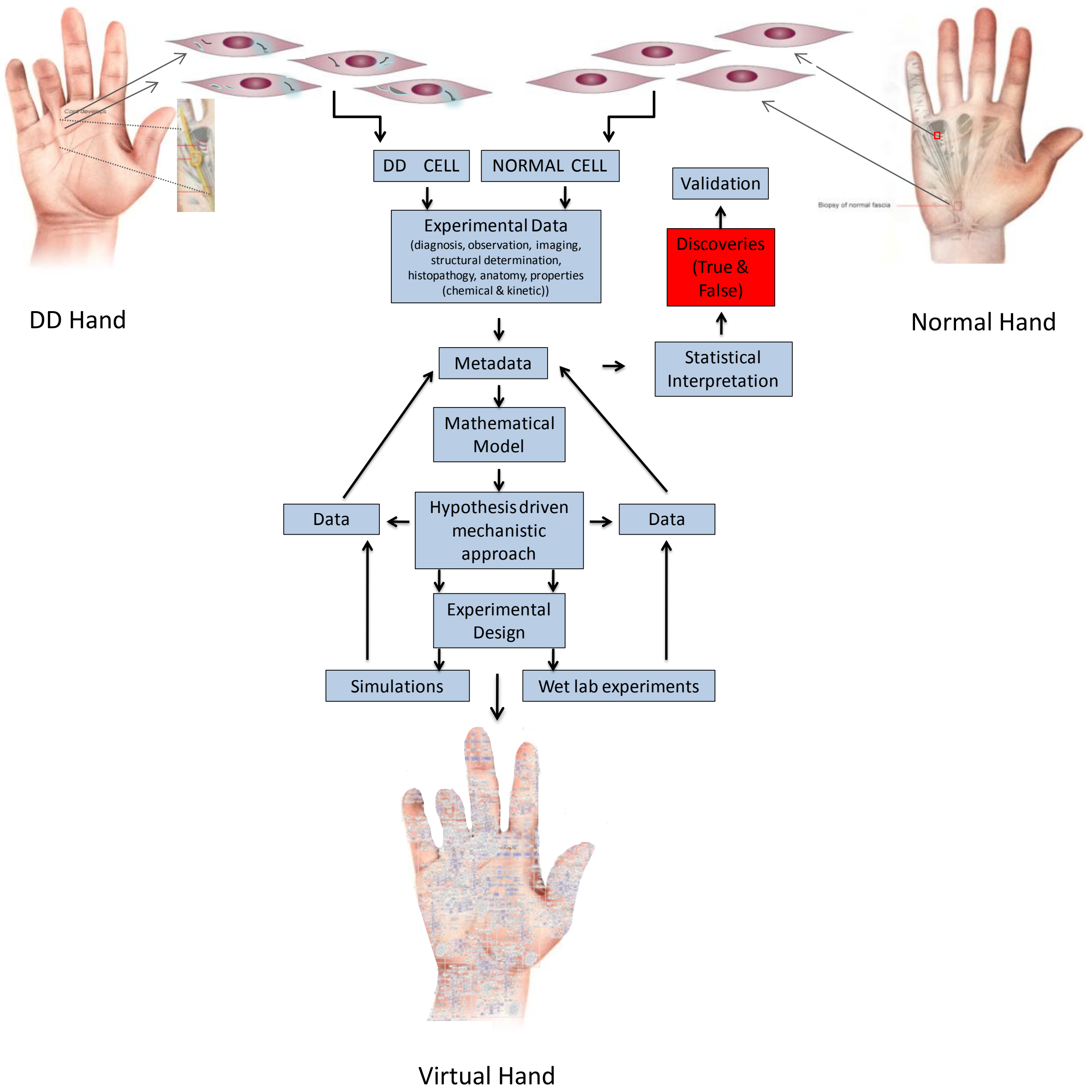
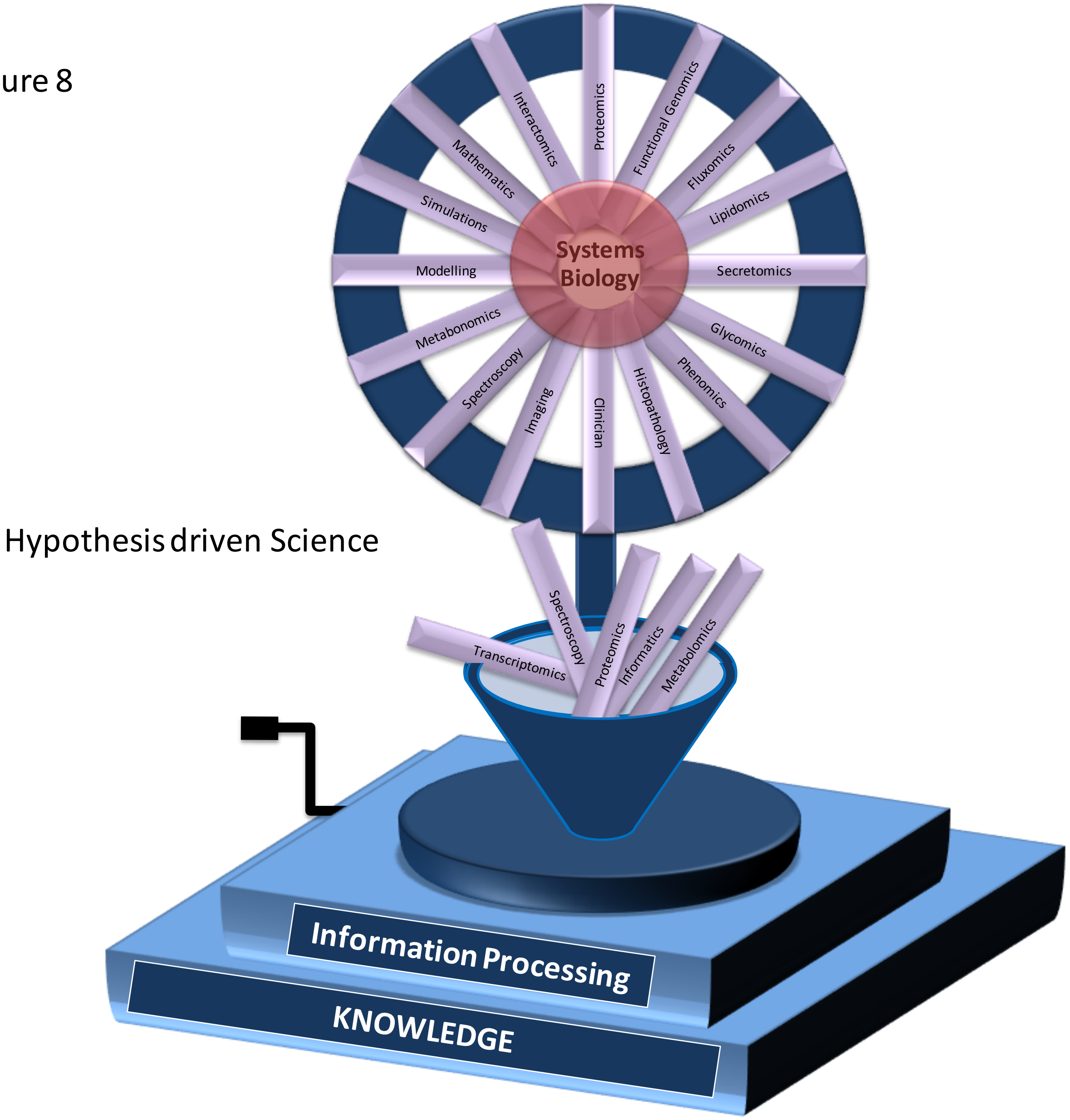

Cord develops
DD Hand
DD CELL
NORMAL CELL
Biopsy of normal fascia
Normal Hand
Experimental Data
(diagnosis, observation, imaging, structural determination, histopathogy, anatomy, properties (chemical & kinetic))
Validation
Discoveries (True & False)
Metadata
Statistical Interpretation
Mathematical Model
Data
Hypothesis driven mechanistic approach
Data
Experimental Design
Simulations
Wet lab experiments
Virtual Hand

Figure 8


Systems Biology
Proteomics
Functional Genomics
Fluxomics
Lipidomics
Secretomics
Glycomics
Phenomics
Histopathology
Clinician
Imaging
Spectroscopy
Metabonomics
Modelling
Simulations
Mathematics
Interactomics
Hypothesis driven Science
Transcriptomics
Spectroscopy
Proteomics
Informatics
Metabolomics
Information Processing
KNOWLEDGE

Appendix B

Table 22 Molecular bonds and their corresponding vibrational modes when molecular orbitals absorb photons certain infrared wavelengths.

Group	Mode	Peak λ (nm)
O - H	stretch	2800
N - H	stretch	3000
C - H	stretch	3400
C = O	stretch	5800
C = C	stretch	6100
N - H	rotation	6500
C - H	rotation	7300
H - C - H	scissors	6800

Table 23 Typical vibrational bonds and their wavenumbers.

