

**UNDERSTANDING SKELETAL MUSCLE  
ADAPTATION IN HEALTH AND CHRONIC  
DISEASE: A MULTI-OMICS BASED SYSTEMS  
BIOLOGY PERSPECTIVE**

by

**PETER KÅRE DAVIDSEN**

A thesis submitted to  
The University of Birmingham  
for the degree of  
DOCTOR OF PHILOSOPHY

Supervisor: Prof. Francesco Falciani

School of Immunity and Infection  
University of Birmingham  
October 2015

## **General Abstract**

Mammalian skeletal muscle has a major impact on whole-body metabolic homeostasis. Hence, maintenance of a metabolically active muscle mass is key for optimal health. Notably, both muscle function and mass are profoundly negatively affected by environmental factors such as chronic smoking and physical inactivity.

RNA abundance integrates genetic, epigenetic and environmental influences. Therefore, while true understanding of physiological adaptation likely require the integration between multi-level datasets, the transcriptome represents a powerful investigative tool in determining the underlying molecular mechanisms behind complex phenotypic traits.

The overarching aim of this thesis was to evaluate, using omics-based systems biology approaches, the global regulation of RNAs during exogenous modulation of mammalian muscle phenotype in order to characterize local homeostatic processes as well as identify robust biomarker signatures.

The **first part** of this thesis deals with smoke-induced peripheral muscle wasting. Initially, biological domain knowledge is used to validate a pre-clinical smoking model. Then, specific cytokines are statistically linked to limb muscle energy metabolism; a testable hypothesis supported by both animal and human data.

The **second part** deals with the development of ‘molecular predictors’ of endurance training adaptability. Two complex clinically relevant traits are considered, namely whole-body insulin sensitivity and plasma triglyceride content. Promisingly, quantitative multi-gene predictors of response to training for both traits of interest were developed.

## **Acknowledgements**

Many people have made contributions to work underlying this thesis. Especially, I would like to thank:

My supervisor, Prof Francesco Falciani, for taking me into his research group. Further, I am grateful for his ideas, guidance and support provided throughout my PhD time.

Also a big thanks you to Dr. Philip Antczak for his programming support and many ideas related to my projects throughout the PhD.

Furthermore, I am grateful to Drs. John Herbert and Kim Clarke for providing support when needed.

My PhD research was fortunate to involve a collaboration in the US. Hence, I wish to thank Prof. Claude Bouchard for providing access to the unique HERITAGE Family Study. Also, I would like to thank Dr. Mark Sarzynski for comments on the predictor papers.

Finally, I owe my special thank to two special girls: my patient girlfriend Louise for support during our stay in UK and my lovely daughter Freya for always putting a smile on my face.

This work was supported by the MRC Arthritis Research UK Centre for Musculoskeletal Ageing Research.

## List of publications during my PhD

- [1] Sarzynski MA, Daividsen PK, Sung YJ, Hesselink MKC, Schrauwen P, Rice TK, Rao DC, Falciani F, Rankinen T & Bouchard C. **Genomic and transcriptomic predictors of triglyceride response to regular exercise.** In Press, *Br J Sports Med.* 2015.
- [2] Daividsen PK, Herbert JM, Antczak P, Clarke K, Ferrer E, Peinado VI, Gonzalez C, Roca J, Egginton S, Barbera JA, & Falciani F. **A systems biology approach reveals a link between systemic cytokines and skeletal muscle energy metabolism in a rodent smoking model and human COPD.** *Genome Med.* 2014. 6:59
- [3] Brina D, Miluzio A, Ricciardi S, Clarke K, Daividsen PK, Viero G, Tebaldi T, Offenhauser N, Rozmann J, Rathkolb B, Neschen S, Klingenspor M, Wolf E, Gailus-Durner V, Fuchs H, de Angelis MH, Quattrone A, Falciani F & Biffo S. **eIF6 coordinates insulin sensitivity and lipid metabolism by coupling translation to transcription.** *Nat Commun.* 2015, 6:8261
- [4] Daividsen PK, Sarzynski MA, Herbert JM, Antczak P, Hesselink MKC, Schrauwen P, Rice TK, Weisnagel SJ, Bergman RN, Rao DC, Bouchard C & Falciani F. **A molecular signature linked to calcium signaling is predictive of training-induced changes in insulin sensitivity.** Manuscript in preparation (see Chapter 4).
- [5] Timmons JA, Baar K, Daividsen PK & Atherton PJ. **Is irisin a human exercise gene?** *Nature.* 2012. 488:E9-10
- [6] Daividsen PK, Turan N, Egginton S & Falciani F. **Multi-level functional genomics data integration as a tool for understanding physiology: A network perspective.** In press, *J Appl Physiol.*

## Table of Contents

<b>General Abstract</b> .....	<b>i</b>
<b>Acknowledgements</b> .....	<b>ii</b>
<b>List of publications during my PhD</b> .....	<b>iii</b>
<b>List of Figures</b> .....	<b>vii</b>
<b>List of Tables</b> .....	<b>xi</b>
<b>Abbreviations</b> .....	<b>xii</b>
<b>1 General Introduction</b> .....	<b>1</b>
<b>1.1 Mammalian skeletal muscle tissue</b> .....	<b>1</b>
1.1.1 Skeletal muscle fibre types and properties .....	1
1.1.2 Skeletal muscle adaptation.....	4
1.1.3 Interindividual variability in skeletal muscle adaptability .....	6
<b>1.2 Application of systems biology approaches to human health</b> .....	<b>8</b>
1.2.1 High Density Data Acquisition .....	9
1.2.2 Microarray data analysis .....	17
<b>1.3 Aims and outline of the thesis</b> .....	<b>28</b>
<b>2 Multi-level functional genomics data integration as a tool for understanding physiology: A network biology perspective</b> .....	<b>30</b>
<b>2.1 ABSTRACT</b> .....	<b>31</b>
<b>2.2 INTRODUCTION</b> .....	<b>32</b>
2.2.1 Modelling in physiological sciences.....	32
2.2.2 The advent of Functional Genomics: a challenge for physiological modelling ..	33
2.2.3 Towards data-driven predictive biology.....	35
2.2.4 COMPUTATIONAL APPROACHES FOR THE ANALYSIS OF COMPLEX DATASETS	37
2.2.5 Inference of biological networks from observational data .....	42
2.2.6 Static vs. dynamic networks.....	42
2.2.7 A primer for network inference methods.....	43
<b>2.3 CASE STUDY: INFERENCE OF OXYGEN-DEPENDENT PATHWAYS IN LIMB SKELETAL MUSCLE</b> .....	<b>49</b>
2.3.1 Step 1. Linking physiological measurements and gene expression data in the COPD cohort.....	52
2.3.2 Step 2. Gene expression dynamics in response to tissue hypoxia.....	58
<b>2.4 CONCLUSIONS</b> .....	<b>65</b>
<b>3 A systems biology approach reveals a link between systemic cytokines and skeletal muscle energy metabolism in a rodent smoking model and human COPD</b> .....	<b>67</b>
<b>3.1 Abstract</b> .....	<b>67</b>
<b>3.2 Background</b> .....	<b>69</b>
<b>3.3 Methods</b> .....	<b>72</b>
3.3.1 Guinea pig smoking model .....	72
3.3.2 RNA isolation from guinea pig samples .....	73
3.3.3 Definition of the guinea pig transcriptome by mRNA sequencing and microarray design.....	73

3.3.4	Guinea pig microarray gene expression profiling .....	74
3.3.5	RT-qPCR validation of custom GP array .....	75
3.3.6	Human COPD clinical studies .....	76
3.3.7	Summarizing the molecular state of skeletal muscle using indices of pathway transcriptional activity .....	79
3.3.8	Inference of biological networks linking lung and skeletal muscles in guinea pigs .....	80
3.3.9	Creating and visualizing a KEGG pathway map .....	82
3.3.10	Measurement of inflammatory mediators in human serum from COPD patients and healthy controls and validation of the guinea pig lung-muscle cross-talk network .....	82
<b>3.4</b>	<b>Results .....</b>	<b>83</b>
3.4.1	Sequencing of the Guinea pig transcriptome and development of a genome-wide guinea pig microarray platform .....	83
3.4.2	Chronic exposure to smoking and/or hypoxia induces transcriptional changes in both lung and skeletal muscle in guinea pigs .....	89
3.4.3	Lung response to the different stressors is comparable in magnitude, but involves different subsets of functional pathways .....	91
3.4.4	Glycolytic and oxidative limb muscles respond differently to either smoking or hypoxia .....	93
3.4.5	The guinea pig smoking model recapitulates the transcriptional changes observed in human COPD skeletal muscles .....	96
3.4.6	Gene expression of lung soluble inflammatory mediators correlate with skeletal muscle gene expression .....	101
3.4.7	Serum cytokine profiling in human COPD patients confirms the predictions of the guinea pig model .....	106
3.4.8	Training did not modulate any of the tested cytokines .....	107
<b>3.5</b>	<b>Discussion .....</b>	<b>112</b>
3.5.1	The role of systemic inflammatory cytokines in controlling the molecular state of skeletal muscles .....	112
3.5.2	Biological significance of the transcriptional response to smoking and/or hypoxia in lungs and muscles .....	114
3.5.3	Gene expression profiling as a tool to assess animal model relevance of human disease .....	116
<b>4</b>	<b>Using an integrative framework combining association data with skeletal muscle gene expression predicts endurance training-induced changes in (whole-body) insulin sensitivity: The HERITAGE Family Study .....</b>	<b>118</b>
<b>4.1</b>	<b>Introduction .....</b>	<b>119</b>
<b>4.2</b>	<b>Research Design and Methods .....</b>	<b>121</b>
4.2.1	HERITAGE Family Study .....	121
4.2.2	GWAS analysis: .....	122
4.2.3	RNA extraction & global gene expression profiling: .....	123
4.2.4	Microarray analysis: .....	123
4.2.5	Pathway-level analyses: .....	124
4.2.6	Data analysis: .....	125
<b>4.3</b>	<b>Results .....</b>	<b>127</b>
4.3.1	Overview of the analysis strategy: .....	127
4.3.2	Improvement potential in SI has a clear heritable component: .....	129

4.3.3	Functional GWAS identifies an association between $\Delta$ SI, carbohydrate metabolism and calcium signalling pathways:.....	130
4.3.4	The calcium-dependent MEF2A transcription factor drives the transcriptional signature associated to $\Delta$ SI.....	136
4.3.5	Basal mRNA abundance of MEF2A interacting targets is predictive of $\Delta$ SI:.....	138
<b>4.4</b>	<b>Discussion.....</b>	<b>142</b>
4.4.1	Predicting changes in insulin sensitivity from gene expression profiling.....	142
4.4.2	Is there a relationship between changes in SI and myofiber interconvertibility? 143	
4.4.3	Conclusion .....	145
<b>5</b>	<b>Combining genomic and transcriptomic predictors of plasma triglyceride response to exercise training: a genome-wide approach.....</b>	<b>146</b>
<b>5.1</b>	<b>Introduction.....</b>	<b>147</b>
<b>5.2</b>	<b>Methods.....</b>	<b>150</b>
5.2.1	Determination of plasma lipids .....	150
5.2.2	GWAS SNP genotyping .....	150
5.2.3	GWAS statistical analyses .....	150
5.2.4	Affymetrix microarray analysis.....	150
5.2.5	Baseline RNA gene signature.....	150
5.2.6	SNP summary score.....	153
<b>5.3</b>	<b>Results .....</b>	<b>154</b>
5.3.1	GWAS associations for TG response to exercise training.....	155
5.3.2	RNA expression-based gene signature of TG response to exercise training ....	156
5.3.3	Association of SNP score and $\Delta$ TG.....	160
<b>5.4</b>	<b>Discussion.....</b>	<b>162</b>
5.4.1	Study limitations.....	164
<b>6</b>	<b>Concluding Remarks and Research Perspective. ....</b>	<b>166</b>
<b>7</b>	<b>Appendices .....</b>	<b>157</b>
<b>7.1</b>	<b>RNA sequencing analysis and the development of a guinea pig microarray platform .....</b>	<b>157</b>
<b>7.2</b>	<b>Cytokine superfamily as defined by commercial PCR array provider .....</b>	<b>167</b>
<b>7.3</b>	<b>ANALYSIS OF SKELETAL MUSCLE TRANSCRIPTIONAL RESPONSE IN A MOUSE SMOKING MODEL .....</b>	<b>168</b>
<b>7.4</b>	<b>Pre-ranked GSEA using transcription factor targets .....</b>	<b>172</b>
<b>7.5</b>	<b>MEF2A interacting targets as defined in the STRING database (confidence score 0.8 or more). ....</b>	<b>174</b>
<b>7.6</b>	<b>List of GWAS SNPs associated with serum triglyceride response to exercise training among Caucasians in the HERITAGE study.....</b>	<b>175</b>
<b>7.7</b>	<b>Result of pathway-based analysis of GWAS associations .....</b>	<b>176</b>
<b>8</b>	<b>Bibliography .....</b>	<b>178</b>

## List of Figures

FIGURE 1-1   OVERVIEW OF THE AFFYMETRIX SNP CHIP TECHNOLOGY.....	11
FIGURE 1-2   PROBE DESIGN ON THE OLDER 3'-BASED AFFYMETRIX MICROARRAYS.....	13
FIGURE 1-3   SCHEMATIC OVERVIEW OF THE DIFFERENCE IN AFFYMETRIX PROBESET DESIGN BETWEEN THEIR NEWER EXON ARRAYS AND THE OLDER 3'-BASED ARRAYS.....	14
FIGURE 1-4   INDIVIDUAL WEB-LAB STEPS OF A MICROARRAY EXPERIMENT.....	16
FIGURE 1-5   HISTOGRAM SHOWING THE DISTRIBUTION OF SIGNAL INTENSITIES BEFORE (LEFT PANEL) AND AFTER (RIGHT PANEL) LOG <sub>2</sub> DATA TRANSFORMATION.....	18
FIGURE 1-6   DETAILS OF THE QUANTILE NORMALIZATION PROCEDURE FOR EXPRESSION ARRAYS.....	20
FIGURE 1-7   FLOWCHART HIGHLIGHTING POTENTIAL BATCH EFFECT SOURCES (GREY BOXES) DURING DATA GENERATION.....	25
FIGURE 1-8   SCHEMATIC OVERVIEW OF THE GENE SET ENRICHMENT ANALYSIS PROCEDURE.....	27
FIGURE 2-1   SCHEMATIC REPRESENTATION OF THE PROCESS INVOLVED IN MODELLING A BIOLOGICAL SYSTEM BY INTEGRATING KNOWLEDGE FROM VARIOUS SOURCES, AND COMPLEX MULTI-LEVEL DATASETS. .	38
FIGURE 2-2   SCHEMATIC REPRESENTATION OF THE ANALYSIS STRATEGY USED IN THE CASE STUDY, HIGHLIGHTING HOW THE INFERRED STATIC MULTI-SCALE NETWORK FROM THE CLINICAL COPD COHORT (FIG. 2- 2A-C) CAN BE BRIDGED TO THE INFERENCE OF A DYNAMICAL NETWORK REPRESENTING THE TEMPORAL PROGRESSION OF EVENTS FOLLOWING AN EXPERIMENTAL CHALLENGE (HYPOXIC EXPOSURE) IN A MURINE ANIMAL MODEL (FIG 2-2D-G).....	51
FIGURE 2-3   GRAPHICAL REPRESENTATION HIGHLIGHTING PUTATIVE REGULATORY ASSOCIATIONS (SIGNIFICANT CORRELATION BETWEEN TWO FACTORS IS SHOWN AS A DOTTED LINE) THAT LIKELY REPRESENT ROBUST INTERACTIONS, BASED ON HIGH MUTUAL INFORMATION VALUES. ....	56
FIGURE 2-4   HIGH-LEVEL REPRESENTATION OF TEMPORAL TRANSCRIPTIONAL CHANGES IN THE MURINE MODEL OF HYPOXIA. ..	60
FIGURE 2-5   THE HIERARCHICAL DYNAMIC STATE-SPACE MODEL IDENTIFIED 4 MODULES (X-AXES DEFINE LENGTH OF HYPOXIC EXPOSURE), EACH CHARACTERISED BY TWO SEPARATE TRANSCRIPTIONAL PROFILES: PLUS AND MINUS, REPRESENTING UP- AND DOWN-REGULATION, RESPECTIVELY.....	62
FIGURE 2-6   A HIGHER RESOLUTION REPRESENTATION OF FIGURE 2-5, HIGHLIGHTING THE MOST SIGNIFICANT GENE INTERACTIONS BETWEEN COMPONENTS IN THE FOUR INFERRED MODULES. ....	64

FIGURE 3-1   BARPLOT HIGHLIGHTING GROUP MEAN DIFFERENCES IN MASS INDEXES (WHOLE-BODY AND FAT-FREE, RESPECTIVELY) AND 6-MIN WALKING DISTANCE. ....	77
FIGURE 3-2   PCA PLOT SHOWING THAT SUBJECTS PROFILED IN THE GSE19407 DATASET (N=127) GROUP BY THE YEAR THEY WERE SCANNED (30,652 PROBESETS INCLUDED). ....	79
FIGURE 3-3   BARPLOT SHOWING LEVEL 1 GENE ONTOLOGY (GO) TERMS OF THE BIOLOGICAL PROCESS CATEGORY. ....	84
FIGURE 3-4   BARPLOT SHOWING LEVEL 1 GENE ONTOLOGY (GO) TERMS OF THE CELLULAR COMPONENT (CC) AND MOLECULAR FUNCTION (MF) CATEGORY, RESPECTIVELY. ....	85
FIGURE 3-5   SCATTERPLOT HIGHLIGHTING THE ASSOCIATION BETWEEN SIGNAL INTENSITY ON THE CUSTOM GUINEA PIG MICROARRAY AND NORMALIZED RNA-SEQ COUNTS USING POOLED RNA FROM SKELETAL MUSCLES. ....	87
FIGURE 3-6   COMPARISON OF ESTIMATED LOG <sub>2</sub> FOLD CHANGES (LUNG/MUSCLE) FROM ILLUMINA RNA-SEQ (Y-AXIS) AND THE CUSTOM MICROARRAY PLATFORM (X-AXIS). ....	88
FIGURE 3-7   PLOT INDICATING THE FRACTIONAL OVERLAP (Y-AXIS) BETWEEN PROFILING TECHNOLOGIES IN TERMS OF GENES BEING CALLED <u>MUSCLE-SPECIFIC</u> BASED ON THE RNA-SEQ DATA (X-AXIS). .	89
FIGURE 3-8   GUINEA PIG WHOLE-BODY WEIGHT GAIN DURING THE EXPERIMENTAL PROTOCOL. ....	90
FIGURE 3-9   DIFFERENTIALLY EXPRESSED GENES IN THE GUINEA PIG EXPERIMENTAL MODEL. ....	91
FIGURE 3-10   OVERLAP ANALYSIS OF ENRICHED KEGG PATHWAYS IN GUINEA PIG LUNG TISSUE. ....	93
FIGURE 3-11   OVERLAP ANALYSIS OF ENRICHED KEGG PATHWAYS IN TWO GUINEA PIG HINDLIMB MUSCLES WITH DISCRETE METABOLIC PROFILES. ....	95
FIGURE 3-12   MUSCLE SPECIFIC PATHWAY-LEVEL COMPARISONS BETWEEN THE GUINEA PIG MODEL AND A CLINICAL COPD STUDY (GSE27536). ....	98
FIGURE 3-13   SPECIFIC PATHWAYS REGULATED IN THE GP MODEL (COLUMN 3–5) WHEN CONTRASTED AGAINST COPD PATIENTS (COLUMN 1) AND HEALTHY CHRONIC SMOKERS (COLUMN 2). ....	100
FIGURE 3-14   NETWORKS REPRESENTING THE TRANSCRIPTIONAL COUPLING BETWEEN LUNG AND SKELETAL MUSCLE IN THE GP SMOKING MODEL. ....	104
FIGURE 3-15   VALIDATION OF SELECTED MICROARRAY RESULTS BY REAL-TIME RT-PCR. ....	105
FIGURE 3-16   HEATMAP VISUALIZATION OF THE PROTEIN EXPRESSION OF 7 SERUM CYTOKINES AMONG HEALTHY CONTROLS AND COPD AT BASELINE AND POST TRAINING (TRAINED). ....	107
FIGURE 3-18   CORRELATION BETWEEN STANDARDISED CXCL9 SERUM PROTEIN LEVELS AND THE TRANSCRIPTIONAL ABUNDANCE OF ATP5J,	

A GENE ENCODING FOR A PROTEIN IN THE FIFTH PROTEIN COMPLEX OF OXIDATIVE PHOSPHORYLATION. ....	110
FIGURE 3-19   CORRELATION BETWEEN STANDARDIZED CXCL10 SERUM PROTEIN LEVELS AND THE TRANSCRIPTIONAL ABUNDANCE OF COX8A, A GENE ENCODING FOR A PROTEIN IN THE FORTH PROTEIN COMPLEX OF OXIDATIVE PHOSPHORYLATION. ....	111
FIGURE 3-20   SCATTERPLOTS HIGHLIGHTING THE CLINICAL ASSOCIATION (SPEARMAN CORRELATION) BETWEEN DISTANCE WALKED IN 6-MIN AND SERUM LEVELS OF CXCL9 (PANEL A) AND CXCL10 (PANEL B), RESPECTIVELY .....	112
FIGURE 4-1   FLOWCHART HIGHLIGHTING THE ANALYSIS STRATEGY. ....	128
FIGURE 4-2   TRAINING-INDUCED CHANGE IN INSULIN SENSITIVITY PLOTTED AGAINST CAUCASIAN FAMILY RANK ( <i>I.E.</i> FAMILIES RANKED BY FAMILY MEAN) IN THE HERITAGE STUDY.....	129
FIGURE 4-3   HISTOGRAM OF THE P-VALUE DISTRIBUTION FOR SNPS FLANKING GENES ANNOTATED TO THE ‘CARDIAC MUSCLE CONTRACTION’ KEGG PATHWAY. ....	131
FIGURE 4-4 (SEE PREVIOUS PAGE)   PANEL A: HISTOGRAM SHOWING THE OBSERVED NUMBER OF SNPS ASSOCIATED WITH BASAL MRNA ABUNDANCE FOR EACH KEGG PATHWAY. THE RED LINE REPRESENTS 3 STANDARD DEVIATIONS FROM THE PERMUTATION-BASED NULL DISTRIBUTUION. SINCE KEGG GENE-SETS ARE NOT INDEPENDENT, THE OVERLAPS BETWEEN SIGNIFICANT ONES WERE ASSESSED USING THE JACCARD’S INDEX OF SIMILARITY. FROM PANEL B IT IS CLEAR THAT MANY OF THE ENRICHED GENE-SETS CAN BE GROUPED INTO THREE OVERALL FUNCTIONAL CATEGORIES: CARBOHYDRATE METABOLISM, CELL COMMUNICATION, AND CALCIUM SIGNALLING. IT IS OF NOTE THAT MULTIPLE TERMS WITHIN ‘CELL COMMUNICATION’ HAVE A STRONG CALCIUM SIGNALLING COMPONENT SUCH AS GAP JUNCTION, VASCULAR SMOOTH MUSCLE CONTRACTION AND WNT SIGNALLING .....	134
FIGURE 4-5   DIAGRAM REPRESENTING THE DETAILED RELATIONSHIPS BETWEEN THE CALCIUM-RELATED KEGG PATHWAYS IDENTIFIED IN TABLE 4-1.....	135
FIGURE 4-6   (A) USING SMALL INTERFERING RNA (SIRNA), WALES ET AL. [207] RECENTLY DEFINED THE GLOBAL TRANSCRIPTIONAL SIGNATURE ASSOCIATED WITH MEF2A MODULATION IN DIFFERENTIATING C2C12S. CORRESPONDING HUMAN ORTHOLOGS FOR THE DYSREGULATED GENES WERE IDENTIFIED USING THE MOUSE GENOME INFORMATICS (MGI) DATABASE (72% MAPPING SUCCESS)..	138
FIGURE 4-7   PLOT OF EXPERIMENTALLY DETERMINED (OBSERVED) VERSUS PREDICTED VALUES OF TRAINING-INDUCED CHANGE IN INSULIN SENSITIVITY (DERIVED FROM EUGLYCAEMIC HYPERINSULINAEMIC CLAMP) IN THE INDEPENDENT TRAINING COHORT.....	141

FIGURE 4-8   COLUMN SCATTER PLOT SHOWING TRAINING-INDUCED CHANGE IN TYPE I MYOFIBER CONTENT WITHIN THE TWO DEFINED RESPONDER GROUPS (CHANGE IN WHOLE-BODY INSULIN SENSITIVITY; $\Delta$ SI).....	144
FIGURE 5-1   HISTOGRAM HIGHLIGHTING THE INDIVIDUAL DIFFERENCES IN TRAINING-INDUCED CHANGE IN SERUM TRIGLYCERIDE LEVELS AMONG CAUCASIANS IN THE HERITAGE STUDY. ....	148
FIGURE 5-2   CHROMOSOME GOAL FITNESS DISTRIBUTION ACROSS ALL 100 SPLITS FOR A) ALL 4,000 PREDICTIVE MODELS, AND B) THE SUBSET (N=512) THAT PERFORMED WELL IN MOST SPLITS (I.E. LOW OVERALL DEVIATION AND HIGH ACCURACY). ....	152
FIGURE 5-3   PERFORMANCE OF THE RNA-BASED REGRESSION MODEL DERIVED FROM THE TRAINING SET (N=37, GRAY DOTS) IN THE TEST SET (N=12, RED DOTS) FOR THE PREDICTION OF EXERCISE TRAINING-INDUCED CHANGES IN SERUM TRIGLYCERIDES IN HERITAGE.....	157
FIGURE 5-4   ALL POSSIBLE SUBSET REGRESSION.....	158
FIGURE 5-5   RESULTS OF THE INGENUITY PATHWAY ANALYSIS (IPA) ON A SUBSET OF THE BASELINE GENE EXPRESSION PREDICTOR MODELS (N=512). HORIZONTAL BARS REPRESENT THE P-VALUE FOR THE TOP 12 CANONICAL PATHWAYS ENRICHED IN GENES WITHIN THE MOST PREDICTIVE GALGO MODELS (SEE FIGURE 5-2B) AND ARE EXPRESSED AS -1 TIMES THE LOG OF THE P-VALUE. THE BLACK VERTICAL LINE REPRESENTS THE PROBABILITY THRESHOLD AT P=0.05. THE ORANGE LINE REPRESENTS THE RATIO OF THE NUMBER OF GALGO MODEL GENES IN A PARTICULAR PATHWAY DIVIDED BY THE TOTAL NUMBER OF GENES THAT MAKE UP THAT PATHWAY.....	160
FIGURE 5-6   ADJUSTED MEAN $\Delta$ TG ACROSS EIGHT SNP SUMMARY SCORE CATEGORIES IN HERITAGE WHITES. ....	161

## List of Tables

TABLE 1-1   CHARACTERISTICS OF THE DIFFERENT HUMAN SKELETAL MUSCLE FIBRE TYPES. ....	3
TABLE 2-1   ANTHROPOMETRIC CHARACTERISTICS DEFINING THE COPD COHORT USED IN THE CASE STUDY.....	50
TABLE 3-1   BASELINE PHYSIOLOGICAL DATA OF THE COPD PATIENTS AND HEALTHY CONTROLS. ....	77
TABLE 3-2   LIST OF GENES THAT ARE ANNOTATED TO THE CYTOKINE SUPERFAMILY AND DIFFERENTIALLY EXPRESSED IN WHOLE-LUNG TISSUE OF TREATED GUINEA PIGS COMPARED TO UNTREATED CONTROLS (N=33; FDR<1%). ....	81
TABLE 4-1   KEGG PATHWAYS ENRICHED FOR DNA VARIANTS ASSOCIATED WITH $\Delta$ SI.....	130
TABLE 4-2   RESULT OF THE MULTIVARIATE REGRESSION MODEL FOR SI TRAINING RESPONSE IN HERITAGE (N=47).....	140
TABLE 5-1   MAJOR CLASSES OF LIPOPROTEINS IN HUMAN PLASMA.....	147
TABLE 5-2   DESCRIPTIVE DATA, INCLUDING PRE- AND POST-TRAINING VALUES FOR LIPID, LIPOPROTEIN AND LIPASE MEASUREMENTS, FOR HERITAGE WHITES WITH VALID GWAS (LEFT) AND GENE EXPRESSION (RIGHT) DATA, RESPECTIVELY. ....	154
TABLE 5-3   RESULTS OF THE RNA-BASED MULTIVARIATE REGRESSION MODEL WITH FORWARD SELECTION FOR TG RESPONSE TO EXERCISE TRAINING IN HERITAGE WHITES (N=37).....	156

## Abbreviations

ARACNE, Algorithm for the Reconstruction of Accurate Cellular NETWORKs;  
ATP, adenosine triphosphate;  
BMI, Body mass index;  
CaMK, Ca<sup>2+</sup>/calmodulin-dependent protein kinase;  
CEL, cell intensity file (Affymetrix);  
CH, chronic hypoxia;  
COPD, chronic obstructive pulmonary disease;  
CS, cigarette smoke;  
CSCH, Chronic smoking and hypoxia combined;  
CVD, cardiovascular disease;  
GO, Gene Ontology;  
GP, guinea pig;  
FA, fatty acid;  
FDR, false discovery rate;  
FFMI, fat free mass index;  
GEO, Gene Expression Omnibus;  
GSEA, Gene Set Enrichment Analysis;  
GWAS, genome-wide association studies  
KEGG, Kyoto Encyclopedia of Genes and Genomes;  
LPL, lipoprotein lipase;  
MM, mismatch  
NGS, next-generation sequencing;  
OXPHOS, oxidative phosphorylation;  
PC, principal component;  
PCA, Principal component analysis;  
RMA, Robust Multichip Average (normalization);  
RNA-Seq, RNA sequencing;  
RPKM, Reads Per Kilobase of transcript per Million mapped reads;  
SAM, Significance Analysis of Microarray  
SEM; standard error of the mean;  
SI, whole-body insulin sensitivity;  
SNP, single nucleotide polymorphism  
T2DM, Type 2 diabetes mellitus;  
TCA, tricarboxylic acid cycle;  
TG, triglycerides;

# **1 General Introduction**

## **1.1 Mammalian skeletal muscle tissue**

Skeletal muscle, as a consequence of its mass and great metabolic capacity, has a major impact on whole-body metabolic homeostasis [1, 2]. Besides the obvious role in locomotion, breathing, and postural maintenance, skeletal muscle is an important thermogenic tissue as well as a significant determinant of the body's overall basal metabolic rate (a reflection of the body's 'idling speed') [3, 4]. As a consequence of this, the maintenance of a metabolically active muscle mass is key for optimal health. In line with this notion, previous epidemiological studies have shown an association between peripheral skeletal muscle weakness and all-cause mortality [5–7].

More recently, less intuitive skeletal muscle functions such as endocrine and possible immune effects have been reported in the literature [1].

### **1.1.1 Skeletal muscle fibre types and properties**

Skeletal muscles are made up of bundles of long aligned myofibers (aka fascicles). The myofibers usually extend the entire length of the muscle. Each myofiber is a multinucleated, terminal differentiated cell formed by the fusion of multiple myoblasts. There are about 600 different muscles in the human body, ranging in size depending on anatomical location; the smallest only contain a few hundred myofibers whereas the powerful thigh muscles contain several hundred thousand myofibers [8].

Myofibers can develop tension and shorten (*i.e.* contract) by moving specialized cellular components; so-called cross-bridge interactions between actin and myosin molecule

complexes leading to a conformational change. The muscle cross-bridge cycling is heavily dependent on ATP, however this energy-rich molecule is only immediately available in tiny amounts (enough to sustain contractions for a few seconds only). As a result, the myofibers rely on several intracellular pathways to supply additional ATP when needed: creatine phosphate (phosphocreatine), oxidative phosphorylation (OxPhos) and glycolysis.

OxPhos, which takes place within the mitochondria (a self-replicating organelle), is a multi-step pathway that effectively harnesses energy from the breakdown of nutrient molecules such as glucose and fatty acids (32 ATP molecules/glucose molecule). Notably, oxygen is required to support the mitochondrial respiration. If the delivery of O<sub>2</sub> from the vasculature to the mitochondria is impaired, whatever the reason might be, the anaerobic glycolytic pathway will immediately take over. Further, glycolysis will also kick in if OxPhos is not able to meet the muscle's demand during intense physical activity. During glycolysis, sugars are chemically broken down to pyruvate aerobically *or* lactate anaerobically, yielding two ATP for each glucose metabolized. Pyruvate is then further degraded by OxPhos to yield more ATP.

Histochemistry (*e.g.* anti-myosin adenosine triphosphatase staining) has shown that mammalian myofibers can have different physiological/biochemical properties. Based on such properties fibers are defined as 'Type I' or 'Type II' [9]. Notably, the latter term is further sub-classified into Type IIa and Type IIx based on ATP-synthesizing ability.

Because of the different properties highlighted in Table 1-1, myofibers have distinct ‘tasks’: Type I fibers are responsible for postural support and prolonged low-intensity activities as a result of the high mitochondrial density. In contrast, Type II fibers are key players during mechanical loading.

	Slow-Oxidative Type I	Type II	
		Fast-Oxidative Type IIa	Fast-Glycolytic Type IIx
<b>General Properties</b>			
Myosin heavy-chain isoform	MHC1	MHC2A	MHC2X
Contractility	Slow twitch	Fast twitch	Fast twitch
Colour of fiber	Red	Red	White
Resistance to fatigue	High	Intermediate	Low
Recruitment threshold	All intensities	>40% VO <sub>2</sub> max	>75% VO <sub>2</sub> max
<b>Morphological properties</b>			
Capillary density	Dense	Dense	Sparse
Mitochondrial density	High	Intermediate-High	Low
<b>Metabolic Properties</b>			
Oxidative phosphorylation capacity	High	Moderate-High	Low
Glycolytic capacity	Low	High	High
Exercise-type dominance	Prolonged low-intensity	Moderate duration, high intensity	Short duration, maximal effort

**Table 1-1 | Characteristics of the different human skeletal muscle fibre types.**  
Adapted from Egan & Zierath [10].

Adult mammalian skeletal muscles contain different proportions of myofiber types depending on the anatomical location. As a result of this diversity, individual muscles overall have different contractile speed, strength and resistance of fatigue (Table 1-1).

Noteworthy, adult skeletal muscles have the capacity to adapt at the myofiber level to various exogenous stimuli such as alterations in hormonal milieu, chronic systemic

disease or long-term exercise training. Adaptations include growing or shrinking as well as switching the expression of myosin heavy chains genes [11, 12].

The *vastus lateralis* thigh muscle, the human muscle tissue of interest throughout the thesis, is composed of ~50% slow-twitch type I fibres in healthy sedentary subjects [13]. However, as extreme examples of the potential for adaptation, in elite sprinters the *vastus* can be made up of >70% fast fibres [14], whereas world-class cyclists and marathon runners can have less than 40% [12].

### **1.1.2 Skeletal muscle adaptation**

Skeletal muscle is a highly plastic organ—2<sup>nd</sup> to the nervous tissue only—that readily changes its phenotype in response to various repeated stimuli (or lack of stimuli due to inactivity or denervation) [15]. Such stimuli include among others changes in mechanical loading (*e.g.* strength training, microgravity) [16], neuromuscular activity (*e.g.* endurance exercise training) [10], and metabolic perturbations (*e.g.* nutrient availability, tissue hypoxia) [17, 18]. This adaptability is an important feature of skeletal muscle tissue as it allows for its economic ‘design’ by maximizing intracellular biochemical processes.

From a molecular perspective, any adaptation can ultimately be viewed as the accretion or loss of key proteins induced by a given stimulus. Accordingly, the altered gene expression response that initiates changes in intracellular protein concentrations—by making more mRNA available for ribosomal translation—is of significant importance. Currently, much of our understanding of skeletal muscle gene expression (as reflected by

mRNA abundance) is focused at the level of transcription, orchestrated by a complex network of transcription factors and their co-regulators.

Theoretically, muscle gene expression can be controlled by *i)* regulating the number of myonuclei (*i.e.* DNA content), *ii)* gene transcription rate, *iii)* post-transcriptional control, *iv)* mRNA translation and *v)* muscle protein degradation, with each of these molecular events being susceptible targets of regulatory influences triggered by the stimulus (*e.g.* a bout of exercise).

The involvement of cell signaling pathways in the transduction of various stimuli into the activation of specific gene expression events is widely recognized [19]. Such changes in muscular gene expression ultimately result in an incremental adaptation in protein level as well as activity (*e.g.* enzymatic activity), thus modifying the structural composition and/or functional properties [10]. Hence, gene expression is an important layer of processing for the integration of various exogenous stimuli into the adjustments of ‘muscle makeup’ necessary to match muscle function to alterations in demand.

Although considerable progress has been made in understanding the functional adaptations of skeletal muscle at the cellular level, the underlying molecular pathways remain obscure, especially those related to translational regulation. In fact, the abundance of a particular mRNA does not necessarily reflect the level of the corresponding protein. This is because the complex mechanisms that regulate protein expression are not dependent on mRNA alone [20]. Finally, the majority of molecular mechanism suggested to govern muscular adaptation in humans, originate from *in vitro* cell models and animal studies (*e.g.* knock-out mice). Hence, more clinical studies are currently needed to help shed more light at the molecular level.

### **1.1.2.1 Skeletal muscle protein turnover**

Skeletal muscle plasticity is, at least in part, due to a dynamic balance between protein synthesis and degradation. Any shift in this balance will lead to qualitative and quantitative alterations in myofibers and associated structures. The muscle protein turnover is known to be influenced by a numerous of extrinsic (*e.g.* feeding, fasting, mechanical loading, immobilization) and intrinsic factors (*e.g.* hormones, disease). Notably, an imbalance between protein synthesis and breakdown (*i.e.* higher degradation or reduced synthesis rates) has been reported to be a key mechanism for the loss of skeletal muscle mass and function [21]. Of particular interest to this thesis are two factors: *i)* chronic obstructive pulmonary disease (COPD)-induced peripheral muscle dysfunction, and *ii)* progressive endurance exercise training.

### **1.1.3 Interindividual variability in skeletal muscle adaptability**

A large variability in the responsiveness and degree of peripheral skeletal muscle adaptation is observed among individuals — despite using carefully controlled exercise training regiments as stimuli [22, 23]. A number of factors have been reported to affect—although only to a minor degree—this adaptive response such as age, nutritional support and genetic predisposition.

Notably, Timmons and co-workers have demonstrated that specific genes might determine, at least in part, how much aerobic fitness increases (*i.e.* gain in  $\text{VO}_{2\text{max}}$ ) following progressive endurance exercise [24]. They demonstrated that a group of 29 genes could categorize individuals into low, medium, and high responders to endurance exercise based on three independent data sets. Similar approaches are also required to be applied to other important clinical endpoints in order to provide initial clues as to what

causes the observed inter-individual variation. In continuation of this, the latter part of this thesis sets out to develop predictive multivariate models of key phenotypic traits known to affect overall human health.

## 1.2 Application of systems biology approaches to human health

It is now clear that much of the complex mammalian muscle physiology or pathophysiology cannot be understood in sufficient detail through a traditional hypothesis-driven reductionist approach alone. Although this approach has proved valuable in explaining broader phenomena and individual mechanisms, linking multiple mechanisms and effects has proved challenging, *e.g.* a disease phenotype is very rarely caused by a single dysfunctional gene or protein [25].

Instead, genetic variability, epigenetic modifications, and post-transcriptional regulation mechanisms etc. all act in concert to determine a specific high-level phenotypic response/adaptation [26]. The potential for such complex interaction makes data interpretation much more complicated than originally envisioned, highlighting the need to move away from the widespread ‘candidate gene approach’ [27].

Triggered by the advent of genome sequencing, inspired by the Human Genome Project [28, 29], dramatic technological advances within the last two decades have led to increased throughput in genome-wide molecular analyses (*i.e.* genomics, epigenomics, transcriptomics, proteomics, and metabolomics). The dense accumulation of data from these tools allows for unbiased and hypothesis-free analysis approaches.

As this thesis revolves around the use of such omics approaches, I will now describe some of them in more detail – with a particular focus on transcriptomic profiling by commercial microarray platforms.

## 1.2.1 High Density Data Acquisition

### 1.2.1.1 GWAS SNP genotyping

Many smaller DNA regions across the entire human genome vary between individuals due to insertions/deletions (indels), copy number variation among others [30, 31]. On average each individual has between 3 and 4 million polymorphic loci within their genome [32]. This genetic variation is what makes you and I unique.

Single nucleotide polymorphisms (SNPs) by far exert the biggest contribution, constituting >85% of the total genetic variation [33]. SNPs are randomly distributed throughout the genome, occurring every 500-1,000 base pair on average [34]. These single base DNA variants may fall within the gene-coding region or in the intergenic regions (*i.e.* regions between genes). Many SNPs outside the protein-coding regions have no effect on gene expression or gene products and thus appear to be phenotypically silent. However, those SNPs that are not silent provide a molecular basis for genetic variation that encompasses susceptibility to diseases and responses to exogenous exposures.

The number of reported common<sup>1</sup> human SNPs, which are deposited in publicly available databases such as the National Center for Biotechnology Information (NCBI) dbSNP, currently exceeds 15 millions (and still growing) [35].

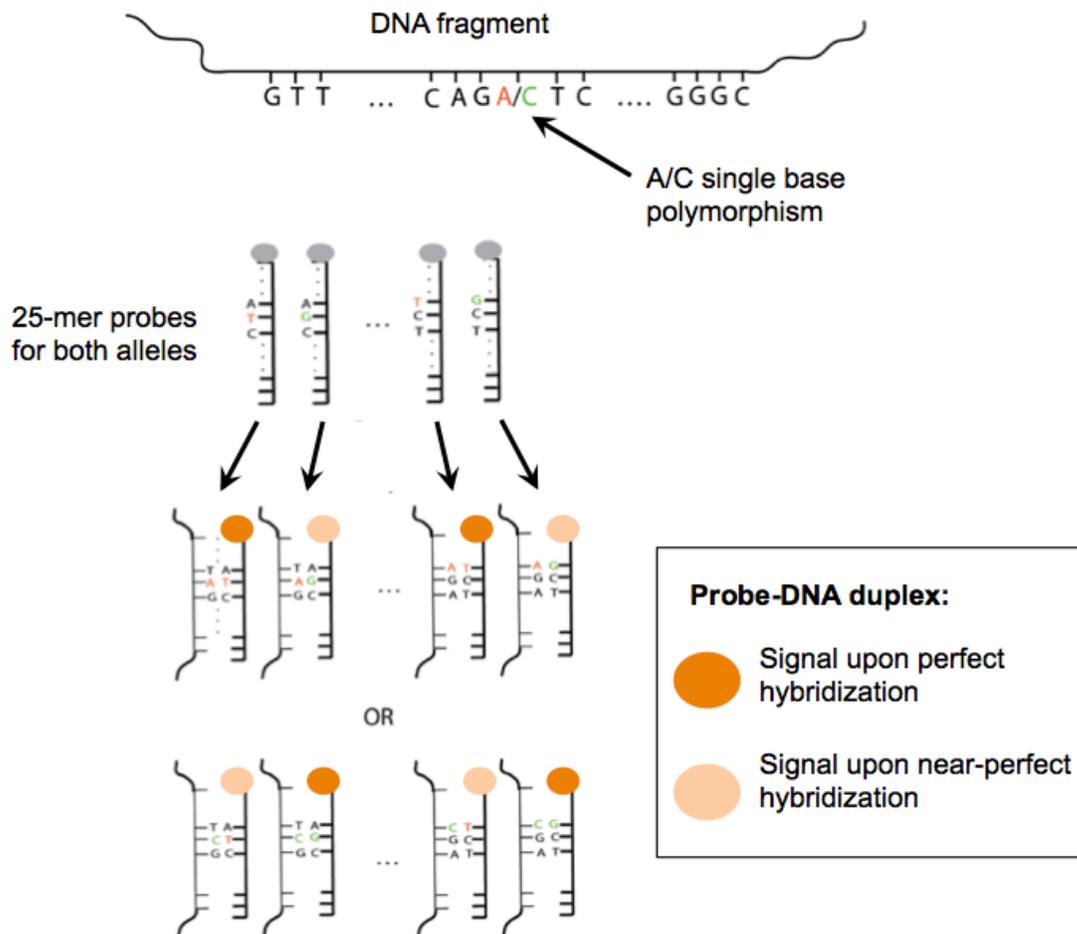
However, today the workflow for SNP-based studies has shifted away from their discovery toward SNP genotype selection — serving as landmarks in the search for genes associated with disease and complex traits.

---

<sup>1</sup> Minor allele frequency 5% or more in the broader population.

The small size of SNPs, as well as their evolutionary stability, has facilitated high-throughput genotyping by use of high-density oligonucleotide array technology (*i.e.* SNP chips/arrays). In addition, as specific combinations of neighboring alleles from different SNPs tend to be inherited together (loci not easily separable by genetic recombination due to their close physical proximity), SNP-chips are in fact able to survey a large portion of the human genetic variation in the form of SNPs [36].

Commercial probe-based SNP-chips are mainly manufactured by two providers: Affymetrix (Santa Clara, CA, USA) and Illumina (San Diego, CA, USA) [36]. Both rely on the Watson-Crick base pairing rules [37]. Notably, the basic principles behind SNP chips are the same as for mRNA microarrays, which is discussed in great detail in the next section. In brief, each surface-immobilized oligonucleotide probe on the array has been designed to target a specific DNA sequence. The signal intensity from the attached fluorescent dye is dependent upon the hybridization affinity between DNA target and capture probe (Figure 1-1). Impeded hybridization—due to a mismatch at the SNP site—will lead to a dimmer signal from the attached fluorescent dye. This base-dependent difference in signal intensity is then used for genotyping.



**Figure 1-1 | Overview of the Affymetrix SNP chip technology.**  
Figure adapted from [38].

### 1.2.1.2 Transcriptome expression profiling

Messenger RNA is far less redundant compared to that of DNA as only 5-10% of the human genome is actively transcribed (of which only around 2% is translated into proteins) [39]. However, mRNA is more dynamic as it integrates genetic [40–42], epigenetic [43, 44] and environmental [45–47] influences, thereby capturing substantially more variance than genomic DNA. In addition, the transcriptome is cell-type specific and time/context dependent.

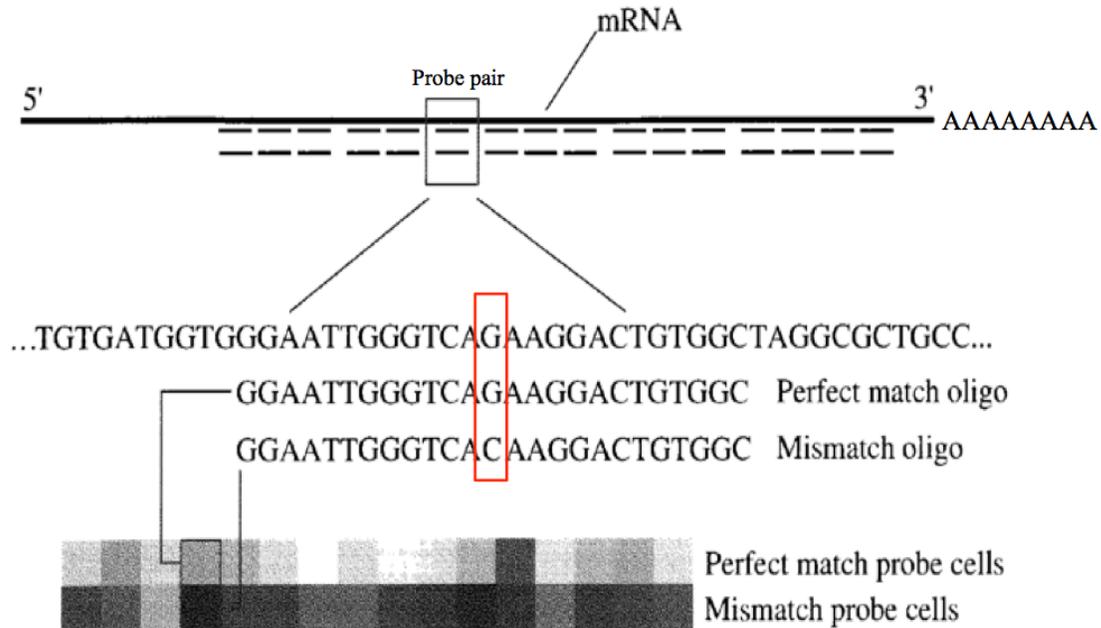
Consequently, the transcriptome, as an intermediate molecular phenotype, is substantially more complex than the genome and allows for interpreting the functional elements in the genome.

The development of whole-genome microarrays has made it possible to interrogate the entire transcriptome of a cell or tissue (*i.e.* the complete set of transcripts expressed). The proven track record of this technology, which spans nearly two decades [48], has helped improve our understanding of molecular processes underlying phenotypes by dissecting natural variation in mRNA abundance.

***Differences in commercial microarray technology solutions:***

Despite the gaining popularity of next-generation sequencing (see next section), microarrays still remain the most popular technology for transcript profiling as they can be readily afforded by most laboratories worldwide; since 2010 ~250,000 arrays have yearly been indexed by the ArrayExpress database [49]. In addition, the short turn-around time and ease of data generation add to the continued popularity of microarrays.

Today commercial microarrays are mainly manufactured by three providers: Affymetrix, Agilent and Illumina. Affymetrix creates high-density short oligonucleotide sequence arrays by use of a photolithographic synthesis process [50]. Each target transcript is represented by a number of ‘probes’ (*i.e.* specific 25-nucleotide sequences) that match to different exonic gene regions, which together make up a ‘probe-set’. Notably, each probe is actually a ‘probe pair’ (at least for the older 3’-based arrays), as it consists of a perfect match and a mismatch (MM) oligo (Figure 1-2). The rationale for the MM oligos is to gauge the level of non-specific cDNA binding (see Section 1.2.2 below for more information on this issue).

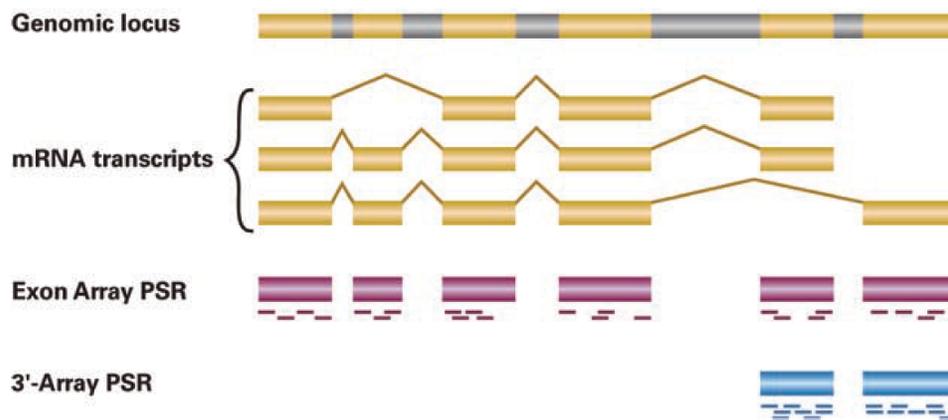


**Figure 1-2 | Probe design on the older 3'-based Affymetrix microarrays.**

For each probe pair, the mismatch oligo has a central base altered (see red square) to measure background. The brighter the colour the more cRNA has hybridized. Figure adapted from [51].

Such a probe design not only allows for unspecific binding estimation, but also evaluation of RNA quality [52]. The latter is particularly important when analyzing data from the public domain, as will be evident from several chapters within this thesis (see Section 3.3.6).

Noteworthy, the newer Affymetrix exon arrays only contain 25 nt-long *perfect* match oligos designed to target individual exons in a given transcript (Figure 1-3). On average each 'exon probeset' includes four probes. Due to the lack of MM oligos, Affymetrix instead added ~1,000 Background Intensity Probes, which are designed *not* to target any expressed sequences in the organism of interest.



**Figure 1-3 | Schematic overview of the difference in Affymetrix probeset design between their newer exon arrays and the older 3'-based arrays.**

Yellow regions represent exons, whereas grey regions represent introns. The short dashes underneath the exon regions indicate individual probes of 25 nt in length – each representing a probe-set. Figure taken from [53].

Agilent, an offshoot of Hewlett-Packard, on the other hand uses a DNA synthesis method based upon inkjet printing technology. This innovative approach provides customers with the ability to quickly order custom designed microarrays at no extra cost. This is particularly useful when working with species for which no commercially available microarrays exists (*e.g.* non-model organisms). In Chapter 3, we take advantage of this when developing the first genome-wide microarray platform for the guinea pig model species.

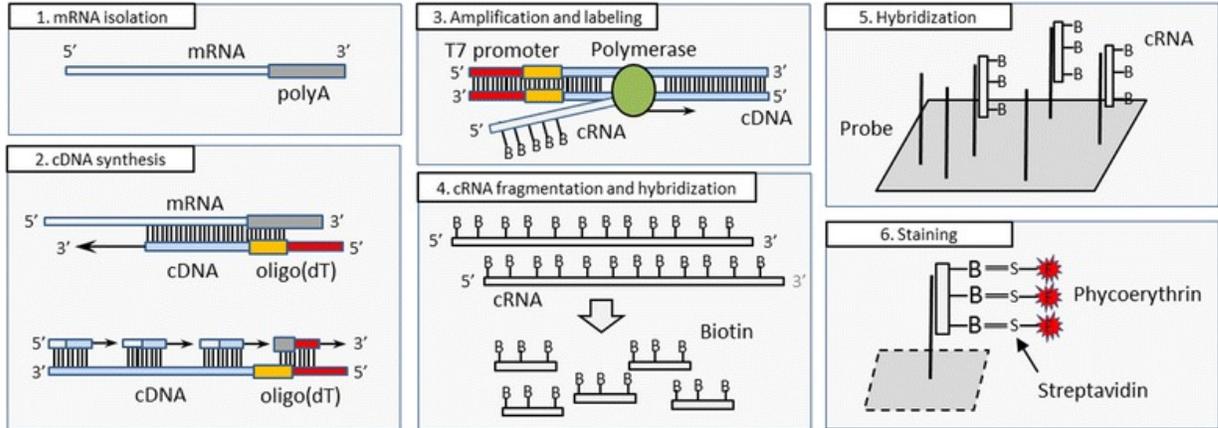
As data generated by mRNA microarrays represent a central part of most chapters throughout this thesis, I will briefly describe a typical sample-to-data workflow using Agilent technology (other technologies will differ slightly).

As depicted in Figure 1-4 below, RNA is first extracted from the biological sample of interest (tissue or cells) using either a spin column-based method or the more traditional phenol-

chloroform based one [54, 55]. Next, excess oligo-dT primers (bearing a T7 promoter) are added which will hybridize to the poly(A) tail of all mRNAs once the temperature is raised to 65°C. Then reverse transcriptase enzyme and deoxynucleotides are added to the mixture allowing the creation of cDNA molecules. Notably, due to this initial linear mRNA amplification step, as little as 25ng of total RNA is sufficient as input.

In order to incorporate fluorescence dye (Cyanine-5 or Cyanine-3), the cDNA is then converted to complementary/copy RNA (cRNA) by a T7 promoter-specific RNA polymerase. The resulting labeled cRNA is fragmented into shorter, less structured fragments before hybridization to the surface-bound capture probes over night. The fragmentation step is imperative in order to reduce structural effects, as secondary and tertiary structures can significantly decrease hybridization efficiency [56, 57].

After washing off remaining unbound cRNA, a high-resolution laser-scanning machine captures the fluorescence signal from labeled cRNA bound to each microarray probe. The amount of fluorescence signal is proportional to the amount of bound cRNA, which obviously is dependent on the transcript abundance within the cells/tissue. Finally, an image analysis procedure, which recognizes the two dimensional position of each spot in the scanned image, is used to extract a numeric representation of each probe on the array based on pixel intensities. These intensity values are representative of the *relative* amounts of mRNA. The relative quantification has to do with microarray intensities being prone to background noise arising from non-specific binding of mRNA species that are only partially complementary to the probe [58]. Once a numeric data table has been acquired the actual data manipulation and analysis can begin



**Figure 1-4 | Individual web-lab steps of a microarray experiment.**  
 Figure taken from [59].

### 1.2.1.3 Next Generation Sequencing (NGS)

As compared to the hybridization-based array technologies presented so far, NGS does not rely on existing knowledge on genome sequence. Here RNA or DNA is sequenced by reading each base/nucleotide, which enables the reconstruction of the genome or transcriptome (depending on biological starting material).

In addition, RNA sequencing (RNA-Seq) provides a more detailed examination of the transcriptome due to its ability to characterize different splice isoforms and broader dynamic range (*i.e.* able to quantify even very lowly expressed transcripts due to a very low background signal) [60]. For the reasons mentioned above, RNA-Seq has a strong potential to replace microarrays for whole-genome transcriptome profiling. The latter statement is supported by the immense number of peer-reviewed research papers citing NGS technologies. Further, RNA-Seq is able to capture both common and rare genetic variants, making this technology a strong alternative to array-based SNP assays mentioned earlier.

Similar to microarrays, RNAs are initially reverse-transcribed to cDNA. This is then fragmented and platform-specific ‘adaptor’ sequences are attached to one or both ends of the cDNA fragments. Notably, an additional size- and type-selection step (*e.g.* ribosomal RNA depletion) is often included in order to increase read depth.

Next, each modified cDNA molecule is sequenced by use of high-throughput sequencing technology. Finally, all of the resulting short sequences (so-called ‘reads’) are assembled into longer fragments and, most often, aligned to a reference transcriptome. Alternatively, all reads can be assembled *de novo* to produce a genome-scale map [60].

The number of reads sequenced per gene sequence is used as a reliable proxy of mRNA abundance (once corrected for the sequence length).

## **1.2.2 Microarray data analysis**

### **1.2.2.1 Raw data pre-processing**

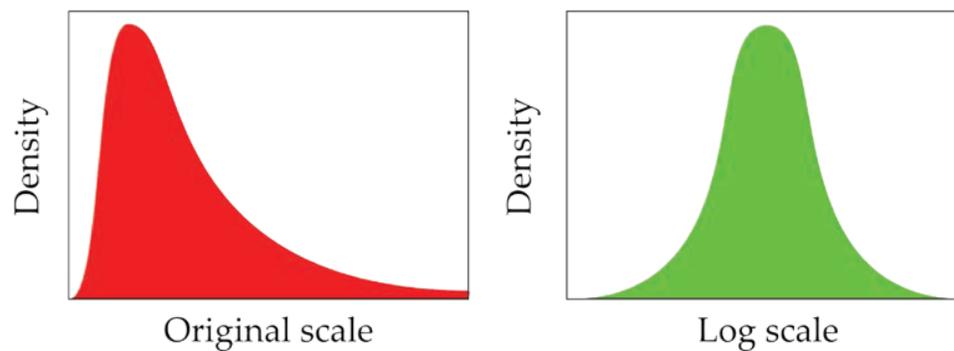
The overall aim of microarray data processing is to remove noise and systematic variability while preserving the biological heterogeneity. In this context, ‘systematic’ refers to effects that affect *all* data points on the array in the *same* way (*e.g.* an overall higher or lower background intensity). The main sources of unwanted interarray variability originate from dye bias (specific to multichannel experiments), differences in sample labeling efficiency as well as differing yields during purification.

#### **Data transformation (usually $\log_2$ ):**

As highlighted in Figure 1-5, the distribution of raw signal intensity values is heavily skewed towards the low signal intensity side. Hence, a data transformation procedure, often logarithm

base two ( $\log_2$ ), is normally applied in order to force the spread of features more evenly across the intensity range. In addition, most statistical tests developed for microarray analysis are parametric and hence assume the data follow a normal distribution.

In brief, the logarithm function will squeeze together larger intensity values (*i.e.* genes abundantly expressed in the tissue of interest) and stretch out those in the low intensity area, thereby causing an overall distributional right-shift (Figure 1-5).



**Figure 1-5 | Histogram showing the distribution of signal intensities before (left panel) and after (right panel)  $\log_2$  data transformation.**

### **Inter-array normalization:**

In an ideal microarray experiment normalization would *not* be required. However, as highlighted in Section 1.2.1.2 below, the microarray experimentation process is fairly complex. As a result of this, systematic errors are frequently introduced during RNA extraction, labeling, hybridization and scanning, which can result in artificial differences between replicated samples. Importantly, such differences may likely confound the biological differences between sample groups in a given experimental setup. Hence, the overall aim of normalization is to remove as much of this unwanted experimental and technical variation while maintaining true biological variability (as much as possible).

Multiple global normalization methods have been developed during the last decade, see [61] for a comprehensive review on this topic. Notably, the choice of method is dependent on the design of array capture probes (*i.e.* 3'-based versus the newer whole transcript based chips), number of channels, as well as manufacturer.

Quantile normalization proposed by Bolstad *et al.* [62], used throughout this thesis, is a popular method likely due to it being mathematically simple (Figure 1-6), fast and easy to implement (needs no user interaction/parameter tuning).

This normalization procedure forces each array in a set of arrays to have the same empirical distribution, as it is assumed that the distributions of intensity values on the slides have the same overall shape. Plotting histograms of the raw probe intensities on a single plot can help establish whether such an assumption is in fact valid. The latter point is important, yet often ignored by researchers (in my experience at least), as the normalization procedure by design always 'works', even if data are bad.

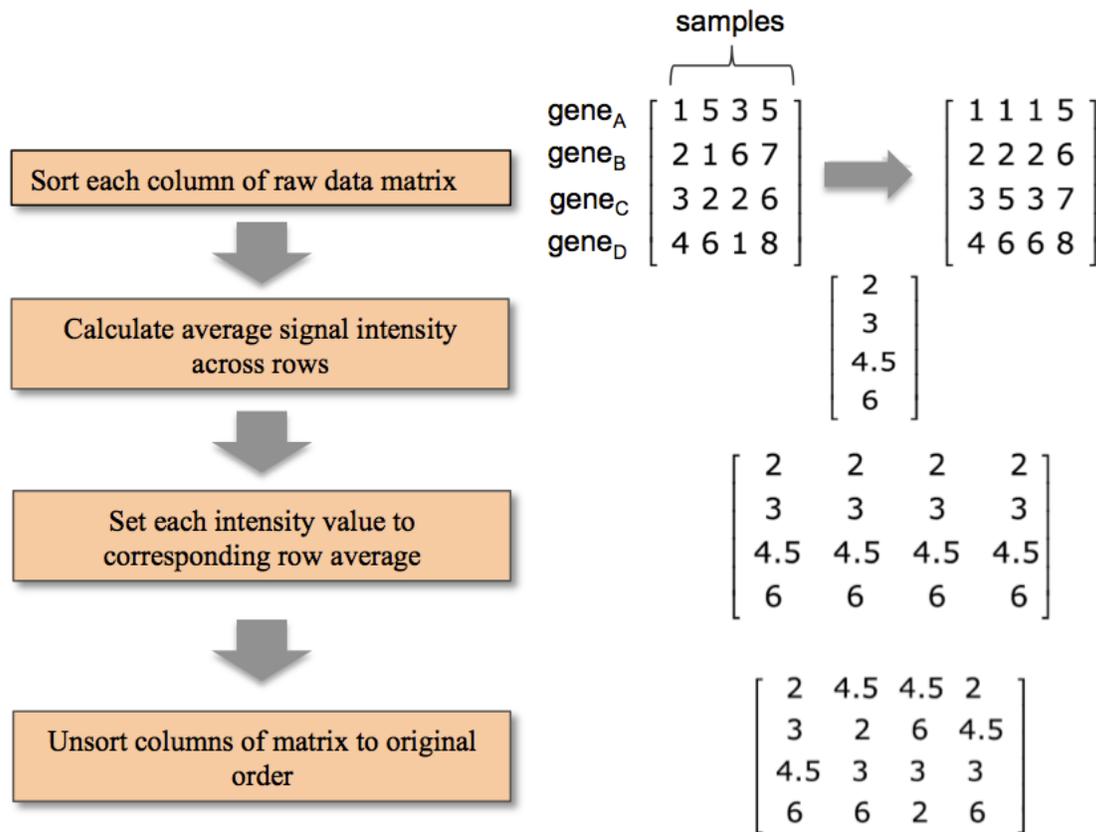


Figure 1-6 | Details of the quantile normalization procedure for expression arrays.

**Gene/Probe-set summarization:**

Due to their dense nature commercial microarrays often have several million features spotted, which allows for multiple features per target gene. The goal of the summarization step is to produce a unique measure for every gene as close to the ‘true’ gene expression as possible.

Although optional, two approaches are commonly preferred:

1. Selecting a single, most representative probe-set from a set of probe-sets (*e.g.* one with the highest expression).
2. Taking the average expression value of all gene specific probe-sets

Noteworthy, Li *et al.* recently published a probe-set scoring method that exploits the unique properties of the older 3'-based<sup>2</sup> Affymetrix arrays [63]. In brief, this method assesses each probeset for its specificity, splice isoform coverage, and the robustness against degraded target transcript.

### **Gene filtering:**

Modern commercial mRNA microarray platforms allow the simultaneous profiling of over 20,000 unique genes (the human genome contains around 21,000 protein-coding genes [64, 65]). However, not all of the profiled transcripts are expected to be expressed in a given experimental condition, as most mammalian tissues only express ~10,000-15,000 protein-coding genes at biologically meaningful levels [66, 67] (and likely even fewer). The recorded signal intensity values from probesets targeting non-expressed transcripts simply represent noise due to non-specific binding — particularly from abundantly expressed RNA species [68]. Hence, in order to increase sensitivity (*e.g.* in finding differentially expressed genes) it is common practice to remove all probesets below or close to the detection limit (*i.e.* background). However, the definition of 'background' is a non-trivial issue, as highlighted by the many different approaches put forward in the literature. To mention a few: signal from anti-genomic probesets (*i.e.* negative controls included on the array by the manufacturer), average intensity signal [69], and percent present-calls [70].

---

<sup>2</sup> Probes are designed to target sequences near the 3'-end of an mRNA transcript to minimize loss of fluorescence signal caused of non-specific RNA degradation.

*MAS 5.0 absent/present calls (specific for Affymetrix chips):*

Throughout this thesis I take advantage of the Affymetrix Human Genome U133 Plus 2.0 GeneChip®. Probes on this array are selected to interrogate the 3'-end of an mRNA species (Figure 1-3). In addition, this platform contains 'mismatch' probes intended to quantify the amount of unspecific binding (Figure 1-2).

A probeset is termed detected/expressed (*i.e.* 'present') by the Affymetrix MAS5 software when the signal intensity of the perfect match oligos is statistically higher ( $p < 0.04$ ) than the corresponding mismatch oligos based on a Wilcoxon signed-rank test (a non-parametric alternative to a paired Student's t-test) [71].

Notably, it has been demonstrated that such a filtering approach increases the correlation to quantitative PCR expression measurements [72].

#### **1.2.2.2 Exploratory analyses to assess data quality**

It is important to realize that effects of degraded RNA, etc. cannot be removed by any normalization procedure.

Exploratory data analysis techniques are vital for outlier detection, identification of sample and gene similarity/dissimilarity and dimension reduction.

The very high dimensionality of microarray data makes direct visualization impossible as the human eye only can handle a maximum of 3 dimensions. Hence, dimensionality reduction techniques are very popular when analyzing omics datasets.

Worth mentioning is principal component analysis (PCA), an unsupervised dimensionality reduction technique, that is widely used within the community to find dominating patterns in

multidimensional data sets [73, 74]. In brief, this well-established mathematical technique exploits that complex data usually is not uniformly distributed, but show strong correlations among groups of measured variables. Such correlation highlights a certain degree of redundancy within a data set. Hence, the number of underlying factors accounting for most of the variation with the data is much smaller than the dimensions of the data itself.

PCA reduces data dimensionality by creating new ‘artificial’ variables (*i.e.* orthogonal axes of best fit), referred to as principal components (PCs). Each of these PCs is a linear combination of the observed variables. Importantly, the different PCs are uncorrelated to each other thereby ensuring that they represent different characteristics of the original dataset [73]. By default, the first PC accounts for as much of the variability in the dataset as possible (and the last PC the least variance). Typically, the first handful of PCs explains the vast majority (>80%) of the variability. Conversely, two points (each represent a biological sample) that are visually separated in a PCA plot by one of the first PCs will have different overall expression profiles.

Other common tools used for exploratory analysis include tree-based visualization and clustering algorithms. Such methods can reveal the similarity between genes and/or samples [75]. Genes with a similar expression profile will cluster together, suggesting that they may represent a coordinated response to an experimental stimulus (*e.g.* a bout of exercise).

**Batch effects:**

Batch effects have been defined as “sub-groups of measurements that have qualitatively different behavior across conditions and *unrelated* to the biological or scientific variables in a study” [76]. As highlighted in Figure 1-7, many independent factors can contribute to the

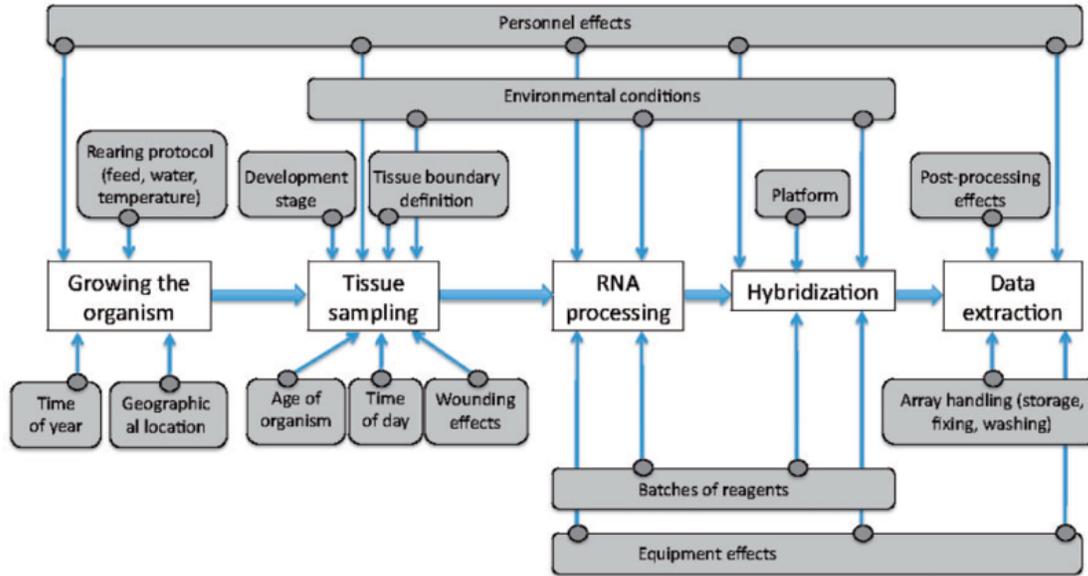
creation of batch effects. Particularly, non-biological sources of variation among microarrays tend to arise when they are processed in separate independent batches (*e.g.* on different days, which is common in larger clinical studies) [76]. Thus, it is important to apply data visualization in order to check for such batch effects. PCA plots are particularly suited as the biggest source of gene expression variability very often associate to technical variables rather than biological groups [76]. In addition, commercial providers add exogenous controls (also known as ‘spike-ins’), which can be useful for checking the quality of the different array processing steps [77].

If ignored, batch effects will likely lead to increased technical noise (which decreases detection power) and number of false positives. To illustrate the latter, if all control samples were processed on day 1 and all treated samples the following day, certain genes might appear differentially expressed by treatment, while this in fact is only due to the confounding batch effect.

Due to the frequent occurrences of batch effects in high-throughput data [76], the development of *in silico* removal methods is a very active research area. Hence, many different adjustment methods now exist based on different statistical models (for a recent review see [78]). One very popular algorithm based on an empirical Bayes framework, which I also take advantage of in Chapters 4 and 5, is ComBat [79]. Notably, a fairly recent systematic evaluation on simulated data identified ComBat as the best performing algorithm [80].

However, it must be acknowledged that *some* of the true biological signal will also be removed by any correction procedure (*i.e.* loss of intra-group biological heterogeneity), due to the difficulty in discerning unwanted variation from the experimental variation of interest [53].

Hence, it is key to reduce the effect of batches by careful experimental design [81, 82], particularly when only subtle differences are expected.



**Figure 1-7 | Flowchart highlighting potential batch effect sources (grey boxes) during data generation.**

Image taken from [78].

### 1.2.2.3 Defining transcriptional signatures of perturbations

With the advent of omics technologies, also surfaced the challenge of interpreting the high volume of data.

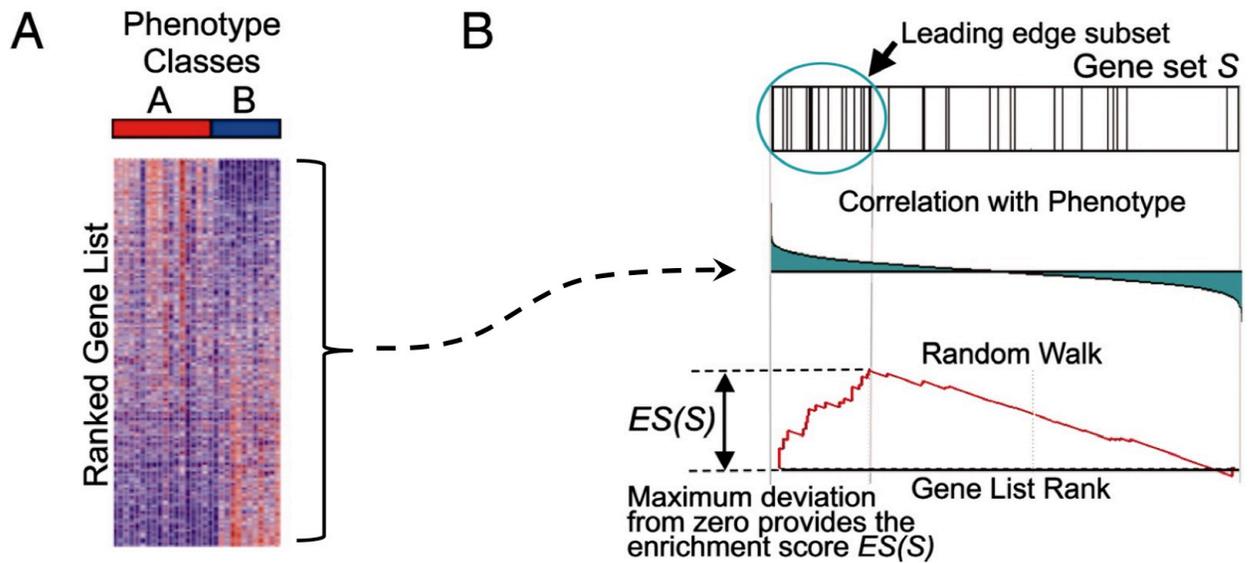
A popular approach is to identify genes that are differentially expressed between experimental conditions based on a user-defined preselected probability cutoff. However, such singular enrichment-based analyses have been criticized for not considering the biological knowledge (*i.e.* genes known to be biologically associated are scored individually) [83].

Hence, novel statistical methods were developed that relies on the principle that genes do not work in isolation, but in an intricate network of interactions.

*Gene Set Enrichment Analysis (GSEA):*

GSEA [84] differs from other overrepresentation analysis tools in that this does not require the user to set an arbitrary probability threshold beforehand. For this reason, the GSEA procedure should *potentially* uncover additional biological data trends (idea is to ‘borrow’ strength), and hence well suited for complex polygenic phenotypes.

In brief, an Enrichment Score (ES) is calculated for each gene set by ranking all genes in the transcriptome based on their association with the phenotype of interest (Figure 1-8). Next, the statistical significance of the observed ES is assessed by permuting phenotype labels many times (*i.e.* the observed ES is the distribution of permuted enrichment scores). Finally, probability values are adjusted for multiple hypothesis testing.



**Figure 1-8 | Schematic overview of the Gene Set Enrichment Analysis procedure.**

Figure adapted from [84]. All genes are ranked according to the fold-change difference in mRNA expression between Phenotype class A and class B (see colour bar at the top of the heatmap in Panel A). Panel B: For a given target gene-set (*e.g.* a KEGG pathway): every present gene (black vertical bar) gives a positive contribution, whereas every absent gene in the ranked list gives a negative contribution to the GSEA enrichment score (ES) as the algorithm walks down the ranked gene list (see red trace in the ES profile at the bottom). The cumulative ES with the biggest deviation from zero proves the final ES (so a high absolute ES denotes a high local enrichment). Distributions of random ES scores from  $N$  permutations is then generated (*i.e.* null-hypothesis distributions) in order to estimate a p-value for every tested gene-set. Finally, probability values are adjusted for multiple hypothesis testing.

### 1.3 Aims and outline of the thesis

#### **Overarching aim**

The overall aim of the thesis was to improve the current understanding of the molecular basis for lower limb skeletal muscle adaptation/dysfunction under two key conditions known to affect its phenotype, namely long-term endurance exercise training and chronic obstructive pulmonary disease (COPD) characterized by peripheral skeletal muscle wasting.

#### **Rationale and specific objectives:**

Despite decades of intense investigation, our understanding of the regulatory molecular mechanisms controlling skeletal muscle homeostasis is still limited. Moreover, biomarkers to support the development of new therapeutic interventions to improve muscle-skeletal functionality in chronic degenerative diseases (*e.g.* COPD and T2DM) are very much needed.

The development of genomics and functional genomics have provided an unprecedented amount of data characterizing the molecular state of skeletal muscles in physiological and pathological conditions. Recently, it has also become apparent that computational approaches to integrate and model omics datasets are fundamental to formulate more robust testable hypotheses.

The current thesis represents a pioneering effort in this direction. More precisely, the current work is based on the **hypothesis** that genetic variants and abnormal or altered expression status/pattern of mRNAs, in lower limb muscles, can be linked to physiology

and that such molecular signatures—when combined with biological domain knowledge—can be exploited to develop robust quantitative predictors.

I have approached the challenge of improving biological understanding of the gene-networks that underlie response to physical exercise and muscle pathophysiology by applying a data-driven computational approach to both animal models and human intervention studies.

This highlights **three main challenges** for this thesis:

- 1) Develop an initial proof of concept to validate the application of data-driven biological network inference in understanding lower limb muscle maladaptation (**addressed in Chapter 2**).
- 2) By using a rodent chronic smoking model, identify and validate candidate inflammatory signals underlying peripheral skeletal muscle wasting in COPD patients (**addressed in Chapter 3**).
- 3) Utilise pre-intervention omics profiling to subsequently build response-predictors of systemic physiological traits (*i.e.* whole-body insulin sensitivity and serum triglyceride content) using one of the biggest exercise training trials to date. Importantly, such quantitative predictors will have to be validated in *independent* clinical studies to avoid model overfitting (**addressed in Chapters 4 & 5**).

## **2 Multi-level functional genomics data integration as a tool for understanding physiology: A network biology perspective**

*The work presented in this chapter represents a collaborative project between the labs of Francesco Falciani and Stuart Egginton (University of Leeds). Prof. Egginton helped conduct the in vivo animal exposure experiment while him and I were both physically situated in Birmingham. Josep Roca (University of Barcelona) provided the anthropometric characteristics related to the clinical COPD cohort.*

*I independently conducted all the murine wet lab work (RNA extraction + array generation) as well as in silico analyses presented in the case study.*

*This chapter (in its current state) is currently under 2<sup>nd</sup> review in Journal of Applied Physiology.*

## 2.1 ABSTRACT

The overall aim of physiological research is to understand how living systems function in an integrative manner. Consequently, the discipline of physiology has since its infancy attempted to link multiple levels of biological organization. Increasingly this has involved mathematical and computational approaches, typically to model a small number of components spanning several levels of biological organization. With the advent of *omics* technologies, which can characterise the molecular state of a cell or tissue, the number of molecular components we can quantify has increased exponentially. Paradoxically, the unprecedented amount of experimental data has made it more difficult to derive conceptual models underlying essential mechanisms regulating mammalian physiology.

In this chapter I present an overview of state-of-the-art methods currently used for identifying biological networks underlying genome-wide responses. These are based on a data-driven approach that relies on advanced computational methods designed to ‘learn’ biology from observational data. Furthermore, I illustrate an application of these computational methodologies **using a case study** (proof of concept) integrating an *in vivo* model representing the transcriptional state of hypoxic skeletal muscle with a clinical study representing muscle wasting in COPD patients. The broader application of these approaches to modelling multiple levels of biological data in the context of modern physiology is discussed.

## 2.2 INTRODUCTION

### 2.2.1 Modelling in physiological sciences

Physiology has evolved as a series of sub-disciplines attempting to understand organismal function as a combination of interacting components and systems. The last decade or so has witnessed the development of Systems Biology as an investigative approach, and its application in different areas of biology, ranging from engineering/synthetic biology (*e.g.* design of bacterial strains with improved properties) to health sciences (*e.g.* disease biomarker identification). Despite the lack of a concise definition acceptable to the majority of the community [85, 86], Systems Biology is frequently understood to be the study of complex regulatory interactions in biological systems using a holistic approach. This is often achieved by integrating different experimental approaches within the conceptual framework of a computational model (*i.e.* a mathematical representation of a system that allows simulation of its behaviour). Physiology is probably one of the few research areas in biological sciences that have traditionally adopted such an approach. It has long sought to understand the behaviour of complex biological processes and cellular systems using an integrative approach, and has extensively adopted mathematical modelling in its tool set. Classical examples include August Krogh's tissue cylinder model of oxygen transport to skeletal muscle [87], and Huxley's two-state cross-bridge model of muscle contraction [88], which are still used by investigators today. Indeed, physiology can be considered a precursor of Systems Biology.

As often happens when a distinct discipline branches out of another, there developed over time a separation of ideas based in part on confusion arising from use of esoteric terminology – similar concepts masked by unfamiliar language. There is therefore a need for an overview of this relatively new discipline, to both emphasise the essential links with basic physiological principles and de-mystify the approach such that the available tools may become more widely adopted in physiological research. The overall aim of this opinion-based review is to describe, using concepts that will be intuitive to physiology researchers, different key methodologies available from the Systems Biology community. In addition, I provide a practical step-by-step guide for integrating multi-level data within an analysis pipeline based around inferred interactions of variables, modelled as a network based on statistical correlations, using a worked example in the field of physiological sciences.

### **2.2.2 The advent of Functional Genomics: a challenge for physiological modelling**

It appears that the function of living mammalian organisms cannot be addressed in satisfactory detail through a traditional hypothesis-driven approach alone. For many years, the attention of molecular biosciences was directed to specific rate-limiting enzymes and key genes with a high impact on physiological trait. However, a disease phenotype is very rarely caused by a single dysfunctional or dysregulated gene or protein. Instead, genetic variability, epigenetic modifications, post-transcriptional regulation mechanisms *etc.* all act in concert to determine a specific complex phenotypic response [26]. The potential for such complex multi-level interaction makes data interpretation

much more complicated than originally envisioned, highlighting the need to move away from the widespread ‘candidate gene’ approach [27].

The advent of DNA sequencing [28, 29] has enabled the development of technological advances, which have led to increased throughput in genome-wide molecular analyses. These technological advances have made it possible to measure gene, protein or metabolite concentration in single experiments at a genome-wide level.

The comprehensive data acquisition tools developed to cope with large datasets have allowed investigators to determine the molecular state of cells, tissues or even entire organs in a single experiment. Such cost-effective *omics* approaches are now becoming prevalent in biological and medical research, and consequently have been responsible for the generation of an incredibly large amount of multivariate molecular data. A large proportion of this data is available in the public domain *via* different online databases (e.g. NCBI Gene Expression Omnibus [89], EBI ArrayExpress [90], and PRIDE [91]).

For example, mRNA microarray technology and more recently mRNA sequencing, has provided insight into the transcriptional response of skeletal muscle to prolonged endurance exercise training, highlighting a pronounced inter-individual variation [19, 24]. The transcriptional signatures identified in such studies likely explain, at least in part, why some people show great improvements in aerobic capacity ( $VO_{2max}$ ) whereas others only experience smaller benefits, despite completing the same supervised exercise training program. Another example of applying *omics* technology to better understand human physiology concerns the quantification of individual levels of different proteins in health and disease; by use of proteomics methodology, Holloway et al. [92]

were the first to investigate adaptations in human muscle protein content to long-term exercise training on a large scale.

While such *omics*-based studies hint at the potential of a data-driven approach, they also illustrate the difficulty in deriving conceptual models underlying the essential mechanisms regulating physiology, as most are restricted to only one aspect of regulation. Perhaps surprisingly, the exponential growth in publicly available *omics* data [93, 94] has not resulted in a paradigm shift in our understanding of biology. The main reason is the continuing challenge of integrating multivariate datasets spanning multiple organization levels in a way that allows the identification of discrete, small biomolecular networks that are truly important in the context of a specific biological response [95]. Such a task cannot be achieved simply using unaided human interpretation. Rather, complex computational techniques are needed that are able to integrate and automatically ‘learn’ the structure of a biological system. Such a modelling framework is very different from what physiological sciences have traditionally employed.

### **2.2.3 Towards data-driven predictive biology**

Although the modelling approach traditionally used by physiologists has been extremely successful, it suffers from severe limitations when challenged with extensive *omics* data. For example, physiological modelling relies to various degrees on a mechanistic understanding of the biological system of interest [96], which automatically limits the number of components that can be included due to gaps in our current knowledge [95, 97]. Moreover, estimation of model parameters, which is usually a challenging task because of experimental limitations (*e.g.* due to limited amount and quality of data),

makes the approach difficult to scale up to a larger number of components and their interactions. Perhaps the most comprehensive example to date is modelling the cardiac cycle based on ion channel kinetics [98].

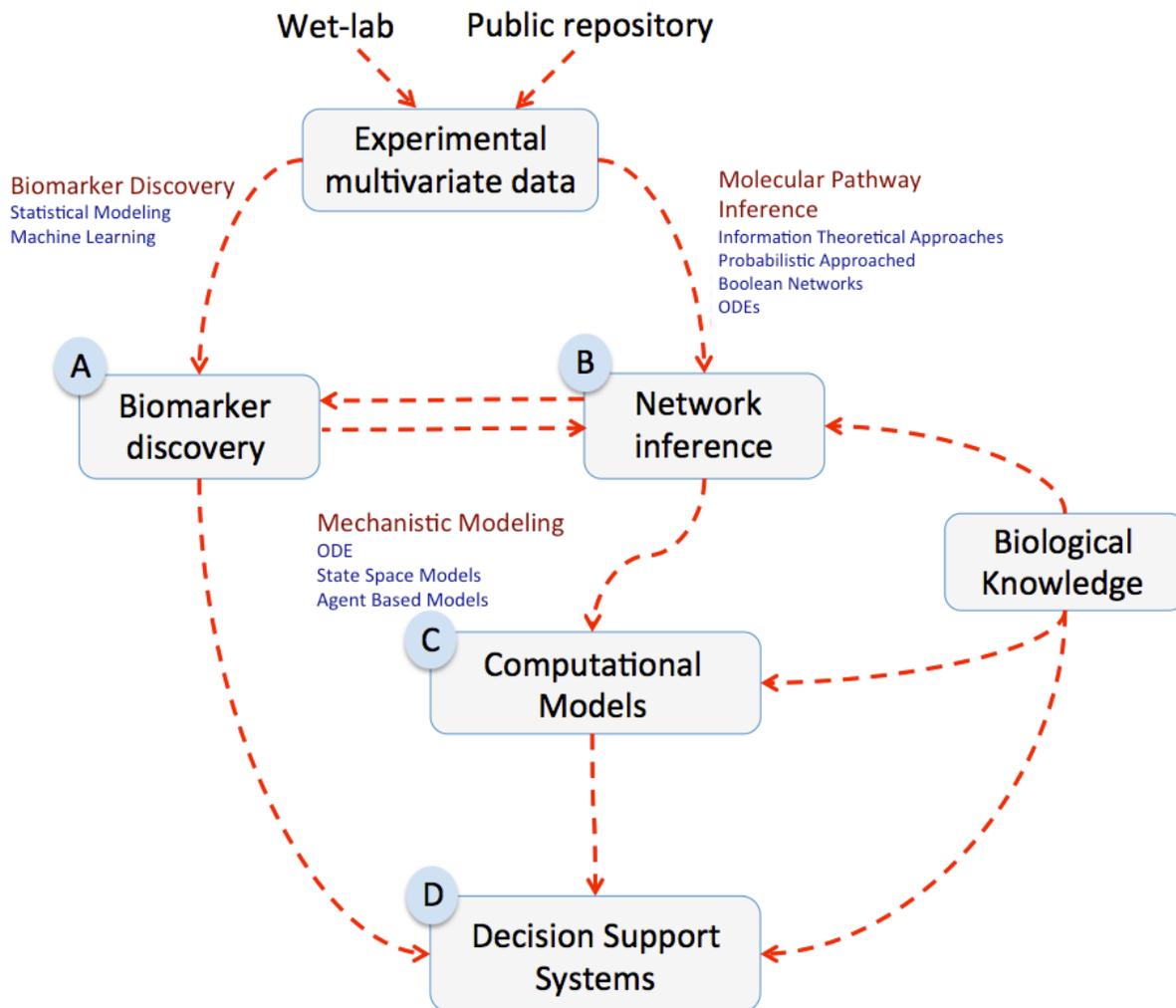
With such large multivariate datasets, and little knowledge about the way biomolecules are connected with each other and to key phenotypic switches, the fundamental question is whether or not we can ‘learn’ the structure of biological interaction networks from high-throughput data. Clearly, there is a need for sophisticated computational tools that are able to *i*) integrate genome-wide measurements spanning multiple levels of biological organization (ranging from subcellular to organ level), *ii*) identify key biomolecular components of the system, and finally *iii*) statistically infer the way that these biomolecules interact in a pairwise manner to generate an observed biological response.

Central to these approaches is the concept of interaction networks, a mathematical representation of a system of biomolecules. Networks are commonly used to describe biological systems at different levels of complexity (e.g. metabolic and signal transduction networks). They can be descriptive models built using a wide spectrum of qualitative data (e.g. biological knowledge of protein-protein interactions, transcription factor binding, etc.) or they can be inferred from quantitative measurements using complex computational models. In this case they can be used to predict the behaviour of the system when perturbed.

In the following section, we summarise specific methodologies that can be applied to achieve such tasks.

#### **2.2.4 COMPUTATIONAL APPROACHES FOR THE ANALYSIS OF COMPLEX DATASETS**

The process of modelling a biological system from complex multi-level datasets can, for the sake of convenience, be divided into four conceptually distinct yet interconnected approaches (Figure 2-1).



**Figure 2-1 | Schematic representation of the process involved in modelling a biological system by integrating knowledge from various sources, and complex multi-level datasets.**

The process can be conceptually subdivided into four distinct yet interconnected approaches (A-D). The experimental data used can either be novel multivariate data generated in your own (wet) laboratory or taken from a public repository. These may then be used to identify predictive biomarkers, i.e. variables that are predictive of a defined outcome (*e.g.* response to exercise training), and also to inform development of important networks that infer such outcomes; experimental data and other source of biological knowledge may also be useful in refining these representation of complex interactions. Such networks may in turn aid biomarker discovery, but are an essential precursor to computational models that are able to explore underlying molecular mechanisms; again, knowledge of specific biological issues may help in their refinement. Finally, incorporation of these models into larger scale analyses offer the potential for *in silico* experimentation, whereby *e.g.* the effect of different therapeutic interventions on disease outcome may be tested.

The first approach is *biomarker discovery* (Figure 2-1A), which perhaps is most widely used in the analysis of functional genomics datasets. Here the objective is to identify measurable variables that are predictive of a given outcome (*e.g.* the response to physical training in a population of individuals). Such measurements can be molecular (*e.g.* gene expression, protein levels, metabolite concentrations, genetic variants) and/or more traditional physiological endpoints (*e.g.* endurance,  $VO_{2max}$ ).

The identification of predictive biomarkers can be achieved by use of univariate and multivariate variable selection strategies that aim to identify the most relevant explanatory measurement(s), while developing a computational model that can accurately predict an outcome [99]. Univariate methods will test every variable (*e.g.* expression of a given gene) on its own, whereas multivariate methods test combinations of variables for their ability to explain a given outcome. Clearly, multivariate approaches better resemble the complex nature of biological networks, and therefore are more likely to provide insights into the mechanisms underlying a complex phenotypic trait. Consistent with this notion, multi-gene biomarkers are often required for robust predictions in independent datasets.

The second approach (Figure 2-1B) consists of ‘reverse engineering’ biomolecular networks from observational data (*i.e.* infer regulatory interactions between quantified biomolecules based on mathematical principles). Here the overall aim is to reconstruct the underlying structure of interactions between biological molecules profiled using *omics* tools (ideally from multiple data sources) and rigorous statistics. Such a *network inference framework* can be achieved by applying a multitude of approaches with varying underlying data assumptions and modelling principles, including ordinary differential-

equation (ODE)-based methods [100], probabilistic modelling techniques (*e.g.* Bayesian theory models) [101, 102], state-space representation models [103], and correlation-based methods. Note, while the first three approaches are able to infer *directed* networks, their capability is currently limited to inferring smaller networks with few variables due to increased computational complexity than possible with correlation approaches.

Importantly, this network inference part may potentially benefit from a biomarker discovery phase, since it has been shown that identified predictive variables are more likely to be directly controlling important physiological processes, and therefore are good candidates to include in a network [95]. Similarly, whole networks can be used as an input for biomarker discovery procedures. It has been shown that often the overall ‘activity’ of a biological network (*e.g.* a specific signalling pathway) is a better predictor than a few key individual genes, proteins and/or metabolites. This implies that in the coming years predictive biomarkers are more likely to consist of a relatively large panel of measurements, possibly spanning multiple levels of complexity within a pathway. Current *omics* platforms are experiencing a rapid development as well as drop in costs, making routine collection of large datasets a feasible option. Once a robust biological network has been inferred this may serve as a good basis for developing a more *conventional modelling approach* to provide explanations for observed phenomena that requires a mechanistic understanding of the system (Figure 2-1C).

Finally, multiple computational models that initially were developed independently can be integrated into a larger and more complex models, which allow responses to physiological/pathological challenges to be simulated, thus integrating effects across multiple organs and/or pathways. These complex models are often referred to as *decision*

*support systems* because of their potential to provide information about the expected outcome of a therapeutic intervention (Figure 2-1D).

Several large international projects aiming at the development of such technology into Systems Medicine integrated frameworks have been established so far, e.g. the Virtual Physiological Human (VPH) project funded by the European Commission 7<sup>th</sup> Framework Programme, which aims to aid clinically relevant research by establishing a framework for handling and integrating various mechanistic models spanning different levels of organizational complexity (ranging from molecular components to organ function). By unifying the modelling languages employed across the different mathematical models included, parameters of a particular model in the hierarchy can be processed by other appropriate models at a lower hierarchical level. These global initiatives should be considered long-term goals, aiming at understanding human physiology quantitatively as a dynamic system.

Developing a comprehensive model of a biological system requires integrating mechanistic and probabilistic inferences. The mathematics for performing such a task is in its infancy, and more development is needed. However, a successful example is illustrated by the anatomically based model of human heart ventricles [98]. In the following sections we aim to provide an overview of some of the methodologies that can be used to infer biomolecular networks, as well as introduce one particular approach we have found useful in our research.

### 2.2.5 Inference of biological networks from observational data

Reverse engineering is an evolving field within network-based Systems Biology. The rapid accumulation of *omics* data in the post-genomic era has made it possible to infer (*aka* ‘reverse engineer’) models of cellular systems with the overall aim of deducing the regulatory structure at a sub-cellular level. Most of the network-based approaches that have been developed are in fact general and can be applied to any type of experimental data. However, because the mRNA expression profiling technology is the most mature *omics* discipline, most applications have been developed to reconstruct *transcriptional* networks (*i.e.* decode the mechanisms of transcriptional control). However, recently it has become apparent that, irrespective of the methodology used to generate data, in order to be able to recapitulate the complex behaviour of a biological system it is essential to integrate multiple types and scales of experimental data (*e.g.* transcriptomic, proteomic, metabolomic).

### 2.2.6 Static vs. dynamic networks

Biological networks can be reconstructed from two different types of experimental studies: either cross-sectional, *e.g.* representing a population of individuals at a given time (*i.e.* steady-state measurements following an experimental perturbation), or prospective, where the experimental data is available across a defined time-course. In reverse engineering, statistical inference of biological causality is an important goal [104]. A simple example of causality could, for example, be a transcription factor regulating the expression of several target genes. Since determining cause and effects implies a direction (*i.e.* the cause precedes the effect), inference of causality from cross-sectional studies presents a challenge due to their static nature, one that is less difficult

when a time-course is available. However, it must be stressed that both approaches are often used in combination to, for example, integrate clinical cross-sectional studies (thereby providing the researcher with a static network representation) and experimental intervention studies that can provide dynamic (prospective) models of the process being studied. At present, most of the developed techniques infer regulatory networks *without* any causality information (likely due to the scarcity of time-course datasets due to their higher costs). However, a small number of causality detection techniques have been proposed in the literature such as dynamic Bayesian networks [105] and Granger causality [106]. It is also important to point out that true time-course datasets can only be developed when the sequence of events is measured within the same cells/tissues. This is for example achieved with imaging techniques that require complex molecular probes, and can typically be only applied to measure a relatively small number of system components [107]. *Omic*s technologies unfortunately are disruptive, so time-course data derived using these approaches are in fact a sequence of independent snapshots, which clearly limits the potential use of dynamical modelling tools.

### **2.2.7 A primer for network inference methods**

The simplest method for inferring statistical relationships between experimental variables is computing the pairwise correlation coefficient across a large collection of heterogeneous samples [108]. Usually such an approach is not able to identify complex non-linear dependencies, and does not discriminate between direct and indirect connections. More complex methods, such as the mutual information (MI) based *Algorithm for the Reconstruction of Accurate Cellular Networks* (ARACNE) [109], also aim at establishing a statistical relationship between pairs of variables but have a stronger

theoretical foundation. Because of the added mathematical complexity they can capture a broader range of biologically relevant dependencies between variables including non-linear, non-monotonic relationships; importantly, they can distinguish between direct and indirect relationships. ARACNE is a free tool for which a Java-based graphical user interface (GUI) exists; hence investigators do not need any programming skills in order to use the software.

ARACNE relies on estimating the probability that a variable (*e.g.* the expression of a gene or a protein) assumes a certain ‘state’ (*i.e.* abundance) given the state of another biomolecule (conditional probability). A number of alternative MI-based implementations have been proposed during the last decade (*e.g.* *Context Likelihood Relatedness* (CLR) [110], *Minimum Redundancy/Maximum Relevance Networks* (MRNET) [111]), which mainly differ by the way inferred *indirect* relationships (so-called ‘edges’) are removed once the dependencies between all pairs of variables have been mathematically formulated. In such analyses, unwanted indirect interactions occur by default if there is strong correlation between biomolecule 1 and biomolecule 2, and between biomolecule 1 and biomolecule 3 in a three-node clique (*i.e.* a triplet of connected variables).

An MI value of zero means that there is no dependency (*i.e.* no information flow) between two variables, whereas an MI value of 1 indicates a perfect association between them, and therefore, a likely strong regulatory interaction between them. For each inferred dependency, a *P*-value is calculated based on the distribution of MI values between random permutations of the original dataset, thereby allowing the elimination of all non-statistically relevant dependencies by thresholding using an appropriate (user-

defined) cut-off level. Importantly, the quality of the inferred interaction network depends on the arbitrarily selected probability cut-off. A small threshold (e.g.  $P=0.05$ ) gives a high recall (*i.e.* fraction of true dependencies that could be inferred) but low precision, whereas a higher threshold (e.g.  $P=10^{-6}$ ) yields better precision (*i.e.* fraction of inferred dependencies that really are in the network) while suffering from a low recall. A further advantage of MI as an information-theoretical measure of dependency between variables concerns its relatively low computational requirements for building an interaction network. Hence, MI is able to handle very large data matrices with thousands of experimental variables, whereas most of the other more advanced techniques mentioned (*e.g.* Bayesian methods) can only deal with much smaller numbers of variables (<100) because of the high computational complexity. However, in order to infer robust statistical associations based on MI a fairly large sample size is required (> 50-100 biological replicates), due to the required estimation of the (joint) frequency distribution of the connectivity. Interaction networks derived from such reverse engineering methodologies can be visualized and further analysed using various freeware software tools such as Cytoscape [112], Pajak [113], and BioLayout [114]. A comprehensive list of visualization tools focused on interaction networks and their web-links has recently been reviewed [115].

Up to now, these information-theoretic approaches have usually been employed on gene expression data only, due to the wealth of such data available. However, as physiologists have known for many decades, biological systems are usually more complex and multi-layered. Indeed, despite some popularist science writing to the contrary, genes on their own are merely permissive elements within biological systems

[26]. Further, it has been shown that when multiple types of data (*e.g.* copy number variants, protein or microRNA expression levels) are incorporated in the network inference pipeline, the accuracy of the learned network topology increases [116]. Hence, at present there is a call for methodologies that can embed multiple data sources in a single computational framework. Our recent work has focused on methods that are able to handle large-scale, multi-dimensional genomic datasets [117, 118].

### **Topological analysis of inferred biological networks provides useful biological insight**

Up to now we have described some of the most widely used methodologies for inferring regulatory networks. However, an immediate challenge arises in interpreting these often large, complex networks that visually present as a ‘hairball’ (*i.e.* too dense a collection of connections to comprehend as a whole) [119]. A simple solution, although not very objective, is to focus the analysis around a favourite gene(s). In this scenario, the investigator typically examines the manually selected sub-network in order to identify unknown or unexpected biological relationships, which in turn may be used to formulate new hypotheses. Such ‘discovery-led’ science may be useful when there is insufficient data to justify an hypothesis-driven approach.

Alternatively, the topological properties of the network can be used to identify interesting genes and sub-networks that can be interpreted. We and others have demonstrated the existence of a higher-level, modular organization in biological networks [95, 120, 121], *i.e.* components of biological systems that act in collaboration to carry out specific biological processes. Consequently, several modularization approaches have now

been developed to help group subsets of cellular components based on a given property, such as topological structure or functional role. Such decomposition of a large complex network into relatively independent sub-networks (or ‘modules’) has been shown to be an effective way to deduce the underlying structure of the fully connected network containing many hundred variables (so-called ‘nodes’), as each module can then be analysed independently. In addition, studies have demonstrated that such identified network modules can serve as better predictors of a physiological response than the classic biomarker discovery approach (see Figure 2-1).

In biomolecular interaction networks, as well as sub-networks, nodes have different levels of connectivity (*i.e.* number of interactions with other nodes). It has been shown that such interaction networks have so-called ‘scale-free’ structure properties, as their node connectivity distribution fits a power law [122]. Such a power law degree distribution implies that most of the connections between biomolecules is linked to a small number of highly connected nodes, such that a large proportion of the molecular state of a cell can be explained by a small subset of biomolecules (so-called ‘hub’ nodes; *e.g.* a transcription factor that regulates many more genes than average). Hence, in biological networks a hub is often assumed to be a key component of a regulatory networks, hence important for the function of a cell/tissue under investigation. This assumption is supported by the fact that random node disruption does not significantly affect the network architecture, whereas deletion of hub nodes leads to a complete breakdown of the network structure [123]. Hence, adjusting the spatial position of each node according to its interconnectivity has been shown to be a simple, yet effective way of visualizing large complex interaction networks [124].

More advanced methods to extract information from complex networks exist that aim to identify *functional* modules (*i.e.* sub-networks of biomolecules that are linked to the same biological function), *e.g.* by integrating both physical interactions (*i.e.* experimentally validated protein-protein interactions) and mRNA expression data [125]. In this context, an identified functional module represents a putative multi-protein complex that is transcriptionally regulated in a specific experimental condition (*e.g.* treatment *vs.* control). Hence, by considering additional data on a different level of organization, one can potentially infer a clearer composite picture of the underlying biological function.

Finally, in order to generate objective hypotheses about biological processes controlled by a specific hub node or sub-network, functional enrichment analysis can be performed on all its direct neighbours (*i.e.* all the adjacent nodes that are directly connected to the hub) [126]. Such enrichment analysis aims at reducing complexity by defining groups of molecules (represented by gene sets) that share similar biological functions (*e.g.* a class of adhesion molecules). To accommodate latest advances in knowledge, the different annotation databases used for this purpose (*e.g.* Gene Ontology [127] and KEGG [128]) are frequently updated by curators. Using software tools like the web-based application DAVID [129] or applications such as BiNGO [130] developed specifically for use with software visualization tools like Cytoscape, one can quickly determine whether any gene sets are statistically over-represented, thus generating hypotheses on the biological processes controlled by those factors outlined above.

### **2.3 CASE STUDY: INFERENCE OF OXYGEN-DEPENDENT PATHWAYS IN LIMB SKELETAL MUSCLE**

The main purpose of this case study is to illustrate in a step-by-step manner the application of reverse engineering to integrate supra-cellular physiological measures and genome-wide expression profiling. From a more biological perspective we aim to identify a clinically relevant signature of skeletal muscle tissue hypoxia.

This analysis uses two different datasets. The first is a publicly available dataset (GSE27536) representing a cohort of COPD patients and healthy controls matched for age and smoking history [131] (see Table 2-1 for subject characteristics), which includes gene expression profiling in *vastus lateralis* limb muscle and whole-body physiological variables (*e.g.*  $VO_{2max}$ , minute ventilation,  $PaO_2$ ) [132][133].

	Healthy controls	COPD, BMI <sub>norm.</sub>	COPD, BMI <sub>low</sub>
Gender (M/F)	10/2	9/0	6/0
Age (years)	65.3±2.9	69.4±1.5	69.2±4.6
BMI (kg/m <sup>2</sup> )	26.3±1.1	27.4±1.4	19.7±1.0 <sup>**,††</sup>
FFMI (kg/m <sup>2</sup> )	21.0±0.8	21.5±0.7	16.7±0.9 <sup>**,††</sup>
VE (L/min)	71.2±5.6	40.5±3.6 <sup>***</sup>	33.0±3.8 <sup>***</sup>
FEV <sub>1</sub> (L)	3.46±0.2	1.41±0.09 <sup>***</sup>	1.21±0.21 <sup>***</sup>
FEV <sub>1</sub> /FVC (%)	75.9±2.4	44.0±2.7 <sup>***</sup>	39.5±4.5 <sup>***</sup>
RV (% of pred.)	103.9±5.2	145.0±13.3	160.0±28.6 <sup>*</sup>
VO <sub>2max</sub> (l•min <sup>-1</sup> •kg <sup>-1</sup> )	22.3±1.4	13.9±1.7 <sup>**</sup>	14.4±1.5 <sup>**</sup>
Peak power (W)	117±8	60±7 <sup>***</sup>	47±9 <sup>***</sup>
6MWD (m)	584±24	469±30 <sup>*</sup>	367±59 <sup>***</sup>
BODE index	0.1±0.1	2.3±0.4 <sup>**</sup>	4.0±1.0 <sup>***,†</sup>
Data are presented as mean±SEM. *: $p < 0.05$ ; **: $p < 0.01$ ; *** $p < 0.001$ versus controls. †: $p < 0.05$ ; †† $p < 0.01$ ; ††† $p < 0.001$ versus COPD patients with a normal BMI. Comparisons were analysed using one-way ANOVA and Tukey's <i>post hoc</i> test. Abbreviations: BMI: body mass index; FFMI: fat-free mass index; VE: lung ventilation; FEV <sub>1</sub> : forced expiratory volume in 1 sec; 6MWD: 6-min walking distance;			

**Table 2-1 | Anthropometric characteristics defining the COPD cohort used in the case study.**

The second dataset represents an unpublished, genome-wide transcriptional response of mouse *soleus* muscle to a gradual decline in atmospheric oxygen concentration (GSE64076; data will be released upon publication).

Using the first dataset, representing the basal transcriptional state of skeletal muscle in a COPD cohort (Figure 2-2), I first show how to infer statistical connections between oxygen availability (e.g. VO<sub>2max</sub>), oxidative stress (protein carbonylation) and gene expression signatures (Figure 2-2A-C).

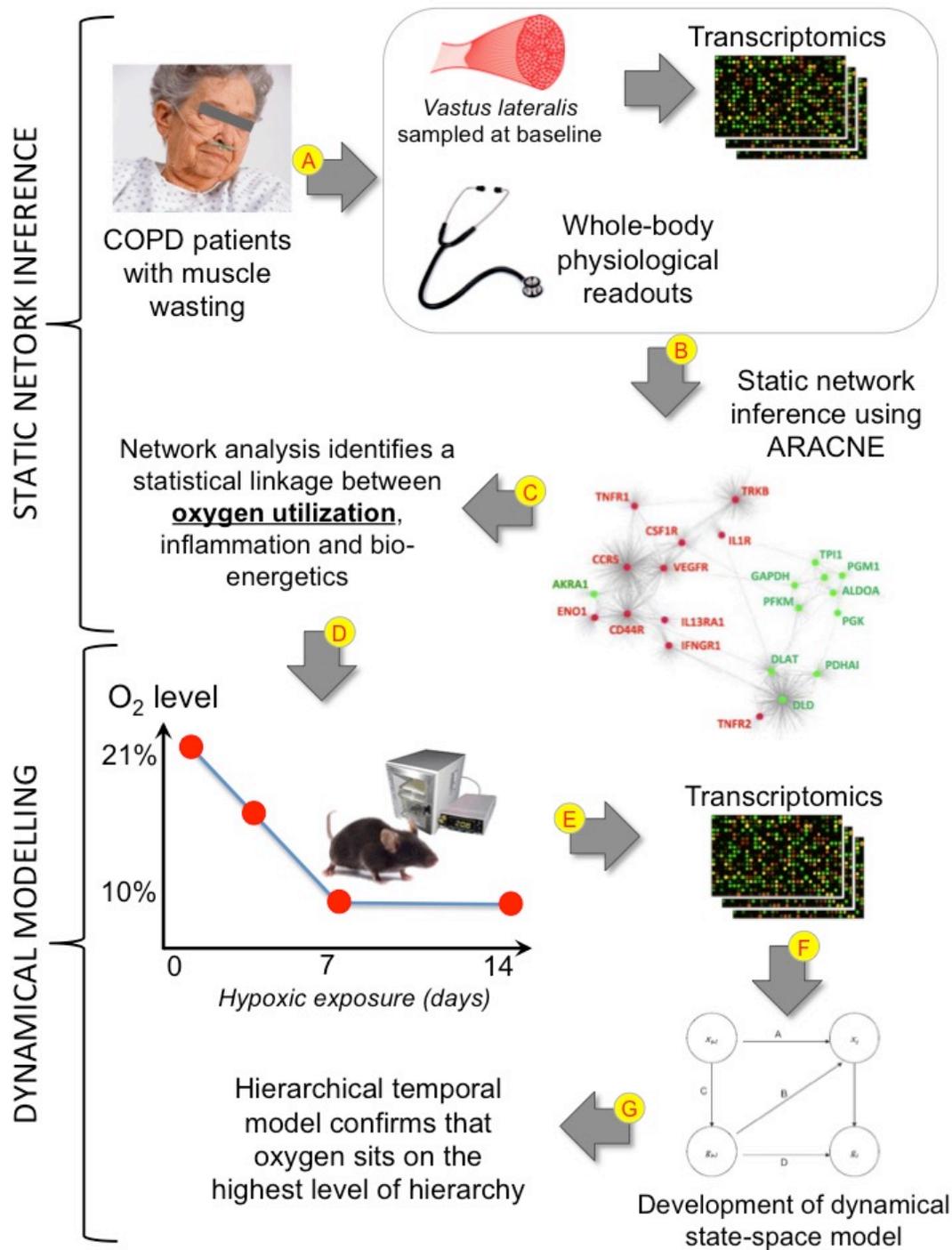


Figure 2-2 | Schematic representation of the analysis strategy used in the case study, highlighting how the inferred static multi-scale network from the clinical COPD cohort (Fig. 2-2A-C) can be bridged to the inference of a dynamical network representing the temporal progression of events following an experimental challenge (hypoxic exposure) in a murine animal model (Fig 2-2D-G).

[figure legend continues on next page]

Having identified a clinical condition with known outcome (exercise intolerance in patients with respiratory disease), we could target unknown mechanisms by focussing on one likely source of functional limitation (skeletal muscle dysfunction  $\pm$  central limitation on O<sub>2</sub> supply), and generate data characterising the phenotype. Both genomic and physiological readouts were used to construct a network of inferred interactions, which was then interrogated to identify statistically robust linkages among broad biological functions. While very useful in providing a list of useful biomarkers, there remains a potential limitation with single point associations. The dynamic nature of relationships is captured by repeated measures across a suitable time scale (which will vary for different molecular, physiological and structural responses) using an animal model of respiratory distress, where the transcriptome-based model demonstrated the central importance of oxygen in the response.

Having defined an oxygen-related signature in the disease setting, we then transpose these findings in a mouse model of gradual hypoxia (second dataset, Figure 2-2D-E). Here we use a different computational approach to develop a hierarchical dynamical model explaining the transcriptional response of oxidative leg muscles to a prolonged gradual reduction in blood oxygenation (hypoxaemia) (Figure 2-2F-G). The model we describe below validates the notion that the signature identified using the clinical study may be truly triggered by changes in oxygen availability. Moreover, the model contributes to the understanding of the transient events following oxygen depletion that cannot be observed using a cross-sectional clinical study.

### **2.3.1 Step 1. Linking physiological measurements and gene expression data in the COPD cohort**

In order to reconstruct an interaction network spanning multiple levels of organization, we have utilised the following strategy that the Falciani lab developed earlier [133].

## 1. Combining measurements from different data sources

In order to combine gene expression data with whole-body physiological readouts, all variables need to have the same units of measurement (as the range of *e.g.* VEGF mRNA expression values is very different from that of  $VO_{2\max}$ ). All such raw scale units can be unified by simply ‘transforming’ each experimental variable to have the same dynamic range, *e.g.* this can be achieved by standardising measurements across samples to have a mean of 0 with a standard deviation of 1. Such an established approach, called z-scoring, enables us to treat the physiological indicators as individual ‘nodes’ in the inferred interaction network with states (just as each gene on the array is treated).

### *Definition of a biological framework for data-driven network inference*

The outcome of data-driven reverse engineering of biological networks, in the absence of any biological assumption(s), often provides results that are difficult to interpret due to the large number of inferred significant interactions. Thus, to reduce complexity of the problem, we decided to focus the analysis on the set of physiological parameters and genes encoding for enzymes in the central bioenergetic pathways (*i.e.* TCA cycle, oxidative phosphorylation, glycolysis). The latter choice is reasonable considering the paramount importance of these molecular pathways in skeletal muscle adaptation. The overall strategy is therefore to identify biomolecules that are highly correlated (based on MI) with biologically important experimental variables. Such a focused analysis will generate multiple network modules of interacting biomolecules, each with a bioenergetic hub gene or physiological measurement at its centre. Two modules will be linked together if a specific gene is statistically linked to both hubs.