Bond	Wavenumber/cm^{-1}
C-H	2840 - 3095
C-C	1610 - 1680
C=O	1680 - 1750
C-O	1000 - 1300
C-Cl	700 - 800
O-H	3233 - 3550 2500 - 3300
N-H	3100 – 3500
C≡C	2000 - 3000

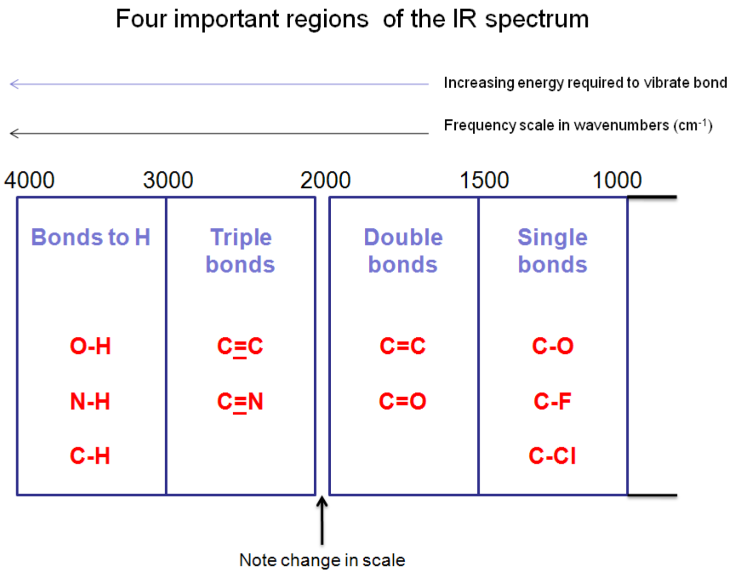

Four important regions of the IR spectrum
Increasing energy required to vibrate bond
Frequency scale in wavenumbers (cm^{-1})
4000
3000
2000
1500
1000
Bonds to H
Triple bonds
Double bonds
Single bonds
O-H
N-H
C-H
C≡C
C≡N
C=C
C=O
C-O
C-F
C-Cl
Note change in scale

Figure 69 Four important regions of the IR spectrum.

Table 24

	Study ID	DOB	Age	Sex	Et bg	DO Operation	Consent	Photo	Proforma	Hospital	Sample type	Site
STUDY 1	DD8	12/11/1931	77	M	W	08/04/2008	Y	N	Y	Derby Hosp	Tissue	Nodule, cord, fascia
	DD9	23/04/1950	58	M	W	08/04/2008	Y	N	Y	Derby Hosp	Tissue	Nodule, cord, fascia, fat, SON
	DD10	20/01/1932	76	M	W	08/04/2008	Y	N	Y	Derby Hosp	Tissue	Nodule, cord, fascia
	DD11	06/05/1940	67	M	W	09/04/2008	Derby	N	Y	Derby Hosp	Blood & tissue	Nodule, cord, fascia, fat, SON
	DD12	03/11/1955	52	M	W	09/04/2008	Derby	N	Y	Derby Hosp	Blood & tissue	Nodule, cord, fascia, fat, SON
	DD13	13/09/1933	74	F	W	09/04/2008	Derby	N	Y	Derby Hosp	Blood & tissue	Nodule, cord, fascia, fat, SON
STUDY 2	DD16	07/08/1962	46	M	W	22/07/2008	Y	ycam	Y		Tissue	Nodule, cord, fascia, fat, SON, Upper Arm skin,
	DD17	18/06/1941	67	M	W	29/07/2008	Y	N	Y		Tissue	Nodule, cord, fascia, fat, SON, Upper Arm skin,
	DD2-R					29/07/2008	Y		Y		Tissue	Nodule, cord, fascia, fat, SON
	DD18	24/02/1940	68	M	W	19/08/2008	Y	N	Y		Tissue	Nodule, cord, fat
	CT4	11/12/1940	67	M	AI	19/08/2008	N-Derby	N	Y	Derby Hosp	Blood & tissue	Fascia, fat, skin
	CT5	21/01/1930	78	M	W	19/08/2008	N-Derby	N	Y	Derby Hosp	Blood & tissue	Fascia, fat, skin
	CT6	14/10/1945	62	F	W	19/08/2008	N-Derby	N	Y	Derby Hosp	Blood & tissue	Fascia, fat, skin
	CT7	18/02/1958	50	F	W	19/08/2008	N-Derby	N	Y	Derby Hosp	Blood & tissue	Fascia, fat, skin
	CT8	08/11/1979	28	M	W	19/08/2008	N-Derby	N	Y	Derby Hosp	Blood & tissue	Fascia, fat, skin

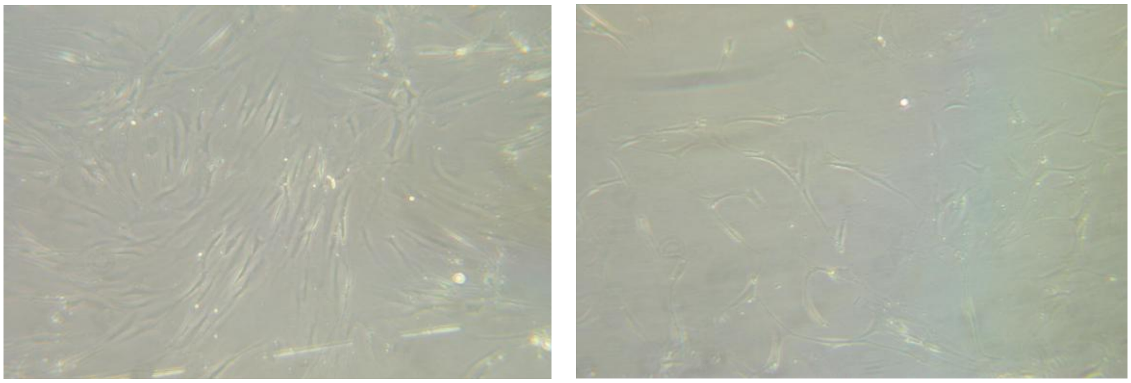


Figure 70 Left (I) cultured fibroblasts from nodules. Right (II) cultured fibroblasts from cord.

Table 25

STUDY 1A					
Study ID	Age	Sex	Site		
DD8	77	M	Nodule	Cord	Fascia
DD9	58	M	Nodule	Cord	Fascia
DD10	76	M	Nodule	Cord	Fascia
DD11	67	M	Nodule	Cord	Fascia
DD12	52	M	Nodule	Cord	Fascia
DD13	74	F	Nodule	Cord	Fascia

Table 26

STUDY 1B							
Study ID	Age	Sex	Site				
DD9	58	M	Nodule	Cord	Fascia	Fat	Skin
DD11	67	M	Nodule	Cord	Fascia	Fat	Skin
DD12	52	M	Nodule	Cord	Fascia	Fat	Skin
DD13	74	F	Nodule	Cord	Fascia	Fat	Skin

Table 27

STUDY 2A					
Study ID	Age	Sex	Site		
DD16	46	M		Cord	Fascia
DD17	67	M	Nodule	Cord	Fascia
DD2-R			Nodule	Cord	Fascia
DD18	68	M	Nodule	Cord	

Table 28

STUDY 2B							
Study ID	Age	Sex	Site				
DD16	46	M		Cord	Fascia		
DD17	67	M	Nodule	Cord	Fascia	Fat	Skin
DD2-R			Nodule	Cord	Fascia	Fat	Skin
DD18	68	M	Nodule	Cord		Fat	

Table 29

STUDY 2C						
Study ID	Age	Sex	Site			
DD16	46	M		Cord	Fascia	
DD17	67	M	Nodule	Cord	Fascia	
DD2-R			Nodule	Cord	Fascia	
DD18	68	M	Nodule	Cord		
CT4	67	M				Fascia*
CT5	78	M				Fascia*
CT6	62	F				Fascia*
CT7	50	F				Fascia*
CT8	28	M				Fascia*

*from CTD patient, external control

Table 30

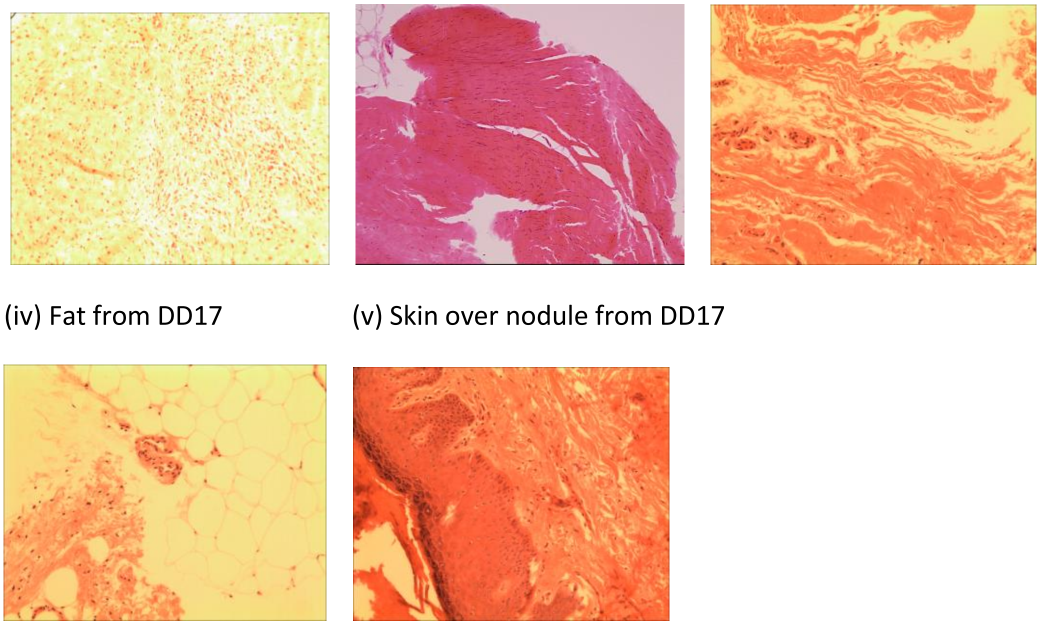
STUDY 2D					
Study ID	Age	Sex		Site	
CT4	67	M	Fascia*	Fat*	Skin*
CT5	78	M	Fascia*	Fat*	Skin*
CT6	62	F	Fascia*	Fat*	Skin*
CT7	50	F	Fascia*	Fat*	Skin*
CT8	28	M	Fascia*	Fat*	Skin*

*from CTD patient, external control

Appendix C

FIGURES 71 (I)-(V) H&E Stains of DD tissue phenotypes.