## 2. Reverse engineering.

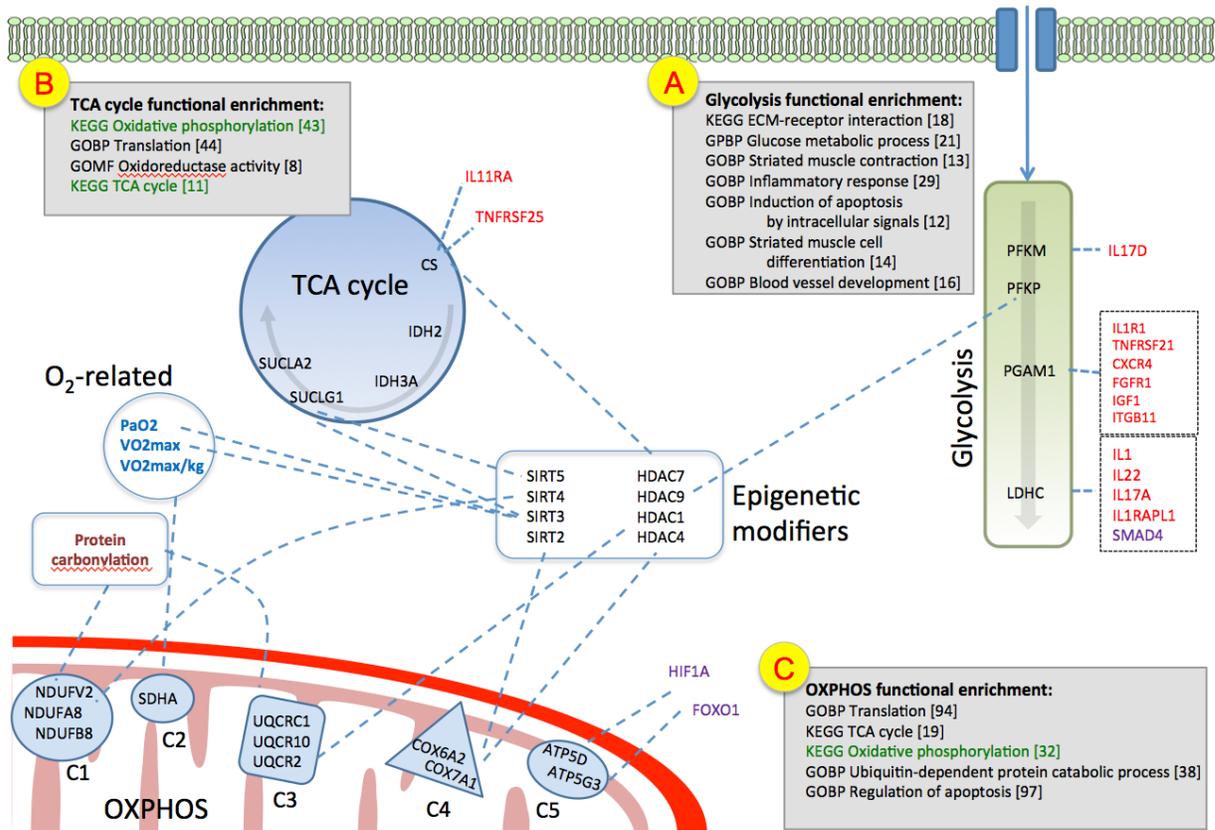
In order to infer robust regulatory relationships between variables in the integrated multi-level dataset, we used the ARACNE algorithm. This choice was based on the large number of measured variables to be considered by the mathematical framework. By combining all genes expressed in human skeletal muscle (>10,000 mRNAs) with the list of physiological variables we far exceed the number of variables that can be handled by more advanced network inference methods (*e.g.* Bayesian methods). Hence, we infer a static network without any obvious hierarchical organization. The result of an ARACNE run is an ‘adjacency matrix’ containing MI values for all pairwise interactions above the specified MI threshold, which can be visualized automatically in Cytoscape.

After calculating MI-based dependencies between all the different variables in our multi-level data matrix, all those inferred regulatory interactions with an MI value below 0.22 (corresponding to a  $P$ -value cut-off of  $10^{-6}$ ) were removed. Such filtering of weaker statistical dependencies is an important step in the generation of a more sparse interaction network, which can more easily be interpreted by the investigator. The stringent  $P$ -value cut-off means that the remaining associations have been inferred with high precision at the cost of a lower recall rate.

## 3. Network visualization

Data visualized as a network are often easier to interpret than long lists of biomolecules and their associated statistical dependencies. Hence, the numeric output of ARACNE, which contains MI values for all pairwise associations, was imported into Cytoscape for

visualization, a conventional way of analysing interaction networks. Briefly, we reconstruct the network neighbourhood of each of the bioenergetic ‘seed’ genes (*i.e.* all variables directly connected to them). The neighbouring variables can either be genes expressed in skeletal muscle and/or physiological variables. Figure 2-3 summaries key regulatory associations (based on MI) between this seed set of genes and their immediate neighbours.



**Figure 2-3 | Graphical representation highlighting putative regulatory associations (significant correlation between two factors is shown as a dotted line) that likely represent robust interactions, based on high mutual information values.**

The focus is on central metabolism pathways (i.e. glycolysis, TCA and OXPHOS, respectively) and their immediate neighbours. The grey boxes define functional enrichment of the different bioenergetic compartments based on direct neighbours. Individual genes of relevance are grouped into modules with others of related function, as are physiological readouts that may be treated in a similar manner for statistical analysis. C1-5: the different complexes in the electron transport chain. The value of such an approach is in providing a detailed overview of a complex interaction network, reducing the huge number of potential factors into groups of defined function, and offering a limited number of candidates whose utility as biomarkers or therapeutic targets may be experimentally verified.

#### 4. Functional analysis of the network hubs

We further explored whether the direct interacting neighbours of each central metabolism pathway mapped to functional categories (*i.e.* GO terms) as well as KEGG pathways. Notably, a marked enrichment of the different bioenergetic compartments was observed (Figure 2-3, boxes A-C) that clearly highlights the interconnected nature of the bioenergetic machinery, *i.e.* functionally related genes appear to be co-expressed.

#### 5. Biological interpretation

The most important finding of the current analysis is that among the direct neighbours to each bioenergetic pathway, particularly the two oxidative ones, we noted a statistical over-representation of genes encoding histone deacetylase enzymes (*i.e.* HDAC and SIRT mRNAs). This observation is consistent with previous studies that have highlighted the importance of sirtuins in regulating metabolism [134–136]. Further, the protein deacetylase *SIRT3* that primarily is localized in the mitochondrial matrix was also significantly positively correlated to both arterial oxygen tension (PaO<sub>2</sub>) and oxygen uptake (VO<sub>2max</sub>). In support of deacetylation being an important control point, it was recently shown that *Sirt3* knockout mice exhibit decreased oxygen consumption, thus affecting cellular respiration [135]. Hence, besides the obvious oxygen-driven effect on aerobic pathways (as indirect measures of oxygen availability such as VO<sub>2max</sub> are linked to key genes in oxidative phosphorylation), the present network-based Systems Biology approach points to tissue hypoxia as being a potential important player in modifying expression of deacetylase modifying enzymes in severe COPD patients with a muscle wasting phenotype. Our Systems Biology approach also negatively links protein

carbonylation (an established proxy measure for oxidative stress; [137]) to Complex 1 and 3 in the electron transport chain (Figure 2-3, bottom left part). The validity of such an association is further strengthened *via* functional enrichment analysis using DAVID, as a significant fraction of direct neighbouring genes to protein carbonylation is statistically associated to gene ontology (GO) terms representing cellular respiration.

If we then focus on the genes in the glycolytic pathway (Figure 2-3, top right part), a high proportion of pro-inflammatory mediators/receptors (e.g. *IL1B*, *IL1RI* and *TNFRSF21*) are among the direct neighbours, as indicated by the enrichment of the ‘inflammatory response’ GO term (Figure 2-3, box A). Hence, hypoxia is pro-inflammatory, as seen by more traditional observation methods [138].

Multi-scale network inference approaches, similar to that illustrated in Figure 3, have proven very effective in generating robust hypotheses (e.g. 45). However, statistical associations may not represent causality, particularly when the inferred associations stem from steady-state measures. Thus, in order to validate our hypothesis that varying oxygen levels (represented by  $VO_{2max}$  and  $PaO_2$ ) control the expression of epigenetic modifiers, we used a more sophisticated network inference algorithm that can learn the structure of networks from time-course data. We applied this dynamic inference approach to a murine model of hypoxia (Step 2).

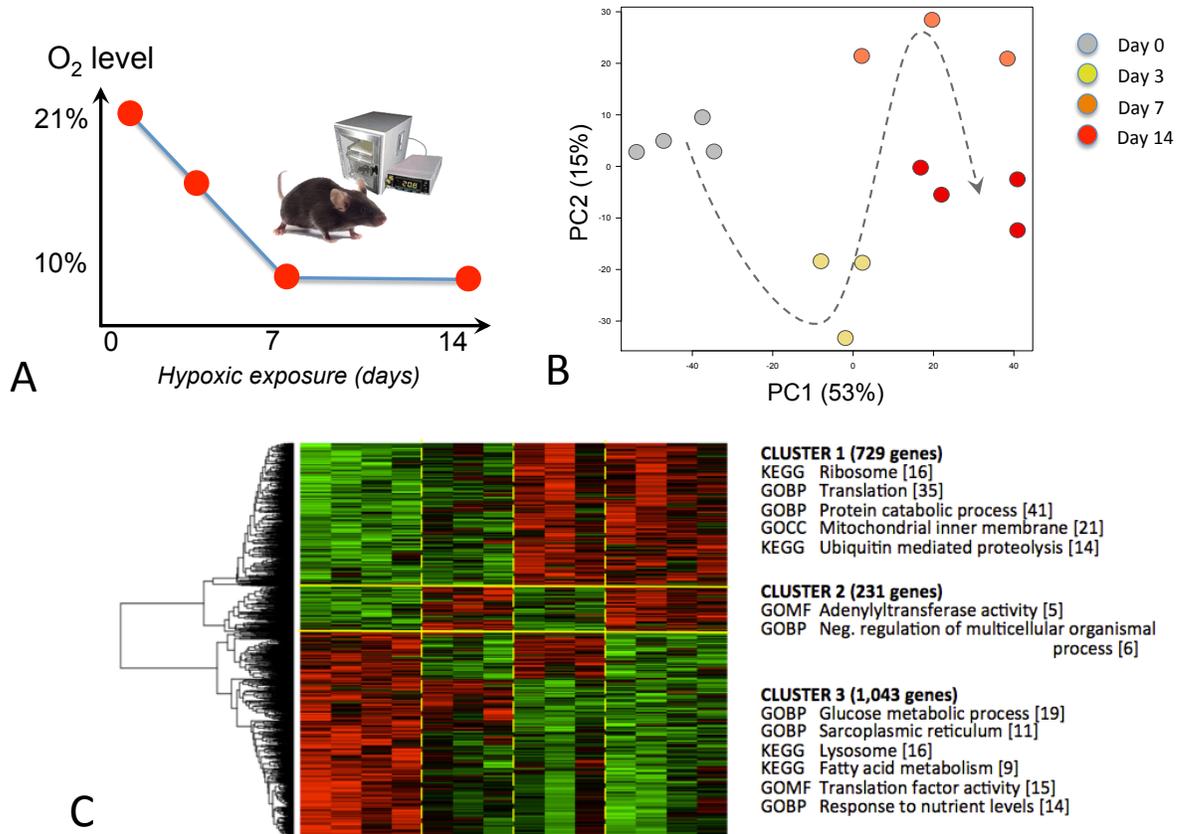
### **2.3.2 Step 2. Gene expression dynamics in response to tissue hypoxia**

Animal models are commonly used for studying the *in vivo* effects of hypoxia, for ethical reasons, where severe or prolonged hypoxaemia is induced and invasive samples are required to explore mechanisms. Importantly, hindlimb skeletal muscles have been

reported to alter metabolic phenotype and reduce fibre size in response to prolonged hypoxic stress in mice [139, 140], highlighting their potential relevance as a pre-clinical model of muscle wasting in COPD patients. In order to experimentally test the hypothesis derived from the clinical COPD network presented in Figure 2-3, we therefore exposed adult male C57/Bl6 mice to chronic systemic hypoxia for up to 2 weeks, in order to simulate levels of hypoxaemia reported in COPD patients with advanced respiratory insufficiency. To capture the temporal effect of reduced oxygen tension on gene regulation, we sampled and gene profiled the soleus muscle (n=4) at 3 different time-points (day 3, 7 and 14) following initiation of the gradual hypoxic insult (*i.e.* the O<sub>2</sub> level was gradually lowered to 10% over the first week and kept stable during the second week) (Figure 2-2, bottom part).

First, a high-level representation of the temporal transcriptional changes was performed using a variable reduction technique called principal component analysis (PCA) (Figure 2-4B). When plotting replicates of two variables against each other, it is relatively easy to see which is a better discriminating factor; visual inspection becomes increasingly difficult as the number of variables increase, hence the need for PCA. In essence, this method aims at ‘tilting’ the axes through the multidimensional data space, such that the first principal component accounts for as much of the variation in the original dataset as possible (the assumption is that the most important dynamics in the dataset are the ones with the largest variation). Our PCA revealed that the early dynamics of hypoxia is captured by the first principal component whereas the 2<sup>nd</sup> most important principal component (in terms of variance captured) separated the later time-points. Further, functional enrichment analysis of the differentially expressed genes (ANOVA,

$P < 0.05$ ) using DAVID (Figure 2-4A), highlighted several important pathways/ontologies. Most striking was the enrichment of protein catabolic process and ubiquitin-mediated proteolysis among genes up-regulated at day 7 and 14, clearly suggestive of a transcriptionally regulated muscle wasting phenotype driven by the experimentally induced hypoxaemic state.



**Figure 2-4 | High-level representation of temporal transcriptional changes in the murine model of hypoxia.**

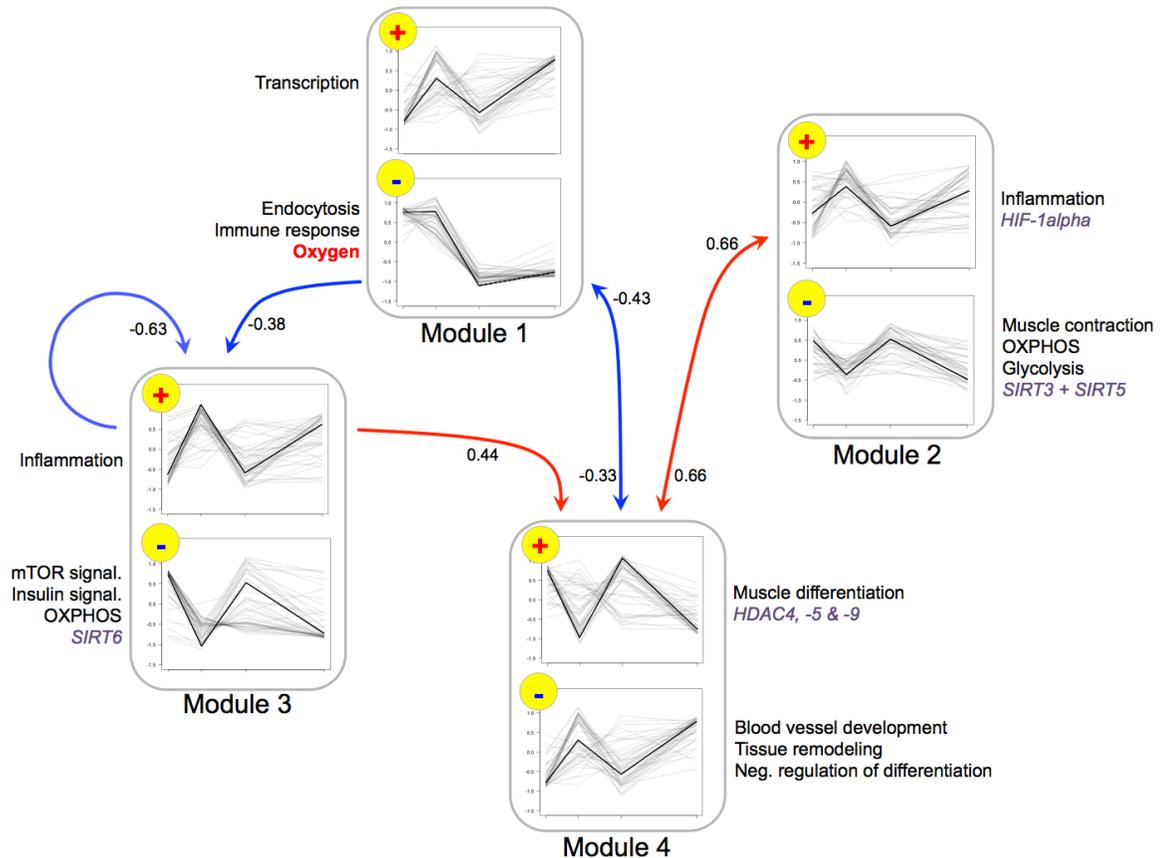
**A)** Graphical representation of the pre-clinical experimental design. The O<sub>2</sub> level was gradually decreased from 21% to 10% during the first week and mice were housed for another week at this O<sub>2</sub> concentration. **B)** PCA plot highlighting the transcriptional dynamics caused by the hypoxic challenge. **C)** Hierarchical clustering using mRNA expression levels of genes modulated by hypoxia ( $P < 0.05$ ). Each row represents a transcript and each column represents a sample. Red and green colours indicate expression levels above and below the median value of the distribution of signal, respectively. Using solid yellow lines we have subdivided genes into overall trends in order to help the reader. Enriched functional terms within these are listed next to the heatmap.

State space models (SSMs) are a class of probabilistic graphical models. SSM provides a general framework for analyzing deterministic and stochastic dynamical systems that can be measured/observed through a stochastic process. The SSM framework has been successfully used for the analysis of gene expression data [103, 141]. In its simpler application the model formalises the effect of hidden, unmeasurable factors in specifying observed gene expression changes over time. The inclusion of these hidden factors is important since we cannot hope to measure all possible factors contributing to genetic regulatory interactions (*e.g.* levels of regulatory proteins as well as effects of mRNA and protein degradation).

The next step was to apply state-space modelling to reverse engineer transcriptional network modules (*i.e.* representing discrete temporal dimensions) from our replicated murine time-course dataset. Such module-based reduction in complexity allows analysis of hundreds or even thousands of genes, as those with a similar temporal expression profile are aggregated into a transcriptional module. To allow construction of a near genome-level model, we took advantage of a newer approach that incorporates this concept of modularization [103].

A SSM can reconstruct the topology of a network representing the systems dynamics, despite a relatively small number of time-points, by using biological replicates for each time-point [103]. In order to reduce complexity, variables that do not change significantly are excluded from the modelling process. In this case study, genes deemed to be significant by ANOVA at a 1% significance level, as well as all hub genes, were included (931 variables in total). The hub genes were chosen to represent the different

components in our interpretative model derived from the clinical COPD dataset (Figure 2-3). Finally, the experimentally set oxygen level was used as an independent variable.



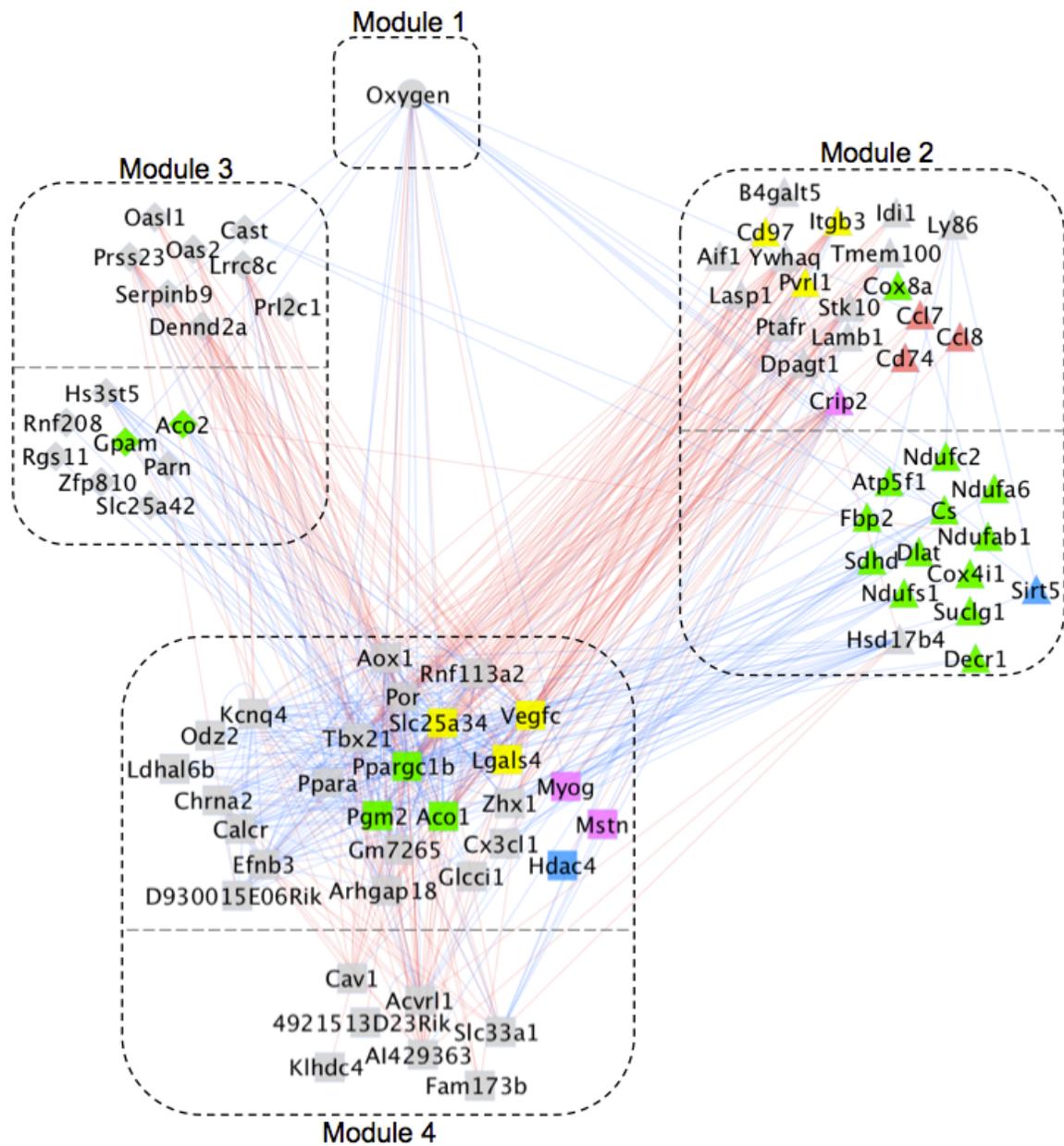
**Figure 2-5 | The hierarchical dynamic state-space model identified 4 modules (x-axes define length of hypoxic exposure), each characterised by two separate transcriptional profiles: plus and minus, representing up- and down-regulation, respectively.**

The hierarchical position of the modules represents the estimated temporal structure of the network. Functionally enriched GO terms (regular text) as well as key genes (italics) are identified next to the relevant module. Blue arrows represent temporal repression whereas red arrows represent temporal induction. The numeric value next to each arrow represents the estimated coefficient.

Based on unsupervised clustering using HOPACH within the software programming environment R [142], we identified 8 distinct gene clusters with similar expression profiles. Hence, to model the effect of hypoxaemia on the skeletal muscle

transcriptome the hidden state dimension was set to 4, as each inferred module contains both a positive (+) and a negative (-) component.

The hierarchical dynamic model in 4 temporal dimensions shows that modules 1 and 2, which sit on the highest level of hierarchy (*i.e.* precede others in time), were enriched in GO terms related to muscle contraction, bioenergetic pathways, and inflammation among others (Figure 2-5). Interestingly, the experimental oxygen concentration was represented in module 1(-) whereas two deacetylases *SIRT3* and *SIRT5* were found in module 2(-). A negative influence is observed of module 1 on module 3, which is located further down the temporal hierarchy. Module 3(+) is highly enriched in inflammatory processes whereas its negative counterpart mainly represents two key signalling pathways (mTOR and insulin). At the lowest temporal level we find module 4, which is enriched in GO terms related to muscle differentiation, tissue remodelling and blood vessel development. Interestingly, three HDACs are represented in module 4(+) (Figure 2-5). Figure 2-6 represents a more focused version of Figure 2-5, highlighting the most significant interactions between components in the four inferred modules from Figure 2-5.



**Figure 2-6 | A higher resolution representation of Figure 2-5, highlighting the most significant gene interactions between components in the four inferred modules.**

Lines represent factor interactions based on mutual information (blue represents temporal repression, red represents temporal induction). Genes are colour coded for broad functional categories (red=cytokines; blue=epigenetic modifiers; green=aerobic metabolism; purple=muscle differentiation; yellow=cell-interaction).

We therefore conclude that the inferred dynamic model using a state space modelling approach appropriately recapitulates the interpretative model advanced in Figure 2-3. In addition, it identifies oxygen at the highest level of hierarchy, whereas key effector functions controlled by oxygen such as inflammation and muscle differentiation are downstream in the temporal hierarchy.

## **2.4 CONCLUSIONS**

The aim of this brief review is to provide an intuitive overview on data-driven ‘learning’ of biological pathways, linking molecular and physiological readouts. We used a case study to make it easier for experimental biologists to see the potential of computational biology to provide interpretative models of complex patterns, and stress that the identification of general properties of a system from a genome wide analysis of a molecular state of a system is a very powerful approach. We also explain that it is complementary, rather than an alternative to hypothesis-driven science. The importance of a robust characterization of the physiological parameters of the biological system, which must be integrated in the model to provide useful testable hypothesis, is emphasised. In support of this concept, we demonstrate the development of an integrative workflow that incorporates measurements from different levels of cellular and molecular organization using a case study representing muscle wasting in COPD. The outline provides an exemplar where individual steps can be modified according to the type of data at hand and addition data types added. For example, in contrast to established gene expression microarrays, techniques for proteomics and especially metabolomics are still under development. Once it is possible to measure the whole proteome and metabolome

of a sample, systems identification pipelines will clearly benefit from these *omics* techniques.

The specific findings in the case study relate to the definition of an oxygen dependent signature in COPD. Such signature (exemplified in Figure 2-3) is static and entirely based on statistical inference. The model is therefore based only on correlation between a series of patient biopsy snapshots, and therefore does not allow any inference of causality. The use of a mouse model of gradual hypoxia allowed us to demonstrate that a signature inferred from the clinical cohort is indeed modulated by experimental reduction in oxygen levels. Moreover, the development of a mathematical model identifies oxygen as the most upstream event as an emergent property. This may appear an obvious finding but, from a methodological perspective, validates the analytical approach.

The data we have used in this case study is gene expression profiling, and as such is representative of available datasets. This has several limitations. The first is that models including multiple levels in the expression of genetics information (*e.g.* epigenetics, microRNA, proteomics, metabolomics, etc.) may better represent biological complexity. However, current computational methods are inadequate to represent properly the interaction between these levels. Moreover, time course data that rely in disruptive sampling strategies are not true time course experiments. As the new functional genomics technologies develop further, as well as novel approaches to model the interaction between different layer of biological organisation, we expect that the efficacy of data-driven approaches will increase further.

### **3 A systems biology approach reveals a link between systemic cytokines and skeletal muscle energy metabolism in a rodent smoking model and human COPD**

*The work presented in this chapter represents a collaborative project between the labs of Francesco Falciani, Joan Barberá (University of Barcelona) and Constancio Gonzalez (University of Valladolid). The two Spanish labs jointly designed and conducted the in vivo animal exposure experiment. Members of the Falciani lab contributed to the development of the custom microarray platform used in this chapter. I carried out all wet-lab genomic analyses as well as all in silico analyses presented herein. Josep Roca (University of Barcelona) provided the raw Affymetrix microarray data as well as anthropometric characteristics related to the clinical COPD cohort.*

*The work presented in this chapter was published in Genome Medicine in 2014 [131].*

#### **3.1 Abstract**

**Background:** A relatively large percentage of patients with chronic obstructive pulmonary disease (COPD) develop systemic comorbidities that negatively affect prognosis, among which peripheral skeletal muscle wasting is particularly debilitating. Despite significant research effort, the pathophysiology of this important extrapulmonary manifestation is still unclear. A key question that remains unanswered is to what extent systemic inflammatory mediators might play a role in this pathology.

Cigarette smoke (CS) is the main risk factor for developing COPD and therefore animal models chronically exposed to CS have been proposed for mechanistic studies and biomarker discovery. Although mice have been successfully used as a pre-clinical *in vivo* model to study the pulmonary effects of acute and chronic CS exposure, data suggest that mice may be inadequate models for studying the effects of CS on peripheral muscle

function. In contrast, recent findings indicate that the guinea pig (*Cavia porcellus*) may better mimic the wasting process observed in human COPD patients.

**Methods:** I have used a systems biology approach to compare the transcriptional profile of hindlimb skeletal muscles from a guinea pig rodent model exposed to CS and/or chronic hypoxia to COPD patients with limb muscle wasting.

**Results:** I show that guinea pigs exposed to long-term CS accurately reflect most of the transcriptional changes observed in dysfunctional limb muscle of severe COPD patients when compared to matched controls. Using network inference, I then show that the expression profile in whole lung of genes encoding for soluble inflammatory mediators is informative of the molecular state of skeletal muscles in the guinea pig smoking model. Finally, I show that CXCL10 and CXCL9, two of the candidate systemic cytokines identified using this pre-clinical model, are indeed detected at significantly higher levels in serum of COPD patients, and that their serum protein level is inversely correlated with the expression of aerobic energy metabolism genes in skeletal muscle.

**Conclusions:** We conclude that CXCL9 and CXCL10 are promising candidate inflammatory signals linked to the regulation of central metabolism genes in peripheral skeletal muscle. On a methodological level, the current work also shows that a systems level analysis of animal disease models can be very effective at generating clinically relevant hypotheses.

## 3.2 Background

Chronic obstructive pulmonary disease (COPD), one of the top five deadliest diseases worldwide [143], is an inflammatory condition of the lungs that predominantly affects people with a long history of cigarette smoking (CS) [144]. In addition to the well-established clinical manifestations in lungs, COPD is also associated with several extra-pulmonary manifestations. Peripheral skeletal muscle wasting and dysfunction is one of the most severe of these pathologies [21]. This muscular deconditioning of the limbs, which is partly independent of the severity of airflow limitation, is a prominent contributor to exercise intolerance [145] as well as being an independent predictor of morbidity and mortality [146]. Long-term CS exposure has a clear potential to contribute to the systemic effects of COPD, as similar findings have been observed in healthy smokers (*e.g.* a decrease in lean muscle mass and force reduction) [147]. However, the direct effects of CS on peripheral muscle function, at the molecular level, are at present poorly understood.

Although we do not still fully understand the mechanisms that contribute to peripheral muscle dysfunction in COPD, there is evidence that multiple factors are likely to influence clinical outcome, such as systemic inflammation, reduced capillary density, tissue hypoxia and subsequent oxidative stress [148, 149]. However, it is currently not known to what extent increased levels of inflammatory cytokines play a role in muscle wasting.

Likewise, hypoxaemia (abnormally low concentration of oxygen in the bloodstream) has been linked to several drivers of skeletal muscle dysfunction in COPD such as down-

regulation of energy consuming processes (*e.g.* protein synthesis, mitochondrial respiration), impaired adult myogenesis (affecting muscle regeneration capacity), fibre type shifting (slow-to-fast myofiber transition) [150], and increased serum levels of cytokines [151], but the exact molecular mechanisms through which chronic or intermitted hypoxia affects muscle maintenance are currently unclear.

Animal models, particularly mouse models, are widely used to study the effects of acute and chronic CS. Importantly, long-term CS exposure in rodent models may be the best approximation of the more *acute* aspects of lung responses in human COPD [152]. However, the current literature suggests that mice chronically exposed to CS develop either none or only mild peripheral muscle dysfunction [147, 153, 154]. For example, only two studies have to our knowledge reported a significant decrease in hindlimb muscle *weight* following long term whole-body exposure using very high smoking doses ( $\geq 20$  cigarettes/day) [153, 155].

The guinea pig (*Cavia porcellus*) (GP), one of the most popular animal models to study infectious diseases [156], has been shown to tolerate CS exposure *without* the rapid weight loss observed in other pre-clinical models [157]. However, promisingly *long-term* CS-exposed guinea pigs fail to appropriately gain body weight compared to age-matched sham controls [147, 158]. Further, in accordance with previous findings in COPD patients [147], guinea pigs demonstrate CS-induced oxidative stress in limb muscles within 3 months of exposure [147, 158], potentially highlighting the relevance of the GP model for studying extrapulmonary comorbidities that characterise human COPD.

However, the GP is a rather challenging model organism to determine the molecular response due to the lack of a fully annotated genome. This paucity of genetic information is unfortunate since gene expression profiling has shown to be a very promising approach to formulate hypotheses on complex molecular mechanisms underlying pathology [159, 160]. Here, we report the development of the first transcriptome-sequencing for guinea pigs representing lung and limb skeletal muscles, as well as the development and validation of a novel *genome-wide* microarray platform. With this novel platform, we profiled the transcriptional response of lung and muscle tissues chronically exposed to CS, hypoxia (CH) or to combined stimuli (CSCH). The overarching aim of this study was to assess whether GP hindlimb muscles (both oxidative and glycolytic) show a transcriptional response to these exposures, and whether such gene signatures can be correlated to the expression of lung secreted proteins.

We discovered that indeed skeletal muscles of GP exposed to all experimental interventions accurately mimic the transcriptional state of human limb muscle sampled from COPD patients with muscle atrophy. Using a relatively simple network inference method we then identified systemic cytokines whose mRNA and serum protein levels are inversely correlated with the transcriptional state of energy metabolism pathways in GP and human COPD patients, respectively.

These results provide further evidence for the utility of the GP smoking model to study muscle wasting in COPD and support the hypothesis that systemic inflammation plays an important role in altering the energy metabolic state in peripheral muscles causing them to dysfunction.

### 3.3 Methods

#### 3.3.1 Guinea pig smoking model

Sixteen male Hartley guinea pigs were divided into four groups: one group was exposed to CS for 3 months (n = 4); a second group was kept in normoxia for 10 weeks and subsequently placed in an hypoxic environment (12% O<sub>2</sub>) for two weeks (n = 4); a third group (n = 4) was CS-exposed for 3 months and to chronic hypoxia the last two smoking weeks; finally we included a fourth group (n = 4) of sham controls remaining in normoxia for the whole study period. To avoid problems related to ageing, young adults were used (8 weeks of age), and to avoid scaling effects body mass was similar (~300 g/animal). All procedures involving animals and their care were approved by the Ethics Committee of the University of Barcelona and by the University of Valladolid Institutional Committee for Animal Care and Use, and were conducted following institutional guidelines that comply with national (Generalitat de Catalunya decree 214/1997, DOGC 2450) and international (Guide for the Care and Use of Laboratory Animals, National Institutes of Health, 85–23, 1985) laws and policies.

Whole lung as well as soleus and lateral gastrocnemius hindlimb muscles were isolated from each animal at the end of the study period. The soleus muscle and the lateral gastrocnemius were selected to represent oxidative and glycolytic muscles, respectively. It should be noted that although the *gastrocnemius* as a whole is a mixed muscle, the lateral portion is predominantly glycolytic.

Animals receiving CS were daily exposed to the smoke of 4 cigarettes (2R4F; Kentucky University Research; Lexington, KY, USA, 11 mg tar, 0.8 mg nicotine per cigarette), 5 days/week using a nose-only inhalation system (Protowrx Design Inc; Langley, British

Colombia, Canada). Sham-exposure to CS was done daily by placing control animals in the nose-only exposure chamber for the same duration (1 hour) without cigarettes being lighted. In this experimental model, neither nutritional status (determined *via* measurements of plasma cholesterol, protein and lipids) nor whole-body weight gain at the end of the study period is significantly different between CS-exposed and sham animals [158]. Moreover, no changes in the proportion of Type I and Type II fibres can be detected between CS-exposed and shams [147]. Detailed information on exposure protocols, pulmonary function data and histological assessments from this study, which demonstrate that observations in lung function and pulmonary structural changes of COPD patients are indeed replicated in the CS-exposed guinea pigs, have been reported in two separate publications [157, 161].

### **3.3.2 RNA isolation from guinea pig samples**

Total RNA was extracted using the RNeasy Mini extraction kits (Qiagen, USA) according to the manufacturer's recommendations. RNA purity and quality was evaluated using a NanoDrop (Thermo Scientific) and a BioAnalyzer 2100 instrument (Agilent Technologies), respectively. All samples had a RIN score >7.

### **3.3.3 Definition of the guinea pig transcriptome by mRNA sequencing and microarray design**

In 2008 the GP genome was sequenced to a depth of ~7X full coverage, and last updated in 2010. However, because of the lack of cDNA and protein resources the GP genome is at present poorly annotated. Thus, in order to address this issue we performed an in-depth mRNA sequencing of the lung and skeletal muscles transcriptomes and used this to annotate the available GP genome for transcribed sequences. We then used this

information to design *and* validate the custom Agilent microarray platform used throughout this chapter.

Transcriptome sequencing was performed using Illumina sequencing. Briefly, NCBI and Ensembl transcripts of GP were combined with transcripts constructed from Illumina paired end reads using the TopHat and Cufflinks algorithms [162, 163]. Microarray probe sequences were then chosen based on the combined transcriptome assembly. The raw RNA-Seq data have been deposited at the Gene Expression Omnibus (GEO) under the reference number GSE56099. A detailed description of the procedure is provided in Appendix 7.1.

### **3.3.4 Guinea pig microarray gene expression profiling**

One hundred nanograms of total RNA from each sample was amplified and converted into labelled cRNA using Agilent's Low Input Quick Amp Labelling Kit according to the manufacturer's recommendations. Cy3-labelled cRNA (600 ng/sample) was hybridized to our custom *Cavia porcellus* oligonucleotide microarray (manufactured by Agilent) in randomized sample order, which generated 61,657 measures per sample (18,073 annotated genes). Hybridization, washing and scanning of arrays were performed according to manufacturer's protocol. Three samples (one from each tissue) were lost during the process of generating raw data. All scanned microarrays passed all eleven of the Agilent's quality metrics. Capture probes that were flagged (*i.e.* did not pass Agilent's 'well above background' condition) on at least 80% of the chips were removed prior to data analysis, such that only those capture probes with a raw signal greater than 99% of the background population signal, for at least 20% of the samples, were retained (29,333 probes were discarded).

Raw microarray data were then normalised against sham controls for each of the three tissues using loess in the ‘marray’ [164] and ‘limma’ [165] Bioconductor packages. Arrays were examined using hierarchical clustering and principal component analysis (PCA) to identify outliers prior to statistical analysis.

The statistical significance of differential expression of each gene was determined using the Significance Analysis of Microarray (SAM) algorithm [166] with a False Discovery Rate (FDR) cut-off of 1%. Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment in differentially expressed genes was examined using the web-based tool DAVID [167]. Disease KEGG pathways were excluded from the analysis to maximise biological interpretability. Therefore the analysis was restricted to KEGG group 1–4 (Metabolism, Genetic Information Processing, Environmental Information Processing, and Cellular Processes). The microarray data have been deposited in GEO under accession number GSE56099.

### **3.3.5 RT-qPCR validation of custom GP array**

Reverse transcription of 1 µg of isolated total RNA from whole-lung tissue (same RNA as were used for the microarray part) was performed using the Tetro cDNA synthesis kit (Bioline) with random hexamer primers following the manufacturer’s instructions. The resulting cDNA was diluted ten-fold and 2.5 µL of this was used to perform qPCR in triplicate (25 µL reaction mixture volume) using the Maxima SYBR green (Thermo Scientific) and 300nM of primers according the manufacturer’s instructions. To adjust for variations in the cDNA synthesis, each gene was normalized to that of 18S ribosomal RNA and beta-actin mRNA, respectively. All reactions were run in singleplex on a StepOnePlus Real Time System (Applied Biosystems) at 95°C for 10 min, followed by

40 cycles at 95°C for 15 sec and 60°C for 1 min. Two-fold dilution series were performed for all primer pairs to verify the linearity of the assay. In addition, dissociation curve analysis was performed after each PCR to check for unspecific signals. Quantification was performed using the comparative cycle threshold ( $2^{-\Delta\Delta C_t}$ ) method [168].

The following primers were used:

CXCL9 fwr: 5'-AGGCACCCCAGTAATGAG-3';

CXCL9 rev: 5'-TGATTTCTGTTTTCTCACACG-3';

CXCL10 fwr: 5'-TCTGAGTGGGACTCAAGGAATACC-3';

CXCL10 rev: 5'-TCCAGACATCTCTTCTCCCCATTC-3';

beta-actin fwr: 5'-GAGGCACCAGGGAGTCATG-3';

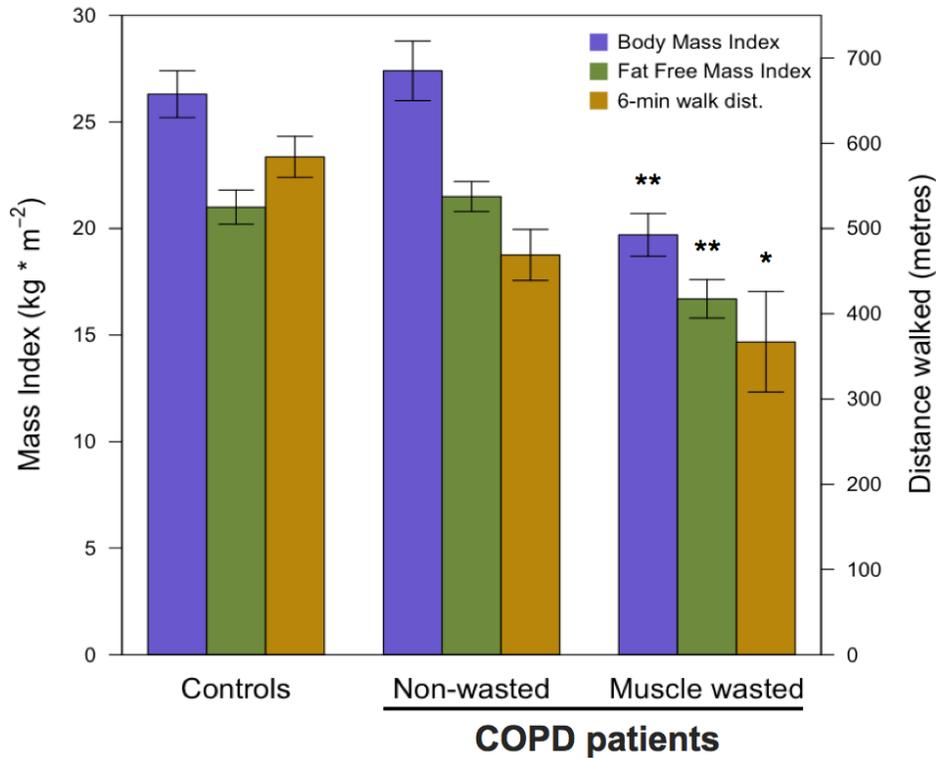
beta-actin rev: 5'-AAGGTGTGGTGCCAGATCTTCTC-3';

18S rRNA fwr: 5'-GTACAGTGAAACTGCGAATGGCTC-3';

18S rRNA rev: 5'-CCGTCGGCATGTATTAGCTCTAG-3'.

### 3.3.6 Human COPD clinical studies

In order to assess the clinical relevance of the findings in respect to the GP dataset, we took advantage of a human microarray dataset we have previously published [133]. This defined the baseline/resting transcriptional state of the *vastus lateralis* muscle in severe COPD patients with either a normal (n = 9) or low (n = 6) body mass index (BMI) and healthy controls matched for age and smoking history (n = 12) (Table 3-1). In addition, the low BMI COPD group also had a significantly lower fat free mass index (FFMI) (on average 16.7 kg/m<sup>2</sup>; Figure 3-1), a clear surrogate for muscle wasting.



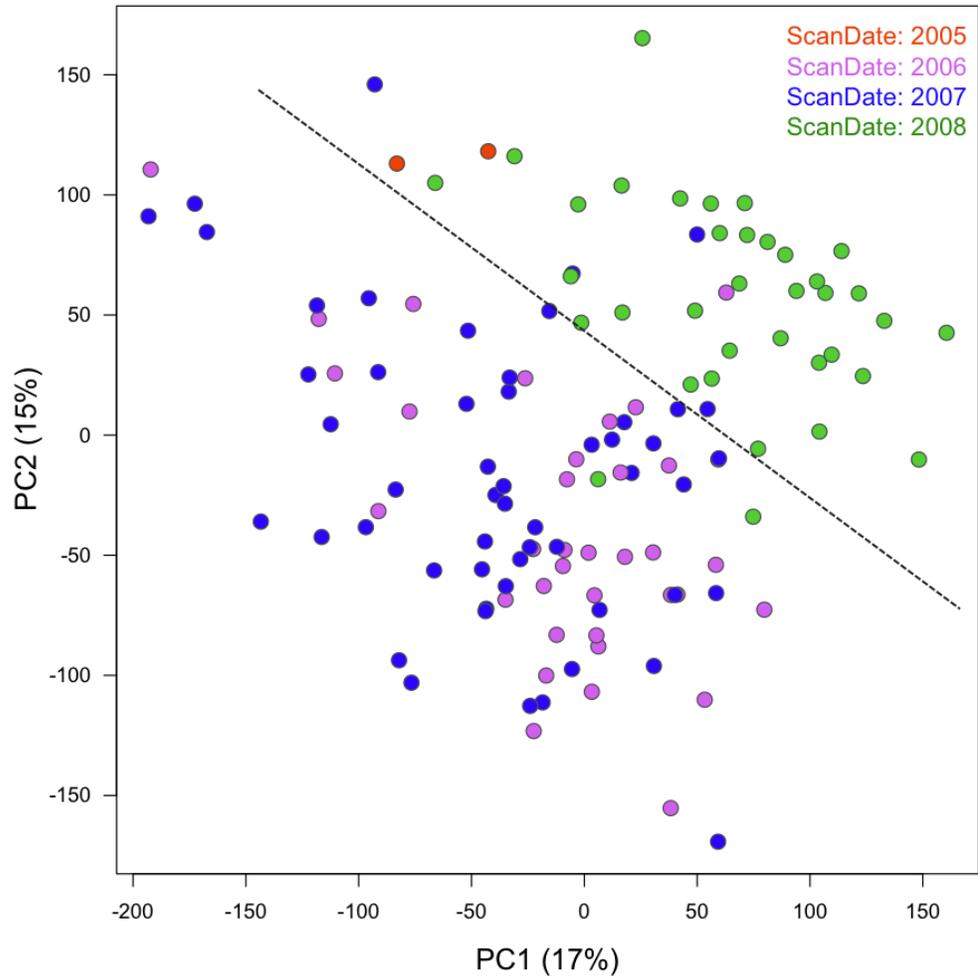
**Figure 3-1 | Barplot highlighting group mean differences in mass indexes (whole-body and fat-free, respectively) and 6-min walking distance.**  
 Error bars represent SEM. \*\*  $p < 0.01$ ; \*  $p < 0.05$ .

	Healthy controls	COPD, BMI <sub>norm.</sub>	COPD, BMI <sub>low</sub>
Gender (M/F)	10/2	9/0	6/0
Age (years)	65.3±2.9	69.4±1.5	69.2±4.6
BMI (kg/m <sup>2</sup> )	26.3±1.1	27.4±1.4	19.7±1.0 <sup>**,††</sup>
FFMI (kg/m <sup>2</sup> )	21.0±0.8	21.5±0.7	16.7±0.9 <sup>**,††</sup>
VE (L/min)	71.2±5.6	40.5±3.6 <sup>***</sup>	33.0±3.8 <sup>***</sup>
FEV <sub>1</sub> (L)	3.46±0.2	1.41±0.09 <sup>***</sup>	1.21±0.21 <sup>***</sup>
FEV <sub>1</sub> /FVC (%)	75.9±2.4	44.0±2.7 <sup>***</sup>	39.5±4.5 <sup>***</sup>
RV (% of pred.)	103.9±5.2	145.0±13.3	160.0±28.6 <sup>*</sup>
VO <sub>2max</sub> (l·min <sup>-1</sup> ·kg <sup>-1</sup> )	22.3±1.4	13.9±1.7 <sup>**</sup>	14.4±1.5 <sup>**</sup>
Peak power (W)	117±8	60±7 <sup>***</sup>	47±9 <sup>***</sup>
6MWD (m)	584±24	469±30 <sup>*</sup>	367±59 <sup>***</sup>
BODE index	0.1±0.1	2.3±0.4 <sup>**</sup>	4.0±1.0 <sup>***,†</sup>
Data are presented as mean±SEM. * $P < 0.05$ ; ** $P < 0.01$ ; *** $P < 0.001$ versus controls. † $P < 0.05$ ; †† $P < 0.01$ ; ††† $P < 0.001$ versus COPD patients with a normal BMI. Comparisons were analysed using one-way ANOVA and Tukey's <i>post hoc</i> test. BMI: body mass index; FEV <sub>1</sub> : forced expiratory volume in 1 s; FFMI: fat-free mass index; VE: lung ventilation; 6MWD: 6-min walking distance			

**Table 3-1 | Baseline physiological data of the COPD patients and healthy controls.**

All participants signed a written, informed consent approved by the Ethics Committee on Investigations Involving Human Subjects at the Hospital Clinic, Universitat de Barcelona, and the study was conducted in accordance with principles of the Declaration of Helsinki. Briefly, raw Affymetrix CEL files were RMA normalized following removal of probes that were termed ‘absent’ in more than 80% of the samples by the MAS5 algorithm inside the ‘affy’ package (26,197 probes were discarded). Following probe summarization, a two-class unpaired SAM analysis was performed using the R package ‘samr’ comparing gene expression levels between COPD patients with a muscle wasting phenotype and matched controls. Enrichment of KEGG terms (group 1 to 4) in the resulting gene lists was assessed using DAVID. Enriched terms used to define the ‘true response’ in the cross-species overlap analysis were defined as having an EASE score p-value < 0.2. The raw microarray .CEL files are deposited under the reference number GSE27536.

In addition, we also analysed a public microarray dataset published by the Ronald Crystal lab [169] examining transcriptional changes in small airway epithelium from healthy non-smokers (n = 47), healthy smokers (n = 58) and smokers with COPD (n = 22), respectively (GSE19407). Due to a clear scan date batch issue (Figure 3-2), which the original authors did not discover, we focused our analysis on the data generated in year 2006 and 2007 (hence excluding the two samples scanned in 2005 as well as the 36 samples processed in 2008). As both human studies were conducted on the Affymetrix U133+2 platform, the data analysis strategy of the raw CEL files representing the pulmonary data was identical to that of the human dataset in skeletal muscle presented in this paper (see above).



**Figure 3-2 | PCA plot showing that subjects profiled in the GSE19407 dataset (n=127) group by the year they were scanned (30,652 probesets included).**

Particularly, a clear separation exists between samples scanned in year 2005/2008 and 2006/2007 (see dotted black line). Noteworthy, the samples profiled in year 2006 and 2007 are intermixed.

### 3.3.7 Summarizing the molecular state of skeletal muscle using indices of pathway transcriptional activity

In order to reduce the complexity of the genome-wide transcriptional state of GP skeletal muscles, thereby increasing statistical power, we computed indices of the overall pathway transcriptional activity [74, 170]. For each of the two GP hindlimb muscles, we

first mapped the thousands of individual gene expression measures onto KEGG pathways using DAVID [128]. We then summarized the transcriptional activity for the enriched pathways (FDR < 10%) by computing the first three principal components (PCs), a procedure that allowed us to retain between 50% and 78% of the total variance (63% on average). Computation of the PCs was performed using the ‘prcomp’ function within the statistical programming environment R.

### **3.3.8 Inference of biological networks linking lung and skeletal muscles in guinea pigs**

An exhaustive list of genes annotated to the cytokine superfamily (n=72) was compiled from the SABiosciences PCR Array Web Portal (see Appendix 1.1 for the complete list). Such an approach has been used previously for compiling gene-lists [171]. Among these candidates we identified 33 genes coding for cytokines, which were differentially expressed in GP lung tissue (Table 3-2). These were selected for further analysis.

Cytokine	Fold change regulation compared to controls		
	CON/CH	CON/CS	CON/CSCH
Bmp2			1.17
Ccl11			1.24
Ccl17	-2.21	-1.31	2.77
Ccl2			2.64
Ccl22		-1.41	3.99
Ccl24	1.54		11.29
Ccl3		-1.77	4.65
Ccl4			1.73
Ccl5			-1.93
Ccl7	1.49	-1.63	1.60
Cntf	-1.47	-1.19	1.11
Csf1			1.30
Csf3	1.07		
Cx3cl1	-1.46	-1.64	
Cxcl1			5.26
Cxcl10	1.30		1.71
Cxcl12	1.52		1.44
Cxcl16	1.40	-1.52	1.39
Cxcl5		-1.37	1.37
Cxcl9	1.21	-1.20	-1.16
Gpi1	-1.33		
Il10		-1.14	-1.16
Il11		-1.11	
Il16		-1.34	
Il17f	-2.31	-2.39	-2.31
Il1a	-1.27		1.72
Il1b		-2.09	3.98
Il1rn	1.12		1.33
Il27	-1.14	-1.12	-1.09
Lif	-1.34	-1.66	
Osm		-2.59	
Ppbbp		-1.16	-1.18
Vegfa			1.26

**Table 3-2 | List of genes that are annotated to the cytokine superfamily AND differentially expressed in whole-lung tissue of treated guinea pigs compared to untreated controls (n=33; FDR<1%).**

Coloured cells denote non-differentially expressed.

Correlation between the expression of these cytokines and skeletal muscle pathway indexes were computed using the Spearman correlations coefficient, which allows the identification of linear and non-linear monotone relationships [108]. Resampling of samples (10,000 permutations) was conducted to obtain *P*-values for each correlation coefficient. Pair-wise associations within the regulatory network were defined as statistically significant when  $P < 0.01$ .

The resulting sparse network was visualised using a force-directed layout as implemented in the network visualization tool Cytoscape v2.8 [112].

### **3.3.9 Creating and visualizing a KEGG pathway map**

To visually represent the relationship between enriched KEGG pathways in the clinical dataset, we computed a pathway similarity matrix based on the Jaccards Index of overlap. This matrix was used as input to a hierarchical clustering procedure (average linkage).

### **3.3.10 Measurement of inflammatory mediators in human serum from COPD patients and healthy controls and validation of the guinea pig lung-muscle cross-talk network**

Previously published multiplex protein profiling data from COPD serum samples (n = 26) and healthy controls (n = 23) were included [133]. Briefly, data were  $\log_2$  transformed followed by imputation of missing values using K nearest neighbours in the R package ‘impute’ [172]. Finally, the full data matrix was Z-scored.

A Mack-Skillings test with two factors (disease and training) was used to identify overall main effects across groups in serum protein levels ( $P < 0.05$ ). A Gene Set Enrichment Analysis (GSEA) was used to establish statistical functional enrichment by ranking all Pearson correlation coefficients between serum protein levels and global muscle mRNA expression [84].

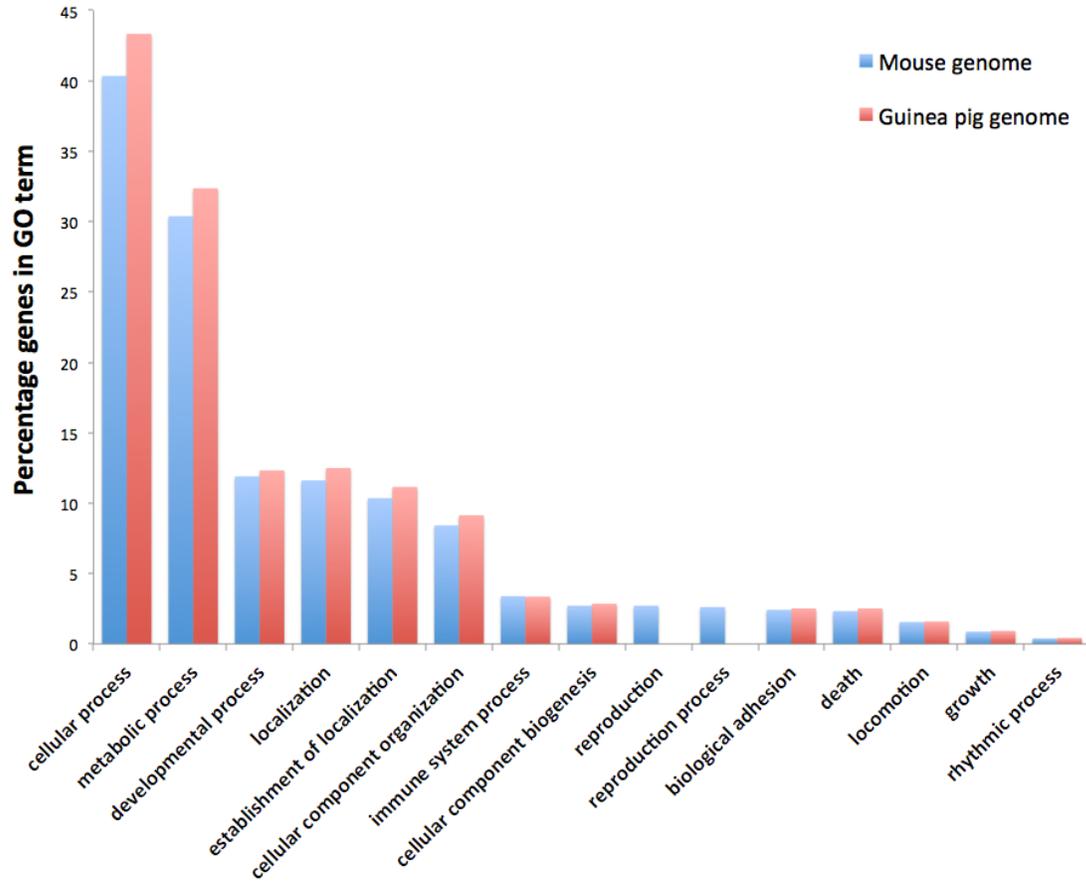
### 3.4 Results

#### 3.4.1 Sequencing of the Guinea pig transcriptome and development of a genome-wide guinea pig microarray platform

Illumina RNA sequencing (RNA-Seq) in this study has facilitated the construction of a comprehensive transcriptome for GP lung and skeletal muscles, with much higher coverage than attainable purely by public available data. In combination with public domain data, Ensemble cDNAs and Genscan gene predictions, we have generated the first comprehensive annotation of the genome-wide transcriptome consisting of 151,072 transcript sequences (of which 81,074 were derived solely from the RNA-Seq data). The number of transcripts annotated with a RefSeq sequence, by stringent BLAST searching against mouse transcripts from NCBI's RefSeq collection, was 97,822. This represented 17,907 non-redundant mouse gene symbols. Annotated genes were classified according to GO categories: cellular component (CC), biological process (BP) and molecular function (MF). Figure 3-3 and Figure 3-4 depicts the distribution of the major GO categories at 'level 1'<sup>3</sup>, for comparison we also included level 1 GO terms for the *mouse* transcriptome. Overall, the GP GO term representation is very comparable to that of the genes annotated in the full mouse genome, highlighting the generality of the assembled GP transcriptome. Only reproduction processes and extracellular region are poorly represented in the guinea pig transcriptome.