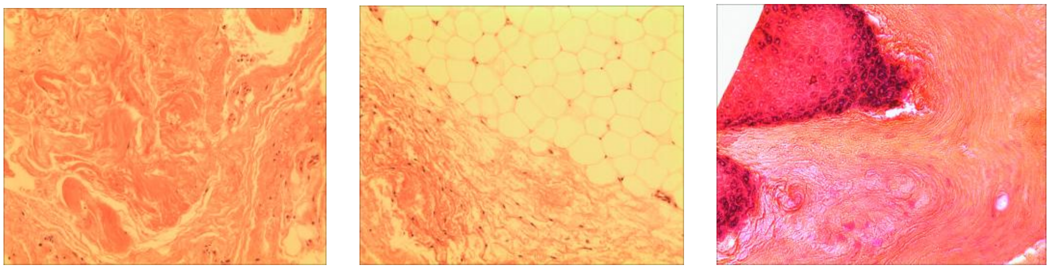
Figure 71 H&E Stains of DD tissue phenotypes and controls



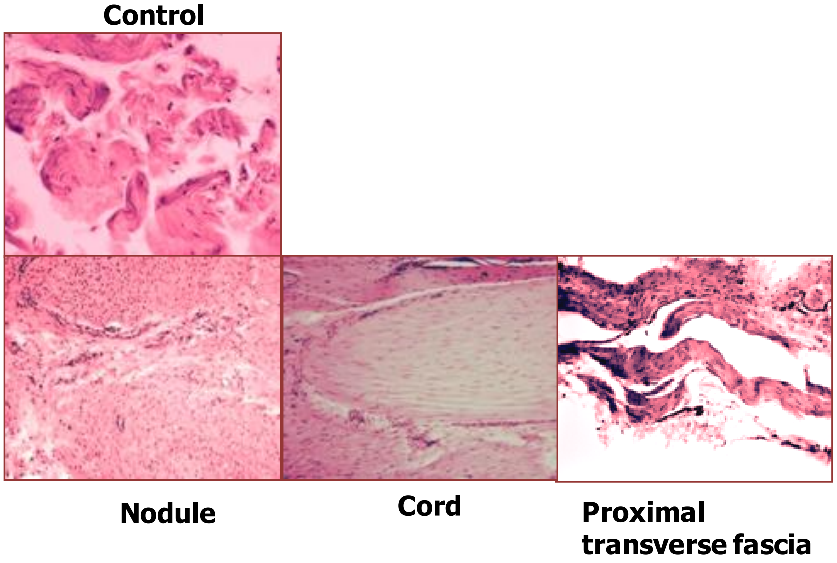
(i) Nodule from DD17 (ii) Cord from DD2 (iii) Palmar fascia from DD17-internal control

(iv) Fat from DD17 (v) Skin over nodule from DD17

FIGURE 71 (VI-VIII) H&E Stains of control tissue.



(vi) Fascia from CT7-external control (vii) Fat from CT7 (viii) Skin from CT6



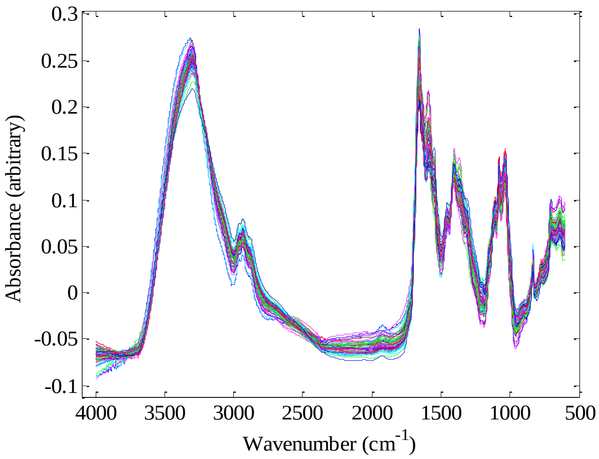


Figure 72 EMSC processed spectra of the secreted metabolites (footprint).

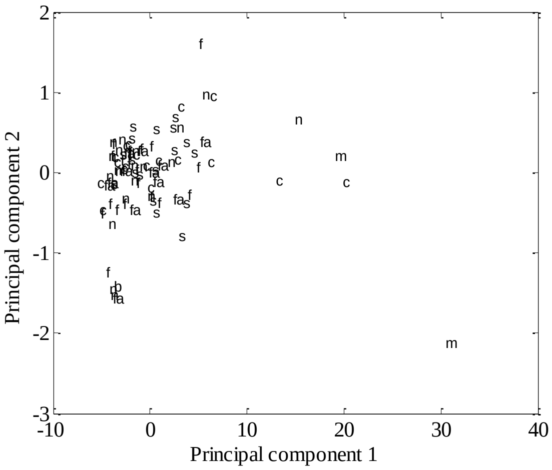
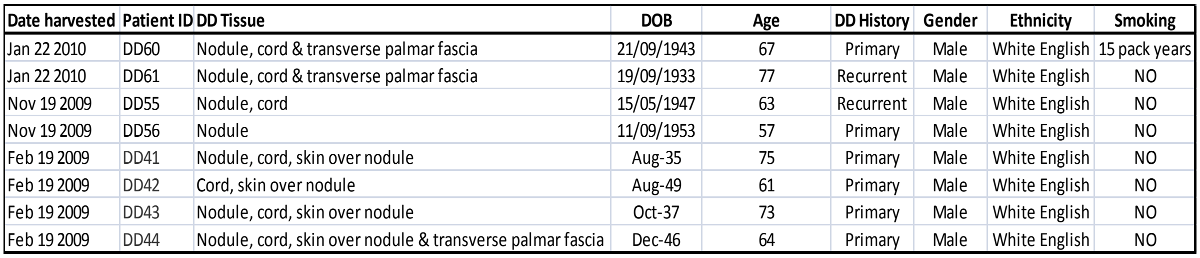


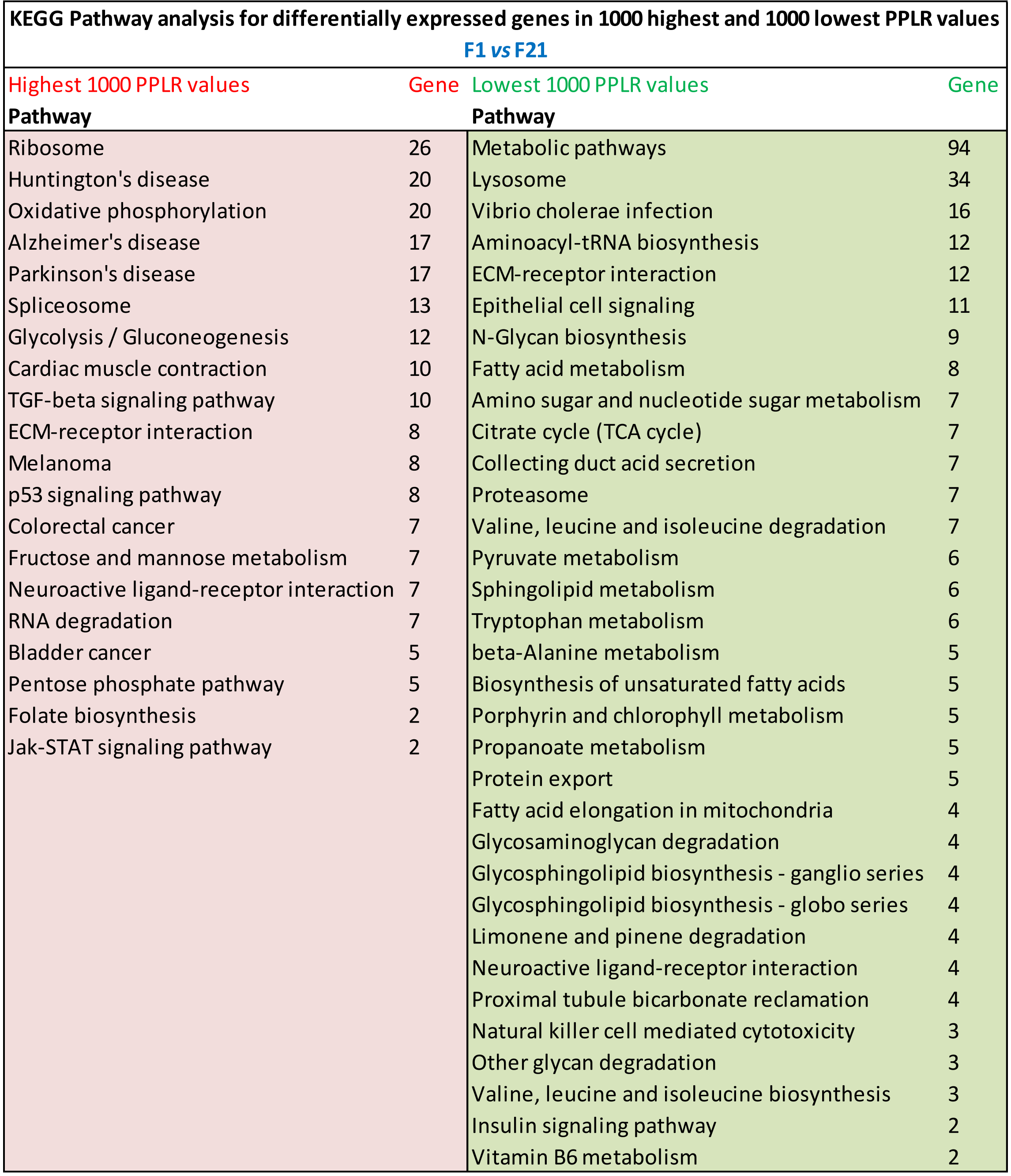
Figure 73 PCA Plot of the FT-IR spectra of metabolic footprint from fibroblasts derived from the DD phenotypes onto the plane defined by PC1 and PC2.

Appendix D



Date harvested	Patient ID	DD Tissue	DOB	Age	DD History	Gender	Ethnicity	Smoking
Jan 22 2010	DD60	Nodule, cord & transverse palmar fascia	21/09/1943	67	Primary	Male	White English	15 pack years
Jan 22 2010	DD61	Nodule, cord & transverse palmar fascia	19/09/1933	77	Recurrent	Male	White English	NO
Nov 19 2009	DD55	Nodule, cord	15/05/1947	63	Recurrent	Male	White English	NO
Nov 19 2009	DD56	Nodule	11/09/1953	57	Primary	Male	White English	NO
Feb 19 2009	DD41	Nodule, cord, skin over nodule	Aug-35	75	Primary	Male	White English	NO
Feb 19 2009	DD42	Cord, skin over nodule	Aug-49	61	Primary	Male	White English	NO
Feb 19 2009	DD43	Nodule, cord, skin over nodule	Oct-37	73	Primary	Male	White English	NO
Feb 19 2009	DD44	Nodule, cord, skin over nodule & transverse palmar fascia	Dec-46	64	Primary	Male	White English	NO

Table 31 KEGG pathways enriched with highest and lowest PPLR values in F1 *vs.* F21.



KEGG Pathway analysis for differentially expressed genes in 1000 highest and 1000 lowest PPLR values

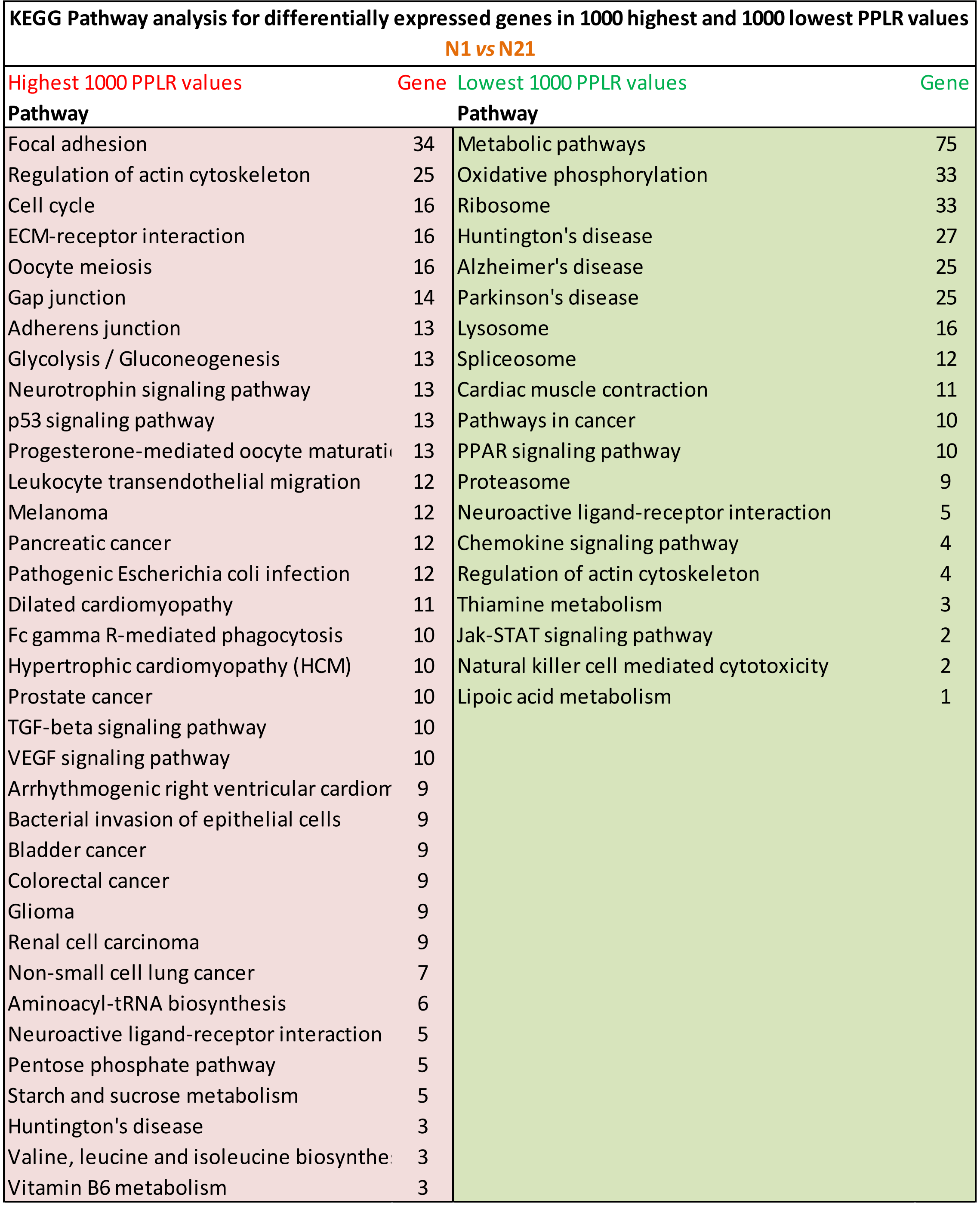
F1 *vs* F21

Highest 1000 PPLR values Pathway	Gene	Lowest 1000 PPLR values Pathway	Gene
Ribosome	26	Metabolic pathways	94
Huntington's disease	20	Lysosome	34
Oxidative phosphorylation	20	Vibrio cholerae infection	16
Alzheimer's disease	17	Aminoacyl-tRNA biosynthesis	12
Parkinson's disease	17	ECM-receptor interaction	12
Spliceosome	13	Epithelial cell signaling	11
Glycolysis / Gluconeogenesis	12	N-Glycan biosynthesis	9
Cardiac muscle contraction	10	Fatty acid metabolism	8
TGF-beta signaling pathway	10	Amino sugar and nucleotide sugar metabolism	7
ECM-receptor interaction	8	Citrate cycle (TCA cycle)	7
Melanoma	8	Collecting duct acid secretion	7
p53 signaling pathway	8	Proteasome	7
Colorectal cancer	7	Valine, leucine and isoleucine degradation	7
Fructose and mannose metabolism	7	Pyruvate metabolism	6
Neuroactive ligand-receptor interaction	7	Sphingolipid metabolism	6
RNA degradation	7	Tryptophan metabolism	6
Bladder cancer	5	beta-Alanine metabolism	5
Pentose phosphate pathway	5	Biosynthesis of unsaturated fatty acids	5
Folate biosynthesis	2	Porphyrin and chlorophyll metabolism	5
Jak-STAT signaling pathway	2	Propanoate metabolism	5
		Protein export	5
		Fatty acid elongation in mitochondria	4
		Glycosaminoglycan degradation	4
		Glycosphingolipid biosynthesis - ganglio series	4
		Glycosphingolipid biosynthesis - globo series	4
		Limonene and pinene degradation	4
		Neuroactive ligand-receptor interaction	4
		Proximal tubule bicarbonate reclamation	4
		Natural killer cell mediated cytotoxicity	3
		Other glycan degradation	3
		Valine, leucine and isoleucine biosynthesis	3
		Insulin signaling pathway	2
		Vitamin B6 metabolism	2

Table 32 KEGG pathways enriched with highest and lowest PPLR values in N21 *vs.* F21

KEGG Pathway analysis for differentially expressed genes in 1000 highest and 1000 lowest PPLR values in N21 *vs* F21			
Highest 1000 PPLR values	Gene	Lowest 1000 PPLR values	Gene
Pathway		**Pathway**	
Ribosome	39	Focal adhesion	37
Oxidative phosphorylation	32	Pathways in cancer	34
Alzheimer's disease	26	Regulation of actin cytoskeleton	23
Huntington's disease	26	ECM-receptor interaction	19
Parkinson's disease	25	Cell cycle	15
Cardiac muscle contraction	13	Oocyte meiosis	14
Lysosome	12	Vascular smooth muscle contraction	14
Spliceosome	11	Gap junction	13
Proteasome	6	p53 signaling pathway	13
Regulation of actin cytoskeleton	3	Pancreatic cancer	11
Thiamine metabolism	2	Small cell lung cancer	11
Lipoic acid metabolism	1	Adherens junction	10
		Arrhythmogenic right ventricular cardiomyopath	10
		Dilated cardiomyopathy	10
		Pathogenic Escherichia coli infection	10
		Progesterone-mediated oocyte maturation	10
		Viral myocarditis	10
		Bacterial invasion of epithelial cells	9
		Glioma	9
		Melanoma	9
		Colorectal cancer	8
		Aminoacyl-tRNA biosynthesis	7
		Bladder cancer	7
		Non-small cell lung cancer	7
		Shigellosis	7
		Malaria	6
		Dorso-ventral axis formation	4
		Pentose phosphate pathway	4
		Thyroid cancer	4
		Huntington's disease	3
		Neuroactive ligand-receptor interaction	3
		One carbon pool by folate	3
		Valine, leucine and isoleucine biosynthesis	3
		Vitamin B6 metabolism	2
		Oxidative phosphorylation	1

Table 33 KEGG pathways enriched with highest and lowest PPLR values in N1 *vs.* N21