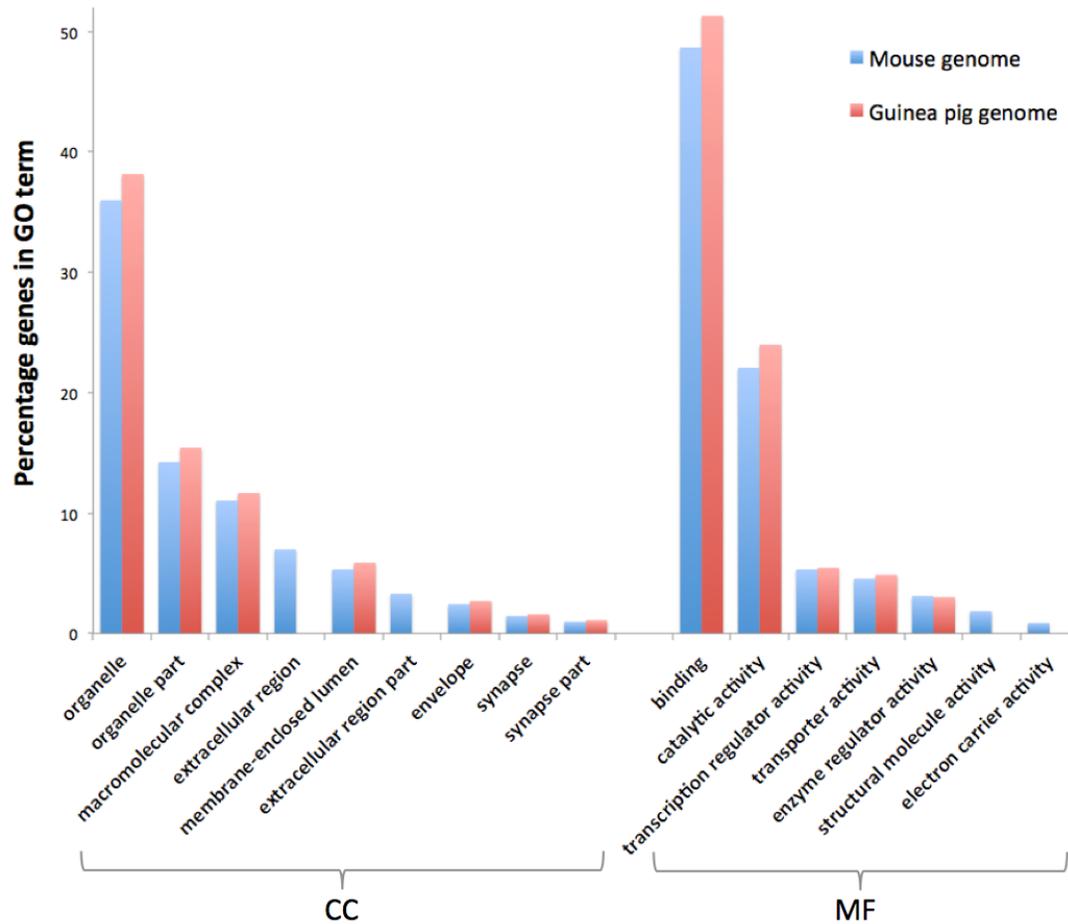
---

<sup>3</sup> Gene Ontology terms are organized hierarchically such that higher level terms are more general.



**Figure 3-3 | Barplot showing level 1 Gene Ontology (GO) terms of the biological process category.**

Blue bars represent all functionally annotated genes from the *Mus musculus* species whereas the red bars represent all annotated genes from the guinea pig RNA-Seq analysis.



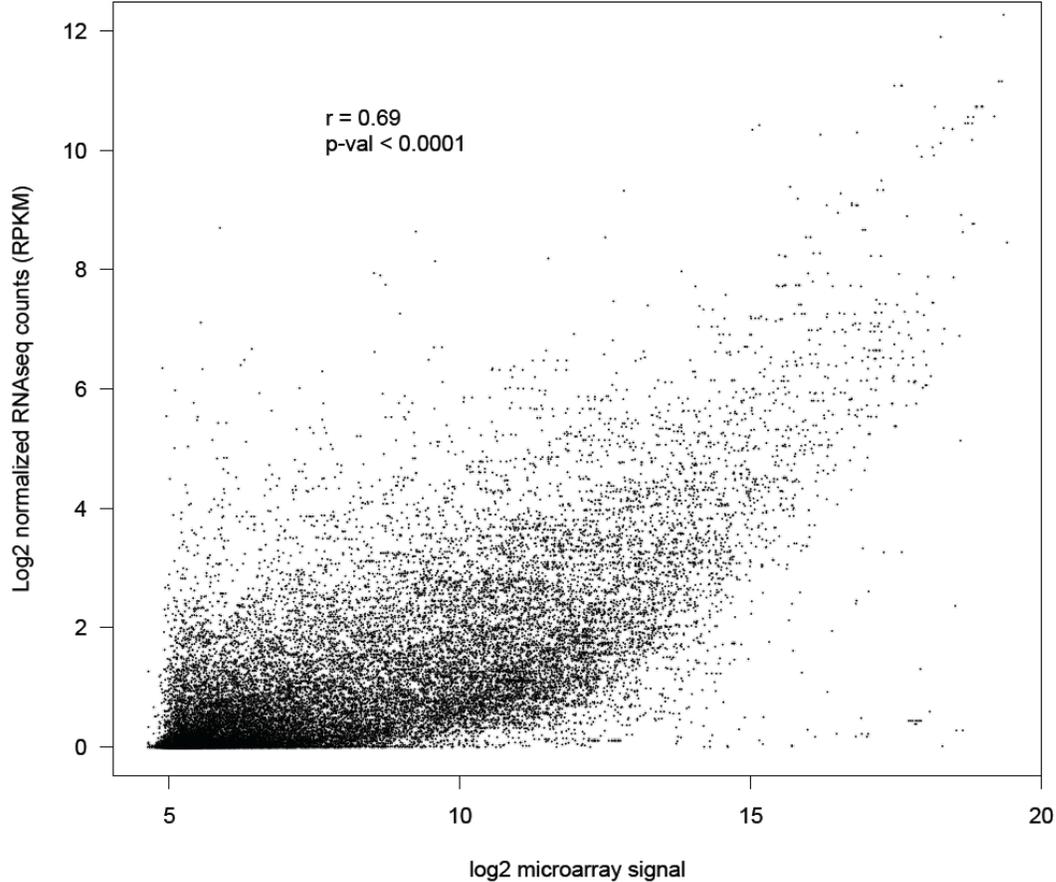
**Figure 3-4 | Barplot showing level 1 Gene Ontology (GO) terms of the cellular component (CC) and molecular function (MF) category, respectively.**

Blue bars represent all functionally annotated genes from the *Mus musculus* species whereas the red bars represent all annotated genes from the guinea pig RNA-Seq analysis.

In addition, by Human ortholog identification, it was shown that the GP transcriptome assembled in this work contained genes included in the entire set of Human KEGG pathways available for download via the Broad Institute’s MSigDB Collections [173] (please see ‘KeggGPandMouseCounts.xlsx’ associated to the [online version](#) of Additional file 1).

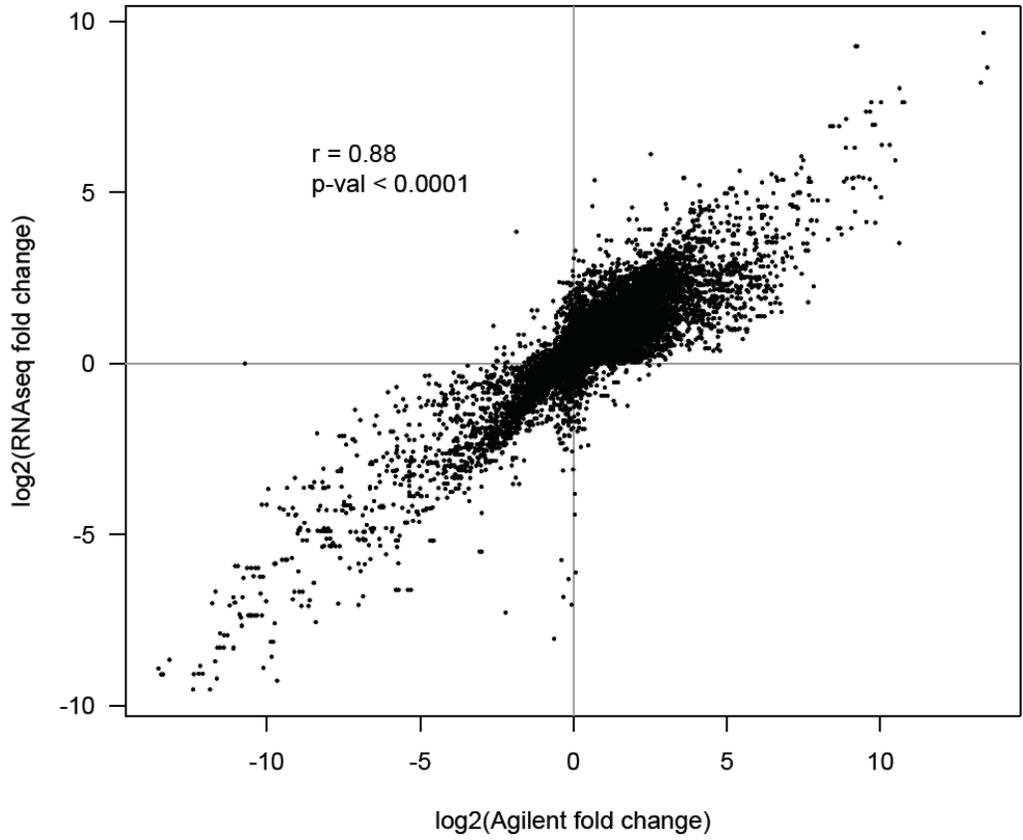
Using the transcript assembly we have developed the first genome-wide microarray platform for the GP model species. Based on the probe performance using an initial 180K array, we developed a 60K custom Agilent microarray, representing 17,896 unique genes. Importantly, we are able to demonstrate a significant concordance between our custom 60K array platform and RNA-Seq data (Figure 3-5 & Figure 3-6), particularly when the ratio between gene expression in lung and muscle tissue is compared (Figure 3-7).

### Guinea pig skeletal muscle tissue

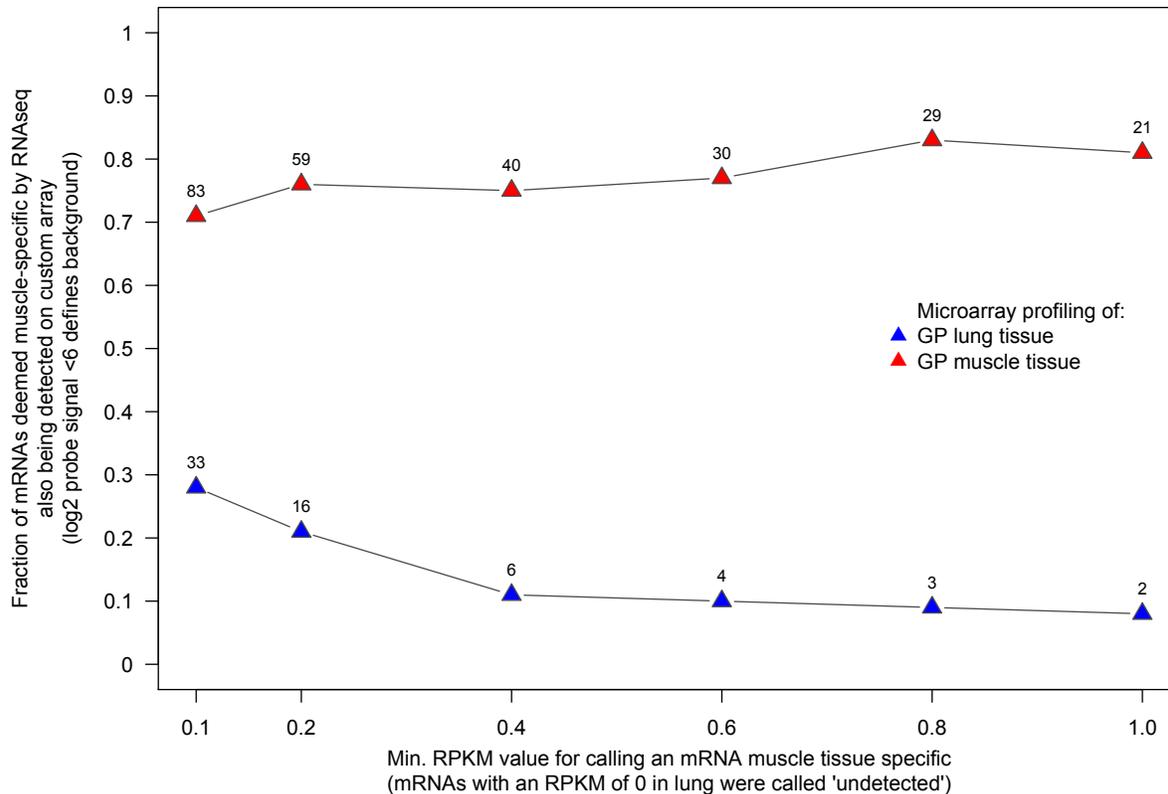


**Figure 3-5 | Scatterplot highlighting the association between signal intensity on the custom guinea pig microarray and normalized RNA-Seq counts using pooled RNA from skeletal muscles.**

The two profiling technologies show a good agreement for genes with an above-median level of expression. However, for transcripts with a  $\log_2$  RPKM value close to zero, the microarray platform assigned a much wider range of intensity values (primarily from 5 to 10 on the  $\log_2$  scale). A possible cause for this reduced correlation for genes with a below-median expression may be related to non-specific binding to the array capture probes, which affect lowly expressed genes to a greater extent. To avoid taking the log base 2 of zero, we added 1 to each of the counts prior to taking the  $\log_2$ .



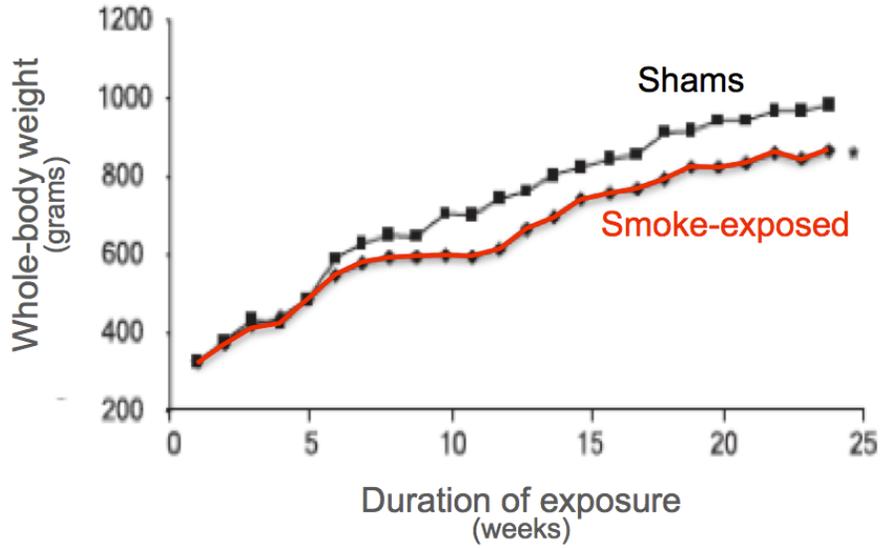
**Figure 3-6** | Comparison of estimated log<sub>2</sub> fold changes (lung/muscle) from Illumina RNA-Seq (y-axis) and the custom microarray platform (x-axis).



**Figure 3-7** | Plot indicating the fractional overlap (y-axis) between profiling technologies in terms of genes being called muscle-specific based on the RNA-Seq data (x-axis). The numeric value above each triangle indicates the number of transcripts termed muscle-specific by RNA-Seq (at a given RPKM threshold) that overlap with genes being expressed above detection background in lung (blue triangles) and muscle tissue (red triangles), respectively, on our 60K custom microarray platform.

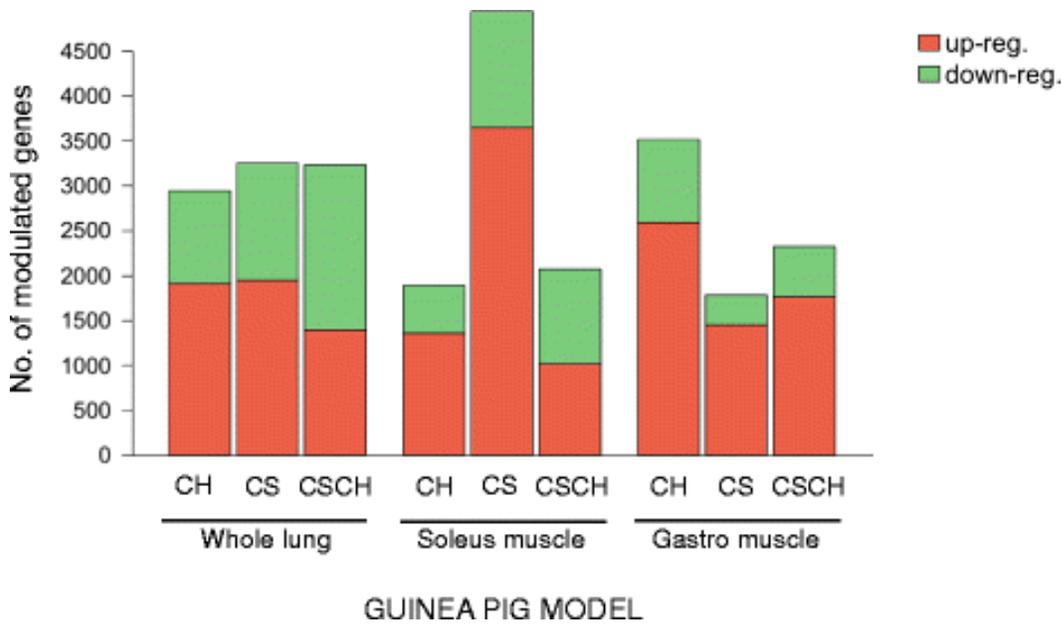
### 3.4.2 Chronic exposure to smoking and/or hypoxia induces transcriptional changes in both lung and skeletal muscle in guinea pigs

To assess whether the current GP model show a transcriptional response, we used our newly developed microarray platform with mRNA extracted from whole lung and two metabolically distinct hindlimb muscles (*gastro* and *soleus*) sampled from shams and nose-only CS exposed animals approximately one month after a significant ( $P < 0.01$ ) decrease in body mass gain could first be detected (Figure 3-8) [157].



**Figure 3-8 | Guinea pig whole-body weight gain during the experimental protocol.**

Indeed, we were able to identify a relatively large number of genes differentially expressed in all three tissues from all experimental groups (Figure 3-9). Notably, both hindlimb muscles examined displayed a marked response, with the oxidative soleus muscle showing the largest number of changes following exposure to CS alone (Figure 3-9).



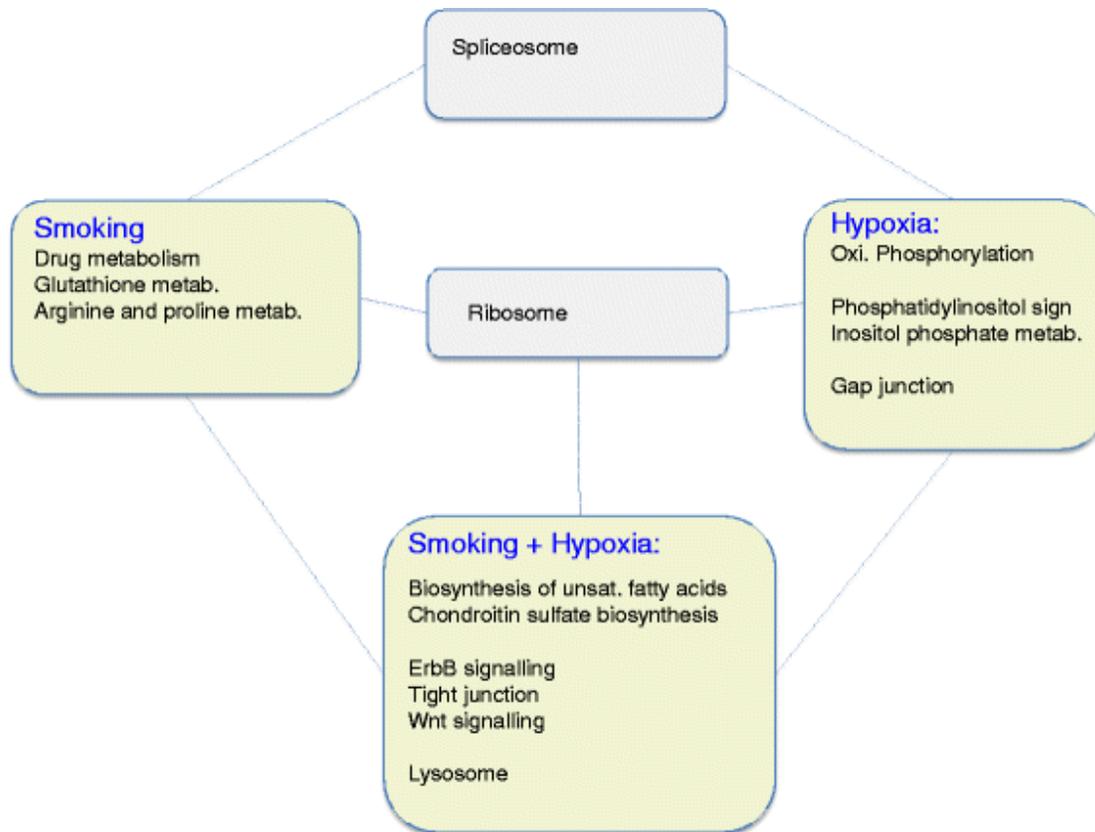
**Figure 3-9 | Differentially expressed genes in the Guinea pig experimental model.**

Barplot summarising the number of differentially expressed genes identified by SAM analysis (FDR < 1%) following long-term exposure to hypoxia (CH), chronic smoking (CS), or smoking followed by hypoxia (CS + CH) in whole lung and skeletal muscles (soleus and gastrocnemius).

### 3.4.3 Lung response to the different stressors is comparable in magnitude, but involves different subsets of functional pathways

Having shown that both lung and skeletal muscles mount a quantitatively comparable transcriptional response following the different experimental challenges, we then compared and characterized such responses at the functional pathway level. We first focused on lung tissue and performed a functional enrichment analysis, which identified 15 KEGG pathways that were enriched in genes differentially expressed in lungs, in at least one experimental condition (Figure 3-10). Only genes annotated to the *ribosome* pathway were significantly modulated by all three experimental conditions. Instead the vast majority of differentially modulated pathways were equally distributed between the specific exposure conditions. The six pathways (40%) unique to the CSCH group could

be grouped into two main functional categories: *i)* biosynthetic pathways, and *ii)* pathways with a strong signalling component such as ErbB- and Wnt signalling as well as tight junction. In contrast, all 3 pathways (*i.e.* cytochrome P450 drug-, glutathione-, and arginine & proline metabolism) enriched in genes up-regulated by long-term smoking *per se* could be associated to metabolic processes primarily involved in detoxification of oxidative stress. Finally, the 4 enriched terms specific to hypoxia can be broken down into 2 main functional classes: *i)* an oxygen-dependent bioenergetic component (Oxidative Phosphorylation) enriched in genes down-regulated by hypoxia, and *ii)* pathways with a strong signalling component also negatively affected by hypoxia (*i.e.* phosphatidylinositol signalling, inositol phosphate metabolism and gap junction).



**Figure 3-10 | Overlap analysis of enriched KEGG pathways in Guinea pig lung tissue.**

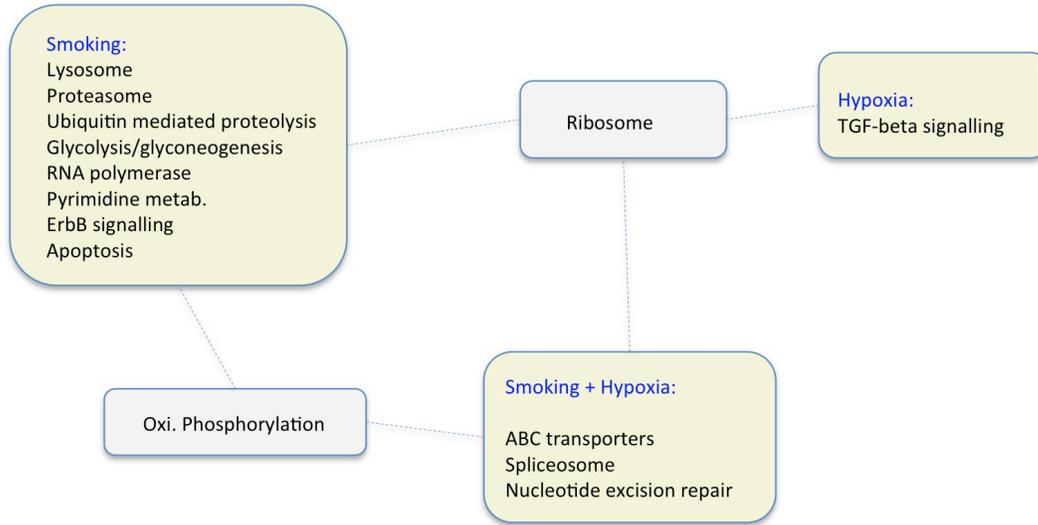
Rectangles coloured in green indicate pathways unique to a specific experimental condition (*i.e.* CS, CH or CSCH). Grey-coloured rectangles indicate pathways that are regulated by more than one condition.

### **3.4.4 Glycolytic and oxidative limb muscles respond differently to either smoking or hypoxia**

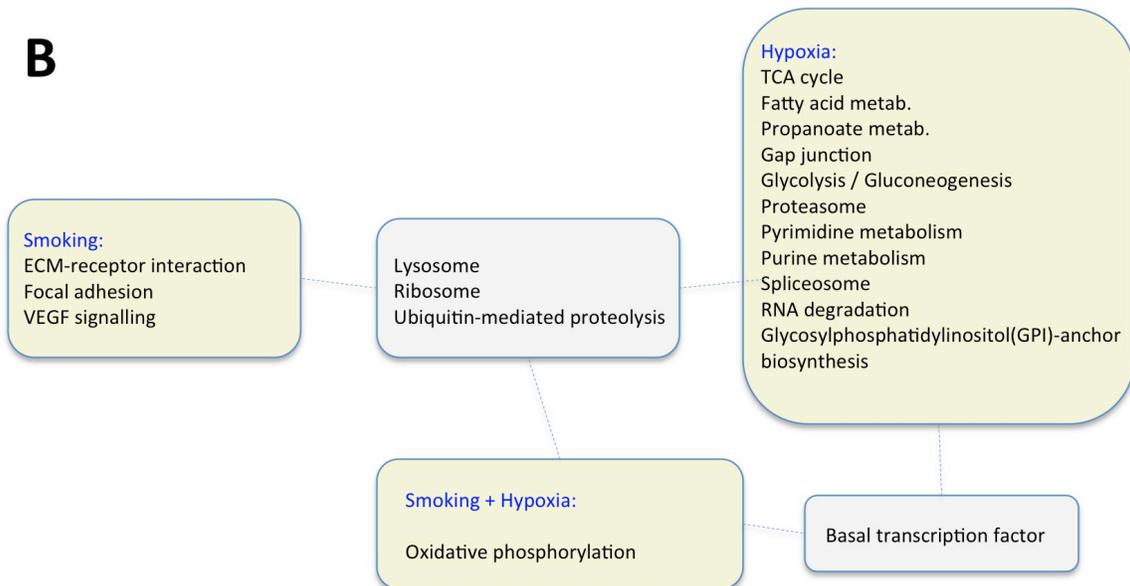
The lateral gastrocnemius modulated a larger number of genes following hypoxia, whereas the soleus (an oxidative muscle) responded preferentially to the smoking challenge (Figure 3-9). This trend was even more evident when testing for functional pathway enrichment. The soleus muscle responded to CS exposure by modulating 10 of the 13 pathways (77%) identified as differentially regulated in at least one sample group,

making this muscle the most sensitive to this stressor (Figure 3-11A). Of these, eight were specifically modulated by CS.

**A**



**B**



**Figure 3-11 | Overlap analysis of enriched KEGG pathways in two Guinea pig hindlimb muscles with discrete metabolic profiles.**

Rectangles coloured in green indicate pathways unique to a specific experimental condition (*i.e.* CS, CH or CSCH). Grey-coloured rectangles indicate pathways that are regulated by more than one condition. **Panel A** is soleus muscle; **Panel B** represents gastrocnemius muscle.

The response to hypoxia only involved 2 pathways of which one was specific to CH (TGF-beta signalling). However, hypoxia has a considerable effect in the CSCH exposure, effectively reducing the impact of CS (only five pathways were enriched when both stimuli were combined).

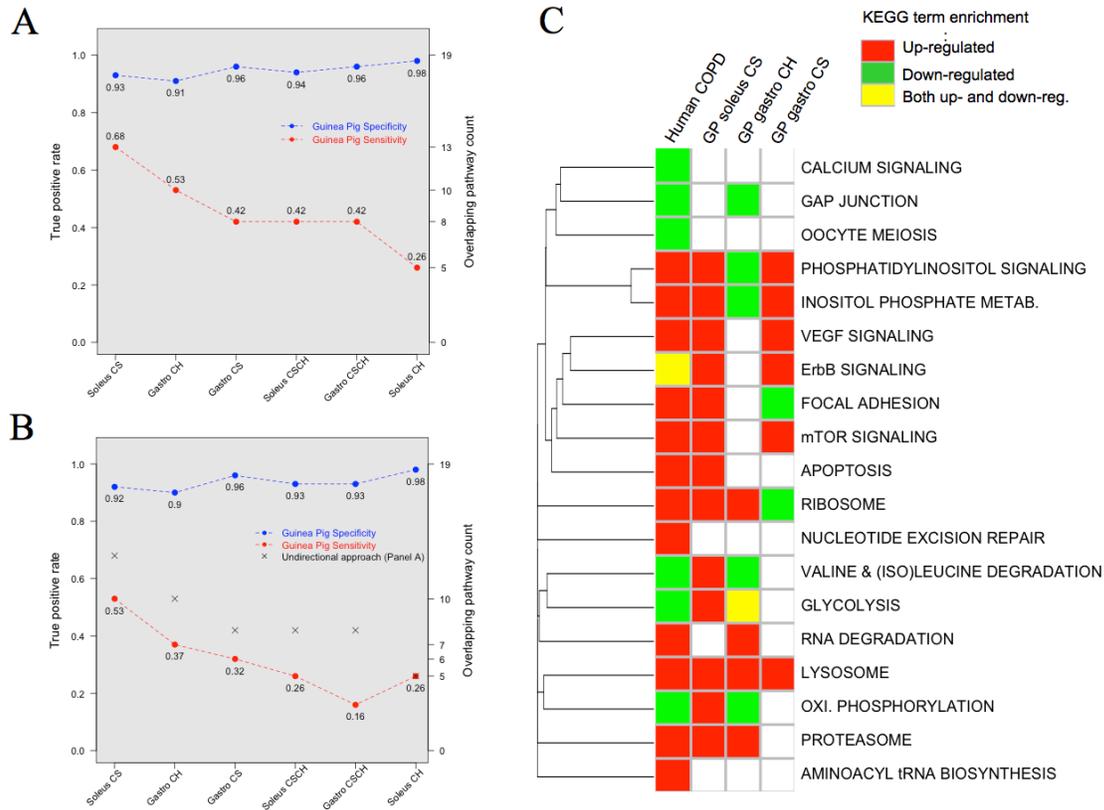
Consistent with the gene-level analysis, the gastrocnemius muscle primarily showed functional pathway enrichment in response to hypoxia (Figure 3-11B). Genes differentially regulated in this condition, were enriched in 15 of the 19 KEGG pathways (79%) modulated in this muscle in at least one of the experimental challenges. These could be grouped in to three main functional clusters: *i*) bioenergetic pathways such as glycolysis and TCA cycle, *ii*) metabolic pathways (e.g. fatty acid-, propanoate-, purine- and pyrimidine metabolism), and *iii*) pathways exerting degradative processes such as proteasome and ubiquitin-mediated proteolysis. Only three pathways, all with a tissue-remodelling component, were specific to the CS group (focal adhesion, VEGF- and ECM-receptor signalling; Figure 3-11B). As a further contrast to the oxidative soleus muscle (Figure 3-11A), aerobic energy metabolism in gastro, represented by OxPhos, was only enriched among down-regulated genes when hypoxia was added on top of the CS-challenge.

#### **3.4.5 The guinea pig smoking model recapitulates the transcriptional changes observed in human COPD skeletal muscles**

Having described the pulmonary as well as extrapulmonary transcriptional response to long-term CS and/or hypoxia, we assessed their clinical relevance with an initial primary focus on peripheral skeletal muscle. This was achieved by comparing the enriched transcription-based functional profiles derived from the GP model with the functional

profile of genes differentially expressed in quadriceps skeletal muscle biopsies from COPD patients relative to matched healthy controls (see online Additional file 8 for the list of regulated transcripts as well as functional enrichment analysis; <http://www.genomemedicine.com/content/6/8/59/additional>) [132].

First, we defined the transcriptional signature representing muscle wasting in human COPD by comparing a cohort of COPD patients with low FFMI to matched healthy individuals. This identified 1,861 differentially regulated genes (1,416 up- and 445 down-regulated), which represented 19 unique functionally enriched KEGG pathways (Figure 3-12C). In order to assess which experimental challenge best mirrored human COPD, we performed a sensitivity and specificity analysis where ‘true response’ was defined by the 19 KEGG terms (KEGG group 1–4) representing dysfunction human COPD muscle. This analysis, performed ignoring the direction of change, showed that a large percentage of enriched KEGG terms in GPs *did* overlap with the COPD pathway signature - irrespective of the exposure (specificity scores ranging from 91% to 98%,  $p < 0.01$ ) (Figure 3-12A). Soleus muscle derived from CS-exposed GP showed the highest sensitivity, with 13 out of 19 (68%) KEGG pathways in common (Figure 3-12A). The *gastro* muscle from hypoxic guinea pigs followed by a short measure with a sensitivity of 53% and 10 KEGG pathways in common. The pathway overlap was statistically significant in all six experimental conditions ( $P < 0.01$ ).



**Figure 3-12 | Muscle specific pathway-level comparisons between the guinea pig model and a clinical COPD study (GSE27536).**

The figure displays the result of the sensitivity and specificity analysis for each of the six experimental conditions at the KEGG pathway level, where “true response” is defined by the 19 enriched KEGG pathways in the *vastus lateralis* of muscle-wasted COPD patients when compared to matched healthy controls. **Panel A** is a plot where the x-axis represents the experimental conditions in the GP model. The two y-axes represent *i)* the true positive rate as a fraction, and *ii)* the actual number of overlapping pathways with the human COPD cohort. In this panel specificity and sensitivity is computed without considering the direction of change in expression (up- and down-regulation). **Panel B** represents the equivalent of the plot in *Panel A* where specificity and sensitivity is computed taking into consideration the direction of change. The grey-colored crosses indicate the sensitivity values from the undirected approach presented in *Panel A*. **Panel C** represents the specific pathways regulated in the GP model (column 2–4) when contrasted against COPD patients with a muscle-wasting phenotype (column 1). Green-coloured cells indicate enrichment among down-regulated transcripts; red-coloured cells indicate enrichment among the up-regulated transcripts; and yellow-coloured cells indicate enrichment for transcripts enriched in both directions.

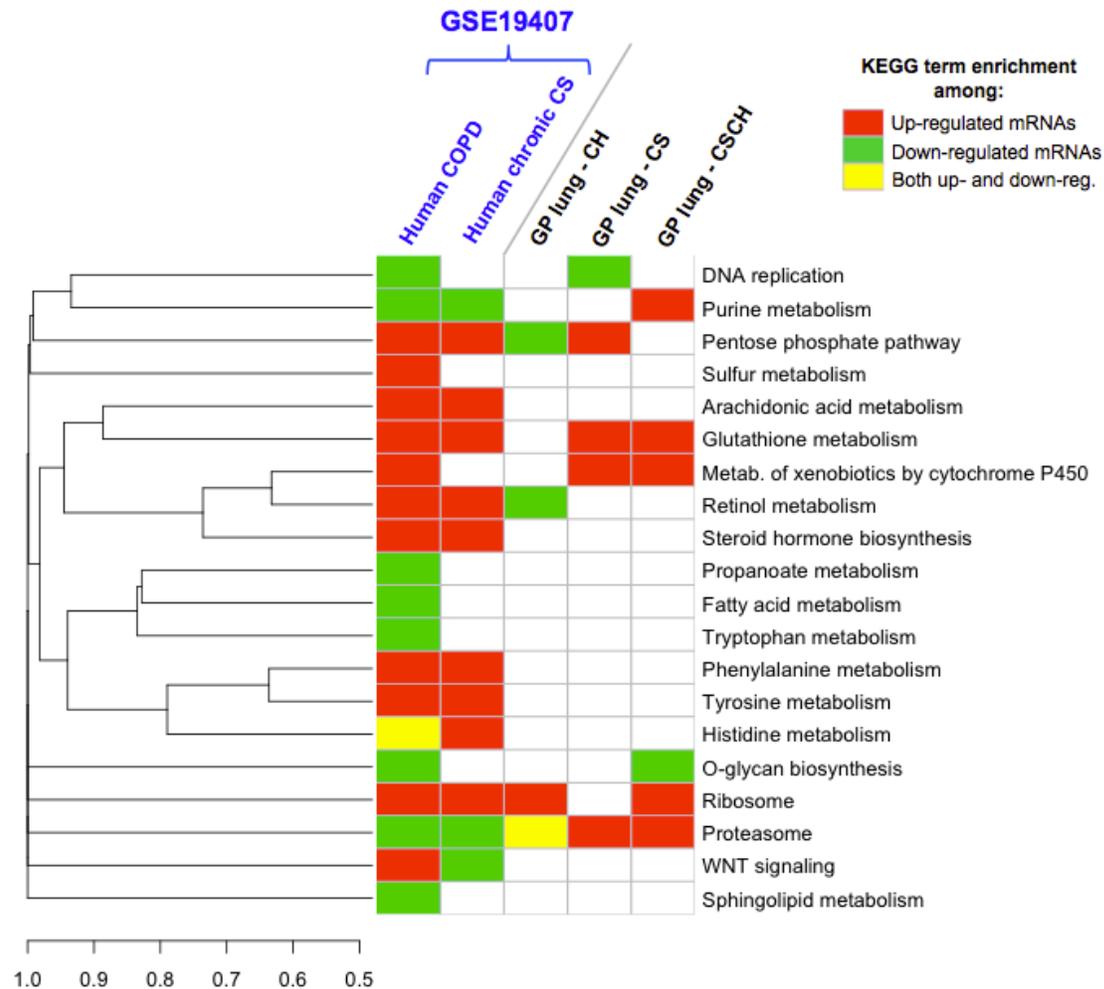
The same comparison, this time only considering those KEGG pathways that were enriched in genes with the same direction of regulation as the human dataset (Figure 3-12B), still revealed a significant overlap with the human dataset for 5 out of 6 experimental challenges (only CSCH-exposed gastro had a  $P > 0.01$ ). On average the sensitivity only decreased by 14%, indicating that most of the response is in the same direction.

We conclude that the transcriptional response of GP limb muscle to long-term CS accurately reflects the transcriptional state of dysfunctional limb muscles in COPD patients. From a biological standpoint, many of the KEGG pathways in common between the GP model and human COPD relate to (Figure 3-12C):

- 1) increased tissue remodelling (VEGF signalling, focal adhesion, apoptosis)
- 2) altered energy metabolism (oxidative phosphorylation, glycolysis)
- 3) increased proteolysis

Thus, we demonstrate the validity of the GP smoking model as the basis for further mechanistic investigation, to facilitate better clinical therapy.

Noteworthy, a similar approach comparing GP and human lungs from healthy and COPD smokers highlighted a less striking, albeit significant response ( $p < 0.05$ ) (Figure 3-13).



**Figure 3-13** | specific pathways regulated in the GP model (column 3–5) when contrasted against COPD patients (column 1) and healthy chronic smokers (column 2). Green-coloured cells indicate enrichment among down-regulated transcripts; red-coloured cells indicate enrichment among the up-regulated transcripts; and yellow-coloured cells indicate enrichment for transcripts enriched in both directions

Of the 2,584 genes regulated in small airway epithelium from smokers with COPD compared to healthy non-smokers (832 up- and 1,752 down-regulated; FDR<5%), which represented 20 unique functionally enriched KEGG pathways (see 1<sup>st</sup> column in Figure 3-13), only 6 pathways (30%) were in common with the CSCH group in our GP model. Whereas this animal model does capture the CS-induced oxidative stress response (*i.e.*

glutathione metabolism and metabolism of xenobiotics by cytochrome P450 are both induced), our functional gene-level comparison indicates a “metabolic gap” in the current GP smoking model, which is particularly pronounced for amino acid metabolism (Figure 3-13).

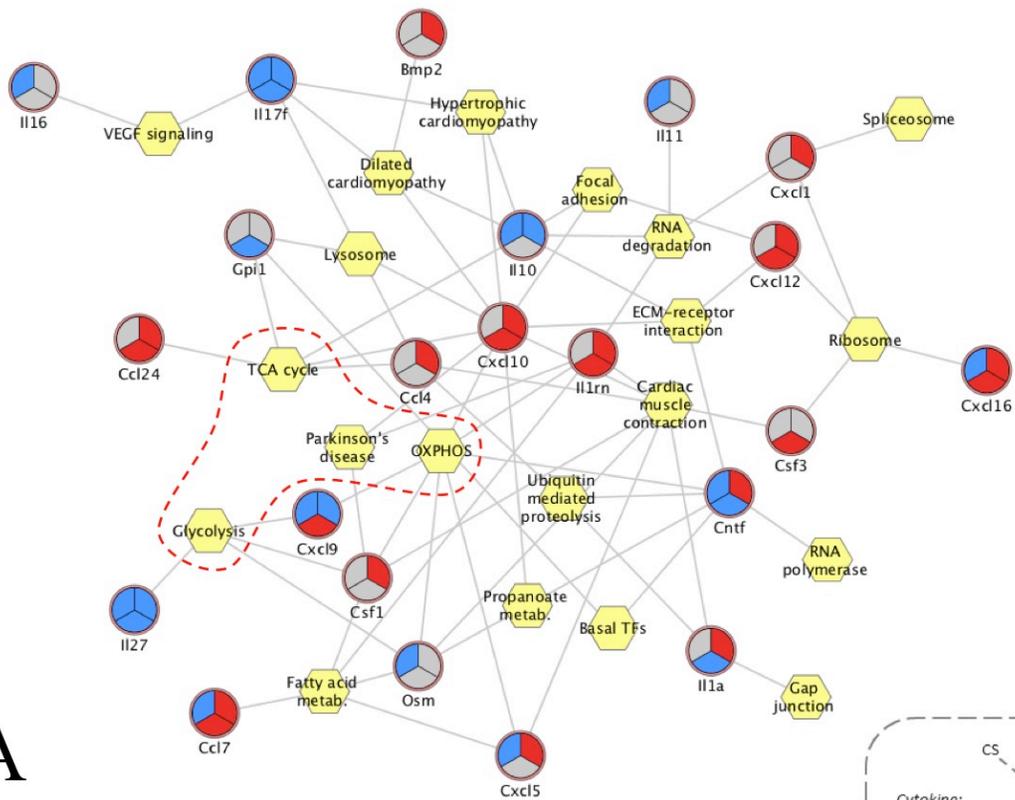
A significant fraction (60%) of the enriched pathways in smokers with COPD are also modulated in healthy smokers *without* COPD. Of the pathways regulated in healthy smokers compared to non-smokers, 3 pathways (25%) overlapped with the smoking GP model (*i.e.* glutathione metabolism, pentose phosphate- and the proteasome pathways) (Figure 3-13).

#### **3.4.6 Gene expression of lung soluble inflammatory mediators correlate with skeletal muscle gene expression**

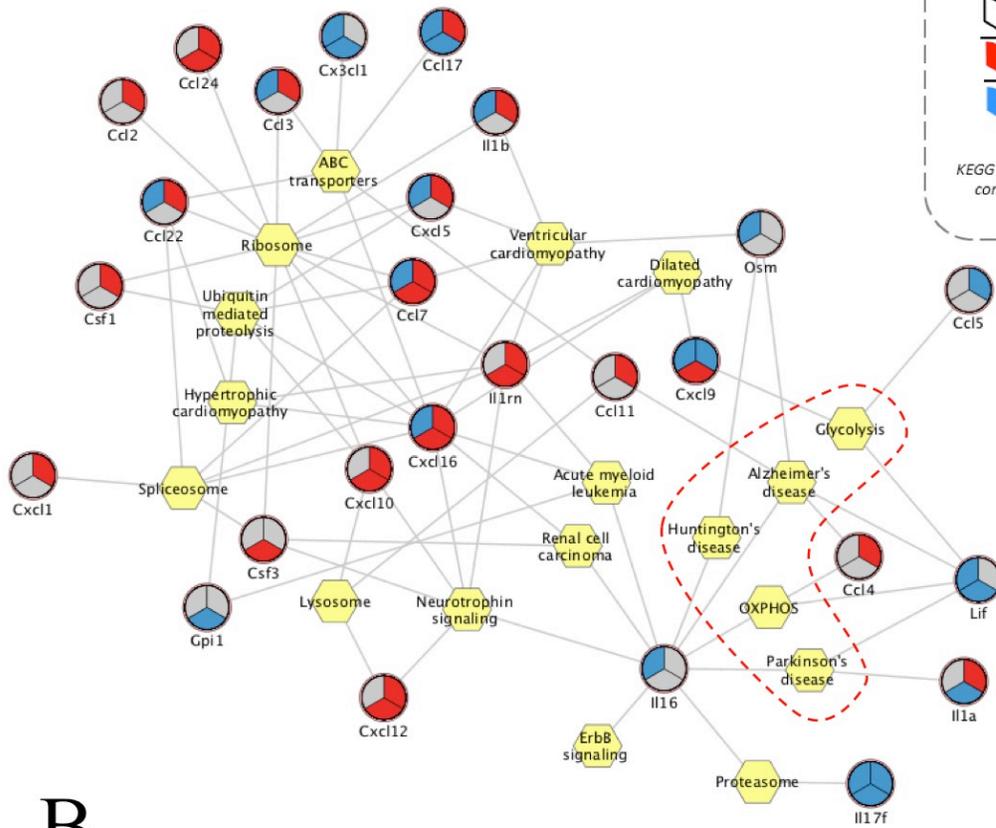
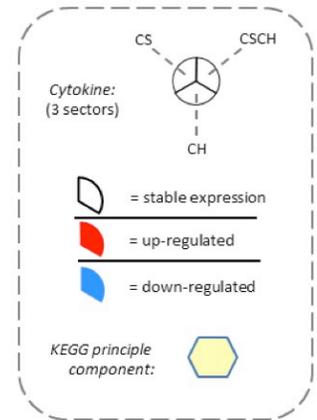
Having demonstrated that the transcriptional state of GP hindlimb muscles following CS exposure represents that of dysfunctional COPD muscle well, we then determined whether we could reverse engineer [108] the structure of a gene regulatory network, linking mRNA abundance of inflammatory mediators expressed in GP lung tissue to indices of the overall KEGG pathway activity in GP hindlimb muscles (see Section 3.3.7 for how these indices were calculated).

Thirty-three of the 72 genes annotated to the cytokine superfamily (46%) were differentially expressed in lung tissue in at least one of the three experimental groups compared to sham-exposed controls and hence used as nodes in the reverse engineering procedure (Table 3-2). We next computed the Spearman correlation coefficients between the profile of expression of each of the 33 candidate factors in lung tissue and indexes of KEGG pathway activity in skeletal muscles. After removing all correlations with  $P >$

0.01, the union of the soluble factors neighbourhoods was visualised using a force driven layout (Figure 3-14).



A

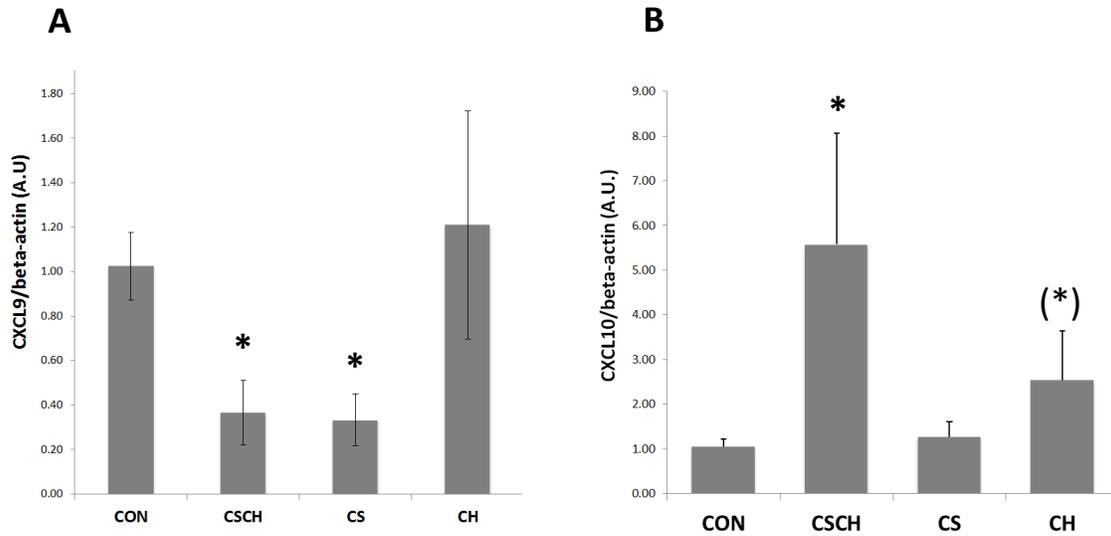


B

**Figure 3-14 | Networks representing the transcriptional coupling between lung and skeletal muscle in the GP smoking model.**

Correlation network linking the expression of soluble factors in GP lung with the transcriptional activity of enriched KEGG pathways in gastrocnemius (**Panel A**) and soleus (**Panel B**) muscles, respectively. Each network edge (grey line) represents the correlation with a given principal component, representative of several genes in each function/KEGG term. Some of the genes will be correlated positively and others will show a negative correlation. Only edges with a bootstrapped  $P$ -value  $< 0.01$  are shown in this figure. Hexagons represent KEGG terms, and circles represent soluble factors that are secreted by lungs. Each circle has been divided into three sectors, representing the CS, CSCH and CH groups. A red sector indicates that the cytokine is significantly up-regulated; a blue sector indicates significant down-regulation; and a grey-coloured sector indicates that the transcriptional level is not significantly affected by the experimental condition.

Analysis of the gastro network (Figure 3-14A) revealed two cytokines each with six edges (*i.e.* Cntf and Cxcl10), indicating that these hub genes could exert an effect on *multiple* pathways within the network. The topological analysis also revealed a dense connected area within the network that was enriched of energy metabolism pathways (OxPhos, TCA cycle and glycolysis). Interestingly, Cxcl10, whose expression level was significantly increased in both the CSCH and CH groups (observed with both microarray and qPCR technology; Figure 3-15), was connected to both members within this energy metabolism dense area of the network that comprised aerobic respiration (*i.e.* OxPhos and TCA cycle). Noteworthy, Cxcl9, which targets the same receptor as Cxcl10, also linked to both oxidative phosphorylation and glycolysis.



**Figure 3-15 | Validation of selected microarray results by real-time RT-PCR.**

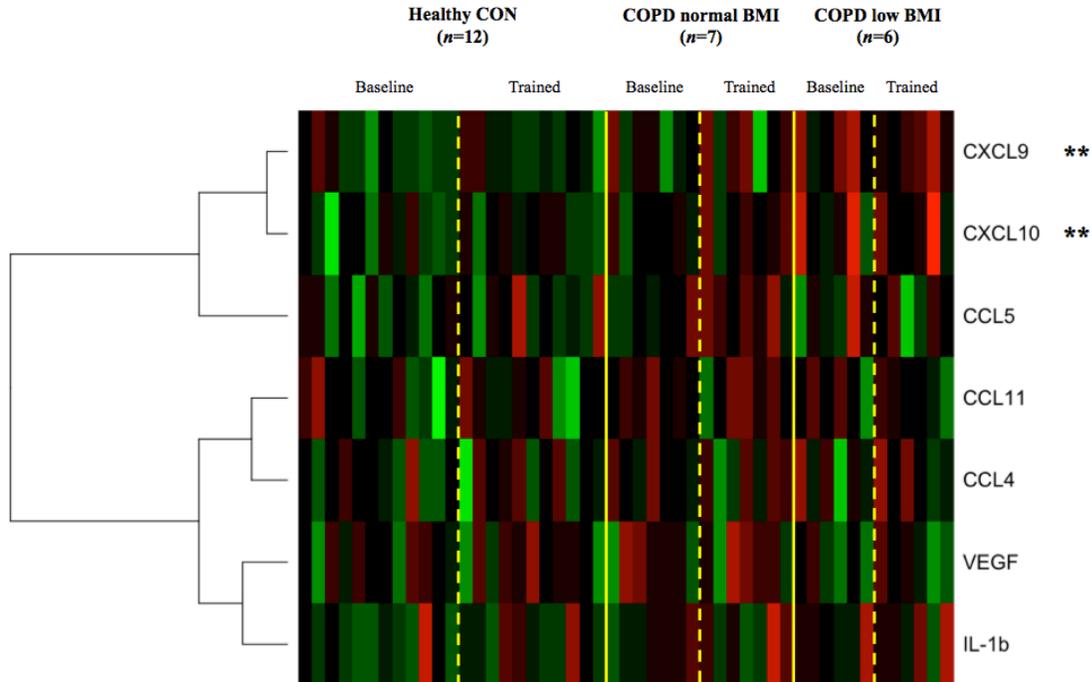
CXCL9 (Panel A) and CXCL10 (Panel B) mRNA levels in lung tissue of sham controls, cigarette smoke-exposed, hypoxic and to combined stimuli. Values (means  $\pm$  SE) in the treated groups are presented as relative to the untreated sham controls (basal level = 1.0). \* Significantly different from the control group ( $p < 0.05$ ). (\*\*) tends to be significantly different among treated ( $0.05 > P < 0.1$ ).

Visual inspection of the soleus network (Figure 3-14B) revealed that Cxcl9 was still linked to bioenergetic processes via the glycolysis pathway. Notably, Cxcl10 was now unconnected to energy metabolism pathways. Instead, this cytokine among others linked to the ribosome component within the network, which had the highest number of connections (11 edges)(Figure 3-14B).

We may thus hypothesise that some of the cytokines we have identified in the GP may also act as systemic signals in COPD patients and be responsible for reducing energy provision in skeletal muscle.

### **3.4.7 Serum cytokine profiling in human COPD patients confirms the predictions of the guinea pig model**

Having shown a remarkable similarity between the transcriptional state of GP and human COPD muscles, we next assessed whether the link between expression of selected pulmonary cytokines and the transcriptional activity of enriched KEGG pathways in peripheral GP muscles was of clinical relevance. For this analysis we took advantage of relevant measures from a COPD serum profiling dataset (*i.e.*, CXCL9, CXCL10, CCL4, CCL5, CCL11, IL1beta and VEGF) used in a previous publication from our group [133] in order to test whether *i)* serum cytokine protein levels were affected by disease and/or prolonged endurance training, and *ii)* if they were correlated with skeletal muscle gene expression. Of the 7 cytokines included, we could indeed verify that the serum protein level of both CXCL9 and CXCL10 were significantly modulated in COPD patients, irrespective of their FFMI, compared to healthy controls (Figure 3-16).



**Figure 3-16 | Heatmap visualization of the protein expression of 7 serum cytokines among healthy controls and COPD at baseline and post training (trained).**

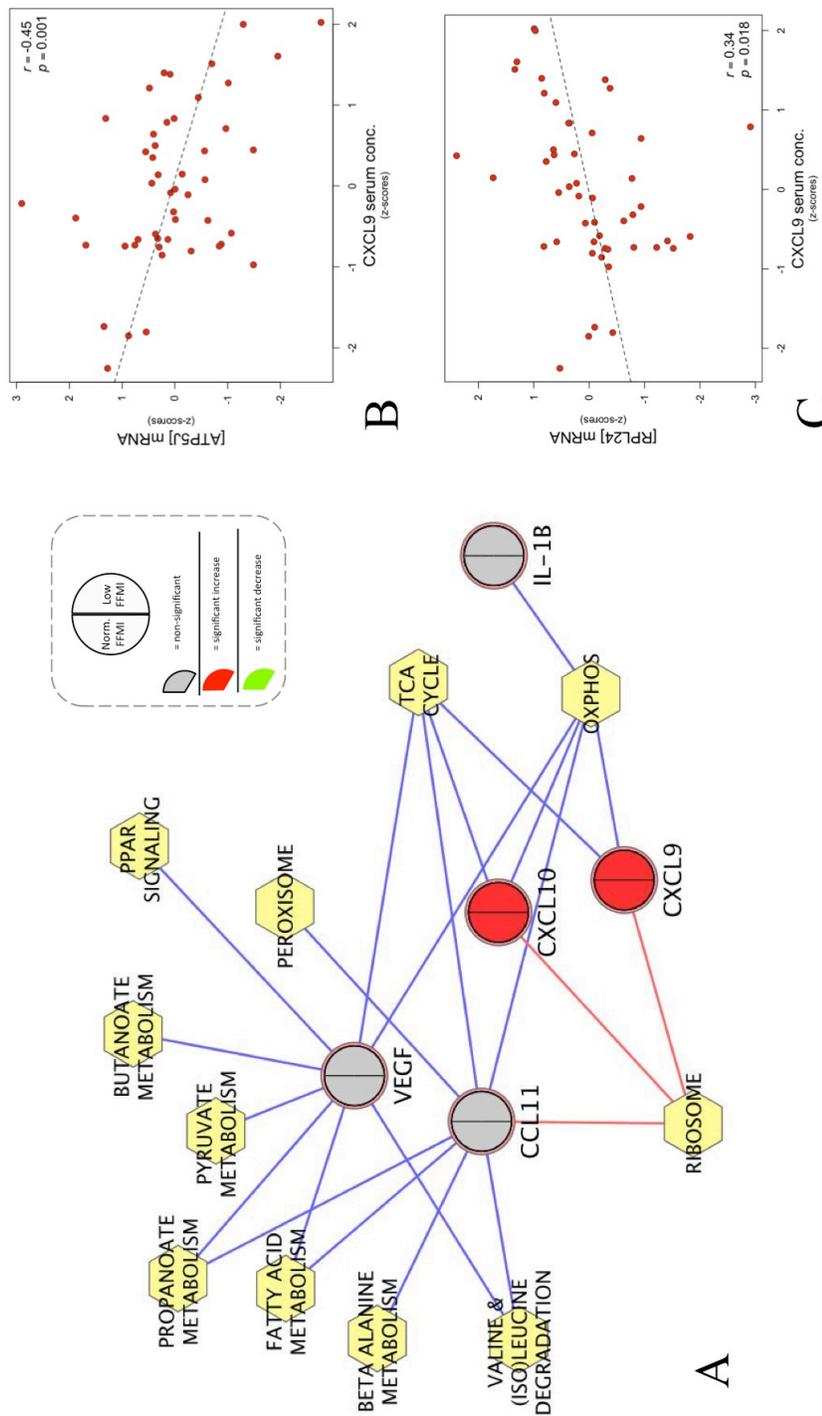
Each row represents a cytokine, whereas each column represents a human subject. Red colours mean increased expression whereas green colours mean decreased expression (relative to the average across subjects). An asterisk denotes significance at  $p < 0.05$  for the disease factor (Mack-Skillings test).