KEGG Pathway analysis for differentially expressed genes in 1000 highest and 1000 lowest PPLR values			
N1 *vs* N21			
Highest 1000 PPLR values	Gene	Lowest 1000 PPLR values	Gene
Pathway		**Pathway**	
Focal adhesion	34	Metabolic pathways	75
Regulation of actin cytoskeleton	25	Oxidative phosphorylation	33
Cell cycle	16	Ribosome	33
ECM-receptor interaction	16	Huntington's disease	27
Oocyte meiosis	16	Alzheimer's disease	25
Gap junction	14	Parkinson's disease	25
Adherens junction	13	Lysosome	16
Glycolysis / Gluconeogenesis	13	Spliceosome	12
Neurotrophin signaling pathway	13	Cardiac muscle contraction	11
p53 signaling pathway	13	Pathways in cancer	10
Progesterone-mediated oocyte maturati	13	PPAR signaling pathway	10
Leukocyte transendothelial migration	12	Proteasome	9
Melanoma	12	Neuroactive ligand-receptor interaction	5
Pancreatic cancer	12	Chemokine signaling pathway	4
Pathogenic Escherichia coli infection	12	Regulation of actin cytoskeleton	4
Dilated cardiomyopathy	11	Thiamine metabolism	3
Fc gamma R-mediated phagocytosis	10	Jak-STAT signaling pathway	2
Hypertrophic cardiomyopathy (HCM)	10	Natural killer cell mediated cytotoxicity	2
Prostate cancer	10	Lipoic acid metabolism	1
TGF-beta signaling pathway	10		
VEGF signaling pathway	10		
Arrhythmogenic right ventricular cardiom	9		
Bacterial invasion of epithelial cells	9		
Bladder cancer	9		
Colorectal cancer	9		
Glioma	9		
Renal cell carcinoma	9		
Non-small cell lung cancer	7		
Aminoacyl-tRNA biosynthesis	6		
Neuroactive ligand-receptor interaction	5		
Pentose phosphate pathway	5		
Starch and sucrose metabolism	5		
Huntington's disease	3		
Valine, leucine and isoleucine biosynthe	3		
Vitamin B6 metabolism	3		

Appendix E

Awards and Grants Associated With This Work/PhD

1. Doctoral Training Centre (BBSRC/EPSRC) Studentship, Oct 2006 - Sep 2010.

2. Biochemical Society travel grant to attend 10th International Conference on Systems Biology, Stanford, USA, Aug 2009.

3. Invitrogen award to present work at University of California, Irvine, USA, May 2010.

4. Analytical Chemistry Trust Fund of the Royal Society of Chemistry, Jan 2011.

5. Federation of European Biochemical Societies Youth Travel Fund Award to present at FEBS-SystemsX-SysBio2011, Austria, Feb 2011.

6. Keystone Symposia Future of Science Fund scholarship for Omics meets cell biology, Alpbach, Austria, May 2011.

Publications and Presentations Arising From This Work

E1 Published Abstracts and presentations

1. “Inferring the metabolic and transcriptional networks specific to Dupuytren’s disease tumours with omics”, Joint FEBS/SystemsX 4th Advanced Lecture Course on Systems Biology; FEBS-SystemsX-SysBio2011: From Molecules to Function, Innsbruck, Austria, Feb 2011.

2. “Parallel analysis of transcript and metabolic profiles of Dupuytren’s Disease fibroblasts in response to altered O2 tension”, Systems Biology of Stem Cells Symposium, UC, Irvine, USA, May 2010.

3. “Systems Biology approaches for identifying cultured Dupuytren's Disease samples for analysis”, The 10th International Conference on Systems Biology, ICSB, Stanford, USA, Aug 2009.

4. “Metabolic Fingerprint Analysis of Fibroblasts Derived from Differential Dupuytren's Disease Tissue Phenotypes”, Research & Innovation Exhibition, Central Manchester and Manchester Children's University Hospitals, NHS Trust, UK, Nov 2008.

5. “Computability in Biology: Metabolic Networks”, FEBS-SysBio2007 From Molecules to Life: March 2007, 2nd FEBS Advanced Lecture Course on Systems Biology, Gosau, Austria.

E2 Published Work

1. **Rehman S**, Salway F, Stanley JK, et al. Molecular Phenotypic Descriptors of Dupuytren's Disease Defined Using Informatics Analysis of the Transcriptome. J Hand Surg Am. 2008;33:359-372.

2. Bevilacqua A, Wilkinson SJ, Dimelow R, Murabito E, **Rehman S**, Nardelli M, van Eunen K, Rossell S, Bruggeman FJ, Blüthgen N, De Vos D, Bouwman J, Bakker BM, Westerhoff HV. SEB Exp Biol Ser. 2008; 61:65-91.Vertical systems biology: from DNA to flux and back.

3. Li K, Zhu W, Zeng K, Zhang Z, Ye J, Ou W, **Rehman S**, Heuer B, Chen S. Proteome characterization of cassava (Manihot esculenta Crantz) somatic embryos, plantlets and tuberous roots. Proteome Sci. 2010 Feb 27;8(1):10.

E3 Submissions for Publication

1. **Rehman S**, Goodacre R, Day PJ, Bayat A, Westerhoff HV. Dupuytren's - A Systems Biology Disease? - review paper submitted to Arthritis Research & Therapy Mar 2011.

2. **Rehman S**, Day PJ. Dupuytren's Depicting accurate relationships in bio-networks from natural language processing approaches –*In preparation.* See abstract and text mining workflow below.

3. **Rehman S**, Day PJ, Xu. Y, Dunn WB, Westerhoff HV. Goodacre R, Bayat A, Metabolic Fingerprint Analysis of Fibroblasts Derived from Differential Dupuytren's Disease Tissue Phenotypes, *In preparation.*

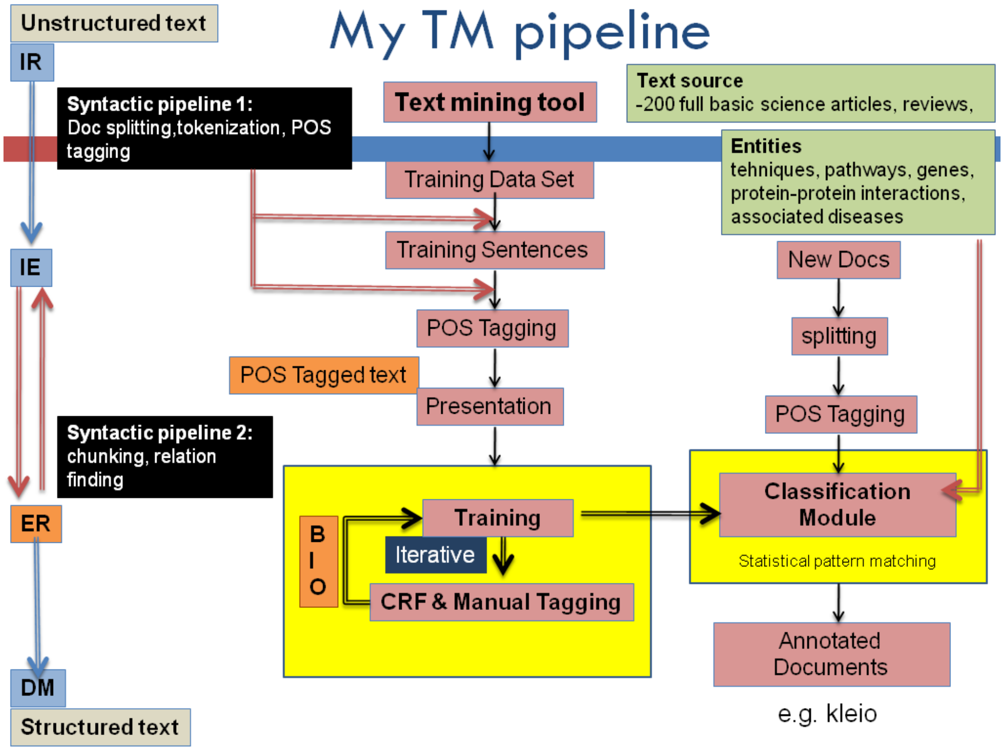
4. **Rehman S**, Day PJ, Xu. Y, Dunn WB. Goodacre R, Bayat A, Westerhoff HV – The effect of Hypoxia in DD metabolomes and transcriptomes investigated through a Systems Approach, *In preparation.*

E3(2)

Abstract

Effective use of published scientific literature plays a crucial role in all stages of research. High-throughput experimental techniques used in ‘omics and integrative systems biology are generating exponential yet unraveled complex data sets. One of the goals of systems biology is to obtain overall quantitative description of dynamic cellular systems. This is currently not achievable as the number of components and interactions involved in these systems is quite large resulting in a very large parameter space, thus the generation of quantitative data sets. The important role of accurate and crucial literature analysis would improve and direct a large number of experimental studies that are initiated by mechanistic and hypothesis driven approaches. Natural Language Processing techniques combined with text mining have been propose as potential new elements for knowledge discovery as part of an application to biological investigation. Nevertheless, current text mining tools do not provide correlations within scientific entities (e.g. gene1-gene2) on the basis of biological relevance but are due to statistical recognition only. Text mining should now play a crucial role not only retrieving commonalities across entities on the basis of co-occurrence but also during a research study itself where the aim would be to validate, annotate and interpret the discovery results generated from analysing the experimentally generated data. In this article we describe our efforts in developing literature analysis and text mining solutions for extracting and documenting known entity-entity (e.g. gene-tissue) interactions by keywords searches based on the experimental methods they are discovered through with respect to the study. We describe an abstract algorithm/pipeline and show how it can be mapped upon our existing workflow FACTA, a concrete service-based text mining workflow using a mix of text processing and data mining components.