More specifically, CXCL10 human serum protein levels were higher in COPD patients. This elevated expression was consistent with the observed up-regulation of Cxcl10 mRNA in lung tissue from smoke and hypoxia exposed GPs compared to shams (see Table 3-2). However, the increase in CXCL9 human serum protein levels (Fig 3-16) only fit with the increase in Cxcl9 mRNA levels in GP lungs exposed to CH compared to shams (Table 3-2).

### 3.4.8 Training did not modulate any of the tested cytokines

We next correlated all mRNA transcripts expressed above background (12,783 genes) in skeletal muscle in the same cohort with serum cytokine levels and tested whether specific

KEGG pathways were significantly enriched between the positively or negatively correlated genes. Encouragingly, these results were remarkably similar to the GP correlation networks shown in Figure 3-14. The similarity was particularly evident in respect to the inverse correlation between CXCL9 and CXCL10 serum protein levels and the expression of aerobic energy metabolism genes in muscle (OxPhos and TCA cycle) (Figure 3-17).

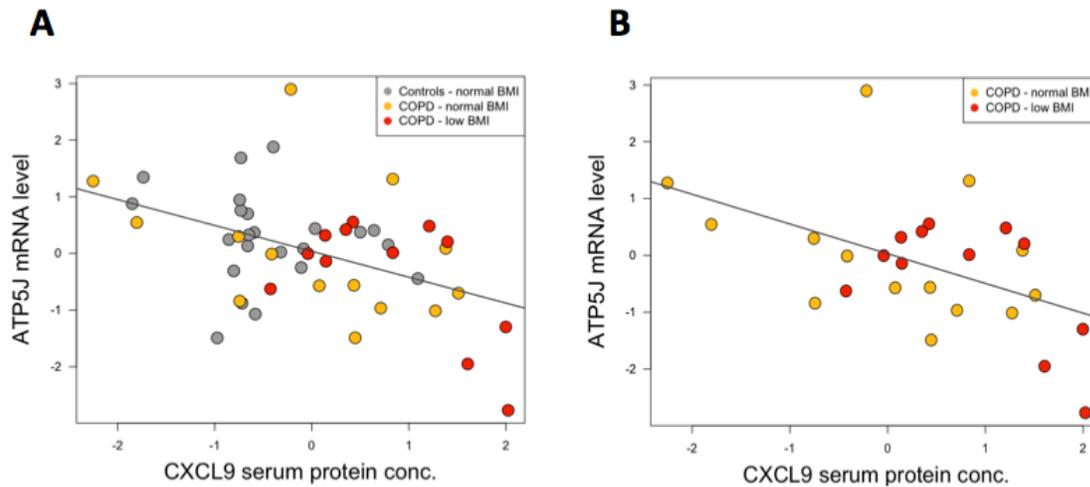


**Figure 3-17 | Network representing the correlation between serum cytokine levels and skeletal muscles gene expression in the human COPD study.**

**Panel A** shows the force-directed network representation linking the serum level of selected chemokines/cytokines (round nodes) with enriched KEGG pathways (yellow hexagons). Each circular node has been divided into two sectors, representing COPD patients with *i)* a normal FFMI and *ii)* a low FFMI. A red sector indicates that the cytokine level is significantly higher compared to matched controls, and a grey-coloured sector indicates that the serum protein level is stable between groups. Blue edges mean negative correlation, whereas red edges mean positive correlation.

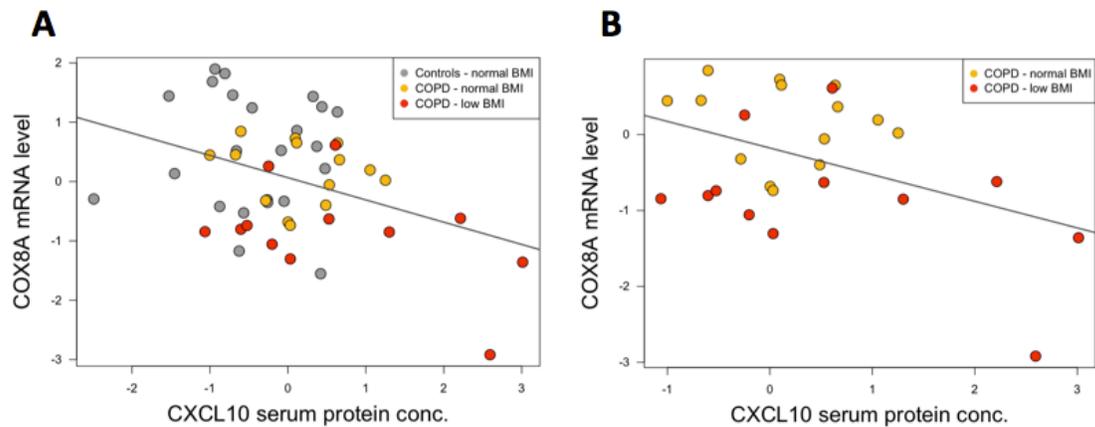
**Panels B-C** are scatterplots representing the association between serum levels of CXCL9 (x-axis) and muscle mRNA expression (y-axis) of examples of genes involved in oxidative phosphorylation (Panel B) and ribosomal biogenesis (Panel C).

Importantly, these strong negative associations to aerobic energy metabolism genes were still present if we ignored the control samples, demonstrating that VO<sub>2</sub>max difference is not a main component of the correlations (Figure 3-18 & Figure 3-19).



**Figure 3-18 | Correlation between standardised CXCL9 serum protein levels and the transcriptional abundance of ATP5J, a gene encoding for a protein in the fifth protein complex of oxidative phosphorylation.**

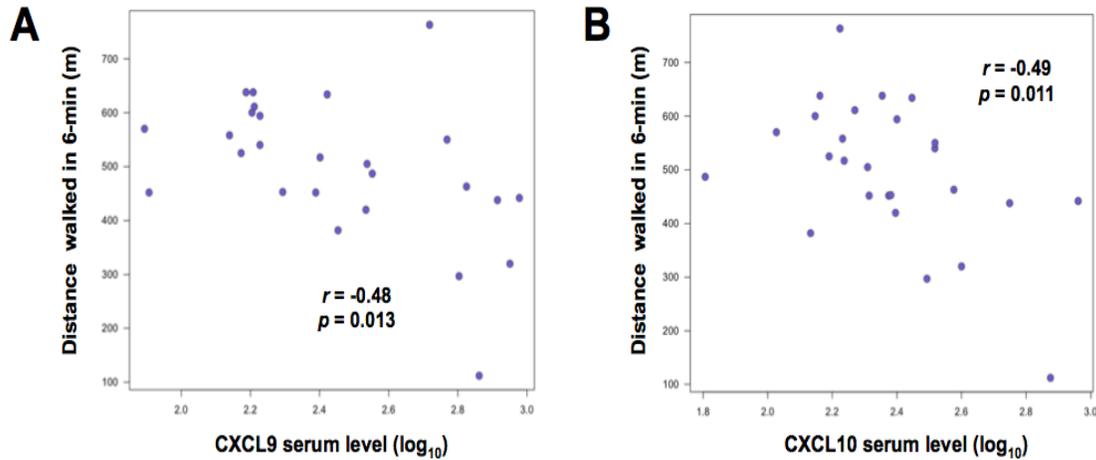
A) All three patient groups were plotted ( $R=46$ ;  $p=0.001$ ). B) When we removed the healthy control samples we still see a significant linear association ( $p=0.01$ ) between CXCL9 and ATP5J expression.



**Figure 3-19 | Correlation between standardized CXCL10 serum protein levels and the transcriptional abundance of COX8A, a gene encoding for a protein in the fourth protein complex of oxidative phosphorylation.**

A) All three patient groups were plotted ( $R=0.37$ ;  $p=0.008$ ). B) When we removed the healthy control samples we still see a significant linear association ( $p=0.03$ ) between CXCL10 and COX8A expression.

Finally, we found a statistically significant negative association between CXCL9 and CXCL10 serum protein levels and the distance walked in 6 minutes, an objective measure of functional exercise capacity.



**Figure 3-20** | Scatterplots highlighting the clinical association (Spearman correlation) between distance walked in 6-min and serum levels of CXCL9 (Panel A) and CXCL10 (Panel B), respectively

### 3.5 Discussion

The most important finding of this study is the discovery that mRNA levels in CS-exposed GP lungs, as well as human serum protein levels of CXCL9 and CXCL10, are significantly inversely correlated to the expression of aerobic energy metabolism genes in skeletal muscles. In addition, we demonstrate that the GP smoking model can mimic many of the transcriptional changes observed in limb muscle of atrophic COPD patients, making this an extremely useful *in vivo* experimental model.

#### 3.5.1 The role of systemic inflammatory cytokines in controlling the molecular state of skeletal muscles

Several human studies have demonstrated that COPD is associated not only to inflammation of the lungs, but also with increased levels of circulating pro-inflammatory cytokines [174, 175]. Notably, there is a clear trend toward significant induction of TNF- $\alpha$  plasma levels of CS-exposed guinea pigs, suggestive of a similar systemic inflammatory process in this model organism [158]. Although elevated levels of pro-

inflammatory cytokines have previously been linked to skeletal muscle dysfunction [176], until now it was unclear whether they are the primary factor driving muscle wasting.

The analyses we have performed provide the first evidence that expression of systemic cytokines significantly correlates with expression of energy metabolism genes (represented by the OxPhos, TCA and glycolytic pathway) in limb muscles (Figure 3-14 & Figure 3-17). This is consistent with the previous observation that dysfunctional limb muscle of COPD patients are unable to co-ordinate the expression of energy metabolism genes [133].

Importantly, there is strong evidence that most of the candidates we have identified in the GP model are also modulated in COPD patients. For example, it has previously been shown that serum levels of CCL24, CSF1 and IL1A (among others) are significantly higher in COPD patients compared to matched controls [177]. Our own validation of selected model predictions demonstrated that serum protein levels of two CXCR3 chemokines (CXCL9 and CXCL10) are significantly higher in the same COPD cohort used for muscle mRNA profiling, and their levels negatively correlate with the expression of energy metabolism genes in human COPD skeletal muscle (see Ox and TCA cycle) (Figure 3-17).

A). Consistent with this observation (see also [178]) we found a statistically significant negative association between CXCL9 and CXCL10 serum protein levels and the distance walked in 6 minutes (Figure 3-20).

Noteworthy, exercise training did not modulate serum levels of these chemokines, indicating that differences in the level of physical activity overall will not affect serum levels of these chronic inflammatory mediators.

Overall, this strongly suggests that systemic inflammation plays a major role in promoting skeletal muscles dysfunction in both smoking GP and human COPD patients.

### **3.5.2 Biological significance of the transcriptional response to smoking and/or hypoxia in lungs and muscles**

In accordance with the (pre)clinical literature, we identified a massive transcriptional response in whole lung tissue (Figure 3-9). As anticipated, the result of the functional enrichment analysis suggests that long-term CS-exposure induces pathways related to the antioxidant defence system (*i.e.* glutathione and cytochrome P450 drug metabolism), most likely in order to try and cope with the increased ROS production. Such responses are consistent with the changes observed in the human airway transcriptome of chronic smokers [179, 180] (Figure 3-13).

Interestingly, the two skeletal muscles showed a distinct pattern of response both at the gene and pathway level, likely reflecting differences in fibre type composition, and hence metabolic profile, between these two hindlimb muscles. Soleus is a postural, primarily oxidative muscle, with a high Type I fibre content, whereas the gastrocnemius is a phasic muscle of mixed Type II composition that is predominantly glycolytic.

Our data suggest that CS *per se* primarily affects the soleus (Figure 3-9) similar to a recent finding in C57 mice, where contractile properties were selectively affected in soleus following chronic CS exposure using a nose-only device [154]. Furthermore, only the soleus showed a tendency towards lower muscle mass after mice had been whole-body exposed to CS for 6 months [153].

The result of the pathway level analysis in GP soleus clearly suggests that long-term smoking modulates the activity of the MAPK pathway as induced signalling pathways (*i.e.* ErbB signalling and apoptosis) represent components of this pathway. This agrees with a recent human study in which several key members of the MAPK pathway were up-regulated at the mRNA level in *vastus lateralis* of patients with COPD [181]. In addition, aerobic energy metabolism is also clearly hit in the CS-group, which is also consistent with findings in skeletal muscle of human COPD patients [182].

One puzzling finding in the current study is that hypoxia appears to exert a ‘protective’ transcriptional effect in the soleus, as only a few enriched pathways were found when the hypoxic challenge was added to the CS intervention. A density plot of fold-changes revealed that CS as a single factor exerts a greater transcriptional effect than either of the two other experimental conditions where CH is present, among the transcripts with an absolute fold-change above 1.4. The reason for this is not clear and highlights the need for further research in this area.

In addition to the demonstrated agreement with published human transcriptional data, we also addressed the clinical relevance of the present animal model by comparing the transcription-based functional profiles related to each experimental challenge with the

functional signature derived from stable yet severe COPD patients when compared to healthy age matched controls. In accordance with results from our gene-level analysis and phenotypic data from the mouse smoking model, the response of the soleus to long-term smoke exposure best mimics expression signatures linked to the effects of COPD in human limb muscle, with 68% KEGG terms in common (Figure 3-12A). However, exposure to CH in *gastro* also yielded a highly statistically significant overlap with the clinical dataset as highlighted by a 53% functional overlap.

### **3.5.3 Gene expression profiling as a tool to assess animal model relevance of human disease**

Most pre-clinical models of CS-induced COPD have for obvious reasons focused on the lung component of the disease [183]. However, with the increasing awareness of the clinically important extrapulmonary manifestations linked to COPD, a number of studies have now begun to try to elucidate the mechanisms governing skeletal muscle dysfunction, whether or not accompanied by loss of muscle mass. Consistent with COPD, previous studies have shown marked reduction in body weight gain [157, 158] as well as increased oxidative stress [147] in GP hindlimb muscles following only 3 months of daily CS exposure. In contrast, macroscopic data suggest that mouse models poorly reproduce the systemic effects of human COPD. For example, long-term exposure to CS only induce *mild* effects in selected skeletal muscles, as defined by fibre redistribution and altered oxidative enzyme activity [153, 154], despite both studies using much longer exposure protocols than that of the present study.

It may be important to assess and compare the relevance of the current GP model with that of other rodent CS models such as the mouse. Although an extensive microarray

analysis of such model has not been published, we have been able to retrieve a dataset from the GEO database (GSE18033). In order to comment on the suitability of this dataset to address this important issue, I analysed this data using the same approach described in this paper (see Appendix 7-3 for the full data analysis).

The experiment performed was limited to the gastrocnemius, which our analysis in the guinea pigs suggests is not the most representative of human COPD. However, before any comparison could be made, our analysis only identified 24 genes that were differentially modulated by chronic CS-exposure (24 weeks) in hindlimb gastrocnemius muscle compared with time-matched sham controls at a reasonable statistical threshold ( $FDR < 15\%$ ). Only by raising the statistical cut-off to 30%, which increased the number of regulated transcripts to 1,020 (Table 3 in Appendix 7-3), could we detect biologically relevant functions, although the maximum sensitivity of 0.11 was substantially lower than any of the experimental conditions involving CS-exposure in the current GP model (Figure 3-12A). Although further analysis of the mouse model is required for reaching any definitive conclusion, these results indicate poor transcriptional response following smoking exposure in mice.

#### **4 Using an integrative framework combining association data with skeletal muscle gene expression predicts endurance training-induced changes in (whole-body) insulin sensitivity: The HERITAGE Family Study.**

*Up to now, this thesis has focused on identifying genes and molecular pathways involved in mammalian peripheral skeletal muscle wasting. Exercise training is essential to skeletal muscle performance and represents a promising intervention strategy.*

*However, of great medical importance, clinical trials have clearly highlighted a heterogeneous ability among participants to physiologically adapt to exercise training (despite being fully standardized and monitored).*

*A major challenge in the analysis of genomic data is to develop predictive statistical models [75]. The next two chapters attempt to address this important clinical aspect.*

*The work presented in this chapter represents a collaborative project between the labs of Francesco Falciani and Claude Bouchard (Pennington Biomedical Research Center, US). The Bouchard lab provided the raw GWAS SNP and mRNA microarray data as well as anthropometric characteristics for all subjects in the HERITAGE exercise training cohort. I have solely conducted all in silico analyses presented throughout the chapter.*

## 4.1 Introduction

Endurance exercise training is a strong physiological stimulus that plays a large role in skeletal muscle homeostasis. It leads to a multitude of functional improvements when performed regularly (*i.e.* training). These include substantial health benefits such as prevention and/or treatment of onset of Type II diabetes [184, 185] by increasing tissue responsiveness to circulating insulin.

Skeletal muscle contraction and peripheral insulin action are highly inter-twined [186]. In fact, up to eighty per cent of the *in vivo* insulin-mediated glucose disposal occurs in skeletal muscle [187], making it a quantitatively important organ.

However, we and others have demonstrated that healthy individuals show a marked heterogeneous ability to improve their peripheral insulin sensitivity (SI) in response to endurance exercise training ( $\Delta$ SI). Notably, a significant percentage of individuals (~20%) show no change in SI and some even demonstrate an adverse response [188, 189].

The molecular mechanisms underlying this heterogeneous ability to improve SI through regular exercise, which likely includes a substantial genetic component [190], are currently not well understood. In support of a molecular basis of this heterogeneous effect, we previously found that sex- and age-matched healthy individuals with high and low SI responses to endurance exercise training have different skeletal muscle gene expression patterns at baseline [188].

These observations led to the hypothesis that training-induced changes in SI may in part be determined by genetic factors. Here, we address this question by applying a computational framework aimed at identifying the chain of events linking genetic

variation in the form of single nucleotide polymorphisms (SNPs) to skeletal muscle gene expression and to phenotypic response ( $\Delta$ SI).

Remarkably, our integrative approach revealed that common genetic variants in close proximity to genes representing calcium signalling and carbohydrate metabolism associate with basal mRNA abundance that are quantitatively predictive of  $\Delta$ SI. We also demonstrate that a significant proportion of the transcriptional variation linked to responder status can be explained by an overall basal difference in the activity of the MEF2 calcium-dependent transcription factor, a well known regulator of skeletal muscle metabolism.

In conclusion, our results provide the first plausible hypothesis on a molecular mechanism underlying the heterogeneous ability to improve SI with training. Moreover, we also provide an independently validated quantitative model, based on a specific functional transcriptional signature that can predict individual SI training response.

## 4.2 Research Design and Methods

### 4.2.1 HERITAGE Family Study

**Participants.** The study design and exercise training protocol of the HERITAGE Family Study have been described elsewhere [191]. Briefly, 834 subjects from 218 families of Blacks and Whites were recruited to participate in an endurance exercise training study. Among them, 483 adults from 99 families of European descent were defined as completers. Blacks are not considered in this and the following chapter. Parents were 65 yr of age or less while offspring ranged in age from 17 to 41 years. Participants were sedentary at baseline, normotensive or mildly hypertensive (<160/<100 mm Hg) without medications for hypertension and dyslipidemia [191]. Most subjects were normoglycemic at baseline but 55 were defined as having impaired fasting glucose (100 mg/dL or 5.6 mmol/L and more but less than 126 mg/dL or 7.0 mmol/L), and 4 were in the diabetic range with fasting glucose of 126 mg/dL or 7.0 mmol/L and more. None of the subjects were on hypoglycemic medications. The study protocol was approved by the Institutional Review Boards at each of the five participating centers of the HERITAGE Family Study consortium. Written informed consent was obtained from each participant.

**Exercise training program.** Each volunteer in HERITAGE exercised three times per week for 20 weeks on cycle ergometers controlled by direct heart rate (HR) monitoring. Details of the exercise program can be found elsewhere [191]. Briefly, subjects exercised at the HR associated with 55% of baseline maximal oxygen uptake ( $VO_{2max}$ ) for 30 minutes per session for the first 2 weeks. The duration and intensity were gradually increased every 2 weeks, until reaching 50 minutes and 75% of the HR associated with

baseline VO<sub>2</sub>max. This level was maintained for the final 6 weeks of training. The protocol was standardized across all clinical centers and supervised to ensure that the equipment was working properly and that the participants were compliant with the protocol.

***Intravenous glucose tolerance test (IVGTT) protocol.*** A frequently sampled IVGTT was performed in all participants after an overnight fast (12 h), at baseline and post-intervention (24-36h after the last exercise bout). In premenopausal women, the test was scheduled to coincide with the follicular phase of the menstrual cycle. The protocol previously defined by Walton *et al.* was followed for the IVGTT [192]. The protocol did not include an injection of insulin or tolbutamide. From the IVGTT data, an SI index (mU / [L×min]), which measures the ability of an increment in plasma insulin to enhance the net disappearance of glucose from plasma and is used as a measure of insulin sensitivity, was derived using the MINMOD Millennium software [193]. Percent changes in SI ( $\Delta$ SI) were calculated as follows:  $([SI_{\text{post}} - SI_{\text{pre}}]/SI_{\text{pre}}) \times 100$ .

#### **4.2.2 GWAS analysis:**

SNP genotyping (~325,000 SNPs, Illumina Human CNV370-Quad v3.0 BeadChips) on genomic DNA from permanent lymphoblastoid cells was done as previously described [194].

SNPs used in association analyses were filtered according to the following criteria: (a) minor allele frequency (MAF) less than 5%, (b) violated Hardy-Weinberg equilibrium (p-value > 0.1), (c) missing values in >10% of individuals, (d) located more than 20kb away from a gene coding sequence, (e) SNP associated gene not expressed in human skeletal

muscle (see below for more details), and (f) redundant SNPs due to strong pairwise linkage disequilibrium ( $R^2 > 0.8$ ; PLINK v1.07 [195]). These criteria reduced the list of candidate SNPs to 68,595 (a 79% reduction).

Genome location of each SNP is based on the latest build in dbSNP (GRCh38/hg38)

SI training response was adjusted for age, weight-adjusted VO<sub>2</sub>max, and baseline SI within sex-by-generation subgroups. Associations between the normalized trait residuals and SNP genotypes were analysed using additive linear mixed effect models that accounted for within family correlations (function *lme* of the ‘nlme’ R package).

#### **4.2.3 RNA extraction & global gene expression profiling:**

Muscle biopsies were taken before and after (~96 h after final training session) exercise training intervention using a percutaneous needle. Skeletal muscle RNA extraction as well as reverse transcription were done as previously described [196]. Affymetrix U133+2 arrays were used to quantitate global mRNA expression levels. The raw microarray CEL files are deposited in the public Gene Expression Omnibus (GEO) database [197] under accession number GSE47874.

#### **4.2.4 Microarray analysis:**

Raw CEL files were Robust Multichip Average (RMA) normalized following removal of probes that were termed ‘absent’ in more than 80% of the samples by the MAS5 algorithm inside the ‘affy’ package (26,151 probesets discarded) [198].

Quality control plots of the polyA-control RNAs (spike-ins added right after RNA purification) highlighted a batch issue/problem that was resolved by applying the ComBat software [79].

The JetSet R package was used to select a single ‘optimal’ probeset to represent each gene based on specificity, robustness against mRNA degradation and MAS5 present call rate [63].

*Validation exercise training dataset:*

We tested our findings in an independent training study, for which Affymetrix Gene 1.1 ST microarray data are publically available (accession: GSE53598), consisting of 15 overweight to obese normoglycemic middle-aged men that underwent 12 weeks of mixed exercise training (2 days/wk aerobic, 1 day/wk resistance training). Demographic data on the subjects as well as the exercise training protocol have been previously described [199]. Whole-body SI was measured using the hyperinsulinemic-euglycemic clamp technique (40 mU/m<sup>2</sup> per min), as previously described [200].

Raw microarray data were downloaded from GEO and .CEL files were RMA normalized for ‘core’ transcripts using the Bioconductor ‘oligo’ package (v. 1.28.3) [201]. RefSeq IDs corresponding to the affy probes were obtained using the ‘hugene11sttranscriptcluster.db’ annotation library in R (v. 8.3.1).

**4.2.5 Pathway-level analyses:**

Biological pathway information was obtained from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [128], and downloaded from the Molecular Signatures

Database collection ([www.broadinstitute.org/gsea/msigdb/index.jsp](http://www.broadinstitute.org/gsea/msigdb/index.jsp)) [173]. To maximise biological interpretability, as previously described [131], the analysis was restricted to KEGG group 1-4 (Metabolism, Genetic Information Processing, Environmental information Processing and Cellular Processes) as well as Endocrine and Circulatory system pathways in group 5. In addition, gene sets with <20 or >200 genes were excluded to protect against spurious associations from very small pathways or the lack of sufficient specificity from very large pathways [202]. A total of 110 pathways were included.

#### *GSEA Pre-ranked analysis:*

The entire skeletal muscle transcriptome was ranked by individually regressing their basal expression against  $\Delta SI$ , with age, gender and type 1 fiber composition included as covariates.

Based on this ranking (Student's *t*-statistic) we performed a pre-ranked gene set enrichment analysis (GSEA; v2.0.12) [84], using the default parameters, to identify candidate gene-sets significantly enriched in genes that are significantly associated with  $\Delta SI$  (either towards the top or bottom). The *a priori* defined gene-set collections used were transcription factor targets (c3.tft.v4.0symbols.gmt) and KEGG pathways, respectively (see above).

#### **4.2.6 Data analysis:**

In order to link gene expression to  $\Delta SI$  we used a regression-based modelling approach allowing for pairwise interactions (function *lm* of the 'stats' R package). More precisely, we define:

$$\Delta SI = a\theta_1 + b\theta_2 + c\theta_3 + d\theta_1\theta_2 + e\theta_1\theta_3 + f\theta_2\theta_3 + sex + VO2_{max} + \varepsilon$$

where mRNA abundance are represented by  $\theta$  and  $\varepsilon$  is the noise model component.

Weight-adjusted aerobic capacity and gender were included as covariates.

## **4.3 Results**

### **4.3.1 Overview of the analysis strategy:**

Identification of prognostic exercise-sensitive biomarkers, ideally of circulating factors, that can tell who will demonstrate the largest improvement in SI is a challenging task.

As outlined in Figure 4-1, this study addresses this issue by applying a proof-of-principle multi-step data mining approach. Due to the multi-omics nature of the HERITAGE study, we took on the challenge of overlaying an individual's genetic background with the tissue-specific gene expression profile against the improvements in whole-body insulin sensitivity to develop a robust muscle-derived RNA-based classifier.

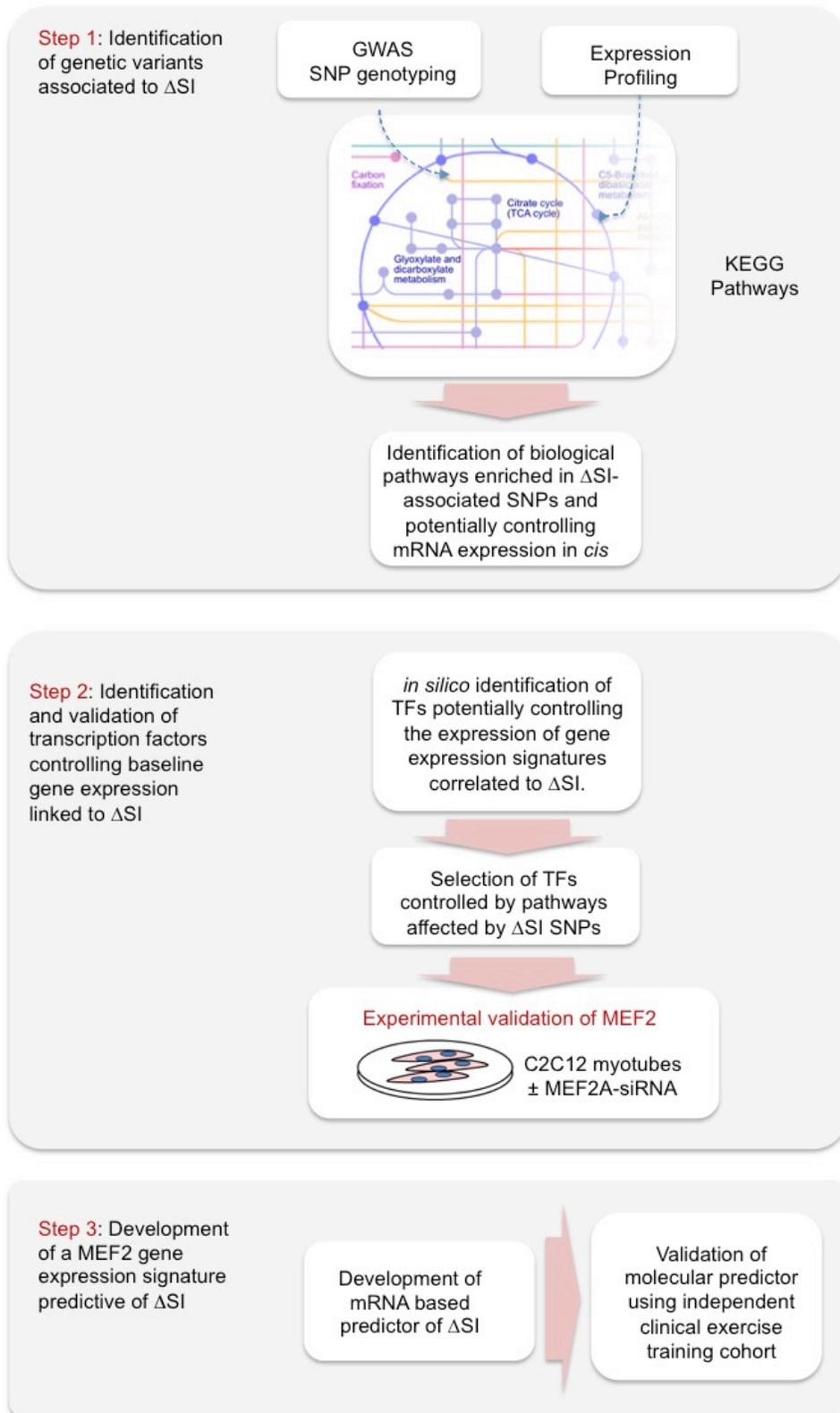
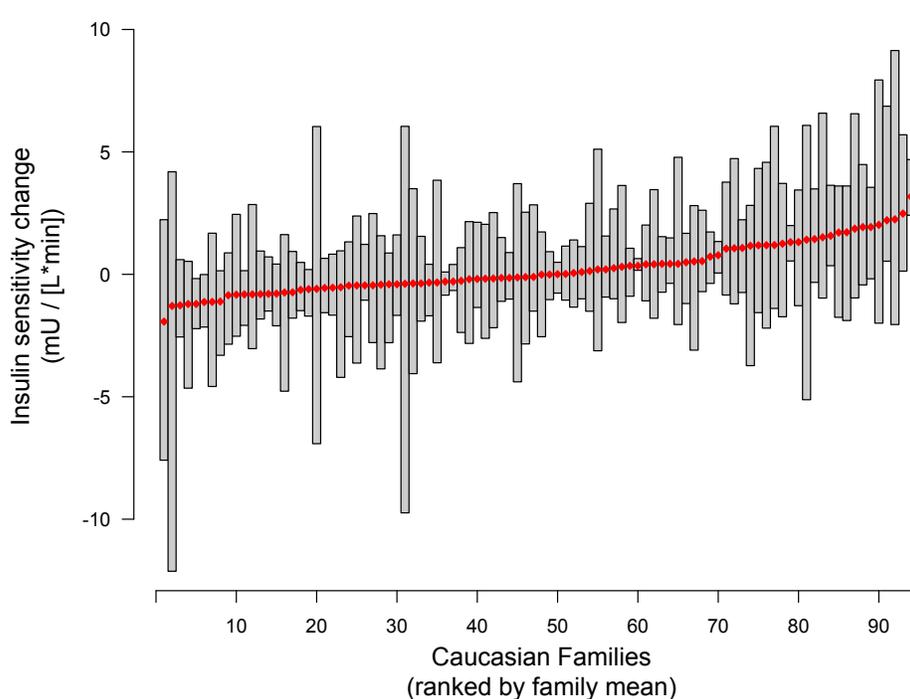


Figure 4-1 | Flowchart highlighting the analysis strategy.

### 4.3.2 Improvement potential in SI has a clear heritable component:

The overarching objective of this study is based on the assumption that molecular mechanisms underlying individual variation in  $\Delta$ SI have a genetic component. We tested this hypothesis by using the HERITAGE Family study, one of the largest endurance training studies to date. Our analysis revealed that 29% of the variance associated to  $\Delta$ SI is accounted for by family membership (Figure 4-2). Moreover, the ANOVA model estimated that there was 40% more variance between families than within families ( $p=0.02$ ), providing first time evidence that the changes in SI in response to exercise training are characterized by a significant genetic component.



**Figure 4-2 | Training-induced change in insulin sensitivity plotted against Caucasian family rank (*i.e.* families ranked by family mean) in the HERITAGE study.**

Only families with at least 3 members are shown (94 families in total); family average was 4.3 members. Each vertical box represents the range of training responses across all family members (parents and offspring). Red horizontal reference line denotes family mean. The F-value from the ANOVA indicates that there is 40% more variance between than within families ( $p=0.02$ ), with 29% of the variance being accounted for by family membership.

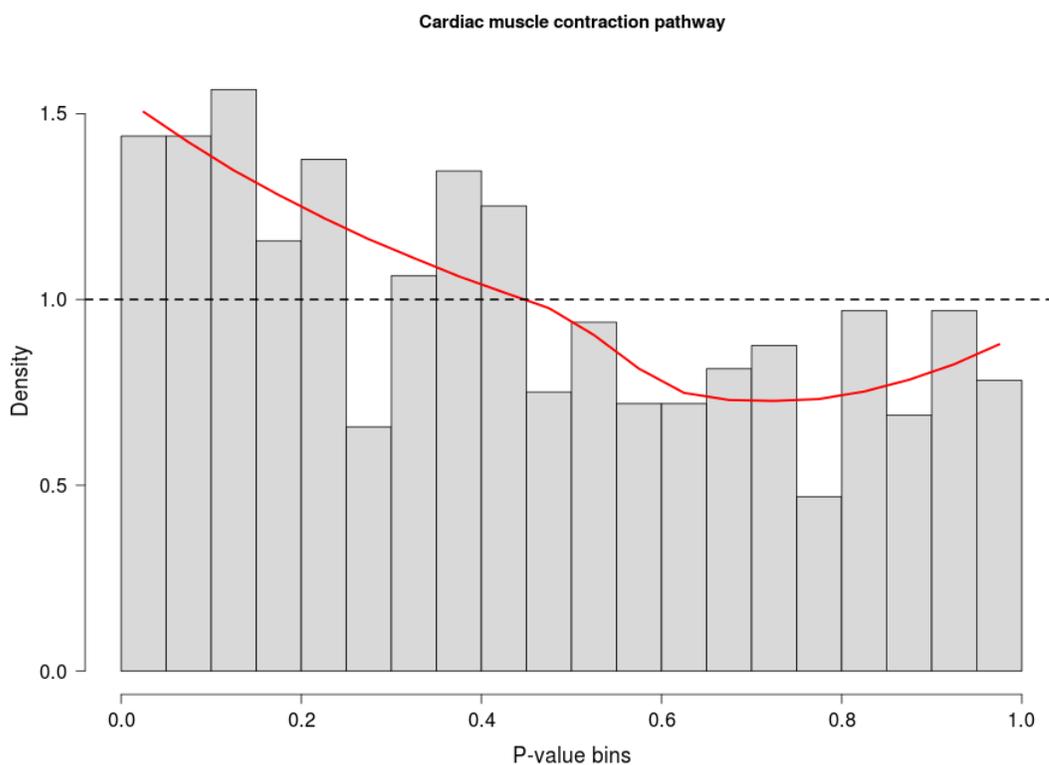
### 4.3.3 Functional GWAS identifies an association between $\Delta$ SI, carbohydrate metabolism and calcium signalling pathways:

Having demonstrated that changes in SI following endurance training show evidence of a genetic component, we set to identify the specific pathways enriched in genetic variants (in the form of SNPs) associated to  $\Delta$ SI. We approached this challenge using a validated analysis technique (functional GWAS), which has been shown to be particularly effective in studies with a relatively small sample size ( $N=424$ ) [203]. This methodology, which considers the effects of multiple SNPs jointly (both within a gene as well as between genes within a pathway), tests whether SNPs that are most associated to  $\Delta$ SI are enriched in specific functional pathways.

Nine KEGG pathways were significantly ( $P_{adj}<0.05$ ) associated with SI improvement potential (Table 4-1 and Figure 4-3 for an example). Overall, the enriched pathways represent three functional categories: calcium signalling, carbohydrate metabolism, and cell communication (see Table 4-1).

KEGG pathway	Gene count	SNP count	FDR	Functional group
Cardiac muscle contraction	77	639	<0.001	Calcium
Inositol phosphate metabolism	57	463	0.017	Calcium
Arachidonic acid metabolism	59	200	0.023	Calcium
Galactose metabolism	27	130	<0.001	Carbohydrate
Starch and sucrose metabolism	55	140	0.002	Carbohydrate
Pentose phosphate pathway	27	132	0.017	Carbohydrate
Fructose and mannose metabolism	36	207	0.017	Carbohydrate
Focal adhesion	200	1575	0.023	Cell communication
Adherens junction	73	658	0.026	Cell communication

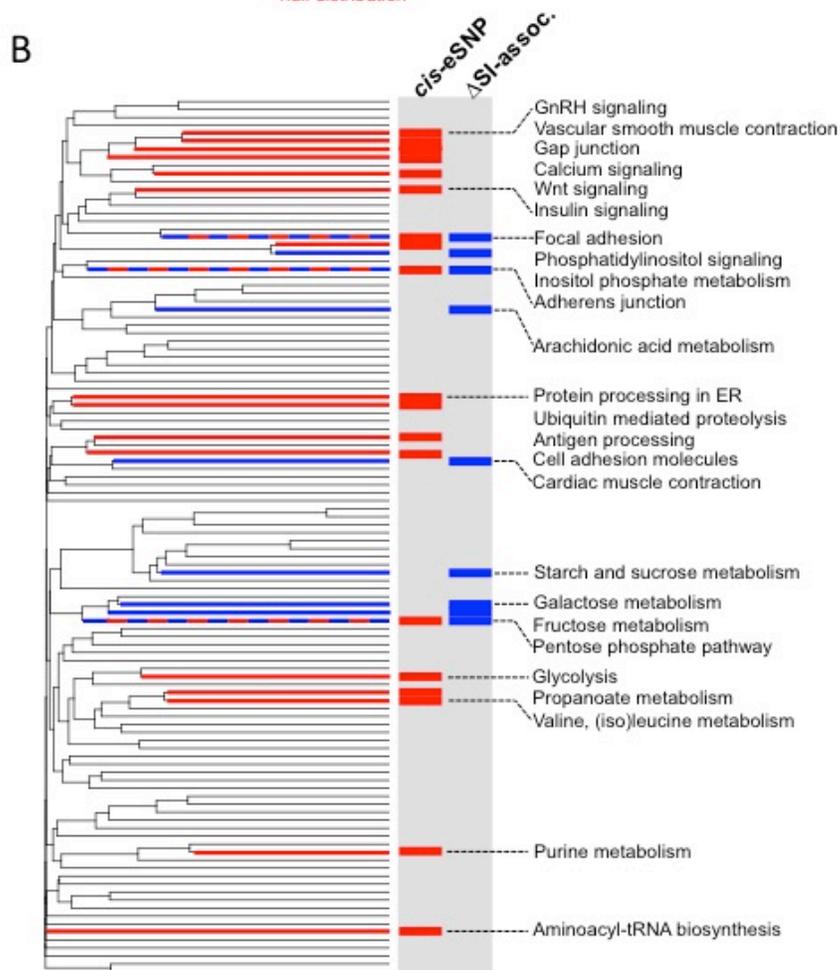
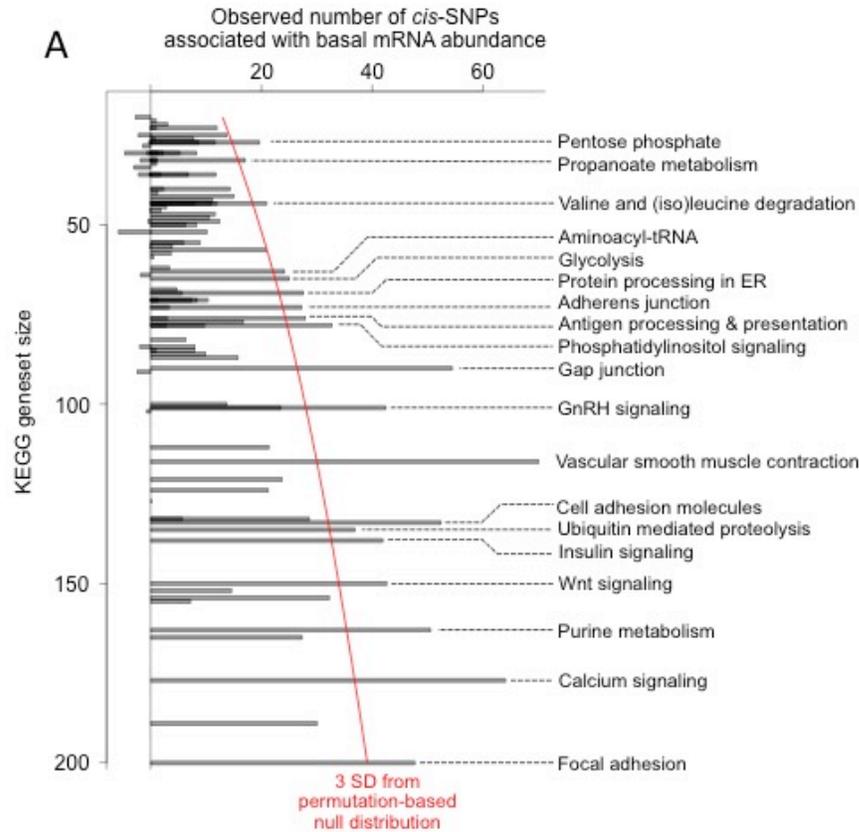
**Table 4-1 | KEGG pathways enriched for DNA variants associated with  $\Delta$ SI.**



**Figure 4-3 | Histogram of the p-value distribution for SNPs flanking genes annotated to the ‘cardiac muscle contraction’ KEGG pathway.**  
The red line is a fitted density curve.

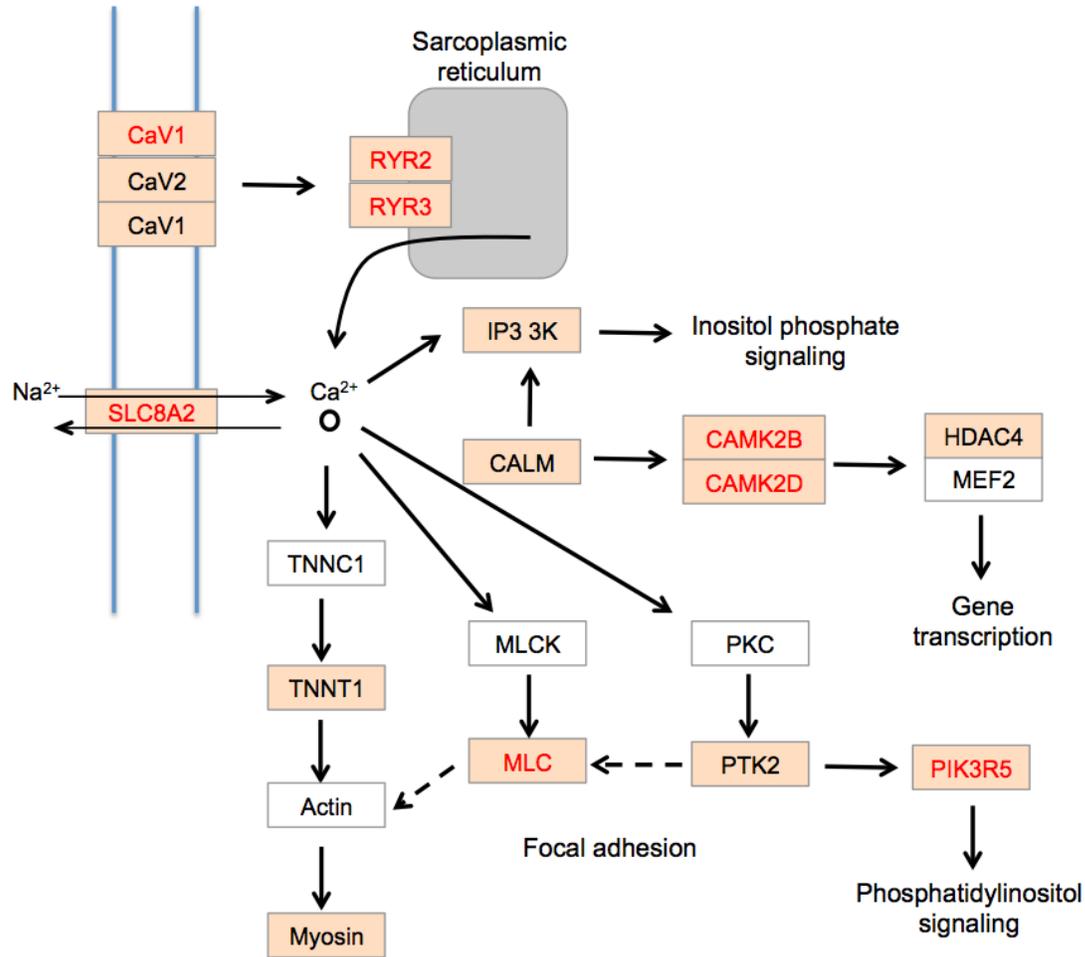
As nearly 90% of all GWAS loci fall outside of protein-coding regions, we sought to identify SNPs that associate with basal mRNA abundance (so-called ‘expression SNPs’) among the subset of White HERITAGE participants (n=47) for which global gene expression data is available using the method detailed by Schadt *et al.* [204]. To refine the interpretation, all 1,280 nominally associated eSNPs (Kruskal-Wallis  $p < 0.01$ ) were mapped to genes with KEGG pathway annotation. As shown in Figure 4-4A, 19 out of 110 KEGG pathways tested had more significant eSNPs associated to them than expected by chance ( $p < 0.01$ ), indicating that mRNA abundance of a significant subset of genes (up to 27%) within these enriched pathways is associated to common DNA sequence variants

with potential regulatory effects. Noteworthy, calcium signaling and vascular smooth muscle contraction were the most enriched pathways.



**Figure 4-4 (see previous page) | Panel A:** histogram showing the observed number of SNPs associated with basal mRNA abundance for each KEGG pathway. The red line represents 3 standard deviations from the permutation-based null distribution. Since KEGG gene-sets are not independent, the overlaps between significant ones were assessed using the Jaccard's index of similarity. From **Panel B** it is clear that many of the enriched gene-sets can be grouped into three overall functional categories: carbohydrate metabolism, cell communication, and calcium signalling. It is of note that multiple terms within 'cell communication' have a strong calcium signalling component such as gap junction, vascular smooth muscle contraction and Wnt signalling

Based on the results presented in Table 4-1 and Figure 4-4A we decided to focus on calcium signalling as it represents the most upstream component of the pathways discovered by the functional GWAS analyses. Initially, we constructed an integrated pathway that connected the calcium-related pathways in Table 4-1 at the gene level. We then mapped all SNPs to genes within this high-level pathway (Figure 4-5). Visual inspection of the resulting diagram was fully consistent with the result of the GWAS analysis, showing that  $\Delta$ SI-associated genetic variants cluster around the cascade of events that link calcium mobilization to the transcriptional response mediated by transcription factor targets of calmodulin-dependent protein kinases (*e.g.* MEF2s, HDACs; see Figure 4-5).



**Figure 4-5 | Diagram representing the detailed relationships between the calcium-related KEGG pathways identified in Table 4-1.**

Genes with an orange background contain DNA variants (within a 20-kb flanking window on either side of the gene) significantly associated to ΔSI. Further, the basal mRNA abundance of genes with a red font are significantly associated with DNA variants.

Interestingly, we also discovered that SNPs identified by our functional GWAS analysis were significantly associated with gene expression variability ( $p < 0.05$ ), suggesting that these have a true regulatory role (Figure 4-5 and Figure 4-5).

#### **4.3.4 The calcium-dependent MEF2A transcription factor drives the transcriptional signature associated to $\Delta$ SI**

The functional pathways enriched in  $\Delta$ SI-associated SNPs represent the initial molecular events that eventually may influence and drive phenotypic response. It is reasonable to expect that downstream events, including changes in mRNA expression/abundance, may be involved in this process. We therefore set to identify transcription factors for which the basal expression of a significant proportion of their target genes is correlated to the percent change in SI post-training. More specifically, by using the transcript abundance of sets of transcription factor (TF) target genes (*i.e.* group of genes sharing a specific TF binding site in the promoter region) we sought to infer the activity of a compendium of TFs at baseline using a GSEA approach.

The gene targets of two transcription factor families, namely splicing factor 1 (V\$SF1\_Q6) and myocyte enhancer factor (MEF)-2 (V\$MEF2\_04), were positively associated with  $\Delta$ SI (FDR < 5%), highlighting that individuals exhibiting high responsiveness of SI to training have an overall higher basal expression of genes co-regulated by SF1 and MEF2, respectively, compared to those demonstrating an adverse response (see Appendix 7.4 for a comprehensive overview of the GSEA).

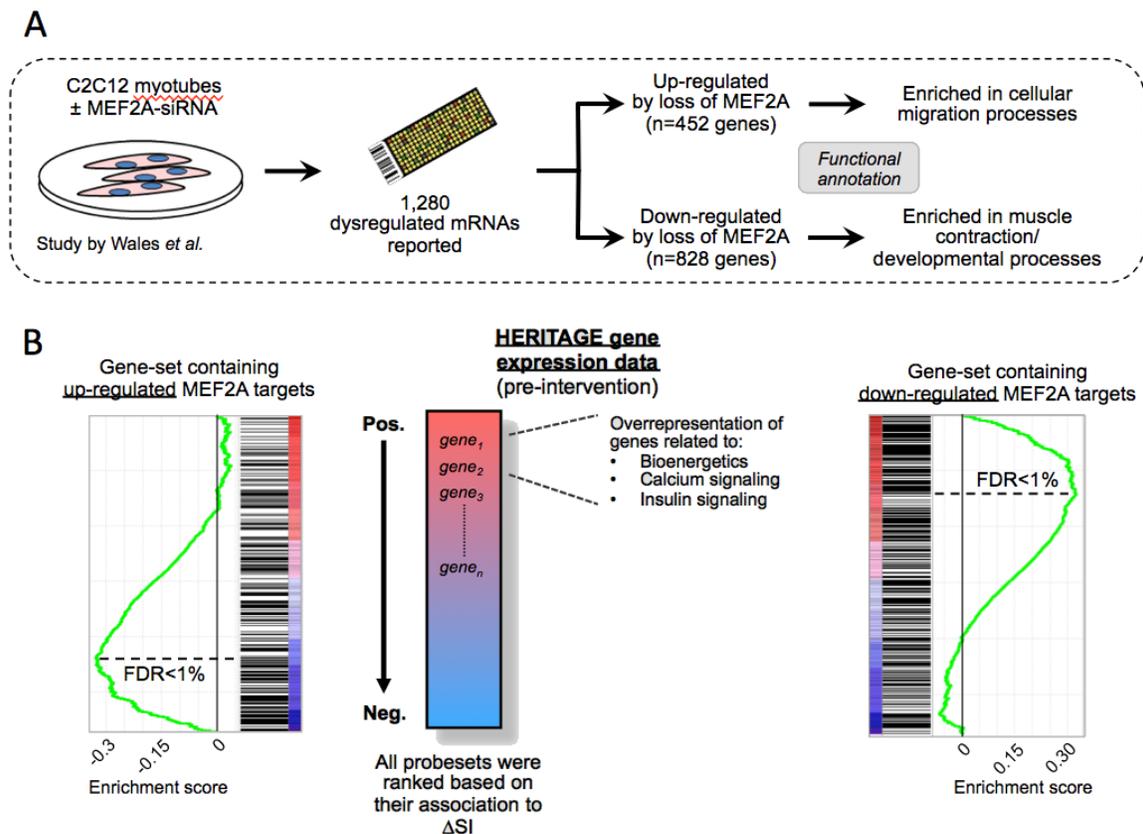
Interestingly, MEF2 has been extensively studied in muscle cells as it is known to directly regulate transcription of multiple muscle-enriched genes (*e.g.* myogenin and troponin I) [205]. Moreover, MEF2 is activated by contraction-induced calcium signalling [206], a pathway whose gene activity overall is under significant regulatory genetic control (see Figure 4-5). We therefore hypothesised that MEF2 could be a key driver of the pleiotropic transcriptional response linked to  $\Delta$ SI.

In order to test this hypothesis we first defined the global transcriptional signature associated to MEF2 knockdown and then asked whether this MEF2-dependent signature can recapitulate the transcriptional signature correlated to  $\Delta$ SI.

Recently, Wales *et al.* reported a comprehensive list of both direct and indirect *MEF2A* mammalian target genes using isoform-specific short interference RNA (siRNA) knockdown in differentiating C2C12 myotubes [207]. Notably, despite the four-membered MEF2 protein family sharing a high sequence homology in their DNA binding domains, genes distinctly sensitive to the A isoform play roles in calcium signalling [208].

Of the 1,280 dysregulated genes reported by Wales *et al.* (828 being down-regulated), we identified human orthologs for 990 (77%) [209]. We then performed a custom GSEA using up- and down-regulated *MEF2A* gene targets, respectively, as gene sets. Intriguingly, as shown in Figure 4-6, genes down-regulated by knockdown of *MEF2A*, which overall relate to ‘muscle function’ [207], were highly enriched (FDR<1%) amongst the most *positively* associated genes to  $\Delta$ SI in HERITAGE. In addition, the 2<sup>nd</sup> gene-set containing up-regulated *MEF2A* target genes, was highly enriched (FDR<1%) amongst the most negatively associated genes to  $\Delta$ SI.

It is of note that the genes within the calcium-signalling pathway are enriched (p=0.02) towards the positive end of the ranked gene-list (Figure 4-6B, middle panel).



**Figure 4-6** | (A) Using small interfering RNA (siRNA), Wales et al. [207] recently defined the global transcriptional signature associated with MEF2A modulation in differentiating C2C12s. Corresponding human orthologs for the dysregulated genes were identified using the Mouse Genome Informatics (MGI) database (72% mapping success). (B) All skeletal muscle expressed genes in HERITAGE were ranked according to their association to  $\Delta SI$  (grey panel in middle). GSEA revealed that the custom geneset containing gene targets up-regulated by MEF2A knockdown was enriched in the negative end of the ranked gene list (left panel), whereas the geneset containing down-regulated MEF2A targets was enriched in the positive end (right panel). The green curves correspond to the running sum of the enrichment score that reflects the degree to which the MEF2A-associated signatures (represented by black horizontal lines) are overrepresented at the top or bottom of the ranked list.

#### 4.3.5 Basal mRNA abundance of MEF2A interacting targets is predictive of $\Delta SI$ :

The above *in vitro* validation of the *in silico* framework prompted us to ask whether we could develop robust statistical models linking the basal mRNA abundance of MEF2A

targets to  $\Delta$ SI. Since our aim was to develop a predictive model that included direct targets of MEF2A—rather than the pleiotropic gene expression signature—we used the STRING database to focus the analysis around physically-functionally interacting MEF2A targets (n=46; see Appendix 7.5 for an exhaustive list) [210]. As most genes do not act in isolation but interact with each other collectively, we developed all possible *multivariate* linear regression models based on the combination of three MEF2A targets, including pairwise interaction components:

$$\frac{46!}{(46 - 3)! (3!)} = 15,180 \text{ models}$$

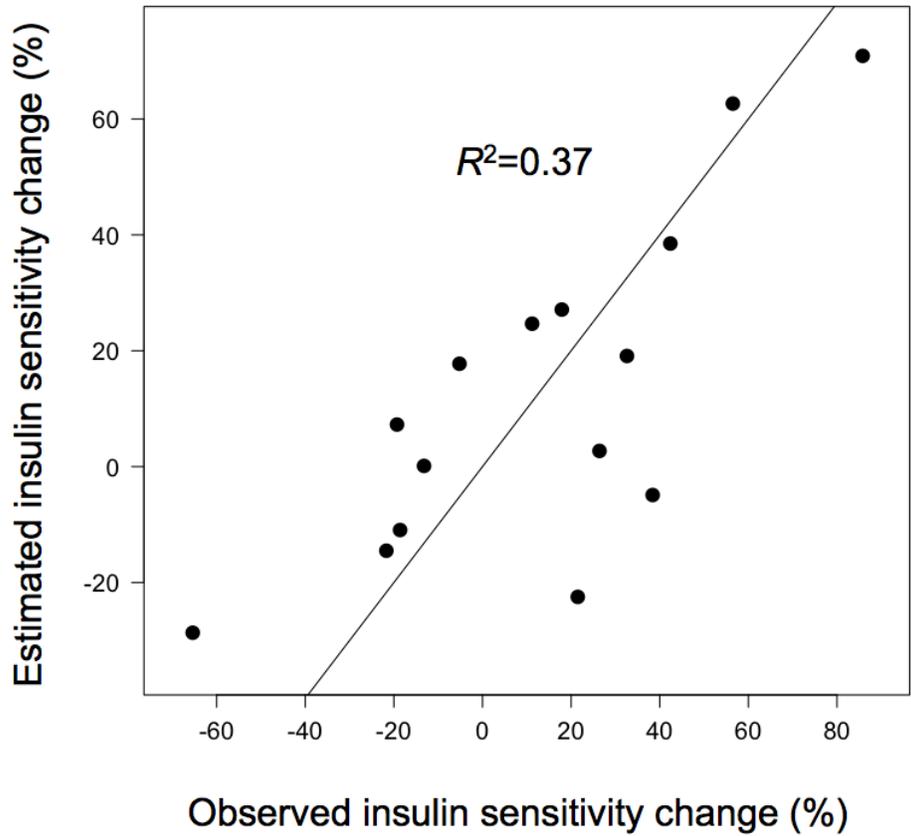
Interestingly, the most predictive statistical model that consisted of *HDAC4*, as well as the delta and gamma chain of the calmodulin-dependent protein kinase (CaMK) II enzyme complex (*i.e.* *CAMK2D* and *CAMK2G*), was able to explain nearly half (48%) of the variance of  $\Delta$ SI in the HERITAGE study (F-value=6.4, p<0.001; Table 4-2). Consistent with the established heteromultimeric nature of CaMKII, the model reports a highly significant interaction between the two isoforms *CAMK2D* and *CAMK2G*. Regression diagnostics confirmed conformity of the residuals to the assumptions of normality, linearity and homoscedasticity.

Variable	Regression coefficient	Standard error	t-value	P-value
VO <sub>2max</sub>	-0.03	0.004	-0.68	0.50
Gender	-2.15	0.83	-2.59	0.013
CAMK2D	40.73	27.98	1.46	0.15
CAMK2G	74.78	32.04	2.33	0.02
HDAC4	-34.94	22.37	-1.56	0.13
CAMK2D: CAMK2G	-12.08	3.41	-3.54	0.001
CAMK2D: HDAC4	6.94	1.63	4.27	<0.001
CAMK2G: HDAC4	-0.40	2.90	-0.14	0.89

**Table 4-2 | Result of the multivariate regression model for SI training response in HERITAGE (n=47).**

Gender and aerobic capacity adjusted for body size were included as covariates in the model. Notably, the abundance of these transcripts was not responsive to the training intervention, but rather higher basal expression levels were associated with greater gains in SI. CAMK2D: 228555\_at; CAMK2G: 212757\_s\_at; HDAC4: 228813\_at

To examine the general applicability of this predictive gene signature, as all HERITAGE samples were used for developing models, we took advantage of a previously published Affymetrix gene expression dataset from a smaller independent mixed exercise training cohort [199]. Importantly, this cohort also spanned a broad range in terms of the training-induced change in SI (-65% to +86%). Intriguingly, as shown in Figure 4-7 below, the multi-gene RNA signature was able to explain 37% of the  $\Delta$ SI among the healthy middle-aged male participants (n=15), a value comparable to the estimated heritability of  $\Delta$ SI (see Figure 4-2). A resampling procedure in which 1,000 random multivariate models were developed confirmed the significance of the regression model (p=0.05), despite the differing gene-chip technology between cohorts (*i.e.* 3'-based versus the newer whole transcript based methodologies).



**Figure 4-7** | Plot of experimentally determined (observed) versus predicted values of training-induced change in insulin sensitivity (derived from euglycaemic hyperinsulinaemic clamp) in the independent training cohort.

## 4.4 Discussion

Having established that the SI training response in part is genetically regulated, we have overlaid an individual's genetic background with the skeletal muscle specific gene expression profile to achieve an integrated view of the SI training response at the molecular level. Notably, such a tandem-omics approach enabled us to develop a quantitative multi-gene predictor of  $\Delta$ SI, which was validated in an independent training cohort. The present analysis thereby extends the notion that transcriptomic profiling captures critical information and process prognostic information about complex physiological processes [24, 211–213].

### 4.4.1 Predicting changes in insulin sensitivity from gene expression profiling

The multi-gene predictor was developed using the HERITAGE cohort, a well-characterized, demographically heterogeneous cohort. Due to the study-design induced interindividual heterogeneity (*e.g.* sex, BMI, VO<sub>2</sub>max), the RNA signature should be of broader applicability (*i.e.* less sensitive to environmental factors). In support of this, we were indeed able to validate the predictor in an independent training cohort of similar ancestry — despite the weekly cycling duration, a primary factor controlling SI training response [214], differed significantly. It remains to be established whether the RNA signature also works across other intervention regimes (*e.g.* interval-based).

Notably, all three predictor genes have many splice variants (between 14 and 17) [215]. However, as the most representative probeset for each gene was selected using the well established JetSet algorithm [63], both the specificity and splice isoform coverage are very high among the predictor probesets. The high coverage is likely also the main

reason behind the successful independent validation, which used exon arrays for profiling (a newer Affymetrix chip generation).

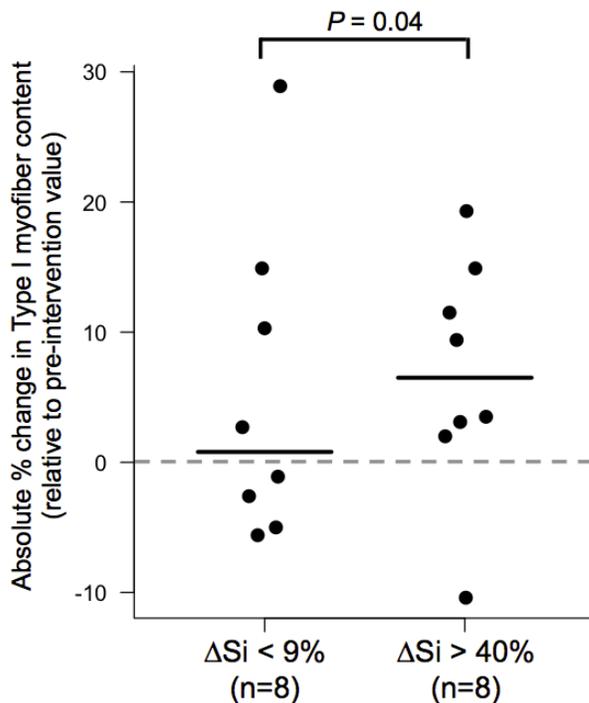
The RNA expression levels of most genes in the KEGG calcium signalling pathway (as well as the actual predictor signature) are stable to the training intervention, which is in accordance with the published ‘training-responsive transcriptome’ database [24]. Hence, the reported genotype/RNA associations within this pathway (see Figure 4-4) appear to be driven by exercise-independent factors.

#### **4.4.2 Is there a relationship between changes in SI and myofiber interconvertibility?**

It is well established that physical activity (*i.e.* muscle contraction) induces the calcium signalling pathway, leading to activation of MEF2 (among others) via CaMKs (CaMKII in particular), evoking a muscle-specific transcriptional response [10]. The non-biased informatics analyses presented here clearly point towards an involvement of genetic variations within the calcium signalling pathway affecting the ability to improve peripheral SI via MEF2. If this hypothesis is true, then it would not be unreasonable to assume that high responders in terms of  $\Delta$ SI would show the biggest training-induced fast-to-slow myofiber transformation, given the key role of MEF2 in adult oxidative type I myofiber formation [216]. Although controversial, multiple clinical studies have in fact reported an increase in type I content following endurance training interventions [217–219].

As age and sex have been shown to significantly affect muscle myofiber composition [11, 220], the analysis was restricted to men under 40 as they made up the largest contribution to the HERITAGE subcohort (~40%).

The result was striking: the group of ‘younger’ males that increased their SI with at least 40% (n=8) showed a significant training-induced increase in Type I myofiber composition (6.5% median increase; p=0.04) (Figure 4-8). In contrast, those with a modest or even negative potential (<9% increase in SI; n=8) showed a very heterogeneous myofiber adaptation and hence failed as a group to increase their type I content (0.8% median increase; p=0.23); this was evident despite the marked influence of an outlier with an abnormally high absolute increase of 29% post-intervention (67% Type I fibers).



**Figure 4-8 | Column scatter plot showing training-induced change in Type I myofiber content within the two defined responder groups (change in whole-body insulin sensitivity;  $\Delta SI$ ).**

Horizontal lines represent the mean value.

#### **4.4.3 Conclusion**

Whole-body SI is likely the most important physiological property that can be manipulated by lifestyle intervention in order to prevent cardiometabolic diseases [221, 222]. Promisingly, muscle RNA abundance can easily be quantified due to the development of pain-free micro-needle biopsy sampling [223, 224]. Further, one-step multiplex real-time RT-PCR assays are a rapid, sensitive and cheap diagnostics option. For these reasons, we believe that our predictive RNA signature potentially—if successfully validated in bigger independent endurance training cohorts--has clinical health implications. This is even more critical as skeletal muscle insulin resistance is one of the earliest hallmarks of the development of type 2 diabetes. All of this emphasizes the importance of developing evidence-based personalized exercise prescription to maximize the health-promoting benefits of a physically active lifestyle.

I would like to point out the importance of the independent validation in order to gain more confidence in the predictor signature put forward in this chapter. Hence, further research is urgently needed to replicate and further test these findings in independent studies and other populations and exercise programs.

## **5 Combining genomic and transcriptomic predictors of plasma triglyceride response to exercise training: a genome-wide approach**

*The work presented in this chapter represents a collaborative project between the labs of Francesco Falciani (University of Liverpool) and Prof. Claude Bouchard (Pennington Biomedical Research Center). The Bouchard lab provided the raw mRNA microarray data as well as all clinical readouts. Dr. Mark Sarzynski from the Bouchard lab conducted the GWAS analysis. I conducted all transcriptomic and pathway-based analyses.*

*The results presented in this chapter were very recently accepted in British Journal of Sports Medicine [226].*

## 5.1 Introduction

Triglycerides (TG) are the major constituent of dietary fats/lipids, composed of a glycerol esterified to three FA chains of varying composition and length. TGs are transported in the blood stream within globular molecular complexes know as lipoproteins. Notably, these circulating lipoproteins differ in size, density and lipid composition (Table 5-1).

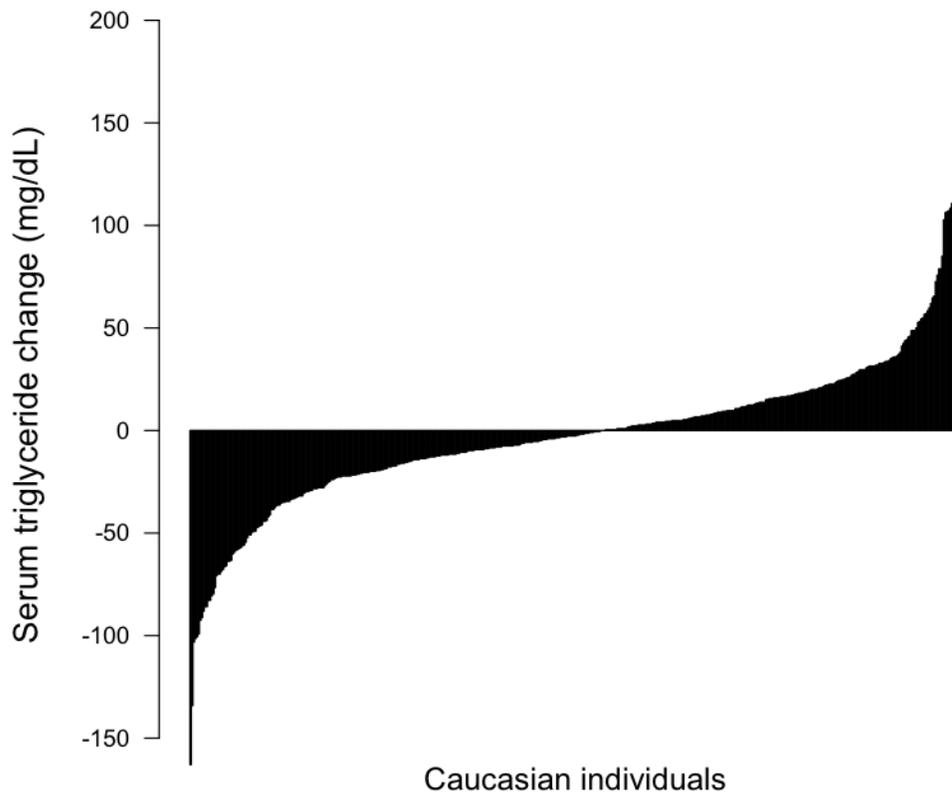
Lipoprotein class	% of	
	TG	ChoL
Chylomicrons	90	3
VLDL	65	15
LDL	10	45
HDL	5	20

**Table 5-1 | Major classes of lipoproteins in human plasma**

Normal resting values of TG are below 150 mg/dL in serum; levels above this threshold constitute hypertriglyceridemia. Very high levels of TGs are defined as serum levels of 500 mg/dL or greater [227]. Epidemiological studies have associated chronic elevated serum TG levels with an increased risk of cardiovascular disease (CVD) [228–230], the leading cause of death in developed countries [231]. Notably, Mendelian randomization studies of genetic variants affecting TG serum levels suggests a causal role of TG on CVD and all-cause mortality [232, 233]. Common *secondary* causes of raised serum TG levels are obesity, physical inactivity, excess alcohol intake and uncontrolled T2DM.

Regular physical activity (*e.g.* fast walking, out-door games) is considered a major target for therapeutic lifestyle changes in the prevention and treatment of elevated TG [234]. Evidence from the existing sports literature highlights a favorable reduction in serum TG levels ranging from 4 to 38 mg/dL follow exercise training [235]. However, we and

others have observed a large inter-individual variation in the magnitude of changes in serum TG levels following endurance training. For example, in the HERITAGE Family Study, the changes in TG after 20 weeks of highly standardized and fully supervised exercise training ranged from a decrease of 163 mg/dL (1.8 mmol/L) to an increase of 207 mg/dL (2.3 mmol/L) among Caucasians (Figure 5-1).



**Figure 5-1** | Histogram highlighting the individual differences in training-induced change in serum triglyceride levels among Caucasians in the HERITAGE study.

In HERITAGE, the heritability estimate for exercise-induced changes in serum TG has been estimated to reach 29% among Caucasians [236]. However, at present the DNA variants predisposing for interindividual variation in TG response remain to be determined.

A limited number of candidate gene studies have provided evidence of the association of DNA sequence variants with TG response to lifestyle and exercise interventions, with nominal associations found for variants in the *APOE*, *LIPC*, and *PGSI* genes [237, 238]. However, these candidate genes explain only a small percentage of the variance in serum TG response to exercise training. Thus, currently there is a clear need for unbiased data-driven approaches in the search for the genes and DNA sequence variants contributing to serum TG response to regular exercise.

Our collaborators, lead by Prof. Bouchard, have demonstrated in 2010 that global mRNA profiling of resting skeletal muscle combined with targeted SNP genotyping can increase the explanatory power of a multi-gene RNA signature for exercise adaptability [24]. Notably, such an approach dramatically improves statistical power and thus allows smaller cohorts of carefully characterized subjects to be studied. Therefore, the overall aim of this chapter is to perform both SNP and mRNA microarray profiling in order to identify a multivariate SNP-based gene signature that accurately predicts the responsiveness of serum TG levels to chronic endurance training.

## 5.2 Methods

### 5.2.1 Determination of plasma lipids

Blood samples for plasma lipid assays were obtained from an antecubital vein into Vacutainer tubes containing EDTA in the morning after a 12-hour fast with participants in a semi-recumbent position. The blood samples were collected twice at baseline (on separate days), and again at 24- and 72-hours after the last exercise session. TG levels were determined in plasma by enzymatic methods using a Technicon RA-500 Analyzer (Bayer Corporation Inc, Tarrytown, NY). The reproducibility of TG measurements in HERITAGE has been previously examined, with a coefficient of variation of 21.8, intraclass correlation of 0.79, and technical error of 0.21 (mmol/L). The response to exercise training was computed as the difference between the average post-exercise training TG measures and the average baseline TG measures.

### 5.2.2 GWAS SNP genotyping

See Section 4.2.2 for more details.

### 5.2.3 GWAS statistical analyses

See online manuscript<sup>4</sup>, as I didn't contribute to this part.

### 5.2.4 Affymetrix microarray analysis

Please refer to Section 4.2.4.

### 5.2.5 Baseline RNA gene signature

We used GALGO, an R package for multivariate variable selection based on a genetic algorithm (GA) methodology [99], to derive a multivariate regression model of  $\Delta$ TG in HERITAGE using baseline mRNA expression levels. In brief, this computational search procedure tries to find the best subset of genes for maximizing the fitness of the

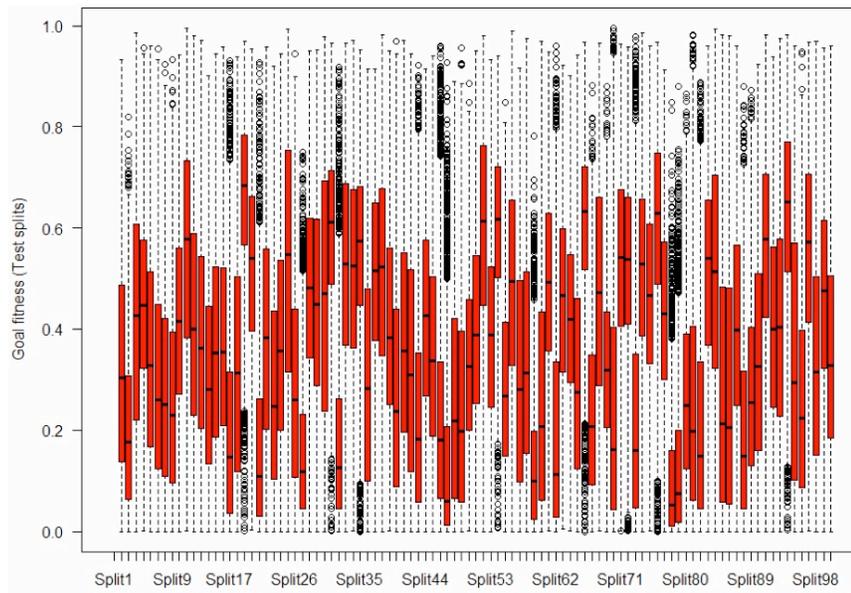
---

<sup>4</sup> <http://www.ncbi.nlm.nih.gov/pubmed/?term=26491034>

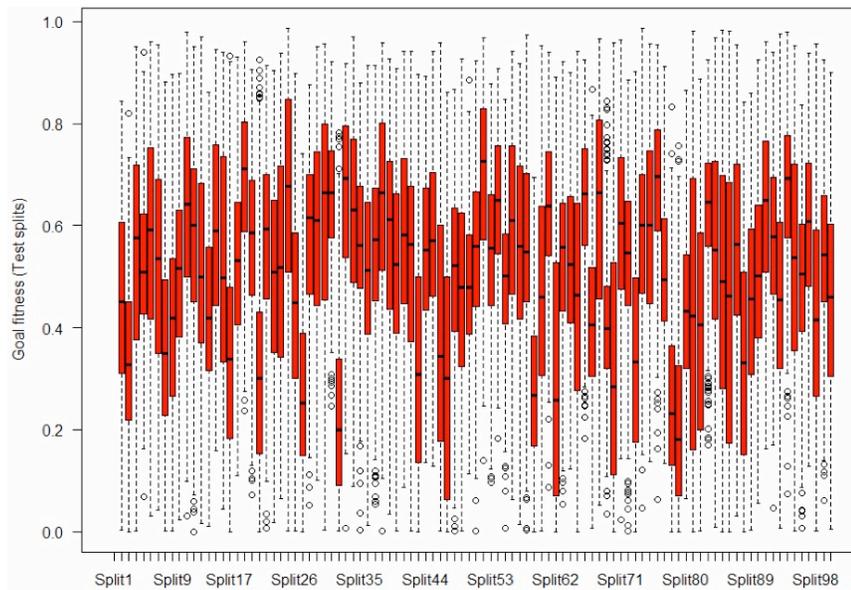
regression model (defined by  $R^2$ ). The fitness function was implemented as a linear model containing three genes with sex as an additional parameter. The fitness value for model *selection* was set to  $R^2 \geq 0.7$  (this particular cutoff was chosen based on thorough parameter estimation in accordance with the GALGO manual). Hence, a GA search is termed ‘successful’ once a model, through generations of evolution, has reached a fitness greater than or equal to the specified goal fitness. To reduce the risk of overfitting, a GA search had to reach the goal fitness within 300 generations in order for it to be selected.

During the GA search procedure  $\frac{3}{4}$  of the samples (N=37) were used for training whereas the remaining 12 samples were completely omitted and used for model validation. In order to better estimate the accuracy of each developed model (due to the modest number of profiled samples) the training dataset was split in 100 different training and test sets. The GA procedure was allowed to run 4,000 *successful* simulations (*i.e.* developed 4,000 regression models with an  $R^2 \geq 0.7$ ).

Due to variability in the model performance across the splits (likely caused by inter-individual differences), I chose to focus on the subset of models (N=512) that performed well in most splits (*i.e.* low overall deviation and high accuracy) (Figure 5-2).



Focus analysis on those models work similarly in most splits (512 models left)



**Figure 5-2** | Chromosome goal fitness distribution across all 100 splits for **A**) all 4,000 predictive models, and **B**) the subset (n=512) that performed well in most splits (i.e. low overall deviation and high accuracy).

### 5.2.6 SNP summary score

A SNP ‘summary score’ for  $\Delta$ TG was constructed using the GWAS SNPs that accounted for all of the estimated heritability in the conditional heritability analyses (N=4: rs222158, rs2722171, rs1906058, rs2593324) along with the top SNPs from the RNA-based classifier genes (N=4: rs1043968, rs3793336, rs594461, rs11646610). The four latter SNPs were included as the heritability estimate reported herein (*i.e.* 29%) is subjected to uncertainty ( $\pm 13\%$ ; see [236]) because of relatively modest number of Whites included in the HERITAGE study ( $n < 500$ ). Hence, the true  $\Delta$ TG heritability could potentially be significantly higher.

Each SNP was re-coded to reflect the number of *positive* or *favorable* response to regular exercise alleles (*e.g.* positive defined as a decrease in TG) in the following manner: 0 for no copy of the positive response allele, 1 for one copy and 2 for two copies of the positive response allele. The sum of the re-coded SNPs was used as the SNP summary score. Associations between SNP summary score and TG response to exercise training in HERITAGE Whites (N=476) were analyzed using general linear models. Covariates in the models included age, sex, baseline BMI, and baseline TG.

### 5.3 Results

The basic characteristics, including mean lipid values at baseline and in response to exercise training, of HERITAGE White participants with valid GWAS or gene expression data are shown in Table 5-2. On average, HDL-C and lipoprotein lipase activity increased with regular exercise, while TG and hepatic lipase activity decreased.

Variable	Subjects with GWAS Data (N = 478)			Subsample with Gene Expression Data (N = 49)	
	Mean (SD)	95% CI	Range	Mean (SD)	95% CI
<b>Age, yrs</b>	35.9 (14.5)		17.0 to 65.2	32.9 (14.3)	
<b>BMI, kg/m<sup>2</sup></b>				25.9 (4.1)	
• Baseline	25.9 (5.0)		17.0 to 47.5		
• Response to exercise training	-0.09 (0.7)	-0.15 to -0.02	-3.5 to 2.9	-0.004 (0.8)	-0.22 to 0.21
<b>Triglycerides, mmol/L</b>					
• Baseline	1.4 (0.8)		0.4 to 6.3	1.4 (0.8)	
• Response to exercise training	-0.02 (0.4)	-0.06 to 0.02	-1.8 to 2.3	-0.2 (0.5)	-0.28 to -0.03
<b>HDL-C, mmol/L</b>					
• Baseline	1.0 (0.3)		0.5 to 2.0	1.1 (0.2)	
• Response to exercise training	0.04 (0.1)	0.03 to 0.05	-0.3 to 0.6	0.07 (0.1)	0.04 to 0.11
<b>LDL-C, mmol/L</b>					
• Baseline	3.0 (0.8)		0.9 to 6.0	2.8 (0.8)	
• Response to exercise training	-0.004 (0.4)	-0.04 to 0.03	-1.2 to 1.5	0.01 (0.4)	-0.09 to 0.12

**Table 5-2 | Descriptive data, including pre- and post-training values for lipid, lipoprotein and lipase measurements, for HERITAGE Whites with valid GWAS (left) and gene expression (right) data, respectively.**

### 5.3.1 GWAS associations for TG response to exercise training

39 SNPs showed associations of  $p < 1 \times 10^{-4}$  (Appendix 7.6). The strongest evidence of association with  $\Delta$ TG was detected at rs2396190 ( $P = 3.3 \times 10^{-7}$ ) located 90kb from *DOCK10* (2q36.2), followed by rs222158 ( $P = 1.8 \times 10^{-6}$ ) located in *CYYRI* (21q21.2). In the final forward regression model, the top 10 SNPs explained 32.0% of the variance in  $\Delta$ TG (Appendix 7.6). *CYYRI* rs222158 was the strongest independent predictor of  $\Delta$ TG in the model, explaining 5.5% of the total variance, followed by *GLT8D2* rs2722171, which explained 4.1% of the variance.

As shown in the last column of Appendix 7.6, the top four SNPs were shown to be sufficient to account for the genetic component of TG response to exercise training in White HERITAGE families. These four SNPs were retained for the  $\Delta$ TG SNP summary score.

*Pathway analysis of GWAS associations.* The result of the pathway-based SNP analyses using the 2<sup>nd</sup>-best p-value and Stouffer's method, respectively, can be found in Appendix 7.7. Briefly, the glycosphingolipid biosynthesis gene set was enriched using both approaches. Glycosphingolipid biosynthesis-related gene sets were the 2<sup>nd</sup>, 5<sup>th</sup>, and 6<sup>th</sup> most enriched positive gene sets using the 2<sup>nd</sup>-best p-value method (FDR: 0.07-0.32), while it was the 3<sup>rd</sup> ranked gene set using the Stouffer method (FDR=0.30). The most enriched positive gene set using the 2<sup>nd</sup>-best p-value method was the heparan sulfate glycosaminoglycan biosynthesis gene set (FDR=0.097), while cell adhesion molecules was the most enriched negative gene set (FDR=0.12).

### 5.3.2 RNA expression-based gene signature of TG response to exercise training

Having shown that genetic variants in lipid biosynthesis pathways are linked to changes in TG levels, we set to test whether and mRNA signature could predict this important clinical variable.

We therefore developed statistical models predictive of  $\Delta$ TG based on a multi-gene RNA signature. The method we used, developed by the Falciani lab, is based on a true multivariate variable selection procedure designed to optimize the prediction of the response variable by means of the smaller number of gene expression profiles available.

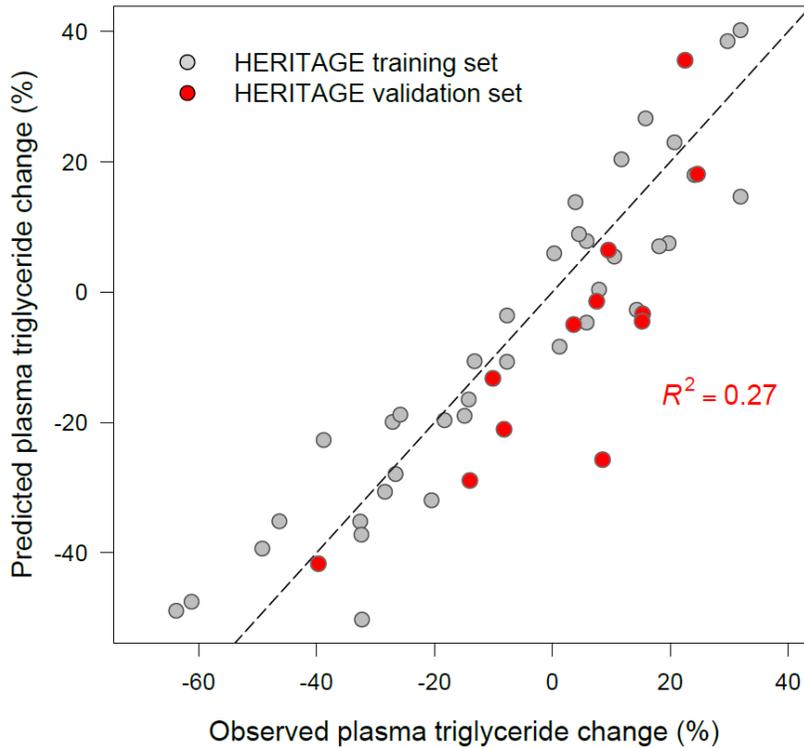
By means of forward stepwise regression, ranking Affymetrix probesets by their selection frequency (high to low) in the different GA-derived predictor models, an 11-gene linear regression model was developed (Table 5-3).

<b>Variable</b>	<b>Beta</b>	<b>SE</b>	<b>t-value</b>	<b>P-value</b>
Gender	0.21	0.17	1.1	0.27
DYX1C1	0.59	0.11	5.6	9.9x10 <sup>-6</sup>
ZNF30	-0.01	0.12	-0.09	0.93
BTG2	0.21	0.12	1.7	0.10
MACROD1	0.10	0.20	0.5	0.62
UBE2L3	0.16	0.15	1.1	0.29
C21orf88	-0.14	0.13	-1.1	0.28
EEF2K	0.21	0.10	2.1	0.05
NCBP2	0.22	0.16	1.5	0.15
FASTK	0.39	0.15	2.5	0.02
C2orf69	-0.03	0.12	-0.3	0.80
NSA2	-0.28	0.10	-2.8	0.009

**Table 5-3 | Results of the RNA-based multivariate regression model with forward selection for TG response to exercise training in HERITAGE Whites (N=37).**

This model was able to explain 80% of the variance in the training set (F-value=13.2, p<0.0001). We next evaluated the model performance on the remaining 12 HERITAGE samples that were completely omitted from the GA search procedure (so-called ‘test set’). As

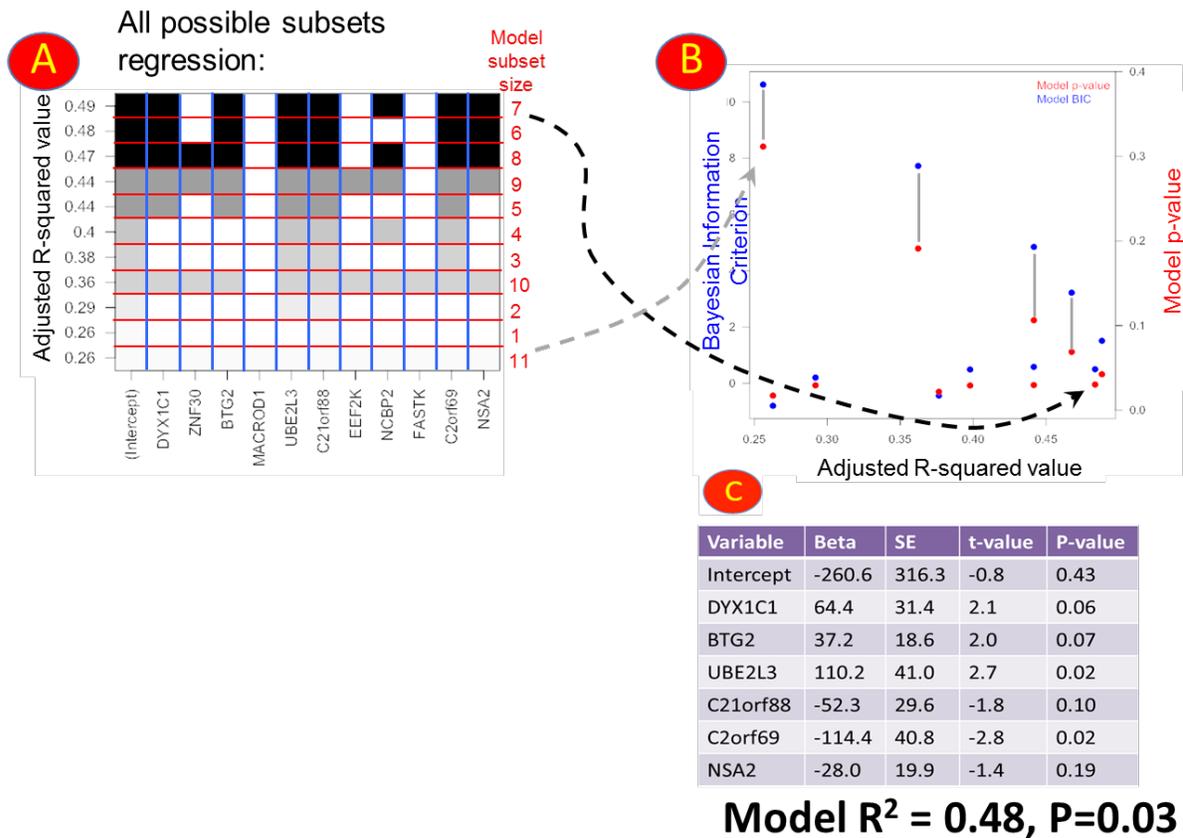
shown in Figure 5-3, the predictor model was able to explain 27% of the variance. Importantly, the test set spanned a broad range in terms of the training-induced changes in TG (range: -40 to +20%).



**Figure 5-3** | Performance of the RNA-based regression model derived from the training set (N=37, gray dots) in the test set (N=12, red dots) for the prediction of exercise training-induced changes in serum triglycerides in HERITAGE.

In order to examine the general applicability of the molecular signature, we took advantage of a previously published Affymetrix gene expression dataset from an independent exercise-training cohort (the same cohort used in Chapter 4). As the gene-chip technology significantly differed between cohorts (*i.e.* 3'-based versus the newer whole transcript based methodologies), we performed all possible subsets regression. We found that the model

containing six genes (*BTG2*, *C2orf69*, *C21orf88*, *DYX1C1*, *NSA2*, *UBE2L3*) performed the best, as this model (F-value=3.7, P=0.03) explained 48% of the variance in TG changes in the independent dataset, while also having the lowest Bayesian Information Criterion (BIC) score (Figure 5-4). Further, a resampling procedure in which 10,000 random multivariate models were developed confirmed the significance of the 6-gene model in terms of  $R^2$  performance.

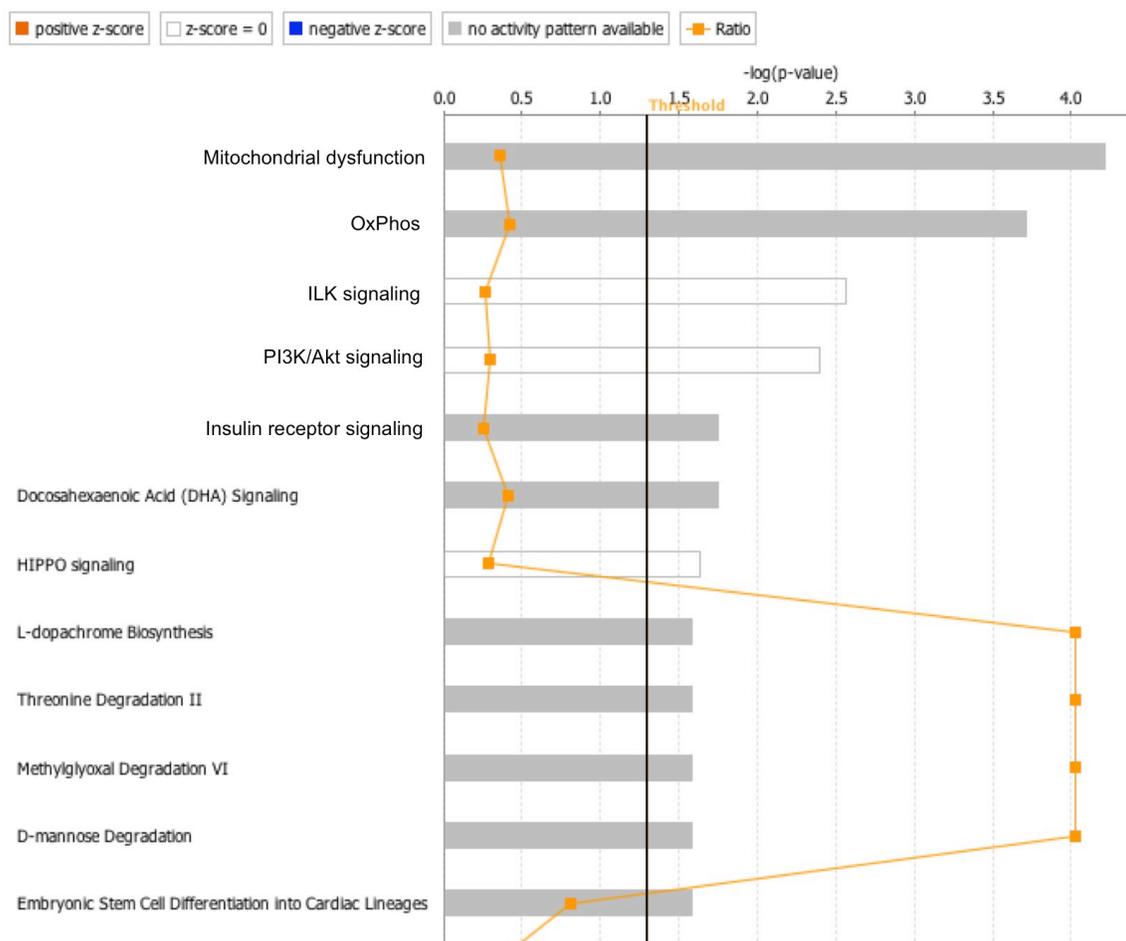


**Figure 5-4 | All possible subset regression.**

(A) Performance of regression models testing subsets of genes from one gene to all 11 genes for ability to predict TG change with exercise training in the independent dataset. (B) Comparison of model performance based on Bayesian Information Criterion (BIC),  $R^2$ , and p-values for each model. Ideally, the goal is to select the model with the lowest BIC and p-value that explains the most variance, which in this case is the model that includes 6 genes. (C) Results of the forward regression model for changes in TG with exercise training in the independent dataset containing six genes.

The association of SNPs (N=498), in or near ( $\pm 20\text{kb}$ ) the 11 predictor genes, with  $\Delta\text{TG}$  was tested in all HERITAGE Whites (N=481). Only SNPs from four genes (*NSA2*, *FASTK*, *MACROD1*, and *EEF2K*) showed nominal ( $p < 0.05$ ) associations with  $\Delta\text{TG}$  (see online Supplementary, Tables S5-S8). The top associated SNP from each of the four genes (rs1043968, rs3793336, rs594461, rs11646610) was used in the SNP summary score.

*Pathway analysis of mRNA predictor models.* We evaluated the enrichment of pathways among genes in the subset of 512 predictive models using Ingenuity Pathways Analysis (IPA). We found that mitochondrial energy metabolism pathways were enriched, including pathways related to mitochondrial dysfunction and oxidative phosphorylation (Figure 5-5).

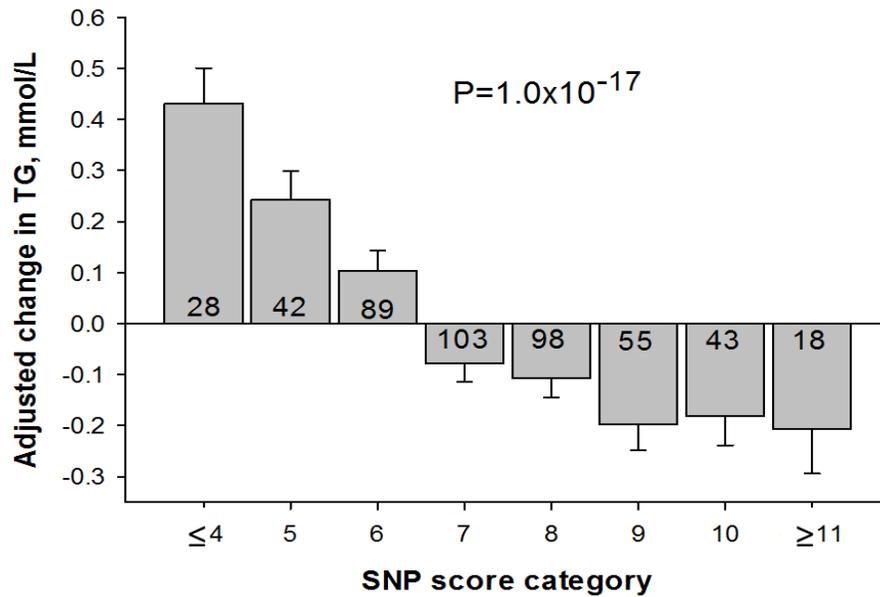


**Figure 5-5 | Results of the Ingenuity Pathway Analysis (IPA) on a subset of the baseline gene expression predictor models (N=512).** Horizontal bars represent the p-value for the top 12 canonical pathways enriched in genes within the most predictive GALGO models (see Figure 5-2B) and are expressed as -1 times the log of the p-value. The black vertical line represents the probability threshold at  $P=0.05$ . The orange line represents the ratio of the number of GALGO model genes in a particular pathway divided by the total number of genes that make up that pathway.

### 5.3.3 Association of SNP score and $\Delta TG$

The SNP score was created by combining the four top SNPs from the GWAS analysis (rs222158, rs2722171, rs1906058, rs2593324) and the four top SNPs from the targeted SNP analysis of the RNA-predictor genes (rs1043968, rs3793336, rs594461, rs11646610). The SNP

score values ranged from 2 to 12 in HERITAGE Whites. The adjusted mean decrease in TG in subjects with 11 or 12 favorable alleles (N=18) was -18.2 mg/dL (-0.21 mmol/L), while those with 4 or less favorable alleles (N=284) experienced an adjusted mean increase of 38.1 mg/dL (0.43 mmol/L) after exercise training (Figure 5-6).



**Figure 5-6 | Adjusted mean  $\Delta$ TG across eight SNP summary score categories in HERITAGE Whites.**

Values were adjusted for age, sex, baseline BMI, and baseline triglyceride level. Number of subjects within each SNP score category is indicated inside each histogram bar.

## 5.4 Discussion

The present work highlights that selected genes and common SNP variants are associated to the pronounced interindividual training-induced variation in serum TG levels. A novel finding is that an 11-gene muscle-derived molecular signature predicted 27% of TG exercise response in HERITAGE; importantly I was able to validate the response predictor model an independent training cohort.

The result of the SNP-derived summary score showcases how parsimonious panels of candidate SNPs from multiple omics platforms potentially can be used as *a priori* predictors of phenotype response.

Pathway-based analysis of the SNP data highlighted that the genetic effect of TG response to regular exercise may be exerted through specific biological pathways related to heparan sulfate glycosaminoglycan and glycosphingolipid biosynthesis and cell adhesion molecules (see Appendix 7.7).

Heparan sulfate has several biological functions, including cell adhesion and notably cell surface binding of lipoprotein lipase (LPL). As such, several studies have shown that heperan sulfate glycosaminoglycan modified proteoglycans act in the hydrolysis of triglyceride-rich lipoproteins [239]. In HERITAGE, LPL activity significantly increased with exercise training, concomitant with decreases in TG levels ( $r = -0.21$ ,  $P < 0.001$ ). Glycosphingolipids mediate and modulate intercellular coordination in multicellular organisms [240]. Glycosphingolipids cluster in lipid rafts, which are enriched in cholesterol and sphingolipids such as sphingomyelin. Lipid rafts are involved in many cellular processes, including membrane sorting and trafficking, cell polarization, and

signal transduction processes. As such, the specialized structure of lipid rafts may play a critical role in the trafficking of lipids between lipoproteins and cells. However, it is unknown how exercise affects lipid rafts and glycosphingolipids.

Pathway-based analysis of the transcriptomic predictor models revealed that mitochondrial dysfunction and oxidative phosphorylation pathways were enriched in relation to TG response to exercise training. Mitochondrial dysfunction has been linked to fat storage abnormalities, including accumulation of TG in various tissues such as muscle and adipose. Additionally, defects in mitochondrial oxidation and phosphorylation are associated with insulin resistance, Type II diabetes, and ectopic TG accumulation (e.g., intramyocellular and intrahepatic lipids), among others. Exercise is known to improve insulin sensitivity and risk of diabetes, while the effects of exercise on intramyocellular and intrahepatic lipids are less clear.

In summary, our bioinformatics analysis suggests that different subsets of pathways are enriched when variants from genomic and transcriptomic data are interrogated for their association with TG response to exercise. These analyses inform on the potential mechanisms involved in TG changes with regular exercise, with the transcriptomics data possibly representing muscle-specific mechanisms.

This study is the first to test systematically whether the TG GWAS loci identified from the most recent meta-analysis modify the response of TG to regular exercise. We did not find any associations between the SNPs associated with TG GWAS loci and the response of TG to regular exercise. Thus, our results suggest that the genes most important in modifying changes in TG in response to regular exercise are likely different from the loci contributing to variation in population TG levels. It is not clear how many

of the markers and genes identified here are functionally related to the response to regular exercise of TG, but, if confirmed in independent studies, they could inform the biology of plasma TG and their adaptation to regular exercise. Given the estimated effect sizes observed in HERITAGE Whites, some of these genomic regions deserve further investigation, such as *GLT8D2*.

#### **5.4.1 Study limitations**

This study is based on a relatively small sample size compared to commonly accepted standards of GWAS. Thus, it is not surprising that we did not find SNPs reaching genome-wide significance. However, it is important to appreciate that human experimental studies are by definition characterized by much smaller sample sizes than epidemiological and observational studies, but are also less likely to be negatively impacted by any number of uncontrolled confounders. HERITAGE remains the largest fully controlled exercise intervention study thus far. The family structure and well-defined and twice-measured phenotypes in HERITAGE help to minimize the influence of confounding factors. Moreover, since one component of the environment (i.e., exercise) has been rigorously controlled in HERITAGE, we had predicted larger effect sizes than is commonly seen in observational studies. Thus, we concluded that it would be useful to undertake hypothesis-free and unbiased GWAS explorations for the response of TG to regular exercise, as it could generate hypotheses and candidate genes deserving further research. Furthermore, our global microarray data from a subset of HERITAGE participants complements our GWAS results and provides a separate unbiased exploration of the genes involved in TG exercise response. We acknowledge that there is still a possibility of false discovery and that the parsimonious SNP score associated with

$\Delta$ TG in HERITAGE is likely to overfit our data, possibly resulting in biased conclusions about the strength of the findings. Therefore, there is an obvious need to replicate our results in other samples and studies.

## 6 Concluding Remarks and Research Perspective.

There are clearly many types of data in a complex biological system that one could choose to quantitate in order to generate a biomarker (signature); ideally of clinical relevance.

I personally believe that alterations in RNA abundance currently are more tightly linked to *true* biological variation as compared to DNA or tissue-based proteomics approaches. For example, protein *abundance* (the readout of most proteomic technologies) is often a poor surrogate for genuine protein *activity* due to the many post-translational modifications, effecting its location—and thereby function—within the compartmentalized eukaryotic cell [241]. Proteomics technology also suffers from the inability to detect accurately low-abundance proteins [242].

In further support of my (admittedly controversial) claim, commercially available personal genetic screening tests are currently too imprecise, highlighted by the recent ban of 23andMe company products ([www.23andMe.com](http://www.23andMe.com)) by the US Food and Drug Administration (FDA).

In support of RNA as a good class of molecules for developing personalized medicine: using omics technologies, particular gene expression microarrays, there have been advances in predicting therapeutic outcomes — especially in the field of cancer chemotherapy [243–246]. This success is in part based on the clear-cut genetic aberrations that can drive tumor growth and aggression. So for example, a given tumor may lose a piece of genomic DNA, producing an unambiguous loss of gene expression in the gene-chip profile for those genes located in the lost DNA region [247].

In the field of metabolic traits/cardiovascular diseases, it is understood that phenotypes are much more complex (*i.e.* polygenic); thus ‘single-gene’ DNA analyses will most likely not, alone, be sufficient to predict susceptibility or treatment responsiveness. In addition, variation in gene expression can be very subtle [24]—yet consistent—and therefore multiple gene markers are needed to provide a more robust diagnostics.

While altered physiological status and altered protein-coding sequence feedback to impact on mRNA abundance, it is certainly also very likely that DNA sequence and epigenetic modifications contribute to this process. Hence it is obvious to conclude that the more of the biological system that is measured (ideally on a global scale) the easier it will be to discern important intracellular pathways etc [248]. In Chapter 4, I took advantage of this concept by combining genomic and transcriptomic data to develop the first RNA signature that can predict whether an individual has a high or low potential for improving whole-body insulin sensitivity.

As highlighted in the Introduction of this thesis, NGS technology is able to provide the customer with a lot more quantitative information including splice variants, DNA sequence variations (*e.g.* SNPs) as well as non-protein coding RNA species (*e.g.* microRNAs and long-non-coding RNAs).

More and more evidence indicates that microRNAs can regulate mRNA stability in cells, thus ultimately affecting gene product abundance [249, 250]. *In vivo* in humans, however, it would appear that microRNAs may impact on protein synthesis rather than mRNA stability [251]. Irrespective of the mechanism of action, non-coding RNA species

represent yet another level of biological variance. Capturing microRNA expression profiles AND integrating them with mRNA profiles should, in theory, allow for the development of more complex statistical models of human metabolic physiology. Certainly, as the cost continues to decrease, NGS will be used more and more in clinical research as a tool for complex biomarker discovery.

Currently, several EU-funded projects are focusing on the development of personalized medicine. In the context of human skeletal muscle biology, the multi-center EU FP7 funded project ‘Metapredict’ ([www.metapredict.eu](http://www.metapredict.eu)) is aiming to develop robust predictors for the personalized clinical response to exercise training. Core to the success of this project is the integration of *sufficient* multi-level omics data (hence the multi-center nature of the project) [225] — with exon array profiling being a vital component!

Epigenetic modifications to DNA bases also convey important biological information. For example, as pointed out by Zierath *et al.* in a recent perspective in *Cell Metabolism*, the interplay between genes and environment may alter our epigenome, which ultimately could impact on health [225]. As NGS depend on PCR amplification, all epigenetic information is lost during sequencing (*i.e.* both 5-methylcytosine and 5-hydroxymethylcytosine are treated as cytosine by the enzymes involved in PCR).

Hence, in the more distant future, targeted DNA sequencing and epigenetic modification profiling using ‘third’ generation sequencing technology platforms may likely become clinically applicable. Currently, the Pacific Biosciences third generation sequencing

system for example allows for the identification of >20 different types of DNA modifications directly *without* any chemical or enzymatic reactions.

However, besides the creation of a multi-layered expression map on the systems level, sample size is obviously a very important aspect when developing robust predictors of complex phenotypes. Hence, further research is needed to replicate and further test the RNA-based signatures that I have presented in Chapters 4 and 5 in independent studies and other populations and exercise programs.

Further, the people behind the HEIRTAGE study are currently doing metabolomics profiling. Hence, by integrating this additional layer of information in the analysis framework put forward in Chapter 4, we might come up with new and improved predictors of training-response.

## 7 Appendices

### 7.1 RNA sequencing analysis and the development of a guinea pig microarray platform

#### 1.0 Developing the Guinea Pig transcriptome

Although GP is a model species, at the time of writing there were only 464 genes found in the Reference sequence project database. To better understanding the transcriptional response of GP in response to smoking, a high coverage transcriptome was constructed using novel RNAseq data derived from lung and skeletal muscle tissues. This was combined with public domain data, Ensembl cDNAs and Genscan gene predictions. The transcript assembly was used to make a comprehensive custom GP microarray for hypoxia and smoking analysis.

#### 1.1 RNAseq transcriptome assembly

There is a genome sequence available for GP ([http://www.ensembl.org/Cavia\\_procellus/Info/Index](http://www.ensembl.org/Cavia_procellus/Info/Index)) and this was used in conjunction with the TopHat [162, 252] and Cufflinks algorithms [163] to assemble transcripts based on genome alignments from the RNAseq data. The first stage was to align reads to the genome with TopHat, which requires the insert size of the RNAseq fragments to be known prior to mapping. This was determined using the SMALT alignment tool with the sampling option [253], aligning reads to publicly available transcript data. **Table 1** lists the number of reads for each tissue and a summary of the fragment size distributions, which were verified using the final assembly.

**Table 1 | Number of RNAseq reads and fragment sizes**

Tissue	Number of paired end reads	length of reads, each pair	Fragment size	Fragment size standard deviation
<b>Muscle</b>	113,438,469	101	173	104
<b>Lung</b>	56,466,360	101	169	92

TopHat v1.3.3 was then used to align the paired-end reads of lung and muscle to the *Cavia porcellus*.cavPor3.64 genome assembly, executed with the following options: insert size, closure-search, butterfly-search and coverage-search. The resulting TopHat genome binary alignment files were separately inputted to Cufflinks v1.1.0 to assemble tissue specific transcriptomes. The multiple read correction, upper quartile normalization and fragment bias correction options were used. Lung and muscle transcriptomes were subsequently combined using the cuffcompare algorithm, which produced an RNAseq transcriptome consisting of 81,074 sequences. A control step was performed to ensure that maximum transcriptome coverage was attained. Unique exon positions from the cufflinks assembly GTF files were extracted and compared to the cuffcompare output. A

number of sequences were not included in the cuffcompare output and these were added into the assembly.

### 1.2 Combining RNAseq and public data

A search on the NCBI Entrez system for the species *Cavia porcellus*, found 19,975 sequences [254]. The Cap3 software program [255] was used to cluster these public domain sequences and 1,762 contigs and 9,569 singletons were obtained. Added to these were the Ensembl GP cDNA and non-coding RNA sequences, from the download files *Cavia\_porcellus.cavPor3.64.cdna.all* and *Cavia\_porcellus.cavPor3.64.ncrna.fa* [256]. Finally, Ensembl Genscan predictions [257] were downloaded and compared to the RNAseq and public data with BLAST [258]. Any Genscan prediction with  $\leq 99\%$  coverage was added into the assembly. The final transcriptome to be used in array design consisted of 151,072 transcripts. **Table 2** presents a breakdown of the different sources of transcripts.

**Table 2 | a numerical breakdown of all the sequences**

Source	Description	Total prior assembly	Post assembly total
<b>RNAseq data</b>	Cufflinks cuffcompare assembled transcriptome sequences from GP lung and muscle poly-A RNA	81,074	81,074
<b>Cufflinks QC sequences</b>	Cufflinks genome position check sequences	1,871	1,871
<b>Ensembl all cDNA</b>	a set of transcript sequences based on EST and other sequence alignment evidence; <i>Cavia_porcellus.cavPor3.64.cdna.all.fa</i>	20,166	20,166
<b>Ensembl all ncRNA</b>	A set of non-coding RNA from Ensembl; <i>Cavia_porcellus.cavPor3.64.ncrna.fa</i>	5,963	5,963
<b>Ensembl Genscan</b>	Ensembl abinitio Genscan gene predictions based on genome alignments	53,615	30,667
<b>NCBI data</b>	Public GP gene, mRNA and EST sequences, clustered with Cap3	19,975	11,331

### 1.3. New transcriptome statistics and annotation

A bullet point summary of the transcriptome statistics is provided below. This is an initial transcriptome that was used to design a custom 180K Agilent microarray. An N50 statistic provides a base pair sequence length (a weighted median value) where 50% of the entire transcriptome sequences are contained within this value, with lengths equal or

greater than this value ([www.broadinstitute.org/crd/wiki/index.php/N50](http://www.broadinstitute.org/crd/wiki/index.php/N50)). The N50 of the full transcript assembly was 3.2Kb.