Appendix



My TM pipeline

Unstructured text

IR

Syntactic pipeline 1: Doc splitting,tokenization, POS tagging

Text mining tool

Text source
-200 full basic science articles, reviews,

Entities
tehniques, pathways, genes, protein-protein interactions, associated diseases

Training Data Set

Training Sentences

IE

POS Tagging

POS Tagged text

Presentation

New Docs

splitting

POS Tagging

Syntactic pipeline 2: chunking, relation finding

ER

B
I
O

Training

Iterative

CRF & Manual Tagging

Classification Module

Statistical pattern matching

Annotated Documents

e.g. kleio

DM

Structured text

Bibliography

1. Rayan GM: **Dupuytren Disease: Anatomy, Pathology, Presentation, and Treatment.** *J Bone Joint Surg Am* 2007, **89:**189-198.
2. Au-Yong ITH, Wildin CJ, Dias JJ, Page RE: **A review of common practice in dupuytren surgery.** *Techniques in Hand and Upper Extremity Surgery* 2005, **9:**178-187.
3. McFarlane RM MD, Flint MH, editors: *Dupuytren's disease: biology and treatment.* New York: Churchill Livingstone; 1990.
4. Gabbiani G, Majno G: **Dupuytren's contracture: fibroblast contraction? An ultrastructural study.** *Am J Pathol* 1972, **66:**131-146.
5. Hill NA, Hurst LC: **Dupuytren's contracture.** *Hand Clin* 1989, **5:**349-357.
6. Orsi R, Zorzi F, Brunelli G, Sacchi G: **Histological and ultrastructural observations in 15 cases of Dupuytren's disease.** *Osservazioni istologiche ed ultrastrutturali in 15 casi di morbo di Dupuytren* 1983, **75:**829-836.
7. Gelberman RH, Amiel D, Rudolph RM, Vance RM: **Dupuytren's contracture. An electronic microscopic, biochemical, and clinical correlative study.** *Journal of Bone and Joint Surgery - Series A* 1980, **62:**425-432.
8. Ulrich D, Ulrich F, Piatkowski A, Pallua N: **Expression of matrix metalloproteinases and their inhibitors in cords and nodules of patients with Dupuytren's disease.** *Arch Orthop Trauma Surg* 2009, **129:**1453-1459.
9. Satish L, Laframboise WA, O'Gorman DB: **Identification of differentially expressed genes in fibroblasts derived from patients with Dupuytren's Contracture.** *BMC Med Genomics* 2008, **1:**1-10.
10. Rehman S, Salway F, Stanley JK, Ollier WER, Day P, Bayat A: **Molecular Phenotypic Descriptors of Dupuytren's Disease Defined Using Informatics Analysis of the Transcriptome.** *J Hand Surg Am* 2008, **33:**359-372.
11. Bayat BSJJBDJATLA: **Differential Gene Expression Analysis of Subcutaneous Fat, Fascia, and Skin Overlying a Dupuytren's Disease Nodule in Comparison to Control Tissue.** *Hand* 2009, **In press**.
12. An G, Hunt CA, Clermont G, Neugebauer E, Vodovotz Y: **Challenges and rewards on the road to translational systems biology in acute illness: four case reports from interdisciplinary teams.** *J Crit Care* 2007, **22:**169-175.
13. Westerhoff HV, Palsson BO: **The evolution of molecular biology into systems biology.** *Nat Biotechnol* 2004, **22:**1249-1252.
14. Ahn AC, Tewari M, Poon C-S, Phillips RS: **The Limits of Reductionism in Medicine: Could Systems Biology Offer an Alternative?** *PLoS Medicine* 2006, **3:**e208.
15. Katagiri F: **Attacking Complex Problems with the Power of Systems Biology.** *Plant Physiol* 2003, **132:**417-419.
16. Spivey A: **Systems Biology: The Big Picture.** *Environ Health Perspect* 2004, **112:** A938–A943.
17. http://www.virtual-liver.de/.
18. Seyhan H, Kopp J, Schultze-Mosgau S, Horch RE: **Increased metabolic activity of fibroblasts derived from cords compared with nodule fibroblasts sampling from patients with Dupuytren's contracture.** *Plast Reconstr Surg* 2006, **117:**1248-1252.
19. Shih B, Wijeratne D, Armstrong DJ, Lindau T, Day P, Bayat A: **Identification of Biomarkers in Dupuytren's Disease by Comparative Analysis of Fibroblasts Versus Tissue Biopsies in Disease-Specific Phenotypes.** *J Hand Surg* 2009, **34:**124-136.
20. Vaughan MB, Howard EW, Tomasek JJ: **Transforming Growth Factor-[beta]1 Promotes the Morphological and Functional Differentiation of the Myofibroblast.** *Experimental Cell Research* 2000, **257:**180-189.
21. Esquenet M, Swinnen JV, Heyns W, Verhoeven G: **LNCaP prostatic adenocarcinoma cells derived from low and high passage numbers display divergent responses not only to androgens but also to retinoids.** *The Journal of Steroid Biochemistry and Molecular Biology* 1997, **62:**391-399.

22. Chakraborty S, Reid S: **Serial Passage of aHelicoverpa armigeraNucleopolyhedrovirus inHelicoverpa zeaCell Cultures.** *Journal of Invertebrate Pathology* 1999, **73:**303-308.

23. Briske-Anderson MJ, Finley JW, Newman SM: **The influence of culture time and passage number on the morphological and physiological development of Caco-2 cells.** vol. 214. pp. 248-257; 1997:248-257.

24. Wenger SLS, Jamie R. Sargent, Linda M. Bamezai, Ramesh Bairwa, Narendra Grant, Stephen G.: **Comparison of Established Cell Lines at Different Passages by Karyotype and Comparative Genomic Hybridization.** *Bioscience Reports* 2004, **24:**631-639.

25. Galligan CL, Baig E, Bykerk V, Keystone EC, Fish EN: **Distinctive gene expression signatures in rheumatoid arthritis synovial tissue fibroblast cells: correlates with disease activity.** *Genes Immun* 2007, **8:**480-491.

26. Moyer KE, Banducci DR, Graham WP, Ehrlich HP: **Dupuytren's disease: physiologic changes in nodule and cord fibroblasts through aging in vitro.** *Plast Reconstr Surg* 2002, **110:**187 - 193.

27. Schöneich C: **Proteomics in gerontological research.** *Experimental Gerontology* 2003, **38:**473-481.

28. James K. Leung OMP-S: **Identification of Genes Involved in Cell Senescence and Immortalization: Potential Implications for Tissue Ageing.** In *Ageing Vulnerability: Causes and Interventions.* Edited by Gregory Bock JAG; 2003: 105-115

29. Benvenuti S CR, Bruce J, Waterfield MD, Jat PS.: **Identification of novel candidates for replicative senescence by functional proteomics.** *Oncogene* 2002, **21(28):**4403-4413.

30. Naumann D, Helm D, Labischinski H: **Microbiological characterizations by FT-IR spectroscopy.** *Nature* 1991, **351:**81-82.

31. Goodacre R, Timmins EM, Burton R, Kaderbhai N, Woodward AM, Kell DB, Rooney PJ: **Rapid identification of urinary tract infection bacteria using hyperspectral whole-organism fingerprinting and artificial neural networks.** *Microbiology* 1998, **144:**1157-1170.

32. Kaderbhai NN, Broadhurst DI, Ellis DI, Goodacre R, Kell DB: **Functional genomics via metabolic footprinting: Monitoring metabolite secretion by Escherichia coli tryptophan metabolism mutants using FT-IR and direct injection electrospray mass spectrometry.** *Comparative and Functional Genomics* 2003, **4:**376-391.

33. Brison DR, Hollywood K, Arnesen R, Goodacre R: **Predicting human embryo viability: The road to non-invasive analysis of the secretome using metabolic footprinting.** *Reproductive BioMedicine Online* 2007, **15:**296-302.

34. Ellis DI, Goodacre R: **Metabolic fingerprinting in disease diagnosis: biomedical applications of infrared and Raman spectroscopy.** *Analyst* 2006, **131:**875-885.

35. Hsu PP, Sabatini DM: **Cancer Cell Metabolism: Warburg and Beyond.** *Cell* 2008, **134:**703-707.

36. DeBerardinis RJ, Sayed N, Ditsworth D, Thompson CB: **Brick by brick: metabolism and tumor cell growth.** *Current Opinion in Genetics & Development* 2008, **18:**54-61.

37. Levine AJ, Puzio-Kuter AM: **The Control of the Metabolic Switch in Cancers by Oncogenes and Tumor Suppressor Genes.** vol. 330. pp. 1340-1344:1340-1344.

38. Warburg O: **On the Origin of Cancer Cells.** vol. 123. pp. 309-314; 1956:309-314.

39. Harris AL: **Hypoxia [mdash] a key regulatory factor in tumour growth.** *Nat Rev Cancer* 2002, **2:**38-47.

40. Vander Heiden MG, Cantley LC, Thompson CB: **Understanding the Warburg Effect: The Metabolic Requirements of Cell Proliferation.** vol. 324. pp. 1029-1033; 2009:1029-1033.

41. Wheaton WW, Chandel NS: **Hypoxia regulates Cell Metabolism.**

42. ARTHUR K. BALIN DBPG, HOWARD RASMUSSEN, VINCENT J. CRISTOFALO: **THE EFFECT OF OXYGEN AND VITAMIN E ON THE LIFESPAN OF HUMAN DIPLOID CELLS IN VITRO.** *ThE JOURNAL OF CELL BIOLOGY* 1977, **74:**58-67.

43. Swan AJ, Tawhai MH: **Evidence for minimal oxygen heterogeneity in the healthy human pulmonary acinus.**

44. Gatenby RA, Gillies RJ: **Why do cancers have high aerobic glycolysis?** *Nat Rev Cancer* 2004, **4:**891-899.

45. Elstrom RL, Bauer DE, Buzzai M, Karnauskas R, Harris MH, Plas DR, Zhuang H, Cinalli RM, Alavi A, Rudin CM, Thompson CB: **Akt Stimulates Aerobic Glycolysis in Cancer Cells.** vol. 64. pp. 3892-3899; 2004:3892-3899.

46. Christofk HR, Vander Heiden MG, Harris MH, Ramanathan A, Gerszten RE, Wei R, Fleming MD, Schreiber SL, Cantley LC: **The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth.** *Nature* 2008, **452:**230-233.

47. Lewis GD, Asnani A, Gerszten RE: **Application of Metabolomics to Cardiovascular Biomarker and Pathway Discovery.** vol. 52. pp. 117-123; 2008:117-123.