- Total transcripts = 151,072
- Longest transcript = 53,613bp
- Shortest transcript = 36bp
- N50 statistic = 3,247bp
- N90 statistic = 902bp
- Mean size = 1,894bp

Annotation of the full transcriptome prior to selection of microarray probes was performed using a BLAST search against the mouse Reference sequence project databases (Refseq) [259]. The number of transcripts annotated with a Refseq sequence were 97,822, consisting of 17,907 non-redundant mouse gene symbols. For the final 60K Agilent microarray design, a re-annotation was carried out due to the delay between full annotation and publication. This was done for the 86,044 transcripts that were represented on the final array, and the Refseq mouse data used was of August 2013. From the 80,044 transcripts of clusters that were mapped to the final array, 40,625 of them were annotated with a mouse gene. In terms of probes, 18,072 of the 60,984 probes on the final array were annotated with a mouse gene. This represents 17,676 non-redundant mouse genes. The Refseq mouse annotation of both the full transcriptome and the final array are accessible online:

[http://pcwww.liv.ac.uk/~herberjm/Davidson\\_et\\_al/Additional\\_file6.zip](http://pcwww.liv.ac.uk/~herberjm/Davidson_et_al/Additional_file6.zip)

(“**Daidsen\_et\_al\_Refseq\_annotation.xlsx**”). Note that 235 probes match transcript clusters that hit more than one mouse gene (the probes matched all transcripts in a cluster but different transcripts of the cluster matched multiple mouse genes). These are highlighted in red in the “ProbeAnnotation” sheet of the file

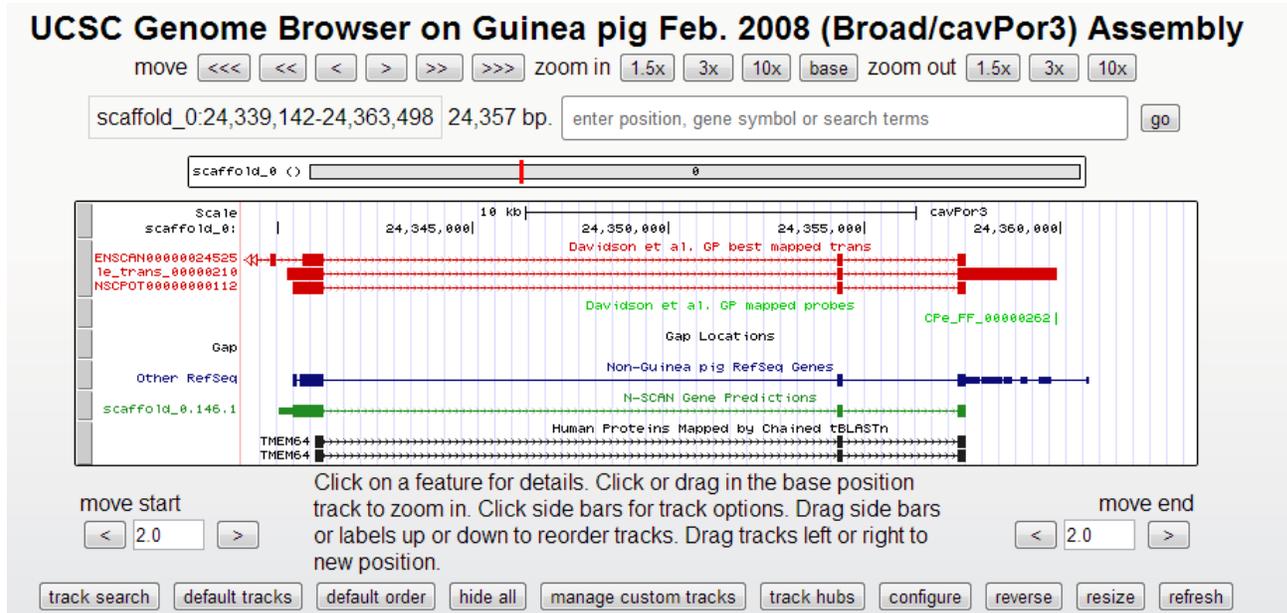
**Daidsen\_et\_al\_Refseq\_annotation.xlsx** (also accessible online<sup>5</sup>).

#### **1.4. Viewing of transcripts and probes in the Guinea pig genome**

A genome overview of both the GP transcripts and microarray probes (see below for details on the development for the development of the microarray platform) has been provided and can be viewed via the University of California Santa Cruz genome browser [260, 261]. This can be done by visiting the GP genome browser page at <http://genome.ucsc.edu/cgi-bin/hgTracks?db=cavPor3> and uploading the custom track file (**Daidsen\_et\_al\_UCSC\_genome\_browser\_custom\_tracks.txt**) provided in the online supporting material<sup>1</sup>. **Figure 1** shows an overview of the TMEM64 ortholog gene, with new transcripts displayed in red, and microarray probes in green (the top two tracks).

---

<sup>5</sup> Supporting material is accessible online at [http://pcwww.liv.ac.uk/~herberjm/Davidson\\_et\\_al/Additional\\_file6.zip](http://pcwww.liv.ac.uk/~herberjm/Davidson_et_al/Additional_file6.zip)



### 1.5. Differential gene expression and a Gene Set Enrichment Analysis of the RNAseq data

Single samples of lung and muscle were subject to RNAseq analysis to identify genes that distinguish between these two tissues. Because of the lack of biological replicates, transcript expression was contrasted using log<sub>2</sub> fold change between lung and muscle. To derive the fold change, the level of expression was measured using the Fragments Per Kilobase of exon model, per Million fragments mapped (FPKM) to the genome, which is the metric calculated by the Cuffdiff program, from the Cufflinks package. This expression is based on genome alignments of RNAseq reads. For a comparison, expression was also measured using the eXpress package [262], which measures Reads Per Kilobase of exon model, per Million reads mapped. eXpress measures levels of reads mapping to the transcriptome and accounts for variation in fragment length distributions, read errors and fragment bias. The log<sub>2</sub> fold change measurements were used to make two ranked list of genes. The correlation of log<sub>2</sub> fold changes between the two methods of measuring RNAseq expression was 0.8. Transcripts were assigned mouse gene symbols based on similarity searching with BLAST.

A Gene Set Enrichment Analysis (GSEA) was carried out using the Broad Institutes command line java tool [84, 263]. This analysis was done using two genesets from the Molecular Signatures Database representing KEGG pathways [128, 264] and Gene Ontology biological processes [265]. The human orthologs of the mouse gene annotation was derived from Homologene [266]. For both KEGG and GO biological process, significant enrichment scores were found. **Table 3a** and **Table 3b** display the most enriched ontologies and pathways for lung and muscle RNAseq data. **Figure 2** overviews examples of the enrichment within the ranked lists, with lung enrichment on the left hand panels and muscle on the right. The results of the GSEA for both gene ontology and KEGG suggest that the muscle sample is enriched for muscle related processes such as muscle respiration, metabolism and contraction. Biological relevance is also found in lung enrichments, showing evidence of immune response, defence and inflammation reactions.

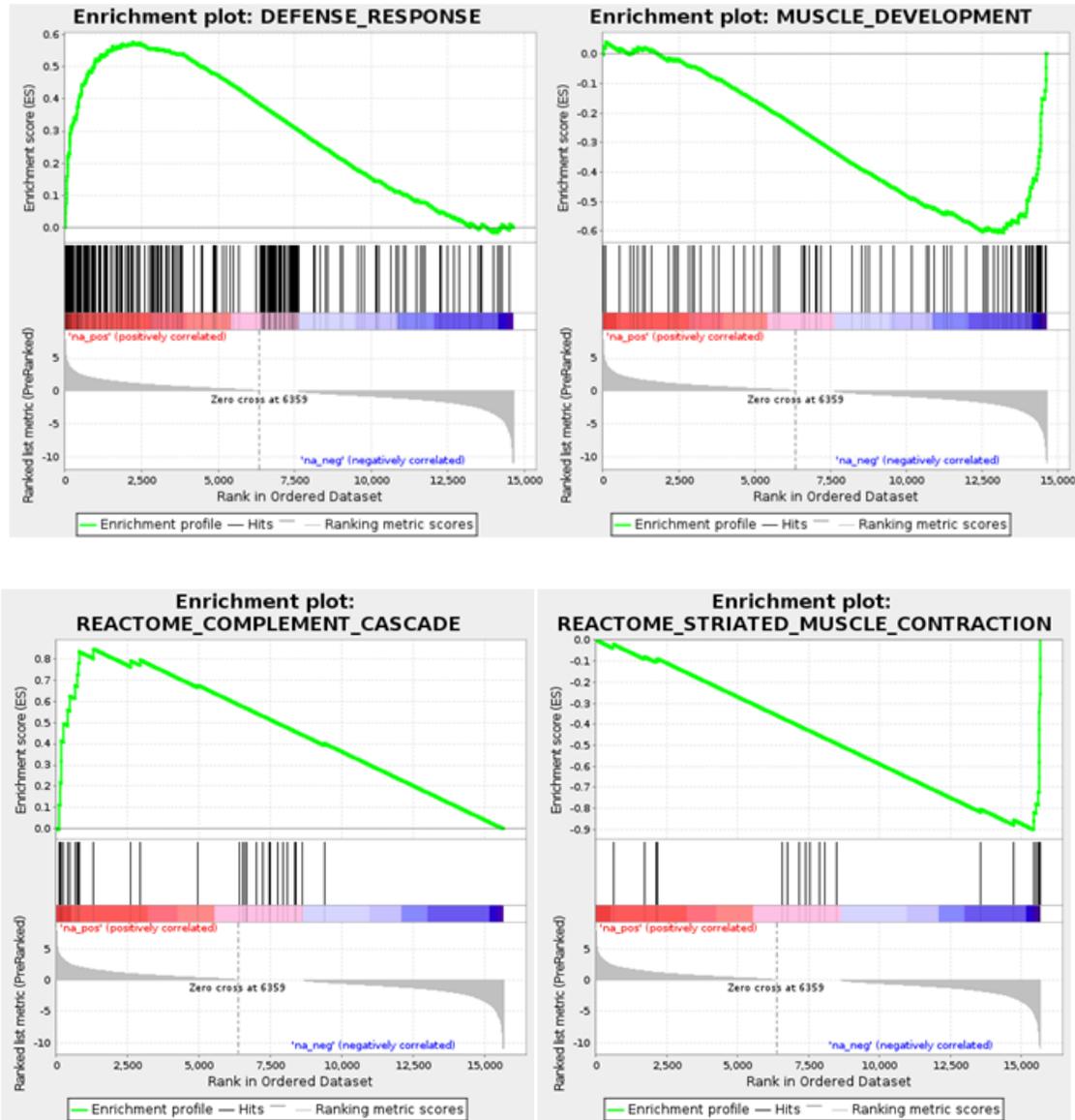
**Table 3A | Gene Set Enrichment Analysis of the RNAseq data (KEGG)**

<b>Gene ontology Biological process</b>	<b>Enrichment Score</b>	<b>FD R</b>	<b>Tissue enriched</b>
REACTOME_COMPLEMENT_CASCADE	0.84	0	lung
REACTOME_INTERFERON_ALPHA_BETA_SIGNALING	0.81	0	lung
REACTOME_CHEMOKINE_RECEPTORS_BIND_CHEMOKINES	0.8	0	lung
KEGG_GRAFT_VERSUS_HOST_DISEASE	0.79	0	lung
KEGG_AUTOIMMUNE_THYROID_DISEASE	0.78	0	lung
KEGG_ALLOGRAFT_REJECTION	0.78	0	lung
KEGG_COMPLEMENT_AND_COAGULATION_CASCADES	0.75	0	lung
REACTOME_CYTOCHROME_P450_ARRANGED_BY_SUBSTRATE_TYPE	0.74	0	lung
REACTOME_IMMUNOREGULATORY_INTERACTIONS_BETWEEN_A_LYMPHOID_AND_A_NON_LYMPHOID_CELL	0.7	0	lung
REACTOME_INTERFERON_GAMMA_SIGNALING	0.7	0	lung
REACTOME_CITRIC_ACID_CYCLE_TCA_CYCLE	-0.94	0	muscle
KEGG_CITRATE_CYCLE_TCA_CYCLE	-0.9	0	muscle
REACTOME_RESPIRATORY_ELECTRON_TRANSPORT	-0.89	0	muscle
REACTOME_RESPIRATORY_ELECTRON_TRANSPORT_ATP_SYNTHESIS_BY_CHEMIOSMOTIC_COUPLING_AND_HEAT_PRODUCTION_BY_UNCOUPLING_PROTEINS	-0.88	0	muscle
REACTOME_TCA_CYCLE_AND_RESPIRATORY_ELECTRON_TRANSPORT	-0.86	0	muscle
REACTOME_MUSCLE_CONTRACTION	-0.83	0	muscle
REACTOME_PYRUVATE_METABOLISM_AND_CITRIC_ACID_TCA_CYCLE	-0.83	0	muscle
KEGG_PARKINSONS_DISEASE	-0.81	0	muscle
KEGG_OXIDATIVE_PHOSPHORYLATION	-0.8	0	muscle
KEGG_CARDIAC_MUSCLE_CONTRACTION	-0.76	0	muscle

**Table 3B | Gene Set Enrichment Analysis of the RNAseq data (GO, biological process)**

Gene ontology Biological process	Enrichment Score	FDR	Tissue enriched
CALCIUM_INDEPENDENT_CELL_CELL_ADHESION	0.78	0	lung
DEFENSE_RESPONSE_TO_BACTERIUM	0.71	0	lung
RESPONSE_TO_BACTERIUM	0.7	0	lung
REGULATION_OF_T_CELL_ACTIVATION	0.67	0	lung
LOCOMOTORY_BEHAVIOR	0.66	0	lung
CELLULAR_DEFENSE_RESPONSE	0.64	0	lung
T_CELL_ACTIVATION	0.64	0	lung
RESPONSE_TO_VIRUS	0.62	0	lung
RESPONSE_TO_OTHER_ORGANISM	0.6	0	lung
CELL_ACTIVATION	0.6	0	lung
LEUKOCYTE_ACTIVATION	0.6	0	lung
DEFENSE_RESPONSE	0.58	0	lung
INFLAMMATORY_RESPONSE	0.58	0	lung
LYMPHOCYTE_ACTIVATION	0.58	0	lung
CELLULAR_RESPIRATION	-0.88	0	muscle
AEROBIC_RESPIRATION	-0.88	0	muscle
ENERGY_DERIVATION_BY_OXIDATION_OF_ORGANIC_COMPOUNDS	-0.83	0	muscle
REGULATION_OF_MUSCLE_CONTRACTION	-0.79	0	muscle
FATTY_ACID_OXIDATION	-0.75	0.003	muscle
ENERGY_RESERVE_METABOLIC_PROCESS	-0.73	0.023	muscle
GENERATION_OF_PRECURSOR_METABOLITES_AND_ENERGY	-0.66	0	muscle
SKELETAL_MUSCLE_DEVELOPMENT	-0.66	0.006	muscle
MITOCHONDRIAL_TRANSPORT	-0.66	0.03	muscle
REGULATION_OF_HEART_CONTRACTION	-0.64	0.029	muscle
MUSCLE_DEVELOPMENT	-0.61	0	muscle
STRIATED_MUSCLE_DEVELOPMENT	-0.6	0.019	muscle
GLUCOSE_METABOLIC_PROCESS	-0.59	0.06	muscle
MITOCHONDRION_ORGANIZATION_AND_BIOGENESIS	-0.57	0.027	muscle
ELECTRON_TRANSPORT_GO_0006118	-0.56	0.027	muscle
CELLULAR_PROTEIN_CATABOLIC_PROCESS	-0.52	0.053	muscle
PROTEIN_CATABOLIC_PROCESS	-0.5	0.054	muscle
CELLULAR_CARBOHYDRATE_METABOLIC_PROCESS	-0.47	0.05	muscle
CELLULAR_CATABOLIC_PROCESS	-0.45	0.05	muscle
CATABOLIC_PROCESS	-0.45	0.051	muscle

Figure 2 | pictorial representations of the GSEA



### 1.6. Grouping of transcripts and evidence of novel transcript/gene discovery

An investigation into transcript redundancy, potential isoforms and if the RNAseq data provides new data was carried out. There is evidence that the February 2008 *Cavia porcellus* draft assembly (Broad Institute cavPor3) contains duplicated regions of chromosomes on different scaffolds. This is shown by the fact that some transcripts map to multiple scaffolds with exactly the same high quality alignment statistics. This seems biologically unlikely, as some polymorphism differences would be expected between genuine paralogs. One example is the transcript *lung\_muscle\_trans\_00000002* that maps to 13 different scaffolds with 100% identity, over the full length of the 943 base pair sequence. This sequence does not span a known repeat. This provides good evidence of redundancy within the genome assembled transcriptome. The full assembly contained

151,072 transcripts and a subset of 86,044 was used in the design of 60,984 probes for the final custom 60K Agilent microarray. The full transcriptome sequences (**Davidson\_et\_al\_full\_transcriptome.txt**) are accessible in the online material: [http://pcwww.liv.ac.uk/~herberjm/Davidson\\_et\\_al/Additional\\_file6.zip](http://pcwww.liv.ac.uk/~herberjm/Davidson_et_al/Additional_file6.zip).

To find potential isoforms of the same gene and to reduce redundancy within the transcriptome, two approaches were taken;

1) Uclust from the Usearch algorithm was used to cluster sequences in a de-novo process [267]. This resulted in a set of 93,902 non-redundant transcripts, which represent the centroid sequence of clusters and any singletons. This was performed using a 90% identity threshold. A set of non-redundant sequences can be found in the online file **Davidson\_et\_al\_uclust\_sequences.txt** and the online file

**Davidson\_et\_al\_Uclust\_Bedtools\_domains\_genes.xlsx** contains sequence identifiers of clusters<sup>6</sup>.

2) Potential isoform groups were attained by genome position, by utilizing the BEDTools [268] cluster algorithm that assigns overlapping transcripts in the genome to locus clusters. The resulting clusters and the transcripts they contain can also be found in the online supporting material as “**Davidson\_et\_al\_Uclust\_Bedtools\_domains\_genes.xlsx**”. From the BEDTools loci, there is evidence that the RNAseq data has identified novel genes or transcripts. From the 33,797 total distinct loci, the numbers of potential novel loci are as follows;

- Total no. of BEDTools cluster loci = 33,797
- No. of loci that contain an Ensembl cDNA = 16,392
- No. of loci that contained either an Ensembl cDNA or a Genscan prediction = 27,374
- No. of loci that contained purely public EST and/or mRNA data = 686
- No. of potential novel loci originating from the RNAseq data (if it contained at least one RNAseq data transcript but no Ensembl cDNAs) = 7,685
- No. of potential novel loci originating from the RNAseq data (i.e. it contained at least one RNAseq transcript and no Ensembl cDNAs or Genscan gene predictions) = 5,737

This provides evidence that the RNA 2<sup>nd</sup> generation sequencing data has produced a more comprehensive GP transcriptome when compared to Ensembl cDNAs and Genscan predictions. This has enabled a more robust investigation into the effects of smoking using this model. A set of 5,737 novel transcript loci were found and are labelled as “RNAseq\_Novel” in the BEDTools file

(**Davidson\_et\_al\_Uclust\_Bedtools\_domains\_genes.xlsx**) available online.

### **1.7 Novel functional domains added to the transcriptome**

Further evidence for RNAseq novelty was found using a comparison of functional domain content between the RNAseq transcriptome data and Ensembl cDNA sequences. This was performed using the Hmmer package [269] with the Pfam protein families

---

<sup>6</sup> Supporting material is accessible online at [http://pcwww.liv.ac.uk/~herberjm/Davidson\\_et\\_al/Additional\\_file6.zip](http://pcwww.liv.ac.uk/~herberjm/Davidson_et_al/Additional_file6.zip)

database A [270]. Initially, open reading frames from all six translation frames of each sequence was derived using the Sixpack tool of EMBOSS, lifted from the Iprscan suite [271, 272]. An expectation value of 1e-6 threshold was employed running hmmsearch. There were 26,129 Ensembl cDNA transcripts in the cavPor3.64 genome, which consisted of 5,397 non-redundant Pfam A functional domains. In contrast, the number of transcripts built from the RNAseq data alone was 82,945 and these transcripts contained 5,478 distinct functional domains. This showed 81 potentially new functional domains as compared to the public data. However, the RNAseq data was derived from two distinct tissues of lung and muscle and some transcripts within the Ensembl cDNA data could be preferentially expressed in other tissues. This was evident when intersect and specific domain frequencies were taken. The number of domains specific to Ensembl cDNAs was 38 and 119 were specific to RNAseq data, with an intersect count of 5,359. A union total number of domains were 5,516 and all domains and classes can be viewed in the Domains sheet in **Dauidsen\_et\_al\_Uclust\_Bedtools\_domains\_genes.xlsx** available online<sup>7</sup>.

### **1.8 New representations of mouse orthologs from RNAseq transcripts**

From the full transcriptome, prior to custom array design, transcripts were assigned mouse orthologs using blastx. 19,867 of 26,129 Ensembl cDNAs were assigned a mouse ortholog and this amounted to 16,537 distinct mouse genes. In comparison, from the 82,945 RNAseq derived transcripts, 52,513 were assigned a mouse ortholog and the number of non-redundant mouse genes was 17,347. The union, intersect and specific frequencies were as follows:

- Total mouse orthologs (union) of Ensembl cDNA and RNAseq data = 17,441
- Ensembl cDNA specific mouse orthologs = 94
- RNAseq specific mouse orthologs = 904
- Intersection between Ensembl cDNA and RNAseq mouse orthologs = 16,443

From this analysis, 904 potential new GP genes were found due to RNAseq sequencing. The actual gene symbols for each of the classes can be seen in the “MouseOrthologs\_RNAseq\_v\_Ensembl” sheet in the Excel file **Dauidsen\_et\_al\_Uclust\_Bedtools\_domains\_genes.xlsx** available in the online supporting material.

### **2.0 Development of a 60K Custom Microarray Design**

A total of 116,623 60-mer oligonucleotide probes were designed from the 151,072 transcript sequences summarised in **Table 2** using the Agilent eArray software. These were used to design an initial 180K custom Agilent microarray in order to test the performance of this large set of capture probes. Pooled mRNA from GP lung and muscle samples (see Method section) were hybridised to these arrays. The selection of probes for the final 60K array design was based on an assessment of the reliability of each probe compared to the sequence read counts in both tissues of interest. This was performed using a two-step process. For each tissue, sequencing read counts

---

<sup>7</sup> Supporting material is accessible online at [http://pcwww.liv.ac.uk/~herberjm/Davidson\\_et\\_al/Additional\\_file6.zip](http://pcwww.liv.ac.uk/~herberjm/Davidson_et_al/Additional_file6.zip)

(reads per kilobase of transcripts per million (RPKM) transformed) and microarray probe intensities were LOESS normalised to generate a ratio representing the relative difference in signal intensity between both platform technologies. The ratios for each probe in lung & muscle were then compared by fitting a linear model (lm, R software), and the residual for each probe was calculated. The residuals represent an overall measure of the probe reliability when compared in both tissues, regardless of signal intensity. This approach assumes that the relationship between sequence counts and probe intensity is not proportional. Probes with similar differences between signal intensity and read counts in lung and muscle, even if this difference is relatively large, were still able to detect relative changes in mRNA concentrations between experimental groups. For each gene symbol mapped to the sequencing data, the selection of a probe to include on the final array design was done by selecting the probe with the lowest residual.

## 7.2 Cytokine superfamily as defined by commercial PCR array provider



Register for Special Offers

Please enter e-mail address

Learn more

Research Area

Complete Array List

Browse by Pathway

- Apoptosis
- Biomarkers
- Cell Cycle
- Cytokine & Inflammation
- ECM & Adhesion
- Neuroscience
- Signal Transduction
- Stem Cell & Development
- Toxicology & Drug ADME

Browse by Disease

- Cancer
- Cardiovascular Diseases
- CNS Disorders
- Immune Disorders
- Infectious Diseases
- Metabolic Diseases

Browse by Epigenetics

Home > Products > PCR Home > PCR Array > Array List > Human Cytokines & Chemokines

### Cytokines & Chemokines

**NOTE:** To access content of RT<sup>2</sup> Profiler PCR Arrays (Prior to May 25, 2012), please click [here](#).  
**Please NOTE:** The RT<sup>2</sup> Profiler PCR Array System was upgraded to Version 4.0 on May 25, 2012. This system has replaced the Version 3.0 system with regards to content and assay design, with the Version 3.0 products set for discontinuation effective December 1, 2012.

Human Mouse Rat Other Species

#### Human Cytokines & Chemokines PCR Array

Price & Ordering

The Human Cytokines & Chemokines RT<sup>2</sup> Profiler PCR Array profiles the expression of 84 key secreted proteins central to the immune response and other functions. Cytokines, small signaling proteins secreted primarily by immune cells, activate inter- and intracellular signaling during immune responses. Historically, cytokines were functionally separated into 2 families: lymphokines/interleukins and chemokines. All cytokines released by immune cells were called lymphokines/interleukins, whereas chemotactic cytokines were called chemokines. However, these family descriptions are not longer accurate because some growth factors and hormones also exhibit cellular effects very similar to cytokine family members. In addition to immune cells, many different cell types express cytokines to stimulate immune response, inflammation, and other processes. The ultimate effect of a cytokine release depends on the activated cell type expressing the specific cytokine receptor. This array includes both families of common cytokines as well as growth factors and hormones with cytokine-like properties. The results of this array should augment understanding of immune response in a variety of cell types. Using real-time PCR, research studies can easily and reliably analyze the expression of a focused panel of key cytokines and chemokines with this array.



Protocol Guide

The RT<sup>2</sup> Profiler PCR Arrays are intended for molecular biology applications. This product is not intended for the diagnosis, prevention, or treatment of a disease.

96-well Plate, 384-well (4 × 96) Plate, and 100-well Disc formats are available.

[Available for cells, tissues, FFPE samples and small samples](#)

Functional Gene Grouping How It Works Manual & Resources Reagents & Software

[Pathway Source References](#) [Modify this Array](#) [Gene Table](#)

**Chemokines:** CCL1, CCL11, CCL13, CCL17, CCL18, CCL19, CCL2, CCL20, CCL21, CCL22, CCL24, CCL3, CCL5, CCL7, CCL8, CX3CL1, CXCL1, CXCL10, CXCL11, CXCL12, CXCL13, CXCL16, CXCL2, CXCL5, CXCL9, PF4, PPBP, XCL1.

**Interleukins:** IL10, IL11, IL12A, IL12B, IL13, IL15, IL16, IL17A, IL17F, IL18, IL1A, IL1B, IL1RN, IL2, IL21, IL22, IL23A, IL24, IL27, IL3, IL4, IL5, IL6, IL7, IL8, IL9.

**Interferons:** IFNA2, IFNG.

**Growth Factors:** BMP2, BMP4, BMP6, BMP7, CNTF, CSF1, CSF2, CSF3, GPI, LIF, MSTN, NODAL, OSM, THPO, VEGFA.

**TNF Superfamily:** CD40LG, FASLG, LTA, LTB, TNF, TNFRSF11B, TNFSF10, TNFSF11, TNFSF13B.

**Other Cytokines:** ADIPOQ, MIF, SPP1, TGFB2.

**Anti-inflammatory Cytokines:** CCL18, CCL19, CCL21, IL10, IL11, IL12A, IL12B, IL13, IL18, IL2, IL22, IL23A, IL24, IL4, IL6, TGFB2.

[Pathway Source References](#) [Modify this Array](#) [Gene Table](#)

### 7.3 ANALYSIS OF SKELETAL MUSCLE TRANSCRIPTIONAL RESPONSE IN A MOUSE SMOKING MODEL

#### SUMMARY

In order to assess whether the extrapulmonary transcriptional response observed in guinea pigs exposed to smoking was conserved in the popular mouse smoking model, we analysed a publicly available microarray dataset representing the temporal transcriptional changes in the gastrocnemius muscle of nose-only CS or clean air (sham controls) exposed adult mice.

#### METHODS

##### Animal Model

In brief, 4 month old female C57/BL6 mice received nose-only exposure of 4% mainstream CS or clean air for 2 hours/day, 5 days a week for 2 (n=6), 12 (n=6) or 24 (n=6) weeks, respectively. Twenty hours following the final smoke/air exposure, mice were anesthetized with isoflurane, exsanguinated by cardiac puncture, and the gastrocnemius muscle was collected. For more information on RNA isolation as well as the labelling and hybridization protocols, see GEO accession: GSE18033.

##### Data processing and Analysis

Data processing of the raw Affymetrix CEL files was undertaken using Bioconductor in R. The arrayMvout package in R [273], which is based on dimension reduction of diverse established quality metrics, identified 8 of the 38 arrays relevant for the cross-species comparison as technical outliers at a nominal outlier flagging rate of  $\alpha = 0.05$  (GSM450983, GSM450998, GSM451003, GSM451006, GSM451012, GSM451013, GSM451014, GSM451018). Hence, these eight samples were removed prior to RMA normalization ('affy' R package). Microarray data were then de-noised by removing probes for which 75% or more of the samples were flagged as 'absent' by the MAS5 algorithm (13,051 probes removed).

#### RESULTS

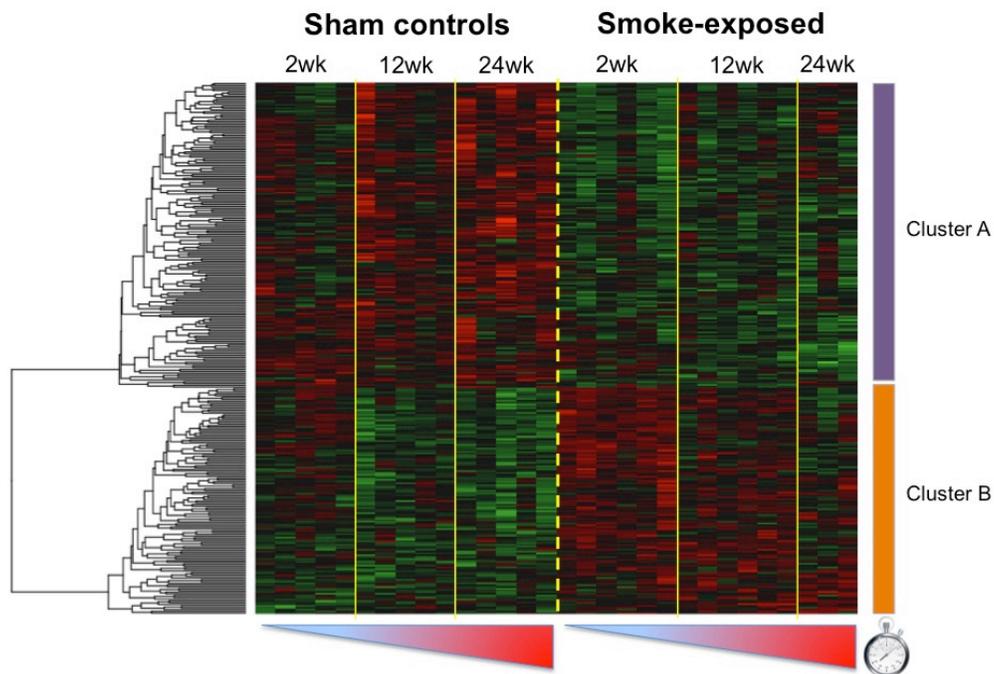
##### Transcriptional regulation in response to prolonged CS exposure can only be detected at a high False Discovery Rate threshold

In order to assess if exposure to long-term CS induced a significant transcriptional response, we set to identify differentially expressed genes. In order to identify genes linked to both time and experimental intervention (*i.e.* CS versus sham controls) we used a two-factor ANOVA (**Table 1**). While we could identify thousands of genes changing across time, only 14 genes were differentially expressed between shams and CS-exposed mice at a standard FDR cut-off (FDR<15%) (**Table 1**). By increasing the cut-off to

FDR<30% we could identify 303 modulated transcripts. These were subdivided in 129 and 174 genes up- and down-regulated, respectively, based on hierarchical clustering (**Figure 2**). We then performed a functional enrichment analysis on these 303 genes using the web based tool DAVID [274].

<i>Factor:</i>	<b>5% FDR</b>	<b>15% FDR</b>	<b>30% FDR</b>
<b>Group</b>	5	14	303
<b>Time</b>	1,774	3,769	6,217
<b>Interaction</b>	0	14	375

**Table 1.** Number of differentially expressed genes at various statistically thresholds identified via a two-way ANOVA in R.



**Figure 2.** Heatmap of the 303 genes that are differentially expressed between shams and smoke-exposed mice at a FDR < 30%. Clustering on the genes clearly identified two main clusters, with transcripts belonging to Cluster B being up-regulated by cigarette smoking.

Functional enrichment analysis of the up- and down-regulated genes, respectively, highlighted several statistically significantly (FDR<10%) enriched ontological processes of biological relevance (**Table 2**).

GO terms	Number of genes	FDR
<b><i>Down-regulated genes (Cluster A)</i></b>		
GO:0006935~chemotaxis	6	4.49
GO:0016331~morphogenesis of embryonic epithelium	5	7.88
GO:0008283~cell proliferation	8	9.98
<b><i>Up-regulated genes (Cluster B)</i></b>		
GO:0044257~cellular protein catabolic process	12	0.24
GO:0007167~enzyme linked receptor protein signalling pathway	7	5.61
GO:0000165~MAPKKK cascade	4	31.07

**Table 2.** Biological processes enriched in genes modulated by the ANOVA Group factor (FDR<30%) within each of the two main gene clusters.

We then decided to analyse each time point (*i.e.* week 2, 12 and 24 after first exposure to CS) separately using SAM analysis [166] (**Table 3**). This analysis only identified 24 genes as being differentially expressed at a reasonable FDR cut-off (FDR<15%) following daily CS exposure for 24 weeks. Only by increasing the statistical cut-off to 30%, which yielded 1,020 regulated transcripts (**Table 3**), could we detect functional enrichment using the web-based tool DAVID (**Table 4**). Notable, both 2 and 12 weeks of exposure, respectively, did not cause any transcriptional changes in the mouse limb muscle (FDR<15%). For the cross-species functional comparison (see **Figure 4** in manuscript) we relied on the SAM analysis (**Table 3**) using a FDR cut-off of 30%.

<i>Duration:</i>	5% FDR	15% FDR	30% FDR
<b>2 weeks</b>	0	0	5
<b>12 weeks</b>	0	0	4
<b>24 weeks</b>	0	24 (21 ↑)	1020 (300 ↑)

**Table 3.** Number of differentially expressed genes identified by a two-class unpaired SAM analysis. The number inside the parenthesis indicates how many genes were up-regulated by cigarette smoking.

<b>KEGG pathways</b>	<b>Number of genes</b>	<b>EASE <i>P</i>-value</b>	<b>FDR</b>
<i>Down-regulated genes</i>			
Jak-STAT signalling	10	0.05	42%
TGF-beta signalling	7	0.05	47%
Insulin signalling	8	0.14	82%
VEGF signalling	6	0.09	65%
Phosphatidylinositol signalling	7	0.03	28%
Acute myeloid leukemia	5	0.10	71%
Endometrial cancer	5	0.08	61%
Neurotrophin signalling	8	0.11	74%
Glycerophospholipid metabolism	5	0.15	86%

**Table 4.** Selected KEGG pathways that are enriched among the differentially expressed genes at week 24 of CS-exposure (FDR<30%).

## CONCLUSIONS

We conclude that the gastrocnemius muscle of C57 mice exposed to long-term CS (up to 24 weeks) does not show any detectable transcriptional changes at a reasonable statistical cut-off (FDR<15%).

## 7.4 Pre-ranked GSEA using transcription factor targets

NAME	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT M
CCAWNWWNNNGGC_UNKNOWN	56	0.522	2.0452	0	0.003	0.0024	2288
V\$SF1_Q6	168	0.406	1.9148	0	0.011	0.0192	3507
V\$MEF2_Q4	18	0.602	1.7974	0.00332226	0.035	0.0868	3475
V\$YY1_Q6	172	0.361	1.6971	3.14E-04	0.089	0.2688	3342
CTAWWWATA_V\$RSRFC4_Q2	251	0.343	1.6882	0	0.078	0.2912	3733
GCCNNWTAAR_UNKNOWN	96	0.386	1.6719	0.00135455	0.079	0.3432	2283
TAAWWATAG_V\$RSRFC4_Q2	106	0.361	1.5924	0.001665	0.170	0.6516	3695
V\$CDP_Q1	52	0.407	1.563	0.01544799	0.206	0.77	4686
V\$FXR_IR1_Q6	67	0.386	1.5613	0.00944669	0.187	0.775	1686
V\$NFMUE1_Q6	177	0.328	1.5534	6.27E-04	0.183	0.804	3159
TGACCTTG_V\$SF1_Q6	166	0.33	1.5403	0.00190235	0.192	0.846	3422
TCCRNRTGC_UNKNOWN	136	0.329	1.4986	0.00387097	0.275	0.943	4150
V\$E2F1_Q6	162	0.321	1.4972	0.00512821	0.257	0.944	4343
GGAMTNNNNNTCCY_UNKNOWN	80	0.353	1.4831	0.01749664	0.276	0.9674	2705
CCGNMNTNACG_UNKNOWN	59	0.374	1.4726	0.0260989	0.289	0.9802	2866
CTCNANGTGNV_UNKNOWN	61	0.369	1.4698	0.02526316	0.279	0.9818	2677
V\$E2F1_Q6_Q1	156	0.313	1.4625	0.00798212	0.283	0.9876	4050
V\$E2F_Q6	156	0.311	1.4515	0.01020083	0.300	0.9918	4185
V\$AP2_Q6_Q1	172	0.306	1.4508	0.00637146	0.286	0.9922	3032
V\$TBP_Q1	162	0.309	1.4454	0.00947269	0.288	0.9952	4152
V\$AP4_Q1	170	0.307	1.4433	0.00983191	0.281	0.9956	2816
V\$E2F1DP2_Q1	160	0.307	1.435	0.00963391	0.291	0.9974	4343
V\$E2F4DP2_Q1	160	0.307	1.4335	0.0131579	0.283	0.9974	4343
V\$E2F1DP1_Q1	160	0.307	1.4322	0.00869565	0.275	0.9974	4343
V\$E2F_Q2	161	0.305	1.4275	0.00919759	0.277	0.998	4343
V\$MYOD_Q6	153	0.308	1.4273	0.01193548	0.267	0.998	3371
ATCMNTCCGY_UNKNOWN	35	0.404	1.4226	0.05633803	0.269	0.9984	1532
V\$MYOD_Q1	176	0.299	1.4154	0.01039698	0.279	0.9988	3371
GCCATNTTG_V\$YY1_Q6	288	0.282	1.4128	0.00420168	0.276	0.9988	3364
V\$E2F_Q4	158	0.301	1.4071	0.01533057	0.283	0.999	4185
V\$ERR1_Q2	180	0.298	1.4068	0.01168298	0.274	0.999	3343
V\$E2F_Q1	51	0.367	1.3981	0.05549881	0.289	0.9996	2198
V\$MEF2_Q1	98	0.323	1.3954	0.02893785	0.287	0.9998	3485
V\$E2F4DP1_Q1	164	0.297	1.3954	0.01254019	0.279	0.9998	4343
SGCGSSAAA_V\$E2F1DP2_Q1	116	0.314	1.3937	0.02702703	0.276	0.9998	2626
V\$E12_Q6	180	0.293	1.3891	0.01318268	0.280	0.9998	3050
V\$SMAD4_Q6	148	0.3	1.389	0.0200526	0.273	0.9998	3873
AGCYRWTTT_UNKNOWN	71	0.337	1.387	0.04233527	0.271	0.9998	2542
V\$RSRFC4_Q2	152	0.297	1.3793	0.02427638	0.284	0.9998	3614
V\$AP2_Q6	178	0.289	1.3671	0.01728473	0.311	0.9998	2988
V\$NFY_Q6_Q1	168	0.29	1.3634	0.02510986	0.314	1	4040
AAGWWRNYGGC_UNKNOWN	83	0.323	1.3628	0.04270343	0.309	1	1830
V\$MYOD_Q6_Q1	172	0.287	1.3595	0.02092581	0.311	1	3385
V\$T3R_Q6	155	0.291	1.3551	0.02410016	0.317	1	4132
V\$MEF2_Q6_Q1	157	0.291	1.3546	0.03229974	0.312	1	3695
V\$AP4_Q6_Q1	167	0.286	1.3466	0.02768274	0.328	1	2580
GGCKCATGS_UNKNOWN	43	0.362	1.3442	0.08260106	0.329	1	1310
TGGNNNNNNKCCAR_UNKNOWN	281	0.268	1.3403	0.01457032	0.334	1	2537
V\$HMX1_Q1	29	0.395	1.3339	0.10550954	0.347	1	1673
V\$MEF2_Q2	156	0.284	1.3245	0.03740409	0.371	1	3499
V\$P53_DECAMER_Q2	162	0.284	1.3228	0.03722558	0.369	1	3879
V\$GATA_C	165	0.28	1.3091	0.04035309	0.409	1	2681
V\$NKX62_Q2	147	0.282	1.3042	0.04556036	0.420	1	3260
V\$E2F1_Q3	168	0.278	1.301	0.04253308	0.424	1	4327
V\$ZIC2_Q1	165	0.277	1.3009	0.04161291	0.417	1	3084
GGCNMMSMYNTTG_UNKNOWN	56	0.329	1.2857	0.10746685	0.468	1	2775
V\$HLF_Q1	164	0.272	1.2827	0.05170199	0.472	1	3762
V\$E47_Q1	176	0.27	1.2824	0.05456286	0.465	1	3314
V\$RF1_Q2	182	0.269	1.2816	0.04529727	0.460	1	3825
V\$GATA1_Q1	151	0.275	1.2738	0.07031499	0.484	1	3084

Positive end of ranked gene list (top 60 gene sets shown)

NAME	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX
V\$PAX8_B	66	-0.404	-1.7045	4.91E-04	0.146	0.1792	1952
V\$ISRE_01	169	-0.305	-1.5125	0.00111173	0.528	0.7612	3345
RYTGCNNRGNAAAC_V\$MIF1_01	59	-0.359	-1.4829	0.02028302	0.472	0.8492	3745
TTCYNRGAA_V\$STAT5B_01	220	-0.283	-1.4616	0.00280426	0.438	0.9074	2881
CCAWWNAAGG_V\$SRF_Q4	65	-0.345	-1.4564	0.02640111	0.369	0.917	2863
V\$IK2_01	176	-0.287	-1.4303	0.00660793	0.397	0.9576	2592
V\$AR_Q2	72	-0.32	-1.3819	0.03986388	0.535	0.9942	1501
RYTAAWNNNTGAY_UNKNOWN	41	-0.364	-1.3757	0.06267806	0.497	0.9964	3126
V\$PBX1_01	157	-0.279	-1.3723	0.01600854	0.456	0.9972	2518
V\$NFKB_Q6	174	-0.273	-1.3655	0.01583949	0.438	0.9978	3555
V\$IRF_Q6	168	-0.274	-1.3621	0.01553294	0.411	0.9982	3802
V\$NFAT_Q4_01	187	-0.267	-1.3414	0.01224944	0.456	0.9996	2211
V\$OCT1_02	129	-0.28	-1.3378	0.03562075	0.434	0.9998	2747
V\$PEA3_Q6	173	-0.266	-1.334	0.01534247	0.417	0.9998	2622
RACCACAR_V\$AML_Q6	169	-0.269	-1.3334	0.01687534	0.391	0.9998	2536
YGACNNYACAR_UNKNOWN	59	-0.319	-1.3228	0.07739212	0.403	1	2619
AACWWCAANK_UNKNOWN	95	-0.285	-1.2861	0.06289926	0.526	1	2530
V\$ICSBP_Q6	176	-0.256	-1.2854	0.03366446	0.500	1	3427
V\$POU3F2_01	62	-0.307	-1.2843	0.09033203	0.479	1	2933
CAGNWMCNNGAC_UNKNOWN	53	-0.312	-1.2716	0.10443491	0.506	1	2772
V\$PAX2_02	174	-0.255	-1.2711	0.04049014	0.484	1	2041
V\$CREL_01	171	-0.255	-1.2645	0.0475927	0.490	1	2182
V\$CIZ_01	156	-0.255	-1.2572	0.06203996	0.500	1	3451
WTGAAAT_UNKNOWN	379	-0.227	-1.2404	0.02186134	0.553	1	2520
V\$STAT1_01	43	-0.317	-1.2278	0.15542112	0.590	1	3883
V\$NFKAPPAB_01	157	-0.249	-1.2259	0.07315761	0.577	1	1608
V\$SRF_Q4	168	-0.247	-1.2227	0.07650273	0.571	1	2723
V\$POU1F1_Q6	148	-0.252	-1.2225	0.08431877	0.551	1	2492
V\$FOXMI_01	155	-0.248	-1.2214	0.08201058	0.537	1	2304
V\$STAT_Q6	179	-0.244	-1.2202	0.08112493	0.524	1	2862
V\$SCOREBINDINGFACTOR_Q6	173	-0.242	-1.2093	0.08144044	0.554	1	3661
V\$FREAC4_01	97	-0.261	-1.1874	0.14484127	0.639	1	1980
V\$STAT5A_01	163	-0.241	-1.1874	0.1050626	0.620	1	2785
V\$NFKB_Q6_01	146	-0.245	-1.1816	0.12459721	0.631	1	3717
RRAGTTGT_UNKNOWN	166	-0.24	-1.1812	0.11561002	0.614	1	2780
V\$HNF3ALPHA_Q6	138	-0.244	-1.1723	0.12342437	0.640	1	2827
V\$IPF1_Q4	147	-0.239	-1.1654	0.13944857	0.657	1	2074
V\$IRF1_Q6	163	-0.236	-1.1564	0.14630225	0.684	1	3764
AAAYWAACM_V\$HFH4_01	173	-0.23	-1.1539	0.13704497	0.680	1	1460
WYAAANRRNNGCG_UNKNOWN	45	-0.296	-1.1532	0.2260431	0.666	1	4313
V\$AREB6_04	169	-0.23	-1.146	0.15504292	0.687	1	2641
V\$FREAC3_01	157	-0.233	-1.1442	0.16256158	0.679	1	2330
V\$FOXO3_01	169	-0.231	-1.144	0.15726027	0.665	1	2931
WGTTNNNNNAAA_UNKNOWN	356	-0.209	-1.1332	0.10817772	0.704	1	2828
V\$AML_Q6	170	-0.227	-1.1304	0.17458564	0.702	1	2622
V\$CEBPB_01	175	-0.225	-1.1303	0.16657797	0.687	1	2881
V\$HFH3_01	138	-0.233	-1.1246	0.1938349	0.701	1	2473
YNTTTNNNANGCARM_UNKNOWN	43	-0.288	-1.1236	0.24873795	0.691	1	2611
V\$IRF1_01	169	-0.226	-1.1183	0.19340897	0.703	1	3764
V\$CDP_02	67	-0.259	-1.1068	0.25502393	0.749	1	2466
V\$HNF3_Q6	126	-0.231	-1.1025	0.24139756	0.757	1	2827
V\$IRF2_01	83	-0.249	-1.0984	0.27348337	0.764	1	3602
V\$TST1_01	163	-0.221	-1.0966	0.22366288	0.759	1	2536
V\$HFH1_01	169	-0.218	-1.0878	0.23883067	0.791	1	2641
V\$PAX4_02	154	-0.222	-1.0848	0.26033935	0.793	1	3606
V\$OCT1_04	145	-0.224	-1.0791	0.27220178	0.809	1	2641
RGAGGAARY_V\$PU1_Q6	320	-0.2	-1.0789	0.21253918	0.796	1	3197
MYAATNNNNNNGGC_UNKNOWN	74	-0.25	-1.0766	0.30717054	0.794	1	737
V\$TEF1_Q6	147	-0.221	-1.0735	0.28938907	0.797	1	2899

Negative end of ranked gene list (top 60 gene sets shown)

## 7.5 MEF2A interacting targets as defined in the STRING database (confidence score 0.8 or more).

Gene Symbol	Gene name	Confidence score
EP300	E1A binding protein p300; Functions as histone acetyltransferase and regulates transcription vi [...] (2414 aa)	0.997
HDAC5	histone deacetylase 5; Responsible for the deacetylation of lysine residues on the N-terminal p [...] (1123 aa)	0.996
MAPK14	mitogen-activated protein kinase 14; Serine/threonine kinase which acts as an essential compone [...] (360 aa)	0.993
HDAC9	histone deacetylase 9; Responsible for the deacetylation of lysine residues on the N-terminal p [...] (1069 aa)	0.991
MEF2D	myocyte enhancer factor 2D; Transcriptional activator which binds specifically to the MEF2 elem [...] (521 aa)	0.99
MAPK7	mitogen-activated protein kinase 7; Plays a role in various cellular processes such as prolifer [...] (816 aa)	0.984
HDAC4	histone deacetylase 4; Responsible for the deacetylation of lysine residues on the N-terminal p [...] (1084 aa)	0.964
HDAC7	histone deacetylase 7 (991 aa)	0.962
MYOD1	myogenic differentiation 1; Involved in muscle differentiation (myogenic factor). Induces fibro [...] (320 aa)	0.955
MYOG	myogenin (myogenic factor 4); Involved in muscle differentiation (myogenic factor). Induces fib [...] (224 aa)	0.955
CABIN1	calcineurin binding protein 1; May be required for replication-independent chromatin assembly. [...] (2220 aa)	0.954
TCF3	transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47); Transcriptional r [...] (654 aa)	0.923
MAPK11	mitogen-activated protein kinase 11; Serine/threonine kinase which acts as an essential compone [...] (364 aa)	0.922
MAPK12	mitogen-activated protein kinase 12; Serine/threonine kinase which acts as an essential compone [...] (367 aa)	0.922
MAPK13	mitogen-activated protein kinase 13; Serine/threonine kinase which acts as an essential compone [...] (365 aa)	0.922
SUMO1	SMT3 suppressor of mif two 3 homolog 1 (S. cerevisiae); Ubiquitin-like protein that can be cova [...] (101 aa)	0.915
AKT1	v-akt murine thymoma viral oncogene homolog 1; AKT1 is one of 3 closely related serine/threonin [...] (480 aa)	0.914
KAT2B	K(lysine) acetyltransferase 2B; Functions as a histone acetyltransferase (HAT) to promote trans [...] (832 aa)	0.905
CREBBP	CREB binding protein; Acetylates histones, giving a specific tag for transcriptional activation [...] (2442 aa)	0.905
ABL1	c-abl oncogene 1, non-receptor tyrosine kinase; Non-receptor tyrosine-protein kinase that plays [...] (1149 aa)	0.902
CTNNA2	catenin (cadherin-associated protein), alpha 2 (905 aa)	0.899
CAMK2A	calcium/calmodulin-dependent protein kinase II alpha; CaM-kinase II (CAMK2) is a prominent kina [...] (489 aa)	0.899
GRIP1	glutamate receptor interacting protein 1; May play a role as a localized scaffold for the assem [...] (1076 aa)	0.899
SLC2A14	solute carrier family 2 (facilitated glucose transporter), member 14; Facilitative glucose tran [...] (520 aa)	0.899
CAMK2B	calcium/calmodulin-dependent protein kinase II beta (666 aa)	0.899
PPP3CA	protein phosphatase 3, catalytic subunit, alpha isozyme; Calcium-dependent, calmodulin-stimulat [...] (521 aa)	0.899
PPP3CB	protein phosphatase 3, catalytic subunit, beta isozyme; Calcium-dependent, calmodulin-stimulate [...] (525 aa)	0.899
JUN	jun proto-oncogene; Transcription factor that recognizes and binds to the enhancer heptamer mot [...] (331 aa)	0.899
CTNNB1	catenin (cadherin-associated protein), beta 1, 88kDa; Key downstream component of the canonical [...] (781 aa)	0.899
MEF2C	myocyte enhancer factor 2C; Transcription activator which binds specifically to the MEF2 elemen [...] (483 aa)	0.899
CAMK2D	calcium/calmodulin-dependent protein kinase II delta; Calcium/calmodulin-dependent protein kina [...] (499 aa)	0.899
NFATC1	nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 1; Plays a role in the [...] (930 aa)	0.899
CAMK2G	calcium/calmodulin-dependent protein kinase II gamma (556 aa)	0.899
CDC42	cell division cycle 42 (GTP binding protein, 25kDa); Plasma membrane-associated small GTPase wh [...] (191 aa)	0.899
CTNNA1	catenin (cadherin-associated protein), alpha 1, 102kDa; Associates with the cytoplasmic domain [...] (906 aa)	0.899
CDH15	cadherin 15, type 1, M-cadherin (myotubule); Cadherins are calcium-dependent cell adhesion prot [...] (814 aa)	0.899
CAMK4	calcium/calmodulin-dependent protein kinase IV; Calcium/calmodulin-dependent protein kinase tha [...] (473 aa)	0.899
CDH2	cadherin 2, type 1, N-cadherin (neuronal); Cadherins are calcium-dependent cell adhesion protei [...] (906 aa)	0.899
BNIP2	BCL2/adenovirus E1B 19kDa interacting protein 2; Implicated in the suppression of cell death. I [...] (435 aa)	0.899
PPARGC1A	peroxisome proliferator-activated receptor gamma, coactivator 1 alpha; Transcriptional coactiva [...] (798 aa)	0.899
CDON	Cdon homolog (mouse); Component of a cell-surface receptor complex that mediates cell-cell inte [...] (1264 aa)	0.899
SPAG9	sperm associated antigen 9 (1321 aa)	0.899
PPP3CC	protein phosphatase 3, catalytic subunit, gamma isozyme; Calcium-dependent, calmodulin-stimulat [...] (512 aa)	0.899
MYF5	myogenic factor 5; Involved in muscle differentiation (myogenic factor). Induces fibroblasts to [...] (255 aa)	0.899
MYF6	myogenic factor 6 (herculin); Involved in muscle differentiation (myogenic factor). Induces fib [...] (242 aa)	0.899
ESRRA	estrogen-related receptor alpha; Binds to an ERR-alpha response element (ERRE) containing a sin [...] (423 aa)	0.899
ASCL1	achaete-scute complex homolog 1 (Drosophila); Transcriptional regulator. May play a role at ear [...] (236 aa)	0.812
NFATC4	nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 4 (964 aa)	0.8
NFATC2	nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 2 (925 aa)	0.8
NFATC3	nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 3; Acts as a regulator [...] (1075 aa)	0.8

## 7.6 List of GWAS SNPs associated with serum triglyceride response to exercise training among Caucasians in the HERITAGE study

SNP	Chromosome	Position*	Gene†	MAF	Regression model			Remaining Heritability‡
					Partial R <sup>2</sup>	Model R <sup>2</sup>	p Value	
rs222158	21	26 794 032	<i>CYR1</i>	0.33	0.055	0.055	2.32×10 <sup>-7</sup>	9.48%
rs2722171	12	102 973 617	<i>GLT8D2</i>	0.20	0.041	0.097	4.70×10 <sup>-6</sup>	6.4%
rs1906058	16	6 084 649	<i>RBFOX1</i>	0.47	0.039	0.135	6.19×10 <sup>-6</sup>	2.6%
rs2593324	3	22 094 225	<i>ZNF385D</i>	0.38	0.037	0.172	6.98×10 <sup>-6</sup>	0%
rs12659606	5	123 591 568	<i>ZNF608</i> (400 kb)	0.11	0.032	0.204	1.82×10 <sup>-5</sup>	NA
rs2190798	19	33 141 651	<i>LOC102724694</i>	0.25	0.028	0.231	5.30×10 <sup>-5</sup>	NA
rs726553	2	225 724 738	<i>DOCK10</i> (100 kb)	0.37	0.025	0.256	8.40×10 <sup>-5</sup>	NA
rs7850237	9	89 643 439	<i>SPATA31C1</i> (75 kb)	0.14	0.027	0.283	3.66×10 <sup>-5</sup>	NA
rs9357234	6	37 138 195	<i>FGD2</i> (33 kb)	0.31	0.020	0.303	0.0003	NA
rs2646822	1	215 602 483	<i>GPATCH2</i> (69 kb)	0.23	0.017	0.320	0.0007	NA
rs3736487	2	189 564 188	<i>COL3A1</i>	0.24	0.012	0.332	0.0038	NA
rs13093483	3	68 376 923	<i>FAM19A1</i>	0.12	0.012	0.344	0.0038	NA
rs1889879	6	69 720 601	<i>BAI3</i>	0.39	0.011	0.356	0.0046	NA
rs10520872	5	21 709 737	<i>LOC105374685</i>	0.14	0.010	0.366	0.0066	NA
rs1452404	4	109 599 17 7	<i>LEF1</i> (290 kb)	0.10	0.011	0.377	0.0053	NA
rs11666431	19	2 904 087	<i>ZNF77</i> (9 kb)	0.42	0.009	0.386	0.0103	NA
rs3861882	9	131 505 125	<i>PRRX2</i>	0.28	0.008	0.394	0.0143	NA
rs9469986	6	11 857 166	<i>ADTRP</i>	0.20	0.008	0.401	0.0165	NA
rs2158244	7	111 610 670	<i>DOCK4</i>	0.31	0.007	0.408	0.0218	NA
rs4742057	9	4 943 916	<i>JAK2</i> (32 kb)	0.42	0.006	0.414	0.0333	NA

\*Positions are relative to Human Genome National Center for Biotechnology Information (NCBI) Build 36.3.  
†The gene located nearest to the SNP. Distance to the gene in kilo bases (1000 bp) is shown in parentheses. If no distance is shown, the SNP is located within the gene locus.  
‡Remaining heritability estimate when a given SNP (plus preceding SNPs) is included as covariate(s) in the MERLIN heritability model.  
GWAS, genome-wide association study; MAF, minor allele frequency; NA, not applicable; SNP, single nucleotide polymorphisms; TG, triglycerides.

## 7.7 Result of pathway-based analysis of GWAS associations

NAME	SIZE	ES	NES	NOM p-value	FDR q-value
LONG-TERM_DEPRESSION	63	0.504	1.587	0.000	0.161
RENAL_CELL_CARCINOMA	62	0.493	1.562	0.000	0.118
GLYCOSPHINGOLIPID_BIOSYNTHESIS_LACTO_AND_NEOLACTO_SERIES	19	0.526	1.464	0.019	0.303
VEGF_SIGNALING_PATHWAY	66	0.458	1.452	0.001	0.263
AXON_GUIDANCE	114	0.434	1.429	0.000	0.280
ENDOMETRIAL_CANCER	45	0.464	1.408	0.002	0.309
ARRHYTHMOGENIC_RIGHT_VENTRICULAR_CARDIOMYOPATHY_ARVC	68	0.444	1.404	0.001	0.280
LONG-TERM_POTENTIATION	60	0.429	1.343	0.012	0.496
CARBOHYDRATE_DIGESTION_AND_ABSORPTION	37	0.444	1.336	0.029	0.485
PORPHYRIN_AND_CHLOROPHYLL_METABOLISM	31	0.452	1.329	0.042	0.472
FC_GAMMA_R-MEDIATED_PHAGOCYTOSIS	80	0.413	1.327	0.008	0.439
B_CELL_RECEPTOR_SIGNALING_PATHWAY	61	0.419	1.319	0.010	0.439
HEPATITIS_C	114	0.394	1.289	0.004	0.585
BACTERIAL_INVASION_OF_EPITHELIAL_CELLS	61	0.412	1.286	0.017	0.558
GLUTAMATERGIC_SYNAPSE	117	0.390	1.282	0.004	0.546
TYPE_II_DIABETES_MELLITUS	42	0.422	1.281	0.034	0.515
HYPERTROPHIC_CARDIOMYOPATHY_HCM	72	0.400	1.281	0.020	0.488
CARDIAC_MUSCLE_CONTRACTION	56	0.411	1.280	0.033	0.465
VASCULAR_SMOOTH_MUSCLE_CONTRACTION	105	0.388	1.268	0.011	0.505
NATURAL_KILLER_CELL_MEDIATED_CYTOTOXICITY	108	0.386	1.264	0.014	0.500

ES, enrichment score; NES, normalized enrichment score; FDR, false discovery rate; FWER, familywise-error rate.

NOM p-value: Nominal p value; that is, the statistical significance of the ES.

FDR q-value: the estimated probability that the normalized enrichment score represents a false positive finding.

GSEA report for top 20 gene sets for  $\Delta$ TG using Stouffer's method to calculate gene-level statistics

NAME	SIZE	ES	NES	NOM p-value	FDR q-value
CELL_ADHESION_MOLECULES_CAMS	121	-0.34	-1.89	0.00	0.12
GRAFT-VERSUS-HOST_DISEASE	31	-0.43	-1.70	0.01	0.39
TYPE_I_DIABETES_MELLITUS	36	-0.39	-1.65	0.01	0.37
LEISHMANIASIS	58	-0.33	-1.58	0.01	0.48
LONG-TERM_DEPRESSION	63	-0.33	-1.55	0.02	0.45
GLUTAMATERGIC_SYNAPSE	117	-0.28	-1.55	0.01	0.40
PROTEASOME	30	-0.39	-1.54	0.04	0.36
INTESTINAL_IMMUNE_NETWORK_FOR_IGA_PRODUCTION	38	-0.36	-1.53	0.03	0.33
GAP_JUNCTION	78	-0.30	-1.52	0.01	0.32
GLYCOSAMINOGLYCAN_BIOSYNTHESIS_CHONDROITIN_SULFATE	17	-0.46	-1.51	0.06	0.31
ASTHMA	23	-0.41	-1.50	0.04	0.29
ALLOGRAFT_REJECTION	30	-0.37	-1.49	0.04	0.29
CALCIUM_SIGNALING_PATHWAY	151	-0.26	-1.48	0.00	0.28
ALANINE_ASPARTATE_AND_GLUTAMATE_METABOLISM	31	-0.37	-1.48	0.04	0.27
ANTIGEN_PROCESSING_AND_PRESENTATION	49	-0.32	-1.48	0.04	0.25
RHEUMATOID_ARTHRITIS	74	-0.30	-1.47	0.04	0.26
AFRICAN_TRYPANOSOMIASIS	30	-0.36	-1.44	0.05	0.29
FC_GAMMA_R-MEDIATED_PHAGOCYTOSIS	80	-0.28	-1.41	0.04	0.32
ECM-RECEPTOR_INTERACTION	74	-0.28	-1.41	0.05	0.32
MATURITY_ONSET_DIABETES_OF_THE_YOUNG	21	-0.40	-1.40	0.10	0.32

ES, enrichment score; NES, normalized enrichment score; FDR, false discovery rate; FWER, familywise-error rate.

NOM p-value: Nominal p value; that is, the statistical significance of the ES.

FDR q-value: the estimated probability that the normalized enrichment score represents a false positive finding.

GSEA report for top 20 gene sets with negative enrichment scores for  $\Delta$ TG using 2nd-best-p-value method to calculate gene-level statistics

## 8 Bibliography

1. Pedersen BK, Febbraio MA: **Muscle as an endocrine organ: focus on muscle-derived interleukin-6.** *Physiol Rev* 2008, **88**:1379–406.
2. Srikanthan P, Karlamangla AS: **Relative muscle mass is inversely associated with insulin resistance and prediabetes. Findings from the third National Health and Nutrition Examination Survey.** *J Clin Endocrinol Metab* 2011, **96**:2898–903.
3. Gallagher D, Belmonte D, Deurenberg P, Wang Z, Krasnow N, Pi-Sunyer FX, Heymsfield SB: **Organ-tissue mass measurement allows modeling of REE and metabolically active tissue mass.** *Am J Physiol* 1998, **275**(2 Pt 1):E249–58.
4. Zurlo F, Larson K, Bogardus C, Ravussin E: **Skeletal muscle metabolism is a major determinant of resting energy expenditure.** *J Clin Invest* 1990, **86**:1423–7.
5. Metter EJ, Talbot LA, Schrage M, Conwit R: **Skeletal muscle strength as a predictor of all-cause mortality in healthy men.** *J Gerontol A Biol Sci Med Sci* 2002, **57**:B359–65.
6. Ruiz JR, Sui X, Lobelo F, Morrow JR, Jackson AW, Sjöström M, Blair SN: **Association between muscular strength and mortality in men: prospective cohort study.** *BMJ* 2008, **337**:a439.
7. FitzGerald S, Barlow C, Kampert J, Morrow J, Jackson A, Blair S: **Muscular Fitness and All-Cause Mortality: Prospective Observations.** *Journal of Physical Activity and Health* 2004:7 – 18.
8. Sherwood L: *Human Physiology: From Cells to Systems, 9th Edition.* Boston: Cengage Learning; 2015.
9. Scott W, Stevens J, Binder-Macleod SA: **Human skeletal muscle fiber type classifications.** *Phys Ther* 2001, **81**:1810–6.
10. Egan B, Zierath JR: **Exercise metabolism and the molecular regulation of skeletal muscle adaptation.** *Cell Metab* 2013, **17**:162–84.
11. Schiaffino S, Reggiani C: **Fiber types in mammalian skeletal muscles.** *Physiol Rev* 2011, **91**:1447–531.
12. Marini M, Veicsteinas A: **The exercised skeletal muscle: a review.** *European Journal of Translational Myology* 2010:105–120.

13. Gouzi F, Maury J, Molinari N, Pomies P, Mercier J, Prefaut C, Hayot M: **Reference values for vastus lateralis fiber size and type in healthy subjects over 40 years old: a systematic review and metaanalysis.** *J Appl Physiol* 2013, **115**:346–354.
14. Trappe S, Luden N, Minchev K, Raue U, Jemiolo B, Trappe TA: **Skeletal muscle signature of a champion sprint runner.** *J Appl Physiol* 2015, **118**:1460–6.
15. Hoppeler H, Flück M: **Normal mammalian skeletal muscle and its phenotypic plasticity.** *J Exp Biol* 2002, **205**(Pt 15):2143–52.
16. Tidball JG: **Mechanical signal transduction in skeletal muscle growth and adaptation.** *J Appl Physiol* 2005, **98**:1900–8.
17. Jeffery Mador M, Bozkanat E: **Skeletal muscle dysfunction in chronic obstructive pulmonary disease.** *Respir Res* 2001, **2**:216–224.
18. Mathieu-Costello O: **Muscle Adaptation to Altitude: Tissue Capillarity and Capacity for Aerobic Metabolism.** 2004.
19. Keller P, Vollaard NBJ, Gustafsson T, Gallagher IJ, Sundberg CJ, Rankinen T, Britton SL, Bouchard C, Koch LG, Timmons JA: **A transcriptional map of the impact of endurance exercise training on skeletal muscle phenotype.** *J Appl Physiol* 2011, **110**:46–59.
20. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L: **Integrated genomic and proteomic analyses of a systematically perturbed metabolic network.** *Science* 2001, **292**:929–34.
21. Gea J, Agusti A, Roca J: **PATHOPHYSIOLOGY OF MUSCLE DYSFUNCTION IN COPD.** *J Appl Physiol* 2013, **114**:1222–34.
22. Timmons JA: **Variability in training-induced skeletal muscle adaptation.** *J Appl Physiol* 2011, **110**:846–53.
23. Mann TN, Lamberts RP, Lambert MI: **High responders and low responders: factors associated with individual variation in response to standardized training.** *Sports Med* 2014, **44**:1113–24.
24. Timmons JA, Knudsen S, Rankinen T, Koch LG, Sarzynski M, Jensen T, Keller P, Scheele C, Vollaard NBJ, Nielsen S, Akerström T, MacDougald OA, Jansson E, Greenhaff PL, Tarnopolsky MA, van Loon LJC, Pedersen BK, Sundberg CJ, Wahlestedt C, Britton SL, Bouchard C: **Using molecular classification to predict gains in maximal aerobic capacity following endurance exercise training in humans.** *J Appl Physiol* 2010, **108**:1487–96.

25. Timmons JA, Helge JW: **A Primer on Systems Biology, as Applied to Exercise Physiology and Metabolism.** In *Genetic and Molecular Aspects of Sports Performance*. Wiley-Blackwell; 2011:309–317.
26. Noble D: *The Music of Life: Biology beyond Genes*. Oxford University Press; 2008.
27. Mattson DL: **Functional Genomics.** In *Integrative Physiology in the Proteomics and Post-Genomics Age*. Edited by Walz W. Humana Press Inc.; 2005:7–26.
28. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, et al.: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**:860–921.
29. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, et al.: **The sequence of the human genome.** *Science* 2001, **291**:1304–51.
30. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C: **Detection of large-scale variation in the human genome.** *Nat Genet* 2004, **36**:949–51.
31. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M: **Large-scale copy number polymorphism in the human genome.** *Science* 2004, **305**:525–8.
32. Bouchard C: **Exercise genomics-a paradigm shift is needed: a commentary.** *Br J Sports Med* 2015.
33. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AWC, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers Y-H, Frazier ME, Scherer SW, Strausberg RL, et al.: **The diploid genome sequence of an individual human.** *PLoS Biol* 2007, **5**:e254.
34. Maresso K, Broeckel U: **Genotyping platforms for mass-throughput genotyping with SNPs, including human genome-wide scans.** *Adv Genet* 2008, **60**:107–39.

35. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467**:1061–73.
36. Ha N-T, Freytag S, Bickeboeller H: **Coverage and efficiency in current SNP chips.** *Eur J Hum Genet* 2014, **22**:1124–30.
37. WATSON JD, CRICK FH: **Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid.** *Nature* 1953, **171**:737–8.
38. LaFramboise T: **Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances.** *Nucleic Acids Res* 2009, **37**:4181–93.
39. Perteua M: **The human transcriptome: an unfinished story.** *Genes (Basel)* 2012, **3**:344–60.
40. Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M: **Global quantification of mammalian gene expression control.** *Nature* 2011, **473**:337–42.
41. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG: **Genetic analysis of genome-wide variation in human gene expression.** *Nature* 2004, **430**:743–7.
42. Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, Lyle R, Hunt S, Kahl B, Antonarakis SE, Tavaré S, Deloukas P, Dermitzakis ET: **Genome-wide associations of gene expression variation in humans.** *PLoS Genet* 2005, **1**:e78.
43. Wiench M, John S, Baek S, Johnson TA, Sung M-H, Escobar T, Simmons CA, Pearce KH, Biddie SC, Sabo PJ, Thurman RE, Stamatoyannopoulos JA, Hager GL: **DNA methylation status predicts cell type-specific enhancer activity.** *EMBO J* 2011, **30**:3028–39.
44. Lindholm ME, Marabita F, Gomez-Cabrero D, Rundqvist H, Ekström TJ, Tegnér J, Sundberg CJ: **An integrative analysis reveals coordinated reprogramming of the epigenome and the transcriptome in human skeletal muscle after training.** *Epigenetics* 2014, **9**:1557–69.
45. Alibegovic AC, Sonne MP, Højbjerg L, Bork-Jensen J, Jacobsen S, Nilsson E, Faerch K, Hiscock N, Mortensen B, Friedrichsen M, Stallknecht B, Dela F, Vaag A: **Insulin resistance induced by physical inactivity is associated with multiple transcriptional changes in skeletal muscle in young men.** *Am J Physiol Endocrinol Metab* 2010, **299**:E752–63.

46. Larrouy D, Barbe P, Valle C, Déjean S, Pelloux V, Thalamas C, Bastard J-P, Le Bouil A, Diquet B, Clément K, Langin D, Viguerie N: **Gene expression profiling of human skeletal muscle in response to stabilized weight loss.** *Am J Clin Nutr* 2008, **88**:125–32.
47. Lammers G, Poelkens F, van Duijnhoven NTL, Pardoel EM, Hoenderop JG, Thijssen DHJ, Hopman MTE: **Expression of genes involved in fatty acid transport and insulin signaling is altered by physical inactivity and exercise training in human skeletal muscle.** *Am J Physiol Endocrinol Metab* 2012, **303**:E1245–51.
48. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray.** *Science* 1995, **270**:467–70.
49. Fasold M, Binder H: **Variation of RNA Quality and Quantity Are Major Sources of Batch Effects in Microarray Expression Data.** *Microarrays* 2014, **3**:322–339.
50. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo M V, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL: **Expression monitoring by hybridization to high-density oligonucleotide arrays.** *Nat Biotechnol* 1996, **14**:1675–80.
51. Knudsen S: *A Biologist's Guide to Analysis of DNA Microarray Data*. New York: John Wiley & Sons; 2002.
52. Fasold M, Binder H: **Estimating RNA-quality using GeneChip microarrays.** *BMC Genomics* 2012, **13**:186.
53. Göhlmann H, Talloen W: *Gene Expression Studies Using Affymetrix Microarrays*. Boca Raton: Chapman & Hall/CRC; 2009.
54. Chomczynski P, Sacchi N: **Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction.** *Anal Biochem* 1987, **162**:156–9.
55. Chomczynski P, Sacchi N: **The single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction: twenty-something years on.** *Nat Protoc* 2006, **1**:581–5.
56. Sykacek P, Kreil DP, Meadows LA, Auburn RP, Fischer B, Russell S, Micklem G: **The impact of quantitative optimization of hybridization conditions on gene expression analysis.** *BMC Bioinformatics* 2011, **12**:73.

57. Koltai H, Weingarten-Baror C: **Specificity of DNA microarray hybridization: characterization, effectors and approaches for data correction.** *Nucleic Acids Res* 2008, **36**:2395–405.
58. Okoniewski MJ, Miller CJ: **Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations.** *BMC Bioinformatics* 2006, **7**:276.
59. Jaksik R, Iwanaszko M, Rzeszowska-Wolny J, Kimmel M: **Microarray experiments and factors which affect their reliability.** *Biol Direct* 2015, **10**:46.
60. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nat Rev Genet* 2009, **10**:57–63.
61. Wu Z: **A review of statistical methods for preprocessing oligonucleotide microarrays.** *Stat Methods Med Res* 2009, **18**:533–41.
62. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.** *Bioinformatics* 2003, **19**:185–93.
63. Li Q, Birkbak NJ, Györfy B, Szallasi Z, Eklund AC: **Jetset: selecting the optimal microarray probe set to represent a gene.** *BMC Bioinformatics* 2011, **12**:474.
64. Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES: **Distinguishing protein-coding and noncoding genes in the human genome.** *Proc Natl Acad Sci U S A* 2007, **104**:19428–33.
65. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489**:57–74.
66. Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, Patapoutian A, Hampton GM, Schultz PG, Hogenesch JB: **Large-scale analysis of the human and mouse transcriptomes.** *Proc Natl Acad Sci U S A* 2002, **99**:4465–70.
67. Jongeneel CV, Iseli C, Stevenson BJ, Riggins GJ, Lal A, Mackay A, Harris RA, O'Hare MJ, Neville AM, Simpson AJG, Strausberg RL: **Comprehensive sampling of gene expression in human cell lines with massively parallel signature sequencing.** *Proc Natl Acad Sci U S A* 2003, **100**:4702–5.
68. Calza S, Raffelsberger W, Ploner A, Sahel J, Leveillard T, Pawitan Y: **Filtering genes to improve sensitivity in oligonucleotide microarray data analysis.** *Nucleic Acids Res* 2007, **35**:e102.

69. Modlich O, Prisack H-B, Munnes M, Audretsch W, Bojar H: **Immediate gene expression changes after the first course of neoadjuvant chemotherapy in patients with primary breast cancer disease.** *Clin Cancer Res* 2004, **10**:6418–31.
70. McClintick JN, Edenberg HJ: **Effects of filtering by Present call on analysis of microarray experiments.** *BMC Bioinformatics* 2006, **7**:49.
71. Liu W, Mei R, Di X, Ryder TB, Hubbell E, Dee S, Webster TA, Harrington CA, Ho M, Baid J, Smeekens SP: **Analysis of high density expression microarrays with signed-rank call algorithms.** *Bioinformatics* 2002, **18**:1593–9.
72. Mieczkowski J, Tyburczy ME, Dabrowski M, Pokarowski P: **Probe set filtering increases correlation between Affymetrix GeneChip and qRT-PCR expression measurements.** *BMC Bioinformatics* 2010, **11**:104.
73. Ringnér M: **What is principal component analysis?** *Nat Biotechnol* 2008, **26**:303–4.
74. Raychaudhuri S, Stuart JM, Altman RB: **Principal components analysis to summarize microarray experiments: application to sporulation time series.** *Pac Symp Biocomput* 2000:455–66.
75. Dopazo J, Zanders E, Dragoni I, Amphlett G, Falciani F: **Methods and approaches in the analysis of gene expression data.** *J Immunol Methods* 2001, **250**:93–112.
76. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA: **Tackling the widespread and critical impact of batch effects in high-throughput data.** *Nat Rev Genet* 2010, **11**:733–9.
77. Tong W, Lucas AB, Shippy R, Fan X, Fang H, Hong H, Orr MS, Chu T-M, Guo X, Collins PJ, Sun YA, Wang S-J, Bao W, Wolfinger RD, Shchegrova S, Guo L, Warrington JA, Shi L: **Evaluation of external RNA controls for the assessment of microarray performance.** *Nat Biotechnol* 2006, **24**:1132–9.
78. Lazar C, Meganck S, Taminau J, Steenhoff D, Coletta A, Molter C, Weiss-Solís DY, Duque R, Bersini H, Nowé A: **Batch effect removal methods for microarray gene expression data integration: a survey.** *Brief Bioinform* 2013, **14**:469–90.
79. Johnson WE, Li C, Rabinovic A: **Adjusting batch effects in microarray expression data using empirical Bayes methods.** *Biostatistics* 2007, **8**:118–27.
80. Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, Liu C: **Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods.** *PLoS One* 2011, **6**:e17238.

81. Kerr MK: **Design considerations for efficient and effective microarray studies.** *Biometrics* 2003, **59**:822–8.
82. Yang H, Harrington CA, Vartanian K, Coldren CD, Hall R, Churchill GA: **Randomization in laboratory procedure is key to obtaining reproducible microarray results.** *PLoS One* 2008, **3**:e3724.
83. Irizarry RA, Wang C, Zhou Y, Speed TP: **Gene set enrichment analysis made simple.** *Stat Methods Med Res* 2009, **18**:565–75.
84. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A* 2005, **102**:15545–50.
85. Joyner MJ, Pedersen BK: **Ten questions about systems biology.** *J Physiol* 2011, **589**(Pt 5):1017–30.
86. Kohl P, Noble D: **Systems biology and the virtual physiological human.** *Mol Syst Biol* 2009, **5**:292.
87. Krogh A: **The number and distribution of capillaries in muscles with calculations of the oxygen pressure head necessary for supplying the tissue.** *J Physiol* 1919, **52**:409–15.
88. HUXLEY AF: **Muscle structure and theories of contraction.** *Prog Biophys Biophys Chem* 1957, **7**:255–318.
89. Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau W-C, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, Edgar R: **NCBI GEO: mining millions of expression profiles--database and tools.** *Nucleic Acids Res* 2005, **33**(Database issue):D562–6.
90. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, Oezcimen A, Rocca-Serra P, Sansone S-A: **ArrayExpress--a public repository for microarray gene expression data at the EBI.** *Nucleic Acids Res* 2003, **31**:68–71.
91. Jones P, Côté RG, Martens L, Quinn AF, Taylor CF, Derache W, Hermjakob H, Apweiler R: **PRIDE: a public repository of protein and peptide identifications for the proteomics community.** *Nucleic Acids Res* 2006, **34**(Database issue):D659–63.
92. Holloway K V, O’Gorman M, Woods P, Morton JP, Evans L, Cable NT, Goldspink DF, Burniston JG: **Proteomic investigation of changes in human vastus lateralis muscle in response to interval-exercise training.** *Proteomics* 2009, **9**:5155–74.

93. Kupersmidt I, Su QJ, Grewal A, Sundaresh S, Halperin I, Flynn J, Shekar M, Wang H, Park J, Cui W, Wall GD, Wisotzkey R, Alag S, Akhtari S, Ronaghi M: **Ontology-based meta-analysis of global collections of high-throughput public data.** *PLoS One* 2010, **5**:e13066.
94. Mah N: **A comparison of oligonucleotide and cDNA-based microarray systems.** *Physiol Genomics* 2004, **16**:361–370.
95. Ortega F, Sameith K, Turan N, Compton R, Trevino V, Vannucci M, Falciani F: **Models and computational strategies linking physiological response to molecular networks from large-scale data.** *Philos Trans A Math Phys Eng Sci* 2008, **366**:3067–89.
96. Gavaghan D, Garny A, Maini PK, Kohl P: **Mathematical models in physiology.** *Philos Trans A Math Phys Eng Sci* 2006, **364**:1099–106.
97. Gomez-Cabrero D, Compte A, Tegner J: **Workflow for generating competing hypothesis from models with parameter uncertainty.** *Interface Focus* 2011, **1**:438–49.
98. Noble D: **Computational models of the heart and their use in assessing the actions of drugs.** *J Pharmacol Sci* 2008, **107**:107–17.
99. Trevino V, Falciani F: **GALGO: an R package for multivariate variable selection using genetic algorithms.** *Bioinformatics* 2006, **22**:1154–6.
100. Bansal M, Della Gatta G, di Bernardo D: **Inference of gene regulatory networks and compound mode of action from time course gene expression profiles.** *Bioinformatics* 2006, **22**:815–22.
101. Yu J, Smith VA, Wang PP, Hartemink AJ, Jarvis ED: **Advances to Bayesian network inference for generating causal networks from observational biological data.** *Bioinformatics* 2004, **20**:3594–603.
102. Neapolitan RE: *Learning Bayesian Networks.* New Jersey: Pearson Prentice Hall; 2004.
103. Hirose O, Yoshida R, Imoto S, Yamaguchi R, Higuchi T, Charnock-Jones DS, Print C, Miyano S: **Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models.** *Bioinformatics* 2008, **24**:932–42.
104. De Smet R, Marchal K: **Advantages and limitations of current network inference methods.** *Nat Rev Microbiol* 2010, **8**:717–29.

105. Perrin B-E, Ralaivola L, Mazurie A, Bottani S, Mallet J, d'Alché-Buc F: **Gene networks inference using dynamic Bayesian networks**. *Bioinformatics* 2003, **19 Suppl 2**:ii138–48.
106. Opgen-Rhein R, Strimmer K: **Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process**. *BMC Bioinformatics* 2007, **8 Suppl 2**:S3.
107. Falati S, Gross P, Merrill-Skoloff G, Furie BC, Furie B: **Real-time in vivo imaging of platelets, tissue factor and fibrin during arterial thrombus formation in the mouse**. *Nat Med* 2002, **8**:1175–1180.
108. Butte AJ, Kohane IS: **Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements**. *Pac Symp Biocomput* 2000:418–29.
109. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A: **ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context**. *BMC Bioinformatics* 2006, **7 Suppl 1**:S7.
110. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS: **Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles**. *PLoS Biol* 2007, **5**:e8.
111. Meyer PE, Kontos K, Lafitte F, Bontempi G: **Information-theoretic inference of large transcriptional regulatory networks**. *EURASIP J Bioinform Syst Biol* 2007:79879.
112. 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**:2498–504.
113. Batagelj V, Mrvar A: **Pajek - Program for large network analysis**. *Connections* 1998, **21**:47–57.
114. Goldovsky L, Cases I, Enright AJ, Ouzounis CA: **BioLayout(Java): versatile network visualisation of structural and functional relationships**. *Appl Bioinformatics* 2005, **4**:71–4.
115. Gehlenborg N, O'Donoghue SI, Baliga NS, Goesmann A, Hibbs MA, Kitano H, Kohlbacher O, Neuweger H, Schneider R, Tenenbaum D, Gavin A-C: **Visualization of omics data for systems biology**. *Nat Methods* 2010, **7**(3 Suppl):S56–68.

116. Poultney CS, Greenfield A, Bonneau R: **Integrated inference and analysis of regulatory networks from multi-level measurements.** *Methods Cell Biol* 2012, **110**:19–56.
117. Gupta R, Stincone A, Antczak P, Durant S, Bicknell R, Bikfalvi A, Falciani F: **A computational framework for gene regulatory network inference that combines multiple methods and datasets.** *BMC Syst Biol* 2011, **5**:52.
118. Cassese A, Guindani M, Tadesse MG, Falciani F, Vannucci M: **A HIERARCHICAL BAYESIAN MODEL FOR INFERENCE OF COPY NUMBER VARIANTS AND THEIR ASSOCIATION TO GENE EXPRESSION.** *Ann Appl Stat* 2014, **8**:148–175.
119. Merico D, Gfeller D, Bader GD: **How to visually interpret biological data using networks.** *Nat Biotechnol* 2009, **27**:921–924.
120. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.** *Nat Genet* 2003, **34**:166–76.
121. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL: **Hierarchical organization of modularity in metabolic networks.** *Science* 2002, **297**:1551–5.
122. Barabasi A, Albert R: **Emergence of scaling in random networks.** *Science* 1999, **286**:509–12.
123. Alderson D, Doyle JC, Li L, Willinger W: **Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications.** *Internet Math* 2005, **2**:431–523.
124. Su G, Kuchinsky A, Morris JH, States DJ, Meng F: **GLay: community structure analysis of biological networks.** *Bioinformatics* 2010, **26**:3135–7.
125. Ideker T, Ozier O, Schwikowski B, Siegel AF: **Discovering regulatory and signalling circuits in molecular interaction networks.** *Bioinformatics* 2002, **18 Suppl 1**:S233–40.
126. Huang DW, Sherman BT, Lempicki RA: **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.** *Nucleic Acids Res* 2009, **37**:1–13.
127. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**:25–9.

128. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M: **KEGG: Kyoto Encyclopedia of Genes and Genomes**. *Nucleic Acids Res* 1999, **27**:29–34.
129. Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery**. *Genome Biol* 2003, **4**:P3.
130. Maere S, Heymans K, Kuiper M: **BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks**. *Bioinformatics* 2005, **21**:3448–9.
131. Davidsen PK, Herbert JM, Antczak P, Clarke K, Ferrer E, Peinado VI, Gonzalez C, Roca J, Egginton S, Falciani F: **A systems biology approach reveals a link between systemic cytokines and skeletal muscle energy metabolism in a rodent smoking model and human COPD**. *Genome Med* 2014.
132. Rabinovich RA, Bastos R, Ardite E, Llinàs L, Orozco-Levi M, Gea J, Vilaró J, Barberà JA, Rodríguez-Roisin R, Fernández-Checa JC, Roca J: **Mitochondrial dysfunction in COPD patients with low body mass index**. *Eur Respir J* 2007, **29**:643–50.
133. Turan N, Kalko S, Stincone A, Clarke K, Sabah A, Howlett K, Curnow SJ, Rodriguez DA, Cascante M, O'Neill L, Egginton S, Roca J, Falciani F: **A systems biology approach identifies molecular networks defining skeletal muscle abnormalities in chronic obstructive pulmonary disease**. *PLoS Comput Biol* 2011, **7**:e1002129.
134. Finkel T, Deng C-X, Mostoslavsky R: **Recent progress in the biology and physiology of sirtuins**. *Nature* 2009, **460**:587–91.
135. Jing E, Emanuelli B, Hirschey MD, Boucher J, Lee KY, Lombard D, Verdin EM, Kahn CR: **Sirtuin-3 (Sirt3) regulates skeletal muscle metabolism and insulin signaling via altered mitochondrial oxidation and reactive oxygen species production**. *Proc Natl Acad Sci U S A* 2011, **108**:14608–13.
136. He W, Newman JC, Wang MZ, Ho L, Verdin E: **Mitochondrial sirtuins: regulators of protein acylation and metabolism**. *Trends Endocrinol Metab* 2012, **23**:467–76.
137. Suzuki YJ, Carini M, Butterfield DA: **Protein carbonylation**. *Antioxid Redox Signal* 2010, **12**:323–5.
138. Gonzalez NC, Wood JG: **Alveolar hypoxia-induced systemic inflammation: what low PO<sub>2</sub> does and does not do**. *Adv Exp Med Biol* 2010, **662**:27–32.

139. Reinke C, Bevans-Fonti S, Drager LF, Shin M-K, Polotsky VY: **Effects of different acute hypoxic regimens on tissue oxygen profiles and metabolic outcomes.** *J Appl Physiol* 2011, **111**:881–90.
140. Willmann G: **Transcriptional Regulation after Chronic Hypoxia Exposure in Skeletal Muscle.** University of Cologne; 2013.
141. Rangel C, Angus J, Ghahramani Z, Lioumi M, Sotheran E, Gaiba A, Wild DL, Falciani F: **Modeling T-cell activation using gene expression profiling and state-space models.** *Bioinformatics* 2004, **20**:1361–1372.
142. Van der Laan MJ, Pollard KS: **Hybrid clustering of gene expression data with visualization and the bootstrap.** *J Stat Plan Inference* 2003, **117**:275–303.
143. Lopez AD, Murray CC: **The global burden of disease, 1990-2020.** *Nat Med* 1998, **4**:1241–3.
144. Fagerström K: **The epidemiology of smoking: health consequences and benefits of cessation.** *Drugs* 2002, **62 Suppl 2**:1–9.
145. Aliverti A, Macklem PT: **How and why exercise is impaired in COPD.** *Respiration* 2001, **68**:229–39.
146. Marquis K, Debigaré R, Lacasse Y, LeBlanc P, Jobin J, Carrier G, Maltais F: **Midthigh muscle cross-sectional area is a better predictor of mortality than body mass index in patients with chronic obstructive pulmonary disease.** *Am J Respir Crit Care Med* 2002, **166**:809–13.
147. Barreiro E, Peinado VI, Galdiz JB, Ferrer E, Marin-Corral J, Sánchez F, Gea J, Barberà JA: **Cigarette smoke-induced oxidative stress: A role in chronic obstructive pulmonary disease skeletal muscle dysfunction.** *Am J Respir Crit Care Med* 2010, **182**:477–88.
148. Decramer M, De Benedetto F, Del Ponte A, Marinari S: **Systemic effects of COPD.** *Respir Med* 2005, **99 Suppl B**:S3–10.
149. Gea J, Pascual S, Casadevall C, Orozco-Levi M, Barreiro E: **Muscle dysfunction in COPD: update on causes and biological findings.** *Journal of Thoracic Disease* 2015.
150. Gosker HR, van Mameren H, van Dijk PJ, Engelen MPKJ, van der Vusse GJ, Wouters EFM, Schols AMWJ: **Skeletal muscle fibre-type shifting and metabolic profile in patients with chronic obstructive pulmonary disease.** *Eur Respir J Off J Eur Soc Clin Respir Physiol* 2002, **19**:617–25.

151. Takabatake N, Nakamura H, Abe S, Inoue S, Hino T, Saito H, Yuki H, Kato S, Tomoike H: **The relationship between chronic hypoxemia and activation of the tumor necrosis factor-alpha system in patients with chronic obstructive pulmonary disease.** *Am J Respir Crit Care Med* 2000, **161**(4 Pt 1):1179–84.
152. Wright JL, Cosio M, Churg A: **Animal models of chronic obstructive pulmonary disease.** *Am J Physiol Lung Cell Mol Physiol* 2008, **295**:L1–15.
153. Gosker HR, Langen RCJ, Bracke KR, Joos GF, Brusselle GG, Steele C, Ward KA, Wouters EFM, Schols AMWJ: **Extrapulmonary manifestations of chronic obstructive pulmonary disease in a mouse model of chronic cigarette smoke exposure.** *Am J Respir Cell Mol Biol* 2009, **40**:710–6.
154. Rinaldi M, Maes K, De Vleeschauwer S, Thomas D, Verbeken EK, Decramer M, Janssens W, Gayan-Ramirez GN: **Long-term nose-only cigarette smoke exposure induces emphysema and mild skeletal muscle dysfunction in mice.** *Dis Model Mech* 2012, **5**:333–41.
155. Caron M-A, Morissette MC, Thériault M-E, Nikota JK, Stämpfli MR, Debigaré R: **Alterations in skeletal muscle cell homeostasis in a mouse model of cigarette smoke exposure.** *PLoS One* 2013, **8**:e66433.
156. Padilla-Carlin DJ, McMurray DN, Hickey AJ: **The guinea pig as a model of infectious diseases.** *Comp Med* 2008, **58**:324–40.
157. Olea E, Ferrer E, Prieto-Lloret J, Gonzalez-Martin C, Vega-Agapito V, Gonzalez-Obeso E, Agapito T, Peinado V, Obeso A, Barbera JA, Gonzalez C: **Effects of cigarette smoke and chronic hypoxia on airways remodeling and resistance. Clinical significance.** *Respir Physiol Neurobiol* 2011, **179**:305–13.
158. Ardite E, Peinado VI, Rabinovich RA, Fernández-Checa JC, Roca J, Barberà JA: **Systemic effects of cigarette smoke exposure in the guinea pig.** *Respir Med* 2006, **100**:1186–94.
159. Ning W, Li C-J, Kaminski N, Feghali-Bostwick CA, Alber SM, Di YP, Otterbein SL, Song R, Hayashi S, Zhou Z, Pinsky DJ, Watkins SC, Pilewski JM, Sciruba FC, Peters DG, Hogg JC, Choi AMK: **Comprehensive gene expression profiles reveal pathways related to the pathogenesis of chronic obstructive pulmonary disease.** *Proc Natl Acad Sci U S A* 2004, **101**:14895–900.
160. Bhattacharya S, Mariani TJ: **Array of hope: expression profiling identifies disease biomarkers and mechanism.** *Biochem Soc Trans* 2009, **37**(Pt 4):855–62.
161. Ferrer E, Peinado VI, Castañeda J, Prieto-Lloret J, Olea E, González-Martín MC, Vega-Agapito M V, Díez M, Domínguez-Fandos D, Obeso A, González C, Barberà JA:

**Effects of cigarette smoke and hypoxia on pulmonary circulation in the guinea pig.** *Eur Respir J* 2011, **38**:617–27.

162. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq.** *Bioinformatics* 2009, **25**:1105–11.

163. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nat Biotechnol* 2010, **28**:511–5.

164. Yang YH, Paquet A, Dudoit S: **marray: Exploratory analysis for two-color spotted microarray data.** 2009.

165. Smyth GK: **Limma: linear models for microarray data.** In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer; 2005:397–420.

166. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Natl Acad Sci U S A* 2001, **98**:5116–21.

167. Huang DW, Sherman BT, Tan Q, Kir J, Liu D, Bryant D, Guo Y, Stephens R, Baseler MW, Lane HC, Lempicki RA: **DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists.** *Nucleic Acids Res* 2007, **35**(Web Server issue):W169–75.

168. Livak KJ, Schmittgen TD: **Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method.** *Methods* 2001, **25**:402–8.

169. Wang R, Ahmed J, Wang G, Hassan I, Strulovici-Barel Y, Hackett NR, Crystal RG: **Down-regulation of the canonical Wnt  $\beta$ -catenin pathway in the airway epithelium of healthy smokers and smokers with COPD.** *PLoS One* 2011, **6**:e14793.

170. Sameith K, Antczak P, Marston E, Turan N, Maier D, Stankovic T, Falciani F: **Functional modules integrating essential cellular functions are predictive of the response of leukaemia cells to DNA damage.** *Bioinformatics* 2008, **24**:2602–7.

171. Chu L-H, Rivera CG, Popel AS, Bader JS: **Constructing the Angiome - a Global Angiogenesis Protein Interaction Network.** *Physiol Genomics* 2012.

172. Hastie T, Tibshirani R, Narasimhan B, Gilbert C: **impute: Imputation for microarray data.** .

173. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP: **Molecular signatures database (MSigDB) 3.0.** *Bioinformatics* 2011, **27**:1739–40.
174. Piehl-Aulin K, Jones I, Lindvall B, Magnuson A, Abdel-Halim SM: **Increased serum inflammatory markers in the absence of clinical and skeletal muscle inflammation in patients with chronic obstructive pulmonary disease.** *Respiration* 2009, **78**:191–6.
175. Nussbaumer-Ochsner Y, Rabe KF: **Systemic manifestations of COPD.** *Chest* 2011, **139**:165–73.
176. Debigaré R, Marquis K, Côté CH, Tremblay RR, Michaud A, LeBlanc P, Maltais F: **Catabolic/anabolic balance and muscle wasting in patients with COPD.** *Chest* 2003, **124**:83–9.
177. Pinto-Plata V, Toso J, Lee K, Park D, Bilello J, Mullerova H, De Souza MM, Vessey R, Celli B: **Profiling serum biomarkers in patients with COPD: associations with clinical parameters.** *Thorax* 2007, **62**:595–601.
178. Sala E, Roca J, Marrades RM, Alonso J, Gonzalez De Suso JM, Moreno A, Barberá JA, Nadal J, de Jover L, Rodriguez-Roisin R, Wagner PD: **Effects of endurance training on skeletal muscle bioenergetics in chronic obstructive pulmonary disease.** *Am J Respir Crit Care Med* 1999, **159**:1726–34.
179. Spira A, Beane J, Shah V, Liu G, Schembri F, Yang X, Palma J, Brody JS: **Effects of cigarette smoke on the human airway epithelial cell transcriptome.** *Proc Natl Acad Sci U S A* 2004, **101**:10143–8.
180. Beane J, Sebastiani P, Liu G, Brody JS, Lenburg ME, Spira A: **Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression.** *Genome Biol* 2007, **8**:R201.
181. Lemire BB, Debigaré R, Dubé A, Thériault M-E, Côté CH, Maltais F: **MAPK signaling in the quadriceps of patients with chronic obstructive pulmonary disease.** *J Appl Physiol* 2012, **113**:159–66.
182. Wüst RCI, Degens H: **Factors contributing to muscle wasting and dysfunction in COPD patients.** *Int J Chron Obstruct Pulmon Dis* 2007, **2**:289–300.
183. Stevenson CS, Birrell MA: **Moving towards a new generation of animal models for asthma and COPD with improved clinical relevance.** *Pharmacol Ther* 2011, **130**:93–105.

184. Knowler WC, Barrett-Connor E, Fowler SE, Hamman RF, Lachin JM, Walker EA, Nathan DM: **Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin.** *N Engl J Med* 2002, **346**:393–403.
185. Tuomilehto J, Lindström J, Eriksson JG, Valle TT, Hämäläinen H, Ilanne-Parikka P, Keinänen-Kiukaanniemi S, Laakso M, Louheranta A, Rastas M, Salminen V, Uusitupa M: **Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance.** *N Engl J Med* 2001, **344**:1343–50.
186. Koval JA, Maezono K, Patti ME, Pendergrass M, DeFronzo RA, Mandarino LJ: **Effects of exercise and insulin on insulin signaling proteins in human skeletal muscle.** *Med Sci Sports Exerc* 1999, **31**:998–1004.
187. Ferrannini E, Simonson DC, Katz LD, Reichard G, Bevilacqua S, Barrett EJ, Olsson M, DeFronzo RA: **The disposal of an oral glucose load in patients with non-insulin-dependent diabetes.** *Metabolism* 1988, **37**:79–85.
188. Teran-Garcia M, Rankinen T, Koza RA, Rao DC, Bouchard C: **Endurance training-induced changes in insulin sensitivity and gene expression.** *Am J Physiol Endocrinol Metab* 2005, **288**:E1168–78.
189. Huffman KM, Koves TR, Hubal MJ, Abouassi H, Beri N, Bateman LA, Stevens RD, Ilkayeva OR, Hoffman EP, Muoio DM, Kraus WE: **Metabolite signatures of exercise training in human skeletal muscle relate to mitochondrial remodelling and cardiometabolic fitness.** *Diabetologia* 2014, **57**:2282–95.
190. Bouchard C: **Genomic predictors of trainability.** *Exp Physiol* 2012, **97**:347–52.
191. Bouchard C, Leon AS, Rao DC, Skinner JS, Wilmore JH, Gagnon J: **The HERITAGE family study. Aims, design, and measurement protocol.** *Med Sci Sports Exerc* 1995, **27**:721–9.
192. Walton C, Godsland IF, Proudler AJ, Felton C, Wynn V: **Evaluation of four mathematical models of glucose and insulin dynamics with analysis of effects of age and obesity.** *Am J Physiol* 1992, **262**(5 Pt 1):E755–62.
193. Boston RC, Stefanovski D, Moate PJ, Sumner AE, Watanabe RM, Bergman RN: **MINMOD Millennium: a computer program to calculate glucose effectiveness and insulin sensitivity from the frequently sampled intravenous glucose tolerance test.** *Diabetes Technol Ther* 2003, **5**:1003–15.
194. Bouchard C, Sarzynski MA, Rice TK, Kraus WE, Church TS, Sung YJ, Rao DC, Rankinen T: **Genomic predictors of the maximal O<sub>2</sub> uptake response to standardized exercise training programs.** *J Appl Physiol* 2011, **110**:1160–70.

195. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81**:559–75.
196. Phillips BE, Williams JP, Gustafsson T, Bouchard C, Rankinen T, Knudsen S, Smith K, Timmons JA, Atherton PJ: **Molecular networks of human muscle adaptation to exercise and age.** *PLoS Genet* 2013, **9**:e1003389.
197. Edgar R: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic Acids Res* 2002, **30**:207–210.
198. Gautier L, Cope L, Bolstad BM, Irizarry RA: **affy--analysis of Affymetrix GeneChip data at the probe level.** *Bioinformatics* 2004, **20**:307–15.
199. Meex RCR, Schrauwen-Hinderling VB, Moonen-Kornips E, Schaart G, Mensink M, Phielix E, van de Weijer T, Sels J-P, Schrauwen P, Hesselink MKC: **Restoration of muscle mitochondrial function and metabolic flexibility in type 2 diabetes by exercise training is paralleled by increased myocellular fat storage and improved insulin sensitivity.** *Diabetes* 2010, **59**:572–9.
200. Mensink M, Hesselink MKC, Russell AP, Schaart G, Sels J-P, Schrauwen P: **Improved skeletal muscle oxidative enzyme activity and restoration of PGC-1 alpha and PPAR beta/delta gene expression upon rosiglitazone treatment in obese patients with type 2 diabetes mellitus.** *Int J Obes (Lond)* 2007, **31**:1302–10.
201. Carvalho BS, Irizarry RA: **A framework for oligonucleotide microarray preprocessing.** *Bioinformatics* 2010, **26**:2363–7.
202. Ghosh S, Vivar JC, Sarzynski MA, Sung YJ, Timmons JA, Bouchard C, Rankinen T: **Integrative pathway analysis of a genome-wide association study of (V)O(2max) response to exercise training.** *J Appl Physiol* 2013, **115**:1343–59.
203. Chai H-S, Sicotte H, Bailey KR, Turner ST, Asmann YW, Kocher J-PA: **GLOSSI: a method to assess the association of genetic loci-sets with complex diseases.** *BMC Bioinformatics* 2009, **10**:102.
204. Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, Kasarskis A, Zhang B, Wang S, Suver C, Zhu J, Millstein J, Sieberts S, Lamb J, GuhaThakurta D, Derry J, Storey JD, Avila-Campillo I, Kruger MJ, Johnson JM, Rohl CA, van Nas A, Mehrabian M, Drake TA, Lusk AJ, Smith RC, Guengerich FP, Strom SC, Schuetz E, Rushmore TH, et al.: **Mapping the genetic architecture of gene expression in human liver.** *PLoS Biol* 2008, **6**:e107.
205. McKinsey TA, Zhang CL, Olson EN: **MEF2: a calcium-dependent regulator of cell division, differentiation and death.** *Trends Biochem Sci* 2002, **27**:40–7.

206. Wu H, Rothermel B, Kanatous S, Rosenberg P, Naya FJ, Shelton JM, Hutcheson KA, DiMaio JM, Olson EN, Bassel-Duby R, Williams RS: **Activation of MEF2 by muscle activity is mediated through a calcineurin-dependent pathway.** *EMBO J* 2001, **20**:6414–23.
207. Wales S, Hashemi S, Blais A, McDermott JC: **Global MEF2 target gene analysis in cardiac and skeletal muscle reveals novel regulation of DUSP6 by p38MAPK-MEF2 signaling.** *Nucleic Acids Res* 2014, **42**:11349–62.
208. Estrella NL, Desjardins CA, Nocco SE, Clark AL, Maksimenko Y, Naya FJ: **MEF2 transcription factors regulate distinct gene programs in mammalian skeletal muscle differentiation.** *J Biol Chem* 2015, **290**:1256–68.
209. Blake JA, Eppig JT, Richardson JE, Davisson MT: **The Mouse Genome Database (MGD): expanding genetic and genomic resources for the laboratory mouse. The Mouse Genome Database Group.** *Nucleic Acids Res* 2000, **28**:108–11.
210. Von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P: **STRING: known and predicted protein-protein associations, integrated and transferred across organisms.** *Nucleic Acids Res* 2005, **33**(Database issue):D433–7.
211. Davidsen PK, Gallagher IJ, Hartman JW, Tarnopolsky MA, Dela F, Helge JW, Timmons JA, Phillips SM, Jw H, Ja T, Sm P: **High responders to resistance exercise training demonstrate differential regulation of skeletal muscle microRNA expression.** 2011:309–317.
212. Sood S, Gallagher IJ, Lunnon K, Rullman E, Keohane A, Crossland H, Phillips BE, Cederholm T, Jensen T, van Loon LJ, Lannfelt L, Kraus WE, Atherton PJ, Howard R, Gustafsson T, Hodges A, Timmons JA: **A novel multi-tissue RNA diagnostic of healthy ageing relates to cognitive health status.** *Genome Biol* 2015, **16**:185.
213. Bouchard C, Rankinen T, Timmons JA: **Genomics and genetics in the biology of adaptation to exercise.** *Compr Physiol* 2011, **1**:1603–48.
214. Houmard JA: **Effect of the volume and intensity of exercise training on insulin sensitivity.** *J Appl Physiol* 2003, **96**:101–106.
215. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Girón CG, Gordon L, Hourlier T, Hunt SE, Janacek SH, Johnson N, Juettemann T, Kähäri AK, Keenan S, Martin FJ, Maurel T, McLaren W, Murphy DN, Nag R, Overduin B, Parker A, Patricio M, Perry E, Pignatelli M, et al.: **Ensembl 2015.** *Nucleic Acids Res* 2014, **43**(Database issue):D662–9.

216. Mu X, Brown LD, Liu Y, Schneider MF: **Roles of the calcineurin and CaMK signaling pathways in fast-to-slow fiber type transformation of cultured adult mouse skeletal muscle fibers.** *Physiol Genomics* 2007, **30**:300–12.
217. Hambrecht R, Fiehn E, Yu J, Niebauer J, Weigl C, Hilbrich L, Adams V, Riede U, Schuler G: **Effects of Endurance Training on Mitochondrial Ultrastructure and Fiber Type Distribution in Skeletal Muscle of Patients With Stable Chronic Heart Failure.** *J Am Coll Cardiol* 1997, **29**:1067–1073.
218. Russell AP, Feilchenfeldt J, Schreiber S, Praz M, Crettenand A, Gobelet C, Meier CA, Bell DR, Kralli A, Giacobino J-P, Deriaz O: **Endurance Training in Humans Leads to Fiber Type-Specific Increases in Levels of Peroxisome Proliferator-Activated Receptor- Coactivator-1 and Peroxisome Proliferator-Activated Receptor- in Skeletal Muscle.** *Diabetes* 2003, **52**:2874–2881.
219. Rico-Sanz J, Rankinen T, Joanisse DR, Leon AS, Skinner JS, Wilmore JH, Rao DC, Bouchard C: **Familial resemblance for muscle phenotypes in the HERITAGE Family Study.** *Med Sci Sports Exerc* 2003, **35**:1360–6.
220. Simoneau JA, Bouchard C: **Human variation in skeletal muscle fiber-type proportion and enzyme activities.** *Am J Physiol* 1989, **257**(4 Pt 1):E567–72.
221. Goodyear LJ, Kahn BB: **Exercise, glucose transport, and insulin sensitivity.** *Annu Rev Med* 1998, **49**:235–61.
222. Pate RR, Pratt M, Blair SN, Haskell WL, Macera CA, Bouchard C, Buchner D, Ettinger W, Heath GW, King AC: **Physical activity and public health. A recommendation from the Centers for Disease Control and Prevention and the American College of Sports Medicine.** *JAMA* 1995, **273**:402–7.
223. Tobina T, Nakashima H, Mori S, Abe M, Kumahara H, Yoshimura E, Nishida Y, Kiyonaga A, Shono N, Tanaka H: **The Utilization of a Biopsy Needle to Obtain Small Muscle Tissue Specimens to Analyze the Gene and Protein Expression.** *J Surg Res* 2009, **154**:252–257.
224. Hayot M: **Skeletal muscle microbiopsy: a validation study of a minimally invasive technique.** *Eur Respir J* 2005, **25**:431–440.
225. Zierath JR, Wallberg-Henriksson H: **Looking Ahead Perspective: Where Will the Future of Exercise Biology Take Us?** *Cell Metab* 2015, **22**:25–30.
226. Sarzynski MA, Davidsen PK, Sung YJ, Hesselink MKC, Schrauwen P, Rice TK, Rao DC, Falciari F, Bouchard C: **Genomic and transcriptomic predictors of triglyceride response to regular exercise.** *Br J Sports Med* 2015.

227. Pejic RN, Lee DT: **Hypertriglyceridemia.** *J Am Board Fam Med* 2006, **19**:310–316.
228. Nordestgaard BG, Varbo A: **Triglycerides and cardiovascular disease.** *Lancet* 2014, **384**:626–635.
229. Hokanson JE, Austin MA: **Plasma triglyceride level is a risk factor for cardiovascular disease independent of high-density lipoprotein cholesterol level: a meta-analysis of population-based prospective studies.** *J Cardiovasc Risk* 1996, **3**:213–9.
230. Sarwar N, Danesh J, Eiriksdottir G, Sigurdsson G, Wareham N, Bingham S, Boekholdt SM, Khaw K-T, Gudnason V: **Triglycerides and the risk of coronary heart disease: 10,158 incident cases among 262,525 participants in 29 Western prospective studies.** *Circulation* 2007, **115**:450–8.
231. Cutler JA, Thom TJ, Roccella E: **Leading causes of death in the United States.** *JAMA* 2006, **295**:383–4; author reply 384.
232. Jørgensen AB, Frikke-Schmidt R, West AS, Grande P, Nordestgaard BG, Tybjærg-Hansen A: **Genetically elevated non-fasting triglycerides and calculated remnant cholesterol as causal risk factors for myocardial infarction.** *Eur Heart J* 2013, **34**:1826–33.
233. Thomsen M, Varbo A, Tybjærg-Hansen A, Nordestgaard BG: **Low nonfasting triglycerides and reduced all-cause mortality: a mendelian randomization study.** *Clin Chem* 2014, **60**:737–46.
234. Berglund L, Brunzell JD, Goldberg AC, Goldberg IJ, Sacks F, Murad MH, Stalenhoef AFH: **Evaluation and treatment of hypertriglyceridemia: an Endocrine Society clinical practice guideline.** *J Clin Endocrinol Metab* 2012, **97**:2969–89.
235. Durstine JL, Grandjean PW, Davis PG, Ferguson MA, Alderson NL, DuBose KD: **Blood lipid and lipoprotein adaptations to exercise: a quantitative analysis.** *Sports Med* 2001, **31**:1033–62.
236. Rice T, Després J-P, Pérusse L, Hong Y, Province MA, Bergeron J, Gagnon J, Leon AS, Skinner JS, Wilmore JH, Bouchard C, Rao DC: **Familial aggregation of blood lipid response to exercise training in the health, risk factors, exercise training, and genetics (HERITAGE) Family Study.** *Circulation* 2002, **105**:1904–8.
237. Bray MS, Hagberg JM, Pérusse L, Rankinen T, Roth SM, Wolfarth B, Bouchard C: **The human gene map for performance and health-related fitness phenotypes: the 2006-2007 update.** *Med Sci Sports Exerc* 2009, **41**:35–73.

238. Huggins GS, Papandonatos GD, Erar B, Belalcazar LM, Brautbar A, Ballantyne C, Kitabchi AE, Wagenknecht LE, Knowler WC, Pownall HJ, Wing RR, Peter I, McCaffery JM: **Do genetic modifiers of high-density lipoprotein cholesterol and triglyceride levels also modify their response to a lifestyle intervention in the setting of obesity and type-2 diabetes mellitus?: The Action for Health in Diabetes (Look AHEAD) study.** *Circ Cardiovasc Genet* 2013, **6**:391–9.
239. Zhang L: *Glycosaminoglycans in Development, Health and Disease*. San Diego: Academic Press; 2010.
240. Schnaar RL, Suzuki A, Stanley P: **Glycosphingolipids.** In *Essentials of Glycobiology. 2nd edition*. New York: Cold Spring Harbor; 2009:129–143.
241. Jensen ON: **Interpreting the protein language using proteomics.** *Nat Rev Mol Cell Biol* 2006, **7**:391–403.
242. Betzen C, Alhamdani MSS, Lueong S, Schröder C, Stang A, Hoheisel JD: **Clinical proteomics: promises, challenges and limitations of affinity arrays.** *Proteomics Clin Appl* 2015, **9**:342–7.
243. Korenberg MJ, Farinha P, Gascoyne RD: **Predicting survival in follicular lymphoma using tissue microarrays.** *Methods Mol Biol* 2007, **377**:255–68.
244. Albain KS, Barlow WE, Shak S, Hortobagyi GN, Livingston RB, Yeh I-T, Ravdin P, Bugarini R, Baehner FL, Davidson NE, Sledge GW, Winer EP, Hudis C, Ingle JN, Perez EA, Pritchard KI, Shepherd L, Gralow JR, Yoshizawa C, Allred DC, Osborne CK, Hayes DF: **Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: a retrospective analysis of a randomised trial.** *Lancet Oncol* 2010, **11**:55–65.
245. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T, Hiller W, Fisher ER, Wickerham DL, Bryant J, Wolmark N: **A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer.** *N Engl J Med* 2004, **351**:2817–26.
246. Dowsett M, Sestak I, Lopez-Knowles E, Sidhu K, Dunbier AK, Cowens JW, Ferree S, Storhoff J, Schaper C, Cuzick J: **Comparison of PAM50 risk of recurrence score with oncotype DX and IHC4 for predicting risk of distant recurrence after endocrine therapy.** *J Clin Oncol* 2013, **31**:2783–90.
247. Fukino K: **Combined Total Genome Loss of Heterozygosity Scan of Breast Cancer Stroma and Epithelium Reveals Multiplicity of Stromal Targets.** *Cancer Res* 2004, **64**:7231–7236.

248. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D: **Methods of integrating data to uncover genotype–phenotype interactions**. *Nat Rev Genet* 2015, **16**:85–97.
249. Krützfeldt J, Rajewsky N, Braich R, Rajeev KG, Tuschl T, Manoharan M, Stoffel M: **Silencing of microRNAs in vivo with “antagomirs”**. *Nature* 2005, **438**:685–9.
250. Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, Johnson JM: **Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs**. *Nature* 2005, **433**:769–73.
251. Gallagher IJ, Scheele C, Keller P, Nielsen AR, Remenyi J, Fischer CP, Roder K, Babraj J, Wahlestedt C, Hutvagner G, Pedersen BK, Timmons JA: **Integration of microRNA changes in vivo identifies novel molecular features of muscle insulin resistance in type 2 diabetes**. *Genome Med* 2010, **2**:9.
252. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biol* 2009, **10**:R25.
253. **SMALT** [<http://www.sanger.ac.uk/resources/software/smalt/>]
254. Baxevanis AD: **Searching the NCBI databases using Entrez**. *Curr Protoc Hum Genet* 2006, **Chapter 6**:Unit 6.10.
255. Huang X, Madan A: **CAP3: A DNA sequence assembly program**. *Genome Res* 1999, **9**:868–77.
256. **Ensembl 2013** [[http://www.ensembl.org/Cavia\\_porcellus/Info/Index](http://www.ensembl.org/Cavia_porcellus/Info/Index)]
257. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA**. *J Mol Biol* 1997, **268**:78–94.
258. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**:403–10.
259. Pruitt KD, Tatusova T, Brown GR, Maglott DR: **NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy**. *Nucleic Acids Res* 2012, **40**(Database issue):D130–5.
260. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC**. *Genome Res* 2002, **12**:996–1006.
261. Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, Raney BJ, Pohl A, Malladi VS, Li CH, Lee BT, Learned K, Kirkup V, Hsu F, Heitner S, Harte RA, Haussler M, Guruvadoo L,

Goldman M, Giardine BM, Fujita PA, Dreszer TR, Diekhans M, Cline MS, Clawson H, Barber GP, et al.: **The UCSC Genome Browser database: extensions and updates 2013.** *Nucleic Acids Res* 2013, **41**(Database issue):D64–9.

262. Roberts A, Pachter L: **Streaming fragment assignment for real-time analysis of sequencing experiments.** *Nat Methods* 2013, **10**:71–3.

263. Mootha VK, Lindgren CM, Eriksson K-F, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstråle M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC: **PGC-1 $\alpha$ -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes.** *Nat Genet* 2003, **34**:267–73.

264. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M: **KEGG for integration and interpretation of large-scale molecular data sets.** *Nucleic Acids Res* 2012, **40**(Database issue):D109–14.

265. **Creating the gene ontology resource: design and implementation.** *Genome Res* 2001, **11**:1425–33.

266. Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, Schuler GD, Schriml LM, Tatusova TA, Wagner L, Rapp BA: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2001, **29**:11–6.

267. Edgar RC: **Search and clustering orders of magnitude faster than BLAST.** *Bioinformatics* 2010, **26**:2460–1.

268. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics* 2010, **26**:841–2.

269. Eddy SR: **Accelerated Profile HMM Searches.** *PLoS Comput Biol* 2011, **7**:e1002195.

270. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer ELL, Eddy SR, Bateman A, Finn RD: **The Pfam protein families database.** *Nucleic Acids Res* 2012, **40**(Database issue):D290–301.

271. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite.** *Trends Genet* 2000, **16**:276–7.

272. Zdobnov EM, Apweiler R: **InterProScan--an integration platform for the signature-recognition methods in InterPro.** *Bioinformatics* 2001, **17**:847–8.

273. Asare AL, Gao Z, Carey VJ, Wang R, Seyfert-Margolis V: **Power enhancement via multivariate outlier testing with gene expression arrays.** *Bioinformatics* 2009, **25**:48–53.

274. Huang DW, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nat Protoc* 2009, **4**:44–57.