48. Broadhurst DI, Kell DB: *Metabolomics* 2006, **2:**171.

49. Wishart DS, Knox C, Guo AC, Eisner R, Young N, Gautam B, Hau DD, Psychogios N, Dong E, Bouatra S, et al: **HMDB: a knowledgebase for the human metabolome.** vol. 37. pp. D603-D610; 2009:D603-D610.

50. Fiehn O: *TrAC, Trends Anal Chem* 2008, **27:**261.

51. Lisec J, Schauer N, Kopka J, Willmitzer L, Fernie AR: *Nat Protoc* 2006, **1:**387.

52. Bruce SJ, Jonsson P, Antti H, Cloarec O, Trygg J, Marklund SL, Moritz T: *Anal Biochem* 2008, **372:**237.

53. De Vos RCH, Moco S, Lommen A, Keurentjes JJB, Bino RJ, Hall RD: *Nat Protoc* 2007, **2:**778.

54. http://genome.wellcome.ac.uk.

55. Derveaux S, Vandesompele J, Hellemans J: **How to do successful gene expression analysis using real-time PCR.** *Methods*, **50:**227-230.

56. Higuchi R, Dollinger G, Walsh PS, Griffith R: **Simultaneous Amplification and Detection of Specific DNA Sequences.** *Nat Biotech* 1992, **10:**413-417.

57. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray.** vol. 270. pp. 467-470; 1995:467-470.

58. http://www.affymetrix.com.

59. http://www.ncbi.nlm.nih.gov.

60. Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Mark P, StÃ¼mpflen V, Mewes H-W, et al: **The MIPS mammalian proteinâ€“protein interaction database.** vol. 21. pp. 832-834; 2005:832-834.

61. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al: **Gene Ontology: tool for the unification of biology.** *Nat Genet* 2000, **25:**25-29.

62. Shannon P, Reiss D, Bonneau R, Baliga N: **The Gaggle: An open-source software system for integrating bioinformatics software and data sources.** vol. 7. pp. 176; 2006:176.

63. http://www.genome.jp/kegg.

64. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks.** *Genome Res* 2003, **13:**2498-2504.

65. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, et al: **The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored.** vol. 39. pp. D561-D568:D561-D568.

66. Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, Carver T, Glover K, Pocock MR, Wipat A, Li P: **Taverna: a tool for the composition and enactment of bioinformatics workflows.** vol. 20. pp. 3045-3054; 2004:3045-3054.

67. www.ingenuity.com.

68. Draghici S, Khatri P, Tarca AL, Amin K, Done A, Voichita C, Georgescu C, Romero R: **A systems biology approach for pathway level analysis.** vol. 17. pp. 1537-1545; 2007:1537-1545.

69. Galperin MY, Koonin EV: **Who's your neighbor? New computational approaches for functional genomics.** *Nat Biotech* 2000, **18:**609-613.

70. Marcotte EM: **Computational genetics: finding protein function by nonhomology methods.** *Current Opinion in Structural Biology* 2000, **10:**359-365.

71. Xia J, Wishart DS: **MetPA: a web-based metabolomics tool for pathway analysis and visualization.** vol. 26. pp. 2342-2344:2342-2344.

72. Benjamini Y, Yekutieli D: **The Control of the False Discovery Rate in Multiple Testing under Dependency.** *The Annals of Statistics* 2001, **29:**1165-1188.

73. Barrow GM: **Introduction to Molecular Spectroscopy.** *McGraw-Hill International Book Company* 1981.

74. Haseth PRGaJAd: **Fourier-Transform Infrared Spectroscopy.** *Wiley-Interscience, New York, Chichester, Brisbane, Toronto, Singapore* 1986.

75. Jonathan Clayden SW, Nick Greeves, Peter Wothers: **Organic Chemistry.** *Oxford University Press* July 2000.

76. Shah V: **Handbook of Plastics Testing and Failure Analysis** *Wiley-Interscience* 2007.

77. D. C. Harris MDB: **Symmetry and Spectroscopy: An Introduction to Vibrational and Electronic Spectroscopy.** In. New York: Dover Publications, INC

78. Hsu CPS: **Infrared Spectroscopy.** In *Handbook of Instrumental Techniques for Analytical Chemistry.* 247-283

79. Fifield FWaK, D: *Principles and Practice of Analytical Chemistry.* 5th edition edn. Oxford, UK: Blackwell.; 2000.

80. Fischer G, Braun S, Thissen R, Dott W: **FT-IR spectroscopy as a tool for rapid identification and intra-species characterization of airborne filamentous fungi.** *Journal of Microbiological Methods* 2006, **64:**63-77.

81. Birkemeyer C, Luedemann A, Wagner C, Erban A, Kopka J: **Metabolome analysis: the potential of in vivo labeling with stable isotopes for metabolite profiling.** *Trends in Biotechnology* 2005, **23:**28-33.

82. **NIST Mass Spectral Database 08.** 2008.

83. Knapp DR: *Handbook of analytical derivatization reactions.* New york: Wiley; 1979.

84. http://www.registech.com/Library/gcderrev.pdf.

85. Leming S, Weiming H, Zhenqiang S, Xianping L, Weida T, Dilip KAaGGK: **Microarrays: Technologies and applications.** In *Applied Mycology and Biotechnology. Volume* Volume 3: Elsevier; 2003: 271-293

86. M. Abhilash HKMRP: **Microarray Technology** *The Internet Journal of Medical Informatics* 2009, **4**.

87. Brown SD, Sum ST, Despagne F, Lavine BK: **Chemometrics.** *Analytical Chemistry* 1996, **68:**21-62.

88. Brown SD: **Chemometrics.** *Analytical Chemistry* 1990, **62:**84R-101R.

89. R.O. Duda PEH, D.G. Stork.: *Pattern Classification* 2nd Edition edn: John Wiley & Sons Inc; 2001.

90. Wold H: **Estimation of principal components and related models by iterative least squares.** *In K R Krishnaiah (ed), Multivariate analysis Academic Press, Inc, New York, NY* 1966**:**p. 391–420.

91. Wold S, Esbensen K, Geladi P: **Principal component analysis.** *Chemometrics and Intelligent Laboratory Systems* 1987, **2:**37-52.

92. B.S. Everitt SL, M. Leese: *Cluster Analysis* 4th Edition edn: Arnold; 2001.

93. Fisher RA, Sir: **The Use of Multiple Measurements in Taxonomic Problems.** *Annals of Eugenics,* 1936, **7:**179-188.

94. Wold H: *In Multivariate Analysis,.* Academic Press; 1966.

95. Valiant LG: *In Proceedings of the American Association for Artifical Intelligence* 1988.

96. Manly BF: **Multivariate statistical methods: a primer.** *J Chapman & Hall, London, United Kingdom* 1994.

97. Massart DL: *Chemometrics: A Textbook* New York: Elsevier Sciences Ltd; 1988.

98. Sinha A, Prazen B, Synovec R: **Trends in chemometric analysis of comprehensive two-dimensional separations.** *Analytical and Bioanalytical Chemistry* 2004, **378:**1948-1951.

99. Kizil R, Irudayaraj J, Seetharaman K: **Characterization of Irradiated Starches by Using FT-Raman and FTIR Spectroscopy.** *Journal of Agricultural and Food Chemistry* 2002, **50:**3912-3918.

100. Yang H, Irudayaraj J, Paradkar MM: **Discriminant analysis of edible oils and fats by FTIR, FT-NIR and FT-Raman spectroscopy.** *Food Chemistry* 2005, **93:**25-32.

101. Johnson HE, Broadhurst D, Goodacre R, Smith AR: **Metabolic fingerprinting of salt-stressed tomatoes.** *Phytochemistry* 2003, **62:**919-928.

102. William Allwood J, Clarke A, Goodacre R, Mur LAJ: **Dual metabolomics: A novel approach to understanding plant-pathogen interactions.** *Phytochemistry*, **71:**590-597.

103. Harrington PdB, Vieira NE, Espinoza J, Nien JK, Romero R, Yergey AL: **Analysis of variance-principal component analysis: A soft tool for proteomic discovery.** *Analytica Chimica Acta* 2005, **544:**118-127.

104. Irizarry RA, Hobbs B, Collin F, Beazerâ€ • Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** vol. 4. pp. 249-264; 2003:249-264.

105. Rousseeuw LKaPJ: **Finding groups in data: an introduction to cluster analysis** *John Wiley & Sons* 1990.

106. MacQueen J: **Some methods for classification and analysis of multivariate observations.** In *Proc Fifth Berkeley Symp on Math Statist and Prob*. Univ. of Calif. Press; 1967 281-297.

107. Storey JD, Tibshirani R: **Statistical significance for genomewide studies.** vol. 100. pp. 9440-9445; 2003:9440-9445.

108. Allocco D, Kohane I, Butte A: **Quantifying the relationship between co-expression, co-regulation and gene function.** vol. 5. pp. 18; 2004:18.

109. Arkin A, Ross J: **Statistical Construction of Chemical Reaction Mechanisms from Measured Time-Series.** *The Journal of Physical Chemistry* 1995, **99:**970-979.

110. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A: **Reverse engineering of regulatory networks in human B cells.** *Nat Genet* 2005, **37:**382-390.

111. Day A, Carlson M, Dong J, O'Connor B, Nelson S: **Celsius: a community resource for Affymetrix microarray data.** vol. 8. pp. R112; 2007:R112.

112. Hibbs MA, Hess DC, Myers CL, Huttenhower C, Li K, Troyanskaya OG: **Exploring the functional landscape of gene expression: directed search of large microarray compendia.** vol. 23. pp. 2692-2699; 2007:2692-2699.

113. Friedman N: **Inferring Cellular Networks Using Probabilistic Graphical Models.** vol. 303. pp. 799-805; 2004:799-805.

114. Friedman N, Linial M, Nachman I, Pe'er D: **Using Bayesian Networks to Analyze Expression Data.** vol. 7. pp. 601-620; 2000:601-620.

115. Lee S-I, Pe'er D, Dudley AeM, Church GM, Koller D: **Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification.** vol. 103. pp. 14062-14067; 2006:14062-14067.

116. Li Z, Chan C: **Inferring pathways and networks with a Bayesian framework.** vol. 18. pp. 746-748; 2004:746-748.

117. Peâ€™er D, Regev A, Elidan G, Friedman N: **Inferring subnetworks from perturbed expression profiles.** vol. 17. pp. S215-S224; 2001:S215-S224.

118. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP: **Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data.** vol. 308. pp. 523-529; 2005:523-529.

119. Pearson R, Liu X, Sanguinetti G, Milo M, Lawrence N, Rattray M: **puma: a Bioconductor package for propagating uncertainty in microarray analysis.** vol. 10. pp. 211; 2009:211.

120. Liu X, Milo M, Lawrence ND, Rattray M: **A tractable probabilistic model for Affymetrix probe-level analysis across multiple chips.** vol. 21. pp. 3637 - 3644; 2005:3637 - 3644.

121. Gentleman R, Carey V, Bates D, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al: **Bioconductor: open software development for computational biology and bioinformatics.** vol. 5. pp. R80; 2004:R80.

122. Sahoo D, Dill D, Gentles A, Tibshirani R, Plevritis S: **Boolean implication networks derived from large scale, whole genome microarray datasets.** vol. 9. pp. R157; 2008:R157.

123. Dudoit S, Shaffer JP, Boldrick JC: **Multiple Hypothesis Testing in Microarray Experiments.** *Statistical Science* 2003, **18:**71-103.

124. http://mathworld.wolfram.com/BonferroniCorrection.html.

125. Savitzky A, Golay MJE: **Smoothing and Differentiation of Data by Simplified Least Squares Procedures.** vol. 36. pp. 1627-1639; 1964:1627-1639.

126. Tom F: **Extended multiplicative scatter correction.** vol. 16. pp. 3-5; 2005:3-5.

127. Jarvis RM, Broadhurst D, Johnson H, O'Boyle NM, Goodacre R: **PYCHEM: a multivariate analysis package for python.** vol. 22. pp. 2565-2566; 2006:2565-2566.

128. http://www.lenntech.com/periodic/water/oxygen/oxygen-and-water.htm.

129. http://en.wikipedia.org/wiki/Torr.

130. http://antoine.frostburg.edu/chem/senese/101/solutions/faq/predicting-DO.shtml.

131. Atkins PW: **Physical Chemistry 1998.** In. 6th edition: Oxford University Press; 1998: 174

132. Logan J: **Investigating Cisplatin Resistance in Neuroblastoma Cell Lines.** *MRes dissertation*.

133. Begley P, Francis-McIntyre S, Dunn WB, Broadhurst DI, Halsall A, Tseng A, Knowles J, Goodacre R, Kell DB: **Development and Performance of a Gas Chromatographyâˆ'Time-of-Flight Mass Spectrometry Analysis for Large-Scale Nontargeted Metabolomic Studies of Human Serum.** *Analytical Chemistry* 2009, **81:**7038-7046.

134. Brown M, Dunn WB, Dobson P, Patel Y, Winder CL, Francis-McIntyre S, Begley P, Carroll K, Broadhurst D, Tseng A, et al: **Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics.** *Analyst* 2009, **134:**1322-1332.

135. http://www.affymetrix.com/support/technical/manuals.affx.

136. Liu X, Milo M, Lawrence ND, Rattray M: **A tractable probabilistic model for Affymetrix probe-level analysis across multiple chips.** vol. 21. pp. 3637-3644:3637-3644.

137. Dennis G, Sherman B, Hosack D, Yang J, Gao W, Lane HC, Lempicki R: **DAVID: Database for Annotation, Visualization, and Integrated Discovery.** vol. 4. pp. P3; 2003:P3.

138. Hosack D, Dennis G, Sherman B, Lane H, Lempicki R: **Identifying biological themes within lists of genes with EASE.** vol. 4. pp. R70; 2003:R70.

139. Aittokallio T, Schwikowski B: **Graph-based methods for analysing networks in cell biology.** vol. 7. pp. 243-255; 2006:243-255.

140. Broadhurst DI, Kell DB: **Statistical strategies for avoiding false discoveries in metabolomics and related experiments.** *Metabolomics* 2006, **2:**171-196.

141. Yoshiaki Shiba TOMS: **Growth and morphology of anchorage-dependent animal cells in a liquid/liquid interface system.** vol. 57. pp. 583-589; 1998:583-589.

142. Kasimir MT RE, Seebacher G, Silberhumer G, Wolner E, Weigel G, Simon P: **Comparison of different decellularization procedures of porcine heart valves.** *Int J Artif Organs* 2003 **26:**421-427.

143. Wilcoxon F: **Individual Comparisons by Ranking Methods.** *Biometrics Bulletin* 1945, **1:**80-83.

144. *Ornithine Biosynthesis,.* School of Biological and Chemical Sciences, Queen Mary, University of London, ; 2007.

145. Sellick CA, Hansen R, Maqsood AR, Dunn WB, Stephens GM, Goodacre R, Dickson AJ: **Effective Quenching Processes for Physiologically Valid Metabolite Profiling of Suspension Cultured Mammalian Cells.** *Analytical Chemistry* 2008, **81:**174-183.

146. Yuan J, Bennett BD, Rabinowitz JD: **Kinetic flux profiling for quantitation of cellular metabolic fluxes.** *Nat Protocols* 2008, **3:**1328-1340.

147. Christen Y: **Oxidative stress and Alzheimer disease1.** vol. 71. pp. 621s-629s; 2000:621s-629s.

148. Beatty S, Koh H-H, Phil M, Henson D, Boulton M: **The Role of Oxidative Stress in the Pathogenesis of Age-Related Macular Degeneration.** *Survey of Ophthalmology*, **45:**115-134.

149. Sykiotis GP, Bohmann D: **Stress-Activated Cap'n'collar Transcription Factors in Aging and Human Disease.** vol. 3. pp. re3-:re3-.

150. Halliwell B: **Oxidants and human disease: some new concepts.** vol. 1. pp. 358-364; 1987:358-364.

151. Xie L, Pandey R, Xu B, Tsaprailis G, Chen Q: **Genomic and proteomic profiling of oxidative stress response in human diploid fibroblasts.** *Biogerontology* 2009, **10:**125-151.

152. Cottingham K: **HUSERMET researchers look to the metabolome for answers.** *Journal of Proteome Research* 2008, **7:**4213-4213.

153. M. Turk AP: **Eigenfaces for Recognition** *Journal of Cognitive Neurosicence* 1991, **3:**71-86.

154. Kroonenberg PM, de Leeuw J: **Principal component analysis of three-mode data by means of alternating least squares algorithms.** *Psychometrika* 1980, **45:**69-97.

155. Shih B, Brown J, Armstrong D, Lindau T, Bayat A: **Differential Gene Expression Analysis of Subcutaneous Fat, Fascia, and Skin Overlying a Dupuytren's Disease Nodule in Comparison to Control Tissue.** *Hand* 2009, **4:**294-301.

156. Raggo C, Ruhl R, McAllister S, Koon H, Dezube BJ, FrÃ¼h K, Moses AV: **Novel Cellular Genes Essential for Transformation of Endothelial Cells by Kaposi's Sarcomaâ€"Associated Herpesvirus.** vol. 65. pp. 5084-5095; 2005:5084-5095.

157. Alberts B BD, Hopin K, Johnson A, Lewis J, Raff M, Roberts K, Walter P *Tissues and Cancer Essential cell biology* New York and London Garland Science; 2004.

158. Spring FA, Dalchau R, Daniels GL, Mallinson G, Judson PA, Parsons SF, Fabre JW, Anstee DJ: **The Ina and Inb blood group antigens are located on a glycoprotein of 80,000 MW (the CDw44 glycoprotein) whose expression is influenced by the In(Lu) gene.** *Immunology* 1988, **64:**37-43.

159. http://biowww.net/gene/gene-COL5A1.html.

160. Entrez Gene: COL5A3 collagen tV, alpha 3.

161. http://ghr.nlm.nih.gov/gene/MAPK10.

162. Germain S, Monnot C, Muller L, Eichmann A: **Hypoxia-driven angiogenesis: role of tip cells and extracellular matrix scaffolding.** vol. 17. pp. 245-251 245-251

163. Gallagher DC, Bhatt RS, Parikh SM, Patel P, Seery V, McDermott DF, Atkins MB, Sukhatme VP: **Angiopoietin 2 Is a Potential Mediator of High-Dose Interleukin 2â€“Induced Vascular Leak.** vol. 13. pp. 2115-2120; 2007:2115-2120.

164. Sanguinetti G, Milo M, Rattray M, Lawrence ND: **Accounting for probe-level noise in principal component analysis of microarray data.** vol. 21. pp. 3748-3754:3748-3754.

165. Calvano SE, Xiao W, Richards DR, Felciano RM, Baker HV, Cho RJ, Chen RO, Brownstein BH, Cobb JP, Tschoeke SK, et al: **A network-based analysis of systemic inflammation in humans.** *Nature* 2005, **437:**1032-1037.

166. Cottret L, Wildridge D, Vinson F, Barrett MP, Charles H, Sagot M-F, Jourdan F: **MetExplore: a web server to link metabolomic experiments and genome-scale metabolic networks.** vol. 38. pp. W132-W137:W132-W137.

167. Tsuruoka Y, Tsujii Ji, Ananiadou S: **FACTA: a text search engine for finding associated biomedical concepts.** vol. 24. pp. 2559-2560; 2008:2559-2560.

168. http://www.mcisb.org.